Package 'MACSr'

September 19, 2024

Title MACS: Model-based Analysis for ChIP-Seq

Version 1.13.0

Description The Model-based Analysis of ChIP-Seq (MACS) is a widely used toolkit for identifying transcript factor binding sites. This package is an R wrapper of the lastest MACS3.

License BSD_3_clause + file LICENSE

Encoding UTF-8

LazyData true

Roxygen list(markdown = TRUE)

RoxygenNote 7.2.3

Depends R (>= 4.1.0)

Imports utils, reticulate, S4Vectors, methods, basilisk, ExperimentHub, AnnotationHub

Suggests testthat, knitr, rmarkdown, BiocStyle, MACSdata

PythonRequirements Python (>= 3.6.0), macs3

VignetteBuilder knitr

biocViews Software, ChIPSeq, ATACSeq, ImmunoOncology

StagedInstall no

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

git_branch devel

git_last_commit d2e5752

git_last_commit_date 2024-04-30

Repository Bioconductor 3.20

Date/Publication 2024-09-18

Author Qiang Hu [aut, cre]

Maintainer Qiang Hu <Qiang.Hu@roswellpark.org>

Contents

`bdgbroadcall` *bdgbroadcall*

Description

Call broad peaks from bedGraph output. Note: All regions on the same chromosome in the bedGraph file should be continuous so only bedGraph files from MACS3 are accpetable.

Usage

```
bdgbroadcall(
  ifile,
  cutoffpeak = 2,
  cutofflink = 1,
  minlen = 200L,
  lvl1maxgap = 30L,
  lvl2maxgap = 800L,
  trackline = TRUE,
  outdir = ".",
  outputfile = character(),
  log = TRUE,
  verbose = 2L
)
```

Arguments

`ifile`	MACS score in bedGraph. REQUIRED.
`cutoffpeak`	Cutoff for peaks depending on which method you used for score track. If the file contains qvalue scores from MACS3, score 2 means qvalue 0.01. DEFAULT: 2
`cutofflink`	Cutoff for linking regions/low abundance regions depending on which method you used for score track. If the file contains qvalue scores from MACS3, score 1 means qvalue 0.1, and score 0.3 means qvalue 0.5. DEFAULT: 1", default = 1
`minlen`	minimum length of peak, better to set it as d value. DEFAULT: 200", default = 200

lvl1maxgap	maximum gap between significant peaks, better to set it as tag size. DEFAULT: 30
lvl2maxgap	maximum linking between significant peaks, better to set it as 4 times of d value. DEFAULT: 800
trackline	Tells MACS not to include trackline with bedGraph files. The trackline is required by UCSC.
outdir	The output directory.
outputfile	The output file.
log	Whether to capture logs.
verbose	Set verbose level of runtime message. 0: only show critical message, 1: show additional warning message, 2: show process information, 3: show debug messages. DEFAULT:2

Value

macsList object.

Examples

```
eh <- ExperimentHub::ExperimentHub()
CHIP <- eh[["EH4558"]]
CTRL <- eh[["EH4563"]]
p1 <- pileup(CHIP, outdir = tempdir(),
             outputfile = "pileup_ChIP_bed.bdg", format = "BED")
p2 <- pileup(CTRL, outdir = tempdir(),
             outputfile = "pileup_CTRL_bed.bdg", format = "BED")
c1 <- bdgcmp(p1$outputs, p2$outputs, outdir = tempdir(),
             oprefix = "bdgcmp", pseudocount = 1, method = "FE")
bdgbroadcall(c1$outputs, cutoffpeak = 2, cutofflink = 1.5,
             outdir = tempdir(), outputfile = "bdgbroadcall")
```

---

bdgcmp *bdgcmp*

Description

Deduct noise by comparing two signal tracks in bedGraph. Note: All regions on the same chromosome in the bedGraph file should be continuous so only bedGraph files from MACS3 are accpetable.

Usage

```
bdgcmp(
  tfile,
  cfile,
  sfactor = 1,
  pseudocount = 0,
 method = c("ppois", "qpois", "subtract", "logFE", "FE", "logLR", "slogLR", "max"),
  oprefix = character(),
  outputfile = list(),
  outdir = ".",
  log = TRUE,
  verbose = 2L
)
```

Arguments

tfile	Treatment bedGraph file, e.g. *_treat_pileup.bdg from MACSv2. REQUIRED
cfile	Control bedGraph file, e.g. *_control_lambda.bdg from MACSv2. REQUIRED
sfactor	Scaling factor for treatment and control track. Keep it as 1.0 or default in most cases. Set it ONLY while you have SPMR output from MACS3 callpeak, and plan to calculate scores as MACS3 callpeak module. If you want to simulate 'callpeak' w/o '–to-large', calculate effective smaller sample size after filtering redudant reads in million (e.g., put 31.415926 if effective reads are 31,415,926) and input it for '-S'; for 'callpeak –to-large', calculate effective reads in larger sample. DEFAULT: 1.0
pseudocount	The pseudocount used for calculating logLR, logFE or FE. The count will be applied after normalization of sequencing depth. DEFAULT: 0.0, no pseudocount is applied.
method	Method to use while calculating a score in any bin by comparing treatment value and control value. Available choices are: ppois, qpois, subtract, logFE, logLR, and slogLR. They represent Poisson Pvalue (-log10(pvalue) form) using control as lambda and treatment as observation, q-value through a BH process for poisson pvalues, subtraction from treatment, linear scale fold enrichment, log10 fold enrichment(need to set pseudocount), log10 likelihood between ChIP-enriched model and open chromatin model(need to set pseudocount), symmetric log10 likelihood between two ChIP-enrichment models, or maximum value between the two tracks. Default option is ppois.",default="ppois".
oprefix	The PREFIX of output bedGraph file to write scores. If it is given as A, and method is 'ppois', output file will be A_ppois.bdg. Mutually exclusive with -o/–ofile.
outputfile	Output filename. Mutually exclusive with –o-prefix. The number and the order of arguments for –ofile must be the same as for -m.
outdir	The output directory.
log	Whether to capture logs.
verbose	Set verbose level. 0: only show critical message, 1: show additional warning message, 2: show process information, 3: show debug messages. If you want to know where are the duplicate reads, use 3. DEFAULT:2

Value

`macsList` object.

