Package ‘gcappc’

April 25, 2017

Title GC Aware Peak Caller

Version 1.0.0

Author Mingxiang Teng and Rafael A. Irizarry

Maintainer Mingxiang Teng <tengmx@gmail.com>

Description Peak calling for ChIP-seq data with consideration of potential GC bias in sequencing reads. GC bias is first estimated with generalized linear mixture models using weighted GC strategy, then applied into peak significance estimation.

Depends R (>= 3.4)

Imports BiocGenerics, GenomeInfoDb, S4Vectors, IRanges, Biostrings, BSgenome, GenomicRanges, Rsamtools, GenomicAlignments, splines, grDevices, graphics, stats, methods

VignetteBuilder knitr

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

URL https://github.com/tengmx/gcappc

License GPL-3

LazyData true

biocViews Sequencing, ChIPSeq, BatchEffect, PeakDetection

RoxygenNote 6.0.1

NeedsCompilation no

R topics documented:

bindWidth ......................................................... 2
gcappcPeaks .................................................... 2
gcEffects ......................................................... 4
peaksCAT ........................................................ 5
read5endCoverage ............................................... 6

Index 8
bindWidth  

**ChIP-seq Binding Width And Peak Window Size Estimation**

**Description**

ChIP-seq experiments usually use crosslinking strategy to capture sequencing fragments. The fragment location is affected by at least but not limited to two factors, protein real binding and crosslinking operation. This function estimate size of binding part in crosslinked DNA-protein complexes, and denoted that as ChIP-seq binding width. Also, the peak detection window half size is estimated based on binding width.

**Usage**

`bindWidth(cov, range = c(50L, 500L), step = 50L, odd = TRUE)`

**Arguments**

- `cov`: A list object returned by function `readSendCoverage`.
- `range`: A non-negative integer vector with length 2. This vector set the range within which binding width and peak window size are estimated. Default `c(50, 500)` represents most ChIP-seq experiments.
- `step`: A non-negative integer to set the resolution of binding width estimation within `range`. This value will be tuned if `auto` is `TRUE`. Default `50` is based on default value of `range`.
- `odd`: A logical vector which, when `TRUE`, only allows return odd number of binding width, which is preferred by the effective GC content estimation. Default: `TRUE`.

**Value**

A numeric vector with 2 elements: Estimated binding width and half size of peak detection window.

**Examples**

```r
bam <- system.file("extdata", "chipseq.bam", package="gcapc")
cov <- readSendCoverage(bam)
bindWidth(cov)
```

---

**gcapcPeaks**  

**GC Effects Aware Peak Calling**

**Description**

This function calls ChIP-seq peaks using potential GC effects information. Enrichment scores are calculated on sliding windows of prefiltered large regions, with GC effects considered. Permutation analysis is used to determine significant binding peaks.
Usage

gcapcPeaks(cov, gcbias, bdwidth, flank = NULL, prefilter = 4L, permute = 5L, pv = 0.05, plot = FALSE, genome = "hg19", method = c("default", "tricube"))

Arguments

cov A list object returned by function read5endCoverage.
gcbias A list object returned by function gcEffects.
bdwidth A non-negative integer vector with two elements specifying ChIP-seq binding width and peak detection half window size. Usually generated by function bindWidth. A bad estimation of bdwidth results no meaning of downstream analysis.
flank A non-negative integer specifying the flanking width of ChIP-seq binding. This parameter provides the flexibility that reads appear in flankings by decreased probabilities as increased distance from binding region. This parameter helps to define effective GC content calculation. Default is NULL, which means this parameter will be calculated from bdwidth. However, if customized numbers provided, there won’t be recalculation for this parameter; instead, the 2nd elements of bdwidth will be recalculated based on flank.
prefilter A non-negative integer specifying the minimum of reads to qualify a potential binding region. Regions with total of reads from forward and reverse strands larger or equivalent to prefilter are selected for downstream analysis. Default is 4.
permute A non-negative integer specifying times of permutation to be performed. Default is 5. When whole large genome is used, such as human genome, 5 times of permutation could be enough.
pv A numeric specifying p-value cutoff for significant binding peaks. Default is 0.05.
plot A logical vector which, when TRUE (default), returns density plots of real and permutation enrichment scores.
genome A BSgenome object containing the sequences of the reference genome that was used to align the reads, or the name of this reference genome specified in a way that is accepted by the getBSgenome function defined in the BSgenome software package. In that case the corresponding BSgenome data package needs to be already installed (see ?getBSgenome in the BSgenome package for the details).
method A character vector specifying choice of method to calculate effective GC content. Default is based on uniformed fragment distribution. A more smoother method based on tricube assumption is also allowed. However, tricube should not be used if estimated peak half size is 3 times or more larger than estimated bind width. The method chosen here needs to be the same as it is when calculating gcbias.

