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

Pedro Madrigal

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom

Received 26 April 2015; **Accepted** 11 June 2015; **Published** 30 July 2015

Madrigal P (2015) *EMBnet.journal* **21**, e837. http://dx.doi.org/10.14806/ej.21.0.837

Competing Interests: none

Abstract

For its unprecedented level of spatial resolution, chromatin immunoprecipitation combined with λ exonuclease digestion followed by high-throughput sequencing (ChIP-exo) has the potential to replace ChIP-seq as the standard approach for genome-wide mapping of protein-DNA interactions. In this assay, the midpoint between the strand-specific paired peaks, formed in the forward and reverse strands, is typically delimited by the exonuclease stop-sites, within which the protein-binding events are located. Although numerous algorithms have been developed for peak-calling in ChIP-seq data, none of them is fully adjusted for the analysis of ChIP-exo. This is because those statistical models do not make use of ChIP-exo's strand-specificity for the identification of protein-DNA binding sites. Here, we present the CexoR algorithm, which aims to ease the analysis of replicated ChIP-exo data in BAM alignment format. The detection algorithm relies on the Skellam distribution (cross-correlation of two Poisson distributions) to calculate probabilities of consecutive punctate-sources of read-enrichment located nearby at Watson-and-Crick strands. ChIP-exo peak-pairs are identified and ranked by their irreproducible discovery rate estimated across biological replicates, and finally reported in BED format files. CexoR can potentially be applied to other ChIP-exo-based protocols, such as ChIP-nexus.

Availability and implementation: CexoR has been implemented in R, and is freely available at http://bioconductor.org.

Introduction

Precisely mapping protein-DNA binding to genomic sites is a pivotal task in order to better understand gene regulation. Chromatin ImmunoPrecipitation (ChIP) followed by microarray hybridisation (ChIP-chip) or sequencing (ChIP-seq) have been extensively used to create maps of Transcription Factor (TF)-binding sites, comparing ChIP-seq favourably with respect to ChIP-chip in terms of resolution and signal-to-noise ratio (Ho *et al.*, 2011). Although ChIP-seq remains the standard, most-used methodology (Furey, 2012), λ exonuclease digestion followed by high-throughput sequencing (ChIP-exo) has recently emerged as a powerful and promising technique able to substitute ChIP-seq, and to circumvent its limitations (Rhee and Pugh, 2011; Mendenhall and Bernstein, 2012). In this protocol, the distribution of ChIP-exo reads is characterised by pairs of two distinct peaks, one at each DNA strand, centred at the λ exonuclease borders and separated frequently at fixed distances. Importantly, the improved resolution of ChIP-exo can provide new insights into protein-DNA interactions (Rhee and Pugh, 2011; Serandour *et al.*, 2013). Furthermore, ChIP-exo allows distinguishing weaker peaks more confidently, and also closely-located binding events that in ChIP-seq are generally deconvolved through computational approaches (*e.g.*, Guo *et al.* (2012)).

Numerous algorithms enable ChIP-seq peak-finding in biological samples considered individually (Bailey *et al.*, 2013). The peak-calling process involves the detection of single regions of significant tag enrichment. However, as underlined in Guo *et al.* (2012), common ChIP-seq peak-finders may fail to identify ChIP-exo single-base-resolution binding if the model they build is not adjusted to the actual distribution of the reads produced by this sequencing technology.

Notably, the offset of top- and bottom-strand reads observed in ChIP-seq is not present in ChIP-exo, and therefore it is not necessary to estimate insert sizes and adjust the positive- and negative-strand reads accordingly (Serandour *et al.*, 2013). For example, some ChIP-seq peak-callers do not account for strand-specific information, while others just compute strand cross-correlation to estimate the fragment length, afterwards shifting the reads with respect to the other strand (Bailey *et al.*, 2013). Software tools like GeneTrack (Albert *et al.*, 2008), GPS-GEM (Guo *et al.*, 2012), peakzilla (Bardet *et al.*, 2013) and MACS (Feng *et al.*, 2012) have been used for peak-calling in ChIP-exo data-sets. However, GeneTrack was designed with ChIP-chip and ChIP-seq in mind, thus requiring manual matching of ChIP-exo peak-pairs located nearby on opposed DNA strands (Rhee and Pugh, 2011). GEM achieved an impressive performance using positional priors based on sequence information. Nevertheless, the presence of a recognisable motif does not guarantee the true discovery of protein-DNA interactions (Bonocora *et al.*, 2013), and these priors should not be used when this premise is not valid. Therefore, non-canonical sites should not be discarded during peak calling, but after, if required for specific downstream analyses, as they might represent cooperativity of the ChIP-ed TFs with other DNA-binding proteins. Furthermore, unlike ChIP-exo, most ChIP-seq peak-calling tools are based on a comparison between a treatment sample and a negative control (which is not available for most ChIP-exo data-sets). Based on this comparison, some of them are able to provide statistical assessments in the form of *p*-values or False Discovery Rates (FDRs) based on different statistical models. As a consequence, default peak-caller stringency cut-offs can generate unreliable FDR estimations (Li *et al.*, 2011; Bailey *et al.*, 2013). Only GEM, mentioned above, and MACE (Wang *et al.*, 2014) have dedicated functionality for ChIP-exo (Zentner and Henikoff, 2014).