Examples

```
eh <- ExperimentHub::ExperimentHub()
CHIP <- eh[["EH4558"]]
CTRL <- eh[["EH4563"]]
p1 <- pileup(CHIP, outdir = tempdir(),
             outputfile = "pileup_ChIP_bed.bdg", format = "BED")
p2 <- pileup(CTRL, outdir = tempdir(),
             outputfile = "pileup_CTRL_bed.bdg", format = "BED")
c1 <- bdgcmp(p1$outputs, p2$outputs, outdir = tempdir(),
             oprefix = "bdgcmp", pseudocount = 1, method = "FE")
```

bdgdiff *bdgdiff*

Description

Differential peak detection based on paired four bedgraph files. Note: All regions on the same chromosome in the bedGraph file should be continuous so only bedGraph files from MACS3 are accpetable.

Usage

```
bdgdiff(
  t1bdg,
  t2bdg,
  c1bdg,
  c2bdg,
  cutoff = 3,
  minlen = 200L,
  maxgap = 100L,
  depth1 = 1,
  depth2 = 1,
  outdir = ".",
  oprefix = character(),
  outputfile = list(),
  log = TRUE,
  verbose = 2L
)
```

Arguments

t1bdg	MACS pileup bedGraph for condition 1. Incompatible with callpeak –SPMR output. REQUIRED
t2bdg	MACS pileup bedGraph for condition 2. Incompatible with callpeak –SPMR output. REQUIRED
c1bdg	MACS control lambda bedGraph for condition 1. Incompatible with callpeak –SPMR output. REQUIRED
c2bdg	MACS control lambda bedGraph for condition 2. Incompatible with callpeak –SPMR output. REQUIRED
cutoff	logLR cutoff. DEFAULT: 3 (likelihood ratio=1000)", default = 3
minlen	Minimum length of differential region. Try bigger value to remove small regions. DEFAULT: 200", default = 200
maxgap	Maximum gap to merge nearby differential regions. Consider a wider gap for broad marks. Maximum gap should be smaller than minimum length (-g). DEFAULT: 100", default = 100
depth1	Sequencing depth (# of non-redundant reads in million) for condition 1. It will be used together with –d2. See description for –d2 below for how to assign them. Default: 1

depth2	Sequencing depth (# of non-redundant reads in million) for condition 2. It will be used together with –d1. DEPTH1 and DEPTH2 will be used to calculate scaling factor for each sample, to down-scale larger sample to the level of smaller one. For example, while comparing 10 million condition 1 and 20 million condition 2, use –d1 10 –d2 20, then pileup value in bedGraph for condition 2 will be divided by 2. Default: 1
outdir	The output directory.
oprefix	Output file prefix. Actual files will be named as PREFIX_cond1.bed, PREFIX_cond2.bed and PREFIX_common.bed. Mutually exclusive with -o/–ofile.
outputfile	Output filenames. Must give three arguments in order: 1. file for unique regions in condition 1; 2. file for unique regions in condition 2; 3. file for common regions in both conditions. Note: mutually exclusive with –o-prefix.
log	Whether to capture logs.
verbose	Set verbose level of runtime message. 0: only show critical message, 1: show additional warning message, 2: show process information, 3: show debug messages. DEFAULT:2

Value

`macsList` object.

Examples

```
eh <- ExperimentHub::ExperimentHub()
CHIP <- eh[["EH4558"]]
CTRL <- eh[["EH4563"]]
c1 <- callpeak(CHIP, CTRL, gsize = 5.2e7, cutoff_analysis = TRUE,
               outdir = tempdir(), name = "callpeak_narrow0", store_bdg = TRUE)
c2 <- callpeak(CHIP, CTRL, gsize = 1e7, nomodel = TRUE, extsize = 250,
               outdir = tempdir(), name = "callpeak_narrow_revert", store_bdg = TRUE)
t1bdg <- grep("treat_pileup", c1$outputs, value = TRUE)
c1bdg <- grep("control_lambda", c1$outputs, value = TRUE)
t2bdg <- grep("treat_pileup", c2$outputs, value = TRUE)
c2bdg <- grep("control_lambda", c2$outputs, value = TRUE)
bdgdiff(t1bdg, t2bdg, c1bdg, c2bdg,
        outdir = tempdir(), oprefix = "bdgdiff")
```

`bdgopt` *bdgopt*

Description

Operations on score column of bedGraph file. Note: All regions on the same chromosome in the bedGraph file should be continuous so only bedGraph files from MACS3 are accpetable.