Value

A GRanges of peaks with meta columns:
es Estimated enrichment score.
pv p-value.
Examples

```r
bam <- system.file("extdata", "chipseq.bam", package="gcapc")
cov <- read5endCoverage(bam)
bdw <- bindWidth(cov)
gcb <- gcEffects(cov, bdw, samp = 0.15)
gcapcPeaks(cov, gcb, bdw)
```

Description

GC effects are estimated based on effective GC content and reads count on genome-wide windows, using generalized linear mixture models. Genome wide windows are randomly sampled with given proportions. GC effects of background and foreground are estimated separately, while effects on background level act more likely to be the real bias.

Usage

```r
gcEffects(cov, bdwidth, flank = NULL, samp = 0.05, plot = TRUE,
gcrange = c(0.3, 0.8), mu0 = 1, mu1 = 50, p = 0.02,
converge = 0.001, emtrace = TRUE, genome = "hg19",
method = c("default", "tricube"))
```

Arguments

- **cov**: A list object returned by function `read5endCoverage`.
- **bdwidth**: A non-negative integer vector with two elements specifying ChIP-seq binding width and peak detection half window size. Usually generated by function `bindWidth`. A bad estimation of bdwidth results no meaning of downstream analysis.
- **flank**: A non-negative integer specifying the flanking width of ChIP-seq binding. This parameter provides the flexibility that reads appear in flankings by decreased probabilities as increased distance from binding region. This paramater helps to define effective GC content calculation. Default is NULL, which means this paramater will be calculated from bdwidth. However, if customized numbers provided, there won't be recalculation for this parameter; instead, the 2nd elements of bdwidth will be recalculated based on flank.
- **samp**: A numeric specifying the proportion of regions to be randomized for GC effects estimation. Default is 0.05.
- **plot**: A logical vector which, when TRUE (default), returns plots of intermediate results.
- **gcrange**: A non-negative numeric vector with length 2. This vector set the range of GC content to filter regions. For human, most regions have GC content between 0.3 and 0.8, which is set as the default. Other regions with GC content beyond this range will be ignored.
- **mu0**: A non-negative numeric initiating read count signals for background regions. This is treated as a starting value for EM algorithm. Default is 1.
peaksCAT

*mu1*  
A non-negative numeric initiating read count signals for foreground regions. This is treated as a starting value for EM algorithm. Default is 50.

*p*  
A non-negative numeric specifying the proportion of foreground regions in all estimated regions. This is treated as a starting value for EM algorithm. Default is 0.02.

*converge*  
A non-negative numeric specifying the condition of EM algorithm termination. EM algorithm stops when the ratio of log likelihood increment to whole log likelihood is less or equivalent to converge.

*emtrace*  
A logical vector which, when TRUE (default), allows to print the trace of log likelihood changes in EM iterations.

*genome*  
A BSgenome object containing the sequences of the reference genome that was used to align the reads, or the name of this reference genome specified in a way that is accepted by the getBSgenome function defined in the BSgenome software package. In that case the corresponding BSgenome data package needs to be already installed (see ?getBSgenome in the BSgenome package for the details).

*method*  
A character vector specifying choice of method to calculate effective GC content. Default is based on uniformed fragment distribution. A more smoother method based on tricube assumption is also allowed. However, tricube should be not used if estimated peak half size is 3 times or more larger than estimated bind width.

