Computational and Structural Biotechnology Journal 16 (2018) 43–53

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Computational Methods for Assessing Chromatin Hierarchy

Pearl Chang, Moloya Gohain, Ming-Ren Yen, Pao-Yang Chen *

Institute of Plant and Microbial Biology, Academia Sinica, Taipei, Taiwan

ARTICLE INFO

Article history:
Received 7 September 2017
Received in revised form 29 January 2018
Accepted 11 February 2018
Available online 15 February 2018

Keywords:
3D genome
Chromatin accessibility
Chromosome conformation capture
3C-technologies
Hi-C
ATAC-seq

ABSTRACT

The hierarchical organization of chromatin is known to associate with diverse cellular functions; however, the precise mechanisms and the 3D structure remain to be determined. With recent advances in high-throughput next generation sequencing (NGS) techniques, genome-wide profiling of chromatin structures is made possible. Here, we provide a comprehensive overview of NGS-based methods for profiling "higher-order" and "primary-order" chromatin structures from both experimental and computational aspects. Experimental requirements and considerations specific for each method were highlighted. For computational analysis, we summarized a common analysis strategy for both levels of chromatin assessment, focusing on the characteristic computing steps and the tools. The recently developed single-cell level techniques based on Hi-C and ATAC-seq present great potential to reveal cell-to-cell variability in chromosome architecture. A brief discussion on these methods in terms of experimental and data analysis features is included. We also touch upon the biological relevance of chromatin organization and how the combination with other techniques uncovers the underlying mechanisms. We conclude with a summary and our prospects on necessary improvements of currently available methods in order to advance understanding of chromatin hierarchy. Our review brings together the analyses of both higher- and primary-order chromatin structures, and serves as a roadmap when choosing appropriate experimental and computational methods for assessing chromatin hierarchy.

1. Introduction

Chromatin is a compact and organized assembly of DNA and proteins [32] that is intricately folded into three dimensions, forming different levels of organization in the nucleus. The highest order of chromatin organization is visible during cell division as a chromosome. In a mammalian chromosome, DNA is condensed approximately 10,000 to 20,000-fold [58], and the structure of chromosomal DNA can be categorized as "higher-order" and "primary-order" according to the folding complexity (See Fig. 1 for an overview and assessment of the hierarchical organization of chromatin).

The higher-order genome structure is most clearly visible during the interphase and mitosis when chromatin fibers extensively fold into chromosomes. An interphase chromosome is formed by a tightly coiled 250 nm chromatid. Microscopic imaging has demonstrated that each chromosome may be confined to genomic compartments [59]. Within these compartments, intra-chromosomal interactions are most frequent within regions known as megabase-sized topologically associating domains (TADs). The active TADs are rich in genes, open chromatin marks, transcription factors and DNase I-hypersensitive sites (DHSs) and show early replication. In contrast, the inactive TADs harbor few genes and DHSs and show late replication [77,80,91].

On the other hand, the primary-order chromatin refers to the unpacked chromatin fiber where 11-nm coils of nucleosomes are exposed. The nucleosome is the fundamental unit of chromatin. Each nucleosome comprises 147 bp of DNA wound 1.65 times around core histones [54,74]. Chromatin can be categorized into two varieties: euchromatin and heterochromatin [31]. They differ in terms of the overall compaction of nucleosomes, numbers of genes and transcription levels. The loosely packed regions form the "euchromatin", whereas the densely-packed regions form the "heterochromatin" and represent the less accessible part of the genome [5]. Typically, euchromatin is enriched in genes, and transcription in this region is active. Heterochromatin usually consists of repetitive sequences and forms structures such as centromeres. However, the condensed structure of some heterochromatin can become loose and transcription may take place when under certain developmental or environmental conditions [38,45].

Gene expression and biological functions intimately rely on the interactions between regions (higher-order structure) and the accessibility of chromatin (primary-order structure), which are mediated by protein complexes and epigenetic modifications [7,79,88]. The set of chromatin-associated proteins and epigenetic modifications at a given time in a genomic region constitutes the "chromatin state". With the latest sequencing techniques followed by computational analysis, it is

* Corresponding author.
E-mail address: paoyang@gate.sinica.edu.tw (P.-Y. Chen).

