Computer analysis of chromosome contacts obtained by ChIA-PET and Hi-C technologies

Ekaterina V. Kulakova,

Novosibirsk State University, 630090, Novosibirsk, Russia, 2 Pirogova Str. kylakovaekaterina@gmail.com

Guoliang Li

National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, 430070, China,
Hubei Province, Wuhan, Hongshan District , No.1, Shizishan Str

Yijun Ruan

The Jackson Laboratory for Genomic Medicine, USA, CT 06032, Farmington, 10 Discovery Drive.
yijun.ruan@jax.org

Yuriy L. Orlov

The Institute of Cytology and Genetics, 630090, Novosibirsk, Russia, Lavrentyeva ave., 10, orlov@bionet.nsc.ru

Transcription regulation is a complex yet well-organized process in eukaryotes, in which chromatin interactions play a critical role for gene expression regulation as well as to further influence other cellular activities. Many technologies have been developed to study the binding of transcription factors (TF) for transcription regulation, such as chromatin immunoprecipitation (ChIP) microarray (ChIP-chip), ChIP-PET and ChIP-Seq, but they are unable to determine the target genes of the distal TF binding sites [1]. Another challenge is to define whether such distal binding sites are functional, i.e. physically proximal to target gene promoters via chromosome loops or attracting RNA polymerase II complex for gene transcription. Therefore, identification of genome-wide distal chromatin interactions that lead the regulatory elements to their target genes may provide novel insights into the study of transcription regulation. Chromatin Interaction Analysis with Paired-End-Tag sequencing (ChIA-PET) method fits these demands still requiring development of specialized high-throughput software for data integration and statistical estimation of the data obtained.

A precise three-dimensional structure of chromatin provides a better landscape of the biological functions. So far, the data of remote interaction is suitable to reconstruct the 3D genome structure. The recent development of genome-wide proximity ligation assays such as Hi-C and its variant TCC [2,3] has significantly facilitated the study of spatial genome organization. The Hi-C technology could capture all the interactions but with low resolution. The ChIA-PET technology [4] greatly enhances the resolution but it can only identify the interactions mediated by a known protein. So, ChIA-PET data can be used to conduct more intense modeling. It is an unbiased, genome-wide, high-throughput and de novo method. Compared with Hi-C, another emerging method for chromatin interactions at a global scale, ChIA-PET is better at its higher resolution associated with a protein of interest for functional study, and lays a solid foundation for studying long-range chromatin interactions in a three-dimensional (3D) manner, as well as provides a more reliable way to determine TF binding sites and identify chromatin interactions.

Compared with other 3C-derived technologies, ChIA-PET protocol is a complex process. It can be summarized into three parts: wet-lab experiments, data analysis and experimental verification. First, the ChIA-PET wet lab complies with the ChIP experiment. Like ChIP-Seq experiment, formaldehyde is used to crosslink DNA-protein complexes in the nucleus and followed by breaking the complexes into fragments with sonication. Then, ChIP is used to enrich DNA fragments bound by a protein of interest. Next, DNA fragments in ChIP-enriched chromatin complexes are ligated with two different half-linker oligonucleotides in two aliquots. Then, the two aliquots are mixed and proximal half-linkers would be ligated with each other. After reversing crosslink, the proteins in the complexes are digested and the DNA fragments are extracted. After digestion with restriction enzyme MmeI, DNA fragments form paired-end tags (PETs) constructs, in "tag-linker-tag" order. Eventually, the PETs are taken to sequencing with new-generation sequencing facilities, like Illumina Hi-Seq. The sequence reads are aligned to the reference genome and further analyses are performed to reveal long-range interactions between functional elements.

While ChIP-Seq is used to analyze the interactions between DNA and protein, ChIA-PET works on the interactions between DNA fragments fundamentally. Fullwood et al. [4] used ChIA-PET technology to construct chromatin interaction network bound by estrogen receptor α (ER-α) from human breast cancer cell line MCF-7 and found long-range ER-α binding sites are mostly located at promoter regions. Handoko et al. [5] found the CTCF-mediated interactions from mouse embryonic pluripotent stem cells. Five distinct chromatin domains revealed by CTCF ChIA-PET raised a new model of CTCF function for chromosome structure organization and linking enhancers to promoters for gene transcription regulation. Li et al. [6] detected promoter-centered distant interactions bound by RNA Polymerase II in cancer cells. In addition to promoter-enhancer and enhancer-enhancer interactions, they found that promoter-promoter interactions are also pervasive in human cells. In all the promoter-nonpromoter interactions, more than 40% of the non-promoter regulatory elements didn't interact with their nearest promoters. This means that the current assumption in ChIP-Seq study – the transcription factor binding sites regulate their nearest genes - is not valid.

Chromatin interaction network is organized into "community", and genes within community perform related functions and respond to external stimuli in a coordinated manner, which means these communities may have been shaped during millions of years' evolution.

The aim of the work was to develop a computer program for statistical data analysis and test it on CTCF binding sites, genes and spatial topological domains. These data have been obtained experimentally by using methods ChIP-seq, Hi-C, ChIA-PET.

We used data on the spatial domains in the genome of the mouse embryonic stem cells and in the human genome, data on the location of CTCF binding sites clusters obtained by ChIA-PET. Gene annotation was obtained from UCSC Genome Browser (http://genome.ucsc.edu). The result of the program was a distribution of CTCF transcription factor binding sites on domains on the human chromosomes. The distributions of human genes relative CTCF binding sites and a randomly generated list of such sites as the program output were used to estimate statistical significance of the associations found.

In addition to ChIA-PET method development, Hi-C method and applications have shown some interesting progress. Gavrilov et al. [7] discussed variability of chromosome contacts between individual cells and suggested in-gel selection of contacting genome fragments. It is expected that the combination of low-resolution structure from Hi-C and high-resolution structure from ChIA-PET will inspire more insights about chromatin structures and their functions in biology. With the rapidly increasing resolution of Hi-C datasets, the size of the chromatin contact map will soon exceed the memory capacity of general computers [7]. The same problem related to ChIA-PET and subsequent data integration to be solved by our software development.

The research has been supported by RFBR (№14-04-01906) and ICG SB RAS budget project

VI.61.1.2. Computing was done at Siberian Supercomputer center SB RAS (SSCC).

References:

1. G. Li et al. (2014) Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application, *BMC Genomics,* **15**(Suppl 12):S11.
2. E. Lieberman-Aiden et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
3. R. Kalhor et al. (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.*, **30**, 90–98.
4. MJ Fullwood, et al. (2009) An oestrogen-receptor-alpha-bound human chromatin interactome, *Nature*, **462**(7269):58-64.
5. L. Handoko et al. (2011) CTCF-mediated functional chromatin interactome in pluripotent cells, *Nat Genet*, **43**(7):630-638.
6. G. Li et al. (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation, *Cell*, **148**(1-2):84-98.
7. A. Gavrilov et al. (2014) Quantitative analysis of genomic element interactions by molecular colony technique. *Nucleic acids research*, **42**(5):e36.
8. W. Li et al. (2015) Hi-Corrector: a fast, scalable and memory-efficient package for normalizing large-scale Hi-C data, *Bioinformatics*, **31**(6): 960-962