

Package ‘AneuFinder’

October 17, 2017

Type Package

Title Analysis of Copy Number Variation in Single-Cell-Sequencing Data

Version 1.4.0

Date 2016-05

Author Aaron Taudt, Bjorn Bakker, David Porubsky

Maintainer Aaron Taudt <aaron.taudt@gmail.com>

Description This package implements functions for CNV calling, plotting, export and analysis from whole-genome single cell sequencing data.

Depends R (>= 3.3), GenomicRanges, cowplot, AneuFinderData

Imports methods, utils, grDevices, graphics, stats, foreach,
doParallel, BiocGenerics, S4Vectors, GenomeInfoDb, IRanges,
Rsamtools, bamsignals, DNACopy, Biostrings, GenomicAlignments,
ggplot2, reshape2, ggdendro, ggrepel, ReorderCluster, mclust

Suggests knitr, BiocStyle, testthat, BSgenome.Hsapiens.UCSC.hg19,
BSgenome.Mmusculus.UCSC.mm10

License Artistic-2.0

LazyLoad yes

VignetteBuilder knitr

biocViews Software, CopyNumberVariation, GenomicVariation,
HiddenMarkovModel, WholeGenome

URL <https://github.com/ataudt/aneufinder.git>

RoxygenNote 5.0.1

NeedsCompilation yes

R topics documented:

AneuFinder-package	3
aneuBiHMM	3
Aneufinder	4
aneuHMM	6
bam2GRanges	7
bed2GRanges	8
biDNACopy.findCNVs	9
binned.data	10

binning	10
binReads	10
bivariate.findCNVs	12
blacklist	14
clusterByQuality	15
collapseBins	16
colors	17
compareMethods	18
compareModels	19
consensusSegments	19
correctGC	20
correctMappability	21
deltaWCalculator	22
DNAcopy.findCNVs	22
estimateComplexity	23
export	23
filterSegments	25
findBreakpoints	25
findCNVs	26
findCNVs.strandseq	28
fixedWidthBins	30
getDistinctColors	31
getQC	32
getSCEcoordinates	33
getSegments	34
heatmapAneuploidies	34
heatmapGenomewide	35
heatmapGenomewideClusters	36
hotspotter	37
importBed	38
initializeStates	38
karyotypeMeasures	39
loadFromFiles	40
plot.aneuBiHMM	41
plot.aneuHMM	42
plot.character	42
plot.GRanges	43
plotHeterogeneity	43
plotHistogram	45
plotKaryogram	46
plotProfile	46
plot_pca	47
print.aneuBiHMM	48
print.aneuHMM	48
qualityControl	49
readConfig	50
simulateReads	50
subsetByCNVprofile	51
transCoord	52
univariate.findCNVs	52
variableWidthBins	54
writeConfig	55

Description

CNV detection in whole-genome single cell sequencing (WGSCS) and Strand-seq data using a Hidden Markov Model. The package implements CNV detection, commonly used plotting functions, export to BED format for upload to genome browsers, and measures for assessment of karyotype heterogeneity and quality metrics.

Details

The main function of this package is [Aneufinder](#) and produces several plots and browser files. If you want to have more fine-grained control over the different steps (binning, GC-correction, HMM, plotting) check the vignette [Introduction to AneuFinder](#).

Author(s)

Aaron Taudt, David Porubsky

Description

The aneuBiHMM object is output of the function [findCNVs.strandseq](#) and is basically a list with various entries. The class() attribute of this list was set to "aneuBiHMM". For a given hmm, the entries can be accessed with the list operators 'hmm[[]]' and 'hmm\$'.

Value

- ID An identifier that is used in various [AneuFinder](#) functions.
- bins A [GRanges](#) object containing the genomic bin coordinates, their read count and state classification.
- segments A [GRanges](#) object containing regions and their state classification.
- weights Weight for each component.
- transitionProbs Matrix of transition probabilities from each state (row) into each state (column).
- transitionProbs.initial Initial transitionProbs at the beginning of the Baum-Welch.
- startProbs Probabilities for the first bin
- startProbs.initial Initial startProbs at the beginning of the Baum-Welch.
- distributions Estimated parameters of the emission distributions.

`distributions.initial`
 Distribution parameters at the beginning of the Baum-Welch.

`convergenceInfo`
 Contains information about the convergence of the Baum-Welch algorithm.

`convergenceInfo$eps`
 Convergence threshold for the Baum-Welch.

`convergenceInfo$loglik`
 Final loglikelihood after the last iteration.

`convergenceInfo$loglik.delta`
 Change in loglikelihood after the last iteration (should be smaller than eps)

`convergenceInfo$num.iterations`
 Number of iterations that the Baum-Welch needed to converge to the desired eps.

`convergenceInfo$time.sec`
 Time in seconds that the Baum-Welch needed to converge to the desired eps.

See Also

`findCNVs.strandseq`

Aneufinder

Wrapper function for the [Aneufinder](#) package

Description

This function is an easy-to-use wrapper to [bin the data](#), [find copy-number-variations](#), [find sister-chromatid-exchange](#) events, plot [genomewide heatmaps](#), [distributions](#), [profiles](#) and [karyograms](#).

Usage

```
Aneufinder(inputfolder, outputfolder, configfile = NULL, numCPU = 1,
  reuse.existing.files = TRUE, binsizes = 1e+06,
  variable.width.reference = NULL, reads.per.bin = NULL,
  pairedEndReads = FALSE, assembly = NULL, chromosomes = NULL,
  remove.duplicate.reads = TRUE, min.mapq = 10, blacklist = NULL,
  use.bamsignals = FALSE, reads.store = FALSE, correction.method = NULL,
  GC.BSgenome = NULL, method = c("dnacopy", "HMM"), strandseq = FALSE,
  eps = 0.1, max.time = 60, max.iter = 5000, num.trials = 15,
  states = c("zero-inflation", paste0(0:10, "-somy")),
  most.frequent.state = "2-somy", most.frequent.state.strandseq = "1-somy",
  resolution = c(3, 6), min.segwidth = 2, bw = 4 * binsizes[1],
  pval = 1e-08, cluster.plots = TRUE)
```

Arguments

`inputfolder` Folder with either BAM or BED files.

`outputfolder` Folder to output the results. If it does not exist it will be created.

`configfile` A file specifying the parameters of this function (without `inputfolder`, `outputfolder` and `configfile`). Having the parameters in a file can be handy if many samples with the same parameter settings are to be run. If a `configfile` is specified, it will take priority over the command line parameters.

numCPU	The numbers of CPUs that are used. Should not be more than available on your machine.
reuse.existing.files	A logical indicating whether or not existing files in outputfolder should be reused.
binsizes	An integer vector with bin sizes. If more than one value is given, output files will be produced for each bin size.
variable.width.reference	A BAM file that is used as reference to produce variable width bins. See variableWidthBins for details.
reads.per.bin	Approximate number of desired reads per bin. The bin size will be selected accordingly. Output files are produced for each value.
pairedEndReads	Set to TRUE if you have paired-end reads in your BAM files (not implemented for BED files).
assembly	Please see fetchExtendedChromInfoFromUCSC for available assemblies. Only necessary when importing BED files. BAM files are handled automatically. Alternatively a data.frame with columns 'chromosome' and 'length'.
chromosomes	If only a subset of the chromosomes should be imported, specify them here.
remove.duplicate.reads	A logical indicating whether or not duplicate reads should be removed.
min.mapq	Minimum mapping quality when importing from BAM files. Set min.mapq=NULL to keep all reads.
blacklist	A GRanges or a bed(.gz) file with blacklisted regions. Reads falling into those regions will be discarded.
use.bamsignals	If TRUE the bamsignals package will be used for binning. This gives a tremendous performance increase for the binning step. reads.store and calc.complexity will be set to FALSE in this case.
reads.store	Set reads.store=TRUE to store read fragments as RData in folder 'data' and as BED files in 'BROWSERFILES/data'. This option will force use.bamsignals=FALSE.
correction.method	Correction methods to be used for the binned read counts. Currently only 'GC'.
GC.BSgenome	A BSgenome object which contains the DNA sequence that is used for the GC correction.
method	Any combination of c('HMM', 'dnacopy'). Option method='HMM' uses a Hidden Markov Model as described in doi:10.1186/s13059-016-0971-7 to call copy numbers. Option 'dnacopy' uses the DNAcopy package to call copy numbers similarly to the method proposed in doi:10.1038/nmeth.3578, which gives more robust but less sensitive results.
strandseq	A logical indicating whether the data comes from Strand-seq experiments. If TRUE, both strands carry information and are treated separately.
eps	Convergence threshold for the Baum-Welch algorithm.
max.time	The maximum running time in seconds for the Baum-Welch algorithm. If this time is reached, the Baum-Welch will terminate after the current iteration finishes. Set max.time = -1 for no limit.
max.iter	The maximum number of iterations for the Baum-Welch algorithm. Set max.iter = -1 for no limit.

<code>num.trials</code>	The number of trials to find a fit where <code>most.frequent.state</code> is most frequent. Each time, the HMM is seeded with different random initial values.
<code>states</code>	A subset or all of <code>c("zero-inflation", "0-somy", "1-somy", "2-somy", "3-somy", "4-somy", ...)</code> . This vector defines the states that are used in the Hidden Markov Model. The order of the entries must not be changed.
<code>most.frequent.state</code>	One of the states that were given in <code>states</code> . The specified state is assumed to be the most frequent one when running the univariate HMM. This can help the fitting procedure to converge into the correct fit. Default is '2-somy'.
<code>most.frequent.state.strandseq</code>	One of the states that were given in <code>states</code> . The specified state is assumed to be the most frequent one when option <code>strandseq=TRUE</code> . This can help the fitting procedure to converge into the correct fit. Default is '1-somy'.
<code>resolution</code>	An integer vector specifying the resolution at bin level at which to scan for SCE events.
<code>min.segwidth</code>	Segments below this width will be removed before scanning for SCE events.
<code>bw</code>	Bandwidth for SCE hotspot detection (see hotspotter for further details).
<code>pval</code>	P-value for SCE hotspot detection (see hotspotter for further details).
<code>cluster.plots</code>	A logical indicating whether plots should be clustered by similarity.

Value

NULL

Author(s)

Aaron Taudt

Examples

```
## Not run:
## The following call produces plots and genome browser files for all BAM files in "my-data-folder"
Aneufinder(inputfolder="my-data-folder", outputfolder="my-output-folder")
## End(Not run)
```

aneuHMM

*Hidden Markov Model***Description**

The `aneuHMM` object is output of the function [findCNVs](#) and is basically a list with various entries. The `class()` attribute of this list was set to "aneuHMM". For a given `hmm`, the entries can be accessed with the list operators `'hmm[[]]'` and `'hmm$'`.

Value

ID	An identifier that is used in various AneuFinder functions.
bins	A GRanges object containing the genomic bin coordinates, their read count and state classification.
segments	A GRanges object containing regions and their state classification.
weights	Weight for each component.
transitionProbs	Matrix of transition probabilities from each state (row) into each state (column).
transitionProbs.initial	Initial transitionProbs at the beginning of the Baum-Welch.
startProbs	Probabilities for the first bin
startProbs.initial	Initial startProbs at the beginning of the Baum-Welch.
distributions	Estimated parameters of the emission distributions.
distributions.initial	Distribution parameters at the beginning of the Baum-Welch.
convergenceInfo	Contains information about the convergence of the Baum-Welch algorithm.
convergenceInfo\$eps	Convergence threshold for the Baum-Welch.
convergenceInfo\$loglik	Final loglikelihood after the last iteration.
convergenceInfo\$loglik.delta	Change in loglikelihood after the last iteration (should be smaller than eps)
convergenceInfo\$num.iterations	Number of iterations that the Baum-Welch needed to converge to the desired eps.
convergenceInfo\$time.sec	Time in seconds that the Baum-Welch needed to converge to the desired eps.

