An Introduction to *exomePeak*

Jia Meng, PhD

Modified: 21 Dec, 2015. Compiled: April 30, 2018

1 Introduction

The *exomePeak* R-package has been developed based on the MATLAB *exomePeak* package, for the analysis of RNA epitranscriptome sequencing data with affinity-based shotgun sequencing approach, such as MeRIP-Seq or m6A-Seq. The exomePeak package is under active development, please don't hesitate to contact me @ jia.meng@hotmail if you have any questions. The inputs of the main function *exomepeak* are the IP BAM files and input control BAM files with optional gene annotation file. The *exomePeak* package fullfills the following two key functions:

- Conduct peak calling to identify the RNA methylation sites under a specific condition
- Conduct peak calling to identify the RNA methylation sites and then identify the differential methylation site which can be differentially regulated at epitranscriptomic layer by RNA modifications.

Gene annotation can be provided as a GTF file, a *TxDb* object, or automatically downloaded from UCSC through the internet.

We will in the next see how the two main functions can be accomplished in a single command.

2 Peak Calling

Let us firstly load the package and get the toy data (came with the package) ready.

```
> library("exomePeak")
> gtf <- system.file("extdata", "example.gtf", package="exomePeak")
> f1 <- system.file("extdata", "IP1.bam", package="exomePeak")
> f2 <- system.file("extdata", "IP2.bam", package="exomePeak")
> f3 <- system.file("extdata", "IP3.bam", package="exomePeak")
> f4 <- system.file("extdata", "IP4.bam", package="exomePeak")
> f5 <- system.file("extdata", "Input1.bam", package="exomePeak")
```

```
> f6 <- system.file("extdata", "Input2.bam", package="exomePeak")
> f7 <- system.file("extdata", "Input3.bam", package="exomePeak")
```

The first main function of `exomePeak` R-package is to call peaks (enriched binding sites) to detect RNA methylation sites on the exome. Inputs are the gene annotation GTF file, IP and Input control samples in BAM format. This function is used when data from only one condition is available.

```
> result <- exomepeak(GENE_ANNO_GTF=gtf,
+                     IP_BAM=c(f1,f2,f3,f4),
+                     INPUT_BAM=c(f5,f6,f7))

[1] "Divide transcriptome into chr-gene-batch sections ..."
[1] "Get Reads Count ..."
[1] "This step may take a few hours ..."
[1] "100 %"
[1] "Get all the peaks ..."
[1] "Get the consistent peaks ..."
[1] "--------------------------------"
[1] "The bam files used:"
[1] "4 IP replicate(s)"
[1] "3 Input replicate(s)"
[1] "--------------------------------"
[1] "Peak calling result: "
[1] "13 peaks detected on merged data."
[1] "Please check 'peak.bed/xls' under C:/Users/biocbuild/bbs-3.7-bioc/tmpdir/RtmpUx0QgQ/Rbu
[1] "10 consistent peaks detected on every replicates. (Recommended list)"
[1] "Please check 'con_peak.bed/xls' under C:/Users/biocbuild/bbs-3.7-bioc/tmpdir/RtmpUx0QgQ

> names(result)

[1] "all_peaks" "con_peaks"
```

The results will be saved in the specified output directory, including the identified peaks and consistent peaks in BED or XLS (tab-delimited) format. The BED format can be visualized in genome browser directly and the peaks may span one or multiple introns. The difference between peak and consistent peak is that, the peaks in the con_|peak file are consitently enriched in all the IP replicates, so indicates higher re-producability.

The first 12 columns in both the BED and the XLS are the same as a standard BED12 format: http://genome.ucsc.edu/FAQ/FAQformat.html#format1

- chrom - The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).
- chromStart - The starting position of the methylation site in the chromosome or scaffold.

