UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Università degli Studi di Padova
Dipartimento di Medicina Molecolare

SCUOLA DI DOTTORATO DI RICERCA IN
MEDICINA MOLECOLARE
CURRICULUM BIOMEDICINA
CICLO XXX

Genome conformation and transcription regulation: methods and applications

Direttore della Scuola: Prof. Stefano Piccolo
Supervisore: Prof. Stefano Piccolo
Co-supervisore: Prof. Silvio Bicciato

Dottoranda: Chiara Nicoletti

Abstract

The 3D organization of chromatin within the nucleus is crucial for genome functionality. This is true at multiple levels of resolution: on a large scale, with chromosomes occupying distinct volumes (chromosome territories), at the level of individual chromatin fibers, organized in compartmentalized domains (as the Topologically Associating Domains, TADs), and down to the formation of short range chromatin interactions (as enhancer-promoter loops). The widespread adoption of high-throughput techniques derived from Chromosome Conformation Capture (3C) has been instrumental in advancing the knowledge of chromatin nuclear organization. In particular, Hi-C has the potential to achieve the most comprehensive characterization of chromatin 3D interactions, as in principle it can detect any pair of restriction fragments connected as a result of ligation by proximity. The analysis of the enormous amount of genomic data produced by Hi-C required the development of ad hoc algorithms and computational procedures. Despite the increasing number of available bioinformatics pipelines, no consensus on the optimal approach to analyze Hi-C data has been reached yet. Therefore, we quantitatively compared several Hi-C data analysis methods for the identification of multi-scale chromatin structures to highlight strengths and weaknesses of the various methods and propose application guidelines and good practices. Specifically, we compared different computational approaches (6 for the characterization of chromatin loops and 7 to identify TADs) using publicly available Hi-C datasets, comprising data from different species and cell lines, Hi-C protocol variations and data resolution. Additionally, the algorithms were tested on simulated Hi-C data to assess sensitivity and precision of each method. The tools differed in terms of implemented analysis steps and strategies adopted for alignment, filtering, normalization, and feature identification (global or local looping interactions calling and single-scale or multi-scale TAD discovery). Results of this comparison indicate that performances of the methods considerably vary, both in quantitative and qualitative terms, and that the tools need extensive optimization of the parameters in order to work properly. Despite, in general, TAD callers resulted riper than algorithms to call interactions, still most of them are characterized by crucial limitations, as for instance the inability to investigate how the 3D organization of chromatin structures evolves over time (as e.g., during differentiation). Although the molecular mechanisms underlying TADs formation are still debated, it is evident that distinct interaction patterns can be observed within individual TADs. In particular, some domains appear to have a very compact structure, while others have a less uniform or weaker interaction frequency within the domain, while showing a strong interaction between the borders. To address these limitations, I developed TAD-AH (TADs Advanced Hierarchy), a four-step sequential procedure coded in R, for the characterization of both static and dynamically changing chromatin domains. As a case study, I analyzed Hi-C data generated prior and post human fibroblasts (IMR90) trans-differentiation into skeletal muscle cells (myoblasts, and, when put in differentiation media, myotubes) by overexpression of muscle stem cells master regulator MyoD.

I integrated Hi-C with epigenomic and transcriptomic data from the same conditions and confirmed that the identified genomic features are consistent with the biological scenario under scrutiny.

Sommario

L’organizzazione tridimensionale della cromatina all’interno del nucleo è alla base della regolazione funzionale del genoma, sia a livello macroscopico, dove i cromosomi occupano spazi distinti (territori cromosomici), sia a livello di singole fibre, dove la cromatina si organizza in domini compartimentalizzati (Topologically Associating Domains, TADs), dentro i quali avviene la formazione di interazioni a corto raggio (come quelle che sussistono tra promotori e regioni regolatrici). Le tecniche denominate Chromosome Conformation Capture (3C) hanno permesso di investigare e caratterizzare i diversi livelli dell’organizzazione strutturale della cromatina all’interno del nucleo. In particolare, l’Hi-C, attraverso la combinazione del protocollo di 3C e del sequenziamento massivo, è in grado di restituire un’immagine completa dell’architettura della cromatina e dei contatti all’interno del genoma. Nonostante in questi ultimi anni siano stati resi disponibili diversi strumenti computazionali per l’analisi dei dati di Hi-C, non esiste tuttora un consenso su quale sia il metodo ottimale da usare. Una valutazione comparativa dei software per l'analisi dei dati Hi-C è quindi necessaria non solo per evidenziare i punti di forza e le debolezze dei vari metodi, ma anche per proporre linee guida utili all’utente medio. Per questo motivo ho applicato diversi approcci computazionali (6 per la caratterizzazione delle interazioni e 7 per identificare i TAD) a 6 set di dati pubblici di Hi-C, relativi a diverse specie e linee cellulari (H1-hESC, IMR90, linee cellulari linfoblastoidi ed embrioni di D. melanogaster), a differenti metodiche sperimentali (standard Hi-C, simplified Hi-C e In situ Hi-C) e analizzati a diverse risoluzioni. Inoltre, gli algoritmi sono stati applicati a dati simulati per determinare sensibilità e precisione di ogni metodo. I software differiscono sia per le fasi di analisi implementate sia per le strategie adottate in ciascun passaggio: l'allineamento della sequenza completa contro quello della sequenza “spezzata”, i filtri applicati, la normalizzazione implicita contro quella esplicita, l’arricchimento di interazione locale contro quello globale e l’individuazione di TAD ad uno o più livelli. I metodi variano molto a livello di prestazioni sia in termini quantitativi sia qualitativi, e richiedono di ottimizzare un’ampia gamma di parametri per funzionare correttamente. Nonostante, in generale, gli algoritmi per identificare i TAD si siano dimostrati più affidabili di quelli per trovare le interazioni, ci sono ancora dei limiti fondamentali nell’identificazione dei TAD, ad esempio nello studio dell’evoluzione di queste strutture nel tempo. Sebbene i meccanismi alla base della formazione dei TAD siano tuttora dibattuti, è innegabile che questi siano caratterizzati da pattern distintivi di interazione: in alcuni TAD possiamo osservare un segnale di interazione più omogeneo, mentre in altri l’interazione è più che altro evidente tra le regioni che lo delimitano. Per superare questi limiti, ho sviluppato un nuovo metodo per l’analisi dei TAD a partire da dati di Hi-C (TAD-AH), atto ad indagare un aspetto finora inesplorato dell'architettura del genoma: la quarta dimensione, ovvero come la struttura si evolve nel tempo in base a stimoli di varia natura (ad esempio durante il differenziamento). Per testare TAD-AH ho analizzato dati di Hi-C generati prima e dopo il trans-differenziamento di fibroblasti umani (IMR90) in cellule muscolari (mioblasti e miotubi) ad opera del principale regolatore delle cellule staminali muscolari, MYOD. L’integrazione dei dati di Hi-C con altri dati epigenomici e trascrittomici ha confermato che la caratterizzazione delle strutture identificate è coerente con lo scenario biologico in esame.

Contents

Chapter 1

Introduction

1.1 Motivation

The 3D organization of chromatin within the nucleus is crucial for genome functionality. This is true at multiple levels of resolution: on a large scale, with chromosomes occupying distinct volumes (chromosome territories), at the level of individual chromatin fibers, organized in compartmentalized domains (as the Topologically Associating Domains, TADs), and down to the formation of short range chromatin interactions (as enhancer-promoter loops; Figure 1.1). Genomic organization varies across cell types and undergoes changes during physiological processes (Pombo and Dillon, 2015) and in pathological conditions (Andrey et al., 2013; Lupianez et al., 2015).

Several laboratories are currently cooperating in an international joint initiative, known as 4D Nucleome Consortium, exactly to understand how the spatial organization of DNA affects genome functionality and study genome topology evolution over time (Dekker et al., 2017).

Traditional studies on genome architecture relied on microscopy associated to molecular biology tools such as Fluorescent In Situ Hybridization (FISH), which highlighted how chromosomes are radially distributed and spatially defined by several factors, comprising replication timing, transcriptional activity and GC content (Bolzer et al., 2005).

Unfortunately, microscopy-based techniques are limited both in throughput, allowing the investigation of few loci at once, and in resolution, as a result of the wavelength of light.

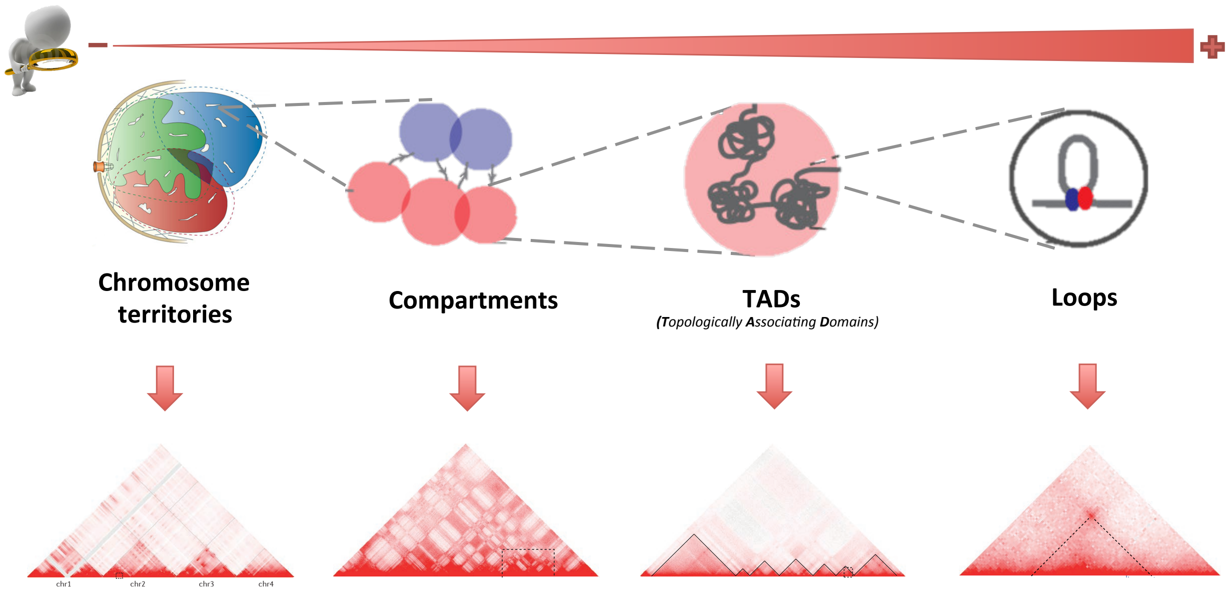

Figure 1.1: Genome 3D structures. Cartoon (above) and as they appear on a Hi-C contact map (below). From left to right: chromosome territories of chromosomes 1-4; euchromatic and heterochromatic compartments (checked pattern); Topologically Associating Domains (triangles); chromatin loops (highly interacting points on top of TADs). Modified from (Bonev and Cavalli, 2016).

1.1.1 Chromosome Conformation Capture techniques

Recently, several genomic strategies to study genome architecture have been developed, collectively known as Chromosome Conformation Capture (3C) techniques (Figure 1.2). These methods share the capacity to translate the three-dimensional information on spatial proximity of DNA into biochemical events, quantifiable by either PCR or next-generation sequencing. Compared to microscopy, 3C-derived methods allow a systematic, high-resolution analysis of DNA topology.

The primary steps, common to all 3C-techniques, are crosslinking with formaldehyde and digestion with a restriction enzyme, with subsequent filling and re-ligation of the resulting sticky ends in order to obtain a circular chimeric molecule comprising the sequences of two spatially close loci. The difference between the 3C-techniques consists only in the way they hybrid DNA molecules resulting from an interaction among a pair of loci in the 3D space.

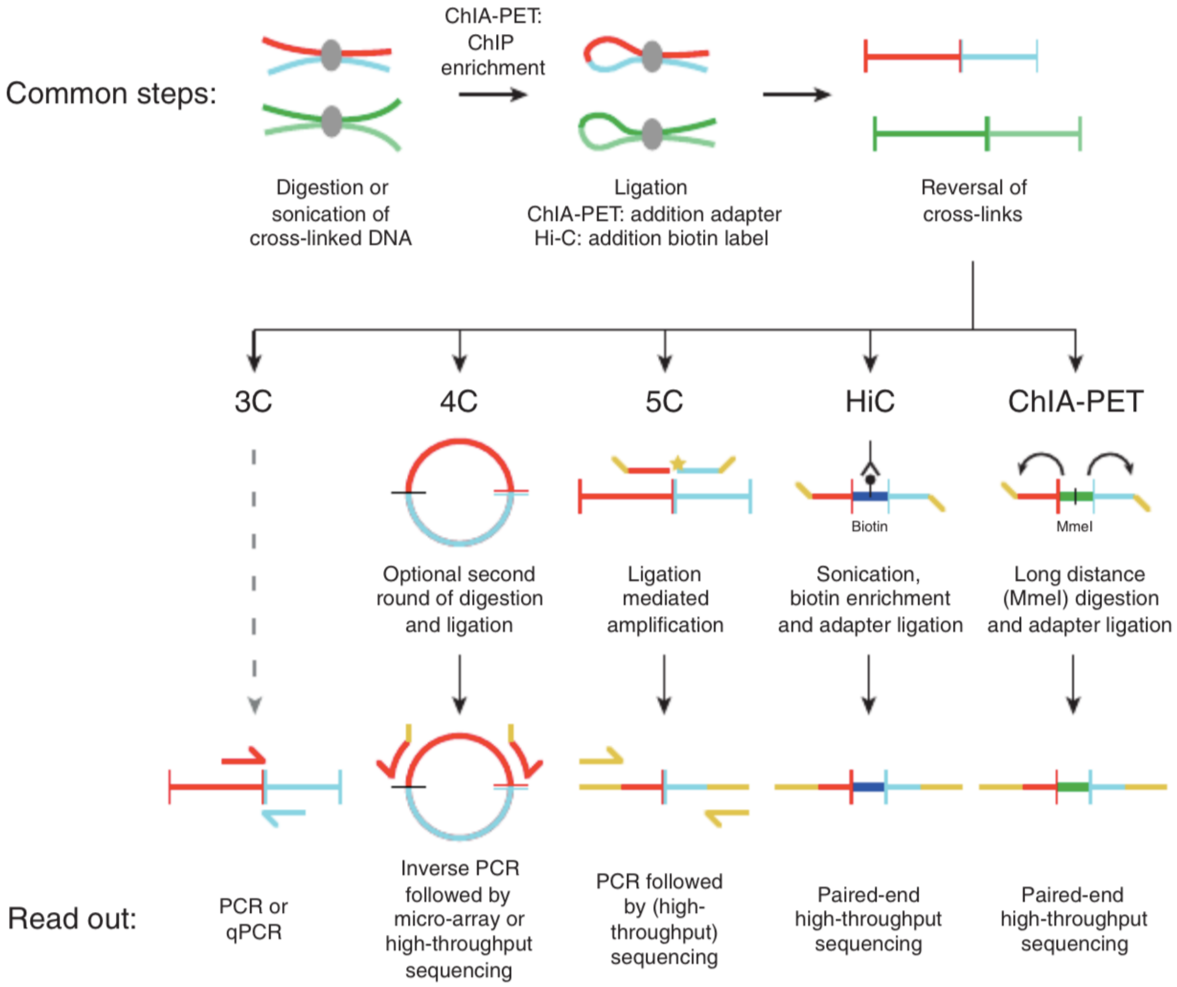

Figure 1.2: 3C-derived techniques experimental protocols (Noordermeer and Duboule, 2013). They all share the initial steps, consisting in chromatin crosslinking, digestion, ligation and reversal of the crosslinking, to create a chimeric molecule containing the sequence of portions of DNA that were close in the nuclear space. Afterwards, steps that are specific to each method result in different readouts: 3C is intended for “one locus versus one locus” inquiries, 4C for “one versus all”, 5C for “many versus many” and finally Hi-C and ChIA-PET for “all versus all” investigations.

In 3C (Chromosome Conformation Capture; Dekker et al., 2002) the use of locus-specific primers leads to the detection of one interaction at a time, covering regions ranging from tens to hundreds of kilobases, whereas in 4C (also known as 3C on chip or Circularized 3C; Simonis et al., 2006) the interactions between one locus and the rest of the genome are profiled through inverse PCR. A variation of the latter method led to 5C (Carbon-Copy Chromosome Conformation Capture; Dostie et al., 2006), which allows the identification of up to millions of interactions in parallel involving two large sets of loci, covering up to tens of megabases, either contiguous or distributed genome-wide. In 5C, the 3C template is hybridized to a mix of oligonucleotides, each of which partially overlaps a different restriction site in the genomic region of interest. Pairs of oligonucleotides that correspond to interacting fragments are juxtaposed on the 3C template and can be ligated together and amplified by multiplexed PCR. Readout of these junctions occurs either on a microarray or by high-throughput sequencing.

Among the 3C-derived techniques, Hi-C (Lieberman-Aiden et al., 2009) is the most promising, being the first method to be completely unbiased and genome-wide.

The traditional Hi-C protocol, known as dilution Hi-C, comprises a single phase in which, after crosslinking with formaldehyde and digestion with a restriction enzyme, the resulting sticky ends are filled with biotinylated nucleotides in order to be subsequently ligated, sheared and purified with streptavidin beads and finally sequenced in paired end mode. The obtained reads are aligned to the genome, filtered and the resulting contact matrix normalized, such that each read pair represents a legitimate ligation junction between two genomic regions that were close to each other in three-dimensional space.

Finally, ChIA-PET (Chromatin Interaction Analysis by Paired-End Tag Sequencing; Fullwood et al., 2009) can be considered a chromatin immuno-precipitation (ChIP) coupled to Hi-C: as in Hi-C, chromatin contacts are captured in a genome-wide fashion, but only those mediated by a protein of interest are retained.

Hi-C produces millions of read pairs (i.e., two sequences of DNA synthetized from the opposite ends of a DNA molecule, in this case represented by a re-ligation fragment) that are used to generate genome-wide maps where each entry x_{ij} accounts for the number of observed interactions between the genomic regions *i* and *j*. The width of such regions (bins) corresponds to the resolution of the dataset, whose choice depends both on the depth of sequencing and on the restriction enzyme adopted: the more restriction sites a genome has, the more fragments the enzyme will generate, providing a more detailed picture of the interacting portions of DNA. For instance, HindIII (a 6-base cutter restriction enzyme) contains around 800,000 cutting sites in the human genome, whereas MboI (a 4-base cutter) can cut in more than seven million different sites.

The last few years saw the flourishing of several Hi-C protocol variants (Table 1.1).

Table 1.1: Hi-C variants and their corresponding resolutions.

Method	Assay type	Resolution	Reference
Capture-C	Multiplexed one to one regions of interest	2kb sliding window in regions of interest	Hughes et al., 2014
Dnase Hi-C	whole genome to whole genome	1-50 kb	Ma et al., 2015
In situ Hi-C	whole genome to whole genome	1-5kb	Rao et al., 2014
single cell Hi-C	whole genome to whole genome	10Mb (single cells)	Nagano et al., 2013

There are four major variants of the Hi-C protocol:

1) *Capture-C* (Hughes et al., 2014) couples Hi-C with oligonucleotide capture technology to target hundreds of regions of interest and uses a 4-base cutter restriction enzyme to achieve a higher resolution, in order to focus on promoters and single nucleotide polymorphisms (SNPs) associated to cancer risk;
2) *Dnase Hi-C* (Ma et al., 2015) protocol envisages the use of the DnaseI instead of a restriction enzyme for the genome fragmentation step (but needs adapters to perform ligation). Thus, the resolution does not depend on the number of restriction sites in the examined genome but is limited only by the sequencing depth;
3) *In situ Hi-C* (Rao et al., 2014) takes advantage of the isolation and independent processing of single nuclei and of the permeabilisation of the nuclear envelope to perform digestion and proximity ligation, thus reaching, as of today, the highest possible resolution (together with Dnase Hi-C);
4) *Single cell Hi-C* (Nagano et al., 2013) relies on isolation and independent processing of single nuclei, but this leads to limited throughput and consequent limited resolution (around 10 Mb). Recent advances in the experimental protocol (Nagano et al., 2017) allow reaching fragment-level resolution.

When several hundred million read pairs are obtained, Hi-C contacts can be detected up to 1 kb resolution (Rao et al., 2014), and, recently, even at sub-kilobase resolution (Eagen et al., 2017). Hi-C thus seems to be the ideal instrument to elucidate the principles behind the three-dimensional architecture of the nucleus and to uncover the link between this architecture and gene expression regulation.

Among the structures that partition the genome, Topologically Associating Domains (from now on named TADs) and chromatin interactions are the most intriguing.

1.1.2 Topologically Associating Domains

TADs were discovered in 2012 by two independent studies (Nora et al., 2012; Dixon et al., 2012), rising from the inspection of Hi-C maps at sufficient resolution (i.e., higher than 100 kb). They appear as highly interacting regions that tend to segregate with respect to the neighboring chromatin and are particularly enriched by boundary elements, as CTCF, at their edges. They often present a hierarchical structure, apparent from high-resolution Hi-C maps, where super-domains of several megabases in length contain smaller ones. TADs are very conserved features of genome topology, with high boundary homology between cell types and even across species (Rao et al., 2014; Dixon et al., 2015), at least when considering just a single layer of TADs. In fact, though TADs can generally be considered conserved, sub-TADs differences are often responsible for cell type-specific chromatin topology (Phillips-Cremins et al., 2013).

Genes inside the same TAD often share similar transcription levels, leading to the assumption that these domains represent fundamental units of the genome expression regulation. TADs have also been linked to DNA replication timing (Pope et al., 2014), corroborated by the observation that TADs are mainly present in the G1 phase of the cell cycle and tend to lose insulation with the entry in S phase (Nagano et al., 2017). They can serve as "niches" in the evolution of pleiotropic loci (Lonfat et al. 2014) and they can also be involved in pathologies (Figure 1.3): the latter result from perturbations in TAD structure, as caused by the loss of a boundary – occurring in various limb malformations (Lupianez et al., 2015) – or through the formation of neo-TADs deriving from genomic translocations – often observed in cancer (Valton and Dekker, 2016).

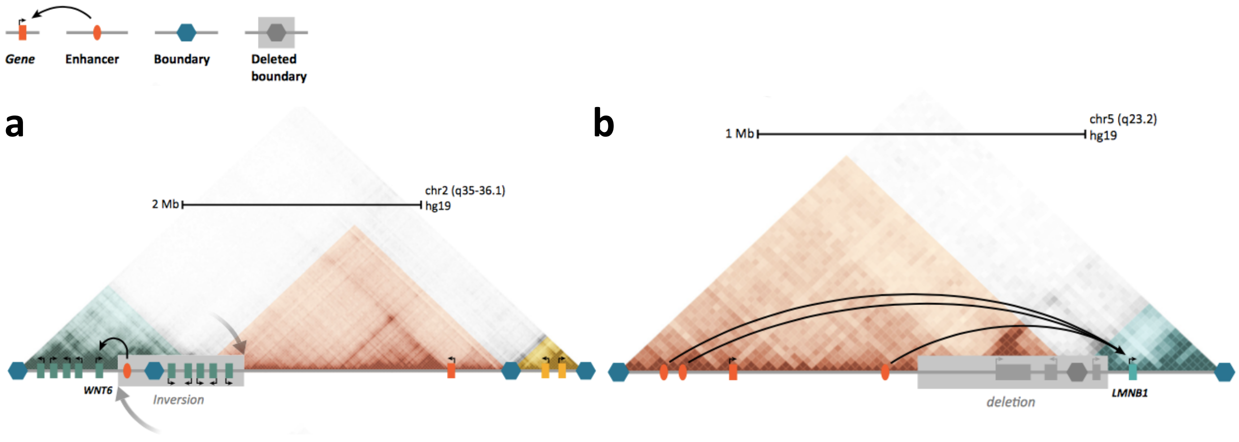

Figure 1.3: Examples of TADs disruption. a) a genomic inversion brings an enhancer cluster in the same TAD as Wnt6, causing its misexpression and resulting in acropectorovertebral dysgenesis; b) a genomic deletion causes the misexpression of LmnB1, resulting in demyelinating leukodystrophy. Adapted from (Lupianez et al., 2016).

1.1.3 Chromatin interactions

Hi-C contacts represent an average ensemble of interactions which can be functional (mediated by protein complexes), bystander (to a functional interaction), random (due to nuclear packaging or random collisions) or due to the co-localization to a sub-nuclear structure (e.g. the lamina associated domains – LADs), relative to a population of cells (Figure 1.4; Dekker et al., 2013).

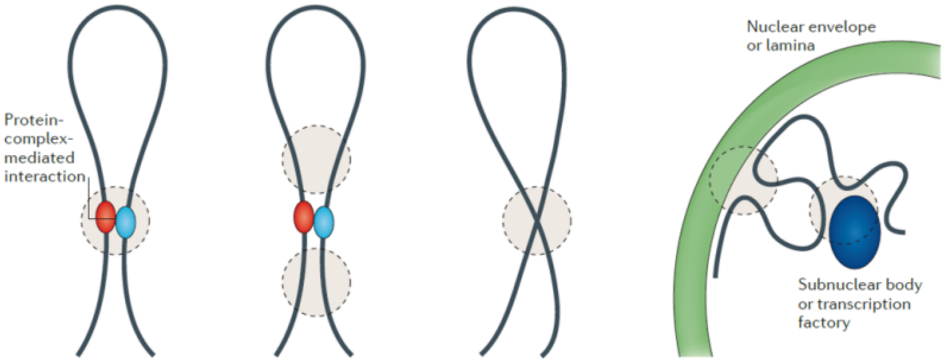
Protein-complex-mediated interaction
Nuclear envelope or lamina
Subnuclear body or transcription factory

Figure 1.4: Types of interactions that can be captured with the Hi-C technique, by (Dekker et al., 2013). From left to right: direct interaction; bystander interaction; random polymer interaction; interaction mediated by sub-nuclear structure.

Examples of functionally relevant interactions involve those connecting distal regulatory elements (enhancers) to their target genes (Apostolou et al., 2013), as well as those mediated by inhibitory complexes such as Polycomb (Eagen et al., 2017).

Among promoter-enhancers contacts, looping interactions represent a conundrum: their anchor regions are often found at TAD boundaries and are enriched in CTCF with a convergent motif orientation (Rao et al., 2014), as well as cohesin, an architectural protein complex mainly involved in chromosome condensation and segregation (Gruber, 2017). Therefore, it has been speculated that looping interactions could be responsible for TAD formation through a loop-extrusion mechanism (Sanborn et al., 2015; Fudenberg et al., 2016): during interphase, cohesin binds DNA to form loops and stops only upon encountering boundary elements, as CTCF (Figure 1.5).

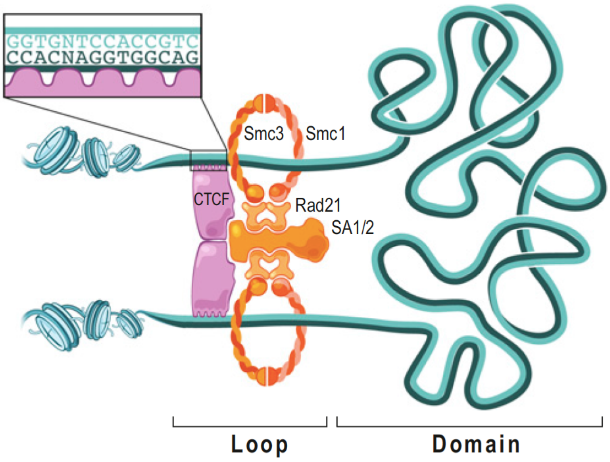

Figure 1.5: Loop extrusion model, from (Sanborn et al., 2015). Cohesin tripartite ring (orange), loaded on DNA, slides along the chromatin fibers and gets stopped by the presence of two CTCF molecules (purple) bound in convergent orientation.

This model is supported by KO experiments of proteins responsible for cohesin loading (SCC4) and unloading (WAPL) from DNA (Haarhuis et al., 2017), which proved that cohesion, if not unloaded, leads to the formation of spurious loops (Figure 1.6b) and that, conversely, loops are disrupted if cohesin is not loaded (Figure 1.6c).

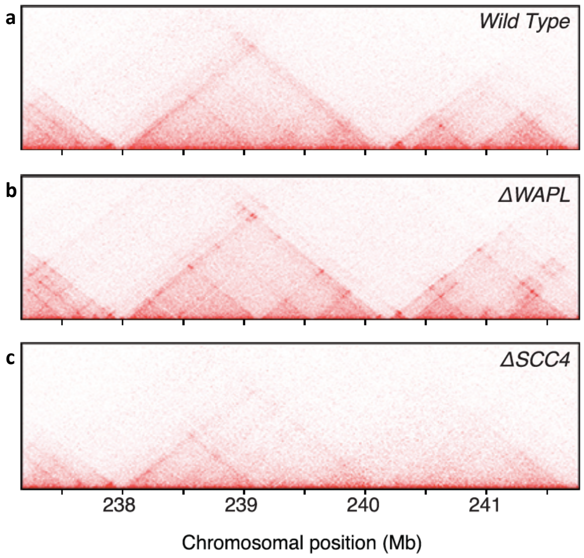

Figure 1.6: Zoom-in of a horizontal Hi-C contact map derived from HAP1 cells (i.e., a haploid fibroblast-like human cell line). In a) we can observe the heat map from wild type cells, whereas b) and c) represent the contact maps resulting from KO of WAPL and SCC4, respectively. Adapted from (Haarhuis et al., 2017).

1.1.4 Hi-C data analysis pipeline

The analysis of the enormous amount of genomic data produced by Hi-C required the development of ad hoc algorithms and computational procedures. Different bioinformatics tools have been implemented to cover the various steps of Hi-C data analysis, e.g. to efficiently preprocess sequence reads (alignment and filtering), remove biases (normalization of contact maps), and infer chromatin structures, as chromatin interactions and TADs (Table 1.2).

Table 1.2: List of available methods for Hi-C data analysis (continues in the next page).

Method	**Category**
3D Genome Browser	Data Visualization
Armatus	TADs identification
AutoChrom3D	Polymer Folding
BNMF	TADs identification
Centurion	De novo genome assembly
CHiCAGO	Chromatin interactions identification
ChromContact	Data annotation/Visualization
chromoR	Normalization/Matrices comparison/TADs identification
ChromSDE	Polymer Folding
CytoHiC	Data Visualization
diffHic	Filtering/Normalization/Chromatin interactions and TADs identification/Comparison
Fast-HiC	Chromatin interactions identification
FisHiCal	Hi-C/FISH data Integration
Fit-Hi-C	Chromatin interactions identification
HiGlass	Data comparison/visualization
GenomicInteractions	Data handling/QC/Filtering/Visualization
GITAR	Data collection/Mapping/Filtering/Normalization/Visualization/TADs identification
GOTHiC	Filtering/Normalization/Chromatin interactions identification
HiBrowse	Data integration/Comparison/Visualization
HiCdat	Data normalization/Correlation analysis/Integration/Comparison/Visualization
HiC-inspector	Data Mapping/Filtering/Matrix generation
Hi-C Data Browser	Data annotation/Visualization
hiclib	Data Mapping/Filtering/Normalization
HiCNorm	Data normalization
Hi-Corrector	Data normalization
hicpipe	Data normalization
HiCPlotter	Data visualization/TADs identification
HiC-Pro	Data Mapping/QC/Matrix generation/Normalization/Allele-specific analysis
HiCseg	TADs identification
HiCUP	Data Mapping/QC/Filtering
HiFive	Data Filtering/Normalization/Visualization
HIPPIE	Data Mapping/Filtering/Normalization/Interactions identification/Annotation
HiTC	Data QC/Normalization/Annotation
HMRFBayesHiC	Chromatin interactions identification
HOMER	Data QC/Filtering/Normalization/Matrix generation/Interactions identification/Visualization/Comparison
HubPredictor	Chromatin interactions identification
Juicebox	Data Visualization
Juicer	Data Mapping/Filtering/Normalization/Interactions and TADs identification/Annotation
LACHESIS	De novo genome assembly

Method	Category
Matryoshka	TADs identification
MOGEN	Polymer Folding
NuChart	Data Annotation/Network analysis
PASTIS	Polymer Folding
TADbit	Data Mapping/QC/Filtering/Normalization/TADs identification/Polymer Folding
TAD_Laplacian	TADs identification
TADlib	Chromatin interactions identification/Annotation
TADtree	TADs identification
WashU Epigenome Browser	Data Visualization

The tools differ in terms of implemented analysis steps and strategies adopted for each step, as e.g., full-read versus split-read alignment, applied filters, implicit versus explicit normalization, global versus local looping interactions calling and single-scale versus multi-scale TAD discovery (Figure 1.7).