To address these inconveniences and allow **C**hIP-**exo** data analysis in **R**, we have developed the Bioconductor package CexoR, which searches peak boundaries in the forward and reverse strands (peak-pairs) rather than strand-agnostic regions for significant enrichment of a treatment compared to a paired negative control. These boundaries are located at the 5′ ends of the ChIP-exo aligned reads, and indicate the location of the λ exonuclease stop-sites (see graphical abstract Figure in Rhee and Pugh (2011)). CexoR is the first R package focusing exclusively on ChIP-exo peak-pair calling, including assessment of reproducibility between biological replicates, and it works without the presence of a control sample. The Irreproducible Discovery Rate (IDR) (Li *et al.*, 2011) analysis, included in the package, has been extensively used in ChIP-seq and RNA-seq data generated by the ENCODE Project (Landt *et al.*, 2012), and it is a recommended approach during ChIP-seq data analysis (Bailey *et al.*, 2013). The analysis of ChIP-exo data is very straightforward, as it only requires a single execution of the function `cexor`.

Implementation

Statistical model

The workflow of CexoR is illustrated in Figure 1.

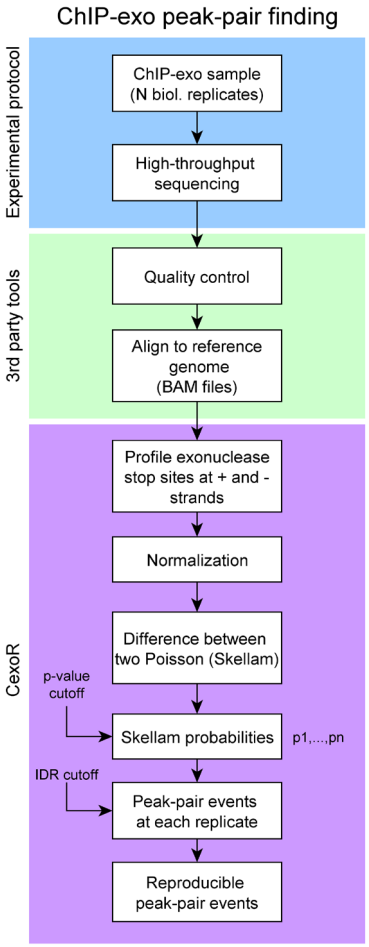

Figure 1. Workflow of ChIP-exo data analysis in R using CexoR.

λ 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 normalised using linear scaling to the sample depth of the smaller data-set. Using the Skellam distribution (Skellam, 1946), CexoR models, at each nucleotide position, the discrete signed difference of two Poisson counts with expected values μ_+ and μ_- in forward and reverse strands. We model the count difference $n_1 - n_2$ at each nucleotide of two statistically independent random variables N_1 (stop-sites in '+' strand) and N_2 (stop-sites in '-' strand), each having Poisson distribution with expected values μ_1 and μ_2. The probability mass function for the Skellam distribution for a count difference $k = n_1 - n_2$ of two Poisson-distributed variables with means μ_1 and μ_2 is given by:

$$f(k;\mu_1,\mu_2) = e^{-(\mu_1+\mu_2)}\left(\frac{\mu_1}{\mu_2}\right)^{k/2} I_k(2\sqrt{\mu_1\mu_2})$$

where k=…,-1,0,1,…, and $I_k(z)$ is the modified Bessel function of the first kind,

