Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biotechnology

Computational methods for analyzing genome-wide chromosome conformation capture data

Chiara Nicoletti, Mattia Forcato and Silvio Bicciato

In all organisms, chromatin is packed to fulfil structural constraints and functional requirements. The hierarchical model of chromatin organization in the 3D nuclear space encompasses different topologies at diverse scale lengths, with chromosomes occupying distinct volumes, further organized in compartments, inside which the chromatin fibers fold into large domains and short-range loops. In the recent years, the combination of chromosome conformation capture (3C) techniques and high-throughput sequencing allowed probing chromatin spatial organization at the whole genome-scale. 3C-based methods produce enormous amounts of genomic data that are analyzed using *ad-hoc* computational procedures. Here, we review the common pipelines and methods for the analysis of genome-wide chromosome conformation capture data, highlighting recent developments in key steps for the identification of chromatin structures.

Address
Department of Life Sciences, University of Modena and Reggio Emilia, Modena 41125, Italy

Corresponding author: Bicciato, Silvio (silvio.bicciato@unimore.it)

Current Opinion in Biotechnology 2018, **54**:98–105

This review comes from a themed issue on **Analytical biotechnology**

Edited by **Hiroshi Shimizu** and **Fumio Matsuda**

https://doi.org/10.1016/j.copbio.2018.01.023

Introduction

The characterization of genome spatial organization is paramount for deciphering how higher-order chromatin structures affect genome functionality. Chromatin structures can be probed using fluorescence-based microscopy and methods based on chromosome conformation capture (3C; [1]), such as 4C and Hi-C [2••]. In Hi-C, the three-dimensional spatial proximity of potentially any pair of genomic loci is first traduced into biochemical events and then quantified in terms of interaction frequencies using high-throughput sequencing [3]. Briefly, cells are first crosslinked with formaldehyde *in vivo*, then the DNA is fragmented with a restriction enzyme (digestion), and fragment ends filled with biotinylated nucleotides and ligated; resulting fragments are sheared, precipitated with streptavidin, and sequenced in pair-end mode (see [4] for a full review of 3C-based methods). Hi-C data typically consist of hundreds of millions, or even billions, of read-pairs generated as the result of the ligation of pairs of DNA fragments that were close to each other in the three-dimensional genome space. Read-pairs are processed to generate a genome-wide interaction map in which each entry corresponds to the number of contacts observed between a pair of genomic regions. The width of such regions, which correspond to a fixed-size partitioning of the linear genome, is selected based on the type of restriction enzyme and on the sequencing depth and determines the resolution of the data set. Theoretically, the maximum resolution corresponds to the restriction fragment size [5]. Hi-C contact maps have been generated and analyzed to confirm the existence of genome features identified with structural, biochemical, and microscopy-based methods; to explore the whole 3D chromatin organization in bacteria [6,7], fungi [8–11], protozoa [12], plants [13–15], and metazoan [3,16–19]; to detect chromosomal re-arrangements in primary tumors [20]; to improve the efficacy of the methods for genome assembly [21–23]; for the deconvolution of metagenomes and the investigation of interactions between different species in complex ecosystems [24–26].

In the last few years, numerous approaches have been developed for the analysis of Hi-C data and understanding their principles and usage is becoming increasingly important to efficiently generate and interpret Hi-C interaction maps [27,28]. In this review, we describe the computational procedures commonly used to pre-process sequence reads and generate, normalize and analyze the contact maps and highlight recent advancements in the identification of chromatin structures. The pre-processing and normalization steps are commonly applied without major adaptations to Hi-C data obtained from any type of organism, while methods for the downstream analysis of contact maps may require some customization considering the different complexity of chromatin and chromosome conformations in different species. Most of these approaches are available in the form of computational tools and a selection is summarized in Table 1. Methods that exploit re-ligation frequencies from Hi-C data to finalize genome assemblies and resolving mixtures of metagenomics communities (contact genomics) have been reviewed in [29•].