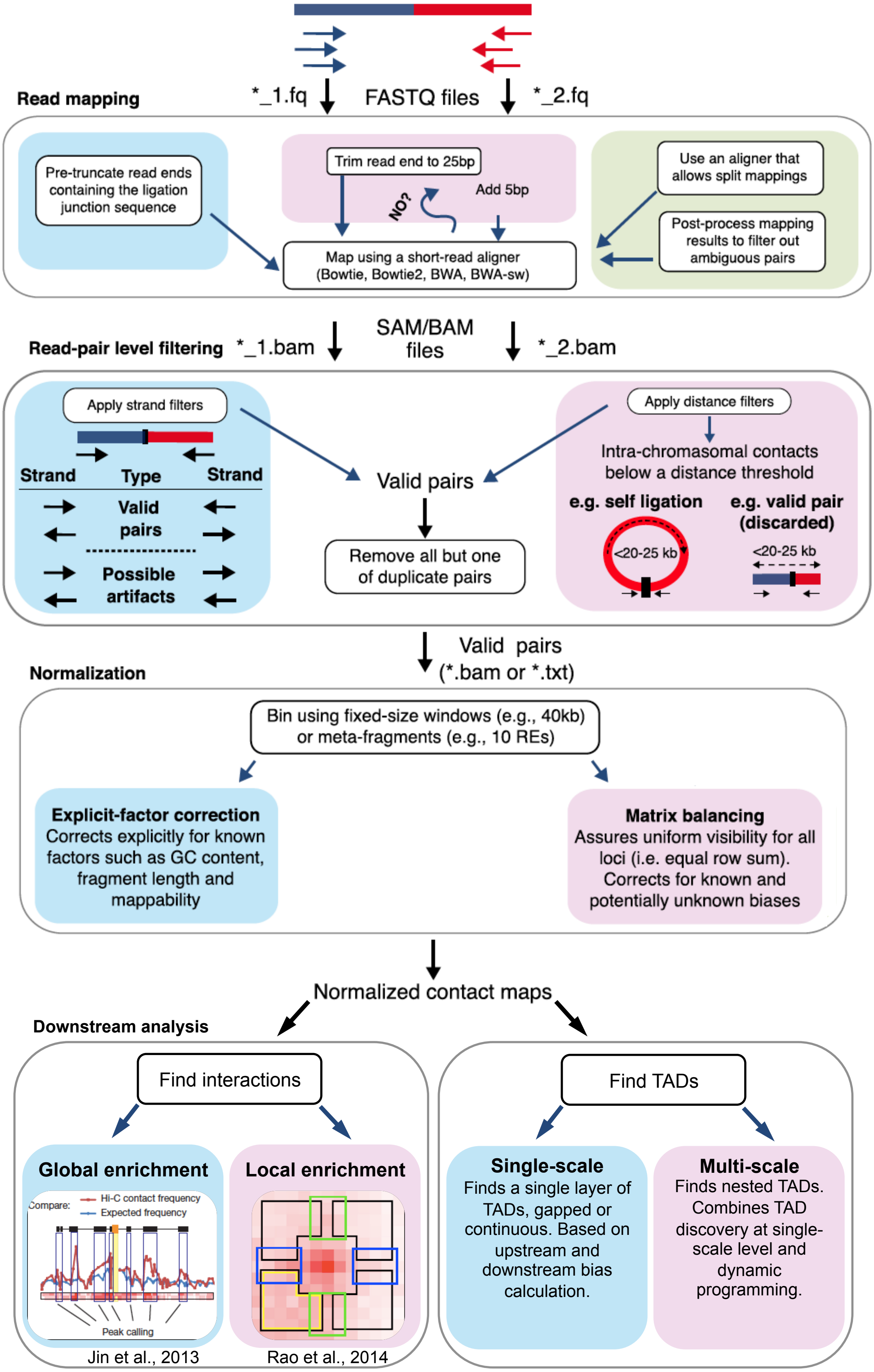
Read mapping
*_1.fq
FASTQ files
*_2.fq
Pre-truncate read ends containing the ligation junction sequence
Trim read end to 25bp
Add 5bp
NO?
Map using a short-read aligner (Bowtie, Bowtie2, BWA, BWA-sw)
Use an aligner that allows split mappings
Post-process mapping results to filter out ambiguous pairs
Read-pair level filtering
*_1.bam
SAM/BAM files
*_2.bam
Apply strand filters
Strand
Type
Strand
Valid pairs
Possible artifacts
Valid pairs
Remove all but one of duplicate pairs
Apply distance filters
Intra-chromasomal contacts below a distance threshold
e.g. self ligation
e.g. valid pair (discarded)
<20-25 kb
<20-25 kb
Normalization
Valid pairs (*.bam or *.txt)
Bin using fixed-size windows (e.g., 40kb) or meta-fragments (e.g., 10 REs)
Explicit-factor correction
Corrects explicitly for known factors such as GC content, fragment length and mappability
Matrix balancing
Assures uniform visibility for all loci (i.e. equal row sum). Corrects for known and potentially unknown biases
Normalized contact maps
Downstream analysis
Find interactions
Global enrichment
Compare:
Hi-C contact frequency
Expected frequency
Peak calling
Jin et al., 2013
Local enrichment
Rao et al., 2014
Find TADs
Single-scale
Finds a single layer of TADs, gapped or continuous. Based on upstream and downstream bias calculation.
Multi-scale
Finds nested TADs. Combines TAD discovery at single-scale level and dynamic programming.

Figure 1.7: Flowchart of Hi-C data analysis main steps. Modified from (Ay and Noble, 2015).

Alignment

Ligation fragments are defined as "chimeras", as they can include two or more non-contiguous genomic loci. Upon sequencing, they can result in reads spanning the ligation junction (especially for longer reads). Such reads would remain unaligned with a standard aligner (e.g., Bowtie), while they can be rescued by other alignment approaches.

There are at least four strategies to rescue chimeras, which would otherwise be discarded, leading to loss of information:

1) scan the reads and trim those that include potential ligation junctions, in order to align each portion of the reads separately;
2) perform iterative mapping (Imakaev et al., 2012), starting with the alignment of the first portion of a read, and adding 5 bp at each iteration until the realization of a unique alignment (i.e. the alignment to a univocal position in the genome);
3) adopt an aligner that allows split mapping (e.g. BWA-mem), which consists in the local alignment of portions of the reads, resulting in multiple reported alignments for chimeras. These reads will be then subject to filtering: if the other mate in the pair aligns univocally to one of the two loci of the chimeric read, the chimera is unambiguous and the read pair is kept, otherwise is discarded;
4) attempt to fully align the reads and find whether the unmapped ones contain exactly one restriction site; if so, split the reads in two pieces and map back each end separately.

Filtering

The second step in Hi-C data analysis, read filtering, is essential for the selection of valid read pairs since, due to the characteristics of the experimental protocol, data account for a lot of spurious contacts (Figure 1.8). Filters can be divided in three main categories:

1) filters applied to single reads. In this category fall filters that discard reads which did not align univocally or had poor alignment score, and reads whose proximity respect to the restriction site is not as expected for a re-ligation event, probably resulting from chromatin random breaks;
2) filters applied to read pairs. This filter category removes read pairs when only one of the two mates aligned, as well as PCR duplicates (i.e., read pairs which share start and end genomic coordinates with other read pairs, resulting from the Hi-C library

polymerase chain reaction amplification step prior to sequencing) and inward/outward read pairs falling in the same restriction fragment, which result from no ligation and self-ligation events, respectively. In order to determine if an inward/outward read pair mapping to neighboring restriction fragments is valid or not, it is possible to apply a distance filter, keeping inward and outward reads separated by more than 1 kb and 25 kb, respectively;

3) filters applied to restriction fragments. These filters discard read pairs mapping to the same restriction fragment, as they would be uninformative, or belonging to fragments characterized by low mappability (i.e., sequence uniqueness).

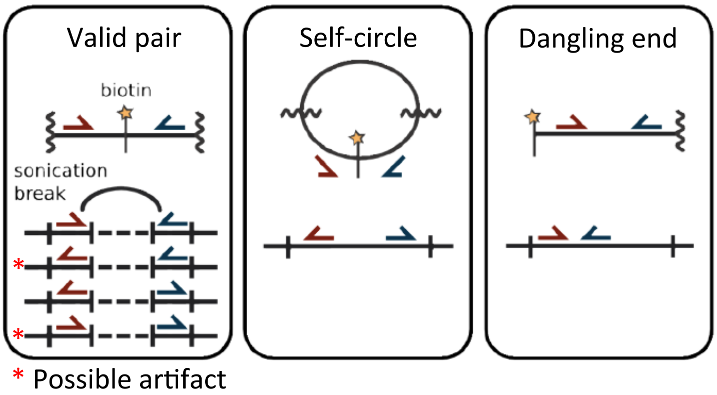

Figure 1.8: Types of read pairs observable after sequencing. Adapted from (Imakaev et al., 2012).

Binning and normalization

After filtering, restriction fragments are grouped into bins defined by fixed genomic windows, in order to have enough contacts covering each region to call the interactions significant for the cell population under scrutiny. Bins are organized in a square contact matrix, where each record represents the contact frequency between two bins in the genome as a proxy of their spatial adjacency.

Normalization helps discriminating real interactions from the background noise represented by random collisions. Two main approaches can be used to address this task, namely explicit-factor correction and matrix balancing. Both methods calculate, genome-wide, a probability of contact (defined as expected count on the basis of the given data) and divide it by the actual contact count for each region (the observed count), leading to a matrix of normalized entries.

Briefly, explicit-factor correction requires an a priori comprehension of the causes behind the biases of Hi-C data. (Yaffe and Tanay, 2011) identified three such factors, i.e. fragment length (affecting ligation efficiency), mappability (sequence uniqueness in the genome, altering Hi-C coverage), and GC content (influencing both PCR amplification and sequencing efficiency), and developed a joint correction strategy that models the probability of observing a contact between two loci given these three genomic features (i.e., explicit factors normalization approach). Subsequently, faster, although equally accurate, variants of this approach were developed, exploiting different regression-based models (Hu et al., 2012). Conversely, matrix balancing does not model biases but relies on the assumption that without biases all the regions of the genome would have the same coverage, thus making normalization a matrix decomposition problem: the algorithm is applied iteratively until all the matrix rows have an equal sum. There are various implementations of matrix balancing (Imakaev et al., 2012; Cournac et al., 2012; Durand et al., 2016) that exploit different balancing strategies (Sinkhorn and Knopp, 1967; Knight and Ruiz, 2012).

Downstream analysis

The normalized contact counts can then be used for downstream analysis, i.e., to extract biologically relevant information from the data, as the characterization of TADs and chromatin interactions. In particular,

- **TADs** became visible for the first time when Hi-C matrices reached the 100 kb resolution as triangles of self-interacting regions with distinct boundaries. The first method to study these genomic features was described in (Dixon et al., 2012) and combines a Hidden Markov Model approach with a directionality index, a simple statistic to quantify the degree of upstream or downstream interaction bias for a genomic region, in order to identify TADs boundaries as points of imbalance. Other strategies have been developed since, some derived from the directionality index, thus identifying a single layer of TADs (i.e., single-scale approach), some others based on dynamic programming, thus allowing the identification of domains that are consistent at different resolutions (i.e., multi-scale approaches);
- **chromatin interactions** can be characterized in many ways, starting with the definition of a background model that takes into account distance scaling factors (i.e. the larger is the genomic distance between two regions, the lower is their probability to

interact and vice versa) and other biases corrected in the normalization step, in order to obtain an observed versus expected ratio and calculate a p-value or a z-score on it. Other approaches are divided into parametric and non-parametric fits and assume that a specific distribution (e.g. Gaussian, Poisson and Negative Binomial distribution) captures the distance dependence of contact counts. Once estimated the parameters of the best fit, the distribution is used together with the distance information and interaction counts to compute an enrichment score for each locus. In the case of non-parametric fits, they capture the distance-dependence relationship directly from the observed counts, using non-parametric methods as the splines, resulting in a distance-dependence changing with resolution and sequencing depth of the data.

In all these cases, the interactions are identified following a global approach, i.e., considering all the contacts engaged by a single genomic window (bin) with all the other bins in which the chromosome is partitioned, mimicking a virtual 4C.

On the contrary, a recently developed method called HiCCUPS (i.e., Hi-C Computational Unbiased Peak Search; Rao et al., 2014; Durand et al., 2016) implements a completely different strategy for calling significant interactions. HiCCUPS computes, for each locus pair, the enrichment of its contact count with respect to the regions all around it (i.e., a local approach). This allows locating regions where contact frequency is substantially higher than its proximal neighborhood and significantly reducing the false positive rate.

1.2 Contribution

The research activity illustrated in this thesis aimed at analyzing Hi-C data from various sources to:

i) compare available methods for the identification of chromatin interactions and TADs from Hi-C data;

ii) develop a new method for Hi-C data analysis, which refines TAD calls and performs differential analysis.

Developed less than a decade ago, Hi-C has seen an advancement of its experimental protocol and is now capable to define genomic contacts at sub-kilobase resolution, representing an unprecedented opportunity to study genome topology and uncover its contribution to gene expression regulation. Hi-C produces an overwhelming amount of data, which required the development of ad hoc algorithms and computational procedures. Currently, there are more than thirty available bioinformatics tools that cover various steps of Hi-C data analysis, implementing different strategies and varying enormously in terms of computational requirements. In such a context, the work of this thesis aimed at comparing all the available tools for chromatin interactions and TADs identification to describe strengths and weaknesses of each approach, and, hopefully, serve as a guide to the user. The pipelines were tested on public datasets chosen from six landmark studies, which differ in many aspects, as they contemplate various organisms, cell lines, Hi-C experimental protocol variants and sequencing depths.

Considering the read alignment phase, the tools implement strategies to perform a full-read alignment as well as split-alignment. For the filtering step, each method adopts different patterns of filters, some employing only basic ones (e.g. alignment quality, presence of PCR duplicates), others supporting more sophisticated approaches (e.g. distance from restriction site, presence of inward/outward read pairs). In the normalization step, adopted strategies basically fall in two categories: methods to normalize the data based on known biases of high-throughput sequencing in general (GC content and mappability) and some specific of the Hi-C protocol (fragment length); and methods to normalize the data for both known and unknown biases, based on the assumption that without biases all the regions of the genome would have equal coverage, thus making normalization a matrix decomposition problem. Finally, downstream analysis, focused alternatively on the identification of TADs or chromatin interactions, differs widely for each approach: methods for TAD discovery can find a single layer of TADs (single-scale analysis) or multiple layers (multi-scale analysis), trying to capture the nested nature of these DNA structures, and can allow the presence of gaps between TADs or find a continuous TAD compartmentalization along chromosomes. Similarly, interaction callers embrace different strategies to find significant contacts: in brief, they either adopt a global or local approach, i.e., establishing the significance of a contact enrichment respect to the contact enrichment of the same region with other regions close on the linear sequence (as in a virtual 4C) or respect to the enrichment of a genomic window surrounding the considered contact.

Results indicate that split-alignment approach should always be preferred (especially when dealing with longer reads) and that *In situ* protocol enables to retain more reads after filtering compared to the other protocols. Performances of the tools vary substantially both in quantitative and qualitative terms, and the tools need extensive optimization of the parameters in order to work properly. In general, we can say that TAD callers give more reproducible results than interaction callers, perhaps also given the more stable nature of domains respect to that of chromatin interactions. Despite TAD callers resulted riper than algorithms to call interactions, still most of them are affected by crucial limitations, as for instance the inability to investigate how the 3D organization of chromatin structures evolves over time (e.g., during differentiation). Moreover, even though the molecular mechanisms underlying TADs generation are still debated, distinct interaction patterns can be observed within individual TADs. In particular, some domains appear to have a very compact structure, while others have a less uniform or weaker interaction frequency within the domain, while showing a strong interaction between the borders. To address these limitations, I developed TAD-AH (TADs Advanced Hierarchy), a four-step sequential procedure coded in R to refine TAD calls at multi-scale level and perform differential analysis. TAD-AH takes as input Hi-C matrices and a list of TADs identified at different resolutions and filters them to obtain hierarchical structures; it classifies TADs based on their inner signal density and integrates the filtered TADs with other omics data such as ChIP-seq and RNA-seq data to further characterize the TADs. Finally, it performs differential analysis to identify TADs that are conserved, acquired or lost between conditions or that just change their characteristics. TAD-AH was tested on a high-resolution Hi-C dataset of human fibroblasts (IMR90) converted to muscle cells upon overexpression of the skeletal muscle stem cell master regulator, MYOD. I integrated Hi-C with epigenomic and transcriptomic data from the same experimental conditions and confirmed that the identified genomic features are consistent with the biological scenario under scrutiny.

1.3 Document organization

Chapter 2 details all information, data and bioinformatics methods used in this thesis; in particular, section 2.1 contains a presentation of Hi-C data as well as other omics data analyzed in this work; section 2.2 describes the creation of simulated Hi-C data; section 2.3 illustrates methods for data pre-processing, while section 2.4 and 2.5 characterize algorithms for TADs (6 methods) and chromatin interactions (7 methods) identification, respectively. Section 2.6 describes the generation of a random set of chromatin interactions, while section 2.7 is about the integration of the interactions found by each algorithm with cell-type specific chromatin states. Section 2.8 focuses on the assessment of CTCF binding profile in order to characterize looping interactions (as described in Rao et al., 2014), while section 2.9 is about computational running times collection. Finally, section 2.10 illustrates a new method (TAD-AH) for TAD calls refinement and differential analysis applied to Hi-C data portraying DNA topology during fibroblasts trans-differentiation into muscle cells. Chapter 3 is dedicated to the presentation of results. Specifically, section 3.1 is about the comparison of methods for data preprocessing (3.1.1), interactions (3.1.2) and TAD (3.1.3) identification, while section 3.2 illustrates the results related to analyses performed with TAD-AH. Finally, conclusions are summarized in Chapter 4.

Chapter 2

Materials and Methods

Chapter 2 details all information, data and bioinformatics methods used in this thesis; in particular, section 2.1 contains a presentation of Hi-C data as well as other omics data analyzed in this work; section 2.2 describes the creation of simulated Hi-C data; section 2.3 illustrates methods for data pre-processing and section 2.4 and 2.5 those for TADs (6 methods) and chromatin interactions (7 methods) discovery, respectively. Section 2.6 describes the generation of a random set of chromatin interactions, while section 2.7 is about the integration of the interactions found by each algorithm with cell-type specific chromatin states. Section 2.8 focuses on the assessment of CTCF binding profile in order to characterize looping interactions (as described in Rao et al., 2014), while section 2.9 is about computational running times collection. Finally, section 2.10 illustrates a new method, TAD-AH, for TAD calls refinement and differential analysis applied to Hi-C data generated during fibroblasts trans-differentiation into muscle cells upon expression of muscle stem cell master regulator MYOD.

2.1 Data collection

2.1.1 Public Hi-C data

Experimental data were obtained from six landmark studies, from which I selected nine data sets for a total of 41 samples (Table 2.1). The datasets cover three experimental protocol variations, comprise several cell types, some of which overlapping to facilitate inter-dataset comparisons, and have been analyzed at different resolutions. Specifically, data have been generated using dilution Hi-C, i.e., the original Hi-C protocol published in (Lieberman-Aiden et al., 2009), simplified Hi-C, introduced in (Sexton et al., 2012), and in situ Hi-C developed by (Rao et al., 2014). The simplified Hi-C differs from the dilution protocol because it does not include the use of biotin to enrich for ligated junctions (thus retaining a lot of spurious fragments), while in situ Hi-C performs DNA re-ligation in intact nuclei instead of under dilute conditions, allowing a more accurate picture of the contacts occurring inside the nucleus.

Table 2.1: Hi-C public experimental data.

	Cell type					Restriction Enzyme							
Study	LCL[a]	H1-hESC	IMR90	Fly Embryo	Hi-C Protocol[b]	HindIII (6bp)	NcoI (6bp)	DpnII (4bp)	MboI (4bp)	Read length (bp)	Median read count (per replicate, in millions)	Resolution (kb)[d]	N° of replicate samples
Lieberman-Aiden[7]	✔				Dilution	✔	✔			76	11	1000	4
Sexton[6]				✔	Simplified			✔		36	362	40	1
Dixon 2012[4]		✔	✔		Dilution	✔				36/100[c]	328	40	4
Jin[8]		✔	✔		Dilution	✔				36/50[c]	440	5/40	7
Rao[9]	✔		✔		In situ			✔	✔	101	240	5/40	23
Dixon 2015[25]		✔			Dilution	✔				36/50[c]	999	5/40	2

[a] LCL: lymphoblastoid cell lines (i.e., GM06990 in Lieberman-Aiden and GM12878 in Rao); [b] Dilution, simplified, and in-situ refer to the Hi-C protocols presented in Lieberman-Aiden et al., (2009), Sexton et al, (2012), and Rao et al.(2014), respectively; [c] Samples have been sequenced with different read length in the same study; [d] Resolution refers to the resolution used in this comparison. In the case of two values, the first refers to the resolution used for chromatin interactions, the second for TADs.

Samples include human cell lines from various tissues (embryonic stem cells: H1-hESC; fetal lung fibroblasts: IMR90; lymphoblastoid cell lines (LCL): GM12878 and GM06990) and D. melanogaster embryos. All data have been obtained using 6 bp (HindIII) or 4 bp (MboI, DpnII) cutter restriction enzymes. Some replicate samples from Lieberman-Aiden and Rao GM12878 have been processed with both restriction enzymes. All biological replicates have been analyzed separately. In particular, the Rao GM12878 dataset contained 26 samples obtained with in-situ protocol and MboI restriction enzyme and divided into a primary (16 technical replicates of 1 biological sample) and a replicate experiment (10 biological and technical replicates; see Supplementary Table 1 of Rao et al. 2014). I decided to select the replicate with the highest number of sequenced reads from the primary experiment (i.e., SRR1658572, originally labelled as HIC003 and renamed here as replicate H) and all the in-situ samples of the replicate experiment, because my main porpoise was to see how consistent the called TADs and interactions were, dealing with biological variation. Moreover, I also analyzed as separate samples the technical replicates of the replicate experiment since the authors defined technical replicates also those samples for which cells were cross-linked together but processed independently. In the (Jin et al., 2013) study, it must be noted that the H1-hESC sample – originally composed of SRR639047, SRR639048, and SRR639049 and here renamed as replicate A – is the same H1-hESC sample presented by (Dixon et al., 2012) (which was composed of SRR442155, SRR442156, and SRR442157 and is here renamed as replicate B). Both H1-hESC samples from (Jin et al., 2013) and (Dixon et al., 2012) were analyzed with chromatin interaction callers at their original resolutions (5 and 40 kb, respectively), but only the H1-hESC sample from (Dixon et al., 2012) was used for the TAD analyses, as they were conducted at 40 kb for all datasets. All the public Hi-C data accession numbers used in this thesis are reported in Table 2.2.

Table 2.2: Details of the 41 samples used in this study. Data were downloaded from SRA (i.e., Sequence Read Archive; ncbi.nlm.nih.gov/sra).

Dataset	Cell Type	Restriction Enzyme	SRA accession number	Replicate ID
Lieberman-Aiden	GM06990	HindIII	SRR027956	A_HindIII
	GM06990	HindIII	SRR027957	A_repeat
	GM06990	HindIII	SRR027958, SRR027959	B
	GM06990	NcoI	SRR027960, SRR027961	A_NcoI
Sexton	Fly Embryo	DpnII	SRR389762, SRR389763, SRR389764, SRR389765, SRR389766, SRR389767, SRR389768	A
Dixon 2012	H1-hESC	HindIII	SRR400260, SRR400261, SRR400262, SRR400263	A
	H1-hESC	HindIII	SRR442155, SRR442156, SRR442157	B
	IMR90	HindIII	SRR400264, SRR400265, SRR400266, SRR400267, SRR400268	A
	IMR90	HindIII	SRR442158, SRR442159, SRR442160	B
Jin	H1-hESC	HindIII	SRR639047, SRR639048, SRR639049	A
	IMR90	HindIII	SRR639025, SRR639026, SRR639027, SRR639028, SRR639029	A
	IMR90	HindIII	SRR639030, SRR639031, SRR639032, SRR639033	B
	IMR90	HindIII	SRR881990, SRR881991, SRR881992	C
	IMR90	HindIII	SRR881993, SRR881994, SRR881995	D
	IMR90	HindIII	SRR881996	E
	IMR90	HindIII	SRR881997	F
Rao	GM12878	MboI	SRR1658592	A
	GM12878	MboI	SRR1658593	B
	GM12878	MboI	SRR1658594, SRR1658595	C1
	GM12878	MboI	SRR1658596, SRR1658597	C2
	GM12878	MboI	SRR1658598	D
	GM12878	MboI	SRR1658599	E1
	GM12878	MboI	SRR1658600	E2
	GM12878	MboI	SRR1658601	F
	GM12878	MboI	SRR1658602	G1
	GM12878	MboI	SRR1658603	G2
	GM12878	MboI	SRR1658572	H
	GM12878	DpnII	SRR1658644	A1
	GM12878	DpnII	SRR1658645	A2
	GM12878	DpnII	SRR1658648	A3
	GM12878	DpnII	SRR1658646	B1
	GM12878	DpnII	SRR1658647	B2
	IMR90	MboI	SRR1658672	A1
	IMR90	MboI	SRR1658673	A2
	IMR90	MboI	SRR1658674	A3
	IMR90	MboI	SRR1658675	A4
	IMR90	MboI	SRR1658676	A5
	IMR90	MboI	SRR1658677	B1
	IMR90	MboI	SRR1658678	B2
Dixon 2015	H1-hESC	HindIII	SRR1030718, SRR1030719, SRR1030720, SRR1030721	A
	H1-hESC	HindIII	SRR1030722, SRR1030723, SRR1030724, SRR1030725, SRR1030726, SRR1030727	B

2.1.2 Genomic data

In order to validate interactions and TADs found with the presented methods, other omics data were used. In particular, Table 2.3 reports the data used to validate chromatin interactions, whereas Table 2.4 lists the ChIP-seq data about boundary elements used both for chromatin interactions and TAD calls.

Table 2.3: Details of the interactions demonstrated to be present (True positive) or absent (True negative) in the same cell types of the Hi-C datasets using 3C, 5C, ChIA-PET, and 3D-FISH and of interactions known to exist in specific cell types at a given physiological state (cell specific evidences used as true positives or true negatives).

List name	Validation technique	Cell type	Interaction type	Number of interactions
3C Hou[1]	3C	Fly embryos	True positive evidence	4
3C Sexton[2]	3C	Fly embryos	True positive	2
3C Jin[3]	3C	IMR90	True positive	6
3C He (TP)[4]	3C	GM12878	True positive	3
5C Ferraiuolo[5]	5C	H1-hESC, IMR90, GM12878	True positive evidence	29
5C Sanyal[6]	5C	H1-hESC	True positive	1237
5C Sanyal[6]	5C	GM12878	True positive	1187
ChIA-PET Ji (TP)[7]	ChIA-PET	H1-hESC	True positive evidence	28
FISH Rao (TP)[8]	3D-FISH	GM12878	True positive	4
3C Woon-Kim[9]	3C	H1-hESC, IMR90, GM12878	True negative evidence	13
3C He (TN)[4]	3C	GM12878	True negative	2
5C Smith[10]	5C	GM12878	True negative	383
ChIA-PET Ji (TN)[7]	ChIA-PET	H1-hESC	True negative evidence	125
FISH Rao (TN)[8]	3D-FISH	GM12878	True negative	4

[1] Hou et al., 2012; [2] Sexton et al., 2012; [3] Jin et al., 2013; [4] He et al., 2014; [5] Ferraiuolo et al., 2010; [6] Sanyal et al., 2012; [7] Ji et al., 2016; [8] Rao et al., 2014; [9] Woon Kim et al., 2011; [10] Smith et al., 2016.

CTCF and BEAF32 ChIP-seq peaks were retrieved from ENCODE and modENCODE. In particular, we considered peaks uniformly generated by the ENCODE Analysis Working Group (ENCODE Project Consortium, 2012; Wang et al., 2012) and peaks obtained from combined replicates for modENCODE data (Celniker et al., 2009).

Table 2.4: Details of CTCF and BEAF32 ChIP-seq peaks used in this study.

Experiment	Cell types	Accession number
CTCF ChIP-seq[1]	H1-hESC, GM12878	GSE29611
CTCF ChIP-seq[1]	IMR90	GSE31477
CTCF ChIP-seq[2]	GM06990	GSE30263
CTCF ChIP-seq[3]	Embryo 14-16hr Oregon-R	GSE47264
BEAF32 ChIP-seq[3]	Embryo 14-16hr Oregon-R	GSE51986

[1] ENCODE Project Consortium, 2012; [2] Wang et al., 2012; [3] Celniker et al., 2009

2.1.3 Hi-C, RNA-seq and ChIP-seq data from IMR90 trans-differentiation

Hi-C. IMR90 cells were grown in growth media (GM) consisting in EMEM supplemented with 10% FBS. Electroporation was performed in proliferating cells at doubling passage 11-15, while all other experiments were performed at doubling passage 23-28.

IMR90 cells were electroporated with helper plasmid and epB-Puro-TT containing or not murine MYOD cDNA. Cells were then selected with 2 ug/ml puromycin. When cells were 60% confluent, MYOD was induced with 200 ng/ml doxycycline in GM (for 24 hours) and cells were collected for the GM time point. For the DM (i.e., differentiation media) time point, MYOD was induced for 24 hours with doxycycline when GM cells were at 95-

100% confluence, then cells were differentiated in EMEM supplemented with 2% horse serum, 1% ITS and 200 ng/ml doxycycline for three days.

In situ Hi-C was performed as previously described (Rao et al., 2014) with minor modifications, and sequencing library size was selected at 200-600 bp. DpnII was chosen as a restriction enzyme for two main reasons: first, being a 4-base cutter, it generates more DNA fragments respect to its 6-base counterparts and thus allows a higher resolution on chromatin contacts; secondly, even though it recognize the same 4-base sequence as MboI, it can cut the genome more frequently as it is the methylation-insensitive isoschizomer of MboI (Rao et al., 2014). HiCPro-v2.7.7 (Servant et al., 2015) was used for alignment on human genome (hg19), valid ligation product detection, quality control, normalization and sparse chromosomal interaction maps, whereas HiTC (Servant et al., 2012) was used to transform sparse matrices to *NbyN* square matrices at 40 kb or 4 kb resolution.

RNA-seq. For each experimental condition, cells were collected from the plate using trypsin, then inhibited by media addiction. Cells were then divided in three tubes: one was processed for cell cycle analysis, one for DNA extraction and one for RNA extraction. Reads were sequenced in single end mode and aligned to the female Homo sapiens hg19 genomes with TopHat2.1.1 (Kim et al., 2013), with options: –g 1 –segment-length 17 –library-type fr-firststrand. HTSeq-0.6.1 (Anders et al., 2015) with –stranded=reverse option was used to assign mapped reads to Homo Sapiens UCSC hg19 genes.

ChIP-seq. Cells were fixed in 1% formaldehyde in PBS for 15 min at RT. Formaldehyde was then quenched with 125mM Glycine for 5 min at RT. Cells were washed in PBS and harvested in PBS supplemented with 1mM PMSF and protease inhibitors. Dry cell pellet was stored at -80°C. Nuclei were then extracted and then lysed in lysis buffer containing 50 mM Tris-HCl, pH 8.0, 150 mM NaCl, 5 mM EDTA, pH 8.0, 0.5% SDS, 0.5% NP-40, 1 mM PMSF and a protease inhibitor. Chromatin was sheared with sonicator to an average DNA fragment length of 200-500 bp. Chromatin was then diluted 5 times in lysis buffer without SDS. DNA amount was measured with the Qubit. DNA was immuno-precipitated either with anti-MYOD, anti-p65, anti-H3K4me3, anti-H3K4me1, H3K27ac or H3K27me3 O/N at 4°C. The immuno-complexes were captured with protein A magnetic beads for 3-4 hr at 4°C. After four washes with buffer containing 50 mM Tris-HCl, pH 8.0, 150 mM NaCl, 5 mM EDTA, pH 8.0, 0.1% SDS, 1% NP-40, 0.5% sodium deoxycholate, one wash with a buffer containing 250 mM LiCl, 100 mM NaCl, 5 mM EDTA, pH 8.0, 1% NP-40, 1% sodium deoxycholate and two washes with TE buffer (10mM Tris-HCl pH=8, 1mM EDTA) chromatin was then eluted and de-crosslinked with 1% SDS in TE O/N at 65°C 600RPM rotation. Also, the input is de-crosslinked with 1% SDS in TE O/N at 65°C 600RPM rotation. After 2 hr digestion at 37°C with 0.2 mg/ml proteinase K, DNA was extracted with phenol/chloroform and ethanol precipitated O/N at -20°C. Prior to sequencing, DNA was suspended in mQ water. The DNA was then analyzed by qPCR calculating the amount of immuno-precipitated DNA relative to the input DNA (i.e., percentage of input). Library preparation and sequencing of immuno-precipitated and input DNA were performed as described in bioinformatics-renlab.ucsd.edu/RenLabLibraryProtocolV1.pdf. Reads were aligned with bowtie2-2.0.5 to the female Homo sapiens hg19 genome with options: --very-sensitive-local. Duplicate reads were removed using samtools1.3. Peaks were called using macs2 (v2.1.1; Zhang et al., 2008) with qvalue<0.01.

2.2 Simulated data generation

Simulated Hi-C data were generated using a modification of the procedure proposed by (Lun and Smyth, 2015). Briefly, in the original approach read counts for a single chromosome contact matrix are generated by random sampling from a negative binomial distribution with dispersion parameter set to 0.01 and mean $\mu_{(x,y)}$. The mean is defined for each pair of genomic bins (x,y) with an additive model summing three components:

1. If the (x,y) pair is inside a TAD, a first signal component with power law decay is added with value equal to $K_t(x-y+p)^c$, where is the distance between the interacting bins in the (x,y) pair, to which a prior p=1 is added. K_t is the baseline for TAD signal in the power law decay equation with exponent c. We estimated the power law decay parameters ($K_t = 28$ and c = -0.69) from a real contact matrix (chr5 in Dixon 2012, IMR90 replicate B) preprocessed with hicpipe at 40kb. When x=y (diagonal of the contact matrix), a fixed value is added instead of K_t. We set the diagonal constant signal value to 35 following the same proportion used by (Lun and Smyth, 2015) for the ratio (1.25) between diagonal constant signal and K_t.
2. A second component is added to account for random noise. For this component a number (Nnoise) of (x,y) pairs is randomly sampled with replacement from the entire contact matrix. The sampling probability for (x,y) pairs (excluding the diagonal) is designed to follow a power law decay with equation $K(x-y+p)^c$, with K=1 and c=-

0.69 as above. Due to the sampling with replacement, any given (x,y) pair could be selected multiple times. Each time an (x,y) pair is selected, a constant noise signal (K_{noise}) value is added to it: we set K_{noise}=2. The number of sampled (x,y) pairs was varied across different simulation settings to account for variable noise levels. Namely the number of sampled pairs was defined as a percentage (4%, 8%, 12%, 16% or 20%) of the total number of data points given the estimated target size of the simulated matrix (estimated target size defined below).

3. A third component is added to account for points of true cis interactions. In our default simulations, we added $N_{contacts}$=205 interactions in the contact matrices, so as to follow the same proportion of true interactions over the total target matrix size used in (Lun and Smyth, 2015) (i.e., 2 contacts every 100k data points). The interaction points were limited to be not too far from the diagonal, following the settings used in the original procedure, i.e., using the maximum TAD size plus one third of it as maximum distance for interacting pairs. In this manner, for a maximum TAD size of 50 bins, the interacting bin pairs were, at maximum, 67 bins apart. The signal component added to interacting (x,y) pairs is defined with a power law decay with equation $K_d(x - y + p)^c$, with p=1 and c=-0.69 as above, and with $K_d = 2K_t = 56$, thus following the same proportion between and used in the original procedure.

The read counts for each (x,y) bin pair, generated with the negative binomial distribution sampling described above, were then formatted and saved as Hi-C count matrices and used as input to those interaction and TAD callers that required raw count as input. For tools requiring observed over expected normalized data, the raw count matrices were converted to Vanilla Coverage matrices, following the procedure described in (Lieberman-Aiden et al., 2009).

