Methods xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Methods

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

Three invariant Hi-C interaction patterns: Applications to genome assembly

Sivan Oddes, Aviv Zelig, Noam Kaplan*

Department of Physiology, Biophysics & Systems Biology, Rappaport Faculty of Medicine, Technion – Israel Institute of Technology, Haifa, Israel

ARTICLE INFO

Keywords:
Computational biology
Genomics
Genome assembly
Genome scaffolding
3D genome
Hi-C

ABSTRACT

Assembly of reference-quality genomes from next-generation sequencing data is a key challenge in genomics. Recently, we and others have shown that Hi-C data can be used to address several outstanding challenges in the field of genome assembly. This principle has since been developed in academia and industry, and has been used in the assembly of several major genomes. In this paper, we explore the central principles underlying Hi-C-based assembly approaches, by quantitatively defining and characterizing three invariant Hi-C interaction patterns on which these approaches can build: Intrachromosomal interaction enrichment, distance-dependent interaction decay and local interaction smoothness. Specifically, we evaluate to what degree each invariant pattern holds on a single locus level in different species, cell types and Hi-C map resolutions. We find that these patterns are generally consistent across species and cell types but are affected by sequencing depth, and that matrix balancing improves consistency of loci with all three invariant patterns. Finally, we overview current Hi-C-based assembly approaches in light of these invariant patterns and demonstrate how local interaction smoothness can be used to easily detect scaffolding errors in extremely sparse Hi-C maps. We suggest that simultaneously considering all three invariant patterns may lead to better Hi-C-based genome assembly methods.

1. Introduction

Since the publication of the draft human genome sequence, advances in DNA sequencing technology have transformed modern biological research [1]. High throughput next-generation sequencing (NGS) technology, based on short reads, has allowed to obtain massive amounts of genomic sequences. In spite of this, standard short read NGS technology alone is not sufficient to produce reference-quality genomes [2,3]. Ironically, the substantial ease of NGS compared to traditional low-throughput methods (e.g. involving cloning), has led to a stark decrease in the quality of published genomes, when compared to traditionally-sequenced genomes such as that of human and mouse.

To understand the problems associated with standard short-read NGS assembly, let us briefly overview how NGS is typically used for *de novo* genome assembly. First, genomic DNA is fragmented and sequenced. Resulting reads that contain unique overlapping sequences are then stitched together to create longer contiguous sequences called *contigs*. Due to the repetitive nature of genomes and their size relative to the read size, many overlaps will be non-unique, resulting in a huge number of contigs (on the order of 10^5–10^6 contigs for a human-sized genome). Next, contigs are grouped and positioned relative to each other in a process called *scaffolding*. Typically, long-insert libraries are used for scaffolding. This consists of genome fragmentation, size-selection to a predetermined size range and paired-end sequencing. Molecules that uniquely map to two different contigs can then be used to estimate the distance between the contigs and position them relative to one another. Each resulting set of associated contigs (including gaps in between) is called a *scaffold*. Ideally, one would like to obtain a single scaffold for each chromosome. However, even with high coverage and multiple-sized long-insert libraries, a human-sized genome may end up having $\sim 10^4$–10^5 scaffolds. While these highly fragmented genomes may be useful for some applications, they have limited utility for the study of long-range phenomena, including large-scale genome evolution, comparative genomics, gene regulation, haplotyping and 3D genome organization. Furthermore, short read NGS cannot accurately reconstruct complex polyploid and rearranged genomes – including cancer genomes. Thus, these limitations seriously hinder our genomic view of many critical biological systems.

Importantly, the dramatic decrease of cost per read produced by NGS will not help alleviate problems associated with non-uniqueness due to read length. The recent development of long-read sequencing technologies, has helped mitigate some of the problems associated with short reads and can reduce the number of scaffolds by one or two orders of magnitude, but is generally expensive and limited in achieving chromosome-scale scaffolds [4,5]. Alternatively, non-sequencing based technologies exist, ranging from classical cloning to newer techniques such as optical mapping [6,7]. Thus, the development of novel simple techniques that can take advantage of short read sequencing remains an

* Corresponding author.
E-mail address: noam.kaplan@technion.ac.il (N. Kaplan).

https://doi.org/10.1016/j.ymeth.2018.04.013
Received 3 November 2017; Received in revised form 9 April 2018; Accepted 13 April 2018

important challenge.

Hi-C is a molecular biology technique which measures spatial physical proximity between pairs of DNA loci genome-wide [8,9]. Hi-C is based on proximity ligation, such that DNA sequences that are in physical proximity are ligated to each other, and are then measured with NGS technology. By sequencing hundreds of millions of these chimeric molecules, data can be aggregated to construct an interaction matrix which provides an interaction frequency for every pair of genomic loci. Interaction patterns observed in the interaction map are then interpreted and used to extract biological knowledge. Hi-C and similar techniques [10,11], all of which are derivatives of Chromosome Conformation Capture (3C) [12], has been used extensively to study genome 3D organization and led to key biological insights in several different species, including bacteria [13], yeast [14,15], plants [16], worms [17], insects [18] and mammals [8,19–22]. Although in this paper we focus on Hi-C, most of this work is relevant for other genome-wide 3C derivatives.

