DRAFT

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

Dongjun Chung[6], Rene Welch[1], Irene Ong[3], Jeffrey Grass[3,4], Robert Landick[3,4,5] and Sündüz Keleş[1,2*]

*Correspondence: keles@stat.wisc.edu
[1]Department of Statistics, University of Wisconsin Madison, 1300 University Avenue, Madison, WI
Full list of author information is available at the end of the article

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, Background of ChIP-exo data is qute 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, hence 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.

In this article, we provide a quality control pipeline which visually asses ChIP-exo biases and calculates a signal-to-noise measure. Also, we updated dPeak [1], which makes a striking balance in sensitivity, specificity, and spatial resolution for ChIP-exo data analysis.

Keywords: ChIP-exo; QC

* to remove later

Contents

Abstract **1**

1 Background **3**

2 Results and discussion **3**

2.1 Characteristics of ChIP exo data 3

2.2 ChIP exo Quality Control pipeline 3

1 Background

ChIP-exo (Chromatin Immunopecipitation followed by exonuclease digestion and next generation sequencing) Rhee and Pugh ([2]) 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. 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 introducted 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 is the step unique to ChIP-exo that 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.

While the number of produced ChIP-exo data keeps increasing, characteristicos 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 deoth. 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 [2]; ER factor in human and FoxA1 factor in mouse from [3]; and generated σ^{70} factor in Escherichia coli (E. Coli) under aerobic ($+O_2$) condition, and treated by rifampicin by 0 and 20 minutes.

In order to archieve 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 [2] discussed that reads in the forward and reverse strand might construct peak apirs around bound protein, of which heights were implicitly assumed to be symmetric. Hence, they used the "peak pair method" taht preducts the midpoint of two modes of pak pairs as potential binding site. Apex, a recetly developed ChIP-exo data analysis method, is also based on this peak pair assumption. However, appropriatness of such assumption was not fully evaluated in the literature yet. Furthermore, it is still unknown which factors could affect protein binding site ientification 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.

2 Results and discussion

2.1 Characteristics of ChIP exo data

2.2 ChIP exo Quality Control pipeline

Methods

Author details

[1]Department of Statistics, University of Wisconsin Madison, 1300 University Avenue, Madison, WI. [2]Department of Biostatistics and Medical Informatics, University of Wisconsin Madison, 600 Highland Avenue, Madison, WI. [3] Great Lakes Bioenergy Research Center, University of Wisconsin Madison, 1552 University Avenue, Madison, WI. [4] Department of Biochemistry, University of Wisconsin Madison, 433 Babcock Drive, Madison, WI. [5] Department of Bacteriology, University of Wisconsin Madison, 1550 Linden Drive, Madison, WI. [6] Department of Public Health Sciences, Medical University of South Carolina, 135 Cannon Street, Charleston, SC.

References

1. Chung, D., Park, D., Myers, K., Grass, J., Kiley, P., Landick, R., Keleş, S.: dpeak, high resolution identification of transcription factor binding sites from pet and set chip-seq data. PlOS, Computational Biology (2013)
2. Rhee, H.S., Pugh, F.: Comprehensive genome-wide protein-dna interactions detected at single-nucleotide resolution. Cell (2011)
3. Serandour, A., Gordon, B., Cohen, J., Carroll, J.: Development of and illumina-based chip-exonuclease method provides insight into foxa1-dna binding properties. Genome Biology (2013)
4. Mendenhall, E., Bernstein, B.: Dna-protein interactions in high definition. Genome Biology (2012)