See Also

[findCNVs](#)

bam2GRanges

Import BAM file into GRanges

Description

Import aligned reads from a BAM file into a [GRanges](#) object.

Usage

```
bam2GRanges(bamfile, bamindex = bamfile, chromosomes = NULL,
  pairedEndReads = FALSE, remove.duplicate.reads = FALSE, min.mapq = 10,
  max.fragment.width = 1000, blacklist = NULL, what = "mapq")
```

Arguments

bamfile	A sorted BAM file.
bamindex	BAM index file. Can be specified without the .bai ending. If the index file does not exist it will be created and a warning is issued.
chromosomes	If only a subset of the chromosomes should be imported, specify them here.
pairedEndReads	Set to TRUE if you have paired-end reads in your BAM files (not implemented for BED files).
remove.duplicate.reads	A logical indicating whether or not duplicate reads should be removed.
min.mapq	Minimum mapping quality when importing from BAM files. Set min.mapq=NULL to keep all reads.
max.fragment.width	Maximum allowed fragment length. This is to filter out erroneously wrong fragments due to mapping errors of paired end reads.
blacklist	A GRanges or a bed(.gz) file with blacklisted regions. Reads falling into those regions will be discarded.
what	A character vector of fields that are returned. Type scanBamWhat to see what is available.

Value

A [GRanges](#) object containing the reads.

Examples

```
## Get an example BAM file with single-cell-sequencing reads
bamfile <- system.file("extdata", "BB150803_IV_074.bam", package="AneuFinderData")
## Read the file into a GRanges object
reads <- bam2GRanges(bamfile, chromosomes=c(1:19,'X','Y'), pairedEndReads=FALSE,
                    min.mapq=10, remove.duplicate.reads=TRUE)
print(reads)
```

bed2GRanges	<i>Import BED file into GRanges</i>
-------------	-------------------------------------

Description

Import aligned reads from a BED file into a [GRanges](#) object.

Usage

```
bed2GRanges(bedfile, assembly, chromosomes = NULL,
            remove.duplicate.reads = FALSE, min.mapq = 10,
            max.fragment.width = 1000, blacklist = NULL)
```


Arguments

bedfile	A file with aligned reads in BED format. The columns have to be c('chromosome','start','end','description')
assembly	Please see fetchExtendedChromInfoFromUCSC for available assemblies. Only necessary when importing BED files. BAM files are handled automatically. Alternatively a data.frame with columns 'chromosome' and 'length'.
chromosomes	If only a subset of the chromosomes should be imported, specify them here.
remove.duplicate.reads	A logical indicating whether or not duplicate reads should be removed.
min.mapq	Minimum mapping quality when importing from BAM files. Set min.mapq=NULL to keep all reads.
max.fragment.width	Maximum allowed fragment length. This is to filter out erroneously wrong fragments.
blacklist	A GRanges or a bed(.gz) file with blacklisted regions. Reads falling into those regions will be discarded.

Value

A [GRanges](#) object containing the reads.

Examples

```
## Get an example BED file with single-cell-sequencing reads
bedfile <- system.file("extdata", "KK150311_VI_07.bam.bed.gz", package="AneuFinderData")
## Read the file into a GRanges object
reads <- bed2GRanges(bedfile, assembly='mm10', chromosomes=c(1:19,'X','Y'),
                    min.mapq=10, remove.duplicate.reads=TRUE)
print(reads)
```

biDNACopy.findCNVs *Find copy number variations (DNACopy, bivariate)*

Description

biDNACopy.findCNVs classifies the binned read counts into several states which represent copy-number-variation using read count information from both strands.

Usage

```
biDNACopy.findCNVs(binned.data, ID = NULL, CNgrid.start = 0.5,
                  count.cutoff.quantile = 0.999)
```

Arguments

binned.data	A GRanges object with binned read counts.
ID	An identifier that will be used to identify this sample in various downstream functions. Could be the file name of the binned.data for example.
CNgrid.start	Start parameter for the CNgrid variable. Very empiric. Set to 1.5 for normal data and 0.5 for Strand-seq data.

`count.cutoff.quantile`

A quantile between 0 and 1. Should be near 1. Read counts above this quantile will be set to the read count specified by this quantile. Filtering very high read counts increases the performance of the Baum-Welch fitting procedure. However, if your data contains very few peaks they might be filtered out. Set `count.cutoff.quantile=1` in this case.

Value

An [aneuHMM](#) object.

<code>binned.data</code>	<i>Binned read counts</i>
--------------------------	---------------------------

Description

A [GRanges](#) object which contains binned read counts as meta data column reads. It is output of the various [binning](#) functions.

<code>binning</code>	<i>Bin the genome</i>
----------------------	-----------------------

Description

Please see functions [fixedWidthBins](#) and [variableWidthBins](#) for further details.

<code>binReads</code>	<i>Convert aligned reads from various file formats into read counts in equidistant bins</i>
-----------------------	---------------------------------------------------------------------------------------------

Description

Convert aligned reads in `.bam` or `.bed(.gz)` format into read counts in equidistant windows.

Usage

```
binReads(file, assembly, ID = basename(file), bamindex = file,
  chromosomes = NULL, pairedEndReads = FALSE, min.mapq = 10,
  remove.duplicate.reads = TRUE, max.fragment.width = 1000,
  blacklist = NULL, outputfolder.binned = "binned_data", binsizes = 1e+06,
  reads.per.bin = NULL, bins = NULL, variable.width.reference = NULL,
  save.as.RData = FALSE, calc.complexity = TRUE, call = match.call(),
  reads.store = FALSE, outputfolder.reads = "data", reads.return = FALSE,
  reads.override = FALSE, reads.only = FALSE, use.bamsignals = FALSE)
```

Arguments

file	A file with aligned reads. Alternatively a GRanges with aligned reads.
assembly	Please see fetchExtendedChromInfoFromUCSC for available assemblies. Only necessary when importing BED files. BAM files are handled automatically. Alternatively a data.frame with columns 'chromosome' and 'length'.
ID	An identifier that will be used to identify the file throughout the workflow and in plotting.
bamindex	BAM index file. Can be specified without the .bai ending. If the index file does not exist it will be created and a warning is issued.
chromosomes	If only a subset of the chromosomes should be binned, specify them here.
pairedEndReads	Set to TRUE if you have paired-end reads in your BAM files (not implemented for BED files).
min.mapq	Minimum mapping quality when importing from BAM files. Set min.mapq=NULL to keep all reads.
remove.duplicate.reads	A logical indicating whether or not duplicate reads should be removed.
max.fragment.width	Maximum allowed fragment length. This is to filter out erroneously wrong fragments due to mapping errors of paired end reads.
blacklist	A GRanges or a bed(.gz) file with blacklisted regions. Reads falling into those regions will be discarded.
outputfolder.binned	Folder to which the binned data will be saved. If the specified folder does not exist, it will be created.
binsizes	An integer vector with bin sizes. If more than one value is given, output files will be produced for each bin size.
reads.per.bin	Approximate number of desired reads per bin. The bin size will be selected accordingly. Output files are produced for each value.
bins	A named list with GRanges containing precalculated bins produced by fixedWidthBins or variableWidthBins . Names must correspond to the binsize.
variable.width.reference	A BAM file that is used as reference to produce variable width bins. See variableWidthBins for details.
save.as.RData	If set to FALSE, no output file will be written. Instead, a GenomicRanges object containing the binned data will be returned. Only the first binsize will be processed in this case.
calc.complexity	A logical indicating whether or not to estimate library complexity.
call	The <code>match.call()</code> of the parent function.
reads.store	If TRUE processed read fragments will be saved to file. Reads are processed according to min.mapq and remove.duplicate.reads. Paired end reads are coerced to single end fragments. Will be ignored if use.bamsignals=TRUE.
outputfolder.reads	Folder to which the read fragments will be saved. If the specified folder does not exist, it will be created.
reads.return	If TRUE no binning is done and instead, read fragments from the input file are returned in GRanges format.

<code>reads.overwrite</code>	Whether or not an existing file with read fragments should be overwritten.
<code>reads.only</code>	If TRUE only read fragments are stored and/or returned and no binning is done.
<code>use.bamsignals</code>	If TRUE the bamsignals package will be used for binning. This gives a tremendous performance increase for the binning step. <code>reads.store</code> and <code>calc.complexity</code> will be set to FALSE in this case.

Details

Convert aligned reads from `.bam` or `.bed(.gz)` files into read counts in equidistant windows (bins). This function uses [countOverlaps](#) to calculate the read counts.

Value

The function produces a `list()` of [GRanges](#) objects with one meta data column 'reads' that contains the read count. This binned data will be either written to file (`save.as.RData=FALSE`) or given as return value (`save.as.RData=FALSE`).

See Also

`binning`

Examples

```
## Get an example BED file with single-cell-sequencing reads
bedfile <- system.file("extdata", "KK150311_VI_07.bam.bed.gz", package="AneuFinderData")
## Bin the BED file into bin size 1Mb
binned <- binReads(bedfile, assembly='mm10', binsize=1e6,
                  chromosomes=c(1:19,'X','Y'))
print(binned)
```

`bivariate.findCNVs` *Find copy number variations (bivariate)*

Description

`bivariate.findCNVs` finds CNVs using read count information from both strands.

Usage

```
bivariate.findCNVs(binned.data, ID = NULL, eps = 0.1, init = "standard",
                  max.time = -1, max.iter = -1, num.trials = 1, eps.try = NULL,
                  num.threads = 1, count.cutoff.quantile = 0.999,
                  states = c("zero-inflation", paste0(0:10, "-somy")),
                  most.frequent.state = "1-somy", algorithm = "EM", initial.params = NULL)
```

Arguments

<code>binned.data</code>	A GRanges object with binned read counts.
<code>ID</code>	An identifier that will be used to identify this sample in various downstream functions. Could be the file name of the <code>binned.data</code> for example.
<code>eps</code>	Convergence threshold for the Baum-Welch algorithm.
<code>init</code>	One of the following initialization procedures: <code>standard</code> The negative binomial of state '2-somy' will be initialized with <code>mean=mean(counts)</code> , <code>var=var(counts)</code> . This procedure usually gives good convergence. <code>random</code> Mean and variance of the negative binomial of state '2-somy' will be initialized with random values (in certain boundaries, see source code). Try this if the <code>standard</code> procedure fails to produce a good fit.
<code>max.time</code>	The maximum running time in seconds for the Baum-Welch algorithm. If this time is reached, the Baum-Welch will terminate after the current iteration finishes. Set <code>max.time = -1</code> for no limit.
<code>max.iter</code>	The maximum number of iterations for the Baum-Welch algorithm. Set <code>max.iter = -1</code> for no limit.
<code>num.trials</code>	The number of trials to find a fit where <code>state most.frequent.state</code> is most frequent. Each time, the HMM is seeded with different random initial values.
<code>eps.try</code>	If code <code>num.trials</code> is set to greater than 1, <code>eps.try</code> is used for the trial runs. If unset, <code>eps</code> is used.
<code>num.threads</code>	Number of threads to use. Setting this to >1 may give increased performance.
<code>count.cutoff.quantile</code>	A quantile between 0 and 1. Should be near 1. Read counts above this quantile will be set to the read count specified by this quantile. Filtering very high read counts increases the performance of the Baum-Welch fitting procedure. However, if your data contains very few peaks they might be filtered out. Set <code>count.cutoff.quantile=1</code> in this case.
<code>states</code>	A subset or all of <code>c("zero-inflation", "0-somy", "1-somy", "2-somy", "3-somy", "4-somy", ...)</code> . This vector defines the states that are used in the Hidden Markov Model. The order of the entries must not be changed.
<code>most.frequent.state</code>	One of the states that were given in <code>states</code> . The specified state is assumed to be the most frequent one. This can help the fitting procedure to converge into the correct fit.
<code>algorithm</code>	One of <code>c('baumWelch', 'EM')</code> . The expectation maximization ('EM') will find the most likely states and fit the best parameters to the data, the 'baumWelch' will find the most likely states using the initial parameters.
<code>initial.params</code>	A aneuHMM object or file containing such an object from which initial starting parameters will be extracted.

