

REVIEW

Advances in computational ChIA-PET data analysis

Chao He^{1,†}, Guipeng Li^{1,†}, Diekidel M. Nadhir¹, Yang Chen^{1,*}, Xiaowo Wang^{1,*} and Michael Q. Zhang^{2,1,*}

¹ MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST, Center for Synthetic and Systems Biology/ Department of Automation, Tsinghua University, Beijing 100084, China

² Department of Biological Sciences, Center for Systems Biology, the University of Texas at Dallas, Richardson, TX 75080-3021, USA

* Correspondence: yc@tsinghua.edu.cn, xwwang@tsinghua.edu.cn, michael.zhang@utdallas.edu

Received May 14, 2016; Revised July 7, 2016; Accepted July 19, 2016

Genome-wide chromatin interaction analysis has become important for understanding 3D topological structure of a genome as well as for linking distal cis-regulatory elements to their target genes. Compared to the Hi-C method, chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) is unique, in that one can interrogate thousands of chromatin interactions (in a genome) mediated by a specific protein of interest at high resolution and reasonable cost. However, because of the noisy nature of the data, efficient analytical tools have become necessary. Here, we review some new computational methods recently developed by us and compare them with other existing methods. Our intention is to help readers to better understand ChIA-PET results and to guide the users on selection of the most appropriate tools for their own projects.

INTRODUCTION

Recently, ChIA-PET [1,2] became one of the few exciting high-throughput genomic technologies that can detect global chromatin interactions at the genome scale at high resolution. Compared to other methods (e.g., Hi-C [3,4]), ChIA-PET allows for mapping of almost all *in vivo* chromatin interactions mediated by a specific protein of interest. At a similar sequencing depth, ChIA-PET can detect chromatin interactions at much higher resolution, which is necessary for studying gene regulation, such as long-range interactions of enhancers and their target gene promoters. Therefore, ChIA-PET is suitable for studying regulatory function of a specific protein factor via its chromatin interactions.

However, a ChIA-PET experiment is relatively complicated and has some characteristics and inevitable noisy sources. It requires effective computational tools to remove this noise in order to detect interactions with high confidence. Below we will first review our new

method MICC [5] which involves a hierarchical mixed probability model for calling statistically significant interactions. Then we will describe how to build a machine learning model that can use multiple 1D ChIP-Seq data to predict 3D chromatin interactions [6] and hence can help to improve ChIA-PET data analysis or to rescue missed/novel interactions. Finally, we will introduce our method 3CPET [7], which allows users to identify possible major protein complexes mediating the ChIA-PET-detected interactions at different loci in the context of 3D genome.

IMPROVED ChIA-PET NOISE MODELS

A common framework for processing ChIA-PET raw data contains five steps. First, one must filter out linkers. For the half-linkers ChIA-PET protocol [1], one has to remove the two types of half-linkers contained in the sequenced paired-end-tags (PETs). The PETs with two different types of half-linkers are designated as chimeric PETs which are used to estimate the frequency of random ligations in solution, while those with the same types of half-linkers are non-chimeric PETs containing both noise and the true signal. For the bridge-linker ChIA-PET protocol [8], the

[†] These authors contributed equally to this work.

This article is dedicated to the Special Collection of Recent Advances in Next-Generation Bioinformatics (Ed. Xuegong Zhang).