Skolkovo Institute of Science and Technology

CHROMATIN FOLDING IN INDIVIDUAL CELLS

Doctoral Thesis

by

ALEKSANDRA GALITSYNA

DOCTORAL PROGRAM IN LIFE SCIENCES

Supervisor
Vice-President for Biomedical Research, Professor, Mikhail Gelfand

Moscow - 2021

I hereby declare that the work presented in this thesis was carried out by myself at Skolkovo Institute of Science and Technology, Moscow, except where due acknowledgement is made, and has not been submitted for any other degree.

Candidate (Aleksandra Galitsyna)

Supervisor (Prof. Mikhail Gelfand)

Chromatin folding in individual cells

by

Aleksandra Galitsyna

Monday 11th October, 2021

Submitted to the Skoltech Center of Life Sciences
on October 2021, in partial fulfillment of the requirements for the
Doctoral Program in Life Sciences

Abstract

Three-dimensional chromatin of *Drosophila* nuclei is constituted of complex structures, such as compartments and topological domains. These structures were revealed by Hi-C, an advanced molecular biology method for probing the architecture of DNA in a population of cells. The readout of this method is a heatmap of averaged interactions in millions of cells, which cannot be readily deconvolved into three-dimensional structures of individual cells. Thus, the emergence and properties of compartments and topological domains remain poorly understood in insects. To overcome this limitation, we study the properties of chromatin in individual cells of *Drosophila*, obtained by single-cell Hi-C (scHi-C). However, we demonstrate that single-cell Hi-C data is profoundly sparse and noisy and does not allow for direct interpretation of its features. Thus, we first study population Hi-C features of the *Drosophila* chromatin. For example, it has been long believed that chromatin forms two compartments, active and repressive, the latter being associated with the nuclear lamina. We question whether lamina is indeed the driver of such segregation and analyze population Hi-C for *Drosophila* cells that are depleted of the lamina. We demonstrate that lamina binding alone cannot be the driver of spatial segregation of domains. On a local scale, the chromatin of *Drosophila* is constituted of insulated domains, and the active chromatin state and binding of insulators have been proposed as their formation factors. We confirm that both these factors are important for domain positioning by training interpretable machine learning models on published epigenetic and Hi-C data for *Drosophila*. We notice that Hi-C data processing requires an essential step of coverage normalization, which effects on the data remain poorly understood. We fill this gap and demonstrate that contacts' uneven coverage of genomic regions is associated with active chromatin states. Importantly, we observe the same bias when we finally investigate the properties of chromatin in individual *Drosophila* cells by single-cell Hi-C. scHi-C is a recent powerful technique that allows studying chromatin folding without averaging over a large number of cells. By studying the data generated by scHi-C, we observe prominent cell-to-cell variability in the long-range contacts between active genomic regions and relatively high conservation on the local scale of domains. We suggest a significant contribution of stochastic processes to the formation of the *Drosophila* 3D genome and propose several possible models explaining our observations. Finally, we summarise

the computational approaches to study chromatin folding in individual cells based on scHi-C and outline future directions for the development of this field.

Publications

Main author

1. Margarita D Samborskaia*, **Aleksandra Galitsyna***, Ilya Pletenev, Anna Trofimova, Andrey A Mironov, Mikhail S Gelfand, and Ekaterina E Khrameeva. Cumulative contact frequency of a chromatin region is an intrinsic property linked to its function. *PeerJ*, 8:e9566, 2020

   * - equal contribution

   Role: performed the computational experiments, analyzed the data, prepared figures, wrote the manuscript. In particular: significantly contributed to all the manuscript sections, performed initial analysis for Fig. 1f-h, S1, S2, S4-11; reproduced and verified the results in these figures; edited the text; significantly contributed to the discussion with the reviewers.

   PeerJ Impact Factor at the date of publication (2019): 2.35, at the date of defense (2020): 2.98

2. Sergey V Ulianov*, Vlada V Zakharova*, **Aleksandra A Galitsyna***, Pavel I Kos*, Kirill E Polovnikov, Ilya M Flyamer, Elena A Mikhaleva, Ekaterina E Khrameeva, Diego Germini, Mariya D Logacheva, et al. Order and stochasticity in the folding of individual Drosophila genomes. *Nature Communications*, 12(1):1–17, 2021

   * - equal contribution

   Role: performed analysis of single-nucleus Hi-C (snHi-C), bulk BG3 in situ Hi-C, and publicly available data, wrote the manuscript. In particular: wrote the "Methods" section starting from "snHi-C raw data processing and contact annotation" to "Robustness of TAD calling"); prepared Fig. 1d-f, 2a-d, 3, 4, 5a,c,f,g; created Supplementary materials Fig. 1 to 20, excluding Fig. 4e-k and 7c-d; processed all the data; posted the publicly available datasets and code; significantly contributed to the "Results" section and discussion with the reviewers.

Nature Communications Impact Factor at the date of publication and the date of defense (2020): 13.78

3. **Aleksandra A Galitsyna** and Mikhail S Gelfand. Single-cell Hi-C data analysis: safety in numbers. *Briefings in Bioinformatics*, 08 2021. bbab316

   Role: analyzed the data, prepared figures, wrote the manuscript.

   Briefings in Bioinformatics Impact Factor at the date of publication and the date of defense (2020): 11.622

Co-author

1. Sergey V Ulianov, Semen A Doronin, Ekaterina E Khrameeva, Pavel I Kos, Artem V Luzhin, Sergei S Starikov, **Aleksandra A Galitsyna**, Valentina V Nenasheva, Artem A Ilyin, Ilya M Flyamer, et al. Nuclear lamina integrity is required for proper spatial organization of chromatin in Drosophila. *Nature Communications*, 10(1):1–11, 2019

   Role: performed Hi-C data analysis. In particular: prepared Hi-C heatmaps, such as displayed in Fig. 3a, 5a; performed TAD calling for Fig. 3b, c; edited the text.

   Nature Communications Impact Factor at the date of publication (2018): 11.878, at the date of defense (2020): 13.78

2. Michal B Rozenwald, **Aleksandra A Galitsyna**, Grigory V Sapunov, Ekaterina E Khrameeva, and Mikhail S Gelfand. A machine learning framework for the prediction of chromatin folding in Drosophila using epigenetic features. *PeerJ Computer Science*, 6:e307, 2020

   Role: conceived and designed the computational experiments, analyzed the data, prepared figures, wrote the manuscript. In particular: wrote "Introduction", "Discussion" and "Conclusions" sections of the paper; created Fig. 1A-C; significantly contributed to the "Machine learning models", "Chromatin marks are reliable predictors of the TAD state" sections, and other sections of the Results; prepared the dataset and created "Data Availability" section;

prepared Table 1 of Supplementary Information; edited the text; significantly contributed to the discussion with the reviewers.

PeerJ Computer Science Impact Factor at the date of publication (2019): 3.091, at the date of defense (2020): 1.39

Acknowledgments

I am thankful to my supervisor, Prof. Mikhail Gelfand, who introduced me to Moscow's chromatin structure research community. Prof. Mikhail Gelfand organized a rare learning environment for a young specialist in computational biology. Regular chromatin seminars lead by Ekaterina Khrameeva (Skoltech Assistant Professor now) and heated discussions with collaborators (Sergey Ulianov and other group members of Prof. Sergey Razin at Institute of Gene Biology) were an essential part of this. The role of Ekaterina Khrameeva, in fact, co-supervising some of my projects, cannot be underestimated.

I want to thank my partners, Michal Rozenwald, Sergey Ulianov, Pavel Kos, and Vlada Zakharova, who supported me through the most challenging moments of the projects. I appreciate the guidance by Prof. Leonid Mirny and fruitful discussions with his lab during the internship at MIT.

I particularly worship the continued patience of my family and Dmitry T.-O., who supported me throughout the years of studies of chromatin architecture.

Finally, I want to acknowledge the members of my Individual Thesis Committee and constructive comments by all PhD Defense Jury Members. I immensely appreciate thoughtful comments during annual reviews by Prof. Petr Sergiev and the recommendation by Prof. Dmitry Pervouchine to dedicate part of my thesis to the literature review.

Contents

Glossary

boundary outer genomic bins of TADs. 16–18, 45, 104

CCF cumulative contact frequency. 12, 71, 72, 104, 163

chromatin DNA-protein complex of the eukaryotic nucleus. 11, 13, 148

chromatin feature a distinguished instance of chromatin structure having the specific pattern of DNA contacts in Hi-C maps. 19, 45, 148

compartment a pattern of long-range contacts of chromosomes observed as a checkerboard in Hi-C maps. 13, 15, 71

compartmentalization a mechanism or process of chromatin compartments formation. 14, 15

contact an event of capturing two DNA fragments in close spatial proximity, observed as a pair of DNA segments in Hi-C read or read pair. 13, 71, 148

CTCF CCCTC-binding factor. 16, 17

DCC Dosage Compensation Complex. 16

epigenetics heritable information stored in the nucleus or cells but not encoded directly in the DNA sequence. Includes histone modifications, methylation of DNA and other information. 15, 45

genomic bin an instance of sequential segments of the genome of equal size used in genomics for simplification of calculations, typically of several kilobase pairs (kb). 14, 23, 71

Hi-C high-throughput chromosomes conformation capture. 11–14, 16, 19–23, 45, 71, 104, 163

IC iterative correction. 20, 71

insulation an effect observed as reduced number of interactions between two genomic segments relative to expected. Happens at the boundaries of TADs. 16, 45

LAD lamina associating domain. 15, 22, 23

lamin-DamID DNA adenine methyltransferase identification for lamina, experimental technique to detect lamina binding in the nucleus. 15, 23

lamina a proteinaceous interior lining of the nucleus. 12, 22

loop extrusion a hypothesized mechanism of chromatin folding involving DNA, the dynamic molecule of loop extruder and (optionally) barrier elements. 14, 15

NL nuclear lamina. 23

nucleus a separate double membrane-bound organelle of the cell containing its DNA in the form of chromatin. 13

scHi-C single-cell Hi-C. 11, 12, 19–21, 72, 104, 148

SMC structural maintenance of chromosome. 16, 17

snHi-C single-nucleus Hi-C. 4, 104

TAD topologically associating domain. 9, 13, 14, 16–18, 22, 23, 45, 46, 72, 104, 105, 163

Chapter 1

Introduction

The thesis is dedicated to studying chromatin folding in individual cells. Most of our knowledge of chromatin folding in individual cells was accumulated for mammalian models. However, the folding of chromatin in fruitfly, *Drosophila melanogaster*, which is an insect, is less well understood and can potentially shed light on general and unique properties of chromatin structure formation across broader domains of life.

A recent and powerful method to probe the chromatin structure is single-cell Hi-C (scHi-C), an adaptation of popular high-throughput chromosomes conformation capture (Hi-C) [1] for individual cells and nuclei. Both Hi-C and scHi-C are sequencing-based methods that require complex processing of the data, which is affected by multiple experimental artifacts. Thus, we dedicate a substantial part of this research to studying conformation capture data processing.

In this work, we focus on the three-dimensional architecture of individual cells of *Drosophila* assayed by scHi-C and discuss the discoveries in the light of chromatin folding models of *Drosophila* obtained from population Hi-C.

1.1 Thesis Structure

Chapter 2 - Background Literature overview on chromatin structure, methods to study it, and features that can be detected.

[1] Usually, several millions of cells serve as input for traditional Hi-C. We will refer to this experimental biology technique as population Hi-C, bulk Hi-C, or Hi-C.

Chapter 3 - Thesis Objectives Here I define the thesis objectives.

Chapter 4 - Role of nuclear lamina in *Drosophila* chromatin formation Here I analyze bulk Hi-C data and assess the nuclear lamina binding as a potentially important factor for chromatin structure formation in individual cells.

Chapter 5 - Structural factors of *Drosophila* bulk chromatin Here I study the protein binding and histone marks that affect the local structure formation in bulk Hi-C.

Chapter 6 - Technical factors affecting Hi-C data Here I investigate the properties of cumulative contact frequency (CCF) on bulk data, which is unavoidable in scHi-C.

Chapter 7 - Chromatin folding in individual cells of *Drosophila* Here I investigate the properties of chromatin in individual cells of *Drosophila* based on scHi-C.

Chapter 8 - Literature review of scHi-C and discussion Here I summarize the computational methods for scHi-C and highlight the future of this research field.

Chapter 9 - Conclusion In the last chapter, I discuss the obtained results.

Chapter 2

Background

DNA of eukaryotic cell bound by an armory of proteins, a complex also called chromatin, folds in a confined space of the nucleus. The three-dimensional architecture of chromatin is not merely a passive consequence of structure formation mechanisms (Fig. 5-1). There is growing evidence that it plays a crucial role in gene regulation and disease [Anania and Lupiáñez, 2020].

Advances in the rapid understanding of chromatin folding principles can be attributed to the development of high-throughput chromosomes conformation capture (Hi-C) and its derivatives, a method to probe genome-wide pairwise DNA contacts in a population of cells [Lieberman-Aiden et al., 2009] (Fig. 2-2). In particular, we know that locally chromatin folds into topologically associating domains (TADs). In contrast, the long-range contacts of chromatin are organized into compartments [Goel and Hansen, 2020, Lieberman-Aiden et al., 2009].

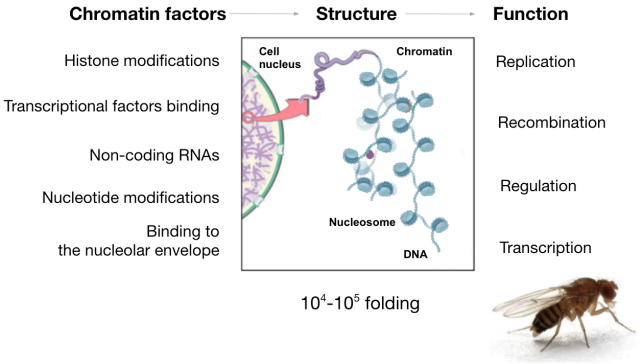


Figure 2-1: Principles of chromatin folding in *Drosophila*. DNA is confined in a limited space of the nucleus, achieving $10^5 - 10^6$ folding. The structure is affected by numerous processes happening in the nucleus. These changes are propagated to the vital biological processes, for which DNA and its architecture are responsible.

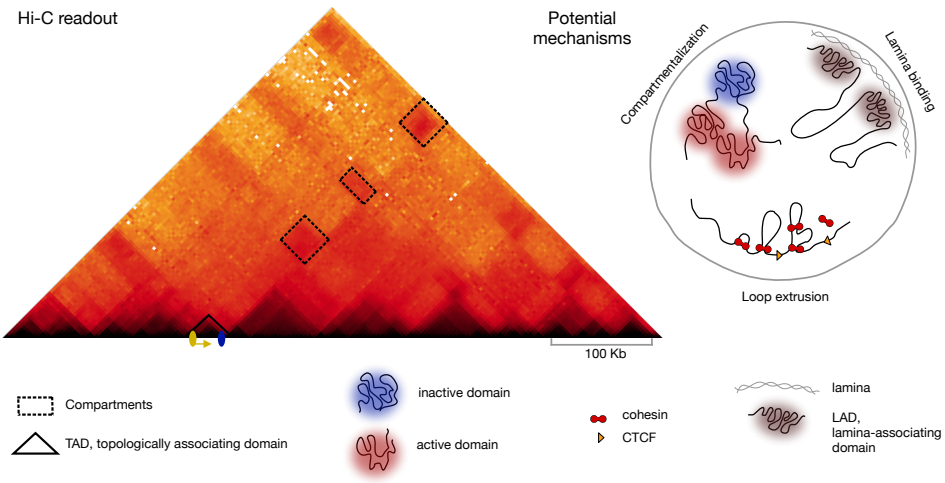


Figure 2-2: Overview of chromatin features detectable by population Hi-C. Left: Hi-C readout for *Drosophila* with marked features. Yellow denotes hypothetical promoter of the gene which is activated by enhancer (blue) in the same TAD. Right: Possible interpretation of the Hi-C features and potential mechanisms explaining their emergence according to the literature. Hi-C datasets collected from [Wang et al., 2018], processed with distiller [by Open Chromosome Collective] at the genomic bin size of 3 Kb and stored on HiGlass web server [Kerpedjiev et al., 2018] by the following link: http://higlass.skoltech.ru/app/?config=KqBRGptVTsm5glGq0djqxQ valid by the date of thesis submission.

Two acknowledged fundamental mechanisms of chromatin structure formation are loop extrusion and compartmentalization [Nuebler et al., 2018] (Fig. 2-3). The first one compacts chromatin on a local scale. In contrast, the second one acts on a large scale. Minor and more targeted mechanisms contribute, such as the formation of Polycomb loops [Du et al., 2020, Eagen et al., 2017]. Extrusion [Banigan et al., 2020] and compartmentalization compact chromatin in the nuclei of a broad range of species, and cross-species comparisons shed light on universal principles of action of these mechanisms. Striking conservation of mechanisms was suggested because TADs are present across a wide range of species, including mammals, insects, and nematodes [Dekker and Heard, 2015]. Moreover, TAD positioning is surprisingly conserved in mammalian evolution, further suggesting the functional importance of structural folding [Rudan et al., 2015]. At the same time, TADs are highly dynamic structures that are lost as a bulk signature during mitosis in each cell division [Naumova et al., 2013, Abramo et al., 2019].

Drosophila melanogaster is one of the popular model species, which chromatin

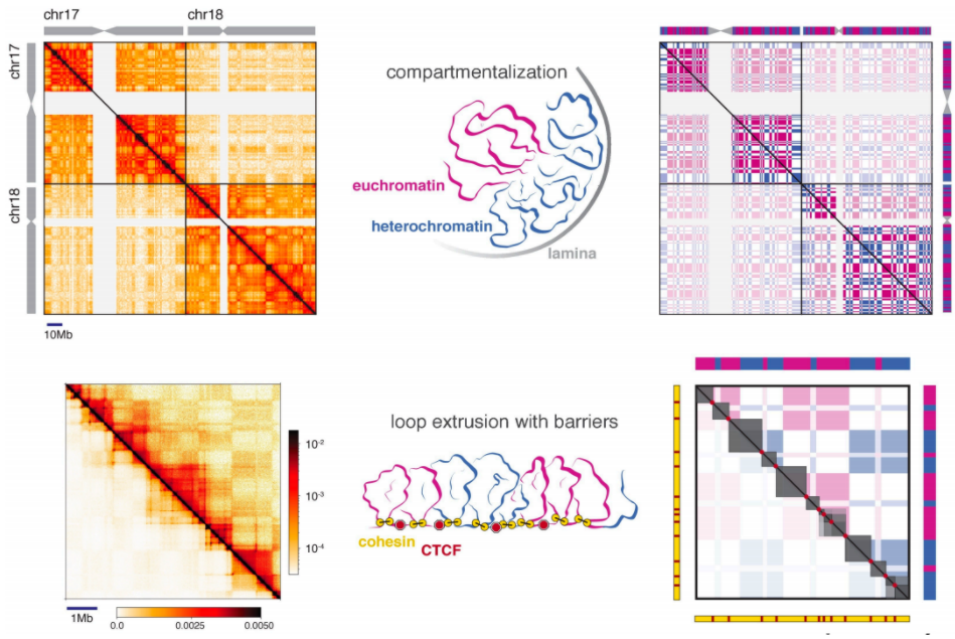


Figure 2-3: Two mechanisms of chromatin structure formation, loop extrusion and compartmentalization, with schematic representation of underlying processes and resulting Hi-C features. Reproduced from [Mirny et al., 2019], with the permission.

structure formation principles are poorly understood. The availability of various datasets on this species' epigenetics presents a unique opportunity to deepen our knowledge of chromatin folding principles [Moretti et al., 2020]. Hi-C based on the population of cells has revealed important averaged patterns of DNA folding in *Drosophila*.

The chromatin of this insect falls into two compartments, active and repressed [Rowley et al., 2017]. The knowledge of drivers of this segregation is still incomplete despite extensive studies in other species, e.g., mammals [Erdel et al., 2020]. In particular, inactive chromatin is associated with lamina, a proteinaceous interior lining of the nucleus [Gruenbaum and Foisner, 2015], typically assayed by DNA adenine methyltransferase identification for lamina, experimental technique to detect lamina binding in the nucleus (lamin-DamID). Extended chromatin regions called lamina associating domains (LADs) are involved in this interaction. Lamina binding is a potential driver for compartments formation. However, while it remains a crucial factor of chromatin formation in *Drosophila*, this hypothesis remains speculative and requires further investigation.

On a local scale, the chromatin of *Drosophila* is constituted of insulated domains, or TADs [Sexton et al., 2012]. It was proposed earlier that active chromatin drives the insulation in *Drosophila* [Ulianov et al., 2016], while in mammals, the dominant mechanism is loop extrusion [Fudenberg et al., 2016] (Fig. 2-4).

In the loop extrusion mechanism, structural maintenance of chromosome (SMC) complexes (typically, cohesin) act as extruding factors, and DNA-binding protein CCCTC-binding factor (CTCF) serves as a barrier element [Rao et al., 2014]. This barrier element stops the translocation of structural maintenance of chromosome (SMC) complex [Li et al., 2020, Fudenberg et al., 2017] leading to the formation of TADs.

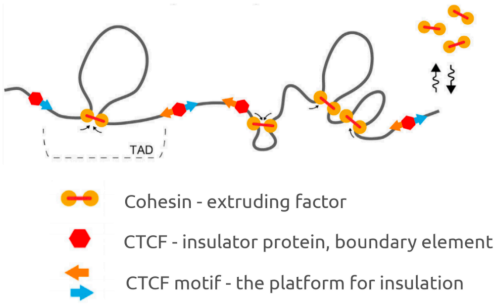


Figure 2-4: Loop extrusion mechanism and its major players in mammals. Reproduced from [Fudenberg et al., 2016], with the permission.

The loop extrusion theory was confirmed by the studies of cohesin and CTCF depletion [Nora et al., 2017, Rao et al., 2017], and multiple molecular mechanisms were proposed for SMC action [Yatskevich et al., 2019]. However, loop extrusion for *Drosophila* is debated [Matthews and White, 2019].

One of early evidence of loop extrusion in mammals was the presence of convergent CTCF binding at the boundaries of TADs [Rao et al., 2014]. Later it was confirmed that positioning of CTCF contributes significantly to the contact probability prediction of the population Hi-C [Rowley et al., 2017, Fudenberg et al., 2020, Belokopytova et al., 2020]. While SMC proteins are notably conserved across the domains of life [Cobbe and Heck, 2004], CTCF has appeared in evolution only recently [Heger et al., 2012]. If loop extrusion is a universal mechanism, this raises the question: what are the barrier elements in species with no CTCF? Various factors have been proposed, including moving polymerase in bacteria [Brandão et al., 2019], Dosage Compensation Complex (DCC) in *C. elegans* [Crane et al., 2015] and cohesin itself in yeast [Costantino et al., 2020].

Figure 2-5: Fragment of alignment of CTCF orthologs in a broad range of mammals and in *Drosophila*. Red square denotes the conserved sequence required for binding with SMC complex. Sequences of CTCF orthologs were obtained from UniProt [Consortium, 2015], and alignment was performed in JalView [Waterhouse et al., 2009] with Muscle algorithm [Edgar, 2004]. Conserved sequence is highlighted in red, as reported in [Li et al., 2020]. Grey demarcates the border between the mammalian and *Drosophila* sequences.

CTCF of *Drosophila* is encoded in its genome, actively expresses and binds DNA in a sequence-specific manner [Heger et al., 2012]. The motif itself is very similar to that in mammals [Holohan et al., 2007]. However, CTCF binding is not enriched at TAD boundaries in most of *Drosophila* cell lines [Ulianov et al., 2016], except neuronal cells [Chathoth and Zabet, 2019]. Instead, there are at least eight other motifs enriched at the TAD boundaries [Ramírez et al., 2018]. Some studies report 12 architectural proteins that contribute to the structure formation [Rowley et al., 2017]. Some propose that pairs of proteins BEAF-32/CP190 and BEAF-32/Chromator can act instead of the usual CTCF/SMC pair [Wang et al., 2018]. A promising approach to unravel TAD-forming mechanisms is to adapt machine learning models that predict the presence of TADs based on the epigenetic markers (such as regression models in [Ulianov et al., 2016]).

Notably, CTCF-SMC complex functioning in mammals requires formation of conserved binding surface [Li et al., 2020] organized by three proteins: CTCF, Scc1 and SA2. Disruption of a small 9-a.a. region in CTCF aminoacid sequence results in disruption of complex formation and loss of CTCF barrier function. In *Drosophila*, the CTCF ortholog also has the putative conserved binding motif (Fig. 2-5).