Value

An [aneuBiHMM](#) object.

blacklist	<i>Make a blacklist for genomic regions</i>
-----------	---------------------------------------------

Description

Produce a blacklist of genomic regions with a high ratio of duplicate to unique reads. This blacklist can be used to exclude reads for analysis in [Aneufinder](#), [bam2GRanges](#) and [bed2GRanges](#). This function produces a pre-blacklist which has to manually be filtered with a sensible cutoff. See the examples section for details.

Usage

```
blacklist(files, assembly, bins, min.mapq = 10, pairedEndReads = FALSE)
```

Arguments

files	A character vector of either BAM or BED files.
assembly	Please see fetchExtendedChromInfoFromUCSC for available assemblies. Only necessary when importing BED files. BAM files are handled automatically. Alternatively a data.frame with columns 'chromosome' and 'length'.
bins	A list with one GRanges with binned read counts generated by fixedWidthBins .
min.mapq	Minimum mapping quality when importing from BAM files. Set min.mapq=NULL to keep all reads.
pairedEndReads	Set to TRUE if you have paired-end reads in your BAM files (not implemented for BED files).

Value

A [GRanges](#) with the same coordinates as bins with metadata columns ratio, duplicated counts and deduplicated counts.

Examples

```
## Get an example BAM file with single-cell-sequencing reads
bamfile <- system.file("extdata", "BB150803_IV_074.bam", package="AneuFinderData")
## Prepare the blacklist
bins <- fixedWidthBins(assembly='mm10', binsizes=1e6, chromosome.format='NCBI')
pre.blacklist <- blacklist(bamfile, bins=bins)
## Plot a histogram to decide on a sensible cutoff
qplot(pre.blacklist$ratio, binwidth=0.1)
## Make the blacklist with cutoff = 1.9
blacklist <- pre.blacklist[pre.blacklist$ratio > 1.9]
```



```
print(cl$parameters)
## Select files from the best 2 clusters for further processing
best.files <- unlist(cl$classification[1:2])
```

collapseBins	<i>Collapse consecutive bins</i>
--------------	----------------------------------

Description

The function will collapse consecutive bins which have, for example, the same combinatorial state.

Usage

```
collapseBins(data, column2collapseBy = NULL, columns2sumUp = NULL,
             columns2average = NULL, columns2getMax = NULL, columns2drop = NULL)
```

Arguments

data	A data.frame containing the genomic coordinates in the first three columns.
column2collapseBy	The number of the column which will be used to collapse all other inputs. If a set of consecutive bins has the same value in this column, they will be aggregated into one bin with adjusted genomic coordinates. If NULL directly adjacent bins will be collapsed.
columns2sumUp	Column numbers that will be summed during the aggregation process.
columns2average	Column numbers that will be averaged during the aggregation process.
columns2getMax	Column numbers where the maximum will be chosen during the aggregation process.
columns2drop	Column numbers that will be dropped after the aggregation process.

Details

The following tables illustrate the principle of the collapsing:

Input data:

seqnames	start	end	column2collapseBy	moreColumns	columns2sumUp
chr1	0	199	2	1 10	1 3
chr1	200	399	2	2 11	0 3
chr1	400	599	2	3 12	1 3
chr1	600	799	1	4 13	0 3
chr1	800	999	1	5 14	1 3

Output data:

seqnames	start	end	column2collapseBy	moreColumns	columns2sumUp
chr1	0	599	2	1 10	2 9
chr1	600	999	1	4 13	1 6

Value

A data.frame.

Author(s)

Aaron Taudt

Examples

```
## Get an example BED file with single-cell-sequencing reads
bedfile <- system.file("extdata", "KK150311_VI_07.bam.bed.gz", package="AneuFinderData")
## Bin the BAM file into bin size 1Mp
binned <- binReads(bedfile, assembly='mm10', binsize=1e6,
                  chromosomes=c(1:19,'X','Y'))
## Collapse the bins by chromosome and get average, summed and maximum read count
df <- as.data.frame(binned[[1]])
# Remove one bin for illustration purposes
df <- df[-3,]
head(df)
collapseBins(df, column2collapseBy='seqnames', columns2sumUp=c('width','counts'),
             columns2average='counts', columns2getMax='counts',
             columns2drop=c('mcounts','pcounts'))
collapseBins(df, column2collapseBy=NULL, columns2sumUp=c('width','counts'),
             columns2average='counts', columns2getMax='counts',
             columns2drop=c('mcounts','pcounts'))
```

colors

AneuFinder *color scheme*

Description

Get the color schemes that are used in the AneuFinder plots.

Usage

```
stateColors(states = c("zero-inflation", paste0(0:10, "-somy"), "total"))
```

```
strandColors(strands = c("+", "-"))
```

Arguments

states A character vector with states whose color should be returned.

strands A character vector with strands whose color should be returned. Any combination of c('+', '-', '*').

Value

A character vector with colors.

Functions

- stateColors: Colors that are used for the states.
- strandColors: Colors that are used to distinguish strands.

Examples

```
## Make a nice pie chart with the AneuFinder state color scheme
statecolors <- stateColors()
pie(rep(1,length(statecolors)), labels=names(statecolors), col=statecolors)

## Make a nice pie chart with the AneuFinder strand color scheme
strandcolors <- strandColors()
pie(rep(1,length(strandcolors)), labels=names(strandcolors), col=strandcolors)
```

compareMethods

Compare copy number calling methods

Description

Compare two sets of [aneuHMM](#) objects generated by different methods (see option method of [findCNVs](#)).

Usage

```
compareMethods(models1, models2)
```

Arguments

models1	A list of aneuHMM objects or a character vector with files that contain such objects.
models2	A list of aneuHMM objects or a character vector with files that contain such objects. IDs of the models must match the ones in models1.

Value

A data.frame with one column 'concordance' which gives the fraction of the genome that is called concordantly between both models.

Author(s)

Aaron Taudt

Examples

```
## Get a list of HMMs
folder <- system.file("extdata", "primary-lung", "hmm", package="AneuFinderData")
files <- list.files(folder, full.names=TRUE)
## Compare the models with themselves (non-sensical)
df <- compareMethods(files, files)
head(df)
```

compareModels	<i>Compare copy number models</i>
---------------	-----------------------------------

Description

Compare two [aneuHMM](#) objects. The function computes the fraction of copy number calls that is concordant between both models.

Usage

```
compareModels(model1, model2)
```

Arguments

model1	An aneuHMM object or file that contains such an object.
model2	An aneuHMM object or file that contains such an object.

Value

A numeric.

Author(s)

Aaron Taudt

consensusSegments	<i>Make consensus segments</i>
-------------------	--------------------------------

Description

Make consensus segments from a list of [aneuHMM](#) or [aneuBiHMM](#) objects.

Usage

```
consensusSegments(hmms)
```

Arguments

hmms	A list of aneuHMM or aneuBiHMM objects or a character vector of files that contains such objects.
------	-----------------------------------------------------------------------------------------------------------------------------------

Details

The function will produce a [GRanges](#) object using the [disjoin](#) function on all extracted `$segment` entries.

Value

A [GRanges](#).

Examples

```
## Get results from a small-cell-lung-cancer
lung.folder <- system.file("extdata", "primary-lung", "hmms", package="AneuFinderData")
lung.files <- list.files(lung.folder, full.names=TRUE)
## Get consensus segments and states
consensusSegments(lung.files)
```

correctGC

GC correction

Description

Correct a list of [binned.data](#) by GC content.

Usage

```
correctGC(binned.data.list, GC.BSgenome, same.binsize = FALSE)
```

Arguments

binned.data.list	A list with binned.data objects or a list of filenames containing such objects.
GC.BSgenome	A BSgenome object which contains the DNA sequence that is used for the GC correction.
same.binsize	If TRUE the GC content will only be calculated once. Set this to TRUE if all binned.data objects describe the same genome at the same binsize.

Value

A list with [binned.data](#) objects with adjusted read counts.

Author(s)

Aaron Taudt

Examples

```
## Get a BED file, bin it and run GC correction
bedfile <- system.file("extdata", "KK150311_VI_07.bam.bed.gz", package="AneuFinderData")
binned <- binReads(bedfile, assembly='mm10', binsize=1e6,
                  chromosomes=c(1:19, 'X', 'Y'))
plot(binned[[1]], type=1)
if (require(BSgenome.Mmusculus.UCSC.mm10)) {
  binned.GC <- correctGC(list(binned[[1]]), GC.BSgenome=BSgenome.Mmusculus.UCSC.mm10)
  plot(binned.GC[[1]], type=1)
}
```

correctMappability	<i>Mappability correction</i>
--------------------	-------------------------------

Description

Correct a list of [binned.data](#) by mappability.

Usage

```
correctMappability(binned.data.list, same.binsize, reference, assembly,  
  pairedEndReads = FALSE, min.mapq = 10, remove.duplicate.reads = TRUE,  
  max.fragment.width = 1000)
```

Arguments

<code>binned.data.list</code>	A list with binned.data objects or a list of filenames containing such objects.
<code>same.binsize</code>	If TRUE the mappability correction will only be calculated once. Set this to TRUE if all binned.data objects describe the same genome at the same binsize.
<code>reference</code>	A file or GRanges with aligned reads.
<code>assembly</code>	Please see fetchExtendedChromInfoFromUCSC for available assemblies. Only necessary when importing BED files. BAM files are handled automatically. Alternatively a data.frame with columns 'chromosome' and 'length'.
<code>pairedEndReads</code>	Set to TRUE if you have paired-end reads in your BAM files (not implemented for BED files).
<code>min.mapq</code>	Minimum mapping quality when importing from BAM files. Set <code>min.mapq=NULL</code> to keep all reads.
<code>remove.duplicate.reads</code>	A logical indicating whether or not duplicate reads should be removed.
<code>max.fragment.width</code>	Maximum allowed fragment length. This is to filter out erroneously wrong fragments due to mapping errors of paired end reads.

Value

A list with [binned.data](#) objects with adjusted read counts.

Author(s)

Aaron Taudt

deltaWCalculator	<i>Calculate deltaWs</i>
------------------	--------------------------

Description

This function will calculate deltaWs from a [GRanges](#) object with read fragments.

Usage

```
deltaWCalculator(frags, reads.per.window = 10)
```

Arguments

frags	A GRanges with read fragments (see bam2GRanges).
reads.per.window	Number of reads in each dynamic window.

Value

The input frags with additional meta-data columns.

Author(s)

Aaron Taudt, David Porubsky, Ashley Sanders

DNAcopy.findCNVs	<i>Find copy number variations (DNAcopy, univariate)</i>
------------------	----------------------------------------------------------

Description

DNAcopy.findCNVs classifies the binned read counts into several states which represent copy-number-variation.

Usage

```
DNAcopy.findCNVs(binned.data, ID = NULL, CNgrid.start = 1.5,
count.cutoff.quantile = 0.999, strand = "*")
```

Arguments

binned.data	A GRanges object with binned read counts.
ID	An identifier that will be used to identify this sample in various downstream functions. Could be the file name of the binned.data for example.
CNgrid.start	Start parameter for the CNgrid variable. Very empiric. Set to 1.5 for normal data and 0.5 for Strand-seq data.
count.cutoff.quantile	A quantile between 0 and 1. Should be near 1. Read counts above this quantile will be set to the read count specified by this quantile. Filtering very high read counts increases the performance of the Baum-Welch fitting procedure. However, if your data contains very few peaks they might be filtered out. Set count.cutoff.quantile=1 in this case.
strand	Run the HMM only for the specified strand. One of c('+', '-', '*').

