UNIVERSITY OF ESSEX

Department of Mathematical Sciences

A Novel Statistical Framework to Detect Complex 3D Genome Organisation Patterns into Topologically Associated Domains

by

Liudmila Mikheeva

A thesis submitted for the degree of

Doctor of Philosophy in Statistics

January 2022

Abstract

Topologically associated domains (TADs) are highly compacted regions of DNA that are suggested to be involved in proper gene regulation and cellular functioning. TADs maintain long-range interactions between distal enhancers and target genes, as well as restrict enhancers contacting genes that are not their target and, consequently, block their inappropriate regulation by these enhancers. The widely used TAD calling tools either restrict TAD borders to be allocated in a “head-to-tail” manner or allow hierarchical TAD folding to be detected. We propose a R-based TAD calling tool that detects start and end TAD border positions separately, so the partial overlapping of TADs as well as large gaps between TADs are also allowed. Using the ratio between the average upstream and downstream Hi-C interaction frequencies, our method detects where the difference between inside-TAD and outside TAD area within the Hi-C matrix is most significant. The novel TAD allocation combined with various genomics data reveals the interplay between architectural proteins and active transcription in the establishment of the TAD border insulation strength and insulation imbalances between neighbouring TADs.

Contents

List of figures

List of tables

List of appendices

List of abbreviations and acronyms

ash-1
: absent, small, or homeotic discs 1

BEAF-32
: boundary element-associated factor of 32kD

BG3
: *Drosophila melanogaster* larval central nervous system cell line

bp
: base pair

BWA-mem
: Burrows-Wheeler aligner with maximal exact matches

C. elegans
: *Caenorhabditis elegans*

ChIP
: chromatin immunoprecipitation

ChIP-chip
: ChIP-on-chip

Chro/Chriz
: chromator

COrTADo
: complex organisation of topologically associated domains

Cp190
: centrosomal protein 190kD

CTCF/dCTCF
: CCCTC-binding factor

CUSUM method
: cumulative sum method, developed by (Page 1954)

DEG
: differentially expressed genes

DESeq2
: DEG analysis based on the negative binomial distribution (Love et al. 2014)

DI
: directionality index (Dixon et al. 2012)

dm3
: release 5 of the *Drosophila melanogaster* genome

dm6
: release 6 of the *Drosophila melanogaster* genome

DNA
: deoxyribonucleic acid

DNase-I
: deoxyribonuclease I

DpnII, HindIII, NcoI
: restriction enzymes

Dref
: DNA replication-related element factor

dRING/Sce
: sex combs extra

Drosophila
: *Drosophila melanofaster*

EPHA4
: ephrin type-A receptor 4 encoded by *Epha4* gene

ESC
: embryonic stem cell

E(z)
: enhancer of zeste

FDR
: false discovery rate

FISH
: fluorescent in situ hybridisation

fs(1)h
: female sterile (1) homeotic

GAF
: GAGA transcription factor, encoded by the Trithorax-like (Trl) gene

GAM
: genome architecture mapping (Beagrie et al. 2017)

GC-content
: guanine-cytosine content

GEO
: The National Center for Biotechnology Information Gene Expression Omnibus

Hi-C
: high-throughput derivation of 3C

HP1a/Su(var)205
: suppressor of variegation 205

HP1b
: heterochromatin protein 1b

HP1c
: heterochromatin protein 1c

HP2/Su(var)2-HP2
: heterochromatin protein 2

HP4
: heterochromatin protein 4

H1, H2Av, H3, H4
: histones H1, H2A variant, H3, H4

H2Bubi
: mono-ubiquitylation of histone H2B

H3K4me1, H3K27me1, H3K36me1, H3K79me1
: mono-methylation of histone 3 on lysine 4, 27, 36, 79

H3K4me2, H3K9me2, H3K27me2, H3K79me2
: di-methylation of histone 3 on lysine 4, 9, 27, 79

H3K4me3, H3K9me3, H3K27me3, H3K36me3, H3K79me3
: tri-methylation of histone 3 on lysine 4, 9, 27, 36, 79

H3K18ac, H3K23ac
: histone H3 acetylated on lysine 18, 23, 27

H4K8ac, H4K16ac
: acetylation of histone H3 on lysine 8, 16

H4K20me1
: mono-methylation of histone 4 on lysine 20

ICE
: iterative correction and eigenvector decomposition (Imakaev et al. 2012)

IP
: immunoprecipitation

IS
: insulation score (Crane et al. 2015)

Iswi/NURF
: imitation SWI

JHDM1/Kdm2
: lysine demethylase 2

JIL-1
: JIL-1 kinase

Kb
: kilobase, 1000 bp

Kc167
: *Drosophila melanogaster* embryonic cell line

KD
: knockdown

KR balancing
: Knight-Ruiz balancing (Knight and Ruiz 2013)

log2
: logarithm to base 2

MA
: moving average

Mb
: megabase, 1000000 bp

MED1
: mediator complex subunit 1

MED30
: mediator complex subunit 30

modENCODE
: Model Organism Encyclopedia of DNA Elements

mod(mdg4)
: modifier of mdg4

mof
: males absent on the first

MRG15
: MORF-related gene 15

mRNA
: messenger RNA

NA
: not available

NURF301/NURF/E(bx)
: enhancer of bithorax

Orc2
: origin recognition complex sub-unit 2

Pc
: Polycomb

PcG
: Polycomb group

Pcl
: Polycomblike

PCR
: polymerase chain reaction

pdf
: probability density function

Pof
: painting of fourth

Pol II/RNAPII
: RNA polymerase II

PRC1
: Polycomb repressive complex 1

PR-Set7
: PR/SET domain containing protein 7

Psc
: posterior sex combs

R
: programming language

Rad21/vtd/cohesin
: verthandi

RNA
: ribonucleic acid

RNAi
: RNA interference

RNA-seq
: RNA sequencing

Rpd3/HDAC1
: histone deacetylase 1

RUNX1
: runt-related transcription factor 1 encoded by *Runx1* gene

SA
: stromalin

SAFE Hi-C
: simplified, amplification-free, and economically efficient process Hi-C (Niu et al. 2019)

SCN
: sequential component normalisation (Cournac et al. 2012)

Sfmbt
: scm-related gene containing four mbt domains

SMC1/cohesin
: structural maintenance of chromosomes 1

SPRITE
: split-pool recognition of interactions by tag extension (Quinodoz et al. 2018)

su(Hw)
: suppressor of Hairy wing

Su(var)3-7
: suppressor of variegation 3-7

Su(var)3-9
: suppressor of variegation 3-9

TAD
: topologically associated domains

TE
: transposable element

Top1
: type 1 topoisomerase

Top2/TopoII
: type 2 topoisomerase

TSS
: transcription start sites

TTS
: transcription termination sites

wds
: will die slowly

WT
: wild-type

Zw5/dwg
: deformed wings

3C
: chromosome conformation capture

3D
: three dimensions

4C
: chromosome conformation capture-on-chip

5C
: chromosome conformation capture carbon copy

Statement of originality

I declare that the content of this thesis was performed by myself, unless otherwise specified. All sources used have been appropriately referenced. The analysis presented in the Chapter 2 was performed by myself and was published as a part of (Chathoth et al. 2022). Results obtained by authors of the paper other than me has been appropriately referenced. The analysis presented in the Chapter 3 and Chapter 4 was performed by myself and is available at the follow GitHub link:

https://github.com/lm17047/COrTADo.git

Chapter 1. Introduction

1.1. General view on chromatin folding

A cell is a fundamental unit of biological organisation. There are two main forms: prokaryotic cells that are found in bacteria and archaea and eukaryotic cells that are found in all other living organisms. Eukaryotic cells differ by the presences of a nucleus with DNA stored inside while in prokaryotic cells DNA is found free in the cytoplasm.

DNA encodes many small regions called genes that act as an instruction for coding proteins. In the human genome, genes represent only 1% of the genome. The non-coding genome contains regulatory regions that provide binding sites for transcription factors which control gene activation and/or repression. The list of such regulatory regions mostly includes promoters, enhancers, silencers and insulators. Promoters are typically allocated ahead of genes and carries binding sites for transcription machinery attachment. Enhancers participate in gene activation: proper physical contacts between transcription factors binding enhancers and transcription machinery binding promoters are essential for transcription initiation (Bjorkegen and Baranello 2018). In contrast with enhancers, silencers are involved in transcription repression: contact with a silencer can block the transcription machinery and stop the transcription of DNA sequencing into RNA. Enhancers and silencers can be distal and the main role of the DNA conformation is to bring them into spatial proximity with their target genes (Cremer and Cremer 2001; Lieberman-Aiden et al. 2009). At the same time, the contacts between enhancers/silencers with non-target genes can result in undesired gene upregulation or downregulation. To do so, insulator regions carries binding sites of proteins that can create barriers to separate improper enhancer-promoter contacts (enhancer-blocking activity), as well as separate repressive chromatin and maintain a sufficient level of accessibility for protein complexes involved in replication or transcription processes (Dali and Blanchette 2017; Zabet and Adryan 2015).

While the diameter of a nucleus is scaled in micrometers, stretched DNA, for example in human, is up to several meters long, which means that DNA is compactly packed inside the nucleus. On the top level of spatial organisation, DNA wraps twice around

proteins called histones and forms structural units called nucleosomes (Bentley et al. 1984; Luger et al. 1997). At the next level, the chromatin (complex of DNA and surrounding proteins) is packed into 30-nm fiber, which is characterised by nucleosome interactions.

1.2. Hi-C is a powerful tool to explore the chromatin architecture

1.2.1. First revelations on chromatin topology were microscopy-based

Historically, the first insights on three-dimensional chromatin organisation, chromosome positioning and DNA-DNA interactions were based on a variety of microscopy techniques (Fraser et al. 2015). The most commonly used microscopy-based method was fluorescent in situ hybridisation (FISH) (Langer-Safer et al. 1982). FISH allowed the extraction of basic features of chromatin topography from a limited number of predetermined loci in individual cells. Based on FISH, some fundamental discoveries of chromatin conformation such as the existence of chromosome territories have been found, e.g. particular chromosomes tend to occupy individual territories within a nucleus with minimal overlapping (Cremer and Cremer 2001).

A great advantage of microscopy-based methods is that they are very powerful: the proximity of DNA segments could be studied with a high level of spatial resolution. However, there are some limitations. For example, for FISH, the analysis is limited to a few loci of interest and does not provide a clear understanding of a genome-wide architecture (Fraser et al. 2015; Bonev and Cavalli 2016). Another recent microscopy-based method called Hi-M overcame this limitation: it allowed to analyse 3D chromatin organisation simultaneously with RNA expression at single nuclei level (Cardozo Gizzi et al. 2019). In addition, high-resolution analysis of both spatial organisation and transcription could become a power source for understanding the connection between the chromatin topology and transcription.

An alternative way to detect chromatin interactions, and study three-dimensional chromatin folding, is to appeal to chromosome conformation capture techniques. Basic 3C experiment quantifies the interactions happening between a single pair of DNA

fragments within the population of cells, so, can be simply characterised as an experiment aimed to detect “one-vs-one” interactions (Dekker et al. 2002). 4C allows to detect the pairwise contacts of single DNA locus with all other loci, reflecting “one-vs-all” manner of interactions (Simonis et al. 2006). 5C detects all pair-wise interactions happening within a given genomic regions, in other words, “many-vs-many” interactions (Dostie et al. 2006). Hi-C (a high-throughput derivation of 3C) quantifies “all-vs-all” pairwise interactions - any two fragments within the studied population of cells that are allocated close to each other in space are detected and counted (Lieberman-Aiden et al. 2009). All 3C-based experimental approaches rely on the same starting steps: cells of interest are exposed to formaldehyde and then DNA fragments in close spatial proximity are cross-linked. Apart from the downstream differences in protocols, the basic idea is common - if the pair of loci cross-linked more often than by chance, it can possibly sustain some particular chromatin architecture that participate in proper cellular functioning.

So, in contrast with FISH, recently developed 3C-based techniques, including Hi-C, allow collecting chromatin contact frequencies at regional, whole chromosome and genome-wide levels. But at the same time, contact frequency in 3C is produced in a bulk manner and, as a consequence, represents the overlap of individual chromatin architecture profiles in a population of cells (Giorgetti and Heard 2016; Fudenberg and Imakaev 2017). Chromatin architecture was found to undergo constant dynamic changes though cell development, so the studied cell population is required to go through similar phasing to display stable organisation with 3C and Hi-C. As an example, in (Gibcus et al. 2017) transition from interphase to late prophase was accompanied by recognisable loss of spatial patterns, as well as in (Nagano et al. 2018) during replication DNA was found to be accompanied with loss of insulation between spatially segregated chromatin domains. As DNA was found to be the highly dynamic structure, 3C and, in particular, Hi-C techniques are limited to represent chromatin organisation of each single cell within the population (Nagano et al. 2013). In addition, despite being powerful source of information of chromatin conformation, 3C-based techniques were

not powerful enough to be conducted in rare cell types due to a lack of biological material: downstream sequencing of interacting fragments required a large amount of DNA. For example, while the zygotic genome activation in flies was using Hi-C, the oocyte-to-zygotes transition in mice required the development of a single-cell Hi-C protocol (Hug et al. 2017; Flyamer et al. 2017; Ing-Simmons et al. 2021).

As expected, a comparison of single-cell and population average-styled Hi-C, as well as other bulk-style experiments capturing genome-wide chromatin architecture like GAM (genome architecture mapping) (Beagrie et al 2017) and SPRITE (split-pool recognition of interactions by tag extension) (Quinodoz et al 2018), demonstrated that chromatin architectural patterns do not always coincide and some frequent single-cell contacts occur across bulk TAD borders (Flyamer et al. 2017; Nagano et al. 2013). Along with the DNA dynamic as a possible source for inconsistency, single cell experiments are characterised by low coverage, which is also expected to lead to large sparcity of contacts.

1.2.2. Understanding the sources of technical biases in Hi-C is important for further processing steps

DNA and proteins are compactly packed inside the nucleus, thereby creating the conditions for intensive DNA-DNA, DNA-protein and protein-protein interactions. Hi-C is a powerful technology that is widely used to address currently challenging questions in the field of chromatin organisation and gene regulation. However, as an experiment-based method Hi-C is sensitive to technical biases that provide imprecise information on chromatin interactions (Yaffe and Tanay 2011; Imakaev et al. 2012; Hu et al. 2012; Servant et al. 2015). The initial Hi-C pipeline was modified in order to increase the contact map resolution, to lessen the effect of biases or to make it simpler and cheaper. Therefore, we focus on biases associated with the general stages of the experiment: cross-linking, cutting with enzyme, ligation, purification and sequencing.

Most 3C-based techniques begin the same way: cells are cross-linked with formaldehyde that creates covalent bonds with macromolecules such as proteins and DNA,

"freezing" the protein-DNA, protein-protein and DNA-DNA interactions. Cross-linking is followed by a fragmentation stage when a restriction enzyme recognise a specific sequence to bind and cut the whole chromatin into short fragments. Original protocols involve HindIII and NcoI restriction enzymes that spot sequences of six nucleobases AAGCTT and CCATGG, respectively (Lieberman-Aiden et al. 2009; Belton et al. 2012). The recently used DpnII (Rao et al. 2014) searches for only four bases sequence GATC, thereby slicing DNA into shorter pieces: the average fragment length for the human genome decreases from 4 Kb to less than 500 bp by switching the HindIII to DpnII enzyme (Belaghzal et al. 2017).

At the further stages, Hi-C is slightly different from other 3C methods. In Hi-C, restriction fragment ends are labeled with biotinylated nucleotide and ligated, creating covalent bonds between proximal ends. Ligated biotin-containing fragments are pulled down, isolating them from non-ligated fragments. At the last step, the generated molecules are sequenced. As each DNA-DNA interaction event is expected to be followed by the ligation event, the generated library of ligated fragments constitutes the collection of pairwise DNA contacts.

Hi-C is a complex and multi-stage technique involving variety of components that need to be understood as sources for technical errors that can reside at all stages (Figure 1.1.A). The ability to distinguish the proper DNA interaction behavior from the spurious one provides a basis for data trimming tools that allow the production of trustful Hi-C contact maps (Yaffe and Tanay 2011; Hu et al. 2012).

Formaldehyde cross-linking. The chromatin architecture is regulated and/or exploited by different protein complexes. Examples include transcription factors that initiate and control transcription or histones and architectural proteins that mediate condensation or relaxation of chromatin fiber (Van Bortle et al. 2014; Stadler et al. 2017; Nora et al. 2017; Nora et al. 2020). Therefore, proteins act as a mediator in DNA-DNA interactions involved in cellular processes. During cross-linking, formaldehyde is reacted with all macromolecules inside the nucleus creating "bridges" between proximal ones (Hoffman et al. 2015). In case of indirect DNA-DNA interaction, the chromatin is

linked through the protein-mediator.

Recent studies showed that formaldehyde cross-linking favors reactions with lysine - the structural unit that makes up proteins (Lu et al. 2010). It results in the covalent bonds that are formed between lysine side chain of a protein and proximal DNA fragment first, leading to the relatively lower cross-linking power in protein-protein interactions (Zeng et al. 2006). Thus, the contacts between DNA fragments via bridges made by the complex of proteins are either rare events or lost due to inefficient cross-linking (Gavrilov et al. 2015). On the other hand, the high density of macromolecules in the presence of formaldehyde can result in the formation of the complex networks between the nuclear elements that do not interact (Gavrilov et al. 2015). The indirect protein binding that brings DNA fragments into spatial proximity may be either over-represented due to artificial formaldehyde cross-linking of proximal non-interacting complexes or under-represented due to low cross-linking power for indirect protein binding.

Restriction enzyme. The Hi-C method detects only pairwise contacts between DNA fragments in each cell from the population, it does not detect higher order structures with three and more loci to be involved simultaneously. Cutting sequences recognised by the selected restriction enzyme affects the fragment size. Longer sequences generate on average longer fragments that have a higher chance to be involved in interactions of more than two fragments.

Biases arising because of inefficient detection of higher order chromatin structures could not be eliminated using standard Hi-C data processing and require improvements of the protocol. Switching to restriction enzymes with shorter binding sequence improves the resolution and decreases the chance to catch complex interactions. An alternative solution is to use methods to study the genome architecture that are not limited by pairwise interactions like GAM (Beagrie et al. 2017), SPRITE (Quinodoz et al. 2018) and Pore-C (Ulahannan et al. 2019).

Despite the fact that interaction of three and more fragments are not detected in general Hi-C, using the methods of polymer physics and pairwise contacts collected

from population of cells we can predict the position of fragments forming simultaneous multi-interaction complexes (Liu et al. 2021). We have to keep in mind that as chromatin is a dynamic structure, we cannot distinguish between such interactions being credible for single cell or being spurious because of the overlap of a large number of cells. However, based on the Hi-C contact map we can select several loci that we suspect are in multi-fragment interactions and check them with microscopy-based techniques.

Ligation. In theory, the number of ligated fragments appearing in a Hi-C library corresponds to the number of interaction events between chromatin fragments in a cell population. However, regions that are not previously cross-linked could also be ligated. These ligation artefacts can include self-ligations when two ends of the same fragment are ligated to each other, or random ligations when ligated fragments belong to different cross-linked pairs (Mifsud et al. 2017). Only intra-molecular ligation when two cross-linked fragments are ligated represents the real chromatin interaction. Also, the length of the restricted fragments itself was found to have different ligation efficiency (Yeffe and Tanay 2011). Several recently developed experiments are ligation-free, so the data produced in accordance with protocols is unaffected by ligation biases. Such tools include GAM (Beagrie et al 2017) and SPRITE (Quinodoz et al. 2018).

GC-content. The restricted fragments are defined as blunt-ended or sticky-ended (Pray 2008). Sticky-ends are characterised by the DNA single-strand overhang while blunt-ends have no unpaired nucleotides. The length and the nucleobase composition of sticky-end overhang may vary depending on the choice of the restriction enzyme, thereby affecting the ligation efficiency. It was previously shown that the sticky-ends that are highly enriched with GC are associated more often, so have higher chance to be ligated (Gao et al. 2015).

PCR amplification. The amount of DNA is required for high-throughput sequencing exceeds the amount of DNA obtained from a cell population. The amplification by PCR allows to produce duplicates of ligated fragments to complete the outstanding amount of DNA required for efficient downstream sequencing. However, introduced duplicates

may skew the interaction profiles, especially in the case of rare cell types (Yeffe and Tanay 2011). Recently developed methods as SAFE Hi-C (Niu et al. 2019) avoid the PCR amplification, so avoid the possibility of skewness arising from the duplicates.

Sequencing and alignment. Ligated fragments are sequenced and referred to as reads - short linear strings of nucleotide bases. The unique sequence of base pairs of a read correspond to its position within the DNA molecule in the reference genome. Alongside the development of next-generation sequencing techniques, which allow to process massive sequence input with low costs, it was shown that depending on methods and platforms used we can face some sequencing errors (Fox et al. 2014). Error rates were found to be relatively low, however, differences between reads and reference genome can produce incorrect alignments. Moreover, even if the fragment is correctly sequenced, it can be removed from the downstream analysis if it cannot be uniquely mapped. Thus, short DNA loci combined from just several nucleotides can be placed in several positions within the genome, so they have a higher chance to be incorrectly aligned. We can avoid this situation by restricting the length of the mapping reads.

Another source of incorrect alignment is transposable elements. Transposable elements (TEs) are the genome fragments that are able to change their position within the genome (McClintock 1940). These regions either can be copied and then integrated elsewhere in the genome in the case of retrotransposons (Boeke et al. 1985), or can move from one location to another in the case of DNA transposons (Greenblatt et al. 1963). In mammals, transposable elements makes up half of the genome (Platt et al. 2018), so the sequenced reads coming from TEs face a danger to be non-uniquely allocated and require specific parameters and/or algorithm to be considered. Interestingly, several recent studies were focused on the functional role of TEs in chromatin architecture (Schmidt et al. 2012; Zhang et al. 2019; Diehl et al. 2020). Thus, architectural protein CTCF, which was suggested to participate in loop extrusion and maintain specific DNA topology in mammalian systems (Racko et al. 2018), obtained many of its binding sites and associated novel chromatin loops from transposons throughout

evolution (Bourque et al. 2008; Schmidt et al., 2012).

1.2.3. Correction and normalisation algorithms reduce read-level biases and possible interaction skewness

The presence of the technical biases makes the extraction of biologically meaningful information from Hi-C data a challenging task. Ligation artefacts and PCR duplicates increase the Hi-C library while they are not coming from the real interactions between DNA fragments. In contrast, complex architectural structures or low mapping quality of reads reduce the set of Hi-C interactions for downstream analysis. Following the mapping to the genome, noise and expected imbalances in Hi-C data require read-level filtering and matrix balancing to be considered. However, all tools and computational methods have their own drawbacks and limitations such as bias trimming efficiency, computational complexity, memory or time consumption. Also, we expect that we did not cover all possible technical errors arising during the experiment as they could vary depending on the protocol modifications or external conditions. However, the correction step is able to correct the general data skewness arising from the combinations of systematic errors.

Balancing algorithms can rely on explicit sources of technical biases: reads are characterised by their length, GC-content and mappability to the genome, which are known and quantified, so can be used as an explanatory factor in statistical modeling. The probabilistic approach proposed by (Yaffe and Tanay 2011) and HiCNorm (Hu et al. 2012) considered three listed sources of systematic biases to model the contact frequency through standard density functions as Bernoulli, Poisson or Negative Binomial. Although the significant technical biases are considered, if we want to test other biases, we face general problems of statistical modeling including the need of quantitative measure for the aimed error source or multicollinearity problem.

In contrast, implicit approaches do not distinguish bias sources and assume that each bin simply accumulates the effect of several biases coming from the experiment conditions, sequencing and mapping. Most widely used methods here are SCN (se-

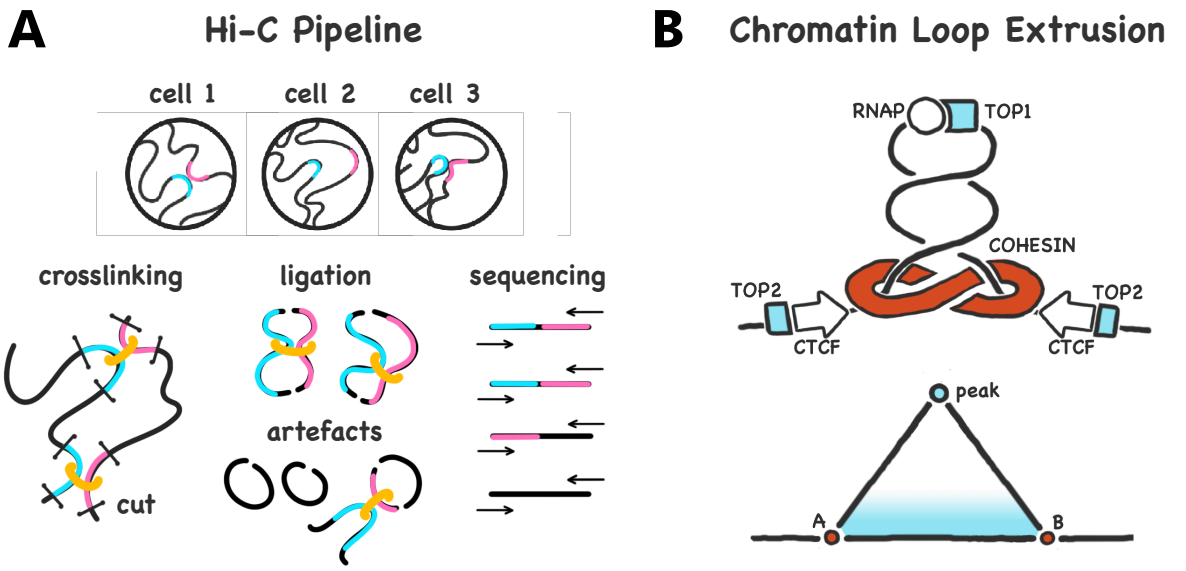


Figure 1.1. Schematic representation of Hi-C pipeline and TAD formation from loop extrusion. **A.** Main steps of basic Hi-C pipeline: pairwise interactions between DNA fragments in population of cells are cross-linked, cut with restriction enzyme, ligated and sequenced. Ligation can generate artefacts including self-ligation event and random ligations which do not represent real loci contacts. The scheme is based on visualisation of Hi-C experiment in (Lieberman-Aiden et al. 2009). **B.** Chromatin loop extrusion model: During transcription, RNA polymerase (RNAP) generates torsional stress, which pushes the supercoiled chromatin through Cohesin rings until Cohesin and CTCF meet. As divergently oriented CTCF does not allow further Cohesin sliding, the anchors of the loop are expected to be in close contact which is visualised on Hi-C interaction map as TAD with peak point. Type 1 (Top 1) and Type 2 (Top 2) topoisomerases bind DNA to drain supercoiling. The scheme is based on visualisation of the model in (Racko et al. 2018).

quential component normalisation) (Cournac et al. 2012), ICE (iterative correction and eigenvector decomposition) (Imakaev et al. 2012), KR (Knight-Ruiz) balancing (Knight and Ruiz 2013) and chromoR (Shavit et al. 2014). Hi-C processing and visualisation tools like Fit-Hi-C (Ay et al. 2014), HiCExplorer (Ramirez et al. 2018) or Fan-C (Kruze et al. 2020) rely on most widely used ICE and KR balancing methods either as a correction step in Hi-C processing or as a source of bias estimation for Hi-C interaction modelling. In particular, matrix-balancing approaches became favoured due to simplicity and being parameter-free. So, they rely on the assumption that all DNA fragments have equal chances to be detected, however the previously shown significance of GC-content or restriction length did not support this assumption (Yeffe and Tanay 2011). Despite this, algorithms that do not require the exact knowledge of the errors origin are more robust against developments in Hi-C and against some variants, such as capture Hi-C (Mifsud et al. 2015). In addition, some recent approaches like GOTHiC (Mifsut et al. 2017) takes intermediate position between explicit and implicit approaches: systematic biased are split into two parts where one part is biases coming from known sources and second part is biases of unknown origin.

1.3. Hi-C revealed the hierarchical folding of chromatin

1.3.1. Chromatin is spatially segregated into compartments depending on transcriptional state

The development of the Hi-C methods allowed a closer look at the genome-wide DNA organisation. The first implementations of Hi-C have been constructed at megabase scale (Lieberman-Aiden et al. 2009) and, despite the resolution that can be considered relatively low today, Hi-C demonstrated results that were consistent with earlier investigations of chromosome territories and other known patterns of inside nuclear positioning in mammals with 3C and FISH (Cremer and Cremer 2001). Thus, distal loci belonging to the same chromosome tended to be near to each other in space as well as specific chromosome pairs showed certain preferences to be close to each other.

Further in-depth analysis provided insights on the next level of hierarchy in chro-

matin folding. Each chromosome inside the nucleus was found to be spatially divided into two minimally intermingling parts with rare cases of DNA contacts between them (Lieberman-Aiden et al. 2009; Rao et al. 2014). These parts named as A and B compartments showed different packing behaviour: A compartment was packed less densely, which marked more accessible DNA for binding of functional elements and, as a consequence, was associated with the active part of chromatin; B compartment was less accessible and marked as an inactive part of the chromatin. High resolution maps recently achieved in lymphoblastoid cells showed the enrichment of active enhancers and promoters: distal enhancers formed the islands of A compartment chromatin surrounded by inactive B compartment chromatin; A compartment was also enriched with gene TSSs (transcription start sites) while the gene body including TTSs (transcription termination sites) belonged mostly to B compartment (Gu et al. 2021). A/B compartments were also found not only in mammalian systems but in other organisms, for example, *Drosophila melanogaster* fly (Rowley et al. 2017) or even in prokaryotic cells in *Sulfolobus* archaea (Takemata and Bell 2020).

The transcriptional status of chromatin, and as a consequence, the affiliation to A or B compartments could differ between cell types as well as during developmental stages (Lieberman-Aiden et al. 2009; Dixon et al. 2015). Despite being tissue-specific and dynamic, A/B compartments showed notable robustness to experimental interference aimed to affect the chromatin architecture, like the depletion of architectural proteins (Kaushal et al. 2021). Besides, A/B compartments were found to be mostly consistent between tumor and normal cells (Johnstone et al. 2020; Adeel et al. 2021). Still, a closer look at the compartmentalisation strength in (Johnstone et al. 2020) revealed the noncanonical regions characterised by intense self-interactions and contacts with both compartments. These regions proposed to reflect the intermediate compartment state and show distinct features between tumor and normal cells.

1.3.2. Topologically associated domains are the next level of chromatin organisation

Later improvements of the Hi-C methods allowed higher levels of resolution and have shed light on the next unit of organisation called topologically associated domains or TADs (Dixon et al. 2012; Nora et al. 2012; Sexton et al. 2012). TADs are characterised by the intense interactions between DNA fragments belonging to the same TAD and less probable contacts between fragments allocated in different TADs. Visually, on Hi-C maps TADs are represented in a form of squares which are continuously spread along the diagonal. TADs were found to be widely conserved across species (Vietri Rudam et al. 2015) as well as during different developmental stages (Ghavi-Helm et al. 2014; Dixon et al 2015), suggesting that they maintain proper cellular functioning and gene regulation. At the same time, some reports showed that TADs can demonstrate cell line-specific allocation within the same organism. For example, in (Chathoth and Zabet 2019) only some of the TAD borders were conserved between embryonic (Kc167) and neuronal (BG3) cell types in *Drosophila melanogaster*. Note, that conserved TAD borders reflected similar binding patterns of architectural proteins. Summing up, topologically associated domains can be defined as a fundamental unit of spatial chromatin organisation.

1.3.3. TAD reorganisation has an ambiguous effect on gene expression

TAD borders have been shown to be enriched in housekeeping genes (Li et al. 2015), developmental enhancers (Cubenas-Potts et al. 2017) and boundaries of highly conserved genomic regulatory blocks (Harmston et al. 2017), which suggested strong association between 3D organisation of DNA and gene regulation. A strong connection between TADs and elements of transcription machinery including enhancer, promoters and various transcription factors suggested that the functional role of TADs is to bring distal regulatory elements closer in space to maintain a physical contact (Rocha et al. 2015). At the same time, a high level of insulation between neighbouring TADs could separate DNA elements from undesired contact and prevent non-target

genes from uncontrolled over-expression. A previously reported presence of architectural proteins and insulators at TAD borders (Van Bortle et at. 2014; Stadler et al. 2017) also supported the insulating function of TADs. Thus, the stability of TAD boundaries is proposed to be essential for establishment of proper cellular development and functioning, hence defective chromatin architecture can bring a possibility of diseases, developmental defects, disorders and/or cancer (Lupianez et al. 2015; Taberlay et al. 2016; Kragesteen et al. 2018). Structural variations in chromatin such as deletions, duplications, inversions and translocations can affect the chromatin topology including the number and sizes of existing domains. For example, in (Lupianez et al. 2015) re-engineered structural variations in mouse near *Epha4* gene, which were considered as a major cause for human limb malformation, led to domain disruptions, *Epha4* misexpression and subsequent developmental defects in limbs.

Any particular domain could unite with neighbouring ones (TAD aggregation), be divided into several small domains (TAD disruption), as well as exhibit complex recompositions. When TADs are aggregated, loci that were outside the domain have a higher chance to interact with loci that were inside the domain. In that way, the loss of insulation drives an enhancer and a non-target promoter to interact, which results in the activation of genes that should be silenced in normal cells. Such defected gene expression during embryonic development could cause irreversible developmental disorders. In (Helmbacher et al. 2000) and (Kragesteen et al. 2018), re-engineered domain disruptions in mice resulted in rare limb malformation followed by the contact between previously insulated DNA fragments and subsequent gene misregulation. In (Nora et al. 2017), TAD border loss due to depletion of protein-insulator CTCF led to clear upregulation of genes which also demonstrated a tendency of colocalisation within the same aggregated TAD. Furthermore, the evidence of aberant gene activation correlated with topology reorganisation was found in cancer cells, for example, in glioma (Flavahan et al. 2015; Flavahan et al. 2016).

Illustrative results were obtained by (Owens et al. 2021) studying the transcription factor RUNX1 which downregulation can cause leukemia. A megabase scale TAD,

where *Runx1* producing gene resided, was formed prior to transcription, supporting the idea of topology-driven gene regulation. However, during differentiation it was found to be separated into two smaller sub-structures which were induced by active transcription of two promoters controlling the gene production. Interestingly, the separation between sub-structures were consistent with enrichment of protein-insulator CTCF, which induced depletion affected the chromatin conformation and declined the production of RUNX1, while the enhancer-promoter contacts were mainly unaffected. In contrast, induction if gene misregulation upon hypoxic treatment in (Nakayama et al. 2021) was accompanied by radial position alterations as well as relative reallocation to each other, however neither strength nor direction of the gene activity violation had a clear relationship with spatial reposition.

However, in some recent reports it was shown that changes in TADs did not always correlate with the changes in gene expression. Thus, in (Ghavi-Helm et al. 2019) rearrangements introduced in *Drosophila* balancer chromosomes were not coupled with the transcriptional changes. Still, only a few TADs were affected possibly due to the small number of rearrangements induced, so it is the high chance that when more TAD borders are lost, shifted or gained would allow the observation the changes in gene expression. The analysis conducted in (Ing-Simmons et al. 2021) on *Drosophila* embryos entirely consisting of ectoderm, neuroectoderm and mesoderm, which were obtained in the result of maternal mutations, also did not show significant differences in TAD organisation in the presence of transcription alterations. However, required sorting of the embryos which resulted in low input Hi-C in addition to strict selection procedure allowing most trustful TAD borders for downstream analysis made viable only strong and massive reorganisations to be detected.

Although we have enough evidence that the 3D organisation of the DNA and gene regulation are related, we still have no clear vision about what exactly drives in this relationship: the chromatin architecture defines the proper gene expression or, conversely, the contacts between regulatory elements establishes the 3D chromatin folding.

1.3.4. TAD formation and maintenance are proposed to be loop extrusion mediated

The way in which domains are formed is still unclear. The correlation between TAD boundaries and anchors of chromatin loops was proposed to shed light on the TAD establishment and maintenance (Rao et al. 2014). DNA loops have been suggested to either mediate contacts between enhancers enriched with transcription factors and target genes to regulate transcription (Pennacchio et al. 2013), or to maintain contacts between insulator elements and isolate chromatin domains to avoid undesired contacts (Phillips and Corces 2009). Recent studies suggested that transcription-based supercoiling could be used as a driving force in a TAD formation mechanism (Bjorkegren and Baranello 2018; Racko et al. 2018; Ruskova and Racko 2021). In general words, the process could be described as follows (Figure 1.1.B). It starts when transcription factors bind enhancer or promoter region and recruit RNA polymerase to transcription start site. At the initiation step, RNA polymerase relaxes double helix strands and initiates synthesis of RNA. Moving along DNA, it continues to unwind DNA strands and elongate RNA chain, then, after polymerase passing, DNA is rewinded into a double helix again. Polymerase does not rotate itself, so when it relaxes DNA helix, it increases torsional stress in chromatin at both sides of the molecule which is drained by Type 1 (Top 1) and Type 2 (Top 2) topoisomerases to allow transcriptional elongation (Pommier et al. 2016). Thus, blocking of topoisomerase activity led to transcription inhibition and chromatin compaction (Neguenbor et al. 2021). At the same time, prolonged stabilisation of Top 1-DNA complexes could generate DNA strand breaks causing cell cycle arrest and cell death - this mechanism was proposed as a treatment of cancer (Gilber et al. 2012).

Torsional stress draining could be incomplete, which facilitates and maintains chromatin supercoiling (Bjorkegren and Baranello 2018). Recent works of (Camela et al. 2019; Gothe et al. 2019) supperted the hypothesis, as loop anchors and TAD boundaries are found to be enriched with topoisomerase. The formation and stabilisation of domains from stressed chromatin is supported by architectural proteins Cohesin and

CTCF. Cohesin is presented in a form of two joint rings that could attach to chromatin - Cohesin binds to part of the DNA that is under torsional stress, supercoiling pushes chromatin through Cohesin rings and forms a loop. Interestingly, transcription inhibition followed by RNA polymerase depletion resulted in DNA compaction with a decrease in colocolisation of Cohesin supporting the idea that Cohesin binding and TAD segregation can be transcription-driven (Neguembor et al. 2021).

The loop growth continues until when Cohesin reaches CTCF that do not allow Cohesin rings to pass through them. In (Nora et al. 2021), depletion of N-tirminus of CTCF affected the TAD boundaries insulation and diminished Cohesin accumulation: chromatin insulated with modified CTCF was proposed to be either completely released of a loop or less insulated due to the incapacity of CTCF to stop loop extrusion. In (Flavahan et al. 2016), hypermethylation found in human glioma blocked the binding of methylation-sensitive CTCF and Cohesin which, as a consequences, reduced the domain insulation and facilitated undesired activation of the glioma oncogene. A similar model was suggested in *C. elegance* with Condensin-mediated loop extrusion (Rowley et al. 2020; Jaminez et al. 2021).

TADs that are formed in accordance with the the loop extrusion models are expected to demonstrate the intense contact between anchor fragments (Racko et al. 2018). However, in the human genome, less than 39% of TADs were accompanied by loop anchor contacts (Rao et al. 2014). Moreover, the depletion of CTCF has as an unpredicted effect in different organisms, cell types or developmental stages. Looking particularly on *Drosophila* associated studies, the absence of CTCF following knockout of CTCF-producing gene at zygotic state can be compensated by protein and mRNA coming from the maternal germline that ensures survival (Gambetta et al. 2021). However, the depletion of maternal material led to death at pupal stage (Kaushal et al. 2021). The same study showed that rescued CTCF transcription avoided lethality and allowed further developmental transition but only in neuronal cells and not in muscle cells. Similar knockout-based research in mammalian systems to study the interplay between TAD reorganisations and gene regulation were difficult to perform due to mas-

sive perturbations and low chances of cells to survive. In general, long-term loss of CTCF led to massive cell death (Yao et al. 2017; Hyle et al 2019; Xu et al. 2021) limiting the time when cells could be treated and studied.

Intriguingly, zebrafish (Franke et al. 2021) was able to survive and progress developmentally through knockout due to fast development similar to flies and it was shown that both transcription and chromatin segregation were enormously affected. Also, recently published the whole transcriptome analysis revealed significant amount of differentially expressed genes in CTCF-deplted cells (Hyle et al 2019). *Drosophila* CTCF knockout also resulted in massive gene trancription changes in neuronal cells (Kaushal el al. 2021). In contrast, partial knockdown of CTCF protein did not reveal a clear effect on transcription in (Bartkuhn et al 2009; Bortle et al. 2012; Schwartz et al. 2012). The low sensitivity of chromatin topology in flies can be explained by lack of CTCF binding at TAD borders - no more than 10% of borders were found to colocolise with CTCF peak (Kaushal et al. 2021). So, some TADs are expected to be associated with formation factors which are different from the transcriptional supercoiling and CTCF binding. Thus, TAD boundaries in *Drosophila melanogaster* genome were found to be enriched with other insulator proteins including BEAF-32, Cp190 and Chro (Van Bortle et al. 2014; Ramirez et al. 2018; Wang et al. 2018; Chathoth and Zabet 2019) reflecting the possible existence of other maintenance mechanisms.

1.4. TAD calling is sensitive to assumptions and model selections

1.4.1. TAD border finders are sensitive to predefined assumptions

Several TAD calling algorithms have been developed allowing extraction of TAD border positions from Hi-C generated data. However, it is still possible that re-constructed domains do not reflect a real chromatin interaction profile. Border allocation, which represents the chromatin packing that does not reflect visually clear TAD geometry on Hi-C interaction matrix or does not have strong connection with any expected epigenetic marks, could result in systematic biases affecting Hi-C data as well as several technical reasons. They include the unrealistic and over-simplified assumptions, sta-

tistical model choice that poorly reflects the observed data or parameter adjustment.

1.4.2. Insulation-based algorithms produce TADs in consecutive manner

The area between two neighbouring TADs is characterised by significant depletion of interactions in comparison with the highly contact enriched areas inside the neighboring TADs. So, under this assumption we technically deal with the goal of detecting the genomic region that demonstrates the insulation of neighbouring downstream and upstream regions from each other.

In (Dixon et al. 2012), a term “topological domain” was initially introduced: topological domains were detected based on a **directionality index** (DI) approach. According to it, for the particular Hi-C bin we measure the number of downstream and upstream pairwise contacts with several adjacent fragments. When the bin demonstrates biases towards upstream interactions (negative DI), this bin is most probably allocated at the end of the TAD. And vice versa, when the bin is biased towards upstream interactions (positive DI), it probably belongs to the start of the TAD. (Dixon et al. 2012) used a Hidden Markov Model (HMM) to infer the statistically significance of short genomic region to demonstrate the DI local minimum right before and the DI local maximum immediately after it. When the region is short it was classified as domain boundary region and represented the region between end of one domain and start of the next one. When the region is relatively long, it was classified as unorganised chromatin. The DI approach was used then in (Pope et al. 2014; Dileep et al. 2015) to study the TAD organisation in mouse and human cells.

Another set of approaches that rely on computation of **insulation score** overcomes the required parameter tuning and computational complexity of DI as well as other statistical based methods. Insulation score (IS) at the particular Hi-C bin quantifies the average interaction intensity between pairs allocated in the proximity to this bin (Crane et al. 2015). When a bin is between two neighbouring TADs, this bin is expected to show the IS drop as the interaction frequency should be significantly lower comparing to inside the TAD region. The TAD finding procedure then counts on the IS local

minimum detection. The IS calling method was adapted and implemented as a part of TopDom (Shin et al. 2015), HiCExplorer (Ramirez et al. 2015; Ramirez et al. 2018), Fan-C (Kruze et al. 2020) tools.

1.4.3. Recent investigations on hierarchical TAD folding require alternative techniques to be considered

Hierarchical packing of chromatin into compartments and then into megabase scale TADs suggested the subsequent partitioning of substructures at sub-megabase scale which plays a critical role in the establishment of cellular functioning (Yu et al. 2017; Norton et al. 2018). Based on the higher resolution 5C data, (Phillips-Cremins et al. 2013) demonstrated the presence of sub-topologies within TADs which were previously annotated in Hi-C. Sub-TADs were hypothesised to facilitate cell and tissue specific enhancer-promoter contacts while large-scale TADs establish distal contacts which are more developmentally conserved. Also, low resolution Hi-C in mouse and human found TADs tended to organise TAD-TAD interactions forming large-scale metaTADs that were characterised by enhanced enrichment of epigenetic features as CTCF, RNA polymerase 2, promoter marks, TFs previously found to be important in TAD border insulation and maintainance (Faser et al. 2015).

Visual inspection of TADs called in Hi-C using consecutive square blocks representation also suggested the possibility of alternative domain allocations (Filippova et al. 2014). It shows the need for calling algorithms to capture not only the visual differences, but the hierarchical folding of TADs as well. Arrowhead algorithm (Rao et al. 2014) that was further implemented as a part of Juicer tool (Durand et al. 2016) is one of the first algorithms that produced the nested TAD organisation based on the detection of corner point of domain and sub-domain squares. TADtree (Weinreb and Raphael 2016) is another well-known tool which allocates *TAD forest*: the collection of TADs with further segmentation on multilevel nested sub-TADs depending on a unique functional relationship of contact enrichment versus the distance between interacting pairs. Despite being able to catch nested TAD organisation, Arrowhead and TADtree

demonstrated dramatic difference in number of TADs called with the same datasets: TADtree produced significantly more TADs for any number of reads retained by the filtering (Forcato et al. 2017). TADtree also was found to be more robust to sequencing depth and resolution, however its computational complexity makes the processing extremely slow and memory heavy (Dali and Blanchette 2017).

1.5. Open research questions in TAD calling provided the basis for the thesis

The high usability made the insulation score based approaches the most popular in the field of TAD research. However, the common part of the various IS-based methods is the consecutive manner of produced TADs. Local minima of the insulation score indicate the genomic positions where the insulation between upstream and downstream interactions is significant and it is the most probable position of a TAD border. According to this definition, the end position of one TAD coincides with the start position of the next one. Some SI-based approaches allow the detection of TAD boundaries instead of a single genomic position, so the neighboring TADs can be separated by short genomic region of several bins. Although, first insights on chromatin segregation into topological domains in mouse ESCs published by (Dixon et al. 2012) based on the DI calling algorithm revealed that approximately 90% of DNA were occupied by TADs, so remaining 10% were split between TAD boundaries and unorganised chromatin. Even the share of outside TAD chromatin is relatively low, the “head-to-tail” TAD allocation assumed in IS-based methods makes the detection of unorganised chromatin questionable.

Recent studies focused on the relationship between spatial chromatin organisation and gene regulation did not provide clear vision on TAD formation mechanics. Observed TAD reorganisations between different experimental conditions, cell types and organisms can be caused by systematic errors like spurious DNA-DNA contact subject to ligation artefacts or loss of significant interactions in case of low mappability of interacting fragments. However, the latest improvements in protocols and developments of alternative genome-wide studies allowed to trace TAD reestablishments with high level of confidence, as well as to look on inner TAD hierarchy. The presence of nested sub-

TADs can be a key factor in understanding and predicting the effect of the short-range perturbations on gene expression and vise versa.

A high level of inconsistency and scarcity of TAD finders recognising nested folding and unorganised territories between TADs makes the validation of complex multilevel chromatin topology a nontrivial task. Moreover, many existing tools are computationally complicated and time/memory consuming to be widely spread to conduct research on various organisms, cells types and experimental conditions. Desired algorithms should ideally be user friendly with simple and intuitive parameter selection, also fast and do not take a lot of memory.

Nested TAD organisation and presence of TAD-free chromatin can be treated as part of the assumptions allowing any kind of domain square positioning within the Hi-C map: squares enclosed one inside another (nested TADs), breaks between squares (TAD-free regions), as well as partial overlap between neighbouring squares. Partially overlapping TADs are geometrically feasible, however we did not face any confirmation about their presence or functional role - they can appear as an artefacts of bulk manner of the Hi-C experiment and as chromatin being dynamic structure, but also can detect the complex architecture representing partial interference between neighbouring chromatin domains.

In this thesis, we aimed to gain deeper understanding on TAD architecture and the validity of its hierarchical folding to regulate and sustain proper cellular functioning.

In Chapter 2, we explore the changes in chromatin architecture on the basis of *Drosophila melanogaster* BG3 neuronal cell associated with the depletion of proteins BEAF-32, Cp190, Chro and Dref, which were previously reported as functionally relevant in flies. We explore changes on TAD level using widely used tools in the field of Hi-C analysis, in particular, we use insulation score-based TAD finder HiCExplorer (Ramirez et al. 2015). We analysed the differences in TAD allocation called in canonical consecutive manner and revealed the connection between chromatin state and ability of domain boundaries to maintain insulation in absence of studied proteins.

A closer look at various genomic regions showing DNA architectural reorganisations

revealed the limited power of HiCExplorer to detect single-sided or imbalanced insulation scenarios. Visual inspection demonstrated the presence of TADs that are allocated in a non-consecutive manner including a sequence of start borders placed one after another without end borders between them (possibly nested or partially overlapping TADs) or interacting fragments that cannot be faithfully placed in a single TAD (possibly partial overlap of break between neighbouring TADs). Some of the TAD boundaries also demonstrated different insulation strength where interactions within the upstream TAD are significantly different from interactions within the downstream TAD. To address the functional diversity of observed insulation patterns, in Chapter 3, a COrTADo approach is introduced. COrTADo (**C**omplex **Or**ganisation of **T**opologically **A**ssociated **Do**mains) is a non-parametric TAD calling algorithm, which allow detection of start and end chromatin domain positions separately as a basis for complex TAD reorganisation reconstruction. As starts and ends are separated, based on COrTADo we can asses the downstream and upstream insulation strength of candidate genomic positions to analyse the functional difference between balanced and imbalanced TAD boundary insulation. In Chapter 4, we analysed the epigenetic differences of TADs in BG3 wild-type cells observed with HiCExplorer and COrTADo.

Chapter 2. 3D chromatin organisation of flies

2.1. Background and motivation

Accumulation of insulator proteins and other epigenetic signatures at topologically associated domain (TAD) borders are proposed to be related to TAD formation and maintenance mechanism. In mammalian systems, the interplay between CTCF and Cohesin is suggested to promote loop extrusion following TAD formation (Bjorkegren and Baranello 2018; Racko et al. 2018; Mirny et al. 2019). So, CTCF and Cohesin have been reported to accumulate at borders and their depletion affects chromatin architecture and TAD disruption (Nora et al. 2017; Nora et al. 2021). However, in *Drosophila melanogaster* the majority of TADs did not show strong contacts between border regions, suggesting the prevalence of compartment domains rather than domains formed by loop extrusion (Matthews and White 2019; Rowley et al. 2019). In addition, Cohesin did not tend to colocolise with CTCF (Bartkuhn et al. 2009).

In flies, other architectural proteins including BEAF-32, Cp190 and Chro demonstrated strong enrichment at TAD borders (Van Bortle and Corces 2012; Van Bortle et al. 2014; Ramirez et al. 2018; Wang et al. 2018; Chathoth and Zabet 2019). Interestingly, CTCF, Cp190 and, to a greater extent, BEAF-32 were found near the gene transcription start sites suggesting their role in regulation of specific proximal genes (Bushey et al. 2009). In the same study, only a few binding sites of another *Drosophila*-specific insulator protein Su(Hw) (Maeda and Karch 2007) did not show any specific preference towards gene locations which signaled of distinct role of this protein in chromatin organisation.

Despite the reported strong colocalisation of BEAF-32 at TAD borders, surprisingly, the knockdown of BEAF-32 in embryonic Kc167 cells did not dramatically affect 3D chromatin architecture (Ramirez et al. 2018). The maintenance of TAD borders during BEAF-32 depletion can possible be explained by the fact that BEAF-32 and another architectural protein called Dref (Mathelier et al. 2014) display the same binding motif and Dref can potentially replace BEAF-32 at TAD borders when BEAF-32 is lost.

The other two proteins, Cp190 and Chro, cannot bind DNA directly but they can

be recruited by BEAF-32 and can be involved in insulation of some chromatin regions (Cubenas-Potts et al. 2017; Wang et al. 2018). In previous research, up to 90% of TAD borders analyzed in another embryonic cell line S2 demonstrated the presence of BEAF-32 with either Cp190 or Chro (Wang et al. 2018). Cp190 and Chro were shown to be recruited by other proteins as well. Thus, the complex formed by the interaction of Chro and JIL-1 was found during interphase to colocolise at polytene chromosome interband regions, and were suggested to be involved in the maintaining of polytine chromosome structures (Rath et al. 2006). Cp190, Su(Hw) and mod(mdg4) are three components of *gypsy* insulator - a fragment of gypsy retrotransposon which is known to regulate gene expression through blocking the activity of nearby enhancers or repressors (Wei et al. 2001; Capelson and Corces 2004; She et al. 2010). During interphase all three proteins separate into bands (inactive condensed chromatin) and interbands (active open chromatin), which can possibly define chromatin domains (Labrador and Corces 2002). However, Cp190 was present together with Su(Hw) and mod(mdg4) in just a few cases (Pai et al. 2004), otherwise Cp190 was found to be recruited by CTCF (Gerasimova et al. 2007; Kaushal et al. 2021).

In this Chapter, we aim to look closer at chromatin topology based on *Drosophila melanogaster* Hi-C data generated on BG3 (neuronal) cells. Also, to take us closer to understanding the connection between spatial chromatin organisation, insulator proteins binding and gene regulation, we analysed the architectural changes resulted in the depletion of BEAF-32, Cp190, Chro and Dref. Note that we based our hypothesis and conclusions by comparing the distribution of genome-wide interactions, A/B compartmentalisation and, especially, TADs which we called in a canonical *head-to-tail* manner using the insulation score-based tool HiCExplorer (Ramirez et al. 2018).

2.2. Hi-C data validation and processing

2.2.1. Protein knockdowns and their efficiency

We analyzed the effect of BEAF-32 single knockdown and Cp190 and Chro double knockdown in BG3 cells followed by *in situ* Hi-C. As Dref can potentially replace the

BEAF-32 due to a similar binding motif, we also investigated the effect of combinatorial knockdown of BEAF-32 and Dref. We used the Hi-C datasets generated by (Chathoth and Zabet 2019; Chathoth et al. 2022).

During the effective RNAi knockdown, the expression of selected proteins significantly reduce. As a consequence, when the knockdown is efficient we expect to observe the reduced level of messenger RNA (mRNA) - mRNA is a molecule that carries the “instruction” for protein synthesis, so when less mRNA is present the less protein can be produced. All three generated mutants demonstrated the reduction of at least 60% of mRNA of depleted proteins (BEAF-32, Cp190, Chro and Dref) in comparison with wild-type (Chathoth et al. 2022). The efficiency of knockdown achieved here is similar to the ones reported by other studies in *Drosophila* cells (Schwartz et al. 2012; Ramirez et al. 2018; Zenk et al. 2021).

Also note, that the DNA topology is expected to undergo dramatic structural metamorphoses every cell cycle. Whether the knockdown demonstrated the cell cycle arrest, we faced the danger of observing differences that are not explained by the protein depletion directly. The distribution of cells in each phase of the cycle in (Chathoth et al. 2022) revealed no effect of knockdowns on cell growth or cell cycle arrest.

2.2.2. Creation and correction of Hi-C matrices using HiCExplorer

We used HiCExplorer to build and correct the contact matrices. HiCExplorer is a widely used tool in the field of 3D chromatin organisation as it addresses the wide range of tasks associated with Hi-C data analysis including the processing, normalisation, data format transformation, TAD calling and visualisation (Ramirez et al. 2018). We performed the following steps before starting the analysis: first, we map the reads to the reference genome; second, we create a contact matrix and then we perform the correction to remove biases and filter bins with poor read coverage.

Reads were aligned to the *Drosophila melanogaster* (dm6) genome (Adams et al. 2000; dos Santos et al. 2015) using BWA-mem (Li and Durbin 2010) with options -t20 -A1 -B4 -E50 -L0. The parameters were selected as recommended in the example

	Library size	Total reads	Mappable reads	Pair used
BG3 wild-type (GSE122603)				
Replicate 1	100%	151,493,083	101,206,098	42,160,814
Replicate 1	80%	121,195,638	81,045,523	34,528,646
Replicate 2	100%	147,730,225	95,703,042	42,999,219
Replicate 2	80%	118,173,473	76,656,728	35,462,491
BG3 BEAF-32 KD (GSE147057)				
Replicate 1	100%	172,597,536	69,859,110	32,240,219
Replicate 1	80%	138,073,566	55,898,357	26,805,854
Replicate 2	100%	156,655,569	81,740,811	41,458,552
Replicate 2	80%	125,326,239	65,443,211	33,860,610
BG3 Cp190 Chro KD (GSE147057)				
Replicate 1	100%	214,736,070	146,124,131	65,278,295
Replicate 1	80%	171,799,624	117,064,857	53,376,413
Replicate 2	100%	167,983,753	104,046,976	45,830,412
Replicate 2	80%	134,385,744	83,338,488	37,245,423
BG3 BEAF-32 Dref KD (GSE147057)				
Replicate 1	100%	197,221,569	107,472,262	37,738,037
Replicate 1	80%	157,771,608	86,029,067	31,775,241
Replicate 2	100%	190,654,840	53,234,566	18,645,676
Replicate 2	80%	152,530,267	42,612,149	15,651,046

Table 2.1. Metrics from analysis of Hi-C library sequencing.

usage in HiCExplorer documentation (Ramirez et al. 2018).

Next, we build matrices using DpnII restriction sites with minimum allowed distance between sites of 150 bp and maximum distance of 1000 bp. So, we obtained matrices with 217638 bins with median width of 529 bp. We worked with two biological replicates in wild-type and in all three mutants. The biological replicates were corrected separately and then merged. Biological replicates can be processed, normalised, and then passed to TAD calling algorithm independently. Then, TADs which are present only in both replicates can be selected for downstream analysis as the most confident ones. However, the TAD calling algorithms and, in particular, HiCExplorer are sensi-

tive to library sizes – when the library size is low, the difference between intra-TAD and inter-TAD contact frequencies is less pronounced, so the algorithm requires less strict thresholds for TAD detection. In addition, TAD border is not punctuated position, it is allocated within the short window – TAD boundary. When the differences between inter-TAD and intra-TAD interactions are not significant, the TAD boundary can be wider as there is no strict separation between neighbouring TADs. In this case, we face a danger that less TAD borders would have exact coincidence. In the current analysis, both replicates and merged matrices after correction step displayed high similarities for contacts-vs-genomic distance relationships, so we can continue for the downstream analysis with merged matrices which are larger in library size then both replicates (Figure 2.1.A). For merged matrices, we filtered the bins which demonstrated unexpectedly low/high counts (Figure 2.1.B and Table 2.2). Note that HiCExplorer performs the correction based on the ICE method (Imakaev et al. 2012) which requires pre-filtering. Bins with extremely low coverage tend to contain repetitive regions, as well as bins with extremely high coverage tend to contain copy number variations, so they should be removed before the matrix is proceeded to correction. The latest versions of HiCExplorer allow KR balancing (Knight and Ruiz 2013) as well - this method does not require filtering.

Note that the number of reads obtained in the result of the processing were consistent with data sets used in previous research in *Drosophila* (Cubenas-Potts et al. 2017; Ramirez et al. 2018; Chathoth and Zabet 2019).

2.3. TAD reorganisation analysis

2.3.1. TAD calling with HiCExplorer

We aim to investigate whether the protein knockdowns lead to the changes in TADs. We used the corrected contact matrices to detect TADs using HiCExplorer (Ramirez et al. 2018). We selected the parameters to ensure that we recover a similar number of TADs as previously reported (Cubenas-Potts et al. 2017; Ramirez et al. 2018; Chathoth and Zabet 2019): minimum TAD width at 5 Kb, P-value threshold at 0.01 with

	Dpn II		10 Kb	
	100%	80%	100%	80%
BG3 wild-type				
Replicate 1	[-1.4; 5]		[-1.4; 5]	
Replicate 2	[-1.4; 5]		[-1.4; 5]	
Merged	[-1.4; 5]	[-1.4; 5]	[-1.4; 5]	
BG3 BEAF-32				
Replicate 1	[-1.0; 5]		[-0.8; 5]	
Replicate 2	[-1.2; 5]		[-1.4; 5]	
Merged	[-1.2; 5]	[-1.2; 5]	[-1.4; 5]	
BG3 Cp190 Chro KD				
Replicate 1	[-1.2; 5]		[-1.4; 5]	
Replicate 2	[-1.2; 5]		[-1.4; 5]	
Merged	[-1.2; 5]	[-1.2; 5]	[-1.4; 5]	
BG3 BEAF-32 Dref KD				
Replicate 1	[-1.2; 5]		[-1.4; 5]	
Replicate 2	[-1.2; 5]		[-1.4; 5]	
Merged	[-1.2; 5]	[-1.4; 5]	[-1.4; 5]	

Table 2.2. Filtering parameters of Hi-C matrices correction with HiCExplorer.