https://doi.org/10.1016/j.csbj.2018.02.003

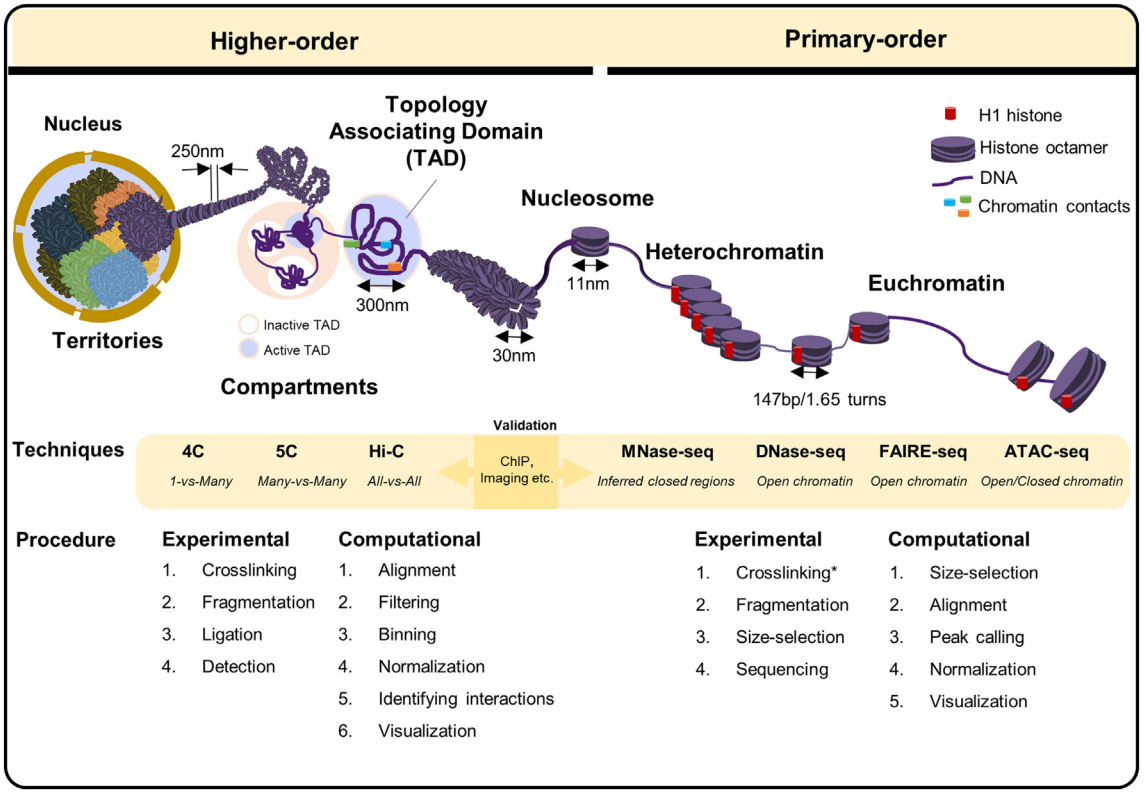

Fig. 1. Genome organization in eukaryotes from higher to primary orders. Features of chromatin organization from higher- to primary-order. Techniques, experimental and computational procedures for assessment of chromatin hierarchy. The active circle represents TADs rich in genes and show early replication. The inactive circle represents TADs that harbor few genes and show late replication. *Among the chromatin accessibility profiling methods, only FAIRE-seq strictly requires crosslinking.

now possible to detect chromatin interactions and its accessibility in the context of functional significance.

This review aims to give a broad overview of NGS-based methods for "higher-order" and "primary-order" chromatin assessment from both experimental and computational aspects. We discuss the characteristics and requirements of each sequencing method together with the computing strategies and bioinformatics tools.

1.1. Assessment of Higher-order Chromatin Structure

1.1.1. Experimental Techniques for the Assessment of Higher-order Chromatin

Microscopy-based imaging tools have been used to observe the higher-order structure of chromatin and its dynamics for over a century [42]. At a resolution of 50–100 nm, light microscopy reveals the shape and distribution of chromosomes in single cells but fails to provide comprehensive detail of the spatial interactions [46]. The development of electron microscopy (EM) and fluorescence *in situ* hybridization (FISH) have provided evidence of chromosomal territories and compartments, organization of TADs and non-random organization of genomic loci within the nuclear periphery [71,104].

Over the past decade, a variety of chromosome conformation capture (3C)-based methods have allowed the detection of higher-order structures of chromatin in unprecedented detail. The conventional 3C method determines the physical interactions of chromatin between two genomic regions (one vs. one) [30,84,102]. The experimental steps include formaldehyde crosslinking to fix *in vivo* contacts, chromatin fragmentation by restriction enzyme digestion and proximity ligation of the digested ends. The restriction enzyme selection depends on the size of target loci; for 3C, frequently cutting enzymes give rise to smaller fragments and hence are more suitable for identifying smaller loci. As a guideline, 4-bp cutters (i.e. frequent cutters) are used when studying small loci sized below 10–20 kb, whereas 6-bp cutters are for loci larger than 20 kb. Ligation junctions are detected in conventional 3C libraries via PCR followed by gel electrophoresis. In combination with next-generation sequencing, the physical interactions of chromatin can be detected with a higher resolution and greater sensitivity [33,56].

More recent 3C-based technologies, such as 4C, 5C, and Hi-C, incorporate next generation sequencing and thereby are capable of providing quantitative measurements for intra (*cis*)- and inter (*trans*)-chromosomal interactions. Circular chromosome conformation capture (4C) uses restriction digestion, followed by inverse PCR, to identify multiple loci interacting with one particular genomic site, referred to as the "bait" or "viewpoint" (one vs. all) [89,93]. The size of a viewpoint is dependent on the primary restriction enzyme used. The optimal size of a viewpoint is approximately 1 kb; viewpoints larger than 1 kb tend to have difficulties to form ligated products, whereas viewpoints that are too short suffer from a lower probability to detect interactions [98]. Furthermore, the reliability of identified close-range (*cis*) or long-range (far-*cis* or *trans*) contact sites depends on experimental setups. Analyses resulted from 4-bp cutter enzymes have been shown to have low reproducibility of 4C signals between replicates, particularly in far-*cis* and *trans* interactions; however, 4-bp cutters are effective in identifying *cis* interacting loci in the vicinity (<10 kb) of the viewpoint [35,99]. In comparison to 4-bp cutters, 6-bp cutters have proven effective in characterizing reliable interactions in distance ranging from 10 kb to 10 Mb [27,73,75]. For extremely long distance interaction (>10 Mb), the signal-to-noise ratios can be improved by *in situ* ligation that occurs inside the nuclei instead of "in solution," thereby decreasing the probability of false inter-chromosomal fusions [98].

Chromosome conformation capture carbon copy (5C) is employed to study all contacts within a particular region (many vs. many), based on highly multiplexed ligation-mediated amplification (LMA) [87]. This technique uses primer pairs that anneal on either side of all ligation junctions in the region of interest in a 3C-based library. These fragments