Value

An [aneuHMM](#) object.

estimateComplexity	<i>Estimate library complexity</i>
--------------------	------------------------------------

Description

Estimate library complexity using a very simple "Michaelis-Menten" approach.

Usage

```
estimateComplexity(reads)
```

Arguments

reads	A GRanges object with read fragments. NOTE: Complexity estimation relies on duplicate reads and therefore the duplicates have to be present in the input.
-------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Value

A list with estimated complexity values and plots.

export	<i>Export genome browser viewable files</i>
--------	---------------------------------------------

Description

Export copy-number-variation state or read counts as genome browser viewable file

Usage

```
exportCNVs(hmms, filename, cluster = TRUE, export.CNV = TRUE,
           export.SCE = TRUE)
```

```
exportReadCounts(hmms, filename)
```

```
exportGRanges(gr, filename, header = TRUE, trackname = NULL, score = NULL,
              priority = NULL, append = FALSE, chromosome.format = "UCSC")
```

Arguments

hmms	A list of aneuHMM objects or a character vector with files that contain such objects.
filename	The name of the file that will be written. The appropriate ending will be appended, either ".bed.gz" for CNV-state or ".wig.gz" for read counts. Any existing file will be overwritten.
cluster	If TRUE, the samples will be clustered by similarity in their CNV-state.
export.CNV	A logical, indicating whether the CNV-state shall be exported.

<code>export.SCE</code>	A logical, indicating whether breakpoints shall be exported.
<code>gr</code>	A GRanges object.
<code>header</code>	A logical indicating whether the output file will have a heading track line (TRUE) or not (FALSE).
<code>trackname</code>	The name that will be used as track name and description in the header.
<code>score</code>	A vector of the same length as <code>gr</code> , which will be used for the 'score' column in the BED file.
<code>priority</code>	Priority of the track for display in the genome browser.
<code>append</code>	Append to filename.
<code>chromosome.format</code>	A character specifying the format of the chromosomes if assembly is specified. Either 'NCBI' for (1,2,3 ...) or 'UCSC' for (chr1,chr2,chr3 ...).# @importFrom utils write.table

Details

Use `exportCNVs` to export the copy-number-variation state from an [aneuHMM](#) object in BED format. Use `exportReadCounts` to export the binned read counts from an [aneuHMM](#) object in WIGGLE format. Use `exportGRanges` to export a [GRanges](#) object in BED format.

Value

NULL

Functions

- `exportCNVs`: Export CNV-state as `.bed.gz` file
- `exportReadCounts`: Export binned read counts as `.wig.gz` file
- `exportGRanges`: Export [GRanges](#) object as BED file.

Author(s)

Aaron Taudt

Examples

```
## Not run:
## Get results from a small-cell-lung-cancer
folder <- system.file("extdata", "primary-lung", "hmms", package="AneuFinderData")
files <- list.files(folder, full.names=TRUE)
## Export the CNV states for upload to the UCSC genome browser
exportCNVs(files, filename='upload-me-to-a-genome-browser', cluster=TRUE)
## End(Not run)
```

filterSegments	<i>Filter segments by minimal size</i>
----------------	----------------------------------------

Description

filterSegments filters out segments below a specified minimal segment size. This can be useful to get rid of boundary effects from the Hidden Markov approach.

Usage

```
filterSegments(segments, min.seg.width)
```

Arguments

segments A [GRanges](#) object.
min.seg.width The minimum segment width in base-pairs.

Value

The input model with adjusted segments.

Author(s)

Aaron Taudt

Examples

```
## Load an HMM
file <- list.files(system.file("extdata", "primary-lung", "hmms",
                             package="AneuFinderData"), full.names=TRUE)
hmm <- loadFromFiles(file)[[1]]
## Check number of segments before and after filtering
length(hmm$segments)
hmm$segments <- filterSegments(hmm$segments, min.seg.width=2*width(hmm$bins)[1])
length(hmm$segments)
```

findBreakpoints	<i>Find breakpoints</i>
-----------------	-------------------------

Description

Breakpoint detection is done via a dynamic windowing approach on read resolution.

Usage

```
findBreakpoints(model, fragments, breakpoint.quantile = 0.99)
```

Arguments

model	An aneuBiHMM object or a file that contains such an object.
fragments	A GRanges object with read fragments.
breakpoint.quantile	A quantile cutoff between 0 and 1 for breakpoint detection. Higher values will result in higher precision but lower sensitivity.

Value

A [GRanges](#) object with breakpoint coordinates.

Author(s)

Aaron Taudt, David Porubsky, Ashley Sanders

findCNVs	<i>Find copy number variations</i>
----------	------------------------------------

Description

findCNVs classifies the binned read counts into several states which represent copy-number-variation.

Usage

```
findCNVs(binned.data, ID = NULL, eps = 0.1, init = "standard",
max.time = -1, max.iter = 1000, num.trials = 15, eps.try = 10 * eps,
num.threads = 1, count.cutoff.quantile = 0.999, strand = "*",
states = c("zero-inflation", paste0(0:10, "-somy")),
most.frequent.state = "2-somy", method = "HMM", algorithm = "EM",
initial.params = NULL)
```

Arguments

binned.data	A GRanges object with binned read counts.
ID	An identifier that will be used to identify this sample in various downstream functions. Could be the file name of the binned.data for example.
eps	Convergence threshold for the Baum-Welch algorithm.
init	One of the following initialization procedures: standard The negative binomial of state '2-somy' will be initialized with mean=mean(counts), var=var(counts). This procedure usually gives good convergence. random Mean and variance of the negative binomial of state '2-somy' will be initialized with random values (in certain boundaries, see source code). Try this if the standard procedure fails to produce a good fit.
max.time	The maximum running time in seconds for the Baum-Welch algorithm. If this time is reached, the Baum-Welch will terminate after the current iteration finishes. Set max.time = -1 for no limit.
max.iter	The maximum number of iterations for the Baum-Welch algorithm. Set max.iter = -1 for no limit.

<code>num.trials</code>	The number of trials to find a fit where <code>state.most.frequent.state</code> is most frequent. Each time, the HMM is seeded with different random initial values.
<code>eps.try</code>	If code <code>num.trials</code> is set to greater than 1, <code>eps.try</code> is used for the trial runs. If unset, <code>eps</code> is used.
<code>num.threads</code>	Number of threads to use. Setting this to >1 may give increased performance.
<code>count.cutoff.quantile</code>	A quantile between 0 and 1. Should be near 1. Read counts above this quantile will be set to the read count specified by this quantile. Filtering very high read counts increases the performance of the Baum-Welch fitting procedure. However, if your data contains very few peaks they might be filtered out. Set <code>count.cutoff.quantile=1</code> in this case.
<code>strand</code>	Run the HMM only for the specified strand. One of <code>c('+', '-', '*')</code> .
<code>states</code>	A subset or all of <code>c("zero-inflation", "0-somy", "1-somy", "2-somy", "3-somy", "4-somy", ...)</code> . This vector defines the states that are used in the Hidden Markov Model. The order of the entries must not be changed.
<code>most.frequent.state</code>	One of the states that were given in <code>states</code> . The specified state is assumed to be the most frequent one. This can help the fitting procedure to converge into the correct fit.
<code>method</code>	Any combination of <code>c('HMM', 'dnacopy')</code> . Option <code>method='HMM'</code> uses a Hidden Markov Model as described in doi:10.1186/s13059-016-0971-7 to call copy numbers. Option <code>'dnacopy'</code> uses the DNAcopy package to call copy numbers similarly to the method proposed in doi:10.1038/nmeth.3578, which gives more robust but less sensitive results.
<code>algorithm</code>	One of <code>c('baumWelch', 'EM')</code> . The expectation maximization ('EM') will find the most likely states and fit the best parameters to the data, the <code>'baumWelch'</code> will find the most likely states using the initial parameters.
<code>initial.params</code>	A aneuHMM object or file containing such an object from which initial starting parameters will be extracted.

Details

findCNVs uses a 6-state Hidden Markov Model to classify the binned read counts: state '0-somy' with a delta function as emission density (only zero read counts), '1-somy', '2-somy', '3-somy', '4-somy', etc. with negative binomials (see [dnbinom](#)) as emission densities. A Baum-Welch algorithm is employed to estimate the parameters of the distributions. See our paper [citation\("AneuFinder"\)](#) for a detailed description of the method.

Value

An [aneuHMM](#) object.

Author(s)

Aaron Taudt

Examples

```
## Get an example BED file with single-cell-sequencing reads
bedfile <- system.file("extdata", "KK150311_VI_07.bam.bed.gz", package="AneuFinderData")
## Bin the BAM file into bin size 1Mp
```

```

binned <- binReads(bedfile, assembly='mm10', binsize=1e6,
                  chromosomes=c(1:19,'X','Y'))
## Fit the Hidden Markov Model
model <- findCNVs(binned[[1]], eps=0.1, max.time=60)
## Check the fit
plot(model, type='histogram')

```

findCNVs.strandseq *Find copy number variations (strandseq)*

Description

findCNVs.strandseq classifies the binned read counts into several states which represent the number of chromatids on each strand.

Usage

```

findCNVs.strandseq(binned.data, ID = NULL, eps = 0.1, init = "standard",
                  max.time = -1, max.iter = 1000, num.trials = 5, eps.try = 10 * eps,
                  num.threads = 1, count.cutoff.quantile = 0.999, strand = "*",
                  states = c("zero-inflation", paste0(0:10, "-somy")),
                  most.frequent.state = "1-somy", method = "HMM", algorithm = "EM",
                  initial.params = NULL)

```

Arguments

binned.data	A GRanges object with binned read counts.
ID	An identifier that will be used to identify this sample in various downstream functions. Could be the file name of the binned.data for example.
eps	Convergence threshold for the Baum-Welch algorithm.
init	One of the following initialization procedures: standard The negative binomial of state '2-somy' will be initialized with mean=mean(counts), var=var(counts). This procedure usually gives good convergence. random Mean and variance of the negative binomial of state '2-somy' will be initialized with random values (in certain boundaries, see source code). Try this if the standard procedure fails to produce a good fit.
max.time	The maximum running time in seconds for the Baum-Welch algorithm. If this time is reached, the Baum-Welch will terminate after the current iteration finishes. Set max.time = -1 for no limit.
max.iter	The maximum number of iterations for the Baum-Welch algorithm. Set max.iter = -1 for no limit.
num.trials	The number of trials to find a fit where state most.frequent.state is most frequent. Each time, the HMM is seeded with different random initial values.
eps.try	If code num.trials is set to greater than 1, eps.try is used for the trial runs. If unset, eps is used.
num.threads	Number of threads to use. Setting this to >1 may give increased performance.

<code>count.cutoff.quantile</code>	A quantile between 0 and 1. Should be near 1. Read counts above this quantile will be set to the read count specified by this quantile. Filtering very high read counts increases the performance of the Baum-Welch fitting procedure. However, if your data contains very few peaks they might be filtered out. Set <code>count.cutoff.quantile=1</code> in this case.
<code>strand</code>	Run the HMM only for the specified strand. One of <code>c('+', '-', '*')</code> .
<code>states</code>	A subset or all of <code>c("zero-inflation", "0-somy", "1-somy", "2-somy", "3-somy", "4-somy", ...)</code> . This vector defines the states that are used in the Hidden Markov Model. The order of the entries must not be changed.
<code>most.frequent.state</code>	One of the states that were given in <code>states</code> . The specified state is assumed to be the most frequent one. This can help the fitting procedure to converge into the correct fit.
<code>method</code>	Any combination of <code>c('HMM', 'dnacopy')</code> . Option <code>method='HMM'</code> uses a Hidden Markov Model as described in doi:10.1186/s13059-016-0971-7 to call copy numbers. Option <code>'dnacopy'</code> uses the DNAcopy package to call copy numbers similarly to the method proposed in doi:10.1038/nmeth.3578, which gives more robust but less sensitive results.
<code>algorithm</code>	One of <code>c('baumWelch', 'EM')</code> . The expectation maximization ('EM') will find the most likely states and fit the best parameters to the data, the <code>'baumWelch'</code> will find the most likely states using the initial parameters.
<code>initial.params</code>	A aneuHMM object or file containing such an object from which initial starting parameters will be extracted.