However, there are multiple substitutions in CTCF, SA2, and Scc1 orthologs that might affect the conserved binding surface (Fig. 2-6). For example, there is a notable substitution of mammalian D326 in SA2 with proline. D326 is an aspartic acid responsible for hydrogen bond formation between proteins. Proline (replacing D326

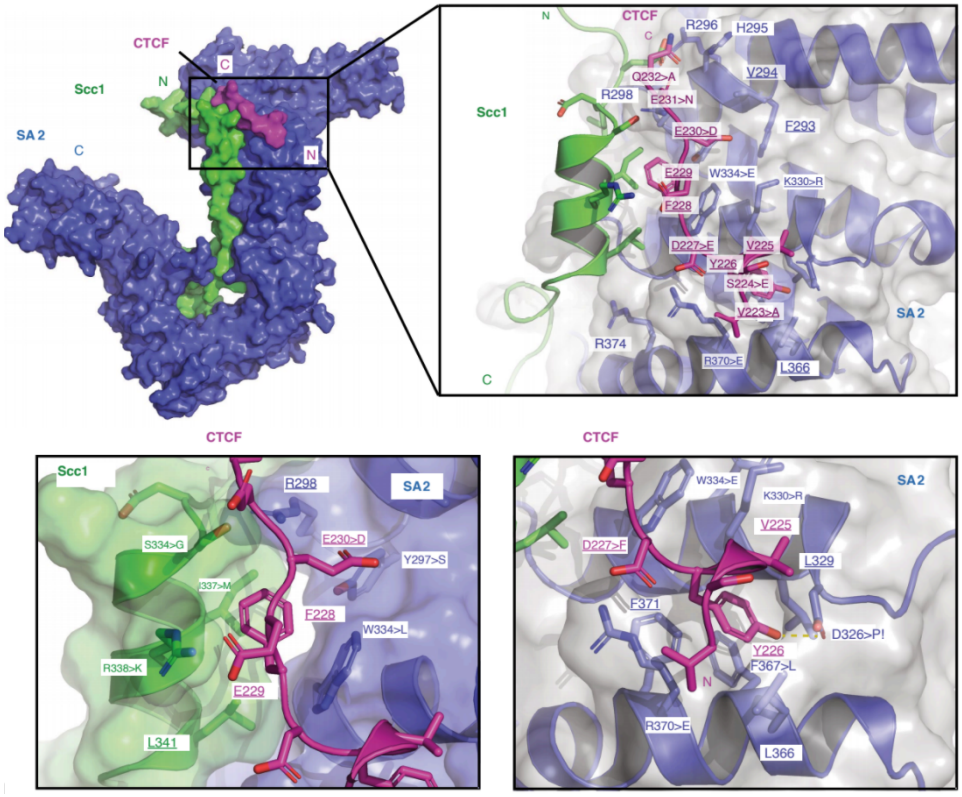


Figure 2-6: Structure of conserved binding surface between CTCF fragment, Scc1 and SA2 (two proteins of SMC complex in mammals), with demonstration of aminoacids substitutions in *Drosophila*. Each amino acid is marked by its position and type in mouse, and the substitutions are marked if present in *Drosophila*. The underlined amino acids are the ones that are either conserved between species or substituted with amino acids with similar properties. Structural alignment is based on [Li et al., 2020], and substitutions are manually marked as described in Fig. 2-5.

in *Drosophila*) is a very rigid aminoacid. Moreover, N-terminal group of proline never forms a hydrogen bond, thus, making proline unfavorable in most regions of alpha-helices and beta-sheets. This suggests that the binding surface between CTCF, SA2, and Scc1 might be disrupted in *Drosophila*. This hypothetical impairment of binding might explain the presence of multiple other barrier elements that are frequently found at TAD boundaries in *Drosophila*. Thus, the studies of chromatin formation mechanisms in insects will shed light on evolutionary mechanisms that explain the formation of chromatin architecture.

The most comprehensive and informative methods to study chromatin formation mechanisms are genome-wide techniques of two principal types: microscopy

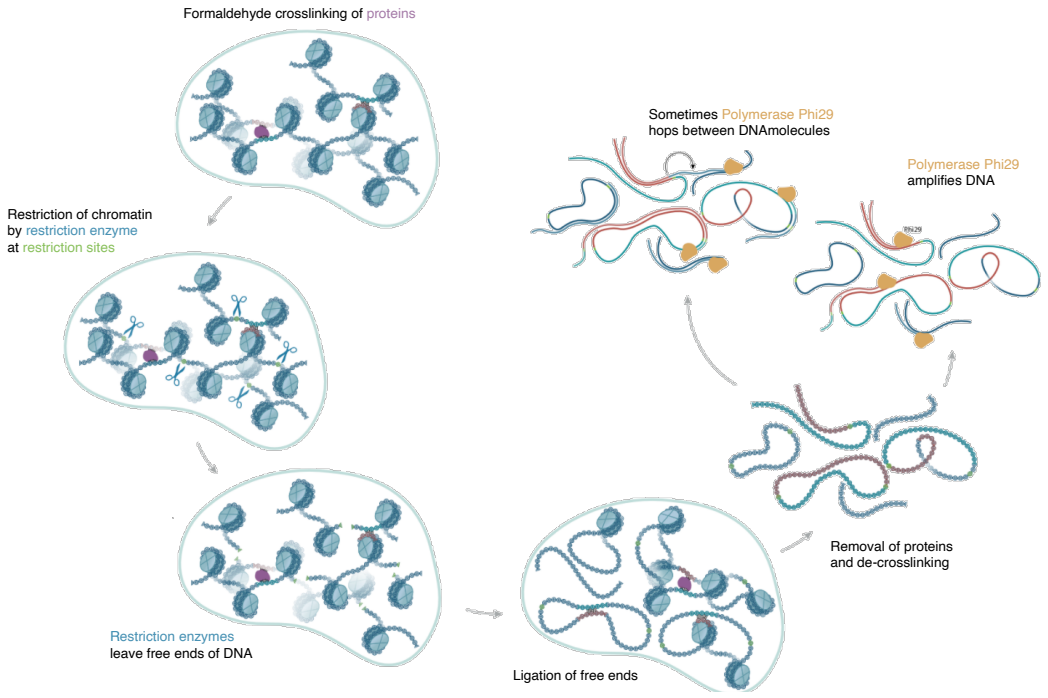


Figure 2-7: ScHi-C technique based on [Ulianov* et al., 2021]. Image by M. Guriev, reproduced with permission.

methods [Szabo et al., 2018] and conformation capture based on the population of cells [Goel and Hansen, 2020]. The first approach has limited resolution and the number of regions and cells that can be studied. The second approach results in the population-average readout, which does not account for the cell-to-cell variability of chromatin features [Fudenberg et al., 2016] and does not allow the observation of the properties of chromatin structure in individual cells. Single-cell Hi-C (scHi-C), Hi-C adapted for individual cells, overcomes this limitation [Nagano et al., 2013] (Fig. 2-7). The power of this method is that the contacts observed together happen in the same 3D conformation of chromatin, which allows the creation and direct interpretation of polymer models describing each cell [Nagano et al., 2013] and the study of mechanisms of structure formation in a more straightforward way [Flyamer et al., 2017, Gassler et al., 2017].

Currently, multiple protocols for scHi-C were proposed [Nagano et al., 2013, 2017, Ramani et al., 2017, Flyamer et al., 2017, Tan et al., 2018], and a limited number of works is dedicated to summarise and compare them [Ulianov et al., 2017,

Lando et al., 2018, Zhou et al., 2021]. Moreover, the scHi-C data properties are not entirely understood, although many are similar to population Hi-C [Nagano et al., 2013, Flyamer et al., 2017]. Importantly, there is little guidance for scHi-C data processing, and most of the software is used *ad hoc*.

Previous single-cell studies of *Drosophila* chromatin were based on microscopy only. The presence of domain structure was demonstrated on several genomic loci [Szabo et al., 2018]. The application of scHi-C will allow to study these properties at the whole-genome level and may shed light on the mechanisms governing chromatin compaction in single cells.

Notably, readouts of Hi-C method are prone to methodological artifacts, such as dependence of the number of contacts on the technical characteristics of the underlying genomic regions [Imakaev et al., 2012]. To normalize out these artifacts, the Hi-C data is iteratively corrected (or ICed, from iterative correction (IC)). These artifacts are most likely to be present in scHi-C as well, which requires investigations on their nature in both bulk and single-cell Hi-C.

Chapter 3

Thesis Objectives

The goals of the research:

- Assess the factors that can affect Hi-C-based readouts
- Create the list of factors affecting *Drosophila* chromatin structure formation based on bulk Hi-C
- Discover the role of lamina binding in *Drosophila* chromatin structure formation
- Design a computational approach to study chromatin in single cells based on scHi-C
- Formulate the principles of chromatin structure formation in *Drosophila*

Chapter 4

Nuclear lamina integrity is required for proper spatial organization of chromatin in *Drosophila*

It is well-established that lamina-binding regions of *Drosophila* form local chromatin domains [Kharchenko et al., 2011, Filion et al., 2010], suggesting that TADs are closely related to LADs. Simple superposition of TADs and LADs demonstrates profound correspondence between these regions across the genome (Fig. 4-1). However, whether this effect is correlative or causative remains poorly understood. One explanation might be that lamina is a driving force of domains or compartments formation in *Drosophila*. An alternative is that substantial overlap between TADs and LADs is a consequence of repositioning of the regions inside the nucleus due to other processes, not related to lamina binding *per se*.

If lamina binding is a crucial mechanism of chromatin formation of *Drosophila*, then it might become the primary target for further studies of folding patterns in individual cells. If it is not, the degree of its contribution to the structure formation should be determined.

The project to answer these questions had started in collaboration with the Institute of Gene Biology (and researchers from other institutions) long before I joined the lab at Skoltech. Our collaborators in Prof. Razin's group created the cells with depleted lamina. They performed the wet-lab part of Hi-C for them.

Processing of Hi-C data was less automatized and understood at these times.

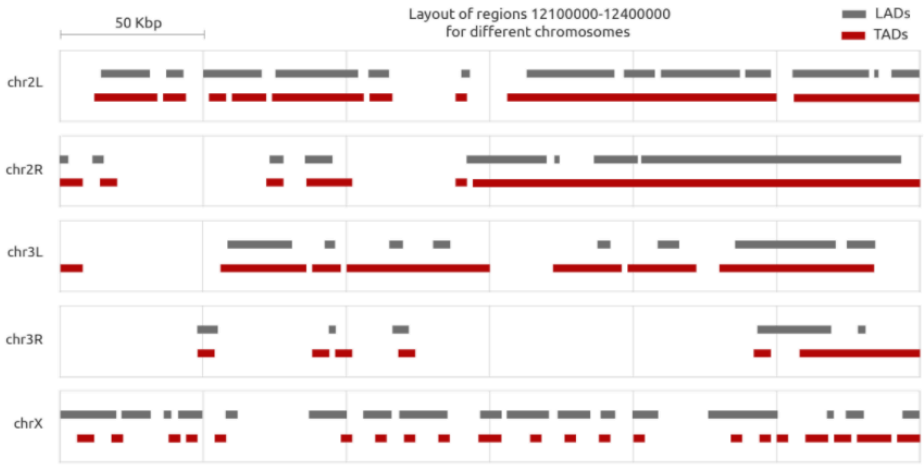


Figure 4-1: Genome-wide annotations for chromosomal fragments of TADs and LADs in *Drosophila* cell line, based on Hi-C from [Ulianov et al., 2016]. We can observe a substantial overlap of annotations. White spaces between segments represent either inter-TADs/inter-LADs, or the boundary bins between two neighboring segments. Inter-TADs are genomic regions between TADs, since the TAD segmentation is not complete. For example, if the TAD found by automatic TAD caller is very small (1-2 genomic bins), it is considered as interTAD. Inter-LADs are genomic regions that were not assigned to LADs by lamin-DamID.

Thus, I implemented data processing (read mapping, quality control, TAD calling) and explored potential experimental artifacts (self-circles, dangling ends, backward ligation, mirror reads). It was essential to develop an in-lab software for this part, and it is further used in other studies (Chapters 6, 7). This study was a primer for my research on bulk and single-cell Hi-C data of *Drosophila*.

Taken together, it was a long path towards understanding the role of nuclear lamina (NL). In the first Hi-C experiments on NL disruption, TADs were less pronounced, and the distance decay was less steep than in wild-type. It turned out that the cells were dying during the treatment. During the first several months of this project, the effect that we observed was chromatin decompaction in the dead cells. Luckily, soon the experiments were repeated on live cells (confirmed by microscopy), and we studied the true effect of chromatin detachment from lamina in *Drosophila* cells. The final version of the manuscript was prepared mostly by the co-authors, although I contributed as well (see "Publications" at p. 4 for details).

The conclusion is that disruption of NL does not force the global reorganization of TADs and compartments in *Drosophila*. TADs do not change their positions

radically, although they become less compact. This observation is in line with the hypothesis that some inactive chromatin domains are attached to the lamina. However, the attachment is not crucial for their formation. This study was a final stop for the research of the lamina binding in *Drosophila* cells. I use this important knowledge in Chapters 5 and 7.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-09185-y OPEN

Nuclear lamina integrity is required for proper spatial organization of chromatin in *Drosophila*

Sergey V. Ulianov[1,2], Semen A. Doronin[3], Ekaterina E. Khrameeva[4,5], Pavel I. Kos[6], Artem V. Luzhin[1], Sergei S. Starikov[7], Aleksandra A. Galitsyna[1,4,5,7], Valentina V. Nenasheva[3], Artem A. Ilyin[3], Ilya M. Flyamer[8], Elena A. Mikhaleva[3], Mariya D. Logacheva[4,9,10], Mikhail S. Gelfand[4,5,11], Alexander V. Chertovich[6], Alexey A. Gavrilov[1], Sergey V. Razin[1,2] & Yuri Y. Shevelyov[3]

How the nuclear lamina (NL) impacts on global chromatin architecture is poorly understood. Here, we show that NL disruption in *Drosophila* S2 cells leads to chromatin compaction and repositioning from the nuclear envelope. This increases the chromatin density in a fraction of topologically-associating domains (TADs) enriched in active chromatin and enhances interactions between active and inactive chromatin. Importantly, upon NL disruption the NL-associated TADs become more acetylated at histone H3 and less compact, while background transcription is derepressed. Two-colour FISH confirms that a TAD becomes less compact following its release from the NL. Finally, polymer simulations show that chromatin binding to the NL can per se compact attached TADs. Collectively, our findings demonstrate a dual function of the NL in shaping the 3D genome. Attachment of TADs to the NL makes them more condensed but decreases the overall chromatin density in the nucleus by stretching interphase chromosomes.

[1] Institute of Gene Biology, Russian Academy of Sciences, Moscow 119334, Russia. [2] Faculty of Biology, M.V. Lomonosov Moscow State University, Moscow 119991, Russia. [3] Institute of Molecular Genetics, Russian Academy of Sciences, Moscow 123182, Russia. [4] Skolkovo Institute of Science and Technology, Skolkovo 143026, Russia. [5] Institute for Information Transmission Problems (the Kharkevich Institute), Russian Academy of Sciences, Moscow 127051, Russia. [6] Faculty of Physics, M.V. Lomonosov Moscow State University, Moscow 119991, Russia. [7] Faculty of Bioengineering and Bioinformatics, M.V. Lomonosov Moscow State University, Moscow 119991, Russia. [8] MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK. [9] Belozersky Institute of Physico-Chemical Biology, M.V. Lomonosov Moscow State University, Moscow 119234, Russia. [10] Russia Extreme Biology Laboratory, Institute of Fundamental Medicine and Biology, Kazan Federal University, Kazan 420012, Russia. [11] Faculty of Computer Science, National Research University Higher School of Economics, Moscow 125319, Russia. These authors contributed equally: Sergey V. Ulianov, Semen A. Doronin, Ekaterina E. Khrameeva, Pavel I. Kos. Correspondence and requests for materials should be addressed to S.V.U. (email: sergey.v.ulyanov@gmail.com) or to Y.Y.S. (email: shevelev@img.ras.ru)

The nuclear lamina (NL)[1] is a meshwork of lamins and lamin-associated proteins lining the nuclear envelope (NE). Several lines of evidence support the idea that the NL is a platform for the assembly of the repressive compartment in the nucleus. In mammals, nematode and *Drosophila*, the lamina-associated chromatin domains (LADs)[2–5] contain mostly silent or weakly expressed genes[2–6]. Activation of tissue-specific gene transcription during cell differentiation is frequently associated with translocation of loci from the NL to the nuclear interior[4,7–11]. The expression level of a reporter gene is ~5-fold lower when it is inserted into LADs compared to inter-LADs[12]. Artificial tethering of weakly expressed reporter genes to the NL results in their downregulation thus indicating that contact with the NL may cause their repression[13–15]. Accordingly, many transcriptional repressors, including histone deacetylases (HDACs) are linked to the NL[16]. The high throughput chromosome conformation capture (Hi-C) technique has revealed the spatial segregation of open (DNase I-sensitive) and closed (DNase I-resistant) chromatin into two well-defined compartments[17]. Importantly, in mammalian cells, the DNase I-resistant compartment is strongly enriched with NL contacts[18,19]. Moreover, a whole-genome DNase I-sensitivity assay in *Drosophila* S2 cells indicated that LADs constitute the densely packed chromatin[20]. Additionally, super-resolution microscopy studies in Kc167 cells show that inactive chromatin domains (including Polycomb (Pc)-enriched regions) are more compact than active ones[21].

The newly developed single-cell techniques demonstrate that LADs, operationally determined in a cell population, may be located either at the NL or in the nuclear interior in individual cells[19,22]. Surprisingly, the positioning of LADs in the nuclear interior barely affects the inactive state of their chromatin[22]. This raises the question as to whether contact with the NL makes the chromatin in LADs compact and inactive. However, few studies directly address this issue. It has been shown that lamin *Dm0* knock-down (Lam-KD) in *Drosophila* S2 cells decreases the compactness of a particular inactive chromatin domain[23]. Accordingly, the accessibility of heterochromatic and promoter regions has been shown to increase upon Lam-KD in *Drosophila* S2R$^+$ cells[24]. However, the impact of the NL on the maintenance of the overall chromatin architecture remains mostly unexplored.

Here we show that upon loss of all lamins, the density of peripheral chromatin is decreased in *Drosophila* S2 cells leading to the slight overall chromatin compaction. At the same time, chromatin in LADs becomes less tightly packed which correlates with the enhancement of initially weak level of histone H3 acetylation and background transcription in these regions.

Results

Lam-KD in S2 cells results in general chromatin compaction. We have studied the effects of NL disruption on global chromatin architecture, histone acetylation and gene expression in *Drosophila*. To select an appropriate experimental model, we first analysed the presence of ubiquitous lamin Dm0 and tissue-specific lamin C proteins[25] in several *Drosophila* cell lines by Western-blotting. Whereas the level of lamin Dm0 is similar in S2, Kc167, and OSC lines, lamin C is robustly present in Kc167 and OSC, but almost completely absent in S2 cells (Fig. 1a). Hence, to remove all lamins, we performed Lam-KD in S2 cells by RNAi (Fig. 1b) and stained the nuclei with anti-histone H4 antibody to visualise the bulk chromatin, and with anti-lamin-B-receptor (LBR[26]) antibody to visualise the NE (Fig. 1c and Supplementary Fig. 1a). Quantification of the fluorescence intensity along the nuclear diameter reveals a slight but statistically significant shift in the radial distribution of total chromatin from the NE towards the nuclear interior upon Lam-KD (Fig. 1d and Supplementary Fig. 1a). To validate this observation, we performed fluorescence in situ hybridization (FISH) with a probe from the cytological region *36C*, which was previously mapped as a LAD in the Kc167 cells[5] (Fig. 1e). The radial position of this region is shifted towards the nuclear interior in Lam-KD S2 cells when compared to control cells (hereinafter treated with dsRNA against bacterial *lacZ* gene) (Fig. 1e). Notably, this observation agrees with previously published results[11] which we reanalysed to demonstrate a shift in the radial position of two other loci (*22A* and *60D*) from the NE upon Lam-KD (Fig. 1f). Moreover, we observed an *en masse* chromatin compaction as a result of NL disruption, since the average volume of total chromatin, reconstructed by DAPI staining, is markedly diminished upon Lam-KD (Fig. 1g and Supplementary Fig. 1b). Remarkably, the average volume of nuclei, reconstructed by LBR-stained NE, was not affected by Lam-KD (Supplementary Fig. 1c). Taken together, these observations indicate that disruption of the NL results in general chromatin compaction and repositioning from the NE.

In mammalian cells, the presence of either lamin A/C or LBR is necessary for proper positioning of the heterochromatin at the nuclear periphery[27]. In contrast, in *Drosophila* S2 cells, where lamin C is not expressed (Fig. 1a), depletion of LBR does not notably affect chromatin positioning relative to the NE (Fig. 1h and Supplementary Fig. 1d). We confirmed this observation by FISH with the probe from the *36C* region examined upon Lam-KD. We found that this region is not repositioned relative to the NE upon LBR depletion (Fig. 1i). These results indicate that the main heterochromatin tethers are different in mammals and *Drosophila* with the lamin Dm0 providing the major impact on LAD attachment, at least in S2 cells.

Lam-KD in S2 cells enhances weak transcription in LADs. To examine which genes are associated with the NL, we have used previously published lamin–DamID data for Kc167 cells[5,28] that are closely related to S2 cells, are of a similar embryonic origin, and have highly correlated transcriptome profiles (Pearson's $R = 0.89$)[29]. As we were not confident that LADs on the X chromosome occupy the same positions in the female Kc167 and in the male S2 cells[30], we excluded X-chromosome from the downstream analysis. We hypothesised that upon Lam-KD, the detachment of LADs from the NE might result in the elevated expression of genes located therein. To test this hypothesis, we performed transcriptome profiling in control and Lam-KD S2 cells using RNA-seq (Supplementary Fig. 2a) and revealed 60 differentially expressed genes (40 up- and 20 downregulated genes) (Fig. 2a). However, the observed increase in gene expression (Supplementary Fig. 2b) does not correlate with the presence of promoters of differentially expressed genes specifically in LADs ($P = 0.21$, permutation test), thus suggesting that either an indirect effect of NL disruption or alterations in chromatin interactions in the nuclear interior are affecting transcription. We then analysed changes in total transcription inside and outside of LADs (i.e. in the inter-LADs). Depletion of lamin Dm0 results in the moderate upregulation of the generally very weak background transcription in LADs (Supplementary Fig. 2c), but not in the inter-LADs (Fig. 2b, c).

