High Resolution Identification of Protein-DNA Binding Events and Quality Control for ChIP-exo data

Rene Welch

Preliminary Examination

Department of Statistics, University of Wisconsin-Madison

December 1st, 2015

Committee Members:

Professor Sündüz Keleş, Department of Statistics, Department of Biostatistics and Medical Informatics

Professor Karl Broman, Department of Biostatistics and Medical Informatics

Professor Colin Dewey, Department of Computer Sciences, Department of Biostatistics and Medical Informatics

Professor Christina Kendziorski, Department of Biostatistics and Medical Informatics

Professor Ming Yuan, Department of Statistics

Contents

List of Figures

Abstract

Recently, ChIP-exo has been developed to investigate protein-DNA interaction in higher resolution compared to popularly used ChIP-Seq. Although ChIP-exo has drawn much attention and is considered as powerful assay, currently no systematic studies have yet been conducted to determine optimal strategies for experimental design and analysis of ChIP-exo. In order to address these questions, we evaluated diverse aspects of ChIP-exo and found the following characteristics of ChIP-exo data. First, the background of ChIP-exo data is quite different from that of ChIP-Seq data. However, sequence biases inherently present in ChIP-Seq data still exist in ChIP-exo data. Second, in ChIP-exo data, reads are located around binding sites much more tightly and hence, it has potential for high resolution identification of protein-DNA interaction sites, and also the space to allocate the reads is greatly reduced. Third, although often assumed in the ChIP-exo data analysis methods, the "peak pair" assumption does not hold well in real ChIP-exo data. Fourth, spatial resolution of ChIP-exo is comparable to that of PET ChIP-Seq and both of them are significantly better than resolution of SET ChIP-Seq. Finally, for given fixed sequencing depth, ChIP-exo provides higher sensitivity, specificity and spatial resolution than PET ChIP-Seq.

We provide a quality control pipeline which visually assess ChIP-exo biases, library complexity and enrichment; and calculates a signal-to-noise measure. Also, we updated dPeak (Chung et al., 2012 [5]), which makes a striking balance in sensitivity, specificity and spatial resolution for ChIP-exo data analysis.

1 Introduction

ChIP-exo (Chromatin Immunopecipitation followed by exonuclease digestion and next generation sequencing) Rhee and Pugh, 2011 ([18]) is the state-of-the-art experiment developed to attain single base-pair resolution of protein binding site identification and it is considered as a powerful alternative to popularly used ChIP-Seq (Chromatin Immunoprecipitation coupled with next generation sequencing) assay.

While the number of produced ChIP-exo data keeps increasing, characteristics of ChIP-exo data and optimal strategies for experimental design and analysis of ChIP-exo data are not fully investigated yet, including issues of sequence biases inherent to ChIP-exo data, choice of optimal statistical methods, and determination of optimal sequencing depth. However, currently the number of available ChIP-exo data is still limited and their sequencing depths are still insufficient for such investigation. To address this limitation we gathered ChIP-exo data from diverse organisms: CTCF factor in human [18]; ER factor in human and FoxA1 factor in mouse (Serandour et al., 2013 [20]); and generated σ^{70} factor in Escherichia coli (E. Coli) under aerobic ($+O_2$) condition, and treated by rifampicin by 0 and 20 minutes.

DNA libraries generated by the ChIP-exo protocol seem to be less complex than the libraries generated by ChIP-Seq (Mahony et al., 2015 [15]). Hence, most of current QC guidelines (Landt et al., 2012 [12]) may not be applicable on ChIP-exo, additionally to our knowledge there are not established quality control pipelines for ChIP-exo. To address this challenge, we suggest a collection of quality control visualizations to understand which biases are present in ChIP-exo data. Previous ChIP-exo analysis used ChIP-Seq samples to compare the resolution between experiments ([18], [19], [20]); Carroll et al., 2014 [3] studied the use of the Strand-Cross Correlation (SCC) (Kharchenko et al., 2008 [10]) and showed that by filtering blacklisted regions the estimation of the SCC is improved. However, this method requires to know blacklisted regions in advance which may not be available, additionally using SCC may not be useful since the peaks that are attained at the read and fragment length are confused in a typical ChIP-exo SCC curve. In our pipeline we propose two out-the-shelf analysis: an enrichment plot and the local normalized SCC cofficient.