Details

`findCNVs.strandseq` uses a Hidden Markov Model to classify the binned read counts: state 'zero-inflation' with a delta function as emission density (only zero read counts), '0-somy' with geometric distribution, '1-somy', '2-somy', '3-somy', '4-somy', etc. with negative binomials (see [dnbinom](#)) as emission densities. A expectation-maximization (EM) algorithm is employed to estimate the parameters of the distributions. See our paper citation ("[AneuFinder](#)") for a detailed description of the method.

Value

An [aneuBiHMM](#) object.

Author(s)

Aaron Taudt

Examples

```
## Get an example BED file with single-cell-sequencing reads
bedfile <- system.file("extdata", "KK150311_VI_07.bam.bed.gz", package="AneuFinderData")
## Bin the file into bin size 1Mp
binned <- binReads(bedfile, assembly='mm10', binsize=1e6,
                  chromosomes=c(1:19,'X','Y'), pairedEndReads=TRUE)
## Fit the Hidden Markov Model
model <- findCNVs.strandseq(binned[[1]], eps=1, max.time=60, method='HMM')
## Check the fit
plot(model, type='histogram')
```

```
plot(model, type='profile')
```

fixedWidthBins	<i>Make fixed-width bins</i>
----------------	------------------------------

Description

Make fixed-width bins based on given bin size.

Usage

```
fixedWidthBins(bamfile = NULL, assembly = NULL, chrom.lengths = NULL,
  chromosome.format, binsizes = 1e+06, chromosomes = NULL)
```

Arguments

bamfile	A BAM file from which the header is read to determine the chromosome lengths. If a bamfile is specified, option assembly is ignored.
assembly	An assembly from which the chromosome lengths are determined. Please see fetchExtendedChromInfoFromUCSC for available assemblies. This option is ignored if bamfile is specified. Alternatively a data.frame generated by fetchExtendedChromInfoFromUCSC
chrom.lengths	A named character vector with chromosome lengths. Names correspond to chromosomes.
chromosome.format	A character specifying the format of the chromosomes if assembly is specified. Either 'NCBI' for (1,2,3 ...) or 'UCSC' for (chr1,chr2,chr3 ...). If a bamfile or chrom.lengths is supplied, the format will be chosen automatically.
binsizes	A vector of bin sizes in base pairs.
chromosomes	A subset of chromosomes for which the bins are generated.

Value

A list() of [GRanges](#) objects with fixed-width bins.

Author(s)

Aaron Taudt

Examples

```
## Make fixed-width bins of size 500kb and 1Mb
bins <- fixedWidthBins(assembly='mm10', chromosome.format='NCBI', binsizes=c(5e5,1e6))
bins
```

getDistinctColors *Get distinct colors*

Description

Get a set of distinct colors selected from `colors`.

Usage

```
getDistinctColors(n, start.color = "blue4", exclude.colors = c("white",  
  "black", "gray", "grey", "\\<yellow\\>", "yellow1", "lemonchiffon"),  
  exclude.brightness.above = 1, exclude.rgb.above = 210)
```

Arguments

`n` Number of colors to select. If `n` is a character vector, `length(n)` will be taken as the number of colors and the colors will be named by `n`.

`start.color` Color to start the selection process from.

`exclude.colors` Character vector with colors that should not be used.

`exclude.brightness.above`
 Exclude colors where the 'brightness' value in HSV space is above. This is useful to obtain a matt palette.

`exclude.rgb.above`
 Exclude colors where all RGB values are above. This is useful to exclude whitish colors.

Details

The function computes the euclidian distance between all `colors` and iteratively selects those that have the furthest closes distance to the set of already selected colors.

Value

A character vector with colors.

Author(s)

Aaron Taudt

Examples

```
cols <- AneuFinder:::getDistinctColors(5)  
pie(rep(1,5), labels=cols, col=cols)
```

`getQC`*Obtain a data.frame with quality metrics*

Description

Obtain a data.frame with quality metrics from a list of `aneuHMM` objects or a list of files that contain such objects.

Usage

```
getQC(models)
```

Arguments

`models` A list of `GRanges` or `aneuHMM` objects or a character vector with files that contain such objects.

Details

The employed quality measures are:

- `total.read.count`: Total read count.
- `avg.binsize`: Average binsize.
- `avg.read.count`: Average read count.
- `spikiness`: Bin-to-bin variability of read count.
- `entropy`: Shannon entropy of read counts.
- `complexity`: Library complexity approximated with a Michaelis-Menten curve.
- `loglik`: Loglikelihood of the Hidden Markov Model.
- `num.segments`: Number of copy number segments that have been found.
- `bhattacharyya distance`: Bhattacharyya distance between 1-somy and 2-somy distributions.
- `sos`: Sum-of-squares distance of read counts to the fitted distributions in their respective segments.

Value

A data.frame with columns

Author(s)

Aaron Taudt

Examples

```
## Get a list of HMMs
folder <- system.file("extdata", "primary-lung", "hmm", package="AneuFinderData")
files <- list.files(folder, full.names=TRUE)
df <- getQC(files)
```

getSCEcoordinates	<i>Get SCE coordinates</i>
-------------------	----------------------------

Description

Extracts the coordinates of a sister chromatid exchanges (SCE) from an [aneuBiHMM](#) object.

Usage

```
getSCEcoordinates(model, resolution = c(3, 6), min.segwidth = 2,  
  fragments = NULL)
```

Arguments

model	An aneuBiHMM object.
resolution	An integer vector specifying the resolution at bin level at which to scan for SCE events.
min.segwidth	Segments below this width will be removed before scanning for SCE events.
fragments	A GRanges object with read fragments or a file that contains such an object. These reads will be used for fine mapping of the SCE events.

Value

A [GRanges](#) object containing the SCE coordinates.

Author(s)

Aaron Taudt

Examples

```
## Get an example BED file with single-cell-sequencing reads  
bedfile <- system.file("extdata", "KK150311_VI_07.bam.bed.gz", package="AneuFinderData")  
## Bin the BAM file into bin size 1Mp  
binned <- binReads(bedfile, assembly='hg19', binsize=1e6,  
  chromosomes=c(1:22,'X','Y'), pairedEndReads=TRUE)  
## Fit the Hidden Markov Model  
model <- findCNVs.strandseq(binned[[1]], eps=0.1, max.time=60)  
## Find sister chromatid exchanges  
model$sce <- getSCEcoordinates(model)  
print(model$sce)  
plot(model)
```

getSegments	<i>Extract segments and cluster</i>
-------------	-------------------------------------

Description

Extract segments and ID from a list of [aneuHMM](#) or [aneuBiHMM](#) objects and cluster if desired.

Usage

```
getSegments(hmms, cluster = TRUE, classes = NULL, exclude.regions = NULL)
```

Arguments

hmms	A list of aneuHMM or aneuBiHMM objects or a character vector of files that contains such objects.
cluster	Either TRUE or FALSE, indicating whether the samples should be clustered by similarity in their CNV-state.
classes	A vector with class labels the same length as hmms. If supplied, the clustering will be ordered optimally with respect to the class labels (see RearrangeJoseph).
exclude.regions	A GRanges with regions that will be excluded from the computation of the clustering. This can be useful to exclude regions with artifacts.

Value

A `list()` with (clustered) segments and SCE coordinates.

heatmapAneuploidies	<i>Plot aneuploidy state</i>
---------------------	------------------------------

Description

Plot a heatmap of aneuploidy state for multiple samples. Samples can be clustered and the output can be returned as `data.frame`.

Usage

```
heatmapAneuploidies(hmms, ylabels = NULL, cluster = TRUE,
  as.data.frame = FALSE)
```

Arguments

hmms	A list of aneuHMM objects or a character vector with files that contain such objects.
ylabels	A vector with labels for the y-axis. The vector must have the same length as hmms. If NULL the IDs from the aneuHMM objects will be used.
cluster	If TRUE, the samples will be clustered by similarity in their CNV-state.
as.data.frame	If TRUE, instead of a plot, a <code>data.frame</code> with the aneuploidy state for each sample will be returned.

Value

A `ggplot` object or a `data.frame`, depending on option `as.data.frame`.

Author(s)

Aaron Taudt

Examples

```
## Get results from a small-cell-lung-cancer
folder <- system.file("extdata", "primary-lung", "hmms", package="AneuFinderData")
files <- list.files(folder, full.names=TRUE)
## Plot the ploidy state per chromosome
heatmapAneuploidies(files, cluster=FALSE)
## Return the ploidy state as data.frame
df <- heatmapAneuploidies(files, cluster=FALSE, as.data.frame=TRUE)
head(df)
```

heatmapGenomewide	<i>Genome wide heatmap of CNV-state</i>
-------------------	-----------------------------------------

Description

Plot a genome wide heatmap of copy number variation state. This heatmap is best plotted to file, because in most cases it will be too big for cleanly plotting it to screen.

Usage

```
heatmapGenomewide(hmms, ylabels = NULL, classes = NULL,
  reorder.by.class = TRUE, classes.color = NULL, file = NULL,
  cluster = TRUE, plot.SCE = TRUE, hotspots = NULL,
  exclude.regions = NULL)
```

Arguments

hmms	A list of <code>aneuHMM</code> objects or a character vector with files that contain such objects.
ylabels	A vector with labels for the y-axis. The vector must have the same length as <code>hmms</code> . If <code>NULL</code> the IDs from the <code>aneuHMM</code> objects will be used.
classes	A character vector with the classification of the elements on the y-axis. The vector must have the same length as <code>hmms</code> .
reorder.by.class	If <code>TRUE</code> , the dendrogram will be reordered to display similar classes next to each other.
classes.color	A (named) vector with colors that are used to distinguish classes. Names must correspond to the unique elements in <code>classes</code> .
file	A PDF file to which the heatmap will be plotted.
cluster	Either <code>TRUE</code> or <code>FALSE</code> , indicating whether the samples should be clustered by similarity in their CNV-state.

plot.SCE Logical indicating whether breakpoints should be plotted.

hotspots A [GRanges](#) object with coordinates of genomic hotspots (see [hotspotter](#)).

exclude.regions A [GRanges](#) with regions that will be excluded from the computation of the clustering. This can be useful to exclude regions with artifacts.

Value

A [ggplot](#) object or NULL if a file was specified.

Examples

```
## Get results from a small-cell-lung-cancer
lung.folder <- system.file("extdata", "primary-lung", "hmms", package="AneuFinderData")
lung.files <- list.files(lung.folder, full.names=TRUE)
## Get results from the liver metastasis of the same patient
liver.folder <- system.file("extdata", "metastasis-liver", "hmms", package="AneuFinderData")
liver.files <- list.files(liver.folder, full.names=TRUE)
## Plot a clustered heatmap
classes <- c(rep('lung', length(lung.files)), rep('liver', length(liver.files)))
labels <- c(paste('lung', 1:length(lung.files)), paste('liver', 1:length(liver.files)))
heatmapGenomewide(c(lung.files, liver.files), ylabels=labels, classes=classes,
                  classes.color=c('blue', 'red'))
```

heatmapGenomewideClusters

Plot heatmaps for quality control

Description

This function is a convenient wrapper to call [heatmapGenomewide](#) for all clusters after calling [clusterByQuality](#) and plot the heatmaps into one pdf for efficient comparison.

