Seminars in Cell & Developmental Biology 23 (2012) 213–221

Contents lists available at SciVerse ScienceDirect

Seminars in Cell & Developmental Biology

journal homepage: www.elsevier.com/locate/semcdb

Review

Unravelling global genome organization by 3C-seq

Hideki Tanizawa, Ken-ichi Noma*

The Wistar Institute, Philadelphia, PA, United States

ARTICLE INFO

Article history:
Available online 18 November 2011

Keywords:
Next-generation sequencing
3C-seq
Genome organization
Genome structure

ABSTRACT

Eukaryotic genomes exist in the cell nucleus as an elaborate three-dimensional structure which reflects various nuclear processes such as transcription, DNA replication and repair. Next-generation sequencing (NGS) combined with chromosome conformation capture (3C), referred to as 3C-seq in this article, has recently been applied to the yeast and human genomes, revealing genome-wide views of functional associations among genes and their regulatory elements. Here, we compare the latest genomic approaches such as 3C-seq and ChIA-PET, and provide a condensed overview of how eukaryotic genomes are functionally organized in the nucleus.

Contents

1. Introduction

Next-generation sequencing (NGS) has been utilized for re-sequencing of various genomes from bacteria to mammals [1,2]. Sequencing of the human genome has already identified millions of single-nucleotide variations [3]. We anticipate that in the near future personal genome projects will yield sequences for a large number of human individuals, ultimately allowing us to statistically assess predispositions toward respective diseases. In addition, the application of NGS offers alternatives for microarray-based technologies. For instance, RNA-seq provides a gene expression profile of the entire transcriptome, while ChIP-seq can profile the genome-wide distributions of histone modifications and chromatin-associated proteins [4–7].

Chromatin fibers are non-randomly organized in the nucleus and three-dimensional (3D) organization of the genome is involved in the various nuclear processes such as transcription, DNA replication, and repair [8]. NGS has been recently applied to understanding global genome organization in yeast and human nuclei. Four groups have independently reported methods to analyze *in vivo* genome organization by combining NGS and the molecular biology procedure called chromosome conformation capture (3C), here referred to as 3C-seq [9–12]. The first report from the O'Sullivan laboratory showed that their method, referred to as genome conformation

* Corresponding author. Tel.: +1 215 898 3933; fax: +1 215 573 7919.
E-mail address: noma@Wistar.org (K.-i. Noma).

1084-9521/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.semcdb.2011.11.003

capture (GCC), successfully combined NGS and 3C, and identified genomic associations throughout the budding yeast genome [9]. Several months later, the Lander and Dekker laboratory reported their modified version of GCC, referred to as Hi-C, and mapped long-range genomic associations throughout the human genome [10]. Subsequently, the Noble laboratory at the University of Washington applied their modified method to the budding yeast genome and modeled the three-dimensional genome structure [11]. Finally, our group has developed a relatively simple approach, named enrichment of ligation products (ELP), and applied it to the fission yeast genome [12]. Our study suggests that chromosomal territories and transcription factories, proposed to exist in mammalian cells, are also present in this model organism.

2. 3C-seq methods to capture global genome organizations: a comparison of the treatments

Four different 3C-seq methods were applied to budding yeast, fission yeast or human cells (Fig. 1). The GCC and biotin-based methods were applied to budding yeast cells [9,11]. The ELP method was applied to fission yeast cells [12]. In each of the studies the yeast cells can be considered as wild type, because mutations are only present at the marker genes and the mating-type loci. The Hi-C approach was applied to human cell lines, lymphoblastoid cells (GM06990) and erythroleukemia cells (K562) [10]. In both the yeast and human experiments, the cells are cultured in nutrient-rich liquid media. For fixation of the freshly growing cells, 1% of formaldehyde is added to the cultures, and cells are cross-linked at room temperature for 10 min. Our ELP method instead employs a Zymolyase treatment, which digests yeast cell walls, before the fixation with 4% paraformaldehyde (pFA) at 18 °C for 30 min. We carefully investigated the different fixation conditions ranging from 1 to 4% of pFA concentration, and it is clear that 4% pFA most efficiently captures genomic associations in the fission yeast cells. In the biotin-based budding yeast method, cells are digested by Zymolyase after the fixation. In the Hi-C method, the human cells are lysed by the Dounce homogenizer. At this point, all the samples contain fixed nuclei and cellular debris.

