The *exomePeak2* user's guide

| Zhen Wei <ZhenWei@xjtlu.edu.cn> | Jia Meng <Jia-Meng@xjtlu.edu.cn> | Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, 215123, China | Institute of Integrative Biology, University of Liverpool, L7 8TX, Liverpool, United Kingdom

2019-11-27

Contents

1 Peak Calling

For peak calling of *MeRIP-Seq* experiment, exomePeak2 demands the reads alignment results in **BAM** files. Users can specify the biological replicates of the IP and input samples by a character vector of the corresponding **BAM** directories at the arguments `bam_ip` and `bam_input` separately.

In the following example, the transcript annotation is provided using GFF files. Transcript annotation can also be provided by the `TxDb` object. exomePeak2 will automatically download the TxDb if the `genome` argument is filled with the corresponding UCSC genome name.

The genome sequence is required to conduct GC content bias correction. If the `genome` argument is missing (`= NULL`), exomPeak2 will perform peak calling without correcting the GC content bias.

```
library(exomePeak2)

GENE_ANNO_GTF = system.file("extdata", "example.gtf", package="exomePeak2")

f1 = system.file("extdata", "IP1.bam", package="exomePeak2")
f2 = system.file("extdata", "IP2.bam", package="exomePeak2")
f3 = system.file("extdata", "IP3.bam", package="exomePeak2")
f4 = system.file("extdata", "IP4.bam", package="exomePeak2")
IP_BAM = c(f1,f2,f3,f4)

f1 = system.file("extdata", "Input1.bam", package="exomePeak2")
f2 = system.file("extdata", "Input2.bam", package="exomePeak2")
f3 = system.file("extdata", "Input3.bam", package="exomePeak2")
INPUT_BAM = c(f1,f2,f3)

exomePeak2(bam_ip = IP_BAM,
           bam_input = INPUT_BAM,
           gff_dir = GENE_ANNO_GTF,
           genome = "hg19",
           paired_end = FALSE)
## class: SummarizedExomePeak
## dim: 31 7
## metadata(0):
## assays(2): counts GCsizeFactors
## rownames(31): peak_11 peak_13 ... control_13 control_14
## rowData names(2): GC_content feature_length
## colnames(7): IP1.bam IP2.bam ... Input2.bam Input3.bam
## colData names(3): design_IP design_Treatment sizeFactor
```

exomePeak2 will export the modification peaks in formats of **BED** file and **CSV** table, the data will be saved automatically under a folder named by `exomePeak2_output`.

Under the default settings, the peak statistics are derived from the $\beta_{i,1}$ terms in the following regression design under the **GLM (Generalized Linear Model)** developed by **DESeq2**:

$$log2(Q_{i,j}) = \beta_{i,0} + \beta_{i,1} I(\rho(j) = IP) + t_{i,j}$$

Where $Q_{i,j}$ is the expected value of reads abundence of the modification peak i under sample j. $\beta_{i,0}$ is the intercept coefficient, $\beta_{i,1}$ is the coefficient for IP/input log2 fold change, $I(\rho(j) = IP)$ is the regression covariate that is the indicator variable for the sample j being IP sample. $t_{i,j}$ is the regression offset that account for the sequencing depth variation and the GC content biases.

Explaination over the columns of the exported table:

- ***chr***: the chromosomal name of the peak.
- ***chromStart***: the start of the peak on the chromosome.
- ***chromEnd***: the end of the peak on the chromosome.
- ***name***: the unique ID of the modification peak.
- ***score***: the -log2 p value of the peak.
- ***strand***: the strand of the peak on genome.
- ***thickStart***: the start position of the peak.
- ***thickEnd***: the end position of the peak.
- ***itemRgb***: the column for the RGB encoded color in BED file visualization.
- ***blockCount***: the block (exon) number within the peak.
- ***blockSizes***: the widths of blocks.
- ***blockStarts***: the start positions of blocks.
- ***geneID***: the gene ID of the peak.
- ***ReadsCount.input***: the reads count of the input sample.
- ***ReadsCount.IP***: the reads count of the IP sample.
- ***log2FoldChange***: the estimates of IP over input log2 fold enrichment (coefficient estimates of $\beta_{i,1}$).
- ***pvalue***: the Wald test p value on the modification coefficient.
- ***padj***: the adjusted Wald test p value using BH approach.

2 Differential Modification Analysis

The code below could conduct differential modification analysis (Comparison of Two Conditions) on exon regions defined by the transcript annotation.

