Package ‘ChIAPoP’

December 21, 2018

Title Identifying Real Chromatin Interaction Sites by the
Zero-Truncated Poisson Model with ChIA-PET Data

Version 0.99.9.6

Description An integrated analysis pipeline for identifying real chromatin interactions from ChIA-PET data (Chromatin Interaction Analysis by Paired-End Tag Sequencing). PoP stands for the Positive Poisson model, i.e., zero-truncated Poisson mode, which is used for assessing statistical significance of an observed chromatic interaction. 'ChIAPoP', by modeling random interaction sites with a Positive Poisson model, is more accurate in calculating p-value of interaction sites from ChIA-PET data than the existing models. POP processes ChIA-PET data in step-wise fashion including linker sequencing trimming, identification chimeric reads, read alignment with Bowtie <http://bowtie-bio.sourceforge.net/index.shtml>, peak calling with MACS2 <https://github.com/taoliu/MACS/>, counting number of interactions, and calculating p-values associated with each interaction.

Depends R (>= 3.4), ShortRead, GenomeInfoDb, GenomicAlignments

Imports BiocGenerics, Biostrings, GenomicRanges, IRanges, S4Vectors,
matrixStats, parallel

SystemRequirements bowtie, macs2

License Artistic License 2.0

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

VignetteBuilder knitr

Suggests BiocStyle, knitr, rmarkdown

NeedsCompilation no

Author Liang Niu [aut, cre],
Weichun Huang [aut]

Maintainer Liang Niu <niulg@ucmail.uc.edu>

R topics documented:

align_reads *align_reads*

Description

Parallel alignment of individual fastq files by bowtie read alignment tool http://bowtie-bio.sourceforge.net/index.shtml.

Usage

```
align_reads(fastqFiles, bowtie.index, bowtie.path = "bowtie")
```

Arguments

fastqFiles	A character vector of names of fastq files to be aligned with bowtie.
bowtie.index	The basename of the reference genome index to be searched by bowtie. For example, the index filename is /idxDIR/referenceIndex if the index is built with the command 'bowtie-build referencGenome.fa /idxDIR/referenceIndex'.
bowtie.path	Full bowtie command path, e.g. /tool_dir/bowtie (default: bowtie). The default assumes that the tool is in a system search path, e.g., /usr/bin/bowtie.

Details

The function generates sam files with names identical to the names of input fastq files, with the file extensions changed to sam.

Examples

```
## Not run:
# input fastq files
fqList<-c("regular_1.fastq","regular_2.fastq",
          "chimeric_1.fastq","chimeric_2.fastq");
# the basename of the hg19 index to be searched by bowtie
bowtieIdx<-"/genomesDir/hg19"
bowtieCmd<-"bowtie"
# if bowtie is not in system search paths, specify bowtie tool path,
# e.g., bowtieCmd<-"/toolDir/bowtie"
# run bowtie alignment of all input fastq files in parallel
align_reads(fqList,bowtieIdx,bowtieCmd);


## End(Not run)
```

bowtie.alignment *bowtie.alignment*

Description

Bowtie reads alignment for a single fastq file in the single-end read alignment mode.

Usage

```
bowtie.alignment(fastq, bowtie.path, bowtie.index,
                 output = gsub(".fastq$", ".sam", fastq),
                 remove.fastq = TRUE)
```

Arguments

`fastq`	Input fastq filename.
`bowtie.path`	Full bowtie command path, e.g. /usr/bin/bowtie.
`bowtie.index`	The basename of the reference genome index to be searched by bowtie. For example, the index filename is /idxDIR/referenceIndex if the index is built with the command 'bowtie-build referencGenome.fa /idxDIR/referenceIndex'.
`output`	Output sam filename. The default is the same fastq filename prefix but with sam filename extension. For example, if the input fastq filename is reads_1.fastq, then the output sam filename is reads_1.sam.
`remove.fastq`	Indicator to specify whether the input fastq file (`fastq`) should be removed or not (default TRUE).

