Package 'GreyListChIP'

October 16, 2019

Title Grey Lists -- Mask Artefact Regions Based on ChIP Inputs

Type Package **Version** 1.16.0

Date 2018-01-21
Author Gord Brown <gdbzork@gmail.com></gdbzork@gmail.com>
Maintainer Gordon Brown <gordon.brown@cruk.cam.ac.uk></gordon.brown@cruk.cam.ac.uk>
Description Identify regions of ChIP experiments with high signal in the input, that lead to spurious peaks during peak calling. Remove reads aligning to these regions prior to peak calling, for cleaner ChIP analysis.
License Artistic-2.0
LazyLoad yes
Depends R (>= 3.5), methods, GenomicRanges
Imports GenomicAlignments, BSgenome, Rsamtools, rtracklayer, MASS, parallel, GenomeInfoDb, SummarizedExperiment, stats, utils
Suggests BiocStyle, BiocGenerics, RUnit
Enhances BSgenome.Hsapiens.UCSC.hg19
biocViews ChIPSeq, Alignment, Preprocessing, DifferentialPeakCalling, Sequencing, GenomeAnnotation, Coverage
git_url https://git.bioconductor.org/packages/GreyListChIP
git_branch RELEASE_3_9
git_last_commit 758bc97
git_last_commit_date 2019-05-02
Date/Publication 2019-10-15
R topics documented: calcThreshold-methods ce10.blacklist
countReads-methods
dm3.blacklist
getKaryotype-methods
grch37.blacklist
1

2 calcThreshold-methods

Index		17
	setRegions-methods	15
	mm9.blacklist	
	mm10.blacklist	14
	makeGreyList-methods	
	loadKaryotype-methods	12
	hg38.blacklist	12
	hg19.blacklist	11
	greyListBS	
	GreyList-class	
	greyList	7

calcThreshold-methods Calculate Read Count Threshold

Description

Based on the counts from countReads, sample counts from the set several times, estimate the parameters of the negative binomial distribution for each sample, then calculate the mean of the parameters (*size* and *mu*). Use these values to calculate the read count threshold, given the specified p-value threshold.

Usage

```
calcThreshold(obj, reps=100, sampleSize=30000, p=0.99,cores=1)
```

Arguments

obj A GreyList object for which to calculate the threshold.

reps The number of times to sample bins and estimate the parameters of the negative binomial distribution.

sampleSize The number of bins to sample on each repetition.

p The p-value threshold for marking bins as "grey".

cores The number of CPU cores (parallel threads) to use when sampling repeatedly

from the set of counts

Details

This method samples from the set of counts generated during the countReads step. Each sample is fitted to the negative binomial distribution, and the parameters estimated. The means of the mu and size parameters is calculated, then used to choose a read count threshold, given the p-value cutoff provided. If cores is given, the process will use that many cores to parallelize the parameter estimation.

Value

The modified GreyList object, with the threshold added.

Author(s)

Gord Brown

ce10.blacklist 3

References

Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth edition. Springer.

Examples

```
# Load a pre-built R object with counts.
data(greyList)
# Calculate the threshold:
gl <- calcThreshold(gl,reps=10,sampleSize=1000,p=0.99,cores=1)</pre>
```

ce10.blacklist

A GRanges object representing ENCODE signal artefact or "black list" regions.

Description

These objects are made from the so-called black list or signal artefact regions defined by Anshul Kundaje Alan Boyle as part of the ENCODE and modENCODE projects. They are regions which show a signal in essentially any ChIP or similar experiment. For this reason it is useful to remove reads aligning to these regions before carrying out any sort of functional genomics analysis.

These objects can be merged with GreyList objects to perform grey list and black list filtering in one step. They are included in the package as a convenience, with the permission of the creators.

Usage

ce10.blacklist

Format

An S4 GRanges object.

Value

GRanges objects defining the black list regions.