In differential modification mode, exomePeak2 will first perform Peak calling on exon regions using both the control and treated samples. Then, it will conduct the differential modification analysis on peaks reported from peak calling using an interactive GLM.

```
f1 = system.file("extdata", "treated_IP1.bam", package="exomePeak2")
TREATED_IP_BAM = c(f1)
f1 = system.file("extdata", "treated_Input1.bam", package="exomePeak2")
TREATED_INPUT_BAM = c(f1)

exomePeak2(bam_ip = IP_BAM,
           bam_input = INPUT_BAM,
           bam_treated_input = TREATED_INPUT_BAM,
           bam_treated_ip = TREATED_IP_BAM,
           gff_dir = GENE_ANNO_GTF,
           genome = "hg19",
           paired_end = FALSE)
## class: SummarizedExomePeak
## dim: 23 9
## metadata(0):
```

```
## assays(2): counts GCsizeFactors
## rownames(23): peak_10 peak_11 ... control_5 control_6
## rowData names(2): GC_content feature_length
## colnames(9): IP1.bam IP2.bam ... treated_IP1.bam treated_Input1.bam
## colData names(3): design_IP design_Treatment sizeFactor
```

In differential modification mode, exomePeak2 will export the differential modification peaks in formats of **BED** file and **CSV** table, the data will also be saved automatically under a folder named by `exomePeak2_output`.

The peak statistics in differential modification setting are derived from the interactive coefficient $\beta_{i,3}$ in the following regression design of the **DESeq2 GLM**:

$$log2(Q_{i,j}) = \beta_{i,0}+\beta_{i,1}I(\rho(j) = IP)+\beta_{i,2}I(\rho(j) = Treatment)+\beta_{i,3}I(\rho(j) = IP\&Treatment)+t_{i,j}$$

Explaination for the additional table columns:

- ***ModLog2FC_control***: the modification log2 fold enrichment in the control condition.
- ***ModLog2FC_treated***: the modification log2 fold enrichment in the treatment condition.
- ***DiffModLog2FC***: the log2 Fold Change estimates of differential modification (coefficient estimates of $\beta_{i,3}$).
- ***pvalue***: the Wald test p value on the differential modification coefficient.
- ***padj***: the adjusted Wald test p value using BH approach.

3 Quantification and Statistical Analysis with Single Based Modification Annotation

exomePeak2 supports the modification quantification and differential modification analysis on single based modification annotation. The modification sites with single based resolution can provide a more accurate mapping of modification locations compared with the peaks called directly from the MeRIP-seq datasets.

Some of the datasets in epitranscriptomics have a single based resolution, e.x. Data generated by the *m6A-CLIP-Seq* or *m6A-miCLIP-Seq* techniques. Reads count on the single based modification sites could provide a more accurate and consistent quantification on *MeRIP-Seq* experiments with single based annotation.

exomePeak2 will automatically initiate the mode of single based modification quantification by providing a sigle based annotation file under the argument `mod_annot`.

The single based annotation information should be provided to the exomePeak2 function in the format of a `GRanges` object.

```
f2 = system.file("extdata", "mod_annot.rds", package="exomePeak2")

MOD_ANNO_GRANGE <- readRDS(f2)

exomePeak2(bam_ip = IP_BAM,
           bam_input = INPUT_BAM,
           gff_dir = GENE_ANNO_GTF,
           genome = "hg19",
           paired_end = FALSE,
           mod_annot = MOD_ANNO_GRANGE)
## class: SummarizedExomePeak
## dim: 171 7
## metadata(0):
## assays(2): '' GCsizeFactors
## rownames(171): peak_1 peak_2 ... control_83 control_84
## rowData names(2): GC_content feature_length
## colnames(7): IP1.bam IP2.bam ... Input2.bam Input3.bam
## colData names(3): design_IP design_Treatment sizeFactor
```

In this mode, exomePeak2 will export the analysis result also in formats of **BED** file and **CSV** table, while each row of the table corresponds to the sites of the annotation `GRanges`.

4 Peak Calling and Visualization in Multiple Steps

The exomePeak2 package can achieve peak calling and peak statistics calulation with multiple functions.