Usage

```
bdgopt(
  ifile,
  method = c("multiply", "add", "p2q", "max", "min"),
  extraparam = numeric(),
```

```
  outputfile = character(),
  outdir = ".",
  log = TRUE,
  verbose = 2L
)
```

Arguments

ifile	MACS score in bedGraph. Note: this must be a bedGraph file covering the ENTIRE genome. REQUIRED
method	Method to modify the score column of bedGraph file. Available choices are: multiply, add, max, min, or p2q. 1) multiply, the EXTRAPARAM is required and will be multiplied to the score column. If you intend to divide the score column by X, use value of 1/X as EXTRAPARAM. 2) add, the EXTRAPARAM is required and will be added to the score column. If you intend to subtract the score column by X, use value of -X as EXTRAPARAM. 3) max, the EXTRAPARAM is required and will take the maximum value between score and the EXTRAPARAM. 4) min, the EXTRAPARAM is required and will take the minimum value between score and the EXTRAPARAM. 5) p2q, this will convert p-value scores to q-value scores using Benjamini-Hochberg process. The EXTRAPARAM is not required. This method assumes the scores are -log10 p-value from MACS3. Any other types of score will cause unexpected errors.", default="p2q"
extraparam	The extra parameter for METHOD. Check the detail of -m option.
outputfile	Output filename. Mutually exclusive with –o-prefix. The number and the order of arguments for –ofile must be the same as for -m.
outdir	The output directory.
log	Whether to capture logs.
verbose	Set verbose level of runtime message. 0: only show critical message, 1: show additional warning message, 2: show process information, 3: show debug messages. DEFAULT:2

Value

`macsList` object.

Examples

```
eh <- ExperimentHub::ExperimentHub()
CHIP <- eh[["EH4558"]]
CTRL <- eh[["EH4563"]]
c1 <- callpeak(CHIP, CTRL, gsize = 5.2e7, cutoff_analysis = TRUE,
               outdir = tempdir(), name = "callpeak_narrow0",
               store_bdg = TRUE)
cfile <- grep("treat_pileup.bdg", c1$outputs, value = TRUE)
bdgopt(cfile, method = "min", extraparam = 10,
       outdir = tempdir(), outputfile = "bdgopt_min.bdg")
```

bdgpeakcall *bdgpeakcall*

Description

Call peaks from bedGraph output. Note: All regions on the same chromosome in the bedGraph file should be continuous so only bedGraph files from MACS3 are accpetable.

Usage

```
bdgpeakcall(
  ifile,
  cutoff = 5,
  minlen = 200L,
  maxgap = 30L,
  call_summits = FALSE,
  cutoff_analysis = FALSE,
  trackline = TRUE,
  outdir = ".",
  outputfile = character(),
  log = TRUE,
  verbose = 2L
)
```

Arguments

ifile	MACS score in bedGraph. REQUIRED.
cutoff	Cutoff depends on which method you used for score track. If the file contains pvalue scores from MACS3, score 5 means pvalue 1e-5. DEFAULT: 5", default = 5.
minlen	minimum length of peak, better to set it as d value. DEFAULT: 200", default = 200.
maxgap	maximum gap between significant points in a peak, better to set it as tag size. DEFAULT: 30", default = 30.
call_summits	If set, MACS will use a more sophisticated approach to find all summits in each enriched peak region DEFAULT: False",default=False.
cutoff_analysis	While set, bdgpeakcall will analyze number or total length of peaks that can be called by different cutoff then output a summary table to help user decide a better cutoff. Note, minlen and maxgap may affect the results. DEFAULT: False", default = False.
trackline	Tells MACS not to include trackline with bedGraph files. The trackline is required by UCSC.
outdir	The output directory.
outputfile	The output file.
log	Whether to capture logs.
verbose	Set verbose level of runtime message. 0: only show critical message, 1: show additional warning message, 2: show process information, 3: show debug messages. DEFAULT:2

Value

macsList object.

Examples

```
eh <- ExperimentHub::ExperimentHub()
CHIP <- eh[["EH4558"]]
CTRL <- eh[["EH4563"]]
p1 <- pileup(CHIP, outdir = tempdir(),
             outputfile = "pileup_ChIP_bed.bdg", format = "BED")
p2 <- pileup(CTRL, outdir = tempdir(),
             outputfile = "pileup_CTRL_bed.bdg", format = "BED")
c1 <- bdgcmp(p1$outputs, p2$outputs, outdir = tempdir(),
             oprefix = "bdgcmp", pseudocount = 1, method = "FE")
bdgpeakcall(c1$outputs, cutoff = 2,
            outdir = tempdir(), outputfile = "bdgpeakcall")
```

---

callpeak *callpeak*

Description

Main MACS3 Function to call peaks from alignment results.

Usage

```
callpeak(
  tfile,
  cfile = NULL,
  gsize = "hs",
  tsize = NULL,
  format = "AUTO",
  keepduplicates = "1",
  outdir = ".",
  name = "NA",
  store_bdg = FALSE,
  do_SPMR = FALSE,
  trackline = FALSE,
  nomodel = FALSE,
  shift = 0,
  extsize = 200,
  bw = 300,
  d_min = 20,
  mfold = c(5, 50),
  onauto = FALSE,
  qvalue = 0.05,
  pvalue = NULL,
  tempdir = "/tmp",
  nolambda = FALSE,
  scaleto = "small",
  downsample = FALSE,
```

```
  slocal = 1000,
  llocal = 10000,
  broad = FALSE,
  broadcutoff = 0.1,
  maxgap = NULL,
  minlen = NULL,
  cutoff_analysis = FALSE,
  fecutoff = 0.1,
  call_summits = FALSE,
  buffer_size = 1e+05,
  verbose = 2L,
  log = TRUE,
  ...
)
```