Simulation of Hi-C matrices for TAD callers. The TADs coordinates, to define which (x,y) pairs are inside a TAD, are randomly simulated sampling TAD sizes from a uniform distribution with minimum value 3 (3 bins TAD size) and maximum value 50 (50 bins TAD size). Each simulation generates the contact matrix of one chromosome, containing a fixed number of TADs and having a variable size determined by the random sampling of TAD sizes. The target size of the simulated matrix can be defined based on the expected average of the uniformly distributed TAD sizes. Using this strategy 171 TADs were simulated, which at 40 kb resolution resulted in a target size of the simulated matrix similar to the size of the human chromosome 5 (180.92 Mb), i.e., the same used to estimate the power law decay parameters. For all of the other parameters we used the values indicated above. Slightly different settings were used to simulate a hierarchy of nested TADs. Namely, while preserving the target chromosome size (180.92 Mb), we simulated a first level of TADs with size ranging from 3 to 20 bins (393 TADs). Then we simulated 2 additional layers of TADs by randomly removing, for each additional layer, 25% of the TAD borders from the preceding (lower) layer. The removal of each border caused the merging of one or more smaller TADs into a larger one. We imposed that the maximum TAD size, obtained after the random merging of smaller TADs, was ≤ 75 bins (3 Mb). This resulted in a hierarchy with 3 layers of simulated nested TADs. Accordingly, in the simulation of nested TADs, we decreased the baseline TAD signal to one third of the value previously estimated (i.e., to $K_t = 9$). In the simulation of nested TADs, we also kept $N_{contacts}$=1000 to have a more complex background in the data.

Simulation of Hi-C matrices for interaction callers. For interaction callers, we generated the simulated data as previously described for TADs, but increasing the number of interactions from $N_{contacts}$=205 to $N_{contacts}$=1000. Moreover, we used as baseline signal for the interactions (K_d) a value with progressively larger ratio to the baseline for TAD signal. Namely we set to 2x, 3x, 4x or 5x the value of (where $K_t = 28$). In addition, to have an even greater signal to noise ratio in the simulated true interactions, we generated a separate set of data in which an additional fixed constant was added to the signal of the true interacting pairs. This additional fixed true interaction constant ($K_{interactions}$) was set equal to K_{noise}=2. We designed these simulation schemas to increase the signal to noise ratio in the simulated true interactions. Indeed, the original method of (Lun and Smyth, 2015) was designed to test differential interaction calls in diffHic. Having multiple replicated samples, a 2-fold signal increase over the TAD baseline signal (i.e., $K_d = 2\ K_t$) was sufficient to detect differential interactions. Instead, we observed that larger signal to noise ratio was generally required to call interaction contacts in individual conditions without replicates. The simulated Hi-C count matrices were used as input to the interaction callers (HiCCUPS, HOMER, diffHic, and Fit-Hi-C) and to HiCseg and TADbit that require raw count as input. For the other TAD callers, requiring observed over expected normalized data, the raw count matrices were converted to Vanilla Coverage matrices, as described in (Lieberman-Aiden et al., 2009).

2.3 Hi-C data pre-processing

2.3.1 General data preprocessing

Except for five methods (diffHic, HIPPIE, TADbit, HiCCUPS and Arrowhead), which comprise read mapping, all the others required the use of an external aligner. Thus, reads have been mapped with a traditional short read aligner, Bowtie (v.1.1.1), for the tools that did not implement the alignment step, and with their specific aligner, if they did. Bowtie was chosen because commonly used and notoriously reliable, and also on the basis of the possibility to keep only uniquely mappable reads, setting the "-m" parameter to 1.

It must be noted that TADbit aligner (GEM mapper; Marco-Sola et al., 2012) was not used in this work, mainly because it is based on full-read alignment as Bowtie, and also, as 5 out of 7 TAD callers did not implement neither alignment nor filtering/binning steps, to use a common pre-processing procedure (hicpipe) for all tools and better appreciate differences in obtained TAD lists ascribable to the various downstream analysis approaches. When possible, the normalization step was common to all TAD callers too.

Subsequently, in order to discriminate the contribution of pre-processing from that of down-stream analysis also for interaction callers, I re-analyzed the data following a common preprocessing procedure, with hiclib.

2.3.2 Common data preprocessing for interaction callers: hiclib

hiclib is a python library containing modules for all Hi-C data processing steps, including iterative alignment, filtering, binning, and normalization as described in (Imakaev et al., 2012). We used a more recent implementation of the tool available in the mirnylib python library, which implements the same methods but adopting compressed binary Hierarchical Data Format (HDF) files for efficient storing of data. Both hiclib and mirnylib are available in the public Bitbucket repository of the authors (bitbucket.org/mirnylab). We used the core modules available within this package to iteratively align (alignment based on Bowtie2 v.2.2.6) and filter the reads. For alignment, we set the minimum sequence size to 20 bp with an increment of 10 bp until full read length. After alignment, modules of the same library were used to convert aligned reads to HDF files. During filtering, we filtered for same fragment reads, PCR duplicates, and read pairs distance sums greater than 500 bp, as described in Imakaev et al.16. Filtered reads in output from mirnylib served as input to all interaction callers, with the only exception of HIPPIE, whose downstream analysis cannot be detached from its preprocessing.

2.3.3 Common data preprocessing for TAD callers: hicpipe

hicpipe (Yaffe and Tanay, 2011; v1.03 - the most updated version available at the beginning of this study, but no more available on wisdom.weizmann.ac.il/~eitany/hicpipe) is an end to end pipeline, consisting of an assortment of scripts designed for automating the various steps of the Hi-C data analysis. The package uses rule definitions within a *makefile* to process Hi-C datasets and is designed to take advantage of high-performance cluster computing with SGE compatible jobs scheduling system. Indeed, almost all of the steps are performed using distributed computing paradigms. As said before, Bowtie (v.1.1.1) was used for alignment. The pipeline digests the genome into corresponding restriction fragment ends (fends) and computes the mappability for these fragments prior to mapping. In the filtering step, same fragment reads, reads pairs with one or both reads mapping precisely over a restriction site, and reads mapping to low mappability fragment ends (mappability <0.5) are discarded. In addition, read pairs based on the sum of distances to the nearest downstream restriction site were filtered out. Specifically, after checking the sum of distances distribution, we used the default 500 bp threshold in all datasets except for Sexton, where the threshold was set at 800 bp. After assigning the reads to their corresponding fends, the pipeline counts the number of unique fend pairs occurring within each genomic bin. As such, each fend-to-fend pair was counted only once even if multiple reads supported that contact. This represented a de facto stringent filter for PCR duplicates. Finally, an expectation model was computed considering the fragment length, GC content, and mappability (Yaffe and Tanay, 2011). This expectation model was used to normalize the observed fend-to-fend contacts. hicpipe was then used to build interactions matrices (either observed contacts or normalized as observed over expected ratios) that served as input for the TAD callers.

2.4 Interaction callers

2.4.1 Fit-Hi-C

Fit-Hi-C (Ay et al., 2015; noble.gs.washington.edu/proj/fit-hi-c) is a Python command-line software that requires Python≥2.7, designed to identify mid-range intra-chromosomal contacts in Hi-C data. Overall, the deployment is straightforward, provided that all python libraries are correctly installed. The tool first models with a spline the observed counts as a function of the genomic distance between all possible pairs. This spline is used to define a threshold to filter out outliers and then a second spline is fit to calculate a refined null model. Using this model, expected contact probabilities are calculated for each locus pair using the genomic distance between the pair. Biases learned by Iterative Correction and Eigenvector decomposition normalization (ICE; Imakaev et al., 2012) are incorporated in the expected contact probability calculation. Finally, p-values for the comparison of observed read counts vs the expected contact probability are calculated using the binomial distribution and are corrected for multiple testing. The algorithm requires as input a file containing the interactions at bin (or fragment) level, a file describing the bins (or fragments), and a file of ICE biases calculated over the interactions in the bin or fragment space. Since Fit-Hi-C does not provide utilities to pair and filter reads, we used interaction files obtained from GOTHiC, which can be easily converted into the format required by Fit-Hi-C. In particular, for all datasets, but Dixon 2015, I used GOTHiC for pairing, filtering, assignment of reads to their restriction fragments, and binning. The same steps have been performed with HOMER to generate the Fit-Hi-C input from the Hi-C data of Dixon 2015. ICE biases were calculated with a script provided by the author of Fit-Hi-C. Since the software required more than 512 GB of RAM for the analysis of datasets at 5kb resolution, all the input files were split per chromosome and the analyses were run separately, as suggested by the author. All Fit-Hi-C parameters were left as default except for the lower bound on the intra-chromosomal distance range (-L), which was set to 1 bin in order to exclude the interactions falling on the diagonal. Only cis interactions are given as output, with contact count, p-value, and FDR. Only interactions with a FDR<0.05 (as suggested in Ay et al.15) were considered for subsequent analyses.

Recently, Fit-Hi-C has been also released as an Bioconductor R package (bioconductor.org/packages/devel/bioc/html/FitHiC.html), which has been reported to be more efficient than the original Python implementation and easier to use for those familiar with R.

2.4.2 GOTHiC

GOTHiC (Mifsud et al., 2017) comes as a Bioconductor package and requires R>3.1.0 (bioconductor.org/packages/release/bioc/html/GOTHiC.html).

The installation is straightforward as for any other R package. The package vignette and manual are very concise, and more precise details on some of the filters and normalization can be obtained only by inspecting the R code.

GOTHiC takes aligned reads as input in BAM or Bowtie format, pairs them (removing PCR duplicates, PD), assigns the pairs to enzyme-specific restriction fragments, discards those coming from restriction fragments separated by less than 10 kb (default value), in order to eliminate undigested chromatin (UC), and bins the interactions into fixed genomic windows. It finally performs a normalization similar to Vanilla Coverage (i.e., a matrix balancing with a single iteration, as seen in Lieberman-Aiden et al., 2009) and a binomial test to identify significant interactions, followed by Benjamini-Hochberg multiple testing correction. p-value is the probability of observing a certain number of reads between two sites by chance and is a function of the coverage of both sites and the total number of reads. Cis and trans interactions (i.e., within and between chromosomes) are reported, together with the log2 of observed over expected interactions, p-value, FDR and the number of read pairs supporting the interaction itself. Only the interactions with a FDR<0.05 and with more than 10 contact counts were considered for subsequent analysis, as described in (Schoenfelder et al., 2015). The user needs to set very few, well-documented parameters: apart from the resolution itself, nothing needs to be changed from one resolution to another. We needed to edit the code of the function that partitions the genome into the fragments generated with the restriction enzyme (i.e., mapReadsToRestrictionSites) to force the removal of random chromosomes and alternative haplotypes, as they would lead to execution failure. Two of three steps of GOTHiC can be parallelized, although parallelization of mapReadsToRestrictionSites did not work in our hands when using the wrapper and, to achieve a full parallelization, we had to launch the three functions one after the other. For larger samples GOTHiC required more than 500GB of RAM and failed to analyze the Dixon 2015 dataset since the read pairing step could not be completed on a machine with 1TB or RAM. The author as well suggested the use of other tools to perform read pairing, as the pairing function of GOTHiC is not memory efficient.

2.4.3 HOMER

HOMER (Heinz et al., 2010; v4.7.2) is a command-line software mainly written in Perl (homer.ucsd.edu/homer/download.html). It is easy to install in a Unix environment, well documented (documentation is provided as html pages) and user-friendly (the user needs to set only few parameters as some are directly estimated from the input data). HOMER takes aligned reads as input, pairs them and removes PCR duplicates and performs quality control (fragment size estimation and distance from restriction site distribution), the latter returning several tables, already formatted for an easy visualization. For the filtering step, which is mainly based on restriction site proximity, the following parameters were used (as suggested in the user guide): -removePEbg -both -removeSelfLigation -removeRestrictionEnds -removeSpikes 10000 5.

In particular, we discarded read pairs separated by less than 1.5 times the estimated fragment length (UC). We also applied several filters accounting for restriction site proximity (RSP), i.e., we keep read pairs only if the distance between both ends and a restriction site was within the estimated fragment length and discarded read pairs where one end started exactly on a restriction site or read pairs if their ends formed a self-ligation with adjacent restriction sites. Finally, we removed read pairs originating from regions with abnormally high read density (spikes, S). With data from (Sexton et al., 2012), we also specified the fragment length (750 bp) since the fragment length distribution showed two peaks. After filtering, HOMER generates a background model, which, as default, normalizes genomic interactions for linear distance and coverage at the chosen bin level (FullModel). The background model is used to estimate the expected read count and a binomial test is applied to call significant chromatin interactions. Finally, the software performs a cumulative binomial test to find significant looping interactions. Interactions within and between chromosomes are reported, together with their modified z-score, p-value, FDR, number of read pairs supporting the interaction (both observed and expected) and the interaction distance. Only the interactions with a p-value<0.001 (default value) were considered for subsequent analysis.

HOMER is not memory intensive and the interaction-calling step can be run on multiple processors. As described in the user-guide, all read pairs are reported twice so the user has to pay attention when calculating the statistics for pairing. Similarly, significant trans interactions are reported twice (e.g., both chr1-binA/chr2-binB and chr2-binB/chr1-binA) and the output file has to be cleaned by the user with custom parsing scripts.

2.4.4 HIPPIE

HIPPIE (Hwang et al., 2015, v0.0.2-beta; wanglab.pcbi.upenn.edu/hippie) relies on R (≥3.1.0), Python and Perl. It calls DNA-DNA interacting regions in Hi-C data, starting from raw reads and identifying interactions at restriction fragment resolution. It has been designed to run on computing clusters with Open Grid Scheduler, or other Sun Grid Engine (SGE) compatible jobs schedulers. The setup of the pipeline is relatively easy, also because accompanied by a detailed documentation. HIPPIE requires the preparation of an initialization file (hippie.sh) and of a separate configuration file (.cfg) for each dataset. HIPPIE performs five different steps: i) reads mapping, ii) quality control, iii) identification of Hi-C peaks, iv) prediction of enhancer-target gene interactions including Hi-C bias correction, and v) analysis of enhancer-target gene interactions. For our purpose, it was sufficient to run the first four phases of the pipeline with default parameters. HIPPIE relies on the chimeric alignment implemented in the STAR aligner (Dobin et al., 2013), which uses the sequential maximum mappable prefix search (MMP). Briefly, the MMP of a read is searched starting from the first base of the read; then, if the read cannot be mapped contiguously to the genome, it is split and MMP search is performed again on the unmapped portion22. HIPPIE takes the aligned reads and filters them for PCR duplicates (PD) and mapping quality (AQ); then read pairs are classified as specific, or non-specific, if the sum of the distances of each mapped read from the nearest restriction site is smaller, or greater, than the given size selection parameter (RSP, as described by Yaffe and Tanay, 2011). Only restriction fragments with coverage of specific reads significantly higher than the coverage of non-specific reads are retained (FLF). Hi-C experimental biases are corrected following the approach described in (Jin et al., 2013), which estimates the expected random contact frequencies accounting for mappability, GC content, fragment length and, for intra-chromosomal read pairs, fragment distance. To estimate the aforementioned biases for different genomes and restriction enzymes, the tool provides an additional package called 'hippie_gc_mapp' (github.com/yihchii/hippie/tree/master/hippie_gc_mapp). We used 'hippie_gc_mapp' to create all the necessary annotation files for the fly genome (dm3; some regular expression had to be adjusted in the Perl scripts) as HIPPIE provides pre-computed annotations only for human (hg19). Significant interactions are detected by fitting a negative binomial distribution, where the mean is estimated from the random expected contact frequencies (defined by the background model), whereas the overdispersion parameter is fixed and taken from Jin et al.4. The output of HIPPIE is a set of restriction fragment-based interactions (inter- and intra-chromosomal) with the associated p-value. In this comparison, we used a more conservative p-value threshold (0.01) as respect to that originally proposed (0.1) to select significant interactions. Indeed, at a p-value≤0.1, HIPPIE called an unrealistically high number of trans interactions (e.g., >300,000 in Rao GM12878). To compare the results of HIPPIE with those of the other tools, significant interactions were mapped from the restriction fragment level to genomic bins.

2.4.5 diffHic

diffHic (Lun et al., 2015; v.1.0.0) is a Bioconductor R package (bioconductor.org/packages/release/bioc/html/diffHic.html) running on R>3.2.0. The user-guide is comprehensive and detailed, although lacking some suggestions on how to tune the various parameters at different data resolutions and on how to set cut-off threshold to call significant interactions. Before alignment reads are split based on the restriction enzyme re-ligation signature, they are then mapped to the genome with Bowtie 2 in single-end mode; finally all aligned reads are organized in a paired-end BAM file. Further processing with Picard Tools (v1.106; FixMateInformation and MarkDuplicates functions) and SAMtools (sort and merge functions) is needed to prepare the input and run diffHic. Statistics regarding the number of mapped, unmapped and chimeric reads are reported for each sample, PCR duplicates were removed and reads with a mapping quality of less than 10 were discarded, as suggested in the user guide. Then, read pairs were classified as inward (potential dangling ends), outward (potential self-circles) or same-strand. diffHic implements a filter based on strand orientation and distance between mates (insert size), to exclude read pairs arising from uncut DNA. Inward and outward reads were filtered out using the same thresholds described in Jin et al. (2013) (i.e., a distance between mates less than 1000 bp for inward and 25000 bp for outward.

Then, the interactions were binned; note that the boundary of each bin is rounded to the closest restriction site. The package features a function implementing implicit normalization (iterative correction) but loop calling has to be performed on not normalized data. diffHic adopts a "local" approach to identify significant loops, looking for bin pairs that have substantially more reads than their neighbors. The enrichment value for each loop is calculated as the log-fold change between the abundance (number of read pairs) of the target bin pair and the region of the neighborhood (donut) with the largest abundance. The size of the donut was fixed at 2 Mb (at 1 Mb bin size), 120 kb (at 40 kb bin size) and 35 kb (at 5 kb bin size) respectively, based on the values described in (Rao et al., 2014) for HICCUPs. Only loops with enrichment over the donut of 0.5, supported by at least 5 read pairs and at a distance from the diagonal of at least two bins, as suggested in the user guide, were considered. Since no statistical test is performed, no significance value is returned. All the interactions (cis and trans ones) were considered for subsequent analysis. To make results more comparable with the other tools, bins were mapped back to fixed size bins.

During this pre-processing step, some Picard functions returned warnings that the author of diffHic suggested could be ignored. Overall, diffHic is very fast and efficient in memory usage, even for high-resolution datasets.

2.4.6 HiCCUPS (Juicer)

HiCCUPS (Rao et al., 2014; github.com/theaidenlab/juicer/wiki/HiCCUPS) is the algorithm for finding chromatin loops of the Juicer pipeline (Durand et al., 2016). Juicer offers an implementation for different jobs schedulers, which makes it easy to adapt the pipeline to most operating systems.

Juicer (version 1.5) is a pipeline that, starting from raw sequencing files, generates normalized contact matrices at several resolutions and includes downstream analysis methods for identifying looping interactions (HiCCUPS; Hi-C Computational Unbiased Peak Search), described here, as well as TADs (Arrowhead), described below among TAD callers. The pipeline aligns raw reads using Burrows-Wheeler Aligner (BWA; Li and Durbin, 2010) algorithm. For the alignment, the authors suggested to use BWA aln for short reads (-r parameter) and BWA mem (default) for long reads. Since Juicer manual does not indicate a specific threshold for defining long as compared to short reads, in our analyses, we followed BWA manual, which recommends BWA mem for 70 bp or longer reads, whereas BWA aln was used for datasets with shorter reads. Juicer pairs the reads, handles chimeras, and merges and sorts the reads to filter out PCR duplicates and read pairs that can be mapped to more than three locations. To generate the restriction site file for a given genome and restriction enzyme, we used the generate_site_positions.py script of Juicer. For the subsequent analysis steps, the pipeline uses three tools, which are part of the Juicer software suite: Juicer Tools Pre, HiCCUPS, and Arrowhead. Juicer Tools Pre is implemented in Java and takes as input the filtered read pairs for binning and normalization. It performs an additional filter removing reads with a mapping quality lower than 1 (or lower than 30; -q option). Here, we used a mapping quality of 1 for this additional filtering. Juicer Pre can perform different normalization strategies, as Vanilla Coverage and Knight-Ruiz matrix balancing, to create the normalized Hi-C contact matrix stored in a .hic binary file. Normalized matrices can be obtained at various resolution setting the –r parameter. By default, HiCCUPS takes in input the contact matrix normalized with Knight-Ruiz matrix balancing (.hic file) to identify cis chromatin interactions. HiCCUPS searches regions that, in the contact matrix, are enriched with respect to the local background implementing the method described in (Rao et al., 2014). Briefly, peaks are identified by detecting pixels enriched with respect to four neighboring areas given the width of the peak (p parameter) and the window size (i parameter). Statistically significant peaks are called using a modified Benjamini-Hochberg FDR and adjacent significant peaks are aggregated into clusters whose centroid constitutes HiCCUPS output. Here, at 5kb resolution, we used the default parameters; at 40kb, we set the peak width p=1, the window size i=3, and the distance for merging to centroid d=80,000; at 1Mb we set p=1, i=2, and d=2,000,000. At all resolutions, we set the FDR threshold f to 0.1. We called interactions at the given resolution, without using HiCCUPS combination of peak annotations at different resolutions. HiCCUPS returns as output the genomic coordinates of interacting loci, which can be directly visualized in another tool from Aiden lab, Juicebox (Durand et al., 2016).

2.5 TAD callers

2.5.1 HiCseg

HiCseg (Lévy-Leduc et al. 2014; v1.1) is an R package (cran.r-project.org/web/packages/HiCseg/index.html) that depends on R≥2.10. It takes as input either raw or normalized Hi-C matrices in square matrix format and performs a 2D-segmentation based on a maximum likelihood approach, in order to partition each chromosome in its constituent TADs. Contrarily from other methods, HiCseg does not summarize Hi-C data in a 1D index which is then segmented, but applies a 2D segmentation directly to the Hi-C matrix, as performed in image processing. Optimal segmentation is obtained using dynamic programming. We used as input raw Hi-C contact matrices processed with hicpipe and a Poisson distribution parameter. The maximum number of change-points (i.e., TAD borders) was set to 1/3 of the matrix size in bins, as suggested by the author. Finally, a block-diagonal (D) model was chosen.

The output consists in a set of intervals representing the estimated change points with their relative values of log-likelihood. HiCseg can process either raw or normalized data changing the type of data distribution accordingly (i.e. Poisson or negative binomial for raw data, Gaussian for normalized data). However, observed over expected normalized data should probably be log transformed to obtain distribution of values closer to Gaussian. We chose instead to run HiCseg on the observed counts matrix because when running HiCseg with Gaussian distribution parameter we obtained unrealistic TAD calling results (either with or without log transformation of normalized values). For instance, when applied on matrixes normalized with hicpipe, we observed that on some datasets (e.g. Rao IMR90 and GM12878) HiCseg identified, in all samples, the maximum number of TADs allowed and, consequently, extremely small TADs, all composed of 2 or 3 bins.

2.5.2 TADbit

TADbit (Serra et al., 2017; alpha version 360) is a Python package (github.com/3DGenomes/TADbit) that includes modules to identify TADs from Hi-C data. The documentation is very comprehensive and the developers responsive and helpful. TADbit takes raw symmetric matrices as input and returns a delimited list of domains spanning from one bin coordinate to another with a confidence score assigned to each domain span. Although TADbit contains an alignment module, which uses GEM Mapper for iterative alignment, here we used only its TAD calling algorithm. As input, we used hicpipe observed interactions since TADbit expects not normalized discrete count values as input. In this case, TADbit employs a separate normalization algorithm termed as "Visibility" normalization, which the authors state is similar to the Iterative Correction algorithm (Imakaev et al., 2012). Moreover, it implements a breakpoint detection method that identifies the optimal segmentation of the chromosome under a BIC-penalized likelihood. The maximum TAD size, which defaults to the entire chromosome length, and the possibility to identify centromeric regions are the two primary parameters that affect the TADs calling. Here, we did not set any maximum limit for TAD sizes and set the parameter to identify centromeric regions to TRUE. The documentation does not explicitly state a procedure for converting these bin coordinates to genomic coordinates, but a closer look at the Python code provides useful hints in this regard. The package is also equipped with multicore capabilities allowing for parallelization.

2.5.3 DomainCaller

DomainCaller (Dixon et al., 2012) is a set of MATLAB and Perl scripts (chromosome.sdsc.edu/mouse/hi-c/download.html).

The method is based on Hidden Markov Model segmentation of the "Directionality Index" (DI). The package takes as input a symmetric interaction matrix and its genomic coordinates (coordinates can be either continuous or discontinuous) to compute the DI, which is then used by the Hidden Markov Model for predicting domains. The DI is a score quantifying the bias in downstream versus upstream contact probabilities for each bin, within a user-defined window of maximum distance. DomainCaller is a single scale algorithm that calls TADs by computing the DI at a specific window size. In many cases the called domains are discontinuous (the entire genome could not be partitioned into domains) as the algorithm allows gaps between TADs. For processing the datasets, we used hicpipe normalized matrices and the default parameters of the package, i.e. a window size of 2Mb for defining the "Directionality Index" and equal probability (0.33) for all three bias states (upstream, downstream, and no bias). We changed the window size to 5Mb only for the Lieberman-Aiden data to compute the DI on a minimum of 5 bins. The usage of each script and the required file formats are described in a readme file. Similar to what reported in (Rao et al., 2014), we found DomainCaller to occasionally fail for large M values (as defined the in DomainCaller code). To solve this problem (Rao et al., 2013) modified the script to stop at the largest M before the failure, if it failed before reaching the default maximum M value (M=20). Instead, we found that the failure was due to a call to a random number generation function (randp) nested within the "mixGaussFit" function. This random number generation occasionally resulted in causing a division by zero and, thus, the script failure. To avoid modifying the original algorithm as done in (Rao et al., 2014), we implemented a try and catch solution where the call to "mixGaussFit" is just repeated if the initial call fails due to the random number generation causing a division by zero. The maximum number of repetition was set to 10,000 attempts. This fix allowed the Hidden Markov Model to finish processing all chromosomes across all datasets.

2.5.4 InsulationScore

InsulationScore (Crane et al., 2015; v1.0.0) is a segmentation algorithm implemented in Perl (github.com/dekkerlab/crane-nature-2015) that identifies TADs within normalized Hi-C matrices. It requires normalized Hi-C matrices in square matrix format as input.

The parameter settings are extensively presented in the original publication and briefly described in accompanying usage documentation.

It uses a sliding square (insulation square), which is moved along the matrix diagonal, computes the mean of the contact signal inside the window and assigns an insulation score to each bin along the diagonal, thus obtaining a 1D insulation vector. Insulation scores are then normalized calculating the log2 ratio between the bin score and the mean across all the insulation vector values. To facilitate boundaries identification, an insulation delta vector is further calculated from the insulation vector. The delta vector is calculated using a second sliding window, the insulation delta window, which quantifies the difference of insulation change on the left and right side of each bin. All the zero-crossing values at valleys in the delta vector are extracted and those with boundary strength >0.1 are called as boundaries. The insulation square was set to 5 Mb for datasets with 1 Mb resolution and to 1Mb for datasets with 40 kb resolution, as in (Schmitt et al, 2016). The insulation delta span was set to 2 Mb at 1 Mb resolution and to 200 kb at 40 kb resolution, as in (Schmitt et al, 2016). Default settings were used for insulation mode, noise threshold (min. depth of valley), and boundary margin of error (with values "mean", 0.1, and 0, respectively), at all resolutions. The software returns several files containing the insulation score and the delta values for each bin and the coordinates and insulation score of all called boundaries.

A limitation of the insulation score is that no boundaries can be called in the first and last portion of the matrix, corresponding in size to the insulation square value.

2.5.5 Arrowhead (Juicer)

Arrowhead (Rao et al., 2014; Durand et al., 2016) is a Java software which can be run as part of Juicer or as a standalone (github.com/theaidenlab/juicer/wiki/Arrowhead). It does not require installation and has a detailed online documentation. In input, it requires a normalized matrix created by Juicer Tools Pre (.hic format). Arrowhead is part of the Juicer suite of tools for Hi-C data analysis and visualization that implements the TAD calling strategy originally presented in (Rao et al., 2014). In particular, this method is based on the Arrowhead transformation of Hi-C contact matrix, which results in translating the patterns of TAD domains from "squares", along the diagonal, to "triangles" of high or low signal, thus resulting in arrows-like patterns. For each pair of loci, potential TAD boundaries, the algorithm computes specific scores (sum of value signs, sum of values and variance) for the "triangles" designed around the pair of loci, thus exploring the definition of TADs at multiple scales. Here, we first converted the hicpipe normalized matrix into the short format with score and then used it in Juicer Tools Pre to create the .hic file imposing no normalization (-n parameter). In Arrowhead, we set the resolution (-r parameter) and the normalization (-k=NONE), while we left the default for the size of the sliding window (-m parameter), as, in our hands, changes in the size of the sliding window did not impact the number and characteristics of the called TADs. In order to run Arrowhead also for sparse matrices (i.e., Hi-C matrices where most of the bins do not have enough sequencing coverage) we added the parameter --ignore sparsity.

In output, Arrowhead returns the genomic coordinates of TADs, which can be directly visualized by another program developed by the Aiden lab, Juicebox (Durand et al., 2016).

2.5.6 TADtree

TADtree (Weinreb et al., 2015) is written in Python, does not require any installation (compbio.cs.brown.edu/projects/tadtree), and is meant to identify hierarchical topological domains from Hi-C data. For each sample, it takes as input the normalized Hi-C matrices in square matrix format and a control file. The control file includes, among other things, the information on the matrices to analyze, the maximum number of TADs to search in each chromosome (based on the chromosomes size), and the maximum TAD size.

TADtree is based on a 1D boundary index similar to the one developed by (Sauria et al., 2015), integrated in an objective function that allows the identification of nested TADs, differently from all the other available tools that find sets of non-overlapping TADs. It is based on the observation that average enrichment of intra-TAD contacts grows linearly with distance, but when one TAD lies inside another, its enrichment grows at a faster rate. The best TAD hierarchy is determined using a dynamic programming algorithm. The program requires as input a normalized contact matrix (in this study normalized with hicpipe) and a control file with various parameters: as suggested by the author, the M, p, q, and gamma parameters were left as in the example control file. The max TAD size (S parameter, expressed in number of bins), at 40 kb resolution, was set to 50 (as in the example control file) and, at 1 Mb resolution, to 5 Mb (as observed in Supplementary Figure S9A of Dixon et al., 2012). Finally, the N parameter (maximum number of TADs to compute for each chromosome) was set to 6 TADs/Mb, as suggested in the original paper. TADtree gives as output, for each chromosome, N duplicate-filtered set of TADs (with TAD borders written as bin numbers), together with a file containing the percentage of duplicates in the unfiltered sets. Duplicate TADs are pairs of TADs with both borders at less than one bin apart from each other. As suggested by the author, the TAD set with the highest N value for which less than 1-2% of all outputted TADs were duplicates, was chosen for subsequent analysis and converted to genomic coordinates (as a side note, the choice of the best list of TADs among the many that are given as output requires some basic coding skills as it is not implemented in TADtree). Since reported TAD borders are zero-based, we added 1 bin before the conversion to genomic coordinates.

2.5.7 Armatus

Armatus (Filippova et al., 2014; v2.0) is a command-line software implemented in C++ (github.com/kingsfordgroup/armatus) that requires C++11 and Boost libraries to be installed. The readme file contains information on installation, input formats, and parameter settings. It adopts a multiscale approach to identify domains conserved across various resolutions by adjusting a single scale parameter (gamma). It is based on a score function that quantifies the quality of a domain based on the local density of interactions. The algorithm calculates TADs at different resolutions (i.e., different values of gamma from zero up to a user defined maximum value) and finds a consensus set of TADs persistent across resolutions. We used Hi-C contact matrices normalized with hicpipe as input and set gamma-max to 0.05 at 1 Mb and to 0.3 at 40 kb resolutions. As suggested by the authors, the values of gamma-max were chosen by analyzing random samples from each dataset with different values of gamma-max, picking the smallest value of gamma that allowed having a median TAD size of at least 3 bins in the consensus set. All other parameters were left as default. Armatus reports a consensus set of TADs, using as TAD border coordinates the start of the bin representing the boundary; therefore, the right boundary of the TAD was adjusted for the bin size in order to consider the bin end. We were not able to obtain TADs for chromosomes 16 to 22 of Lieberman-Aiden dataset since the matrices of these chromosomes (processed at 1 Mb of resolution) contain less than 101 rows and, although not specified in any documentation, we realized that Armatus does not process matrices with less than 101 rows.

2.6 Generation of random sets of chromatin interactions

Empirical p-values were estimated with random permutations of interactions. Briefly, for each dataset, cell type, and data analysis method, we defined, for each sample, a random set of cis interactions by keeping constant the sample-specific number of interactions and the sample-specific distribution of distances between anchoring points. The first of the two anchoring points for each interaction was randomly selected from the pool of detectable anchoring points, defined as any genomic bin that was called as anchoring point in any sample from the same dataset and cell type. The second anchoring point was randomly defined by sampling from the observed distribution of anchoring point distances. The resulting sets of random interactions were then used to compute random Jaccard Index values in pairwise comparisons. The random sampling of interactions was repeated 1000 times to obtain a null distribution of randomly expected Jaccard Index values for each pairwise comparison. The empirical p-value is estimated as the probability of observing a random Jaccard Index value larger than or equal to the observed one. Almost all of the observed Jaccard Index values in the pairwise comparisons are significantly larger than expected by chance. Stacked bars lower than the maximum value are used for samples including one or more replicates with no detected interactions (which were not included in the pairwise comparisons).

2.7 Integration with chromatin states

Chromatin states for IMR90, H1-hESC and GM12878 (15-states model) were downloaded from (Roadmap Epigenomics Consortium et al., 2015). We merged chromatin states into 4 major classes: promoter (Active TSS, Flanking Active TSS, Bivalent/poised TSS Flanking bivalent TSS), enhancers (Enhancers, Genic enhancers, Bivalent enhancers), repressed Polycomb (Repressed Polycomb, Weak repressed Polycomb), and heterochromatin/Low (Heterochromatin, Quiescent/low). We did not consider chromatin states related to transcription and the state ZNF genes + repeats.

Chromatin states for fly late embryos (16 chromatin states) were retrieved from modENCODE35. We merged chromatin states into 4 major classes: promoter (Promoter), enhancer (Enhancer 1, Enhancer 2), repressed Polycomb (PC repressed 1, PC repressed 2), and heterochromatin/Low (Heterochromatin 1, Heterochromatin 2, Low signal 1, Low signal 2, Low signal 3). We did not consider chromatin states related to transcription.

Each bin was classified based on the overlap with chromatin states, requiring a minimum overlap of 50 bp. A bin can fall entirely in a category or be assigned to multiple categories. Interactions between a bin classified entirely as Promoter or entirely as enhancer and a bin classified entirely as Heterochromatin/Low were considered less likely to occur and labelled as "not expected". Interactions between bins classified as promoter and bins classified as enhancer were counted as promoter-enhancer interactions, even if the bins contained other chromatin states. Interactions defined as Heterochromatin/Low-Heterochromatin/Low comprise only bins classified entirely as Heterochromatin/Low.