To confirm RNA-seq results, we applied RT-qPCR to analyse the transcription level of 14 randomly selected genes whose promoters are located in different LADs (Supplementary Table 1). Almost all of these genes are expressed in S2 cells at a very low level. 12 out of 14 genes appeared to be upregulated upon Lam-KD (~2 fold on average) when compared to control S2 cells (Fig. 2d, top panel). It has previously been shown that Lam-KD in *Drosophila* S2 cells results in increased DNase I sensitivity and the derepression of several testis-specific genes in the silent chromatin domain from the *60D* chromosomal region[11]. We

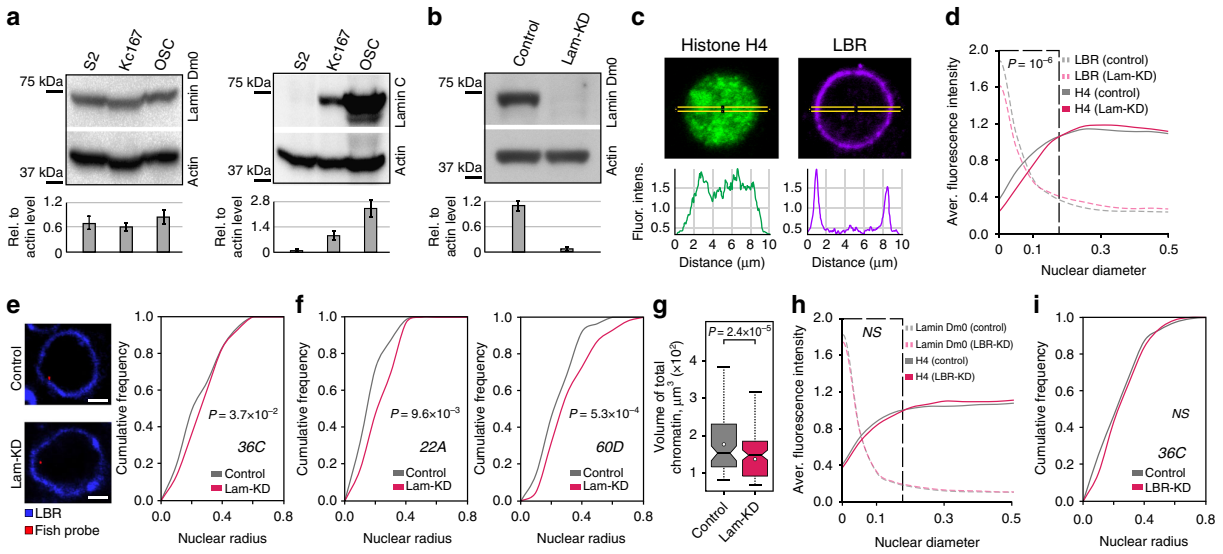


Fig. 1 Chromatin is released from the NE and becomes denser upon Lam-KD in S2 cells. **a**, **b** Western-blot analysis of lamin Dm0 and lamin C protein levels in *Drosophila* cell lines (**a**), or in Lam-KD and control S2 cells (**b**). Band intensity quantitation is presented below. **c** A representative example of nuclei immunostained with antibodies against histone H4 and LBR. Fluorescence intensity along the yellow-framed zone was measured using ImageJ software. **d** Averaged fluorescence intensity profiles along the nuclear diameter in Lam-KD ($n = 120$) and control ($n = 180$) S2 cells immunostained with antibodies against histone H4 and LBR. *P* value for the framed region of histone H4 profiles was estimated in a Wilcoxon test. **e** Cumulative frequency of radial positions of the FISH probe located in the LAD from cytological region *36C* in Lam-KD ($n = 175$) or control ($n = 175$) S2 cells (right panel). *P* value was estimated in a Kolmogorov-Smirnov one-sided test. Representative examples of FISH signals in nuclei stained with anti-LBR antibody are shown on the left panels. Scale bar 1 μm. **f** Cumulative frequency of radial positions of FISH probes to cytological regions *22A* (left) or *60D* (right) in Lam-KD ($n = 100$ for *22A* and $n = 99$ for *60D*) or control ($n = 115$ for *22A* and $n = 72$ for *60D*) S2 cells. *P* value was estimated in a Kolmogorov-Smirnov one-sided test. Data for analysis were taken from ref. [11]. **g** Total chromatin volume measured by DAPI fluorescence in Lam-KD ($n = 125$) or control ($n = 83$) S2 cells. *P* value was estimated in a Wilcoxon test. Thick black lines and white dots represent median and average values, upper and lower ends of boxplot show the upper and lower quartiles, the whiskers indicate the maximum and minimum values. **h** Averaged fluorescence intensity profiles along the nuclear diameter in LBR-depleted (LBR-KD, $n = 120$) or control ($n = 120$) S2 cells immunostained with antibodies against histone H4 and lamin Dm0. *NS* – non-significant difference ($P > 0.05$, Wilcoxon test). **i** Cumulative frequency of radial positions of the FISH probe to cytological region *36C* in LBR-KD ($n = 150$) or control ($n = 150$) S2 cells. *NS* – non-significant difference ($P > 0.05$, Kolmogorov-Smirnov one-sided test)

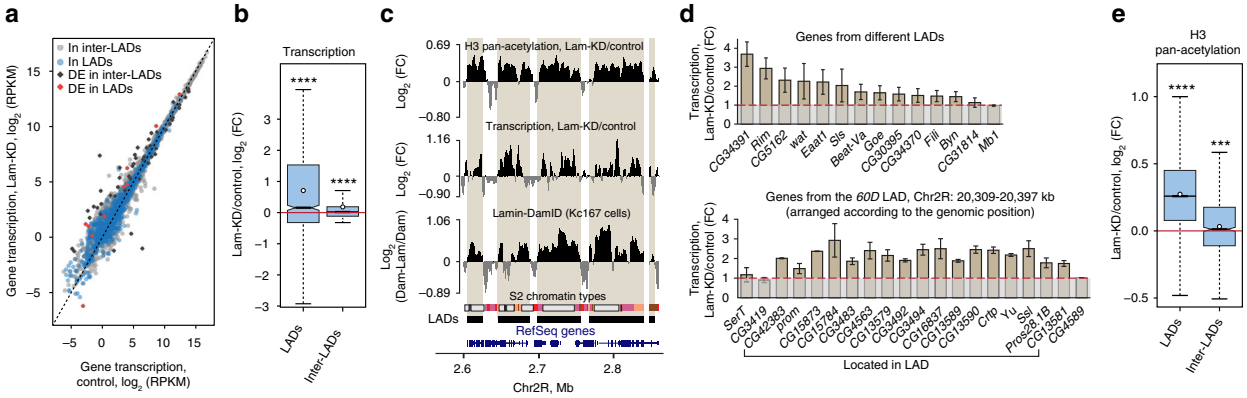


Fig. 2 Nuclear lamina disruption leads to the increased H3 pan-acetylation and transcriptional upregulation in LADs. **a** Gene expression in Lam-KD and control S2 cells according to the RNA-seq data. Genes located in LADs and inter-LADs are designated by light-blue and light-grey circles, respectively. Differentially expressed (DE) genes, located in LADs and inter-LADs, are designated by black and red rectangles, respectively. **b** Changes of total transcription according to the RNA-seq data between Lam-KD and control S2 cells in LADs and in inter-LADs. **c** A representative screenshot from UCSC Genome Browser showing the Lamin-DamID profile in Kc167 cells[5] and $\log_2$(FC) profiles of H3 pan-acetylation (according to ChIP-seq) and transcription (according to RNA-seq) in Lam-KD/control S2 cells. Chromatin annotation in S2 cells[40], LAD annotation in Kc167 cells[28], and RefSeq gene positions are shown below. **d** Changes in the transcription level in Lam-KD compared to control S2 cells by RT-PCR analysis for the randomly chosen genes from different LADs (top panel), and for all the genes from the *60D* LAD (bottom panel). In the bottom panel, genes *SerT, CG3419* and *CG4589* are located outside the LAD. Expression data of *Crtp, Yu, Ssl, Pros28.1B* and *CG13581* genes upon Lam-KD in S2 cells are from ref. [11]. Error bars show SEM between two independent biological replicates. **e** Changes of H3 pan-acetylation according to ChIP-seq data between Lam-KD and control S2 cells in LADs and in inter-LADs. In panels (**b**, **e**), $^{****}P < 0.0001$, $^{***}P < 0.001$ in a Wilcoxon test. See Fig. 1g legend for description of boxplot elements

found that all the genes located in this LAD are almost uniformly upregulated upon Lam-KD in S2 cells (Fig. 2d, bottom panel). Thus, NL disruption results in the partial derepression of chromatin in LADs leading to the increased background transcription.

It has been shown that pan-acetylation of histones H3 and H4 coupled with general DNase I-sensitivity was elevated in the *60D* LAD upon Lam-KD in S2 cells[23]. To check whether the repression of transcription in LADs may be caused by histone deacetylation, we determined histone H3 pan-acetylation level across the entire genome by chromatin immunoprecipitation (ChIP-seq) (Supplementary Fig. 2d). We found that the general level of histone H3 acetylation is markedly elevated in LADs, but not in the inter-LADs upon Lam-KD when compared to control cells (Fig. 2c, e). Thus, we suggest that a fraction of HDACs associated with the NL[31,32] may be at least partially responsible for the low level of histone H3 acetylation and for the transcriptional repression in LADs making their chromatin less accessible for spurious binding by *trans*-acting factors.

Lam-KD in S2 cells leads to decompaction of inactive TADs. To explore genome-wide effects of NL disruption on the spatial organization of chromatin, we applied the Hi-C technique[17] to control and Lam-KD S2 cells and identified topologically-associating domains (TADs)[33–35] (Fig. 3a) using previously described two-step procedure[36]. The strong similarity between Hi-C map data obtained in this work with that previously published for S2 cells[37] (Supplementary Fig. 3a), as well as the high correlation between Hi-C replicates (Supplementary Fig. 3b) demonstrates the high quality and reliability of the data. Furthermore, in agreement with the conservation of TAD boundaries in unrelated *Drosophila* cell types[36,38] and upon different biological conditions[39], pairwise comparison of TAD positions between Lam-KD and control cells does not show statistically significant alterations (Supplementary Fig. 3c). We conclude that NL disruption in S2 cells does not affect the overall TAD profile genome-wide. This allows us to compare the average contact frequency (ACF, see Methods) within each TAD between control and Lam-KD S2 cells. We argue that differences in ACF should reflect changes in the physical density of a TAD.

Figure 3b shows a clear negative trend between LAD coverage within a TAD and intra-TAD ACF changes upon Lam-KD relative to control cells. This trend is absent when LAD coverage is plotted against ACF variability between control replicates (Supplementary Fig. 3d). Remarkably, the opposite trend is revealed between intra-TAD ACF changes and the proportion

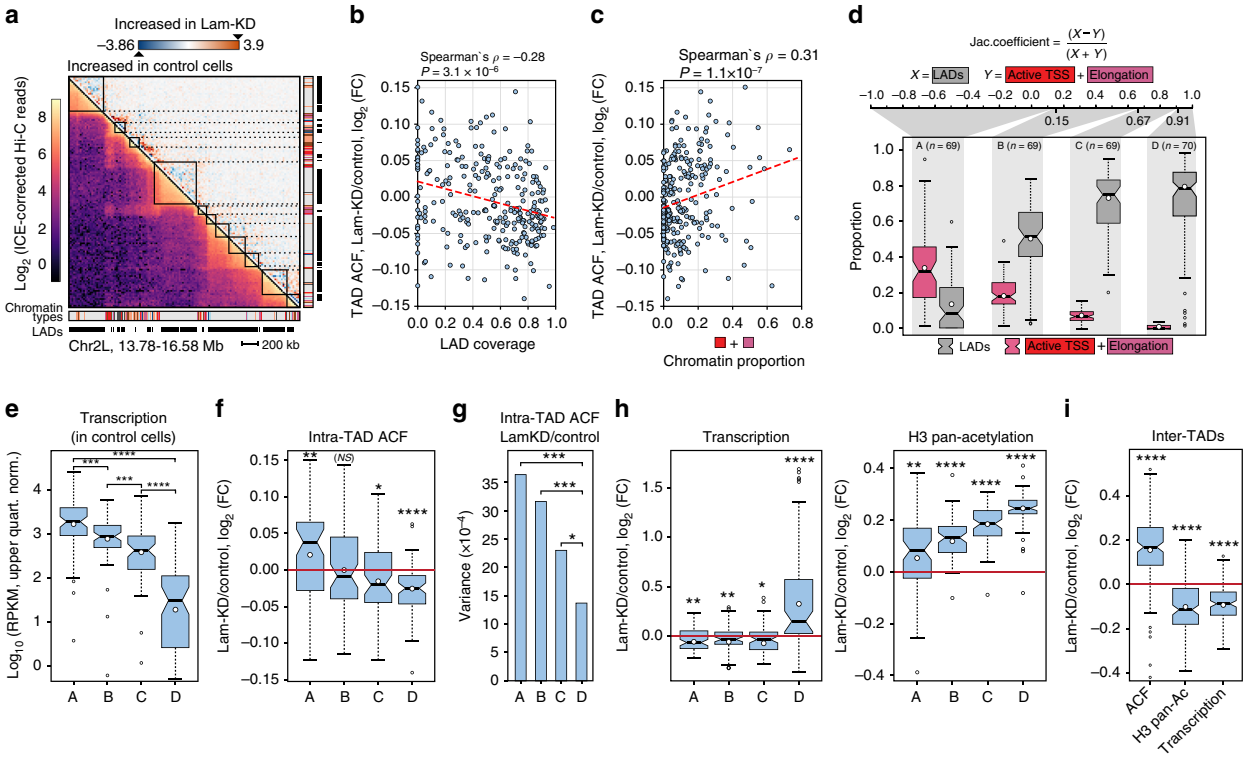


Fig. 3 TADs respond differentially to NL disruption. **a** Hi-C map showing a total chromatin interaction profile (left half of the map), or subtraction map (Lam-KD - control) (right half of the map) at a 2.8-Mb region of chromosome 2 L. Chromatin annotation in S2 cells[40] and LADs annotation in Kc167 cells[28] are shown below. **b**, **c** Dependence of intra-TAD ACF changes upon Lam-KD on the LAD coverage (**b**), or on the proportion of "red" and "purple" chromatin types[40] (**c**) within these TADs. Trend line is in red. **d** Separation of TADs into four groups according to the Jaccard similarity coefficient. Box plots show the proportion of active chromatin ("red" (active TSS) plus "purple" (elongation) chromatin types[40]) and LAD coverage[28] within each group. **e** Transcription level in the four groups of TADs according to RNA-seq in control S2 cells. **f** Changes of intra-TAD ACF between Lam-KD and control cells in the four groups of TADs. **g** Variances of $\log_2$(FC) of the intra-TAD ACF upon Lam-KD compared to control cells in the four groups of TADs. ***$P < 0.001$, *$P < 0.05$ in a Levene's test. **h** Changes of total transcription (left panel) and H3 pan-acetylation (right panel) levels between Lam-KD and control cells in the four groups of TADs. **i** Changes of ACF values, H3 pan-acetylation and total transcription in the inter-TAD regions between Lam-KD and control cells. See Fig. 1g legend for description of boxplot elements represented on panels (**d**–**f**), (**h**) and (**i**). In panels **e**, **f**, **h**, **i**, ****$P < 0.0001$, ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$, NS non-significant difference ($P > 0.05$) in a Wilcoxon test

of "red" plus "purple", but not "coral" plus "brown" active chromatin types (according to 9-type chromatin annotation in S2 cells[40]; Fig. 3c and Supplementary Fig. 3e). To simultaneously account for the LAD coverage and the proportion of "red" plus "purple" chromatin types, we calculated the Jaccard coefficient between these two metrics for each TAD. Based on this, we then divided the ranked TADs into four equal-sized groups A, B, C and D (Fig. 3d), with TADs in group A being relatively enriched in active chromatin and depleted of LADs, and TADs in group D being depleted of active chromatin and enriched in LADs. Consistent with chromatin type annotation, transcription level in control S2 cells appears to be highest in TADs from group A, and lowest in TADs from group D (Fig. 3e).

Strikingly, we observed the opposite changes of ACF values upon Lam-KD in the TADs from groups A and D having polar metrics. ACF values are increased in the TADs from group A containing the highest proportion of active chromatin and the lowest LAD coverage and are decreased in the TADs from group D with the lowest proportion of active chromatin and the highest LAD coverage (Fig. 3f). TADs from groups B and C, which preserve (group B) or slightly decrease (group C) their ACFs upon Lam-KD (Fig. 3f), likely represent the mixture of chromatin increasing and decreasing its density. In support of this idea, the variance of ACF changes is the lowest within group D TADs (Fig. 3g) which strongly correspond to LADs, when compared to other groups containing the mixture of active and inactive chromatin types (Fig. 3d).

Consistent with the transcriptional derepression in LADs (Fig. 2a, b), the overall level of transcription is markedly elevated in the group D TADs upon Lam-KD when compared to control cells (Fig. 3h, left panel). In contrast, TADs from group A demonstrate a weak decrease in transcription upon Lam-KD. Moreover, we found that upon Lam-KD, the histone H3 acetylation level is enhanced in TADs in a strong quantitative manner dependent on their LAD coverage (Supplementary Fig. 3f), with the most pronounced increase of acetylation observed in the group D TADs (Fig. 3h, right panel).

We then asked how Lam-KD influences ACF, transcription and histone H3 acetylation in the inter-TADs which represent the most active genome regions[36]. In agreement with the observations for the group A TADs, we found an increase of ACF and a decrease in transcription within inter-TAD regions upon Lam-KD (Fig. 3i). However, contrary to group A TADs, total histone H3 acetylation level appears to be decreased upon Lam-KD in the inter-TADs (Fig. 3i).

Collectively, these findings indicate that upon NL disruption, chromatin becomes more densely packed in the active, and less densely packed in the inactive genomic regions.

Lam-KD in S2 cells impairs spatial chromatin segregation. In mammals, TADs belonging to the same epigenetic type (active or inactive) tend to interact with each other across large genomic distances, thus partitioning the interphase chromatin into A and B compartments[17]. The molecular mechanisms driving such interactions are largely unknown, but a role for the NL has been suggested[6]. To identify chromatin compartments in control and Lam-KD S2 cells, we applied principal component analysis (PCA), which is commonly used for compartment calling[17]. The first principal component (PC1) profile clearly correlates with the transcription profile, where the positive values of PC1 correspond to the transcriptionally active loci (Fig. 4a and Supplementary Fig. 4a and b). Surprisingly, and contrary to the findings in *Drosophila* embryos[35], the spatially distant interactions in control S2 cells appear to be enhanced, relative to those expected, for only the genomic regions with a PC1 > 0, i.e. within the active A compartment (Fig. 4b). To verify that this is not due to technical problems in our analysis, we applied PCA to the previously published Hi-C data[35] and confirmed the existence of A and B compartments in the embryos (Supplementary Fig. 4c). Upon Lam-KD in S2 cells, interaction frequency is markedly decreased within the A compartment and is increased for the regions with a PC1 < 0, i.e. within the inactive chromatin (Fig. 4c and d and Supplementary Fig. 4d). Importantly, we observed the notable gain of interactions between A compartment and the rest of the genome upon Lam-KD (Fig. 4d). These results indicate that NL disruption leads to partial "blurring" of chromatin compartmentalisation.

Chromatin density is decreased upon TAD release from the NE. One of the most striking observations of our study is the decrease of chromatin density in a fraction of NL-attached TADs upon their release from the NE. To confirm this observation by an alternative approach, we performed two-colour FISH using a pair of probes positioned at the borders of a long TAD/LAD which is located at the cytological region *36C* and which reduces its ACF upon Lam-KD (Fig. 5a). We found that the inter-probe distances in this TAD normalised to the nuclear radius (radial-normalised distance, RND) are increased upon Lam-KD compared to those in control cells (Fig. 5b), thus indicating that the chromatin density of this TAD decreases. Moreover, in a fraction

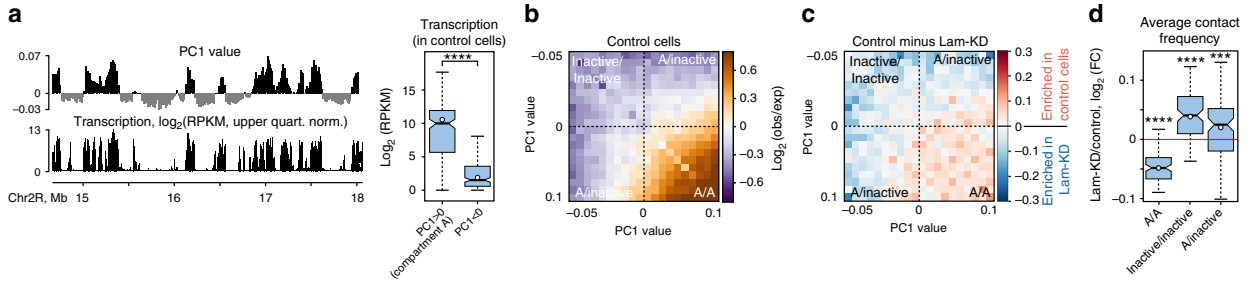


Fig. 4 NL disruption partially impairs spatial segregation of active and inactive chromatin. **a** A representative screenshot from the UCSC Genome Browser showing correspondence between the positive PC1 values and the transcriptionally active genome regions. Box plots show transcription level in 20-kb genomic bins possessing positive ($n = 4073$) and negative ($n = 1868$) PC1 values. **** – $P < 0.0001$ in a Wilcoxon test. **b** Heatmap showing $\log_2$ values of contact enrichment for the intra-chromosomal contacts in the control cells between genomic regions as a function of their PC1 values (saddle plot). **c** Subtraction of saddle plots in Lam-KD cells from that in control cells. **d** Changes of averaged contact frequency between and within active and inactive fractions of chromatin in Lam-KD compared to control cells. ****$P < 0.0001$, ***$P < 0.001$ in a Wilcoxon test. See Fig. 1g legend for description of boxplot elements represented on panels (**b**) and (**d**)

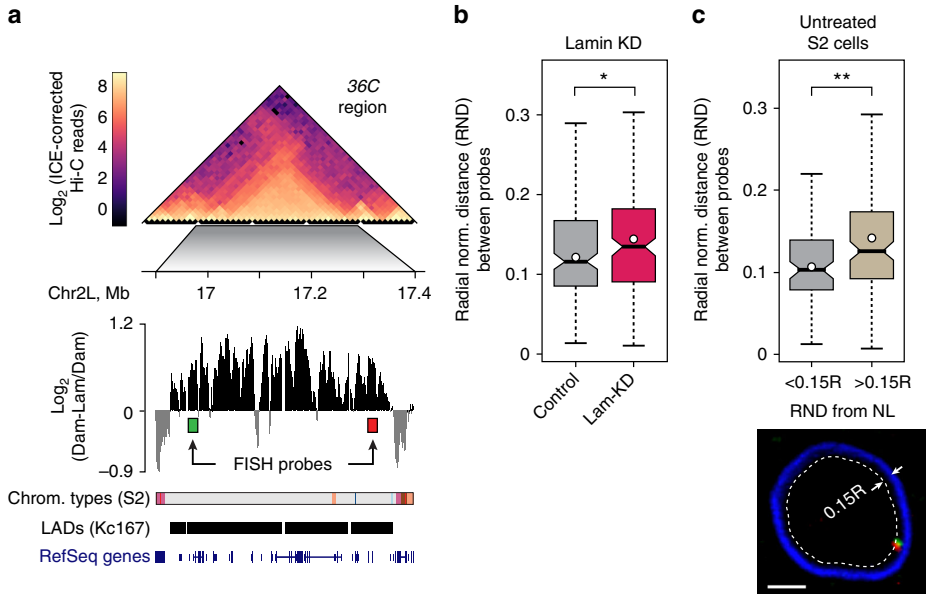


Fig. 5 Chromatin density is decreased in a TAD following its detachment from the NE. **a** Genomic region *36C* carrying probes for two-colour FISH analysis. The Hi-C map from untreated S2 cells, lamin-DamID profile from Kc167 cells[5], the occupancy of 9 chromatin types from S2 cells[40], LAD annotation from Kc167 cells[28], and RefSeq gene positions are shown. **b** Radial-normalised distance (RND) between two FISH probes in Lam-KD ($n = 175$) and control ($n = 175$) S2 cells. *$P < 0.05$ in a Kolmogorov-Smirnov one-sided test. **c** RND between two FISH probes in the untreated S2 cells. RND < 0.15, $n = 84$, RND > 0.15, $n = 124$. **$P < 0.01$ in a Kolmogorov-Smirnov one-sided test. A representative example of two-colour FISH signals (red and green) in a nucleus stained with anti-LBR antibody is shown below. Nuclear shell of 0.15 R width (R nuclear radius) where the signals are in visible contact with the NL is outlined by the dotted line. Scale bar 1 µm. See Fig. 1g legend for description of boxplot elements represented on panels (**b**, **c**)

of untreated S2 cells, where both FISH probes are confined within the shell adjoining to the NL, the RND between probes are smaller than between probes located more distally from the NE (Fig. 5c). Thus, detachment from the NL appears to be sufficient for LAD decompaction even if the NL is intact.

NL is able to mechanically compact LADs. To find out whether it is an inherent feature of the chromatin in LADs to become loosely packed after the detachment from the NL, we performed polymer modelling of chromatin-NE interactions. We employed dissipative particle dynamics (DPD)-based simulation of a model polymer (MP) whose folding pattern closely recapitulates formation of globular chromatin domains (such as TADs and LADs) built up from non-acetylated nucleosomes[36]. Here, the MP folding is simulated in the presence of a surface mimicking the NL due to its ability to interact with globular domains of the MP (blue blocks, containing "non-acetylated" sticky particles, Fig. 6a). Each blue block of the MP adopts two alternative states: it can be considered as a LAD when attached to the surface by at least one particle, or as a non-LAD when none of its particles are in contact with the surface (Fig. 6a). To obtain a dataset large enough for statistical analysis, we performed ten independent simulations. Firstly, we observe a clear TAD profile at the ensemble distance map indicating that the presence of a surface does not influence the overall folding pattern of the MP (Supplementary Fig. 5). We then ranked blue blocks from all runs according to the number of their contacts with the surface and plotted these values against the number of spatial interactions between particles in each block. We revealed a positive correlation between the number of intra-block contacts and the number of particles within this block interacting with the surface (Fig. 6b). Accordingly, the volume of a block decreases (Fig. 6c) and the shape of a block gradually changes from a sphere to a "pancake" with an increasing number of surface contacts (Fig. 6d). These results indicate that chromatin attachment to the NL per se is sufficient to compact LADs, likely confining interactions between nucleosomes in a LAD from a 3D volume to a 2D surface.