Arguments

tfile	ChIP-seq treatment files.
cfile	Control files.
gsize	Effective genome size. It can be 1.0e+9 or 1000000000, or shortcuts:'hs' for human (2.7e9), 'mm' for mouse (1.87e9), 'ce' for C. elegans (9e7) and 'dm' for fruitfly (1.2e8), Default:hs.
tsize	Tag size/read length. This will override the auto detected tag size. DEFAULT: Not set
format	Format of tag file, "AUTO", "BED" or "ELAND" or "ELANDMULTI" or "ELAND-EXPORT" or "SAM" or "BAM" or "BOWTIE" or "BAMPE" or "BEDPE".
keepduplicates	It controls the behavior towards duplicate tags at the exact same location – the same coordination and the same strand.
outdir	If specified all output files will be written to that directory.
name	Experiment name, which will be used to generate output file names.
store_bdg	Whether or not to save extended fragment pileup, and local lambda tracks (two files) at every bp into a bedGraph file.
do_SPMR	If True, MACS will SAVE signal per million reads for fragment pileup profiles.
trackline	Tells MACS to include trackline with bedGraph files.
nomodel	Whether or not to build the shifting model.
shift	The arbitrary shift in bp. Use discretion while setting it other than default value.
extsize	The arbitrary extension size in bp.
bw	Band width for picking regions to compute fragment size.
d_min	Minimum fragment size in basepair. Any predicted fragment size less than this will be excluded.
mfold	Select the regions within MFOLD range of high-confidence enrichment ratio against background to build model.
onauto	Whether turn on the auto pair model process.
qvalue	Minimum FDR (q-value) cutoff for peak detection.
pvalue	Pvalue cutoff for peak detection. DEFAULT: not set.
tempdir	Optional directory to store temp files.

nolambda	If True, MACS will use fixed background lambda as local lambda for every peak region.
scaleto	When set to 'small', scale the larger sample up to the smaller sample.
downsample	When set, random sampling method will scale down the bigger sample. By default, MACS uses linear scaling.
slocal	The small nearby region in basepairs to calculate dynamic lambda.
llocal	The large nearby region in basepairs to calculate dynamic lambda.
broad	If set, MACS will try to call broad peaks using the –broad-cutoff setting.
broadcutoff	Cutoff for broad region. This option is not available unless –broad is set.
maxgap	Maximum gap between significant sites to cluster them together. The DEFAULT value is the detected read length/tag size.
minlen	Minimum length of a peak. The DEFAULT value is the predicted fragment size d.
cutoff_analysis	While set, MACS2 will analyze number or total length of peaks that can be called by different p-value cutoff then output a summary table to help user decide a better cutoff.
fecutoff	When set, the value will be used to filter out peaks with low fold-enrichment.
call_summits	If set, MACS will use a more sophisticated signal processing approach to find subpeak summits in each enriched peak region.
buffer_size	Buffer size for incrementally increasing internal array size to store reads alignment information. DEFAULT: 100000.
verbose	Set verbose level of runtime message. 0: only show critical message, 1: show additional warning message, 2: show process information, 3: show debug messages. DEFAULT:2
log	Whether to capture logs.
...	More options for macs2.

Value

`macsList` object.

Examples

```
eh <- ExperimentHub::ExperimentHub()
CHIP <- eh[["EH4558"]]
CTRL <- eh[["EH4563"]]
res <- callpeak(CHIP, CTRL, gsize = 5.2e7,
                cutoff_analysis = TRUE,
                outdir = tempdir(),
                name = "callpeak_narrow0")
```

callvar *callvar*

Description

Call variants in given peak regions from the alignment BAM files.

Usage

```
callvar(
  peakbed,
  tfile,
  cfile,
  outputfile = character(),
  GQCutoffHetero = 0,
  GQCutoffHomo = 0,
  Q = 20,
  maxDuplicate = 1L,
  fermi = "auto",
  fermiMinOverlap = 30L,
  top2allelesMinRatio = 0.8,
  altalleleMinCount = 2L,
  maxAR = 0.95,
  np = 1L,
  verbose = 2L,
  log = TRUE
)
```

Arguments

peakbed	Peak regions in BED format, sorted by coordinates. REQUIRED.
tfile	ChIP-seq/ATAC-seq treatment file in BAM format, containing only records in peak regions, sorted by coordinates. Check instruction on how to make the file using samtools. REQUIRED.
cfile	Control file in BAM format, containing only records in peak regions, sorted by coordinates. Check instruction on how to make the file using samtools.
outputfile	Output VCF file name.
GQCutoffHetero	Genotype Quality score (-10log10((L00+L11)/(L01+L00+L11))) cutoff for Heterozygous allele type. Default:0, or there is no cutoff on GQ.
GQCutoffHomo	Genotype Quality score (-10log10((L00+L01)/(L01+L00+L11))) cutoff for Homozygous allele (not the same as reference) type. Default:0, or ther is no cutoff on GQ.
Q	Only consider bases with quality score greater than this value. Default: 20, which means Q20 or 0.01 error rate.
maxDuplicate	Maximum duplicated reads allowed per mapping position, mapping strand and the same CIGAR code. Default: 1. When sequencing depth is high, to set a higher value might help evaluate the correct allele ratio.

`fermi` Option to control when to apply local assembly through Fermi. By default (set as 'auto'), while SAPPER detects any INDEL variant in a peak region, it will utilize Fermi to recover the actual DNA sequences to refine the read alignments. If set as 'on', Fermi will be always invoked. It can increase specificity however sensivity and speed will be significantly lower. If set as 'off', Fermi won't be invoked at all. If so, speed and sensitivity can be higher but specificity will be significantly lower. Default: auto

`fermiMinOverlap`
The minimal overlap for fermi to initially assemble two reads. Must be between 1 and read length. A longer fermiMinOverlap is needed while read length is small (e.g. 30 for 36bp read, but 33 for 100bp read may work). Default:30

`top2allelesMinRatio`
The reads for the top 2 most frequent alleles (e.g. a ref allele and an alternative allele) at a loci shouldn't be too few comparing to total reads mapped. The minimum ratio is set by this optoin. Must be a float between 0.5 and 1. Default:0.8 which means at least 80%% of reads contain the top 2 alleles.