For each class of chromatin states, we computed the ratio of observed count of interactions over random expectation taking into account the number of interacting bin pairs and the number of bin pairs belonging to each of the considered chromatin state classes. We considered only genomic bins annotated with a chromatin state. Using the total number of possible bin pairs annotated with a chromatin state as grand total, we constructed a two by two contingency table by counting the number of i) bin pairs belonging to the specific the chromatin state classes (each of the three classes above considered separately), and ii) bin pairs involved in an interaction. The contingency table was used to estimate the background model of randomly expected number of bin pairs belonging to the chromatin state class and involved in an interaction, which was then compared to the observed number to obtain the ratio of observed/expected counts.

Regarding TAD callers, ChIPpeakanno R package was used to compare chromatin interactions with chromatin states and TAD boundaries with CTCF ChIP-seq peaks form ENCODE and BEAF32 peaks from modENCODE (Table 2.4).

Enrichment analysis. For each class of chromatin states, we computed the ratio of observed count of interactions over random expectation taking into account the number of interacting bin pairs and the number of bin pairs belonging to each of the considered chromatin state classes. Briefly, we considered only genomic bins annotated with a chromatin state. Using the total number of possible bin pairs annotated with a chromatin state as grand total, we constructed a 2x2 contingency table by counting the number of i) bin pairs belonging to the specific the chromatin state classes (each of the three classes above considered separately), and ii) bin pairs involved in an interaction. The contingency table was used to estimate the background model of randomly expected number of bin pairs belonging to the chromatin state class and involved in an interaction, which is then compared to the observed number to obtain the ratio of observed/expected counts. The dashed line marks observed over expected ratio equal to 1. With the exception of Sexton and Jin H1-hESC datasets (that contain a single replicate), only interactions conserved in at least 2 replicates within each dataset were classified using the chromatin states.

2.8 Assessment of CTCF-binding motifs orientation

The assessment of the convergent orientation of CTCF-binding motifs was performed only for cis interactions identified in datasets at 5 kb resolution, as a larger binning would result in the presence of too many CTCF motifs to discriminate their orientation. Cell line-specific CTCF ChIP-seq peaks were retrieved from ENCODE (Table 2.4) and further processed with HOMER motif analysis to obtain the coordinates and orientation of CTCF motifs found inside CTCF ChIP-seq peaks (CTCF-binding motifs). For each tool, the cis interactions conserved in at least 2 replicates within each dataset (with the exception of Jin H1-hESC that contains a single replicate) were intersected with the list of CTCF-binding motifs. Finally, as described in (Rao et al., 2014), an interaction was labelled as having a convergent CTCF orientation if the interacting bin closer to the p-terminus of the chromosome contained one CTCF motif on the forward strand (+ orientation) and the interacting bin closer to the q-terminus of the chromosome contained one CTCF motif on the reverse strand (- orientation). Among interactions with a single CTCF-binding motif in each of the two interacting bins, the various methods identified the following percentages of interactions with convergent orientation of CTCF-binding motifs: 96% for HiCCUPS, 52.1% for GOTHiC, 78.2% for HOMER, 52.9% for diffHic, 45.6% for HIPPIE, and 66.7% for Fit-Hi-C (median values computed on all datasets at 5kb resolution).

2.9 Runtimes comparison

To compare the running times, we used IMR90 repB from Dixon2012 dataset and IMR90 repA5 of Rao dataset. The two samples have a similar number of total reads but differ for the protocol used (dilution Hi-C versus In situ), restriction enzyme (6 bp cutter HindIII versus 4 bp cutter MboI) and read length (36 versus 101 bp). IMR90 repB from Dixon2012 was analyzed at 40 kb for both interactions and TADs, whereas IMR90 repA5 from Rao was analyzed at 5 kb for interactions and at 40 kb for TADs. Runtimes were quantified using the "time" function for all those tools running in Bash and with the function "system.time" for R packages. All analyses were run one at a time on a single CPU in a 2.3 GHz Hexadeca-Core AMD Opteron Processor 6276 equipped with LINUX distributions.

2.10 TAD-AH: Topologically Associating Domains Advanced Hierarchy

The majority of available methods for TADs detection call just a single layer of TADs and the algorithms that can detect nested TADs fail to fully gather TADs hierarchical nature. Moreover, none of the existing methods takes into consideration differences in TAD characteristics (e.g., their signal density), which can reflect differences in their biological role as recently speculated by (Rowley et al., 2017; Rao et al., 2017). In the proposed model, chromatin 3D organization is driven by two opposite mechanisms: one, promoted by epigenetics/transcription through phase separation, segregates chromatin into active and inactive compartments. Another mechanism, driven by cohesion/CTCF through loop extrusion, tends to counteract compartments aggregation. Finally, there is still no strategy to perform differential analysis of TADs found in different biological settings.

To address these limitations, I developed TAD-AH (i.e., TADs Advanced Hierarchy), a four-step sequential procedure coded in R (tested on R-3.1.3). In the first step, TAD-AH takes as input TAD calls from two conditions and applies several filters to retrieve the TAD lists that best fit the data. In TAD-AH, TADs are identified using Armatus (Filippova et al., 2014; v2.0) on ICE normalized Hi-C data. Armatus strategy consists in calling TADs in a multi-scale fashion and giving as output a consensus set of TADs, which are conserved across different resolutions. The consensus set portrays just a single layer of TADs though, thus failing to capture the hierarchical nature of these entities. For this reason, I chose to keep all the TADs called at different resolutions (i.e., multi-scale), and use them as input for TAD-AH (Fig. 2.1).

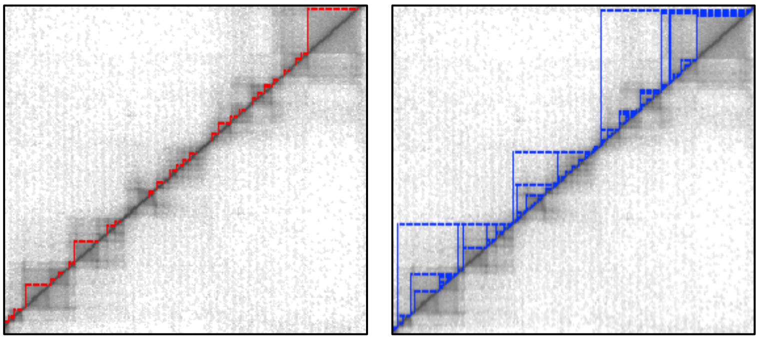

Fig. 2.1: Heatmap of a Hi-C contact matrix at 40 kb resolution, with Armatus consensus (red) and multi-scale (blue) TAD calls.

However, Armatus multi-scale TAD calls cannot be used as is, since Armatus was not designed to consider the different layers of TADs together and it can occur that some TADs cross each other or present almost overlapping boundaries. Thus, the first step of TAD-AH is intended to filter out artifacts and redundant information.

TAD-AH takes as input TAD lists (chromosome - start coordinate - end coordinate format) and NxN Hi-C matrices – preferentially ICE normalized, as this normalization enables the comparison of the signal coming from matrices from different conditions, as described in (Imakaev et al., 2012).

First, TADs that are too small (yet indistinguishable from the signal coming from the diagonal) or too big (i.e., exceeding a size of 4 Mb) are discarded. Secondly, TAD-AH removes TAD calls generated from Armatus misinterpretation of the presence of empty rows (e.g. from genomic regions with low mappability) in the Hi-C matrix as a variation in the insulation value used to determine TADs presence. Then, TAD-AH gets rid of those TADs whose boundaries cross other domains boundaries. It does so by either discarding i) TADs that cross more than one TAD; ii) TADs with higher insulation scores at their boundaries – estimated as in Crane et al., 2015 – or, if the two overlapping TADs are near the start/end of the Hi-C matrix (i.e., where insulation score computation is not possible), iii) TADs with lower inner contact signal. Another imprecision in TAD calls is represented by the presence of TADs that differ just for a shift of one of its boundaries coordinate (i.e., duplicates): in this case, the TAD with lower tip signal compared to its surrounding regions (i.e., HiCCUPS strategy implementation – Rao et al., 2014) is filtered out. After filtering, TADs are classified into either dense or loop-mediated: dense TADs tend to have a tip mean signal comparable to that of its surrounding genomic regions, whereas loop-mediated TADs have much more signal at the tip than in the mid portion of the TAD. TADs classification is based on the ratio between tip mean signal and the signal coming from the middle portion of the TAD, which I refer to as TAD density score. Both the size of the tip region and that of the middle TAD portion depend on the size of the considered TAD: the tip region is set to 1/6 of the TAD size, while the considered middle portion is set to 1/2 of the tip size, with the restriction that, for TADs whose size exceeds 300 kb, the tip region is set to 40 kb. This choice was based on the inspection of high-resolution Hi-C matrices, where it is improbable to find loop-mediated TADs with a tip signal enrichment extending beyond 40 kb. The threshold to divide TADs into dense and loop-mediated is set to a density score equal or higher than 1.4, which can be considered quite stringent, but should avoid misclassification of small TADs. Once classified into dense or loop-mediated, TADs hierarchy is reconstructed.

The third step in TAD-AH analysis is the integration of the TAD lists with other omics data, as gene expression (RNA-seq) and histone modifications or transcription factors binding (ChIP-seq) data. For the integration with gene expression data, which requires as input a GTF file with genes start-end coordinates and a table with the gene counts from the different conditions, genes are overlapped to TADs coordinates and assigned to the smaller TAD in the hierarchy in which the gene is fully contained. Gene expression levels are then averaged for each TAD considering all the genes inside it. For the integration with ChIP-seq data, which requires as input a BED file with peaks coordinates, ChIP-seq peaks are overlapped to TADs coordinates and enrichment of binding around TAD boundaries is computed considering the ChIP-seq peaks summit (i.e., the peak coordinate with the highest signal density; for broad peaks, as those coming from some kinds of histone modifications, the enrichment around the boundaries can be calculated considering the peak middle point as the summit).

In the last step, TAD-AH performs differential analysis, which identifies TADs that are retained, lost or acquired between conditions, or that change from dense to loop-mediated and vice versa. Small shifts in TAD boundaries coordinates (e.g., up to 3 bins or 1 bin at 4 kb or 40 kb resolution, respectively) are tolerated, and TADs are considered conserved between conditions.

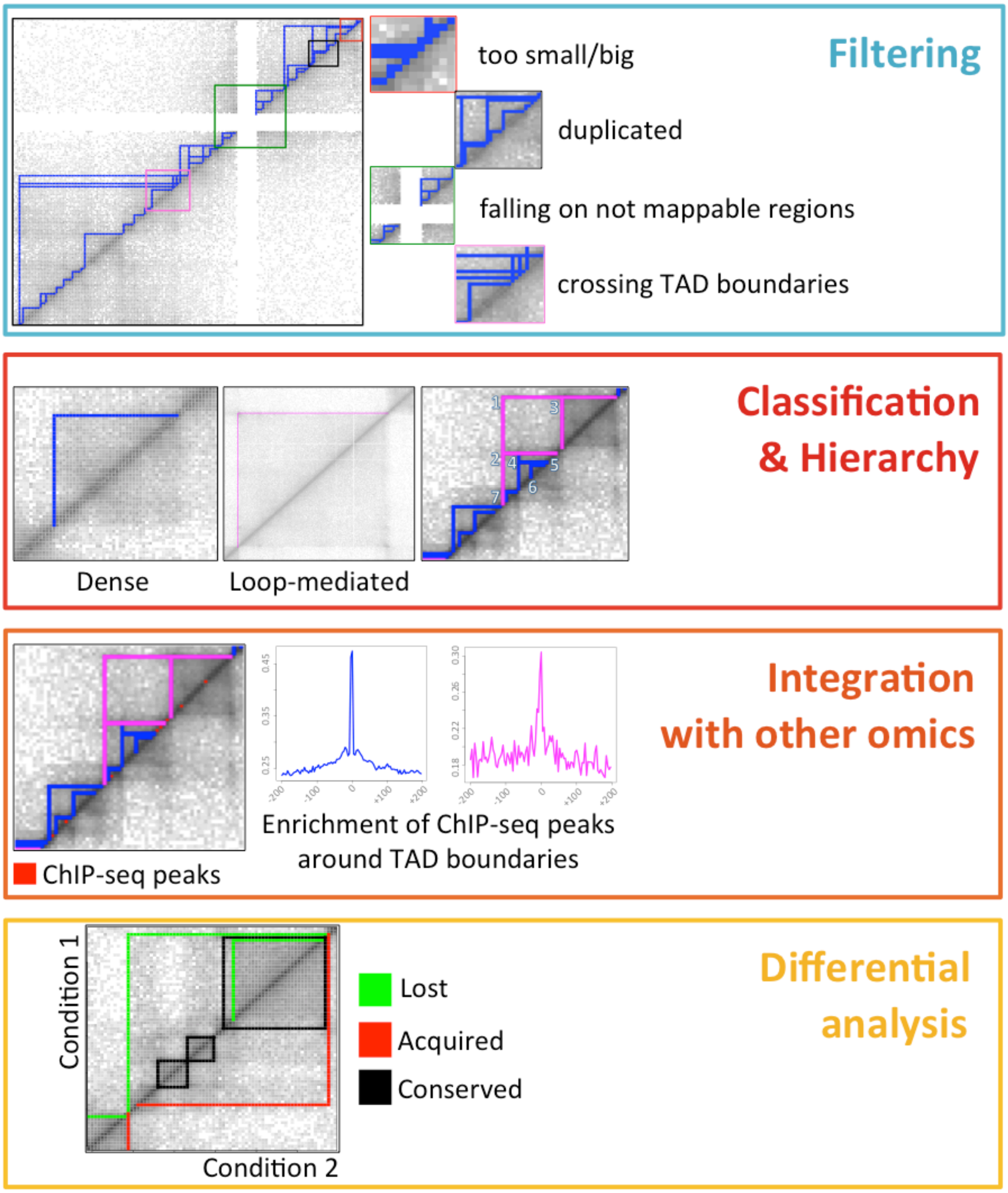

Figure 2.2: TAD-AH analysis steps. TADs are initially filtered to remove artifacts and duplicated TAD calls. Secondly, TADs are divided in dense or loop-mediated based on their tip to mid TAD signal ratio, and TADs hierarchies are reconstructed. In the third step, TADs are integrated to other genomic data, as RNA-seq and ChIP-seq. Finally, TADs from different conditions are compared, to find domains that are conserved, lost or acquired.

I tested TAD-AH performance on Hi-C data at 4 kb resolution derived from IMR90 cells reprogrammed to myoblasts and then differentiated to myotubes, following the overexpression of the myogenic master regulator, MYOD (Fig. 2.3; see section 2.1.3 of Materials and Methods for details on the generation of Hi-C data). Results were integrated with gene expression and MYOD binding data (RNA-seq and ChIP-seq data, respectively; see section 2.1.3 of Materials and Methods for details on data generation).

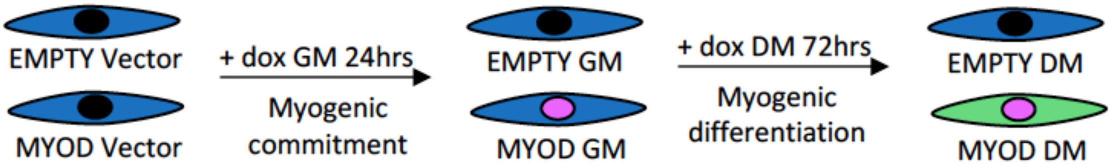

Figure 2.3: Human fibroblasts (IMR90 cells; left), transfected either with an empty vector or a vector containing muscle cells master regulator gene MYOD, were put in growth media (GM) with doxycycline for 24 hours. Cells carrying the MYOD vector reprogrammed into myoblasts (middle) and, when put in differentiation media (DM), differentiated to myotubes (right).

Chapter 3

Results

The results illustrated in this chapter are divided in two parts. The first section describes the analyses performed for the comparison of 6 pipelines to find chromatin interactions and 7 for TADs identification from Hi-C data, applied to six public datasets and to simulated Hi-C data. The results relatively to this part of my project have been published in (Forcato et al., 2017). The second part shows the results relatively to TAD characterization in human fibroblasts (IMR90 cells) before and after trans-differentiation by muscle stem cells master regulator, MYOD, with a new method for TAD calls refinement, characterization and differential analysis, called TAD-AH.

3.1 Comparison of methods for Hi-C data analysis

3.1.1 Hi-C data preprocessing

The methods described below preprocess Hi-C data using different alignment and filtering strategies (Fig. 3.1). Only few of the selected pipelines implement alignment (namely, HIPPIE, diffHic and Juicer, which covers the preprocessing steps for both HiCCUPS and Arrowhead), while almost all the interaction callers implement filtering and normalization, with Fit-Hi-C representing the sole exception. On the other hand, all the TAD callers, apart from TADbit and Arrowhead, just perform downstream analysis. In order to maximize comparability, we then decided to use a common full-read aligner, Bowtie, to map the reads for the tools that did not implement alignment, and a common pre-processing pipeline, hicpipe, to filter and normalize the matrices later used as input for the TAD callers.

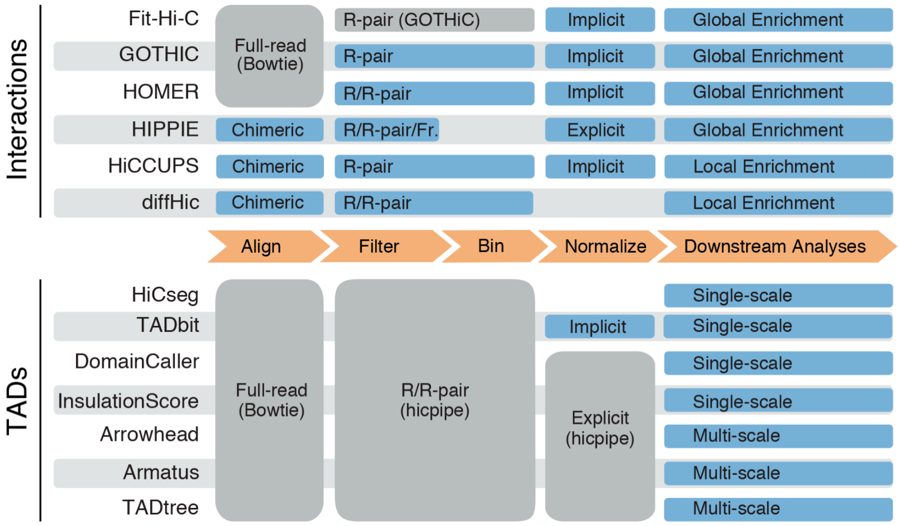

Figure 3.1: Tools for the identification of chromatin interactions and TADs from Hi-C data and key analysis steps (orange arrows). Blue boxes detail the strategy used in each analysis step by each tool. A grey box is used when an external tool is required for a preprocessing step. HIPPIE does not include binning and diffHic and HiCseg do not require the normalization step. Although TADbit and Arrowhead (through Juicer) can perform alignment, filtering, and binning, a uniform preprocessing procedure was used for all TAD callers to maximize comparability. Since most tools perform filtering and binning together, a blue or grey box spanning both steps is used in the schematic workflow. For filtering the following abbreviations are used: read level filtering (R); read-pair level filtering (R-pair); fragment level filtering (Fr.).

For all datasets chimeric alignment performed best, always showing a higher alignment rate respect to the full-read approach. Specifically, HIPPIE aligned on average 18.4% more reads than Bowtie, Juicer 27.4% and diffHic 40.1%. These differences are exacerbated by the read length: when we deal with datasets comprising longer reads we observe how the full-read alignment performance decreases while the chimeric approach see an increase in the number of aligned reads; as an example, diffHic aligner, chimeric Bowtie2, aligns from 30.9% (at 36 bp) to 55.4% (at 101 bp) more reads than Bowtie (Fig. 3.2a).

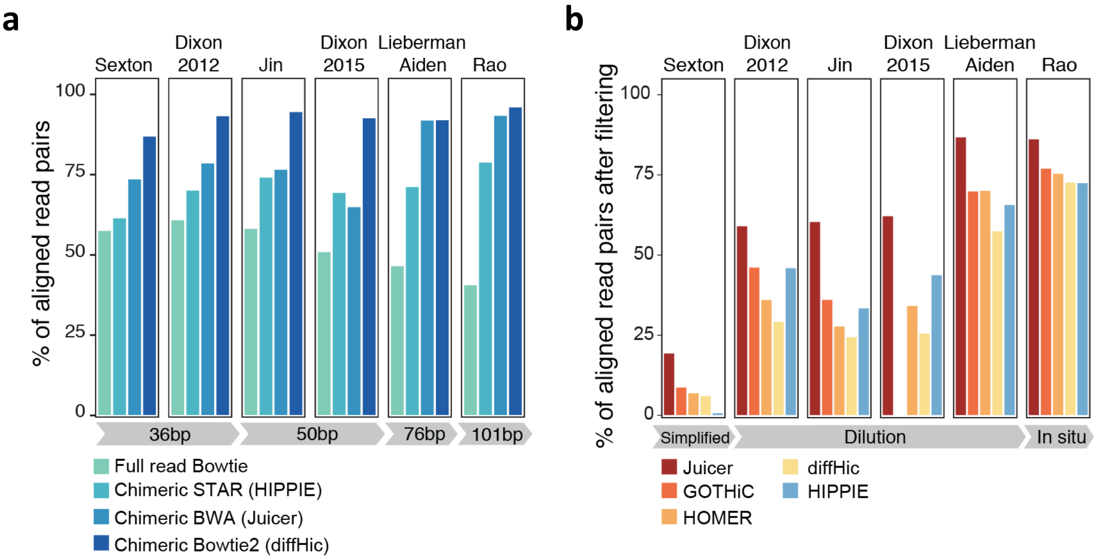

Figure 3.2: a) Median percentage of aligned read pairs (alignment rate) for all datasets ordered by read length (grey arrows at the bottom). At increasing read length, the chimeric approach (chimeric STAR, chimeric BWA, and chimeric Bowtie 2 implemented by HIPPIE, Juicer, and diffHic, respectively) leads to better alignment performance, when compared to full read alignment (Bowtie). Some datasets contain samples with different or mixed read length that were not used when calculating the alignment rate of this figure (i.e., one sample from Dixon 2012; 3 samples from Jin and 1 sample from Dixon 2015; n=36 samples).
b) Median percentage of mapped reads retained after filtering (fraction of usable reads) in each dataset, ordered by experimental protocol (grey arrows at the bottom). This percentage varies from dataset to dataset, depending on the experimental protocol used. GOTHiC could not be applied to Dixon 2015 since the read-pairing step required an amount of memory larger than that commonly available in standard computing servers (i.e., 1 TB of RAM). All n=41 samples were used to calculate the median percentage of mapped reads retained after filtering.

Regarding the filtering step, in all cases Juicer retained the largest number of aligned reads (Fig. 3.2b), both because it aligns more reads (thanks to the chimeric alignment strategy) and as it only filters for PCR duplicates, without discarding other potential artifacts (Fig. 3.3).

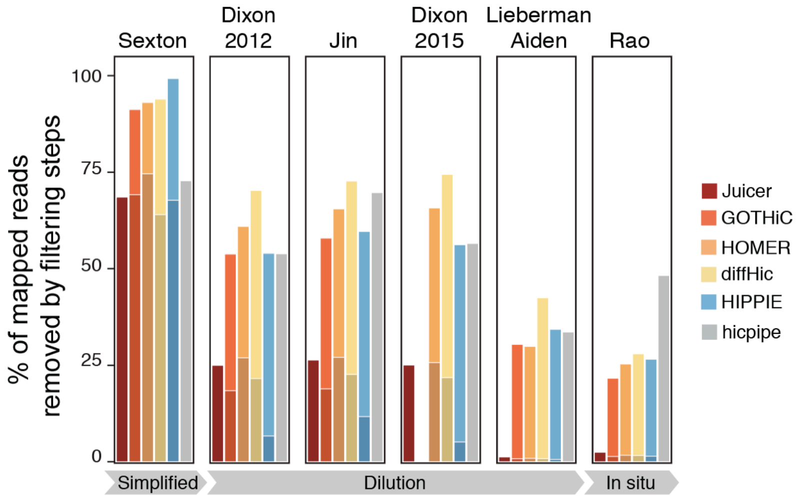

Figure 3.3: Median percentage of mapped reads removed in the filtering step in each study, grouped by experimental protocol (grey arrows). In each bar, the dark shaded portion shows the fraction of reads filtered during PCR duplicates removal; the light shaded part is the fraction of reads discarded by all other filters. Juicer applies only PCR duplicates removal. Filtering results for hicpipe (the method used to generate input matrices to TAD callers) are reported for comparison. Although addressing the filtering in a different way, hicpipe performs similarly to the other tools. GOTHiC could not be applied to Dixon 2015 study since the read-pairing step required more than 1 TB of RAM, i.e. well beyond the amount of memory available in standard computing servers, and prevented the tool from completing the analysis.

diffHic, which, together with HOMER, is the one implementing the major number of filters, generally filtered the highest proportion of aligned reads (from 27% to 94% depending on the dataset; Fig. 3.3), but, given its higher alignment rate, still retained a large number of reads (Fig. 3.4).

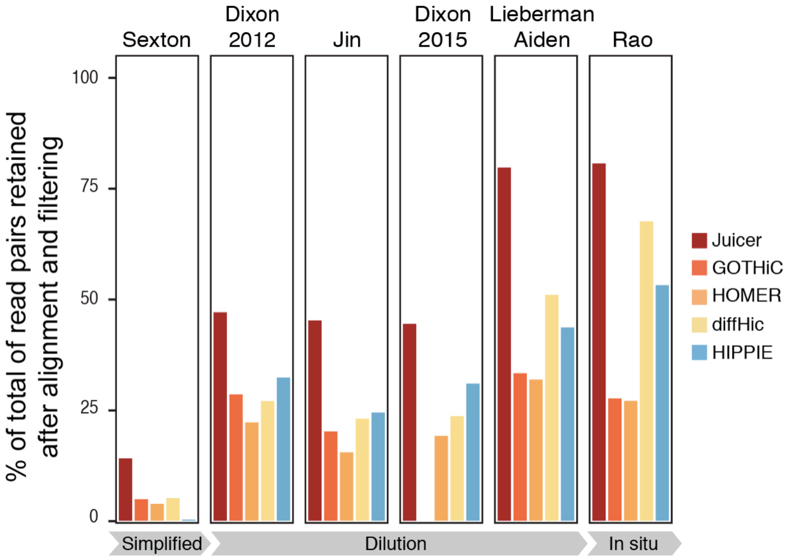

Figure 3.4: Median percentage of total reads retained after alignment and filtering in each study, grouped by experimental protocol (grey arrows). GOTHiC could not be applied to Dixon 2015 study (see Fig. 3.3).

The number of reads retained after the filtering steps was not, as one would expect, much impacted by the different filtering approaches implemented by each tool, but instead by the experimental protocol used to generate the dataset, with in situ Hi-C returning more reads passing the filtering step (>76%; Fig. 3.4). On the other hand, simplified Hi-C protocol performed poorly, with less that 9% of the initial Hi-C reads retained after filtering, mostly due by the abundance of PCR duplicates (>68% of aligned reads; Fig. 3.3).

After filtering, reads were summarized into genomic bins of a fixed size, which was selected based on the one adopted in the original publications. Specifically, interaction callers were tested on datasets analyzed at 1 Mb, 40 kb or 5 kb bin resolution, whereas TAD callers where tested on matrices binned either at 1 Mb or 40 kb (see Table 2.1 of Materials and Methods for further details).

Normalization was performed according to the strategy implemented by each interaction caller, whereas for TAD callers that required a normalized Hi-C matrix as input, data were commonly preprocessed using hicpipe, which implements explicit normalization (Fig. 3.1).

In all cases, we did not assess the effect of different normalization approaches, as comprehensive evaluations of normalization methods have already been reported (Yaffe and Tanay, 2011; Imakaev et al., 2012; Sauria et al., 2015).

3.1.2 Chromatin interactions callers

Several metrics were taken into consideration to measure the efficiency of the methods for chromatin interactions identification. The adopted metrics included the number of found interactions, as well as the distance between the interacting loci; the concordance of the interactions identified for the same cell line in different biological replicates of the same study or across studies at equal resolution; and the enrichment in cell type-specific chromatin states. An additional basis for comparison was the enrichment for interactions found in literature that are known to occur in the cell lines under study or reported to be specific of given cell types at a given physiological state (interaction evidences), validated with 3C-derived (3C, 5C, ChIA-PET) techniques or based on imaging (3D-FISH).

Moreover, we calculated the sensitivity (true positive rate) and precision of the methods in identifying interactions from simulated data.

Starting from experimental data, we quantified the total number of interactions called by each method as a function of the number of reads retained by the filtering step (Fig. 3.5).

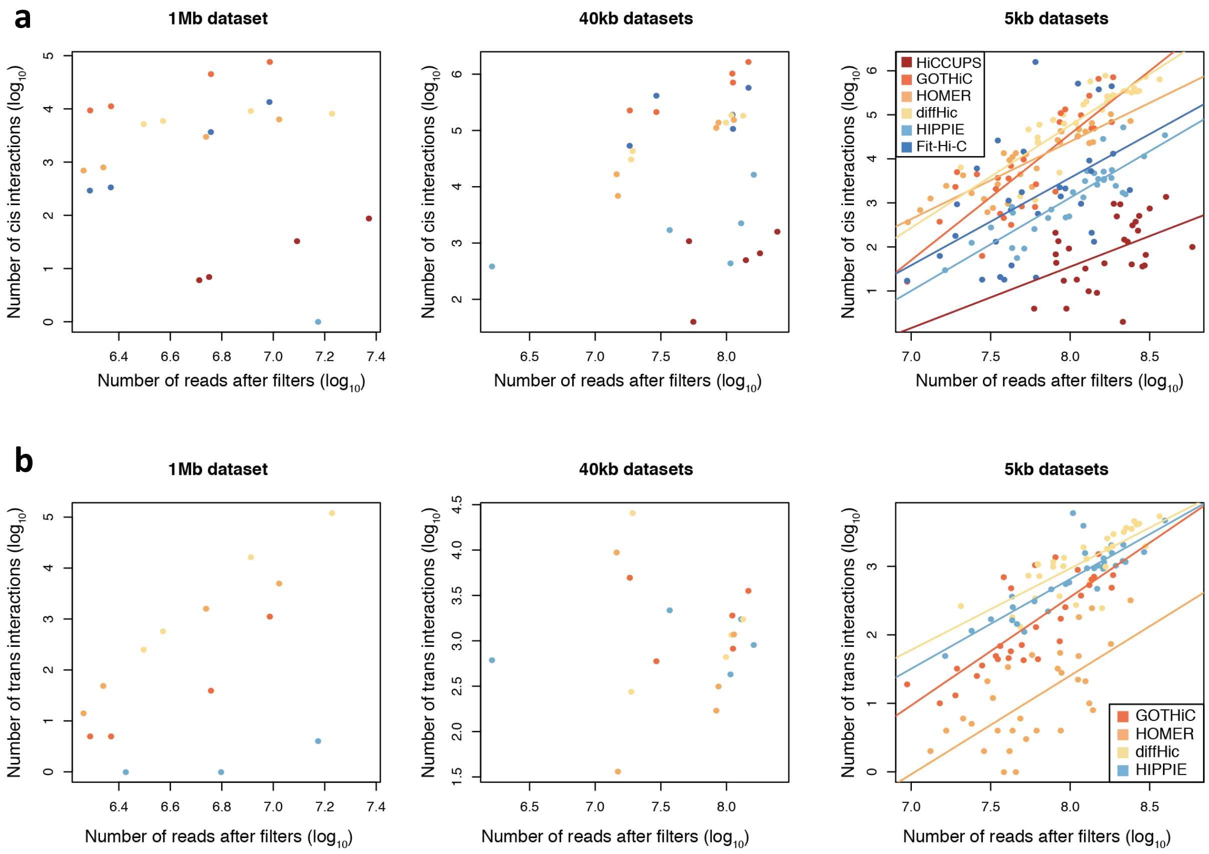

Figure 3.5: Scatter plot of total number of cis interactions called by each method versus the number of reads retained by the filtering step in all datasets at 1 Mb, 40 kb, and 5 kb resolutions. Different points represent sample replicates. Linear interpolation (of log transformed data) is shown as solid line only for datasets at 5 kb, where more data points are available. b) Same as in a) for trans interactions. Trans interactions are not returned by Fit-Hi-C and HiCCUPS.

The number of interactions increased with the number of retained reads for all tools at any resolution, although the rate of increase varied from tool to tool (Fig. 3.5a). For all tools, the rate of increase of the number of found interactions according to the number of retained reads was higher for cis than for trans interactions (Fig. 3.5a and 3.5b, respectively). Consistent with the expectation that 3D interactions mostly occur within chromosomes (cis) rather than between chromosomes (trans), all methods detected more cis than trans interactions. In most datasets, the number of called cis interactions was highest for GOTHiC at all resolutions, followed by Fit-Hi-C at 40 kb and by diffHic at 5 kb (Fig. 3.6). In general, diffHic found the largest number of trans interactions (Fig. 3.6).

HiCCUPS, which combines adjacent points of enriched contact on the Hi-C matrix into a single interaction, identified fewer cis interactions than other tools (Fig. 3.6).