Source

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74. doi: 10.1038/nature11247. https://sites.google.com/site/anshulkundaje/projects/bl

4 countReads-methods

countReads-methods

Count reads from a BamFile

Description

Given the tiling of the genome created when the GreyList object was created (or replaced via getKaryotype), count reads overlapping the bins, in preparation for estimating the threshold for grey-listing bins.

Usage

```
countReads(obj,bamFile)
```

Arguments

obj A GreyList object on which to count reads.

bamFile A BamFile from which to count reads.

Details

This method counts reads contained within the bins that make up the genome tiling. Bins are overlapping (by default 1Kb bins at 512b intervals) so reads are counted once for each bin that wholly contains them.

Value

The modified GreyList object, with added counts.

Author(s)

Gord Brown

See Also

```
GreyList, BamFile
```

Examples

```
# Load a pre-built GreyList object.
data(greyList)

path <- system.file("extra", package="GreyListChIP")
## Not run: fn <- file.path(path, "sample_chr21.bam")
## Not run: gl <- countReads(gl,fn)</pre>
```

dm3.blacklist 5

dm3.blacklist	GRanges objects representing ENCODE signal artefact or "black list" regions.

Description

These objects are made from the so-called black list or signal artefact regions defined by Anshul Kundaje Alan Boyle as part of the ENCODE and modENCODE projects. They are regions which show a signal in essentially any ChIP or similar experiment. For this reason it is useful to remove reads aligning to these regions before carrying out any sort of functional genomics analysis.

These objects can be merged with GreyList objects to perform grey list and black list filtering in one step. They are included in the package as a convenience, with the permission of the creators.

Usage

dm3.blacklist

Format

An S4 GRanges object.

Value

GRanges objects defining the black list regions.

Source

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74. doi: 10.1038/nature11247. https://sites.google.com/site/anshulkundaje/projects/bl

```
{\tt getKaryotype-methods} \quad \textit{Replace the karyotype of a GreyList object}
```

Description

Though a BSgenome object (or a karyotype file) is supplied when the GreyList object is created, it is conceivable that the user might want to replace it. This method allows that.

Usage

```
getKaryotype(obj, genome, tileSize=1024)
```

Arguments

obj A GreyList object.

genome A BSgenome object, from which to take the karyotype.

tileSize The size in nucleotides of each tile. Overlapping tiles will be generated, spaced

at 1/2 the width of the tiles.

6 grch37.blacklist

Value

Returns the GreyList object with a new genome and tiling.

Author(s)

Gord Brown

See Also

```
GreyList, BSgenome
```

Examples

```
# Load a pre-built GreyList objct.
data(greyList)
library(BSgenome.Hsapiens.UCSC.hg19)

# Replace the karyotype, updating the genome tiling.
## Not run: gl <- getKaryotype(gl,BSgenome.Hsapiens.UCSC.hg19)</pre>
```

grch37.blacklist

GRanges objects representing ENCODE signal artefact or "black list" regions.

Description

These objects are made from the so-called black list or signal artefact regions defined by Anshul Kundaje Alan Boyle as part of the ENCODE and modENCODE projects. They are regions which show a signal in essentially any ChIP or similar experiment. For this reason it is useful to remove reads aligning to these regions before carrying out any sort of functional genomics analysis.

These objects can be merged with GreyList objects to perform grey list and black list filtering in one step. They are included in the package as a convenience, with the permission of the creators.

Usage

```
grch37.blacklist
```

Format

An S4 GRanges object.

Value

GRanges objects defining the black list regions.

Source

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome.

Nature. 2012 Sep 6;489(7414):57-74. doi: 10.1038/nature11247. https://sites.google.com/site/anshulkundaje/projects/bl

grch38.blacklist 7

grch38.blacklist	GRanges objects representing ENCODE signal artefact or "black list"
	regions.

Description

These objects are made from the so-called black list or signal artefact regions defined by Anshul Kundaje Alan Boyle as part of the ENCODE and modENCODE projects. They are regions which show a signal in essentially any ChIP or similar experiment. For this reason it is useful to remove reads aligning to these regions before carrying out any sort of functional genomics analysis.