`altalleleMinCount`
The count of the alternative (non-reference) allele at a loci shouldn't be too few. By default, we require at least two reads support the alternative allele. Default:2

`maxAR` The maximum Allele-Ratio allowed while calculating likelihood for allele-specific binding. If we allow higher maxAR, we may mistakenly assign some homozygous loci as heterozygous. Default:0.95

`np` CPU used for mutliple processing. Please note that, assigning more CPUs does not guarantee the process being faster. Creating too many parrallel processes need memory operations and may negate benefit from multi processing. Default: 1

`verbose` Set verbose level of runtime message. 0: only show critical message, 1: show additional warning message, 2: show process information, 3: show debug messages. DEFAULT:2

`log` Whether to capture logs.

Value

`macsList` object.

Examples

```
## Not run:
callvar(
"PEsample_peaks_sorted.bed",
"PEsample_peaks_sorted.bam",
"PEcontrol_peaks_sorted.bam",
"/tmp/test.vcf")

## End(Not run)
```

cmbreps *cmbreps*

Description

Combine BEDGraphs of scores from replicates. Note: All regions on the same chromosome in the bedGraph file should be continuous so only bedGraph files from MACS3 are accpetable.

Usage

```
cmbreps(
  ifiles = list(),
  weights = 1,
  method = c("fisher", "max", "mean"),
  outputfile = character(),
  outdir = ".",
  log = TRUE,
  verbose = 2L
)
```

Arguments

ifiles	MACS score in bedGraph for each replicate. Require at least 2 files such as '-i A B C D'. REQUIRED
weights	Weight for each replicate. Default is 1.0 for each. When given, require same number of parameters as IFILE.
method	to use while combining scores from replicates. 1) fisher: Fisher's combined probability test. It requires scores in ppois form (-log10 pvalues) from bdgcmp. Other types of scores for this method may cause cmbreps unexpected errors. 2) max: take the maximum value from replicates for each genomic position. 3) mean: take the average value. Note, except for Fisher's method, max or mean will take scores AS IS which means they won't convert scores from log scale to linear scale or vice versa.", default="fisher"
outputfile	Output filename. Mutually exclusive with –o-prefix. The number and the order of arguments for –ofile must be the same as for -m.
outdir	The output directory.
log	Whether to capture logs.
verbose	Set verbose level of runtime message. 0: only show critical message, 1: show additional warning message, 2: show process information, 3: show debug messages. DEFAULT:2

Value

macsList object.

Examples

```
eh <- ExperimentHub::ExperimentHub()
CHIP <- eh[["EH4558"]]
CTRL <- eh[["EH4563"]]
c1 <- callpeak(CHIP, CTRL, gsize = 5.2e7, cutoff_analysis = TRUE,
               outdir = tempdir(), name = "callpeak_narrow0",
               store_bdg = TRUE)
cmbreps(ifiles = list(c1$outputs[1], c1$outputs[7]),
        method = "max", outdir = tempdir(), outputfile = "cmbreps")
```

`filterdup` *filterdup*

Description

filterdup

Usage

```
filterdup(
  ifile,
  gsize = "hs",
  format = "AUTO",
  tsize = NULL,
  pvalue = 1e-05,
  keepduplicates = "auto",
  outputfile = character(),
  outdir = ".",
  verbose = 2L,
  buffer_size = 10000,
  dryrun = FALSE,
  log = TRUE
)
```

Arguments

`ifile`	Input file(s).
`gsize`	Effective genome size. It can be 1.0e+9 or 1000000000, or shortcuts:'hs' for human (2.7e9), 'mm' for mouse (1.87e9), 'ce' for C. elegans (9e7) and 'dm' for fruitfly (1.2e8), Default:hs.
`format`	Input file format.
`tsize`	Tag size. This will override the auto detected tag size.
`pvalue`	Pvalue cutoff for binomial distribution test. DEFAULT:1e-5.
`keepduplicates`	It controls the behavior towards duplicate tags at the exact same location – the same coordination and the same strand. The 'auto' option makes MACS calculate the maximum tags at the exact same location based on binomal distribution using 1e-5 as pvalue cutoff; and the 'all' option keeps every tags. If an integer is given, at most this number of tags will be kept at the same location. Note, if you've used samtools or picard to flag reads as 'PCR/Optical duplicate' in bit 1024, MACS2 will still read them although the reads may be decided by MACS2

	as duplicate later. If you plan to rely on samtools/picard/any other tool to filter duplicates, please remove those duplicate reads and save a new alignment file then ask MACS2 to keep all by '–keep-dup all'. The default is to keep one tag at the same location. Default: 1".
outputfile	The output file.
outdir	The output directory.
verbose	Set verbose level of runtime message. 0: only show critical message, 1: show additional warning message, 2: show process information, 3: show debug messages. DEFAULT: 2.
buffer_size	Buffer size for incrementally increasing internal array size to store reads alignment information. In most cases, you don't have to change this parameter. However, if there are large number of chromosomes/contigs/scaffolds in your alignment, it's recommended to specify a smaller buffer size in order to decrease memory usage (but it will take longer time to read alignment files). Minimum memory requested for reading an alignment file is about # of CHROMOSOME * BUFFER_SIZE * 8 Bytes. DEFAULT: 100000.
dryrun	When set, filterdup will only output numbers instead of writing output files, including maximum allowable duplicates, total number of reads before filtering, total number of reads after filtering, and redundant rate. Default: not set.
log	Whether to capture logs.

Value

`macsList` object.

Examples

```
eh <- ExperimentHub::ExperimentHub()
CHIP <- eh[["EH4558"]]
res <- filterdup(ifile = CHIP, outputfile = "test.bed", outdir = tempdir())
```

`hmmratac` *hmmratac*

Description

Dedicated peak calling based on Hidden Markov Model for ATAC-seq data.