Discussion

Here, using a variety of approaches we explored what happens to chromatin upon NL disruption. Using immunostaining and FISH experiments, we revealed that Lam-KD in *Drosophila* S2 cells leads to a slight reduction in total chromatin volume and, as a result, an increase in chromatin packaging density (Fig. 1 and Supplementary Fig. 1). However, the stronger compaction of chromatin is not homogeneous and depends on the epigenetic state and scale. Our Hi-C analysis clearly indicates two opposite trends in chromatin behaviour. The contact frequency in the active chromatin increases over short distances (i.e. within the "active" TADs and the inter-TADs) and decreases over long distances (i.e. within the A compartment). Whereas in the inactive chromatin it, inversely, decreases over short distances (i.e. within the TADs mostly corresponding to LADs), but increases at the chromosomal scale (Figs. 3 and 4).

We suggest a model explaining general chromatin stretching as well as the condensation of inactive chromatin in TADs mediated by the NL (Fig. 6e). If chromatin mobility is constrained by its tethering to the NL, then the release from this tethering will lead to chromatin shrinkage due to macromolecular crowding[41] and inter-nucleosomal interactions[42,43]. Therefore counterintuitively, the NL appears not to restrict chromatin expansion but provides an anchoring surface necessary to keep interphase chromosomes slightly stretched. At the same time, inactive chromatin may become additionally condensed due to the deacetylation by HDACs, linked to the NL[31,32], and/or mechanically, due to chromatin binding with the NL (Fig. 6d).

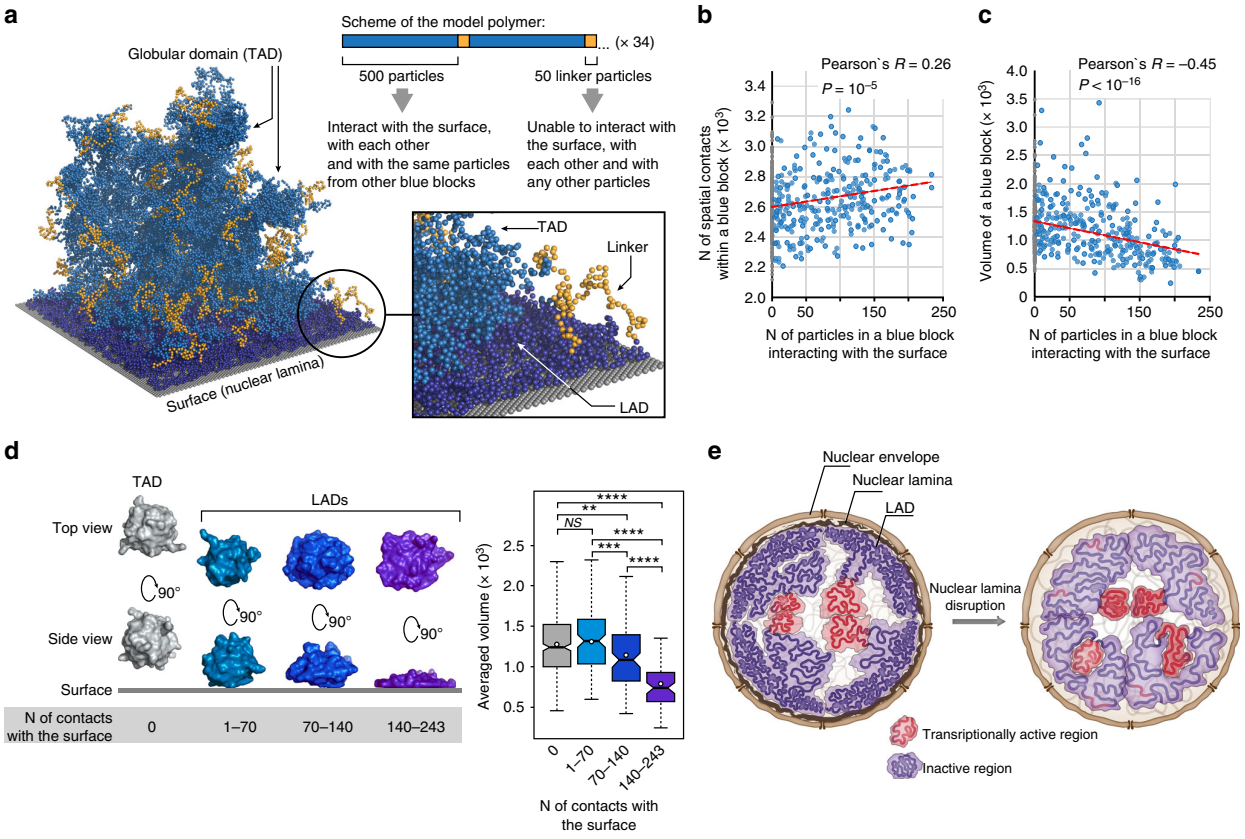


Fig. 6 Model polymer simulation shows that the attachment to the NL is sufficient for chromatin compaction. **a** Scheme of the model polymer (MP) and visualisation of its conformation derived from a simulation run. Blue particles are able to interact with each other and with the surface recapitulating properties of non-acetylated nucleosomes in the inactive chromatin. Orange particles are unable to interact with any particles and with the surface recapitulating properties of acetylated nucleosomes in active chromatin. MP is composed of 68 blue and 68 orange particle blocks. Blue blocks with at least one particle contacting the surface are coloured with dark blue and considered as LADs. **b** The number of spatial contacts between particles in a blue block depends positively on the number of particles in this block contacting the surface. Trend line is in red. **c** The volume of a blue block depends inversely on the number of particles in this block contacting the surface. Trend line is in red. **d** Averaged 3D structures of blue blocks with a different number of contacts with the surface (left panel). Box plots (see Fig. 1g legend for description of their elements) show the decrease of LAD volume with the increasing number of contacts with the surface (right panel). ****$P < 0.0001$, ***$P < 0.001$, **$P < 0.01$, NS non-significant difference ($P > 0.05$) in a Wilcoxon test. **e** Illustrative representation of key observations made in this work

A recently published study analysed the 3D genome organization upon NL disruption in mouse embryonic stem cells (mESCs)[44]. It is interesting to compare our results from *Drosophila* with those from mice. Upon loss of all lamins, the general TAD profile is still preserved in both species, however, intra- and inter-TAD interactions are altered. Strikingly, upon loss of all lamins, a fraction of NL-attached TADs becomes less condensed in both species (Fig. 3; ref. [44]). However, in contrast to *Drosophila* S2 cells, this is not accompanied by a general detachment of chromatin from the NE in mESCs[44]. Additionally, distant interactions within the inactive chromatin are mostly increased in both species upon lamin loss (Fig. 4d; ref. [44].). Finally, while some genes located at the nuclear periphery and in the nuclear interior have changed their expression both in mESCs[44] and in *Drosophila* S2 cells (Fig. 2a), an increase in the background transcription upon lamin loss is detected specifically in *Drosophila* LADs (Fig. 2b), and this was not reported for mESCs[44]. Taken together, these findings indicate that both in mammals and *Drosophila* the NL not only makes nearby chromatin more compact and repressed, but also affects chromatin interactions and gene expression in the nuclear interior.

The diversity of mechanisms of chromatin attachment to the NL in *Drosophila* and mammals may explain the differences in chromatin behaviour in response to the lack of all lamins. For example, it was shown that LBR and PRR14 proteins participate in the tethering of the H3K9-methylated chromatin to the NE in mammals[27,45]. Whereas in mammalian ESCs LADs are strongly enriched with the H3K9me2/3[2,4,46], in *Drosophila* Kc167 and, likely, in S2 cells this modification is not present in LADs[5,28]. Accordingly, our results indicate that LBR is not required to keep chromatin at the nuclear periphery in S2 cells (Fig. 1h, i). Therefore, the removal of all lamins may not be sufficient to detach all LADs from the NE in mESCs, but can release LADs in *Drosophila* S2 cells.

In conclusion, using different approaches we revealed that NL disruption in *Drosophila* S2 cells leads to general chromatin compaction, accompanied by the impaired spatial segregation of total chromatin into active and inactive types, and the decompaction of a fraction of NL-attached TADs linked to partial derepression of their chromatin. Importantly, the observed phenomena may be related to the abnormal expression of genes in lamin-associated diseases[1].

Methods

Cell cultures and RNAi. The *Drosophila melanogaster* S2 cell line (from the collection of IMG RAS) and Kc167 cell line (from the Drosophila Genomics Resource Center) were grown at 25 °C in Schneider's Drosophila Medium (Gibco) supplemented with 10% heat-inactivated fetal bovine serum (FBS, Gibco), 50 units/ml penicillin, and 50 µg/ml streptomycin. OSCs[47] kindly provided by M. Siomi were cultured in Shields and Sang M3 insect medium (Sigma-Aldrich) supplemented with 10% heat-inactivated FBS (Gibco), 10% fly extract (http://biology.st-andrews.ac.uk/sites/flycell/flyextract.html), 10 µg/ml insulin (Sigma-Aldrich), 0.6 mg/ml glutathione (Sigma-Aldrich), 50 units/ml penicillin, and 50 µg/ml streptomycin. dsRNAs against *lacZ* or lamin *Dm0* for RNAi treatment of S2 cells were prepared as previously described[11]. dsRNAs against *LBR* were prepared in the same fashion using the *Drosophila* genome DNA as a template for PCR amplification and primers provided in the Supplementary Table 1. Treatment of cells with dsRNA was performed over four days using a previously described protocol[48].

Western-blot analysis. Proteins were extracted with 8 M urea, 0.1 M Tris-HCl, pH 7.0, 1% SDS, fractionated by SDS-PAGE (12% acrylamide gel) and transferred to a PVDF membrane (Immobilon-P, Millipore). Blots were developed using alkaline phosphatase-conjugated secondary antibodies (Sigma) and the Immun-Star AP detection system (Bio-Rad). The following antibodies were used for detection: murine monoclonal anti-lamin Dm0 (1:2000; ADL67[49]), rabbit polyclonal anti-lamin C[25] (1:10000), murine monoclonal anti-beta Actin (1:3000; ab8224, Abcam).

Chromatin visualisation by histone H4 or DAPI staining. Lam-KD, LBR-KD or control S2 cells were seeded on coverslips for 30 min. After rinsing with PBS, cells were fixed in 100% methanol for 5 min at room temperature (for further examination of chromatin distribution based on the immunostaining of histone H4) or in 4% formaldehyde in PBS for 25 min at room temperature (for further estimation of chromatin volume based on DAPI staining), rinsed with PBS three times, blocked with PBTX (PBS with 0.1% Tween-20 and 0.3% Triton X-100) containing 3% normal goat serum (Invitrogen) for 1 h at room temperature. The remaining immunostaining procedure was performed as previously described[50]. As primary antibodies we used murine monoclonal anti-histone H4 (1:200; ab31830, Abcam), guinea pig polyclonal anti-LBR[26] (1:1000), rabbit polyclonal anti-lamin Dm0[51] (1:500). As the secondary antibodies we used Alexa Fluor 546-conjugated goat anti-rabbit IgG (Invitrogen) or Alexa Fluor 488-conjugated goat anti-mouse IgG (Invitrogen), or Alexa Fluor 633-conjugated goat anti-guinea pig IgG (Invitrogen).

ImageJ quantitation of chromatin distribution. Using ImageJ, we measured histone H4, LBR and lamin Dm0 profiles across the nucleus diameter of the equatorial focal plane of nuclei of Lam-KD, LBR-KD or control S2 cells. Fluorescent intensities were extracted, individual profiles were first normalised on the average intensity, then on the diameter of the nucleus (delimited by peaks of LBR fluorescence for Lam-KD, or by peaks of lamin Dm0 fluorescence for LBR-KD) and further aligned to determine the averaged profile. Nuclei from 2–3 independent experiments (60 nuclei per experiment) were analysed.

Estimation of the volume of DAPI-stained chromatin. Confocal images containing 20–30 DAPI-stained formaldehyde-fixed Lam-KD or control S2 cells were processed and analysed with the same parameters using IMARIS 7.4.2 software (Bitplane AG). Only nuclei with the lowest residual lamin Dm0 staining were used for analysis in Lam-KD cells, whereas in control cells, conversely, the nuclei with poor lamin Dm0 staining were not taken for analysis. For background subtraction, images were thresholded to ~15% of the maximal intensity of the channel so that the generated nuclear surfaces would not expand beyond the peak of LBR fluorescence intensity. With these parameters, the surfaces of nuclei, appropriate for analysis, were automatically reconstructed. Finally, the volumes of ~100 reconstructed nuclei were retrieved from the Statistics tab for the analysis.

Two-colour FISH. ~20-kb FISH probes were generated using a long-range PCR kit (Encyclo Plus PCR (Evrogen)) by PCR-amplification of 4 tiling genome fragments covering either the region 2 L:16964000–16982000 or 2 L:17310000–17328000, with the use of primer pairs provided in the Supplementary Table 1. 1 µg of template DNA for hybridization was labelled by random primed synthesis with the DIG DNA labelling kit (Roche) or by ChromaTide Alexa Fluor 546–14-dUTP (Life Technologies). Probes were further combined and hybridized with S2 cells as described previously[23]. For NL or FISH probe detection, as the primary antibodies we used guinea pig polyclonal anti-LBR[26] (1:1000), or rabbit polyclonal anti-lamin Dm0[51] (1:500) and sheep polyclonal anti-DIG-FITC (1:500, Roche). As the secondary antibodies we used Alexa Fluor 633-conjugated goat anti-guinea pig IgG (Invitrogen), or Alexa Fluor 546-conjugated goat anti-rabbit IgG (Invitrogen) and Alexa Fluor 488-conjugated goat anti-FITC IgG (Invitrogen).

Measurement of distances between FISH probes and the NE. Three-dimensional image stacks were recorded with a confocal LSM 510 Meta laser scanning microscope (Zeiss). Optical sections with 0.4-µm intervals along the Z-axis were captured. Images were processed and analysed by using IMARIS 7.4.2 software (Bitplane AG) with the blind experimental setup. Distances between both probes or between the probes and the NE were counted as previously described[23]. Briefly, we were unable to fully reconstruct the surfaces of nuclei automatically based on their LBR or lamin Dm0 immunostaining. Therefore, the nuclear rim of a particular nucleus was manually outlined in all optical sections of the stack by the middle of its LBR or lamin Dm0 staining to further reconstruct the surface of this nucleus automatically. To determine the distance between FISH signals and the NE, the instrument "measurement point" was positioned on the brightest voxel of the FISH probe and another "measurement point" was positioned on the reconstructed nuclear surface at the point of its earliest intersection with a progressively growing sphere from the first "measurement point". The distance between two "measurement points" (i.e. the shortest distance between the centre of the FISH probe and the middle of the NE) was measured for each nucleus. Distances between two FISH probes were measured correspondingly. Data were obtained in two independent experiments for 75–100 nuclei per experiment. In parallel, volumes of nuclei were retrieved, and radii of nuclei were calculated considering nuclei to be spherical. Finally, distances were normalised to the nuclear radii.

Analysis of gene expression. Total RNA was isolated from Lam-KD or control S2 cells using Trizol reagent (Invitrogen), and contaminating DNA was removed by DNase I treatment. RNA quality was assessed using capillary electrophoresis with a Bioanalyzer 2100 (Agilent). Poly(A)$^+$ RNA was extracted from total RNA using oligo(dT) magnetic beads (Thermo Fisher Scientific). NEBNext Ultra II RNA library preparation kit (New England Biolabs) was used for preparation of libraries following manufacturer's instructions. Libraries from two biological replicates of Lam-KD or control S2 cells were quantified using a Qubit fluorometer and quantitative PCR, and sequenced on the Illumina NextSeq resulting in 8.4–9.4 × 10^6 80-nt single-end reads. Reads were mapped to the *D. melanogaster* reference genome (version dm3) using HISAT[52] v2.1.0 with option –max-intronlen 50,000. Reads with low mapping quality were removed using SAMtools[53] with option -q 30. We calculated $\log_2$ transcription levels in 20-kb genomic bins using BEDtools[54] v2.16.2 with option -split, and then applied the hclust function in R to cluster the replicates using 1-Spearman's correlation coefficient as a distance metric. Gene expression was quantified with StringTie[52] for the reference annotation version r5.12. We filtered out genes with zero expression in more than two replicates. Among the remaining 10,076 genes, the differentially expressed genes were defined using the edgeR[55] package with trimmed mean of M values (TMM) normalisation at FDR = 0.05 cutoff. Genes were assigned to the LADs if their TSSs were located within LADs, while genes were assigned to the inter-LADs if their TSSs were at least 1-kb distant from LADs. A pseudocount was added to all expression values to get rid of zeros. The pseudocount was calculated as the minimal value in the gene expression table after normalisation. Then, we averaged the replicates and calculated $\log_2$(FC) values between Lam-KD and control samples.

Real-time RT-qPCR assay for the randomly selected genes from different LADs was performed on cDNAs synthesised with oligo(dT) primers on the poly(A)$^+$ RNA isolated from 3 biological replicates of Lam-KD or control S2 cells, using EvaGreen chemistry (Jena Bioscience) and the CFX96 hardware (BioRad). Expression levels of genes were normalised on the *act5C* gene expression. For semi-quantitative RT-PCR, applied for the analysis of genes from the 60D LAD, the reverse transcription of RNA was performed using SuperScript II reverse transcriptase (Invitrogen) in the presence of hexamer random primers. PCR amplification of cDNAs was performed with the addition of ^{33}P-dATP. Probes after PCR were separated in 5% PAAG, which was then fixed, dried and exposed to the storage phosphor screen (Amersham Biosciences). The signals were scanned with a Phosphorimager Storm-820 (Molecular Dynamics). For each primer pair the number of PCR cycles were optimised to fit the exponential phase of amplification which was controlled by two-fold cDNA dilution. The expression levels of genes were normalised to the ubiquitous *CG4589* gene expression. Sequences of gene-specific primers are presented in the Supplementary Table 1.

ChIP-seq procedure and data analysis. ChIP-seq from two biological replicates of control and Lam-KD S2 cells with anti-H3-pan acetylated antibodies (Active Motif, #39139) was performed as previously described[56], with some modifications. After rinsing with PBS, ~2 × 10^7 cells were fixed with 1.8% formaldehyde in PBS containing 0.5 mM DTT for 20 min at room temperature. Cross-linking was stopped by adding glycine to 225 mM for 5 min and washing in PBS containing 0.5 mM DTT three times for 5 min. Cells were washed once in the A2 buffer (140 mM NaCl, 15 mM HEPES pH 7.6, 1 mM EDTA, 0.5 mM EGTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.5 mM DTT, complete EDTA-free protease inhibitor cocktail (Roche)). Cells were lysed in the A2 buffer containing 1% SDS for 10 min at room temperature, after that the lysate was 20-fold diluted by the A2 buffer and incubated for 2 min at 4 °C. After sonication with VCX 400 Vibra-Cell Processor (Sonics; 30 pulses of 10 sec with 10-sec intervals at 15% max power) and 10-min high-speed centrifugation, the fragmented chromatin (with the average DNA fragment size ~0.5 kb) was recovered in the supernatant. For each immunoprecipitation, ~10 µg of chromatin (~700 µl) was pre-incubated in the presence of 100 µl of Protein A-Sepharose (PAS, 50% w/v, GE Healthcare) for 1 h at 4 °C. PAS was removed by centrifugation, 5% of chromatin was isolated as an "Input"

material, after that 2 µl anti-H3-pan acetylated antibodies (Active Motif, #39139) were added to the rest chromatin and samples were incubated overnight at 4 °C in a rotating wheel. Then, 100 µl of PAS was added and incubation was continued for 4 h at 4 °C. Samples were centrifuged at maximum speed for 1 min and the supernatant was discarded. Samples were washed four times in the A2 buffer containing 0.05% SDS and twice in 1 mM EDTA, 10 mM Tris (pH 8), 0.5 mM DTT buffer (each wash for 5 min at 4 °C). Chromatin was eluted from PAS in 100 µl of 10 mM EDTA, 1% SDS, 50 mM Tris (pH 8) at 65 °C for 10 min, followed by centrifugation and recovery of the supernatant. PAS material was re-extracted in 150 µl of TE, 0.67% SDS. To reverse cross-links, the combined eluate (250 µl) was incubated 6 h at 65 °C and treated by Proteinase K for 3 h at 50 °C. Samples were phenol-chloroform extracted and isopropanol precipitated in the presence of 20 µg glycogen. DNA was dissolved in 100 µl of water. ChIP samples containing ~25 ng of precipitated DNA, as well as "Input" samples were prepared for next-generation sequencing using a NEBNext Ultra II DNA library prep kit for Illumina (New England Biolabs). Libraries were sequenced on the Illumina HiSeq 2000 resulting in $3.1–3.4 \times 10^6$ 75-bp single-end reads. Reads were mapped to the *D. melanogaster* reference genome (version dm3) using Bowtie 2 v2.2.1 (with the –very-sensitive option)[57]. Reads with low mapping quality were removed using SAMtools[53] with option -q 30. Duplicate reads were removed using SAMtools rmdup. We calculated $\log_2$ ChIP and input signals in 1-kb genomic bins using BEDtools[54] v2.16.2, and then applied hclust function in R to cluster the replicates using 1-Spearman's correlation coefficient as a distance metric. Reads were assigned to LADs if they overlapped LADs, while reads were assigned to inter-LADs if they were at least 1-kb distant from LADs. We calculated read numbers within each LAD and inter-LAD, normalised the values for the sum of read coverage per replicate, excluded zero-covered LADs and inter-LADs from further analysis, averaged the replicates, and calculated $\log_2$(FC) values between Lam-KD and control ChIP samples.

Hi-C procedure and data analysis. Hi-C libraries from two independent biological replicates of control and Lam-KD S2 cells were prepared essentially as described previously[36] using the *HindIII-HF* restriction enzyme (NEB). Libraries were sequenced on the Illumina HiSeq 2000 platform resulting in $3–4 \times 10^7$ paired-end reads. Reads were mapped to the *D. melanogaster* reference genome (version dm3) using Bowtie 2 v2.2.1 (with the –very-sensitive option)[57]. The Hi-C data were processed using the ICE pipeline v0.9 (20 iterations of iterative correction) as described[58]. Hi-C interaction maps with 20-kb resolution were obtained. TADs were predicted using the Armatus software[59] v1.0, in which the average size and the number of TADs is determined by the scaling parameter γ. TAD annotation was performed in two steps as described[36]. First, we manually selected parameter γ to achieve good partitioning of TADs ($\gamma = 1.20$ for Lam-KD cells and $\gamma = 1.12$ for control cells). Then, TADs larger than 600 kb were split into smaller TADs with the scaling parameter γ multiplied by 2. After that, the smallest TADs (equal or less than 60 kb) were annotated as inter-TADs due to their poorly resolved internal structure. As a result, 576 (in control) and 588 (in Lam-KD) TADs were revealed. To examine whether TAD positions are altered upon Lam-KD, we analysed the degree of overlap of each TAD in the merged replicates of control and Lam-KD cells with that in the control replicates or in the Lam-KD replicates and did not find statistically significant difference ($P > 0.05$ in a two-sided Wilcoxon test). ACF within each TAD was calculated as an average value of iteratively corrected read numbers between all genomic bins belonging to the TAD, excluding boundary bins from both TAD sides. ACF within each inter-TAD was calculated as an average value of iteratively corrected read numbers between all genomic bins belonging to the inter-TAD and the boundary bins of adjacent TADs. For each TAD, we calculated the ratio between ACF value in each Lam-KD replicate and ACF value in each control replicate (four ratios in total). TADs with at least three ratios of the same sign were used for the downstream analysis. We note that when we selected TADs according to more strict criterion (i.e. all four ratios were changed in the same direction), it did not affect the results of analysis (Supplementary Fig. 6). Chromatin compartments were annotated using the principal component analysis as described[17]. Saddle plots were generated as described[58]. Briefly, we used the observed/expected Hi-C maps, which we calculated from 20-kb iteratively corrected interaction maps of *cis*-interactions by dividing each diagonal of a matrix by its chromosome-wide average value. In each observed/expected map, we rearranged the rows and the columns in the order of increasing PC1 values (which we calculated for the control matrices). Finally, we aggregated the rows and the columns of the resulting matrix into 20 equally sized aggregated bins, thus obtaining a saddle plot of compartmentalization.

Analysis of published data. We employed chromatin type annotation for S2 cells[40]. Proportions of chromatin types in 20-kb bins were calculated. Annotation of LADs was obtained from ref. [28]. We calculated the proportion of LAD length in each 20-kb TAD bin.