Usage

```
heatmapGenomewideClusters(cl, file = NULL, ...)
```

Arguments

cl The return value of [clusterByQuality](#).

file A character specifying the output file.

... Further parameters passed on to [heatmapGenomewide](#).

Value

A [cowplot](#) object or NULL if a file was specified.

Examples

```
## Get a list of HMMs and cluster them
folder <- system.file("extdata", "primary-lung", "hmm", package="AneuFinderData")
files <- list.files(folder, full.names=TRUE)
cl <- clusterByQuality(files)
heatmapGenomewideClusters(cl)
```

hotspotter

Find hotspots of genomic events

Description

Find hotspots of genomic events by using kernel [density](#) estimation.

Usage

```
hotspotter(gr.list, bw, pval = 1e-08)
```

Arguments

<code>gr.list</code>	A list with GRanges object containing the coordinates of the genomic events.
<code>bw</code>	Bandwidth used for kernel density estimation (see density).
<code>pval</code>	P-value cutoff for hotspots.

Details

The hotspotter uses [density](#) to perform a KDE. A p-value is calculated by comparing the density profile of the genomic events with the density profile of a randomly subsampled set of genomic events. Due to this random sampling, the result can vary for each function call, most likely for hotspots whose p-value is close to the specified `pval`.

Value

A [GRanges](#) object containing coordinates of hotspots with p-values.

Author(s)

Aaron Taudt

importBed	<i>Read bed-file into GRanges</i>
-----------	-----------------------------------

Description

This is a simple convenience function to read a bed(.gz)-file into a [GRanges](#) object. The bed-file is expected to have the following fields: chromosome, start, end, name, score, strand.

Usage

```
importBed(bedfile, skip = 0, chromosome.format = "NCBI")
```

Arguments

bedfile	Filename of the bed or bed.gz file.
skip	Number of lines to skip at the beginning.
chromosome.format	Desired format of the chromosomes. Either 'NCBI' for (1,2,3 ...) or 'UCSC' for (chr1,chr2,chr3 ...).

Value

A [GRanges](#) object with the contents of the bed-file.

Author(s)

Aaron Taudt

Examples

```
## Get an example BED file with single-cell-sequencing reads
bedfile <- system.file("extdata", "KK150311_VI_07.bam.bed.gz", package="AneuFinderData")
## Import the file and skip the first 10 lines
data <- importBed(bedfile, skip=10)
```

initializeStates	<i>Initialize state factor levels and distributions</i>
------------------	---------------------------------------------------------

Description

Initialize the state factor levels and distributions for the specified states.

Usage

```
initializeStates(states)
```

Arguments

states	A subset of <code>c("zero-inflation", "0-somy", "1-somy", "2-somy", "3-somy", "4-somy", ...)</code> .
--------	-------------------------------------------------------------------------------------------------------

Value

A list with \$labels, \$distributions and \$multiplicity values for the given states.

karyotypeMeasures	<i>Measures for Karyotype Heterogeneity</i>
-------------------	---------------------------------------------

Description

Computes measures for karyotype heterogeneity. See the Details section for how these measures are defined.

Usage

```
karyotypeMeasures(hmms, normalChromosomeNumbers = NULL, regions = NULL,
  exclude.regions = NULL)
```

Arguments

hmms	A list with aneuHMM objects or a list of files that contain such objects.
normalChromosomeNumbers	A named integer vector or matrix with physiological copy numbers, where each element (vector) or column (matrix) corresponds to a chromosome. This is useful to specify male or female samples, e.g. <code>c('X'=2)</code> for female samples or <code>c('X'=1, 'Y'=1)</code> for male samples. Specify a vector if all your hmms have the same physiological copy numbers. Specify a matrix if your hmms have different physiological copy numbers (e.g. a mix of male and female samples). If not specified otherwise, '2' will be assumed for all chromosomes.
regions	A GRanges object containing ranges for which the karyotype measures will be computed.
exclude.regions	A GRanges with regions that will be excluded from the computation of the karyotype measures. This can be useful to exclude regions with artifacts.

Details

We define x as the vector of copy number states for each position. The number of HMMs is S . The measures are computed for each bin as follows:

Aneuploidy: $D = \text{mean}(\text{abs}(x - P))$, where P is the physiological number of chromosomes at that position.

Heterogeneity: $H = \text{sum}(\text{table}(x) * 0 : (\text{length}(\text{table}(x)) - 1)) / S$

Value

A list with two data.frames, containing the karyotype measures \$genomewide and \$per.chromosome. If region was specified, a third list entry \$regions will contain the regions with karyotype measures.

Author(s)

Aaron Taudt

Examples

```

### Example 1 ###
## Get results from a small-cell-lung-cancer
lung.folder <- system.file("extdata", "primary-lung", "hmms", package="AneuFinderData")
lung.files <- list.files(lung.folder, full.names=TRUE)
## Get results from the liver metastasis of the same patient
liver.folder <- system.file("extdata", "metastasis-liver", "hmms", package="AneuFinderData")
liver.files <- list.files(liver.folder, full.names=TRUE)
## Compare karyotype measures between the two cancers
normal.chrom.numbers <- rep(2, 23)
names(normal.chrom.numbers) <- c(1:22, 'X')
lung <- karyotypeMeasures(lung.files, normalChromosomeNumbers=normal.chrom.numbers)
liver <- karyotypeMeasures(liver.files, normalChromosomeNumbers=normal.chrom.numbers)
print(lung$genomewide)
print(liver$genomewide)

### Example 2 ###
## Construct a matrix with physiological copy numbers for a mix of 5 male and 5 female samples
normal.chrom.numbers <- matrix(2, nrow=10, ncol=24,
                               dimnames=list(sample=c(paste('male', 1:5), paste('female', 6:10)),
                                               chromosome=c(1:22, 'X', 'Y')))
normal.chrom.numbers[1:5, c('X', 'Y')] <- 1
normal.chrom.numbers[6:10, c('Y')] <- 0
print(normal.chrom.numbers)

### Example 3 ###
## Exclude artifact regions with high variance
consensus <- consensusSegments(c(lung.files, liver.files))
variance <- apply(consensus$copy.number, 1, var)
exclude.regions <- consensus[variance > quantile(variance, 0.999)]
## Compare karyotype measures between the two cancers
normal.chrom.numbers <- rep(2, 23)
names(normal.chrom.numbers) <- c(1:22, 'X')
lung <- karyotypeMeasures(lung.files, normalChromosomeNumbers=normal.chrom.numbers,
                          exclude.regions = exclude.regions)
liver <- karyotypeMeasures(liver.files, normalChromosomeNumbers=normal.chrom.numbers,
                           exclude.regions = exclude.regions)
print(lung$genomewide)
print(liver$genomewide)

```

loadFromFiles

Load AneuFinder objects from file

Description

Wrapper to load **AneuFinder** objects from file and check the class of the loaded objects.

Usage

```
loadFromFiles(files, check.class = c("GRanges", "aneuHMM", "aneuBiHMM"))
```


Arguments

files	A list of GRanges , aneuHMM or aneuBiHMM objects or a character vector with files that contain such objects.
check.class	Any combination of <code>c('GRanges', 'aneuHMM', 'aneuBiHMM')</code> . If any of the loaded objects does not belong to the specified class, an error is thrown.

Value

A list of [GRanges](#), [aneuHMM](#) or [aneuBiHMM](#) objects.

Examples

```
## Get some files that you want to load
folder <- system.file("extdata", "primary-lung", "hmms", package="AneuFinderData")
files <- list.files(folder, full.names=TRUE)
## Load and plot the first ten
hmms <- loadFromFiles(files[1:10])
lapply(hmms, plot, type='profile')
```

plot.aneuBiHMM

Plotting function for [aneuBiHMM](#) objects

Description

Make different types of plots for [aneuBiHMM](#) objects.

Usage

```
## S3 method for class 'aneuBiHMM'
plot(x, type = "profile", ...)
```

Arguments

x	An aneuBiHMM object.
type	Type of the plot, one of <code>c('profile', 'histogram', 'karyogram')</code> . You can also specify the type with an integer number. profile An profile with read counts and CNV-state. histogram A histogram of binned read counts with fitted mixture distribution. karyogram A karyogram-like chromosome overview with CNV-state.
...	Additional arguments for the different plot types.

Value

A [ggplot](#) object.

plot.aneuHMM *Plotting function for aneuHMM objects*

Description

Make different types of plots for [aneuHMM](#) objects.

Usage

```
## S3 method for class 'aneuHMM'
plot(x, type = "profile", ...)
```

Arguments

x	An aneuHMM object.
type	Type of the plot, one of c('profile', 'histogram', 'karyogram'). You can also specify the type with an integer number. karyogram A karyogram-like chromosome overview with CNV-state. histogram A histogram of binned read counts with fitted mixture distribution. karyogram An profile with read counts and CNV-state.
...	Additional arguments for the different plot types.

Value

A [ggplot](#) object.

plot.character *Plotting function for saved AneuFinder objects*

Description

Convenience function that loads and plots a [AneuFinder](#) object in one step.

Usage

```
## S3 method for class 'character'
plot(x, ...)
```

Arguments

x	A filename that contains either binned.data or a aneuHMM .
...	Additional arguments.

Value

A [ggplot](#) object.

plot.GRanges	<i>Plotting function for binned read counts</i>
--------------	-------------------------------------------------

Description

Make plots for binned read counts from [binned.data](#).

Usage

```
## S3 method for class 'GRanges'
plot(x, type = "profile", ...)
```

Arguments

x	A GRanges object with binned read counts.
type	Type of the plot, one of <code>c('profile', 'histogram', 'karyogram')</code> . You can also specify the type with an integer number. karyogram A karyogram-like chromosome overview with read counts. histogram A histogram of read counts. profile An profile with read counts.
...	Additional arguments for the different plot types.

Value

A [ggplot](#) object.

plotHeterogeneity	<i>Heterogeneity vs. Aneuploidy</i>
-------------------	-------------------------------------

Description

Make heterogeneity vs. aneuploidy plots using individual chromosomes as datapoints.

Usage

```
plotHeterogeneity(hmms, hmms.list = NULL, normalChromosomeNumbers = NULL,
  plot = TRUE, regions = NULL, exclude.regions = NULL)
```

Arguments

hmms	A list of aneuHMM objects or a character vector with files that contain such objects.
hmms.list	Alternative input for a faceted plot. A named list() of lists of aneuHMM objects. Alternatively a named list() of character vectors with files that contain aneuHMM objects. List names serve as facets for plotting. If specified, <code>normalChromosomeNumbers</code> is assumed to be a list() of vectors (or matrices) instead of a vector (or matrix).

normalChromosomeNumbers	A named integer vector or matrix with physiological copy numbers, where each element (vector) or column (matrix) corresponds to a chromosome. This is useful to specify male or female samples, e.g. <code>c('X'=2)</code> for female samples or <code>c('X'=1, 'Y'=1)</code> for male samples. Specify a vector if all your hmms have the same physiological copy numbers. Specify a matrix if your hmms have different physiological copy numbers (e.g. a mix of male and female samples). If not specified otherwise, '2' will be assumed for all chromosomes. If you have specified <code>hmms.list</code> instead of <code>hmms</code> , <code>normalChromosomeNumbers</code> is assumed to be a <code>list()</code> of vectors (or matrices), with one vector (or matrix) for each element in <code>hmms.list</code> .
plot	A logical indicating whether to plot or to return the underlying data.frame.
regions	A GRanges object containing ranges for which the karyotype measures will be computed.
exclude.regions	A GRanges with regions that will be excluded from the computation of the karyotype measures. This can be useful to exclude regions with artifacts.

Value

A [ggplot](#) object or a `data.frame` if `plot=FALSE`.