Details

The alignment setting neither allows mismatch (-v 0) nor allows multiple alignment (-m 1). The function generates a sam file with name `output`.

call.peaks *call.peaks*

Description

Macs2 (https://github.com/taoliu/MACS) peak calling in all/selected chromosomes.

Usage

```
call.peaks(macs2.path, input.sams, peaks.file = "macs2",
           extsize = 150,  qvalue = 0.05, gsize = "hs",
           chr.include = "all", chunkSize = 1e+06)
```

Arguments

`macs2.path`	Full macs2 command path, e.g., /usr/bin/macs2.
`input.sams`	A charcter vector of input sam alignment filenames.
`peaks.file`	Prefix of output peak data file (default macs2).
`extsize`	The value that macs2 uses to extend reads in 5' to 3' direction to fix-size fragments (default: 150).
`qvalue`	The qvalue cutoff that macs2 uses to call peaks (default: 0.05).
`gsize`	The size of reference genome from which ChIA-PET data are derived. It can be either a number (of bases) or a known abbreviation of the species of genome reference, e.g, "hs" for the human genome, "mm" for the mouse genome (default: "hs", the size of the human genome).
`chr.include`	Either "all" or a character vector of names of chromosomes included for peak calling analysis (default: "all", all chromosomes are included). The chromosome names should be consistent with the input sam files.
`chunkSize`	Number of reads to be processed each time (default: 1e6).

Details

The function generates a file with filename NAME_peaks.narrowPeak, where NAME is `peaks.file`. The file is a BED6+4 format file which contains the peak locations together with peak summit, pvalue and qvalue.

`count.interactions` *count.interactions*

Description

The function creates a regular count table, a chimeric count table, and a table of anchor regions with sequencing depth.

Usage

```
count.interactions(regular.1, regular.2, chimeric.1, chimeric.2,
                   peaks.bed.file, regular.count.file,
                   chimeric.count.file, anchor.info.file,
                   distance.strand.file = NULL, anchor.length.min = NULL,
                   fragment.length = NULL, tag.length = 20,
                   ref.genome = "hg19", chunkSize = 1e+06,
                   remove.sam = TRUE, remove.peaks.bed.file = TRUE)
```

Arguments

`regular.1`	Filename of the input sam alignment file for the first read of the regular (non-chimeric) read pairs.
`regular.2`	Filename of the input sam alignment file for the second read of the regular (non-chimeric) read pairs.
`chimeric.1`	Filename of the input sam alignment file for the first read of the chimeric read pairs.

`chimeric.2` Filename of the input sam alignment file for the second read of the chimeric read pairs.

`peaks.bed.file`
Filename of the peak data file in the BED format, e.g., the NAME_peaks.narrowPeak file from macs2 (`https://github.com/taoliu/MACS`).

`regular.count.file`
Filename of the output file with counts of regular fragment pairs that overlap with pairs of anchor regions. It is a tab-delimited text file with 7 columns: chr1, start1, end1, chr2, start2, end2 and count. The first 6 columns are of chromosome locations for both anchor regions, and the last one is the corresponidng count of regular fragment pairs. The file has no header and uses 1-based chromosome coordinate system.

`chimeric.count.file`
Filename of the output file with count of chimeric fragment pairs that overlap with pairs of anchor regions. It has the same format as the file `regular.count.file`.

`anchor.info.file`
Filename of the output file with anchor regions and their sequencing bias. It is a tab-delimited text file containing anchor's chromosome location and non-self read coverage depth with 4 columns: chr, start, end, non.self.depth. The file has no header and uses 1-based chromosome coordinate system.

`distance.strand.file`
Filename of the optional input file of genomic distance and strand directions for filtered intra-chromosomal read pairs (default: NULL). This file can be generated from `process.sam`. The file is required if `anchor.length.min` and/or `fragment.length` are not provided by users.

