Methods xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Methods

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

The Hitchhiker's guide to Hi-C analysis: Practical guidelines

Bryan R. Lajoie, Job Dekker *, Noam Kaplan

Program in Systems Biology, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, 368 Plantation Street, Worcester, MA 01605-0103, USA

ARTICLE INFO

Article history:
Received 5 August 2014
Received in revised form 28 October 2014
Accepted 30 October 2014
Available online xxxx

Keywords:
Chromosome conformation capture
Deep sequencing
Chromatin structure
Bioinformatics

ABSTRACT

Over the last decade, development and application of a set of molecular genomic approaches based on the chromosome conformation capture method (3C), combined with increasingly powerful imaging approaches, have enabled high resolution and genome-wide analysis of the spatial organization of chromosomes. The aim of this paper is to provide guidelines for analyzing and interpreting data obtained with genome-wide 3C methods such as Hi-C and 3C-seq that rely on deep sequencing to detect and quantify pairwise chromatin interactions genome-wide.

"Don't panic" – Hitchhiker's Guide to the Galaxy, Douglas Adams.

1. Introduction

The human genome consists of over 3 billion nucleotides and is contained within 23 pairs of chromosomes. If the chromosomes were aligned end to end and the DNA stretched, the genome would measure roughly 2 m long. Yet the genome functions within a sphere smaller than a tenth of the thickness of a human hair (10 μm). This suggests that the genome does not exist as a simple one-dimensional polymer; instead the genome folds into a complex compact three-dimensional structure.

It is increasingly appreciated that a full understanding of how chromosomes perform their many functions (e.g. express genes), replicate and faithfully segregate during mitosis, requires a detailed knowledge of their spatial organization. For instance, genes can be controlled by regulatory elements such as enhancers that can be located hundreds of Kb from their promoter. It is now understood that such regulation often involves physical chromatin looping between the enhancer and the promoter [28,41,15,30,39,52,49]. Further, recent evidence suggests chromosomes appear to be folded as a hierarchy of nested chromosomal domains [33,16,38,44,24,7], and these are also thought to be involved in regulating genes, e.g. by limiting enhancer–promoter interactions to only those that can occur within a single chromosomal domain [21,13,42,23,50].

The chromosome conformation capture methodology (3C) is now widely used to map chromatin interaction within regions of interest and across the genome. Chromatin interaction data can then be leveraged to gain insights into the spatial organization of chromatin, e.g. the presence of chromatin loops and chromosomal domains. The various 3C-based methods have been described extensively before and are not discussed here in detail [5,37]. We first discuss methods and considerations that are important for using deep sequencing data to build bias-free genome-wide chromatin interaction maps. We then describe several approaches to analyze such maps, including identification of patterns in the data that reflect different types of chromosome structural features and their biological interpretations.

2. Comprehensive genome-wide measurement of chromatin interactions

Indiscriminate methods such as microscopy or FISH can study the 3D genome, but have limited resolution and are limited in their capacity to measure multiple discrete contacts simultaneously. The Chromosome Conformation Capture (3C) method was the first molecular method to interrogate physical chromatin interactions in an unbiased manner [14]. 3C has since been further developed into various other derivatives including 4C [46,55], 5C [17] and Hi-C [33]. These methods use 3C as the principal methodology by which they capture genomic interactions. They differ in the actual method by which the captured interactions are measured, e.g. by

* Corresponding author.
E-mail addresses: Bryan.lajoie@umassmed.edu (B.R. Lajoie), Job.dekker@umassmed.edu (J. Dekker), noam.kaplan2@gmail.com (N. Kaplan).

http://dx.doi.org/10.1016/j.ymeth.2014.10.031
1046-2023/© 2014 Published by Elsevier Inc.

PCR in 3C and by unbiased deep sequencing in Hi-C and 3C-seq. Though the 3C method does capture genome-wide data, it was not until the era of deep sequencing came about that one was able to survey all genome wide interactions in a single experiment, as in Hi-C and 3C-seq.

In 3C, cells are cross-linked using formaldehyde, lysed and the chromatin is then digested with a restriction enzyme of choice (typically HindIII or EcoRI). The chromatin is then extracted and the restriction fragments are ligated under very dilute conditions to favor intra-molecular ligation over inter-molecular ligation. The crosslinks are then reversed, proteins are degraded and DNA is purified. The newly generated chimeric DNA ligation products represent pairwise interactions (physical 3D contacts) and can then be analyzed by a variety of down-stream methods. This results in a collection of chimeric DNA fragments consisting of a ligation of DNA sequences from two interacting loci.