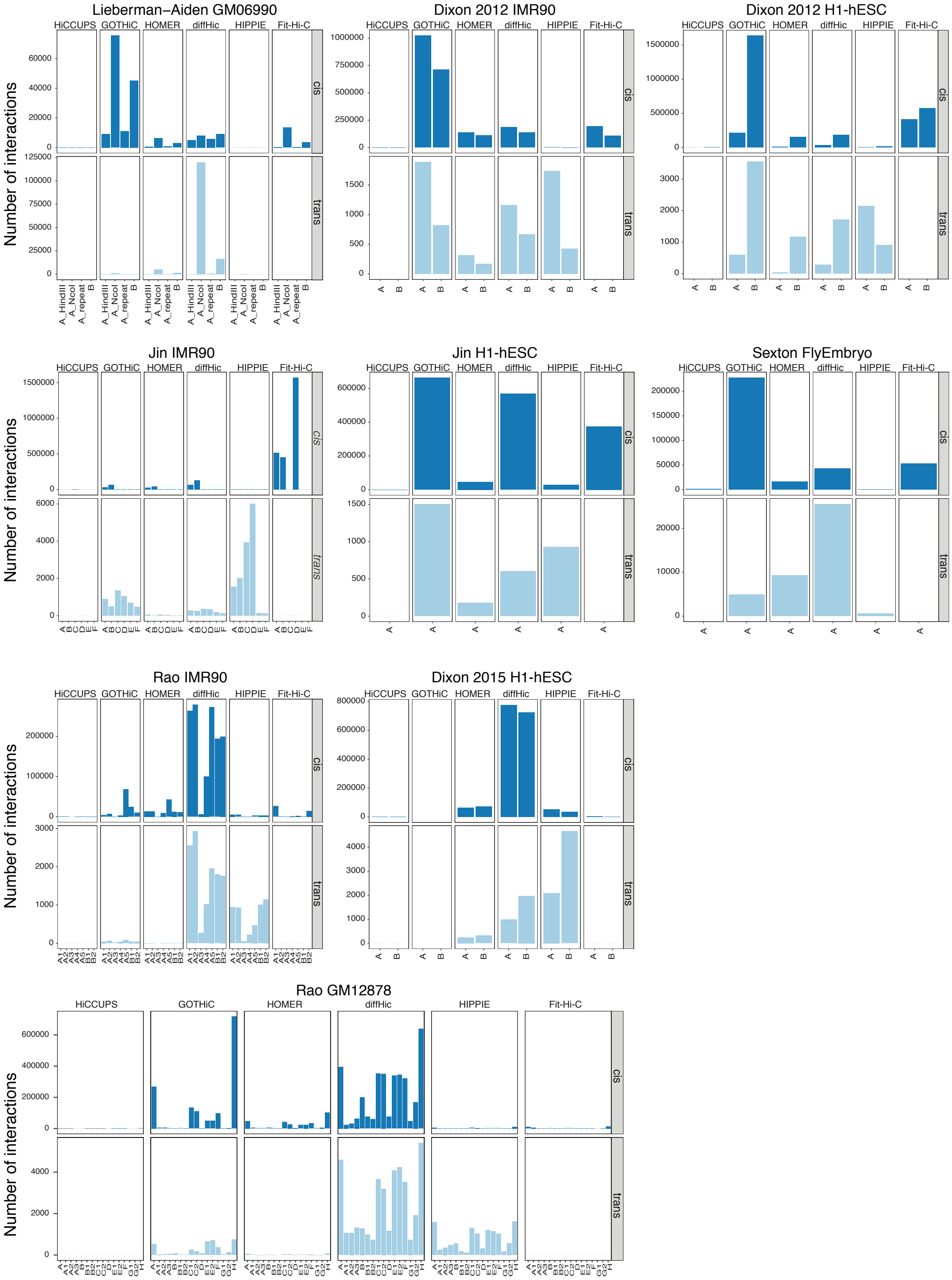

Figure 3.6: Number of cis and trans interactions identified by the various tools in each replicate of each dataset. GOTHiC was not applied to Dixon 2015 (see Fig. 3.3).

When considering the distance between the interacting points in cis, GOTHiC found interactions at shorter mean distance with respect to all other tools, both at 5 and 40 kb resolution (Fig. 3.7). At 5 kb, Fit-Hi-C called interactions at an average distance of more than 10 Mb, which could easily be expected if we consider that Fit-Hi-C is designed to call mid-range interactions (Ay et al., 2014). At a resolution of 1 Mb, with the exception of HIPPIE, all tools detected interactions with an average distance comprised between 10 Mb (GOTHiC and HiCCUPS) and 53 Mb (diffHic) (Fig. 3.7).

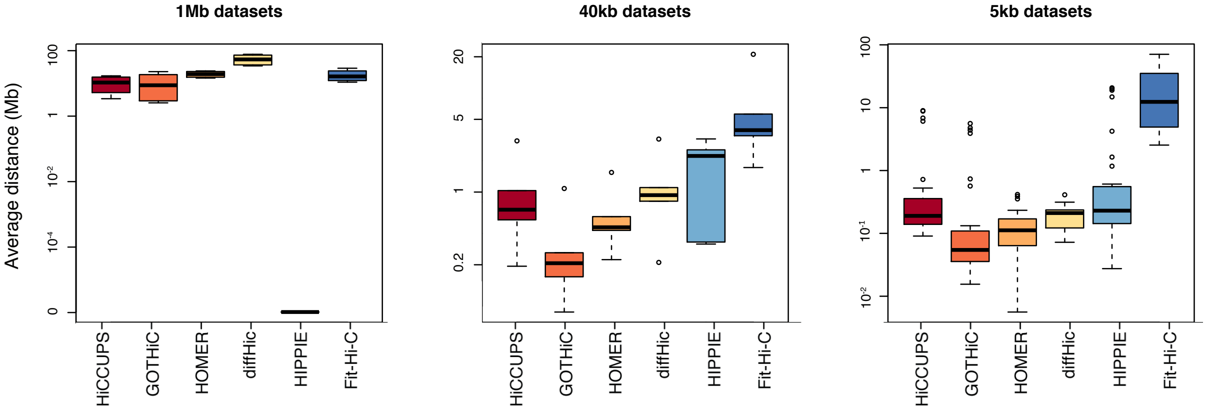

Figure 3.7: Boxplot of average distances between anchoring points in cis interactions (log scale) in sample replicates of all datasets at 1 Mb, 40 kb and 5 kb resolutions. At 1 Mb (Lieberman-Aiden dataset), HIPPIE found just 1 interaction between two adjacent bins.

The diversity in the number of interactions and in the distance between the interacting points found by the different approaches is straightforward in the visual inspection of the contact maps (Fig. 3.8).

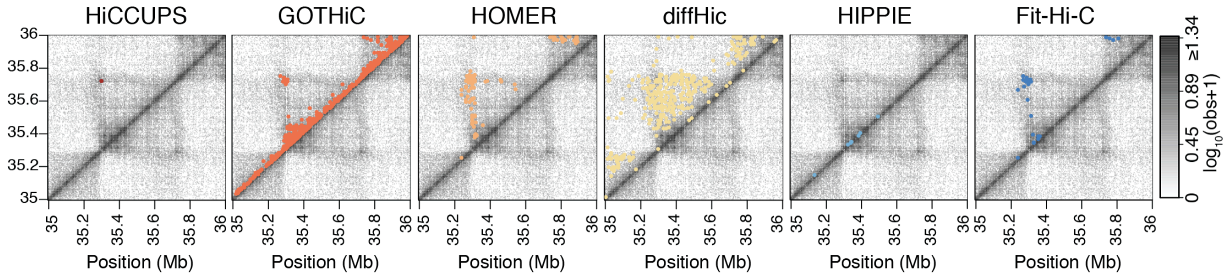

Figure 3.8: Heatmap of the contact matrix of Rao GM12878 replicate H (chr21:35,000,000-36,000,000) at 5 kb resolution. Identified peaks are marked in different colors for the various methods. GOTHiC and diffHic recalled the largest number of interactions while HIPPIE identified fewer interactions than all other tools. HiCCUPS finds only one interaction at the top of the TAD because of its centroid aggregation of nearby peaks. Distances between interactions are smaller for GOTHiC and larger for Fit-Hi-C.

To evaluate the consistency of interactions called in different biological replicates, we computed the similarity coefficient of Jaccard (Jaccard Index, JI), as a measure of the agreement between sets of interactions.

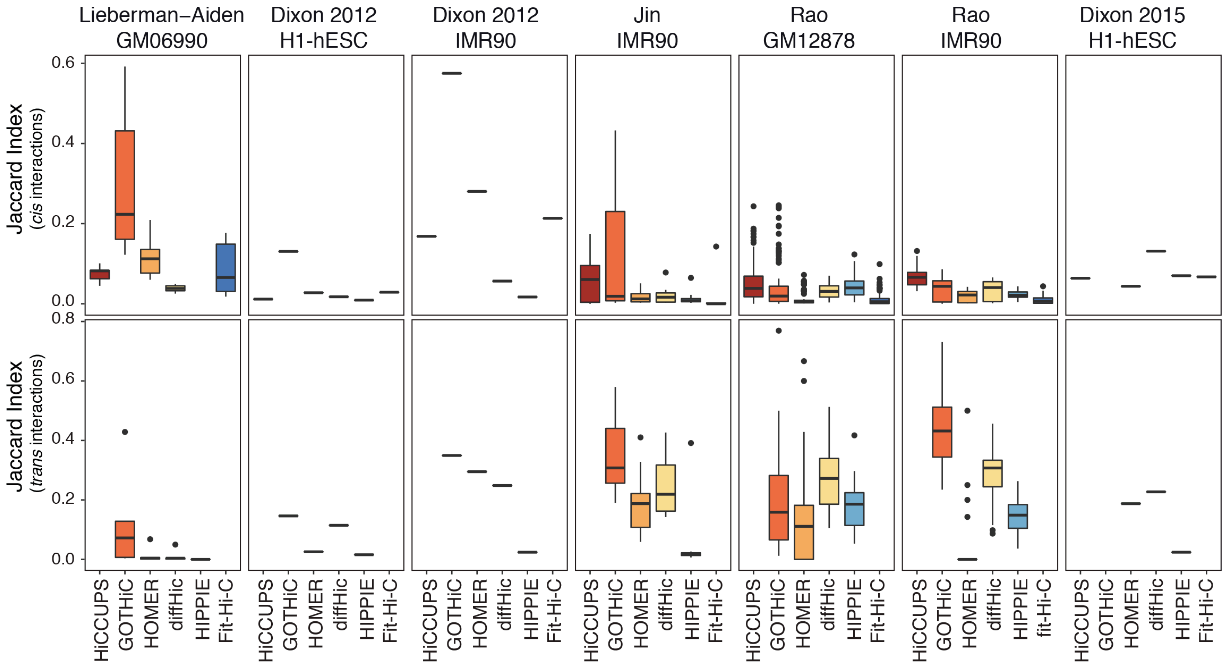

Figure 3.9: Box plots of the Jaccard Index for concordance of cis (upper panels) and trans (lower panels) interaction calls between sample replicates in any dataset (intra-dataset concordance). Jaccard Index was not calculated for GOTHiC in Dixon 2015 (see Supplementary Figure 1) and for HIPPIE in cis interactions of Lieberman-Aiden (see Supplementary Figure 4).

Overall, the reproducibility between biological replicates of the same dataset (intra-dataset) was modest at all resolutions (Fig. 3.9), but still appreciably higher in comparison with random sets of interactions (p-values≤0.001; Fig. 3.10).

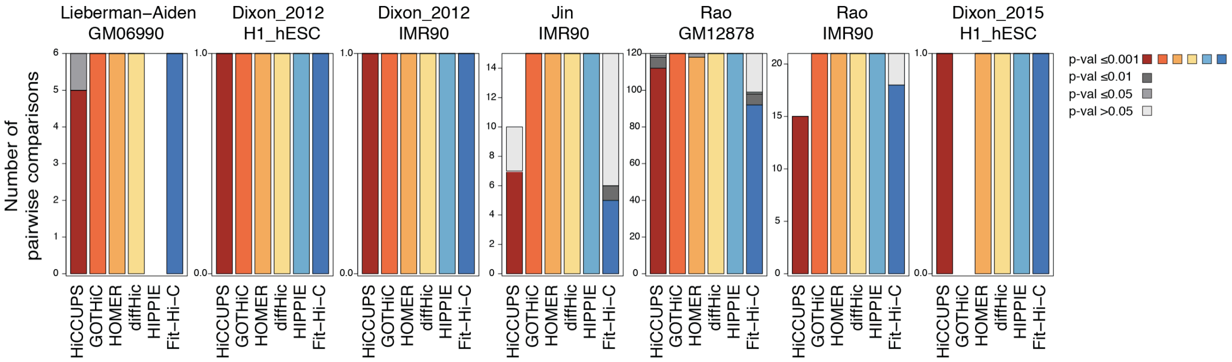

Figure 3.10: Stacked bar plot for the number of pairwise comparisons of cis interactions between replicates stratified by significance. The y-axis scale depends on the number of pairwise comparison per dataset. Bars are colored according to the tool if the comparisons have a Jaccard Index p-value ≤0.001 and with shades

of grey for comparisons with Jaccard Index p-value >0.001. For details regarding the generation of the random sets of interactions, see section 2.6 of Materials and Methods.

Unexpectedly, the concordance was higher for trans (median JI of 0.19) than for cis interactions (median JI<0.03). At low resolution GOTHiC scored the highest concordance, probably due to the fact that it called a large number of short-range interactions in each replicate, thus increasing the chances of overlap. On the contrary, interactions found by HiCCUPS in datasets at 5 kb resolution were among the most conserved across replicates.

Since the number of interactions was extremely variable between replicates, we recalculated the Jaccard Index considering only the top 1000 cis interactions called by each method in every replicate of Rao IMR90. However, except for Fit-Hi-C, this approach produced no overall substantial increase of the concordance (resulting q-values were above 0.05 in a one-tail Wilcoxon with Benjamini-Hochberg correction; Fig. 3.11a).

Instead, the reproducibility improved with the number of reads, especially for HiCCUPS and GOTHiC, when grouping samples based on increasing number of reads (Fig. 3.11b).

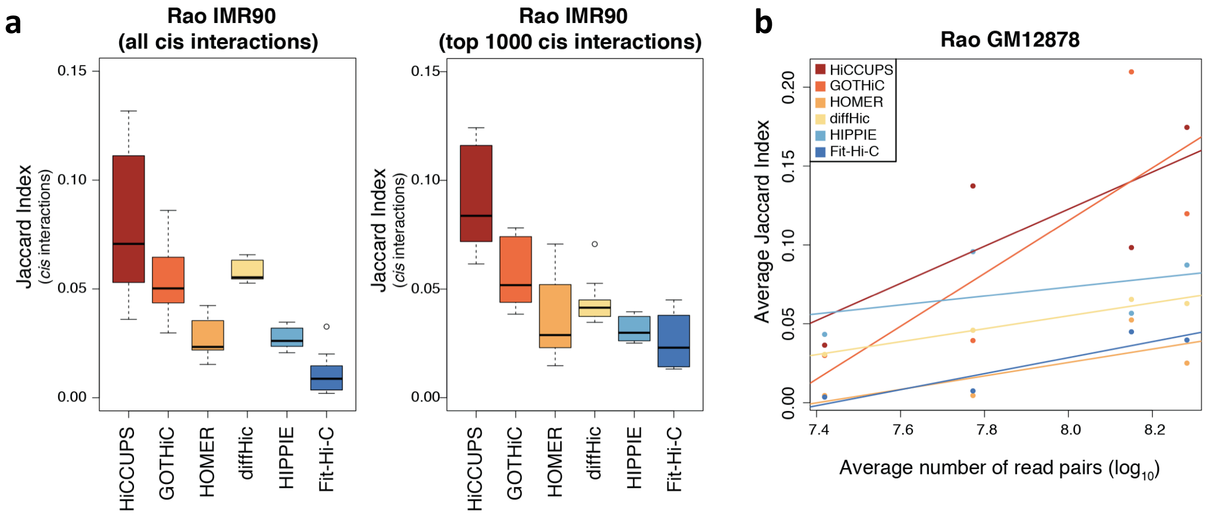

Figure 3.11: a) Box plots of Jaccard Index of all (left) and top 1000 (right) cis interaction calls between replicates A1, A2, A5, B1, and B2 of IMR90 samples in Rao dataset. The top 1000 interactions were defined based on the False Discovery Rate (FDR) for HiCCUPS, GOTHiC, and Fit-Hi-C, on the p-value for HOMER and HIPPIE, and using the enrichment score in diffHic. b) Scatter plot and linear interpolation of average Jaccard Index (y-axis) versus average number of read pairs (x-axis in log scale) in Rao GM12878 replicates stratified by number of reads (see Online Methods). The plot shows that for HiCCUPS and GOTHiC the Jaccard Index has a stronger increase in pairwise comparisons between samples in groups with larger number of reads.

Interactions found by GOTHiC and HiCCUPS were the most conserved also when considering the overlap coefficient, a measure of similarity less sensitive to unbalanced number of interactions among compared replicates (Fig. 3.12).

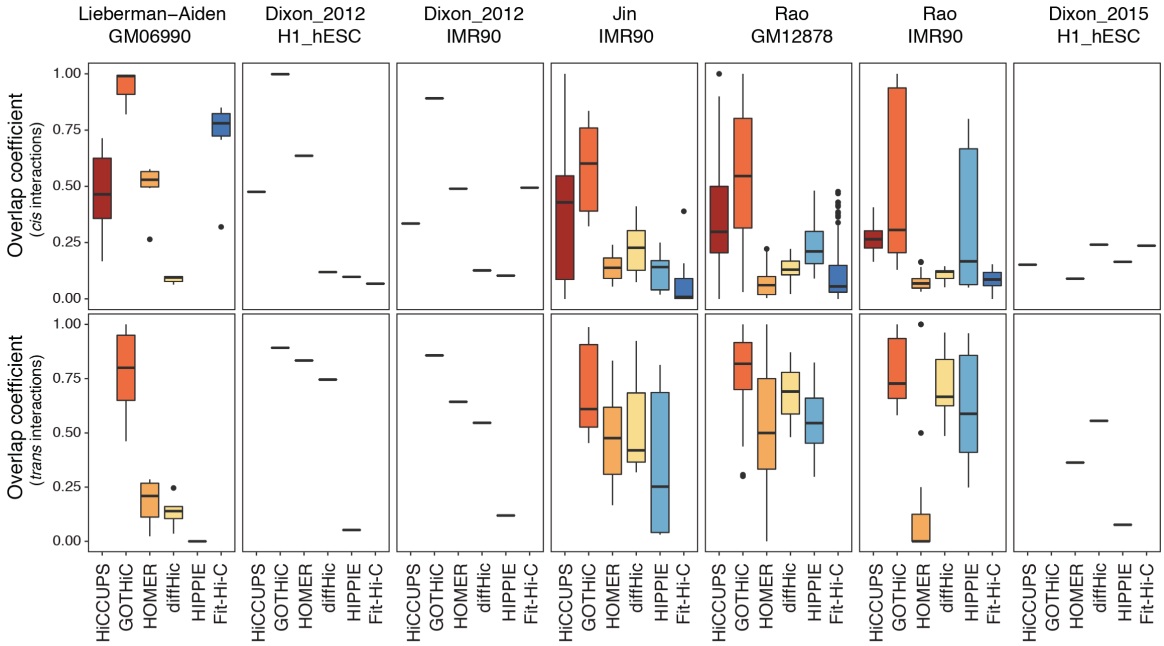

Figure 3.12: Box plots of the overlap coefficient for concordance of cis (upper panels) and trans (lower panels) interaction calls between sample replicates in any dataset (intra-dataset concordance). The overlap coefficient is measured as the size of the common set of interactions in a pairwise comparison, divided by the size of the smallest between the two compared sets. The overlap coefficient was not calculated for GOTHiC in Dixon 2015 (see Fig. 3.3) and for HIPPIE in cis interactions of Lieberman-Aiden (see Fig. 3.7).

The intra-dataset concordance remained unaltered when considering replicates of the same cell line processed using different restriction enzymes (Rao GM12878 with DpnII and MboI and Lieberman-Aiden GM06990 with HindIII and NcoI; Fig. 3.13a-b).

Conversely, the reproducibility between interactions identified in samples of the same cell line in different datasets (generated adopting different experimental protocols and restriction enzymes), was much lower (median JI<4×10-4; Fig. 3.13c).

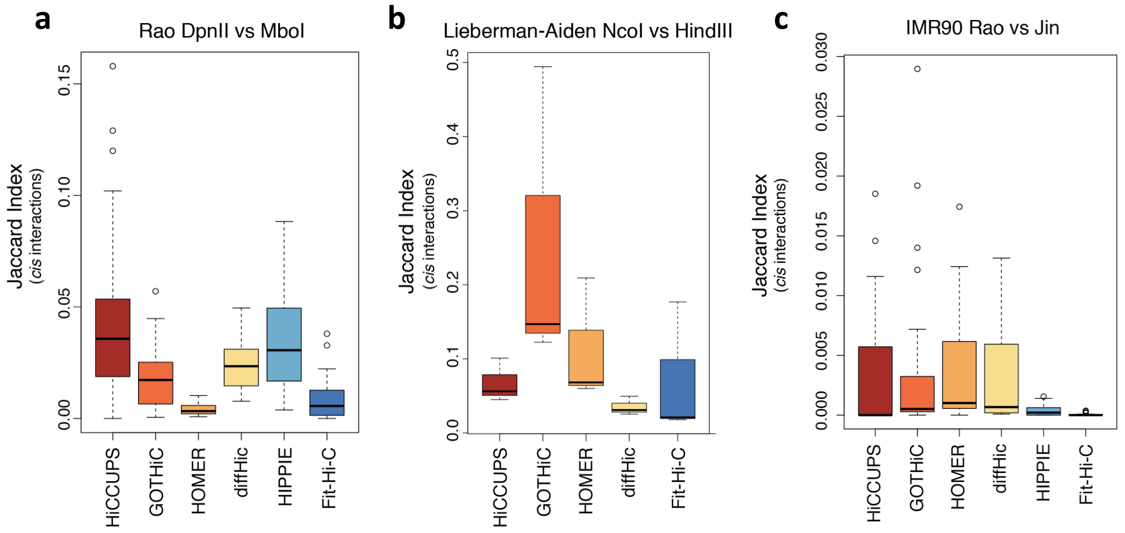

Figure 3.13: a) Box plots of the Jaccard Index of cis interaction calls between all pairs of DpnII - MboI Rao GM12878 processed replicates. b) Box plots of the Jaccard Index of cis interaction calls between all pairs of HindIII - NcoI Lieberman-Aiden GM06990 processed replicates. Jaccard Index was not calculated for HIPPIE (see Fig. 3.7); c) Box plots of the Jaccard Index of cis interaction calls between all pairs of Rao IMR90 - Jin IMR90 replicates.

The next step was the assessment of the ability of each tool to find interactions enriched in chromatin states associated to transcriptional regulation.

Specifically, for each dataset and cell type, interactions common to at least two biological replicates were classified based on cell type-specific chromatin states present in each of the two interacting bins (Roadmap Epigenomics Consortium, Nature 2015; Ho et al., 2014).

Regarding datasets at 40 kb resolution, all approaches identified large proportions of promoter-enhancer cis interactions (46.5% on average), given the greater likelihood for wide genomic bins to comprise a promoter or an enhancer (Fig. 3.14a). On the other hand, at 5 kb resolution, an average of only 16% of all cis interactions were classified as promoter-enhancer, 23% as interactions connecting heterochromatin or quiescent states, and 3% as biologically less expected, i.e., connecting promoter or enhancer to heterochromatin or quiescent states (Fig. 3.14b). At this resolution, HiCCUPS and HOMER found the highest percentage of promoter-enhancer interactions, even though not the highest absolute number, which was scored by diffHic.

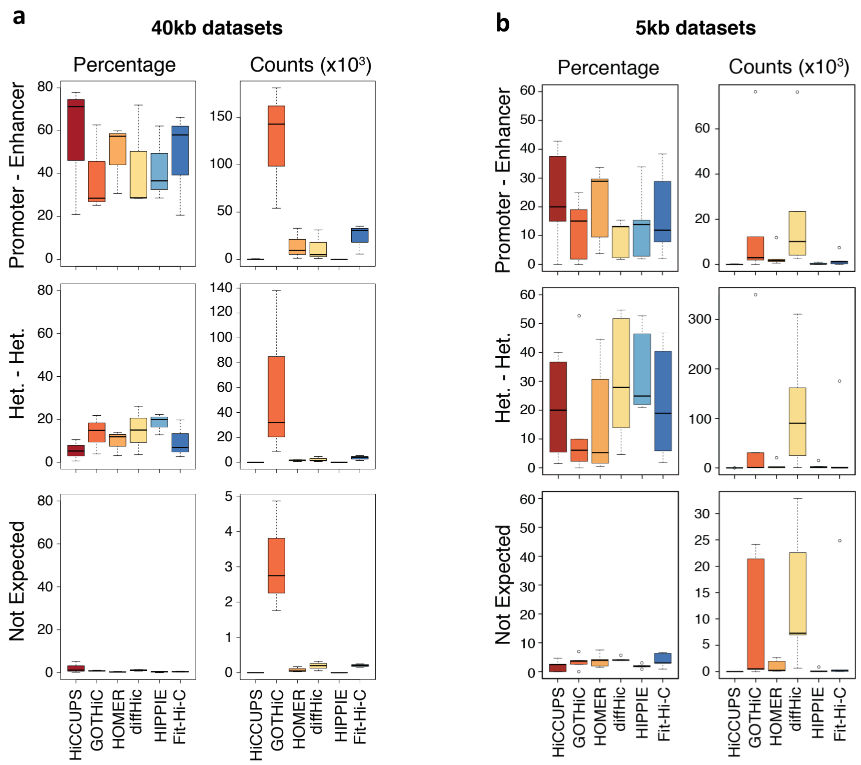

Figure 3.14: a) Proportion (left) and absolute number (right) of cis interactions classified on the base of the chromatin states at their anchoring points as promoter-enhancer (upper), heterochromatin/quiescent to heterochromatin/quiescent (middle), and less expected (lower) in all datasets at 40 kb (data not shown for interactions classified as other combinations of chromatin states). With the exception of Jin H1-hESC (that contains a single replicate), only cis interactions conserved in at least 2 replicates within each dataset were classified using the chromatin states. b) Same as a, at 5 kb resolution.

On the contrary, the percentage of interactions occurring between chromosomes classified as promoter-enhancer was very low for all tools in almost all cases (Fig. 3.15).

diffHic registered the highest quantity and proportion of interactions connecting heterochromatin or quiescent states, even though for all methods the percentage of this type of interaction was particularly high in some datasets. Regardless of the approach and of the resolution, fewer than 8% of all cis interactions were considered as biologically less plausible.

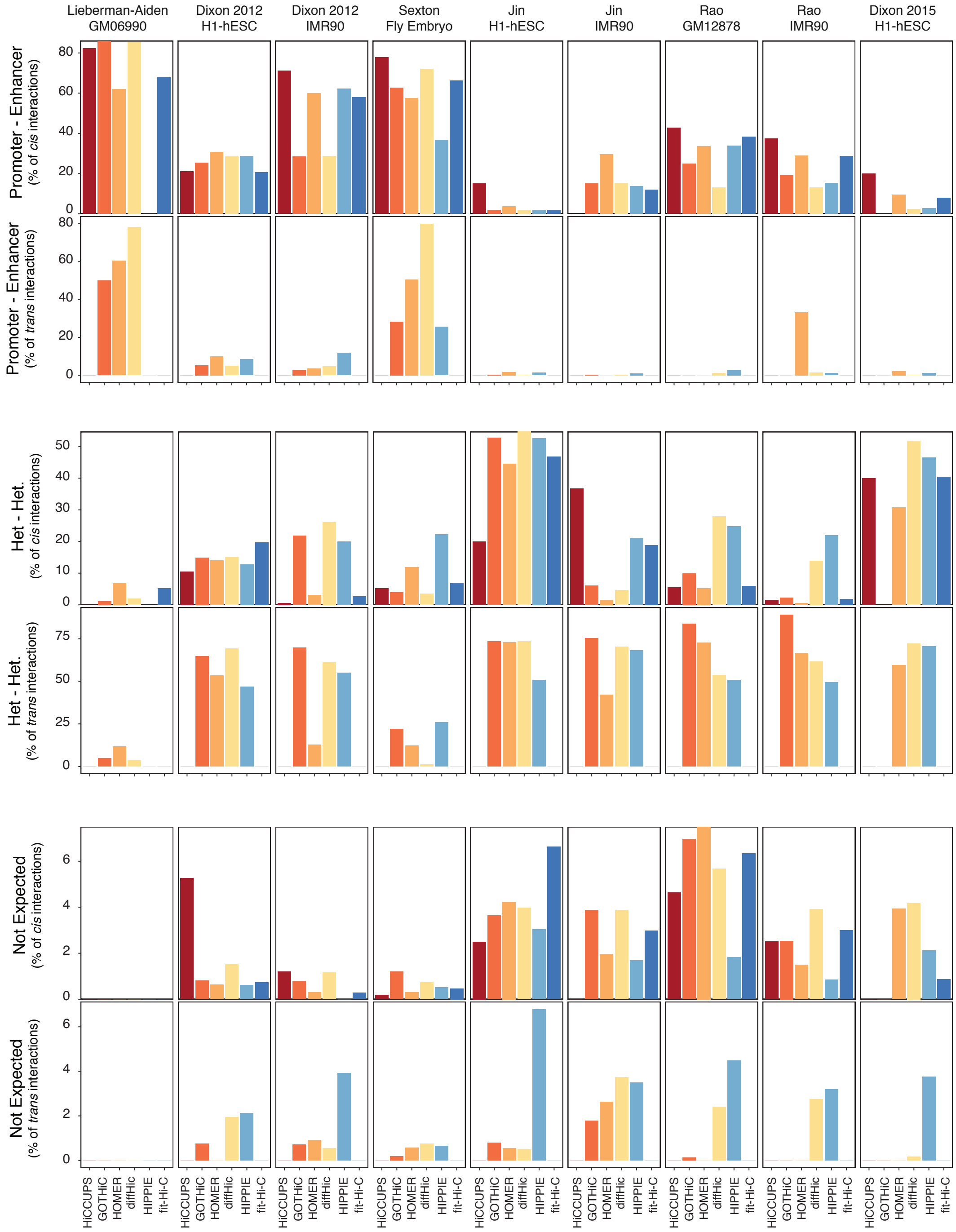

Figure 3.15: Proportion of cis and trans interactions classified on the base of the chromatin states at their anchoring points as promoter-enhancer (upper), heterochromatin/quiescent to heterochromatin/quiescent (middle), and not expected (lower) in each dataset (data not shown for interactions classified as other combinations of chromatin states). With the exception of Sexton and Jin H1-hESC datasets (that contain a single replicate), only interactions conserved in at least 2 replicates within each dataset were classified using the chromatin states. GOTHiC was not applied to Dixon 2015 (see Fig. 3.3 legend). HIPPIE identified no conserved interactions in Lieberman-Aiden.

In general, the enrichment in the number of promoter-enhancer interactions found over random expectation has a tendency to be higher in datasets at 5 kb resolution (p-value≤0.01 in a hypergeometric test in almost all datasets; Fig. 3.16).

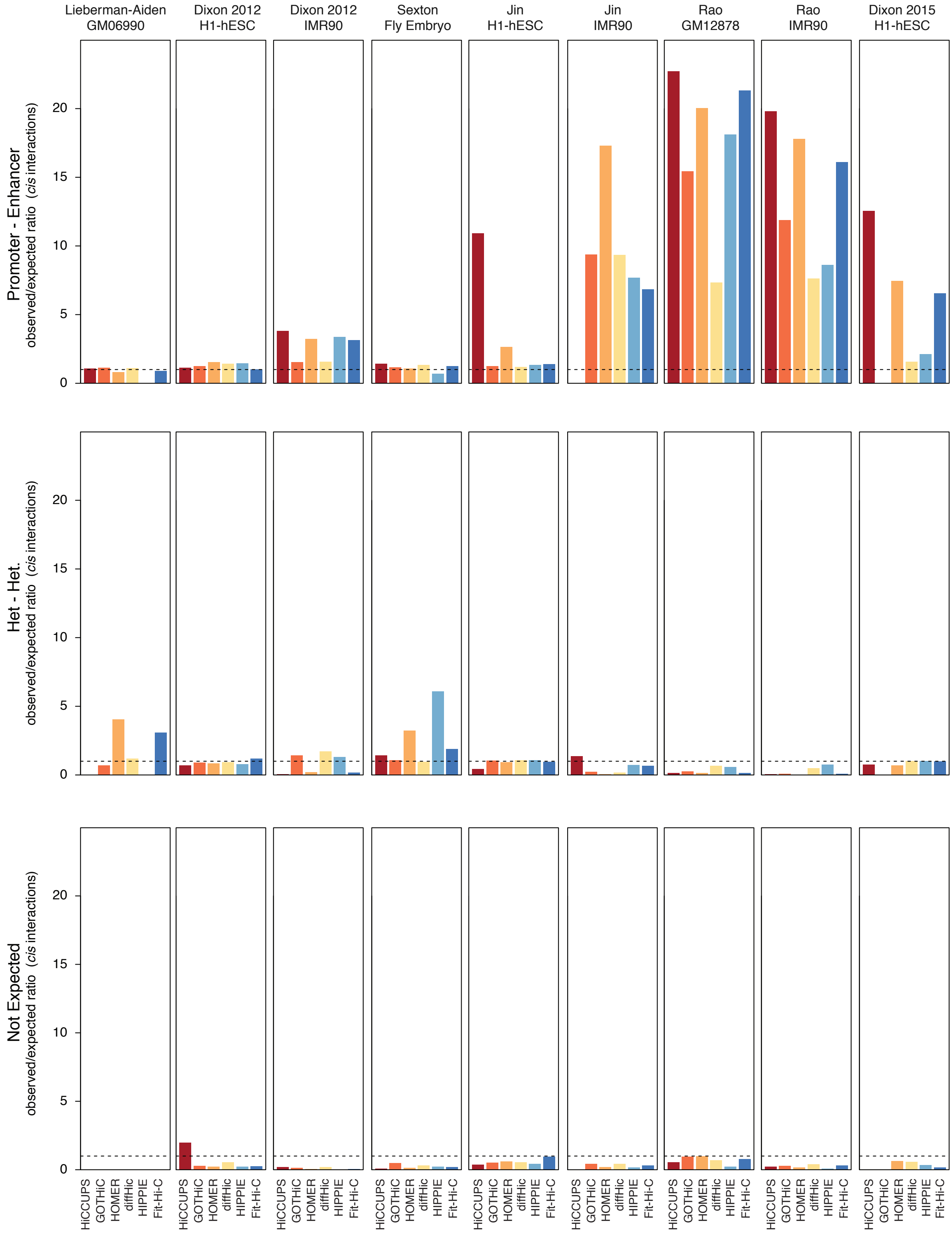

Figure 3.16: Ratio of observed/expected counts of bin pairs classified as promoter-enhancer (upper), heterochromatin/quiescent to heterochromatin/quiescent (middle), and not expected (lower) in all datasets. GOTHiC was not applied to Dixon 2015 (see Fig. 3.3 legend). HIPPIE identified no conserved interactions in Lieberman-Aiden dataset. For further details on how random expectation was computed see section 2.7 of Materials and Methods.

Since the convergent orientation of CTFC motifs has been reported as a distinctive feature of a specific type of chromatin contacts, i.e., looping interactions (Rao et al., 2014), we also quantified the orientations of CTCF-binding motifs among interactions with a single CTCF-binding motif in each of the two interacting bins. All methods identified large proportions (from 45.6% in HIPPIE to 96% in HiCCUPS) of convergent motif pairs among interactions containing CTCF at both sides (Table 3.1).

Table 3.1: CTCF ChIP-seq enrichment and motif orientation at Hi-C interactions anchor regions.

Comparison	**Method**	**Median no. of interactions (%)**
CTCF both sides **vs.** Total interactions	HiCCUPS	25.8
	GOTHiC	1.0
	HOMER	6.1
	diffHic	1.4
	HIPPIE	0.5
	Fit-Hi-C	3.7
At least one convergent pair **vs.** Total interactions	HiCCUPS	24.3
	GOTHiC	0.6
	HOMER	5.1
	diffHic	0.8
	HIPPIE	0.3
	Fit-Hi-C	2.8
At least one convergent pair **vs.** CTCF both sides	HiCCUPS	98.5
	GOTHiC	53.9
	HOMER	79.4
	diffHic	56.3
	HIPPIE	46.0
	Fit-Hi-C	76.5
Only one convergent CTCF pair **vs.** Only one CTCF pair	HiCCUPS	96.0
	GOTHiC	52.1
	HOMER	78.2
	diffHic	52.9
	HIPPIE	45.6
	Fit-Hi-C	66.7

Moreover, we compared the performance of each tool in recalling validated cis interaction evidences (see Table 2.3 of Materials and Methods for the complete list of interactions).