Examples

```
### Example 1: A faceted plot of lung and liver cells ###
## Get results from a small-cell-lung-cancer
lung.folder <- system.file("extdata", "primary-lung", "hmms", package="AneuFinderData")
lung.files <- list.files(lung.folder, full.names=TRUE)
## Get results from the liver metastasis of the same patient
liver.folder <- system.file("extdata", "metastasis-liver", "hmms", package="AneuFinderData")
liver.files <- list.files(liver.folder, full.names=TRUE)
## Make heterogeneity plots
plotHeterogeneity(hmms.list = list(lung=lung.files, liver=liver.files))

### Example 2: Plot a mixture sample of male and female cells ###
## Get results from a small-cell-lung-cancer
folder <- system.file("extdata", "primary-lung", "hmms", package="AneuFinderData")
files <- list.files(folder, full.names=TRUE)
## Construct a matrix with physiological copy numbers for a mix of 48 male and 48 female samples
normal.chrom.numbers <- matrix(2, nrow=96, ncol=24,
                              dimnames=list(sample=c(paste('male', 1:48), paste('female', 49:96)),
                                             chromosome=c(1:22, 'X', 'Y')))
normal.chrom.numbers[1:48,c('X','Y')] <- 1
normal.chrom.numbers[49:96,c('Y')] <- 0
head(normal.chrom.numbers)
## Make heterogeneity plots
plotHeterogeneity(hmms = files, normalChromosomeNumbers = normal.chrom.numbers)

### Example 3: A faceted plot of male lung and female liver cells ###
## Get results from a small-cell-lung-cancer
lung.folder <- system.file("extdata", "primary-lung", "hmms", package="AneuFinderData")
lung.files <- list.files(lung.folder, full.names=TRUE)
## Specify the physiological copy numbers
chrom.numbers.lung <- c(rep(2, 22), 1, 1)
names(chrom.numbers.lung) <- c(1:22, 'X', 'Y')
```

```

print(chrom.numbers.lung)
## Get results from the liver metastasis of the same patient
liver.folder <- system.file("extdata", "metastasis-liver", "hmms", package="AneuFinderData")
liver.files <- list.files(liver.folder, full.names=TRUE)
## Specify the physiological copy numbers
chrom.numbers.liver <- c(rep(2, 22), 2, 0)
names(chrom.numbers.liver) <- c(1:22, 'X', 'Y')
print(chrom.numbers.liver)
## Make heterogeneity plots
plotHeterogeneity(hmms.list = list(lung=lung.files, liver=liver.files),
                  normalChromosomeNumbers = list(chrom.numbers.lung, chrom.numbers.liver))

### Example 4 ###
## Exclude artifact regions with high variance
consensus <- consensusSegments(c(lung.files, liver.files))
variance <- apply(consensus$copy.number, 1, var)
exclude.regions <- consensus[variance > quantile(variance, 0.999)]
## Make heterogeneity plots
plotHeterogeneity(hmms.list = list(lung=lung.files, liver=liver.files),
                  exclude.regions=exclude.regions)

```

plotHistogram

Plot a histogram of binned read counts with fitted mixture distribution

Description

Plot a histogram of binned read counts from with fitted mixture distributions from a [aneuHMM](#) object.

Usage

```
plotHistogram(model, state = NULL, strand = "*", chromosome = NULL,
              start = NULL, end = NULL)
```

Arguments

model	A aneuHMM object.
state	Plot the histogram only for the specified CNV-state.
strand	One of c('+', '-', '*'). Plot the histogram only for the specified strand.
chromosome, start, end	Plot the histogram only for the specified chromosome, start and end position.

Value

A [ggplot](#) object.

plotKaryogram	<i>Karyogram-like chromosome overview</i>
---------------	-------------------------------------------

Description

Plot a karyogram-like chromosome overview with read counts and CNV-state from a [aneuHMM](#) object or [binned.data](#).

Usage

```
plotKaryogram(model, both.strands = FALSE, plot.SCE = FALSE, file = NULL)
```

Arguments

model	A aneuHMM object or binned.data .
both.strands	If TRUE, strands will be plotted separately.
plot.SCE	Logical indicating whether breakpoints should be plotted.
file	A PDF file where the plot will be saved.

Value

A [ggplot](#) object or NULL if a file was specified.

plotProfile	<i>Read count and CNV profile</i>
-------------	-----------------------------------

Description

Plot a profile with read counts and CNV-state from a [aneuHMM](#) object or [binned.data](#).

Usage

```
plotProfile(model, both.strands = FALSE, plot.SCE = TRUE, file = NULL,
  normalize.counts = NULL)
```

Arguments

model	A aneuHMM object or binned.data .
both.strands	If TRUE, strands will be plotted separately.
plot.SCE	Logical indicating whether breakpoints should be plotted.
file	A PDF file where the plot will be saved.
normalize.counts	An character giving the copy number state to which to normalize the counts, e.g. '1-somy', '2-somy' etc.

Value

A [ggplot](#) object or NULL if a file was specified.

plot_pca	<i>Perform a PCA for copy number profiles</i>
----------	-----------------------------------------------

Description

Perform a PCA for copy number profiles in [aneuHMM](#) objects.

Usage

```
plot_pca(hmms, PC1 = 1, PC2 = 2, colorBy = NULL, plot = TRUE,  
         exclude.regions = NULL)
```

Arguments

hmms	A list of aneuHMM objects or a character vector with files that contain such objects.
PC1	Integer specifying the first of the principal components to plot.
PC2	Integer specifying the second of the principal components to plot.
colorBy	A character vector of the same length as hmms which is used to color the points in the plot.
plot	Set to FALSE if you want to return the data.frame that is used for plotting instead of the plot.
exclude.regions	A GRanges with regions that will be excluded from the computation of the PCA. This can be useful to exclude regions with artifacts.

Value

A [ggplot](#) object or a data.frame if plot=FALSE.

Examples

```
## Get results from a small-cell-lung-cancer  
lung.folder <- system.file("extdata", "primary-lung", "hmms", package="AneuFinderData")  
lung.files <- list.files(lung.folder, full.names=TRUE)  
## Get results from the liver metastasis of the same patient  
liver.folder <- system.file("extdata", "metastasis-liver", "hmms", package="AneuFinderData")  
liver.files <- list.files(liver.folder, full.names=TRUE)  
## Plot a clustered heatmap  
classes <- c(rep('lung', length(lung.files)), rep('liver', length(liver.files)))  
labels <- c(paste('lung', 1:length(lung.files)), paste('liver', 1:length(liver.files)))  
plot_pca(c(lung.files, liver.files), colorBy=classes, PC1=2, PC2=4)
```

print.aneuBiHMM	<i>Print aneuBiHMM object</i>
-----------------	-------------------------------

Description

Print aneuBiHMM object

Usage

```
## S3 method for class 'aneuBiHMM'  
print(x, ...)
```

Arguments

x	An aneuBiHMM object.
...	Ignored.

Value

An invisible NULL.

print.aneuHMM	<i>Print aneuHMM object</i>
---------------	-----------------------------

Description

Print aneuHMM object

Usage

```
## S3 method for class 'aneuHMM'  
print(x, ...)
```

Arguments

x	An aneuHMM object.
...	Ignored.

Value

An invisible NULL.

qualityControl	<i>Quality control measures for binned read counts</i>
----------------	--------------------------------------------------------

Description

Calculate various quality control measures on binned read counts.

Usage

```
qc.spikiness(counts)
```

```
qc.entropy(counts)
```

```
qc.bhattacharyya(hmm)
```

```
qc.sos(hmm)
```

Arguments

counts A vector of binned read counts.

hmm An [aneuHMM](#) object.

Details

The Shannon entropy is defined as $S = -\sum(n * \log(n))$, where $n = counts/sum(counts)$.

Spikyness is defined as $K = \sum(abs(diff(counts)))/sum(counts)$.

Value

A numeric.

Functions

- qc.spikiness: Calculate the spikiness of a library
- qc.entropy: Calculate the Shannon entropy of a library
- qc.bhattacharyya: Calculate the Bhattacharyya distance between the '1-somy' and '2-somy' distribution
- qc.sos: Sum-of-squares distance from the read counts to the fitted distributions

Author(s)

Aaron Taudt

readConfig	<i>Read AneuFinder configuration file</i>
------------	-------------------------------------------

Description

Read an AneuFinder configuration file into a list structure. The configuration file has to be specified in INI format. R expressions can be used and will be evaluated.

Usage

```
readConfig(configfile)
```

Arguments

configfile	Path to the configuration file
------------	--------------------------------

Value

A list with one entry for each element in configfile.

Author(s)

Aaron Taudt

simulateReads	<i>Simulate reads from genome</i>
---------------	-----------------------------------

Description

Simulate single or paired end reads from any **BSgenome** object. These simulated reads can be mapped to the reference genome using any aligner to produce BAM files that can be used for mappability correction.

Usage

```
simulateReads(bsgenome, readLength, bamfile, file,
  pairedEndFragmentLength = NULL, every.X.bp = 500)
```

Arguments

bsgenome	A BSgenome object containing the sequence of the reference genome.
readLength	The length in base pairs of the simulated reads that are written to file.
bamfile	A BAM file. This file is used to estimate the distribution of Phred quality scores.
file	The filename that is written to disk. The ending .fastq.gz will be appended.
pairedEndFragmentLength	If this option is specified, paired end reads with length readLength will be simulated coming from both ends of fragments of this size. NOT IMPLEMENTED YET.
every.X.bp	Stepsize for simulating reads. A read fragment will be simulated every X bp.

Details

Reads are simulated by splitting the genome into reads with the specified readLength.

Value

A fastq.gz file is written to disk.

Author(s)

Aaron Taudt

Examples

```
## Get an example BAM file with single-cell-sequencing reads
bamfile <- system.file("extdata", "BB150803_IV_074.bam", package="AneuFinderData")
## Simulate 51bp reads for at a distance of every 5000bp
if (require(BSgenome.Mmusculus.UCSC.mm10)) {
  simulateReads(BSgenome.Mmusculus.UCSC.mm10, bamfile=bamfile, readLength=51,
               file=tempfile(), every.X.bp=5000)
}
```

subsetByCNVprofile *Get IDs of a subset of models*

Description

Get the IDs of models that have a certain CNV profile. The result will be TRUE for models that overlap all specified ranges in profile by at least one base pair with the correct state.

Usage

```
subsetByCNVprofile(hmms, profile)
```

Arguments

hmms	A list of aneuHMM objects or a character vector with files that contain such objects.
profile	A GRanges object with metadata column 'expected.state' and optionally columns 'expected.mstate' and 'expected.pstate'.

Value

A named logical vector with TRUE for all models that are concordant with the given profile.

Examples

```
## Get results from a small-cell-lung-cancer
lung.folder <- system.file("extdata", "primary-lung", "hmms", package="AneuFinderData")
lung.files <- list.files(lung.folder, full.names=TRUE)
## Get all files that have a 3-somy on chromosome 1 and 4-somy on chromosome 2
profile <- GRanges(seqnames=c('1','2'), ranges=IRanges(start=c(1,1), end=c(195471971,182113224)),
                  expected.state=c('3-somy','4-somy'))
ids <- subsetByCNVprofile(lung.files, profile)
print(which(ids))
```

transCoord	<i>Transform genomic coordinates</i>
------------	--------------------------------------

Description

Add two columns with transformed genomic coordinates to the [GRanges](#) object. This is useful for making genomewide plots.

Usage

```
transCoord(gr)
```

Arguments

gr A [GRanges](#) object.

Value

The input [GRanges](#) with two additional metadata columns 'start.genome' and 'end.genome'.

univariate.findCNVs	<i>Find copy number variations (univariate)</i>
---------------------	-------------------------------------------------

Description

univariate.findCNVs classifies the binned read counts into several states which represent copy-number-variation.