- chromEnd - The ending position of the RNA methylation site in the chromosome or scaffold.
- name - Defines the name of gene on which the RNA methylation site locates
- score - p-value of the peak
- strand - Defines the strand - either '+' or '-'. This is inferred based on gene annotation.
- thickStart - The same as "chromStart". The starting position of the methylation site in the chromosome or scaffold.
- thickEnd - The same as "chromEnd". The ending position of the methylation site in the chromosome or scaffold.
- itemRgb - always 0
- blockCount - The number of blocks (exons) the RNA methylation site spans.
- blockSizes - A comma-separated list of the block sizes. The number of items in this list should correspond to blockCount.
- blockStarts - A comma-separated list of block starts. All of the blockStart positions should be calculated relative to chromStart. The number of items in this list should correspond to blockCount.

The meaning of the last 3 columns of the xls file is:

- lg.p - log10(p-value) of the peak, indicating the significance of the peak as an RNA methylation site
- lg.fdr - log10(fdr) of the peak, indicating the significance of the peak as an RNA methylation site after multiple hypothesis correction
- fold_enrichment - fold enrichment within the peak in the IP sample compared with the input sample.

```
> recommended_peaks <- result$con_peaks # consistent peaks (Recommended!)
> peaks_info <- mcols(recommended_peaks) # information of the consistent peaks
> head(peaks_info)
```

```
DataFrame with 6 rows and 3 columns
       lg.p    lg.fdr fold_enrchment
  <numeric> <numeric>      <numeric>
1     -47.8     -46.5           8.05
2     -15.1       -14           9.55
3       -15     -13.9           3.78
4      -221      -219           15.5
5     -14.6     -13.6           5.81
6      -163      -161           17.5
```

or to get all the peak detected (some of them do not consistently appear on all replicates:

```
> all_peaks <- result$all_peaks # get all peaks
> peaks_info <- mcols(all_peaks) # information of all peaks
> head(peaks_info)
```

```
DataFrame with 6 rows and 3 columns
       lg.p    lg.fdr fold_enrchment
  <numeric> <numeric>      <numeric>
1      -6.9     -6.04           2.66
2     -47.8     -46.5           9.55
3       -15     -13.9           3.78
4      -221      -219           15.5
5     -14.6     -13.6           5.81
6      -163      -161           17.5
```

3 Peak Calling and Differential Methylation Analysis

When there are MeRIP-Seq data available from two experimental conditions, the `exomePeak` function may can unveil the dynamics in post-transcriptional regulation of the RNA methylome. In the following example, the function will report the sites that are post-transcriptional differentially methylated between the two tested conditions (TREATED vs. UNTREATED).

Again, let us firstly load the package and get the toy data (came with the package) ready.

```
> library("exomePeak")
> gtf <- system.file("extdata", "example.gtf", package="exomePeak")
> f1 <- system.file("extdata", "IP1.bam", package="exomePeak")
> f2 <- system.file("extdata", "IP2.bam", package="exomePeak")
> f3 <- system.file("extdata", "IP3.bam", package="exomePeak")
> f4 <- system.file("extdata", "IP4.bam", package="exomePeak")
> f5 <- system.file("extdata", "Input1.bam", package="exomePeak")
> f6 <- system.file("extdata", "Input2.bam", package="exomePeak")
> f7 <- system.file("extdata", "Input3.bam", package="exomePeak")
> f8 <- system.file("extdata", "treated_IP1.bam", package="exomePeak")
> f9 <- system.file("extdata", "treated_Input1.bam", package="exomePeak")
```