Usage

```
hmmratac(
  bam,
  outdir = ".",
  name = "NA",
  verbose = 2L,
  log = TRUE,
  cutoff_analysis_only = FALSE,
  em_skip = FALSE,
  em_means = list(50, 200, 400, 600),
  em_stddevs = list(20, 20, 20, 20),
```

```
  min_frag_p = 0.001,
  hmm_binsize = 10L,
  hmm_lower = 10L,
  hmm_upper = 20L,
  hmm_maxTrain = 1000,
  hmm_training_flanking = 1000,
  hmm_file = NULL,
  hmm_training_regions = NULL,
  hmm_randomSeed = 10151,
  hmm_modelonly = FALSE,
  prescan_cutoff = 1.2,
  openregion_minlen = 100,
  pileup_short = FALSE,
  keepduplicates = FALSE,
  blacklist = NULL,
  save_digested = FALSE,
  save_likelihoods = FALSE,
  save_states = FALSE,
  save_train = FALSE,
  decoding_steps = 1000,
  buffer_size = 1e+05,
  ...
)
```

Arguments

bam	Sorted BAM files containing the ATAC-seq reads. If multiple files are given as '-t A B C', then they will all be read and pooled together. REQUIRED.
outdir	If specified all output files will be written to that directory. Default: the current working directory
name	Name for this experiment, which will be used as a prefix to generate output file names. DEFAULT: "NA"
verbose	Set verbose level of runtime message. 0: only show critical message, 1: show additional warning message, 2: show process information, 3: show debug messages. DEFAULT:2
log	Whether to capture logs.
cutoff_analysis_only	Only run the cutoff analysis and output a report. After generating the report, the process will stop. The report will help user decide the three crucial parameters for `-l`, `-u`, and `-c`. So it's highly recommanded to run this first! Please read the report and instructions in `Choices of cutoff values` on how to decide the three crucial parameters.
em_skip	Do not perform EM training on the fragment distribution. If set, EM_MEANS and EM.STDDEVS will be used instead. Default: False
em_means	Comma separated list of initial mean values for the fragment distribution for short fragments, mono-, di-, and tri-nucleosomal fragments. Default: 50 200 400 600
em_stddevs	Comma separated list of initial standard deviation values for fragment distribution for short fragments, mono-, di-, and tri-nucleosomal fragments. Default: 20 20 20 20

`min_frag_p` We will exclude the abnormal fragments that can't be assigned to any of the four signal tracks. After we use EM to find the means and stddevs of the four distributions, we will calculate the likelihood that a given fragment length fit any of the four using normal distribution. The criteria we will use is that if a fragment length has less than MIN_FRAG_P probability to be like either of short, mono, di, or tri-nuc fragment, we will exclude it while generating the four signal tracks for later HMM training and prediction. The value should be between 0 and 1. Larger the value, more abnormal fragments will be allowed. So if you want to include more 'ideal' fragments, make this value smaller. Default = 0.001

`hmm_binsize` Size of the bins to split the pileup signals for training and decoding with Hidden Markov Model. Must >= 1. Smaller the binsize, higher the resolution of the results, slower the process. Default = 10

`hmm_lower` Upper limit on fold change range for choosing training sites. Default: 20

`hmm_upper` Lower limit on fold change range for choosing training sites. Default: 10

`hmm_maxTrain` Maximum number of training regions to use. Default: 1000

`hmm_training_flanking`
Training regions will be expanded to both side with this number of basepairs. The purpose is to include more background regions. Default: 1000

`hmm_file` A JSON file generated from previous HMMRATAC run to use instead of creating new one. When provided, HMM training will be skipped. Default: NA

`hmm_training_regions`
Filename of training regions (previously was BED_file) to use for training HMM, instead of using foldchange settings to select. Default: NA

`hmm_randomSeed` Seed to set for random sampling of training regions. Default: 10151

`hmm_modelonly` Stop the program after generating model. Use this option to generate HMM model ONLY, which can be later applied with `--model`. Default: False

`prescan_cutoff` The fold change cutoff for prescanning candidate regions in the whole dataset. Then we will use HMM to predict states on these candidate regions. Higher the prescan cutoff, fewer regions will be considered. Must > 1. Default: 1.2

`openregion_minlen`
Minimum length of open region to call accessible regions. Must be larger than 0. If it is set as 0, it means no filtering on the length of the open regions called. Please note that, when bin size is small, setting a too small OPENREGION_MINLEN will bring a lot of false positives. Default: 100

`pileup_short` By default, HMMRATAC will pileup all fragments in order to identify regions for training and candidate regions for decoding. When this option is on, it will pileup only the short fragments to do so. Although it sounds a good idea since we assume that open region should have a lot of short fragments, it may be possible that the overall short fragments are too few to be useful. Default: False

`keepduplicates` Keep duplicate reads from analysis. By default, duplicate reads will be removed. Default: False

`blacklist` Filename of blacklisted regions to exclude (previously was BED_file). Examples are those from ENCODE. Default: NA

`save_digested` Save the digested ATAC signals of short-, mono-, di-, and tri- signals in three BedGraph files with the names NAME_short.bdg, NAME_mono.bdg, NAME_di.bdg, and NAME_tri.bdg. DEFAULT: False

`save_likelihoods`
Save the likelihoods to each state annotation in three BedGraph files, named with NAME_open.bdg for open states, NAME_nuc.bdg for nucleosomal states, and NAME_bg.bdg for the background states. DEFAULT: False

save_states	Save all open and nucleosomal state annotations into a BED file with the name NAME_states.bed. DEFAULT: False
save_train	Save the training regions and training data into NAME_training_regions.bed and NAME_training_data.txt. Default: False
decoding_steps	Number of candidate regions to be decoded at a time. The HMM model will be applied with Viterbi to find the optimal state path in each region. bigger the number, 'possibly' faster the decoding process, 'definitely' larger the memory usage. Default: 1000.
buffer_size	Buffer size for incrementally increasing internal array size to store reads alignment information. In most cases, you don't have to change this parameter. However, if there are large number of chromosomes/contigs/scaffolds in your alignment, it's recommended to specify a smaller buffer size in order to decrease memory usage (but it will take longer time to read alignment files). Minimum memory requested for reading an alignment file is about # of CHROMOSOME * BUFFER_SIZE * 8 Bytes. DEFAULT: 100000
...	More options for macs2.