Currently, there are two 3C-based methods to obtain genome-wide chromatin interaction data: Hi-C and 3C-seq. In the Hi-C protocol one includes a step to introduce biotinylated nucleotides at ligation junctions which enables specific purification of these junctions [33]. This has the important advantage that it prevents sequencing DNA molecules that do not contain such junctions and are thus mostly uninformative. In 3C-seq one employs the classical 3C protocol and often a more frequently cutting enzyme (e.g. DpnII) followed by intra-molecular ligation without biotin incorporation [44]. The ligated DNA is then directly sequenced to identify pairwise chromatin interactions genome-wide. The 3C-seq methodology sequences all molecules including un-ligated molecules which can complicate the processing/filtering steps and can reduce the percentage of usable reads. However, experimental techniques exist to help minimize uninformative (un-ligated, self-ligated, etc.) Hi-C products (e.g. exonuclease treatment to remove unligated biotinylated ends).

We propose guidelines for analyzing genome-wide chromatin interaction maps generated by Hi-C, but many of these considerations also apply to 3C-seq or other equivalent data.

3. Hi-C data resolution

The space of all possible interactions, which is surveyed by Hi-C experiments, is very large. For example, consider the human genome. Using a 6-bp cutting restriction enzyme, there are almost 10^6 restriction fragments, leading to an interaction space on the order of 10^{12} possible pairwise interactions. Thus, achieving sufficient coverage to support maximal resolution is a significant challenge. However, once can reduce the interaction space, and thus the resolution, by aggregating restriction fragments into fixed-size bins which in turn increases the effective coverage (see Section 5.4).

In light of this, it is critical to establish the goals of the experiment, meaning whether one is most interested in either large-scale genomic conformations (e.g. genomic compartments) or specific small-scale interaction patterns (e.g. promoter–enhancer looping).

If the goal is to measure large scale structures, such as genomic compartments, then a lower resolution will often suffice (1–10 MB). Here, Hi-C using a traditional 6 bp-cutting enzyme could be used. However if the goal is to measure specific interactions of a small region, e.g. promoter–enhancer looping, then one should choose to use a restriction enzyme that cuts more frequently (e.g. 4 bp) and a method that does not measure the entire genome, but instead focuses on exploring only a subset of the genome (i.e. 3C/4C/5C).

In Hi-C the maximal effective resolution of a dataset is determined by several factors, first and foremost is coverage. Given increasing amounts of reads, one will cover more of the interaction space and thus improve the maximal resolution. Library complexity is another factor. Library complexity is defined as the total number of unique chimeric molecules that exist in a Hi-C library, which is a factor of both the number of cells and the quality of the library. A library with a low complexity level will saturate quickly with increasing sequencing depth, e.g. less information will be gained from additional sequencing. The saturation curve can be estimated from a dataset by plotting the cumulative number of unique interactions observed versus increasing read depth.

In our experience, an adequately complex Hi-C dataset for the human genome with roughly 100 million mapped/valid junction reads, is sufficient to support a 40 kb data resolution. Data below 40 kb may be usable, though it will suffer from a higher level of noise. It is important to note that effective resolution scales with genomic distance, such that short-range interactions will typically have higher coverage and thus higher effective resolution.

4. Computational considerations

Hi-C data produced by deep sequencing is no different than other genome-wide deep sequencing datasets. The data starts out as genomic reads in the traditional FASTQ file format (containing a DNA read string and a phred quality (QV) score string). Hi-C libraries are traditionally sequenced using paired-end technology, where a single read is produced from each 5′ end of the molecule. However, Hi-C ligation products can also be sequenced using single end reads, assuming reads are sufficiently long to cover both parts of the chimeric molecule (ligation product) and are handled appropriately during the mapping steps (see Section 5.1).

The data storage requirements for Hi-C datasets are almost solely driven by the sequencing depth needed to achieve the desired resolution and the size of the FASTQ files. The processed Hi-C data will normally be order(s) of magnitude smaller than the size of the FASTQ files. It is easy to parallelize the steps needed to map the reads to the genome, and thus achieve a significant speedup in the Hi-C processing steps. The majority of Hi-C-specific filtering and processing steps are independent and can therefore also be parallelized.

5. Hi-C workflow

We describe the major steps needed to process a Hi-C dataset (Fig. 1):

1. Read mapping
2. Fragment assignment
3. Fragment filtering
4. Binning
5. Bin level filtering
6. Balancing

5.1. Read mapping

Reads can be aligned using any standard read alignment software (i.e. Bowtie [31]) to the genome of interest. Any aligner can be used for mapping Hi-C reads – the goal is to simply find a unique alignment for each read. Even though Hi-C data is sequenced using paired-end reads, the reads are not mapped using the paired-end mode of most aligners. The paired-end mode for most aligners assumes that the ends of a single continuous genomic fragment are being sequenced, and the distance between these two ends fits a known distribution. Since the insert size of the Hi-C ligation product can vary between 1 bp to hundreds of megabases (in terms of linear genome distance), it is difficult to use most paired-end alignment modes as is. One straightforward solution