Please note that, this time we have two additional bam files obtained under a different "Treated" condition, i.e., `f8` and `f9`.

```
> result <- exomepeak(GENE_ANNO_GTF=gtf,
+                     IP_BAM=c(f1,f2,f3,f4),
+                     INPUT_BAM=c(f5,f6,f7),
```

```
+                     TREATED_IP_BAM=c(f8),
+                     TREATED_INPUT_BAM=c(f9))

[1] "Divide transcriptome into chr-gene-batch sections ..."
[1] "Get Reads Count ..."
[1] "This step may take a few hours ..."
[1] "100 %"
[1] "Comparing two conditions ..."
[1] "Get all the peaks ..."
[1] "Get the consistent peaks ..."
[1] "---------------------------------"
[1] "The bam files used:"
[1] "4 IP replicate(s)"
[1] "3 Input replicate(s)"
[1] "1 TREATED IP replicate(s)"
[1] "1 TREATED Input replicate(s)"
[1] "---------------------------------"
[1] "Peak calling and differential analysis result: "
[1] "13 peaks detected."
[1] "Please check 'diff_peak.bed/xls' under C:/Users/biocbuild/bbs-3.7-bioc/tmpdir/RtmpUx0Qg
[1] "---------------------------------"
[1] "0 significantly differential methylated peaks are detected."
[1] "Please check 'sig_diff_peak.bed/xls' under C:/Users/biocbuild/bbs-3.7-bioc/tmpdir/RtmpU
[1] "---------------------------------"
[1] "0 consistent significantly differential methylated peaks are detected.(Recommended list
[1] "Please check 'con_sig_diff_peak.bed/xls' under C:/Users/biocbuild/bbs-3.7-bioc/tmpdir/R
[1] "---------------------------------"
```

The algorithm will firstly identify reads enriched binding sites or peaks, and then check whether the sites are differentially methylated between the two experimental conditions. The results will be saved in the specified output directory, including the identified (consistent) peaks in BED and tab-delimited formats, along with the differential information indicating whether the site is hyper- or hypo-methylated under the treated condition. Similar to the peak calling case, the BED format can be visualized in genome browser directly and the peaks may span one or multiple introns.

Similar to the peak calling case, the function will report a set of consistent differentially methylated peaks (con_sig_diff_peak.xls) saved in the specified folder, which is the recommended set.

- diff_peak.xls - all the detected peaks and their differential methylation information
- sig_diff_peak.xls - all the differentially methylated peaks
- con_sig_diff_peak.xls - all the consistently differentially methylated peaks. There are peaks are consistently differentially methylated among all repli-

cates, indicating highly confidence. This set of differential methylation peaks is highly suggested.

Along with the XLS files, the matched BED files are also generated for visualization purpose.

Similar to before,

- The first 12 columns in both the BED and the XLS are the same following the standard BED12 format: http://genome.ucsc.edu/FAQ/FAQformat.html#format1 For more details, please previous section for detailed description of the first 12 columns.
- lg.p, lg.fdr, fold_enrchment are results from peak detection step, i.e., log10(pvalue), log10(fdr) and fold enrichment of the detected peak as a true methylation site. Specifically, when dealing with two experimental conditions, the fold enrichment indicates whether reads are more enriched in the pooled IP sample under both conditions than in the pooled Input sample under both conditions. The enrichment_change needs to be greater than 1 to be considered being enriched as an RNA methylation site.
- diff.lg.fdr, diff.lg.p, diff.log2.fc are results from differential methylation analysis, i.e., log10(fdr), log10(pvalue) and log2(odds ratio) of the peak as a differential methylation site between the two experimental conditions tested. If diff.log2.fc is larger than 0, the site is hypermethylated under the treated condition, otherwise if smaller than 0, it is hypomethylated under the treated condition.

The function also returns 3 GRangesList object, containing all the peaks, the differentially methylated peaks with the given threshold on the merged data, consistently differentially methylated peaks. The consistent differentially methylated peaks in the last appear to be differential for all the replicates and is thus recommended. The information of the identified peaks and the differential analysis are stored as metadata, which can be extracted.

```
> names(result)

[1] "diff_peaks"         "sig_siff_peaks"
[3] "con_sig_diff_peaks"

> is.na(result$con_sig_diff_peaks) # no reported consistent differnetial peaks

[1] TRUE
```