FDR correction for multiple testing and a minimum threshold of the difference between the TAD separation score of 0.04. TAD separation score, in simple, represents the average interaction frequency between each locus and its nearby DNA fragments. Within the TAD, the contact frequency is average higher than the frequency at the TAD border (between two neighbouring TAD). The difference between TAD separation score at the locus and neighbouring regions should be more than pre-selected threshold to be selected as a candidate position for TAD border. The Mann-Whitney U test then performed to estimate the significance of this difference and as multiple tests are performed, the multiple-testing correction Is required. FDR correction is not as strict as, for example, Bonferroni correction, but it is helpful when we require more exploratory analysis rather than extremely strict results. Under selected thresholds, we identified

between 1417 and 2260 TADs (Figure 2.2.A).

HiCExplorer as any computational method is sensitive to parameter selection. This raises the difficulties for downstream analysis. When the weak set of parameters are selected we face a danger of detecting TADs that are real and can be further analysed, as well as spurious TADs that are detected due to experimental and processing biases. So, selection of stricter parameters helps us to retrieve more reliable TADs and remove TADs that are probably spurious ones. Given this motivation, we also called strong borders using a stringent value of the threshold of the difference between the TAD separation score of 0.08. This value ensures that we retrieved the strongest borders - we identified between 1136 and 441 strong TADs (Figure 2.2.B).

Differences in library sizes can also introduce uncertainty in comparative analysis of TAD organisation between wild-type and mutants. Thus, lower number of reads would make the differences between interaction frequencies within inter- and intra-TAD areas to be less pronounced and, as a result, the insulation strength between chromatin domains would be less significant. In the current study, to investigate the robustness of detected TAD borders to difference in the size of Hi-C libraries, we downsampled all Hi-C libraries by 20% and repeated the analysis (Table 2.1)

TAD borders identified with downsampled Hi-C libraries exhibited reassuring overlap with TAD borders detected in the full data set for both weak and strong calling parameters (Figure 2.2). After downsampling, we recovered not less than 69% of weak and 66% of strong borders - these borders we defined as robust meaning they are recovered in both full and downsampled datasets.

2.3.2. Differential gene expression and TAD reorganisation

TADs were found to correlate with gene expression and proposed to establish contacts between distal enhancers/silencers and target genes, as well as insulate the undesired connection which can lead to gene misregulation. Previous studies either induced the structural variations altering chromatin topology or depleted architectural proteins enriched at TAD boundaries to decrease the insulation between chromatin do-

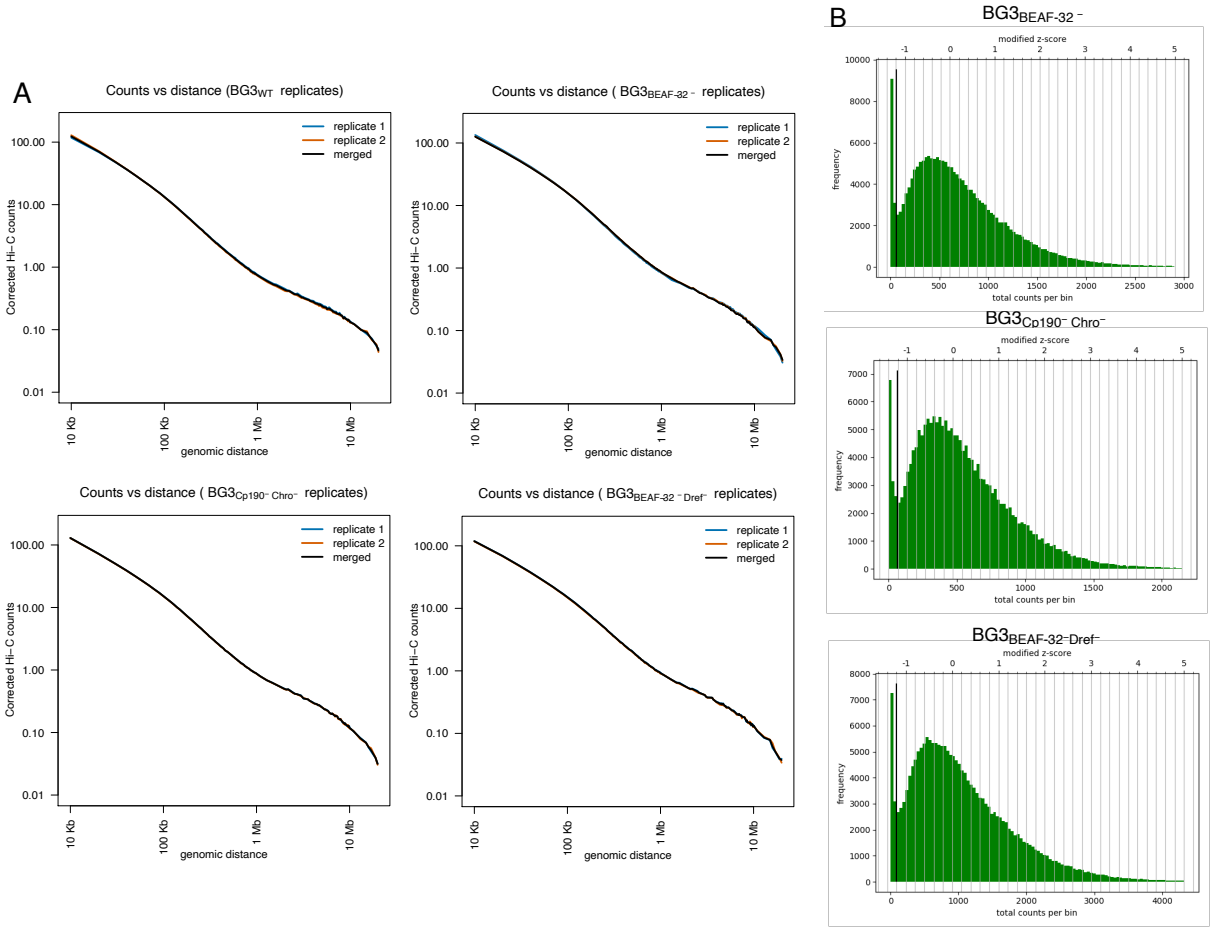


Figure 2.1. Hi-C processing and visualisation. **A.** Hi-C contacts versus distance plots for two replicates and merged datasets on BG3 wild-type, BEAF-32 single knockdown, Cp190 Chro double knockdown and BEAF-32 Dref double knockdown. **B.** Diagnostic histograms for Hi-C corrections in BEAF-32 single knockdown, Cp190 Chro double knockdown and BEAF-32 Dref double knockdown. The vertical black line represents the lower threshold for removing low read bins.

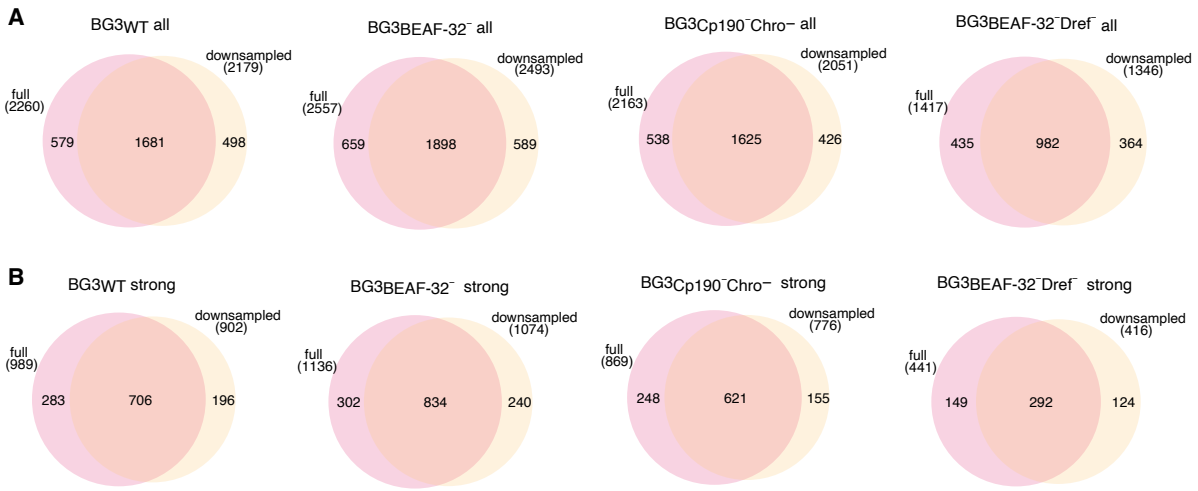


Figure 2.2. Robustness of TAD borders. We defined as robust the borders which were detected on both the full dataset and a downsampled dataset where we randomly removed 20% of the reads. We consider the case of all (B) and strong (A) borders separately for Hi-C datasets in wild-type and BEAF-32 single knockdown, Chro and Cp190 double knockdown and BEAF-32 and Dref double knockdown.

mains (Lupianez et al. 2015; Kragesteen et al. 2018; Owens et al. 2021). Here, we perturbed a large number of TADs by knocking down architectural proteins and investigated whether that leads to changes in gene expression. We grouped possible rearrangements depending on relative position of genes at wild-type and mutant TADs: TAD border loss, gain or shift can, theoretically, happen within a gene body (Figure 2.3.A), or gene can fully reside within both wild-type and knockdown TADs (Figure 2.3.B). According to the second scenario, TAD borders can stay absolutely conserved or fuzzy conserved (less than 2 Kb shifts of both borders) which possibly reflects the perturbation of inside-TAD contacts (possibly indicating the existence of sub-structures), as well as distal within-TAD contacts while the TAD still stays insulated from neighbouring chromatin (possibly indicating the existence of higher-order TAD network). The second scenario also include massive rearrangements that possibly disrupt old or organise new contacts with the gene.

Genes guide the protein production: they store the information that is transcribed into RNA and then translated into proteins (Crick 1958; Crick 1970). The alterations

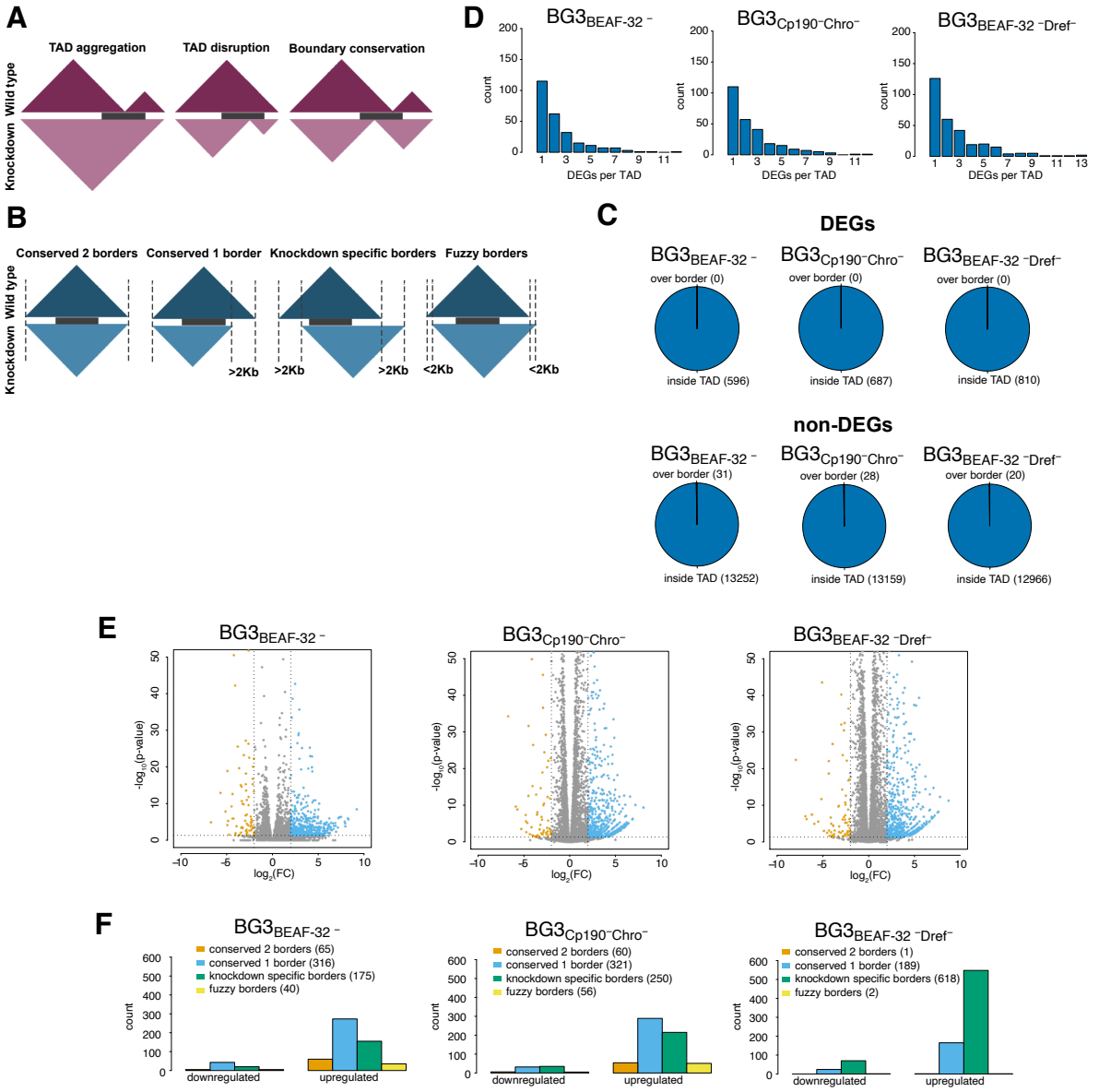


Figure 2.3. The effects of TAD reorganisation on transcription. **A.** Scenarios for relative position of genes, wild-type and mutant TADs when gene spans over TAD borders. **B.** Same as A, but gene is within the both wilt-type and mutant TADs. **C.** Distribution of differentially expressed genes (DEGs, top panel) and non-differentially expressed genes (non-DEGs, bottom panel) between (A) and (B) scenarios. **D.** Histogram represents the number of TADs containing specified number of DEGs (x-axis). **E.** Volcano plots for the RNA-seq analysis (orange represents downregulated genes, blue upregulated and grey non-DEG) in BEAF-32 knockdown, Cp190 Chro double knockdown and BEAF-32 Dref double knockdown. **F.** Distribution of downregulated and upregulated genes between scenarios in (A): orange – both TAD borders are conserved, blue – only one of the TAD border is conserved, green – none of the TAD border is conserved and yellow – TAD borders are shifted within 2 Kb.

in gene expression result in overproduction or underproduction of RNA molecules in comparison with the normal state (control). RNA sequencing (RNA-seq) is a high-throughput sequencing method which, in simple words, quantifies the amount of RNA present in a population of cells. It allows to spot any relative changes in the RNA expression level when comparing data generated in two conditions, for example, in wild-type cells and in protein knockdown. We used the RNA-seq data produced by (Chathoth et al. 2022) (GSE147059) to identify transcriptional changes upon BEAF-32, Cp190, Chro and Dref knockdown. There were three replicates for each condition: wild-type, BEAF-32 KD, Cp190 Chro KD and BEAF-32 Dref KD. We processed RNA-seq as in (Chathoth et al. 2022). We trimmed reads using Trimmomatic v0.39 (Bolger et al. 2014), then aligned to the *Drosophila melanogaster* (dm6) genome (Adams et al. 2000; dos Santos et al. 2015) using TopHat v2.1.2 (Kim et al. 2013) with Bowtie2 v2.3.4.1 (Langmead and Salzberg 2012) and finally removed duplicated reads with the Picard tool (http://broadinstitute.github.io/picard/). We counted reads using HTseq (Andres et al. 2015) and calculated the significance of the difference in expression between wild-type and knockdowns using the DESeq2 algorithm (Love et al. 2014) for genes with at least 10 reads aligned. A gene was defined as differentially expressed when the adjusted p-value generated in DESeq2 was less than the threshold of 0.05 and the absolute value of normalised log2 fold change between reads of knockdown over wild-type exceed the threshold at 2.0 (at least 4 times increase/decrease in reads during knockdown comparatively to the wild-type).

We found significant changes in gene expression with 598, 688 and 814 differentially expressed genes (DEG) in BEAF-32 KD, Cp190 Chro KD and BEAF-32 Dref KD, respectively. None of the DEGs spanned the robust TAD borders in either wild-type or knockdowns (Figure 2.3.C) and several TADs contained more than one DEG (Figure 2.3.D). A larger number of genes were upregulated in knockdowns compared to WT with most of them associated with significant architectural changes as single border shift or knockdown specific reorganisations (Figure 2.3.F). Very few DEGs belonged to TADs that had both borders conserved or fuzzy conserved in the knockdowns. In-

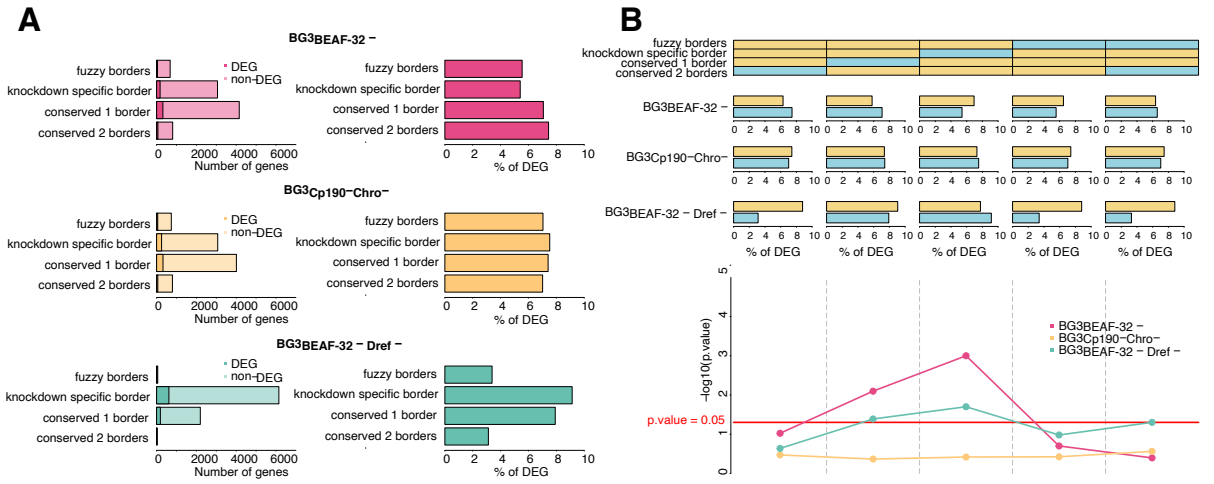


Figure 2.4. Association between architectural rearrangement and ratio between differentially and non-differentially expressed genes. **A.** Count and proportion of DEG and non-DEG in the three knockdowns BEAF-32 knockdown, Cp190 Chro double knockdown and BEAF-32 Dref double knockdown within reorganised TADs which are fully conserved, lose one or both borders or have slightly shifted borders in the knockdowns. **B.** We applied permutation test to investigate whether DEGs overlap with any of the class of architectural changes more than expected by chance. Different combinations of TAD rearrangements were checked: blue and yellow colors identify the subgroups compared. Scatter plot represents corresponding -log10 p-values at each combination.

terestingly, some of non-differentially expressed genes can span robust TAD border (Figure 2.3.C). If TAD border is allocated within gene body, it can potentially prevent, for example, the contact between the promoter which belong to upstream TAD from enhancer which belong to the downstream TAD. Note that the amount of such genes is not significant in comparison to the genes which were found within TADs.

Despite the fact that we observed the gene alterations mostly associated with dramatic domain reorganisations, we also detected many genes that were not affected by the same architectural changes (Figure 2.4.A). The proportion of genes which altered their expression patterns in result of knockdowns were about 10% in each reorganisation scenario. Given these findings, we suggest that TADs do not follow a simple scenario where the boundary disruption or reduction of insulation between domains lead to dramatic gene activation.

We also decided to perform a permutation test to check whether the ratio of differentially and non-differentially expressed genes was significantly different in some specific group of architectural rearrangements than anywhere within the genome (Figure 2.4.B). For the permutation test, we applied regionR package with 1000 permutations (Gel et al. 2016). The whole genome in wild type was split on short genomic regions where each region was classified as one of four proposed architectural rearrangements. Then, we tested whether there any significant association between differentially expressed genes and regions belonging to pre-selected architectural changes (Figure 2.4.B, classes defined with blue colour versus classes defined with yellow colour). We found that the proportion of DEG was higher than expected in comparison with the genome-wide distribution in two cases: either when a single TAD border was conserved or when the TAD borders were dramatically reallocated ("knockdown specific border" group). This tendency, however, was BEAF-32 KD and BEAF-32 Dref KD specific - double knockdown of Cp190 and Chro did not show noticeable changes in differential expression associated with specific topological changes.

In (Nora et al. 2017) domain reorganisations caused by the depletion of the architectural protein CTCF also resulted in misregulation of genes which promoters were found to be allocated too close to disrupted borders. We also wanted to look at the allocation of DEGs to spot any preference of altered genes (Figure 2.5). As we know that genes fully lie within TADs, we introduce the start and end ratios: the start ratio is computed as distance between TAD start to gene start (if gene is on positive strand; when gene is on negative strand, it is the gene end) over the half TAD distance; similarly we compute the end ratio but at the end half of the TAD. When the start ratio is larger than 1, it means that the gene is allocated closer to the end of the TAD, when the end ratio is larger than 1, the gene is allocated closer to the start of the TAD. When both start and end ratios are less than 1, the gene spans the centre of the TAD (Figure 2.5, top panel). Note, that ratios are computed with respect to the gene allocation within wild-type TADs. Just a few DEGs were found to cross the TAD centre, with the majority of genes randomly distributed inside TADs with no specific localisation near or

away from TAD border for all reorganisation scenarios in all three mutants.

Depletion of architectural proteins caused the large perturbations in genome-wide chromatin conformation along with changes in gene expression. Genes which changed their expression demonstrated upregulation possibly indicating the roles of TADs in maintaining of repressing state of that genes. We suggest that massive rearrangements happened mostly within heterochromatin while the borders within euchromatin were less sensitive to the protein depletion. In case of BEAF-32 single knockdown and BEAF-32 and Dref double knockdown, significantly more genes altered their expression pattern when were allocated within either TADs that lost one border or both borders demonstrated dramatic shift - more than 2 Kb away from their WT position. However, depletion of Cp190 and Chro did not significant statistical association between DEGs and any TAD reorganisation pattern. The difference can possibly be explained by the direct (BEAF-32 and Dref) and indirect (Cp190 and Chro) binding of the depleted proteins to the DNA. Cp190 and Chro can be recruited by other proteins which were not significantly involved in TAD organisation, so we did not observe significant association between gene expressional patterns and chromatin architecture.

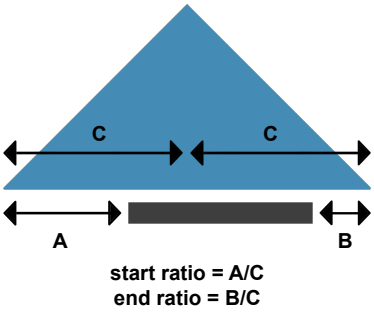
2.3.3. Comparative analysis of wild-type versus mutants

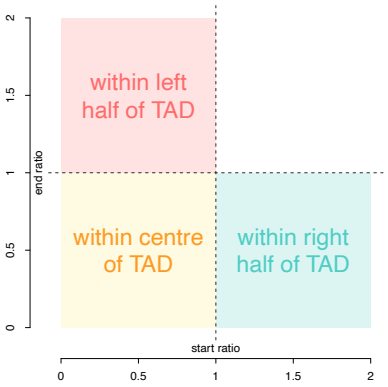
We detected more robust TAD borders, both weak and strong, with BEAF-32 KD (Figure 2.6.A). Cp190 and Chro double knockdown did not bring a significant difference in the number of TADs. So, the depletion of BEAF-32, Cp190 and Chro was suggested to affect not only the number of TAD borders but their allocation as well. We compared robust TAD border positions between WT and mutants and defined 10 classed of changes:

1. **strong** $\rightarrow$ **strong:** border in WT had exactly the same position in mutant and annotated as strong in both WT and mutant.

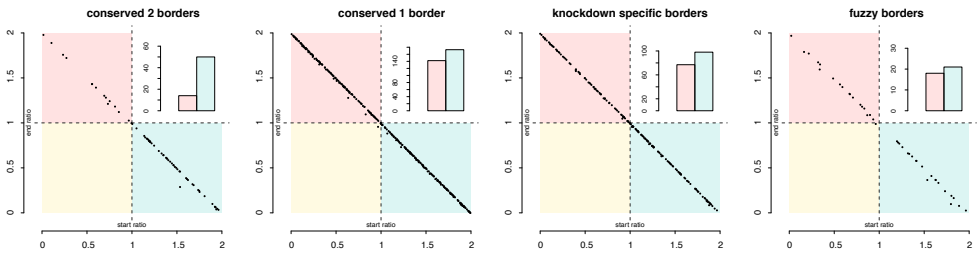
2. **weak** $\rightarrow$ **weak:** border in WT had exactly the same position in mutant and annotated as weak in both WT and mutant.

3. **strong** $\rightarrow$ **weak:** strong border in WT had exactly the same position in mutant but

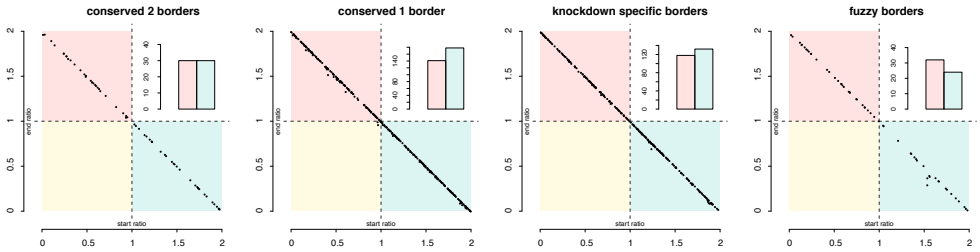




BG3$_{BEAF\text{-}32^-}$



BG3$_{Cp190^-Chro^-}$



BG3$_{BEAF\text{-}32^-DREF^-}$

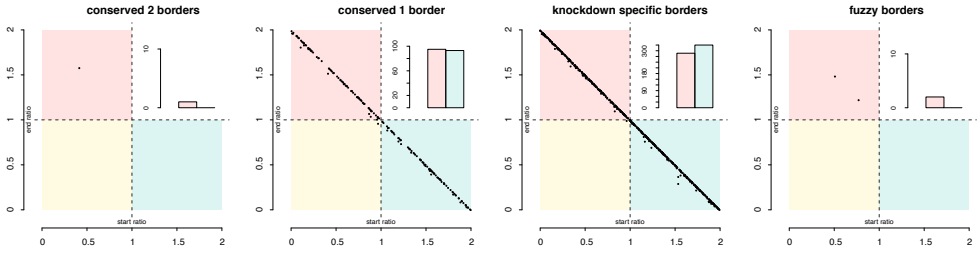


Figure 2.5. The allocation of differentially expressed genes within robust TADs in BEAF-32 knockdown, Cp190 and Chro double knockdown, BEAF-32 Dref double knockdown. The start ratio is defined as a distance from the left border of the TAD to the start position of the gene divided by the half of TAD size, where a start ratio bigger than 1 means that the gene is allocated on the right side of the TAD (green square). The end ratio is defined as a distance from the right border of the TAD to the end position of the gene divided by the half of TAD size, where an end ratio bigger than 1 indicates the gene allocated on the left side of the TAD (red square). Genes having both start and end ratio less than 1 are allocated within TAD centre (yellow square). The majority of differentially expressed genes occupy less than half of the TAD. Only couple of genes are allocated within TAD centre – they are very close to point (1,1) indicating relatively short genes. Majority of genes are allocated either on the left or the right half of the TAD with no strong bias towards TAD borders.

became weak.

4. **weak** $\rightarrow$ **strong:** weak border in WT had exactly the same position in mutant but became strong.

5. **strong** $\rightarrow$ **fuzzy:** strong border in WT was found within 2 Kb window in mutant.

6. **weak** $\rightarrow$ **fuzzy:** weak border in WT was found within 2 Kb window in mutant.

7. **strong** $\rightarrow$ **no:** strong border in WT was not found within 2 Kb window in mutant.

8. **weak** $\rightarrow$ **no:** weak border in WT was not found within 2 Kb window in mutant.

9. **no** $\rightarrow$ **strong:** strong border in mutant was not found within 2 Kb window in WT.

10. **no** $\rightarrow$ **weak:** weak border in mutant was not found within 2 Kb window in WT.

Note that for weak borders we considered only unique weak borders that were the ones that did not belong to the set of strong borders. Also, the 2 Kb window to define fuzzy borders was selected based on minimum width parameter: we treated the shift in TAD borders between WT and mutant being less than half of minimum TAD width as a product of negligible matrix differences rather than a consequence of the proteins depletion.

In Figure 2.6.B, we represented the distribution of TAD border changes based on the classification above. The largest number of rearrangements in wild-type could be described as loosing or gaining of weak borders. It was not surprising, as small

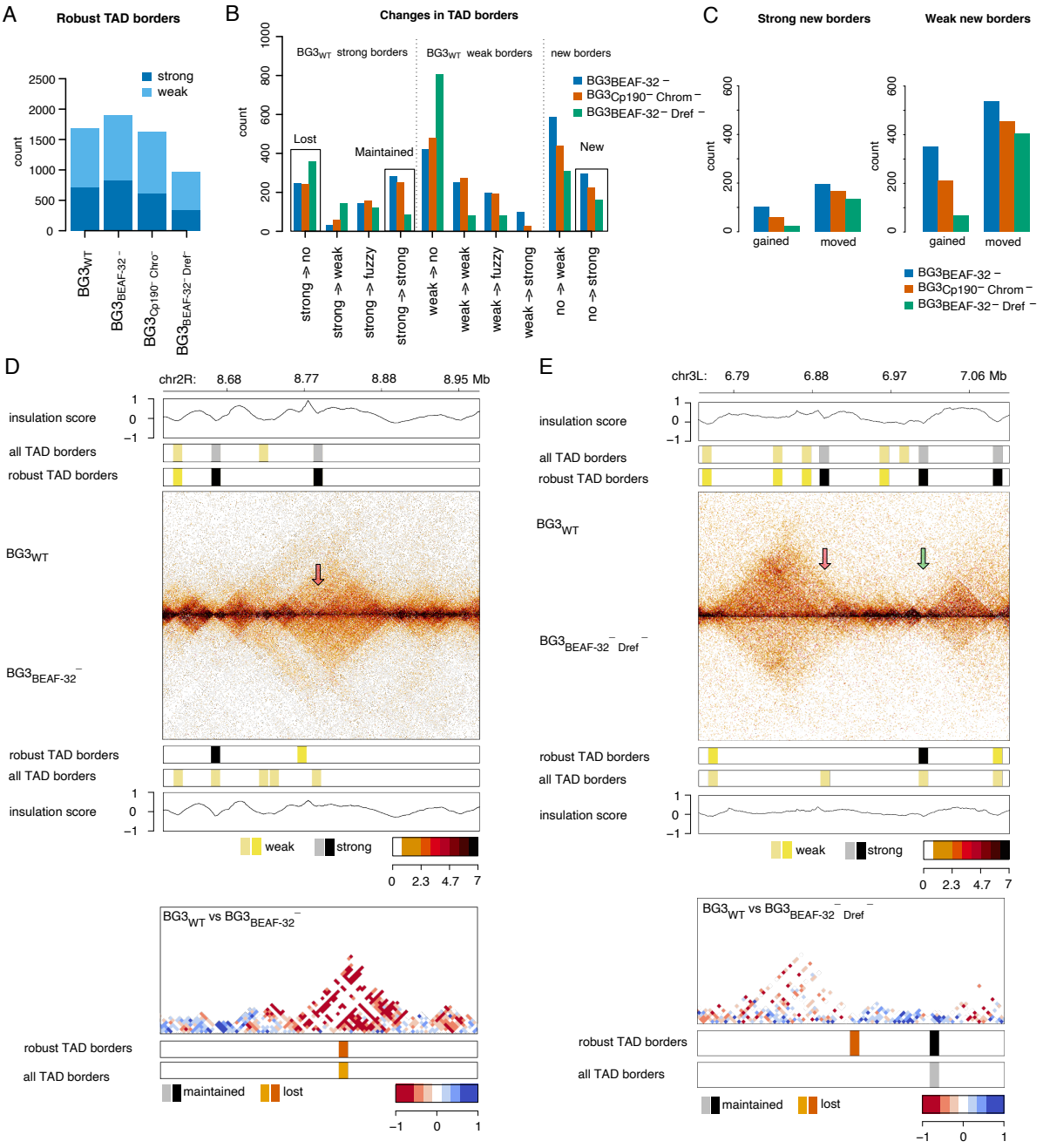

A
Robust TAD borders
strong
weak
count
B
Changes in TAD borders
Lost
Maintained
New
C
Strong new borders
Weak new borders
gained
moved
D
E
insulation score
all TAD borders
robust TAD borders
weak
strong
maintained
lost

Figure 2.6. TAD border reorganisations in the knockdowns. **A.** Number of robust TAD borders in wild-type cells, BEAF-32 knockdown, Cp190 Chro double knockdown and BEAF-32 Dref double knockdown. We split each class of TAD border into strong borders and weak borders, depending on whether the TAD borders can still be detected when increasing the stringency of the TAD calling algorithm. **B.** Classification of TAD border based on classes and positions as described in the main text. We focus our attention on lost, maintained and new borders. **C.** Distribution of new TAD borders in the knockdowns between gained, which appear inside the wild-type TAD, and moved, which correspond to relocation dramatic reallocation of wild-type TAD border. **D-E.** Examples of a borders classified as lost and maintained in BEAF-32 knockdown (C) and Cp190 Chro double knockdown (D). From top to bottom we plot the insulation score, TAD borders in full dataset (grey are strong and yellow are weak), TAD borders recovered both in full and downsampled dataset (black are strong and yellow are weak) and contact map in wild-type cell and then mirror plots in the knockdowns. Darker colours on heat maps indicate more contacts retrieved by Hi-C. Green arrows indicate maintained borders and red arrows lost borders. We also plot log2fold change between wild-type and knockdowns in 5 Kb bins build with diffHiC (Lun and Smyth 2015) and edgeR (Robinson et al. 2010) packages.

shifts in TAD separation score could affect whether the border could not be detected or detected as weak. The changes associated with strong borders were of the greatest interest to us as they have a high chance to reflect real and pronounced changes in chromatin architecture. We detected significant amount of borders that we classified as lost ("strong $\rightarrow$ no" class), maintained ("strong $\rightarrow$ strong" class) and new ("no $\rightarrow$ strong" class) borders.

While the maintenance and loss of TAD borders in absence of proteins-insulators are expected, the noticeable amount of new strong borders was quite interesting. We found new borders formation ranging from 200 to 300 strong borders (Figure 2.6.C). To be classified as a new, a mutant TAD border should be allocated more than 2 Kb away from the nearest TAD border in WT. So, this definition covers both actual new borders which were organised by splitting the original TAD in several separate sub-structures or borders which appeared due to a move from WT border by more than 2 Kb. We detected new borders which were organised by both splitting and movement with the

slightly larger preference towards new borders organised by movement (Figure 2.6.C). Both gained new borders and new borders resulted in shift of the wild-type ones can probably be explained by the reallocation of remaining proteins after the knockdown. Also, other insulation proteins can replace the role of BEAF-32, Dref, Cp190 and Chro and maintain the insulation at the position of new borders in knockdowns. However, as there is no publicly available ChIP data on various architectural proteins in knockdowns, we don't have enough evidence to support this hypothesis. The alternative explanation could be related to the differences in Hi-C libraries between wild-type and knockdowns: if the border does not pass the downsampling in wild-type but passes in knockdown, it will be defined as new. Also, TAD border is not a strict punctuated position, it is allocated within the window (TAD boundary), so the fluctuations in the position of TAD border are possible. When two neighboring TADs are not strictly segregated from each other, the TAD boundary can be large enough to show differences in the same TAD border position between wild-type and mutant. However, in absence of ChIP data we cannot properly access whether these differences are the downstream results of Hi-C and robustness analysis or related to protein binding.

The difference between maintained and lost borders can shed light on the interplay between BEAF-32, Cp190, Chro and Dref with other epigenetic machineries in the establishment of specific chromatin architecture. Technically, a strong TAD border could be lost due to two reasons. First, the border could be completely lost in mutant due to aggregation of neighbouring TADs or due to dramatic shift at more than 2 Kb which signaled a weakening of insulation between neighbouring genomic regions as well. Second, the strong TAD border in mutant lost or shifted during downsampling, while the same TAD border in wild-type survived the downsampling. So, again we faced a significant difference in insulation strength between wild-type and mutant occurring due to proteins knockdown.

The second scenario was represented in Figure 2.6.D and E. The red arrow indicated the lost border that was in wild-type classified as strong both before and after the downsampling (panels above the heat map), while in mutant the same genomic

position was detected as a weak border before downsampling only (panels below the heat map). Here, this genomic position represented the case where we observe dramatic weakening of insulation between neighbouring TADs and TAD border disruption. To confirm the difference in Hi-C interactions brought by the proteins depletion at the selected region, we computed log2 fold change between wild-type and knockdown Hi-C map following the steps and parameters recommended in diffHiC package (Lun and Smyth 2015). Briefly, we considered individual replicates and used edgeR package (Robinson et al. 2010) to compute the log2 fold change between maps using 5 Kb bins. The difference heat map (Figure 2.6.D and E, bottom panel) confirmed the loss of interactions in wild-type within the region of lost TAD border indicating the loss of insulation.

Note that we did not use the statistical testing for detecting differential TADs. As an insulation score-based approach, HiCExplorer defines the genomic region to be, most probably, a TAD border when it demonstrates significantly different interaction frequency in comparison with neighbouring upstream and downstream regions. The threshold of the difference between the TAD separation score of 0.04 and 0.08 (see Section 2.3.1) sets how significant this difference should be for the TAD border to proceed further in the analysis. However, depending on experimental set up and library sizes, the TAD separation score as well as the differences in it can vary between different data sets. The statistical tests which compare the TAD separation scores and how dramatic it changes at TAD border, can give us a clear understanding how significant are the differences in insulation strength of TAD borders between wild-type and knockdowns. Also, tests can provide more information about the borders that were lost during the downsampling. For example, when the changes in TAD separation scores were not dramatic initially, then the TAD borders would be more sensitive to the reduction of the library size during the downsampling and the borders which should be considered in the analysis would be lost, which, in turn, generates spurious lost and new borders. So, skipping the statistical testing for differential TAD borders affect the number of maintained, lost and new borders which we considered for further analysis

of protein occupancy and biological functioning.

2.3.4. Separation of direct and indirect effect

Depletion of proteins induced a significant rearrangement of TAD organisation. However, some of the topological changes could resulted from downstream effects, so TAD border rearrangements, especially gain and loss of insulation between domains, could be indirect targets of protein depletion which made the observed effects unclear. To distinguish between the direct and indirect targets, we are interested in borders that in wild-type cells were colocolised with BEAF-32, Cp190, Chro and Dref, so were directly affected by protein depletion.

The protein binding that happens more often than by chance potentially plays a functional role in the maintenance of chromatin architecture and cellular functioning. The chromatin immunoprecipitation (ChIP) technology enables the quantification of protein-DNA interactions of a specific protein or associated factor. We used ChIP-chip datasets, in particular, ChIP peaks called by modENCODE Consortium. To dissect the direct effect of protein knockdowns, we were interested in occupancy of BEAF-32 (GSE32775, GSE20811), Chro (GSE20761) and Cp190 (GSE32776, GSE20814) in *Drosophila* BG3 cells. Unfortunately, there was no Dref data generated in the required cell type, so the mechanics of changes in TADs in BEAF-32 Dref KD was not as clear as in other two mutants. Based on this, we made a decision to continue the downstream analysis based on BEAF-32 KD and Cp190 Chro KD only.

Compared to wild-type cells, out of all 706 strong borders, 188 borders were maintained and 149 were lost in both BEAF-32 KD and Cp190 Chro KD (Figure 2.7.A), the rest of strong borders either maintained/lost uniquely in specified mutants, defined as fuzzy or moderately weakening. Note that new borders, in contrast, were mostly knockdown-unique. The majority of common maintained borders (94%) are direct targets of at least one of three proteins depleted during knockdown (Figure 2.7.B and C). We also observed that at maintained borders Chro and Cp190 almost everywhere colocolises with BEAF-32. However, at borders where BEAF-32 was not found, we

detect some of Cp190 together with CTCF. Note that CTCF ChIP-chip peaks were also generated and processed by modENCODE Consortium (GSE20767 and GSE32783). In contrast, only 47% of lost borders were direct targets of protein loss suggesting that the other half of lost borders were affected by downstream perturbations. Interestingly, borders that were organised after protein depletion in both mutants were also enriched with BEAF-32, Cp190 and Chro in wild-type. However, a particular border can be defined as new in the case when it was completely undetected in wild-type after the downsampling and it was detected as strong in mutant. So, it was still possible that the border was detected in the full set.

Here, we called TADs on Hi-C data generated in (Chathoth and Zabet 2019; Chathoth et al. 2022) while comparing it with ChIP protein enrichment data generated by modENCODE Consortium, so we can face some inconsistencies due to different experimental conditions. Also, we know that the proteins were significantly depleted, but not in full and in absence of ChIP data generated on knockdown cells we are restricted in how we can properly check it. However, BEAF-32 and Cp190 single RNAi (GSE32773, GSE32774, GSE32816) knockdowns were generated in BG3 cells in (Schwartz et al. 2012), so analysing these data we found that the majority of maintained TAD borders (70%) retain BEAF-32 or Cp190 upon knockdown while most of the lost borders (70%) lose binding of these architectural proteins after knockdown (Figure 2.7.D-E). Despite the fact that the Hi-C and ChIP data were generated on different samples and in different experimental conditions, we still obtained the strong difference, which supports, that the borders that are lost are true direct targets of the architectural proteins.

Interestingly, we observed 79 (53%) of lost borders which were lost without direct binding of architectural proteins (Figure 2.7.C). From the technical point of view, the indirect lost borders can result in the robust analysis, when in the knockout the border did not pass the downsampling when in the wild-type it did. However, the lost borders were defined as common lost borders in two mutants (BEAF-32 knockdown and Cp190 and Chro double knockdown) which can happen if the insulation score of TAD borders in both mutants was generally lower than in wild-type. From the biological point of

view, the indirect lost borders were possibly in contact with direct lost borders forming the complex architectural structures. In that case, the direct lost border can cause further perturbations in nearby chromatin topology.

The detection of TAD borders and further comparative analysis of TAD border allocation was based on insulation score-based approach implemented as a part of HiC-Explorer tool. A comparison of insulation scores between different conditions, from a technical point of view, can give an unclear picture on the seriousness of architectural changes. The insulation score measures the average Hi-C interactions around a selected genomic position. While the relative contact intensity between neighbouring regions under the same Hi-C map can shed a light on differences between inside-TAD and outside-TAD interactions, the absolute insulation score reflects such characteristics as library size, so the comparison of insulation strength between wild-type and mutant can be questionable. So, the detection of the maintained and lost TAD borders was performed following a robust analysis using five filtering steps. First, we used two threshold values to distinguish between strong and weak TAD border insulation. Second, to account the differences in Hi-C libraries, we performed the downsampling of the libraries by 20% and repeated the TAD calling procedure. We select the borders that were retained upon the downsampling as robust borders. Third, we defined as maintained and lost only such borders that demonstrated the most pronounced switch in insulation strength while their position stays the same. Fourth, for the downstream analysis we left only borders that were common between BEAF-32 single and Cp190 and Chro double knockdowns. Fifth, the only direct targets of protein depletion were retained. Overall, after so many trimming steps we still observed strong difference between maintained and lost epigenetic signatures. Altogether, this analysis supports that the observed result are robust.

2.3.5. Adjusting the ChIP profiles for further comparative analysis

ChIP data generated and processed by modENCODE Consortium include a variety of epigenetic factors like architectural proteins, transcription, replication and accessibility

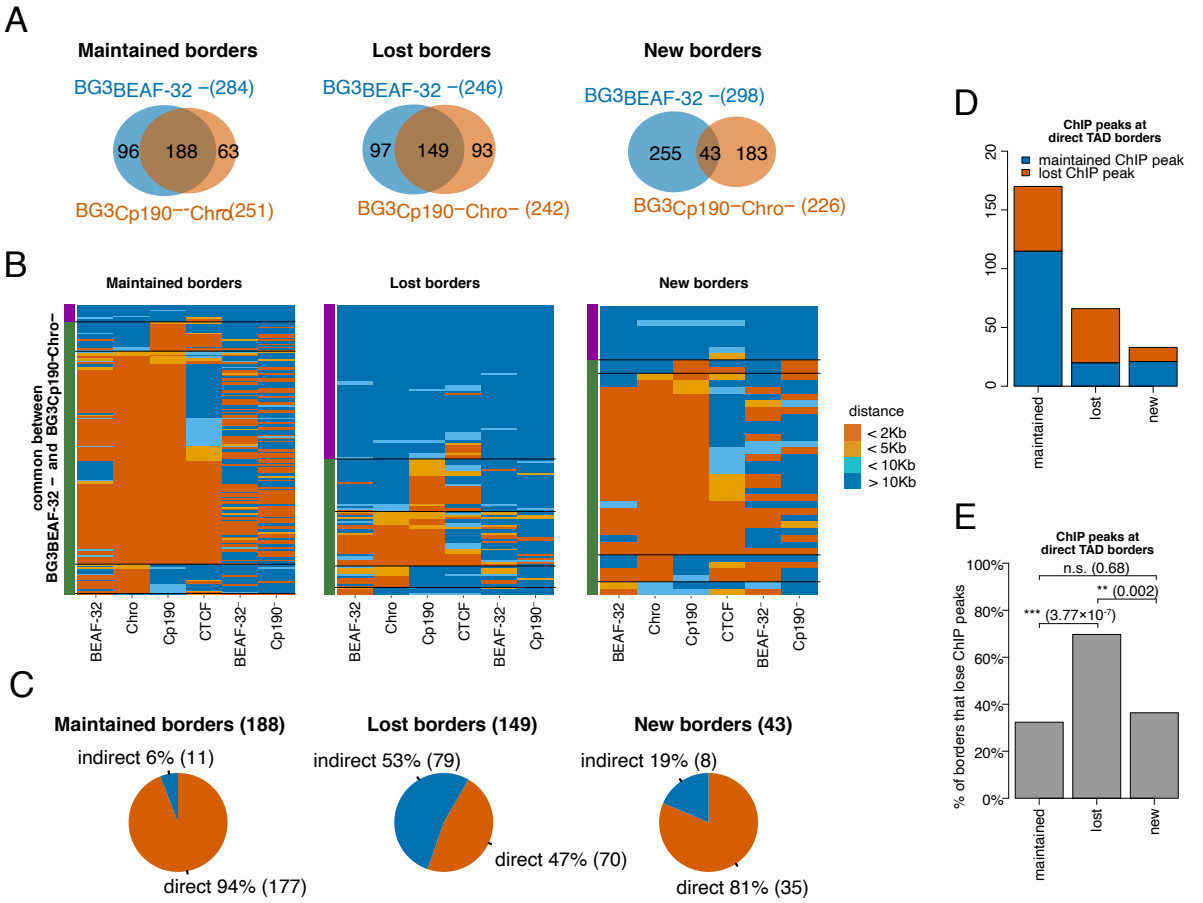


Figure 2.7. Direct and indirect TAD borders. **A.** Overlap of lost, maintained and new robust borders which are common in BEAF-32 KD and Cp190 Chro KD. **B.** Heatmaps represent the distance of the closest ChIP peak from a maintained, lost and new border which are common in both mutants. ChIP peaks obtained for BEAF-32 (wild-type and BEAF-32 knockdown), Chro (wild-type), Cp190 (wild-type and Cp190 knockdown), CTCF (wild-type). Green bar on the side of each heatmap marks the direct borders (borders that show binding of BEAF-32, Chro or Cp190 in wild-type cells), while purple indirect borders (all other borders). **C.** Number and percentage of maintained, lost and new borders that have direct binding of BEAF-32, Cp190 or Chro. **D-E.** Number and percentage of TAD borders that have BEAF-32 or Cp190 ChIP peak in wild-type cells and lose or maintain those peaks in BEAF-32 and Cp190 single knockdowns. We performed a Fisher's exact test and the corresponding p-value is displayed inside the plots.

related complexes. We obtained most of the available modENCODE Consortium and (Pherson et al. 2019) data (M-values smoothed over 500 bp) on BG3 cells to determine the key differences in occupancy profiles between common maintained and lost borders in BEAF-32 KD and Cp190 Chrom KD. The full list of datasets used is provided in Appendix 2.1.

During ChIP, the fragments enriched of specific binging sites are isolated from DNA using appropriate protein-specific antibody (IP samples). The number of fragments obtained in IP samples are generally compared to the input control samples where the DNA fragments are either (1) cross-linked and sonicated under the consistent conditions as the IP samples, or (2) selected using IgG antibody that picks fragments without specific preference. For IP samples, it is expected that the fragments enriched with specific protein are detected more often than the non-enriched fragments. For input control samples, the detection is expected to be approximately flat for both enriched and non-enriched fragments. The log2 ratio values (M-values) computed between the intensity of fragments obtained from IP and input control samples are generally used to indicate the enrichment: the positive values indicate fragments detected more in IP than in input control samples (factor enrichment) while the negative values indicate fragments detected more in input control than in IP samples (factor depletion).

A comparative analysis of maintained and lost borders for each selected epigenetic factor is relatively straightforward as we compare profiles within the same biological sample. However, when comparing ChIP profiles within the same class of borders but produced for different proteins enriched, additional pre-processing schemes are required as signal-to-noise ratio varies between samples. We proposed the following algorithm which includes the profile extraction, cleaning and normalisation in order to further make trustful conclusions about protein enrichment differences.

1. We start with the extraction of M-values within 5 Kb window binned at 100 bp around selected wild-type TAD borders. A TAD border is represented as a single position when it is more like a boundary, so functionally relevant protein or factor can bind border within the region.

2. Negative values indicate non-enrichment independently on the exacted value. One of the possibilities is to replace all negative signals with zero for further analysis simplification. However, we decided to include negative values into processing as we want to use them for further clustering.

3. As M-values are computed based on log2 ratio, the large difference in IP signal intensity and input control signal intensity may produce outliers. Some regions cannot be uniquely mapped to the genome, for example, because of short length or association with transposable elements. These regions may be considered to be less functional and some researchers exclude them from the analysis. Instead of ignoring regions that produce outliers, we prefer to winsorise them. We select the cut-off points for positive and for negative signals separately: we define 5%-quantile of negative signals as down cut-off point and 95%-quantile of positive signals as up cut-off point. There is the possibility that enrichment signals are distributed around zero, so cut-off points will be selected to be 0 or very close to zero (lies within the -1 to 1 interval). For such ChIP profiles, we replace the up and down cut-off points with maximum and minimum, respectively.

4. Positive signals are scaled by the reciprocal of the up cut-off point and negative signals by the reciprocal of the down cut-off point. The signals of the resulting datasets belong to the -1 to 1 interval. Thus, the distribution within the profiles can shed light on occupancy strength: when a large amount of signals within the window are allocated next to 1 it means that the factor is strongly enriched; when large amount of signals within the window are allocated next to 0 and below it means that factor is strongly depleted.

2.3.6. Clustering plots visualise comparisons between enrichment profiles

We obtained occupancy profiles for maintained and lost borders at 50 epigenetics factors (Figure 2.9, 2.13, 2.14 and 2.15). As the number of profiles is huge and complicated to analyse simultaneously, we introduced an intuitive and simple clustering procedure that assigns the ChIP enrichment to the one of six groups: no, extra low,

low, medium, high and extra high enrichment (Figure 2.8.A). We extract ChIP signals summarized over 5 Kb window for each selected TAD border across all profiles and define 50%-quantile of positive summarized signals as positive cut-off value that distinguishes enrichment level from medium and high. The set of rules to annotate the enrichment class to the profile is the following:

1. **no enrichment:** both median and 3rd quartile are less than zero.
2. **extra low:** median is less than zero while 3rd quartile is greater than zero (3rd quartile may be greater than positive cut-off as well).
3. **low:** median and 3rd quartile are between zero and positive cut-off (1st quartile may be as negative as positive).
4. **medium:** 3rd quartile is greater than positive cut-off while the median is between 0 and positive cut-off (1st quartile may be as negative as well as positive).
5. **high:** median is greater than a positive cut-off while the 1st quartile is less than a positive cut-off (1st quartile may be less than 0 as well).
6. **extra high:** 1st quartile is greater than positive cut-off.

We provided the example of clustering algorithm performance for some selected histone modifications (Figure 2.8.B). We can notice that clustering plot (heat map) efficiently reflects the significant differences in box plot distributions. Corresponding cluster plots are also added to (Figure 2.9, 2.13, 2.14 and 2.15) in order to simplify the comparative analysis of epigenetic factors enrichment at maintained and lost TAD borders.

2.3.7. Enrichment patterns are distinct for maintained and lost TAD borders

Architectural proteins. We considered maintained and lost TAD borders that were present in both BEAF-32 knockdown and Cp190 and Chro double knockdown mutants. Selected borders demonstrated enrichment of at least one of the insulation proteins BEAF-32, Cp190 and Chro in accordance with our expectations as we left only direct targets (Figure 2.9). BEAF-32 was found at half of the maintained borders while Cp190 and Chro were present in approximately all borders. This finding supports the idea of

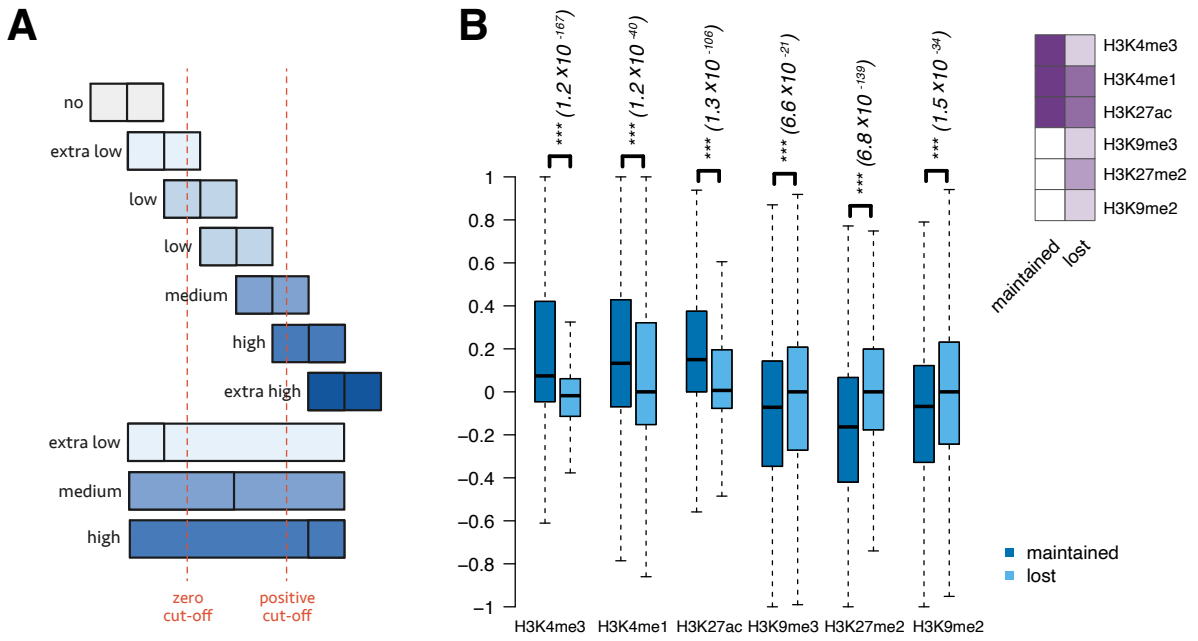


Figure 2.8. ChiP occupancy clustering algorithm. **A.** Schematic representation of relative allocation of average ChIP occupancy signal within 5 Kb window extracted at each TAD border in different enrichment clusters as described in the main text. **B.** Boxplots of average ChIP occupancy signal of selected histone modifications extracted within 5 Kb window at maintained and lost TAD borders. Heatmap represents the clustering plot produced by algorithm. Darker color indicate higher ChIP enrichment. We performed the Mann-Whitney U test. We denoted p-values as: n.s. ≥ 0.05, * p-value < 0.05, ** < 0.01 and *** < 0.001.

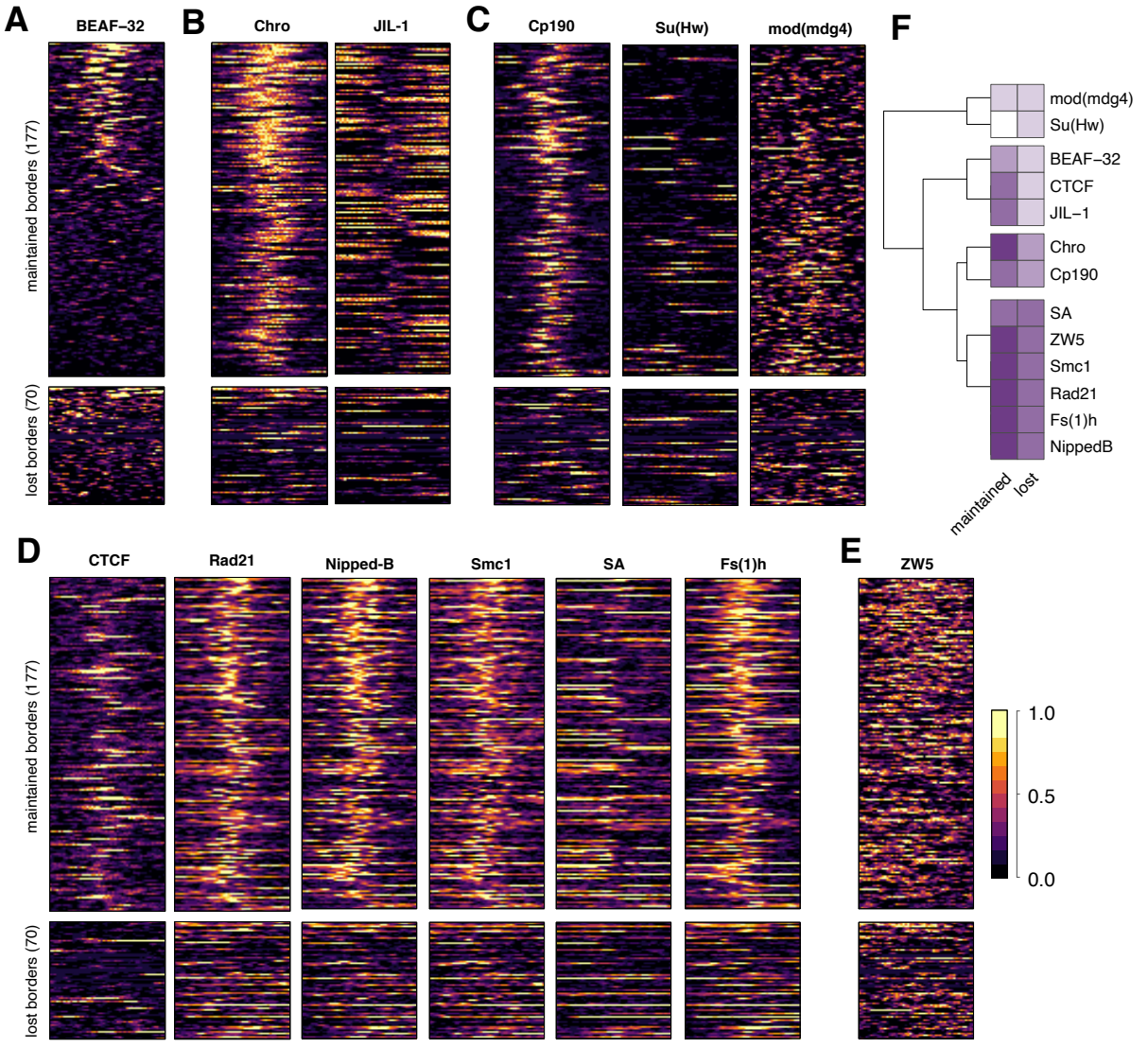


Figure 2.9. Architectural proteins enriched at TAD borders. **A-E.** Profiles of architectural proteins around direct maintained and lost TAD borders that were common in BEAF-32 knockdown and Cp190 Chro double knockdown. Profiles represent 5 Kb region around selected borders. **F.** Clustering of the signal at direct maintained and lost TAD borders as described in the main text.

another recruiter protein for Cp190 and/or Chro to maintain chromatin topology. We did not observe connection between Cp190 and JIL-1 binding the same way as they colocolise at interbands regions during interphase (Rath et al. 2006), however, JIL-1 showed the depletion strictly at maintained TAD borders while it was present within the 5 Kb window. In (Wang et al. 2001), it was proposed that JIL-1 activity establish more open chromatin to facilitate gene transcription (Figure 2.9.B). Cp190 also did not show the strong colocolisation with Su(Hw) and mod(mdg4) as in *gypsy* insulator complex - only a weak connection between Cp190 and mod(mdg4) at the maintained borders depleted of BEAF-32 (Figure 2.9.C). At the same time, protein mod(mdg4) can directly bind DNA, recruit Su(Hw) and mask its repression activity (Melnikova et al. 2003). So, absence of Su(Hw) and presence of JIL-1 indicate about the possible association between active chromatin state and maintenance of the borders when BEAF-32, Cp190 and/or Chro are removed. Within the lost borders, BEAF-32, Cp190 and Chro were also enriched in wild-type cells but did not show strong colocalisation at TAD borders and were allocated around the borders. Also, we did not observe significant enrichment of JIL-1, Su(Hw) or mod(mdg4).

