CexoR : An R package to uncover high-resolution protein-DNA interactions in ChIP-exo replicates

Pedro Madrigal

Last Modified: June, 2015. Compiled: October 13, 2015

Wellcome Trust Sanger Institute
Hinxton, Cambridge, UK

Contents

1 Citation

Madrigal P (2015) CexoR: an R/Bioconductor package to uncover high-resolution protein-DNA interactions in ChIP-exo replicates. **EMBnet.journal** 21, e837. http://dx.doi.org/10.14806/ej.21.0.837.

2 Introduction

For its unprecedented level of resolution, chromatin immunoprecipitation combined with lambda exonuclease digestion followed by sequencing (ChIP-exo) is a potential candidate to replace ChIP-seq as the standard approach for high-confidence mapping of protein-DNA interactions. Numerous algorithms have been developed for peak calling in ChIP-seq data. However, adjusting the statistical models to ChIP-exo making use of its strand-specificity can improve the identification of protein-DNA binding sites. The midpoint between the strand-specific paired peaks formed at its forward and reverse strands is delimited by the exonuclease stop sites, within the protein binding event is located (Rhee and Pugh, 2011).

3 Methodology

Lambda exonuclease stop site (5' end of the reads) counts are calculated separately for both DNA strands from the alignment files in BAM format using the Bioconductor *Rsamtools*. Counts are then normalized using linear scaling to the same sample depth of the smaller dataset. Using the Skellam distribution (Skellam, 1946), *CexoR* models at each nucleotide position the discrete signed difference of two Poisson counts at forward and reverse strands, respectively. Then, detecting nearby located significant count differences of opposed sign (peak-pairs) at both strands allows *CexoR* to delimit the flanks of the protein binding event location at base pair resolution. A one-sided p-value is obtained for each peak using the complementary cumulative Skellam distribution function, and a final p-value for the peak-pair (default cut-off $1e-12$) is reported as the sum of the two p-values. To account for the reproducibility of replicated peak-pairs, which central point must be located at a user-defined maximum distance, p-values are submitted for irreproducible discovery rate estimation (Li et al., 2011). Stouffer's and Fisher's combined p-values are given for the final peak-pair calls. Finally, BED files containing reproducible binding event locations formed within peak-pairs are reported, as well as their midpoints.

More information can be found in Madrigal (2015).