`anchor.length.min`
Minimum length of an anchor region (default: NULL). It is used to construct anchor regions from peaks. It needs to be specified if `distance.strand.file` is not given.

`fragment.length`
Typical fragment length (default: NULL). It is used to construct anchor regions from peaks. It needs to be specified if `distance.strand.file` is not given.

`tag.length` Typical tag (non-linker part of a read) length (default: 20).

`ref.genome` Reference genome name (default: hg19). It is required for calculating distance between two anchor regions, especially for circular chromosomes/genomes. The reference name should be accepted by the GenomeInfoDb package. The current supported genome reference names are: hg38, hg19, hg18, panTro4, panTro3, panTro2, bosTau8, bosTau7, bosTau6, canFam3, canFam2, canFam1, musFur1, mm10, mm9, mm8, susScr3, susScr2, rn6, rheMac3, rheMac2, galGal4, galGal3, gasAcu1, danRer7, apiMel2, dm6, dm3, ce10, ce6, ce4, ce2, sacCer3, sacCer2.

`chunkSize` Number of reads to be processed each time (default: 1e6).

`remove.sam` Indicator to specify whether the input sam files (`regular.1`, `regular.2`, `chimeric.1`, `chimeric.2`) should be removed or not (default: TRUE).

`remove.peaks.bed.file`
Indicator to specify whether the input peak data file (`peaks.bed.file`) should be removed or not (default: TRUE).

Details

This function creates tables of counts of regular and chimeric fragment pairs that overlap with anchor region pairs, using regular and chimeric read alignments and peak calling data. It also estimates non-self read coverage depth of each anchor region. The non-self coverage depth is defined as sequencing depth-2*(number of regular fragment pairs with both fragments overlap with the anchor region).

pop.pipeline *pop.pipeline*

Description

The ChIAPoP pipeline to run the full analysis of a ChIA-PET dataset from original protocol.

Usage

```
pop.pipeline(fastq.1, fastq.2,
             bowtie.index, bowtie.path = "bowtie",
             macs2.path = "macs2", gsize = "hs",
             chr.include = "all", ref.genome = "hg19")
```

Arguments

`fastq.1` Fastq filename of the first reads.

`fastq.2` Fastq filename of the second reads.

`bowtie.index` The basename of the reference genome index to be searched by bowtie. For example, the index filename is /idxDIR/referenceIndex if the index is built with the command 'bowtie-build referencGenome.fa /idxDIR/referenceIndex'.

`bowtie.path` Full bowtie command path, e.g. /tools/bin/bowtie. The default assumes that bowtie is in the system search paths.

`macs2.path` Full macs2 command path, e.g., /tools/bin/macs2. The default assumes that macs2 is in the system search paths.

`gsize` The size of reference genome from which ChIA-PET data are derived. It can be either a number (of bases) or a known abbreviation of the species of genome reference, e.g, "hs" for the human genome, "mm" for the mouse genome (default: "hs", the size of the human genome).

`chr.include` Either "all" or a character vector of names of chromosomes included for peak calling analysis (default: "all", all chromosomes are included). The chromosome names should be consistent with the bowtie index.

`ref.genome` Reference genome name (default: hg19). It is required for calculating distance between two anchor regions. The reference name should be accepted by the GenomeInfoDb package. The current supported genome reference names are: hg38, hg19, hg18, panTro4, panTro3, panTro2, bosTau8, bosTau7, bosTau6, canFam3, canFam2, canFam1, musFur1, mm10, mm9, mm8, susScr3, susScr2, rn6, rheMac3, rheMac2, galGal4, galGal3, gasAcu1, danRer7, apiMel2, dm6, dm3, ce10, ce6, ce4, ce2, sacCer3, sacCer2.