We observe the high CTCF enrichment at some of the maintained borders (approximately half of them) while Cohesin subunits Rad21, Nipped-B and Smc1 are highly enriched at the majority of maintained borders (Figure 2.9.D). It possibly reflects the situation when Cohesin is loaded on chromatin which is stressed due to active transcription, but, in absence of CTCF that stops the loop extrusion, Cohesin dissociates and the loop anchor fragments would break the contact. As a result, the amount of peaked TADs are expected to be less than non-peaked ones, in consistent with previously published reports (Matthews and White 2019; Rowley et al. 2019). However, we are limited in making a proper conclusion on this hypothesis as we require the orientation of CTCF to understand whether Cohesin stops at maintained borders or not. Interestingly, another Cohesin subunit SA also shows a high level of enrichment within 5Kb window but the signal is not centered as for other subunits. Regions where we observe the enrichment of SA seems to be enriched with Rad21, Nipped-B and Smc1

as well. The difference in patterns possibly can be explained by the presence of active promoters and enhancers at maintained borders: it has been previously shown that SA does not occupy most of the active promoters, in contrast with Rad21, Nipped-B and Smc1, but tends to occupy enhancers together with them (Pherson et al. 2019). It is also supported by the enrichment of Fs(1)h which facilitates Nipped-B and Rad21 association with enhancers (Pherson et al. 2019). In addition, the enrichment profile of insulator protein ZW5 does not reveal any specific patters (Figure 2.9.E). ZW5 was previously shown to interact with BEAF-32 (Blanton et al, 2003) and colocolise with Cohesin (Zolotarev et al, 2016). As for BEAF-32, Cp190 and Chro, lost borders are enriched with other architectural proteins from low to high level but without proper colocalisation at the centre (Figure 2.9.F).

At maintained borders, we observed the intensive colocalisation of proteins which form the Cohesin complex (Rad21, Nipped-B, Smc1, SA) along with Fs(1)h which facilitates the loading of Cohesin sub-units to enhancer regions. Strong association with enhancers indicated the role of enhancer-promoter contacts and chromatin looping in TAD formation.

Transcription and replication. Chromatin supercoiling caused by torsional stress accompanying Pol-II transcription activity is a potential driver for loop extrusion and TAD formation. Maintained borders have a high enrichment of Pol-II indicating the association with active state (Figure 2.10.A). As well, MED1 and MED30 show strong colocalisation at the centre of the profiles for maintained borders which also indicates active transcription initiation. In addition, Orc2, which is essential for the initiation of DNA replication, seems to occupy the centre of the profiles as well, but the difference in occupancy strength is not significantly different from the lost borders. We also observe enrichment of Topo-II that surrounds TAD border regions but does not bind them (Figure 2.10.A). Topo-II is a Type 2 topoisomerase that relaxes torsional stress allowing intersegmental chromatin passages. In HeLa cells, it was shown that Type 2 Beta topoisomerase interacts with CTCF and Cohesin at TAD borders (Uuskula-Reimand et al. 2016) and we observe the same pattern here. These results indicates the potential

role of supercoiling and active transcription at maintained borders.

We also notice moderate binding of GAF which is not strictly centered (Figure 2.10.B). GAF (also known as GAGA factor) is a pioneer sequence-specific binding factor that recognise (GA) repeated elements, binds introns and participate in transcription elongation (van Steensel et al. 2003). In addition, GAF can recruit components of chromatin-remodeling complexes such as NURF and ISWI to drive DNA opening (Chetverina et al. 2021), however, not necessarily, as shown by (Tang et al. 2021) on live *Drosophila* hemocytes. Note that in our case, binding of NURF301 and ISWI seems stronger in both maintained and lost borders in comparison with GAF (Figure 2.12.A). Interestingly, joint binding of GAF and Pol-II can indicate the presence of Pol-II pausing (Chetverina et al. 2021). Thus, mutations in GAF sequencing studied in (Lee et al. 1992) led to a reduction in Pol-II pausing. However, the Pol-II pausing index computed at maintained, lost and new borders confirmed only negligible differences in Pol-II pausing (Chathoth et al. 2022).

The low level of enrichment of histones (H1, H2Av, H3 and H4) at maintained borders indicates the highly accessible DNA while slightly higher signal at lost borders indicates less accessible DNA (Figure 2.10.B). Surprisingly, H2Av shows the moderate presence around maintained borders without binding them. The histone H2A variant was shown to be involved in euchromatic silencing and formation of heterochromatin (Swaminathan et al. 2005). All together, the histones allocation around TAD borders indicates open chromatin at the maintained borders which is surrounded by dense chromatin.

The normalised RNA-seq signal around both maintained and lost borders did not demonstrate noticeable changes in two knockdowns (Figure 2.10.C). In case of maintained borders this result was expected as we did not observe significant changes in gene expression associated with conserved borders (Figure 2.3 and 2.6). In case of lost borders, we might expect to notice some changes. However, as differentially expressed genes did not demonstrate specific preference in position within reorganised TADs (Figure 2.5), loss of TAD border can correlate with changes in gene expression

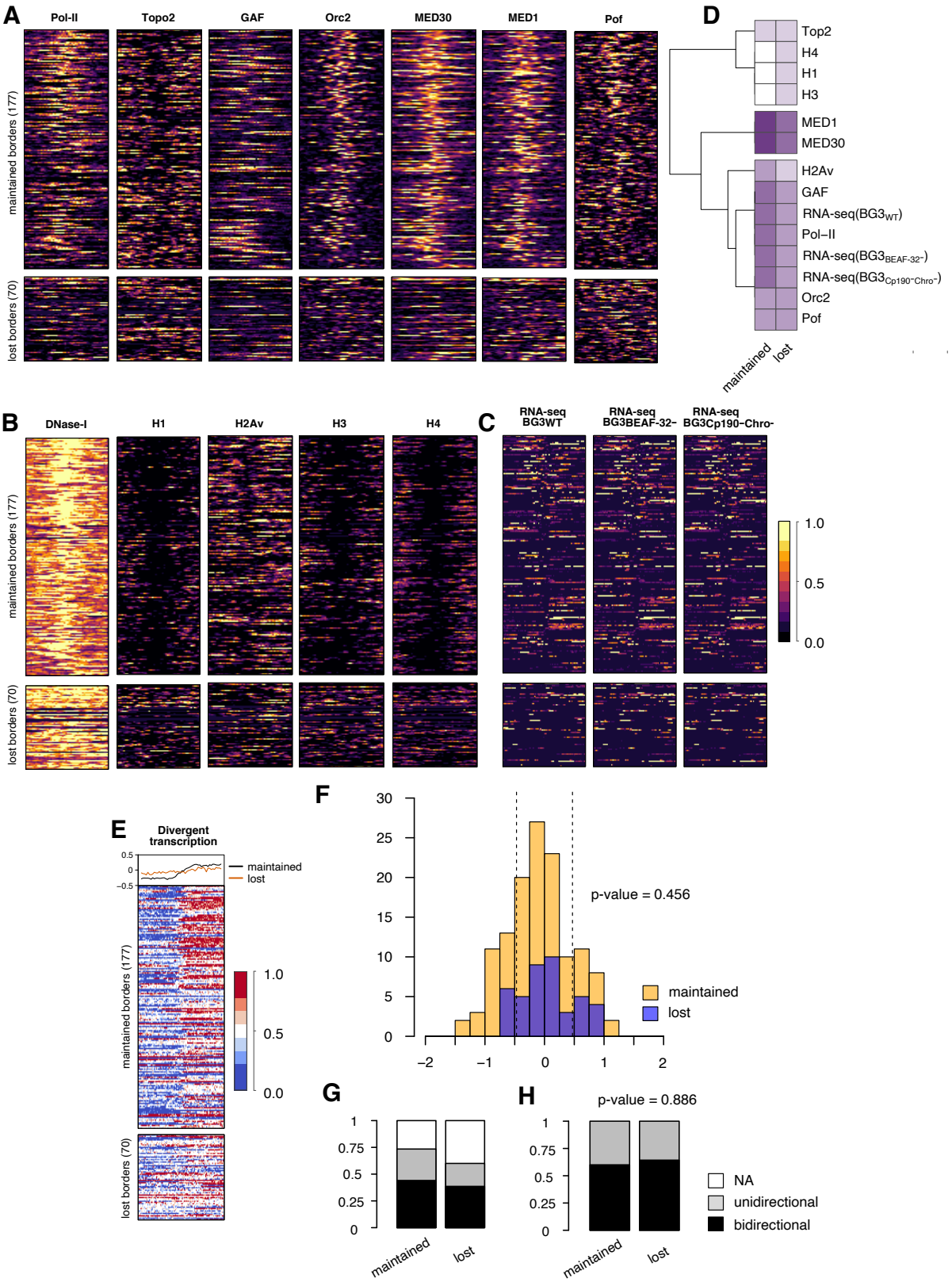

A
Pol-II
Topo2
GAF
Orc2
MED30
MED1
Pof
maintained borders (177)
lost borders (70)
D
Top2
H4
H1
H3
MED1
MED30
H2Av
GAF
RNA-seq(BG3WT)
Pol–II
RNA-seq(BG3BEAF-32-)
RNA-seq(BG3Cp190-Chro-)
Orc2
Pof
maintained
lost
B
DNase-I
H1
H2Av
H3
H4
C
RNA-seq BG3WT
RNA-seq BG3BEAF-32-
RNA-seq BG3Cp190-Chro-
1.0
0.5
0.0
E
Divergent transcription
maintained
lost
F
p-value = 0.456
G
H
p-value = 0.886
NA
unidirectional
bidirectional

Figure 2.10. Transcription and replication associated factors enriched at TAD borders. **A-C.** Profiles of architectural proteins around direct maintained and lost TAD borders that were common in BEAF-32 knockdown and Cp190 Chro double knockdown. Profiles represent 5 Kb region around selected borders. **D.** Clustering of the signal at direct maintained and lost TAD borders as described in the main text. **E.** The nascent RNA profiles. The signal represent the log10 of average signal retrieved from positive over negative strand. The red color represents the transcription occured mostly on positive strand, the blue color represents the transcription occured mostly on negative strand. **F.** The distribution of directionality scores computed as described in the main text. Orange color represents the scores extracted at maintained borders, blue color represents the score extracted at lost border. We performed a Mann-Whitney U test to confirm the difference in distributions. **G.** The distribution of bidirectional, unidirectional borders and borders with no transcription. **H.** The proportion of unidirectional and bidirectional borders excluding the borders with no transcription indicated. We performed a Fisher's exact test and corresponding p-value is specified.

at a larger distance from this border.

Depending on where the nascent RNA is found - either on the positive or negative strand - we can conclude whether we detect bidirectional transcription or unidirectional one. In the case of bidirectional transcription, transcription machinery binds the genomic region but it can go downstream or upstream, while in unidirectional transcription the machinery moves in single direction. We analysed the distribution of sifnals computed as log10 of amount of nascent RNA retrieved from positive over negative strand. Visually, the maintained borders seem to associate mostly with divergent transcription while lost borders are more unidirectional (Figure 2.10.E). The line plots representing the averaged signal within maintained and lost borders also confirms the association of upstream and downstream regions from the maintained TAD borders with transcription on positive and negative strands, respectively.

We also computed the directionality scores (Figure 2.10.F). We computed the mean nascent RNA levels considering 500 bp window that were 500 bp downstream away on the positive strand and 500 bp upstream away on the negative strand. Then, we computed the directionality score as log10 ratio of mean nascent RNA on the posi-

tive over negative strand. Borders with directionality score being lower than 0.47 were classified as bidirectional (Chathoth and Zabet 2019). The value of 0.47 represents slightly less than three times more transcription on positive strand than on negative strand. Large share of lost borders did not show presence of any transcription, this share was slightly less for maintained borders (Figure 2.10.G). However, the proportion of bidirectional borders seems slightly higher than share of unidirectional ones with no significant difference between maintained and lost borders (a Fisher's exact test p-value > 0.05)(Figure 2.10.H). The inconsistency between visual inspection and statistical test can be associated with the relatively small amount of lost borders - we detected only 24 out of 70 lost borders being unidirectional.

Active transcription seems to be a key player in the maintenance and formation of TADs (Ulianov et al. 2016; Li et al. 2015; Rowley et al. 2017). In consistence with the past research, we noticed that borders that survived BEAF-32, Cp190 and Chro depletion were closely accompanied by the binding of other insulator proteins, transcription factors, as well as open chromatin, while lost borders either were depleted with the analysed factors or were from low to moderately enriched but not strongly colocolised at the centre of the 5 Kb window (Figure 2.9 and 2.13). Given these observations, we suspect that active chromatin state can possibly maintain the chromatin topology after removal of important insulator proteins. In contrast, the lost borders did not have the same association with transcription, so depletion of insulators can reduce the segregation strength between kilobase-scale chromatin domains.

Histone modifications. We can get a better understanding of the functional difference between maintained and lost borders looking at the presence or absence of specific histone modifications (Figure 2.11). According to the clustering (Figure 2.11.D), we have modifications that demonstrate: (1) high enrichment in both maintained and lost borders (H4K8ac, H3K18ac, H3K27me1, H3K79me2, H3K4me2, H3K27ac, H3K4me1); (2) high enrichment at maintained but significantly reduced enrichment at lost borders (H3K79me3, H3K4me3, H3K79me1, H2Bubi, H3K36me3); (3) moderate and low enrichment at both maintained and lost borders (H3K36me1, H4K16ac, H4K20me1); (4)

low or full depletion at maintained borders while the lost ones start to demonstrate low enrichment (H3K23ac, H3K27me2, H3K9me2, H3K9me3); (5) both maintained and lost borders are depleted with H3K27me3. However, histone modifications are not found exactly at TAD borders and they occupy the surrounding regions.

The histone modification within the first three groups are mostly associated with active chromatin state, so indicating the euchromatic regions. Histone modifications H4K8ac, H3K18ac, H3K4me3, H3K27ac mark transcription start sites, so can be classified as promoter marks (Wang et al. 2008; Yang et al. 2012; Dong et al. 2012; Taberley et al. 2016). Histone modification H3K4me2 marks active enhancers. Actually, histone H3 lysine K4 methylation (H3K4me1, H3K4me2, H3K4me3) is generally associated with euchromatin and ongoing gene expression (Pekowska et al. 2011). Then, H3K27ac can mark different states of enhancer regions depending on colocalisation with H3K4me1: when both modifications are present, we observe a signal from active enhancer; when H3K4me1 is present but H3K27ac is not we observe a signal from primed enhancer (inactive enhancer that is primed for future activation) (Calo and Wysocka 2013). Another two modifications, H3K79me1 and H3K79me2, did not show any specific preference for either active or silent genes so possibly indicating its bivalent function depending on presence or absence of specific factors (Steger et al. 2008). Interestingly, the modification H3K18ac which were found to be enriched at enhancers (Wolfe et al. 2021) seem to show moderate enrichment at both maintained and lost borders.

In contrast, histone modifications mostly depleted in maintained borders are associated with inactive, heterochromatic regions. H3K27me2 was involved in enhancer silencing which prevents its transcriptional activity (Ferrari et al. 2014; Lee et al. 2015). Enrichment of H3K27me2 together with H3K9me2 and H3K9me3 at lost borders indicates the association with inactive heterochromatin. We also notice that H3K27me3 was approximatley everywhere depleted. H3K27me3 is a Polycomb mark: it allows the loading of Polycomb repressive complex PRC1 which includes Pc and dRING proteins, so leading to chromatin compaction and transcription pausing (Min et al. 2003;

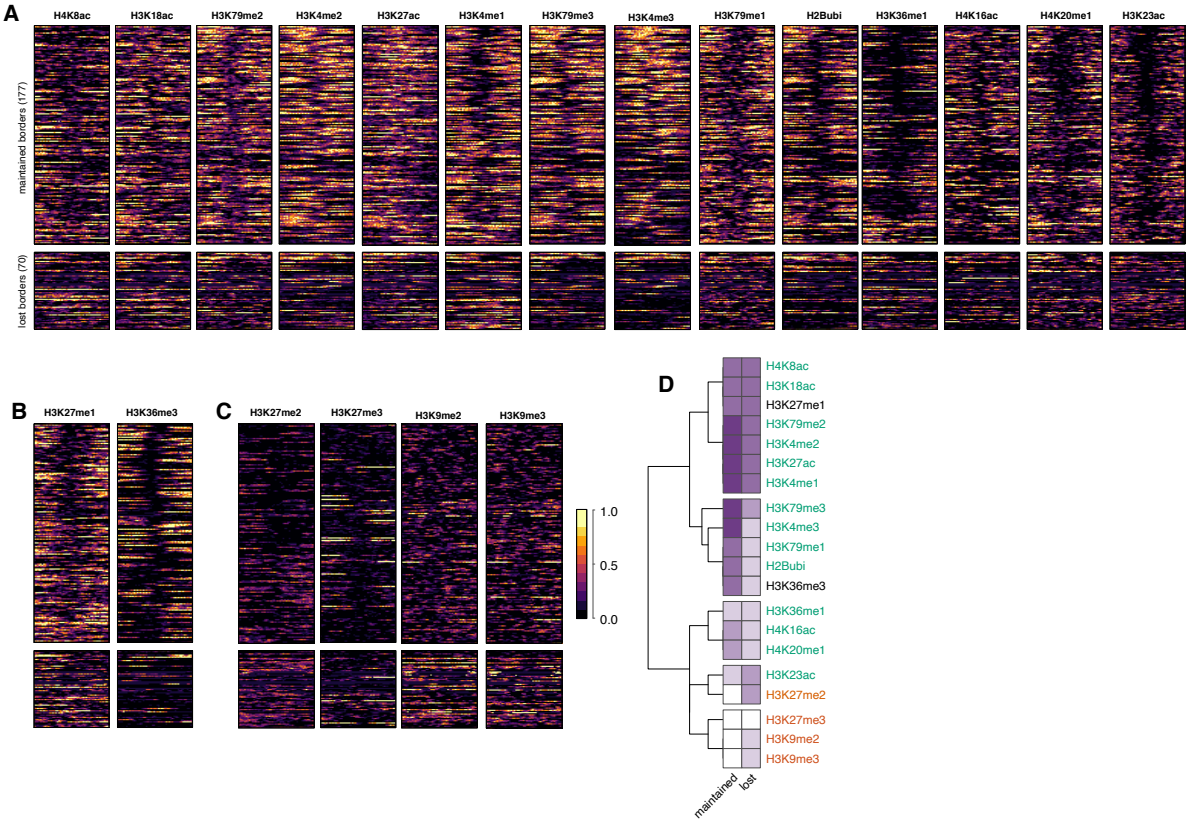


Figure 2.11. Histone modifications enriched at TAD borders. **A-E.** Profiles of architectural proteins around direct maintained and lost TAD borders that were common in BEAF-32 knockdown and Cp190 Chro double knockdown. Profiles represent 5 Kb region around selected borders. **F.** Clustering of the signal at direct maintained and lost TAD borders as described in the main text.

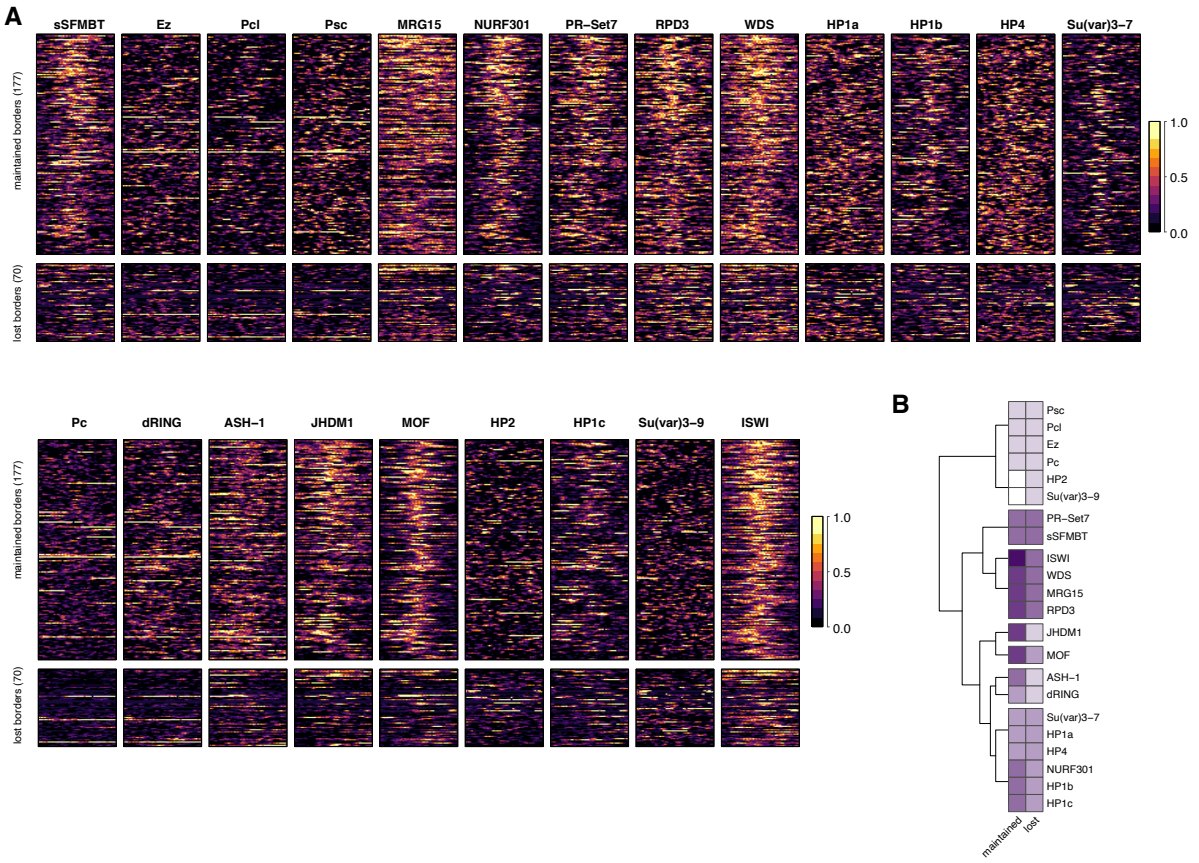


Figure 2.12. Remodelling, heterochromatin and Polycomb-associated factors enriched at TAD borders. **A-E.** Profiles of architectural proteins around direct maintained and lost TAD borders that were common in BEAF-32 knockdown and Cp190 Chro double knockdown. Profiles represent 5 Kb region around selected borders. **F.** Clustering of the signal at direct maintained and lost TAD borders as described in the main text.

Lehmann et al. 2012). We do not observe enrichment of H3K27me3 as well as Pc and dRING at both maintained and lost borders.

Lost borders are highly enriched with repressive histone modifications as well as they are enriched with histones. Altogether, they are signature of silent chromatin: transcription there is stopped by either dense heterochromatin which does not allow transcription factors to bind and initiate the transcription, or by silencing of enhancers which also blocks the transcriptional activity. As heterochromatin prevents the transcription, in absence of architectural proteins without the support of transcriptional machinery the chromatin loses its conformation and TAD borders are lost. Histones are also found not at maintained TAD borders, but they are enriched around, forming dense chromatin around open and actively transcribed euchromatin “islands”. Absence of histones at maintained borders specifically is suggested to allow the transcription factors to bind and initiate transcription there, supporting TAD formation through the enhancer-associated contacts.

Remodelling, heterochromatin and Polycomb. To further support the connection of lost borders with heterochromatin, we looked at some nucleosome remodelling factors, heterochromatic and Polycomb marks (Figure 2.12). Lost borders were found to be slightly enriched with heterochromatic marks HP2 and Su(vaw)3-9, however, we do not obseve any noticable presence of Polycomb marks like Pc and dRING which form Polycomb repression complex (Figure 2.12.B). Note that the protein Su(var)3-9 was previously reported to have a role in maintenance of TADs allocated at heterochromatin (Saha et al. 2020).

Summing up, maintained TAD borders demonstrated strong correlation with factors which were found to actively participate in transcription, so showing allocation within euchromatic chromatin. Lost borders did not demonstrate the same tendency, either showing the presence of active marks randomly allocated within 5 Kb window around borders, or showing presence of heterochromatin marks which are slightly/moderately higher than at maintained borders. The existence of two classes of TAD borders in *Drosophila*, active and repressed borders, display different mechanisms of their main-

tenance. A similar classification into active and repressed domains in *Drosophila* has been previously proposed (Ogiyama et al. 2018; Ramirez et al. 2018; Szabo et al. 2018). On the other hand, there were some research reported association of TADs mainly with euchromatin (Sexton et al. 2012; Ulianov et al. 2016; Hug et al. 2017). Possible reason why we observe heterochromatic lost borders and not euchromatic borders is that we used 5 Kb window around the borders instead of single genomic position. The BG3 chromatin state analysis, which was performed in (Chathoth et al. 2022) on single genomic positions where maintianed, lost and new TAD borders were allocated, confirmed the enrichment of borders in enhancer and active TSS states, as well as the depletion in heterochromatin. In contrast, assuming 5 Kb window instead of single position, we observe the enrichment of Polycomb and heterochromatin in euchromatin state at lost TAD borders, consistent with ChIP occupancy analysis. Given these findings, the borders that were lost in both BEAF-32 KD and Cp190 Chro KD can be defined as the ones allocated in euchromatic islands surrounded by heterochromatin.

2.4. Summary and discussion

The colocolisation of architectural proteins at borders of topologically associated domains (TADs) raises the question of whether they have a functional role in establishment and maintenance of chromatin topology. For example, the loop extrusion coupled with Cohesin-CTCF interplay was proposed to conduct the TAD formation in mammalian systems (Zuin et al. 2014; Racko et al. 2018; Nora et al. 2017). In *Drosophila*, there are several architectural proteins which previously showed the accumulation at TAD borders. These proteins include BEAF-32, Cp190 and Chro. In this Chapter, we access the TAD reorganisations caused by the knockdowns of these proteins plus Dref which shares similar binding motif with BEAF-32, so can potentially replace it at TAD borders.

We identified between 600 and 800 genes which altered their expression in the result of the knockdowns. The majority of those genes were located within TADs

which have significantly shifted the position of one or both borders. The differentially expressed genes were associated with these rearrangements more often than by chance which indicates the that strong TAD reorganisations are coupled with significant changes in gene expression. However, this association was not detected at Cp190 and Chro double knockdown. Cp190 and Chro are the proteins that cannot bind DNA independently and they are recruited by other proteins. If proteins are not involved in TAD border organisation, the depletion of Cp190 and/or Chro would not affect the gene expression through TAD reorganisation, so the association of DEGs and changes in TAD allocation would be reduced.

The protein depletion resulted in perturbation of TAD borders. We found that borders that were lost during the knockdown were mostly associated with silenced regions of the genome: they displayed the moderate enrichment of heterochromatin and the depletion of active histone modifications. The borders that were maintained upon the knockdowns, in opposite, were mostly associated with euchromatin and active transcription. These borders were also enriched in Cohesin, CTCF, Mediator complexes and Trithorax-group (Fs(1)h, NURF301, ISWI, mod(mdg4), ASH-1, GAF). So, these proteins were possibly able to compensate the loss of BEAF-32, Cp190 and Chro to maintain the TAD topology. On the other hand, the knockdown removed the significant share of proteins but not in full. Therefore, we observed that maintained borders retained the some binding of these architectural proteins upon the depletion.

Depletion of BEAF-32, Cp190 and Chro seems to affect both genome-wide chromatin architecture and gene expression patterns. Genomic regions which maintained the chromatin topology even if they were direct targets of the depletion of proteins-insulators demonstrated the association with enhancers, promoters, active transcription, and open chromatin. The regions which were lost in the knockdowns, in opposite, were found to be associated with heterochromatin, so were not supported by the transcription machinery and were not able to maintain the topology in absence of the BEAF-32, Cp190 or Chro. We can suggest that, based on this finding, the chromatin conformation is established by the binding of architectural proteins and transcription,

which further can maintain the TADs even in absence of proteins. However, without the transcription in place, the TAD borders cannot be maintained in absence of the proteins. To have more support for this hypothesis, we can be interested to look at occupancy of the proteins and transcription factors in the knockdowns as it would shed the light whether the transcription is a necessary factor to maintain the TAD borders or there is a complex interactions between BEAF-32, Cp190, Chro and other proteins which can takes place. Moreover, even if we observe the interplay between insulator proteins and transcription at the maintained TAD borders, it is not clear whether these proteins form the chromatin architecture which functions as a “guide”for transcription and ensures the proper gene regulation, which, in turn, can maintain the topology even if the proteins are depleted. The dynamical changes in the topology and protein occupancy can provide enough support or, in opposite, contradict with the described hypothesis, but would shed the light on the important and the widely discussed question in the field of chromatin architecture – whether the transcription maintains the chromatin topology or chromatin topology ensures the proper gene regulation.

While the depletion of the proteins affected the TAD organisation, the A/B compartmentalisation seemed mostly unaffected (Chathoth et al. 2022). Approximately half of the chromatin was affiliated to the A compartment and the remaining half belonged to the B compartment for all datasets. Knockdowns did not lead to dramatic switches between compartments: only 4-5% of active chromatin were silenced and only 4-5% of inactive chromatin were activated. However, we found that most of the maintained borders are localised in A compartments (acive chromatin) while the lost and new borders are localised in B compartment (inactive chromatin). This finding is consistent with the result of the ChIP occupancy analysis in a way that the most of the lost borders were found in silent chromatin while the maintained borders were found in active chromatin. In addition, we noticed slightly enhanced A-A interactions in wild-type that possibly indicated more intense interactions between active regions and these interactions were slightly diminished in knockdowns. However, the ratio of homotypic A-A and B-B interactions to heterotypic A-B and B-A interactions were similar with only a small decrease

for BEAF-32 KD. Altogether, our results indicated that BEAF-32, Cp190, Chro and Dref had little effect on the organisation of compartments in *Drosophila*.

Chapter 3. Statistical framework for complex chromatin organisation analysis

3.1. Introduction

Recent studies on the 3D chromatin organisation relied on the computational methods that called topologically associated domains (TADs) in a “head-to-tail”, non-hierarchical manner (Dixon et al. 2012, Crane et al. 2015, Ramirez et al. 2015, Kruze et al. 2020). Researchers started to overcome the limitations of widely used TAD callers, so TADs were demonstrated to have more complex, hierarchical folding, i.e. large TADs were combined from nested sub-TADs organisation. It was also previously shown that many TADs did not display a sharp separation, so we observe a more like smooth transition from one TAD to the neighbouring one. (Chang et al. 2020) (Figure 3.1). However, the exact mechanism is unclear. For example, in mouse embryonic stem cells the absence of punctuated TAD boundaries was associated with clusters of CTCF binding peaks within fuzzy TAD boundaries which additively contribute to the domain insulation strength (Chang et al. 2021).

Overall, we face two possible scenarios for these “transition zones”. On the one hand, we observe just the artefacts of the cell-to-cell variations, dynamic architectural reorganisations, experimental noise or low sequencing depth. On the other hand, “transition zones” can reflects the chromatin regions that either belong to both TADs (partial overlap) or do not belong to TADs at all (TAD breaks) (Figure 3.1). These suggestions leads us to the idea that we possibly can face more complex topology and we need new TAD detection methods.

In previous Chapters, we discussed the background and previous research associated with 3D chromatin organisation. We also analysed the *Drosophila* Hi-C data and revealed the functional roles of specific architectural proteins found in flies. *Drosophila* is the one of the commonly used model organisms and the relatively small size of the genome makes it possible to study the chromatin organisation with high resolution Hi-C maps. For the 3D chromatin conformation analysis, we looked at the consequences

associated with the depletion of the named proteins on TADs level. For TAD analysis we used the HiCExplorer tool that considers the insulation score for finding TADs. However, visual inspection of some example genomic regions reveal the presence of complex chromatin pattern which were not detected by the tool.

In this Chapter, we provide the statistical framework that leads us to the new TAD calling algorithm named **COrTADo** (**C**omplex **Or**ganisation of **T**opologically **A**ssociated **Do**mains). Using COrTADo, we can call start and end TAD border positions separately that can serve as a basis for further reconstruction of complex chromatin architecture.

3.2. Materials and methods

Data. We used the Hi-C dataset generated by (Chathoth and Zabet 2019) in *Drosophila melanogaster* wild-type BG3 cells at DpnII resolution. We used the matrix without the

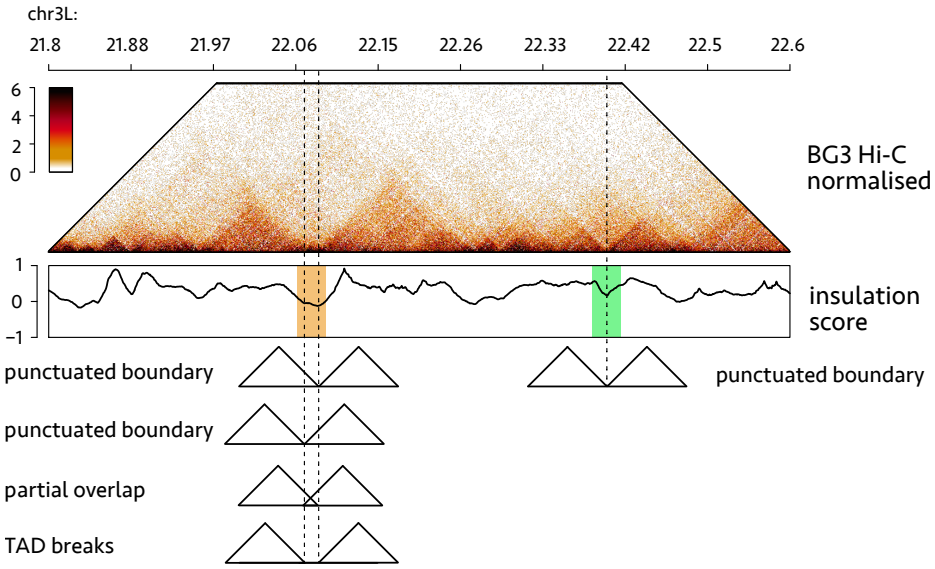


Figure 3.1. Examples of smooth transition between neighbouring TADs (left, color with orange) and visually clear punctuated TAD boundary (right, color with green). Top panel represents the heat map of the *Drosophila* BG3 cells at the genomic region chr3L: 21.8 - 22.6 Mb (Chathoth and Zabet 2019). The line plot represents the insulation score computed using HiCExplorer (Ramirez et al. 2018), the local minima indicate the regions with high insulation between neighbouring genomic regions and most probably allocation of TAD borders. For fuzzy boundary (orange), the TAD borders can be placed either as a single position, partial overlap or separated by the break.

correction method which is usually recommended before TAD calling (see Chapter 1 for more details). We decided to work with raw Hi-C interaction matrix while exploring the statistical framework to call complex architectural structures as we aim to compare the performance of the method on raw and corrected Hi-C dataset further in the Chapter.

HiCExplorer. We called TADs using HiCExplorer with parameters similar to (Chathoth et al. 2021) using minimum width at 5 Kb, FDR correction for multiple testing, p-value threshold of 0.01 and minimum threshold of the difference between insulation score of 0.04 with the only difference being that we used the raw matrix instead of being corrected with KR balancing method. We used these TAD borders for checking the hypotheses which we propose and explore in this Chapter. We suggest that HiCExplorer as any other TAD calling tool has its limitations that do not allow to catch all visual architectural patterns or suggested complex chromatin structures like breaks between TADs, nested TADs or partially overlapping TADs. However, together with previously published research we demonstrated that HiCExplorer is able to call structures that correlate with chromatin epigenetic mechanisms, so can be called trustful for hypotheses testing.

3.3. Statistical framework

3.3.1. Hi-C interaction frequency significantly changes when crossing TAD edge

In Hi-C, we generate the matrix where each matrix entry $X_{(i,j)}$ represents the number of interactions between the DNA locus i and DNA locus j. Suppose, we focus on a single row locus i. All entries before the diagonal with index numbers 1, 2, ..., $i-1$ represent loci allocated upstream from locus i. All entries after the diagonal with index numbers $i+1$, $i+2$, ..., $n-1$, n represent loci allocated downstream from locus i (Figure 3.2.A). During Hi-C experiment, the fixed locus i can be cross-linked and ligated only once with other locus j. However, the diploid genomes, like *Drosophila*, mouse or human, have two copies of the DNA. It means that within the population of T cells, locus i can participate in up to $2T$ pair-wise interactions (Figure 3.2.B).

Topologically associated domains (TADs) are visually detectable regions in Hi-C

contact map. They are represented in the form of consecutive squares allocated along the Hi-C diagonal where intra-TAD interactions happen more frequently than inter-TAD interactions. When loci i and j are allocated within the same TAD, they interact more frequently than in case when i and j belong to different TADs. Formally, if we have loci i and j from the same TAD and any locus j^* outside this TAD, we expect that $E\left(X_{(i,j)}\right) > E\left(X_{(i,j^*)}\right)$. Suppose that any locus j starting from the position j_s up to the position j_e belongs to the same TAD with locus i (k_s and k_e represent start and end of the TAD, respectively). Then, for $\forall i \in [k_s; k_e], \forall j \in [k_s; k_e]$ and $\forall j^* \notin [k_s; k_e]$, we expect $E\left(X_{(i,j)}\right) > E\left(X_{(i,j^*)}\right)$. The positions k_s and k_e are the same for all pairs of loci within the same TAD. For Hi-C interaction map, inside-TAD interactions form the square with corners at $(k_s; k_s)$ (top left corner), $(k_s; k_e)$ (bottom left corner), $(k_e; k_e)$ (bottom right corner), $(k_e; k_s)$ (top right corner) (Figure 3.2.A). Visually, inside-TAD area separated from outside-TAD area by four square edges. In particular, the column positions of the left and right edges represent the TAD start position k_s and end position k_e, respectively. For the aim of the analysis, at this and further Sections we focus on left TAD edge (TAD start) and, as the Hi-C matrix is symmetric, only on the lower triangle of the matrix. Next, we formally introduce the term left (start) TAD edge k:

For any locus i, the **start TAD edge** is a locus $k \leq i-1$ such that for $\forall j \in [k; i-1]$ and $\forall j^* \leq k-1$, we have $E\left(X_{(i,j)}\right) > E\left(X_{(i,j^*)}\right)$.

Formally, TAD edge definition and canonical TAD border definition are the same – it is the genomic position that separates the intra-TAD interacting DNA fragments from the fragments which do not belong to the same TAD. We introduced the different definition to show that we aim to detect the genomic position that is the same for all neighbouring Hi-C rows belonging to the same TAD, technically, detecting a line segment on the Hi-C contact map which separated the inside-TAD interaction frequencies from outside-TAD ones and not a single coordinate that separates two neighbouring TADs. Also note, that further in the method development, we introduce the procedure to estimate the TAD edge length which measures the number of consecutive loci i

which has the start TAD edge at the same locus k. Technically, we aim to measure the length of the line segment of the Hi-C contact map which represents the edge of the TAD. So, more geometrically related term “TAD edge” suits the algorithm better than the term “TAD border”.

We do not rely on genomic distances in this definition, only on the coordinates (i, j) of the Hi-C matrix entries. So, we base all derivations only on the position of locus j relative to locus i, independently on the Hi-C resolution. It reflects the situation when Hi-C experiment and further pre-processing generates loci with the same genomic lengths. It is common situation for large genomes like human or mouse. For organisms like *Drosophila* we can work efficiently (in terms of computational memory and time) with high resolution maps, for example, generated with DpnII restriction enzyme. However, the DNA fragments then would have non-homogeneous bin sizes.

According to the definition of the TAD, loci within the same TAD on average interact more frequently than loci that do not belong to the same TAD. Then, the selected locus i is expected to interact less frequently with fragments that are allocated upstream from the left TAD edge, i.e. with fragments $j = 1$, ..., $k - 1$, than with fragments that are allocated downstream from the left TAD edge, i.e. with fragments $j = k$, ..., $i - 1$ (Figure 3.2.A). If the expected number of interactions between locus i and locus j is $\mu_{(i,j)}$ then we have:

$$(\mu_{(i,1)}, \mu_{(i,2)}, ..., \mu_{(i,k-1)}) < (\mu_{(i,k)}, \mu_{(i,k+1)}, ..., \mu_{(i,i-1)})$$

The mean is expected to be low before reaching the left TAD edge and is expected to be high after crossing the left TAD edge. Any position k within the Hi-C row i, that demonstrates all mean estimates on the left-hand side, i.e. $(\mu_{(i,1)}, ..., \mu_{(i,k-1)})$ being less that all mean estimates on the right-hand side, i.e. $(\mu_{(i,k)}, ..., \mu_{(i,i-1)})$, is a candidate for being a left TAD edge position. For Hi-C data, the DNA fragments that are allocated close to each other on chromatin strand should interact more frequently than distal DNA fragments. So, whether TADs are present or not, the intensive interactions happen when the Hi-C matrix entry is close to diagonal. Technically, within the Hi-C row, mean estimates smoothly increase when we move closer to the diagonal, so,

all of the left-hand side means automatically are less than all of the right-hand side means. Then, we need to slightly modify the criteria for TAD edge candidate search: the change from $\mu_{(i,k-1)}$ to $\mu_{(i,k)}$ should be sufficiently larger than any other change between neighbouring means if k is a candidate to be a left TAD edge position.

The proposed criteria relies on the estimation of the mean parameters. The main limitation here is the sample size: we work with a single row i of Hi-C matrix from the column $j = 1$ up to the diagonal excluding the diagonal element, i.e. $j = i - 1$. The natural choice for the mean estimation under the described conditions can be a Moving Average (MA): we create a series of subsets of fixed size and compute averages within them. Mathematically speaking, if we have a dataset $(x_{(i,1)}, x_{(i,2)}, ..., x_{(i,i-1)})$, the mean over the first MA with window size w is defined as

$$MA_{(i,1)} = \frac{1}{w}(x_{(i,1)} + x_{(i,2)} + ... + x_{(i,w)})$$

When calculating the next mean MA_2 the range from 2 to $(w + 1)$ is considered:

$$MA_{(i,2)} = \frac{1}{w}(x_{(i,2)} + x_{(i,3)} + ... + x_{(i,w+1)})$$

We continue to slide the MA window while we can create subset of size w. The last mean $MA_{(i-w)}$ for the given dataset can be defined as:

$$MA_{(i,i-w)} = \frac{1}{w}(x_{(i,i-w)} + x_{(i,i-w+1)} + ... + x_{(i,i-1)})$$

We have to keep in mind, that the MA algorithm is sensitive to the window size w. The wider the window the smoother the result: the mean becomes less responsive to the dramatic changes in observations. The shorter windows, in opposite, allow the detection of sharp increases/decreases quicker, but, at the same time, the mean becomes highly disperse for noisy data which is the case for Hi-C experiment (Yaffe and Tanay 2011, Imakaev et al. 2012). For the preliminary analysis, we can motivate MA window size selection by the minimum TAD that can be detected.

We continue to slide the MA window until we reach the last element $(i - 1)$ of the given dataset. It means that the last mean that we are able to estimate is based on the observations from the range $[(i - w); (i - 1)]$. As we cannot compute mean when $j > (i-w)$, we cannot access changes between $\mu_{(i,j-1)}$ to $\mu_{(i,j)}$, and, as a consequence, we cannot place TAD edge candidate when $j > (i - w)$. Summing up, if the TAD edge

position is allocated within the range $[(i - w + 1); (i - 1)]$, it cannot be detected. So, there is the connection between maximum MA window size w and the minimum TAD size that we are able to detect.

The minimum TAD size can be introduced through the minimum TAD width - the length between TAD borders, or, geometrically, the diagonal of the TAD square (Figure 3.2.A). Then, applying the Pythagoras Theorem, the minimum edge length is

$$edge = \frac{\sqrt{2}}{2} width$$

This formula is a good edge estimation when we work with genomic distances. If the minimum TAD width is given in bins, then the minimum edge length is exactly the same as minimum width

$$edge = width$$

The minimum MA window w should be less than the edge length in order to detect the TAD edge. It is important to note here, that selecting w to be the same as minimum edge length allows only the matrix row at the TAD bottom edge to be detected. The reason is that for all above rows belonging to the same TAD, the distance between left TAD edge and diagonal element is shorter than the MA window. As a consequence, for these rows the left TAD edge cannot be detected (Figure 3.2.A).

We aim to formalise the criteria of TAD edge detection. We start from checking whether the stated criteria is consistent with TADs allocated by widely used TAD calling tool in the field of Hi-C analysis - HiCExplorer (Ramirez et al. 2018). The details of the data used and parameters selected to call TADs with HiCExplorer are described in the Section 3.2. Methods and Materials. Note that the allocation of TADs can vary depending on parameters and TAD calling algorithms used. We use here HiCExplorer as a standard for TAD allocation at the selected Hi-C dataset because earlier we demonstrated the connection between HiCExplorer TAD borders and their epigenetic features (see Chapter 2 for more details). We believe that HiCExplorer and other widely used tools have limitations that do not allow to catch complex topological structured that we can visually reveal and we aim to explore. However, we start with genomic positions that we can call TAD borders with high level of confidence.

	Min TAD width		Min TAD edge	
	Kb	Bins	Kb	Bins
Genome-wide	9.366	3	6.623	3
chr3R: 21.3 - 21.7 Mb	28.263	38	19.985	38

Table 3.1. Minimum values of TAD geometry based on TADs called by HiCExporer genome-wide and on test region.

The *Drosophila* BG3 (neuronal cells) line) matrix was built using the DpnII restriction sites, so we obtained the matrix bins with the median width of 575 bp. When we called TADs using HiCExplorer tool, we set the minimum TAD width to be 5 Kb (in order to remove small TADs that can appear because of the noise) that is, in average, 8.7 or, roughly, 9 bins. In fact, the minimum HiCExplorer detected TAD was larger in width than 5 Kb but significantly shorter than 9 bins (Table 3.1). So, we can mark the first limitation - the TAD edge detection algorithm that we are formalising in this Chapter cannot call relatively small TADs that possibly can be called by the alternative TAD calling tools, for example, HiCExplorer. Note that this limitation remains unaddressed in the Thesis.

The genomic region chr3R: 21.3 - 21.7 Mb, that we selected for the testing, contains 8 relatively large TADs (Figure 3.2.F, top panel). As a consequence, the recommended MA window size w can be any number below 38 bins. For simplicity, we can select $w = 10$ when analysing the testing region (multiple of 10, which is below 38 and close to 9 bins). However, for the genome-wide analysis more accurate selection procedure is required, otherwise we face a danger to exclude small TADs which can be relevant for downstream analysis.

We expect to observe the significant increase in estimated mean when we cross the left TAD edge as the interactions are expected to happen more frequently inside a TAD than outside a TAD. Suppose that we want to visually detect this dramatic change looking on the line plot $MA_{(i,j)}$ versus window j of a single Hi-C row i. Unfortunately, due to the large number of NAs and high level of noise in the data, we expect the mean

estimates to be significantly disperse. So, the visually observed dramatic change in mean value cannot be distinguished between real change and spurious change that arises because of fluctuations.

As we already mentioned above, we use the highly reliable TAD border positions (and corresponding TAD edges positions) called using HiCExplorer tool. Within a single Hi-C row, known TAD edge position k can be treated as a zero point (Figure 3.2.C). First, we compute MAs moving to the right-hand side from the position k, i.e.

$$MA_{(i,1)} = \frac{1}{w}(x_{(i,k+1)} + x_{(i,k+2)} + ... + x_{(i,k+w)})$$
$$MA_{(i,2)} = \frac{1}{w}(x_{(i,k+2)} + x_{(i,k+3)} + ... + x_{(i,k+w+1)})$$
$$MA_{(i,3)} = \frac{1}{w}(x_{(i,k+3)} + x_{(i,k+4)} + ... + x_{(i,k+w+2)})$$

and so on. Second, we define MAs moving to the left-hand side from the position k, i.e.

$$MA_{(i,-1)} = \frac{1}{w}(x_{(i,k-1)} + x_{(i,k-2)} + ... + x_{(i,k-w)})$$
$$MA_{(i,-2)} = \frac{1}{w}(x_{(i,k-2)} + x_{(i,k-3)} + ... + x_{(i,k-w-1)})$$
$$MA_{(i,-3)} = \frac{1}{w}(x_{(i,k-3)} + x_{(i,k-4)} + ... + x_{(i,k-w-2)})$$

and so on. Then, visualising all MAs using the line plot for all Hi-C rows we observe the dramatic shift in estimated mean when we cross zero position that corresponds to the TAD edge positions (Figure 3.2.C). This pattern is not noticeable if we select another position to be the zero point. For example, if we start to compute MAs moving to the right-hand side and to the left-hand side from the Hi-C diagonal positions instead of TAD edge positions, there is no significant “jumps” in mean estimates (Figure 3.2.D).

There are several interesting features to note about the mean. The first observation is that, when we center the mean signal at Hi-C diagonal, there is a smooth increase in mean on the left-hand side from the diagonal and more noisy but still smooth decrease on the right-hand side from the diagonal (Figure 3.2.F). This finding is consistent with the expected change in mean that is related to the change in genomic distance between interacting fragments - close fragments tend to interact more frequently.

The second observation is, when we center the mean signal at TAD edge positions, the mean estimates on the right-hand side, i.e. interactions inside TAD, are more scat-

tered than mean estimates on the left-hand side, i.e. interactions before crossing TAD edge (Figure 3.2.C). This observation can be critical in case if we want to make an assumption about the Hi-C interaction distribution. In simple words, we deal here with change-point detection problem: within the Hi-C row we search for the column position where the interaction mean dramatically changes, which indicates the TAD edge position. As the first choice, we can suggest widely used sequential change detection algorithm, in particular, cumulative sum (CUSUM) method developed by (Page 1954). In short, the main idea of the algorithm is as follows. Under the null hypothesis we assume no change happens and assume some distribution (probability density function) for the signal that relies on the mean parameter. Under the alternative hypothesis there is one change happens at some specific moment of time and the mean parameter is shifted. Observations are ordered in time and at each observation we accumulate log-likelihood ratios - the measure that evaluates how good the observation fits the null hypothesis versus the alternative hypothesis. When the accumulated sum beats the pre-selected threshold, we detect the point where the mean parameter is shifted. The critical step here is the probability density function selection. In Hi-C, in addition to the mean changing, we observe also changes in the variance. So, we either need to modify the CUSUM algorithm to deal with shift in two parameters, or introduce the functional relationship between mean and variance that fits the observed Hi-C data. For more details about CUSUM algorithm and its adaptation to TAD detection problem, see Appendix 3.1.

We propose the non-parametric solution that relies on log mean ratio computation. The dramatic change happening between neighbouring MAs is suggested to be a unique feature of the mean estimates allocated close by the TAD edge position. So, within the single Hi-C row the pairs of neighbouring MAs that do not demonstrate the significant change from one to another are expected to belong both to the area outside TAD or to the area inside TAD. Speaking formally, for any position $j = 1, ..., i-1$ we define single $MA_{(i,j+1)}$ on the right-hand side from j as

$$MA_{(i,j+1)} = \frac{1}{w}(x_{(i,j+1)} + x_{(i,j+2)} + ... + x_{(i,j+w)}) \tag{3.1}$$

and we define single MA_{j-1} on the left-hand side from j as

$$MA_{(i,j-1)} = \frac{1}{w}(x_{(i,j-1)} + x_{(i,j-2)} + ... + x_{(i,j-w)}) \quad (3.2)$$

Then, the $log_2(MA_{(i,j+1)}/MA_{(i,j-1)})$ ratio reaches its local maximum only when the position j is a candidate to be a left TAD edge position. Given that there are several Hi-C rows that belong to the same TAD and cross the TAD edge at exactly the same position, the line plot of log2 mean ratio versus locus j position should demonstrate several local peaks coinciding at the same locus j position that are TAD edge candidates. It is not a big surprise that for the selected testing Hi-C region, the TAD edge positions called by HiCExplorer tool nearly coincide with local peaks of $log_2(MA_{(i,j+1)}/MA_{(i,j-1)})$ ratio (Figure 3.2.F). As the local maximum positions perfectly mimics the reliable TAD edge positions, it provides us enough evidence to move to the next step in this study - to adapt the described criteria to become MA-based TAD edge calling tool.

Log2mean seems quite similar with the directionality index (DI) introduced in (Dixon et al. 2012). First, we estimate the average interaction frequency in the vicinity of the DNA fragment upstream and downstream, then we compute the DI which measures the bias towards the downstream or upstream interactions. On the TAD border, when locus belongs to the upstream TAD, it tends to demonstrate bias towards upstream interactions; when locus belongs to the downstream TAD, it tends to demonstrate bias towards downstream interactions. So, at the end of one TAD, DI demonstrates local minimum, and, at the beginning of the next TAD, DI demonstrates local maximum. The same as for DI, log2mean ratio requires the computation of average interaction frequencies in the vicinity of the DNA fragment. However, in log2mean ratio is computed not per single locus but for each pair of loci i and j. In particular, we compute the average interaction frequencies between locus i and loci which are allocated in the vicinity of locus j in downstream and upstream directions. At the start of the TAD, when loci i and j belongs to the same TAD, log2mean ratio demonstrates the preference towards downstream interactions. At the end of the TAD, when loci i and j belongs to the same TAD, log2mean ratio demonstrates the preference towards upstream interactions. To sum up, even if the decision rules about the end and start of the TAD detection are

similar for log2mean ratio and DI, DI is one-dimensional measure while log2mean ratio is two-dimensional measure. So, the local minimum/maximum detection for log2mean ratio requires that the local extrema should be allocated at the same j locus for several neighbouring i loci which all belong to the same TAD. Also note, DI considers only interactions which happen in the vicinity of the selected locus while log2mean ratio considers all pairs of i and j which can be both distal and proximal.

The amount of positive and close to zero log2 mean ratios is prevailing (Figure 3.2.G). However, we also observe a significant amount of negative log2 mean ratios. The negative signal at the position j means that the left-hand side mean estimate $MA_{(i,j-1)}$ is higher than right-hand side mean estimate $MA_{(i,j+1)}$, indicating the "drops" in interaction intensity when we move within Hi-C row in the direction of the diagonal. The values of negative signals are not so high in absolute terms and they do not cross the value of -1: it means that the interaction mean does not fall by more than half of the mean computed one window before. Positive signals, in opposite, reach higher values that is expected in presence of TADs.

Topologically associated domains can be defined as consecutive squares along the diagonal of Hi-C contact map enriched with interactions inside. Consecutive here means that TADs are allocated "head to tail": the end of the one TAD has the same position as the start of the next one. This definition eliminates the existence of breaks and complex structures such as nested or partially overlapping TADs.

The existence of consecutive TADs only implies the presence of single left TAD edge position for each Hi-C row. As a consequence, it should result in the global maximum of $log_2(MA_{(i,j+1)}/MA_{(i,j-1)})$ ratio at the TAD edge candidate position rather than local maximum. In fact, we observe several clear peaks between the HiCExplorer TAD edge positions (Figure 3.2.F). Based on the criteria introduced above we can suggest that these peaks are either noise, or real signals that reveal the presence of edges appeared because of nested or partially overlapping TADs. More evidence for the second explanation comes from the heat map that represents the $log_2(MA_{(i,j+1)}/MA_{(i,j-1)})$ signals within the Hi-C matrix (Figure 3.2.H). The local maxima mentioned above are not

randomly allocated within the matrix, they are organised in visibly clear stripes meaning that several neighbouring Hi-C rows show the same TAD edge candidate position.

To sum up, the MA-based TAD edge calling method potentially allows us to detect more complex architectural structures such as nested or partially overlapping TADs. This weakening of the consecutive TAD assumption creates the opportunity to explore the chromatin architectural hierarchy that is one of the critical topics in the field of TAD calling tools.

3.3.2. TAD edges are visualised as highly intensive stripes at log2 mean ratio heat map

Log2 mean ratio is proposed to reach its local maxima at the positions that are most probable candidates to become the left TAD edge positions. All neighbouring Hi-C rows that demonstrate the same left TAD edge position are expected to belong to the same TAD and the last such row is a candidate to become the bottom TAD edge position (limiting the length of left TAD edge). These two criteria seem to be quite simple, but the realisation is not so clear - due to the data noise we expect that candidate position can highly vary within Hi-C rows belonging to the same TAD.

In the previous Section S3.3.1. we stated that the $log_2(MA_{(i,j+1)}/MA_{(i,j-1)})$ ratio reaches its local maximum only when the position j is a candidate to be a left TAD edge position, where $MA_{(i,j+1)}$ and $MA_{(i,j-1)}$ are defined according to Formula (3.1) and (3.2), respectively. In addition, when we use the “local maximum” instead of “global maximum”, we allow several left TAD edge positions to exist within single Hi-C row, meaning that we want to detect complex architectural structures such as nested or partially overlapping TADs.