Recently, we and others have proposed that Hi-C measurements can be used to solve outstanding challenges related to genome assembly [23–33]. This is based on the notion that all Hi-C interaction maps share common features which relate 3D interaction frequencies to the 1D ordering of the genome. For example, loci which are nearby in the genomic sequence tend, on average, to interact more frequently than loci that are located far away on the chromosome, which in turn interact more frequently than loci that are located on different chromosomes [23,34]. Thus, given a set of contigs and Hi-C measurements on a new genome, we can build on these shared principles, consolidating interaction frequencies between contigs to estimate the relative positions of contigs and scaffold them. Advantages of this approach include robustness to very large gaps, relatively low sequencing coverage requirement and applicability to any species. Additionally, a major feature of this approach is that it is based on short-read technology, and can thus directly benefit from the rapid increase in short-read sequencing power. Since the initial proposal of this notion [23,24], Hi-C has been used to scaffold several major genomes including the frog [35], quinoa [36], goat [37], mosquito [29], barley [38], house spider [39], alligator [40], durian [41], lettuce [42] and cassava [42] genomes. Still, Hi-C is typically used for scaffolding in conjunction with intermediate techniques, such as long-read sequencing.

Hi-C is also useful for addressing other challenges related to genome assembly. (1) *Haplotype phasing:* Large-scale haplotype phasing is difficult with short reads, since Single Nucleotide Polymorphisms (SNPs) may be sparse and only reads that map to two SNPs are useful. Due to the physical separation of homologous chromosomes in the nucleus, the probability of observing intrachromosomal interactions is much higher than observing interchromosomal interactions. Thus, Hi-C can be used to reconstruct chromosome-scale scaffolds by providing SNP linkage information over large distances that are not spanned by normal reads [31,43]. (2) *Metagenome deconvolution:* In metagenome and microbiome samples, sequencing typically results in a set of contigs and the challenge is to group contigs that come from the same species. Hi-C provides long-distance linkage information and can thus be used to connect contigs, while the problem of observing genomic interactions between cells is extremely low [26–28]. Hi-C can also be used to associate plasmids with their respective genomes. (3) *Cancer genomes:* Cancer genomes are often highly rearranged and thus pose a major challenge to assemble *de novo*. Large structural variations are especially difficult to measure with short reads, since the rearrangement will only be reflected in a small fraction of the reads (i.e. those that map to the edges of the rearranged region). In Hi-C, structural variations can be detected since they appear to deviate from standard Hi-C patterns [32,33].

2. Invariant Hi-C patterns

Since the 3D organization of a genome reflects its functional state, it is not surprising that 3D genome organization differs between species. In fact, 3D genome organization varies between cell-types [8,20], along different stages of the cell-cycle [44–46], and even within homogenous populations of synchronized cells [46]. Despite this, certain aspects of 3D genome organization, as measured by Hi-C, are universal [23,34]. We refer to these as *invariant patterns*, since they have been observed across species, cell types and conditions. These patterns have even observed in single-cell interaction maps [46,47]. Importantly, other biological patterns, which are specific to the biological system at hand, will lead to local deviations from these general rules, but by large the invariant patterns are dominant in Hi-C data. In fact, these patterns are so robust and ubiquitous that they are used to evaluate the quality of Hi-C experiments and check for experimental artifacts [34]. Importantly, these patterns have previously been studied quantitatively mainly at the level of genome averages, but less at the level of individual loci. In this paper, we set out to explore the central principles underlying Hi-C-based assembly, by quantitatively defining and characterizing three invariant Hi-C interaction patterns on which these approaches can build: Intrachromosomal interaction enrichment, distance-dependent interaction decay and local interaction smoothness. Specifically, we evaluate to what degree each invariant pattern holds on a single locus level in different species, cell types and Hi-C map resolutions/sequencing depths. Evaluation at the single-locus level is important since loci which deviate from these patterns may lead to assembly errors. We note that a general limitation of these analyses is that Hi-C does not directly measure interaction probabilities, meaning that due to the random sampling and sequencing-depth limitation inherent to the experiment, the estimation of small interaction probabilities will be unreliable.

In our analyses we use Hi-C maps from human (Hap1 [48], IMR90 [19], HESC [19]), mouse (MESC [19], cortex [19]), worm (*C. elegans* [17]) and bacteria (*C. crescentus* [13]). All maps were processed using the Dekker lab Hi-C pipeline and balanced unless specified otherwise [49,50]. All code required to reproduce the results in this paper is available at https://github.com/KaplanLab/Invariants.

2.1. Invariant pattern I: Intrachromosomal interaction enrichment

The first invariant pattern is intrachromosomal interaction enrichment. In Hi-C interaction maps, this is observed as a tendency of loci to interact more frequently with loci within the same chromosome (*cis*-interactions) than with loci on different chromosomes (*trans*-interactions). Two major components underlie this pattern. The first component is a phenomenon known as *chromosome territories*, in which chromosomes occupy distinct volumes throughout cell cycle, leading to physical separation between chromosomes [51]. The second component is the random positioning of chromosomes in the nucleus. Although some chromosomes show a tendency to be located near the center of the nucleus while others tend to be located more peripherally (this is known as radial positioning), the relative positions of chromosomes with respect to each other is largely random [52]. This may be due to the inability of relative chromosome positioning to be inherited through cell-cycle, which could lead to such variation in a cell population [53]. Thus, on a population average, as measured in a standard Hi-C experiment, the probability of any specific pair of chromosomes to interact over the entire population is low. Notably, this phenomenon may not hold in a single-cell Hi-C interaction map, but chromosome territories will be present in single-cell interaction maps [46,47]. The combination of these two components yields a strong bias towards intrachromosomal interactions in Hi-C interaction maps.

This invariant is explicitly used to evaluate the quality of Hi-C libraries. Typically the ratio between intrachromosomal (cis) and interchromosomal (trans) is used as a quality metric for Hi-C, but, depending on how it is calculated, may be genome-specific as it can depend on the number and sizes of chromosomes. The underlying logic is that general random noise (such as that caused by background ligation) will affect the interaction matrix uniformly, and thus cause cis and trans to be