$$I_k(z) = \left(\frac{z}{2}\right)^k \sum_{j=0}^{\infty} \frac{\left(\frac{z^2}{4}\right)^j}{j!\Gamma(k+j+1)}$$

where $\Gamma(a)$ is the gamma function. This is done under the assumption that the λ exonuclease digests each DNA strand independently, and that digested DNA sites are random (Rhee and Pugh, 2011). Then, detecting adjacent significant count differences of opposed sign (peak-pairs) in both strands, CexoR delimits the flanks of the protein-binding events at base-pair (bp) resolution (Figure 2). The range of distances allowed between peak-pairs located in opposed strands in a replicate is user settable (parameter `dpeaks`). A one-sided *p*-value is obtained for each peak using the complementary cumulative Skellam distribution function, and a conservative *p*-value for the peak-pair (default cut-off $p \leq$ 1E-12) is reported as the sum of the two *p*-values. Then, peak-pairs across replicates, whose midpoint is located at a user-defined maximum distance (parameter `dpairs`), are selected for further analysis (Figure 2).

It is extremely important to select the parameters `dpeaks` and `dpairs` carefully, for example taking into account the expected length of the footprint of the ChIPed TF, or if the binding events typically cluster nearby along the genome. To account for the reproducibility of signal values of replicated peak-pairs, $\log_{10}$ *p*-values of each replicate are submitted for IDR analysis (Li *et al.*, 2011). Finally, the locations of reproducible binding events formed within peak-pairs are reported, as well as their midpoints. Additionally, Stouffer's and Fisher's combined *p*-values are given for the final peak-pair calls.

Installation

To install CexoR, start R and enter:

```
R> source("http://bioconductor.org/biocLite.R")
R> biocLite("CexoR")
```

Example of use

We downloaded three replicates of human CCCTC-binding factor (CTCF) ChIP-exo data from NCBI Short Read Archive accession number SRA044886 (Rhee and Pugh, 2011), and

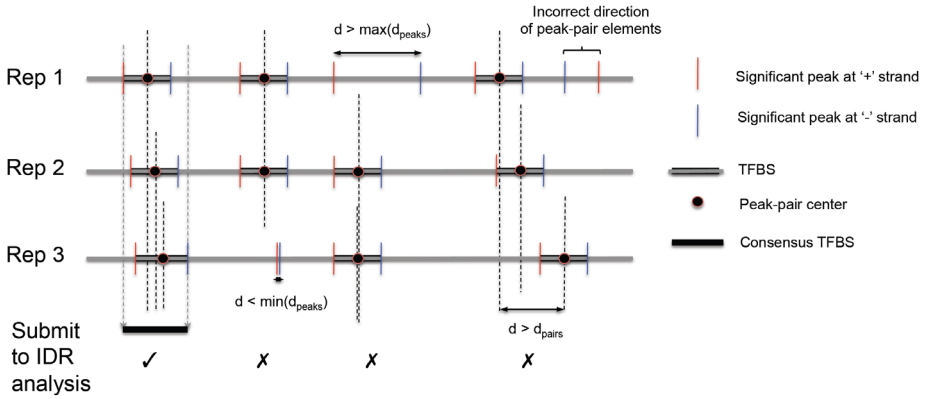

Figure 2. Illustration of the definition of ChIP-exo peak-pairs and overlap criteria between replicates.