Finding the local maxima is not a trivial exercise. When working with noisy data like Hi-C, we would prefer to smooth signal or fit the data to a high degree polynomial function. The problem here is the computational efficiency - each Hi-C row should be processed separately that requires time and memory. In addition, the smoothing should be done in a way to have local peaks of log2 mean ratio coinciding at the same

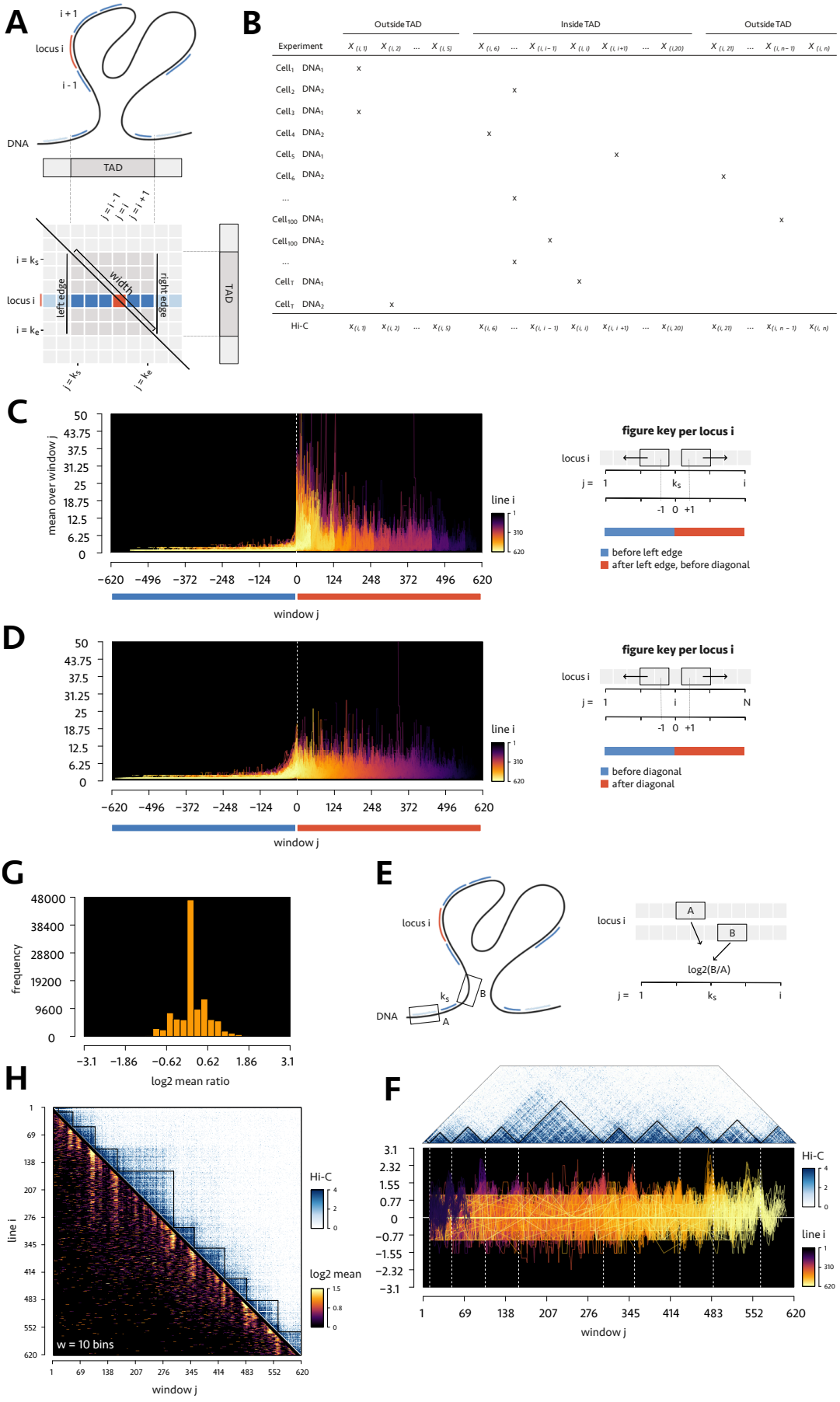

A
i + 1
locus i
i - 1
DNA
TAD
i = k_s
locus i
i = k_e
left edge
width
right edge
TAD
j = k_s
j = k_e
B
Outside TAD
Inside TAD
Outside TAD
Experiment
Hi-C
C
mean over window j
window j
line i
figure key per locus i
locus i
before left edge
after left edge, before diagonal
D
window j
line i
figure key per locus i
locus i
before diagonal
after diagonal
G
frequency
log2 mean ratio
E
locus i
DNA
A
B
log2(B/A)
H
line i
window j
w = 10 bins
Hi-C
log2 mean
F
Hi-C
line i
window j

Figure 3.2. The TAD edge definition and detection criteria. **A.** Selecting the particular locus i, the 3D architecture (top panel) is represented in form of the matrix (bottom panel) produced by Hi-C. Interactions between locus i and other loci are written within matrix row, upstream loci are allocated before the diagonal, downstream loci are allocated after the diagonal; diagonal element is a self-interaction event (orange), it is removed from the analysis. Dark grey colour represents the intra-TAD interactions, light grey colour represents the inter-TAD interactions. Dark blue represents the intra-TAD interactions with locus i, light blue represents the inter-TAD interactions with locus i. The left ($j = k_s$) and right edges ($j = k_e$) are the elements of intra-TAD interactions, which separate the intra-TAD and inter-TAD area Width in the size parameter of the TAD. **B.** For selected locus i, single cell experiment detects a single pairwise interaction. The interaction frequency generated by Hi-C (row i in matrix), represents the sum of all single pairwise interactions in the population of cells. **C.** MA mean estimates centred at the TAD borders called with HiCExplorer. Figure key explains the computation window and x-coordinates for each MA mean estimates. On the positive side of x-axis there are windows calculated on the right-hand side from the selected center point (coloured with orange), on the negative side of x-axis there are windows calculated on the left-hand side from the selected center point (coloured with blue). Each Hi-C row is represented by individual line with individual colour: from the purple line corresponding to the row 1 and to the yellow line corresponding to the row 620. **D.** The same as C, but MA estimates are centred at the diagonal. E. Explanatory figure on log2 mean computation within single Hi-C row i (right panel) and its 3D representation (left panel). Within window B the interactions with locus i are expected to be higher than within window A, so log2 mean ratio is expected to have a peak at k_s which is a TAD edge coordinate. **F.** Top panel represents the heat map of the genomic region chr3R: 21.3 - 21.7 Mb selected for testing the TAD edge criteria. Dark blue colours indicate intense pair-wise contacts retrieved by Hi-C. Black lines indicate the TADs called using HiCExplorer. Bottom panel represents the distribution of log2 ratio of one MA window on the right-hand side from the column position j to one MA window on the left-hand side from the column position j. White lines indicate the TAD edge positions called with HiCExplorer. **G.** The histogram of log2 ratios from F. **H.** Lower triangle represents the allocation of log2 ratios of one MA window on the right-hand side from the position j to one MA window on the left-hand side from the position j within Hi-C matrix. Upper triangle represents the Hi-C interaction frequency. Black lines indicate the TADs called using HiCExplorer.

position within neighbouring Hi-C rows belonging to the TAD. So, before exploring the statistical framework to detect local peaks we can start with naive and simple approach to allocate left TAD edges in order to understand possible limitations and specific details that we should keep in mind when calling local maxima of log2 mean ratio.

The naive threshold-based method can be described as following. Before the algorithm starts, a user selects a particular threshold value that defines the minimum log2 mean ratio. During the algorithm processing, when log2 mean ratio exceeds the given threshold at some column position we define it as the potential left TAD edge position. When the threshold is selected to be relatively low, we expect to detect several consecutive locations to be the potential TAD edge. So, in this case we deal with more like a window rather than the exact position - the TAD edge can be allocated somewhere within the detected window. From a biological point of view, the window makes sense - the chromatin is a dynamic structure, so the interactions detected on TAD border in some cells can be not detected in other cells, and, as a result, TAD border position can vary within several bins. When the threshold is selected to be relatively high, a smaller amount of TAD edges is expected to be found.

When the log2 mean ratio exceeds the threshold, several column positions could be:

- consecutive, meaning that we can identify the window where the possible TAD edge is allocated,
- single peaks separated from each other by large gaps, meaning that, most probable, we detect jumps in signal because of noise,
- the mix of consecutive regions, gaps and peaks.

In addition, the consecutive region observed in one Hi-C row possibly does not fully coincide with consecutive region observed in neighbouring Hi-C row, meaning that these two rows do not have the same TAD edge position. Also, the relatively short gaps between long consecutive regions are possibly produced by noise in the data and, thus, could be combined. Overall, we require the introduction of a set of rules that will guide us when gaps and consecutive regions should be aggregated in order to detect stripes of intense log2 mean ratios signal. The whole threshold-based TAD edge

calling algorithm with required set of rules is completely formalised in the Appendix 3.2, in this Section we focus on the results.

In Figure 3.3, we demonstrate how log2 mean stripes are detected under the different threshold values: the rectangle areas on Hi-C maps surround the bright stripes (Figure 3.3.A-D). As expected, the areas of polygons are expected to decrease when the threshold value increases: there is the lower chance for log2 mean ratio to beat the higher threshold values, so the fewer column and row candidate positions to be included. For some extreme values of threshold, the majority of polygons completely disappears and only small areas are left (Figure 3.3.A-D, bottom panel).

We allocated these polygons as a technical feature that allow to make a decision about the performance of the proposed algorithm: polygons should cover log2 mean ratio stripes that are visible on Hi-C heat map and overlap with reliable TADs called by the HiCExplorer tool. Based on these two criteria, the TAD edge squares are performing well for threshold values being around 1.00 - 1.20 (Figure 3.3.E-G).

The main limitation of the method is the set of rules for gaps and consecutive regions aggregation. They do not take into account the specific features of the data, for example, the variation in the log2 mean ratios, for example, when signal is more disperse we would expect to have larger gaps between consecutive regions. However, the threshold-based TAD edge detection procedure revealed several key features. First, the TAD edge is not a single position conserved within several neighbouring rows but more like a genomic region with intense log2 mean ratio signal. Second, within log2 mean stripe the signal is not homogeneous: the signal is more intense in the centre of the stripe and the signal is blurry on the edges of the stripe. Third, stripe loses its intensity when the signal is far from the diagonal. It means that the TAD edge is more clear when it is closer to the diagonal and it starts to disappear when it is far from the diagonal. It is consistent with the fact that even if the DNA fragments belong to the same TAD, they interact more frequently when they are closer to each other on the chromatin. In addition, as log2 mean ratio signal disappears away from the diagonal, with a threshold-based approach we get low detection of the end of the TAD. It leads

us to the idea that the log2 mean stripe length can be a good feature to allocate to the TAD end position, but it is not sufficient.

3.3.3. TAD edge positioning is based on identification of log2 mean ratio local extrema and its statistical validation

For the selected pair of DNA fragments (i, j), we compute the log2 mean ratio according to the formula $log_2(MA_{(i,j+1)}/MA_{(i,j-1)})$ where $MA_{(i,j+1)}$ and $MA_{(i,j-1)}$ are computed based on the pre-defined window size w according to Formula (3.1) and (3.2), respectively.

Assume that we select a single Hi-C row i and start to compute log2 mean ratios, moving from the left Hi-C matrix edge (from the column position $j = w + 1$) towards the Hi-C matrix diagonal (to the column position $j = i - 1 - w$) (Figure 3.4.A). When we are outside TAD, we expect to observe the interactions with the same intensity in average. Even if we assume that we have genomic distance versus interaction intensity functional relationship in addition to the presence of TADs, we can make an assumption that moving average mean $MA_{(i,j+1)}$ computed on the right-hand-side from the position j is not significantly different from the mean $MA_{(i,j-1)}$ computed on the left-hand-side due to the relatively small size of estimation window w. So, the log2 mean ratio is expected to fluctuate around zero when we are at outside TAD region. When we move closer to the TAD edge, the right-hand-side mean estimating window catches more interactions from inside TAD region. As a result, the log2 mean ratio is expected to rise till the moment when the moving average mean $MA_{(i,j+1)}$ computed based only on interactions inside TAD, in other words, when the column position j is one bin before the TAD edge. After crossing the TAD edge, the left-hand-side mean estimation window catches more interactions from inside TAD region and less from outside TAD region. It means that the mean on the right-hand-side and mean on the left-hand-side from the column position j becomes not significantly different from each other, so the log2 mean ratio start to decrease to zero. Theoretically speaking, we expect to observe the local maximum at the left TAD edge position and the width of the peak is expected

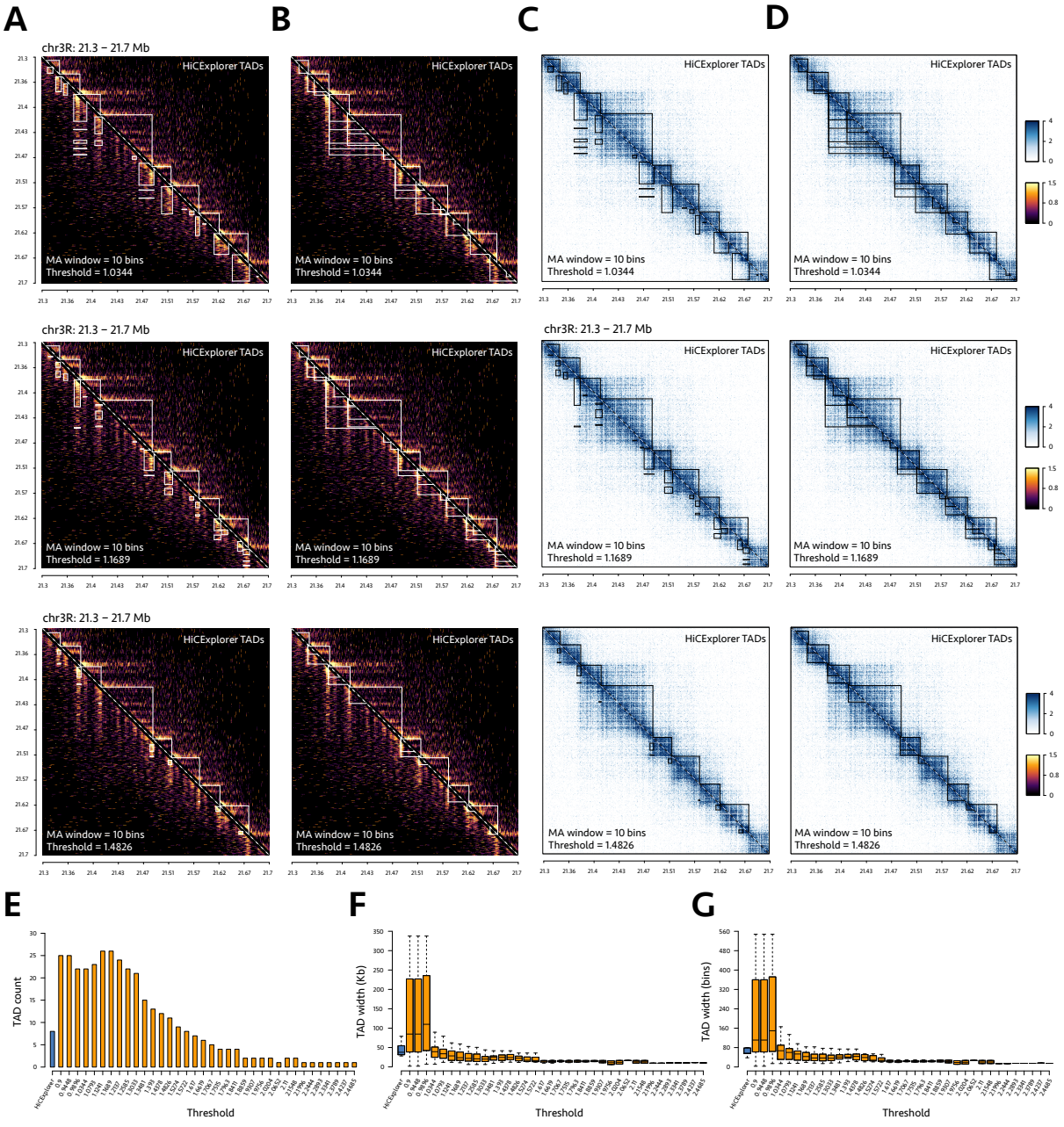

A
B
C
D
chr3R: 21.3 – 21.7 Mb
HiCExplorer TADs
MA window = 10 bins
Threshold = 1.0344
Threshold = 1.1689
Threshold = 1.4826
E
F
G
TAD count
TAD width (kb)
TAD width (bins)
Threshold

Figure 3.3. Performance of the threshold-based TAD edge calling algorithm. **A.** Heat map represents the allocation of log2 mean ratios within Hi-C matrix on the testing genomic region. White lines on upper triangle indicate the TADs called using HiCExplorer. White lines on lower triangle indicate TAD edge areas detected with 1.0344 (top panel), 1.1689 (middle panel) and 1.4826 (bottom panel) thresholds. **B.** Same as A, but white lines on lower triangle indicate reconstructed TAD edges on the basis of TAD edge squares. **C-D.** The Hi-C interaction heat map. Black lines represented the same HiCExplorer TADs, TAD edge squares and reconstructed TAD edges as on A-B. **E.** The number of reconstructed TADs called on the testing genomic region with different thresholds (x-axis). The blue bar represents the number of TADs called by HiCExplorer. **F.** The distribution of widths measured in Kb of reconstructed TADs. The blue bar represents the width of TADs called by HiCExplorer. **G.** Same as F, but widths are measured in bins.

to be $(2w + 1)$. So, instead of point position we observe the stripe that surrounds the left TAD edge candidate position (Figure 3.4.B). The stripe is expected to be non-homogeneous: the log2 mean ratio reaches its maximum intensity around the centre of the stripe.

We worked only with lower triangle of Hi-C matrix as we were exploring the criteria of left TAD edge detection, i.e. the allocation of the start of the TAD. Moving further within Hi-C row i, beyond the diagonal, we deal with interaction between locus i and locus $j = i+1$, $i+1$, ..., n where n is the total number of fragments. As the Hi-C matrix is symmetric, when we cross the left TAD edge within the lower triangle of Hi-C map, we expect to cross the right edge of the same TAD within the upper triangle of Hi-C map (Figure 3.4.C). The right TAD edge represents the end of the TAD. Thus, when we move from the position $j = i + 1$ to the position $j = n$, we observe the interactions within inside TAD area, then cross the edge and we observe the dramatic “drop” of interaction intensity when we are outside TAD. So, we get the similar situation with the left TAD edge but the difference is that instead of the local maximum we observe local minimum at the end TAD edge position. The local minimum appears because when right-hand side MA window is already outside TAD and left-hand side MA window is still inside TAD, we get the negative log2 mean ratio as when calculating the ratio we

get numerator being lower than denominator. When both MA windows are both either inside or outside, the log2 mean ratio should fluctuate around zero.

From the memory efficiency point of view, we may prefer to store only either lower or upper triangle instead of storing the full Hi-C matrix. Again, as Hi-C matrix is symmetric, moving within the Hi-C row i at the upper triangle from the diagonal is analogical to moving within Hi-C column i at the lower triangle from the diagonal (Figure 3.4.D). Mathematically, for the selected pair of DNA fragments (j, i), we define log2 mean ratio as $log_2(MA_{(j+1,i)}/MA_{(j-1,i)})$ where $MA_{(j+1,i)}$ and $MA_{(j-1,i)}$ are computed based on the pre-defined window size w as following

$$MA_{(j+1,i)} = \frac{1}{w}(x_{(j+1,i)} + x_{(j+1,i)} + ... + x_{(j+w,i)}) \tag{3.3}$$

$$MA_{(j-1,i)} = \frac{1}{w}(x_{(j-1,i)} + x_{(j-2,i)} + ... + x_{(j-w,i)}) \tag{3.4}$$

Then, we aim to detect local minima position of log2 mean ratio as the end TAD edge candidates. Note that from here we refer left TAD edge as start TAD edge and right TAD edge as end TAD edge. When we visualise the computation of log2 mean ratio column-wide in a form of a heat map, we observe the stripes of negative log2 mean ratios analogically to row-wise computed log2 mean ratios when detecting the start TAD edge (Figure 3.4.E). Note that these stripes mimic the visually detectable ends of TAD triangles.

We suggest that the width of the peak (i.e. the width of log2 mean ratio stripe) is defined through the MA window size w as $(2w+1)$. As we expect to observe one single peak with the mentioned width, we expect that no other TAD edge can be detected within the distance of $(2w + 1)$. So, the following limitation is arising: there is the connection between minimum distance between neighbouring TAD edges and selected MA window size w. If the window size is too large, we would catch several TAD edges within one stripe. Still, if the window size is too short, we would catch many fluctuations due to biases in Hi-C data.

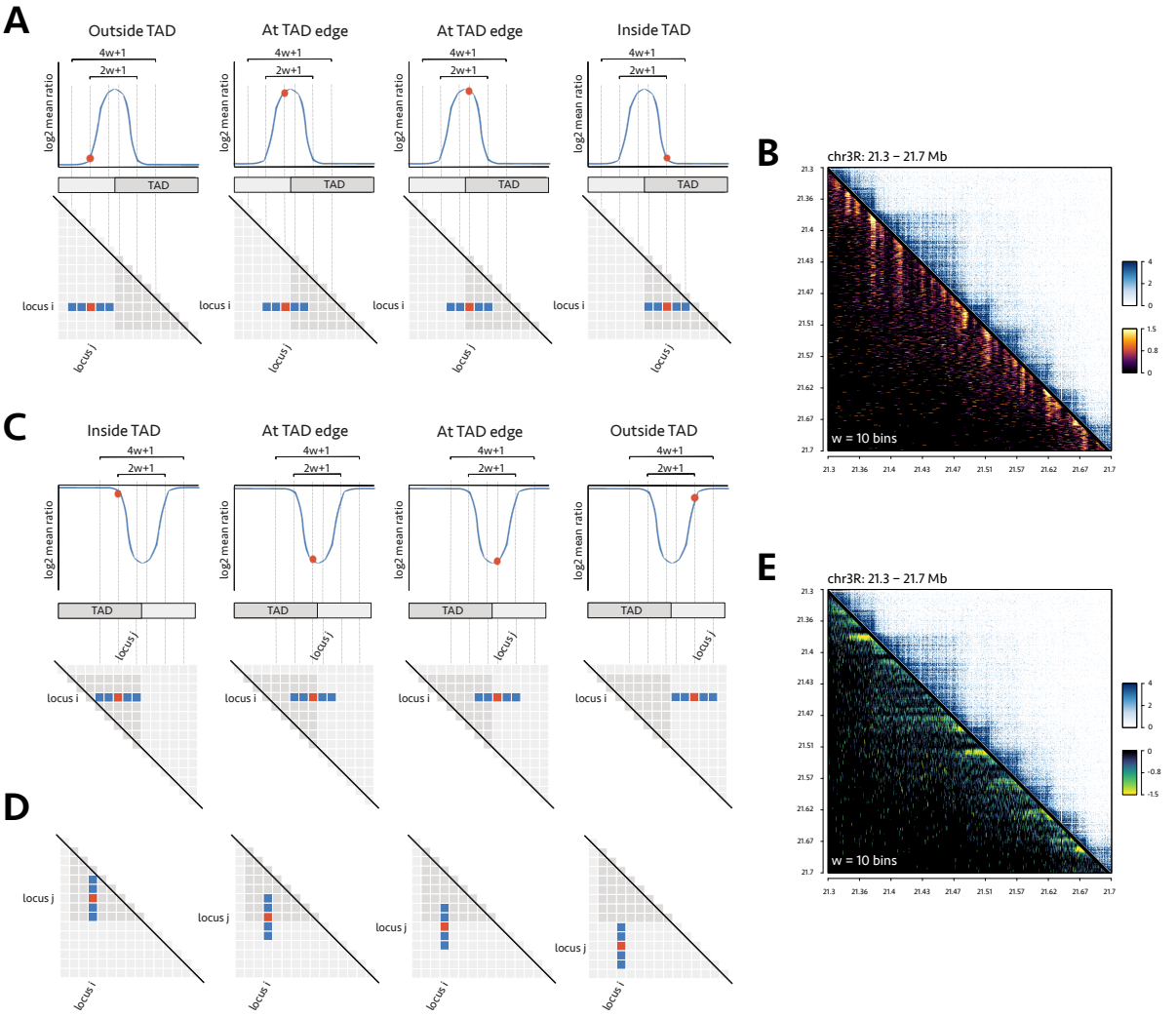


Figure 3.4. Local extrema searching for start and end TAD edge allocation. **A.** Log2 mean ratio is expected to follow the bell-shaped function near the start (left) TAD edge position within single Hi-C row. Orange cell represents the position of log2 mean ratio of the loci pair (i, j), the blue cells on the right indicates the right hand side moving average mean estimate window, the blue cells on the left indicates the left hand side moving average mean estimate window. B. Lower triangle represents the allocation of log2 mean ratios within Hi-C matrix, computed row-wise according to explanation in A. Upper triangle represents the Hi-C interaction frequency. C. Same as A, but the function is near the end (right) TAD edge. D. For computational simplicity, because of symmetric Hi-C matrix the end TAD edge expected function can be defined within the lower triangle. E. Same as B, but the ratios are computed column-wise according to description in D.

3.3.4. Moving Average window size balances the number of valid log2 mean ratios

Generally, Hi-C data contains large amount of non-available interactions. Some pairs of DNA fragments are not present due to low mappability: if they are too short or enriched with repetitive elements and cannot be uniquely mapped to the genome. Also fragments that are involved in complex interactions cannot be counted: if, for example, there are more than two DNA fragments interact simultaneously, only single pair is counted. All these experimental and processing limitations generates NAs in Hi-C data. On the other hand, NAs could appear if the pair of loci did not interact in population of cells in Hi-C experiment. Then, the contact count in this case can be replaced by zero. Some processing procedures and Hi-C-related modelling algorithms propose to replace all NAs with zeros. However, if we aim to work with estimation of average frequencies, presence of large number of zeros can underestimate the averages.

Short MA windows have a higher chance to be filled by lost data points (NAs) only, while long MA windows more probably contain both observations and NAs. The danger of MA subsets either containing mainly NAs or only NAs is that the estimated mean is expected to be either biased or not to be computed at all, respectively. So, the window w should be large enough to have reasonable amount of the windows with observations. At the same time, larger window size leads to wider log2 mean ratio stripe that, as a consequence, affects the accuracy of allocation of the exact TAD edge position within the stripe and reduce the number of stripes that we are able to detect.

Most of the NAs appear far from the diagonal representing extremely distal interactions. Also, as we can see from TADs called with HiCExplorer, the largest detected TAD width within the whole genome is around 400 Kb or 600 bins, at the testing region the maximum width is around 75 Kb and 140 bins (Figure 3.5.C). It means that the storing the whole matrix is not required, it is enough to store the restricted amount of bins from the diagonal (Figure 3.5.A). Thus, when we work genome-wide, we can restrict the data up to 1000 bins as we do not expect TADs to be detected with edges at length more than 1000 bins. At DpnII resolution with median bin size of 575 bp, for each in-

vestigating locus i 1000 bins restriction means that we deal with interactions that are not far than 500-600 Kb upstream from the locus i if we extract signal row-wise (start TAD edge detection) and downstream from the locus i if we extract signal column-wise (end TAD edge detection) (Figure 3.5.A and B).

Increasing the MA window size will reduce the number of data observations used for log2 mean computation. Assume, that we denote the restriction length as X and the total number of fragments as N. On Figure 3.5.A and B, the area representing the total number of interacting pairs used for log2 mean ratio computation is coloured with blue and orange. For selected window size w, the amount of log2 mean ratios are visualised with orange colour only and can be formally computed as

$(X - 2w)(N - X)$

The log2 mean ratio turns into NA when at least one of the MA windows is completely filled with NAs, in other words, when we have at least w consecutive NAs.

1. When w consecutive NAs are at the beginning of the Hi-C row i, for column position $j = w + 1$ according to Formula (3.2)

$MA_{(i,j-1)} = NA$, so $log_2 MA_{(i,j+1)}/MA_{(i,j-1)} = NA$.

2. When w consecutive NAs are somewhere within the Hi-C row i, i.e. for column positions $j = j_1, j_2$, ..., j_w where $j_1 \neq 1$ and $j_w \neq i - 1$, at $j = j_1 - 1$ according to Formula (3.1)

$MA_{(i,j+1)} = NA$, so $log_2 MA_{(i,j+1)}/MA_{(i,j-1)} = NA$

and at $j = j_w + 1$ according to Formula (3.2)

$MA_{(i,j-1)} = NA$, so $log_2 MA_{(i,j+1)}/MA_{(i,j-1)} = NA$

3. When w consecutive NAs are at the end of the Hi-C row i, for column position $j = i - 1 - w$ according to Formula (3.1)

$MA_{(i,j+1)} = NA$, so $log_2 MA_{(i,j+1)}/MA_{(i,j-1)} = NA$.

Overall, the amount of non-available log2 mean ratios when the MA window size is w is between the number of w consecutive NAs and the number of w consecutive NAs multiplied by 2. Also, each w consecutive NAs produce non-available log2 mean ratios with proposed window sizes at $w-1$, $w-2$, $w-3$ and so on. Thus, computing the num-

ber of consecutive NAs within the given restricted Hi-C dataset, we can estimate the upper and lower boundary of the number of non-available NAs and, as a consequence, the number of available NAs (Figure 3.5.D and E). Convex shape of the amount of valid log2 mean ratios versus the window size is expected. When we first increase the short MA window we reduce the chance to catch the MA window consisting of NAs only, so we get fewer non-valid log2 mean ratios and get more valid ones. However, further increase of window size requires more interaction for log2 mean ratio computation, so we get fewer ratios computed. So, for large windows the advantage of getting less non-valid log2 mean ratios is outweighed by the disadvantage of getting fewer signals for further analysis. The minimum estimated number of valid log2 mean ratios reaches its maximum at $w = 70$. This value is too large from the biological point of view as it is expected to produce the log2 mean ratio stripe of approximately 70 Kb. The marginal increase of the convex function reaches 5% at $w = 10$, the same value as we proposed to use when we tested the TAD edge detection criteria in the Section S3.1.

3.3.5. The Mann-Whitney U test statistics is used to detect TAD edge stripes

Suppose that we have a formal procedure that at the end gives us the set of locus positions that are the TAD edge candidates, i.e. local maxima of log2 mean ratio for start TAD edge detection and local minima of log2 mean ratio for end TAD edge detection. These candidates include both the peaks that can be called local maxima/minima with some degree of certainty and peaks that are actually spurious because of the noise. We aim to formalise the procedure that can distinguish between these two scenarios. We focus at start TAD edge only at the current stage as the procedure is expected to be approximately the same for the end TAD edge detection with the only difference that we need to work with local minima detection instead of local maxima.

While we build and explore the approach, we can use the following procedure to select local maxima candidates quickly. We, first, extract log2 mean ratios per each Hi-C row and then use scatter plot with log2 mean ratios versus Hi-C row coordinates (Figure 3.6.A). Then, we apply the Kernel Regression Smoother (the Nadaraya–Watson

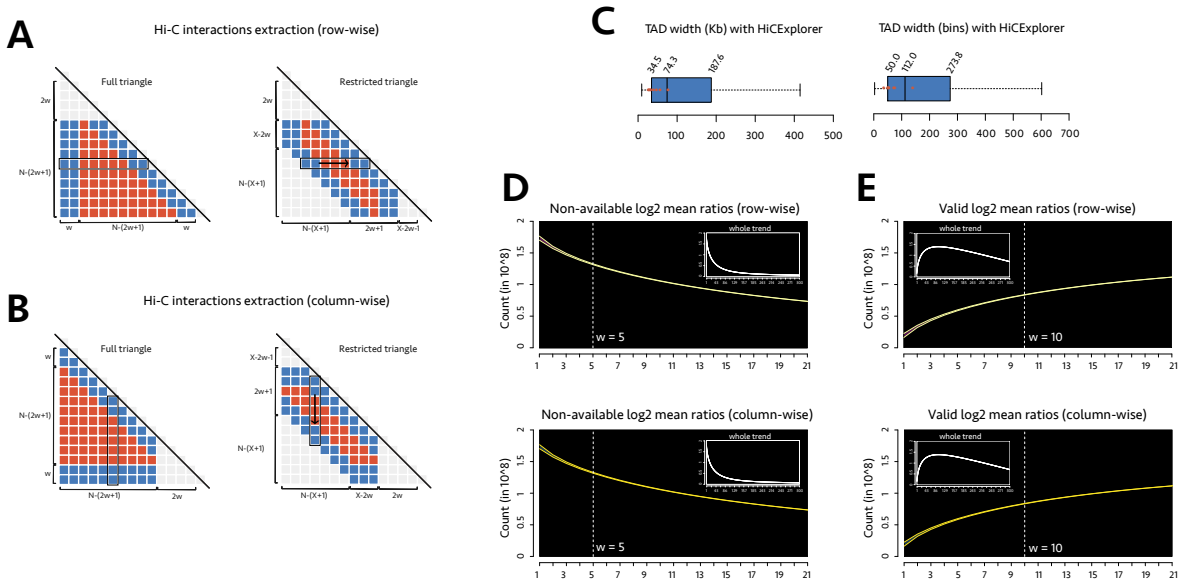


Figure 3.5. Optimal MA window size diagnostics. **A.** Schematic representation of the restricted and non-restricted Hi-C matrices. On the left panel we represent the unrestricted matrix with N interacting loci and MA window w. On the right panel we represent the restricted dataset when the restricted length is X. The blue and orange colours represent the matrix elements that contain the interacting frequencies, the orange colour represents the positions for which we can compute the log2 mean ratios with MA window w. The arrow represents the direction of computing log2 mean ratio, for start TAD edge we compute the ratios row-wise. **B.** Same as A, but we compute log2 mean ratio column-wise for end TAD edge detection. **C.** The distribution of widths (in Kb and in bins) of TADs detected with HiCExplorer genome-wide. The orange dots represent the TADs allocated within the selected test region. **D.** The minimum and maximum estimates of the number of log2 mean ratios being NA computed row-wise (top) and column-wise (bottom) within the genome. The vertical line represents the MA window w starting from which we observe marginal decrease being less than 5%. In the corner we represent the function at larger window sizes. Note that we estimated the numbers based on restricted matrix with $X = 1000$. **E.** Same as D, but we estimated number of log2 mean ratios that do not equal to NA.

kernel regression estimate implemented in R as a part of `stats` package) with normal kernel and various bandwidth parameters and defined the local maxima points. We use the Hi-C matrix restricted by 1000 bins from the diagonal and we still get many NAs and also next to zero log2 mean ratios which are produced far from the diagonal. Including these distal log2 mean ratios can over-weight the smoothing curve closer to the zero line and over-smooth visually detectable peaks of log2 mean ratios. To avoid these problem, at each Hi-C row we extracted only log2 mean ratios that exceed the pre-selected quantile, so we ensure that we apply smoothing only on top log2 mean ratios that are most probably allocated closer to the diagonal. We set two quantile limits at 0.9 and 0.5 and bandwidth parameter at 10. We do not require high accuracy at this stage, as before formalising the procedure we aim to learn nuances that we have to keep in mind.

Even if smoothing curve at 0.5-quantile limit is visually over-smoothed, we still call all visually detectable local maxima (Figure 3.6.A). Actually, we get more local maxima candidates with 0.5-quantile than with 0.9-quantile. The reason is the presence of peaks that are allocated too close to each other (for example, close to rows 100, 180 or 280), so the smoothing curve creates patterns that we can name as “plateau regions”: parts of the smoothing curve that are approximately flat. Given these findings, we can conclude that when we call local maxima candidates we do not require the data set to be restricted but the log2 mean ratios extracted from the whole Hi-C row makes differences to be less pronounced. So, using weighted average of log2 mean ratios when we give larger weights to observations next to the diagonal and lower weights to observations far from the diagonal can make the local maxima more pronounced. Also note that the smoothing is a useful tool when we work with large datasets and we can get too many local maxima candidates that should be examine further. However, it is sensitive to the bandwidth parameter: when the bandwidth is too large we can accidentally lose some significant peaks because of over-smoothing (like in case of plateau regions) but when the bandwidth is small, the signal becomes noisy and we again should deal with extreme number of candidates. With the relatively small *Drosophila* genome, we can

continue to work without smoothing but if we want to perform the TAD edge calling on large organisms like human we may require the smoothing step.

If several neighbouring Hi-C rows share the same start TAD edge, we expect the local maximum to be detected at approximately the same column position (to be accurate, within the stripe of $(2w + 1)$ width). So, we should observe not a single, but a set of log2 mean ratios within several neighboring Hi-C rows that have in average higher value close to the local maximum and lower value when we are far away from the local maximum position. So, there are three main features that we need to formalise. The first one is that we have to define the genomic region that surrounds the local maximum and demonstrate in average higher log2 mean ratios. The second feature is that we have to define the genomic regions that are remote enough to demonstrate in average lower log2 mean ratios. And the third one is that we have to define the way to conclude whether the difference in average log2 mean ratios is significant or not. So, it make sense to introduce the window here to define the region around the local maximum candidate position. Actually, we introduce three windows called left, middle and right windows (Figure 3.6.B-D). The middle window is the region centered at the maximum candidate position, the left and right widows are the regions on the left-hand-side and right-hand-side, respectively, from the middle window and they describe area that is far away enough from maximum candidate position.

The values of the log2 mean ratio covered by the left, middle and right windows are used to validate the candidate position to be real rather than spurious local maximum. If it is the local maximum, the log2 mean ratios are expected to be in average higher within the middle window than within the right or left window (Figure 3.6.C). The Mann-Whitney U test is the first natural choice for a statistical non-parametric test that allows us to say with some certainty level that two sets of observations come from different populations.

The Mann-Whitney U test.

We have a sample of n_x observations $(x_1, x_2, ..., x_{n_x})$ and a sample of n_y observations $(y_1, y_2, ..., y_{n_y})$.). All x_i where $i = 1, \ldots, n_x$ are independent, all y_j where $j = 1, \ldots, n_y$ are independent and x_i and y_j are independent from each other. We aim to test whether two samples come from the same population:

$H_0 : P(x_i > y_j) = 1/2 \ \forall \ i = 1, ..., n_x; \ \ j = 1, ..., n_y;$

$H_1 : P(x_i > y_j) \neq 1/2$ (two-sided)

Simply to say, we aim to compare the medians: under the null hypothesis the medians are the same, under the alternative hypothesis the medians are different. When the null hypothesis is rejected, it means that we obtained two sets of observations most probably representing two different distributions. When the null hypothesis is not rejected, it means that the datasets did not demonstrate enough evidence to conclude that they obtained from different distributions.

The test statistics is calculated as follows:

$U_x =$ number of times when $x_i > y_j \ \forall \ i = 1, ..., n_x; \ \ j = 1, ..., n_y;$

$U_y =$ number of times when $x_i < y_j \ \forall \ i = 1, ..., n_x; \ \ j = 1, ..., n_y;$

$U = \min(U_x, U_y)$

Then, use the statistical tables for the Mann-Whitney U test to find the probability of observing the value lower than U. If test is two-sided, double the probability.

We apply the one-sided Mann-Whitney U test comparing, at first, log2 mean values from the middle and left windows, then from the middle and right windows. Thus, low p-values (lower than pre-defined cut-off) means that the log2 mean ratios around the candidate position are on average higher than ratios that are far enough from the candidate position and this candidate position can be called a local maximum. So, we say that the candidate position is the local maximum when log2 mean ratios within both left and right window are on average lower than within middle window. Technically, we compare the maximum of two p-values with the pre-defined cut-off to conclude that the candidate is a local maximum. This is reasonable when we observe a clear stripe that is remote enough from neighbouring stripes. When two stripes are close enough

we have a high chance that we catch a neighbouring peak within right or left window. It is exactly the situation of plateau regions: we observe two clear peaks which are too close to each other (Figure 3.6.D). So, as the consequence of the Mann-Whitney U test, neither of the peaks is determined as local maxima. To avoid the problem, either the width of the left, middle and right windows should be selected in a way that two neighbouring candidates do not belong to neighbouring testing windows, or instead of both left and right windows being on average filled with lower signals, it is enough to have the significant difference at only one side - either left or right. The first solution creates a danger of performing the test with relatively small sample sizes that decreases its power and trustfulness in the results. The second solution allows both clear peaks and plateau regions to be detected.

So, we have left, middle and right windows to perform a test and if the p-value in at least one direction is significant, we confirm the local maximum and, as consequence, start TAD edge position.However, the number of observations which we include within each window can affect the result of the validation test. Log2mean ratio demonstrates the distance decay: log2mean ratio is intense within the TAD edge and then decreases when it is far away from the diagonal. At the long distance, there are either no contacts detected or contact frequencies are very low, so we accumulate more log2mean ratios which are zeros or NAs. In particular, visually the log2mean stripes do not go further than 150 bins (Figure 3.5.C). If we include all zeros and NAs appearing at the long distances, we face a risk that the distributions would almost be composed of zeros, so the medians at left, middle and right would be approximately the same and the Mann-Whitney U test would not recognise the candidate position as local maxima even if at shorter distances it is. For example, the peak that is under 0.9-quantile limit is defined as significant, but when we include more observations with 0.5-quantile limit the same peak will not be defined as significant anymore (Figure 3.6.B and C). The number of observations included into the test can be restricted by the sizes of left, middle and right windows. The number of observations included into the test is defined by the sizes of left, middle and right windows.

The Mann-Whitney U test is non-parametric, so it is helpful tool when we don't assume any distribution. However, the test assumes the independence of the observations between the groups. In case of Hi-C data, the contacts are not independent. When two DNA fragments interact to each other, due to ligation neighboring fragments will also demonstrate intensive interactions in Hi-C dataset. So, we face a spatial dependency of neighboring chromatin interactions. However, when we compute log2mean ratio, we extract the average long-range interactions between fragment i and two windows – upstream and downstream from fragment j. In this Chapter, the window is 10 bins which is approximately 5 Kb – the distance is long enough to diminish the spatial dependence. Two neighbouring log2mean ratios computed row-wise within the same Hi-C row are based on approximately the same contact observations and different only by two contact frequencies: downstream windows are different in contacts x_{j+1} and x_{j+w+1} and upstream windows are different in contacts x_{j-1} and x_{j-w-1}. These interactions are relatively far from each other, so can be treated as independent for some extent. Overall, on log2mean ratios within the same Hi-C row can be treated as independent. At column-wise perspective, we face a distance decay. When row-wise log2mean ratios within the same column are computed, we look on the average interaction frequencies between fragments from $(j-w)$ to $(j+w)$ and fragments i, $i+1$, $i+2$ and etc. For i and $i+1$, for example, the spatial dependency between log2mean ratios exist as these fragments are too close to each other. However, the spatial relationship between log2mean ratios within validation test windows is expected to be different for left, middle and right windows if TAD edge is present within the middle window. The log2mean ratio decay should be more pronounced within middle window as ratios face both distance effect and the presence of TAD edge while within left and right windows ratios face only distance effect. So, we suggest that the distance decay is not critical for independence between left, middle and right validation datasets. Although, for future research it worth to carefully analyse the functional relationship between log2mean ratio and genomic distance to better understand the distance decay effect and how it can affect the TAD edge length estimation.

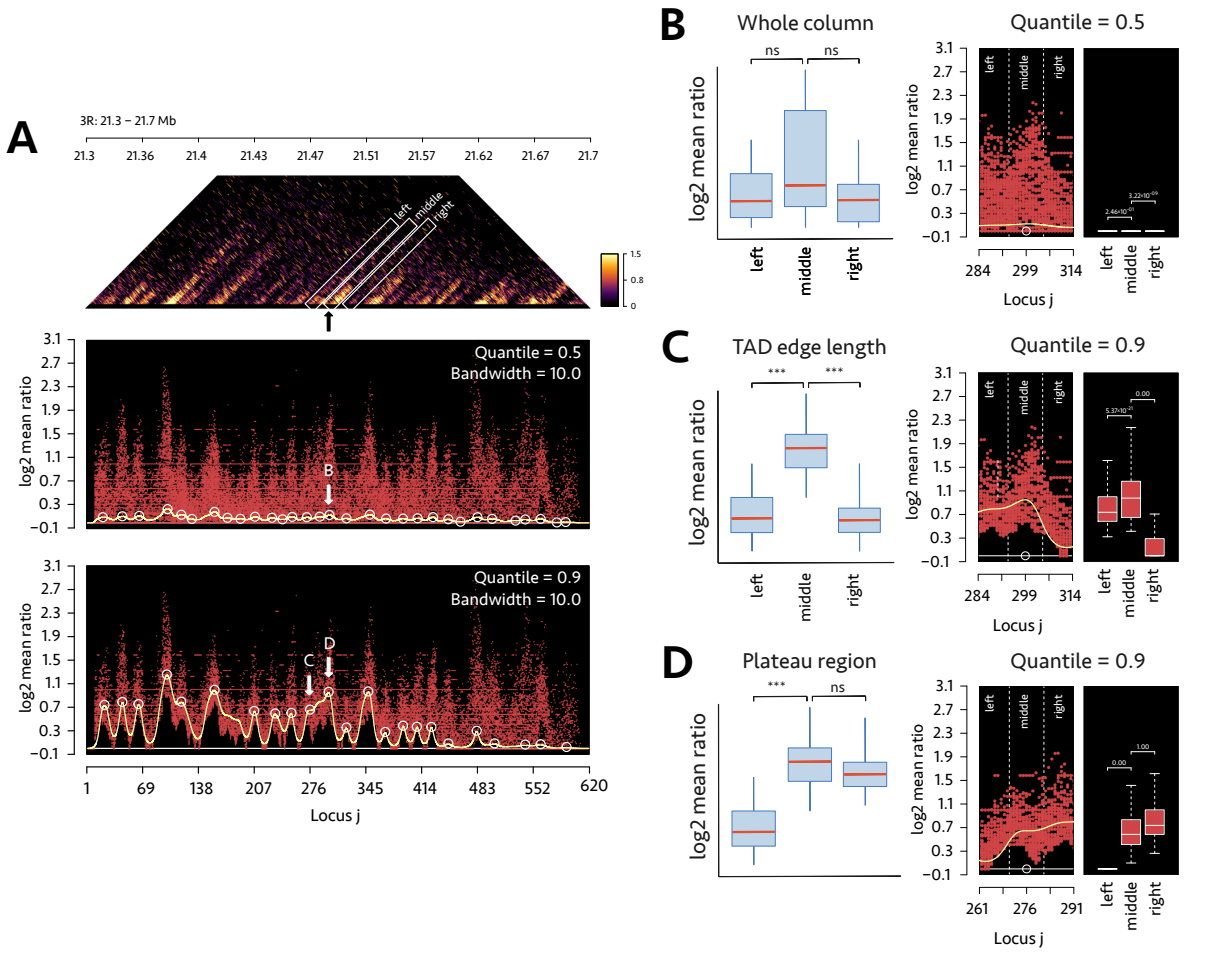


Figure 3.6. The Mann-Whitney U test as a tool for local maxima validation. **A.** The allocation of left, middle and right Mann-Whitney U testing windows. Top panel represents the heat map of log2 mean ratios computed row-wise rotated to simplify the representation of start TAD edge. Middle panel represented the log2 mean ratios extracted within each column candidate position, we expect to observe more higher dots at start TAD edge candidate positions. We select only ratios that are higher than 0.5-quantile limit computed at each x-coordinate. The yellow line smooths the signal based on the Kernel Regression Smoother with normal kernel and bandwidth parameter set at 10. Bottom panel represents the same as top panel but with 0.9-quantile limit. **B.** The visualisation of the scenario if we select all (or majority) of log2 mean ratios within the candidate position. On the right, we represented the described scenario based on 0.5-quantile limit (middle panel A). The numbers represent the p-values of one-sided Mann-Whitney U tests. **C.** Same as B, but we represent the scenario when we select only ratios belonging to the TAD edge stripe length. On the right, we represented the described scenario based on 0.9-quantile limit (bottom panel A). **D.** Same as C, but we represented the plateau region scenario.

3.3.6. Computing the effect size statistics may be a base for log2 mean stripe length identification

We need to define the proper left, middle and right window sizes selection procedure. Using the Mann-Whitney U test, we can compare the median of two data sets that are not required to be the same size. So, in theory, the width and the length of testing windows can be different. However, for simplicity and homogeneity of motivation we select the sizes of windows according to similar rules.

We start with testing windows width. We expect that the width of log2 mean stripe, which position we aim to detect, is approximately $(2w + 1)$, so the distance between two neighbouring peaks is at least $(2w+1)$. If we assume that all three testing windows have the same widths $\hat{w}$, the distance between centre of middle window and the last column position belonging to the right window is approximately $1.5\hat{w}$ (Figure 3.7.A). This distance should be equal or less than minimum distance to the neighbouring peak:

$1.5\hat{w} \leq 2w + 1$

$\hat{w} \leq \frac{2}{3}(2w + 1)$

For simplicity, when selecting the testing window to be the same width as the MA window, i.e. $\hat{w} = w$, we insure that candidate peaks are allocated far enough from each other to be detected. In addition, we also stated that we aim to detect the significant difference between average log2 mean ratios in at least one direction - either left-hand side or right-hand side - so, we do not face a danger to accidentally cover two neighbouring local maxima candidates within two neighbouring testing windows.

In the previous section, we stated that we have to limit the length of testing window in order to avoid the risk of incorrect validation testing due to the distance decay of log2mean ratios within the stripe. The length of the testing window should reflect the length of log2 mean ratio stripe - when the stripe is finished, there is no more difference in signals between left, middle and right windows. As different TADs are expected to have different length of TAD edges, we should be accurate if we decide to select the single testing window length for all candidate positions. Otherwise, we can introduce the procedure that selects the optimal window length individually for each

candidate position based on some statistics. The **effect size** measure can be used as this statistic. The effect size can be use when we need to measure the strength of the relationship between two variables in a population. In particular for the Mann-Whitney U test, the effect size can measure the strength of the difference between medians of two investigated samples.

The effect size of the Mann-Whitney U test.

The effect size r for the Mann-Whitney U test with sample sizes n_x and n_y and test statistics U is computed as:

$r = \frac{z}{\sqrt{n_x+n_y}}$

where $z = \frac{|U-\mu|}{\sigma}$, $\mu = \frac{n_x n_y}{2}$ and $\sigma^2 = \frac{n_x n_y (n_x+n_y+1)}{12}$

The programming implementation can use the following formula where _val is the Mann-Whitney U test p-value and the factor `k = 2` if test is two-sided and `k = 1`:

```
r <- abs(qnorm(p_val/k))/sqrt(n_x + n_y)
```

When a stripe is fully covered, the difference between median signals within neighbouring testing regions is maximised. Then, increasing the testing window length further than the end of the stripe, we catch more close to zero signals and the overall difference is expected to fall. Following this logic, we expect the difference between medians to be strong when the testing window covers the part of the stripe or the stripe in full and the strength of the difference should fall when we set testing window be longer than the stripe. Considering distance decay of log2 mean, the effect size is expected to smoothly fall even within the log2 mean stripe. On Figure 3.7.D, where demonstrated the relationship between p-value, effect size and depth for start TAD edge candidate positions on the testing genomic region, we observed that the effect size generally decreases when depth increases and the speed of change is also different: first, effect size changes slowly and p-value dramatically decreases, then, effect size falls faster and p-value reaches its minimum, and, at the end, effect size again slowly decreases together with the increase in p-value. As p-values themselves are not comparable, we can implement the optimal stripe depth estimation based only on

the effect size decrease speed as follows. With particular local maxima candidate, we start with left, middle and right windows with size being $w \times w$ and compute both the Mann-Whitney U test p-values and effect sizes comparing middle versus left testing window and middle versus right testing window. Then, we select the single p-value that is lower (as we require only one p-value to be lower than pre-selected cut-off) and single corresponding effect size. At the next step, we expand the length of the testing window by w, so the testing windows have size of $2w \times w$ (Figure 3.7.A). We repeat the procedure until we reach the end of the restricted matrix. For each candidate position, we fix the effect size one step before its largest "drop" as it reflects the moment when the marginal increase in testing window length brings the signals that significantly reduce the strength of the difference in medians. Together with the effect size we fix the corresponding p-value.

Each candidate position is characterised by the p-value and effect size. The p-value informs that the difference in log2 mean ratios between the candidate position and neighbouring regions exists or not and the effect size reports the strength of this difference. We require one more measure that indicates the actual strength of TAD edge, i.e. the quantity representing the average increase in signal when we cross the TAD edge. As we define the optimal stripe length with effect size, the sample average computed on the middle window log2 mean ratio can be a proper estimate for this measure. Generally, TAD edge strength can be treated as an insulation strength - how clear is the insulation of DNA fragments belonging the TAD from the fragments that are outside the TAD edge. Here we face a difference between the proposed TAD edge calling procedure from other insulation score based techniques like HiCExplorer. For starts we indicate how each particular locus is insulated when we move along DNA downstream, for ends we indicate the insulation in upstream direction. So, we indicate one-sided insulation cases. HiCExplorer is an insulation score (IS) based TAD calling tool (Ramirez et al. 2018). The IS indicates the average interaction frequency around the genomic position and when the IS is significantly lower than IS at surrounding regions, we detect the TAD border (Crane et al. 2015). Overall, HiCExplorer searches

for the regions that demonstrate significantly lower interaction frequency in comparison with regions allocated in both directions - downstream and upstream. So, HiCExplorer detects only two-sided insulation. When one-sided insulation is allowed, we give more freedom for the algorithm to detect more complex structures as now each start does not require the paired end at the same position and each end does not require the paired start.

On Figure 3.7.B, we demonstrate all discussed above features of the TAD edge calling algorithm. To make it easier to refer to the algorithm, we name it as **COrTADo** (**C**omplex **Or**ganisation of **T**opologically **A**ssociated **Do**mains). COrTADo combines the data import, transformation in log2 mean ratio matrix, statistics extraction and validation analysis. As an input for statistics extraction step, we get the matrix of log2 mean ratios computed row-wise. Start TAD edge is defined as the straight line segment that starts on the diagonal and goes within several neighbouring rows down at single column position. Given the definition, at first, we extract log2 mean ratios per each column. Then, we compute the weighted average within each column where the weight is the reciprocal of the distance to the diagonal (in bins). So, we get the largest weight to the signals that are allocated close to the diagonal and then weights decrease when we far from the diagonal. At the next step we detect the local maxima coordinates at log2 mean weighed average line as potential TAD edge positions. We define the optimal stripe length, extract the p-value, effect size and strength statistics. The coincidence between the estimated stripe lengths and stripes is highly visible. It is not perfect when the stripe has clear reduction in the intensity, but after the reduction the signal is still present. For example, the start TAD edge at Figure 3.7.B at the position chr3R: 21.38 (4th stripe from the left) shows the estimated stripe length being shorter than visually clear stripe. It possibly signals about the presence of nested TADs, in particular, when two or more TADs have the same start position but different edge lengths. The proposed criteria catches the maximum “fall” in effect size and not the moment when stripe signal absolutely disappear. So, the criteria can efficiently catch the TAD edge start but if we aim to reconstruct the TAD patterns (and not only edge positions)

we have to keep in mind that detected start position can be the start for several TADs.

For the end TAD edge detection, the procedure is approximately the same, the only couple of corrections have to be considered (Figure 3.7.C). The first difference is that when we computed the log2 mean weighted average line we search for local minima coordinates to select TAD edge candidates. The second difference is that when we perform a test, we expect that the log2 mean ratios within the middle window is in average lower than within the left and right windows. Also note that when we compute the insulation strength, we take the absolute average value of the log2 mean ratios within the middle window. The log2 mean ratios at end TAD edge are expected to be negative as we detect a "drop" of interaction intensity when crossing TAD edge. In order to be consistent with the start TAD edge characteristics, we take the absolute value.

Note, that the whole COrTADO algorithm is formalised in the Appendix 3.3 and the implementation in R is available on this GitHub repository:

https://github.com/lm17047/COrTADo.git

3.3.7. Threshold selection affects the number of confident TAD edges

We got approximately 7200 candidate positions for start TAD edges and 7200 candidate positions for end TAD edges. Each candidate position was characterised by three parameters: p-value, effect size and signal strength. Based on the combination of these measures we could distinguish between different classes of TAD edges such as insignificant insulation, weak insulation and strong insulation. Insignificant insulation related to situations when we detected candidate position either due to biases in Hi-C data or because the difference between interaction frequency inside TAD and outside TAD was negligible. On the other side, strong insulation took place when this difference is dramatic and we could visually distinguish the presence of the TAD edge on Hi-C heat map. We also require the weak insulation scenario, when the TAD edge was relatively weaker and not so pronounced as in strong insulation case but still intra-TAD interactions were more statistically significant than inter-TAD interactions.

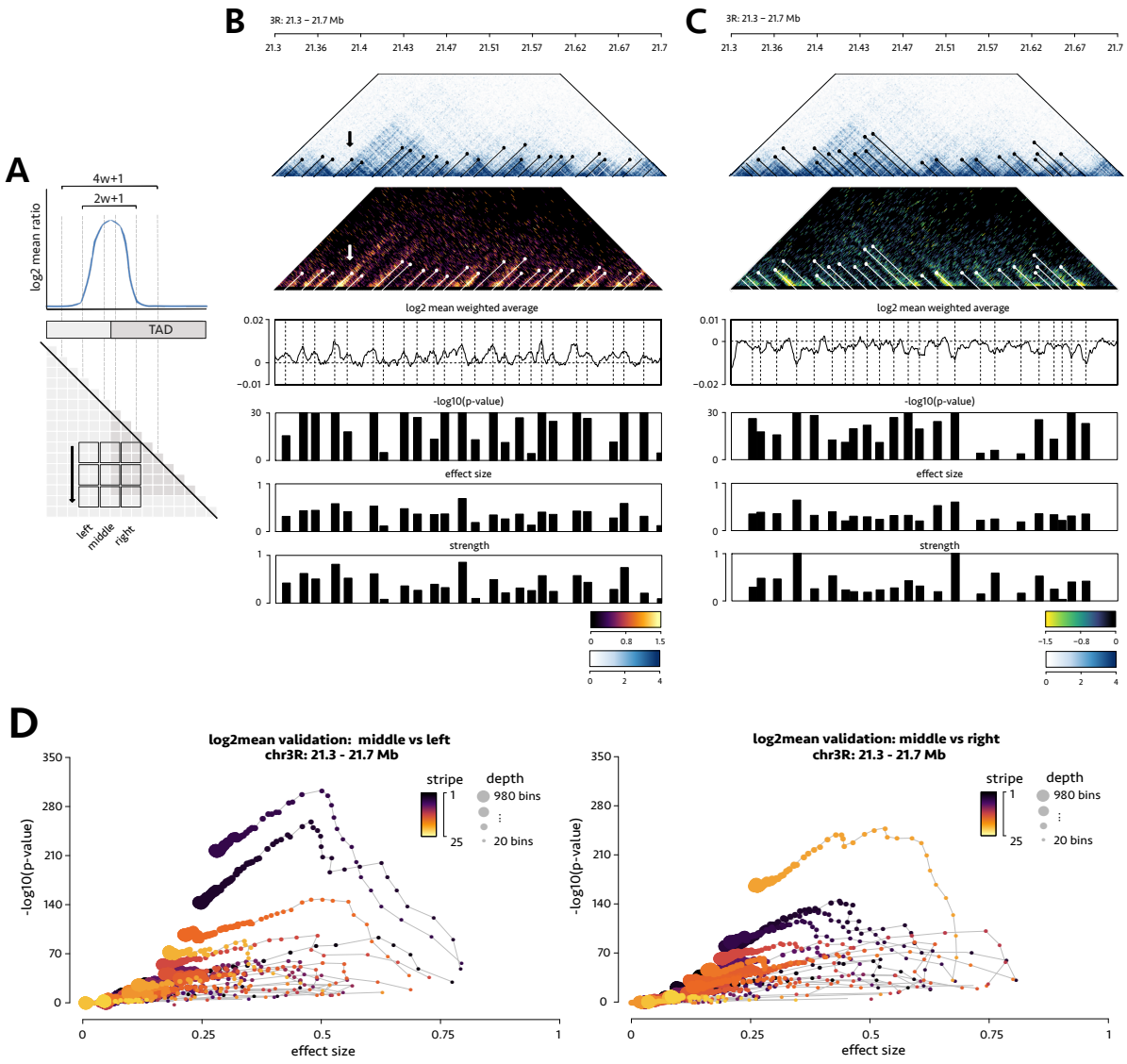

A
4w+1
2w+1
log2 mean ratio
TAD
left
middle
right
B
3R: 21.3 – 21.7 Mb
21.3 21.36 21.4 21.43 21.47 21.51 21.57 21.62 21.67 21.7
log2 mean weighted average
0.02
0
−0.01
-log10(p-value)
30
0
effect size
1
0
strength
1
0
0 0.8 1.5
0 2 4
C
3R: 21.3 – 21.7 Mb
21.3 21.36 21.4 21.43 21.47 21.51 21.57 21.62 21.67 21.7
log2 mean weighted average
0.01
0
−0.02
-log10(p-value)
30
0
effect size
1
0
strength
1
0
−1.5 −0.8 0
0 2 4
D
log2mean validation: middle vs left
chr3R: 21.3 - 21.7 Mb
stripe
1
25
depth
980 bins
20 bins
-log10(p-value)
350
280
210
140
70
0
0 0.25 0.5 0.75 1
effect size
log2mean validation: middle vs right
chr3R: 21.3 - 21.7 Mb
stripe
1
25
depth
980 bins
20 bins
-log10(p-value)
350
280
210
140
70
0
0 0.25 0.5 0.75 1
effect size

Figure 3.7. Effect size statistics for optimal TAD edge length estimation and COrTADo summary. **A.** The step-wise procedure of checking the candidate length of TAD edge stripe based on the Mann-Whitney U test effect size. Top panel represents the expected behaviour of log2 mean ratios within the start TAD edge stripe. Bottom panel represents the allocation of left, middle and right testing windows. The arrow represents the direction in which we expand the testing window. **B.** The COrTADo start selection summary. First panel is the heat map that represents the Hi-C interaction matrix. Second panel represents the allocation of log2 mean ratios computed row-wise. Heat maps are rotated to simplify the representation of start TAD edge. Black and white lines represent the estimated TAD edge stripe length based on effect size statistics. Second panel represents the log2 mean weighted average. Local maxima positions identify the TAD edge candidate positions. Next three panels represent p-value, effect size and strength statistics computed at TAD edge candidate positions. **C.** Same as B, but instead of start TAD edge summary we represent end TAD edge summary. Note that as strength is expected to be negative (as ratios within stripe are mostly negative), we represent absolute value of strength to be consistent with start TAD edge summary. **D.** The relationship between validation test non-adjusted p-values for middle-vs-left windows (left panel) and middle-vs-right windows (right panel) and effect sizes at different depths (log2mean ratio stripe length). Each line represents the test for each start TAD edge candidate shown at (B), lines are distinguishable by colors. Size of the points represents different depths, selected with step of 10 bins from 20 to 980.