Polymer modelling. We used Dissipative Particle Dynamics (DPD) to perform computer simulations, as previously described[36] with some modifications. Briefly, macromolecules are represented in terms of the bead-and-spring model, with the particles interacting by a conservative force (repulsion), a dissipative force (friction), and a random force (heat generator). A detailed description of the implementation of this technique was provided earlier[60]. The simulated cell volume was $50 \times 50 \times 50$ DPD units, density equals 3, so the total number of particles in the system is 375,000. We assume that a particle corresponds to a nucleosome. In addition, we introduce special boundary conditions, which are periodic for the solvent and impermeable for other particles. The surface consists of immobile, hexagonally positioned particles. In our simulations, particles mimic either "active" or "inactive" nucleosome types, while the surface mimics the NL. "Inactive" particles may create reversible "saturating" bonds[61,62] with each other as well as with the particles of a surface. Each "inactive" particle may have only one additional bond per moment, which simulates an interaction of a positively charged histone tail of non-acetylated nucleosome with the "acidic patch" of another nucleosome[42,43,63]. Our copolymer chain is represented by 64 blocks each consisting of 500 "inactive" and 50 "active" particles. The probability of creating an association between two "inactive" particles was set to 0.001, between the "inactive" particle and the surface – 0.007, while the probability to break such association was set to 0.01. During simulations, all particles were checked every 200 DPD steps, when the local equilibration was obtained. We performed 10 independent runs on the MSU supercomputer "Lomonosov-2" using our own implementation for the domain decomposition parallelised DPD code which is available at GitHub [https://github.com/KPavelI/dpd].

Statistical analysis. We applied the Wilcoxon test to check whether the distribution of $\log_2$(FC) values was symmetric around zero, as well as to test whether two distributions of $\log_2$(FC) values differed by a location shift of zero.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw and processed Hi-C, RNA-seq and ChIP-seq data were deposited in the GEO NCBI under the accession number GSE110082. DPD code is available at GitHub [https://github.com/KPavelI/dpd]. The source data underlying Fig. 1a, b, d–f, h, i, 2d, 3, 5b, c and Supplementary Fig. 1c are provided as a Source Data file. All other data supporting the findings of this study are available from the corresponding authors upon request. A reporting summary for this Article is available as a Supplementary Information file.

Received: 16 February 2018 Accepted: 21 February 2019
Published online: 12 March 2019

References

1. Gruenbaum, Y. & Foisner, R. Lamins: nuclear intermediate filament proteins with fundamental functions in nuclear mechanics and genome regulation. *Annu. Rev. Biochem.* **84**, 131–164 (2015).
2. Guelen, L. et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948–951 (2008).
3. Ikegami, K., Egelhofer, T. A., Strome, S. & Lieb, J. D. Caenorhabditis elegans chromosome arms are anchored to the nuclear membrane via discontinuous association with LEM-2. *Genome Biol.* **11**, R120 (2010).
4. Peric-Hupkes, D. et al. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol. Cell* **38**, 603–613 (2010).
5. van Bemmel, J. G. et al. The insulator protein SU(HW) fine-tunes nuclear lamina interactions of the Drosophila genome. *PLoS ONE* **5**, e15013 (2010).
6. van Steensel, B. & Belmont, A. S. Lamina-associated domains: links with chromosome architecture, heterochromatin, and gene repression. *Cell* **169**, 780–791 (2017).
7. Kosak, S. T. et al. Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development. *Science* **296**, 158–162 (2002).
8. Zink, D. et al. Transcription-dependent spatial arrangements of CFTR and adjacent genes in human cell nuclei. *J. Cell. Biol.* **166**, 815–825 (2004).
9. Ragoczy, T., Bender, M. A., Telling, A., Byron, R. & Groudine, M. The locus control region is required for association of the murine β-globin locus with engaged transcription factories during erythroid maturation. *Genes Dev.* **20**, 1447–1457 (2006).
10. Williams, R. R. et al. Neural induction promotes large-scale chromatin reorganisation of the Mash1 locus. *J. Cell. Sci.* **119**, 132–140 (2006).
11. Shevelyov, Y. Y. et al. The B-type lamin is required for somatic repression of testis-specific gene clusters. *Proc. Natl Acad. Sci. USA* **106**, 3282–3287 (2009).
12. Akhtar, W. et al. Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell* **154**, 914–927 (2013).
13. Finlan, L. E. et al. Recruitment to the nuclear periphery can alter expression of genes in human cells. *PLoS Genet.* **4**, e1000039 (2008).

14. Reddy, K. L., Zullo, J. M., Bertolino, E. & Singh, H. Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature* **452**, 243–247 (2008).
15. Dialynas, G., Speese, S., Budnik, V., Geyer, P. K. & Wallrath, L. L. The role of Drosophila lamin C in muscle function and gene expression. *Development* **137**, 3067–3077 (2010).
16. Barton, L. J., Soshnev, A. A. & Geyer, P. K. Networking in the nucleus: a spotlight on LEM-domain proteins. *Curr. Opin. Cell Biol.* **34**, 1–8 (2015).
17. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
18. Zhu, J. et al. Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642–654 (2013).
19. Kind, J. et al. Genome-wide maps of nuclear lamina interactions in single human cells. *Cell* **163**, 134–147 (2015).
20. Milon, B. C. et al. Map of open and closed chromatin domains in Drosophila genome. *BMC Genom.* **15**, 988 (2014).
21. Boettiger, A. N. et al. Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature* **529**, 418–422 (2016).
22. Kind, J. et al. Single-cell dynamics of genome-nuclear lamina interactions. *Cell* **153**, 178–192 (2013).
23. Milon, B. C. et al. Role of histone deacetylases in gene regulation at nuclear lamina. *PLoS ONE* **7**, e49692 (2012).
24. Verboon, J. M. et al. Wash interacts with lamin and affects global nuclear organization. *Curr. Biol.* **25**, 804–810 (2015).
25. Riemer, D. et al. Expression of Drosophila lamin C is developmentally regulated: analogies with vertebrate A-type lamins. *J. Cell. Sci.* **108**, 3189–3198 (1995).
26. Wagner, N., Weber, D., Seitz, S. & Krohne, G. The lamin B receptor of Drosophila melanogaster. *J. Cell. Sci.* **117**, 2015–2028 (2004).
27. Solovei, I. et al. LBR and lamin A/C sequentially tether peripheral heterochromatin and inversely regulate differentiation. *Cell* **152**, 584–598 (2013).
28. Filion, G. J. et al. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* **143**, 212–224 (2010).
29. Cherbas, L. et al. The transcriptional diversity of 25 Drosophila cell lines. *Genome Res.* **21**, 301–314 (2011).
30. Lee, H. et al. DNA copy number evolution in Drosophila cell lines. *Genome Biol.* **15**, R70 (2014).
31. Somech, R. et al. The nuclear-envelope protein and transcriptional repressor LAP2β interacts with HDAC3 at the nuclear periphery, and induces histone H4 deacetylation. *J. Cell. Sci.* **118**, 4017–4025 (2005).
32. Holaska, J. M. & Wilson, K. L. An emerin "proteome": purification of distinct emerin-containing complexes from HeLa cells suggests molecular basis for diverse roles including gene regulation, mRNA splicing, signaling, mechanosensing, and nuclear architecture. *Biochem* **46**, 8897–8908 (2007).
33. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
34. Hou, C., Li, L., Qin, Z. S. & Corces, V. G. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Mol. Cell* **48**, 471–484 (2012).
35. Sexton, T. et al. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458–472 (2012).
36. Ulianov, S. V. et al. Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Res.* **26**, 70–84 (2016).
37. Ramírez, F. et al. High-affinity sites form an interaction network to facilitate spreading of the MSL complex across the X chromosome in Drosophila. *Mol. Cell* **60**, 146–162 (2015).
38. Eagen, K. P., Hartl, T. A. & Kornberg, R. D. Stable chromosome condensation revealed by chromosome conformation capture. *Cell* **163**, 934–946 (2015).
39. Li, L. et al. Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing. *Mol. Cell* **58**, 216–231 (2015).
40. Kharchenko, P. V. et al. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471**, 480–485 (2011).
41. Hancock, R. Packing of the polynucleosome chain in interphase chromosomes: evidence for a contribution of crowding and entropic forces. *Semin. Cell. Dev. Biol.* **18**, 668–675 (2007).
42. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251–260 (1997).
43. Schalch, T., Duda, S., Sargent, D. F. & Richmond, T. J. X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature* **436**, 138–141 (2005).
44. Zheng, X. et al. Lamins organize the global three-dimensional genome from the nuclear periphery. *Mol. Cell* **71**, 802–815.e7 (2018).
45. Poleshko, A. et al. The human protein PRR14 tethers heterochromatin to the nuclear lamina during interphase and mitotic exit. *Cell Rep.* **5**, 292–301 (2013).
46. Wen, B., Wu, H., Shinkai, Y., Irizarry, R. A. & Feinberg, A. P. Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nat. Genet.* **41**, 246–250 (2009).
47. Niki, Y., Yamaguchi, T. & Mahowald, A. P. Establishment of stable cell lines of Drosophila germ-line stem cells. *Proc. Natl Acad. Sci. USA* **103**, 16325–16330 (2006).
48. Clemens, J. C. et al. Use of double-stranded RNA interference in Drosophila cell lines to dissect signal transduction pathways. *Proc. Natl Acad. Sci. USA* **97**, 6499–6503 (2000).
49. Stuurman, N., Maus, N. & Fisher, P. A. Interphase phosphorylation of the Drosophila nuclear lamin: site-mapping using a monoclonal antibody. *J. Cell. Sci.* **108**, 3137–3144 (1995).
50. Ilyin, A. A. et al. Piwi interacts with chromatin at nuclear pores and promiscuously binds nuclear transcripts in Drosophila ovarian somatic cells. *Nucleic Acids Res.* **45**, 7666–7680 (2017).
51. Osouda, S. et al. Null mutants of Drosophila B-type lamin Dm(0) show aberrant tissue differentiation rather than obvious nuclear shape distortion or specific defects during cell proliferation. *Dev. Biol.* **284**, 219–232 (2005).
52. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).
53. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
54. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
55. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
56. Chanas, G., Lavrov, S., Iral, F., Cavalli, G. & Maschat, F. Engrailed and polyhomeotic maintain posterior cell identity through cubitus-interruptus regulation. *Dev. Biol.* **272**, 522–535 (2004).
57. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
58. Imakaev, M. et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012).
59. Filippova, D., Patro, R., Duggal, G. & Kingsford, C. Identification of alternative topological domains in chromatin. *Algorithms Mol. Biol.* **9**, 14 (2014).
60. Groot, R. D. & Warren, P. B. Dissipative particle dynamics: Bridging the gap between atomistic and mesoscopic simulation. *J. Chem. Phys.* **107**, 4423 (1997).
61. Lifshitz, I. M., Grosberg, A. Y. & Khokhlov, A. R. Structure of a polymer globule formed by saturating bonds. *J. Exp. Theor. Phys.* **44**, 855–860 (1976).
62. Chertovich, A. V., Ivanov, V. A., Khokhlov, A. R. & Bohr, J. Monte Carlo simulation of AB-copolymers with saturating bonds. *J. Phys. Condens. Matter* **15**, 3013–3027 (2003).
63. Shogren-Knaak, M. et al. Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science* **311**, 844–847 (2006).

Acknowledgements

We thank Paul Fisher for anti-lamin C and anti-lamin Dm0 antibodies, Georg Krohne for anti-LBR antibody, Alexey A. Gavrilov (Faculty of Physics, M.V. Lomonosov Moscow State University) for DPD computer code. The research was carried out using the equipment of the shared research facilities of HPC computing resources at M.V. Lomonosov Moscow State University. This work was supported by the Russian Science Foundation (grant number 16–14–10081), by the Russian Foundation for Basic Research (grant number 17–00–00183) and by Foundation for the Advancement of Theoretical Physics "BASIS" (grant number 17–21–2101–1 to P.I.K.).

Author contributions

S.V.U., S.A.D., E.E.K. and P.I.K contributed equally. Y.Y.S. and S.V.R. conceived the project. S.V.U. performed Hi-C and gene expression analysis. S.A.D. carried out lamin Dm0 and LBR depletions, western-blot analysis, ChIP, biological material preparation for RNA-seq, immunostaining and FISH experiments. E.A.M. maintained cell cultures. E.E.K. processed Hi-C data. E.E.K., S.V.U., A.A.I. and S.S.S. analysed various data sets (RNA-seq, ChIP-seq, Lamin DamID, chromatin type annotations). E.E.K., S.V.U., S.S.S. (supervised by E.E.K.), A.A.Galilsyna (supervised by M.S.G.) and I.M.F. performed Hi-C data analysis. V.V.N. performed FISH data analysis. P.I.K. and A.V.C.

performed polymer simulations. A.V.L. (supervised by A.A.Gavrilov) and P.I.K. performed analysis of polymer simulation data. A.V.L. (supervised by S.V.U.) performed PCA analysis. M.D.L. prepared libraries for ChIP-seq and RNA-seq, and carried out NGS of Hi-C, RNA-seq and ChIP-seq libraries. Y.Y.S. and S.V.U. wrote the manuscript with the input from all authors.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-09185-y.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

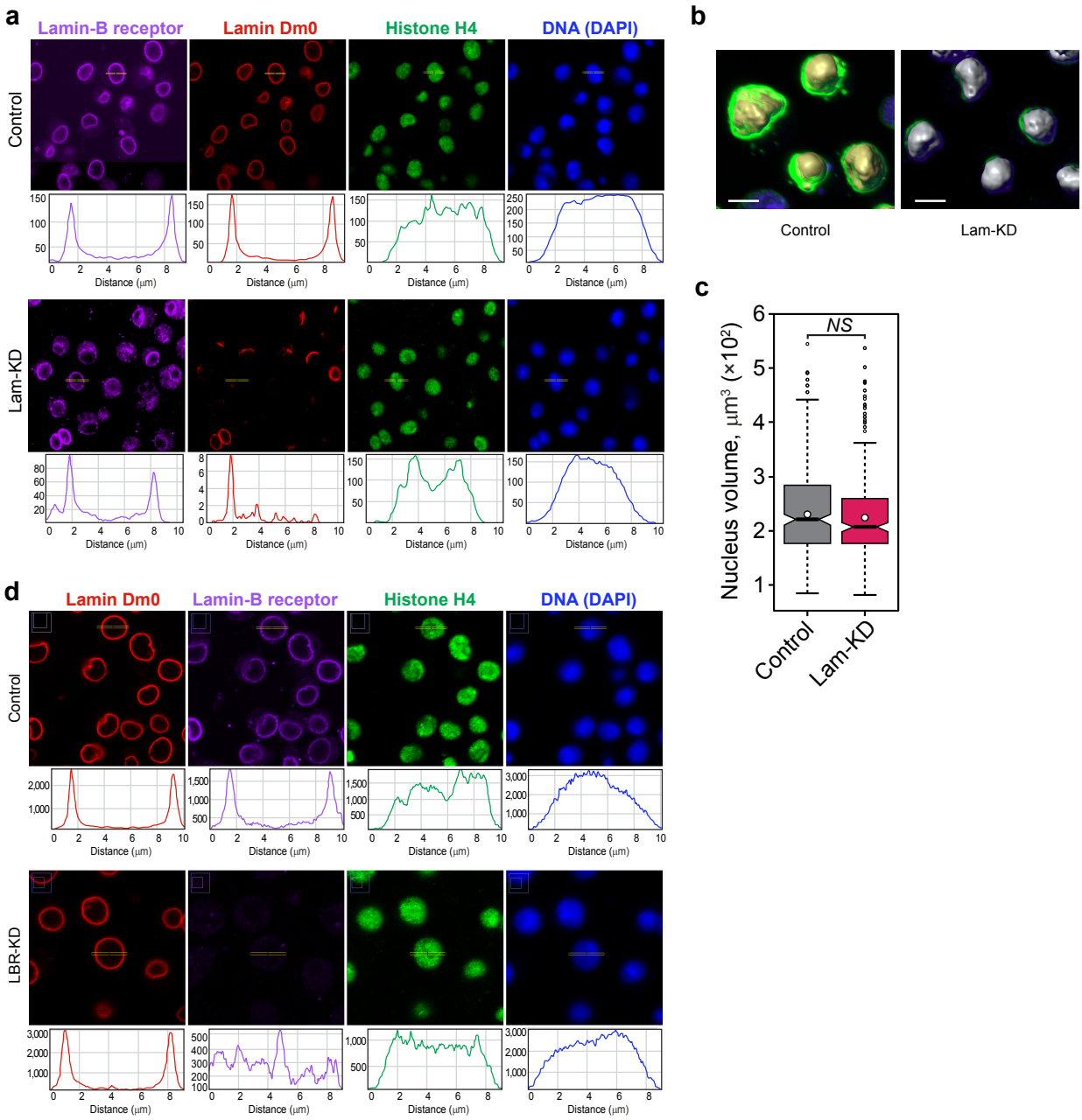
Journal peer review information: *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

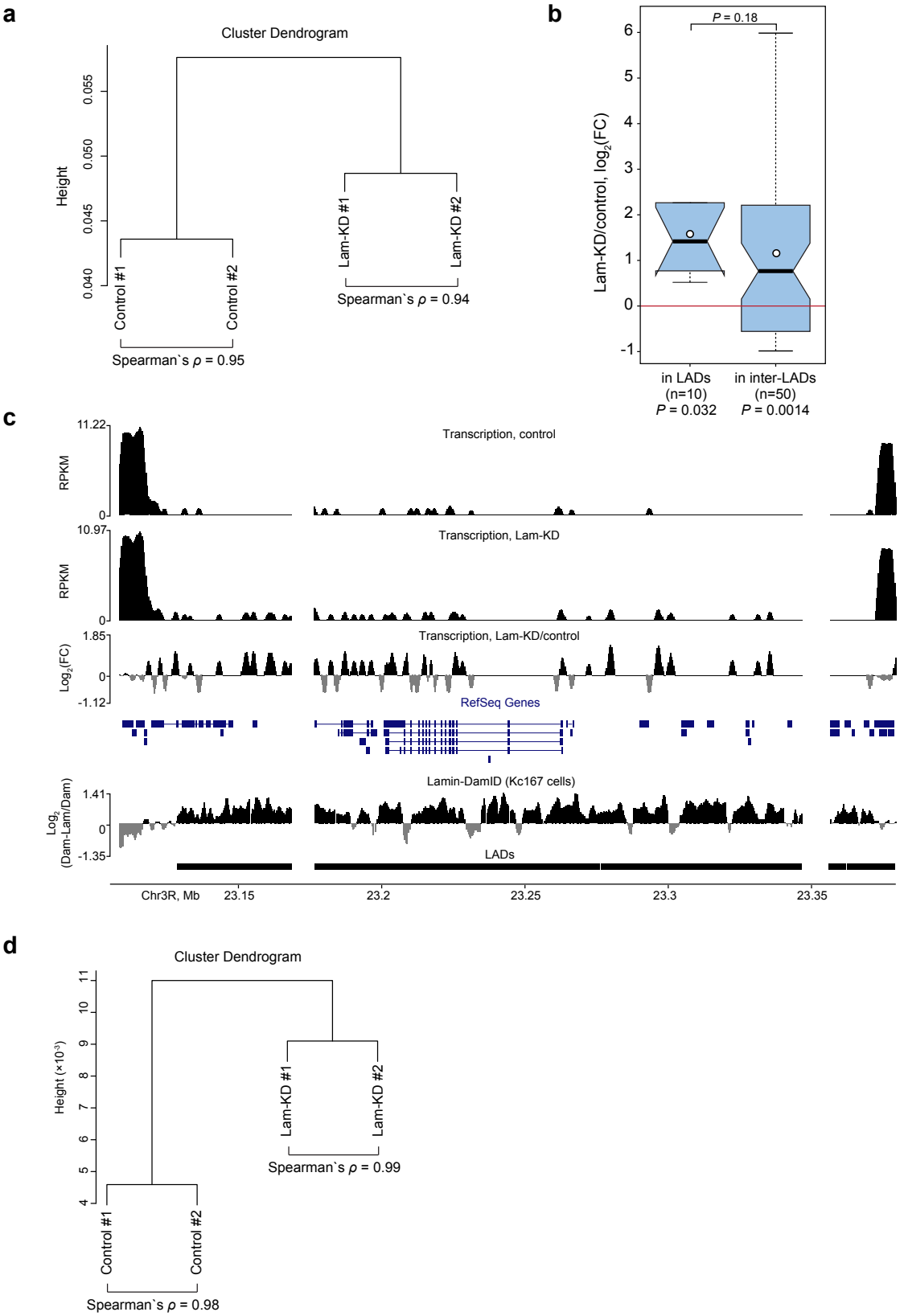
Supplementary Information

Nuclear lamina integrity is required for proper spatial organization of chromatin in *Drosophila*

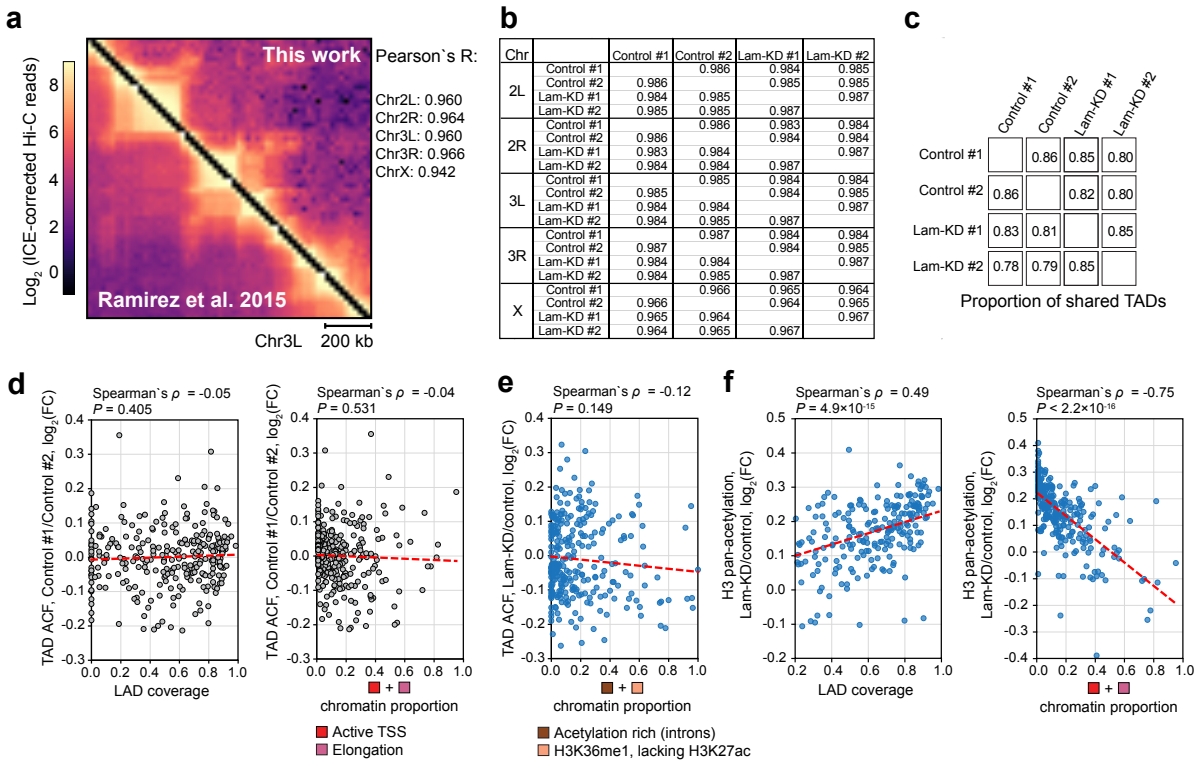
Ulianov et al.



Supplementary Fig. 1 Immunostaining of chromatin in Lam-KD, LBR-KD and control cells. **a** A representative example of nuclei immunostained with antibodies against histone H4, LBR and lamin Dm0 in Lam-KD and control cells. Fluorescence intensity along the yellow-framed zone was measured using ImageJ software and presented below the images. **b** A representative example of nuclei immunostained with antibodies against lamin Dm0 (green) and counterstained by DAPI (blue) with the subsequent automatic reconstruction of the chromatin surface by DAPI staining using IMARIS software in Lam-KD or control S2 cells. Scale bar 5 μm. **c** Distribution of the volume of nuclei in Lam-KD (n=275) or control (n=275) S2 cells reconstructed by the LBR-stained NE using IMARIS software. *NS* – non-significant difference ($P > 0.05$ in a Wilcoxon test). Thick black lines and white dots represent median and average values, respectively. **d** A representative example of nuclei immunostained with antibodies against histone H4, LBR and lamin Dm0 in LBR-KD and control cells. Fluorescence intensity along the yellow-framed zone was measured using ImageJ software and presented below the images.