Details

The output from this pipeline includes the main result file `result.txt`. The result file contains 9 columns: chr1, start1, end1, chr2, start2, end2, count, p.value, p.fdr. The file has a header and uses 1-based chromosome coordinate system. p.value and p.fdr are p-value and false discovery rate adjusted p-value (Benjamini-Hochberg method) of the potentially interactive pair of anchor regions. The pipeline also generates two model fitting plot files: 1) "truncated_poisson_fit.pdf" for truncated Poisson model fitting, and 2) "logistic_regression_fit.pdf" for logistic regression fitting.

Value

NULL

Examples

```
#this example takes two fastq files (read1 and read2) from a human ChIA-PET
#data, and uses the default the human genome hg19 to run the full analysis.
#This example assumes that both MacS2 and Bowite tools are in system search
#directories.
## Not run:
# fastq file of the first reads
fq1<-"input_fastq_read1.fastq";
# fastq file of the second reads
fq2<-"input_fastq_read2.fastq";
#run the full analysis
pop.pipeline(fq1,fq2)

## End(Not run)
```

`pop.test` *pop.test*

Description

ChIAPoP analysis for the ChIA-PET data from the original protocol.

Usage

```
pop.test(regular.count.file, chimeric.count.file, anchor.info.file,
         result.file, tp.fit.pdf = "truncated_poisson_fit.pdf",
         logistic.fit.pdf = "logistic_regression_fit.pdf",
         ref.genome = "hg19")
```

Arguments

`regular.count.file`
: Filename of the input file with counts of regular fragment pairs that overlap with anchor region pairs. It is a tab-delimited text file with 7 columns: chr1, start1, end1, chr2, start2, end_2, count. The first 3 columns are chromosome, start position, and end position of the first anchor region of the pair, and the next 3 columns are those of the second anchor region of the pair, and the last column is the observed count of regular fragment pairs that overlap with the pair. The file should have no header and use 1-based chromosome coordinate system. It can be generated by `count.interactions`.

`chimeric.count.file`
: Filename of the input file with counts of chimeric fragment pairs that overlap with anchor region pair. It should have the same format as `regular.count.file`. It can be generated by `count.interactions`.

`anchor.info.file`
: Filename of the intput file with anchor regions and their sequencing bias (depth). It is a tab-delimited text file containing anchor's chromosome location and sequencing bias with 4 columns: chr, start, end, sequencing bias. The file has no header and uses 1-based chromosome coordinate system. It can be generated by `count.interactions`.

`result.file`
: File name of the output result file. It is a tab-delimited text file with 9 columns: chr1, start1, end1, chr2, start2, end2, count, p.value, p.fdr. p.value and p.fdr are p-value and false discovery rate (Benjamini-Hochberg method) of the pair of anchor regions. The file has a header and uses 1-based chromosome coordinate system.

`tp.fit.pdf`
: Filename of a PDF graph file for the fitting of truncated Poisson regression for estimating the noise level of inter-chromosomal pairs. The default filename is "truncated_poisson_fit.pdf".

`logistic.fit.pdf`
: Filename of a PDF graph file for the fitting of logistic regression for estimating the noise level of intra-chromosomal pairs. The default filename is "logistic_regression_fit.pdf".

`ref.genome`
: Reference genome name (default: hg19). It is required for calculation distance between anchor regions that are on the same chromosome. The reference name should be accepted by the GenomeInfoDb package. The current supported genome reference names are: hg38, hg19, hg18, panTro4, panTro3, panTro2, bosTau8, bosTau7, bosTau6, canFam3, canFam2, canFam1, musFur1, mm10, mm9, mm8, susScr3, susScr2, rn6, rheMac3, rheMac2, galGal4, galGal3, gasAcu1, danRer7, apiMel2, dm6, dm3, ce10, ce6, ce4, ce2, sacCer3, sacCer2.

Details

This function estimates null-distributions, depending on by sequencing coverage bias and intra-chromosomal distance. It then uses estimated null-distributions to assess statistical significance of potentially interactive pairs of anchor regions.

Value

A data.frame that contains anchor region pairs sorted by p.value. The unsorted result is saved to the output file `result.file`.