In order to archive the potential benefits of ChIP-exo on protein binding site identification, it is critical to understand which are the important characteristics of ChIP-exo data and to use algorithms that could fully utilize information available in ChIP-exo data. Rhee and Pugh, 2011 [18] discussed that reads in the forward and reverse strand might construct peak pairs around bound protein, of which heights were implicitly assumed to be symmetric. Hence, they used the

"peak pair method" that predicts the midpoint of two modes of peak pairs as potential binding site. Mace (Wang et al., 2014 [21]), CexoR (Madrigal, 2015 [14]) and peakzilla (Bardet et al., 2013 [1]), recently developed ChIP-exo data analysis methods, are also based on this peak pair assumption. However, appropriateness of such assumption was not fully evaluated in the literature yet. Furthermore, it is still unknown which factors could affect protein binding site identification using ChIP-exo data. In order to address this problem, we investigated various aspects of ChIP-exo data by contrasting them with their respective ChIP-Seq experiments.

Currently, research on statistical methods for ChIP-exo data is still in its very early stage. Although many methods have been proposed to identify protein binding sites from ChIP-Seq data (reviewed by Wilbanks and Facciotti, 2012 [22] and Pepke and Wold, 2009 [17]), such as MACS (Zhang et al., 2008 [24]), CisGenome (Hongkai et al., 2008 [9]) and MOSAiCS (Kuan et al., 2009 [11]), these approaches reveal protein binding sites in lower resolution, i.e., at an interval of hundreds to thousands of base pairs. Furthermore, they report only one "mode" or "predicted binding location" per peak. Hence, these methods are not appropriate to evaluate the potential of ChIP-exo data for high resolution identification of protein binding sites. More recently, deconvolution algorithms such as Deconvolution (Lun et al., 2009 [13]), GEM (Guo et al., 2012 [6], an improved version of Guo et al., 2010 [7]) and PICS (Zhang et al., 2010 [23]) have been proposed to identify binding sites in higher resolution using ChIP-Seq data. However, most of them are still not tailored for ChIP-exo and PET and SET ChIP-Seq data in a unified framework and as a result, currently available methods are not appropriate for fair comparison between ChIP-exo and ChIP-Seq. To address these limitations, we developed an improved of dPeak (Chung et al., 2013 [5]), a high resolution binding site identification (deconvolution) algorithm that we previously developed for PET and SET ChIP-Seq data, so that it can also handle ChIP-exo data. The dPeak algorithm implements a probabilistic model that accurately describes the ChIP-exo and ChIP-Seq data generation process.

In this work, we demonstrate that the "peak pair" assumption of Rhee and Pugh [18] does not hold well in real ChIP-exo data. Furthermore, we found that when we analyze ChIP-exo data from eukaryotic genomes, it is important to consider sequence biases inherent to ChIP-exo data, such as mappability and GC content in order to improve sensitivity and specificity of binding site identification. We evaluated several method to identify binding events and dPeak outperforms or performs competitively respecto to GEM and MACE when analyzing ChIP-exo data. More importantly, when comparable number of reads is used for both ChIP-exo and ChIP-Seq, dPeak coupled with ChIP-exo data provides resolution comparable to PET ChIP-Seq and both significantly improve

the resolution of protein binding site identification compared to SET-based analysis with any of the available methods.

