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

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.

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.

1 Introduction

2 Statistical Framework for ChIP-exo

3 Results

4 Planned work

References

[1] D. Chung, D. Park, K. Myers, J. Grass, P. Kiley, R. Landick, and S. Keleş. dpeak, high resolution identification of transcription factor binding sites from pet and set chip-seq data. *PlOS, Computational Biology*, 2013.