These objects can be merged with GreyList objects to perform grey list and black list filtering in one step. They are included in the package as a convenience, with the permission of the creators.

Usage

grch38.blacklist

Format

An S4 GRanges object.

Value

GRanges objects defining the black list regions.

Source

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74. doi: 10.1038/nature11247. https://sites.google.com/site/anshulkundaje/projects/bl

greyList A sample GreyList object for use in examples.

Description

This is a sample GreyList object, covering only human chromosome 21 (from genome version hg19). The input library used to generate this grey list can be found in the European Nucleotide Archive, under accession number ERR336953.

The library was made from a culture of the MCF-7 cell line, bought from ATCC. The library was sequenced to a depth of 35,716,191 reads on an Illumina Genome Analyzer IIx. The reads were aligned to human reference genome hg19 (GRCh37) using BWA version 0.7.5a with default parameters. Approximately 96% of reads aligned to the reference genome. Reads aligning to chromosome 21 were extracted using Samtoo1s. The chromosome 21-only karyotype file was created by deleting all lines except chromosome 21, in a file generated by fetchChromSizes as described in the vignette. This package was then used to create the GreyList sample object.

When printed, the object displays several important slots in the object (if they have been filled with calculated values). For example, this object has all its slots filled, indicating that the analysis is complete:

8 GreyList-class

```
GreyList on karyotype file karyotype_chr21.txt
  tiles: 94004
  files: jc899_chr21.bam
  size (mean): 0.370332362145541
  mu (mean): 10.2330008719269
  params: reps=10, sample size=1000, p-value=0.99
```

threshold: 81 regions: 118 coverage: 4.45%

The fields are described in the class's documentation, but briefly, we can see:

- 1. the name of the karyotype file (or BSgenome object),
- 2. the number of tiles (overlapping by 1/2 the tile width),
- 3. the BAM file(s) used for read counting (currently only 1 is allowed),
- 4. the two estimated parameters of the negative binomial distribution, size and mu,
- 5. the input parameters,
- 6. the calculated read depth cutoff (over 1kb tiles),
- 7. the number of distinct regions, and
- 8. the percentage coverage of the reference genome.

The fact that all the fields are present indicates that the regions have been generated; otherwise fields still without values would be omitted. Of course any stage can be re-run with different parameters.

Usage

```
data(greyList)
```

Format

An S4 GreyList object.

Value

A GreyList object named gl.

GreyList-class

"GreyList" Objects

Description

Regions of high signal in the input samples of a ChIP experiment can lead to artefacts in peak calling. This class generates "grey lists" of such regions, for use in filtering reads before peak calling (or filtering peaks after peak calling, though it is generally safer to filter first).

Objects from the Class

Objects can be created by calls of the form new("GreyList",genome,...), where genome is a "BSgenome" object describing a genome, such as BSgenome.Hsapiens.UCSC.hg19. Alternatively, a karyotype file can be provided explictly: new("GreyList",karyoFile=fn,...). Either genome or karyoFile must be provided; if both are present, the BSgenome object takes precedence. Alternatively, an explicit list of regions may be provided as a GRanges object.

GreyList-class 9

Slots

genome: The BSgenome object corresponding with the genome the reads are aligned to

karyotype: The Seqinfo object from the BSgenome object, or made from the karyo_file

karyo_file: The name of a file containing chromosome sizes for the reference genome of interest, one per line, as "chromName chromLength" pairs.

genomeRegions: a GRanges object that defines which regions of the genome should be used to build the grey list. This is to allow the list to be built on just part of the genome.

tiles: A GRanges object with an overlapping tiling of the genome (by default 1Kb tiles every 512b)

counts: A numeric vector holding the counts corresponding to the tiling and the BAM file provided.

files: A vector of BAM filenames that were used to generate the counts (currently only accepts one).

