Identification of Transcription Factor Binding Sites in ChIP-exo using R/Bioconductor

Pedro Madrigal[1,2]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

[2]Wellcome Trust-MRC Cambridge Stem Cell Institute, Anne McLaren Laboratory for Regenerative Medicine, Department of Surgery, University of Cambridge, Cambridge, CB2 0SZ, UK

Email feedback to: pm12@sanger.ac.uk

Last reviewed: 19 May 2015 by Gordon Brown, Cancer Research UK Cambridge Institute, University of Cambridge

Keywords: ChIP-exo; peak calling; next generation sequencing; transcription factors

Introduction

Precisely mapping protein-DNA binding to genomic sites is a pivotal task in order to understand gene regulation. Chromatin immunoprecipitation (ChIP) followed by microarray hybridization (ChIP-chip) or sequencing (ChIP-seq) have been extensively used to map transcription factor binding sites (TFBSs), with ChIP-seq comparing favourably with respect to ChIP-chip in terms of resolution and signal-to-noise ratio (Ho et al., 2011). While ChIP-seq remains the standard, most-used methodology (Furey, 2012), λ exonuclease digestion followed by high-throughput sequencing, or 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 mapped 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 (Rhee and Pugh, 2011). Importantly, the improved resolution of ChIP-exo can provide novel insights into protein-DNA interactions (Rhee and Pugh, 2011; Serandour et al., 2013). Furthermore, ChIP-exo distinguishes weaker peaks more confidently, and also closely-located binding events, that in ChIP-seq are generally unresolved or deconvolved through computational approaches (e.g., Guo et al. (2012)).

In this protocol, first I describe the differences between ChIP-seq and ChIP-exo data analysis pipelines, and then concentrate on peak calling using the R/Bioconductor package CexoR. Unlike (for example) the popular ChIP-seq peak caller MACS (Feng et al., 2012), CexoR analyses multiple ChIP-exo replicates together, allowing a better identification of narrow peaks and simpler downstream analysis.
CexoR is able to locate reproducible protein-DNA interaction in ChIP-exo datasets with no need of 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.

The computational analysis of ChIP-exo

Figure 1 illustrates the proposed bioinformatics pipeline for ChIP-exo data analysis.

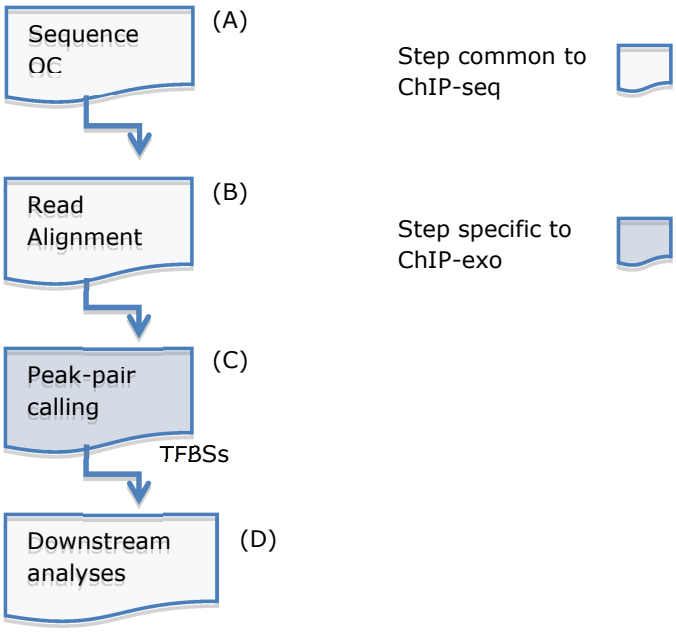

Figure 1. Workflow of ChIP-exo data analysis

While steps (A) Quality control of sequence data, (B) Read alignment, and (D) Downstream analyses, are common to ChIP-seq data analysis (see **A pipeline for ChIP-seq data analysis** (Epigenesys protocol #56), or recommended guidelines in Bailey et al., 2013)), very few software tools allow the identification of TFBSs in ChIP-exo. Only GEM (Guo et al., 2012), MACE (Wang et al., 2014), and CexoR have dedicated functionality for ChIP-exo (Zentner and Henikoff, 2014).