**Value**

A list of objects

<table>
<thead>
<tr>
<th>Object</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>glm0</td>
<td>Estimated generalized linear model for background GC effects.</td>
</tr>
<tr>
<td>glm1</td>
<td>Estimated generalized linear model for foreground GC effects.</td>
</tr>
<tr>
<td>mu0</td>
<td>Predicted signals by GC content if belongs to background.</td>
</tr>
<tr>
<td>mu1</td>
<td>Predicted signals by GC content if belongs to foreground.</td>
</tr>
<tr>
<td>z</td>
<td>Estimated probability of being foreground for input windows.</td>
</tr>
<tr>
<td>p</td>
<td>Estimated proportion of foreground regions.</td>
</tr>
<tr>
<td>ll</td>
<td>Converged log likelihood.</td>
</tr>
</tbody>
</table>

**Examples**

```r
bam <- system.file("extdata", "chipseq.bam", package="gcpc")
cov <- readSendCoverage(bam)
bdw <- bindWidth(cov)
gcb <- gcEffects(cov, bdw, samp = 0.15)
```

---

**Description**

Plot the consistancy between two peak lists.
Usage

```r
peaksCAT(x, y, ranks = seq(200, 20000, 50), esx = 1, esy = 1,
    add = FALSE, ...)
```

Arguments

- `x`: A GRanges of identified peaks from one method or one replicate. At least one
  meta column should be included to allow for significance ranking of peaks.
- `y`: A GRanges of identified peaks from compared method or another replicate. At
  least one meta column should be included to allow for significance ranking of
  peaks.
- `ranks`: A non-negative integer vector specifying the ranks to be used for CAT plot.
- `esx`: A non-negative integer specifying which meta column of `x` to be used to rank
  peak significance. Larger values in this column should indicate higher significance.
- `esy`: A non-negative integer specifying which meta column of `y` to be used to rank
  peak significance. Larger values in this column should indicate higher significance.
- `add`: A logical vector which, when TRUE, adds the current plotting line to existing
  plots. FALSE will generate a new plot.
- `...`: Other parameters passed to `plot` or `lines`.

Value

A CAT plot.

Examples

```r
bam <- system.file("extdata", "chipseq.bam", package="gcapc")
cov <- read5endCoverage(bam)
bdw <- bindWidth(cov)
gcb1 <- gcEffects(cov, bdw, samp=0.15, plot=FALSE)
peaks1 <- gcapcPeaks(cov, gcb1, bdw)
gcb2 <- gcEffects(cov, bdw, samp=0.2, plot=FALSE)
peaks2 <- gcapcPeaks(cov, gcb2, bdw)
peaksCAT(peaks1, peaks2, ranks=seq(100,200,5), ylim=c(0,1))
```

Description

Reads coverage in single base pair resolution using only 5-prime end of BAM file records. Coverage is reported for forward and reverse strands separately. Options for customized filtering of BAM records are provided.

Usage

```r
read5endCoverage(bam, chroms = NULL, mapq = 30L, duplicate = FALSE,
    flag = scanBamFlag(isUnmappedQuery = FALSE, isSecondaryAlignment = FALSE,
    isNotPassingQualityControls = FALSE))
```
Arguments

bam The path to a BAM file, which is sorted and indexed.

chroms NULL or a vector of chromosome names that compatible with the provided
BAM file. Reads coverage will be generated for these chromosomes. Default
(NULL) will use all chromosomes in BAM file.

mapq A non-negative integer specifying the minimum mapping quality to include.
BAM records with mapping qualities less than mapq are discarded.

duplicate A logical vector which, when FALSE (Default), returns maximum coverage of 1
for every base pair. Reads that start at the same position but on different strands
are not treated as duplicates.

flag A returned object by Rsamtools::scanBamFlag. Additional options for BAM
records filtering.

Value

A list of two objects by GenomicRanges::coverage

fwd Coverage object for forward strand.

rev Coverage object for reverse strand.

Examples

bam <- system.file("extdata", "chipseq.bam", package="gcpc")
read5endCoverage(bam)
Index

bindWidth, 2
BSgenome, 3, 5

gcapcPeaks, 2
gcEffects, 4
getBSgenome, 3, 5

peaksCAT, 5

read5endCoverage, 6