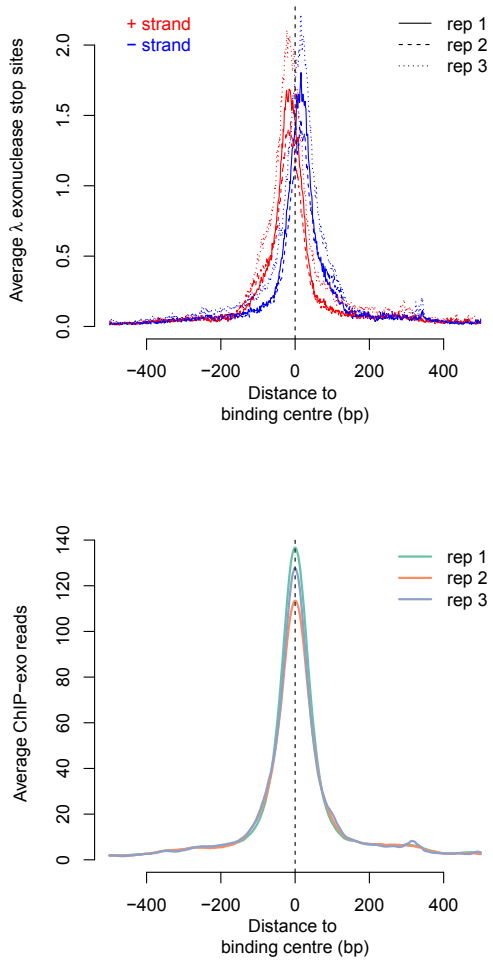

Figure 3. Graphical output of the example code used to identify 2,200 CTCF-binding sites in chromosome 1. (Top) Average λ exonuclease stop-sites. (Bottom) Average ChIP-exo profile of mapped reads. Only the final 2,200 regions are considered in the plots.

aligned the reads to the human reference genome (hg19) using Bowtie 1.0.0 (Langmead *et al.*, 2009). Reads not mapping uniquely were discarded. ChIP-exo data analysis in CexoR is straightforward, as it only requires a single execution of the function `cexor`. For example, to find TF-binding sites in the first chromosome, we run:

```
R> library(CexoR)
R> system("wget http://genome.ucsc.edu/
goldenpath/help/hg19.chrom.sizes")
R> genome <- read.table("hg19.chrom.sizes",
head=F)
R> chipexo <- cexor(bam=c('CTCF_rep1.bam',
'CTCF_rep2.bam', 'CTCF_rep3.bam'), chrN=as.
character(genome$V1[1]), chrL= genome$V2[1],
idr=0.01, p=1e-12, dpeaks=c(5,100),
dpairs=50, bedfile=TRUE)
```

We find >16,000 peak-pairs for each replicate, but only 2,200 reproducible TF-binding events after IDR analysis (p-value<1e-12; IDR<0.01)

```
R> for(i in 1:3){print(length(chipexo$paired
PeaksRepl[[i]]))}
[1] 18624
[1] 16188
[1] 20394
R> length(chipexo$bindingEvents)
[1] 2200
```

We can now plot the mean profile of λ exonuclease stop-sites and reads, 500 bp around the central position of reproducible peak-pair locations, by running the function `plotcexor`

```
R> plotcexor(bam= c('CTCF_rep1.bam', 'CTCF_
rep2.bam', 'CTCF_rep3.bam'), peaks=chipexo,
EXT=500)
```

The output is shown in Figure 3.

These visualisation plots are obtained using the Bioconductor package genomation (Akalin *et al.*, 2015).

Full details and examples are given in the manual and vignette of the package, release version 1.6.[1] and devel version 1.7.2[2].

Conclusions

Here, we present a new software package to analyse ChIP-exo data-sets. This is an alternative to the recently developed model-based analysis of ChIP-exo (MACE) (Wang *et al.*, 2014). The major differences between MACE and CexoR are: i) MACE detects peak-pairs using the Chebyshev inequality for outlier detection, making no assumption about the distribution of the coverage signal, while CexoR considers the cross-correlation of two Poisson distributions at each DNA strand; ii) MACE matches the borders using the Gale-Shapley stable matching algorithm, which performs an optimisation procedure to estimate border pair sizes, while CexoR uses a 'closest principle' to match peak-pairs within an allowed distance between significant peaks located in opposed strands in a replicate; iii) MACE incorporates an optional step of sequence-bias correction, which shows very little improvement when applied; and iv) MACE computes Shannon's entropy before border detection to consolidate a signal across multiple replicates, while CexoR runs IDR analysis across previously detected

1 bioconductor.org/packages/release/bioc/html/CexoR.html
2 bioconductor.org/packages/devel/bioc/html/CexoR.html

peak-pairs whose central positions are located at close distances in the replicates. Paired-end read information is not used by any of the packages.

In summary, the Bioconductor package CexoR is able to locate reproducible protein-DNA interactions in ChIP-exo data-sets with no need for genome sequence information, manual matching of peak-pairs, paired control data (inputs), or downstream assessment of replicate reproducibility. In addition, the R statistical environment allows integration with other pipelines and downstream analyses via other R and Bioconductor packages. We hope that our software tool will speed up the analysis of forthcoming ChIP-exo data-sets.