size_param: The computed estimates of the "size" parameter of the negative binomial distribution, estimated by MASS::fitdistr from repeated sampling from the counts.

size_stderr: The standard errors of the "size" parameters, as estimated by MASS::fitdistr.

size mean: The mean of the "size" estimates.

mu_param: Computed estimates of the "mu" parameter of the negative binomial distribution, estimated by MASS::fitdistr from repeated sampling from the counts.

mu_stderr: The standard errors of the "mu" parameter.

mu_mean: The mean of the "mu" estimates.

reps: How many samples from the counts were taken.

sample_size: How many values were sampled from the counts, for each estimate of "size" and "mu".

pvalue: The requested p-value threshold.

threshold: The calculated threshold, based on the p-value.

max_gap: The largest gap to consider when merging nearby regions (i.e. if there are "grey" regions up to this many nucleotides apart, merge them into one long region).

regions: A GRanges object defining the final grey list regions.

coverage: The percentage of the genome covered by the grey list regions.

Methods

calcThreshold signature(obj = "GreyList"): Calculate the cutoff for reads in bins, based on fitting the counts to a negative binomial distribution.

countReads signature(obj = "GreyList"): Count reads in bins across the genome.

export signature(object = "GreyList", con = "character", format = "missing"): Write the
 grey list to a file.

initialize signature(.Object = "GreyList"): Create an initial object (invoked automatically by new("GreyList",...)).

loadKaryotype signature(obj = "GreyList"): Load a genome description from a file. The file
format is one line per chromosome, with the name of the chromosome followed by white
space followed by an integer indicating the length of the chromosome.

getKaryotype signature(obj = "GreyList"): Get the karyotype of a genome from a BSgenome
 object.

10 greyListBS

```
setRegions signature(obj = "GreyList"): Set the region(s) of a genome to use in making the
GreyList object.
```

makeGreyList signature(obj = "GreyList"): Compute the actual grey list, after calculating
the threshold.

```
show signature(object = "GreyList"): Display the grey list.
```

Author(s)

```
Gord Brown (<gdbzork@gmail.com>)
```

See Also

```
BSgenome, Seqinfo, GRanges
```

Examples

```
showClass("GreyList")

# Load a karyotype file:
path <- system.file("extra", package="GreyListChIP")
fn <- file.path(path, "karyotype_chr21.txt")

# Create a GreyList object:
gl <- new("GreyList", karyoFile=fn)</pre>
```

greyListBS

Construct a grey list with default arguments

Description

This function is a convenience function, wrapping several steps of grey list construction into one step. If you are content to accept the package's defaults, and are using a BSgenome object to supply the karyotype, this function might be of use to you.

Usage

```
greyListBS(genome, bam)
```

Arguments

genome a BSgenome object, for the relevant genome.

bam a BAM file to use for making the grey list.

Value

An object of class GreyList.

Author(s)

Gord Brown

hg19.blacklist

See Also

```
GreyList
```

Examples

```
# If you want to accept the defaults for everything, you can create the
# GreyList in one step using a BSgenome object:
library(BSgenome.Hsapiens.UCSC.hg19)
path <- system.file("extra", package="GreyListChIP")
## Not run: fn <- file.path(path, "sample_chr21.bam")
## Not run: gl <- greyListBS(BSgenome.Hsapiens.UCSC.hg19,fn)</pre>
```

hg19.blacklist

GRanges objects representing ENCODE signal artefact or "black list" regions.

Description

These objects are made from the so-called black list or signal artefact regions defined by Anshul Kundaje Alan Boyle as part of the ENCODE and modENCODE projects. They are regions which show a signal in essentially any ChIP or similar experiment. For this reason it is useful to remove reads aligning to these regions before carrying out any sort of functional genomics analysis.

These objects can be merged with GreyList objects to perform grey list and black list filtering in one step. They are included in the package as a convenience, with the permission of the creators.

Usage

```
hg19.blacklist
```

Format

An S4 GRanges object.