Genomic region that was allocated in between two TADs can demonstrate two-sided balanced insulation when both upstream and downstream chromatin domains are strongly insulated from each other. However, the genomic region can also demonstrate two-sided insulation but insulation strength from upstream and downstream domains would be different. We call this scenario as two-sided imbalanced insulation. In the case of weak insulation parameter thresholds, both upstream insulation and downstream insulation would be detected, so this genomic region would be included in both COrTADo start and COrTADo end sets. In case of strong insulation parameter thresholds, insulation in only one direction would be detected and we would get either COrTADo start or COrTADo end. Note, that in this scenario of unbalanced insulation other two-sided insulation based TAD calling tools, such as HiCExplorer or Fan-C, would not recognise the genomic region of interest if the threshold parameters would be too stringent.

The first parameter was **p-value**. The p-value measured the probability that there was no difference between median of log2 mean signal at candidate position and median of log2 mean signal within neighbouring regions. Lower probability meant greater difference in median signals. We would reject the hypothesis of median equality when the probability is less then pre-selected threshold - significance level α. However, we could not compare p-values straightforward as we had many candidate positions and the problem of multiple testing arose.

In general, when we perform several hypothesis tests simultaneously, the overall chance to detect at least one false positive, i.e. when we reject null hypothesis while it is true, increases with the number of tests done. In particular for the COrTADo algorithm, when we test so many candidate positions the chance to incorrectly accept the uninsulated TAD edge is extremely high. So, the multiple testing correction implies that individual test p-values should be adjusted in some way to ensure that the overall probability of observing at least one significant result due to a chance remains below the pre-selected threshold α. There are several correction methods that deal with this problem, the most widely used are Bonferroni correction and Benjamini-Hochberg cor-

rection (also known as FDR correction). Bonferroni adjustment treats all simultaneous tests equally and independently, so if the probability to observe at least one false positive across all tests is set at α, each individual test result is rejected when it is less than α divided by the number of tests performed. Despite being computationally simple, the disadvantage of Bonferroni correction is that it can be too strict. This method, when corrects the false positive rate, the same time increases the false negative rate. For COrTADo it means that there is the high chance to remove TAD edge candidates that are actually highly insulated. This characteristic may be useful when the cost of error is too high and we cannot pay the risk. When the research is more exploratory, we would prefer more relaxed method, for example, FDR correction. Instead of decreasing the chance of incorrectly selecting an insignificant result within all tests, FDR correction selects the individual cut-off based on the rank of individual p-value among others. It implies that the proportion of false positives, on average, would be less than overall cut-off α. FDR correction is widely used in genomic association studies.

The second parameter was **effect size**. The parameter p-value itself was not enough to detect the difference in median signals between candidate position and neighbouring regions as it was sensitive to sample size. A larger sample used in the test tended to generate lower p-value and demonstrated more significant result even if the difference was not actually significant. So, long TAD edge stripes with negligible insulation had a high chance to be accepted. The effect size was the measure that, independently on sample size, identified the proportion of observations that supported the median differences direction. Effect size of 0 detected complete overlap in distributions of candidate position and neighbouring regions, effect size of 1.0 detected the absence of any overlap in two sets. Summing up, we would need to accept candidate positions with adjusted p-values being less than selected significance level and then exclude from them the ones that demonstrated low effect size values.

The third parameter was TAD edge insulation **strength**. Even if the difference in median signals was significant between the TAD edge candidate position and neighbouring regions (low p-value and high effect size), the strength of insulation could be

relatively low. Most of the time we would expect that effect size and strength were highly correlated. However, careful analysis of these two measures together would allow us to distinguish between TAD with weak but visually sharp edge and TAD with dramatic increase in interactions inside but blurred edge.

We analysed the distribution of three described parameters at all candidate start and end positions called with COrTADo. Although we explained the preference of FDR multiple testing correction over Bonferroni, we visualised both adjustment methods (Figure 3.8.A and B). As expected, Bonferroni produced less significant candidates than FDR at the same significance level of 0.01; i.e. Bonferroni removed 1934 start and 1914 end candidates while FDR removed only 821 start and 849 end candidates.

Even as FDR allowed more candidates, at the next step we removed the ones that demonstrated low insulation by selecting proper effect size and strength thresholds (Figure 3.8.C). To classify the candidates based on the effect size, we used Cohen's conventions (Cohen 1988). Effect size of 0.5 and larger represented a large level of insulation, effect size from 0.3 to 0.5 represented the moderate insulation, effect size from 0.1 to 0.3 represented weak insulation and below 0.1 was negligible insulation. We selected possible strength thresholds from 0.2 to 1.0 with 0.2 step and analysed the number of candidates that fell into each cluster. Note that as strength was computed as the average log2 mean ratio within TAD edge stripe, it estimated the log2 ratio of mean interaction frequency of intra-TAD area over mean interaction frequency of inter-TAD area. So, possible strength thresholds referred to approximately 1.15, 1.32, 1.52, 1.74 and 2 times increase in interaction frequency when crossing TAD edge. Two largest clusters that contained 20-22% of valid candidates (FDR adjusted p-value $<$ than 0.01) were characterised by small effect size (from 0.1 to 0.3) and strength up to 0.4. Approximately half of valid candidates demonstrated either effect size being greater than 0.3 or strength being greater than 0.4 - we defined these borders as confident ones and we would use them for further analysis (Figure 3.7.C, yellow and green area). There were 3284 start and 3320 end positions (Figure 3.8.D). Also, we defined more stringent set of thresholds - either effect size cut-off at 0.5 or strength cut-off at 0.6 - to

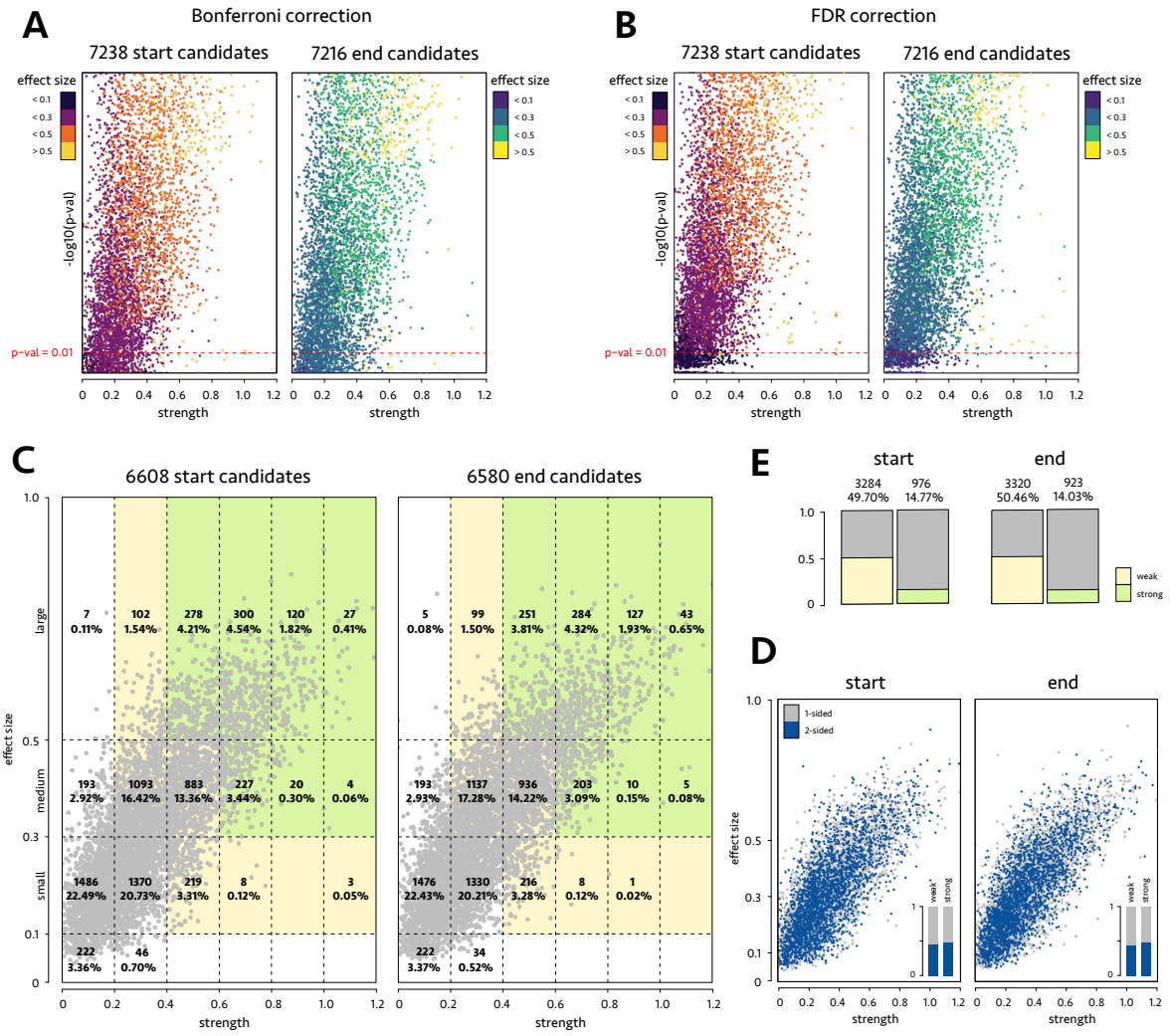

A
Bonferroni correction
7238 start candidates
7216 end candidates
effect size
< 0.1
< 0.3
< 0.5
> 0.5
-log10(p-val)
p-val = 0.01
strength
B
FDR correction
7238 start candidates
7216 end candidates
C
6608 start candidates
6580 end candidates
large
medium
small
7 0.11%
102 1.54%
278 4.21%
300 4.54%
120 1.82%
27 0.41%
193 2.92%
1093 16.42%
883 13.36%
227 3.44%
20 0.30%
4 0.06%
1486 22.49%
1370 20.73%
219 3.31%
8 0.12%
3 0.05%
222 3.36%
46 0.70%
5 0.08%
99 1.50%
251 3.81%
284 4.32%
127 1.93%
43 0.65%
193 2.93%
1137 17.28%
936 14.22%
203 3.09%
10 0.15%
5 0.08%
1476 22.43%
1330 20.21%
216 3.28%
8 0.12%
1 0.02%
222 3.37%
34 0.52%
E
start
end
3284 49.70%
976 14.77%
3320 50.46%
923 14.03%
weak
strong
D
start
end
1-sided
2-sided

Figure 3.8. Diagnostics and classification of TAD edges based on strong or weak insulation. **A-B.** Distribution of -log10 of adjusted p-value (Bonferroni correction in A and FDR correction on B) versus strength. Darker colours indicate lower effect size, brighter colour indicate higher effect size. Left panel indicated the values computed at left TAD edge candidates that represent TAD start positions (7459 candidates). Right panel indicated the values computed at bottom TAD edge candidates that represent TAD end positions (7462 candidates). Red horizontal line indicates the adjusted p-value threshold of 0.01 to remove insignificant TAD edge candidates. **C.** Distribution of effect size versus strength at candidates with FDR adjusted p-value < 0.01. Effect size divided into 4 classes: large when ≥ 0.5, medium when < 0.5, small when < 0.3 and no effect when < 0.1. Vertical lines represent possible strength thresholds of 0.2, 0.4, 0.6, 0.8 and 1.0. These values mean, respectively, 1.15, 1.32, 1.52, 1.74 and 2.0 times increase of interaction frequency when crossing TAD edge. Numbers and percentages correspond to the candidates that belong to each specified strength and effect size intervals. Green area represents set of intervals that defines strong COrTADo borders. Yellow area represents additional set of intervals that together with green area defines weak COrTADo borders. **E.** Bar plot represents the number and share of strong (green) and weak (yellow) COrTADo borders within all valid candidates (adjusted p-value < 0.01). **D.** Distribution of effect size versus strength at candidates with FDR adjusted p-value < 0.01. Dark blue colour demonstrates COrTADo start candidates that allocated within 5 Kb from nearest end candidate (and vice versa) and can be defined as candidates with two-sided insulation. Grey colour demonstrates other candidates that can be defined as candidates with one-sided insulation. Bar plot at right bottom corner represents the share of candidates with two-sided and one-sided insulation selected with weak and strong thresholds.

separate the borders with more significant and stronger insulation (Figure 3.8.C and D, green area). There were 976 start and 923 end positions defined as strong COrTADo borders (Figure 3.8.D).

We suggested that one-sided and two-sided insulation might be maintained by different epigenetic mechanisms and their roles in cellular functioning might be different. To study possible imbalances in insulation at TAD borders, we ensured that selected thresholds did not introduce any biases in proportions of two-sided and one-sided insulation cases. We defined start candidate as two-sided insulation when it was allocated

within 5 Kb from the nearest end candidate. Analogically, we defined two-sided end candidates. We did not observed any specific preferences of either two-sided or one-sided candidates in terms of effect size and strength, both classes were approximately uniformly distributed on strength versus effect size plane (Figure 3.8.E).

3.4. Robustness of COrTADo algorithm under different conditions

3.4.1 Normalised versus non-normalised Hi-C map

The whole framework was based on *Drosophila* BG3 wild-type Hi-C data which was not trimmed and not normalised. Usually, before TAD calling, Hi-C data should be corrected to avoid the influence of technical biases (see Chapter 1 for more details). Along with TAD allocation, HiCExplorer uses the ICE correction (iterative correction and eigenvector decomposition) (Imakaev et al. 2012). The ICE procedure relies on the assumption that all DNA fragments have the same chance to be detected, so there is no special preference towards some pair-wise interactions during a Hi-C experiment. As a result, the corrected Hi-C interaction matrix has a higher chance to represent the contacts that are consistent with the real 3D chromatin architecture and which are not caused by ligation artefacts or low mappability of the fragments.

Here, we compare the performance of COrTADo when we use raw Hi-C data and when we use the data where we performed ICE correction before TAD edge calling. First, we balanced the Hi-C matrix with the same parameters as we did in Chapter 2 to filter low quality reads and remove the systematic biases. Then, we called start and end TAD edges using COrTADo with MA window w set at 10 bins, FDR multiple testing correction and adjusted p-value threshold at 0.01. The window size was selected based on MA window size as described in the Section 3.3.4. We got approximately 6300 candidate positions for start TAD edges and the same number for end TAD edges, which is slightly less than for raw Hi-C (Figure 3.9.A, C and D). In addition, the distribution of effect sizes and strengths is visually indistinguishable between raw and corrected datasets (Figure 3.9.A). The Mann-Whitney U test also confirmed that the differences are not statistically significant or very small (Figure 3.9.E and F, boxplots labelled with

"all"). The high similarity makes it possible to apply the same set of rules as before to separate candidate positions with weak and strong insulation profiles. The number of weak start and end COrTADo borders was also similar when we corrected the data, while the number of strong borders were slightly reduced (Figure 3.9.B-D).

At first, it seems that the reduction is reasonable: the correction should remove the skewness towards some bins, so the Hi-C interaction matrix will demonstrate more smooth signal and, as a consequence, the sharp TAD edges that appeared due to noise would not be detected anymore and only clear edges would remain. The heat maps of raw and corrected Hi-C matrices plotted at the testing genomic regions confirmed this - we notice more sharp and scattered interactions with raw data and more smooth interactions with corrected data (Figure 3.8.H, bottom panel). However, the allocation of borders detected on raw and corrected datasets mostly was not the same - even if they were placed relatively close to each other, they demonstrated the different insulation strength (Figure 3.9.H).

At the genome-wide level, the raw and corrected borders did not mostly overlap as well (Figure 3.9.I). However, while we provide the punctuated positions (same restriction site) of the start and end TAD borders, the actual border is allocated within the log2 mean ratio stripe, so the position of the border can vary, in our particular case, within $\pm w$, so ± 10 bins. When we looked at any overlaps between stripes, we got overall more common borders - approximately 67% of borders detected at raw Hi-C data were also detected at corrected Hi-C data (Figure 3.9.J, left panel). As the medium bin size at DpnII resolution is around 500-600 bp, the 10 bins window on average represents the 5 Kb window. However, we get less commonly detected borders when we use 5 Kb window instead of 10 bins window (Figure 3.9.K, left panel) meaning that the distance between neighbouring borders that we treat as common between raw and corrected datasets is slightly more than 5 Kb for at least 1000 common borders. However, expanding the window up to 10 Kb the overlap becomes approximately full, so borders survive the Hi-C matrix correction (Figure 3.9.L, left panel).

Surprisingly, the percentage of common borders decreases when we compare sets

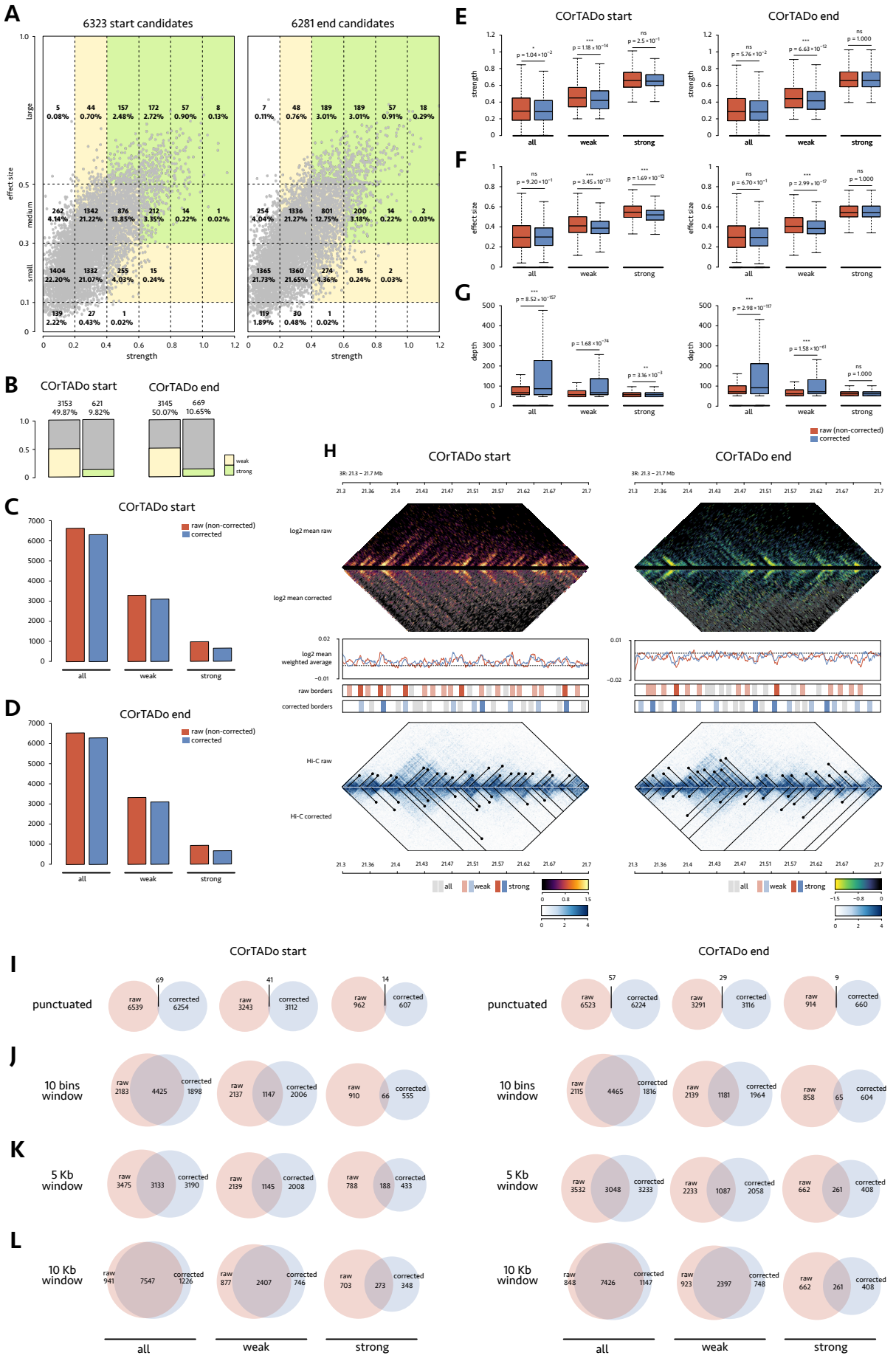

A
6323 start candidates
6281 end candidates
effect size
strength
large
medium
small
B
COrTADo start
COrTADo end
weak
strong
C
COrTADo start
raw (non-corrected)
corrected
all
weak
strong
D
COrTADo end
E
COrTADo start
COrTADo end
F
G
depth
H
log2 mean raw
log2 mean corrected
log2 mean weighted average
raw borders
corrected borders
Hi-C raw
Hi-C corrected
I
punctuated
J
10 bins window
K
5 Kb window
L
10 Kb window

Figure 3.9. COrTADo performance with raw and corrected Hi-C data. **A.** Distribution of effect size versus strength at candidates with FDR adjusted p-value < 0.01 for TAD edges called on the corrected Hi-C data. Effect size divided into 4 classes: large when ≥ 0.5, medium when < 0.5, small when < 0.3 and no effect when < 0.1. Vertical lines represent possible strength thresholds of 0.2, 0.4, 0.6, 0.8 and 1.0. These values mean, respectively, 1.15, 1.32, 1.52, 1.74 and 2.0 times increase of interaction frequency when crossing TAD edge. Numbers and percentages correspond to the candidates that belong to each specified strength and effect size intervals. Green area represents set of intervals that defines strong COrTADo borders. Yellow area represents additional set of intervals that together with green area defines weak COrTADo borders under the same set of rules as described in the main text for raw Hi-C data. **B.** Bar plot represents the number and share of strong (green) and weak (yellow) COrTADo borders within all valid candidates (adjusted p-value < 0.01) at corrected Hi-C data. **C-D.** Bar plots represent the number of all valid, weak and strong start and end COrTADo borders called on raw (orange) and corrected (blue) Hi-C data.**E-F.** Box plots represent the distribution of insulation strength (E), effect size (F) and log2 mean stripe length or depth (G) for all valid, weak and strong start and end COrTADo borders called on raw (orange) and corrected (blue) Hi-C data. **H.** COrTADo summary at testing genomic region. First panel is the heat map that represents the allocation of log2 mean ratios computed at raw (top) and corrected (bottom) Hi-C data. Second panel represents the log2 mean weighted average for raw (orange) and corrected (blue) datasets. Next two panel represent the allocation of COrTADo borders: with grey we indicate valid borders that are not in weak and not in strong sets, transparent color indicates the weak borders and bright color indicates the strong borders. Last panel is the heat map that represents the allocation of raw (top) and corrected (bottom) Hi-C interactions. Black lines represent the estimated TAD edge stripe length based on effect size statistics. **I-K.** Venn Diagramms indicate the overlap between COrTADo borders called at raw and corrected datasets for different classes (all, weak and strong) assuming the punctuated positions (I) and 10 bins (J), 5 Kb (K) and 10 Kb (L) windows around TAD borders.

of weak and strong borders (Figure 3.9.J-L, middle and right panels). We expect that, for example, strong borders are the most insulated ones, so they should be most probably detected in both raw and corrected data. In order to understand the source of this inconsistency, we decided to look closer on the distribution of strength, effect size and log2 mean stripe length (depth) (Figures 3.9.E-G). We noticed that for both COrTADo

starts and ends, strength and effect size in case of corrected dataset are slightly lower than in case of raw dataset indicating a modest decrease in insulation strength. However, significant changea were detected only for weak borders, for set of all borders and strong borders the decrease was either insignificant or little significant. Depth distributions, in turn, were extremely different when looking at all detected TAD borders. In fact, after the Hi-C correction, the optimal log2 mean stripe lengths became much longer.At the testing region, some of the stripes were indeed longer than the visually noticeable TAD edges (Figure 3.9.H). Overall, the Hi-C matrix correction seems to smooth interactions in the way that TAD edges became more blurred, so the optimal stripe length then fixed at the larger distances from the diagonal. As a consequence, the average insulation strength within the stripe is computed, first, based on in average slightly lower log2 mean ratios as the “jump” from inter-TAD to intra-TAD interactions is not as sharp as for raw Hi-C data, and, second, based on more interactions which are far from the diagonal. This explains the reason why we observe many common borders within all valid raw and corrected datasets, but the share becomes lower with more stringent set of thresholds. The decrease in average insulation strength followed by the matrix correction means that the borders may have possibly changed the insulation strength class.

The COrTADo calling algorithm relies on an estimation of optimal stripe length based on effect size. The current implementation of the procedure computes the effect size at different stripe lengths and then searches for the point where the “drop” in effect size is the most dramatic, indicating the possible end of a log2 mean stripe. The algorithm seems computationally simple, but requires the visually sharp stripe, otherwise we face a danger of overestimating the stripe length. In accordance with this, we prefer to use the current version of COrTADo on raw Hi-C data as the estimated stripe length seems more realistic and better mimics the visually clear TAD edges. However, we have to keep in mind that we have a higher chance to detect spurious edges along with the actual ones. During Hi-C, some of the bins tends to interact more than others, so they would show more interactions and possibly create spurious edges. As

an advantage, COrTADo compares the interactions within and outside the stripe using windows rather than punctuated genomic positions, so skewed interactions of a single bin should not affect TAD edge detection dramatically.

3.4.2. Replicate stability

The TAD calling procedure is sensitive not only to the systematic biases but also to the library size. When the library is small, we observe less interactions, so the insulation between chromatin domains is expected to be not so pronounced in comparison with data based on a large Hi-C library. In this Chapter, we work with the merged Hi-C matrix which represents the sum of two matrices obtained from two biological replicates. So, each replicate generates some proportion of total Hi-C reads. In the case of our data, both replicates are approximately the same size (see Chapter 2 and Table 2.1 for more details), which means that each replicate represents approximately 50% of a merged library. We used COrTADo to call start and end TAD edges on both replicates to access the robustness of the algorithm depending on the library size.

Using FDR multiple testing correction and p-value threshold of 0.01, we obtained around 6000 start and 6000 end TAD border positions (Figure 3.10.A-C). Looking at the distribution of effect size versus strength parameters of the selected borders, we notice that the points are more concentrated at lower values and there are slightly less points showing a strong insulation strength. In consistence with this observation, when we apply the same set of rules as before to call weakly and strongly insulated TAD borders, we get significantly less borders belonging to each class in comparison with merged data. It supports the suggestion that as the library size decreases, we get less interactions and insulation between chromatin domain is not so pronounced.

As COrTADo allocates the TAD edges based on the change between intra-TAD and inter-TAD interaction frequencies observed in several neighbouring genomic positions, the TAD edges have a high chance to be detected at the same positions even with the reduced library. A comparison of punctuated TAD border positions did not show massive overlap between merged and both replicates, however, the significant amount

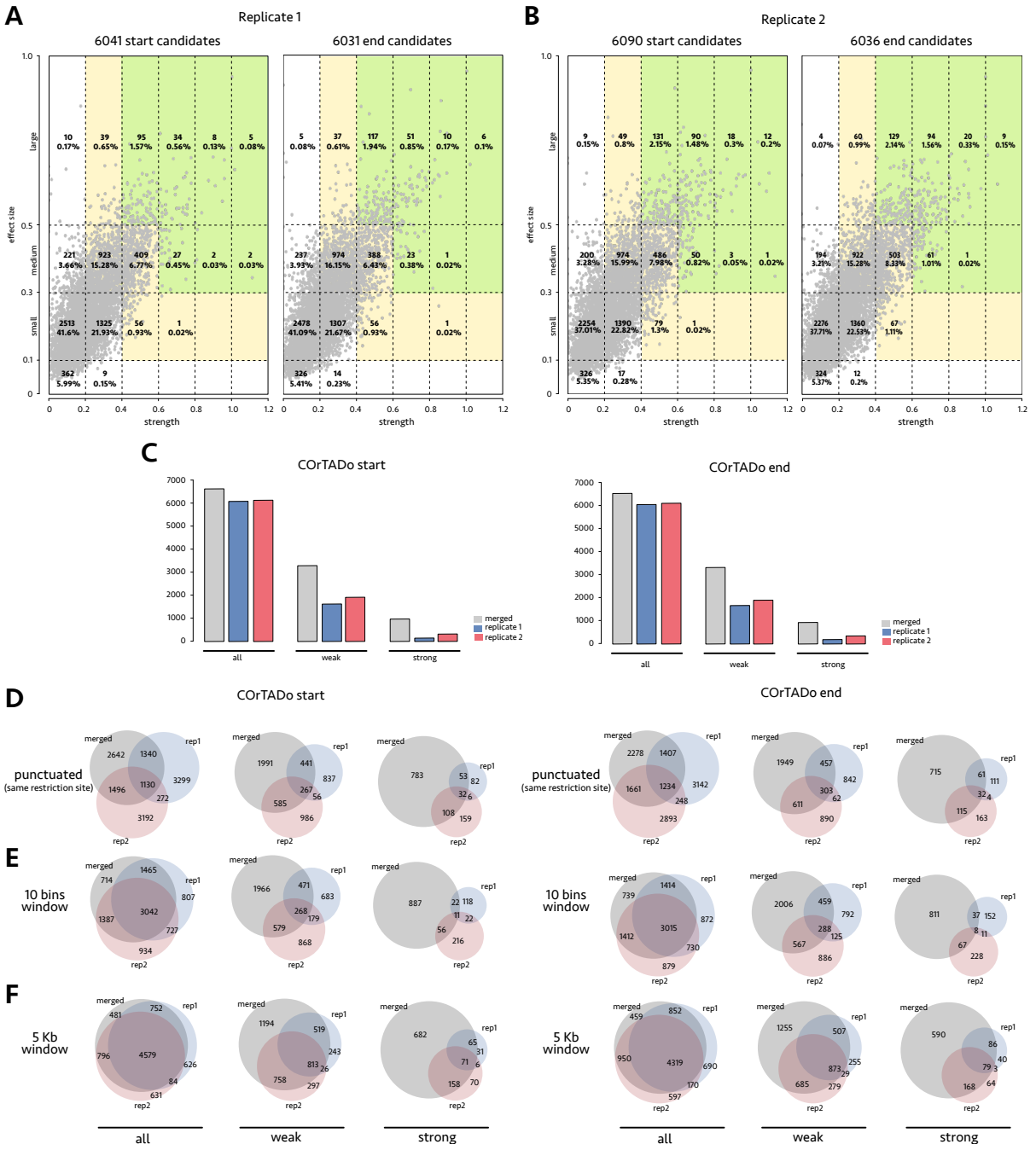

A
Replicate 1
6041 start candidates
6031 end candidates
B
Replicate 2
6090 start candidates
6036 end candidates
effect size
strength
large
medium
small
C
COrTADo start
COrTADo end
merged
replicate 1
replicate 2
all
weak
strong
D
COrTADo start
COrTADo end
punctuated
(same restriction site)
E
10 bins
window
F
5 Kb
window
rep1
rep2
merged

Figure 3.10. COrTADo performance with merged matrix and matrices per biological replicate. **A-B.** Distribution of effect size versus strength at candidates with FDR adjusted p-value $<$ 0.01 for TAD edges called on the replicate 1 (A) and replicate 2 (B). Effect size divided into 4 classes: large when $\geq$ 0.5, medium when $<$ 0.5, small when $<$ 0.3 and no effect when $<$ 0.1. Vertical lines represent possible strength thresholds of 0.2, 0.4, 0.6, 0.8 and 1.0. These values mean, respectively, 1.15, 1.32, 1.52, 1.74 and 2.0 times increase of interaction frequency when crossing TAD edge. Numbers and percentages correspond to the candidates that belong to each specified strength and effect size intervals. Green area represents set of intervals that defines strong COrTADo borders. Yellow area represents additional set of intervals that together with green area defines weak COrTADo borders under the same set of rules as described in the main text for merged Hi-C data. **C.** Bar plots represent the number of all valid, weak and strong start and end COrTADo borders called on merged data (grey), replicate 1 (blue) and replicate 2 (red) data. **D-F.** Venn Diagramms indicate the overlap between COrTADo borders called at merged matrix and replicate matrices for different classes (all, weak and strong) assuming the punctuated positions (D) and 10 bins (E) and 5 Kb (F) windows around TAD borders.

of borders were common between either replicate 1 and replicate 2, or merged data set and one of the replicates (Figure 3.10.D). Moving from the punctuated position to the genomic window of 10 bins (Figure 3.10.E) or 5 Kb window (Figure 3.10.F) increases the overlap, so many borders become common between two sets and all three sets as well. However, comparison of the borders defined as weak and strong did not reveal huge coincidence between merged, replicate 1 and replicate 2 data sets as expected.

3.4.3. Homogeneous and non-homogeneous bins

The COrTADo algorithm does not rely on the genomic distances and only on the relative distances between the bins within Hi-C matrix. When bins are homogeneous, mathematically speaking, there is not difference between genomic distances and relative distances. Up to this point, we analysed the Hi-C data at DpnII resolution - use of the DpnII restriction enzyme enables the generation of the high resolution Hi-C matrix with non-homogeneous bins with the median size between 500-600 bp. We aim

to look at the differences in TAD edge calling procedure between Hi-C matrices with non-homogeneous (Dpn II resolution) and homogeneous (1 Kb resolution) bins.

Note that a switch to 1 Kb bins changes the ratio between valid and non-available log2 mean ratios: when a particular bin with the size less than 1 Kb shows NA, there is a high chance to lose this NA after the aggregation of this bin with neighbouring ones to create a single bin of 1 Kb; when a particular bin is larger than 1 Kb shows NA, it will be possibly segregated into several NAs which would affect the total number of valid interactions. When we perform the window size diagnostic, we confirmed the optimal window size to be at least 7 bins which is analogical to 7 Kb window (Figure 3.11.A).

We run the COrTADo calling with FDR multiple testing correction and p-value threshold of 0.01 as well as with DpnII Hi-C data. Interestingly, the distribution of edge parameters was significantly different when we use 1 Kb bins. While the total numbers of valid TAD edges were comparable, the 1 Kb borders showed the insulation strengths that were much higher (Figure 3.11.B and C). Around 70% of TAD borders called on 1 Kb resolution matrix demonstrated either a medium and large effect size or insulation strength being higher than 0.4 (than is at least 1.3 times increase of inside-TAD interactions in comparison with outside TAD area). Also, under the same set of rules to define weak and strong borders for DpnII and 1 Kb data sets, the amount of strong borders at 1 Kb resolution was more than 2 times higher than at DpnII resolution. A visually clear increase in insulation strength indicates that the TAD edges at 1 Kb resolution maps are more pronounced than the same edges at DpnII resolution maps.

There is nearly no overlap between DpnII and 1 Kb TAD edges when we look at single genomic positions which is reasonable taking into account different genomic coordinates of DNA fragments (Figure 3.11.D). However, looking at the overlaps between 10 bins window for DpnII data and 7 bins window for 1 Kb data around the valid TAD edges, approximately half of the DpnII and half of 1 Kb borders were common (Figure 3.11.E). As the distribution of insulation strength parameters have changed, the overlap for weak and strong borders was not so pronounced. The large differences in classification into weak and strong borders with homogeneous and non-homogeneous bin

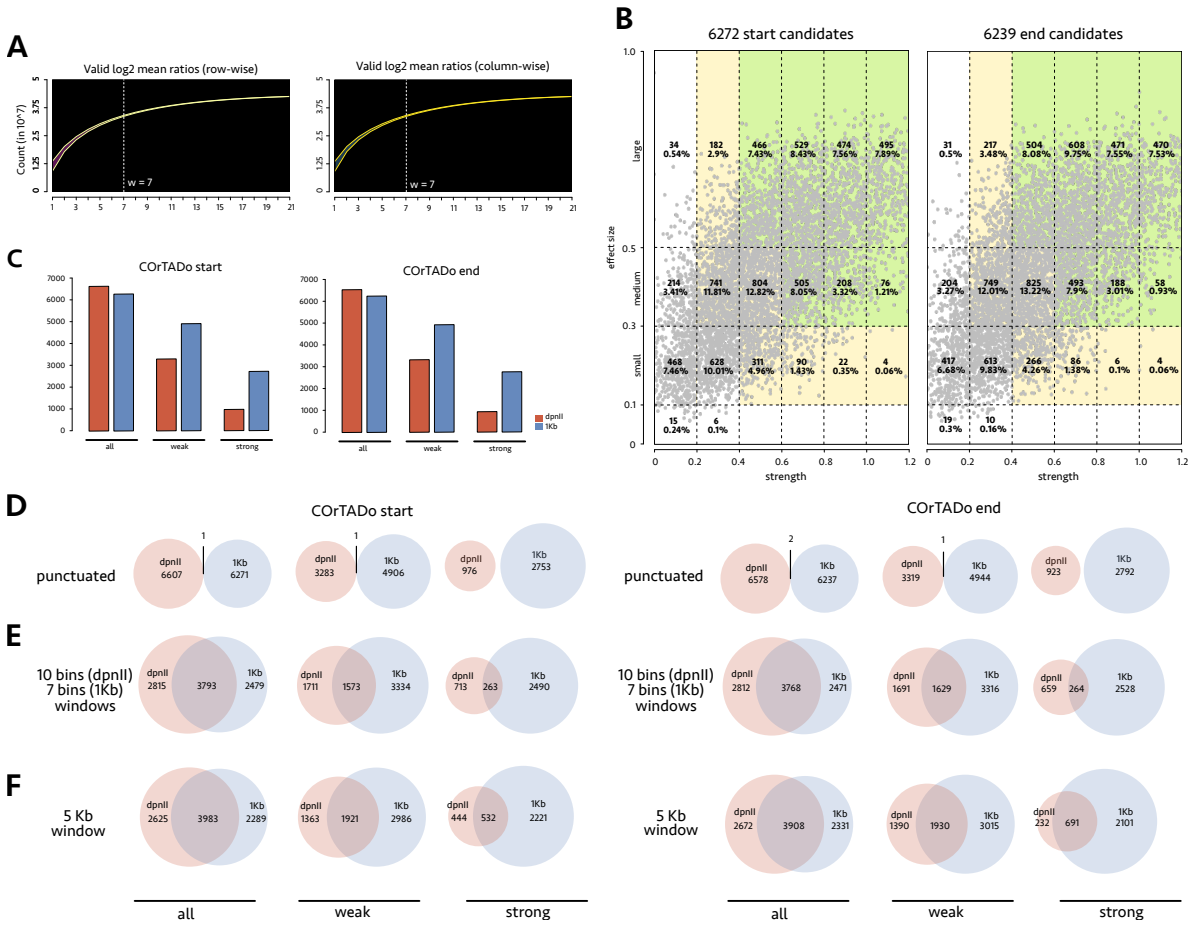


Figure 3.11. COrTADo performance with homogeneous and non-homogeneous bin sizes. **A.** MA window size diagnostics as described in the main text. **B.** Distribution of effect size versus strength at candidates with FDR adjusted p-value < 0.01 for TAD edges called on 1 Kb resolution Hi-C matrix. Effect size divided into 4 classes: large when ≥ 0.5, medium when < 0.5, small when < 0.3 and no effect when < 0.1. Vertical lines represent possible strength thresholds of 0.2, 0.4, 0.6, 0.8 and 1.0. These values mean, respectively, 1.15, 1.32, 1.52, 1.74 and 2.0 times increase of interaction frequency when crossing TAD edge. Numbers and percentages correspond to the candidates that belong to each specified strength and effect size intervals. Green area represents set of intervals that defines strong COrTADo borders. Yellow area represents additional set of intervals that together with green area defines weak COrTADo borders under the same set of rules as described in the main text for DpnII resolution Hi-C data. **C.** Bar plots represent the number of all valid, weak and strong start and end COrTADo borders called on DpnII (orange) and 1 Kb (blue) resolution Hi-C data. **D-F.** Venn Diagramms indicate the overlap between COrTADo borders called at DpnII and 1 Kb resolution matrices for different classes (all, weak and strong) assuming the punctuated positions (D) and 10 (DpnII) and 7 (1 Kb) bins (E) and 5 Kb (F) windows around TAD borders.

sizes indicate that the set of rules to distinguish between different insulation strength classes should be revised for each particular data set. Selection of common rules for all conditions, which we aim to compare, can dramatically affect the downstream conclusions about the differences in insulation of chromatin domains.

3.5. Summary and discussion

We introduced the COrTADo. COrTADo is a TAD calling tool which uses the Hi-C generated data to detect the border positions of topologically associated domains (TADs). The algorithm allocates the start and end TAD borders separately, so allowing complex chromatin architectural patterns to be detected. The list of such patterns include partially overlapping and nested TADs, breaks between TADs, as well as TAD borders demonstrating different insulation strengths in upstream and downstream directions. TAD borders detected using COrTADo can be further used to reconstruct the complex chromatin topology. TAD borders detected using COrTADo can be further used to reconstruct the complex chromatin topology. Complex architectural patterns can include partial overlapping, breaks between neighboring TADs and nested TAD. All these patterns rely on the reconstruction of TADs based on TAD borders and it is a separate research question which should be address in the future. One of the natural assumptions that could be made is that the start anchor locus should interact with fragments within the TAD with approximately the same level of intensity as end anchor locus, so the start and end TAD edges which have approximately the same length and the same insulation strength can be connected to form a TAD. However, under this assumption start or end edges which have significantly different insulation strength and/or stripe length profiles cannot be connected. The above suggestion requires more investigation whether the TADs can have different contact intensity at the start and the end. In theory, it could be the case when single enhancer can contact several promoters. Then, as Hi-C is performed in bulk-manner, the enhancer can interact with promoter one and another two with different intensities as, within the population, the number of cells demonstrating contacts with promoter one and demonstrating contacts with pro-

moter two are not the same. At the same time, the detection of partial overlappings between neighboring TADs seems less complicated question. Partial overlapping requires the end TAD edge to be allocated downstream from the start TAD edge and their stripe lengths should be long enough for edges to intersect. Also, the interaction frequency can be assumed to be uniformly distributed within the overlapping area to exclude the complicated scenario of many partial overlappings happen within the vicinity. This direction for further research can be interesting in order to understand whether the partial overlapping is biologically relevant, or it is the algorithm artefacts arising from the intensive interactions between proximal fragments so the neighboring TADs cannot be strictly separated. COrTADo is a novel tool which can also reveal the insulation disbalances between chromatin domains, so making it possible to analyse the epigenetic differences between TADs with distinct interaction densities.

To be precise, in COrTADo we allocate the TAD edges rather than the TAD borders. A TAD edge is defined as a straight line on the Hi-C interaction matrix, which lies on the border of TAD square. So, in simple words, it separates the inside-TAD area from outside-TAD area. COrTADo relies on this definition of TAD edge in a way that TAD edge is most probably is located where we observe the dramatic increase in Hi-C interactions occured within several neighbouring genomic regions (Figure 3.1). So, the TAD edge finding turns into change-point detection problem, where we aim to detect the genomic position associated with the noticeable change in interaction density. Basic change-point algorithms, for example, CUSUM, require specific assumptions including the probability density function to model the Hi-C interactions or decision thresholds (Appendix 3.1). We observed that when we cross the TAD edge, not only the average interaction density changes but the deviation of Hi-C interactions as well. In this case, the change-point algorithm should be modified in order to search for either changes in both variability and average contact frequency, or we need to make an assumption about the relationship between mean and deviation. COrTADo, in contrast, is computationally simple: it is data-oriented and its current implementation in R gives a freedom to a user to manipulate most of the parameters operated during TAD finding.

In the current state, COrTADo is limited to detect only increases in interaction density occurred in upstream direction (COrTADo end) or in downstream direction (COrTADo start). However, with some extend the other architectural patterns can be detected. Recent studies on Hi-C demonstrated the presence of stripe domains – loop anchors which are involved in high frequency interactions with the entire domain probably indicating the positions of super-enhancers which contacts several promoters (Vian et al. 2018; Kraft et al. 2019). On Hi-C contact map, the stripe domain is visually like the TAD with intensive interactions on the edge and significantly less interactions within inside-TAD area. If the DNA fragment which is associated with stripe domain is wide enough, the COrTADo would be able to identify its position when log2mean ratio reaches its local maximum and then its local minimum within the width of the stripe. Also, any chessboard patterns, when interaction frequencies significantly increase and then shortly decrease while moving along the DNA downstream, can be detected by COrTADo after slight modification. The robustness analysis also revealed several other limitations of COrTADo. The algorithm starts from the computation of the log2 mean ratios representing the average difference between neighbouring downstream and upstream genomic regions. When we cross the TAD edge, the average interaction frequency changes in one of the directions. The log2 mean ratio, then, should reach its local maxima or local minima within several genomic positions allocated at the same TAD edge. Due to a data noise and the fact that we use Moving Average approach to estimate the average contact intensity, we are not able to select single genomic position to be a candidate for downstream validation procedure (Figure 3.4, Appendix 3.2). In COrTADo, the window, where most probably the TAD edge candidate position is allocated, is defined under the assumption that TAD edge is sharp, so the local minimum of log2 mean ratio should also be sharp (Figure 3.4). The computation of the log2 mean ratio is based on two MA mean estimation windows: when one of them is fully inside the TAD while another one is still outside, the log2 mean ratio reaches its local extrema. When the TAD edge is more blurry, the difference between MA windows would be less pronounced as both windows would cover interactions on the edge before one of them

reached the inner TAD area. In case of corrected Hi-C matrices, the interaction signal is more smooth: Hi-C matrix balancing removes the variations in the interactions caused by differences in DNA fragments mappability, as well as by other experimental and systematic biases. So, we would observe TAD edges which are visually less sharp, so there is the higher chance that the window where TAD edge is allocated would expand. However, in COrTADo implementation a user has freedom to select the window size width which is then used in local maxima/minima validation procedure. Still, the parameter is single for all TAD edge candidate positions, so if within the single Hi-C interaction matrix we observe both well and poorly insulated TAD borders, we face a danger to remove the significant blurry TAD edges when using narrow window width, as well as we possibly exclude many significant sharp TAD edges when using the wide window width.

Another important point which is associated with MA window is the influence of NAs. Different Hi-C processing pipelines has different solution how they consider non-available data points. Some of the fragments demonstrate NAs when they are associated with low quality reads or the reads are poorly mappable due to, for example, short length or repetitive elements. So, it is not clear, whether some pairs of DNA fragments did not ligated within the Hi-C cell population or they were removed from the analysis. Such NAs can be replaced with zeros, or pairs with NAs can be completely removed from the analysis. In Appendix 3.4, we show that the inclusion of zeros instead of NAs reduces the sample means estimate. It leads to the fact that the MA mean estimates also would be decreased, as a consequence, it is possible that the MA mean differences would become less pronounced and more TAD edges would not survive the validation procedure. Also note, that the inclusion of zeros does not show straightforward effect on variability, so it should not solve the problem of the difference in Hi-C interaction variability between inside-TAD and outside-TAD area.

During the validation, we have to estimate the optimal TAD edge length. The estimation procedure relies on the estimation of the effect size: it is expected to be at maximum when we collect the observations within the TAD edge, when we start to ac-

cumulate observations outside TAD edge (so, the TAD edge length is overestimated), the effect size should decrease. If the length is too long, we face a danger to catch the interactions which are outside the TAD edge, so the insulation strength would be defined incorrectly and, as consequence, TAD edge would be removed as being insignificantly insulated (Figure 3.6). If the length is too short, we face a danger to compare differences in contact intensities between TAD edge and neighbouring genomic regions based on interactions next to diagonal which also leads to incorrect estimation of insulation strength. The current implementation of COrTADo searches for the first dramatic "drop" in effect size to fix the TAD edge length. However, the comparative analysis between borders called on non-corrected versus corrected Hi-C matrices revealed the low robustness of this approach. When the TAD edges are not sharp enough, the effect size would possibly decrease more smoother, which leads to TAD edge length overestimation.

In the Section 3.4, we analysed the COrTADo performance on *Drosophila* BG3 dataset under different conditions. In the next Chapter, we aim to compare the COrTADo performance with another TAD calling tool HiCExplorer (Ramirez et al. 2018). HiCExplorer assumes "head-to-tail" TAD allocation, so the comparative analysis is expected to shed the light on the existence of the unbalanced TAD borders – borders which are insulated either upstream or downstream, so cannot be detected by canonical TAD calling algorithms. However, the analysis presented in the Section 3.4 and Chapter 4 relies on the single dataset and single alternative algorithm. When using single dataset, we face a danger of making conclusions based on the results containing certain biases, in particular, in Hi-C some fragments can interact more frequently than others due to technical biases. Even when we apply the correction methods, we can reduce the effect of systematic errors, but we do not remove them completely. Using different datasets obtained in different laboratories and in different experimental conditions allows us to reduce the effect of the noise and obtain the results which are truthful not on a single but in many datasets. Also, we would be able to better understand the factors that can create inconsistencies in the results between different

datasets – the same way as, for example, the comparison between different replicates demonstrated the role of library size on the insulation strength of COrTADo borders. Also, using other alternative TAD calling methods can improve the research. All algorithms are sensitive to the parameters selection and rely on different assumptions, so only TAD borders which are found by several methods can be called as reliable. However, the algorithms allowing hierarchical TADs and algorithms allowing canonical "head-to-tail" TADs can produce significant differences in the allocation, so comparing COrTADo with more tools – canonical and hierarchical – can improve the confidence of called TAD borders and downstream conclusions.

We also validated the robustness of the COrTADo algorithm by calling TAD borders on merged dataset versus Hi-C generated on each biological replicate, which are accounted for approximately 50% of merged library. COrTADo was able to sufficiently recall most of the border, however, the insulation strength profiles were not similar, so the borders which were defined as strong were mostly unique. Given this findings, the set of rules which we aim to apply in order to distinguish between strong and weak insulation should be selected individually for each data conditions.

We performed the statistical framework based on *Drosophila melanogaster* genome which is noticeably smaller than human or mouse. It means that we were able to validate large amount of TAD edge candidate positions within the reasonable amount of time. In case of human genome, for example, before moving to validation step, we require to remove less probable candidates. Otherwise, we face a danger to validate the number of local extrema which is greater than in case of fly genome in several times. To call local maxima, COrTADo uses the weighted log2 mean average computed within single Hi-C row (for start positions) or single Hi-C column (for end positions). To remove fluctuations which can lead to spurious local extrema candidates, we can implement smoothing the weighted log2 mean average. In particular, the current R implementation of COrTADo also allows the Kernel Regression Smoother. However, the selection of bandwidth parameter is another critical point. Bandwidth define the width of the window within which the signal will be average. With large bandwidth we face a danger to

over-smooth neighboring local extrema and lose significant TAD edges, with the short bandwidth the amount of candidate positions can be excessive. There is no unique way how to define the optimal bandwidth. In order to smooth the signal during the algorithm development, we used the bandwidth that produced the marginal decrease in number of candidate positions which was less than 5%. Although, we did not use signal smoothing when we called COrTADo borders genome-wide. Without proper examination of the TAD edge candidates robustness under the different bandwidth parameters we would not implement it for the downstream analysis.

Chapter 4. Functional differences of balanced and imbalanced insulation in TADs

4.1. Introduction

The non-hierarchical TAD organisation was under the question in recent research in the field of 3D chromatin architecture. Several novel computational approaches were developed to study more complex TAD folding. The major concern is the functional role of the complex architectural structures. Computational methods based on Hi-C data face the challenges of working with noise that comes from the large population of cells. Adding the fact that the chromatin is the dynamic structure, which changes its conformation through the time, we are in danger to detect spurious complex structures that can be experiment artefacts and can show insignificant connection with cellular functioning.

In the Chapter 3, we introduced the COrTADo: the TAD calling algorithm that detects the significant changes in Hi-C interaction intensity moving downstream or upstream along the chromatin. Technically speaking, with COrTADo we can detect the regions demonstrating at least one-sided insulation: the parts of the chromatin that seems to be topologically separated from upstream DNA fragments (for COrTADo starts) and from downstream fragments (for COrTADo ends). So, COrTADo has the freedom to detect any kind of complex TAD topology: nested TADs (hierarchical folding), partially overlapping TADs or breaks between TADs. We also discussed the robustness of the algorithm in different conditions. Based on the analysis from the previous Chapters, we got the motivation for the particular parameters selected to call COrTADo borders based on Hi-C data generated in *Drosophila* BG3 cells. In this Chapter, we sought to identify the difference in epigenetic states of COrTADo borders and borders called by other tool that assumes the “head-to-tail” allocation - HiCExplorer. Using publicly available ChIP data in *Drosophila* BG3 cells (larval central nervous system), we demonstrated the ability of both algorithms to detect two classes of TAD borders - allocated within either active or silent chromatin regions.

4.2. Materials and methods

Data. We used the Hi-C datasets generated by (Chathoth and Zabet 2019) in *Drosophila melanogaster* wild-type BG3 cells at DpnII resolution (median width of 575 bp).

COrTADo. We used the Hi-C dataset that contained approximately 80 M valid pairs (see Chapter 2 and Table 2.1 for more details). We selected the following parameters and thesholds based on statistical framework provided in Chapter 3. We selected pairs that were no more than 1000 bins apart (approximately 500 Kb). We called TADs using Moving Average window size of 10 bins (approximately 5 Kb), FDR correction for multiple testing and p-value threshold of 0.01. For weak borders, we selected candidates with either strength $>$ 0.4 (medium strength) and effect size $>$ 0.1 (small effect) or strength $>$ 0.2 (low strength) and effect size $>$ 0.3 (medium effect). For strong borders, we selected candidates with either strength $>$ 0.6 (high strength) and effect size $>$ 0.3 (medium effect) or strength $>$ 0.4 (medium strength) and effect size $>$ 0.5 (large effect). The window size parameter was selected from the diagnostic plots. The threshold values and multiple testing correction method were selected based on the diagnostic plots and to ensure that we recover a comparable number of TADs as previously reported (Chathoth and Zabet 2019).

HiCExplorer. We used the pre-processed (ICE correction) Hi-C datasets. We called TADs using HiCExplorer with parameters similar to (Chathoth and Zabet 2019) using FDR correction for multiple testing, p-value threshold of 0.01 and a minimum threshold of the difference between the insulation score of 0.04 for weak and 0.08 for strong TADs.

4.3. Preliminary results of comparative analysis between COrTADo and HiCExplorer TAD borders

We identified approximately 3284 start and 3320 end TAD edge positions using COrTADo at DpnII resolution with the average strength of 0.48, which reflected the average 1.4-times increase of intra-TAD interactions compared to inter-TAD. 976 start and 923 end positions were classified as strong and demonstrated the average strength of 0.68

corresponding the average 1.6-times intra-TAD interactions increase (see Methods, Figure 4.1.A). The number of detected TAD borders with weak set of thresholds was significantly higher than reported in other studies (Cubenãs-Potts et al. 2017; Ramirez et al. 2018; Chathoth and Zabet 2019). In particular, when we called TADs using HiCExplorer on the same Hi-C data at DpnII resolution we obtained 2253 TAD borders and 982 of them were classified as strong (Figure 4.1.A).

Although the number of borders detected with stringent parameters was consistent between COrTADo and HiCExplorer, both tools generated the large share of unique borders. Only 33-34% of weak COrTADo borders were allocated within 5 Kb from weak HiCExplorer borders and could be classified as common (Figure 4.1.B, first and second panels). This overlap was even lower under strong set of parameters - the share of common borders was only 24% (Figure 4.1.B, third and fourth panels). Additional 26% of strong COrTADo borders were found within weak HiCExplorer borders but other 50% were absolutely unique.

Most algorithms, including HiCExplorer, detects TADs based on “insulation”: the region that insulates upstream and downstream interactions is most probably the region between domains, i.e. TAD border (see Chapter 2 for more details). As a consequence, TADs are called in “start-to-end” manner: the start of one TAD coincides with the end of the previous one. In COrTADo, the “insulation” can be one-sided - TAD start insulates at least downstream interactions and TAD ends insulates at least upstream interactions. If we suggest the presence of imbalanced insulation where the insulation strength in one direction is greater than in other one, these genomic regions are less probable to be detected by algorithms with two-sided insulation assumption under strong set of parameters. This limitation is clear if we look on the particular genomic regions represented in Figure 4.1.D. The triangular shapes which reflect TADs are visually clear, however the separation between them is not sharp and looks more like “transition regions” where one TAD smoothly goes to another one. HiCExplorer allocates the TAD borders based on TAD-separation score (which is technically the same as insulation score introduced by (Crane et al. 2015)). In simple words, the score represents the

average interaction frequency at each genomic position and when it reaches its local minima we observe the region that demonstrates the “drop” of interactions in comparison with neighbouring upstream and downstream regions. When this “drop” is larger than pre-selected delta parameter (in our case, 0.04 for weak and 0.08 for strong borders), HiCExplorer allocates the TAD border. However, when the “transition region” is relatively long, we would not see a sharp local minimum at TAD-separation score and it would be more like “plateau” - at the end of one TAD there would be a decrease in score followed by approximately monotone signal and then there is an increase at the start of the next TAD (Figure 4.1.D, top panels). COrTADo, in contrast, has high chance to detect such imbalances at least in one direction (Figure 4.1.D, middle and bottom panels). In addition, COrTADo provides the insulation strength statistics that reflects the difference between inside-TAD and outside-TAD average interaction frequencies associated with each particular TAD edge. So, we can detect the TAD borders that show significantly different interaction strength profiles in downstream and upstream directions. Consequently, it raises the question whether the imbalanced insulation is associated with specific chromatin and epigenetic mechanisms or is is an artefact of bulk Hi-C experiment.

HiCExplorer also generated unique borders that were not detected by COrTADo under both strict and relaxed set of thresholds (Figure 4.1.C). Thus, approximately 40% of strong HiCExplorer borders were allocated within 5 Kb from strong COrTADo borders, so were treated as common under strong threshold values (Figure 4.1.C, bottom panel). The share of common borders increased up to approximately 70% when we also searched within weak COrTADo borders (Figure 4.1.C, middle panel). It means that we observed more strong borders which were not common with strong COrTADo borders but were common with weak ones. This increase possibly revealed the difference in stringency of selected thresholds in COrTADo and HiCExplorer. So, the COrTADo thresholds did not allow some of the borders defined as strong with HiCExplorer, while they were possibly detected within all valid COrTADo borders. According to this, approximately 30% of weak HiCExplorer borders, which were not found within weak

COrTADo borders, could be not unique but just removed as the insulation strength of these borders was low to be defined as weak with COrTADo (Figure 4.1.C, top panel).

4.4. Results of the comparative analysis between COrTADo and HiCExplorer

4.4.1. COrTADo and HiCExplorer display the existence of active and silent borders

To investigate the functional role of TAD borders that were obtained using either COrTADo or HiCExplorer, we analysed the presence of several factors: architectural proteins (BEAF-32, Cp190, Chro, CTCF, Rad21 and MED1), polycomb marks (Pc and dRING), accessibility marks (H3 and H4), transcription factors (Pol-II and Trl), nascent RNA and histone modifications (H3K4me1, H3K4me3, H3K9me2, H3K9me3, H3K27me3 and H3K27ac). We used the ChIP-chip datasets generated and pre-processed (M values smoothed over 500 bp) by the modENCODE Consortium. We also used the pre-processed MED1 and Rad21 ChIP-chip (GSE118484) from (Pherson et al. 2019) and 3'NT-seq (GSE100545) from (Pherson et al. 2017). The full list of data GEO accession numbers used can be found in Appendix 2.1. For selected factors, we used ChIP peaks called by modENCODE Consortium. We also merged ChIP peaks datasets for BEAF-32 (GSE20811, GSE32775), Cp190 (GSE20814, GSE32778) and CTCF (GSE20767, GSE32783). For MED1 and Rad21, we used ChIP peaks called in (Chathoth et al. 2021). For nascent RNA, we first summarised 3'NT-seq data obtained from positive and negative strands, then smoothed the score using the Kernel Regression Smoother with normal kernel and bandwidth set at 80. We selected the bandwidth that generates the marginal decrease in number of peaks to be less than 5% when we increased bandwidth by 1.

We combined all TAD borders called by different tools (COrTADo start, COrTADo end, HiCExplorer) under the weak set of thresholds, computed the distance to nearest occupancy peak and split them into 4 categories: less than 2 Kb, 5 KB, 10 Kb and more than 10 Kb. We re-ordered TAD borders and and selected factors based on Hierarchical Clustering with Euclidean distance measure and complete linkage clustering method

for plotting (Figure 4.1.F and G). We grouped factors into five categories that reflected different epigenetic states (Figure 4.1.F).

1. H3K27me3, Pc and dRING are known to be enriched in **Polycomb-associated** regions. Pc and dRING are proteins from PcG (Polycomb group) that are both parts of PRC1 complex. PRC1 binds H3K27me3, leads to chromatin compaction and Pol-II pausing (Min et al. 2003; Lehmann et al. 2012).