Peak calling in ChIP-seq vs. Peak calling in ChIP-exo

Numerous algorithms enable ChIP-seq peak finding in biological samples considered separately (Bailey et al., 2013). The peak-calling process consists of the detection of single regions of significant tag enrichment. However, as underlined in Guo et al. (2012), common ChIP-seq peak finders may fail in the identification of ChIP-exo single-base resolution binding if the statistical model considered 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, shifting afterwards the reads with respect to the other strand (Bailey et al., 2013). Software tools such as 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 datasets. However, GeneTrack was designed with ChIP-chip and ChIP-seq in mind, thus requiring a 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 recognizable motif does not guarantee the truly 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 datasets). Based on this comparison, some of them are able to provide statistical assessment in form of *p*-values or false discovery rates (FDR) 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).

Identification of transcription factor binding sites in ChIP-exo in R

To address these inconvenients, and allow ChIP-exo data analysis in R, the Bioconductor package CexoR searches peak boundaries at the forward and reverse strands (peak-pairs) rather than strand-agnostic regions of significant enrichment of a treatment compared to a paired negative control (See Note 1). 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 have 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).

Installation

To install CexoR, start R and enter:

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

R> biocLite("CexoR")
```

Example of use

ChIP-exo data analysis in CexoR is straightforward, as it only requires a single execution of the function `cexor`, e.g.:

```
R> library(CexoR)

R> chipexo <- cexor(bam=c('CTCF_rep1.bam', 'CTCF_rep2.bam',
'CTCF_rep3.bam'), chrN='chr22', chrL=51304566, idr=0.01)
```

Details of input parameter choices and output/results interpretation are given in the **manual** and **vignette** of the package at:

Release version:

http://www.bioconductor.org/packages/release/bioc/html/CexoR.html

Development version:

http://www.bioconductor.org/packages/devel/bioc/html/CexoR.html

References

Albert I, Wachi S, Jiang C, Pugh BF (2008) GeneTrack-a genomic data processing and visualization framework. *Bioinformatics* **24**, 1305–1306. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 10.1371/journal.pcbi.1002638

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. doi: 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. doi: 10.1101/gr.136184.111

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

Mendenhall EM, Bernstein BE (2012) DNA-protein interactions in high definition. *Genome Biol* **13**, 139. doi: 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. doi: 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. doi: 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. doi: 10.1093/nar/gku846

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

Author notes:

Note 1: Algorithm

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 normalized using linear scaling to the sample depth of the smaller dataset. Using the Skellam distribution (Skellam, 1946), CexoR models at each nucleotide position a discrete signed difference of two Poisson counts with expected values μ_+ and μ_- at forward and reverse strands. We model the count difference $n_1 - n_2$ of two statistically independent random variables N_1 (stop sites in '+' strand) and N_2 (stop sites in '-' strand), each one 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,

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

and Γ(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) at both strands CexoR delimits the flanks of the protein binding events at base pair resolution. 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-6) is reported as the sum of the two *p*-values. To account for the reproducibility of replicated peak-pairs, which midpoint must be located at a user-defined maximum distance, $\log_{10}$ *p*-values of each replicate are submitted for irreproducible discovery rate (IDR) analysis (Li et al., 2011). Initial estimates for the four parameters needed by IDR (mu, sigma, rho, prop) can be provided as input in the `cexor` function. Finally, the locations of reproducible binding events formed within peak-pairs are reported, as well as their midpoints. Stouffer's and Fisher's combined *p*-values are also given for the final peak-pair calls.

Reviewer comments:

Reviewed by Gordon Brown (Gordon.Brown@cruk.cam.ac.uk)
Cancer Research UK Cambridge Institute, University of Cambridge

The reviewer had three main comments:
To emphasised more clearly the benefits of CexoR.
To reconsider the use of irreproducible discovery rate (IDR) parameters.
To better describe the input parameters and output/results of the package.

Author's modifications and response to the comments are included in the submitted protocol version as well as in the updated software package.