Value

GRanges objects defining the black list regions.

Source

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74. doi: 10.1038/nature11247. https://sites.google.com/site/anshulkundaje/projects/bl

hg38.blacklist	GRanges objects representing ENCODE signal artefact or "black list" regions.

Description

These objects are made from the so-called black list or signal artefact regions defined by Anshul Kundaje Alan Boyle as part of the ENCODE and modENCODE projects. They are regions which show a signal in essentially any ChIP or similar experiment. For this reason it is useful to remove reads aligning to these regions before carrying out any sort of functional genomics analysis.

These objects can be merged with GreyList objects to perform grey list and black list filtering in one step. They are included in the package as a convenience, with the permission of the creators.

Usage

hg38.blacklist

Format

An S4 GRanges object.

Value

GRanges objects defining the black list regions.

Source

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74. doi: 10.1038/nature11247. https://sites.google.com/site/anshulkundaje/projects/bl

loadKaryotype-methods Load a karyotype from a file

Description

Load a karyotype from a file, (re)generate the tiling, for a GreyList object.

Usage

loadKaryotype(obj, karyoFile, tileSize=1024)

Arguments

obj	A GreyList object to set a new karyotype for.
karyoFile	A text file describing a genome's karyotype. The format is one line per chromosome (or contig or whatever), with the name of the chromosome, some white space, and an integer giving the length of the chromosome in nucleotides.
tileSize	The width of tiles on which to count. Tiles will be placed every tileSize/2 nucleotides, to catch regions of high signal that might otherwise be split across (non-overlapping) tiles and hence missed.

Value

Returns the GreyList object with a new genome and tiling, loaded from the provided file.

Author(s)

Gord Brown

See Also

GreyList

Examples

```
# load a pre-built GreyList object:
data(greyList)

# Get a karyotype file:
path <- system.file("extra", package="GreyListChIP")
fn <- file.path(path, "karyotype_chr21.txt")

# Replace the karyotype in the GreyList:
gl <- loadKaryotype(gl,fn)</pre>
```

makeGreyList-methods General

Generate a grey list from a GreyList object

Description

Create the actual grey list, based on the threshold calculated by calcThreshold and the read counts from countReads.

Usage

```
makeGreyList(obj,maxGap=16384)
```

Arguments

obj The GreyList object to create the list for.

maxGap If the distance between neighbouring grey regions is less than or equal to maxGap,

the regions will be merged into one big region.

Details

Create the grey list as a GRanges object. Merge grey regions if they are separated by up to maxGap bp.

Value

The modified GreyList, with the regions added.

Author(s)

Gord Brown

14 mm10.blacklist

See Also

```
GreyList, GRanges
```

Examples

```
# load a pre-built GreyList object:
data(greyList)

# calculate the actual regions:
gl <- makeGreyList(gl)</pre>
```

mm10.blacklist

GRanges objects representing ENCODE signal artefact or "black list" regions.

Description

These objects are made from the so-called black list or signal artefact regions defined by Anshul Kundaje Alan Boyle as part of the ENCODE and modENCODE projects. They are regions which show a signal in essentially any ChIP or similar experiment. For this reason it is useful to remove reads aligning to these regions before carrying out any sort of functional genomics analysis.

These objects can be merged with GreyList objects to perform grey list and black list filtering in one step. They are included in the package as a convenience, with the permission of the creators.

Usage

mm10.blacklist

Format

An S4 GRanges object.

Value

GRanges objects defining the black list regions.

Source

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74. doi: 10.1038/nature11247. https://sites.google.com/site/anshulkundaje/projects/bl

mm9.blacklist

mm9.blacklist	GRanges objects representing ENCODE signal artefact or "black list" regions.
---------------	--

Description

These objects are made from the so-called black list or signal artefact regions defined by Anshul Kundaje Alan Boyle as part of the ENCODE and modENCODE projects. They are regions which show a signal in essentially any ChIP or similar experiment. For this reason it is useful to remove reads aligning to these regions before carrying out any sort of functional genomics analysis.