2. Pol-II, H3K4me3 (active promoters) (Koch et al.2007; Pekowska et al. 2011; Dong et al. 2012) and nascent RNA together with histone H4 (compact DNA) are strong signatures of **active transcription**. Both insulation proteins BEAF-32 and Chro can also be associated with transcription. Binding sites of BEAF-32 are allocated near TSS, in particulr, BEAF-32 was shown to separate head-to-head genes with different transcription patterns (Yang et al. 2012). Chro cannot not bind DNA independently but it can be recruited by BEAF-32. Note that BEAF-32 and Chro are previously found to be strongly enriched at TAD borders in *Drosophila* (Bushey et al. 2009; Cubenas-Potts et al. 2017). Architectural protein Rad21 is a Cohesin subunit, which binds active promoters and it is suggested to play a role in enhancer-promoter looping (Pherson et al. 2019).

3. The combination of CTCF, histone H3 (compact DNA), GAF and H3K27ac we classify as **bivalent** as we can distinguish between active or silence states when these molecular signals are present together with other factors. So, the enrichment of both H3K27ac and H3K4me1 is a signature of active enhancer, while the enrichment of H3K4me1 in absence of H3K27ac is associated with primed enhancers - inactive enhancers that are primed for future activation (Calo and Wysocka 2013). CTCF is another insulation protein that is mainly known for its interplay with Cohesin in TAD formation in mammals (Racko at al. 2018; Nora et al. 2021). GAF (known also as GAGA factor) promotes promoter proximal pausing in *Drosophila*, so joint enrichment of Pol-II and GAF signals about paused transcription (Chetverina et al. 2021).

4. H3K9me2 and H3K9me3 are both marks of **heterochromatin**. As well as the Polycomb-associated regions, heterochromatin-associated regions indicate the silenced

chromatin but through different mechanisms. Heterochromatin is highly packed DNA regions that is inaccessible for transcriptional machinery while Polycomb repressive complexes block transcription initiation but the region remains accessible for transcription factors binding.

5. In addition to H3K4me1 which role we already mentioned above, we observed enrichment of Cp190 and MED1 at the fifth class. MED1 is a subunit of large Mediator complex that links enhancer-bound transcriptional factors with Pol-II and transcriptional machinery (Immarigeon et al. 2019). Cp190 as an insulator, on the one hand, can block functionally distinct enhancers and promoters from improper interaction, and, on the other hand, promotes formation of chromatin loops that bring distal enhancers to their target promoters. Altogether, the presence of these factors can be referred to signatures of **enhancer** activity.

We applied K-Means Clustering method and revealed that TAD borders detected with either of considered algorithms can be classified into two large classes (Figure 4.1.E and G).

The borders associated with the first class were allocated within the black bar on Figure 4.1.G. These TAD borders were occupied mostly with signatures of active transcription and enhancer activity. We also observed partial enrichment of heterochromatin and bivalent factors. Even if the most of the borders were enriched with H3K4me1, we could distinguish borders that were enriched with H3K27ac and classified as active enhancer regions, but also borders that were depleted with H3K27ac which indicated the presence of primed enhancers. Interestingly, we also observed small sub-group of borders that were highly enriched with Polycomb marks (repression state) while enriched with active transcription marks like Pol-II, nascent RNA and H3K4me3. These signatures indicated that this sub-group belonged to poised enhancers - poised enhancers were found to target major developmental genes that stayed inactive and were activated during differentiation. Overall, we associated the first class of TAD borders with the active chromatin state.

The second class was more associated with the features of silenced chromatin

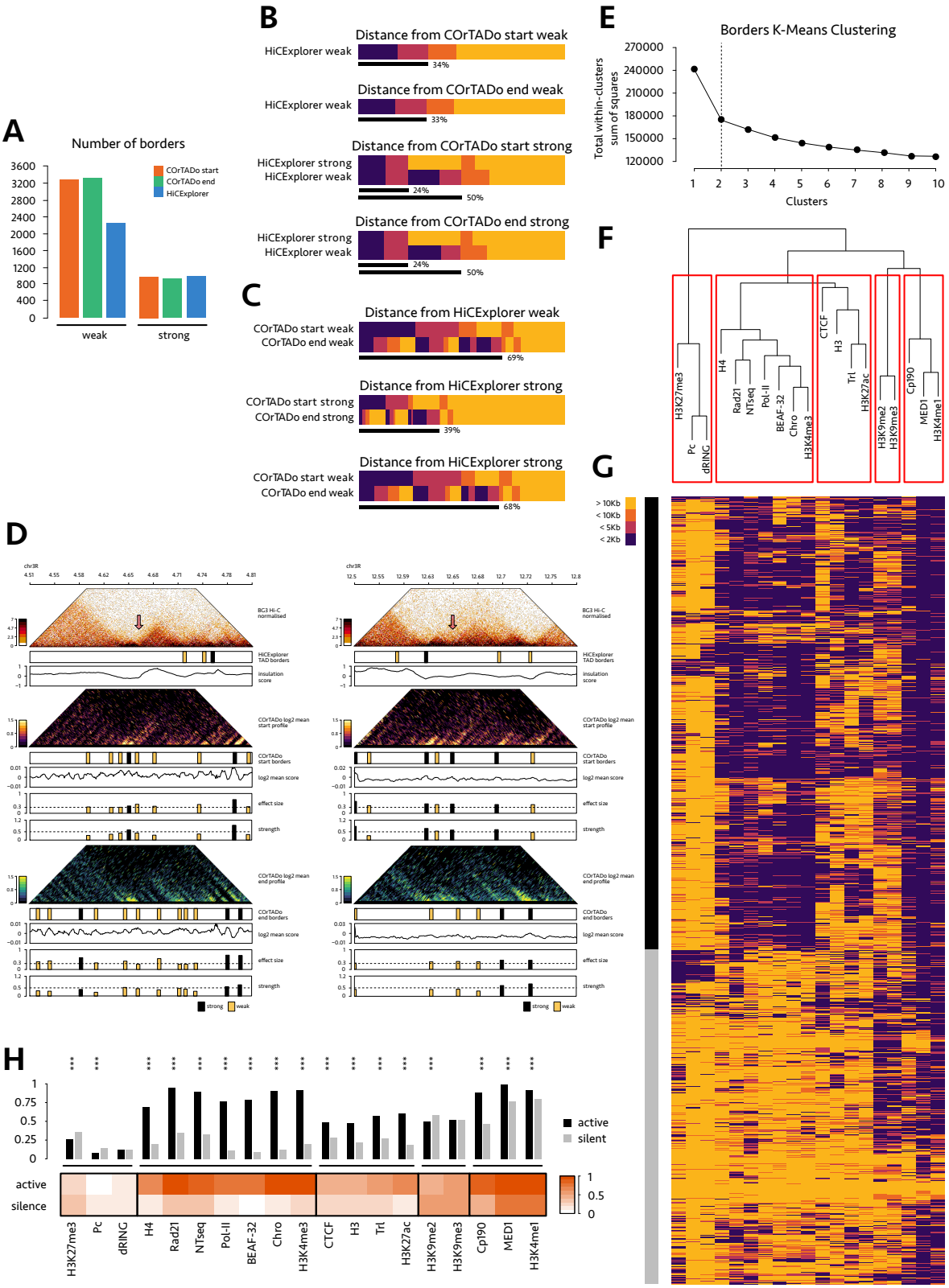

A
Number of borders
COrTADo start
COrTADo end
HiCExplorer
weak
strong
B
Distance from COrTADo start weak
HiCExplorer weak
34%
Distance from COrTADo end weak
HiCExplorer weak
33%
Distance from COrTADo start strong
HiCExplorer strong
HiCExplorer weak
24%
50%
Distance from COrTADo end strong
HiCExplorer strong
HiCExplorer weak
24%
50%
C
Distance from HiCExplorer weak
COrTADo start weak
COrTADo end weak
69%
Distance from HiCExplorer strong
COrTADo start strong
COrTADo end strong
39%
Distance from HiCExplorer strong
COrTADo start weak
COrTADo end weak
68%
D
E
Borders K-Means Clustering
Total within-clusters sum of squares
Clusters
F
H3K27me3
Pc
dRING
H4
Rad21
NTseq
Pol-II
BEAF-32
Chro
H3K4me3
CTCF
H3
Trl
H3K27ac
H3K9me2
H3K9me3
Cp190
MED1
H3K4me1
G
> 10Kb
< 10Kb
< 5Kb
< 2Kb
H
active
silent
active
silence

Figure 4.1. Classification of TAD borders detected by COrTADo and HiCExplorer. **A.** The number of TAD borders detected by different tools. COrTADo identified more TAD borders than HiCExplore under weak set of thresholds. With more stringent thresholds the number of borders was consistent. **B.** Distance from COrTADo borders to nearest HiCExplorer border. Approximately 33-34% of COrTADo weak borders were allocated withing 5 Kb distance. Approximately 24% and 50% of COrTADo strong borders were allocated close to HiCExplorer strong and weak borders, respectively. **C.** Distance from HiCExplorer borders to nearest COrTADo border. Approximately 69% of HiCExplorer weak borders were classified as common with COrTADo as they were allocated within 5 Kb. 39% and 68% of HiCExplorer borders were allocated within 5 Kb from COrTADo weak and strong borders, respectively. D. Representative examples of COrTADo and HiCExplorer TAD calling. Red arrow marked the genomic region which was detected either as COrTADo start or COrTADo end but was not detected by HiCExplorer. Black bard identified borders called with weak set of thresholds, orange bars identified borders called with strong set of thresholds. Panels represented measures that were used as a basis for classification on weak and strong borders (see Chapters 3 and 4 for more details). **E.** K-Means Clustering analysis at joint COrTADo and HiCExplorer borders. The location of a bend is an indicator of the optimal number of clusters. **F.** Dendrogam represented the selected factors that were ordered based on the Hierarchical Clustering and divided into 5 classes considering their epigenetic signatures. **G.** Heat map indicates the distance from joint TAD borders to nearest peak of selected factors. Side bar demonstrates the division of TAD borders into active (black) and silence (grey) regions. **H.** Barplots (top) and heat map (bottom) represent the percentage of active and silent borders which distance to closest peak is no more than 5 Kb. We performed a Fisher's exact test and the corresponding p-values were displayed above the barplots (n.s. $p \geq 0.05$, * $p < 0.05$, ** $p < 0.01$ and *** $p < 0.001$).

(Figure 4.1.G, grey bar). We observed the sub-group of borders with pronounced enrichment of Polycomb-associated factors indicating genomic regions in repressed state that were controlled mostly by Polycomb repression machinery. Majority of borders within the group demonstrated high enrichment of heterochromatin associated marks also indicating repression. In addition, the absence of H3K27ac coupled with the partial presence of H3K4me1 indicates the association of these TAD borders with primed enhancers.

We also compute the percentages of active and silent borders which were allocated no more than 5 Kb from each particular factor (Figure 4.1.H). We performed the Fisher's exact test which confirmed that two classes show significantly different occupancy profiles. The only two signatures that were present in approximately the same share of borders are dRING (Polycomb) and H3K9me3 (heterochromatin). However, other two Polycomb-associated factors H3K27me3 and Pc as well as heterochromatin signature H3K9me2 were present more within silent chromatin class. All other factors were mostly enriched within active chromatin class.

4.4.2. Active borders seemed to be mostly common between COrTADo and HiCExplorer while silent ones are mostly algorithm specific

We separated active and silent TAD borders based on the calling algorithm and analysed the differences in factors binding, insulation strength and the distribution of common and unique borders. Signatures of active and silent chromatin states in COrTADo and HiCExplorer profiles were not visually different: active borders were strongly depleted in Polycomb marks, partially depleted in bivalent and heterochromatin marks and strongly enriched with active transcription and enhancer marks, while silent borders contained small group of Polycomb enriched borders, strong depletion of active transcription and partial enrichment of heterochromatic regions (Figure 4.2., heat maps).

We also visualised the position of borders detected by both tools under the same stringency of the thresholds (Figure 4.2., red bars). Within the COrTADo borders, the common ones demonstrated the tendency to be mostly active, while the unique borders seemed to be associated more with silent state. The difference is more pronounced at weak borders, among the strong borders the connection was not as visually clear. The same separation of common and unique borders was detected for HiCExplorer borders as well.

As COrTADo detects start and end TAD borders separately, we can face several insulation scenarios. First, borders can show **two-sided insulation** when COrTADo start

and COrTADo end were allocated relatively close to each other (we selected the distance of 5 Kb) and defined both as strong or both as weak. This scenario represent the situation when interaction frequencies upstream and downstream from the TAD border are approximately the same. Second, borders can show **imbalanced insulation** when neighbouring COrTADo start and COrTADo end belong to different classes meaning that the average contact frequencies within the neighbouring TADs are significantly different. Third, when the neighbouring COrTADo borders were allocated at more than 5 Kb from each other, we observed the break between two neighbouring TADs and we defined it as **one-sided insulation**. As HiCExplorer detects TADs in "start-to-end" manner, the two-sided and imbalanced insulation COrTADo borders are expected to be common between both tools while the one-sided borders have high chance to be undetected with HiCExplorer, so to be unique COrTADo borders. In accordance with this, the blue bar, which represents the allocation of two-sided and imbalanced insulation borders, seemed to mimic the allocation of common borders (Figure 4.2., blue bar).

In addition, the insulation strength statistics of COrTADo and HiCExplorer borders appeared to correlate with the active/silent chromatin state (Figure 4.2., line plot). In COrTADo, insulation strength represents the average difference in interaction frequency between inside-TAD and outside-TAD areas. So, the greater insulation strength score means the greater segregation of chromatin domain from upstream (for COrTADo start) or downstream (for COrTADo end) fragments. Under the weak set of thresholds, the insulation strength at active borders seemed higher than at silent borders suggesting the contribution of active transcription to the insulation strength of TAD boundaries.

Note that HiCExplorer is an insulation score-based approach which defines the genomic region to be, most probably, a TAD border when it demonstrates significant both upstream and downstream insulation. Technically, it searches for the locus that shows the decline in contacts with close neighbouring DNA fragments comparatively to downstream and upstream regions. So, a low insulation score indicates poorly interacting

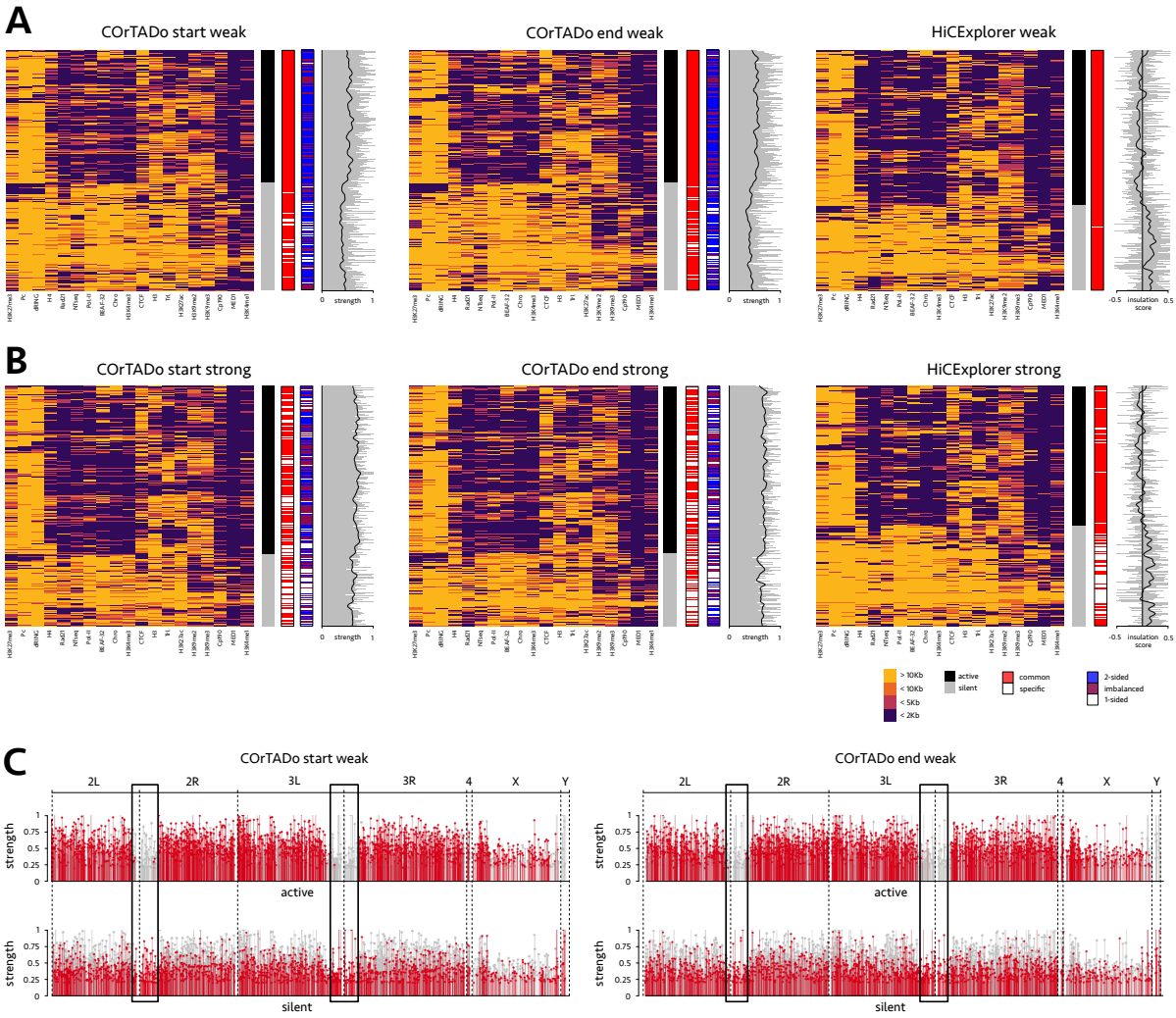


Figure 4.2. Individual TAD border profiles called by COrTADo and HiCExplorer under the weak (**A**) and strong (**B**) set of thresholds. **Heat maps** represent the distance to closest ChIP peak of selected factors, distances are split into four groups: $>$ 10 Kb, $<$ 10 Kb, $<$ 5 Kb and $<$ 2 Kb. **First bar** represents the classification into active (black) and silent (grey) borders. **Second bar** represents the allocation of common (red) and algorithm specific (white) borders. A border is defined as common if it is detected by both tools under the same threshold (weak or strong) stringency. **Third bar** (present for COrTADo profiles only) represents the allocation of 2-sided (dark blue) and imbalanced (violet) insulated borders as described in the main text. **Line plot** represents the smoothed strength (for COrTADo) and insulation score (for HiCExplorer) profiles. **C.** Allocation of weak COrTADo borders within the genome. Black squares indicate the regions visually depleted with active borders but enriched with silent borders.

regions (again, comparatively to surrounding chromatin) and a high insulation score indicates highly interacting regions. As a consequence, lower insulation score indicates the TAD borders with higher insulation, and higher insulation score indicates the TAD borders with lower insulation. According to this, active borders detected by HiCExplorer were associated mostly highly insulated while the silent borders were less insulated.

We also visualised the positions of active and silent COrTADo borders within the genome (Figure 4.2.C). We noticed some genomic regions which demonstrated visually clear depletion of active borders while the silent borders were present there. These were the regions belonged to chromosome centromeres - they were expected to consist of heterochromatin to ensure sister chromatid cohesion and proper chromosome segregation (Przewloka and Glover 2009).

4.4.3. Active borders detected with COrTADo were more robust to tightening of thresholds than ones detected with HiCExplorer

Although the profiles of closest ChIP peaks seemed not significantly different between COrTADo and HiCExplorer, the proportions of active and silent borders were not the same (Figure 4.3.A and B). Under the weak set of thresholds, COrTADo borders were defined as active in slightly more than 50% of cases while for HiCExplorer this share was approximately 60%. We performed Fisher's exact test and confirmed that HiCExplorer demonstrated more pronounced dis-balance between active and silent regions in comparison with COrTADo. Interestingly, with stringent parameters, COrTADo demonstrated more preference towards active regions while HiCExplorer lost more active regions than silent ones. These findings raised the following suggestion. Active borders could be characterised by comparatively high insulation strength in at least one direction, so they would be more robust to tightening of thresholds. Given that HiCExplorer tended to lose more active borders than silent, either active borders would possibly demonstrate lower two-sided insulation strength than silent regions, or active borders would possibly demonstrated more imbalanced insulation and would not "survive" the threshold tightening. The first hypothesis is inconsistent with the fact that

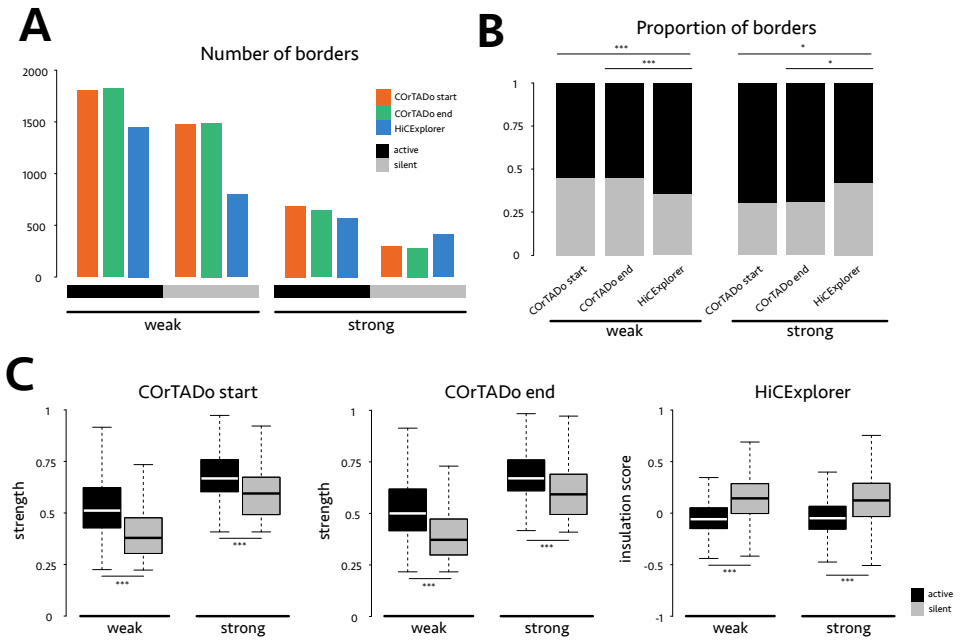


Figure 4.3. Comparative analysis between COrTADo and HiCExplorer borders. **A-B.** Bar plots represent the numbers and the proportions of active and silent borders detected under different conditions. We performed a Fisher's exact test and the corresponding p-values were displayed above the barplots (n.s. $p \geq 0.05$, * $p < 0.05$, ** $p < 0.01$ and *** $p < 0.001$). **C.** Distribution of COrTADo strengths and HiCExplorer insulation scores at active versus silent borders. We performed a Mann-Whitney U test and the corresponding p-values were displayed below the boxplots (n.s. $p \geq 0.05$, * $p < 0.05$, ** $p < 0.01$ and *** $p < 0.001$).

we observed active borders being significantly more insulated in all possible conditions (Figure 4.3.C). We observed the decrease in insulation strength of silent COrTADo borders in comparison with active ones. The insulation score computed with HiCExplorer was higher for silent borders indicating their lower insulation strength. We confirmed the significance of the insulation differences with the Mann-Whitney U test (p-value $<$ 0.001 in all cases). In contrast, share of active COrTADo borders increased with the threshold tightening (Figure 4.3.B). As COrTADo allows the imbalanced insulation and as active borders demonstrated higher insulation strength than silence ones, it was expected that active borders had higher chance to pass the strong set of thresholds.

4.4.4. Selected COrTADo thresholds tended to be more strict than HiCExplorer ones

COrTADo detected approximately 33-34% of borders which were also detected by HiCExplorer, while the remaining 66-67% were COrTADo unique (Figure 4.1.B). The presence of unique COrTADo borders was expected as COrTADo allows the one-sided and imbalanced insulation scenarios which have a low chance to be detected by HiCExplorer. HiCExplorer also produced some unique borders - they were accounted for approximately 31% of all weak HiCExplorer borders. The presence of these HiCExplorer-specific borders is not clear.

Interestingly, in Figure 4.2, we noticed the association between common borders and their active state. For the whole mass of active borders detected by either of tools, most of them were generated by COrTADo only (Figure 4.4.A). The share of HiCExplorer-specific active borders were significantly lower. This disbalance with the preference towards COrTADo-specific borders possibly indicated the presence of numerous active borders with one-sided or imbalanced insulation.

In Figure 4.2, we noticed the association between borders defined as common and borders defined as active for both COrTADo and HiCExplorer, as well as association of common borders with higher insulation. We performed a Fisher's exact test and confirmed that borders, which were found by both tools, demonstrated in average higher insulation strength than the borders which were algorithm specific (Figure 4.4.B). So, the most insulated borders had a higher chance to be detected by both tools.

Both COrTADo and HiCExplorer displayed that common borders were mostly active while the algorithm-specific were mostly silent. Interestingly, the HiCExplorer-specific borders show more balanced separation into active and inactive borders than COrTADo (Figure 4.4.C). When we call borders using COrTADo we initially generate much larger set of all valid borders. Even these borders were validated using statistical test, they can contain some TAD edges which appeared to be valid due to incorrect estimation of optimal TAD edge length (see Chapter 3 for mote details). So, the borders that did not pass the weak set of COrTADo thresholds can be common with HiCExplorer borders.

In agreement with the suggestion, when we compare the allocation of HiCExplorer weak and all COrTADo valid borders, the share of common borders increased up to 90% (Figure 4.4.D). So, the HiCExplorer borders can be defines as common, common which did not pass the COrTADo thresholds but were defined within valid borders and strictly HiCExplorer-specific (Figure 4.4.E).

The borders which were common between COrTADo and HiCExplorer but did not passed the weak set of COrTADo thresholds were found in approximately 30% of HiC-Explorer borders (Figure 4.4.E). Half of such borders was defined as active and another half as inactive, so the weak COrTADo thresholds did not show preference towards active or silent borders. Occupancy profiles of the common borders that did not passed the thresholds are approximately the same with HiCExplorer common borders: we observe the significant difference only in histones H3 and H4, BEAF-32, Chro, CTCF, Cp190 and the modest difference with heterochromatin signature H3K9me3. Aside from H3K9me3 and histones, all factors can be characterised as architectural proteins which tend to colocolise at TAD borders in flies (see Chapter 2 for more details). Altogether, the borders that did not “survive” the weak COrTADo thresholds were less occupied with these architectural proteins. Given this result, we can suggest that the binding of BEAF-32, Chro, CTCF and or/Chro defines the TAD boundary insulation strength. Assuming that the common between HiCExplorer and COrTADo borders probably demonstrate two-sided insulation, these architectural proteins can establish strong insulation in both upstream and downstream directions. Interestingly, BEAF-32 was shown to separate head-to-head genes with different transcription patterns (Yang et al. 2012). It rises the question of the connection between insulation balance and the directionality of transcription. It is possible that when the transcription machinery demonstrates the preference toward unidirectional transcription at at the specific TAD border, we would observe more imbalanced or unidirectional transcription.

HiCExplorer borders, which were not detected within all valid COrTADo borders, accounted for only 10% (Figure 4.4.E) indicating the strictly algorithm-specific borders. These borders were mostly silent - only 25% of them were defined as active. Despite

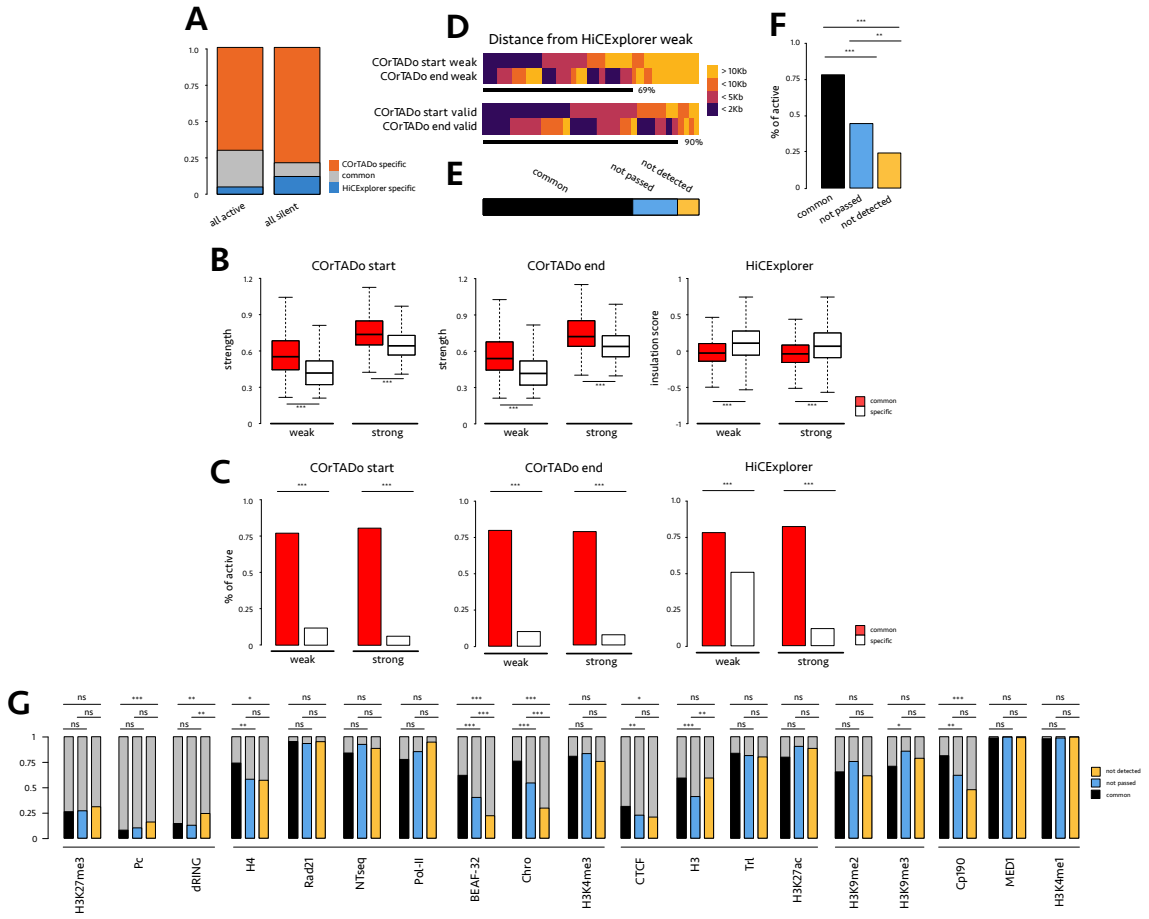


Figure 4.4. Association of common and algorithm-specific borders with the active state. **A.** The share of common, COrTADo-specific and HiCExplorer-specific borders within all active and silence borders detected with either of tools. **B.** Distribution of COrTADo strengths and HiCExplorer insulation scores for common versus algorithm specific borders. We performed a Mann-Whitney U test and the corresponding p-values were displayed below the boxplots (n.s. $p \geq 0.05$, * $p < 0.05$, ** $p < 0.01$ and *** $p < 0.001$). **C.** Barplots represent the share of active borders within common (red) and algoritm-specific (white) borders. Corresponding p-values of a Fisher's exact test displayed above the barplots. **D.** Distance from HiCExplorer weak borders to nearest COrTADo border. We consider both weak borders, as well as all valid COrTADo borders (adjusted p-value < 0.001). **E.** The distribution of HiCExplorer common, not passed and not detected borders as described in the main text. **F.** The proportion of active borders within each of specified class of weak HiCExplorer borders. Corresponding p-values of a Fisher's exact test displayed above the barplots. **G.** Barplots represent the percentage of common, not passed and not detected HiCExplorer borders which distance to closest peak is no more than 5 Kb. We performed a Fisher's exact test and the corresponding p-values were displayed above the barplots.

the silent state, these borders were strongly enriched with the signatures of active transcription including Rad21, nascent RNA, Pol-II, H4K4me3 (active promoters), Trl and MED1, as well as with joint presence of H3K27ac and H3K4me1 indicating the active enhancer regions (Figure 4.4.G). At the same time, the association of the undetected borders with the Polycomb-associated factors (Pc and dRING) is significantly higher in comparison with common borders. The presence of Polycomb marks coupled with strong enrichment of active transcription factors indicated that some of the HiCExplorer-specific borders belong to poised enhancers (inactive enhancers that wait for the activation).

COrTADo and HiCExplorer thresholds aim to remove the borders which demonstrate insignificant difference between inter- and intra-TAD interactions. In HiCExplorer, the TAD separation score represents the average interaction frequency between the genomic region and regions in proximity. So, the minimum difference between the TAD separation score of 0.04 (insulation strength threshold for HiCExplorer weak borders) means the TAD separation score at the genomic region corresponding to TAD border should be different from the score at the neighboring regions by 0.04. In simple, the difference in interaction frequency between intra- and inter-TAD interactions should be at least 0.04. In COrTADo, insulation strength is computed only at candidate genomic positions and not genome-wide. So, the insulation strength threshold of 0.4 (insulation strength threshold for COrTADo weak borders) means that the intra-TAD interaction frequency should be at least 1.4 times more the inter-TAD interaction frequency. So, the difference between interaction frequencies in HiCExplorer is measured through the subtruction (delta) of interaction frequencies while in COrTADo this difference is measured through the division (fraction) of interaction frequencies. COrTADo and HiCExplorer thresholds can become comparable only if we introduce another parameter like inter-TAD interaction frequency. For example, at each level of inter-TAD interaction frequency we can select HiCExplorer and COrTADo insulation strength thresholds which would limit the difference between inter- and intra-TAD interaction frequencies in similar manner. However, it makes the threshold selection even more complicated

procedure.

We demonstrated that when we weakened the COrTADo thresholds we obtained the larger overlap with HiCExplorer borders. If we aim to get more TAD borders which are consistent between both methods, we can possibly review the thresholds and set less strict ones. However, along with the increasing of overlap between HiCExplorer and COrTADo borders, we increase the number of COrTADo-unique borders which are less significant than the borders under the stricter thresholds and which can affect the confidence of the downstream analysis.

4.4.5. Most of the COrTADo borders could be described as active unique borders with imbalanced insulation

As COrTADo detects the insulation of the chromatin domains in either upstream or downstream direction, as well as provides the insulation strength statistics, we can closer look on the sources of balanced in imbalanced insulation. Under the weak set of thresholds, the difference in insulation strengths was pronounced in all pair-wised comparisons of two-sided, imbalanced and one-sided insulation (Figure 4.5.A). In case of imbalanced insulation, each weak border has its pair within strong borders. The increased insulation strength of weak imbalanced borders indicates that they are more closer to strong borders then to the weak ones in terms of insulation strength. One-side insulated weak borders, in contrast, are less insulated indicating that they are just slightly above the insulation strength allowed to for weak COrTADo borders. Borders with balanced insulation are somewhere between them. For strong borders, the allocation is different. The strongest insulation is observed at two-side insulated borders. Then, there are borders with imbalanced and then with one-sided insulation. Based on these observations, we suggest that the borders with weak two-sided insulation are mostly affected by parameter stringency.

Two-sided balance and imbalanced insulation borders did not show significant difference in the share of common borders with the exception for weak COrTADo ends. In particular, approximately 50% of two-sided and slightly more for imbalanced borders

were defined as common between HiCExplorer and COrTADo borders under the weak set of parameters (Figure 4.5.B). However, the share drops to approximately 30% of common borders when the thresholds tightened. As expected, the one-sided insulation is mostly COrTADo-unique, so it supports the idea that COrTADo is able to detect the TAD borders with imbalanced insulation which cannot be detected by tools assuming "head-to-tail" TAD border allocation.

Most of the two-sided and imbalanced borders were associated with active state while one-sided borders did not show specific presence between active and silent chromatin (Figure 4.5.C). However, when we consider strong borders, all insulation classes demonstrate approximately the same preference towards active borders. As the two-sided insulation borders did not change the proportion of active borders when moved from weak to strong, there was no specific preference towards active or silent state within the lost borders. The proportion of active one-sided borders increased with the strong thresholds, indicating the loss mostly of the borders associated with silent state. Imbalanced borders were mostly active. Note that all imbalanced weak borders do not "survive" the threshold tightening. They are defined as weak borders that have their pair within strong border insulated in opposite direction (weak start and strong end and visa versa). So, all borders that are weak, imbalanced and active would not be present within strong borders. On the other hand, borders that are imbalanced, strong and active are the ones which successfully switch the class from weak to strong.

Theoretically, an imbalanced border can be detected in two cases. First, when it is a real insulation imbalance, i.e. the end of the one TAD and start of the next TAD coincide but they demonstrate significant differences in insulation strengths. Second, when it is a spurious insulation imbalance, i.e. the end of the one TAD and the start of the next TAD coincide but the insulation strength was incorrectly estimated. The incorrect estimation can happen because of incorrect estimation of TAD edge length: when the length is overestimated, we consider more long-range contacts which can underestimate insulation strength. Analogically, when the length is underestimated, the insulation strength is overestimated as only intense short-range contacts are con-

sidered. Underestimation most probably takes place when we have nested TADs. We fix the TAD edge length at the level when inclusion of more long-range contacts shows the first dramatic decrease in effect size. Effect size decreases dramatically in case when we include more long-range contact with significantly different contact frequency. When TADs are nested and have coinciding start or end positions, the contact frequency would significantly change when we move from inner TAD to outer TAD, so we expect that the edge of outer TAD would be not detected by COrTADo. In that case, even if the neighboring TAD may have similar average interaction frequency with outer TAD, the border between them would be imbalanced with inner TAD. Overestimation of TAD edge length can also happen when distance decay of log2mean ratio is smooth, and TAD is fading towards its peak. In this case, the effect size would decrease smoothly as well and estimated TAD edge would be too long. If the neighboring TAD has stricter edges, there is high chance of imbalanced insulation detected even if the interaction frequencies are visually approximately the same. The current criterion for TAD edge length estimation is not strict and create possibilities for predilection towards detecting imbalanced borders, so in future research it should be improved. However, using the threshold on absolute value of effect size to remove insignificant borders we lessen the effect of overestimation of TAD edge length which can affect the corresponding insulation strength and analyse only significant border imbalances.

4.4.6. Imbalanced insulation did not revealed possible preference towards specific direction of transcription machinery

The emergence of nascent RNA on either positive or negative strand highlights the direction of the transcription machinery moving. Thus, presence of nascent RNA on positive strand indicates the transcription going downstream while the presence of nascent RNA on negative strand indicates the transcription going upstream. It was previously found that TAD borders in *Drosophila* were associated with divergent transcription: Pol-II could bind the DNA region and transcribe in either downstream or upstream direction. Given these findings, imbalanced insulation could be also possibly

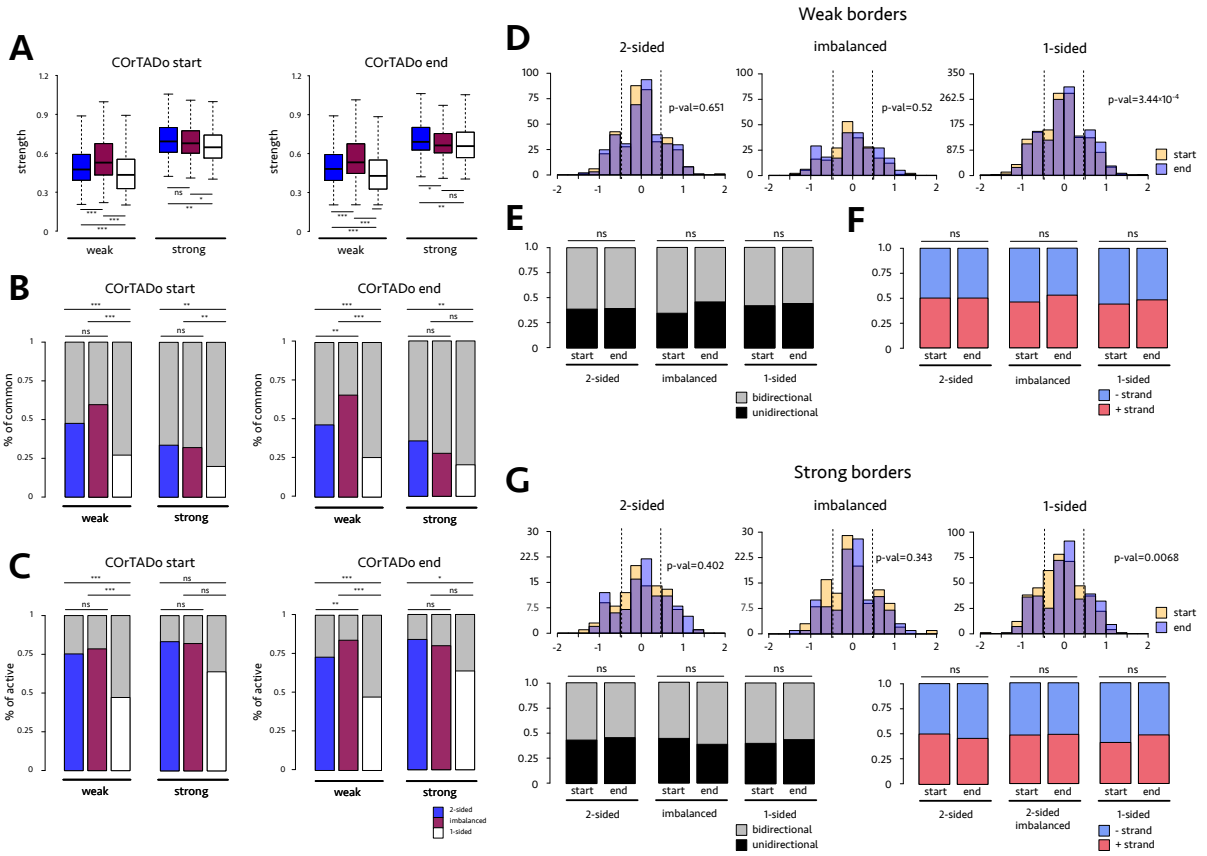


Figure 4.5. Balanced and imbalanced insulation detected by COrTADo. **A.** Distribution of COrTADo insulation strengths for two-sided, imbalanced and one-sided insulation. We performed a Mann-Whitney U test and the corresponding p-values were displayed below the boxplots (n.s. $p \geq 0.05$, * $p < 0.05$, ** $p < 0.01$ and *** $p < 0.001$). **B.** Barplots represent the share of common borders within each specified insulation class. **C.** Barplots represent the share of active borders within each specified insulation class. Corresponding p-values of a Fisher's exact test displayed above the barplots. **D.** Histogram represented the distribution of directionality scores in two-sided, imbalance and one-sided insulation case. Orange color represented COrTADo starts and violet color represented COrTADo ends. Directionality score was calculated as log10 of average nascent RNA positive strand signal within 500 bp window 500 bp away downstream from the COrTADO border over average negative strand signal within 500 bp window 500 bp away upstream from the COrTADo border. The vertical lines represent cut-off values at -0.47 and 0.47 to distinguish between unidirectional and bidirectional borders.**E.** Bar plots represented the shares of unidirectional (grey) and bidirectional (black) borders in each specified scenario. **F.** We split unidirectional borders depending on preference towards positive strand (red, directionality score > 0.47) and towards negative strand (blue, directionality score < -0.47). **G.** Same as (D-F), for strong COrTADo borders.

associated with preference towards specific transcription direction. So, high insulation at COrTADo start represented regions with Pol-II preference towards downstream transcription while highly insulated COrTADo end could be associated with transcription of upstream regions.

We first computed the mean nascent RNA levels considering 500 bp window that were 500 bp downstream away on the positive strand and 500 bp upstream away on the negative strand. Then, we computed directionality score as log10 ratio of mean nascent RNA on the positive over negative strand. Borders with directionality score being lower than 0.47 were classified as bidirectional. The value of 0.47 represents slightly less than three times more transcription on positive strand than on negative strand.

We computed the directionality scores for borders classified as two-side insulation, imbalance insulation and one-sided insulation (Figure 4.5.D). When borders did not show specific choice of transcription direction we defined them as bidirectional. The proportion of bidirectional borders was slightly dominated by unidirectional ones in all scenarios (Figure 4.5E.B). The imbalanced start and end borders displayed slightly different distribution of bidirectional and unidirectional borders. Interestingly, when the weak imbalanced starts showed slightly less share of the unidirectional borders and ends showed slightly more, within the strong borders the distributions were opposite. However, the differences were not significant. Also, we observed slight preference of upstream transcription (negative strand) at COrTADo starts and downstream transcription (positive strand) at COrTADo ends (Figure 4.4.F). We would expect to see more pronounced association in one-sided insulation case. Although, all the differences seem to be insignificant. Overall, we did not get enough evidence to support the idea that the imbalanced insulation correlates with the specific direction of transcription machinery. On the other hand, the insignificance of the results can be related to the fact that, first, the nascent RNA and Hi-C datasets were obtained in different experimental condition, so the association can be not so pronounces. Also, we did not exclude the silent borders and borders with no transcription from the analysis. One-sided insulation

was shown to be associated with both active and silent borders while imbalanced and two-sided insulation showed more preference towards active borders (Figure 4.5.C). Therefore, the expected disbalance between unidirectional and bidirectional borders was not clear.

4.4.7. Polycomb-associated borders were mostly COrTADo-specific

Although the separation between active and silent borders was visually clear, there was a group of silent borders that demonstrated some distinctive features. These borders were strongly enriched with Polycomb marks while other silent borders were approximately everywhere Polycomb depleted. Using methods of Hierarchical Clustering with one additional cluster we separated these borders for further analysis. We performed Fisher's exact test and confirmed that these borders had significantly different association of chromatin binding factors (Figure 4.6.B). In comparison with other silent borders, Polycomb-associated borders showed the significantly different level of occupancy in case of Pol-II (transcription), BEAF-32 (insulation protein) and H3K9me3 (heterochromatin) enrichment. In case of other factors these borders showed the largest difference in H3K27me3, Pc and dRING. Based on these results, we suggested that the Polycomb-associated group was combined of regions that were associated with joint Polycomb and heterochromatin repression machinery and regions that were associated with Polycomb repression only, while other silent borders were mainly heterochromatic.

The number of Polycomb-associated borders called by COrTADo was higher than by HiCExplorer under weak set of thresholds and approximately the same under strong set of thresholds (Figure 4.6.A). The proportion of Polycomb borders was moderately higher for COrTADo under the weak threshold values, for the strong borders the proportion was the same for both tools (Figure 4.6.C).

Most of Polycomb borders called by COrTADo were algorithm-specific, while for HiCExplorer the share of unique Polycomb borders accounted for approximately 25% (Figure 4.6.E). At the same time, strong Polycomb borders were both COrTADo and

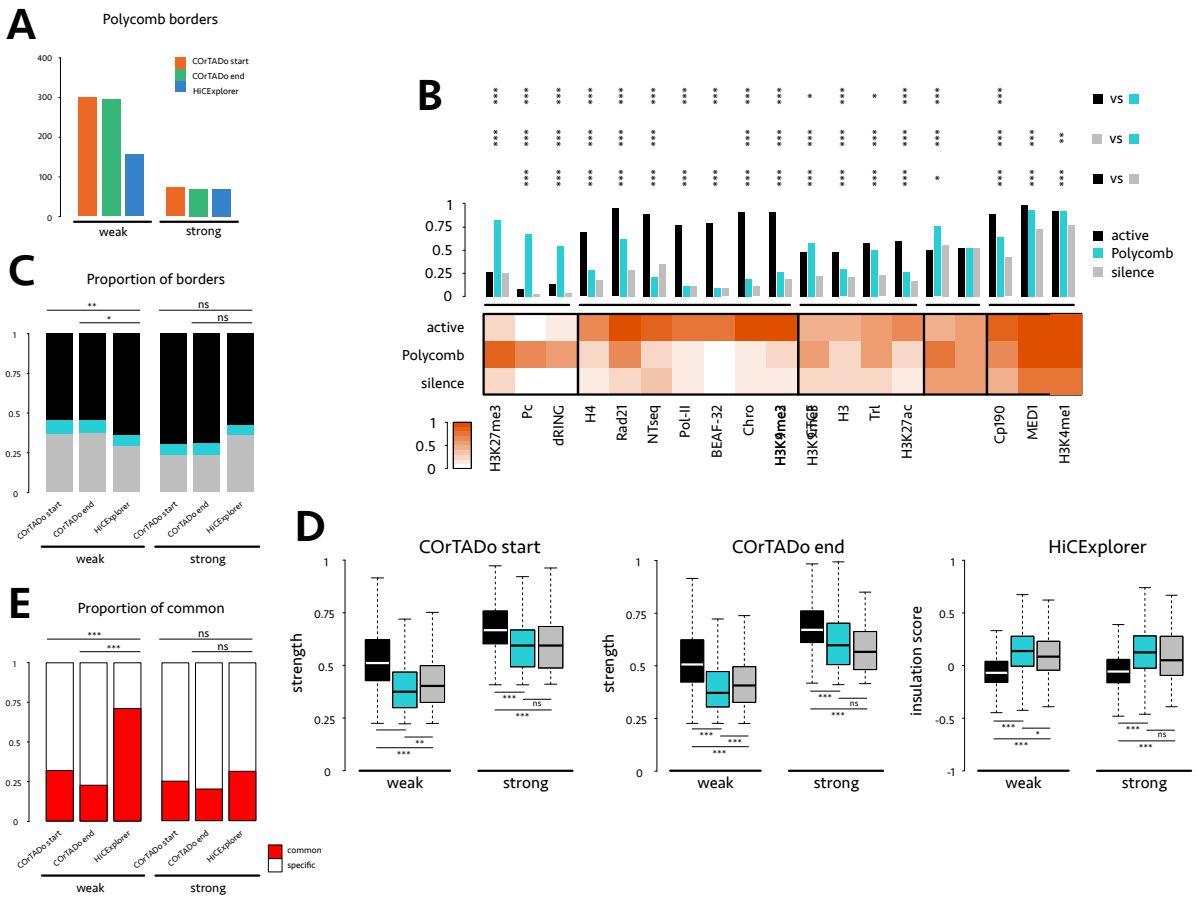


Figure 4.6. Polycomb-associated borders and their difference from remaining silent borders. **A.** Bar plot represents the number of Polycomb-associated borders called with strong and weak thresholds using COrTADo and HiCExplorer. **B.** Bar plot (top) and heat map (bottom) represent the percentage of active, Polycomb-associated and remaining silent borders which distance to closest peak is not more than 5 Kb. We performed a Fisher's exact test and the corresponding p-values were displayed above the bar plots (n.s. $p \geq 0.005$, * $p < 0.05$, ** $p < 0.01$ and *** $p < 0.001$). **C.** Bar plot represents the proportions of active, Polycomb-associated and remaining silent borders. **D.** Distribution of COrTADo insulation strengths for active, Polycomb and remaining silent insulation. We performed a Mann-Whitney U test and the corresponding p-values were displayed below the boxplots (n.s. $p \geq 0.05$, * $p < 0.05$, ** $p < 0.01$ and *** $p < 0.001$). **E.** Barplots represent the share of common borders within Polycomb-associated borders.

HiCExplorer-specific. So, HiCExplorer lose more common Polycomb borders with the threshold tightening which can possibly indicate the association of Polycomb borders with imbalanced insulation.

4.5. Summary and discussion

In this Chapter, we performed the comparative analysis between TAD borders called using the insulation score-based approach implemented as a part of HiCExplorer tool and borders called by COrTADo. We introduced the COrTADo as we aim the investigate the functional role of complex chromatin architectural patterns such as nested TADs, partially overlapping TADs, breaks and imbalances in TAD boundary insulation. The first three groups of complex topologies required the reconstruction of TAD edges. However, the imbalanced insulation between neighbouring chromatin domains can be studied straightforward. COrTADo calls start and end TAD borders separately, so we can explore the upstream and downstream insulation and classify borders into balanced (or two-sided insulation), imbalanced (where the insulation strengths of upstream and downstream borders are significantly different) and one-sided insulation (when either start or end is present). However, due to bulk manner of Hi-C data, the imbalances can appear due to either the differences in mappability of regions colocolised with TAD borders or the dynamic changes of chromatin architecture within the population of cells.

Most of the borders indicated by HiCExplorer were either common with COrTADo, common but these borders did not pass through the weak set of COrTADo thresholds and were HiCExplorer-specific. The borders were common in 90% of HiCExplorer borders which is not surprising as COrTADo allows both balanced and imbalanced insulation detection while HiCExplorer can indicate only balanced insulation scenario. In particular, HiCExplorer-specific seemed to mostly associate with primed enhancers. The borders that did not passed the weak COrTADo thresholds showed approximately the the same epigenetic features as common ones, except the depletion of architectural proteins BEAF-32, Cp190, CTCF and Chro was noticeable. It reveals the importance

of these proteins binding in maintaining the insulation strength.

When the threshold values tightened, COrTADo lost mostly silent borders which is expected as they demonstrated comparatively low insulation strength. However, HiCExplorer border lost mostly active borders. Given this, we suggested the association of active borders with imbalanced insulation - with more stringent parameters only one of the borders would be able to cross the thresholds while another one would be lost. HiCExplorer, in turn, would not be able to detect this imbalances. We supported this observations when looked at the association between active borders and insulation balance.

Imbalanced insulation coupled with active chromatin state can possibly reveal the role of transcription directionality in maintaining the one-sided insulation. However, the nascent RNA data did not show enough support to this hypothesis.

We observed the large group of borders that were active and one-side insulated as well, however we detected these borders not only with COrTADo but also with HiCExplorer. The ability of HiCExplorer to detect so many borders that we classified as one-sided signals about different stringency of thresholds selected for HiCExplorer and COrTADo. In case of imbalanced insulation, COrTADo removed candidate with lower insulation while HiCExplorer allowed this level of insulation to be detected. So, further weakening of COrTADo thresholds could possibly lead to better overlap between HiCExplorer and COrTADo. Also note here that HiCExplorer strong and weak thresholds did not produce the borders with significant difference in insulation score. Delta thresholds that were used to distinguish between weak and strong borders allowed to select genomic regions that demonstrated the desired difference in interaction frequency with neighbouring regions, so this parameter did not affect the strength of insulation itself. In contrast, COrTADo had two parameters that controlled both difference with neighbouring regions (effect size) and insulation strength itself. Based on this, we could expect that thresholds in COrTADo were stricter than thresholds in HiCExplorer.

Discussion and future research

5.1. Summary and discussion

Chromatin architecture and, in particular, topologically associated domains (TADs) are highly conserved between cell types, developmental stages and even different organisms which suggests their importance for proper gene regulation maintenance and cell functioning. Previously it was demonstrated that the chromatin topology disruptions, which were caused by changes architectural protein binding, affected the changes in gene expression patterns (Lupianez et al. 2015; Taberlay et al. 2016; Kragesteen et al. 2018). In particular, in Drosophila we have shown that the knockdown of BEAF-32, Cp190, Chro and Dref caused rearrangements in normal chromatin architecture and changes in gene expression. Also, we found that the genes altered their expression when they were associated with massive perturbations in TAD allocation.

The cause-and-effect relationship between gene expression and chromatin topology is not trivial. Along with the research showing the disruption of chromatin architecture coupled with the changes in gene expression, there is also some evidence on conformational rearrangements which were not coupled with transcriptional changes, and vice versa – no changes in TAD organisation in the presence of transcriptional alterations (Ghavi-Helm et al. 2019; Ing-Simmons et al. 2021). Based on the analysis on BEAF-32 knockdown and Cp190 and Chro double knockdown presented in Chapter 2, we showed that the TADs were found both in active and silent chromatin which is consistent with previous studies on Drosophila (Ramirez et al. 2018). However, the TAD borders which were maintained in the mutants were mainly associated with active state while the borders which were lost in the result of knockdowns were associated with silent state. It suggests that in the absence of architectural proteins the chromatin architecture still can be sustained with the support from the transcription machinery but, in absence of such support, the chromatin architecture is disrupted when the architectural proteins are not present. These results and suggestions are consistent with the hypothesis that architectural proteins can maintain the topology which can “guide” the transcription machinery and maintain proper contacts between regulatory elements

and/or insulate the improper ones. Then, further in the development, the transcription can maintain the architecture even without such 'guidance' from the architectural proteins. This hypothetical model is consistent with the dynamics of Runx1 gene expression and topological changes associated with CTCF binding published in (Owens et al. 2021). Although, it is unclear whether the transcription is enough when the topology is already established or for how long the transcription can maintain the chromatin architecture. These questions are still not addressed. In addition, significant inconsistencies between different studies whether the gene expression defines the chromatin segregation between TADs or TAD organisation establishes the proper gene regulation still does not provide clear understanding on TAD formation mechanism.

TAD allocation is sensitive to the several factors including the experiment condition, data processing and TAD calling algorithm. Some of the TADs can be detected in the result of the noise when the associated DNA fragments demonstrate significantly higher interaction frequency due to improper ligation, PCR amplification or mapping to the genome. So, the additional steps in order to remove TADs which were allocated because of the biases are required, otherwise the analysis of the biological functions of the TADs would not be clear. However, we can face a risk that along with spurious TADs we can remove the real ones which also can affect the downstream analysis. For example, excluding excessive number of TADs can affect the comparative analysis between wild-type and mutant datasets which probably would have similar TAD allocations, so small rearrangements between two conditions would be missing. However, in our analysis in Chapter 2 even with multi-stage robust analysis which removes the many TADs which can be affected by the TAD calling tool parameter selection and differences in library sizes between datasets, we still observe massive rearrangements between wild-type and different knockdowns.

Canonically, TADs are assumed to have "head-to-tail" allocation meaning that the end of the one TAD should coincide with the start of the next one. This assumption ignores the complex architectural patterns as breaks, nested TADs, or partial overlapping between neighbouring TADs. Recent TAD calling algorithms started to overcome

this limitation allowing hierarchical folding of TADs (Rao et al. 2014; Durand et al. 2016; Weinreb and Raphael 2016; Forcato et al. 2017). Other complex architectural structures require TAD starts and ends to be known, so we require the algorithm, which can allocate these starts and ends separately from each other. To be more specific, we require the allocation of start and end TAD edges – the set of loci which interacts within the same TAD and separates the inter- and intra-TAD interactions. Visually, these edges are described as imaginary line segments on the border of TADs on the Hi-C contact maps. Reconstruction of complex topology from the TAD edges can potentially shed the light on whether the visually detectable complex TAD folding is inconsistent with canonical “head-to-tail” TAD allocation because the “head-to-tail” assumption unable to describe complex chromatin organisation or it is an artefact of the bulk manner of Hi-C experiment. Further in the Section 5.2.1, we propose simple basic model which has a potential to separate real complex chromatin interactions from spurious ones in case of partial overlap between neighbouring TADs.

Understanding of complex chromatin architecture and accurate TAD allocation which represents the real DNA-DNA interactions is also important for Hi-C pre-processing and removing the biases. In general, we assume the genomic distance between interacting fragments to be an important explanatory variable at Hi-C interaction frequency modelling. There are other important factors which include GC-content, DNA accessibility or presence of transposable elements. However, these other factors seem to affect the chromatin architecture and presence of TADs. So, using the affiliation of DNA fragments to the same TAD can possibly cover all these factors along with other hidden factors which were not presented in the Hi-C associated studies up to date. Further in the Section 5.2.2, we highlighted some potential scenarios which can possibly describe the functional relationship between Hi-C contact frequency, genomic distance, and TAD affiliation, which, in turn, can become a basis for different statistical modelling problems associated with Hi-C data such as removing the biases or detecting interactions which are significantly different from the expected Hi-C contacts and can possibly indicate the presence of the loops.

Although, the reconstruction of complex TAD folding based on TAD edges requires, first, an algorithm to detect the TAD edges and in the Chapter 3 we introduced COrTADo. COrTADo detects TAD edges based on the changes in insulation strength which measures the differences in contact intensity between DNA fragments belonging to the same TAD and between fragments outside single TAD. As COrTADo detects TAD starts and ends separately from each other, it creates a basis for downstream analysis on insulation imbalances. Under the canonical “head-to-tail” TAD allocation, when neighbouring TADs demonstrate the significantly different inter-TAD average contact frequencies, we face a risk that TAD border would not be allocated even if they are visually clear. Also, if neighbouring TADs are not clearly separated and have long “transition” region, there is a high chance to lose such TAD border as well.