Examples

```
## Not run:
#input file of regular reads count
regular.count.file<-"regular_count.txt"
#input file of chimeric reads count
chimeric.count.file<-"chimeric_count.txt"
#input file of anchor information
anchor.info.file<-"anchor_info.txt"
#run the statistical test of significance of potentially interactive pairs
#of anchor regions
```

```
results<-pop.test(regular.count.file,chimeric.count.file,anchor.info.file,
"results.txt")
#show top significant results
head(results)

## End(Not run)
```

pop.test.lr *pop.test.lr*

Description

ChIAPoP analysis for ChIA-PET data from new protocol (long read).

Usage

```
pop.test.lr(count.file, anchor.info.file, result.file,
            inter.fit.pdf = "inter_fit.pdf",
            intra.fit.pdf = "intra_fit.pdf",
            ref.genome = "hg19")
```

Arguments

`count.file` Filename of of the input file with counts of fragment pairs that overlap with anchor region pairs. It is a tab-delimited text file with 7 columns: chr1, start1, end1, chr2, start2, end2, count. The first 3 columns are chromosome, start position, and end position of the first anchor region of the pair, and the next 3 columns are those of the second anchor of the pair, and the last column is the observed count of fragment pairs that overlap with the pair. It should have no header and use 1-based chromosome coordinate system.

`anchor.info.file` Filename of the intput file with anchor regions and their sequencing bias (depth). It is a tab-delimited text file containing anchor's chromosome location and sequencing bias with 4 columns: chr, start, end, sequencing bias. The file should have no header and use 1-based chromosome coordinate system.

`result.file` File name of the output result file. It is a tab-delimited text file with 9 columns: chr1, start1, end1, chr2, start2, end2, count, p.value, p.fdr. p.value and p.fdr are p-value and false discovery rate (Benjamini-Hochberg method) of the pair of anchor regions. The file has a header and uses 1-based chromosome coordinate system.

`inter.fit.pdf` Filename of a PDF graph file for the fitting of logistic regression for estimating the noise level of inter-chromosomal pairs. The default filename is inter_fit.pdf.

`intra.fit.pdf` Filename of a PDF graph file for the fitting of logistic regression for estimating the noise level of intra-chromosomal pairs. The default filename is intra_fit.pdf.

`ref.genome` Reference genome name (default: hg19). It is required for calculation distance between anchor regions that are on the same chromosome. The reference name should be accepted by the GenomeInfoDb package. The current supported genome reference names are: hg38, hg19, hg18, panTro4, panTro3, panTro2, bosTau8, bosTau7, bosTau6, canFam3, canFam2, canFam1, musFur1, mm10, mm9, mm8, susScr3, susScr2, rn6, rheMac3, rheMac2, galGal4, galGal3, gasAcu1, danRer7, apiMel2, dm6, dm3, ce10, ce6, ce4, ce2, sacCer3, sacCer2.

Details

This function estimates null-distributions, depending on by sequencing coverage bias and intra-chromosomal distance. It then uses estimated null-distributions to assess statistical significance of potentially interactive pairs of anchor regions. This function is for the new long-reads ChIA-PET data, which does not generate chimeric-reads data.

Value

A data.frame that contains anchor region pairs sorted by p.value. The unsorted result is saved to the output file `result.file`.

`process.sam` *process.sam*

Description

Process the sam files to remove unaligned pairs and duplicate pairs. It also reverses the strand orientation for the reads.

Usage

```
process.sam(sam.1, sam.2,
            sam.new.1 = gsub(".sam$", "_processed.sam", sam.1),
            sam.new.2 = gsub(".sam$", "_processed.sam", sam.2),
            PET.strand.info = TRUE, remove.old.sam = TRUE,
            chunkSize = 1e+06)
```

Arguments

`sam.1` Filename of the input alignment sam file for the first reads.

`sam.2` Filename of the input alignment sam file for the second reads.