4 Example

We downloaded the 3 replicates of human CTCF ChIP-exo data from GEO (SRA044886) (Rhee and Pugh, 2011), and aligned the reads to the human reference genome (hg19) using Bowtie 1.0.0. Reads not mapping uniquely were discarded. We can search reproducible binding events between peak-pairs in the first million bp of Chr2 in the 3 biological replicates by:

```
R> options(width=60)
R> ## hg19. chr2:1-1,000,000
R>
R> owd <- setwd(tempdir())
R> library(CexoR)
R> rep1 <- "CTCF_rep1_chr2_1-1e6.bam"
R> rep2 <- "CTCF_rep2_chr2_1-1e6.bam"
R> rep3 <- "CTCF_rep3_chr2_1-1e6.bam"
R> r1 <- system.file("extdata", rep1, package="CexoR",mustWork = TRUE)
R> r2 <- system.file("extdata", rep2, package="CexoR",mustWork = TRUE)
R> r3 <- system.file("extdata", rep3, package="CexoR",mustWork = TRUE)
R> peak_pairs <- cexor(bam=c(r1,r2,r3), chrN="chr2", chrL=1e6, idr=0.01, N=3e4)
R> peak_pairs$bindingEvents

GRanges object with 13 ranges and 6 metadata columns:
       seqnames           ranges strand   |       IDR
          <Rle>        <IRanges>  <Rle>   | <numeric>
   [1]     chr2 [ 11501,  11701]      *   |         0
   [2]     chr2 [ 18785,  18886]      *   |         0
   [3]     chr2 [142184, 142371]      *   |         0
   [4]     chr2 [172170, 172354]      *   |         0
   [5]     chr2 [332699, 332870]      *   |         0
   ...      ...               ...    ... ...       ...
   [9]     chr2 [662610, 662783]      *   |         0
  [10]     chr2 [667465, 667634]      *   |         0
  [11]     chr2 [714362, 714545]      *   |         0
  [12]     chr2 [715918, 716096]      *   |         0
  [13]     chr2 [850211, 850402]      *   |         0
      rep1.neg.log10pvalue rep2.neg.log10pvalue
                 <numeric>            <numeric>
   [1]     30.4163126907231     23.8800099446662
   [2]     17.1722792988759     23.6215053779463
```

```
  [3]     14.0090264037776     14.1269192828087
  [4]     17.1729114065705     17.3140464218041
  [5]     13.7082602274835     20.3044099969383
  ...                  ...                  ...
  [9]     30.4173675386899     34.1026478409501
 [10]      30.416312746351     44.8250721483685
 [11]     27.0319320795335     20.6462827639871
 [12]     17.1725341291709     30.6523068792823
 [13]     30.1158098266283     23.9692667994076
      rep3.neg.log10pvalue Stouffer.pvalue Fisher.pvalue
                 <numeric>       <numeric>     <numeric>
  [1]     23.5193517784223               0             0
  [2]     23.5194902250265               0             0
  [3]     20.0153779476671               0             0
  [4]     23.5194901677857               0             0
  [5]     13.9013686692951               0             0
  ...                  ...             ...           ...
  [9]     26.8375668123792               0             0
 [10]      27.015922891878               0             0
 [11]      26.616622618567               0             0
 [12]     23.6777302494668               0             0
 [13]     13.9016067615724               0             0
 -------
 seqinfo: 1 sequence from an unspecified genome
```

R> peak_pairs$bindingCentres

```
GRanges object with 13 ranges and 6 metadata columns:
       seqnames           ranges strand   |       IDR
          <Rle>        <IRanges>  <Rle>   | <numeric>
   [1]     chr2 [ 11601,  11602]      *   |         0
   [2]     chr2 [ 18836,  18837]      *   |         0
   [3]     chr2 [142278, 142279]      *   |         0
   [4]     chr2 [172262, 172263]      *   |         0
   [5]     chr2 [332784, 332785]      *   |         0
   ...      ...              ...    ... ...       ...
   [9]     chr2 [662696, 662697]      *   |         0
  [10]     chr2 [667550, 667551]      *   |         0
  [11]     chr2 [714454, 714455]      *   |         0
  [12]     chr2 [716007, 716008]      *   |         0
  [13]     chr2 [850306, 850307]      *   |         0
      rep1.neg.log10pvalue rep2.neg.log10pvalue
                 <numeric>            <numeric>
  [1]     30.4163126907231     23.8800099446662
  [2]     17.1722792988759     23.6215053779463
  [3]     14.0090264037776     14.1269192828087
  [4]     17.1729114065705     17.3140464218041
  [5]     13.7082602274835     20.3044099969383
  ...                  ...                  ...
  [9]     30.4173675386899     34.1026478409501
 [10]      30.416312746351     44.8250721483685
 [11]     27.0319320795335     20.6462827639871
 [12]     17.1725341291709     30.6523068792823
 [13]     30.1158098266283     23.9692667994076
      rep3.neg.log10pvalue Stouffer.pvalue Fisher.pvalue
                 <numeric>       <numeric>     <numeric>
  [1]     23.5193517784223               0             0
  [2]     23.5194902250265               0             0
  [3]     20.0153779476671               0             0
```

```
  [4]     23.5194901677857              0             0
  [5]     13.9013686692951              0             0
  ...                  ...            ...           ...
  [9]     26.8375668123792              0             0
 [10]      27.015922891878              0             0
 [11]      26.616622618567              0             0
 [12]     23.6777302494668              0             0
 [13]     13.9016067615724              0             0
 -------
 seqinfo: 1 sequence from an unspecified genome
```

```
R> setwd(owd)
R>
```