Usage

```
univariate.findCNVs(binned.data, ID = NULL, eps = 0.1, init = "standard",
                    max.time = -1, max.iter = -1, num.trials = 1, eps.try = NULL,
                    num.threads = 1, count.cutoff.quantile = 0.999, strand = "*",
                    states = c("zero-inflation", paste0(0:10, "-somy")),
                    most.frequent.state = "2-somy", algorithm = "EM", initial.params = NULL)
```

Arguments

<code>binned.data</code>	A GRanges object with binned read counts.
<code>ID</code>	An identifier that will be used to identify this sample in various downstream functions. Could be the file name of the <code>binned.data</code> for example.
<code>eps</code>	Convergence threshold for the Baum-Welch algorithm.
<code>init</code>	One of the following initialization procedures: <code>standard</code> The negative binomial of state '2-somy' will be initialized with <code>mean=mean(counts)</code> , <code>var=var(counts)</code> . This procedure usually gives good convergence. <code>random</code> Mean and variance of the negative binomial of state '2-somy' will be initialized with random values (in certain boundaries, see source code). Try this if the <code>standard</code> procedure fails to produce a good fit.
<code>max.time</code>	The maximum running time in seconds for the Baum-Welch algorithm. If this time is reached, the Baum-Welch will terminate after the current iteration finishes. Set <code>max.time = -1</code> for no limit.
<code>max.iter</code>	The maximum number of iterations for the Baum-Welch algorithm. Set <code>max.iter = -1</code> for no limit.
<code>num.trials</code>	The number of trials to find a fit where <code>most.frequent.state</code> is most frequent. Each time, the HMM is seeded with different random initial values.
<code>eps.try</code>	If code <code>num.trials</code> is set to greater than 1, <code>eps.try</code> is used for the trial runs. If unset, <code>eps</code> is used.
<code>num.threads</code>	Number of threads to use. Setting this to <code>>1</code> may give increased performance.
<code>count.cutoff.quantile</code>	A quantile between 0 and 1. Should be near 1. Read counts above this quantile will be set to the read count specified by this quantile. Filtering very high read counts increases the performance of the Baum-Welch fitting procedure. However, if your data contains very few peaks they might be filtered out. Set <code>count.cutoff.quantile=1</code> in this case.
<code>strand</code>	Run the HMM only for the specified strand. One of <code>c('+', '-', '*')</code> .
<code>states</code>	A subset or all of <code>c("zero-inflation", "0-somy", "1-somy", "2-somy", "3-somy", "4-somy", ...)</code> . This vector defines the states that are used in the Hidden Markov Model. The order of the entries must not be changed.
<code>most.frequent.state</code>	One of the states that were given in <code>states</code> . The specified state is assumed to be the most frequent one. This can help the fitting procedure to converge into the correct fit.
<code>algorithm</code>	One of <code>c('baumWelch', 'EM')</code> . The expectation maximization ('EM') will find the most likely states and fit the best parameters to the data, the 'baumWelch' will find the most likely states using the initial parameters.
<code>initial.params</code>	A aneuHMM object or file containing such an object from which initial starting parameters will be extracted.

Value

An [aneuHMM](#) object.

variableWidthBins	<i>Make variable-width bins</i>
-------------------	---------------------------------

Description

Make variable-width bins based on a reference BAM file. This can be a simulated file (produced by [simulateReads](#) and aligned with your favourite aligner) or a real reference.

Usage

```
variableWidthBins(reads, binsizes, chromosomes = NULL)
```

Arguments

reads	A GRanges with reads. See bam2GRanges and bed2GRanges .
binsizes	A vector with binsizes. Resulting bins will be close to the specified binsizes.
chromosomes	A subset of chromosomes for which the bins are generated.

Details

Variable-width bins are produced by first binning the reference BAM file with fixed-width bins and selecting the desired number of reads per bin as the (non-zero) maximum of the histogram. A new set of bins is then generated such that every bin contains the desired number of reads.

Value

A `list()` of [GRanges](#) objects with variable-width bins.

Author(s)

Aaron Taudt

Examples

```
## Get an example BED file with single-cell-sequencing reads
bedfile <- system.file("extdata", "KK150311_VI_07.bam.bed.gz", package="AneuFinderData")
## Read the file into a GRanges object
reads <- bed2GRanges(bedfile, assembly='mm10', chromosomes=c(1:19,'X','Y'),
                    min.mapq=10, remove.duplicate.reads=TRUE)
## Make variable-width bins of size 500kb and 1Mb
bins <- variableWidthBins(reads, binsizes=c(5e5,1e6))
## Plot the distribution of binsizes
hist(width(bins[['1e+06']]), breaks=50)
```

writeConfig	<i>Write AneuFinder configuration file</i>
-------------	--------------------------------------------

Description

Write an AneuFinder configuration file from a list structure.

Usage

```
writeConfig(conf, configfile)
```

Arguments

conf	A list structure with parameter values. Each entry will be written in one line.
configfile	Filename of the outputfile.

Value

NULL

Author(s)

Aaron Taudt

zinbinom	<i>The Zero-inflated Negative Binomial Distribution</i>
----------	---------------------------------------------------------

Description

Density, distribution function, quantile function and random generation for the zero-inflated negative binomial distribution with parameters w , $size$ and $prob$.

Usage

```
dzinbinom(x, w, size, prob, mu)
pzinbinom(q, w, size, prob, mu, lower.tail = TRUE)
qzinbinom(p, w, size, prob, mu, lower.tail = TRUE)
rzinbinom(n, w, size, prob, mu)
```

Arguments

x	Vector of (non-negative integer) quantiles.
w	Weight of the zero-inflation. $0 \leq w \leq 1$.
size	Target for number of successful trials, or dispersion parameter (the shape parameter of the gamma mixing distribution). Must be strictly positive, need not be integer.
prob	Probability of success in each trial. $0 < \text{prob} \leq 1$.
mu	Alternative parametrization via mean: see ‘Details’.
q	Vector of quantiles.
lower.tail	logical; if TRUE (default), probabilities are $P[X \leq x]$, otherwise, $P[X > x]$.
p	Vector of probabilities.
n	number of observations. If $\text{length}(n) > 1$, the length is taken to be the number required.

Details

The zero-inflated negative binomial distribution with $\text{size} = n$ and $\text{prob} = p$ has density

$$p(x) = w + (1 - w) \frac{\Gamma(x + n)}{\Gamma(n)x!} p^n (1 - p)^x$$

for $x = 0, n > 0, 0 < p \leq 1$ and $0 \leq w \leq 1$.

$$p(x) = (1 - w) \frac{\Gamma(x + n)}{\Gamma(n)x!} p^n (1 - p)^x$$

for $x = 1, 2, \dots, n > 0, 0 < p \leq 1$ and $0 \leq w \leq 1$.

Value

dzinbinom gives the density, pzinbinom gives the distribution function, qzinbinom gives the quantile function, and rzinbinom generates random deviates.

Functions

- dzinbinom: gives the density
- pzinbinom: gives the cumulative distribution function
- qzinbinom: gives the quantile function
- rzinbinom: random number generation

Author(s)

Matthias Heinig, Aaron Taudt

See Also

[Distributions](#) for standard distributions, including [dbinom](#) for the binomial, [dnbinom](#) for the negative binomial, [dpois](#) for the Poisson and [dgeom](#) for the geometric distribution, which is a special case of the negative binomial.

Index

- aneuBiHMM, [3](#), [13](#), [19](#), [26](#), [29](#), [33](#), [34](#), [41](#), [48](#)
- AneuFinder, [3](#), [4](#), [7](#), [40](#), [42](#)
- AneuFinder (AneuFinder-package), [3](#)
- Aneufinder, [3](#), [4](#), [14](#)
- AneuFinder-package, [3](#)
- aneuHMM, [6](#), [10](#), [13](#), [15](#), [18](#), [19](#), [23](#), [24](#), [27](#), [29](#), [32](#), [34](#), [35](#), [39](#), [41–43](#), [45–49](#), [51](#), [53](#)

- bam2GRanges, [7](#), [14](#), [22](#), [54](#)
- bamsignals, [5](#), [12](#)
- bed2GRanges, [8](#), [14](#), [54](#)
- biDNACopy.findCNVs, [9](#)
- bin the data, [4](#)
- binned.data, [10](#), [20](#), [21](#), [42](#), [43](#), [46](#)
- binning, [10](#), [10](#)
- binReads, [10](#)
- bivariate.findCNVs, [12](#)
- blacklist, [14](#)
- BSgenome, [50](#)

- clusterByQuality, [15](#), [36](#)
- collapseBins, [16](#)
- colors, [17](#), [31](#)
- compareMethods, [18](#)
- compareModels, [19](#)
- consensusSegments, [19](#)
- correctGC, [20](#)
- correctMappability, [21](#)
- countOverlaps, [12](#)
- cowplot, [36](#)

- dbinom, [56](#)
- deltaWCalculator, [22](#)
- density, [37](#)
- dgeom, [56](#)
- disjoin, [19](#)
- Distributions, [56](#)
- distributions, profiles and karyograms, [4](#)
- DNACopy, [5](#), [27](#), [29](#)
- DNACopy.findCNVs, [22](#)
- dnbinom, [27](#), [29](#), [56](#)
- dpois, [56](#)
- dzinbinom (zinbinom), [55](#)

- emControl, [15](#)
- estimateComplexity, [23](#)
- export, [23](#)
- exportCNVs (export), [23](#)
- exportGRanges (export), [23](#)
- exportReadCounts (export), [23](#)

- fetchExtendedChromInfoFromUCSC, [5](#), [9](#), [11](#), [14](#), [21](#), [30](#)
- filterSegments, [25](#)
- find copy-number-variations, [4](#)
- find sister-chromatid-exchange, [4](#)
- findBreakpoints, [25](#)
- findCNVs, [6](#), [18](#), [26](#)
- findCNVs.strandseq, [3](#), [28](#)
- fixedWidthBins, [10](#), [11](#), [14](#), [30](#)

- genomewide heatmaps, [4](#)
- GenomicRanges, [11](#)
- getDistinctColors, [31](#)
- getQC, [15](#), [32](#)
- getSCEcoordinates, [33](#)
- getSegments, [34](#)
- ggplot, [35](#), [36](#), [41–47](#)
- GRanges, [3](#), [5](#), [7–14](#), [19](#), [21–26](#), [28](#), [30](#), [32–34](#), [36–39](#), [41](#), [43](#), [44](#), [47](#), [51–54](#)

- heatmapAneuploidies, [34](#)
- heatmapGenomewide, [35](#), [36](#)
- heatmapGenomewideClusters, [36](#)
- hotspotter, [6](#), [36](#), [37](#)

- importBed, [38](#)
- initializeStates, [38](#)

- karyotypeMeasures, [39](#)

- loadFromFiles, [40](#)

- Mclust, [15](#)
- mclust, [15](#)

- plot.aneuBiHMM, [41](#)
- plot.aneuHMM, [42](#)
- plot.character, [42](#)

plot.GRanges, 43
plot_pca, 47
plotHeterogeneity, 43
plotHistogram, 45
plotKaryogram, 46
plotProfile, 46
print.aneuBiHMM, 48
print.aneuHMM, 48
pzinbinom (zinbinom), 55

qc.bhattacharyya (qualityControl), 49
qc.entropy (qualityControl), 49
qc.sos (qualityControl), 49
qc.spikiness (qualityControl), 49
qualityControl, 49
qzinbinom (zinbinom), 55

readConfig, 50
RearrangeJoseph, 34
rzinbinom (zinbinom), 55

scanBamWhat, 8
simulateReads, 50, 54
stateColors (colors), 17
strandColors (colors), 17
subsetByCNVprofile, 51

transCoord, 52

univariate.findCNVs, 52

variableWidthBins, 5, 10, 11, 54

writeConfig, 55

zinbinom, 55