Applying COrTADo on *Drosophila* BG3 wild-type Hi-C data, we demonstrated that COrTADo was able to detect the majority of TAD borders detected by other tool, HiCExplorer, which assumes “head-to-tail” allocation. However, COrTADo also detects many unique borders which were expected assuming the presence of borders with imbalanced insulation. Unique COrTADo borders were actually detected with other tool but demonstrated lower insulation strength. They also demonstrated significantly lower binding of several architectural proteins including BEAF-32, Chro, Cp190, CTCF. Based on these findings, we suggest that the architectural proteins binding can affect the insulation strength of TAD borders – more protein presented nearby the TAD border, more segregated the corresponding chromatin into domains. We also showed the importance of some of these proteins for TAD border maintenance in Chapter 2. Borders, which were classified as lost during BEAF-32, Cp190 and Chro knockdowns, demonstrated higher TAD separations score, which is the same as lower insulation strength, and low or no binding of architectural proteins which does not contradicts with the results in Chapter 4.

Most of the COrTADo borders were classified as active unique borders with imbalanced insulation. As being active, we can expect that most of COrTADo borders would be maintained in the result of BEAF-32 single knockdown or Cp190 and Chro double

knockdown based on the results in Chapter 2. However, we predict that their insulation strength profiles would be affected – the maintained borders can significantly lose their insulation strength in absence of architectural proteins. As we did not analyse whether the protein binding have any preference towards COrTADo end or start border to affect the significance of insulation disbalance, we don't have any prediction on the distribution of balanced and imbalanced border after the protein knockdowns. The position preference has a potential to become an important factor for establishment of insulation imbalance because, first, we showed that the binding is important for insulation strength and, second, the directionality of transcription was not found to be significantly important to define the insulation imbalance. Altogether, the comparison of COrTADo borders in wild-type and BEAF-32, Cp190 and Chro knockdowns can help us to understand the exact mechanism of TAD border insulation establishment and its connection to transcription without being biased towards only borders with balanced insulation only.

5.2. Future research

5.2.1. Reconstruction of complex TADs is required for single cell architecture prediction

We assume that a simple chromatin loop could have a toroidal or a plectonemic shape (Bjorkegren and Baranello, 2018) (Figure 5.1). In the case of a toroidal loop, anchor loci interact most of the time, while loci that are in between do not interact with each other. So, we expect that a sub-TAD that represents a simple toroidal loop could be treated as a triangle having a peak point and no interactions inside. It is unrealistic representation, but its simplicity helps to understand the behaviour of plectonemic loop and how the aggregation of supercoiled loops is visualised at Hi-C matrix. In the case of a plectonemic loop, anchor loci and all loci in between interact more frequently than by chance. Sub-TAD that represent this kind of loop could be treated as a triangle with no peak but having approximately the same level of interaction intensity inside (we assume that a simple plectonemic loop is small enough to have approximately the

same number of interactions between loci that belong to the loop).

At the next step we want to understand how two simple loops could be aggregated into more complicated chromatin structure. As we matched loop shapes and triangle representations, we can imagine how the chromatin architecture will look like when: (i) two triangles have one common point; (ii) two triangles partially overlap. If two triangles have one tangent point, it means that two simple loops should have one common anchor locus. As a consequence, all three anchor fragments should be in close spatial proximity. It means that the whole structure could be aggregated into one large TAD with a peak on the top.

If two triangles partially overlap, it means that there are several loci that belong to two simple loops. These loci are located between the anchor fragment from one loop and the anchor fragment from another loop. This chromatin region is expected to have a higher interaction intensity because it faces interactions from two loops at the same time. The aggregating TAD in this case is expected to be non-peaked as the furthest anchor loci are not in close spatial proximity, so they will not interact more often than by chance.

We described simple structures that could be theoretically observed in chromatin architecture in a single cell. The next step is to understand what we expect to see on the resulting Hi-C heat map, based on the ideas described above.

On the one hand, we keep in mind that the Hi-C experiment takes a huge amount of cells and, as a consequence, produces the contact map in a population average manner. It means that it is still possible that some loops are not tightly conserved: one particular loop could appear on some fixed chromatin region but have some location variation (Figure 5.2., model 1). Because of the aggregation of single profiles within the population, the Hi-C contact map is expected to have two partially overlapping triangles. If we assume that the first loop has some contact intensity λ_b and the second loop has some contact intensity λ_p, we expect that the overlapping region should have the λ_v contact intensity where $\lambda_v = \lambda_b + \lambda_p$ because of the population averaging manner of the experiment. So, this is spurious overlapping sub-TADs.

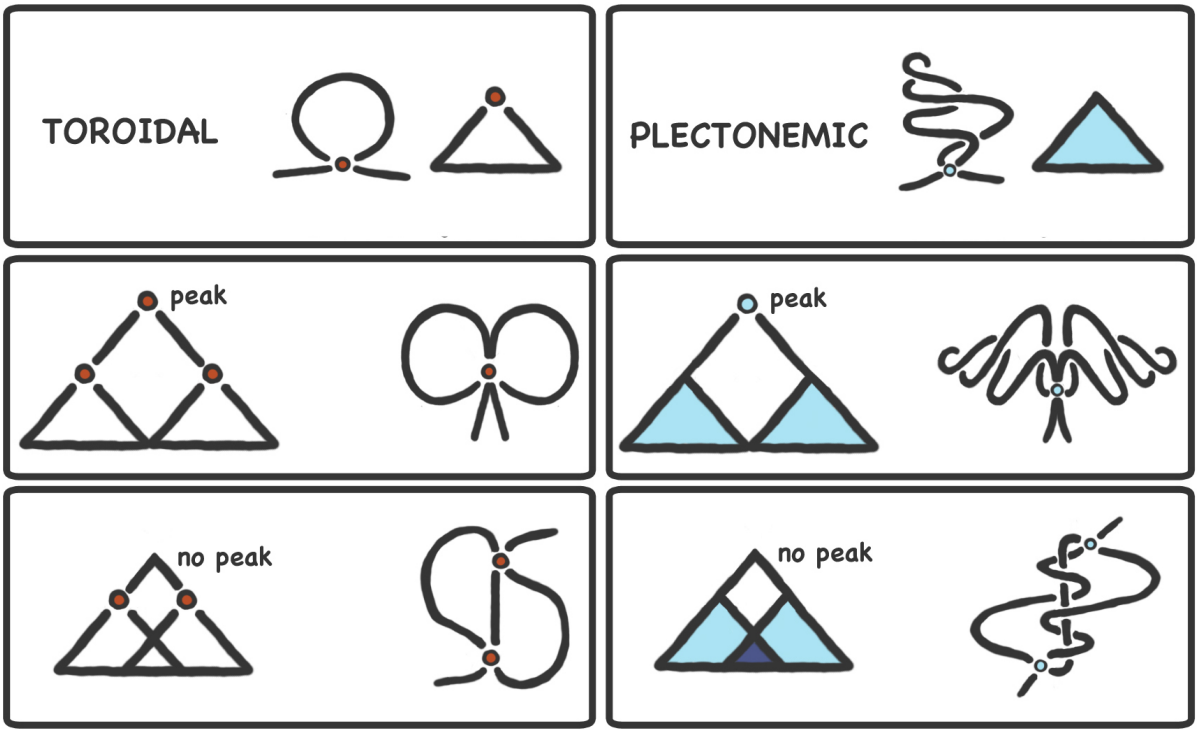


Figure 5.1. Schematic representation of loop aggregation.

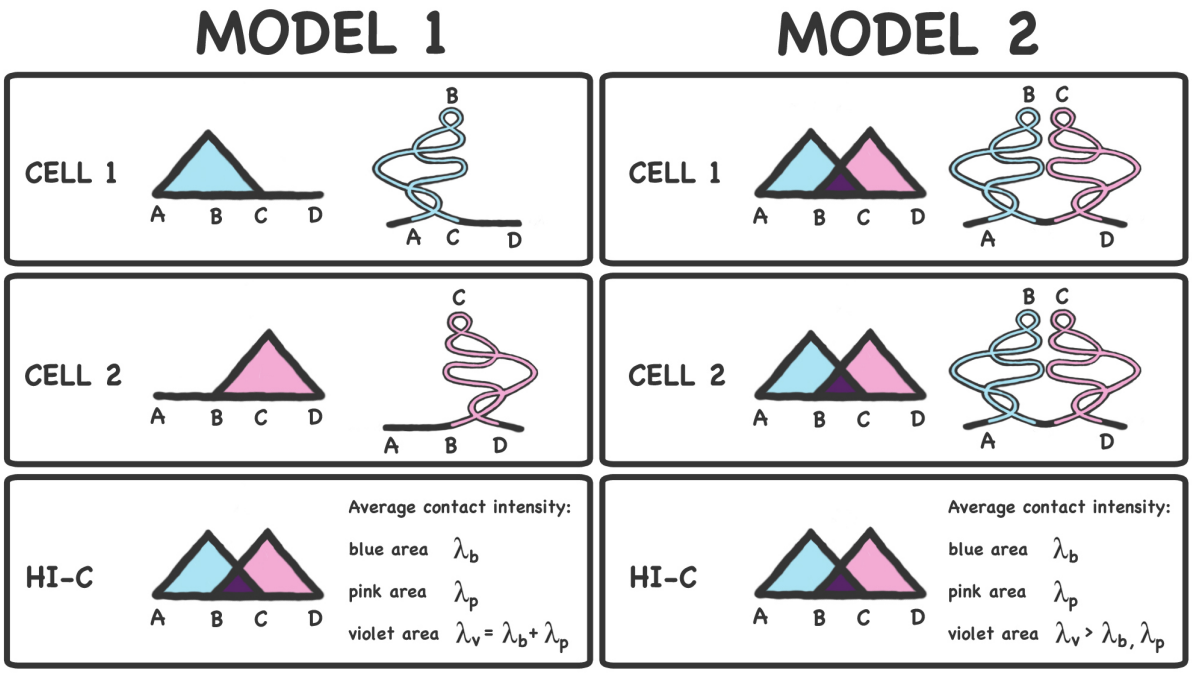


Figure 5.2. Schematic representation of single cell scenario.

On the other hand, in theory, the combination of two simple loops that are represented by two partially overlapping triangles could be conserved across cells (Figure 5.2, model 2). We expect that the models should be different by the contact intensity value: in real overlapping sub-TADs the contact intensity λ_v is defined only by the specific loops architecture (and not by the population averaging manner like for spurious overlapping) and, as a consequence, may be different from $\lambda_b + \lambda_p$.

In order to access these different scenarios, we require a TAD calling tool which allows the partial overlapping to be detected. COrTADo is able to allocate start and end positions separately which can become a basic for reconstruction of complex chromatin topology. COrTADo is able to allocate start and end positions separately which can become a basic for reconstruction of complex chromatin topology. Complex patterns such as partial TAD overlapping, breaks and nested TADs can be produced by either chromatin contacts which can have a specific biological function and it was not studied properly before, or it can be a product of noisy interactions from population of cells in Hi-C experiment and it can help to improve the existing TAD calling algorithms to obtain more robust TAD allocation. There is still important methodological question how we can move from separate start and end TAD border positions to complex chromatin architecture, but COrTADo created a basis for this move, as well as it can help to answer other research questions like the existence of imbalanced insulation between neighbouring TAD borders and reasons behind such phenomenon.

5.2.2. Importance of topologically associated domain allocation for Hi-C interaction modelling

If the locus i interacts with another DNA locus j for randomly selected cell, it is expected to be cross-linked and ligated during Hi-C only once with this specific locus j, meaning that locus i may be involved only in single interaction in each cell. Note that it is not true for diploid organisms which has two copied of DNA. So, the correct way is to say “may be involved in not more than two interactions in each cell”. However, from the mathematical point of view, one diploid cell can be treated here analogically to two

haploid cells. If the locus i does not interact with any other locus genome-wide, we can call it as self-interaction event and to assign the number of such interactions to the diagonal entry (i, i). Note that self-interactions are expected to happen rarely and commonly treated as experiment artefacts rather than real absence of interactions with other DNA fragments.

Mathematically speaking, for a randomly selected cell, the DNA fragment i successfully interacts with exactly one of n other fragments. Suppose, the contact probability between locus i and locus $j = 1, ..., n$ is defined as p_j. If the random variable $X_{(i,j)}$ indicates the number of times when locus i contacts locus j in a population of T independent cells in Hi-C experiment, the vector $\mathbf{X}_i = (X_{(i,1)}, X_{(i,2)}, ..., X_{(i,n)})$ follows a Multinomial distribution with parameters T and $\mathbf{p}_i = (p_{(i,1)}, p_{(i,2)}, ..., p_{(i,n)})$. Then, the expected number of interactions between locus i and locus $j = 1, ..., n$ is

$$\mu_{(i,j)} = E(X_{(i,j)}) = Tp_{(i,j)}$$

It is important to note here that the interaction outcomes $X_{(i,1)}, X_{(i,2)}, ..., X_{(i,n)}$ are dependent as they must be summed to T. The first limitation of the proposed model comes from the fact that the sum of observations $x_{(i,1)}, x_{(i,2)}, ..., x_{(i,n)}$ most of the time is less than the population size T. The reason is that we lose huge amount of pair-wise interactions during Hi-C experiment and during following pre-processing. In order to implement the proposed Multinomial modelling, additional assumptions for lost interactions has to be made.

In Hi-C associated studies, we assume that the interaction frequency between DNA fragments depends on the genomic distance between them - fragments that are close to each other on the DNA strand tend to interact more frequently than fragments that are far away from each other. It is the reasonable assumption for Hi-C maps generated using cells in late prophase when chromosomes become condensed and chromatin architectural features are lost. Heat maps of late prophase Hi-C are close to “gradient” pattern - interaction intensity smoothly changes from high to low when we move from the diagonal to the top of the map.

This pattern is violated during the interphase. We observe the appearance of differ-

ent architectural structures like TADs and loops. Interactions inside TADs happen more frequently than interactions outside TADs. If the pair of loci belongs to the same TAD, we expect to observe more interactions between them comparing to the pair of loci that do not belong to the same TAD but located on the same genomic distance from each other. Generally speaking, when we model either the Hi-C interaction frequency or contact mean, both the genomic distance between fragments and their affiliation to the same TAD should be taken into account.

Note, that existing statistical models for Hi-C data highlights other factors besides the genomic distance to be significant. Such factors include, for example, GC-content, DNA accessibility or presence of transposable elements (TEs). Along with existing studies, we demonstrated that TADs were found to be associated with complex interplay between transcription activity, architectural proteins and other DNA-binding molecules, which, in term, could correlate with DNA accessibility or transposable elements. Thus, many CTCF binding sites were found within TEs, so there could be a link between TEs and chromatin architecture. Or, another example, the binding of transcription machinery and transcription initiation could be blocked within highly compact, inaccessible chromatin regions. We experience the lack of understanding of cause-and-effect relationship between cell functioning and chromatin architecture, which makes the process of parameter selection in Hi-C modelling unclear. So, further analysis, to define whether the TAD affiliation and genomic distance are factors that are highly significant themselves or we have to take into account other important factors, is required.

Suppose, we focus on the Hi-C matrix row i which corresponds to the interactions with locus i and there is single start TAD edge which has a position k - all entries before the position k (entries with index numbers 1, 2, ..., $k-1$) represent DNA fragments outside TAD; all entries from the position k to the diagonal (entries with index numbers k, $k+1$, ..., $i-1$) represent fragments that are inside the same TAD as the locus i. The probability of interaction between fragments belonging to the same TAD is higher than the probability of interaction between fragments which do not belong to the same TAD.

Mathematically speaking,

$$(p_{(i,1)}, p_{(i,2)}, ..., p_{(i,k-1)}) < (p_{(i,k)}, p_{(i,k+1)}, ..., p_{(i,i-1)})$$

As interaction mean is the direct ratio of the interaction probability, we have

$$(\mu_{(i,1)}, \mu_{(i,2)}, ..., \mu_{(i,k-1)}) < (\mu_{(i,k)}, \mu_{(i,k+1)}, ..., \mu_{(i,i-1)})$$

Next, we propose some basic scenarios to model the vector of means $(\mu_{(i,1)}, \mu_{(i,2)}, ..., \mu_{(i,i-1)})$ that, first, take into account the TAD edge position, and, second, are reasonable and computationally feasible.

Scenario 0. As discussed above, the mean was previously modelled as a decreasing function of the distance between locus i and locus j. Analogically, we model mean as an increasing function of the locus j position. The function is increasing because locus $j = 1$ indicates the fragment that is at the greatest distance from locus i on the analysing Hi-C matrix row, while the locus $j = i-1$ indicates the fragment neighbouring to the locus i. Mathematically speaking,

$$\mu_{(i,j)} = \beta j^{\alpha}$$

with parameters $\beta > 0$ and $\alpha > 0$. According to the scenario, TAD edge does not affect the interaction mean. We can use this model as a control model in order to decide whether the TAD edge is a significant factor in Hi-C interaction modelling or not.

Scenario 1. The mean count is modelled as a constant function that changes its level after reaching the TAD edge:

$$\mu_{(i,j)} = \begin{cases} \beta_0 & j < k \text{ (outside TAD)} \\ \beta_i & j \geq k \text{ (inside TAD)} \end{cases}$$

with parameters $\beta_i > \beta_0 > 0$ for any $i = 1, ..., n$. According to the scenario, the only significant factor is TAD edge. In addition, the parameter β_i is proposed to be the same for rows i belonging to the same TAD and it can be different for rows belonging to different TADs. The scenario does not look realistic, but it is computationally fast and easy to implement when TAD edges are known. This scenario can help to assess the accuracy of the method in comparison with other scenarios.

Scenario 2. The mean count is modelled as a increasing function of the locus j posi-

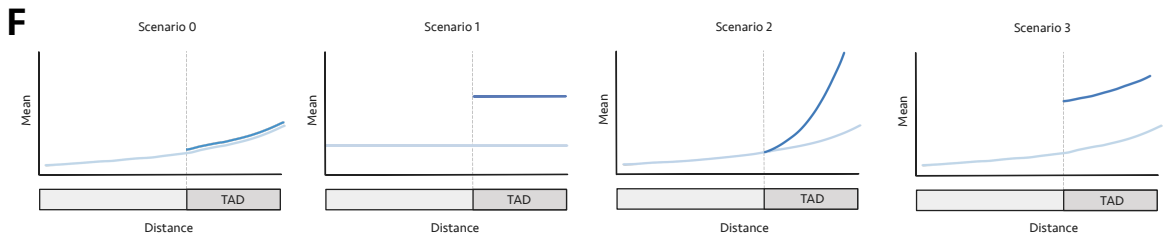


Figure 5.3. Schematic representation of Hi-C modelling scenarios.

tion. The functional relations inside and outside TAD are different:

$$\mu_{(i,j)} = \begin{cases} \beta_0 j^{\alpha_0} & j < k \text{ (outside TAD)} \\ \beta_i j^{\alpha_i} & j \geq k \text{ (inside TAD)} \end{cases}$$

with parameters $\beta_0 > 0$, $\beta_i > 0$, $\alpha_0 > 0$, $\alpha_i > 0$ and $(\beta_i \alpha_0)/(\beta_0 \alpha_i) > 0$ for any $i = 1, ..., n$. Similar to Scenario 1, the parameters β_i and α_i are proposed to be different for rows belonging to different TADs. On the one hand, the Scenario 2 seems to be more realistic than Scenario 0 and Scenario 1. On the other hand, the Scenario 2 is computationally heavy. Also, it requires TAD to be large enough to have observations for parameters estimation.

Scenario 3. The mean count is modelled as an increasing function of the locus j position, inside TAD the mean value is shifted up by constant value γ:

$$\mu_{(i,j)} = \begin{cases} \beta j^{\alpha} & j < k \text{ (outside TAD)} \\ \beta j^{\alpha} + \gamma_i & j \geq k \text{ (inside TAD)} \end{cases}$$

with parameters $\beta > 0$, $\alpha > 0$ and $\gamma_i > 0$ for any $i = 1, ..., n$. The parameter γ_i is expected to be different for rows that belong to different TADs.

All scenarios described above rely on known TAD edge positions as we need to define the point, where the mean is expected to change its functional relation. It allows us to divide observations into inside and outside TAD subsets in order to estimate the functional parameters. In addition, in all scenarios we assume single TAD edge to be present within the selected Hi-C row and the possibility of nested and overlapping TADs

violates this assumption. Allowing more TAD edges, we make Scenarios more complicated and, at the same time, reduce number of observations available for parameter estimation.

Improper TAD edge allocation possibly could happen when using TAD calling tools that generated TADs in “head-to-tail” manner. When TAD edge is longer than expected (or, in other words, visually detectable) we include more observations that do not follow the predicted functional relationship between interaction frequency, genomic distance and TAD affiliation and introduce biases in parameter estimation. When TAD edge is shorter than expected, we result in scarcity of observation for estimation. So, implementation of TAD calling tool, which allow complex architectural structures and, as a consequence, proper TAD edge detection, is the required step before proposed Hi-C model.

References

Adams, Mark D et al. (2000). “The genome gequence of *Drosophila melanogaster*”. In: *Science* 287.5461, pp. 2185–2195. DOI: 10.1126/science.287.5461.2185.

Adeel, Muhammad Muzammal et al. (2021). “Structural variations of the 3D genome architecture in cervical cancer development”. In: *Front Cell Dev Biol* 9, p. 706375. DOI: 10.3389/fcell.2021.706375.

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber (2015). “HTSeq – a Python framework to work with high-throughput sequencing data”. In: *Bioinformatics* 31.2, pp. 166–169. DOI: 10.1093/bioinformatics/btu638.

Ay, Ferhat, Timothy L Bailey, and William Stafford Noble (2014). “Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts”. In: *Genome Res* 24.6, pp. 999–1011. DOI: 10.1101/gr.160374.113.

Bartkuhn, Marek and Rainer Renkawitz (2008). “Long range chromatin interactions involved in gene regulation”. In: *Biochimica et Biophysica Acta* 1783.11, pp. 2161–2166. DOI: 10.1016/j.bbamcr.2008.07.011.

Beagrie, Robert A et al. (2017). “Complex multi-enhancer contacts captured by genome architecture mapping”. In: *Nature* 543.7646, pp. 519–524. ISSN: 1476-4687. DOI: 10.1038/nature21411.

Belaghzal, Houda, Job Dekker, and Johan H Gibcus (2016). “Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation”. In: *Methods* 176.3, pp. 139–148. DOI: 10.1016/j.ymeth.2017.04.004.

Belton, Jon Matthew et al. (2012). “Hi-C: a comprehensive technique to capture the conformation of genomes”. In: *Methods* 58.3, pp. 268–276. DOI: 10.1016/j.ymeth.2012.05.001.

Bentley, Graham A et al. (1984). “Crystal structure of the nucleosome core particle at 16A resolution”. In: *J Mol Biol* 176.1, pp. 55–75. DOI: 10.1016/0022-2836(84)90382-6.

Björkegren, Camilla and Laura Baranello (2018). “DNA supercoiling, topoisomerases, and cohesin: partners in regulating chromatin architecture?” In: *Int J Mol Sci* 19.3. DOI: 10.3390/ijms19030884.

Boeke, Jef D et al. (1985). “Ty elements transpose through an RNA intermediate”. In: *Cell* 40.3, pp. 491–500. DOI: 10.1016/0092-8674(85)90197-7.

Bolger, Anthony M, Marc Lohse, and Bjoern Usadel (2014). “Trimmomatic: a flexible trimmer for Illumina sequence data”. In: *Bioinformatics* 30.15, pp. 2114–2120. DOI: 10.1093/bioinformatics/btu170.

Bonev, Boyan and Giacomo Cavalli (2016). “Organization and function of the 3D genome”. In: *Nat Rev Genet* 17, pp. 661–678. DOI: 10.1038/nrg.2016.112.

Bourque, Guillaume et al. (2008). “Evolution of the mammalian transcription factor binding repertoire via transposable elements”. In: *Genome Res* 18.11, pp. 1752–1762. DOI: 10.1101/gr.080663.108.

Bushey, Ashley M, Edward Ramos, and Victor G Corces (2009). “Three subclasses of a *Drosophila* insulator show distinct and cell type-specific genomic distributions”. In: *Genes Dev* 23.11, pp. 1338–1350. DOI: 10.1101/gad.1798209.

Calo, Eliezer and Joanna Wysocka (2013). “Modification of enhancer chromatin: what, how, and why?” In: *Mol Cell* 49.5, pp. 825–837. DOI: 10.1016/j.molcel.2013.01.038.

Canela, Andres et al. (2017). “Genome organization drives chromosome fragility”. In: *Cell* 170.3, 507–521.e18. DOI: 10.1016/j.cell.2017.06.034.

Capelson, Maya and Victor G Corces (2004). “Boundary elements and nuclear organization”. In: *Biol Cell* 96.8, pp. 617–629. DOI: 10.1016/j.biolcel.2004.06.004.

Cardozo Gizzi, Andrés M et al. (2019). “Microscopy-based chromosome conformation capture enables simultaneous visualization of genome organization and transcription in intact organisms”. In: *Molecular Cell* 74.1, pp. 212–222. DOI: 10.1016/j.molcel.2019.01.011.

Chang, Li-Hsin, Sourav Ghosh, and Daan Noordermeer (2020). “TADs and their borders: free movement or building a wall?” In: *J Mol Biol* 432.3, pp. 643–652. ISSN: 0022-2836. DOI: https://doi.org/10.1016/j.jmb.2019.11.025.

Chathoth, Keerthi T, Liudmila A Mikheeva, et al. (2021). “The role of insulators and transcription in 3D chromatin organisation of flies”. In: *bioRxiv*. DOI: 10.1101/2021.04.26.441424.

Chathoth, Keerthi T and Nicolae Radu Zabet (2019). “Chromatin architecture reorganization during neuronal cell differentiation in *Drosophila* genome”. In: *Genome Res* 29.4, pp. 613–625. DOI: 10.1101/gr.246710.118.

Chetverina, Darya, Maksim Erokhin, and Paul Schedl (2021). “GAGA factor: a multifunctional pioneering chromatin protein”. In: *Cell Mol Life Sci* 78.9, pp. 4125–4141. DOI: 10.1007/s00018-021-03776-z.

Cournac, Axel et al. (2012). “Normalization of a chromosomal contact map”. In: *BMC Genomics* 13.1, p. 436. DOI: 10.1186/1471-2164-13-436.

Crane, Emily et al. (2015). “Condensin-driven remodelling of X chromosome topology during dosage compensation”. In: *Nature* 523.7559, pp. 240–244. DOI: 10.1038/nature14450.

Cremer, T and C Cremer (2001). “Chromosome territories, nuclear architecture and gene regulation in mammalian cells”. In: *Nat Rev Genet* 2, pp. 292–301. DOI: 10.1038/35066075.

Crick, F (1958). “On protein synthesis”. In: *Symp Soc Exp Biol* 12, pp. 138–163.

— (1970). “Central dogma of molecular biology”. In: *Nature* 227.5258, pp. 561–563. DOI: 10.1038/227561a0.

Cubeñas-Potts, Caelin et al. (2017). “Different enhancer classes in *Drosophila* bind distinct architectural proteins and mediate unique chromatin interactions and 3D architecture”. In: *Nucleic Acids Res* 45.4, pp. 1714–1730. DOI: 10.1093/nar/gkw1114.

Dali, Rola and Mathieu Blanchette (2017). “A critical assessment of topologically associating domain prediction tools”. In: *Nucleic Acids Res* 45.6, pp. 2994–3005. DOI: 10.1093/nar/gkx145.

Dekker, Job et al. (2002). “Capturing chromosome conformation”. In: *Science* 295, pp. 1306–1311.

Diehl, Adam G, Ningxin Ouyang, and Alan P Boyle (2020). “Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes”. In: *Nat Commun* 11, pp. 1–18. DOI: 10.1038/s41467-020-15520-5.

Dileep, Vishnu et al. (2015). “Topologically associating domains and their long-range contacts are established during early G1 coincident with the establishment of the replication-timing program”. In: *Genome Res* 25.8, pp. 1104–1113. DOI: 10.1101/gr.183699.114.

Dixon, Jesse R, Inkyung Jung, et al. (2015). “Chromatin architecture reorganization during stem cell differentiation”. In: *Nature* 518.7539, pp. 331–336. DOI: 10.1038/nature14222.

Dixon, Jesse R, Siddarth Selvaraj, et al. (2012). “Topological domains in mammalian genomes identified by analysis of chromatin interactions”. In: *Nature* 485.7398, pp. 376–380. DOI: 10.1038/nature11082.

Dong, Xianjun et al. (2012). “Modeling gene expression using chromatin features in various cellular contexts”. In: *Genome Biol* 13.9, R53. DOI: 10.1186/gb-2012-13-9-r53.

dos Santos, Gilberto et al. (2015). “FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations”. In: *Nucleic Acids Res* 43, pp. D690–7. DOI: 10.1093/nar/gku1099.

Dostie, Josée et al. (2006). “Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements”. In: *Genome Res* 16.10, pp. 1299–1309. DOI: 10.1101/gr.5571506.

Ferrari, Karin J et al. (2014). “Polycomb-dependent H3K27me1 and H3K27me2 regulate active transcription and enhancer fidelity”. In: *Mol Cell* 53.1, pp. 49–62. DOI: 10.1016/j.molcel.2013.10.030.

Filippova, Darya et al. (2014). “Identification of alternative topological domains in chromatin”. In: *Algorithms Mol Biol* 9.1, p. 14. DOI: 10.1186/1748-7188-9-14. URL: https://doi.org/10.1186/1748-7188-9-14.

Flavahan, William A et al. (2016). “Insulator dysfunction and oncogene activation in IDH mutant gliomas”. In: *Nature* 529.7584, pp. 110–114. DOI: 10.1038/nature16490.

Forcato, Mattia et al. (2017). “Comparison of computational methods for Hi-C data analysis”. In: *Nat Methods* 14.7, pp. 679–685. DOI: 10.1038/nmeth.4325.

Franke, Martin et al. (2021). “CTCF knockout in zebrafish induces alterations in regulatory landscapes and developmental gene expression”. In: *Nat Commun* 12.1, p. 5415. DOI: 10.1038/s41467-021-25604-5.

Fraser, James et al. (2015). “Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation.” In: *Mol Syst Biol* 11.12, p. 852. DOI: 10.15252/msb.20156492.

Fudenberg, Geoffrey et al. (2017). “Emerging evidence of chromosome folding by loop extrusion”. In: *Cold Spring Harb Symp Quant Biol* 176.3, pp. 139–148. DOI: 10.1101/sqb.2017.82.034710.Emerging.

Gambetta, Maria Cristina and Eileen E M Furlong (2018). “The insulator protein CTCF is required for correct Hox gene expression, but not for embryonic development in *Drosophila*”. In: *Genetics* 210.1, pp. 129–136. DOI: 10.1534/genetics.118.301350.

Gao, Tina et al. (2015). “Increasing overhang GC-content increases sticky-end ligation efficiency”. In: *J Exp Microbiol Immunol*, pp. 1–8.

Gavrilov, Alexey, Sergey V Razin, and Giacomo Cavalli (2015). “*In vivo* formaldehyde cross-linking: it is time for black box analysis”. In: *Brief Funct Genomics* 14.2, pp. 163–165. DOI: 10.1093/bfgp/elu037.

Gel, Bernat et al. (2016). “regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests”. In: *Bioinformatics* 32.2, pp. 289–291. DOI: 10.1093/bioinformatics/btv562.

Gerasimova, Tatiana I et al. (2007). “Coordinated control of dCTCF and gypsy chromatin insulators in *Drosophila*”. In: *Mol Cell* 28.5, pp. 761–772. DOI: 10.1016/j.molcel.2007.09.024.

Ghavi-Helm, Yad, Aleksander Jankowski, et al. (2019). “Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression”. In: *Nat Genet* 51.8, pp. 1272–1282. DOI: 10.1038/s41588-019-0462-3.

Ghavi-Helm, Yad, Felix A Klein, et al. (2014). “Enhancer loops appear stable during development and are associated with paused polymerase”. In: *Nature* 512.7512, pp. 96–100. DOI: 10.1038/nature13417.

Gibcus, Johan H et al. (2018). “A pathway for mitotic chromosome formation”. In: *Science* 359.6376, pp. 1–29. DOI: 10.1126/science.aao6135.

Giorgetti, Luca and Edith Heard (2016). “Closing the loop: 3C versus DNA FISH”. In: *Genome Biol* 17, pp. 1–9. DOI: 10.1186/s13059-016-1081-2.

Golloshi, Rosela, Jacob T. Sanders, and Rachel Patton McCord (2018). “Iteratively improving Hi-C experiments one step at a time”. In: *Methods* 142, pp. 47–58. DOI: 10.1016/j.ymeth.2018.04.033.

Gothe, Henrike Johanna et al. (2019). “Spatial chromosome folding and active transcription drive DNA fragility and formation of oncogenic MLL translocations”. In: *Mol Cell* 75.2, 267–283.e12. DOI: 10.1016/j.molcel.2019.05.015.

Granjon, Pierre (2013). “The CuSum algorithm - a small review”. In: *HAL* hal-00914697.

Harmston, Nathan et al. (2017). “Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation”. In: *Nat Commun* 8.1, p. 441. DOI: 10.1038/s41467-017-00524-5.

Helmbacher, F et al. (2000). “Targeting of the EphA4 tyrosine kinase receptor affects dorsal/ventral pathfinding of limb motor axons”. In: *Development* 127.15, pp. 3313–3324. DOI: 10.1242/dev.127.15.3313.

Hoffman, Steven J. and Charlie Tan (2015). “Overview of systematic reviews on the health-related effects of government tobacco control policies”. In: *BMC Public Health* 15.1, pp. 1–11. DOI: 10.1186/s12889-015-2041-6.

Hu, Ming et al. (2012). “HiCNorm: removing biases in Hi-C data via Poisson regression”. In: *Bioinformatics* 28.23, pp. 3131–3133. DOI: 10.1093/bioinformatics/bts570.

Hug, Clemens B et al. (2017). “Chromatin architecture emerges during zygotic genome activation independent of transcription”. In: *Cell* 169, pp. 216–228. DOI: 10.1016/j.cell.2017.03.024.

Hyle, Judith et al. (2019). “Acute depletion of CTCF directly affects MYC regulation through loss of enhancer–promoter looping”. In: *Nucleic Acids Res* 47.13, pp. 6699–6713. DOI: 10.1093/nar/gkz462.

Imakaev, Maxim et al. (2012). “Iterative correction of Hi-C data reveals hallmarks of chromosome organization”. In: *Nature* 9.10, pp. 999–1003. DOI: 10.1038/nmeth.2148.Iterative.

Immarigeon, Clément et al. (2020). “Mediator complex subunit Med19 binds directly GATA transcription factors and is required with Med1 for GATA-driven gene regulation *in vivo*”. In: *J Biol Chem* 295.39, pp. 13617–13629. DOI: 10.1074/jbc.RA120.013728.

Ing-Simmons, Elizabeth et al. (2021). “Independence of chromatin conformation and gene regulation during *Drosophila* dorsoventral patterning”. In: *Nat Genet* 53.4, pp. 487–499. DOI: 10.1038/s41588-021-00799-x.

Jimenez, David Sebastian et al. (2021). “Condensin DC spreads linearly and bidirectionally from recruitment sites to create loop-anchored TADs in *C. elegans*”. In: *bioRxiv*. DOI: 10.1101/2021.03.23.436694.

Johnstone, Sarah E et al. (2020). “Large-scale topological changes restrain malignant progression in colorectal cancer”. In: *Cell* 182.6, 1474–1489.e23. DOI: 10.1016/j.cell.2020.07.030.

Kaushal, Anjali et al. (2021). “CTCF loss has limited effects on global genome architecture in *Drosophila* despite critical regulatory functions”. In: *Nat Commun* 12.1, p. 1011. DOI: 10.1038/s41467-021-21366-2.

Kim, Daehwan et al. (2013). “TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions”. In: *Genome Biol* 14.4, R36. DOI: 10.1186/gb-2013-14-4-r36.

Knight, Philip A and Daniel Ruiz (2012). “A fast algorithm for matrix balancing”. In: *IMA J Numer Anal* 33.3, pp. 1029–1047. DOI: 10.1093/imanum/drs019.

Koch, Christoph M et al. (2007). “The landscape of histone modifications across 1% of the human genome in five human cell lines”. In: *Genome Res* 17.6, pp. 691–707. DOI: 10.1101/gr.5704207.

Kraft, Katerina et al. (2019). “Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations”. In: *Nat Cell Biol* 21, pp. 305–310. DOI: 10.1038/s41556-019-0273-x.

Kragesteen, Bjørt K et al. (2018). “Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis”. In: *Nat Genet* 50.10, pp. 1463–1473. DOI: 10.1038/s41588-018-0221-x.

Kruse, Kai, Clemens B Hug, and Juan M Vaquerizas (2020). “FAN-C: a feature-rich framework for the analysis and visualisation of chromosome conformation capture data”. In: *Genome Biol* 21.1, p. 303. DOI: 10.1186/s13059-020-02215-9.

Labrador, Mariano and Victor G Corces (2002). “Setting the boundaries of chromatin domains and nuclear organization”. In: *Cell* 111.2, pp. 151–154. DOI: 10.1016/s0092-8674(02)01004-8.

Langer-Safer, Pennina R, Michael Levine, and David C Ward (1982). “Immunological method for mapping genes on *Drosophila* polytene chromosomes”. In: *Proc Natl Acad Sci USA* 79.14, pp. 4381–4385. DOI: 10.1073/pnas.79.14.4381.

Langmead, Ben and Steven L Salzberg (2012). “Fast gapped-read alignment with Bowtie 2”. In: *Nat Methods* 9.4, pp. 357–359. DOI: 10.1038/nmeth.1923.

Lee, Jong-Hee et al. (2015). “Single transcription factor conversion of human blood fate to NPCs with CNS and PNS developmental capacity”. In: *Cell Rep* 11.9, pp. 1367–1376. DOI: 10.1016/j.celrep.2015.04.056.

Lehmann, Lynn et al. (2012). “Polycomb repressive complex 1 (PRC1) disassembles RNA polymerase II preinitiation complexes”. In: *J Biol Chem* 287.43, pp. 35784–35794. DOI: 10.1074/jbc.M112.397430.

Li, Heng and Richard Durbin (2010). “Fast and accurate long-read alignment with Burrows–Wheeler transform”. In: *Bioinformatics* 26.5, pp. 589–595. DOI: 10.1093/bioinformatics/btp698.

Li, Li et al. (2015). “Widespread rearrangement of 3D chromatin organization underlies Polycomb-mediated stress-induced silencing”. In: *Mol Cell* 58.2, pp. 216–231. DOI: 10.1016/j.molcel.2015.02.023.

Liu, Zhihua et al. (2011). “*Drosophila* Acyl-CoA synthetase long-chain family member 4 regulates axonal transport of synaptic vesicles and is required for synaptic development and transmission”. In: *J Neurosci* 31.6, pp. 2052–2063. DOI: 10.1523/JNEUROSCI.3278-10.2011.

Love, Michael I, Wolfgang Huber, and Simon Anders (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biol* 15.12, p. 550. DOI: 10.1186/s13059-014-0550-8.

Lu, Junjie et al. (2010). “G2 phase chromatin lacks determinants of replication timing”. In: *J Cell Biol* 189.6, pp. 967–980. DOI: 10.1083/jcb.201002002.

Luger, Karolin et al. (1997). “Crystal structure of the nucleosome core particle at 2.8 A resolution”. In: *Nature* 389, pp. 251–260. DOI: 10.1038/38444.

Lupiáñez, Darío G et al. (2015). “Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions”. In: *Cell* 161.5, pp. 1012–1025. DOI: 10.1016/j.cell.2015.04.004.

Maeda, Robert K and François Karch (2007). “Making connections: boundaries and insulators in *Drosophila*”. In: *Curr Opin Genet Dev* 17.5, pp. 394–399. DOI: 10.1016/j.gde.2007.08.002.

Martin, Patrick C.N. and Nicolae Radu Zabet (2020). “Dissecting the binding mechanisms of transcription factors to DNA using a statistical thermodynamics frame-

work". In: *Comput Struct Biotechnol J* 18, pp. 3590–3605. DOI: 10.1016/j.csbj.2020.11.006.

Mathelier, Anthony et al. (2014). "JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles". In: *Nucleic Acids Res* 42, pp. D142–7. DOI: 10.1093/nar/gkt997.

Melnikova, L S et al. (2019). "Functional properties of the Su(Hw) complex are determined by its regulatory environment and multiple interactions on the Su(Hw) protein platform". In: *Vavilovskii Zhurnal Genet Selektsii* 23.2, pp. 168–173. DOI: 10.18699/VJ19.477.

El-Metwally, Sara, Osama M. Ouda, and Mohamed Helmy (2014). "Novel next-generation sequencing applications". In: *Next generation sequencing technologies and challenges in sequence assembly*, pp. 61–70. DOI: 10.1007/978-1-4939-0715-1_7.

Mifsud, Borbala, Inigo Martincorena, et al. (2017). "GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data". In: *PLoS One* 12.4, pp. 1–15. DOI: 10.1371/journal.pone.0174744.

Mifsud, Borbala, Filipe Tavares-Cadete, et al. (2015). "Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C". In: *Nat Genet* 47.6, pp. 598–606. DOI: 10.1038/ng.3286.

Min, Jinrong et al. (2003). "Structure of the catalytic domain of human DOT1L, a non-SET domain nucleosomal histone methyltransferase". In: *Cell* 112.5, pp. 711–723. DOI: 10.1016/s0092-8674(03)00114-4.

Mirny, Leonid A, Maxim Imakaev, and Nezar Abdennur (2019). "Two major mechanisms of chromosome organization". In: *Curr Opin Cell Biol* 58, pp. 142–152. DOI: 10.1016/j.ceb.2019.05.001.

Nagano, Takashi et al. (2013). "Single-cell Hi-C reveals cell-to-cell variability in chromosome structure". In: *Nature* 502.7469, pp. 59–64. DOI: 10.1038/nature12593.

Nagano, Tatsuya, Motoko Tachihara, and Yoshihiro Nishimura (2018). "Mechanism of resistance to epidermal growth factor receptor-tyrosine kinase inhibitors and a potential treatment strategy". In: *Cells* 7.11, pp. 1–16. DOI: 10.3390/cells7110212.

Nakayama, Jun-ichi et al. (2001). “Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly”. In: *Science* 292.5514, pp. 110–113. DOI: 10.1126/science.1060118.

Neguembor, Maria Victoria et al. (2021). “Transcription-mediated supercoiling regulates genome folding and loop formation”. In: *Mol Cell* 81.15, 3065–3081.e12. DOI: https://doi.org/10.1016/j.molcel.2021.06.009.

Niu, Longjian et al. (2019). “Amplification-free library preparation with SAFE Hi-C uses ligation products for deep sequencing to improve traditional Hi-C analysis”. In: *Commun Biol* 2.1, p. 267. DOI: 10.1038/s42003-019-0519-y.

Nora, Elphège P, Anton Goloborodko, et al. (2017). “Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization”. In: *Cell* 169.5, 930–944.e22. DOI: 10.1016/j.cell.2017.05.004.

Nora, Elphège P, Bryan R Lajoie, et al. (2012). “Spatial partitioning of the regulatory landscape of the X-inactivation centre”. In: *Nature* 485.7398, pp. 381–385. DOI: 10.1038/nature11049.

Nora, Elphège P. et al. (2020). “Molecular basis of CTCF binding polarity in genome folding”. In: *Nat Commun* 11.1, pp. 1–13. DOI: 10.1038/s41467-020-19283-x.

Norton, Heidi K et al. (2018). “Detecting hierarchical genome folding with network modularity”. In: *Nat Methods* 15.2, pp. 119–122. DOI: 10.1038/nmeth.4560.

Ogiyama, Yuki et al. (2018). “Polycomb-dependent chromatin looping contributes to gene silencing during *Drosophila* development”. In: *Mol Cell* 71.1, 73–88.e5. DOI: 10.1016/j.molcel.2018.05.032.

Pai, Chi-Yun et al. (2004). “The centrosomal protein CP190 is a component of the gypsy chromatin insulator”. In: *Mol Cell* 16.5, pp. 737–748. DOI: 10.1016/j.molcel.2004.11.004.

Pekowska, Aleksandra et al. (2018). “Gain of CTCF-Anchored chromatin loops marks the exit from naive pluripotency”. In: *Cell Syst* 7.5, 482–495.e10. DOI: 10.1016/j.cels.2018.09.003.

Pennacchio, Len A et al. (2013). “Enhancers: five essential questions”. In: *Nat Rev Genet* 14.4, pp. 288–295. DOI: 10.1038/nrg3458.

Pherson, Michelle et al. (2019). “Cohesin occupancy and composition at enhancers and promoters are linked to DNA replication origin proximity in *Drosophila*”. In: *Genome Res* 29.4, pp. 602–612. DOI: 10.1101/gr.243832.118.

Phillips, Jennifer E and Victor G Corces (2009). “CTCF: master weaver of the genome”. In: *Cell* 137.7, pp. 1194–1211. DOI: 10.1016/j.cell.2009.06.001.

Phillips-Cremins, Jennifer E. et al. (2013). “Architectural protein subclasses shape 3D organization of genomes during lineage commitment”. In: *Cell* 153.6, pp. 1281–1295. DOI: 10.1016/j.cell.2013.04.053.

Platt, Roy N, Michael W Vandewege, and David A Ray (2018). “Mammalian transposable elements and their impacts on genome evolution”. In: *Chromosom Res* 26.1-2, pp. 25–43. DOI: 10.1007/s10577-017-9570-z.

Pommier, Yves et al. (2016). “Roles of eukaryotic topoisomerases in transcription, replication and genomic stability”. In: *Nat Rev Mol Cell Biol* 17.11, pp. 703–721. DOI: 10.1038/nrm.2016.111.

Pope, Benjamin D et al. (2014). “Topologically associating domains are stable units of replication-timing regulation”. In: *Nature* 515.7527, pp. 402–405. DOI: 10.1038/nature13986.

Quinodoz, Sofia A et al. (2018). “Higher-order inter-chromosomal hubs shape 3D genome organization in the Nucleus”. In: *Cell* 174.3, 744–757.e24. DOI: 10.1016/j.cell.2018.05.024.

Racko, Dusan et al. (2018). “Chromatin loop extrusion and chromatin unknotting”. In: *Polymers* 10.10, pp. 1–11. DOI: 10.3390/polym10101126.

Ramirez, Fidel et al. (2018). “High-resolution TADs reveal DNA sequences underlying genome organization in flies”. In: *Nat Commun* 9.1, p. 189. DOI: 10.1038/s41467-017-02525-w.

Ramírez, Fidel et al. (2015). “High-affinity sites form an interaction network to facilitate spreading of the MSL complex across the X chromosome in *Drosophila*”. In: *Mol Cell* 60.1, pp. 146–162. DOI: 10.1016/j.molcel.2015.08.024.

Rao, Suhas S P et al. (2014). “A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping”. In: *Cell* 159, pp. 1665–1680. DOI: 10.1016/j.cell.2014.11.021.

Rath, G M et al. (2006). “The C-terminal CD47/IAP-binding domain of thrombospondin-1 prevents camptothecin- and doxorubicin-induced apoptosis in human thyroid carcinoma cells”. In: *Biochim Biophys Acta* 1763.10, pp. 1125–1134. DOI: 10.1016/j.bbamcr.2006.08.001.

Rocha, Pedro P et al. (2015). “Breaking TADs: insights into hierarchical genome organization”. In: *Epigenomics* 7.4, pp. 523–526. DOI: 10.2217/epi.15.25.

Rowley, M Jordan, Xiaowen Lyu, et al. (n.d.). “Condensin II counteracts Cohesin and RNA Polymerase II in the establishment of 3D chromatin organization”. In: *Cell Rep* 11 (), 2890–2903.e3. DOI: 10.1016/j.celrep.2019.01.116.

Rowley, M Jordan, Michael H Nichols, et al. (2017). “Evolutionarily conserved principles predict 3D chromatin organization”. In: *Mol Cell* 67.5, 837–852.e7. DOI: 10.1016/j.molcel.2017.07.022.

Rusková, Renáta and Dušan Račko (2021). “Entropic competition between supercoiled and torsionally relaxed chromatin fibers drives loop extrusion through pseudo-topologically bound cohesin”. In: *Biology* 10.2. DOI: 10.3390/biology10020130.

Saha, Parna et al. (2020). “Interplay of pericentromeric genome organization and chromatin landscape regulates the expression of *Drosophila melanogaster* heterochromatic genes”. In: *Epigenetics Chromatin* 13.1, p. 41. DOI: 10.1186/s13072-020-00358-4.

Schmitt, Anthony D, Ming Hu, Inkyung Jung, et al. (2016). “A compendium of chromatin contact maps reveals spatially active regions in the human genome”. In: *Cell Rep* 17.8, pp. 2042–2059. DOI: 10.1016/j.celrep.2016.10.061.

Schmitt, Anthony D, Ming Hu, and Bing Ren (2016). “Genome-wide mapping and analysis of chromosome architecture”. In: *Nat Rev Mol Cell Biol* 17.12, pp. 743–755. DOI: 10.1038/nrm.2016.104.

Servant, Nicolas et al. (2015). “HiC-Pro: An optimized and flexible pipeline for Hi-C data processing”. In: *Genome Biol* 16.1, pp. 1–11. DOI: 10.1186/s13059-015-0831-x.

Sexton, Tom et al. (2012). “Three-dimensional folding and functional organization principles of the *Drosophila* genome”. In: *Cell* 148.3, pp. 458–472. DOI: 10.1016/j.cell.2012.01.010.

Shavit, Yoli and Pietro Lio (2014). “Combining a wavelet change point and the Bayes factor for analysing chromosomal interaction data”. In: *Mol Biosyst* 10.6, pp. 1576–1585. DOI: 10.1039/c4mb00142g.

Shin, Hanjun et al. (2016). “TopDom: an efficient and deterministic method for identifying topological domains in genomes”. In: *Nucleic Acids Res.* 44.7, e70. DOI: 10.1093/nar/gkv1505.

Simonis, Marieke et al. (2006). “Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)”. In: *Nat Genet* 38.11, pp. 1348–1354. DOI: 10.1038/ng1896.

Stadler, Michael R, Jenna E Haines, and Michael B Eisen (2017). “Convergence of topological domain boundaries, insulators, and polytene interbands revealed by high-resolution mapping of chromatin contacts in the early *Drosophila melanogaster* embryo”. In: *Elife* 6, pp. 1–29. DOI: 10.7554/eLife.29550.

Steensel, Bas van and Eileen E M Furlong (2019). “The role of transcription in shaping the spatial organization of the genome”. In: *Nat Rev Mol Cell Biol* 20.6, pp. 327–337. DOI: 10.1038/s41580-019-0114-6.

Steger, David J et al. (2008). “DOT1L/KMT4 recruitment and H3K79 methylation are ubiquitously coupled with gene transcription in mammalian cells”. In: *Mol Cell Biol* 28.8, pp. 2825–2839. DOI: 10.1128/MCB.02076-07.

Sullivan, Gail M and Richard Feinn (2012). “Using effect size - or why the p value is not enough”. In: *J Grad Med Educ* 4.3, pp. 279–282. DOI: 10.4300/JGME-D-12-00156.1.

Swaminathan, Jyothishmathi, Ellen M Baxter, and Victor G Corces (2005). “The role of histone H2Av variant replacement and histone H4 acetylation in the establishment of *Drosophila* heterochromatin”. In: *Genes Dev* 19.1, pp. 65–76. DOI: 10.1101/gad.1259105.

Szabo, Quentin et al. (2018). “TADs are 3D structural units of higher-order chromosome organization in *Drosophila*”. In: *Sci Adv* 4.2, eaar8082. DOI: 10.1126/sciadv.aar8082.

Taberlay, Phillippa C et al. (2016). “Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations”. In: *Genome Res* 26.6, pp. 719–731. DOI: 10.1101/gr.201517.115.

Takemata, Naomichi and Stephen D Bell (2021). “Multi-scale architecture of archaeal chromosomes”. In: *Mol Cell* 81.3, 473–487.e6. DOI: 10.1016/j.molcel.2020.12.001.

Tang, Xiaona et al. (2021). “Kinetic principles underlying pioneer function of GAGA transcription factor in live cells”. In: *bioRxiv*. DOI: 10.1101/2021.10.21.465351.

Ulahannan, Netha et al. (2019). “Nanopore sequencing of DNA concatemers reveals higher-order features of chromatin structure”. In: *bioRxiv*. DOI: 10.1101/833590.

Uusküla-Reimand, Liis et al. (2016). “Topoisomerase II beta interacts with cohesin and CTCF at topological domain borders”. In: *Genome Biol* 17.1, p. 182. DOI: 10.1186/s13059-016-1043-8.

Van Bortle, Kevin and Victor G Corces (2012). “Nuclear organization and genome function”. In: *Annu Rev Cell Dev Biol* 28, pp. 163–187. DOI: 10.1146/annurev-cellbio-101011-155824.

Van Bortle, Kevin, Michael H Nichols, et al. (2014). “Insulator function and topological domain border strength scale with architectural protein occupancy”. In: *Genome Biol* 15.6, R82. DOI: 10.1186/gb-2014-15-5-r82.

Van Bortle, Kevin, Edward Ramos, et al. (2012). "*Drosophila* CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains". In: *Genome Res* 22.11, pp. 2176–2187. DOI: 10.1101/gr.136788.111.

Vian, Laura et al. (2018). "The energetics and physiological impact of cohesin extrusion". In: *Cell* 173.5, pp. 1165–1178. DOI: 10.1016/j.cell.2018.03.072.

Vietri Rudan, Matteo et al. (2015). "Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture". In: *Cell Rep* 10.8, pp. 1297–1309. DOI: 10.1016/j.celrep.2015.02.004.

Wang, Qi et al. (2018). "Sub-kb Hi-C in *D. melanogaster* reveals conserved characteristics of TADs between insect and mammalian cells". In: *Nat. Commun* 9.1, p. 188. DOI: 10.1038/s41467-017-02526-9.

Wei, Xiaolu et al. (2021). "Heterochromatin-dependent transcription of satellite DNAs in the *Drosophila melanogaster* female germline". In: *bioRxiv*. DOI: 10.1101/2020.08.26.268920.

Weinreb, Caleb and Benjamin J Raphael (2016). "Identification of hierarchical chromatin domains". In: *Bioinformatics* 32.11, pp. 1601–1609. DOI: 10.1093/bioinformatics/btv485.

Wolfe, Jareth C et al. (2021). "An explainable artificial intelligence approach for decoding the enhancer histone modifications code and identification of novel enhancers in *Drosophila*". In: *Genome Biol* 22.1, p. 308. DOI: 10.1186/s13059-021-02532-7.

Xu, Beisi et al. (2021). "Acute depletion of CTCF rewires genome-wide chromatin accessibility". In: *Genome Biol* 22.1, p. 244. DOI: 10.1186/s13059-021-02466-0.

Yaffe, Eitan and Amos Tanay (2011). "Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture". In: *Nat Genet* 43.11, pp. 1059–1065. DOI: 10.1038/ng.947.

Yang, Jingping, Edward Ramos, and Victor G Corces (2012). "The BEAF-32 insulator coordinates genome organization and function during the evolution of *Drosophila* species". In: *Genome Res* 22.11, pp. 2199–2207. DOI: 10.1101/gr.142125.112.

Yu, Miao and Bing Ren (2017). “The three-dimensional organization of mammalian genomes”. In: *Annu Rev Cell Dev Biol* 33, pp. 265–289. DOI: 10.1146/annurev-cellbio-100616-060531.

Zabet, Nicolae Radu and Boris Adryan (2015). “Estimating binding properties of transcription factors from genome-wide binding profiles”. In: *Nucleic Acids Res* 43.1, pp. 84–94. DOI: 10.1093/nar/gku1269.

Zeng, Kai et al. (2006). “Statistical tests for detecting positive selection by utilizing high-frequency variants”. In: *Genetics* 174.3, pp. 1431–1439. DOI: 10.1534/genetics.106.061432.

Zenk, Fides et al. (2021). “HP1 drives de novo 3D genome reorganization in early *Drosophila* embryos”. In: *Nature* 593.7858, pp. 289–293. DOI: 10.1038/s41586-021-03460-z.

Zhang, Xiangbo and Yongwen Qi (2019). “The landscape of copia and gypsy retrotransposon during maize domestication and improvement”. In: *Front Plant Sci* 10, pp. 1–8. DOI: 10.3389/fpls.2019.01533.

Appendix 2.1. ChIP in flies

	Source	GEO accession	Occupancy	Peak
Architectural proteins				
BEAF-32	modENCODE_921	GSE20811	+	+
BEAF-32	modENCODE_3665	GSE32775		+
BEAF-32 (BEAF-32 KD)	modENCODE_3663	GSE32773		+
BEAF-32 (Cp190 KD)	modENCODE_3664	GSE32774		+
CTCF	modENCODE_3673	GSE32783	+	+
CTCF	modENCODE_282	GSE20767		+
Cp190	modENCODE_924	GSE20814	+	+
Cp190	modENCODE_3668	GSE32778		+
Cp190 (Cp190 KD)	modENCODE_3747	GSE32816		+
Chro	modENCODE_275	GSE20761	+	+
JIL-1	modENCODE_3035	GSE27754	+	
mod(mdg4)	modENCODE_3064	GSE20802	+	
Su(Hw)	modENCODE_951	GSE20833	+	
ZW5	modENCODE_3064	GSE25373	+	
Fs(1)h	Pherson et al. 2019	GSE118484	+	
Nipped-B	Pherson et al. 2019	GSE118484	+	
Rad21	Pherson et al. 2019	GSE118484	+	+
SA	Pherson et al. 2019	GSE118484	+	
Smc1	Pherson et al. 2019	GSE118484	+	
Transcription and replication				
GAF	modENCODE_2651	GSE23466	+	+
MED1	Pherson et al. 2019	GSE118484	+	+
MED30	Pherson et al. 2019	GSE118484	+	
Nascent RNA	Pherson et al. 2017	GSE100545	+	+
Orc2	modENCODE_2754	GSE20888	+	
Pof	modENCODE_3052	GSE27808	+	
Pol-II	modENCODE_950	GSE20832	+	+
Topo-II	modENCODE_5058	GSE45069	+	

Appendix Table 2.1. Datasets for epigenetic factors used in the Thesis.

	Source	GEO accession	Occupancy	Peak
DNA accessibility				
DNase-I	Kharchenko et al. 2011		+	
H1	modENCODE_3299	GSE32767	+	
H2Av	modENCODE_6073	GSE45110	+	
H3	modENCODE_3302	GSE32769	+	+
H4	modENCODE_3303	GSE32770	+	+
Histone modifications				
H2Bubi	modENCODE_288	GSE20771	+	
H3K18ac	modENCODE_291	GSE20774	+	
H3K23ac	modENCODE_293	GSE20776	+	
H3K27ac	modENCODE_295	GSE20778	+	+
H3K27me1	modENCODE_3941	GSE51965	+	
H3K27me2	modENCODE_2999	GSE27789	+	
H3K27me3	modENCODE_297	GSE20780	+	
H3K36me1	modENCODE_299	GSE20782	+	
H3K36me3	modENCODE_301	GSE20783	+	
H3K4me1	modENCODE_2653	GSE23468	+	+
H3K4me2	modENCODE_2654	GSE23469	+	
H3K4me3	modENCODE_967	GSE20839	+	+
H3K79me1	modENCODE_3005	GSE32736	+	
H3K79me2	modENCODE_306	GSE20788	+	
H3K79me3	modENCODE_4934	GSE45062	+	
H3K9me2	modENCODE_310	GSE20791	+	+
H3K9me3	modENCODE_312	GSE20793	+	+
H4K16ac	modENCODE_316	GSE20795	+	
H4K20me1	modENCODE_3286	GSE32755	+	
H4K8ac	modENCODE_5060	GSE45070	+	

Appendix Table 2.1. (continue) Datasets for epigenetic factors used in the Thesis.

	Source	GEO accession	Occupancy	Peak
Polycomb				
Pc	modENCODE_325	GSE20803	+	+
dRING	modENCODE_927	GSE20817	+	+
sSFMBT	modENCODE_2986	GSE27728	+	
Ez	modENCODE_2650	GSE23465	+	
Pcl	modENCODE_948	GSE20830	+	
Psc	modENCODE_3055	GSE25370	+	
Heterochromatin				
HP1a	modENCODE_4126	GSE44515	+	
HP1b	modENCODE_3016	GSE44462	+	
HP1c	modENCODE_942	GSE20824	+	
HP2	modENCODE_3026	GSE27747	+	
HP4	modENCODE_4185	GSE44521	+	
Su(var)3-7	modENCODE_2671	GSE23486	+	
Su(var)3-9	modENCODE_952	GSE20834	+	
Nucleosome remodellers				
ASH-1	modENCODE_3279	GSE32748	+	
JHDM1	modENCODE_5145	GSE45092	+	
ISWI	modENCODE_3030	GSE27750	+	
MRG15	modENCODE_3045	GSE25365	+	
NURF301	modENCODE_5063	GSE45072	+	
PR-Set7	modENCODE_5065	GSE45074	+	
RPD3	modENCODE_4188	GSE44523	+	
WDS	modENCODE_5148	GSE45094	+	
MOF	modENCODE_3041	GSE27803	+	

Appendix Table 2.1. (continue) Datasets for epigenetic factors used in the Thesis.