If the assumptions are valid (imbalance of forward- and reverse-read distribution in peak-pairs at the boundaries of a TF-binding site), the package can also be used with other next-generation sequencing data, such as ChIP-nexus (He *et al.*, 2015). It is important to note that CexoR can only be used with $\geq$ 2 samples. Further validation and benchmarks of advanced peak-detection methods will be necessary in the new generation of protocols profiling TF binding at high resolution.

Key Points

- ChIP-exo data analysis involves more complex bioinformatics than standard ChIP-seq.
- CexoR (ChIP-exo data analysis in R) is a new Bioconductor package able to locate reproducible protein-DNA interactions in ChIP-exo data-sets.
- CexoR is among the first bioinformatics tools allowing peak-pair calling in ChIP-exo, and the algorithm considers a cross-correlation of two Poisson distributions.
- CexoR could potentially be used with other sequencing data-sets, such as ChIP-nexus, and it includes functionality for the visualisation of the results.

Acknowledgements

The author would like to thank the two anonymous reviewers for their helpful comments.

References

Akalin A, Franke V, Vlahoviček K, Mason CE, Schübeler D. Genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics* **31**, 1127-1129. http://dx.doi.org/10.1093/bioinformatics/btu775

Albert I, Wachi S, Jiang C, Pugh BF (2008) GeneTrack-a genomic data processing and visualization framework. *Bioinformatics* **24**, 1305–1306. http://dx.doi.org/10.1093/bioinformatics/btn119

Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q *et al.* (2013) Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput Biol* **9**, e1003326. http://dx.doi.org/10.1371/journal.pcbi.1003326

Bardet AF, Steinmann J, Bafna S, Knoblich JA, Zeitlinger J *et al.* (2013) Identification of transcription factor binding sites from ChIP-seq data at high resolution. *Bioinformatics* **29**, 2705–2713. http://dx.doi.org/10.1093/bioinformatics/btt470

Bonocora RP, Fitzgerald DM, Stringer AM, Wade JT (2013) Non-canonical protein-DNA interaction identified by ChIP are not artifacts. *BMC Genomics* **14**, 254. http://dx.doi.org/10.1186/1471-2164-14-254

Feng J, Liu T, Qin B, Zhang Y, Liu XS (2012) Identifying ChIP-seq enrichment using MACS. *Nat Protoc* **7**, 1728–1740. http://dx.doi.org/10.1038/nprot.2012.101

Furey TS (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet* **13**, 840–852. http://dx.doi.org/10.1038/nrg3306

Guo Y, Mahony S, Gifford DK (2012) High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol* **8**, e1002638. http://dx.doi.org/10.1371/journal.pcbi.1002638

He Q, Johnston J, Zeitlinger J (2015) ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature Biotech* **33**, 395-401. http://dx.doi.org/10.1038/nbt.3121

Ho JW, Bishop E, Karchenko PV, Nègre N, White KP *et al.* (2011) ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC Genomics* **12**, 134. http://dx.doi.org/10.1186/1471-2164-12-134

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**, 1813-1831. http://dx.doi.org/10.1101/gr.136184.111

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25. http://dx.doi.org/10.1186/gb-2009-10-3-r25

Li Q, Brown JB, Huang H, Bickel PJ (2011) Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* **5**, 1751–1779. http://dx.doi.org/10.1214/11-AOAS466

Mendenhall EM, Bernstein BE (2012) DNA-protein interactions in high definition. *Genome Biol* **13**, 139. http://dx.doi.org/10.1186/gb-2012-13-1-139

Rhee HS, Pugh BF (2011) Comprehensive genome-wide protein-DNA interactions at single-nucleotide resolution. *Cell* **147**, 1408–1419. http://dx.doi.org/10.1016/j.cell.2011.11.013

Serandour AA, Brown GD, Cohen JD, Carroll JS (2013) Development of an Illumina-based ChIP-exonuclease method provides insight into FoxA1-DNA binding properties. *Genome Biol* **14**, R147. http://dx.doi.org/10.1186/gb-2013-14-12-r147

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.

Wang L, Chen J, Wang C, Uusküla-Reimand L, Chen K (2014) MACE: model based analysis of ChIP-exo, *Nucleic Acids Res* **42**, e156. http://dx.doi.org/10.1093/nar/gku846

Zentner GE, Henikoff S (2014) High-resolution digital profiling of the epigenome. *Nat Rev Genet* **15**, 814-827. http://dx.doi.org/10.1038/nrg3798