In general, GOTHiC retrieved the highest number of true-positive interactions. Fit-Hi-C and HOMER performances were comparable to GOTHiC, but it should be noted that they were able to achieve the same result calling less interactions (Fig. 3.17a). At 5 kb resolution diffHic was the best performing method, while HOMER found more true-positives than any other approach, at similar numbers of found interactions (Fig. 13.18). All methods recalled low percentages of true negatives in nearly every dataset, although GOTHiC seemed more prone to call false positive interactions at 40 kb resolution (Fig. 3.17b).

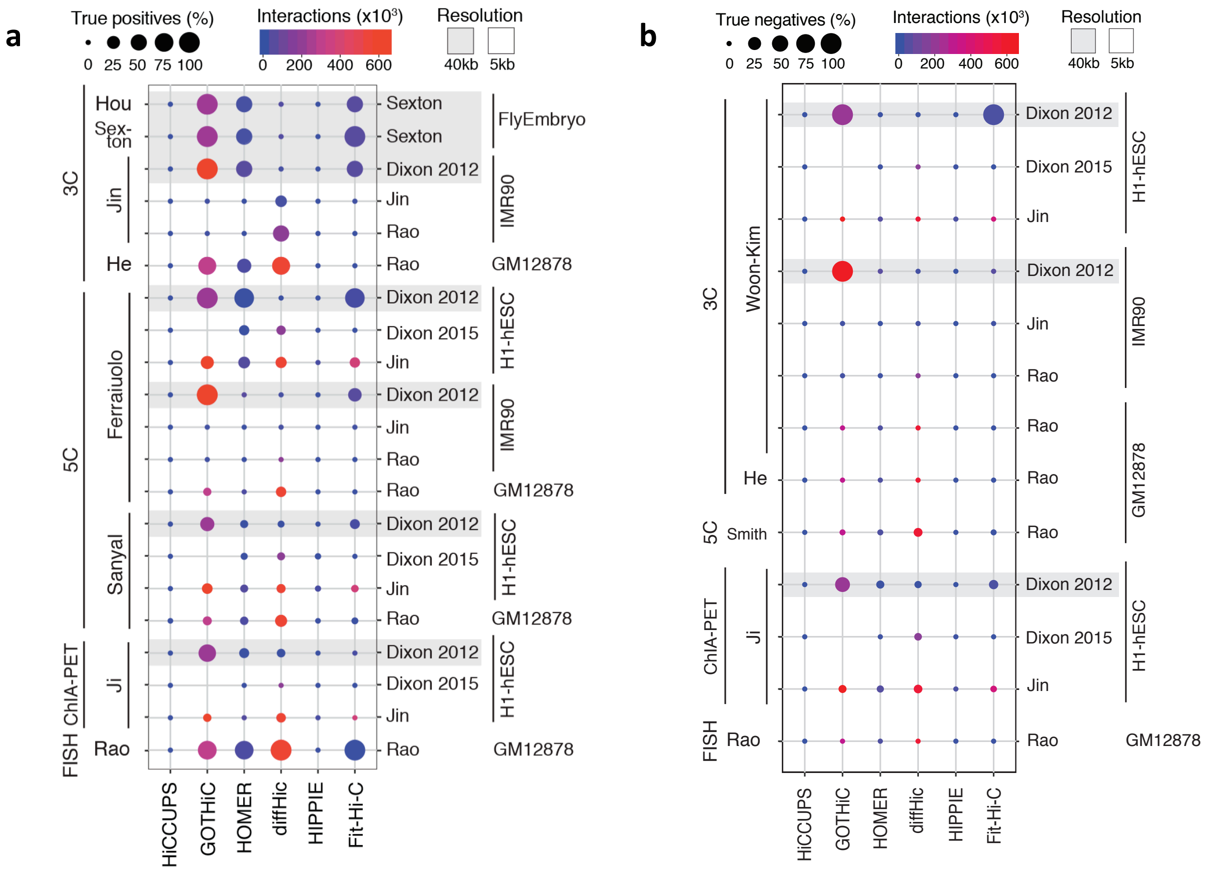

Figure 3.17: a) Performances in the identification of true positive validated evidences of cis interactions. Each row represents the comparison between a list of true positives and the interactions called by each method in each dataset. The dot size is proportional to the percentage of recalled true positives and the dot color accounts for the number of total called interactions. The validation technique and the name of true positive lists are showed on the left side. The dataset used to call interactions are on the right, shaded in grey if at 40 kb resolution. True-positive interactions were searched among cis interactions conserved in at least 2 replicates within each dataset, with the exception of Jin H1-hESC and Sexton (both containing a single replicate). GOTHiC was not applied to Dixon 2015 (see legend of Fig. 3.3); b) same as a, but with the list of true negatives.

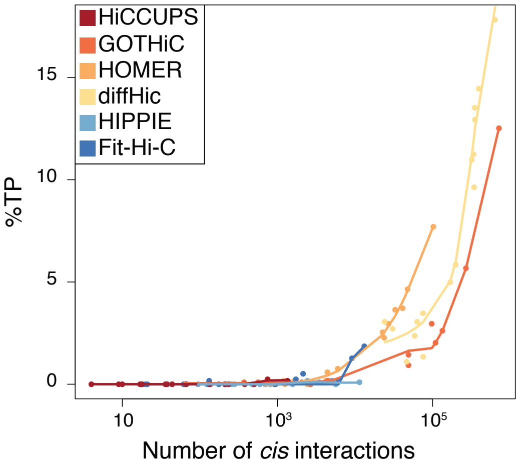

Figure 3.18: Percentage of true-positive interactions (%TP) from 5C data of Sanyal *et al.*Sanyal et al., "The Long-Range Interaction Landscape of Gene Promoters." recalled, in each replicate of Rao GM12878 dataset (at 5 kb resolution), by each method as a function of the total number of called *cis* interactions (x-axis in log scale). We used data from Rao GM12878 since Rao dataset contained the largest number of replicates for GM12878 cell line and GM12878 was characterized by a large number of known true positives.

Finally, to evaluate how downstream results are affected by the alignment and filtering strategy, we compared the performances of the methods starting from input data generated from a common preprocessing procedure. In particular, we used hiclib iterative alignment and filtering to create contact maps then used as input for all tools (Fig. 3.19a-b; for details on the analysis, see section 2.3.1 of Materials and Methods).

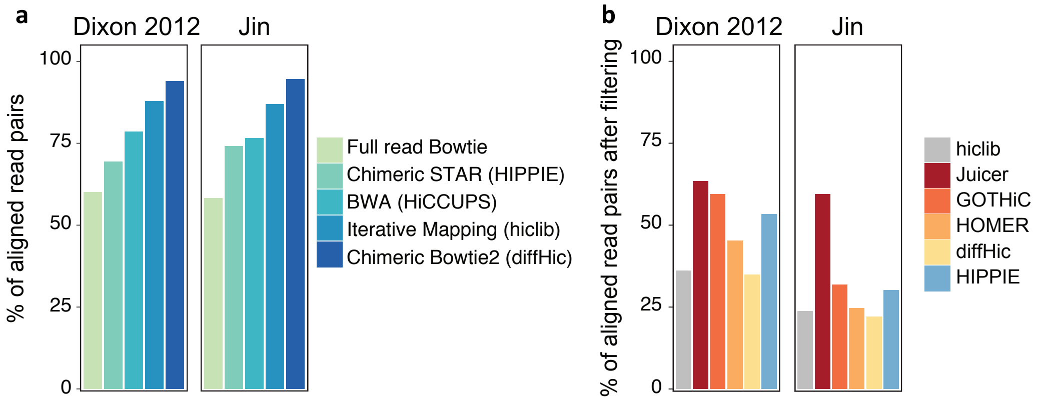

Figure 3.19: We applied hiclib as a common preprocessing procedure to align and filter reads from Dixon2012 IMR90 and Jin IMR90. These data were then used as input to all tools, with the exception of HIPPIE, for which it is not possible to disentangle preprocessing and downstream analysis. Normalization and downstream analysis were performed using each tool proprietary procedures. We used Juicer Tools Pre to convert hiclib output into the .hic input file for HiCCUPS. a) Median percentage of aligned read pairs (alignment rate) for all approaches, including hiclib iterative mapping. b) Median percentage of mapped reads retained after filtering (fraction of usable reads) for all tools, including hiclib.

With the exception of Fit-Hi-C, all tools returned comparable number of interactions and chromatin state classification, irrespective of the preprocessing procedure. Whereas the reproducibility slightly decreased for HiCCUPS when used on data preprocessed with hiclib (Fig. 3.20).

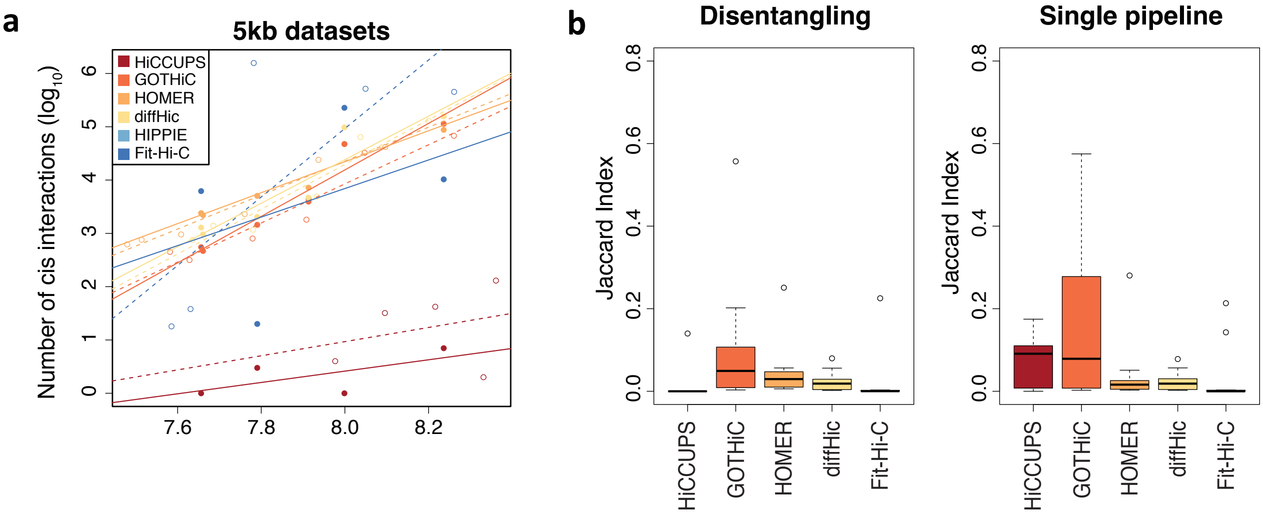

Figure 3.20: a) Scatter plot of total number of cis interactions called by each method versus the number of reads retained by the filtering steps in Jin IMR90 dataset. Different points represent sample replicates analyzed using hiclib common preprocessing (filled dots) or the preprocessing of each tool (open circles). Linear interpolation (of log transformed data) is shown as solid line for hiclib common preprocessing and as dashed line for each tool preprocessing. b) Box plots of the Jaccard Index of cis interaction calls between sample replicates in Dixon2012 IMR90 and Jin IMR90 commonly preprocessed using hiclib (left panel) or using each single tool (right panel).

Interactions found starting from data preprocessed by hiclib (disentangling) or, independently by each tool (single pipeline), were reasonably conserved for most methods with a median overlap coefficient of 53% (Fig. 3.21), although using a different preprocessing might impact the results of some downstream approaches, like HiCCUPS and Fit-Hi-C.

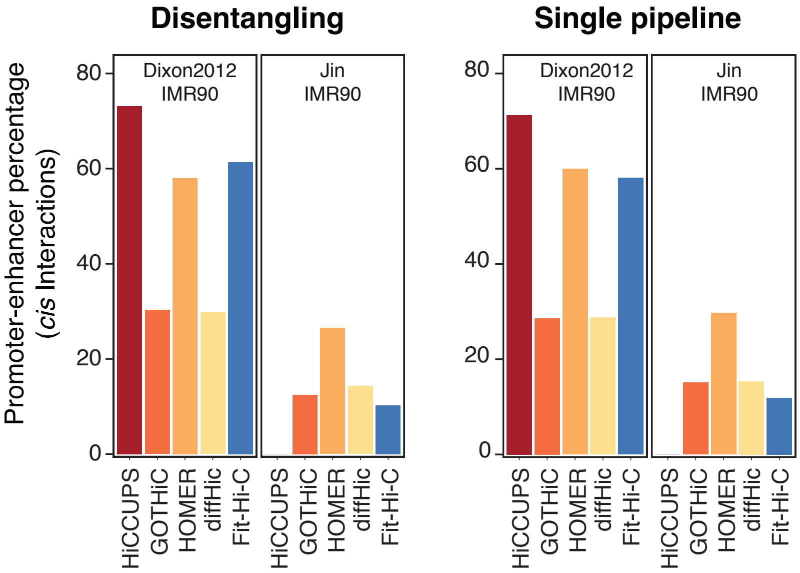

Figure 3.21: Proportion of cis interactions classified as promoter-enhancer in Dixon2012 IMR90 and Jin IMR90, commonly preprocessed using hiclib (left panel) or using each single tool (right panel).

To assess sensitivity and precision of the methods, we modified the model to generate simulated interaction matrices originally proposed by (Lun et al., 2015) and analyzed the simulated data with HiCCUPS, HOMER, diffHic, and Fit-Hi-C, the only tools among the ones in analysis that can take as input an interaction matrix. For a set of 40 simulated samples and 8 levels of base interaction strength, all tools called a much larger number of interactions than the 1000 true interactions (Fig. 3.22a).

As for experimental data, Fit-Hi-C called interactions at larger mean distance (Fig. 3.22b).

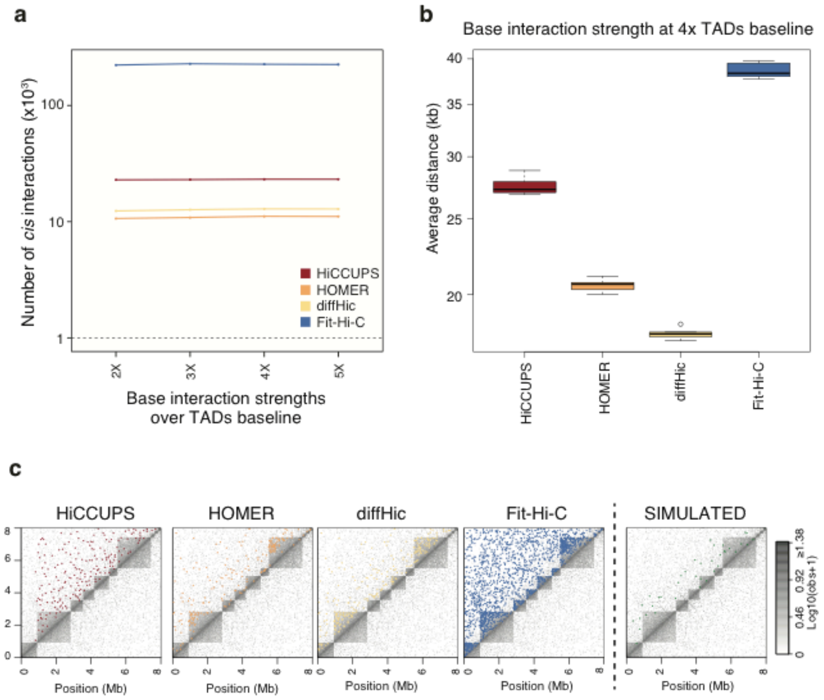

Figure 3.22: a) Average number of cis interactions called by each method as a function of the base interaction strength without the additional fixed constant ($K_{interactions}$, see Supplementary note 3). The number of true interactions (1000) is shown as a dashed line. Data are shown as mean±standard error of the mean (SEM). Similar results were obtained using the additional fixed constant (data not shown). b) Boxplot of average distances between anchoring points in cis interactions (log scale) in 5 replicates generated at a base interaction strength equal to 4 times the baseline of simulated TADs. c) Heatmap of the contact matrix generated with base interaction strength equal to 2 times the baseline of simulated TADs (simulated chr:0-8,000,000). True simulated interaction peaks are in green, identified peaks are marked in different colors for the various methods.

Fit-Hi-C reached the highest sensitivity, though all tools presented a particularly high FDR (i.e., low precision), as expected given the difference between the number of true and of total called interactions (Fig. 3.23a-b).

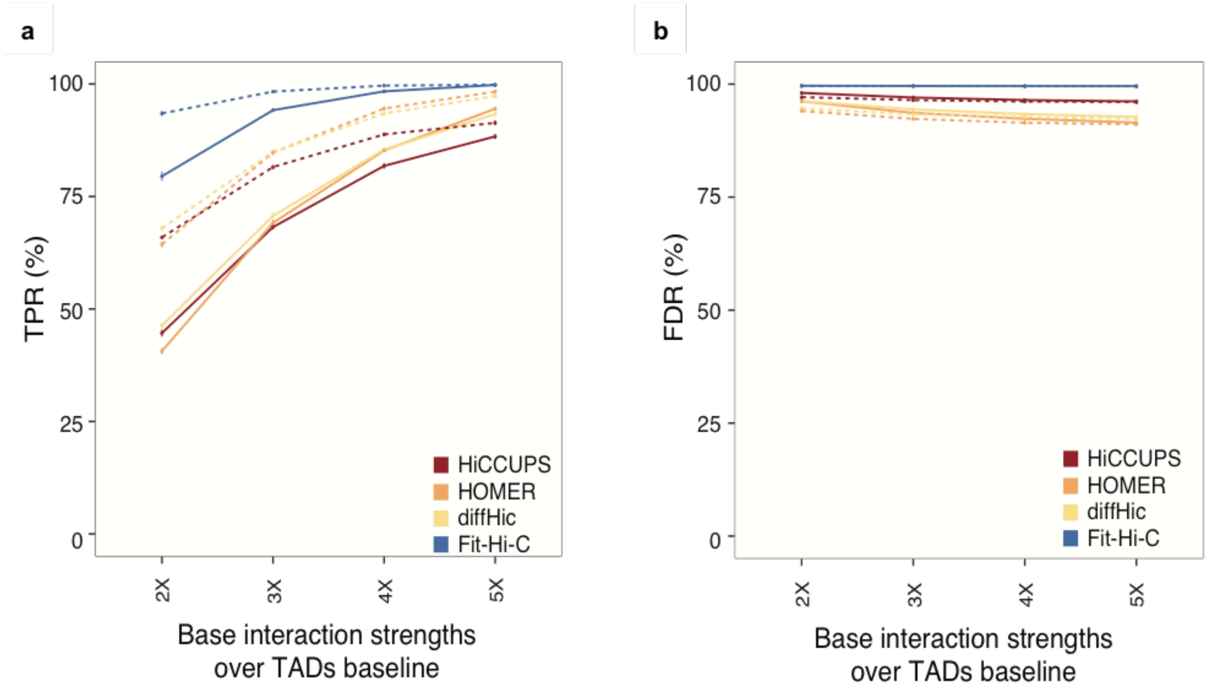

Figure 3.23: a) True positive rate (sensitivity) as a function of the base interaction strength with (dashed line) and without (solid line) the $K_{interactions}$ constant. Data are shown as mean±standard error of the mean; b) False Discovery Rate (1-precision) as a function of the base interaction strength without the $K_{interactions}$ constant. Data are shown as mean±standard error of the mean. Similar results were obtained using the additional fixed constant (data not shown). For further details, see

3.1.3 TAD callers

As for chromatin interactions, to compare TAD callers on experimental data we considered several metrics, including: the total number of called TADs; the TAD size; the concordance of TAD boundaries within and between datasets when analyzing biological replicates; and the enrichment at TAD boundaries of known boundary elements (i.e., CTCF and BEAF32).

Differently from interaction callers, the number of TADs found by each tool did not vary according to the number of post-filtering reads, with the sole exception of Arrowhead, which shows a distinct linear relationship (Fig. 3.24a). On the contrary, it seems that at low numbers of post-filtering reads most of the tools (except Arrowhead) tend to call a higher number of TADs, probably due to the sparsity of these matrices. The number of identified TADs was distinctive of each method and was, generally, inversely proportional to the TAD size (Fig. 3.24b).

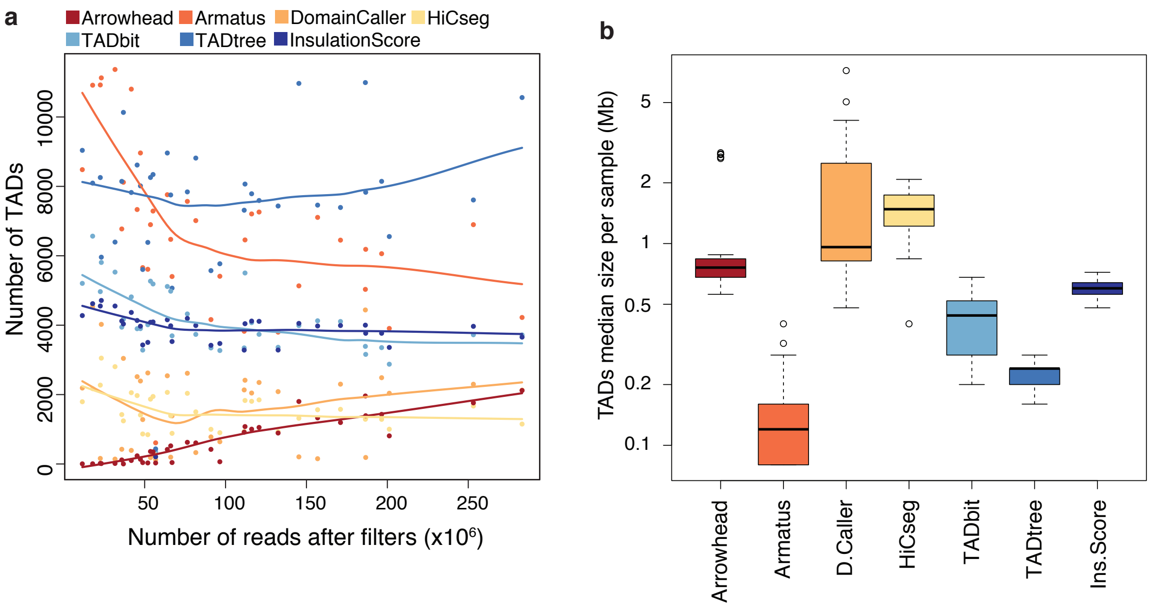

Figure 3.24: a) Scatter plot of total number of TADs called by each method as a function of the number of reads retained by the filtering step in all datasets except Lieberman-Aiden (n=36; Table 2.2). Different points represent sample replicates. Loess interpolation for each method is shown as solid line. Due to elevated computational time and memory issues, datasets originally binned at 40kb or less were analyzed at 40kb. b) Boxplot of median TAD size in all replicates of all datasets (analyzed at 40kb) except Lieberman-Aiden dataset (n=36).

On average, at 40 kb resolution, TADtree identified the highest (7638) and Arrowhead the lowest (636) number of TADs, while InsulationScore called the largest number of domains at 1 Mb resolution (Fig. 3.25).

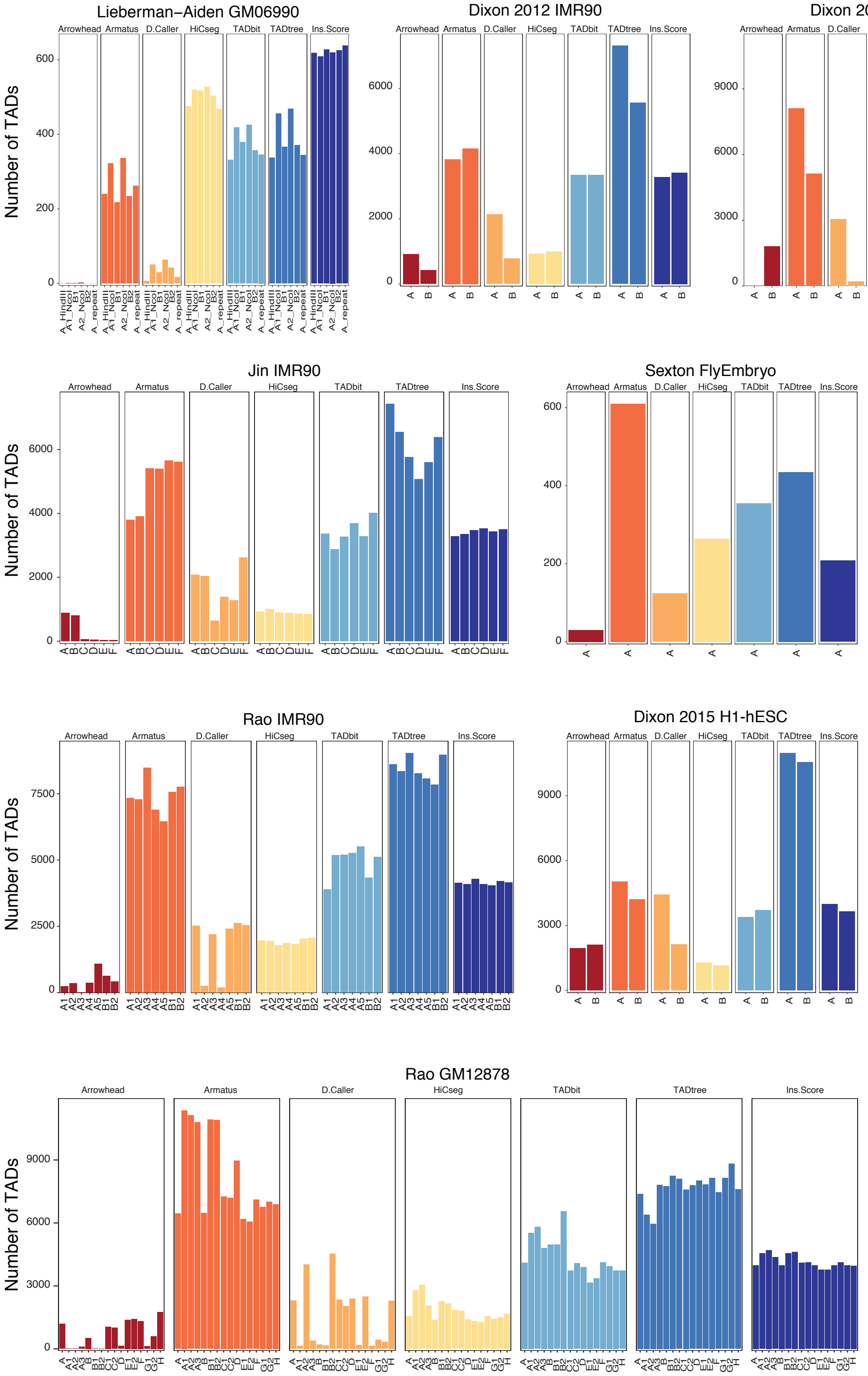

Figure 3.25: Number of TADs identified by the various tools in each replicate of each dataset.

The distinctive features of the TADs found by each approach are illustrated in the heatmap representation of the contact matrices (Fig. 3.26).

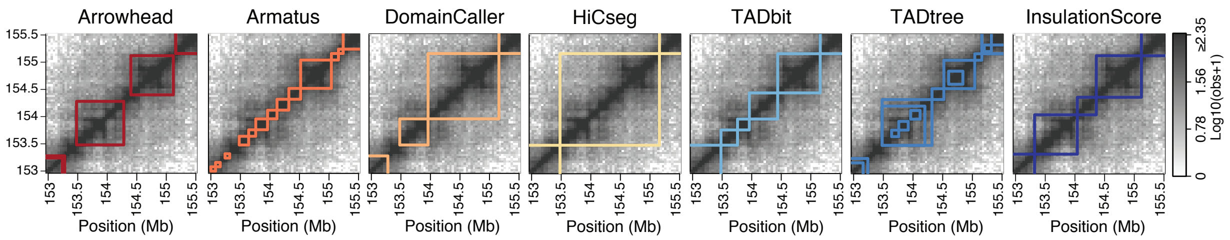

Figure 3.26: Heatmap of the contact matrix of Rao GM12878 replicate H (chr1:153,000,000-155,500,000) at 40 kb resolution. Identified TADs are depicted in different colors for the various methods.

It is immediately evident how the various methods differ in terms of chromosome partitioning into TADs: some find a continuous set of domains, as HiCseg, TADbit and InsulationScore, while the others allow the presence of "gaps" (i.e., unorganized regions of chromatin; Dixon et al., 2012) between TADs. Moreover, TADtree and Arrowhead, which implement multi-scale approaches, return nested TADs.

To evaluate the conservation of TADs found by the different tools, we estimated the Jaccard Index as a degree of the similarity between TAD borders across biological replicates. In general, HiCseg presented the highest intra-dataset reproducibility (i.e., among replicates of the same dataset). Altogether, TAD boundaries displayed a higher conservation (median JI of 0.25) compared to what observed for chromatin interactions (Fig. 3.27).

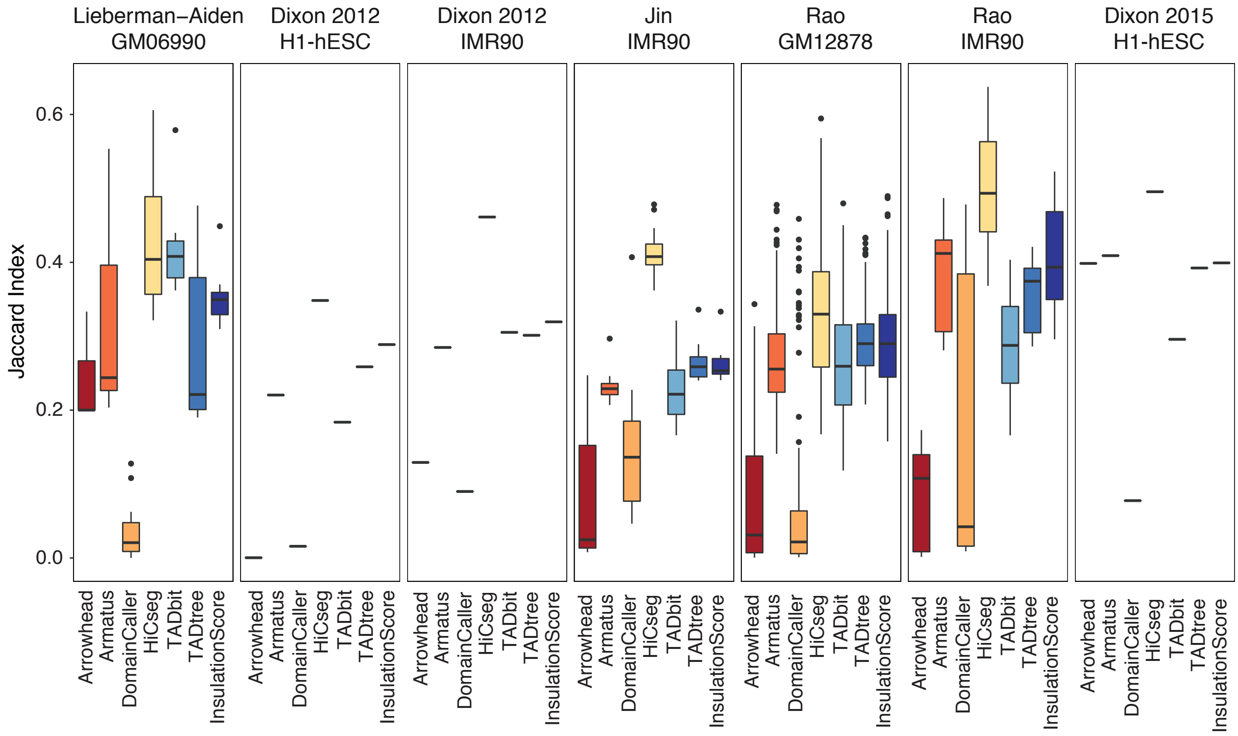

Figure 3.27: Box plots of the Jaccard Index for concordance of TAD boundaries between pairs of sample replicates in each dataset (intra-dataset).

We also noted that reproducibility increased with the number of reads for all methods if samples were divided in groups according to their post-filtering read count (Fig. 3.28a). HiCseg resulted to be the method with the highest reproducibility also when employing the overlap coefficient, i.e. a measure robust to differences in the number of TADs called across replicates (Fig. 3.28b).

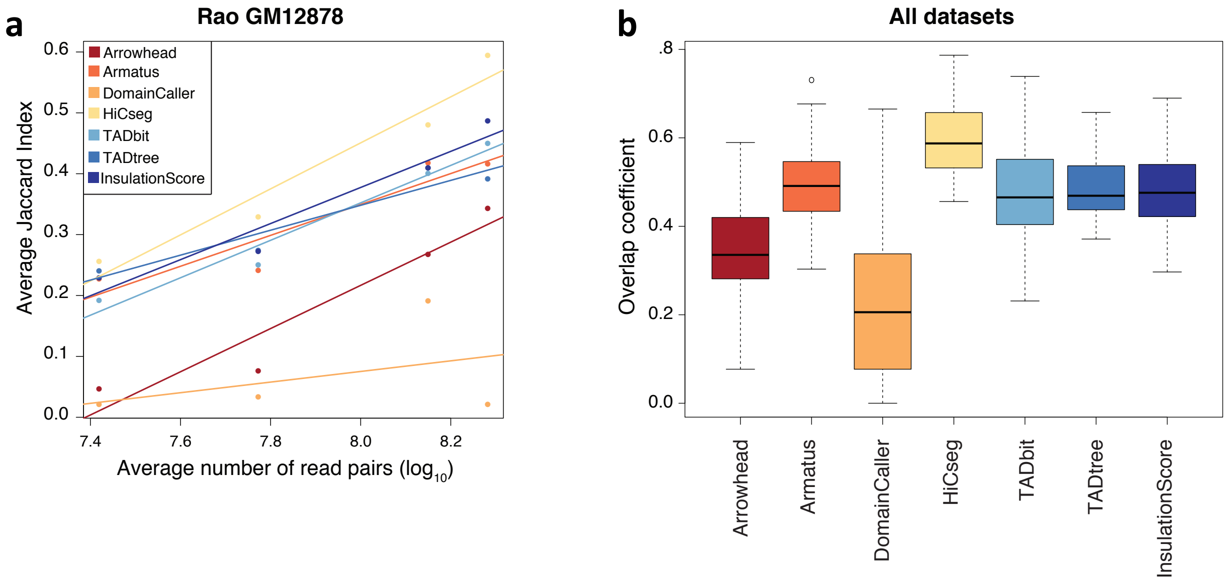

Figure 3.28: a) Scatter plot and linear interpolation of average Jaccard Index (y-axis) versus average number of read pairs (x-axis in log scale) in Rao GM12878 replicates stratified by number of reads. Rao GM12878 dataset was chosen because it was the only dataset with a sufficient number of replicates to perform this analysis. Specifically, replicates B2, B1, A2, A1, and G1 constituted the group of samples with less than 40 million reads; A3, D, B, and G2 the group with more than 40 and less than 100 million reads; C2, C1, F, and A the group of samples with a number of filtered reads comprised between 100 and 180 million reads; and E1 and E2 the group of samples with more than 180 million reads. Replicate H was not included in any of the above groups; b) Box plots of the overlap coefficient for concordance of TAD calls between sample replicates in any dataset (intra-dataset concordance). The overlap coefficient is measured as the size of the common set of TADs in a pairwise comparison, divided by the size of the smallest between the two compared sets.