`sam.new.1` Filename of the ouput alignment sam file for the first reads, after filtering out unwanted reads from `sam.1` and reversing the read orientations (default: the same filename of input `sam.1` but with extension "_processed.sam").

`sam.new.2` Filename of the ouput alignment sam file for the second reads, after filtering out unwanted reads from `sam.2` and reversing the read orientations (default: the same filename of input `sam.2` but with extension "_processed.sam").

`PET.strand.info`
Indicator to specify whether to save strand orientations and genomic distance information for filtered intra-chromosomal read pairs to a file named PET_distance_strand.txt (default: TRUE). Possible strand orientations for read pairs are:

1. type 0: –> –>
2. type 1: <– <–
3. type 2: –> <–

4. type 3: <−−>

`remove.old.sam`
Indicator to specify whether the input sam files (`sam.1` and `sam.2`) should be removed or not (default: TRUE).

`chunkSize` Number of reads to be processed each time (default: 1e6).

Details

This function reads in two paired alignment files in the sam format, and removes read pairs if one or both reads are unmapped. It also removes PCR-duplicated read pairs. In addition, the function reverses the strand orientation of the remaining read pairs to be used for macs2 peak calling. These orientation-reversed read pairs are then saved to two new sam files designated by `sam.new.1` and `sam.new.2` (for the first and second aligned reads, respectively). When `PET.strand.info` is TRUE (e.g., when the input SAM files are for regular pairs; default), the function calculates genomic distance between the two reads, records (original) strand orientation of the two reads, for each filtered intra-chromosomal read pair, and saves the information to the file `PET_distance_strand.txt`.

`trim.linkers` *trim.linkers*

Description

Trim linkers from input ChIA-PET data.

Usage

```
trim.linkers(fl.1, fl.2,
             linkers = c("GTTGGATAAGATATC", "GTTGGAATGTATATC"),
             destination.regular = c("regular_1.fastq",
                                     "regular_2.fastq"),
             destination.chimeric = c("chimeric_1.fastq",
                                      "chimeric_2.fastq"),
             tag.end.ub = 18, mh = 1)
```

Arguments

`fl.1` Fastq filename of the first reads.

`fl.2` Fastq filename of the second reads.

`linkers` A character vector with two linker sequences. The default linkers are GTTGGATAAGATATC and GTTGGAATGTATATC.

`destination.regular`
A character vector of two new fastq filenames for regular (non-chimeric) first and second reads, respectively. The default is c("regular_1.fastq","regular_2.fastq").

`destination.chimeric`
A character vector of two new fastq filenames of chimeric first and second reads, respectively. The default is c("chimeric_1.fastq","chimeric_2.fastq").

`tag.end.ub` Minimum length of tag (i.e., non-linker) sequences (default 18). A read pair with at least one tag sequence shorter than `tag.end.ub` will be removed from further analysis.

`mh` Maximum number of mismatches allowed in searching for linker sequences.

Details

This function trims linker sequences in sequencing reads, and separates chimeric read pairs (read pairs with two different types of linker sequences) from regular read pairs (read pairs with both linker sequences of the same type) based on their linker types. The chimeric read pairs are saved to the new chmmeric fastq files designated by `destination.chimeric`, and regular read pairs are saved to the new fastq files designated by `destination.regular`.

Examples

```
fq1<-system.file("extdata", "toy_fastq_1.fastq", package = "ChIAPoP")
fq2<-system.file("extdata", "toy_fastq_2.fastq", package = "ChIAPoP")
outputDIR<-tempdir();
regularReads<-paste(outputDIR,
                    c("regular_1.fastq","regular_2.fastq"),sep="/");
chimericReads<-paste(outputDIR,
                     c("chimeric_1.fastq","chimeric_2.fastq"),sep="/");
trim.linkers(fq1,fq2,
             destination.regular=regularReads,
             destination.chimeric=chimericReads);
##remove output example files
#file.remove(c(regularReads,chimericReads));
```

Index