`macsList-class` *macsList*

Description

macsList

Arguments

arguments	The arguments used in the function.
outputs	The outputs from the function.
log	The run logs.

`MACSr` *MACSr*

Description

The Model-based Analysis of ChIP-Seq (MACS) is a widely used toolkit for identifying transcript factor binding sites. This package is an R wrapper of the lastest MACS3.

pileup *pileup*

Description

Pileup aligned reads with a given extension size (fragment size or d in MACS language). Note there will be no step for duplicate reads filtering or sequencing depth scaling, so you may need to do certain pre/post-processing.

Usage

```
pileup(
  ifile,
  outputfile = character(),
  outdir = ".",
 format = c("AUTO", "BAM", "SAM", "BED", "ELAND", "ELANDMULTI", "ELANDEXPORT", "BOWTIE",
    "BAMPE", "BEDPE"),
  bothdirection = FALSE,
  extsize = 200L,
  buffer_size = 100000L,
  verbose = 2L,
  log = TRUE
)
```

Arguments

ifile	Alignment file. If multiple files are given as '-t A B C', then they will all be read and combined. Note that pair-end data is not supposed to work with this command. REQUIRED.
outputfile	Output bedGraph file name. If not specified, will write to standard output. REQUIRED.
outdir	The output directory.
format	Format of tag file, \"AUTO\", \"BED\", \"ELAND\", \"ELANDMULTI\", \"ELANDEXPORT\", \"SAM\", \"BAM\", \"BOWTIE\", \"BAMPE\", or \"BEDPE\". The default AUTO option will let '%(prog)s' decide which format the file is. DEFAULT: \"AUTO\", MACS3 will pick a format from \"AUTO\", \"BED\", \"ELAND\", \"ELANDMULTI\", \"ELANDEXPORT\", \"SAM\", \"BAM\" and \"BOWTIE\". If the format is BAMPE or BEDPE, please specify it explicitly. Please note that when the format is BAMPE or BEDPE, the -B and –extsize options would be ignored.
bothdirection	By default, any read will be extended towards downstream direction by extension size. So it's [0,size-1] (1-based index system) for plus strand read and [-size+1,0] for minus strand read where position 0 is 5' end of the aligned read. Default behavior can simulate MACS3 way of piling up ChIP sample reads where extension size is set as fragment size/d. If this option is set as on, aligned reads will be extended in both upstream and downstream directions by extension size. It means [-size,size] where 0 is the 5' end of a aligned read. It can partially simulate MACS3 way of piling up control reads. However MACS3 local bias is calculated by maximizing the expected pileup over a ChIP fragment size/d estimated from 10kb, 1kb, d and whole genome background. This option will be ignored when the format is set as BAMPE or BEDPE. DEFAULT: False

extsize	The extension size in bps. Each alignment read will become a EXTSIZE of fragment, then be piled up. Check description for -B for detail. It's twice the shiftsize in old MACSv1 language. This option will be ignored when the format is set as BAMPE or BEDPE. DEFAULT: 200
buffer_size	Buffer size for incrementally increasing internal array size to store reads alignment information. In most cases, you don't have to change this parameter. However, if there are large number of chromosomes/contigs/scaffolds in your alignment, it's recommended to specify a smaller buffer size in order to decrease memory usage (but it will take longer time to read alignment files). Minimum memory requested for reading an alignment file is about # of CHROMOSOME * BUFFER_SIZE * 8 Bytes. DEFAULT: 100000
verbose	Set verbose level. 0: only show critical message, 1: show additional warning message, 2: show process information, 3: show debug messages. If you want to know where are the duplicate reads, use 3. DEFAULT:2
log	Whether to capture logs.

Value

macsList object.

Examples

```
eh <- ExperimentHub::ExperimentHub()
CHIP <- eh[["EH4558"]]
p <- pileup(CHIP, outdir = tempdir(), outputfile = "pileup_bed.bdg", format = "BED")
```

---

predictd *predictd*

Description

Predict d or fragment size from alignment results. In case of PE data, report the average insertion/fragment size from all pairs. *Will NOT filter duplicates*

Usage

```
predictd(
  ifile,
  gsize = "hs",
  format = "AUTO",
  plot = normalizePath(tempdir(), "predictd_mode.pdf"),
  tsize = NULL,
  bw = 300,
  d_min = 20,
  mfold = c(5, 50),
  buffer_size = 1e+05,
  verbose = 2L,
  log = TRUE
)
```

Arguments

ifile	Input file(s).
gsize	Effective genome size. It can be 1.0e+9 or 1000000000, or shortcuts:'hs' for human (2.7e9), 'mm' for mouse (1.87e9), 'ce' for C. elegans (9e7) and 'dm' for fruitfly (1.2e8), Default:hs.
format	Input file format.
plot	PDF path of peak model and correlation plots.
tsize	Tag size. This will override the auto detected tag size.
bw	Band width for picking regions to compute fragment size. This value is only used while building the shifting model. DEFAULT: 300
d_min	Minimum fragment size in basepair. Any predicted fragment size less than this will be excluded. DEFAULT: 20
mfold	Select the regions within MFOLD range of high-confidence enrichment ratio against background to build model. Fold-enrichment in regions must be lower than upper limit, and higher than the lower limit. Use as "-m 10 30". DEFAULT:5 50
buffer_size	Buffer size for incrementally increasing internal array size to store reads alignment information. DEFAULT: 100000.
verbose	Set verbose level of runtime message. 0: only show critical message, 1: show additional warning message, 2: show process information, 3: show debug messages. DEFAULT:2
log	Whether to capture log.

