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

Pedro Madrigal*

July, 2013

Department of Biometry and Bioinformatics, Institute of Plant Genetics
Polish Academy of Sciences
Poznan, Poland

1 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).

2 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 package 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

*pm@engineering.com

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). Finally, BED files containing reproducible binding event locations formed within peak-pairs are reported, as well as their midpoints.

3 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 then can searched for reproducible binding events between peak-pairs in the first million bp of Chr2 in the 3 biological replicates by:

```
R> options(width=40)
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 with 13 ranges and 1 metadata column:
       seqnames           ranges strand
          <Rle>        <IRanges>  <Rle>
   [1]     chr2 [ 11501,  11701]      *
   [2]     chr2 [ 18785,  18886]      *
   [3]     chr2 [142184, 142371]      *
   [4]     chr2 [172170, 172354]      *
   [5]     chr2 [332699, 332870]      *
   ...      ...              ...    ...
   [9]     chr2 [662610, 662783]      *
  [10]     chr2 [667465, 667634]      *
  [11]     chr2 [714362, 714545]      *
  [12]     chr2 [715918, 716096]      *
  [13]     chr2 [850211, 850402]      *
          |     value
          | <numeric>
   [1]    |         0
   [2]    |         0
   [3]    |         0
   [4]    |         0
   [5]    |         0
   ... ...       ...
   [9]    |         0
  [10]    |         0
  [11]    |         0
  [12]    |         0
  [13]    |         0
  ---
  seqlengths:
      chr2
   1000000

R> peak_pairs$bindingCentres
```

```
GRanges with 13 ranges and 1 metadata column:
       seqnames           ranges strand
          <Rle>        <IRanges>  <Rle>
   [1]     chr2 [ 11601,  11602]      *
   [2]     chr2 [ 18836,  18837]      *
   [3]     chr2 [142278, 142279]      *
   [4]     chr2 [172262, 172263]      *
   [5]     chr2 [332784, 332785]      *
   ...      ...               ...    ...
   [9]     chr2 [662696, 662697]      *
  [10]     chr2 [667550, 667551]      *
  [11]     chr2 [714454, 714455]      *
  [12]     chr2 [716007, 716008]      *
  [13]     chr2 [850306, 850307]      *
          |     value
          | <numeric>
   [1]    |         0
   [2]    |         0
   [3]    |         0
   [4]    |         0
   [5]    |         0
   ... ...       ...
   [9]    |         0
  [10]    |         0
  [11]    |         0
  [12]    |         0
  [13]    |         0
  ---
  seqlengths:
      chr2
   1000000

R> setwd(owd)
R>
```

13 reproducible peak-pair events are reported for the established thresholds (p-value $\leq 1e-12$, IDR ≤ 0.01).

Important note: For the correct estimation of the IDR (Li et al., 2011) peak-pair calling should be relaxed (e.g., p-value=1e-3), enabling the noise component be present in the data, therefore allowing the peak-pairs be separated into the reproducible and the irreproducible group. 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. For more information about using IDR in high-throughput sequencing datasets see Land et al. (2012), and Bailey et al. (in press).

4 References

- Bailey TL, et al. (in press). Practical guidelines for the comprehensive analysis of ChIP-seq data. **PLoS Comput Biol**.
- 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.

- Madrigal P, et al. (in preparation).
- 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.

5 Details

This document was written using:

```
R> sessionInfo()

R version 3.0.2 (2013-09-25)
Platform: x86_64-unknown-linux-gnu (64-bit)

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

attached base packages:
[1] parallel  stats     graphics
[4] grDevices utils     datasets
[7] methods   base

other attached packages:
[1] CexoR_1.0.0
[2] IRanges_1.20.0
[3] BiocGenerics_0.8.0

loaded via a namespace (and not attached):
 [1] BSgenome_1.30.0
 [2] Biostrings_2.30.0
 [3] GenomicRanges_1.14.0
 [4] RCurl_1.95-4.1
 [5] Rsamtools_1.14.0
 [6] XML_3.98-1.1
 [7] XVector_0.2.0
 [8] bitops_1.0-6
 [9] idr_1.1.1
[10] rtracklayer_1.22.0
[11] stats4_3.0.2
[12] tools_3.0.2
[13] zlibbioc_1.8.0
```