Novel ChIP-Seq simulation software (isChIP) for data analysis optimization

Tatiana Subkhankulova

University of Bath, Bath, UK
subkhankul@hotmail.com

Irina I. Abnizova

Welcome Trust Sanger Institute, Hinxton, Cambridge, UK
ia1@sanger.ac.uk

Rene te Boekhorst

University of Hertfordshire, Hatfield, UK

Fyodor M. Naumenko, Yuriy L. Orlov

Novosibirsk State University, Novosibirsk, Russia
fedor.naumenko@gmail.com; orlov@bionet.nsc.ru

Key words: sequencing technologies, ChIP-seq, gene expression, software

Transcriptional regulation of genes expression, chromatin dynamics and epigenetic modifications plays crucial role in normal development, complex diseases such as cancer, metabolic and cardiovascular diseases, and regeneration. Chromatin immunoprecipitation followed by next generation sequencing (ChIP-Seq) is recognized as an extremely power tool to study numerous factors interacting with DNA at a genome-wide level, shedding light on cell differentiation, proliferation or silencing mechanisms involved in environmental and pathological responses [1]. Learning about the epigenetic modifications to chromatin is crucial for establishing the link between the chromatin states and actively transcribed genes; and, therefore, the dynamic interactions between the epigenetics and gene regulatory networks.

Reads obtained from a ChIP-seq experiment are mapped to a reference genome following by reconstruction of the enrichment peaks with a peak-calling program. About a hundred peak-calling tools are currently known (MACS, PeakRanger, SICER, etc). They are based on different mathematical algorithms, program languages and utilize multiple parameters, yet should effectively recognize putative binding regions. In fact, however, they demonstrate a high inconsistency in finding chromatin-enriched peaks, particularly, if the data are represent rather not canonical transcription factor binding events, but chromatin

modification or Polymerase II wider binding regions. It makes the choice of appropriate software to achieve the highest standard in ChIP-Seq analysis an immensely challenging task.

Although many bio-informatics tools are constantly being developed using novel algorithms and sophisticated computational approaches, little consideration has been given to a consistent assessment for the efficiency and capability of these programs to reconstruct the "true" protein-DNA interacting regions. In fact, most of program-testing tools were developed presumably to assess the efficiency of alignment software. They typically randomly model a number of reads across the reference genome, permitting simulation of structural, SNP and INDEL variations, mutations, and sequencing errors (reviewed by Escalona et al. [2]. The problem of mapping is related to analysis of technological noise in sequencing platforms (in particular, Illumina) [3].

To overcome these problems, we developed a new ChIP-Seq simulation algorithm, implemented in a unique in silico ChIP-Seq software (isChIP). We demonstrated that isChIP closely approximate real ChIP-Seq protocol, and capable to model ChIP-Seq data similar to those obtained from experimental Illumina sequencing. We showed that, indeed, peak-calling software can dramatically affect the ChIP Seq results. Moreover, we demonstrate that isChIP is applicable for optimization of the experimental parameters of the ChIP-Seq process.

The work has been supported by RFBR (17-54-80103, 17-54-560028) and ICG SB RAS budget project 0324-2016-0003.

References

1. Y. Orlov et al. (2012) Computer and Statistical Analysis of Transcription Factor Binding and Chromatin Modifications by ChIP-seq data in Embryonic Stem Cell, *J Integr Bioinform*, 9(2):211.

2. M. Escalona, S. Rocha, D. Posada (2016) A comparison of tools for the simulation of genomic next-generation sequencing data, *Nature Review Genetics*, **17**:459–469.

3. te Boekhorst R. et al. (2016) Computational problems of analysis of short next generation sequencing reads, *Vavilov journal of selection and breeding*, 20(6):746-755.