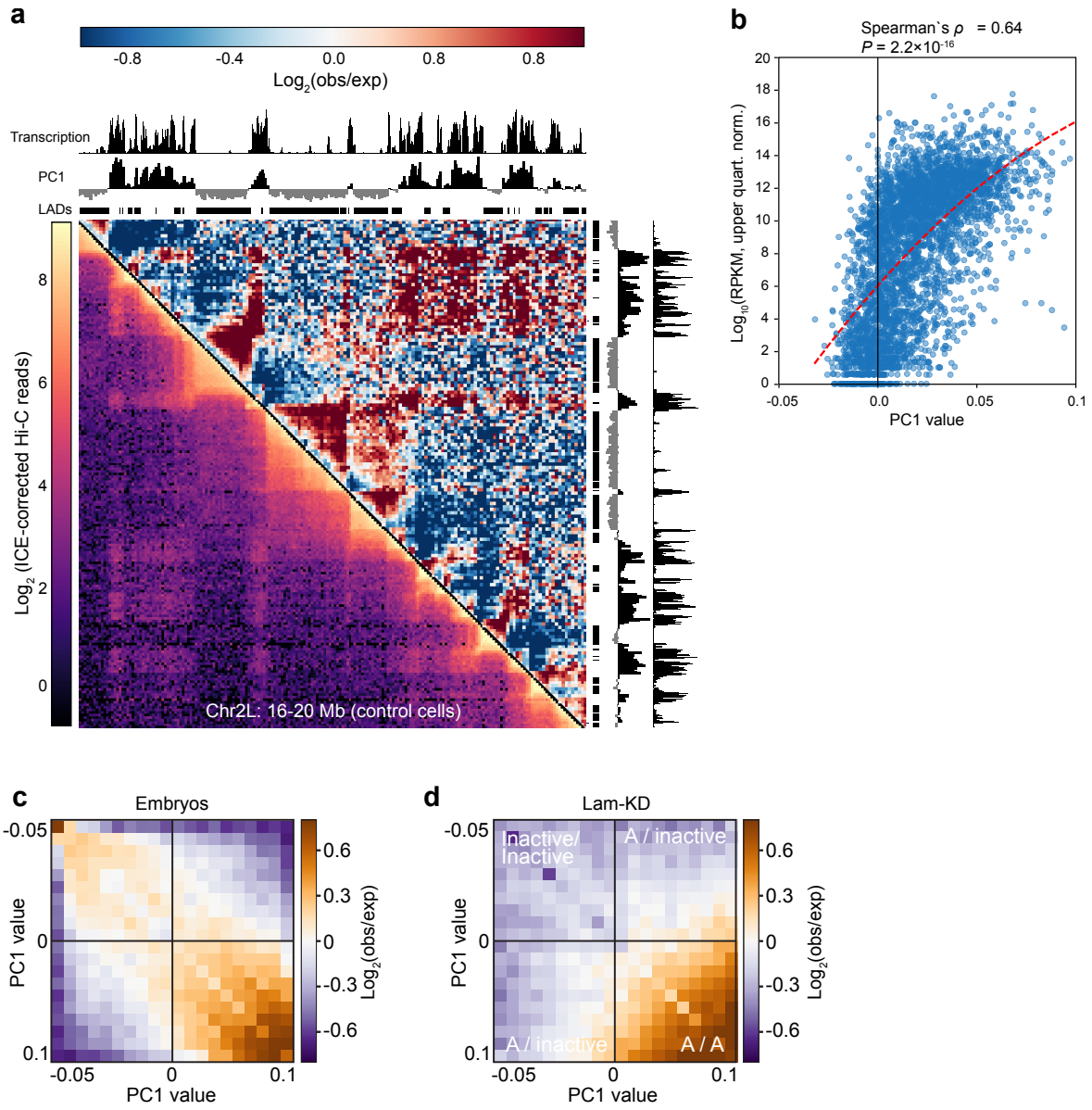
Supplementary Fig. 2 RNA-seq and ChIP-seq profiling in Lam-KD and control S2 cells. **a** Cluster analysis of biological replicates of the RNA-seq experiment. **b** Changes of gene expression for the differentially expressed genes in Lam-KD relative to control cells. The *P*-values for the comparison between “in LADs” and “in inter-LADs” groups, as well as for testing that average values in distributions exceed zero (the latter are shown below the box plots) were estimated in a Wilcoxon test. **c** A representative screenshot from the UCSC Genome Browser showing up-regulation of background transcription in LADs. **d** Cluster analysis of biological replicates of the ChIP-seq experiment.



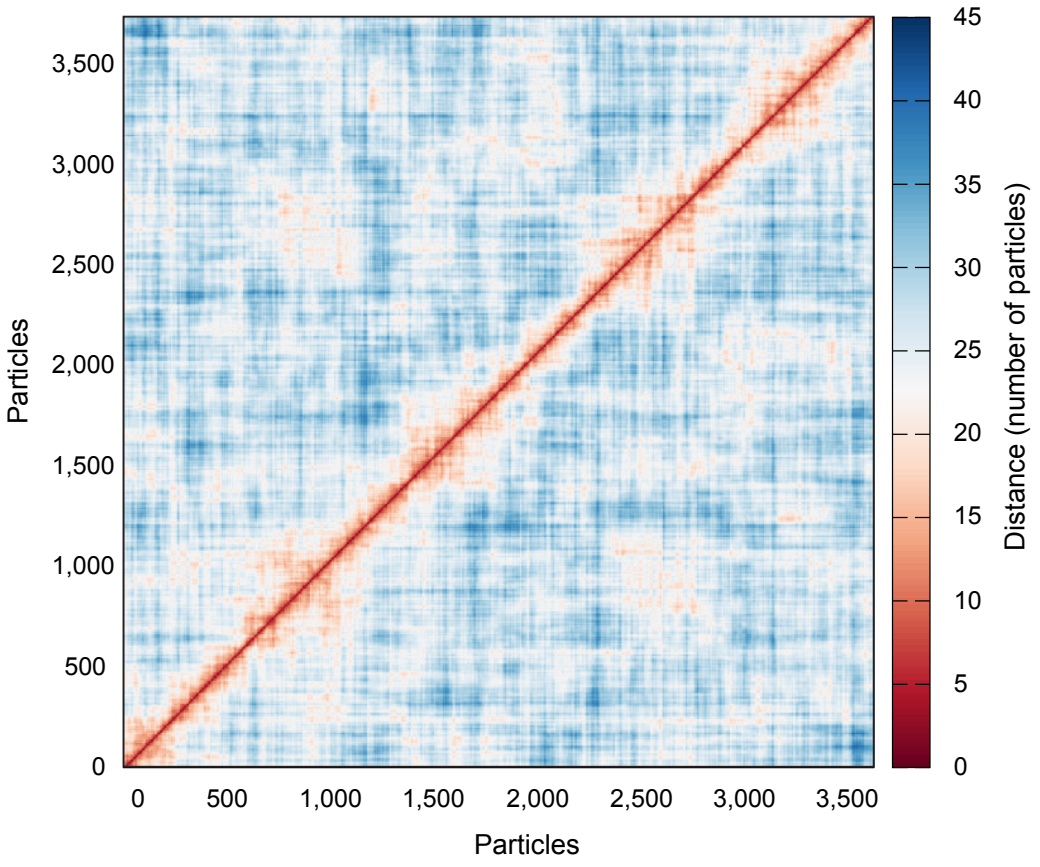
Chr		Control #1	Control #2	Lam-KD #1	Lam-KD #2
2L	Control #1		0.986	0.984	0.985
	Control #2	0.986		0.985	0.985
	Lam-KD #1	0.984	0.985		0.987
	Lam-KD #2	0.985	0.985	0.987	
2R	Control #1		0.986	0.983	0.984
	Control #2	0.986		0.984	0.984
	Lam-KD #1	0.983	0.984		0.987
	Lam-KD #2	0.984	0.984	0.987	
3L	Control #1		0.985	0.984	0.984
	Control #2	0.985		0.984	0.985
	Lam-KD #1	0.984	0.984		0.987
	Lam-KD #2	0.984	0.985	0.987	
3R	Control #1		0.987	0.984	0.984
	Control #2	0.987		0.984	0.985
	Lam-KD #1	0.984	0.984		0.987
	Lam-KD #2	0.984	0.985	0.987	
X	Control #1		0.966	0.965	0.964
	Control #2	0.966		0.964	0.965
	Lam-KD #1	0.965	0.964		0.967
	Lam-KD #2	0.964	0.965	0.967	

	Control #1	Control #2	Lam-KD #1	Lam-KD #2
Control #1		0.86	0.85	0.80
Control #2	0.86		0.82	0.80
Lam-KD #1	0.83	0.81		0.85
Lam-KD #2	0.78	0.79	0.85	

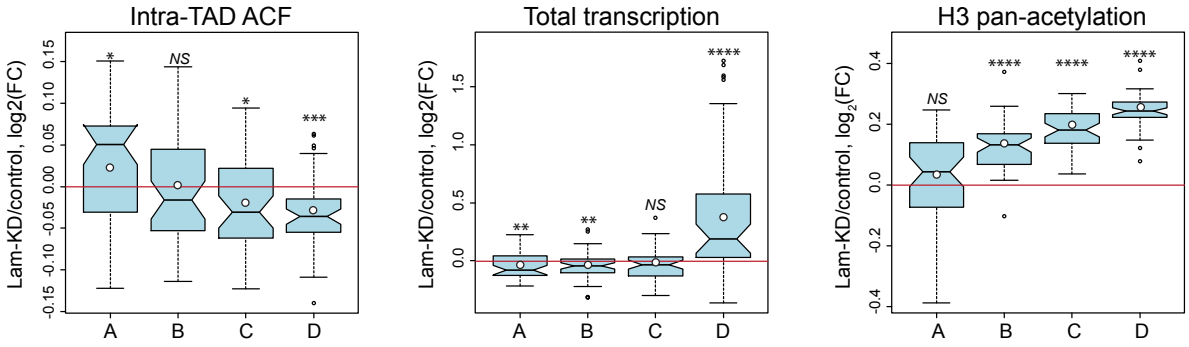
Supplementary Fig. 3 Hi-C analysis and characteristics of TADs from the four groups. **a** Comparison of heatmaps generated based on Hi-C data for control S2 cells in this work and for S2 cells in supplem. ref. 1. Pearson's correlation coefficients between ICE-corrected Hi-C matrices for each chromosome are shown to the right. **b** Pearson's correlation coefficients between ICE-corrected Hi-C matrices obtained in biological replicates of Hi-C experiments performed in this work. **c** Proportions of TADs located at the same position (±1 bin) in control and Lam-KD cells. **d** Intra-TAD ACF variability between replicates for control cells does not correlate with the LAD coverage (left panel) and the proportion of active chromatin (right panel) within TADs. **e** Intra-TAD ACF changes upon Lam-KD do not correlate with the proportion of "coral" plus "brown" chromatin types within TADs. **f** Changes of histone H3 pan-acetylation level in TADs upon Lam-KD positively correlate with the LAD coverage (left panel) and negatively correlate with the proportion of "red" plus "purple" chromatin types (supplem. ref. 2; right panel). Trend lines are in red.



Supplementary Fig. 4 Identification of chromatin compartments. **a** ICE-corrected Hi-C map (left half) and observed/expected Hi-C matrix (right half) demonstrating the presence of A compartment manifested in the enhanced interactions between the transcriptionally active loci. **b** Transcription level within genomic bin correlates positively with the PC1 value. **c** Heatmap of intra-chromosomal contacts in embryos (from supplem. ref. 3) between genomic regions as a function of their PC1 values (saddle plot); heatmap shows $\log_2$ values of contact enrichment. **d** Saddle plot for Lam-KD S2 cells.



Supplementary Fig. 5 The presence of an interacting surface does not prevent the formation of TADs in a model polymer. Distance heat map of a model polymer obtained by the averaging of distance maps from 10 independent simulation runs.



Supplementary Fig. 6 (related to Fig. 3) Strict criterion of TAD selection (i.e. all four ratios of replicates were changed in the same direction upon Lam-KD) does not affect the results of analysis. Changes of intra-TAD ACF (left panel), total transcription (middle panel) and H3 pan-acetylation (right panel) between Lam-KD and control cells in the four groups of TADs. The thick black lines and white dots represent median and average values, respectively. **** – $P < 0.0001$, *** – $P < 0.001$, ** – $P < 0.01$, * – $P < 0.05$, *NS* – non-significant difference ($P > 0.05$) in a Wilcoxon test.

Supplementary Table 1. Primers used in this study.

Gene or primer name	Direct primer (5'->3')	Reverse primer (5'->3')
Primers for RT-PCR of genes from 60D LAD		
SerT	GACATGACCACGCCCAGGTTCACG	TCCGATACTGATCTTCCGACGGGCA
CG3419	CATCGCACATATACAGCCTGCCCTTGC	TTGGTCCCGACTTCTTCACCTCGCA
CG42383	CCCGGAAACAGTTCATGGGACCAGG	GCGGAAAGGCGGTCATCAGGACAAA
prom	TGCTTCTCTGGGCTATTGGCACTCCG	GCTTGTCGCTGAACTGCTCGTCCGT
CG15873	CGGAACCCGTGGGAGAAAGCATGAA	CGCAGTCGGAAAGGGATTGGATGGT
CG15874	CAAAGGTTGCGGGAAATGGGCTTGT	GTCCAGGCTGCGGGCACTCCTCT
CG3483	AAGCAGTTCGCCCAGTTCGCCTACG	GCCACCGTGATCCTTTTGCGCTTGT
CG4563	TCAACCTGGGCGACCTGGGCTACTT	ATCTCGCATCCCGCACGCCTACC
CG13579	CTGCGAGCCCTTCTACAGCAAGCCA	GGTCGGATCGTTTAGCCGGAAGCG
CG3492	GCAGTGAAGCGCGGACCCTACACAA	TGTGGGATCGGATCTTGGGCATCG
CG3494	TGCCCAAAGATCTGCCCTTGCTCTCC	CAGCGTGGCAACTGGCGGAAGCGA
CG16837	CGTTCGTCAATCTCTTCGCCCATCG	CCAGGGTCATGTGGATTTCGCCCAT
CG13589	CACCCCTTGCTAAAATGACGAATGC	TTTCGCTGGTTCTATGAATGTGGCA
CG13590	CCTTACTGCCCCTTCTGTGCTTCGAGC	CCACCATCAACCACACCTGTCCCTATGAG
CG4589	CGGAACAATATGGGACACACGAGCAACA	CGACCCTGTCCACGACGATAACGGC
Primers for RT-PCR of genes from other LADs		
Eaat1	TACATTGGCATCATAAACTCATC	AACATCACAAGACCCAGGAC
CG30395	AGTTATACTTCAATGCACCTGTTT	ATGGGAGTCTTCGGGCTTAC
CG34391	GGCTAATGCTGCTTGAATGC	TGTGGGTAACCTGCTTGGAT
CG5162	ATCCTAACACCTACTGGCATAA	CAGGGTATCAACGAAACGAG
CG34370	AATCATCAGCCAATTCTAACTACC	TCTTCCTTAGCATCGCCCAC
Rim	AGCCGACACCATTACCACCT	CGAATGTTTGTCAGAATCCCT
wat	CGGCACCAGAGCTAATGTAT	CACCCTGAACACCCTTACGC
beat-Va	ATCCGTCACAAACAGAGCAT	TCTTTGGGGAAAACAACATC
Sls	CCACCATGATGTTGTTGCAC	CACTTCCGCTACCATCCATA
Byn	ACATTGGCGCTCACTATTTG	GAGGCACTGATCTTCACGAC
Goe	CTGTAGGACGACCAGAACCC	CATGATCCCACTAATTTGAGC
CG31814	CATTAGAGCATCTCGACCCA	GGGAATTGAAAAGGACTAAGTAAA
Fili	GGCAATGTGATGAGCGAACT	TGATTAAGGGCAGATATGAAAA
Mb1	TCAGTTATTGATAAATGGACGCA	AGTGGATAGCGGATGGAATG
Primers for 2L 16,964,000-16,982,000 amplification (green FISH probe)		
1	CCTCCATTTCCACCCACAGTTTCCCA	GCCCAAGTGCCACGAGCCTCAAATAA
2	TTATTTGAGGCTCGTGGCACTTGGGC	GCGATTTTCAGGACTCGGGGACTGG
3	CCAGTCCCCGAGTCCTGAAAATCGC	TTTTTGCTTTGACAACCCTGCCGCA
4	TGCGGCAGGGTTGTCAAAGCAAAAA	CCTTGTCCAGAGGATAAAAAACGGTGCCC
Primers for 2L 17,310,000-17,328,000 amplification (red FISH probe)		
1	GCCCACCACCCACTTTTTGGCTTTG	CCCTCTGACCCAACAGCACGTTTTTCA
2	GGGAGGGCGAACATTGTGGGATCAG	TTTGTCATTGTGGGTGCGTTGCTGC
3	GCAGCAACGCACCCACAATGACAAA	GAGAGCGAGCAAAAAGGCCGTGGAA
4	TTCCACGGCCTTTTTGCTCGCTCTC	GGAGCTCTTGTGAGGCCCGAACCAA
Primers for lamin Dm0 dsRNA preparation		
Direct primer	GAATTAATACGACTCACTATAGGGAGAATGTCGAGCAAATCCCGACGT	
Reverse primer	GAATTAATACGACTCACTATAGGGAGAGCGACTGCTTCAACTTGGCATC	
Primers for LBR dsRNA preparation		
Direct primer	GAATTAATACGACTCACTATAGGGAGACCCAGTCCAAGCAGCCCAGCC	
Reverse primer	GAATTAATACGACTCACTATAGGGAGAGGCAAAGGCACCCACCACTCGT	

Supplementary References

1. Ramírez, F. *et al*. High-affinity sites form an interaction network to facilitate spreading of the MSL complex across the X chromosome in Drosophila. *Mol. Cell* **60,** 146–162 (2015).
2. Kharchenko, P. V. *et al*. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471,** 480–485 (2011).
3. Sexton, T. *et al*. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148,** 458–472 (2012).

Chapter 5

A machine learning framework for the prediction of chromatin folding in *Drosophila* using epigenetic features

On a local scale, *Drosophila* chromatin is constituted of Topologically Associating Domains, or TADs, visible in bulk Hi-C. TAD is an insulated neighborhood of the genome, with more contacts within it than with surrounding regions. The boundaries of these neighborhoods are associated with the binding of insulating proteins (proteins that lead to the insulation effect when bound to DNA). However, a previous study by our group [Ulianov et al., 2016] has demonstrated that they are not the primary factors demarcating the boundaries.

Thus, it was important to study the problem of TAD boundaries-forming factors in depth in bulk Hi-C before

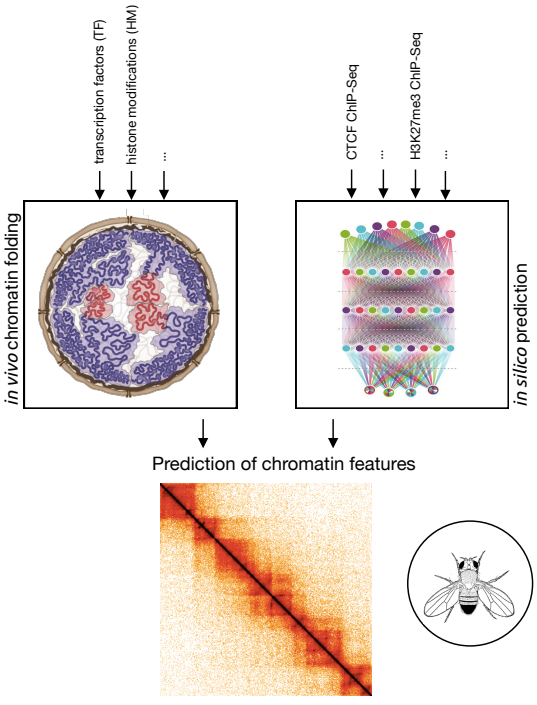


Figure 5-1: Black-box models for prediction of chromatin features of *Drosophila*. Inputs for the models serve as a measure of binding of transcription factors and histone modifications, factors of epigenetics that can drive the structure formation *in vivo*. These mechanisms are mimicked by black-box models *in silico*.

studying it at the level of individual
cells. We applied an interpretable machine learning model to predict TADs in bulk data based on epigenetic features, including insulators and histone modifications. The results suggested that a protein of the insulator type (Chriz) and active histone modification (H3K4me3) are the most relevant for the prediction across multiple cell types of *Drosophila*.

I want to emphasize that these results further guided the search of factors responsible for TADs formation in single cells in Chapter 7. Although the model was developed by the first author of this study, I significantly contributed to the text and prepared some of the figures of this paper. I also designed some of the computational experiments which were then implemented by the first author.

PeerJ Computer Science

A machine learning framework for the prediction of chromatin folding in *Drosophila* using epigenetic features

Michal B. Rozenwald[1], Aleksandra A. Galitsyna[2], Grigory V. Sapunov[1,3], Ekaterina E. Khrameeva[2] and Mikhail S. Gelfand[2,4]

[1] Faculty of Computer Science, National Research University Higher School of Economics, Moscow, Russia
[2] Skolkovo Institute of Science and Technology, Moscow, Russia
[3] Intento, Inc., Berkeley, CA, USA
[4] A.A. Kharkevich Institute for Information Transmission Problems, RAS, Moscow, Russia

ABSTRACT

Technological advances have lead to the creation of large epigenetic datasets, including information about DNA binding proteins and DNA spatial structure. Hi-C experiments have revealed that chromosomes are subdivided into sets of self-interacting domains called Topologically Associating Domains (TADs). TADs are involved in the regulation of gene expression activity, but the mechanisms of their formation are not yet fully understood. Here, we focus on machine learning methods to characterize DNA folding patterns in *Drosophila* based on chromatin marks across three cell lines. We present linear regression models with four types of regularization, gradient boosting, and recurrent neural networks (RNN) as tools to study chromatin folding characteristics associated with TADs given epigenetic chromatin immunoprecipitation data. The bidirectional long short-term memory RNN architecture produced the best prediction scores and identified biologically relevant features. Distribution of protein Chriz (Chromator) and histone modification H3K4me3 were selected as the most informative features for the prediction of TADs characteristics. This approach may be adapted to any similar biological dataset of chromatin features across various cell lines and species. The code for the implemented pipeline, Hi-ChIP-ML, is publicly available: https://github.com/MichalRozenwald/Hi-ChIP-ML

Subjects Bioinformatics, Computational Biology, Molecular Biology, Data Mining and Machine Learning, Data Science
Keywords Topologically Associating Domains (TADs), Recurrent Neural Networks (RNN), Hi-C experiments, Linear Regression, Gradient Boosting, Chromatin, DNA folding patterns, Machine Learning

INTRODUCTION

Machine learning has proved to be an essential tool for studies in the molecular biology of the eukaryotic cell, in particular, the process of gene regulation (*Eraslan et al., 2019*; *Zeng, Wang & Jiang, 2020*). Gene regulation of higher eukaryotes is orchestrated by two primary interconnected mechanisms, the binding of regulatory factors to the promoters and enhancers, and the changes in DNA spatial folding. The resulting binding patterns and chromatin structure represent the epigenetic state of the cells. They can be assayed

Submitted 28 August 2020
Accepted 30 September 2020
Published 30 November 2020

Corresponding authors
Michal B. Rozenwald, mbrozenvald@edu.hse.ru, michal.rozenwald@gmail.com
Mikhail S. Gelfand, m.gelfand@skoltech.ru, michal.rozenwald@gmail.com

Academic editor
Alexander Bolshoy

Additional Information and Declarations can be found on page 15

DOI **10.7717/peerj-cs.307**

How to cite this article Rozenwald MB, Galitsyna AA, Sapunov GV, Khrameeva EE, Gelfand MS. 2020. A machine learning framework for the prediction of chromatin folding in *Drosophila* using epigenetic features. *PeerJ Comput. Sci.* **6**:e307 http://doi.org/10.7717/peerj-cs.307

by high-throughput techniques, such as chromatin immunoprecipitation (*Ren et al., 2000*; *Johnson et al., 2007*) and Hi-C (*Lieberman-Aiden et al., 2009*). The epigenetic state is tightly connected with inheritance and disease (*Lupiáñez, Spielmann & Mundlos, 2016*; *Yuan et al., 2018*; *Trieu, Martinez-Fundichely & Khurana, 2020*). For instance, disruption of chromosomal topology in humans affects gliomagenesis and limb malformations (*Krijger & De Laat, 2016*). However, the details of underlying processes are yet to be understood.