The intra-dataset reproducibility did not show substantial changes for most methods when comparing results obtained in replicates of the same cell line processed with different restriction enzymes (Rao GM12878 with DpnII and MboI and Lieberman-Aiden GM06990 HindIII and NcoI; Fig. 3.29a-b). Instead, the inter-dataset concordance (i.e., between TAD boundaries called in samples of the same cell line in different datasets, generated following different experimental protocols and restriction enzymes) was lower (median JI of 0.16) than the intra-dataset concordance, with TADtree exhibiting the highest and Arrowhead the lowest inter-dataset reproducibility (Rao and Jin IMR90; Fig. 3.29c).

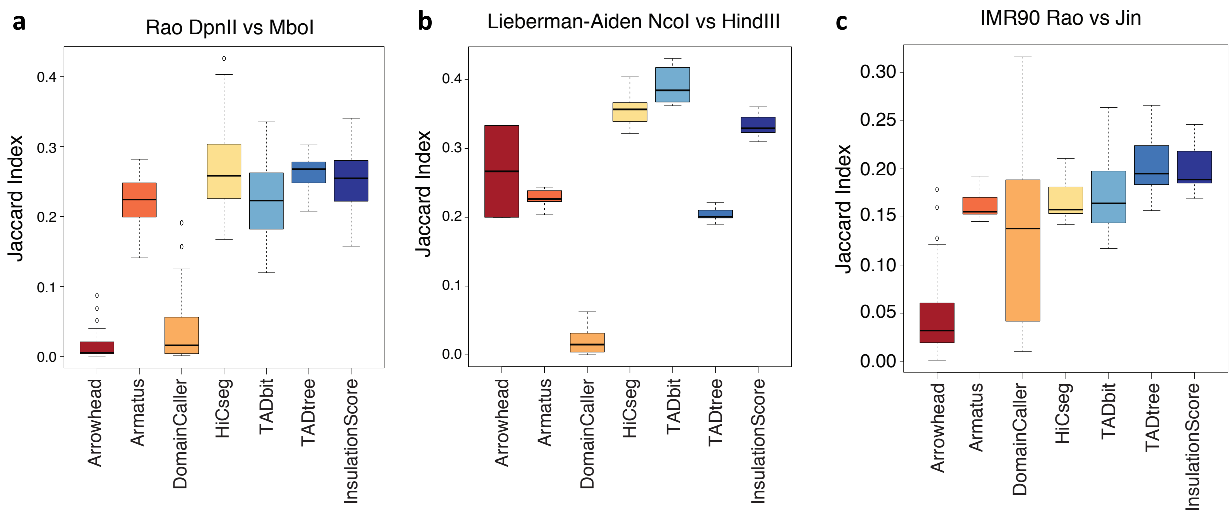

Figure 3.29: a) Box plots of the Jaccard Index of TAD boundaries between all pairs of DpnII - MboI Rao GM12878 processed replicates. b) Box plots of the Jaccard Index of TAD boundaries between all pairs of HindIII - NcoI Lieberman-Aiden GM06990 processed replicates. c) Box plots of the Jaccard Index of TAD boundaries between all pairs of Rao IMR90 - Jin IMR90 replicates.

All the approaches identified TADs with comparable levels of enrichment in insulator proteins (e.g. CTCF or BEAF32) at their borders. Approximately, more than 50% of TAD boundaries comprised CTCF ChIP-seq peaks (Fig. 3.30).

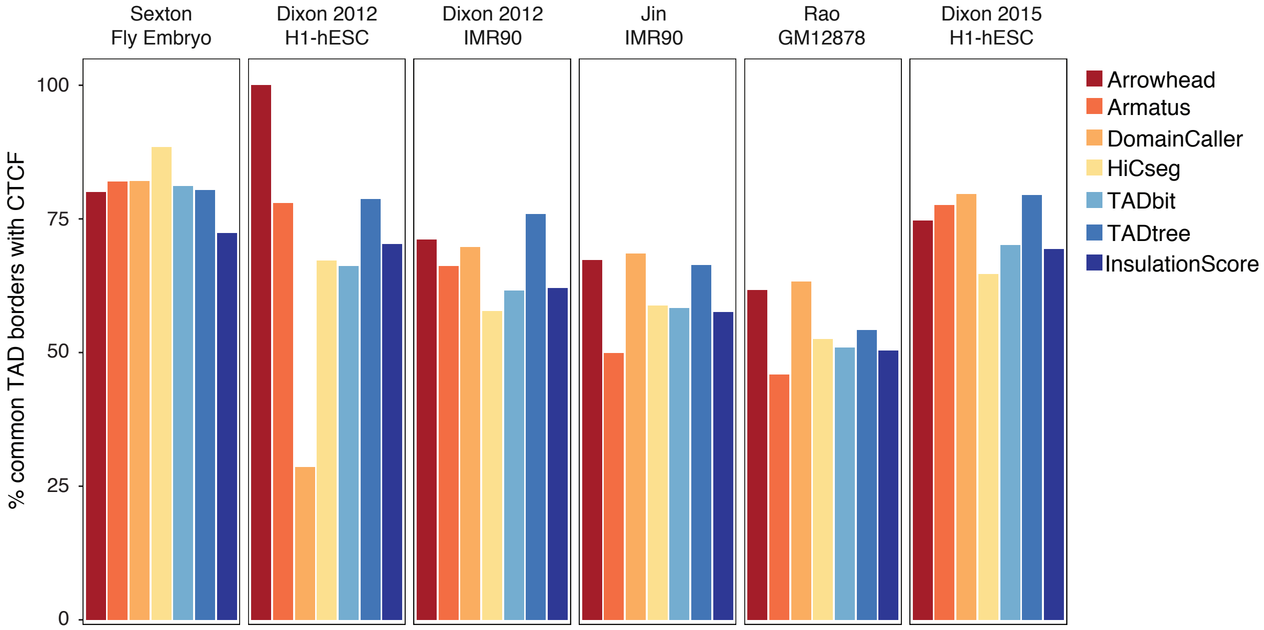

Figure 3.30: Percentage of TAD boundaries overlapping CTCF binding regions in a window of 40 kb in all datasets. With the exception of Sexton dataset, that contains a single replicate, only TAD boundaries conserved in at least 2 replicates within each dataset were used to calculate the overlap with CTCF ChIP-seq peaks.

The only exception was represented by DomainCaller, which performed poorly in the H1-hESC cell line from Dixon 2012 (<30% of boundaries overlapping CTCF) due to large differences of calls between the two H1-hESC replicates. All tools found domains with a substantial presence of CTCF peaks at their boundaries, with TADs called by Armatus and TADtree exhibiting a sharper CTCF enrichment at TAD borders compared to those found by the other methods (Fig. 3.31).

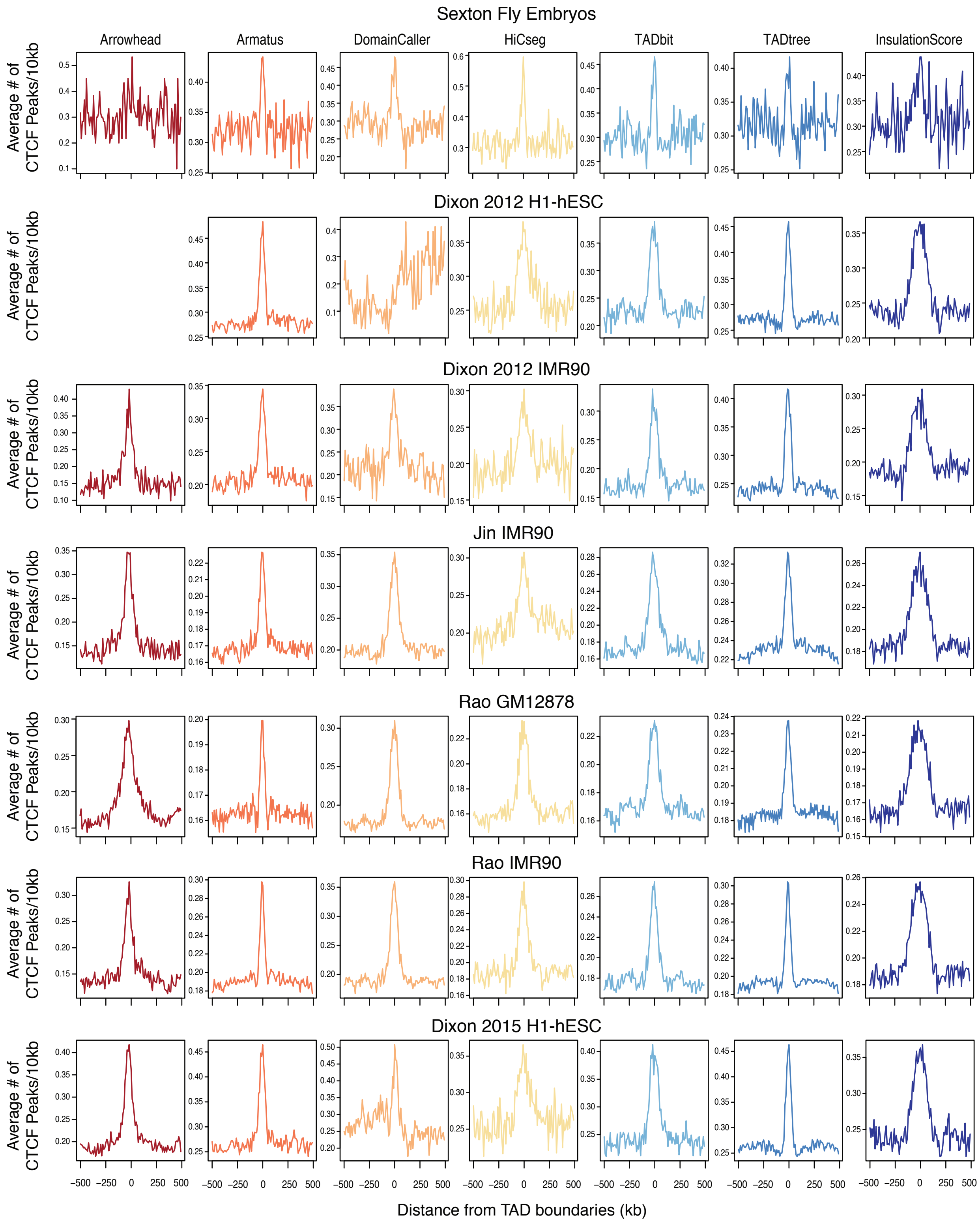

Figure 3.31: Enrichment of CTCF binding (ChIP-seq peaks) in a window of 1 Mb around the TAD boundaries (all datasets). With the exception of Sexton dataset, that contains a single replicate, only TAD boundaries conserved in at least 2 replicates within each dataset were used to calculate the CTCF binding enrichment. The enrichment for Arrowhead in Dixon 2012 H1-hESC was not calculated since Arrowhead found only one conserved TAD boundary in this dataset.

The less spiked enrichment of CTCF peaks at TAD boundaries identified by InsulationScore may be partly explained by the observation made by the authors in (Crane et al., 2015) that the boundary position determined by InsulationScore should be defined as a interval around the insulation minimum rather than as a single bin position. For fly embryos data (i.e., Sexton dataset) we also checked the enrichment of BEAF32 binding, an architectural protein reportedly more common at TAD borders compared to CTCF in Drosophila (Sexton et al., 2012). Most tools returned TADs with a strong enrichment of BEAF32 at their boundaries in Sexton dataset (Fig. 3.32).

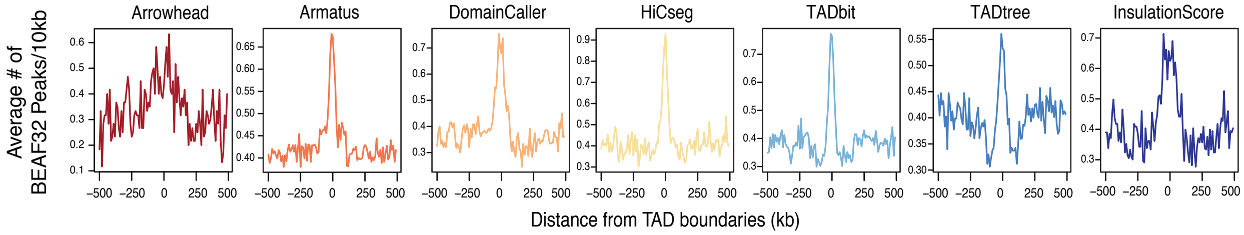

Figure 3.32: Enrichment of BEAF32 binding (ChIP-seq peaks) in a window of 1 Mb around the TAD boundaries (Sexton dataset).

As previously done for the interactions callers, to assess sensitivity and precision of the methods we generated simulated interaction matrices with a modified version of the model originally proposed by (Lun et al., 2015) and analyzed the synthetic data with the various TAD callers. When using simulated Hi-C data of 25 samples at increasing noise level, only DomainCaller, TADbit and InsulationScore identified almost all 171 simulated not overlapping TADs, regardless of the noise (Fig. 3.33a). Comparably to what observed with experimental Hi-C data, HiCseg identified few TADs of large size, while TADtree found many small domains (Fig. 3.33a-c). Armatus, HiCseg and TADtree were deeply affected by the noise levels in the synthetic data: in particular, Armatus tended to call more TADs at higher levels of noise, whereas HiCseg and TADtree showed an opposite trend. This was associated with a steep increase in the False Discovery Rate for Armatus and a drop of the same metric for TADtree, at increasing noise levels (Fig. 3.34b).

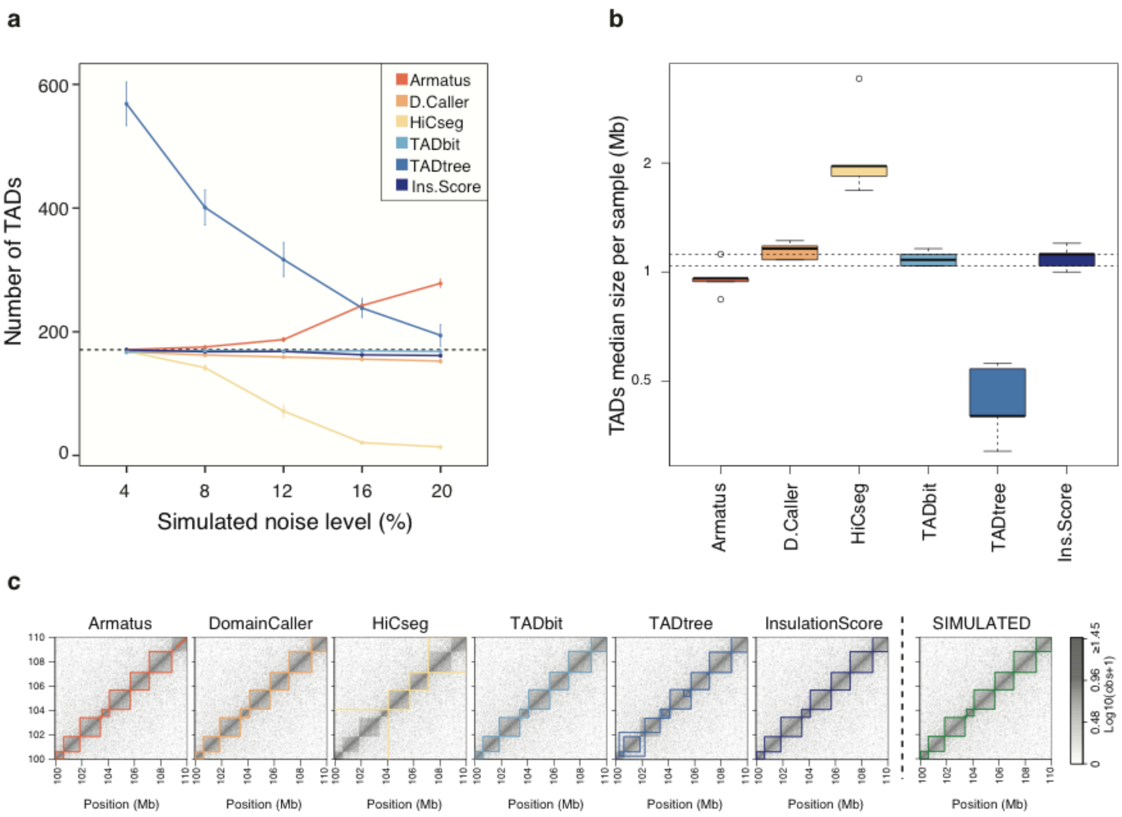

Figure 3.33: a) Average number of TADs called by each method as a function of the simulated noise level. The number of true TADs (171) is shown as a dashed line. Data are shown as mean±standard error of the mean (SEM). Arrowhead identified only 1 TAD in 1 simulated matrix and thus results for this tool are not reported here. b) Boxplot of median TAD sizes called by each method in 5 replicates generated at a noise level equal to the 12% of the total number of data points of the simulated matrices. The 1st and 3rd quartile of the distribution of median true TAD sizes are shown as dashed lines. c) Heatmap of the contact matrix generated at a noise level equal to the 12% of the total number of data points of the simulated matrices (simulated chr:100,000,000-110,000,000). True simulated TADs are in green, identified TADs are marked in different colors for the various methods.

TADbit and Armatus had the highest sensitivity in recovering TAD boundaries, although TADbit displayed a higher precision (low FDR) at all noise levels (Fig. 3.34a-b).

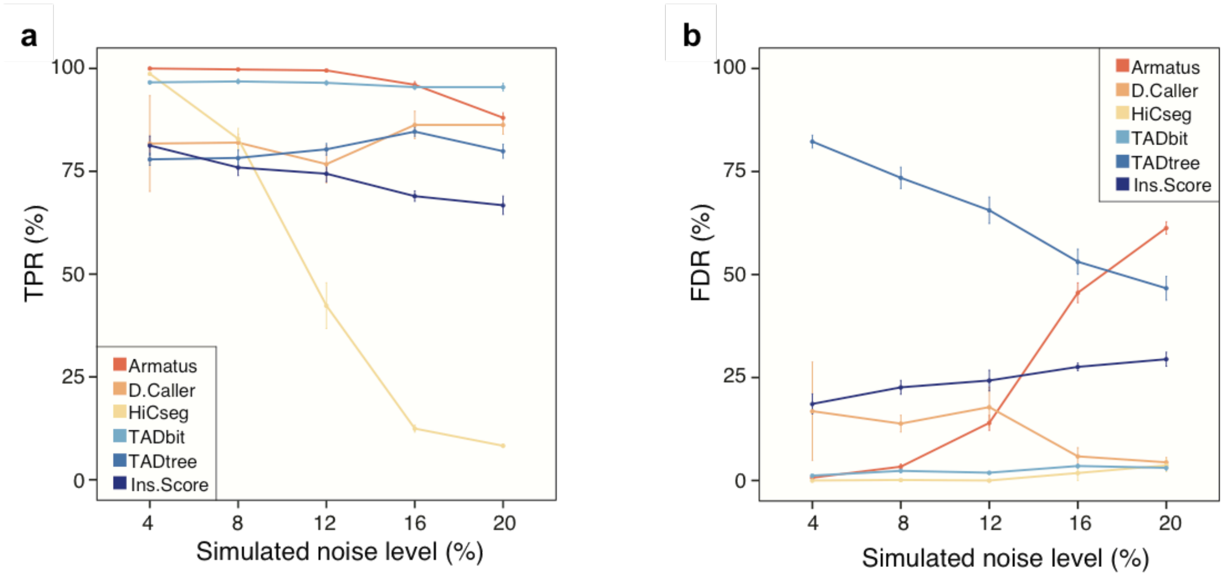

Figure 3.34: a) True positive rate in the identification of TAD boundaries as a function of the noise level (sensitivity). Data are shown as mean±standard error of the mean. b) False Discovery Rate (1-precision) in the identification of TAD boundaries as a function of the noise level. Data are shown as mean±standard error of the mean.

Similar conclusions can be drawn when challenging the methods with synthetic data comprising a hierarchy of nested domains, with the difference that TADtree shows an increase in both sensitivity and precision, probably imputable to the fact that TADtree is specifically designed to recognize nested TADs (Fig. 3.35).

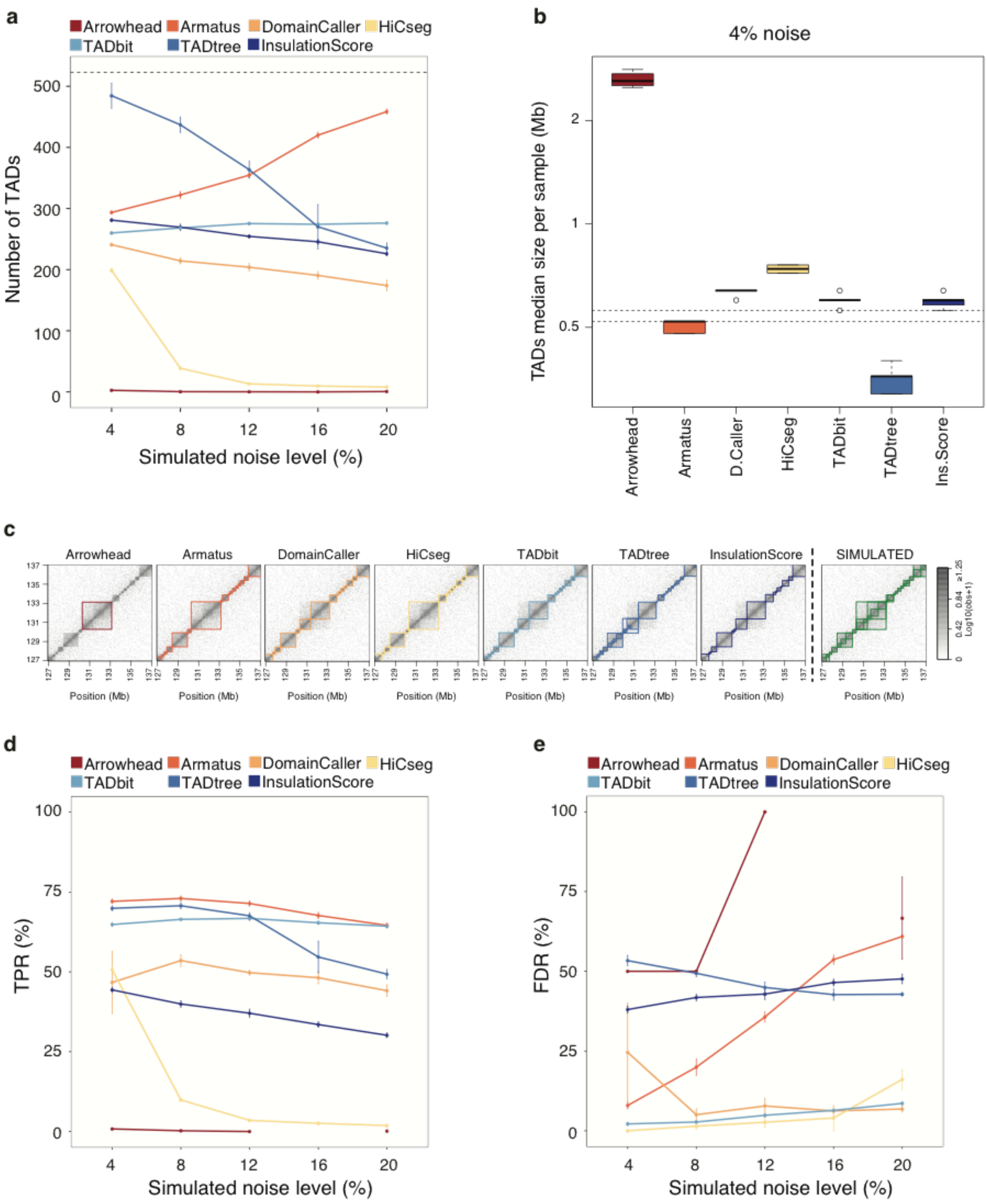

Figure 3.35: a) Average number of TADs called by each method as a function of the simulated noise level. The average number of true TADs (523) is shown as a dashed line. Data are shown as mean±standard error of the mean (SEM). Arrowhead did not identify any TAD at 16% noise level. b) Boxplot of median TAD sizes called by each method in 5 sample replicates generated at a noise level equal to the 4% of the total number of data points of the simulated matrices. The 1st and 3rd quartile of the distribution of median true TAD sizes are shown as dashed lines. c) Heatmap of the contact matrix generated at a noise level equal to the 4% of the total number of data points of the simulated matrices (simulated chr:127,000,000-137,000,000). True simulated TADs are in green, called TADs are marked in different colors for each method. d) True positive rate in the identification of TAD boundaries as a function of the noise level (sensitivity). Data are shown as mean±standard error of the mean. e) False Discovery Rate in the identification of TAD boundaries as a function of the noise level (1-precision). Data are shown as mean±standard error of the mean.

3.1.4 Computational running times

The various methods presented large differences in terms of computational resources needed (running time and memory usage). Among interaction callers, Fit-Hi-C took the longest running time at all resolutions, while the downstream analysis performed with TADtree was the slowest among TAD callers (Fig. 3.36). GOTHiC and Fit-Hi-C are the most demanding methods in terms of memory usage requiring more than 512 GB of RAM for the analysis of samples at 5 kb resolution. In particular, the analysis of the H1-hESC replicates from Dixon 2015 (accounting for 504 and 922 millions of aligned read pairs, respectively) could not be completed using GOTHiC on a machine equipped with 1 TB of RAM. Additional details on the usability of the tools are reported in Section 2.9 of Materials and Methods.

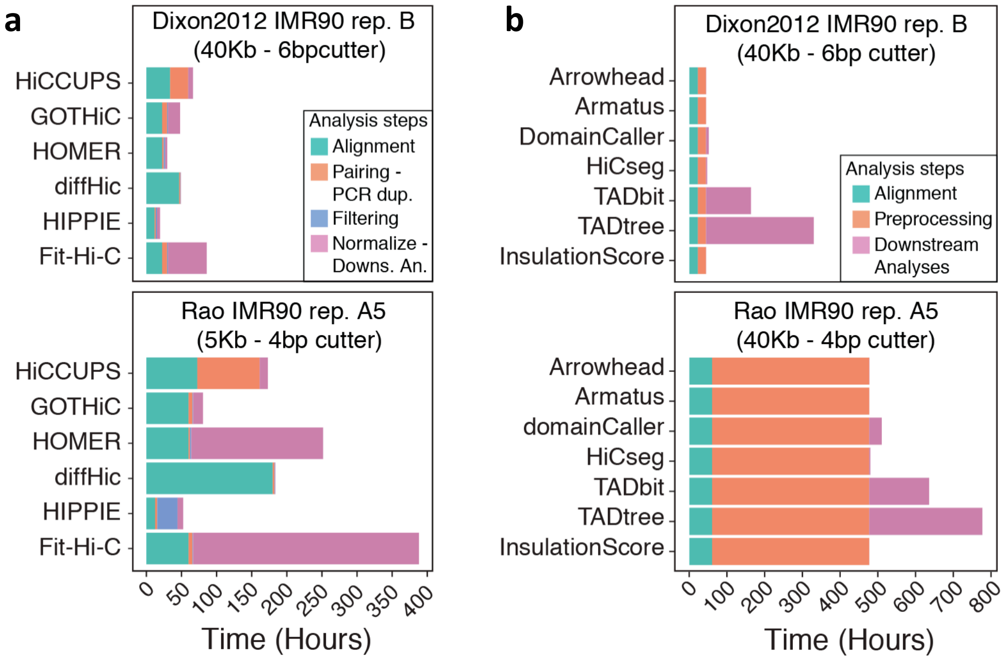

Figure 3.36: a) Time required by the various methods to perform alignment, reads pairing and PCR duplicates removal, other filtering, and normalization-downstream analysis for calling interactions in single replicates at different resolutions (replicate B of Dixon IMR90 at 40 kb and replicate A5 of Rao IMR90 at 5 kb; n=2 samples). The analyses were run on a single CPU and on a GPU for HiCCUPS. For GOTHiC, HOMER, and Fit-Hi-C the alignment time is relative to Bowtie. The time for reads pairing and PCR duplicates removal and other filtering of Fit-Hi-C corresponds to that of GOTHiC. b) Time required by the various methods to perform alignment, preprocessing (pairing, filtering, and normalization) and downstream analysis for TAD calling in replicates B of Dixon IMR90 and A5 of Rao IMR90 (n=2 samples). Alignment and preprocessing time are the same for all tools since all methods have been applied to a matrix generated by hicpipe. For TADbit, the time of downstream analysis also accounts for the normalization step. Both samples were analyzed at 40 kb resolution. However, Rao IMR90 replicate A5 required a higher preprocessing running time due to the large number of restriction fragments generated by the 4 bp cutter restriction enzyme.

3.2 TAD-AH

I tested TAD-AH on Hi-C data derived from 3 experimental conditions: i) human fibroblasts (IMR90), ii) human fibroblasts converted to myoblasts through MYOD overexpression and iii) differentiated into myotubes. TADs were identified using Armatus (Filippova et al., 2014) in a multi-scale fashion, on normalized Hi-C matrices at 4 kb resolution.

3.2.1 Filtering

The first step in TAD-AH analysis is filtering out the possible artifacts (e.g., TADs with boundaries overlapping regions characterized by low mappability) and duplicates (i.e., TADs with almost overlapping boundaries) among the domains given as input. Altogether, 28,791 out of 47,379 (60.8%) TADs were filtered from fibroblasts, 26,979 out of 44,866 (60.1%) from myoblasts and 21,648 out of 35,282 (61.4%) from myotubes. On average, 35.6% of the initial domains from each condition were discarded according to the filter on size (i.e., TADs smaller than 20 kb or bigger than 4 Mb were discarded), 14% were filtered as they represented duplicated TAD calls and 8.7% were discarded because their boundaries crossed other TADs. Only a small proportion of TADs were filtered because their boundaries were called near empty rows, resulting on average in 2.6% of the initial TADs for each condition. An example of TAD-AH filtering statistics is reported in Fig. 3.37).

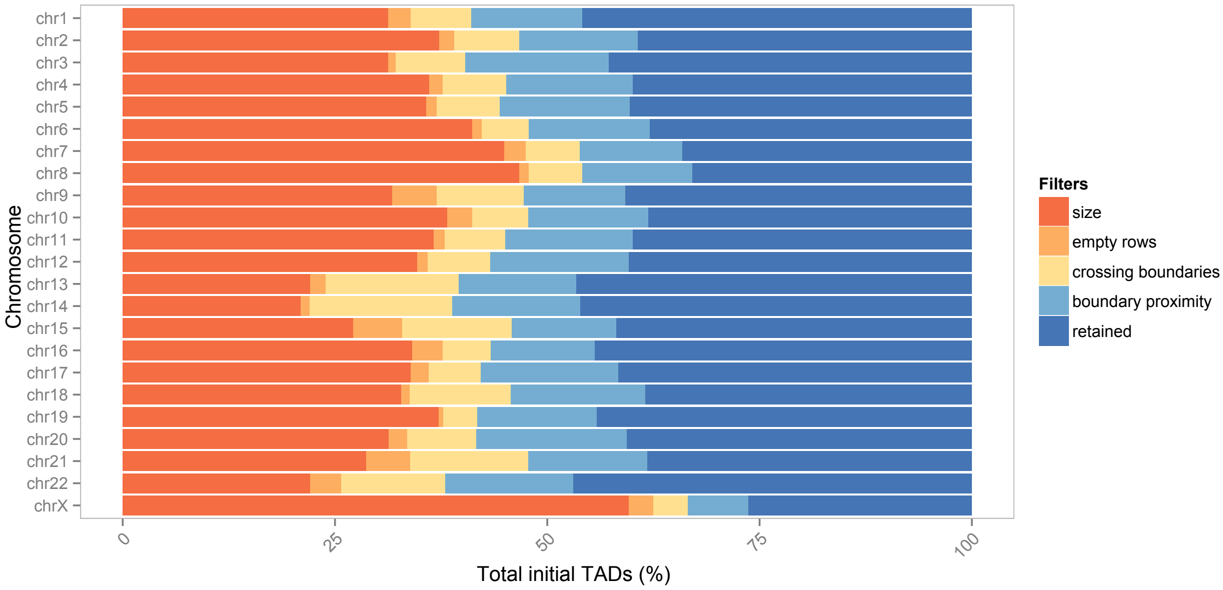

Figure 3.37: Barplot of TAD-AH filtering statistics of TAD called in fibroblasts. For each chromosome, the proportion of filtered and retained TADs is reported. Color legend: orange= TADs discarded by the size filter; light orange= TADs discarded because boundaries falling near empty rows; yellow= TADs discarded because crossing other TADs; light blue= TADs discarded because almost overlapping other TADs (duplicates); blue= TADs retained after filtering.

3.2.2 Classification and hierarchy reconstruction

Once obtained the refined TAD lists, TADs are classified into dense or loop-mediated, based on the ratio between the contact signal at the tip and at the middle portion of each TAD. Generally, dense TADs outnumber loop-mediated ones, which represent on average only the 10.8% of the TADs identified in each condition (Fig. 3.38a). The two categories also show distinct size distributions, with an average size of 199,100 bp for dense and 458,100 bp for loop-mediated TADs, respectively (Fig. 3.38b).