Value

predicted fragment sizes.

Examples

```
eh <- ExperimentHub::ExperimentHub()
CHIP <- eh[["EH4558"]]
predictd(CHIP, d_min = 10, gsize=5.2e+7, plot = NULL)
```

---

randsample *randsample*

Description

Randomly sample number/percentage of total reads.

Usage

```
randsample(
  ifile,
  outdir = ".",
  outputfile = character(),
  percentage = numeric(),
  number = numeric(),
```

```
  seed = -1L,
  tsize = NULL,
 format = c("AUTO", "BAM", "SAM", "BED", "ELAND", "ELANDMULTI", "ELANDEXPORT", "BOWTIE",
    "BAMPE", "BEDPE"),
  buffer_size = 100000L,
  verbose = 2L,
  log = TRUE
)
```

Arguments

ifile	Alignment file. If multiple files are given as '-t A B C', then they will all be read and combined. Note that pair-end data is not supposed to work with this command. REQUIRED.
outdir	The output directory.
outputfile	Output bedGraph file name. If not specified, will write to standard output. REQUIRED.
percentage	Percentage of tags you want to keep. Input 80.0 for 80%%. This option can't be used at the same time with -n/–num. REQUIRED
number	Number of tags you want to keep. Input 8000000 or 8e+6 for 8 million. This option can't be used at the same time with -p/–percent. Note that the number of tags in output is approximate as the number specified here. REQUIRED
seed	Set the random seed while down sampling data. Must be a non-negative integer in order to be effective. DEFAULT: not set
tsize	Tag size. This will override the auto detected tag size. DEFAULT: Not set
format	Format of tag file, \"AUTO\", \"BED\" or \"ELAND\" or \"ELANDMULTI\" or \"ELANDEXPORT\" or \"SAM\" or \"BAM\" or \"BOWTIE\" or \"BAMPE\" or \"BEDPE\". The default AUTO option will %(prog)s decide which format the file is. Please check the definition in README file if you choose ELAND/ELANDMULTI/ELANDEXPORT/SAM/BAM/BOWTIE or BAMPE/BEDPE. DEFAULT: \"AUTO\""
buffer_size	Buffer size for incrementally increasing internal array size to store reads alignment information. In most cases, you don't have to change this parameter. However, if there are large number of chromosomes/contigs/scaffolds in your alignment, it's recommended to specify a smaller buffer size in order to decrease memory usage (but it will take longer time to read alignment files). Minimum memory requested for reading an alignment file is about # of CHROMOSOME * BUFFER_SIZE * 8 Bytes. DEFAULT: 100000
verbose	Set verbose level. 0: only show critical message, 1: show additional warning message, 2: show process information, 3: show debug messages. If you want to know where are the duplicate reads, use 3. DEFAULT:2
log	Whether to capture logs.

Value

`macsList` object.

Examples

```
eh <- ExperimentHub::ExperimentHub()
CHIP <- eh[["EH4558"]]
randsample(CHIP, number = 1000, outdir = tempdir(), outputfile = "randsample.bed")
```

`refinepeak` *refinepeak*

Description

(Experimental) Take raw reads alignment, refine peak summits and give scores measuring balance of waston/crick tags. Inspired by SPP.

Usage

```
refinepeak(
  bedfile,
  ifile,
 format = c("AUTO", "BAM", "SAM", "BED", "ELAND", "ELANDMULTI", "ELANDEXPORT", "BOWTIE"),
  cutoff = 5,
  windowsize = 200L,
  buffer_size = 100000L,
  verbose = 2L,
  outdir = "./",
  outputfile = character(),
  log = TRUE
)
```

Arguments

`bedfile`	Candidate peak file in BED format. REQUIRED.
`ifile`	ChIP-seq alignment file. If multiple files are given as '-t A B C', then they will all be read and combined. Note that pair-end data is not supposed to work with this command. REQUIRED.
`format`	Format of tag file, \"AUTO\", \"BED\" or \"ELAND\" or \"ELANDMULTI\" or \"ELANDEXPORT\" or \"SAM\" or \"BAM\" or \"BOWTIE\". The default AUTO option will let '%(prog)s' decide which format the file is. Please check the definition in README file if you choose ELAND/ELANDMULTI/ELANDEXPORT/SAM/BAM DEFAULT: \"AUTO\""
`cutoff`	Cutoff DEFAULT: 5
`windowsize`	Scan window size on both side of the summit (default: 100bp)
`buffer_size`	Buffer size for incrementally increasing internal array size to store reads alignment information. In most cases, you don't have to change this parameter. However, if there are large number of chromosomes/contigs/scaffolds in your alignment, it's recommended to specify a smaller buffer size in order to decrease memory usage (but it will take longer time to read alignment files). Minimum memory requested for reading an alignment file is about # of CHROMOSOME * BUFFER_SIZE * 8 Bytes. DEFAULT: 100000

verbose	Set verbose level. 0: only show critical message, 1: show additional warning message, 2: show process information, 3: show debug messages. If you want to know where are the duplicate reads, use 3. DEFAULT:2
outdir	The output directory.
outputfile	Output bedGraph file name. If not specified, will write to standard output. REQUIRED.
log	Whether to capture logs.

Value

macsList object.

Examples

```
eh <- ExperimentHub::ExperimentHub()
CHIP <- eh[["EH4558"]]
CTRL <- eh[["EH4563"]]
res <- callpeak(CHIP, CTRL, gsize = 5.2e7, cutoff_analysis = TRUE,
                outdir = tempdir(), name = "callpeak_narrow0")
refinepeak(grep("narrowPeak", res$outputs, value = TRUE), CHIP,
           outdir = tempdir(), outputfile = "refine")
```

Index