The study of Hi-C maps of genomic interactions revealed the structural and regulatory units of eukaryotic genome, topologically associating domains, or TADs. TADs represent self-interacting regions of DNA with well-defined boundaries that insulate the TAD from interactions with adjacent regions (*Lieberman-Aiden et al., 2009*; *Dixon et al., 2012*; *Rao et al., 2014*). In mammals, the boundaries of TADs are defined by the binding of insulator protein CTCF (*Rao et al., 2014*). However, *Drosophila* CTCF homolog is not essential for the formation of TAD boundaries (*Wang et al., 2018*). Contribution of CTCF to the boundaries was detected in neuronal cells, but not in embryonic cells of *Drosophila* (*Chathoth & Zabet, 2019*). At the same time, up to eight different insulator proteins have been proposed to contribute to the formation of TADs boundaries (*Ramírez et al., 2018*).

Ulianov et al. (2016) demonstrated that active transcription plays a key role in the *Drosophila* chromosome partitioning into TADs. Active chromatin marks are preferably found at TAD borders, while repressive histone modifications are depleted within inter-TADs. Thus, histone modifications instead of insulator binding factors might be the main TAD-forming factors in this organism.

To determine factors responsible for the TAD boundary formation in *Drosophila*, *Ulianov et al. (2016)* utilized machine learning techniques. For that, they formulated a classification task and used a logistic regression model. The model input was a set of ChIP-chip signals for a genomic region, and the output, a binary value indicating whether the region was located at the boundary or within a TAD. Similarly, *Ramírez et al. (2018)* demonstrated the effectiveness of the lasso regression and gradient boosting for the same task.

However, this approach has two substantial limitations. First, the prediction of TAD state as a categorical output depends on the TAD calling procedure. It requires setting a threshold for the TAD boundary definition and it is insensitive to sub-threshold boundaries.

Alternatively, the TAD status of a region may be derived from a Hi-C map either by calculation of local characteristics of TADs such as Insulation Score (*Crane et al., 2015*), D-score (*Stadhouders et al., 2018*), Directionality Index (*Dixon et al., 2012*), or by dynamic programming methods, such as Armatus (*Filippova et al., 2014*). Methods assessing local characteristics of TADs result in assigning a continuous score to genomic bins along the chromosome. Dynamic programming methods are typically not anchored to a local genomic region and consider Hi-C maps of whole chromosomes. The calculation of *transitional gamma* has the advantages of both approaches (*Ulianov et al., 2016*). It runs dynamic programming for whole-chromosome data for multiple parameters and assesses the score for each genomic region.

The second limitation is that regression and gradient boosting in *Ulianov et al. (2016)* and *Ramírez et al. (2018)* account for the features of a given region of the genome, but

ignore the adjacent regions. Such contextual information might be crucial for the TAD status in *Drosophila*.

For a possible solution, one may look at instructive examples of other chromatin architecture problems, such as improvement of Hi-C data resolution (*Gong et al., 2018*; *Schwessinger et al., 2019*; *Li & Dai, 2020*), inference of chromatin structure (*Cristescu et al., 2018*; *Trieu, Martinez-Fundichely & Khurana, 2020*), prediction of genomic regions interactions (*Whalen, Truty & Pollard, 2016*; *Zeng, Wu & Jiang, 2018*; *Li, Wong & Jiang, 2019*; *Fudenberg, Kelley & Pollard, 2019*; *Singh et al., 2019*; *Jing et al., 2019*; *Gan, Li & Jiang, 2019*; *Belokopytova et al., 2020*), and, finally, TAD boundaries prediction in mammalian cells (*Gan et al., 2019*; *Martens et al., 2020*).

The machine learning approaches used in these works include generalized linear models (*Ibn-Salem & Andrade-Navarro, 2019*), random forest (*Bkhetan & Plewczynski, 2018*; *Gan et al., 2019*), other ensemble models (*Whalen, Truty & Pollard, 2016*), and neural networks: multi-layer perceptron (*Gan et al., 2019*), dense neural networks (*Zeng, Wu & Jiang, 2018*; *Farré et al., 2018*; *Li, Wong & Jiang, 2019*), convolutional neural networks (*Schreiber et al., 2017*), generative adversarial networks (*Liu, Lv & Jiang, 2019*), and recurrent neural networks (*Cristescu et al., 2018*; *Singh et al., 2019*; *Gan, Li & Jiang, 2019*).

Among these methods, recurrent neural networks (RNNs) provide a comprehensive architecture for analyzing sequential data (*Graves, Jaitly & Mohamed, 2013*), due to the temporal modeling capabilities. A popular implementation of RNN long short-term memory (LSTM) models (*Hochreiter & Schmidhuber, 1997*) creates informative statistics that provide solutions for complex long-time-lag tasks (*Graves, 2012*). Thus, the application of LTSM RNNs to problems with sequential ordering of a target, such as DNA bins characteristics, is a promising approach. Moreover, this feature is particularly relevant for the TAD boundary prediction in *Drosophila*, where the histone modifications of extended genomic regions govern the formation of boundaries (*Ulianov et al., 2016*).

Here, we analyze the epigenetic factors contributing to the TAD status of the genomic regions of *Drosophila*. As opposed to previous approaches, we incorporate information about the region context on two levels. First, we utilize the context-aware TAD characteristic *transitional gamma*. Second, we use the advanced method of recurrent neural network that preserves the information about features of adjacent regions.

MATERIALS AND METHODS

Data

Hi-C datasets for three cultured *Drosophila melanogaster* cell lines were taken from *Ulianov et al. (2016)*. Cell lines Schneider-2 (S2) and Kc167 from late embryos and DmBG3-c2 (BG3) from the central nervous system of third-instar larvae were analysed. The *Drosophila* genome (dm3 assembly) was binned at the 20-kb resolution resulting in 5950 sequential genomic regions of equal size. Each bin was described by the start coordinate on the chromosome and by the signal from a set of ChIP-chip experiments. The ChIP-chip data were obtained from the modENCODE database (*Waterston et al., 2009*) and processed as in *Ulianov et al. (2016)*.

As chromatin architecture is known to be correlated with epigenetic characteristics in *Drosophila* (*Ulianov et al., 2016*; *Hug et al., 2017*; *Ramírez et al., 2018*), we selected two sets of epigenetic marks, i.e., transcription factors (TF), and insulator protein binding sites, and histone modifications (HM), for further analysis. The first set included five features (Chriz, CTCF, Su(Hw), H3K27me3, H3K27ac), which had been reported as relevant for TAD formation in previous studies (*Ulianov et al., 2016*). The second set contained eighteen epigenetic marks in total, extending the first set with thirteen potentially relevant features chosen based on the literature (RNA polymerase II, BEAF-32, GAF, CP190, H3K4me1, H3K4me2, H3K4me3, H3K9me2, H3K9me3, H3K27me1, H3K36me1, H3K36me3, H4K16ac). To normalize the input data, we subtracted the mean from each value and then scaled it to the unit variance using the preprocessing scale function of the Sklearn Python library (*Pedregosa et al., 2011*). We standardized each feature independently; the mean and variance were calculated per each feature (chromatin mark) separately across all input objects (bins), see Fig. S2. For the full list of chromatin factors and their modENCODE IDs, see Table S1.

Target value

TADs are calculated based on Hi-C interactions matrix. As a result of TAD calling algorithm, TADs are represented as a segmentation of the genome into discrete regions. However, resulting segmentation typically depends on TAD calling parameters. In particular, widely used TAD segmentation software Armatus (*Filippova et al., 2014*) annotates TADs for a user-defined scaling parameter *gamma*. Gamma determines the average size and the number of TADs produced by Armatus on a given Hi-C map.

Following *Ulianov et al. (2016)*, we avoided the problem of selection of a single set of parameters for TADs annotation and calculated the local characteristic of TAD formation of the genome, namely, *transitional gamma*. The calculation of transitional gamma includes the TAD calling for a wide range of reasonable parameters gamma and selection of characteristic gamma for each genomic locus. This procedure is briefly described below.

When parameter gamma is fixed, Armatus annotates each genomic bin as a part of a TAD, inter-TAD, or TAD boundary. The higher the gamma value is used in Armatus, the smaller on average the TADs sizes are. We perform the TAD calling with Armatus for a set of parameters and characterize each bin by transitional gamma at which this bin switches from being a part of a TAD to being a part of an inter-TAD or a TAD boundary. We illustrate the TADs annotation and calculation of transitional gamma in Figs. 1A–1C.

Whole-genome Hi-C maps of *Drosophila* cells were collected from *Ulianov et al. (2016)* and processed using Armatus with a gamma ranging from 0 to 10 with a step of 0.01. We then calculated the transitional gamma for each bin. The resulting distribution of values can be found in Fig. 1D. We note that the value 10 is corresponding to the bins that form TAD regions that we have never observed as being TAD boundary or inter-TAD. These bins might switch from TADs with the further increase of gamma. However, they represent a minor fraction of the genome corresponding to strong inner-TAD bins.

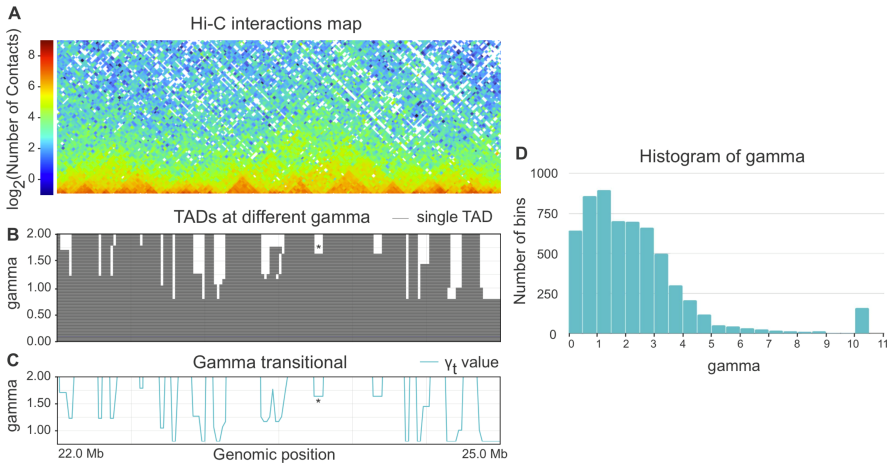


Figure 1 **(A–C) Example of annotation of chromosome 3R region by transitional gamma. For a given Hi-C matrix of Schneider-2 cells (A), TAD segmentations (B) are calculated by Armatus for a set of gamma values (from 0 to 10, a step of 0.01). Each line in B represents a single TAD. Then gamma transitional (C) is calculated for each genomic region as the minimal value of gamma where the region becomes inter-TAD or TAD boundary. The blue line in C represents the transitional gamma value for each genomic bin. The plots (B) and (C) are limited by gamma 2 for better visualization, although they are continued to the value of 10. Asterisk (*) denotes the region with gamma transitional of 1.64, the minimal value of gamma, where the corresponding region transitions from TAD to inter-TAD. (D) The histogram of the target value transitional gamma for Schneider-2 cell line. Note the peak at 10.**

Full-size DOI: 10.7717/peerjcs.307/fig-1

Problem statement

To avoid ambiguity, we formally state our machine learning problem:

- **objects** are genomic bins of 20-kb length that do not intersect,
- **input features** are the measurements of chromatin factors' binding,
- **target value** is the transitional gamma, which characterizes the TAD status of the region and, thus, the DNA folding,
- **objective** is to predict the value of transitional gamma and to identify which of the chromatin features are most significant in predicting the TAD state.

Selection of loss function

The target, transitional gamma, is a continuous variable ranging from 0 to 10, which yields a regression problem (*Yan & Su, 2009*). The classical optimization function for the regression is *Mean Square Error (MSE)*, instead of precision, recall or accuracy, as for binary variables. However, the distribution of the target in our problem is significantly unbalanced (see Fig. 1D) because the target value of most of the objects is in the interval between 0 and 3. Thus, the contribution of the error on objects with a high true target value may be also high in the total score when using MSE.

We note that the biological nature of genomic bins with high transitional gamma is different from other bins. Transitional gamma equal to 10 means that the bin never transformed from being a part of a TAD to an inter-TAD or TAD boundary. To solve this

contradiction, we have introduced a custom loss function called modified *weighted Mean Square Error (wMSE)*. It might be reformulated as MSE multiplied by the weight (penalty) of the error, depending on the true value of the target variable.

$$wMSE = \frac{1}{N}\sum_{i=1}^{N}(y_{\text{true}_i} - y_{\text{pred}_i})^2 \frac{\alpha - y_{\text{true}_i}}{\alpha},$$

where N is the number of data points, y_{true_i} is the true value for data point number i, y_{pred_i} is the predicted value for data point number i. Here, α is the maximum value of y_{true} increased by 1 to avoid multiplying the error by 0. The maximum value of the transitional gamma in our dataset is 10, thus in our case, α equals 11. With wMSE as a loss function, the model is penalized less for errors on objects with high values of transitional gamma.

Machine learning models

To explore the relationships between the 3D chromatin structure and epigenetic data, we built linear regression (LR) models, gradient boosting (GB) regressors, and recurrent neural networks (RNN). The LR models were additionally applied with either L1 or L2 regularization and with both penalties. For benchmarking we used a constant prediction set to the mean value of the training dataset.

Due to the DNA linear connectivity, our input bins are sequentially ordered in the genome. Neighboring DNA regions frequently bear similar epigenetic marks and chromatin properties (*Kharchenko et al., 2011*). Thus, the target variable values are expected to be vastly correlated. To use this biological property, we applied RNN models. In addition, the information content of the double-stranded DNA molecule is equivalent if reading in forward and reverse direction. In order to utilize the DNA linearity together with equivalence of both directions on DNA, we selected the bidirectional long short-term memory (biLSTM) RNN architecture (*Schuster & Paliwal, 1997*). The model takes a set of epigenetic properties for bins as input and outputs the target value of the *middle bin*. The middle bin is an object from the input set with an index i, where i equals to the floor division of the input set length by 2. Thus, the transitional gamma of the middle bin is being predicted using the features of the surrounding bins as well. The scheme of this model is presented in Fig. 2.

We exploited the following parameters of the biLSTM RNN in our experiments.

The sequence length of the RNN input objects is a set of consecutive DNA bins with fixed length that was varied from 1 to 10 (*window size*).

The numbers of LSTM units that we tested for were 1, 4, 8, 16, 32, 64, 128, 256, 512.

The weighted Mean Square Error loss function was chosen and models were trained with a stochastic optimizer Adam (*Kingma & Ba, 2014*).

Early stopping was used to automatically identify the optimal number of training epochs. The dataset was randomly split into three groups: train dataset 70%, test dataset 20%, and 10% data for validation.

To explore the importance of each feature from the input space, we trained the RNNs using only one of the epigenetic features as input. Additionally, we built models in which columns from the feature matrix were one by one replaced with zeros, and all other features

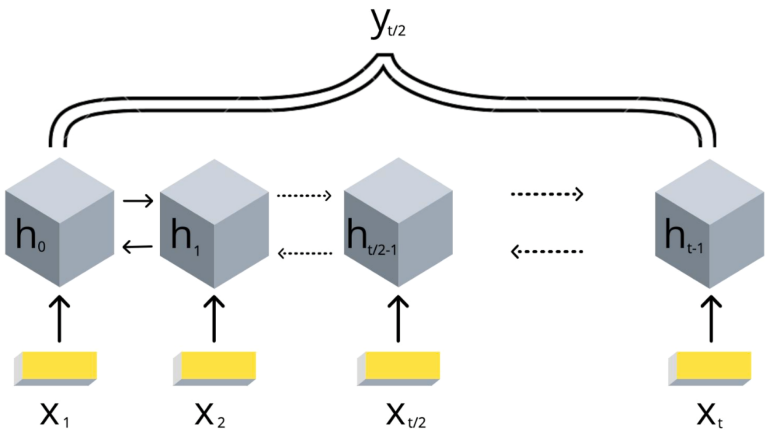


Figure 2 Scheme of the implemented bidirectional LSTM recurrent neural networks with one output. The values of $\{x_1, .., x_t\}$ are the DNA bins with input window size t, $\{h_1, .., h_t\}$ are the hidden states of the RNN model, $y_{t/2}$ represents the corresponding target value transitional gamma of the middle bin $x_{t/2}$. Note that each bin xi is characterized by a vector of chromatin marks ChIP-chip data.

Full-size DOI: 10.7717/peerjcs.307/fig-2

were used for training. Further, we calculated the evaluation metrics and checked if they were significantly different from the results obtained while using the complete set of data.

RESULTS

Chromatin marks are reliable predictors of the TAD state

First, we assessed whether the TAD state could be predicted from the set of chromatin marks for a single cell line (Schneider-2 in this section). The classical machine learning quality metrics on cross-validation averaged over ten rounds of training demonstrate strong quality of prediction compared to the constant prediction (see Table 1).

High evaluation scores prove that the selected chromatin marks represent a set of reliable predictors for the TAD state of *Drosophila* genomic region. Thus, the selected set of 18 chromatin marks can be used for chromatin folding patterns prediction in *Drosophila*.

The quality metric adapted for our particular machine learning problem, wMSE, demonstrates the same level of improvement of predictions for different models (see Table 2). Therefore, we conclude that wMSE can be used for downstream assessment of the quality of the predictions of our models.

These results allow us to perform the parameter selection for linear regression (LR) and gradient boosting (GB) and select the optimal values based on the wMSE metric. For LR, we selected alpha of 0.2 for both L1 and L2 regularizations.

Gradient boosting outperforms linear regression with different types of regularization on our task. Thus, the TAD state of the cell is likely to be more complicated than a linear combination of chromatin marks bound in the genomic locus. We used a wide range of variable parameters such as the number of estimators, learning rate, maximum depth of the individual regression estimators. The best results were observed while setting the

Table 1 Evaluation of classical machine learning scores for all models, based on 5-features and 18-features inputs.

Model type	MSE Train	MSE Test	MAE Train	MAE Test	R^2
Constant prediction	3.71	3.72	1.36	1.31	0
Using 5 features:					
LR + L1	2.91	2.91	1.11	1.11	0.21
LR + L2	2.92	2.93	1.12	1.12	0.21
LR + L1 + L2	2.86	2.87	1.11	1.11	0.23
GB-250	2.45	2.67	1.10	1.11	0.28
biLSTM RNN	2.36	2.90	0.92	1.01	0.33
Using 18 features:					
LR + L1	2.77	2.77	1.09	1.09	0.25
LR + L2	2.69	2.69	1.08	1.08	0.27
LR + L1 + L2	2.67	2.68	1.07	1.07	0.28
GB-250	2.22	2.53	1.06	1.07	0.32
biLSTM RNN	**2.03**	**2.45**	**0.85**	**0.90**	**0.43**

Table 2 Weighted MSE of all models, based on 5-features and 18-features inputs.

	5 features		18 features	
	Train	Test	Train	Test
Constant prediction	1.61	1.62	1.61	1.62
Linear Regression	1.20	1.20	1.13	1.14
Linear regression + L1	1.17	1.17	1.12	1.12
Linear regression + L2	1.18	1.19	1.11	1.12
Linear regression + L1 + L2	1.17	1.16	1.11	1.11
Grad boosting 100 estimators	1.11	1.13	1.08	1.10
Grad boosting 250 estimators	1.06	1.11	0.95	1.07
biLSTM 64 units & 6 bins	**0.83**	**0.88**	**0.79**	**0.84**

'n_estimators': 100, 'max_depth': 3 and n_estimators': 250, 'max_depth': 4, both with 'learning_rate': 0.01. The scores are presented in Tables 1 and 2.

The context-aware prediction of TAD state is the most reliable

The alternative model that we studied was biLSTM neural network, which provides explicit accounting for linearly ordered bins in the DNA molecule.

We have investigated the hyperparameters set for biLSTM and assessed the wMSE on various input window sizes and numbers of LSTM units. As we demonstrate in Fig. 3, the optimal sequence length is equal to the input window size 6 and 64 LSTM units. This result has a potential biological interpretation as the typical size of TADs in *Drosophila*, being around 120 kb at 20-kb resolution Hi-C maps which equals to 6 bins.

The incorporation of sequential dependency improved the prediction significantly, as demonstrated by the best quality scores achieved by the biLSTM (Table 2). The selected

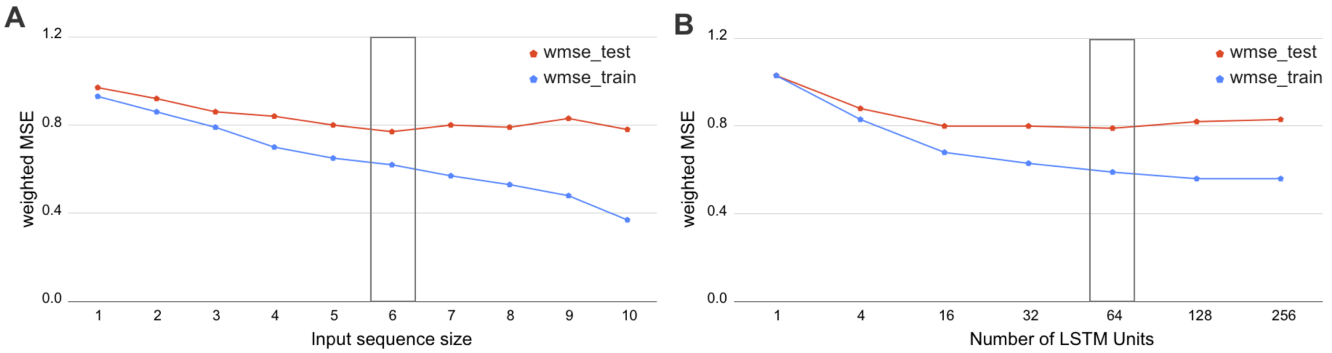


Figure 3 Selection of the biLSTM parameters. Weighted MSE scores for the train and test datasets are presented. (A) Results of RNN with 64 units for different sizes of sequence length. The sequence size corresponds to the input window size of the RNN or number of bins used together as an input sequence for the neural network. (B) Results of RNN with an input sequence of six bins for the different number of LSTM units. The box highlights the best scores. The biLSTM with six input bins and 64 LSTM units was used throughout this study if not specified otherwise.

Full-size DOI: 10.7717/peerjcs.307/fig-3

biLSTM with the best hyperparameters set performed two times better than the constant prediction and outscored all trained LR and GB models, see Tables 1 and 2. We note that the proposed biLSTM model does not take into account the target value of the neighboring regions, both while training and predicting. Our model uses the input values (chromatin marks) solely for the whole window and target values for the central bin in the window for training and assessment of validation results. Thus, we conclude that biLSTM was able to capture and utilize the sequential relationship of the input objects in terms of the physical distance in the DNA.

Reduced set of chromatin marks is sufficient for a reliable prediction of the TAD state in *Drosophila*

Next, we used an opportunity to analyse feature importance and select the set of factors most relevant for chromatin folding. For an initial analysis, we selected a subset of five chromatin marks that we considered important based on the literature (two histone marks and three potential insulator proteins, 5-features model).

The 5-features model performed slightly worse than the initial 18-features model (see Tables 1 and 2). The difference in quality scores is rather small, supporting the selection of these five features as biologically relevant for TAD state prediction.

We note that the small impact of shrinking of the number of predictors might indicate the high correlation between chromatin features. This is in line with the concept of chromatin states when several histone modifications and other chromatin factors are responsible for a single function of DNA region, such as gene expression (*Filion et al., 2010*; *Kharchenko et al., 2011*).

Feature importance analysis reveals factors relevant for chromatin folding into TADs in *Drosophila*

We have evaluated the weight coefficients of the linear regression because the large weights strongly influence the model prediction. Chromatin marks prioritization of 5-features LR

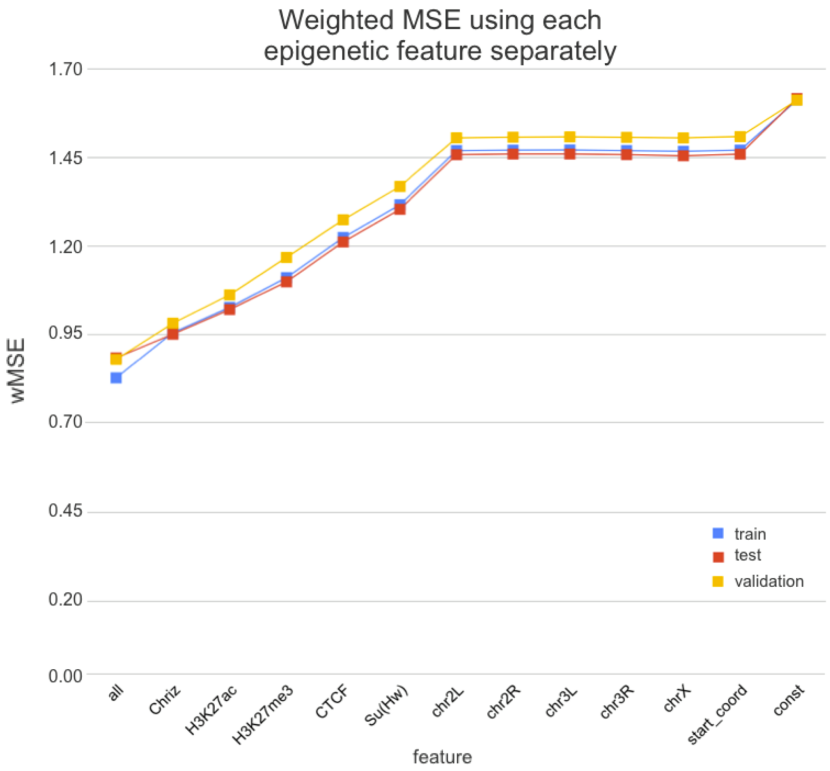


Figure 4 Weighted MSE using one feature for each input bin in the biLSTM RNN. The first mark (‘*all*’) corresponds to scores of NNs using the first dataset of chromatin marks features together, the last mark (‘*const*’) represents wMSE using constant prediction. Note that the lower the wMSE value the better the quality of prediction.