Unfortunately, there is no reported consistent differnetial peaks on the toy data, to get the information of all the peaks and the differential analysis information:

```
> diff_peaks <- result$diff_peaks # consistent differential peaks (Recommended!)
> peaks_info <- mcols(diff_peaks) # information of the consistent peaks
> head(peaks_info[,1:3]) # peak calling information
```

```
DataFrame with 6 rows and 3 columns
       lg.p    lg.fdr fold_enrchment
  <numeric> <numeric>      <numeric>
1     -5.91     -5.05           2.47
2     -50.7     -49.5           9.19
3     -24.7     -23.5              5
4      -222      -220           14.4
5     -15.4     -14.4           5.97
6      -171      -170           14.7
```

```
> head(peaks_info[,4:6]) # differential analysis information
```

```
DataFrame with 6 rows and 3 columns
  diff.lg.fdr diff.lg.p diff.log2.fc
    <numeric> <numeric>    <numeric>
1      -0.262    -0.402        -1.18
2      -0.262    -0.296        0.775
3      -0.274    -0.689         1.78
4       -1.28     -2.09        -1.75
5      -0.262    -0.342        -1.05
6       -1.28      -2.2        -1.45
```

4 Download Gene Annotation Directly from Internet

Gene annotation may be alternatively downloaded directly from internet, but will take a really long time due to the downloading time and huge transcriptome needed to be scanned.

```
> result <- exomepeak(GENOME="hg19",
+                     IP_BAM=c(f1,f2,f3,f4),
+                     INPUT_BAM=c(f5,f6,f7),
+                     TREATED_IP_BAM=c(f8),
+                     TREATED_INPUT_BAM=c(f9))
```

Please make sure to use the right genome assembly.

5 Handling Paired-end Reads

Unfortunately, exomePeak currently supports pair-end data in a na??ve mode, i.e., treat pair of reads as two independent reads rather than a single fragment. "treat pair of reads as two independent reads rather than a single fragment" refers to the internal process of exomePeak package, not how you align the reads. Even if the paired end data is aligned as paired end data, the pairing information will still be ignored when analyzed by exomePeak.

6 Session Information

```
> sessionInfo()

R version 3.5.0 (2018-04-23)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows Server 2012 R2 x64 (build 9600)

Matrix products: default

locale:
[1] LC_COLLATE=C
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

attached base packages:
[1] stats4    parallel  stats     graphics  grDevices
[6] utils     datasets  methods   base

other attached packages:
 [1] exomePeak_2.14.0            GenomicAlignments_1.16.0
 [3] SummarizedExperiment_1.10.0 DelayedArray_0.6.0
 [5] BiocParallel_1.14.0         matrixStats_0.53.1
 [7] rtracklayer_1.40.0          GenomicFeatures_1.32.0
 [9] AnnotationDbi_1.42.0        Biobase_2.40.0
[11] Rsamtools_1.32.0            Biostrings_2.48.0
[13] XVector_0.20.0              GenomicRanges_1.32.0
[15] GenomeInfoDb_1.16.0         IRanges_2.14.0
[17] S4Vectors_0.18.0            BiocGenerics_0.26.0

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.16           compiler_3.5.0
 [3] prettyunits_1.0.2      bitops_1.0-6
 [5] tools_3.5.0            zlibbioc_1.26.0
 [7] progress_1.1.2         biomaRt_2.36.0
 [9] digest_0.6.15          bit_1.1-12
[11] lattice_0.20-35        RSQLite_2.1.0
[13] memoise_1.1.0          pkgconfig_2.0.1
[15] Matrix_1.2-14          DBI_0.8
[17] GenomeInfoDbData_1.1.0 stringr_1.3.0
[19] httr_1.3.1             grid_3.5.0
[21] bit64_0.9-7            R6_2.2.2
[23] XML_3.98-1.11          blob_1.1.1
[25] magrittr_1.5           assertthat_0.2.0
```

```
[27] stringi_1.1.7          RCurl_1.95-4.10
```