Table 1

A selection of tools for the analysis of Hi-C data grouped in pipelines for a complete analysis, methods mostly used for pre-processing, domain and point interaction callers, and web applications for visualization (see [27,28] for detailed descriptions of the most widely used software/tools). Procedures and tools for genome assembly and deconvolution of metagenomics mixtures have been reviewed in [26]. Comprehensive catalogues of methods for Hi-C data analysis are available at OMICtools (https://omictools.com) and 4D Nucleome (https://www.4dnucleome.org/software.html) web portals

	Tool	Alignment	Filtering	Normalization	Downstream analysis	Interactive visualization	Reference
Complete pipelines	diffHic	✓	✓	✓	✓		[33]
	Juicer	✓	✓	✓	✓		[38]
	HIPPIE	✓	✓	✓	✓		[39]
	TADbit	✓	✓	✓	✓		[40]
	HOMER		✓	✓	✓		[63]
	HiC-bench	✓	✓	✓	✓		[65]
Pre-processing	hiclib	✓	✓	✓			[32••]
	HiCUP	✓	✓				[34]
	HiC-Pro	✓	✓	✓			[35]
	HiFive		✓	✓			[43]
	HiCapp	✓	✓	✓			[48]
TAD callers	DomainCaller				✓		[18••]
	insulationScore				✓		[19]
	Armatus				✓		[60]
	TADtree				✓		[61]
Interaction callers	SHAMAN			✓	✓		[42]
	Binless			✓	✓		[44]
	GOTHiC		✓	✓	✓		[62]
	Fit-Hi-C				✓		[64]
Interactive visualization	3D Genome Browser					✓	[66]
	WashU Epigenome Browser					✓	[67]
	Juicebox					✓	[68•]
	HiGlass					✓	[69•]

Preprocessing: alignment and filtering of read-pairs

The first step in the analysis of Hi-C data is the alignment of the reads, usually obtained from a paired-end sequencing experiment in the form of FASTQ files (Figure 1a). Since Hi-C libraries are composed of chimeric fragments assembled from distinct genomic loci through ligation, each mate of the read-pairs has to be aligned separately to the reference genome. However, standard (full-read) aligners, as Bowtie [30], fall short in mapping chimeric reads (i.e. reads spanning the ligation junction) that, especially with read-length longer than 75 bp, may represent a substantial fraction of all sequenced reads [28,31••]. To overcome this limitation, it is possible to apply an iterative mapping approach [32••], use strategies that search for the ligation junction to trim or split chimeric reads [33–35], or algorithms for split alignment [36,37] followed by a post-processing of the alignment results [38,39]. In iterative mapping the first portion of the reads (e.g. 25 bp starting at the 5' end) is aligned to the reference genome and those reads that do not map uniquely are extended of 5 bp and re-mapped. The extension-remapping cycle is repeated iteratively until either both reads are uniquely aligned or reads are extended to their full length.

Once aligned, reads are filtered to separate valid, high-quality read-pairs from those accounting for spurious contacts, technical noise, or from reads with low mapping quality (Figure 1b). In principle, Hi-C reads can be filtered with the same basic filters commonly used in other high-throughput sequencing experiments, that is, filters that eliminate reads with multiple alignments or with poor alignment score, read-pairs with only one aligned mate, and PCR duplicates [38]. However, given the distinctive nature of a Hi-C experiment, some computational methods implement additional filtering criteria [33–35,40]. The aim of these filters is to discard read-pairs that are uninformative or, potentially, arise from imperfect or non-specific digestion and ligation. Uninformative reads are read-pairs that fall within the same restriction fragment and are discarded because they can either represent no-ligation fragments (inward-oriented pairs), self-ligation events (outward-oriented pairs) or simply errors (reads with the same orientation). Reads falling on neighbouring restriction fragments and potentially associated to re-ligation of fragments or to no-ligation or self-ligation of un-digested fragments are detected and eliminated based on the strand orientation and on the distance between mates. Reads that map at a distance from the closest restriction site larger than the library size after the shearing step are eliminated as nonspecific products of digestion and ligation. A final additional filter may be applied to reads mapping to restriction fragments that are too small, too large, highly represented or belonging to fragments characterized by low mappability.