2 ChIP-exo's review

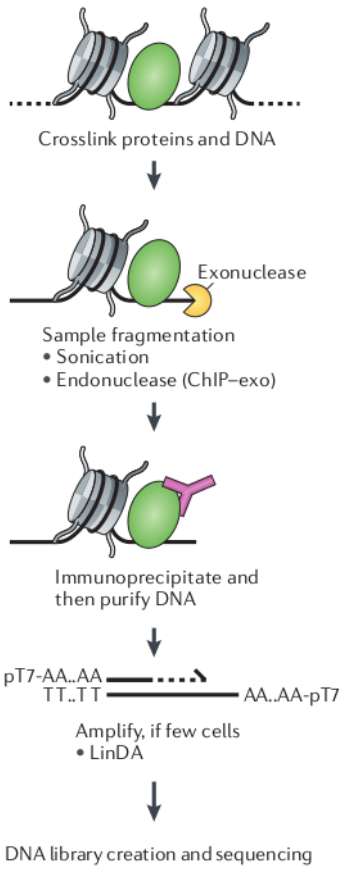

Figure 1: ChIP-exo diagram. The only difference with ChIP-Seq is the second step, where an exonuclease enzyme is added and it diggest the DNA fragments starting from the 5' end until it finds a crosslink protein (this figure is taken from Furey, 2012 [4]).

ChIP-exo is based in a modification to the ChIP-seq protocol, where an exonuclease enzyme is added and it diggests the fragments starting by the 5' ends and stops until it finds a protein that is crosslinked to the DNA.

ChIP-exo experiments first capture millions of DNA fragments (150 - 250 bp in length) that the protein under study interacts with using random fragmentation of DNA and a protein-specific antibody. Then, exonuclease is introduced to trim 5' end of each DNA fragment to a fixed distance from the bound protein. As a result, boundaries around the protein of interest constructed with 5' ends of fragments are located much closed to bound protein compared to ChIP-Seq. This step is unique to ChIP-exo and could potentially provide significantly higher spatial resolution compared to ChIP-Seq. Finally, high throughput sequencing of a small region (25 to 100 bp) at 5' end of each fragment generates millions of reads or tags.

Since this protocol is a modification to ChIP-Seq, some of its characteristics are still maintained in ChIP-exo while several new are being found. Rhee and Pugh, 2011 [18], showed that ChIP-exo generated fragments are located around the binding events more tightly than ChIP-Seq generated fragments. Hence, it is of utmost importance to understand how this new protocol works, specifically we focused on understanding which biases are mantained from ChIP-Seq to ChIP-exo, which biases are specific to ChIP-exo, how the current ChIP-Seq QC guidelines behave in ChIP-exo and we developed a new QC pipeline specific to ChIP-exo. A typical

example is shown in figure 2, which shows the σ^{70} and β'_f factors (and the β factor only on top) on a E. Coli genome's region: First, both regions show that the σ^{70} sample is enriched; second the σ^{70} sample in the ChIP-exo panel shows that its respective form in the ChIP-Seq panel may be form of more binding events; and third, either of the β factors show that the complexity of the ChIP-exo libraries may be lower in some cases, i.e. the peak is mapped with fewer positions.

2.1 Current Quality Control Measures for ChIP-Seq

Despite that not specific QC pipeline exists for ChIP-exo.