Appendix 3.1. CUSUM-based change-point detection algorithm

We fix the single Hi-C row $i = 1$, ..., N where N is the number of DNA fragments. We observe $x_{(i,j)}$ Hi-C interactions between locus i and locus $j = 1$, ..., $i - 1$. We estimate the mean interaction frequency between locus i and locus j based on a Moving Average (MA) as

$$MA_{(i,j)} = \frac{1}{w}(x_{(i,j)} + x_{(i,j+1)} + ... + x_{(i,j+(w-1))})$$

where w is MA estimation window size. So, within the Hi-C row i we get a sequence of mean estimates $(MA_{(i,1)}, MA_{(i,2)}, ..., MA_{(i,i-w)})$. When the particular column position j represents the TAD edge, we expect the increase in contact mean. So, we assume the following behaviour. The mean estimates before the position j are approximately monotone, then there is a rise when we reach the TAD edge, and again monotone inside the TAD. If there are nested or partially overlapped TADs, within the single Hi-C row we get several increases in contact mean, each of them represent the TAD edge position. This assumption is quite unrealistic as we ignore the relationship between the genomic distance and interaction frequency - fragments which are allocated close to each other on DNA strand interacts more frequently than fragments which are far away from each other. However, due to simplicity of the suggested behaviour we can refer to the change-point detection algorithms in order to detect the coordinates of contact mean changes.

In this section we introduce the general steps of CUSUM (cumulative sum) algorithm and its possible modifications to make it appropriate specifically to Hi-C data and TAD edge allocation problem.

Appendix 3.1.1. General form of a sequential change detection algorithm

Suppose that Hi-C interactions within the row i are modelled as independent and identically distributed random variables $X_{(i,j)}$ where $j = 1, 2, ..., i-1$. Each variable follows the probability density function $f_X(x|\mu)$ depending on a parameter μ that is the same for all variables before the change time $j_c < (i - 1)$ and increases by δ at the change time j_c. Therefore, $X_{(i,1)}$, $X_{(i,2)}$, ..., $X_{(i,j_c-1)}$ follows pdf depending on the parameter

$\mu = \mu_0$ (before the shift) and $X_{(i,j_c)}$, $X_{(i,j_c+1)}$, ..., $X_{(i,i-1)}$ follows pdf with $\mu = \mu_1$ where $\mu_1 = \mu_0 + \delta$ (after the shift). Suppose we are in the case when the shift does not happen within the variables $X_{(i,j)}$ (null hypothesis H_0), then the whole pdf is defined as

$$f_{X|H_0} = \prod_{j=1}^{j=i-1} f_X(x_{(i,j)}|\mu_0)$$

If we are in the case when one change in parameter happens at change time j_c (alternative hypothesis H_1), then the whole pdf is defined as

$$f_{X|H_1} = \prod_{j=1}^{j=j_c-1} f_X(x_{(i,j)}|\mu_0) \prod_{j=j_c}^{j=i-1} f_X(x_{(i,j)}|\mu_1)$$

We aim to construct the rule in order to decide between two hypotheses H_0 and H_1. The detection theory provides us with the solution – the likelihood ratio test.

Likelihood ratio test. The log-likelihood ratio is defined as

$$L_X = ln\left(\frac{f_{X|H_1}}{f_{X|H_0}}\right)$$

Then, decide H_1 if $L_X > h$ or decide H_0 otherwise, where h is a test threshold.

In our case, the log-likelihood ratio is defined as

$$L_X(j_c) = ln\left(\prod_{j=j_c}^{j=i-1} \frac{f_X(x_{(i,j)}|\mu_1)}{f_X(x_{(i,j)}|\mu_0)}\right) = \sum_{j=j_c}^{j=i-1} ln\left(\frac{f_X(x_{(i,j)}|\mu_1)}{f_X(x_{(i,j)}|\mu_0)}\right)$$

However, j_c is unknown and we aim to detect it. One way to efficiently define the change time is to use the value of j_c that maximises the likelihood $f_{X|H_1}$, i.e. the distribution most probably fits the observed data given that the alternative hypothesis is true (the shift in mean is present):

$$\widehat{j_c} = \underset{1 \le j_c \le i-1}{\arg\max}\ f_{X|H_1}(j_c) = \underset{1 \le j_c \le i-1}{\arg\max}\ L_X(j_c) = \underset{1 \le j_c \le i-1}{\arg\max} \sum_{j=j_c}^{j=i-1} ln\left(\frac{f_X(x_{(i,j)}|\mu_1)}{f_X(x_{(i,j)}|\mu_0)}\right)$$

Then, the likelihood ratio test turns into the generalised likelihood ratio test.

Generalised likelihood ratio test. The generalised log-likelihood ratio is defined as

$$G_X = \max_{1 \le j_c \le i-1} L_X(j_c) = \max_{1 \le j_c \le i-1} ln\left(\frac{f_{X|H_1}}{f_{X|H_0}}\right)$$

Then, decide H_1 if $G_X > h$ or decide H_0 otherwise, where h is a test threshold.

In our case, the generalised log-likelihood ratio is defined as

$$G_X = \max_{1 \le j_c \le i-1} \sum_{j=j_c}^{j=i-1} ln\left(\frac{f_X(x_{(i,j)}|\mu_1)}{f_X(x_{(i,j)}|\mu_0)}\right)$$

Appendix 3.1.2. CUSUM algorithm and its recursive form

The decision rule for the generalised likelihood ratio test could be rewritten in a form of recursive algorithm. Suppose at position j we observe the experiment value $x_{(i,j)}$. Then, we can define the **instantaneous log-likelihood ratio** at j as

$$s(j) = ln\left(\frac{f_X(x_{(i,j)}|\mu_1)}{f_X(x_{(i,j)}|\mu_0)}\right) \tag{A3.1}$$

and **cumulative log-likelihood ratio** which is the cumulative sum of instanteneous ratios up to the position j, i.e.

$$S(j) = \sum_{t=1}^{t=j} s(t) = \sum_{t=1}^{t=j} ln\left(\frac{f_X(x_{(i,t)}|\mu_1)}{f_X(x_{(i,t)}|\mu_0)}\right)$$

So, when the coordinate j passed through the change time j_c, i.e. $j > j_c$, we have

$$S(j) = \sum_{t=1}^{t=j_c-1} ln\left(\frac{f_X(x_{(i,t)}|\mu_1)}{f_X(x_{(i,t)}|\mu_0)}\right) + \sum_{i=j_c}^{t=j} ln\left(\frac{f_X(x_{(i,t)}|\mu_1)}{f_X(x_{(i,t)}|\mu_0)}\right)$$

$$S(j) = S(j_c - 1) + L_X(j_c)$$

$$L_X(j_c) = S(j) - S(j_c - 1)$$

As generalised log-likelihood ratio is the same as maximised likelihood ratio, we can rewrite it as

$$G_X(j) = \max_{1 \le j_c \le j-1} L_X(j_c) = S(j) - \min_{1 \le j_c \le j-1} S(j_c - 1)$$

Analogically, the optimal change position $\widehat{j_c}$ turns into

$$\widehat{j_c} = \underset{1 \le j_c \le j-1}{\arg\max}\ L_X(j_c) = \underset{1 \le j_c \le j-1}{\arg\min}\ S(j_c - 1) \tag{A3.2}$$

The cumulative log-likelihood ratio and the generalised log-likelihood ratio then can be rewritten in the recursive manner as

$$S(j) = \sum_{t=1}^{t=j} s(t) = \sum_{t=1}^{t=j-1} s(t) + s(j) = S(j-1)\ + s(j) \tag{A3.3}$$

$$G_X(j) = \max_{1 \le j_c \le j-1} (S(j-1) + s(j) - S(j_c - 1))$$

$$G_X(j) = \max\{G_X(j-1); G_X(j-1) + s(j)\} \tag{A3.4}$$

Overall the CUSUM algorithm can be described as follows. We have an ordered set of experimental values $x_{(i,j)}$ where $j = 1, 2$, ..., $i - 1$ and for each j we compute the instantaneous log-likelihood ratio based on Formula A3.1. In simple words, this ratio represents which hypothesis, null or alternative, fits the observed data point most probably: when the ratio is positive, the data better fits the alternative hypothesis (shift in

mean is present); when the ratio is negative, the data better fits the null hypothesis (no shift in mean). Then, we compute cumulative and generalised log-likelihood ratios according to the Formulas A3.3 and A3.4, respectively. When $G_X(j)$ cross the pre-selected threshold value h, it means that there is a high chance that we accumulated large amount of data points that better fits the alternative hypothesis rather than the null hypothesis. At this moment we fix the detection time j_d. According to the Formula A.3.2, the change point j_c can be fixed one step before the moment when $S(j)$ reaches its minimum such as $j_c < j_d$ (change point happens before the detection point). Note that as the pre-selected threshold h should be a positive number, the Formula A3.4. can be simplified as

$$G_X(j) = \max\{0; G_X(j-1) + s(j)\} \tag{A3.5}$$

Appendix 3.1.3. Dealing with unknown parameters

The algorithm relies on several assumptions that should be made. Assumptions deal with pdf $f_X(x|\mu)$ that defines the value of instantaneous log-likelihood ratio the decision threshold h that is used to decide between two hypotheses H_0 and H_1.

Normal probability density function. Each particular observation $x_{(i,j)}$ in Hi-C matrix represents the number of pair-wise interactions counted during the Hi-C experiment in a population of cells. However, we require the low quality read filtering and matrix correction to reduce the influence of experimental biases which should be made before the TAD edge detection. As a result, the discrete count data in raw Hi-C turns into continuous data in corrected data. So, we would select Poisson density function if we decide to skip the correction step, and we would select Normal distribution if we apply the matrix correction step.

The Normal distribution has one advantage over the Poisson distribution. Normal pdf depends on two parameters – mean μ and varianceσ^2 – that can be set independently from each other. The Poission pdf, in contrast, depends on single mean parameter and models variance to be equal to the mean. This finding can become critical particularly for Hi-C data: Hi-C interactions tend to demonstrate the variance

being usually greater than mean. So, assuming the Poisson distribution would lead to overdispersion: the observations would demonstrate variation which is higher than expected.

At the moment, we assume that the Normal pdf function to implement the algorithm for TAD edge detection. For simplicity, we also assume the variance parameter is known and stays the same for all samples.

Replacement alternative mean parameter by introducing shift parameter. Instantaneous log-likelihood ratio is defined by the form of pdf, which explicit form, in turn, is defined by the mean before the shift $\mu = \mu_0$ and mean after the shift $\mu = \mu_1$. In practice, we may have some prior information about the mean value before the shift but the mean value after the shift is mostly unknown. In this case, the simplest solution is to introduce the shift parameter δ that represents the difference in mean value before and after the change, in other words, $\mu_1 = \mu_0 + \ \delta$. In case of Normal distribution, the instantaneous log-likelihood ratio becomes

$s(j) \ = \frac{\delta}{\sigma^2} \left(x_{(i,j)} - \mu_0 - \frac{\delta}{2} \right)$

Note that shift parameter is priory unknown as well as the mean before the shift μ_0 and the variance σ^2.

Estimate null mean parameter using maximum likelihood estimation. The unknown mean before shift $\mu = \mu_0$ can be replaced by its maximum likelihood estimate based on the available samples at time j, i.e.

$\widehat{\mu_0}(j) = \frac{1}{j} \sum_{t=1}^{t=j} x_{(i,t)}$ and $\widehat{\sigma^2}(j) = \frac{1}{t-1} \sum_{t=1}^{t=j} (x_{(i,t)} - \widehat{\mu_0})^2$

Note that in case of Normal distribution we are able to run the algorithm starting with the coordinate $t = 2$ only as variance estimate is required to be different from zero. The maximum likelihood-based algorithm is fromalised in Appendix Algorithm 3.1.1.

Large fluctuations in data may dramatically affect the maximum likelihood estimate of mean parameter, especially when the coordinate j is not large and not so many samples are available. It is not critical for samples that are expected to have no shifts in mean for long time and could be controlled through the detection threshold h: large increase in mean estimate at the beginning of the algorithm (when number of obser-

vations is not large) will increase the value $G_X(j)$ and may beat the threshold h while the shift in observations appeared because of the noise and not because of real shift in mean; increasing the threshold allows to move the detection time from the beginning to the moment when the real change in mean takes place and more observations that are most probably indicate shift in mean would be accumulated. Although, for samples where the time periods with no shift are relatively short the maximum likelihood estimation may be not appropriate. For example, it is the case of nested TADs. If we assume that within the Hi-C row i we should detect at least two TAD edge, when the first change point j_c is detected, we need to reset the algorithm starting now from $j = j_c$. The distance between two neighbouring TAD edges can be relatively short, so there is a high chance that values accumulated $G_X(j)$ represents the spurious "jumps" in observations.

Appendix 3.1.4. Estimate null mean parameter using Bayesian update method

In case of samples with short periods of stable mean to be expected, we may use the method that relies on some prior knowledge about the mean parameter before any sample value $x_{(i,j)}$ is observed. Then, the mean parameter is updated with respect to the new information coming from the new data point. This method will reduce the effect of the large data fluctuations on the mean estimate, providing us with more stable mean parameter independently on the length of the stability period.

Suppose that $X_{(i,j)}$ for $j = 1$, ... $i - 1$ are independent random variables that follows Normal distribution. Mean parameter μ_0 is the same for all variables before the change time j_c and increases by δ at the change time j_c. Therefore, $X_{(i,1)}$, ..., $X_{(i,j_c-1)}$ follow Normal distribution with mean $\mu = \mu_0$ and $X_{(i,j_c)}$, ..., $X_{(i,i-1)}$ follows Normal distribution with mean $\mu = \mu_1$ where $\mu_1 = \mu_0 + \delta$. We assume that the variance parameter is known and stays the same for all samples.

Suppose we observed the experiment values $\mathbf{x_0} = (x_{(i,1)}, ..., x_{(i,j_c-1)})$ before the change time, they can be viewed as known values. Then, the likelihood function is written as

$$L\left(\mu_0|\mathbf{x_0}\right) \propto \prod_{j=1}^{j=j_c-1} exp\left\{\frac{1}{2\sigma^2}(x_{(i,j)}-\mu_0)^2\right\}$$

$$L\left(\mu_0|\mathbf{x_0}\right) \propto \exp\left\{\frac{1}{2\sigma^2}\sum_{j=1}^{j=j_c-1}\left(x_{(i,j)}-\mu_0\right)^2\right\}$$

$$L\left(\mu_0|\mathbf{x_0}\right) \propto \exp\left\{\frac{1}{2\frac{\sigma^2}{j_c-1}}\left(\mu_0-\frac{\sum_{j=1}^{j=j_c-1}x_{(i,j)}}{j_c-1}\right)^2\right\}$$

So, μ_0 follows Normal distribution with mean $\frac{\sum_{j=1}^{j=j_c-1}x_{(i,j)}}{j_c-1}$ and variance $\frac{\sigma^2}{j_c-1}$. As $\mu_1 = \mu_0 + \delta$, we can conclude that μ_1 follows Normal distribution with mean $\frac{\sum_{j=1}^{j=j_c-1}x_{(i,j)}}{j=j_c-1} + \delta$ and variance $\frac{\sigma^2}{j_c-1}$.

Then, we observe the new experiment value $x_{(i,j_c)}$ where $X_{(i,j_c)}$ follows Normal distribution with mean μ_1 and variance σ^2. It is the information brought in by the data and could be used in order to update the prior knowledge about the mean parameter μ_1. As we deal with normally distributed prior and normally distributed likelihood, we can conclude that the posterior distribution of μ_1 is Normal distribution with

$$\text{mean} = \frac{\left(\frac{\sum_{j=1}^{j=j_c-1}x_{(i,j)}}{j_c-1}+\delta\right)\sigma^2+x_{(i,j_c)}\frac{\sigma^2}{j_c-1}}{\sigma^2+\frac{\sigma^2}{j_c-1}} = \frac{\sum_{j=1}^{j=j_c-1}x_{(i,j)}+x_{(i,j_c)}+\delta(j_c-1)}{j_c}$$

$$\text{variance} = \frac{\sigma^2\frac{\sigma^2}{j_c-1}}{\sigma^2+\frac{\sigma^2}{j_c-1}} = \frac{\sigma^2}{j_c}$$

Before the new experiment value $x_{(i,j_c+1)}$ is observed, the Normal posterior distribution of μ_1 could be treated as prior knowledge that we want update with the new information. The posterior distribution of μ_1 then is Normal with

$$\text{mean} = \frac{\sum_{j=1}^{j=j_c-1}x_{(i,j)}+x_{(i,j_c)}+x_{(i,j_c+1)}+\delta(j_c-1)}{j_c+1}$$

$$\text{variance} = \frac{\sigma^2}{j_c+1}$$

Generalising the procedure, when the new experiment value $x_{(i,k)}$ where $k > j_c$ is observed, the posterior distribution of μ_1 is Normal with

$$\text{mean} = \frac{\sum_{j=1}^{j=k}x_{(i,j)}+\delta(j_c-1)}{k}$$

$$\text{variance} = \frac{\sigma^2}{k}$$

Detailed above derivation describes the procedure that allows to simulate the μ_1 with new-coming observations after the first shift is detected. The advantage of the method is that the simulated values of μ_1 rely on the whole history of observations

which were before the change point j_c. So, it is expected that on relatively short datasets, simulated value of μ_1 would be less variable than the value of μ_1 estimated with maximum likelihood.

These derivations also can be used to estimate μ_0 before the first shift detected. Pairs of loci which are allocated far away from the diagonal (for example, on the corner of Hi-C matrix build for Mb-scale genomic regions) demonstrate the interactions which are approximately uniform and close to zero (name it as background noise). First, we use maximum likelihood estimation for mean and variance parameters based on these distal interactions. Then, when we start the algorithm at $j = 1$, we treat this moment as the first observation after the zero shift from the Hi-C background noise.

Appendix 3.2. Threshold-based TAD edge detection algorithm

In this section we introduce the threshold-based TAD edge detection algorithm allowing to call the visually clear log2 mean ratio stripes which indicate the area of most probable allocation of left TAD edges within Hi-C interaction matrix (see Chapter 3 for more details). This algorithm generates the rectangle areas covering the stripes on log2 mean ratio heat map. Note that the algorithm was built for the purpose of hypotheses testing and as a basis for downstream modeling. The algorithm has very little applicability for real data analysis because of its sensitivity to the noise.

Appendix 3.2.1. Algorithm notations and assumptions

We fix the single Hi-C row $i = 1$, ..., N where N is the number of DNA fragments. We observe $x_{(i,j)}$ Hi-C interactions between locus i and locus $j = 1$, ..., $i - 1$. We introduce log2 mean ratio at position j as $log_2(MA_{(i,j+1)}/MA_{(i,j-1)})$ with $MA_{(i,j+1)}$ and $MA_{(i,j-1)}$ are defined as:

$$MA_{(i,j+1)} = \frac{1}{w}(x_{(i,j+1)} + x_{(i,j+2)} + ... + x_{(i,j+w)})$$

$$MA_{(i,j-1)} = \frac{1}{w}(x_{(i,j-1)} + x_{(i,j-2)} + ... + x_{(i,j-w)})$$

where w is MA estimation window size. The log2 mean ratio represents the ratio of downstream over upstream average interaction frequency. We aim to select the candidate positions $\hat{j}$ such that the change in contact mean from upstream to downstream is at least at some pre-selected level that we denote as $threshold$ (Appendix Figure 3.2.A-B). So, the following inequality should be satisfied:

$$log_2(MA_{(i,\hat{j}+1)}/MA_{(i,\hat{j}-1)}) > threshold$$

Set of column candidate positions $(\hat{j}) = (\hat{j_1}, \hat{j_2}, \hat{j_3}, ...)$ consists of smaller sub-sets (Appendix Figure 3.2.C), which can be characterised as:

- consecutive, meaning that we can identify the window where the possible TAD edge is allocated,
- single peaks separated from each other by large gaps, meaning that, most probable, we detect jumps in signal because of noise,
- the mix of consecutive regions, gaps and peaks.

To be a TAD edge position, the log2 mean ratio at $\hat{j}$ should exceed the $threshold$ in several Hi-C rows, so should be detected more than ones within the analysed Hi-C interaction matrix (Appendix Figure 3.2.C). So, we introduce the **Frequency Rule** in order to remove column positions that do not cross the threshold or cross it in a single Hi-C row only.

Frequency Rule. We collect all column candidate positions, which have log2 mean ratio being greater or equal than pre-selected threshold, from all Hi-C rows within the investigating genomic region. Then, we compute how many times we observe each candidate position. The computed frequency represents in how many Hi-C rows we observe each column candidate position. If the frequency is 1, then we exclude this candidate position from the analysis. If the frequency is greater than 1, we leave the candidate position. It allow us to leave only column positions that are observed in more than one Hi-C row.

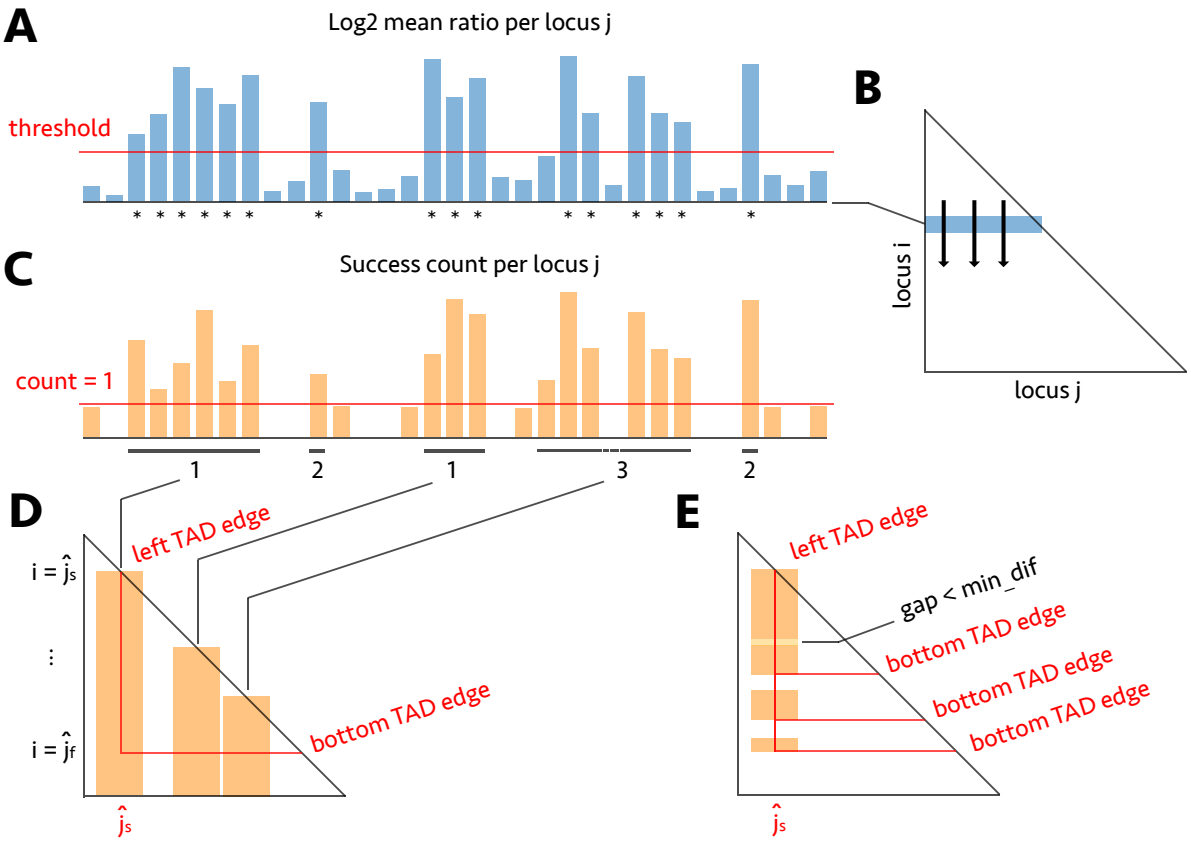
We assume that single TAD edge position can be found within subset of consecutive column candidate positions (Appendix Figure 3.2.C, scenario 1). It means that when the single column candidate position $\hat{j}$ is separated from other candidate positions by long gaps we possibly detects this $\hat{j}$ because of the data noise (Appendix Figure 3.2.C, scenario 2). As well as single column candidates, short gaps have to be removed. When two subsets, which members are placed continuously, are separated by short gap, it is the high chance that we observe the gap because of the noise, so we need to aggregate these subsets into one (Appendix Figure 3.2.C, scenario 3). Altogether, we aim to detect only subsets with consecutive column positions. We introduce the **Consecutive Rule** in order to select consecutive positions including the case when we need to aggregate two consecutive subsets if they are separated by single column position.

Consecutive Rule. We are interested in only consecutive column positions, meaning that if $(\hat{j}_1, \hat{j}_2, ..., \hat{j}_5)$ is consecutive subset then $\hat{j}_2 = \hat{j}_1 + 1$, $\hat{j}_3 = \hat{j}_2 + 1$ and so on. If we observe a break between candidate positions, for example, between positions $\hat{j}_4$ and $\hat{j}_3$, that, mathematically, can be written as $\hat{j}_4 > \hat{j}_3 + 1$, we have two options. In first option, we observe that after the position $\hat{j}_4$ there is a new consecutive region, i.e. $\hat{j}_5 = \hat{j}_4 + 1$, so we leave the whole subset $(\hat{j}_1, \hat{j}_2, ..., \hat{j}_5)$. In second option, we observe that after the position $\hat{j}_4$ there is a large break again, i.e. $\hat{j}_5 > \hat{j}_4 + 1$, so we remove $\hat{j}_4$ and $\hat{j}_5$ from the analysis. We continue to apply the rule until only consecutive regions remain.

First of all, we have to note that according to Consecutive Rule, at this point we treat as a "short gap" only gaps which lengths are 1 position only. Using this assumption, we face a danger to generate too many long consecutive subsets separated by relatively short gaps, so we have to either modify the Consecutive Rule, or take into account such gaps in further steps of the algorithm. Second thing is that the order of rules to apply is critical. The Frequency Rule applied after the Consecutive Rule would remove some elements from consecutive subsets and, as a consequence, generate single column candidates with computed frequencies to be higher than 1, but separated from consecutive regions by large gaps. However, the Consecutive Rule applied after the Frequency Rule insures that consecutive regions mostly (not all) formed by column positions that are observed in more than one row in Hi-C contact map. The start and the end of each consecutive region are, respectively, the left and the right coordinates of the window that contains the TAD edge position (Appendix Figure 3.2.D).

Appendix 3.2.2. Detection of bottom TAD edge requires the usage of Frequency and Consecutive Rules as well

We defined the set which includes several subsets of consecutive column positions and each of them represents the window around particular left TAD edge. Suppose that there is TAD edge at the position $\hat{j}_s$ and it is included into subset $(..., \hat{j_{s-1}}, \hat{j}_s, \hat{j_{s+1}}, ...)$. The same $\hat{j}_s$ is expected to be a left TAD edge position for several neighbouring Hi-C



Appendix Figure 3.2. Step-by-step performance of the threshold-based TAD edge detection algorithm. **A-B.** At each Hi-C row, we define column positions at which log2 mean ratio exceeds the threshold. Red line represents the threshold value and starts represent the success event (log2 mean ratio beats the threshold). **C.** We sum number of successes per each column. According to frequency rule we select column positions which demonstrate the number of successes being greater than 1. There are three scenarios: (1) consecutive column positions; (2) single position separated from neighbouring consecutive positions by large gaps (removed according to Consecutive Rule); (3) Consecutive regions separated by short gaps of 1 position (aggregated according to Consecutive Rule). **D.** Scenarios 1 and 3 form the left and right coordinates of TAD edge squares (coloured with orange). Each rectangle area defines the window where left TAD edge is allocated. **E.** Applying Frequency and Consecutive Rules at Hi-C rows within each subset of consecutive column positions, we generate TAD edge squares where top and bottom coordinates represent the window where bottom TAD edge is allocated. TAD edge squares separated by less than minimum with difference between neighbouring TAD borders (user-selected parameter) are aggregated. Left TAD edge is placed at the centre between left and right coordinates, bottom TAD edge is placed at bottom coordinate.

rows $i = \hat{j_s}, \hat{j_{s+1}}$,..., $\hat{j_f}$ where $\hat{j_f}$ is bottom TAD edge position. This definition represents the left TAD edge as a straight line starting at the Hi-C diagonal and finishing at specific Hi-C row $i = \hat{j_f}$ (Appendix Figure 3.2.D).

As the TAD edge is defined as a straight line, sequence of Hi-C rows between the diagonal and bottom TAD edge position should be continuous. So, for each subset of consecutive column positions (..., $\hat{j_{s-1}}, \hat{j_s}, \hat{j_{s+1}}$, ...) we aim to define the subsets of continuous row positions $\hat{i}$ such as

$log_2(MA_{(\hat{i},\hat{j}+1)}/MA_{(\hat{i},\hat{j}-1)}) > threshold$

Technically speaking, we need to apply similar Frequency and Consecutive Rules within the subset of columns (..., $\hat{j_{s-1}}, \hat{j_s}, \hat{j_{s+1}}$, ...) and call consecutive subsets within Hi-C rows. Despite the fact that TAD edge is defined as a straight line and there should be single consecutive set of rows, we used a term “subsets” instead and there are two reasons. First reason is technical ((Appendix Figure 3.2.E). Due to a noise, we expect several gaps within consecutive row position subsets. If we remove only gaps of 1 position, according to the Consecutive Rule, and select the first subset to represent continuous TAD edge, we face a danger to fix the TAD edge with much shorter length that expected. Second reason is associated with nested TADs. If we assume that we do not allow nested TADs, then the bottom TAD edge position must be the only one. In opposite, taking into account nested TADs we can have several TADs with the same start edge and different end edges. So, we expect to observe several consecutive regions separated by large gaps.

The Frequency Rule would allow only Hi-C rows that demonstrate more than one column position $\hat{j}$ where log2 mean ratio cross the threshold. So, if the number of such positions $\hat{j}$ is exactly 1, we exclude this candidate row from the analysis. If the number of such positions $\hat{j}$ is greater than 1, we leave the candidate row. After removing unreliable row candidates with the Frequency Rule, the set of row candidates can be:

- consecutive, meaning that all candidate rows belong to the start edge and the end of consecutive region represents the position of end edge,
- single peaks separated from each other by large gaps, meaning that, most prob-

able, we caught single jumps in signal because of noise,

- the mix of consecutive regions, gaps and peaks.

After applying the Consecutive Rule, within each column consecutive subset we generate several row consecutive subsets. The start and end of each row consecutive region are, respectively the top and the bottom coordinates of the window that contains the bottom TAD edge position.

Appendix 3.2.3. The neighbouring consecutive row candidates can be aggregated, otherwise, nested TADs are allocated too close to each other

The gaps between consecutive row candidates are not restricted. After applying the Consecutive Rule, the consecutive regions may be separated by both short and large gaps. Imagine the following scenario: there are two neighboring relatively short consecutive regions, containing just 3 rows each, and they are separated by the short gap, just 3 rows as well. In this case, we define two nested TADs and their end edges are allocated from 3 to 9 rows apart (as the end edge should be allocated somewhere within the consecutive region). So, the difference in these TAD widths is expected to be from 3 to 9 bins.

There are different ways to define the criteria or rule to aggregate two neighbouring consecutive regions separated by the relatively short gap into one large region. At the current stage, we are mainly interested in the ability of the proposed algorithm to detect the TAD edges that are, first, coincide with visually clear TAD edges in Hi-C interaction matrix map, and, second, coincide with the reliable TADs called by other tools (we used HiCExplorer in Chapter 3). As a result, the criteria to aggregate the neighbouring consecutive regions or not can be stated in a simple way by user-controlled parameter to set the minimum TAD width difference (Appendix Figure 3.2.E). If the gap between two neighbouring consecutive regions is less than the minimum TAD width difference we aggregate to regions into one, if the gap is greater than the minimum TAD width difference we leave two regions to be separated.

Appendix 3.2.4. Threshold-based TAD edge detection algorithm produces the TAD edge squares which can be further reconstructed into TADs

We got the left/right coordinates of the windows containing the left TAD edges and top/bottom coordinates of the windows containing bottom TAD edges. The combination of these 4 coordinates creates the structure that we name as “TAD edge square”. The TAD edge squares are expected to shrink when the threshold value increases: there is the lower chance for log2 mean ratio to cross the higher threshold values, so the less column positions $\hat{j}$ and less row positions $\hat{i}$ to be included.

We also can reconstruct TAD edges based on squares to make them more canonical to the Hi-C user’s eyes (Appendix Figure 3.2.E). At the current stage, we propose the left TAD edge to be allocated exactly in the centre of the TAD edge square. With respect to the bottom TAD edge, we propose to allocate it in the bottom of the TAD edge square - all row candidates within the TAD edge square should belong to the same TAD, so the last row belonging to the square represents the last row belonging to the TAD.

Appendix 3.2.5. Algorithm formalisation suitable for programming language implementation

Here we provide the algorithm in its formalised format. The whole procedure can be split in six main stages. Line numbers below refer to the specific lines in formalised algorithm, see Appendix Algorithm 3.2 for details.

- Stage 1 (lines 1-6) describes the selection of candidate column positions that are potential left TAD edge positions.
- Frequency Rule (lines 7-8) is applied further on column candidate positions collected in Stage 1. In simple words, the Rule removes all the candidate column positions that are observed only in single Hi-C row and cannot be reliable left TAD edge position candidate.
- Consecutive Rule (lines 9-17) allow us to leave only consecutive column candidate positions and remove single positions that are separated from others by

gaps. The start of each consecutive region represents the most left TAD edge position, the end of each consecutive region represents the most right TAD edge position.

- For each pair left/right position, during Stage 2 (lines 23-31) we collect candidate row positions that have log2 mean ratio exceeding the threshold value in at least two column positions. So, in this stage we at the same time search for row candidate positions and remove the unreliable ones with Frequency Rule.
- We want to leave only consecutive row candidates, so we apply the Consecutive Rule (lines 32-39) on row candidate positions.
- Then, if two row consecutive regions are closer than minimum allowed break, we join two consecutive regions (lines 41-45). The start and end of aggregated region represents top and bottom positions, respectively, of TAD edge squares. If two regions are not aggregated, we have two separate TAD edge squares with top side position at starts of the regions and bottom side position at ends of the regions.

Depending on the threshold value provided by the user, we can either have no TAD edge squares if the threshold value is extremely high, or have TAD edge squares represented by four coordinates each, where first coordinate is the left side position of the square, second coordinate is for the right side, third coordinate is for top side and last coordinate is for bottom square side position.

Appendix Algorithm 3.2. Threshold-based TAD edge detection algorithm.

Data: $N \times N$ Hi-C matrix where $x_{(i,j)}$ is a matrix entry on row i and column j

Result: 4 coordinates of TAD edge squares

User Inputs:

$w =$ MA estimation window width

$threshold =$ threshold value that log2 mean ratio should exceed

$min\_dif =$ minimum width difference in nested TADs

Initialize:

$\hat{I} =$ set of row candidate positions, $\hat{I} \leftarrow \{\}$

$\hat{J} =$ set of column candidate positions, $\hat{J} \leftarrow \{\}$

$LEFT =$ set of left side positions of TAD edge squares, $LEFT \leftarrow \{\}$

$RIGHT =$ set of right side positions of TAD edge squares, $RIGHT \leftarrow \{\}$

$TOP =$ set of top side positions of TAD edge squares, $TOP \leftarrow \{\}$

$BOTTOM =$ set of bottom side positions of TAD edge squares, $BOTTOM \leftarrow \{\}$

for $i \leftarrow (2w+1)$ **to** n **do**
    for $j \leftarrow (1+w)$ **to** $(i-1-w)$ **do**
        $M_{(i,j+1)} =$ mean over non-NA elements $(x_{(i,j+1)}, x_{(i,j+2)}, ..., x_{(i,j+w)})$
        $M_{(i,j-1)} =$ mean over non-NA elements $(x_{(i,j-1)}, x_{(i,j-2)}, ..., x_{(i,j-w)})$
        if log2$(MA_{(i,j+1)}/MA_{(i,j-1)}) > threshold$ **then**
            collect j into set $\hat{J}$
if element in set $\hat{J}$ is observed only once **then**
    remove it from set $\hat{J}$
order elements in set $\hat{J}$
$k = 1$
while $k < |\hat{J}|$ **do**
    if $\hat{j}_{k+1} - \hat{j_k} = 1, (j_k, j_{k+1}) \in \hat{J}$ **then**
        collect j_k into set $LEFT$
    else
        if $k > 1$ **then**
            collect j_k into set $RIGHT$
            $k = k + 1$

if $LEFT$ and $RIGHT$ sets are empty **then**
    stop the algorithm
order elements in set $LEFT$ and set $RIGHT$
$t = 1$
while $t < |LEFT|$ **do**
    for $i \leftarrow (2w + 1)$ **to** n **do**
        $count = 0$
        for $j \leftarrow left_t \in LEFT$ **to** $right_t \in RIGHT$ **do**
            $M_{(i,j+1)}$ = mean over non-NA elements $(x_{(i,j+1)}, x_{(i,j+2)}, ..., x_{(i,j+w)})$
            $M_{(i,j-1)}$ = mean over non-NA elements $(x_{(i,j-1)}, x_{(i,j-2)}, ..., x_{(i,j-w)})$
            if log2$(MA_{(i,j+1)}/MA_{(i,j-1)}) > threshold$ **then**
                $count = count + 1$
        if $count > 1$ **then**
            collect i into set $\hat{I}$
    $k = 1$
    while $k < |\hat{I}|$ **do**
        if $\hat{i}_{k+1} - \hat{i_k} = 1, (i_k, i_{k+1}) \in \hat{I}$ **then**
            collect k into set TOP
        else
            if $k > 1$ **then**
                collect k into set $BOTTOM$
                $k = k + 1$
    order elements in set TOP and set $BOTTOM$
    if TOP contains more than 1 element **then**
        for $s \leftarrow 1$ **to** $(|TOP| - 1)$ **do**
            if $top_{s+1} - bottom_s \leq min\_dif + 1, top_{s+1} \in TOP, bottom_s \in BOTTOM$ **then**
                remove top_{s+1} from set TOP
                remove $bottom_s$ from set $BOTTOM$
    if TOP and $BOTTOM$ sets are not empty **then**
        for $s \leftarrow 1$ **to** $|TOP|$ **do**
            combination of $(left_t, right_t, top_s, bottom_s)$ forms singe TAD edge square
stop the algorithm

Appendix 3.3. COrTADo R-based implementation

In this section we provide the detailed technical overview on the TAD edge calling procedure that is currently implemented in R as a collection of functions. To call TAD edges it is necessary to perform the following basic steps:

1. Transform Hi-C input data into a list format.
2. Compute log2 mean ratios row-wise and column-wise.
3. Allocate start and end TAD edge candidates.
4. Perform edges depth estimation.
5. Perform validation test and multiple test correction.
6. Select edges which satisfy the pre-selected thresholds.

Note that the usage example provided below is based on the parameters selected for the genome-wide COrTADo calling presented in Chapter 3 based on *Drosophila melanogaster* BG3 cells at DpnII resolution. We have run the analysis per chromosome, therefore in this Section we would specify the single chromosome 3R.

Appendix 3.3.1. Prepare for analysis, load the file and transform into list

The input format is a .bed or .tsv, the file contains the table with seven columns where first three represents the interacting row locus i (chromosome, start and end positions), next three represents the interacting column locus j (chromosome, start and end positions) and the last one represents the interaction frequency (discrete if Hi-C data is non-normalised and continuous if Hi-C data is normalised). Before transforming into the list we have to take into account two things. First, the table represents only the upper triangle of Hi-C matrix, i.e. interactions between pair of loci (i, j) where $j \geq i$. Second, the table omits non-interacting pairs, i.e. if reads for (i, j) interactions are not mapped to the genome, the pair (i, j) is not represented in the table.

Assume that input .tsv file has a name `Example.tsv` and it can be loaded into R as data.frame object, as shown below:

```
df <- read.table(file = "Example.tsv", header = FALSE, sep = "\t")
```

Then, we want to focus the further analysis on specific genomic region. It has to be given in the following format `chr:start-end`. If the genomic region is the whole chromosome, then the format is simply `chr`.

```
region = "3R"
chr <- as.character(unlist(strsplit(region, ":"))[1])
start <- as.numeric(unlist(
        strsplit(unlist(strsplit(region, ":"))[2], "-"))[1])
end <- as.numeric(unlist(
        strsplit(unlist(strsplit(region, ":"))[2], "-"))[2])
if (is.na(start) | is.na(end)){
    select_df <- df[df[,1] == chr & df[,4] == chr, ]
} else {
    select_df <- df[
        df[,1] == chr & df[,4] == chr &
        df[,2] >= start & df[,5] >= start &
        df[,3] <= end & df[,6] <= end, ]}}
```

The variable `select_df` contains the Hi-C contact data restricted by the provided region only.

We transform the table into two lists: the list where each element represents the interactions happen within single Hi-C row (for COrTADo start) and the list where each element represents the interactions happen within single Hi-C column (for COrTADo end). So, further computations can be performed separately (and in parallel) within each Hi-C row/column. In case when we call TADs genome-wise (or in single chromosome), we do not need the whole row to be extracted as it requires more memory and more processing time - we need only the part of the matrix that is close to the main diagonal and the size of this part is related to the maximum TAD width that we expect to detect (see Chapter 3 for more details). We restrict the matrix through the parameter `limit_size`.

To transform the table into the list where each element represents the interactions within single Hi-C row/column, we, first, need to identify the gaps - fragments that do not interact with each other, so they are not included into the table, then, when we identified the full list of loci within the analysing region (including gaps), we fill the list either with observed interactions where they are present in the input table or with NAs/zeros where we observe gaps. To do this, we run the following function:

```
hic_region_list_per_row <- transform_hictable2list(
        data_table = select_df, direction = "row",
        fill_empty = FALSE, replace_zero = FALSE,
        limit_size = NA, resolution = NA, cores = select_n_cores)
```

```
hic_region_list_per_col <- transform_hictable2list(
        data_table = select_df, direction = "column",
        fill_empty = FALSE, replace_zero = FALSE,
        limit_size = NA, resolution = NA, cores = select_n_cores)
```

The function has two required arguments `data_table` and `direction`, other four arguments are optional. The detailed description of the parameters is summarised in Appendix Table 3.3.1. In the Chapter 3, we specified the following parameters:

`select_limit_size = 1000`

`select_n_cores = 30`

When we run the function with `direction = "row"`, the function produces the list where each element named as row locus and contains a data.frame with three columns: column one is `col_locus` is the column locus name, column two is `col_ind` is the column locus index within the selected Hi-C dataset (note, not a whole Hi-C dataset) and `contact` is the interaction frequency between column locus and row locus. Note that the locus names are stored in the following format `chr_start`. When we run the function with `direction = "column"`, the format is the same but each element of the list represents the data extracted within single Hi-C column.

Appendix 3.3.2. Compute log2 mean ratios

We compute the log2 mean ratios that are used to identify the difference between the mean values on the regions that are on the right and on the left from each Hi-C bin. When we call COrTADo start, for each column position (where appropriate), we extract the interactions within the window of selected size on the right-hand-side and on the left-hand-side and compute mean values withing these windows (see Formula 3.3 and 3.4). When we call COrTADo end, we do the same computations column-wise for each row position. Then, we divide right-hand-side mean over left-hand-side mean and take log2 of the result to get the log2 mean ratio at the analysed position. To do so, we run the function `compute_log2mean` on the Hi-C data that is extracted row-wise and column-

wise and stored in variables log2mean_list_per_row and log2mean_list_per_col, respectively:

```
log2mean_list_per_row <- compute_log2mean(
                data_list = hic_region_list_per_row,
                window_size = select_window_size,
                replace_zero = FALSE, cores = select_n_cores)
```

```
log2mean_list_per_col <- compute_log2mean(
                data_list = hic_region_list_per_col,
                window_size = select_window_size,
                replace_zero = FALSE, cores = select_n_cores)
```

The list of required and optional parameters is specified in Appendix Table 3.3.2. In the Chapter 3, we specified the following parameters:

```
select_window_size = 10
```

```
select_n_cores = 30
```

Function produces the list where each element named as row locus and contains a data.frame with three columns: column one is col_locus is the column locus name, column two is col_ind is the column locus index within the Hi-C matrix and log2mean is the log2 mean ratio computed at column locus position at specified row locus. If the specific data.frame contains only single row (NA, NA, NA) it means that the number of column positions within the row is not enough to compute the log2 mean ratios. Note that the locus names are stored in the following format chr_start.

Appendix 3.3.3. Call start and end TAD edges

Those column positions that have a local maximum of the log2 mean ratio within several neighbouring Hi-C rows are more likely to be start TAD edge positions. Then, start TAD edge positioning is based on four steps. First, extract all log2 mean ratios computed row-wise and store them column-wise, then detect the column positions that are candidates to be local maximum. Smoothing can be applied here - we can smooth the log2 mean ratios using Gaussian Kernel smoothing and call local maxima. The probability of a candidate position to be a local maximum is assessed by comparing the distributions of log2 mean ratios at the candidate position with the regions on the right and left by Wilcoxon Rank-Sum Test. We use the effect size statistics to access the

optimal length of the TAD edge - the number of neighbouring Hi-C rows that demonstrate approximately the same average log2 mean ratio before log2 mean ratio started to fade (decrease to zero) due to a distance decay. When effect size demonstrate first dramatic “drop”, we fix it as the optimal TAD edge length (depth), perform validation test, extract p-value, effect size and insulation strength (the average log2 mean ratio within the middle testing window). As the multiple tests performed on all candidate positions, the Bonferroni correction is applied. The positions that have adjusted p-values to be below the stated threshold are most likely to be start TAD edges. We also can apply the effect size threshold to remove the candidates with low effect size and which are most probably insignificant TAD edges.

Note that as in the usage example is build based on the analysis of single chromosome we did not specified the multiple testing correction method and thresholds. In order to complete the full, genome-wide analysis, we need to run the whole algorithm on each chromosome, join all the results and run the correction and remove candidates that did not pass the thresholds.

Also note that this is the detailed description for running the `call_startCOrTADo`. The procedure to call the COrTADo ends is exactly the same, the only difference is that the log2 mean ratios are computed column-wise and indicate the decrease in Hi-C contact frequencies, so we start with storing the log2 mean ratio row-wise and searching for local minima instead of maxima.

All stages are summarised in functions `call_startCOrTADo` and `call_endCOrTADo`:

```
startCOrTADo_df <- call_startCOrTADo(
        data_list = log2mean_list_per_row, replace_zero = FALSE,
        window_size = select_window_size,
        test_depth_step = select_window_size,
        bandwidth_size = NA, do_weighted = TRUE,
        prob_limit = NA, es_limit = NA,
        do_onesided = TRUE, do_prob_correction = FALSE,
        correction_method = NA, cores = select_n_cores)
```

```
endCOrTADo_df <- call_endCOrTADo(
        data_list = log2mean_list_per_col, replace_zero = FALSE,
        window_size = select_window_size,
        test_depth_step = select_window_size,
        bandwidth_size = NA, do_weighted = TRUE,
```

```
        prob_limit = NA, es_limit = NA,
        do_onesided = TRUE, do_prob_correction = FALSE,
        correction_method = NA, cores = select_n_cores)
```

The list of required and optional parameters is specified in Appendix Table 3.3.3. In the Chapter 3, we specified the following parameters:

`select_window_size = 10`

`select_n_cores = 30`

Required arguments

`data_table`	The input data.frame object, consists of seven columns where first three represents the interacting row locus i (chromosome, start and end positions), next three represents the interacting column locus j (chromosome, start and end positions) and the last one represents the interaction frequency (discrete if Hi-C data is non-normalised and continuous if Hi-C data is normalised).
`direction`	Data transformed into a list where each list represents single Hi-C row if `direction = "row"`. Each list represents single Hi-C column if `direction = "column"`.

Optional arguments

`fill_empty`	Logical value indicating whether the gaps should be considered in the data set or skipped. Default: TRUE.
`replace_zero`	Logical value indicating whether the gaps with non-available interactions should be replaced with zeros. Default: FALSE.
`limit_size`	The maximum distance between interacting pairs. Should be specified in bins. If not specified, whole Hi-C matrix is considered. Default: NA.
`resolution`	If bins are the same size, we need to specify here a bin size. If the parameter is not specified, the gaps will be filled with single NA/zero interaction independently on the number of bins stored within the gap. Skip if `replace_zero = FALSE`. Should be specified in bp. Default: NA.
`cores`	A non-negative integer. The number of cores to use to run processes in parallel. If NA, apply non-parallel computing. Default: NA.

Appendix Table 3.3.1. Parameters of `transform_hictable2list()` function.

Required arguments	
`data_list`	The output of `transform_hictable2list()` function produced at the previous stage. A list object where each element is a data.frame indicating Hi-C contacts extracted from specified row (`direction = "row"`) or column (`direction = "column"`).
`window_size`	A non-negative integer. The desired length of MA estimation window. Should be specified in bins.
Optional arguments	
`replace_zero`	Logical value indicating whether the log2 mean ratio resulting in NA should be replaced with zero. Default: FALSE.
`cores`	A non-negative integer. The number of cores to use to run processes in parallel. If NA, apply non-parallel computing. Default: NA.

Appendix Table 3.3.2. Parameters of `compute_log2mean()` function.

Required arguments	
`data_list`	The output of `compute_log2mean()` function produced at the previous stage. A list object where each element is a data.frame indicating log2 mean ratios computed row-wise (for `call_startCOrTADo()`) or column-wise (for `call_endCOrTADo()`) .
`window_size`	A non-negative integer. The desired width of validation test windows. Should be specified in bins.
`test_depth_step`	A non-negative integer. Indicates the increment of validation test window length. Should be specified in bins.
Optional arguments	
`replace_zero`	Logical value indicating whether the log2 mean ratio resulting in NA should be replaced with zero. Default: FALSE.
`bandwidth_size`	A kernel bandwidth smoothing parameter (see `ksmooth()` R documentation for more details). If NA the simple average of log2 mean ratio at the column position is used. Default: NA.
`do_weighted`	Logical value indicating whether the log2 mean ratios should be weighted or not. If TRUE, weights are computed as reciprocal of the distance between corresponding pairs of loci. Default: TRUE.
`prob_limit`	P-value threshold for the candidate position to be a local maximum (for `call_startCOrTADo()`) or local minimum ((for `call_endCOrTADo()`). Default: NA.
`es_limit`	Effect size threshold for the candidate position to be a local maximum (for `call_startCOrTADo()`) or local minimum ((for `call_endCOrTADo()`). Default: NA.

Appendix Table 3.3.3. Parameters of `call_startCOrTADo()` and `call_endCOrTADo()` functions.

Optional arguments

`do_onesided`	Logical value indicating the alternative hypothesis at Wilcoxon Rank-Sum test (validation test). If TRUE, the alternative is ”greater”, the validation test checks whether the distribution within middle window is greater than within left and right windows. If FALSE, the alternative is ”two.sided”, the validation test checks whether the distribution within middle window is different from the distributions within left and right windows. Default: TRUE.
`do_prob_correction`	Logical value indicating whether the multiple testing correction is applied. Default: TRUE.
`correction_method`	If the multiple correction method is applied, the method (”fdr” or ”bonferroni”) should be specified. Default: NA.
`cores`	A non-negative integer. The number of cores to use to run processes in parallel. If NA, apply non-parallel computing. Default: NA.

Appendix Table 3.3.3 (continue). Parameters of `call_startCOrTADo()` and `call_endCOrTADo()` functions.

Appendix 3.4. Inclusion of zero observations instead of NAs in the MA estimation procedure

Suppose we use the MA estimation method to estimate sample mean and sample variance. We select one particular window with width w. In this window only n observations are available, other $(w-n)$ observations are lost (NAs). We have two opportunities: either exclude all NAs from the estimation and use only n available data points, or include all NAs as zeros, meaning that we treat them as zero contacts between corresponding DNA fragments. In order to compare two opportunities, we can estimate mean and variance using only available data first, and, then, introduce the formulas to recalculate sample mean and sample variance when we add zero observations to the available data points. The formulas allow us to describe the relationship between sample mean and variance when we exclude zero observations and when we include them in the estimation procedure. We start with derivation of general formula that allow us to re-estimate mean and variance when one single observation arrives.

Appendix 3.4.1. Sample mean and variance can be simply recalculated when new observation appears

We estimate sample mean $\overline{x_n}$ and sample variance σ_n^2 based on n available data points. The sample mean $\overline{x_n}$ is defined according to the formula

$\overline{x_n} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$\sum_{i=1}^{n} x_i = n\overline{x_n}$

The standard formula for sample variance σ_n^2 is slightly modified

$\sigma_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x_n})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - n\overline{x_n}^2)$

$\sum_{i=1}^{n} x_i^2 = (n-1)\sigma_n^2 + n\overline{x_n}^2$

Then, the new observation x_{n+1} arrives. The sample mean and variance of the full data set $\overline{x_{n+1}}$ and σ_{n+1}^2, resectively, are defined as

$\overline{x_{n+1}} = \frac{1}{n+1}\sum_{i=1}^{n+1} x_i = \frac{1}{n+1}(\sum_{i=1}^{n} x_i + x_{n+1}) = \frac{1}{n+1}(n\overline{x_n} + x_{n+1})$

$\sigma_{n+1}^2 = \frac{1}{n}\left(\sum_{i=1}^{n+1} x_i^2 - (n+1)\overline{x_{n+1}}^2\right) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 + x_{n+1}^2 - \frac{1}{n+1}(n\overline{x_n} + x_{n+1})^2\right) =$

$$= \frac{1}{n}\left((n-1)\sigma_n^2 + n\overline{x_n}^2 + x_{n+1}^2 - \frac{1}{n+1}(n\overline{x_n} + x_{n+1})^2\right) = \frac{n-1}{n}\sigma_n^2 + \frac{1}{n+1}(\overline{x_n} - x_{n+1})^2$$

Appendix 3.4.2. Sample mean and variance estimated on full data set have exact formulas when new appearing data points are zeros

Assume that the new arriving point is zero, so $x_{n+1} = 0$. Then,

$$\overline{x_{n+1}} = \frac{1}{n+1}(n\overline{x_n} + 0) = \frac{n}{n+1}\overline{x_n}$$
$$\sigma_{n+1}^2 = \frac{n-1}{n}\sigma_n^2 + \frac{1}{n+1}(\overline{x_n} - 0)^2 = \frac{n-1}{n}\sigma_n^2 + \frac{1}{n+1}\overline{x_n}^2$$

So, we can re-estimate the sample mean and sample variance with new zero data point. Then, at the next step, we have another data point appearing that is again zero, i.e. $x_{n+2} = 0$. The new sample mean and variance then equal to

$$\overline{x_{n+2}} = \frac{1}{n+2}\left((n+1)\overline{x_{n+1}} + 0\right) = \frac{n}{n+2}\overline{x_n}$$
$$\sigma_{n+2}^2 = \frac{n}{n+1}\sigma_{n+1}^2 + \frac{1}{n+2}(\overline{x_{n+1}} - 0)^2 = \frac{n-1}{n+1}\sigma_n^2 + \frac{2n}{(n+1)(n+2)}\overline{x_n}^2$$

Repeating the same procedure with the third zero observation, i.e. $x_{n+3} = 0$, the sample mean and variance of full data set are defined as

$$\overline{x_{n+3}} = \frac{1}{n+3}\left((n+2)\overline{x_{n+2}} + 0\right) = \frac{n}{n+3}\overline{x_n}$$
$$\sigma_{n+3}^2 = \frac{n+1}{n+2}\sigma_{n+2}^2 + \frac{1}{n+3}(\overline{x_{n+2}} - 0)^2 = \frac{n-1}{n+2}\sigma_n^2 + \frac{3n}{(n+2)(n+3)}\overline{x_n}^2$$

Looking carefully at the pattern, we can generalise the formulas for k zero observations:

$$\overline{x_{n+k}} = \frac{n}{n+k}\overline{x_n}$$
$$\sigma_{n+k}^2 = \frac{n-1}{n-1+k}\sigma_n^2 + \frac{kn}{(n-1+k)(n+k)}\overline{x_n}^2$$

When we use the estimation window of width w and only n observations are available, other $(w-n)$ observations are not lost and may be treated to be zeros. So, if we estimate sample mean $\overline{x_n}$ and sample variance σ_n^2 based on n available data points and then we want to re-estimate them including $k = w - n$ data points that are not available as zeros, then the new mean and variance will be

$$\overline{x_w} = \frac{n}{n+w-n}\overline{x_n} = \frac{n}{w}\overline{x_n}$$
$$\sigma_w^2 = \frac{n-1}{n-1+w-n}\sigma_n^2 + \frac{(w-n)n}{(n-1+w-n)(n+w-n)}\overline{x_n}^2 = \frac{n-1}{w-1}\sigma_n^2 + \frac{n(w-n)}{w(w-1)}\overline{x_n}^2$$

Appendix 3.4.3. Inclusion of zeros in the estimation procedure reduces sample mean while sample variance behaves differently depending on the underlying conditions

In order to analyse the effect of zeros inclusion, we can, first, look at the difference between sample means with and without zeros

$$\overline{x_w} - \overline{x_n} = \frac{n}{w}\overline{x_n} - \overline{x_n} = \frac{n-w}{w}\overline{x_n} < 0 \text{ as } n \leq w$$

So, we expect to observe the reduction in the sample mean value when we include zeros when the observations are not available. Then, looking at the difference in variance, we have

$$\sigma_w^2 - \sigma_n^2 = \frac{n-1}{w-1}\sigma_n^2 + \frac{n(w-n)}{w(w-1)}\overline{x_n}^2 - \sigma_n^2 = \frac{w-n}{w-1}\left(\frac{n}{w}\overline{x_n}^2 - \sigma_n^2\right)$$

Note, that we expect to see the decrease in sample variance when $\left(\frac{n}{w}\overline{x_n}^2 - \sigma_n^2\right) < 0$ or, alternatively, $\frac{n}{w}\overline{x_n}^2 < \sigma_n^2$. When we have the opposite, i.e. $\frac{n}{w}\overline{x_n}^2 > \sigma_n^2$, we expect to see the increase in sample variance.

Appendix 3.4.4. Variance being greater than mean when we assume only the available data is present as well as when we include zeros

We assume the variance larger than mean, i.e. $\sigma_n^2 > \overline{x_n}$, when we exclude NAs from the estimation procedure and use only n available data points. We aim to investigate whether the variance is larger than mean when we include zeros instead NAs. In order to give an answer, we need to look at the value of the difference between sample variance and sample mean based on full data set including zeros:

$$\sigma_w^2 - \overline{x_w} = \frac{n(w-n)}{w(w-1)}\overline{x_n}^2 - \frac{n}{w}\overline{x_n} + \frac{n-1}{w-1}\sigma_n^2$$

We can treat this expression as the second degree polynomial of $\overline{x_n}$. Then, the graph of this function will be a parabola that is opened upwards as $\frac{n(w-n)}{w(w-1)} > 0$ and has its lowest point at $\overline{x_n} = \frac{w-1}{2(w-n)}$. Then, the lowest possible value of $(\sigma_w^2 - \overline{x_w})$ is

$$\frac{n-1}{w-1}\sigma_n^2 - \frac{n(w-1)}{4w(w-n)}$$

As $\sigma_n^2 > \overline{x_n}$, at the minimum point we have $\sigma_n^2 > \frac{w-1}{2(w-n)}$. It leads to the fact that

$$\frac{n-1}{w-1}\sigma_n^2 - \frac{n(w-1)}{4w(w-n)} > \frac{n-1}{w-1}\frac{w-1}{2(w-n)} - \frac{n(w-1)}{4w(w-n)} = \frac{1}{4(w-n)}\left(n\left(1+\frac{1}{w}\right) - 2\right) > 0 \text{ for } n \geq 2$$

In order to estimate sample variance, we need at least two available observations,

so we always have $n \geq 2$. So, we showed that the minimum value of the difference $(\sigma_w^2 - \overline{x_w})$ is always positive, meaning that this difference is everywhere positive. In other words, we will observe sample variance being greater than sample mean even if we include all non-available observations as zeros.