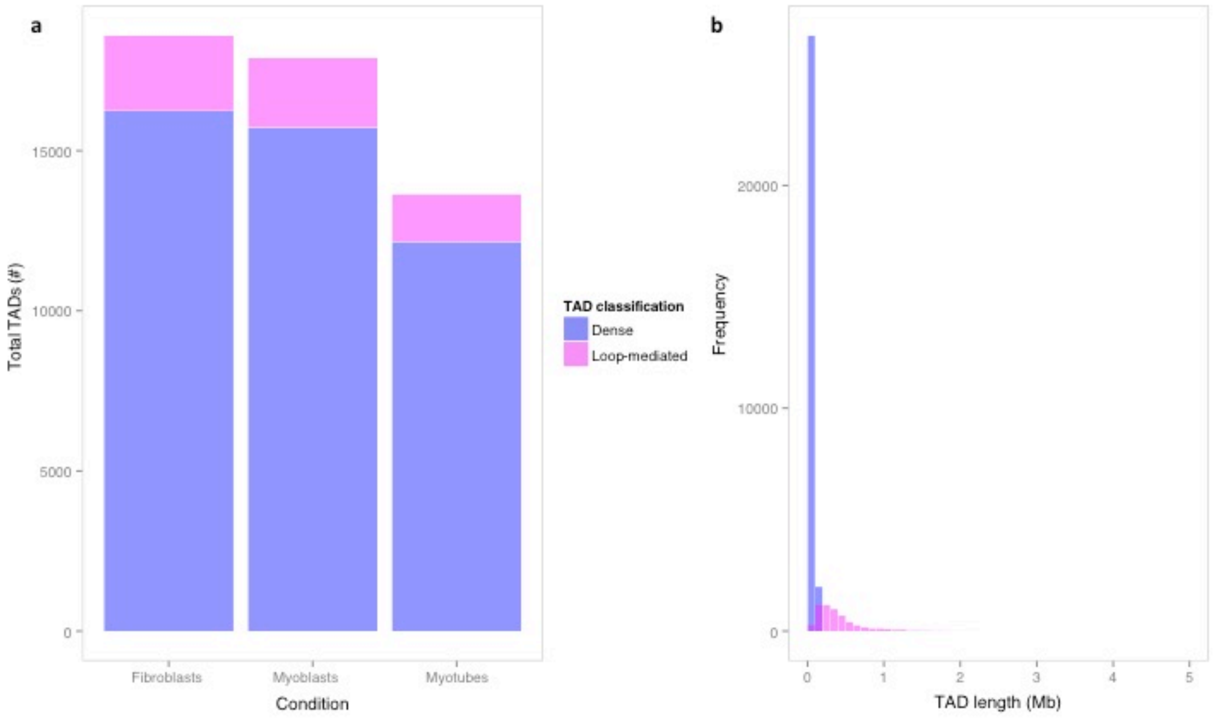

Figure 3.38: a) For each experimental condition, number of TADs classified as dense or loop-mediated; b) Size distribution for all the TADs found in the 3 experimental conditions. Color legend: blue= dense TADs; magenta= loop-mediated TADs.

When considering TADs hierarchical organization, only a small proportion of them (on average, 13.5% of post-filtering TADs) are found in a nested structure, and those who do are often found in trees composed by few TADs (Fig. 3.39).

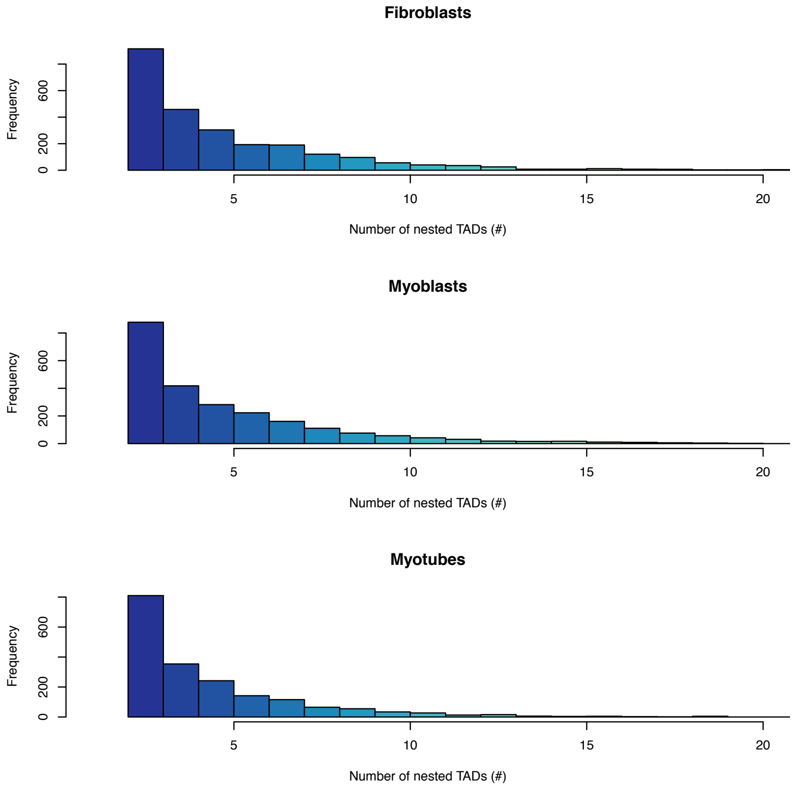

Figure 3.39: Frequency of TAD trees at increasing numbers of nested TADs, for the 3 experimental conditions.

3.2.3 Integration with other *omics*

After hierarchy reconstruction, TAD-AH integrates TAD positional information with other omics data, as ChIP-seq and RNA-seq. For this analysis, I used ChIP-seq data of MyoD binding in myoblasts and myotubes and RNA-seq data from all 3 experimental conditions.

Regarding the integration with ChIP-seq data, 42,397 out of 74,283 (57.1%) MyoD peaks bind to 9,585 TADs in myoblasts and 44,761 out of 110,676 (40.4%) MyoD peaks bind to 7,583 TADs in myotubes, respectively. Generally, MyoD preferentially binds dense TADs (31,852 out of 42,397 peaks for myoblasts and 34,296 out of 44,761 for myotubes). When plotting the distribution of the ChIP-seq peaks around the TAD boundaries, it can be appreciated how MyoD binding resembles the binding of CTCF, characterized by a sharp enrichment around the boundaries of both dense and loop-mediated TADs (Fig. 3.40).

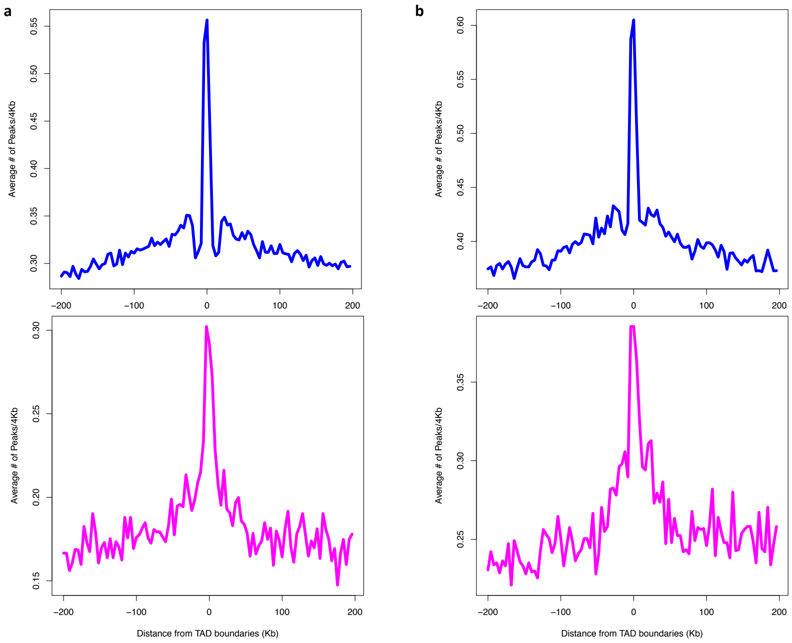

Figure 3.40: Frequency of MyoD ChIP-seq peaks in a window of 400 kb around TAD boundaries in a) myoblasts and b) myotubes. Boundaries that were shared by dense and loop-mediated TADs were excluded from the analysis. Color legend: blue= Dense TADs; magenta= loop-mediated TADs.

The expression levels of genes contained in TADs with MyoD bound on both boundaries (on average, 12.7% of TADs bound by MyoD in the two experimental conditions) are higher than the expression levels of genes contained in TADs with MyoD bound on just one boundary or in TADs with MyoD bound in the middle portions of the TAD or in TADs not bound by MyoD (Fig. 3.41a). In general, considering the gene expression in dense and loop-mediated TADs, it can be noticed that genes inside dense TADs tend to be more expressed than those inside loop-mediated TADs, irrespective of MyoD binding (Fig. 3.41b).

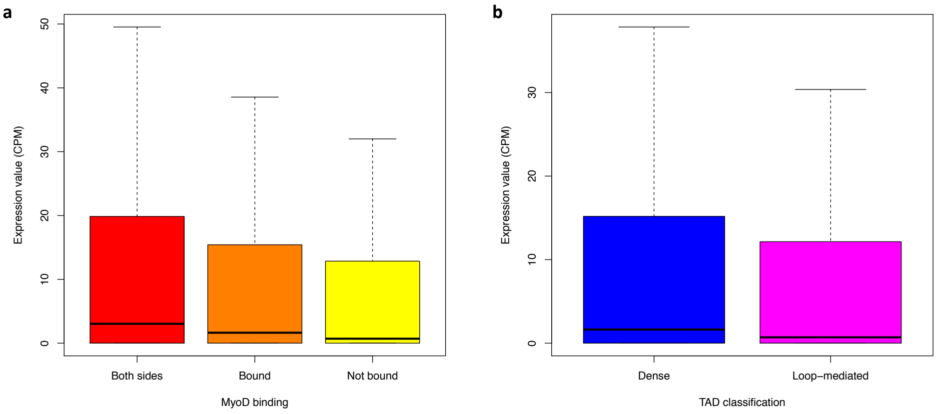

Figure 3.41: a) Expression (presented as counts per million) of genes inside TADs either bound at both boundaries by MyoD (red), bound on one side or inside the TAD (orange) or not bound by MyoD (yellow);

b) Expression of genes inside dense (blue) or loop-mediated TADs (magenta). Gene expression values are from myoblasts, similar results were obtained also in myotubes (not shown).

3.2.4 Differential analysis

The last step in TAD-AH analysis consists in identifying TADs that are conserved, acquired or lost between conditions and that change their characteristics (i.e., switch from dense to loop-mediated classification). On average, less than 20% of the total TADs found in the two conditions are conserved (Fig. 3.42a), and 94.6% of them maintain the same classification across conditions. In all cases, the major shifts are from loop-mediated to dense (Fig. 3.42b).

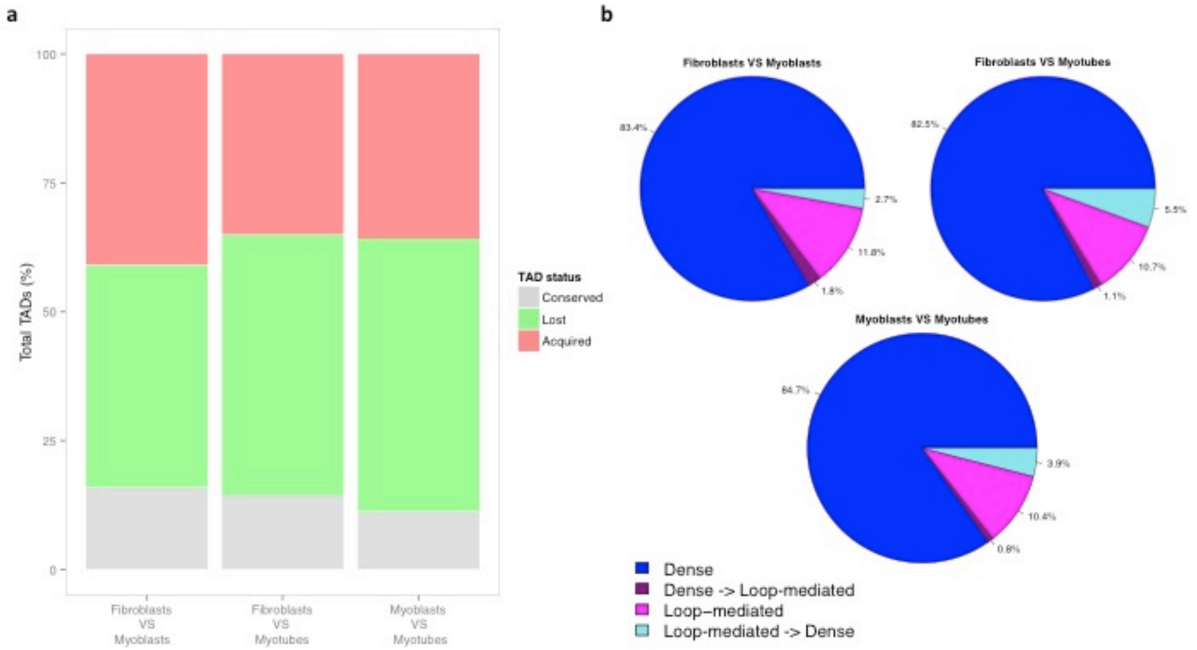

Figure 3.42: Differential TAD analysis for all pair-wise comparisons. a) Proportion of total TADs identified that were conserved, acquired or lost in the compared conditions; b) Proportion of conserved TADs that maintained or changed classification in each comparison. Color legend: grey= conserved TADs; green= lost TADs; red= acquired TADs; blue= dense TADs; magenta= loop-mediated TADs; light blue= switch from loop-mediated (condition 1) to dense (condition 2); dark magenta= switch from dense (condition 1) to loop-mediated (condition 2).

Chapter 4

Conclusions

In the last few years the study of genome topology has experienced an extraordinary push forward, thanks to the advent – and constant improvement – of the 3C-derived techniques. Initially designed to study single loci in detail, these techniques (in particular, Hi-C) now succeed to produce a whole picture of the chromatin contacts that occur inside the nucleus, at unprecedented resolution and even at single cell level. As the amount and resolution of generated data increase, we start to comprehend how chromatin structure is exploited to establish a further level of transcription regulation, which will ultimately shed light on the sequence of events that leads to disease when chromatin organization is perturbed. Indeed, it has already been observed how alterations of well-established chromatin structures (e.g., chromatin contacts and topologically associating domains), caused by mutations in regulatory sequences, drive the epigenetic and, ultimately, the transcriptional perturbations responsible for inherited diseases as well as cancer onset (Lupianez et al., 2015; Valton and Dekker, 2016).

The enormous amount of sequencing data produced by Hi-C required the development of ad-hoc computational analysis strategies, which differ at various levels, from the number of implemented analysis steps to the approaches adopted for each phase. Nevertheless, the reliability of the chromatin structures identified by these methods remains an open question, and attempts to compare the performance of tools for chromatin interactions and TADs discovery consisted only in semi-quantitative approaches (Lun et al., 2015; Lèvy-Leduc et al., 2014; Weinreb et al., 2016; Filippova et al., 2014).

Certainly, robust comparisons in terms of sensitivity and specificity are precluded by the absence of a substantial amount of validated chromatin contacts to use as positive and negative controls and by the conceptual challenge in generating simulated Hi-C data. In the attempt to overcome these limitations, we developed a computational framework that uses a set of metrics to quantitatively compare the performance of various approaches for Hi-C data analysis, applied to a large set of experimental as well as simulated data.

The methods we considered implement different preprocessing strategies to align the sequencing reads and to filter out reads affected by sequencing and protocol artifacts. Results showed that the chimeric alignment could map many more reads than a full-read aligner, with this effect becoming more prominent as the read length increases. Indeed, longer reads can be expected to cross through the ligation junction thus becoming chimeric. Interestingly, the number of reads retained after filtering is influenced more by the experimental protocol than by the type of filter performed by the various tools. Irrespective of the filtering strategy, the combination of the in-situ Hi-C protocol and a chimeric aligner yielded the largest quantity of reads for downstream analysis.

Results suggest that no method can be defined as the gold standard to identify chromatin interactions and that the choice of the algorithm affects the number and features of the called interactions. For example, with experimental data, GOTHiC called the largest number of cis interactions at all resolutions. However, most of these interactions are at short distance, including interaction between adjacent bins, which, although potentially informative, might be difficult to isolate from the background noise. On the other extreme, HiCCUPS was the tool that identified the smallest absolute number of significant interactions. However, it is worthwhile noting that, differently from any other tool, HiCCUPS aggregates nearby peaks into a single interaction. Moreover, for a fair comparison, we used HiCCUPS at a single resolution (i.e., 5 kb or 40 kb), whereas the method default settings involve combining interactions called at multiple resolutions. This, as we also verified, would yield a larger number of interactions. Hi-C replicates are usually combined before the analysis to generate a unique sample, thus increasing the number of reads. Here, to quantitatively measure the conservation of the called interactions, we kept replicates separated. Unexpectedly, interactions identified in one sample were poorly conserved when considering other replicates from the same cell type of the same study. Although limited, the intra-dataset reproducibility was, in almost all cases, significantly higher than what expected from random sets of interactions. The reproducibility could be only marginally ameliorated selecting the top significant interactions identified by each method or replicates with a similar number of reads and interactions. Moreover, irrespectively of the tool, the inter-dataset concordance was even scarcer.

These results, though far from being satisfactory, are rather predictable if we consider that Hi-C contact maps derive from a cell population, which consists of cells in different biological states (e.g., cell cycle phase, response to extracellular stimuli, to name a few) that inevitably will vary from one another on many local chromatin arrangements, and that, in conjunction with an inadequate sequencing depth, will fill the Hi-C contact matrix with noisy interactions, from which the truly conserved ones from the population of cells will struggle to emerge.

Aside from the limited reproducibility of chromatin contacts, all algorithms found equivalent, statistically significant amounts of cis promoter-enhancer looping interactions and very few considered as biologically less plausible. The various methods performed differently when considering the ability to recall previously validated or reported cell type-specific cis interactions. In particular, diffHic and GOTHiC are the methods that recall more true-positives with high- and low-resolution data, respectively. This could be ascribed to the fact that both methods called, in general, the largest number of total interactions and also explains the presence of some true negatives among the interactions called by GOTHiC. In most cases, HOMER performed similarly to GOTHiC, although calling fewer interactions. Synthetic data indicated that all tools have an extremely low precision in the identification of simulated interactions.

Differently from interaction callers, the methods for the identification of TADs had similar performances when using experimental data, although returning different numbers of TADs with different mean size. These results are consistent with what recently described by (Dali and Blanchette, 2017). Almost all tools predicted concordant domains characterized by a significant reproducibility in TAD boundaries and enrichment in binding sites of known architectural proteins. Simulated data of not nested TADs highlighted that some methods (as DomainCaller, TADbit, and InsulationScore) are more robust than others in the identification of the correct number and size of TADs even at high levels of noise. TADbit showed the best balance between sensitivity and precision.

In general, this study indicates that, while no single method outperforms others in every setting, TAD callers produce more comparable results, thus being, from a methodological point of view, riper than interaction callers. Amid TAD callers, TADbit, Armatus, and TADtree performed analogously for most metrics in experimental and simulated data. Regarding interaction callers, HOMER and HiCCUPS returned the highest proportion of interactions with a potential biological significance, though HiCCUPS full potential can only be realized in the analysis of very high-resolution Hi-C data, as it requires a remarkable read depth to call a substantial number of interactions.

Results obtained from experimental data frequently diverged from what observed with synthetic data, particularly for interaction callers. This is most likely due to the difficulty in modeling sound approaches to simulate Hi-C data including specific elements, as TADs and interactions, which serve as univocal true positives and negatives. Despite the availability of different promising strategies from the biophysics of polymer folding modeling (Imakaev et al., 2015), no method has yet been suggested to simulate the creation of reads that accurately reproduce the distribution and biases distinctive of real Hi-C data.

The development of simulated data will also be fundamental to rationally adjust each algorithm parameter, hence reducing the heuristics now intrinsic in the determination of the best analysis setting.

The different tools significantly vary also in terms of stability of the implementation, interoperability, usability, and required computing resources. HOMER does not necessitate an in-depth bioinformatics expertise and presents little computational requirements, while HiCCUPS, through Juicer tools, provides effective approaches for data storage and visualization, but it is more demanding in terms of computational resources. Considering the rate of data production, at increasing resolution levels, developers should focus on the implementation of methods capable to analyze larger datasets within reasonable amounts of time and on the choice of shared data formats to let an easy conversion of inputs and outputs between the various approaches (Dekker et al., 2017).

The comparison of the computational methods for TADs identification let emerging that little has been done to fully comprehend the nature of TADs, often described only by the boundaries that separate one domain from another, ignoring their hierarchical structure or differences in their characteristics, as, importantly, the presence of domains with more diffuse contact signal (here defined as dense TADs) and of TADs where the signal is essentially confined to the TAD tip and edges (here defined as loop-mediated TADs). Moreover, no algorithm addresses the differential analysis between domains found in different biological settings.

For these reasons, I developed TAD-AH, a method to refine TAD calls obtained at multi-scale level, which applies several filters to discard artifacts and duplicated TAD calls, classifies domains based on their signal distribution and saves a hierarchical TAD list. Moreover, TAD-AH integrates the resulting list with other omics data (as RNA-seq and ChIP-seq data) to further characterize TADs, and performs differential analysis between TAD calls from different conditions.

I tested TAD-AH on Hi-C data at 4 kb resolution derived from the trans-differentiation of human fibroblasts into muscle cells, upon expression of myogenic master regulator MYOD, which comprises 3 states: fibroblasts, myoblasts and myotubes. TADs were identified with Armatus in multi-scale mode, i.e., collecting TAD calls at each resolution instead of considering only the consensus set of domains found at all resolution. TAD-AH filtered out more than half of the original TAD calls for each experimental condition, mainly because Armatus tends to find very small domains, and also because many TADs were duplicates. Few TADs were discarded because their boundaries fell near empty rows (i.e., low mappability regions) or because they crossed other domains. When considering TADs classification, TAD-AH found that, for all experimental conditions, dense TADs outnumber loop-mediated ones and are characterized by a lower average size. Only a small proportion of TADs (around 13% of post-filtering TADs) appear to be organized in nested structures, which usually include less than 10 TADs. This result is probably due to the inability of Armatus to identify all nested TADs in the first place.

TADs integration with ChIP-seq and RNA-seq data revealed that MyoD peaks are frequently found at TAD boundaries and that genes inside domains with MyoD bound at both boundaries tend to be more expressed than those inside domains where MyoD binds in other portions of the TAD or that are not bound by MyoD. Furthermore, irrespective of MyoD binding, genes inside dense TADs are generally more expressed than those inside loop-mediated ones. Regarding the differential analysis, only a small proportion (around

20%) of total called TADs seems to be conserved across conditions, even though it must be noted that, at 4 kb resolution, the impact of small shifts in the boundary coordinates had a great impact in determining such a low conservation: the same analyses, carried out at 40 kb resolution, resulted in a conservation of almost 50% of the TADs for all comparisons (results not shown). Among the conserved TADs, the majority maintains the same classification across conditions, and the main changes concern the switch of loop-mediated TADs into dense ones.

In conclusion, TAD-AH analysis revealed that TAD characterization can be helpful for the interpretation of the biological role of topologically associating domains, but its potential is hampered by two major limitations, i.e., the genome coverage, and the precision in TAD identification. Briefly, data at very high genome coverage is mandatory to really appreciate TAD hierarchies extending far from the Hi-C matrix diagonal. Indeed, if the coverage is not sufficiently high, regions away from the main diagonal, where most of the signal concentrates, remain sparse in signal and this hampers the ability of any method to detect differences in higher hierarchy structures. Moreover, a higher genome coverage would allow reaching higher resolution, thus further improving the identification of TAD hierarchies. Lastly, higher genome coverage would facilitate TAD calling algorithms in finding shifts in intra-contact versus inter-contact signal, thus enabling a more precise identification of boundaries and – hopefully – leading to the identification of TADs (especially nested ones) that, even if visible from the heatmap of the Hi-C matrix, are still not recognized.

Chapter 5

References

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Andrey, G., Montavon, T., Mascrez, B., Gonzalez, F., Noordermeer, D., Leleu, M., Trono, D., Spitz, F., and Duboule, D. (2013). A switch between topological domains underlies HoxD genes collinearity in mouse limbs. Science *340*, 1234167.

Apostolou, E., Ferrari, F., Walsh, R.M., Bar-Nur, O., Stadtfeld, M., Cheloufi, S., Stuart, H.T., Polo, J.M., Ohsumi, T.K., Borowsky, M.L., et al. (2013). Genome-wide chromatin interactions of the Nanog locus in pluripotency, differentiation, and reprogramming. Cell Stem Cell *12*, 699–712.

Ay, F., and Noble, W.S. (2015). Analysis methods for studying the 3D architecture of the genome. Genome Biol. *16*, 183.

Ay, F., Bailey, T.L., and Noble, W.S. (2014). Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res. *24*, 999–1011.

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Müller, S., Eils, R., Cremer, C., Speicher, M.R., et al. (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. PLoS Biol. *3*, e157.

Bonev, B., and Cavalli, G. (2016). Organization and function of the 3D genome. Nature Reviews Genetics *17*, nrg.2016.112.

Cavalli, G., and Misteli, T. (2013). Functional implications of genome topology. Nat. Struct. Mol. Biol. *20*, 290–299.

Celniker, S.E., Dillon, L.A.L., Gerstein, M.B., Gunsalus, K.C., Henikoff, S., Karpen, G.H., Kellis, M., Lai, E.C., Lieb, J.D., MacAlpine, D.M., et al. (2009). Unlocking the secrets of the genome. Nature *459*, 927–930.

Cournac, A., Marie-Nelly, H., Marbouty, M., Koszul, R., and Mozziconacci, J. (2012). Normalization of a chromosomal contact map. BMC Genomics *13*, 436.

Crane, E., Bian, Q., McCord, R.P., Lajoie, B.R., Wheeler, B.S., Ralston, E.J., Uzawa, S., Dekker, J., and Meyer, B.J. (2015). Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature *523*, 240–244.

Dali, R., and Blanchette, M. (2017). A critical assessment of topologically associating domain prediction tools. Nucleic Acids Res.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306–1311.

Dekker, J., Marti-Renom, M.A., and Mirny, L.A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat. Rev. Genet. *14*, 390–403.

Dekker, J., Belmont, A.S., Guttman, M., Leshyk, V.O., Lis, J.T., Lomvardas, S., Mirny, L.A., O'Shea, C.C., Park, P.J., Ren, B., et al. (2017). The 4D nucleome project. Nature *549*, 219–226.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331–336.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. *16*, 1299–1309.

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst *3*, 95–98.

Eagen, K.P., Aiden, E.L., and Kornberg, R.D. (2017). Polycomb-mediated chromatin loops revealed by a subkilobase-resolution chromatin interaction map. Proc. Natl. Acad. Sci. U.S.A. *114*, 8764–8769.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Ferraiuolo, M.A., Rousseau, M., Miyamoto, C., Shenker, S., Wang, X.Q.D., Nadler, M., Blanchette, M., and Dostie, J. (2010). The three-dimensional architecture of Hox cluster silencing. Nucleic Acids Res. *38*, 7472–7484.

Filippova, D., Patro, R., Duggal, G., and Kingsford, C. (2014). Identification of alternative topological domains in chromatin. Algorithms Mol Biol *9*, 14.

Forcato, M., Nicoletti, C., Pal, K., Livi, C.M., Ferrari, F., and Bicciato, S. (2017). Comparison of computational methods for Hi-C data analysis. Nature Methods 14, 679–685.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L.A. (2016). Formation of Chromosomal Domains by Loop Extrusion. Cell Rep *15*, 2038–2049.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., et al. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. Nature *462*, 58–64.

Gruber, S. (2017). Shaping chromosomes by DNA capture and release: gating the SMC rings. Curr. Opin. Cell Biol. *46*, 87–93.

Haarhuis, J.H.I., van der Weide, R.H., Blomen, V.A., Yáñez-Cuna, J.O., Amendola, M., van Ruiten, M.S., Krijger, P.H.L., Teunissen, H., Medema, R.H., van Steensel, B., et al. (2017). The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. Cell *169*, 693–707.e14.

He, B., Chen, C., Teng, L., and Tan, K. (2014). Global view of enhancer-promoter interactome in human cells. Proc. Natl. Acad. Sci. U.S.A. *111*, E2191-2199.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Ho, J.W.K., Jung, Y.L., Liu, T., Alver, B.H., Lee, S., Ikegami, K., Sohn, K.-A., Minoda, A., Tolstorukov, M.Y., Appert, A., et al. (2014). Comparative analysis of metazoan chromatin organization. Nature *512*, 449–452.

Hou, C., Li, L., Qin, Z.S., and Corces, V.G. (2012). Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. Mol. Cell *48*, 471–484.

Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B., and Liu, J.S. (2012). HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics *28*, 3131–3133.

Hughes, J.R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., De Gobbi, M., Taylor, S., Gibbons, R., and Higgs, D.R. (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat. Genet. *46*, 205–212.

Hwang, Y.C., Lin, C.F., Valladares, O., Malamon, J., Kuksa, P.P., Zheng, Q., Gregory, B.D., and Wang, L.-S. (2015). HIPPIE: a high-throughput identification pipeline for promoter interacting enhancer elements. Bioinformatics *31*, 1290–1292.

Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J., and Mirny, L.A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat. Methods *9*, 999–1003.

Imakaev, M.V., Fudenberg, G., and Mirny, L.A. (2015). Modeling chromosomes: Beyond pretty pictures. FEBS Lett. *589*, 3031–3036.

Ji, X., Dadon, D.B., Powell, B.E., Fan, Z.P., Borges-Rivera, D., Shachar, S., Weintraub, A.S., Hnisz, D., Pegoraro, G., Lee, T.I., et al. (2016). 3D chromosome regulatory landscape of human pluripotent cells. Cell Stem Cell *18*, 262–275.

Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.-A., Schmitt, A.D., Espinoza, C.A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature *503*, 290–294.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

Knight, P.A., and Ruiz, D. (2013). A fast algorithm for matrix balancing. IMA J Numer Anal *33*, 1029–1047.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lévy-Leduc, C., Delattre, M., Mary-Huard, T., and Robin, S. (2014). Two-dimensional segmentation for analyzing Hi-C data. Bioinformatics *30*, i386-392.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Lonfat, N., Montavon, T., Darbellay, F., Gitto, S., and Duboule, D. (2014). Convergent evolution of complex regulatory landscapes and pleiotropy at Hox loci. Science *346*, 1004–1006.

Lun, A.T.L., and Smyth, G.K. (2015). diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. BMC Bioinformatics *16*, 258.

Lupiáñez, D.G., Spielmann, M., and Mundlos, S. (2016). Breaking TADs: How Alterations of Chromatin Domains Result in Disease. Trends Genet. 32, 225–237.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012–1025.

Ma, W., Ay, F., Lee, C., Gulsoy, G., Deng, X., Cook, S., Hesson, J., Cavanaugh, C., Ware, C.B., Krumm, A., et al. (2015). Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. Nat. Methods *12*, 71–78.

Marco-Sola, S., Sammeth, M., Guigó, R., and Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. Nat. Methods *9*, 1185–1188.

Mifsud, B., Martincorena, I., Darbo, E., Sugar, R., Schoenfelder, S., Fraser, P., and Luscombe, N.M. (2017). GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data. PLoS ONE 12, e0174744.

Mora, A., Sandve, G.K., Gabrielsen, O.S., and Eskeland, R. (2016). In the loop: promoter-enhancer interactions and bioinformatics. Brief. Bioinformatics *17*, 980–995.

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature *502*, 59–64.

Nagano, T., Lubling, Y., Várnai, C., Dudley, C., Leung, W., Baran, Y., Mendelson Cohen, N., Wingett, S., Fraser, P., and Tanay, A. (2017). Cell-cycle dynamics of chromosomal organization at single-cell resolution. Nature *547*, 61–67.

Noordermeer, D., and Duboule, D. (2013). Chromatin looping and organization at developmentally regulated gene loci. Wiley Interdiscip Rev Dev Biol *2*, 615–630.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Phillips-Cremins, J.E., Sauria, M.E.G., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S.K., Ong, C.-T., Hookway, T.A., Guo, C., Sun, Y., et al. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell *153*, 1281–1295.

Pombo, A., and Dillon, N. (2015). Three-dimensional genome architecture: players and mechanisms. Nat. Rev. Mol. Cell Biol. *16*, 245–257.

Pope, B.D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D.L., Wang, Y., Hansen, R.S., Canfield, T.K., et al. (2014). Topologically associating domains are stable units of replication-timing regulation. Nature *515*, 402–405.

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

Rao, S.S.P., Huang, S.-C., Glenn St Hilaire, B., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.-R., Sanborn, A.L., Johnstone, S.E., Bascom, G.D., Bochkov, I.D., et al. (2017). Cohesin Loss Eliminates All Loop Domains. Cell *171*, 305–320.e24.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Rowley, M.J., Nichols, M.H., Lyu, X., Ando-Kuri, M., Rivera, I.S.M., Hermetz, K., Wang, P., Ruan, Y., and Corces, V.G. (2017). Evolutionarily Conserved Principles Predict 3D Chromatin Organization. Mol. Cell *67*, 837–852.e7.

Sanborn, A.L., Rao, S.S.P., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. U.S.A. *112*, E6456-6465.

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109–113.

Sauria, M.E.G., Phillips-Cremins, J.E., Corces, V.G., and Taylor, J. (2015). HiFive: a tool suite for easy and efficient HiC and 5C data analysis. Genome Biol. *16*, 237.

Schmitt, A.D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C.L., Li, Y., Lin, S., Lin, Y., Barr, C.L., et al. (2016a). A compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell Rep *17*, 2042–2059.
Schmitt, A.D., Hu, M., and Ren, B. (2016b). Genome-wide mapping and analysis of chromosome architecture. Nat. Rev. Mol. Cell Biol. *17*, 743–755.

Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.-M., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S.W., et al. (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res. *25*, 582–597.

Serra, F., Baù, D., Goodstadt, M., Castillo, D., Filion, G.J., and Marti-Renom, M.A. (2017). Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. PLoS Comput. Biol. 13, e1005665.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. *16*, 259.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. Cell *148*, 458–472.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat. Genet. *38*, 1348–1354.

Smith, E.M., Lajoie, B.R., Jain, G., and Dekker, J. (2016). Invariant TAD boundaries constrain cell-type-specific looping interactions between promoters and distal elements around the CFTR locus. Am. J. Hum. Genet. *98*, 185–201.

Teng, L., He, B., Wang, J., and Tan, K. (2015). 4DGenome: a comprehensive database of chromatin interactions. Bioinformatics *31*, 2560–2564.

Valton, A.-L., and Dekker, J. (2016). TAD disruption as oncogenic driver. Curr. Opin. Genet. Dev. *36*, 34–40.

Wang, H., Maurano, M.T., Qu, H., Varley, K.E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., Sandstrom, R., et al. (2012). Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome Res. *22*, 1680–1688.

Weinreb, C., and Raphael, B.J. (2016). Identification of hierarchical chromatin domains. Bioinformatics *32*, 1601–1609.

Woon Kim, Y., Kim, S., Geun Kim, C., and Kim, A. (2011). The distinctive roles of erythroid specific activator GATA-1 and NF-E2 in transcription of the human fetal γ-globin genes. Nucleic Acids Res. *39*, 6944–6955.

Yaffe, E., and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat. Genet. *43*, 1059–1065.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.