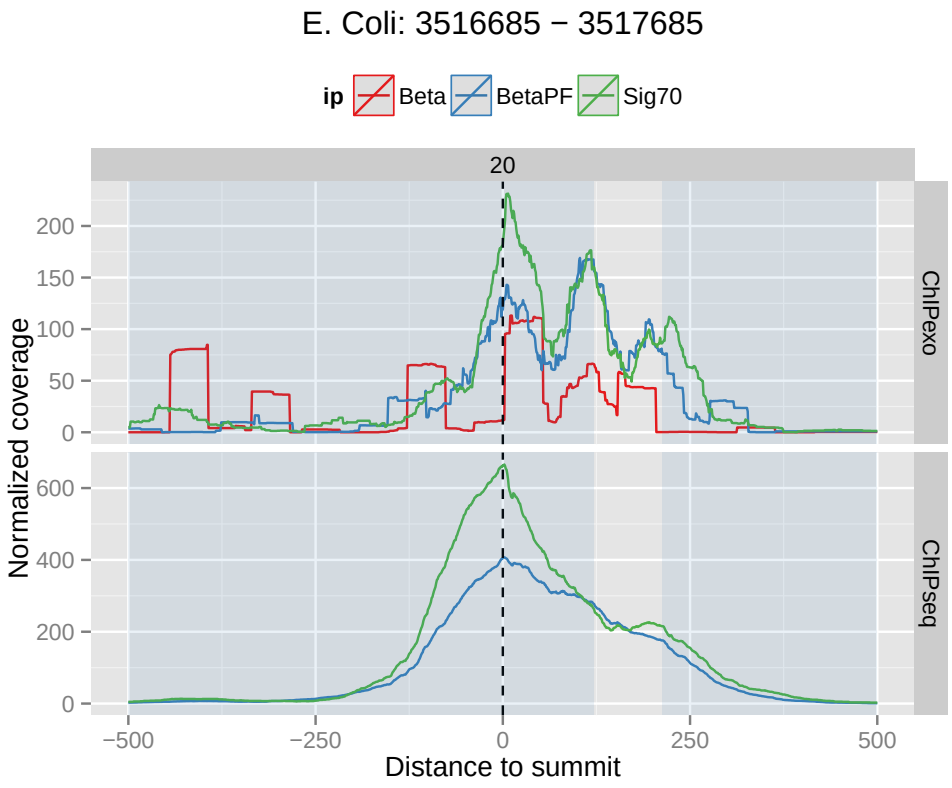

Figure 2: An example of a PET ChIP-Seq peak compared to ChIP-exo. σ^{70}, β and β'_f are subunits of the transcription initiation complex in E. Coli. Gene regions on the reverse strand are highlighted in light blue and the region was centered respect to ChIP-Seq σ^{70} summit. The library complexity is lower in ChIP-exo than in ChIP-Seq but the resolution is increased, this can be observed at the σ^{70} peak that for ChIP-exo was created as the combination of at least three ChIP-exo "sub-peaks".

3 Results

4 Planned work

References

[1] Anaïs F. Bardet, Jonas Steinmann, Sangeeta Bafna, Juergen A. Knoblich, Julia Zeitlinger, and Alexander Stark. Identification of transcription factor binding sites from chip-seq data at high resolution. *Bioinformatics*, 2013.

[2] Benjamin Bolstad, Rafael Irizarry, Magnus Âstrand, and Terence Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 2003.

[3] Thomas Carroll, Ziwei Liang, Rafik Salama, Rory Stark, and Ines de Santiago. Impact of artifact removal on chip quality metrics in chip-seq and chip-exo data. *Frontiers in Genetics, Bioinformatics and Computational Biology*, 2014.

[4] ChIP-seq, beyond: new, improved methodologies to detect, and characterize protein DNA interactions. Furey, terrence s. *Nature Reviews: Genetics*, 2012.

[5] Dongjun Chung, Dan Park, Kevin Myers, Jeffrey Grass, Patricia Kiley, Robert Landick, and Sündüz Keleş. dpeak, high resolution identification of transcription factor binding sites from pet and set chip-seq data. *PlOS, Computational Biology*, 2013.

[6] Yuchun Guo, Shaun Mahony, and David K. Gifford. High resolution genome wide binding event finding and motif discovery reals transcription factor spatial bindings constraints. *PlOS, Computational Biology*, 2012.

[7] Yuchun Guo, Georgios Papachristoudis, Robert C. Altshuler, Georg K. Gerber, TOmmi S. Jaakkola, David K. Gifford, and Shaun Mahony. Discovering homotypic binding events at high spatial resolution. *Bioinformatics*, 2010.

[8] Rafael Irizarry, Bridget Hobbs, Francois Collin, Yasmin Beazer-Barclay, Kristen Antonellis, Uwe Scherf, and Terence Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 2003.

[9] Hongkai Ji, Hui Jiang, Wenxiu Ma, David S. Johnson, Richard M. Myers, and Wing H. Wong. An integrated software system for analyzing chip-chip and chip-seq data. *Nature biotechnology*, 2008.