13 reproducible peak-pair events are reported for the established thresholds (p-value $\leq 1e-12$, IDR ≤ 0.01). We can now plot the mean profile of lambda exonuclease stop sites and reads, 500 bp around the central position of reproducible peak-pair locations, by running the function "plotcexor":

```
R> options(width=60)
R> plotcexor(bam=c(r1,r2,r3), peaks=peak_pairs, EXT=500)
```

Important notes:

- For the correct estimation of the IDR (Li et al., 2011) peak-pair calling should be relaxed (e.g., p-value=1e-3, or smaller depending on the sequencing depth), enabling the noise component be present in the data and therefore allowing the peak-pairs to be separated into a reproducible and an irreproducible groups. In the example shown above, as the dataset is very small and peaks are highly reproducible, IDR in the overlapped peak-pairs across the 3 replicates is zero.
- IDR calculation could produce varying results depending on the choice of initial estimates for four parameters needed by the algorithm (mu, sigma, rho, prop). Li et al. (2011) recommend trying several choices, so that the parameter estimation does not get trapped in a local maximum.
- For more information about using IDR in high-throughput sequencing datasets see Land et al. (2012) and Bailey et al. (2013), or the mathematical description in Li et al. (2011).

5 References

- Madrigal P (2015) CexoR: an R/Bioconductor package to uncover high-resolution protein-DNA interactions in ChIP-exo replicates. **EMBnet.journal** 21: e837.
- Bailey TL, et al. (2013). Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. **PLoS Comput Biol** 9: e1003326.
- Landt SG, et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. **Genome Res** 22: 1813-1831.
- Skellam JG (1946) The frequency distribution of the difference between two Poisson variates belonging to different populations. **J R Stat Soc Ser A** 109: 296.
- Li Q, Brown J, Huang H, Bickel P (2011) Measuring reproducibility of high-throughput experiments. **Ann Appl Stat** 5: 1752-1779.
- Rhee HS, Pugh BF (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. **Cell** 147: 1408-1419.

6 Details

This document was written using:

```
R> sessionInfo()

R version 3.2.2 (2015-08-14)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 14.04.3 LTS

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

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

other attached packages:
[1] CexoR_1.8.0         IRanges_2.4.0
[3] S4Vectors_0.8.0     BiocGenerics_0.16.0

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.1                RColorBrewer_1.1-2
 [3] futile.logger_1.4.1        GenomeInfoDb_1.6.0
 [5] plyr_1.8.3                 XVector_0.10.0
 [7] bitops_1.0-6               futile.options_1.0.0
 [9] tools_3.2.2                zlibbioc_1.16.0
[11] digest_0.6.8               gtable_0.1.2
[13] BSgenome_1.38.0            gridBase_0.4-7
[15] proto_0.3-10               rtracklayer_1.30.0
[17] stringr_1.0.0              Biostrings_2.38.0
[19] grid_3.2.2                 data.table_1.9.6
[21] impute_1.44.0              Biobase_2.30.0
[23] plotrix_3.5-12             XML_3.98-1.3
[25] BiocParallel_1.4.0         seqPattern_1.2.0
[27] genomation_1.2.0           readr_0.1.1
[29] idr_1.2                    ggplot2_1.0.1
[31] reshape2_1.4.1             lambda.r_1.1.7
[33] magrittr_1.5               matrixStats_0.14.2
[35] Rsamtools_1.22.0           scales_0.3.0
[37] GenomicAlignments_1.6.0    GenomicRanges_1.22.0
[39] MASS_7.3-44                SummarizedExperiment_1.0.0
[41] BiocStyle_1.8.0            colorspace_1.2-6
[43] KernSmooth_2.23-15         stringi_0.5-5
[45] RCurl_1.95-4.7             munsell_0.4.2
[47] chron_2.3-47
```