1. **Check the bam files of MeRIP-seq data before peak calling.**

```
MeRIP_Seq_Alignment <- scanMeripBAM(
                         bam_ip = IP_BAM,
                         bam_input = INPUT_BAM,
                         paired_end = FALSE
                        )
```

For MeRIP-seq experiment with interactive design (contain control and treatment groups), use the following code.

```
MeRIP_Seq_Alignment <- scanMeripBAM(
    bam_ip = IP_BAM,
    bam_input = INPUT_BAM,
    bam_treated_input = TREATED_INPUT_BAM,
    bam_treated_ip = TREATED_IP_BAM,
    paired_end = FALSE
  )
```

2. **Conduct peak calling analysis on exons using the provided bam files.**

```
SummarizedExomePeaks <- exomePeakCalling(merip_bams = MeRIP_Seq_Alignment,
                                         gff_dir = GENE_ANNO_GTF,
                                         genome = "hg19")
```

Alternatively, use the following code to quantify MeRIP-seq data on single based modification annotation.

```
SummarizedExomePeaks <- exomePeakCalling(merip_bams = MeRIP_Seq_Alignment,
                                         gff_dir = GENE_ANNO_GTF,
                                         genome = "hg19",
                                         mod_annot = MOD_ANNO_GRANGE)
```

3. **Estimate size factors that are required for GC content bias correction.**

```
SummarizedExomePeaks <- normalizeGC(SummarizedExomePeaks)
```

4. **Report the statistics of modification peaks using Generalized Linear Model (GLM).**

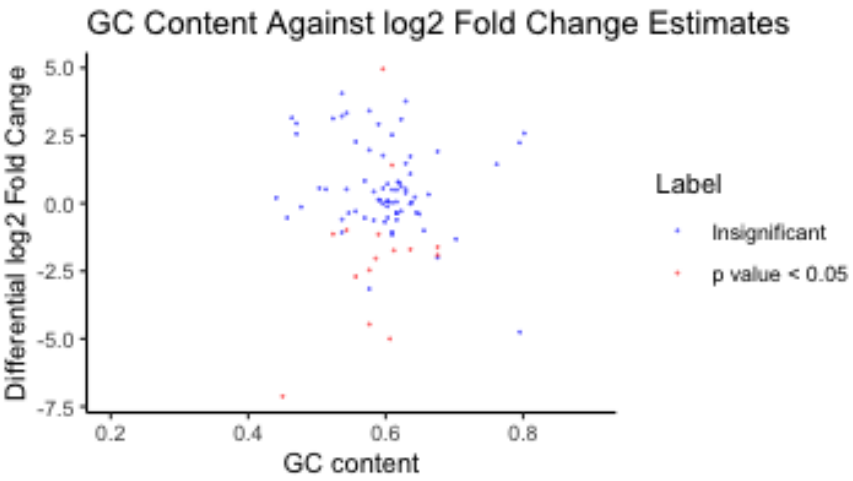
```
SummarizedExomePeaks <- glmM(SummarizedExomePeaks)
```

Alternatively, If the treated IP and input bam files are provided, `glmDM` function could be used to conduct differential modification analysis on modification Peaks with interactive GLM.

```
SummarizedExomePeaks <- glmDM(SummarizedExomePeaks)
```

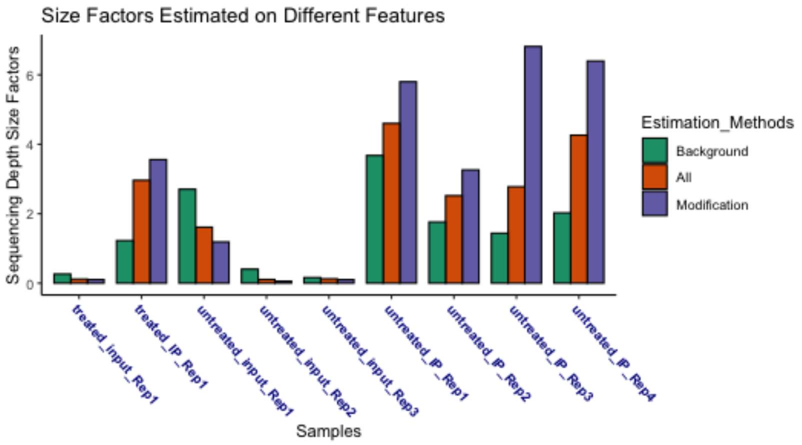
5. **Generate the scatter plot between GC content and log2 Fold Change (LFC).**

```
plotLfcGC(SummarizedExomePeaks)
```



6. **Generate the bar plot for the sequencing depth size factors.**

```
plotSizeFactors(SummarizedExomePeaks)
```



7. **Export the modification peaks and the peak statistics with user decided format.**

```
exportResults(SummarizedExomePeaks, format = "BED")
```

5 Contact

Please contact the maintainer of exomePeak2 if you have encountered any problems:

ZhenWei : zhen.wei@xjtlu.edu.cn

Please visit the github page of exomePeak2:

https://github.com/ZhenWei10/exomePeak2

6 Session Info

```
sessionInfo()
## R version 3.5.3 (2019-03-11)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Sierra 10.12.6
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRlapack.dylib
##
## locale:
## [1] zh_CN.UTF-8/zh_CN.UTF-8/zh_CN.UTF-8/C/zh_CN.UTF-8/zh_CN.UTF-8
##
## attached base packages:
##  [1] splines   parallel  stats4    stats     graphics  grDevices utils
##  [8] datasets  methods   base
##
## other attached packages:
##  [1] BSgenome.Hsapiens.UCSC.hg19_1.4.0 BSgenome_1.50.0
##  [3] rtracklayer_1.42.2                Biostrings_2.50.2
##  [5] XVector_0.22.0                    exomePeak2_0.9.9
##  [7] cqn_1.28.1                        quantreg_5.51
##  [9] SparseM_1.77                      preprocessCore_1.44.0
## [11] nor1mix_1.3-0                     mclust_5.4.5
## [13] SummarizedExperiment_1.12.0       DelayedArray_0.8.0
## [15] BiocParallel_1.16.6               matrixStats_0.54.0
## [17] Biobase_2.42.0                    GenomicRanges_1.34.0
## [19] GenomeInfoDb_1.18.2               IRanges_2.16.0
## [21] S4Vectors_0.20.1                  BiocGenerics_0.28.0
## [23] BiocStyle_2.10.0
##
## loaded via a namespace (and not attached):
##  [1] colorspace_1.4-1         htmlTable_1.13.1
##  [3] base64enc_0.1-3          rstudioapi_0.10
##  [5] MatrixModels_0.4-1       bit64_0.9-7
##  [7] AnnotationDbi_1.44.0     apeglm_1.4.2
##  [9] geneplotter_1.60.0       knitr_1.23
## [11] zeallot_0.1.0            Formula_1.2-3
## [13] Rsamtools_1.34.1         annotate_1.60.1
## [15] cluster_2.1.0            BiocManager_1.30.4
## [17] compiler_3.5.3           httr_1.4.0
## [19] backports_1.1.5          assertthat_0.2.1
## [21] Matrix_1.2-17            lazyeval_0.2.2
## [23] acepack_1.4.1            htmltools_0.3.6
## [25] prettyunits_1.0.2        tools_3.5.3
## [27] coda_0.19-3              gtable_0.3.0
## [29] glue_1.3.1               GenomeInfoDbData_1.2.0
## [31] reshape2_1.4.3           dplyr_0.8.3
## [33] Rcpp_1.0.2               bbmle_1.0.20
## [35] vctrs_0.2.0              xfun_0.8
## [37] stringr_1.4.0            XML_3.98-1.20
## [39] zlibbioc_1.28.0          MASS_7.3-51.4
## [41] scales_1.0.0             hms_0.5.0
## [43] RMariaDB_1.0.6           RColorBrewer_1.1-2
## [45] yaml_2.2.0               memoise_1.1.0
## [47] gridExtra_2.3            ggplot2_3.2.1
## [49] emdbook_1.3.11           biomaRt_2.38.0
## [51] rpart_4.1-15             latticeExtra_0.6-28
## [53] stringi_1.4.3            RSQLite_2.1.2
## [55] genefilter_1.64.0        checkmate_1.9.4
## [57] GenomicFeatures_1.34.8   rlang_0.4.1
## [59] pkgconfig_2.0.3          bitops_1.0-6
## [61] evaluate_0.14            lattice_0.20-38
## [63] purrr_0.3.2              labeling_0.3
## [65] GenomicAlignments_1.18.1 htmlwidgets_1.3
## [67] bit_1.1-14               tidyselect_0.2.5
## [69] plyr_1.8.4               magrittr_1.5
## [71] bookdown_0.12            DESeq2_1.22.2
## [73] R6_2.4.0                 Hmisc_4.2-0
## [75] DBI_1.0.0                pillar_1.4.2
## [77] foreign_0.8-71          survival_2.44-1.1
## [79] RCurl_1.95-4.12         nnet_7.3-12
## [81] tibble_2.1.3            crayon_1.3.4
## [83] rmarkdown_1.14          progress_1.2.2
## [85] locfit_1.5-9.1          grid_3.5.3
## [87] data.table_1.12.2       blob_1.2.0
## [89] digest_0.6.22           xtable_1.8-4
## [91] numDeriv_2016.8-1.1     munsell_0.5.0
```