[10] Peter Kharchenko, Michael Tolstorukov, and Peter Park. Design and analysis of chip-seq experiments for dna-binding proteins, 2008.

[11] Pei Fei Kuan, Dongjun Chung, Guangjin Pan, James A. Thomson, Ron Stewart, and Sündüz Keleş. A statistical framework for the analysis of chip-seq data. *Journal of the American Statistical Association*, 2009.

[12] Stephen Landt, Georgi Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley Bernstein, Peter Bickel, James Brown, Philip Cayting, Yiwen Chen, Gilberto DeSalvo, Charles Epstein, Catherine Fisher-Aylor, Ghia Euskirchen, Mark Gerstein, Jason Gertz, Alexander Hartemink, Michael Hoffman, Vishwanath Iyer, Youngsook Jung, Subhradip Karmakar, Manolis Kellis, Peter Kharchenko, Qunhua Li, Tao Liu, Shirley Liu, Lijia Ma, Aleksandar Milosavljevic, Richard Myers, Peter Park, Michael Pazin, Marc Perry, Debasish Raha, Timothy Reddy, Joel Rozowsky, Noam Shoresh, Arend Sidow, Matthew Slattery, John Stamatoyannopoulos, Michael Tolstorukov, Kevin White, Simon Xi, Peggy Farnham, Jason Lieb, Barbara Wold, and Michael Snyder. *Genome Research*, 2012.

[13] Desmond S. Lun, Ashley Sherrid, Brian Weined, David R. Sherman, and James E. Galagan. A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from chip-seq data. *Genome Biology*, 2009.

[14] Pedro Madrigal. *EMBnet.journal*, 2015.

[15] Shaun Mahony and Pugh B. Franklin. Protein-dna binding in high-resolution. *Critical Reviews in Biochemistry and Molecular Biology*, 2015.

[16] Eric M. Mendenhall and Bradley E. Bernstein. Dna-protein interactions in high definition. *Genome Biology*, 2012.

[17] Shirley Pepke, Barbara Wold, and Mortazavi Ali. Computation for chip-seq nad rna-seq studies. *Nature*, 2009.

[18] Ho Sung Rhee and Franklin Pugh. Comprehensive genome-wide protein-dna interactions detected at single-nucleotide resolution. *Cell*, 2011.

[19] Ho Sung Rhee and Franklin Pugh. Chip-exo a method to identify genomic location of dna-binding proteins at near single nucleotide accuracy. *Current Protocols in Molecular Biology*, 2012.

[20] Aurelien Serandour, Brown Gordon, Joshua Cohen, and Jason Carroll. Development of and illumina-based chip-exonuclease method provides insight into foxa1-dna binding properties. *Genome Biology*, 2013.

[21] Liguo Wang, Junshend Chen, Chen Wang, Liis Uusküla-Reimand, Kaifu Chen, Alejandra Medina-Rivera, Edwin J. Young, Michael T. Zimmermann, Huihuang Yan, Zhifu Sun, Yuji Zhang, Stephen T. Wu, Haojie Huang, Michael D. Wilson, Jean-Pierre A. Kocher, and Wei Li. Mace: model based analysis of chip-exo. *Nucleic Acids Research*, 2014.

[22] Elizabeth Wilbanks and Marc Facciotti. Evaluation of algorithm performance in chip-seq peak detection. *PlOS One*, 2012.

[23] Xuekui Zhang, Gordon Robertson, Martin Krzwinski, Kaida Ning, Arnaud Droit, Steven Jones, and Raphael Gottardo. Pics: Probabilistic inference for chip-seq. *Biometrics*, 2010.

[24] Yong Zhang, Tao Liu, Clifford A. Meyer, Jérˆome Eeckhoute, David Johnson, Bradley Bernstein, Chad Nausbam, Richard M. Myers, Myles Brown, Wei Li, and X. Shirley Liu. Model-based analysis of chip-seq (macs). *Genome Biology*, 2008.