After removing the uncross-linked proteins from the nuclei by treating the samples with SDS solution, permeable nuclei are subjected to restriction enzyme digestion. The GCC uses 4 bp cutter *Msp*I, and DNA is digested for 2 h. The biotin-based budding yeast method uses *Hin*dIII or *Eco*RI for overnight digestion. The ELP employs *Hin*dIII for 2 h digestion. The Hi-C uses *Hin*dIII or *Nco*I for overnight digestion. The shorter digestion time is likely better if the restriction enzymes efficiently cut a major population of genomic DNA in the fixed samples, because endogenous nucleases derived from the samples and exonucleases co-purified with restriction enzymes can potentially cause deleterious DNA digestion. In the Hi-C method, the overhangs derived from the restriction enzyme digestion are filled with biotin-labeled nucleotides. Samples are diluted 16–20 times with T4 DNA ligase buffer, followed by DNA ligation at 16 °C for 1–4 h. More dilution is theoretically better, because it prevents inter-molecular random ligation between non-associating DNA fragments. The cross-links in the ligation products are then reversed by incubating the samples containing proteinase K at 65 °C overnight, followed by phenol extraction and ethanol precipitation.

At this stage, the samples contain purified hybrid DNA molecules reflecting genomic associations. To obtain short DNA molecules suitable for NGS, DNA samples must be digested into smaller sizes ranging from approximately 100–700 bp. To accomplish this the GCC and Hi-C uses sonication and the ELP and biotin-based budding yeast method employs a restriction enzyme with a 4 bp sequence specificity. After shearing the hybrid DNA molecules into small fragments, a majority of DNA fragments do not correspond to hybrid DNA fragments including the restriction enzyme site. For example, if sample DNA is sheared into around 200 bp in length and *Hin*dIII sites appear, on average, every 4 kb, then only 1 of 20 DNA molecules (~5%) contains *Hin*dIII sites, indicating genomic associations. Therefore, small hybrid DNA molecules containing restriction enzyme sites must be enriched before being applied to NGS. This enrichment can be achieved by three methods (Fig. 1). In brief, the Hi-C method uses streptavidin beads to purify the biotin-labeled fragments. The ELP method employs the simple trick using self-ligation and further treatment with another restriction enzyme used for 3C (*Hin*dIII in our experiment). The biotin-based budding yeast method takes additional steps compared to the ELP. It involves ligation of biotinylated adaptor containing the *Eco*P15I restriction enzyme site and *Eco*P15I digestion, followed by purification of biotin-labeled DNA fragments using streptavidin beads. We note here that the biotin–streptavidin purification and the ELP method appear to be comparable in terms of their efficiency in recovering hybrid DNA molecules.

After concentrating the small hybrid DNA molecules, DNA fragments are fused to adaptors and sequenced by NGS with the Paired End (PE) module. The NGS with PE module can determine the sequences present at both ends of the single DNA molecule, and can also determine whether or not they are derived from the same contiguous DNA fragment. Therefore, the PE sequencing allows us to detect hybrid DNA molecules reflecting genomic associations.

3. Processing of sequencing data to detect genomic associations

A single sequencing run of NGS can provide more than 10 million paired reads for hybrid DNA molecules generated by the processes described above. Those paired reads are first mapped to genomic positions, followed by several filtering processes (Fig. 2). To obtain a sufficient number of paired reads to cover the entire genomes, the samples were sequenced 3–10 times. The number of paired reads recovered determines the resolution of the genomic data. It is clear that more sequencing data increases the mapping resolution. For example, the ELP sample was sequenced three times and its derived genomic data achieved a 20 kb resolution for the fission yeast genome. Two independent experiments showed a clear correlation at a 20 kb resolution. The biotin-based budding yeast sample was sequenced 4–10 times for each library, which was sufficient for a 5 kb resolution. The Hi-C sample was sequenced twice, which offered a 1 Mb resolution for the human genome. The resolution is dependent upon the amount of PE sequencing data and the size of the target genome (Fig. 2). Every combination of genomic sections, whose sizes are dependent upon the level of resolution, is scored according to the number of assigned paired reads. The number of paired reads assigned to the combination of genomic sections correlates with the average association frequency in the cell population. Therefore, the numbers of the paired reads assigned to the genomic combinations can be used to estimate genomic associations (Fig. 2).

We performed a control experiment, in which purified genomic DNA that was not cross-linked was digested by a restriction enzyme and ligated, allowing ligation to occur between any DNA fragments. This randomly ligated (RL) control sample, which does not reflect *in vivo* associations between genomic loci, was also processed with the ELP method. We noticed that paired reads from the RL control are not evenly distributed throughout the genome, indicating that there are obvious sequencing biases which also appear to affect the distribution of paired reads from the 3C samples. Therefore, numbers of paired reads from the 3C samples need to be normalized based on the RL data.