These objects can be merged with GreyList objects to perform grey list and black list filtering in one step. They are included in the package as a convenience, with the permission of the creators.

Usage

mm9.blacklist

Format

An S4 GRanges object.

Value

GRanges objects defining the black list regions.

Source

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74. doi: 10.1038/nature11247. https://sites.google.com/site/anshulkundaje/projects/bl

setRegions-methods Replace the set of regions of a GreyList object

Description

Though a BSgenome object (or a karyotype file, or a GRanges) is supplied when the GreyList object is created, it is conceivable that the user might want to replace it. This method allows that.

Usage

```
setRegions(obj, regions, tileSize=1024)
```

Arguments

obj A GreyList object.

regions A GRanges object, from which to take the karyotype.

tileSize The size in nucleotides of each tile. Overlapping tiles will be generated, spaced

at 1/2 the width of the tiles.

setRegions-methods

Value

Returns the GreyList object with a new tiling.

Author(s)

Gord Brown

See Also

```
GreyList, GRanges
```

Examples

```
# Load a pre-built GreyList objct.
data(greyList)

# Replace the karyotype, updating the genome tiling.
regions=GRanges(seqnames=Rle(c('chr21','chr21','chr22')),ranges=IRanges(c(1,20000,30000),end=c(10000,30000))
gl <- setRegions(gl,regions)</pre>
```

Index

*Topic classes	GRanges, $3, 5-16$
GreyList-class, 8	grch37.blacklist,6
*Topic datasets	grch38.blacklist,7
ce10.blacklist,3	GreyList, <i>3-8</i> , <i>10-16</i>
dm3.blacklist,5	<pre>GreyList (GreyList-class), 8</pre>
grch37.blacklist,6	greyList,7
grch38.blacklist,7	GreyList-class, 8
greyList,7	greyListBS, 10
hg19.blacklist, 11	
hg38.blacklist, 12	hg19.blacklist, 11
mm10.blacklist, 14	hg38.blacklist, 12
mm9.blacklist, 15	1
*Topic grey list	initialize, GreyList, BSgenome-method
greyListBS, 10	(GreyList-class), 8
*Topic methods	loadKaryotyna (loadKaryotyna-mathada)
calcThreshold-methods, 2	loadKaryotype (loadKaryotype-methods), 12
countReads-methods, 4	loadKaryotype,GreyList-method
<pre>getKaryotype-methods, 5</pre>	(loadKaryotype-methods), 12
loadKaryotype-methods, 12	loadKaryotype-methods, 12
makeGreyList-methods, 13	Todukai yotype methods, 12
setRegions-methods, 15	<pre>makeGreyList(makeGreyList-methods), 13</pre>
/	makeGreyList,GreyList-method
BamFile, 4	(makeGreyList-methods), 13
BSgenome, 5, 6, 8–10, 15	makeGreyList-methods, 13
calcThreshold(calcThreshold-methods), 2	mm10.blacklist, 14
calcThreshold, GreyList-method	mm9.blacklist, 15
(calcThreshold-methods), 2	,
calcThreshold-methods, 2	Seqinfo, <i>9</i> , <i>10</i>
ce10.blacklist, 3	setRegions(setRegions-methods), 15
countReads (countReads-methods), 4	setRegions,GreyList-method
countReads, GreyList-method	(setRegions-methods), 15
(countReads-methods), 4	setRegions-methods, 15
countReads-methods, 4	show, GreyList-method (GreyList-class), 8
countricads methods, 4	
dm3.blacklist,5	
export,GreyList,character,missing-method	
(GreyList-class), 8	
<pre>getKaryotype (getKaryotype-methods), 5</pre>	
getKaryotype,GreyList-method	
(getKaryotype-methods), 5	
<pre>getKaryotype-methods, 5</pre>	
gl (greyList), 7	