Full-size DOI: 10.7717/peerjcs.307/fig-4

model demonstrated that the most valuable feature was Chriz, while the weights of Su(Hw) and CTCF were the smallest. As expected, Chriz factor was the top in the prioritization of the 18-features LR model. However, the next important features were histone marks H3K4me1 and H3K27me1, supporting the hypothesis of histone modifications as drivers of TAD folding in *Drosophila*.

We used two approaches for the feature selection of RNN: use-one feature and drop-one feature. When each single chromatin mark was used as the only feature of each bin of the RNN input sequence for training, the best scores were obtained for Chriz and H3K4me2 (Figs. 4, 5 and 6), similarly to the LR models results. When we dropped out one of the five features, we got scores that are almost equal to the wMSE using the full dataset together. This does not hold for experiment with excluded Chriz, where wMSE increases. These results align with the outcome of use-one approach and while applying LR models.

Similar results were obtained while using the broader dataset. The results of applying the same approach of omitting each feature one by one using the second dataset of features allowed the evaluation of the biological impact of the features. The corresponding wMSE

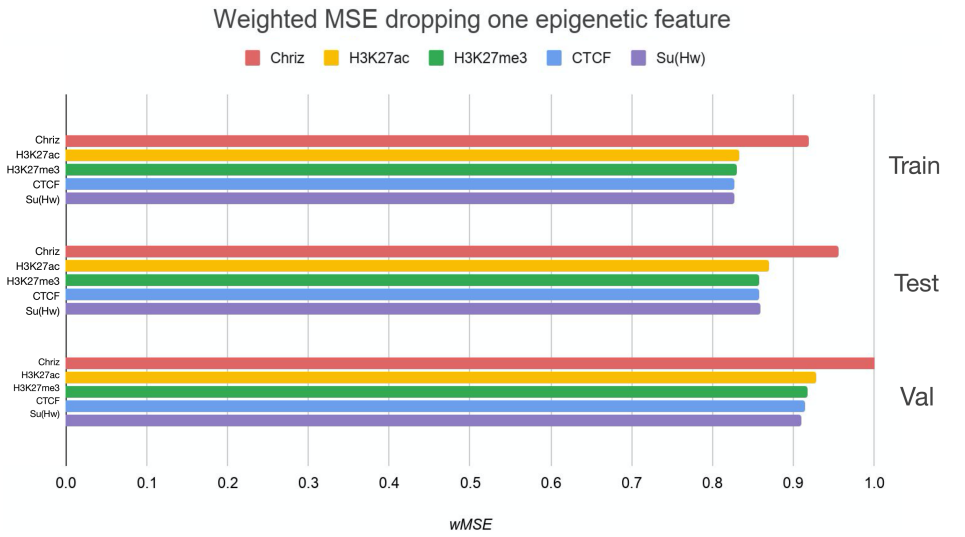


Figure 5 Weighted MSE using four out of five chromatin marks features together as the biLSTM RNN input. Each colour corresponds to the feature that was excluded from the input. Note that the model is affected the most when Chriz factor is dropped from features.

Full-size DOI: 10.7717/peerjcs.307/fig-5

scores are presented in Fig. 6 as well as the result of training the model on all features together.

The results of omitting each feature one by one while using the second dataset of features are almost identical as we expected. It could be explained by the fact that most of the features are strongly correlated.

TAD state prediction models are transferable between cell lines of *Drosophila*

In order to explore the transferability of the results between various *Drosophila* cell lines, we have applied the full pipeline for Schneider-2 and Kc167 cells from late embryos and DmBG3-c2 (BG3) cells from the central nervous system of third-instar larvae. Across all cell lines, the biLSTM model has gained the best evaluation scores (Table 3). On average, the smallest errors were produced on the test set of the BG3 cell line.

Notably, the selected top features are robust between cell lines. The results of the usage of each feature separately for each of the cell lines can be found in Fig. S1. Chriz was identified as the most influencing feature for Schneider-2 and BG3 while being in the top four features for Kc167. Histone modifications H3K4me2 and H3K4me3 gain very high scores on each dataset. However, CTCF was found in the top of the influencing chromatin marks only on the Kc167, while insulator Su(Hw) constantly scores almost the worst wMSE across all cell lines.

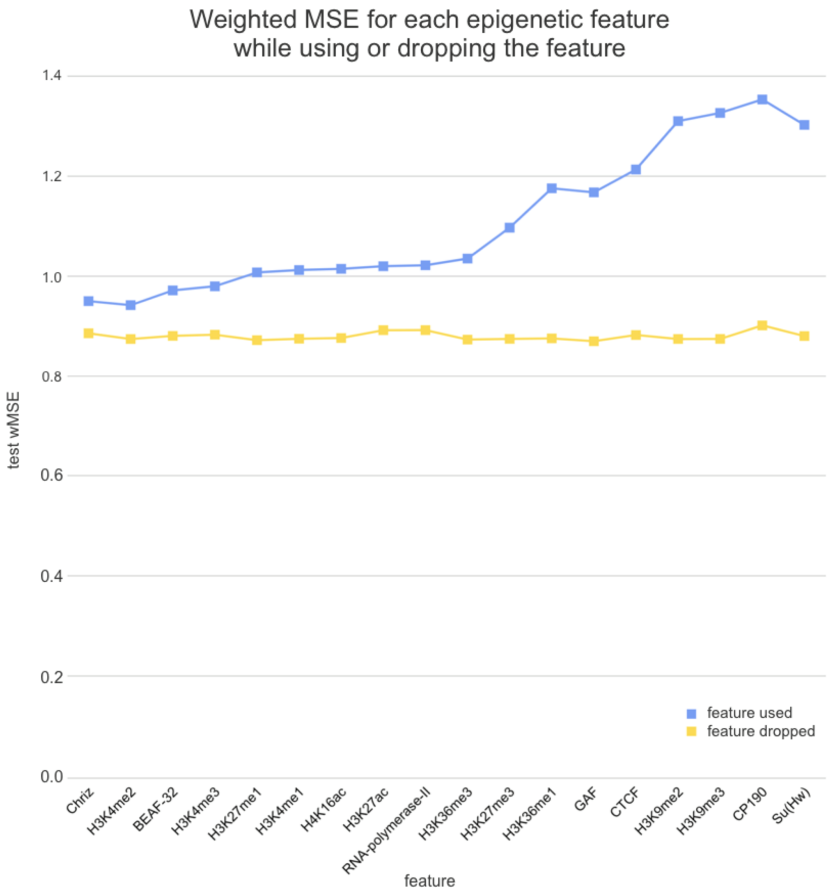


Figure 6 **Weighted MSE on the test dataset while using each chromatin mark either as a single feature (blue line) or excluding it from the biLSTM RNN input (yellow line).**
Full-size DOI: 10.7717/peerjcs.307/fig-6

Table 3 **Weighted MSE on cross-validation of all methods for each cell line and while using them together.** Lower wMSE orresponds to better quality of prediction.

Method	Schneider-2	Kc167	DmBG3-c2	All
Constant prediction	1.62 ± 0.09	1.53 ± 0.06	1.36 ± 0.05	1.51 ± 0.04
Linear regression	1.14 ± 0.08	1.01 ± 0.06	0.91 ± 0.08	1.04 ± 0.04
Linear regression + L1	1.12 ± 0.07	1.04 ± 0.06	0.95 ± 0.07	1.05 ± 0.04
Linear regression + L2	1.12 ± 0.07	1.01 ± 0.06	0.9 ± 0.08	1.03 ± 0.04
Linear regression + L1 + L2	1.11 ± 0.07	1.02 ± 0.06	0.91 ± 0.07	1.03 ± 0.04
Gradient boosting	1.07 ± 0.06	0.98 ± 0.07	0.86 ± 0.08	0.96 ± 0.04
biLSTM 64 units & 6 bins	**0.86 ± 0.04**	**0.83 ± 0.04**	**0.73 ± 0.01**	**0.78 ± 0.01**

The all-cell-lines model improves prediction for most cell lines

Finally, we tested the improvement of the prediction models that can be achieved by merging the information about all cell lines. For that, we merged all three cell lines as the input dataset and used the all-cell-lines model for the prediction on each cell line.

The gain of scores was the highest for Schneider-2 and Kc167, while BG3 demonstrated a slight decline in the prediction quality. We also note that biLSTM was less affected by the addition of cross-cell-line data among all models.

In general, the quality of the prediction has mostly improved, suggesting the universality of the biological mechanisms of the TAD formation between three cell lines (two embryonic and one neuronal) of *Drosophila*.

DISCUSSION

Here, we developed the Hi-ChIP-ML framework for the prediction of chromatin folding patterns for a set of input epigenetic characteristics of the genome. Using this framework, we provide the proof of concept that incorporation of information about the context of genomic regions is important for the TAD status and spatial folding of genomic regions. Our approach allows for diverse biological insights into the process of TAD formation in *Drosophila*, identified using the features importance analysis.

Firstly, we found that chromodomain protein Chriz, or Chromator (*Eggert, Gortchakov & Saumweber, 2004*), might be an important player of the TAD formation mechanism. Recurrent neural networks that used only Chriz as the input produced the highest scores among all RNNs using single epigenetic marks (Figs. 4, 6). Moreover, the removal of Chriz strongly influenced the prediction scores when four out of five selected ChIP features were together (Fig. 5). All linear models assigned the highest regression weight to the Chriz input signal. Further, with the L1 regularization Chriz was the only feature that the model selected for prediction. This chromodomain protein is known to be specific for the inter-bands of *Drosophila melanogaster* chromosomes (*Chepelev et al., 2012*), TAD boundaries and the inter-TAD regions (*Ulianov et al., 2016*), while profiles of proteins that are typically over-represented in inter-bands (including Chriz) correspond to TAD boundaries in embryonic nuclei (*Zhimulev et al., 2014*). The binding sites of insulator proteins Chriz and BEAF-32 are enriched at TAD boundaries (*Hou et al., 2012*; *Hug et al., 2017*; *Ramírez et al., 2018*; *Sexton et al., 2012*). *Wang et al. (2018)* reported the predictor of the boundaries based on the combination of BEAF-32 and Chriz. This might explain BEAF-32 achieving the third rank of the predictability score.

Secondly, the application of the recurrent neural network using each of the selected chromatin marks features separately (Fig. 6) has revealed a strong predictive power of active histone modifications such as H3K4me2. This result aligns with the fact that H3K4me2 defines the transcription factor binding regions in different cells, about 90% of transcription factor binding regions (TFBRs) on average overlap with H3K4me2 regions, and use H3K4me2 together with H3K27ac regions to improve the prediction of TFBRs (*Wang, Li & Hu, 2014*). Histone modifications H3K4me3, H3K27ac, H3K4me1, H3K4me3, H4K16ac, and other active chromatin marks are also enriched in inter-TADs and TAD boundaries (*Ulianov et al., 2016*). In addition, H3K27ac and H3K4me1 distinguish poised and active enhancers (*Barski et al., 2007*; *Creyghton et al., 2010*; *Rada-Iglesias et al., 2011*).

Thirdly, models using Su(Hw) and CTCF perform as expected given that, for the prediction of TAD boundaries, the binding of insulator proteins Su(Hw) and CTCF have

performed worse than other chromatin marks (*Ulianov et al., 2016*). In *Drosophila*, the absence of strong enrichment of CTCF at TAD boundaries and preferential location of Su(Hw) inside TADs implies that CTCF- and Su(Hw)-dependent insulation is not a major determinant of TAD boundaries. Our results also demonstrate that the impact of Su(Hw) and CTCF is low for both proteins.

Thus, our framework not only accurately predicts positions of TADs in the genome but also highlights epigenetic features relevant for the TAD formation. Importantly, the use of adjacent DNA bins created a meaningful biological context and enabled the training of a comprehensive ML model, strongly improving the evaluation scores of the best RNN model.

We note that there are few limitations to our approach. In particular, the resolution of our analysis is 20 kb, while TAD properties and TAD-forming factors can be different at finer resolutions (*Wang et al., 2018*; *Rowley et al., 2017*; *Rowley et al., 2019*). On the other hand, the use of coarse models allowed us to test the approach and select the best parameters while training the models multiple times efficiently. The training of the model for Hi-C with the resolution up to 500 bp presents a promising direction for future work, leading to the clarification of other factors' roles in the formation of smaller TAD boundaries that are beyond the resolution of our models.

We also note that transitional gamma is just one of multiple measures of the TAD state for a genomic region. We motivate the use of transitional gamma by the fact that it is a parameter-independent way of assessing TAD prominence calculated for the entire map. This guarantees the incorporation of the information about the interactions of the whole chromosome at all genomic ranges, which is not the case for other approaches such as the Insulation Score (*Crane et al., 2015*), D-score (*Stadhouders et al., 2018*), and Directionality Index (*Dixon et al., 2012*). On the other hand, the presented pipeline may be easily transferred to predict these scores as target values, which is an important direction for the extension of the work.

Here we selected features that had been reported to be associated with the chromatin structure. We note there might be other factors contributing to the TAD formation that were not included in our analysis. The exploration of a broader set of cell types might be a promising direction for this research, as well as the integration of various biological features, such as raw DNA sequence, to the presented models. We also anticipate promising outcomes of applying our approach to study the chromatin folding in various species except for *Drosophila*.

The code is open-source and can be easily adapted to various related tasks.

CONCLUSIONS

To sum up, we developed an approach for analysis of a set of chromatin marks as predictors of the TAD state for a genomic locus. We demonstrate a strong empirical performance of linear regression, gradient boosting, and recurrent neural network prediction models for several cell lines and a number of chromatin marks. The selected set of chromatin marks can reliably predict the chromatin folding patterns in *Drosophila*.

Recurrent neural networks incorporate the information about epigenetic surroundings. The highest prediction scores were obtained by the models with the biologically interpretable input size of 120 kb that aligns with the average TAD size for the 20 kb binning in *Drosophila*. Thus, we propose that the explicit accounting for linearly ordered bins is important for chromatin structure prediction.

The top-influencing TAD-forming factors of *Drosophila* are Chriz and histone modification H3K4me2. The chromatin factors that influence the prediction most are stable across the cell lines, which suggests the universality of the biological mechanisms of TAD formation for two embryonic and one neuronal *Drosophila* cell line. On the other hand, the training of models on all cell lines simultaneously generally improves the prediction.

The implemented pipeline called Hi-ChIP-ML is open-source. The methods can be used to explore the 3D chromatin structure of various species and may be adapted to any similar biological problem and dataset. The code is freely available at: https://github.com/MichalRozenwald/Hi-ChIP-ML.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This study was supported by the Russian Science Foundation, grant number 19-74-00112, and Skoltech Fellowship in Systems Biology for Aleksandra A. Galitsyna. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Russian Science Foundation: 19-74-00112.
Skoltech Fellowship in Systems Biology.

Competing Interests

Mikhail Gelfand is an Academic Editor for PeerJ. Grigory V. Sapunov is employed by Intento, Inc.

Author Contributions

- Michal B. Rozenwald conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Aleksandra A. Galitsyna conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Grigory V. Sapunov, Ekaterina E. Khrameeva and Mikhail S. Gelfand conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

1. The code and the data are available at GitHub: https://github.com/MichalRozenwald/Hi-ChIP-ML
2. The chromatin marks are available at modEncode using the following IDs:

name Schneider-2 Kc167 DmBG3-c2
1 Chriz 279 277 275
2 CTCF 3749 3749 3671
3 Su(Hw) 5147 3801 3717
4 BEAF-32 922 3745 3663
5 CP190 925 3748 3666
6 GAF 3753 3753 2651
7 H3K4me1 3760 5138 2653
8 H3K4me2 965 4935 2654
9 H3K4me3 3761 5141 967
10 H3K9me2 311 938 310
11 H3K9me3 4183 3013 312
12 H3K27ac 3757 3757 295
13 H3K27me1 3943 3942 3941
14 H3K27me3 298 5136 297
15 H3K36me1 3170 3003 299
16 H3K36me3 303 302 301
17 H4K16ac 320 318 316
18 RNA-polymerase-II 329 328 950

3. The Hi-C data is available at NCBI GEO: GSE69013.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.307#supplemental-information.

REFERENCES

Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129(4)**:823–837 DOI 10.1016/j.cell.2007.05.009.

Belokopytova PS, Nuriddinov MA, Mozheiko EA, Fishman D, Fishman V. 2020. Quantitative prediction of enhancer–promoter interactions. *Genome Research* **30(1)**:72–84 DOI 10.1101/gr.249367.119.

Bkhetan ZA, Plewczynski D. 2018. Three-dimensional epigenome statistical model: genome-wide chromatin looping prediction. *Scientific Reports* **8**:5217 DOI 10.1038/s41598-018-23276-8.

Chathoth KT, Zabet NR. 2019. Chromatin architecture reorganization during neuronal cell differentiation in *Drosophila* genome. *Genome Research* **29(4)**:613–625 DOI 10.1101/gr.246710.118.

Chepelev I, Wei G, Wangsa D, Tang Q, Zhao K. 2012. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Research* **22(3)**:490–503 DOI 10.1038/cr.2012.15.

Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, Uzawa S, Dekker J, Meyer BJ. 2015. Condensin-driven remodelling of x chromosome topology during dosage compensation. *Nature* **523(7559)**:240–244 DOI 10.1038/nature14450.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America* **107(50)**:21931–21936 DOI 10.1073/pnas.1016071107.

Cristescu B-C, Borsos Z, Lygeros J, Martínez MR, Rapsomaniki MA. 2018. Inference of the three-dimensional chromatin structure and its temporal behavior. ArXiv preprint. arXiv:1811.09619.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485(7398)**:376–380 DOI 10.1038/nature11082.

Eggert H, Gortchakov A, Saumweber H. 2004. Identification of the *Drosophila* interband-specific protein Z4 as a DNA-binding zinc-finger protein determining chromosomal structure. *Journal of Cell Science* **117(18)**:4253–4264 DOI 10.1242/jcs.01292.

Eraslan G, Avsec Ž, Gagneur J, Theis FJ. 2019. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics* **20(7)**:389–403.

Farré P, Heurteau A, Cuvier O, Emberly E. 2018. Dense neural networks for predicting chromatin conformation. *BMC Bioinformatics* **19(1)**:1–12 DOI 10.1186/s12859-018-2286-z.

Filion GJ, Van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, De Castro IJ, Kerkhoven RM, Bussemaker HJ, Van Steensel B. 2010. Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell* **143(2)**:212–224 DOI 10.1016/j.cell.2010.09.009.

Filippova D, Patro R, Duggal G, Kingsford C. 2014. Identification of alternative topological domains in chromatin. *Algorithms for Molecular Biology* **9(1)**:14 DOI 10.1186/1748-7188-9-14.

Fudenberg G, Kelley DR, Pollard KS. 2019. Predicting 3D genome folding from DNA sequence. *bioRxiv* 800060 DOI 10.1101/800060.

Gan M, Li W, Jiang R. 2019. EnContact: predicting enhancer-enhancer contacts using sequence-based deep learning model. *PeerJ* **2019(9)**:1–19 DOI 10.7717/peerj.7657.

Gan W, Luo J, Li YZ, Guo JL, Zhu M, Li ML. 2019. A computational method to predict topologically associating domain boundaries combining histone Marks and sequence information. *BMC Genomics* **20(13)**:1–12 DOI 10.1186/s12864-018-5379-1.

Gong Y, Lazaris C, Sakellaropoulos T, Lozano A, Kambadur P, Ntziachristos P, Aifantis I, Tsirigos A. 2018. Stratification of TAD boundaries reveals preferential insulation

of super-enhancers by strong boundaries. *Nature Communications* **9(1)**:542 DOI 10.1038/s41467-018-03017-1.

Graves A. 2012. Supervised sequence labelling. In: *Supervised sequence labelling with recurrent neural networks. Studies in computational intelligence, vol 385*. Berlin: Springer, 5–13 DOI 10.1007/978-3-642-24797-2_2.

Graves A, Jaitly N, Mohamed A-R. 2013. Hybrid speech recognition with deep bidirectional LSTM. In: *2013 IEEE workshop on automatic speech recognition and understanding*. IEEE, 273–278.

Hochreiter S, Schmidhuber J. 1997. Long short-term memory. *Neural Computation* **9(8)**:1735–1780 DOI 10.1162/neco.1997.9.8.1735.

Hou C, Li L, Qin ZS, Corces VG. 2012. Gene density, transcription, and insulators contribute to the partition of the *Drosophila* genome into physical domains. *Molecular Cell* **48(3)**:471–484 DOI 10.1016/j.molcel.2012.08.031.

Hug CB, Grimaldi AG, Kruse K, Vaquerizas JM. 2017. Chromatin architecture emerges during zygotic genome activation independent of transcription. *Cell* **169(2)**:216–228 DOI 10.1016/j.cell.2017.03.024.

Ibn-Salem J, Andrade-Navarro MA. 2019. 7C: computational chromosome conformation capture by correlation of ChIP-seq at CTCF motifs. *BMC Genomics* **20(1)**:777 DOI 10.1186/s12864-019-6088-0.

Jing F, Zhang S, Cao Z, Zhang S. 2019. An integrative framework for combining sequence and epigenomic data to predict transcription factor binding sites using deep learning. In: *IEEE/ACM transactions on computational biology and bioinformatics*. Piscataway: IEEE.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316(5830)**:1497–1502 DOI 10.1126/science.1141319.

Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T, Linder-Basso D, Plachetka A, Shanower G, Tolstorukov MY, Luquette LJ, Xi R, Jung YL, Park RW, Bishop EP, Canfield TK, Sandstrom R, Thurman RE, MacAlpine DM, Stamatoyannopoulos JA, Kellis M, Elgin SCR, Kuroda MI, Pirrotta V, Karpen GH, Park PJ. 2011. Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature* **471(7339)**:480–485 DOI 10.1038/nature09725.

Kingma DP, Ba J. 2014. Adam: a method for stochastic optimization. ArXiv preprint. arXiv:1412.6980.

Krijger PHL, De Laat W. 2016. Regulation of disease-associated gene expression in the 3D genome. *Nature Reviews Molecular Cell Biology* **17(12)**:771–782 DOI 10.1038/nrm.2016.138.

Li Z, Dai Z. 2020. SRHiC: a deep learning model to enhance the resolution of Hi-C data. *Frontiers in Genetics* **11**:353 DOI 10.3389/fgene.2020.00353.

Li W, Wong WH, Jiang R. 2019. DeepTACT: Predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Research* **47(10)**:e60 DOI 10.1093/nar/gkz167.

Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**(**5950**):289–293 DOI 10.1126/science.1181369.

Liu Q, Lv H, Jiang R. 2019. hicGAN infers super resolution Hi-C data with generative adversarial networks. *Bioinformatics* **35**(**14**):i99–i107 DOI 10.1093/bioinformatics/btz317.

Lupiáñez DG, Spielmann M, Mundlos S. 2016. Breaking TADs: how alterations of chromatin domains result in disease. *Trends in Genetics* **32**(**4**):225–237 DOI 10.1016/j.tig.2016.01.003.

Martens LD, Faust O, Pirvan L, Bihary D, Samarajiwa SA. 2020. Identifying regulatory and spatial genomic architectural elements using cell type independent machine and deep learning models. *bioRxiv*. DOI 10.1101/2020.04.19.049585.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. 2011. Scikit-learn: machine learning in python. *Journal of Machine Learning Research* **12**:2825–2830.

Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. 2011. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**(**7333**):279–283 DOI 10.1038/nature09692.

Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, Habermann B, Akhtar A, Manke T. 2018. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature communications* **9**(**1**):1–15 DOI 10.1038/s41467-017-02088-w.

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**(**7**):1665–1680 DOI 10.1016/j.cell.2014.11.021.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA. 2000. Genome-wide location and function of DNA binding proteins. *Science* **290**(**5500**):2306–2309 DOI 10.1126/science.290.5500.2306.

Rowley MJ, Lyu X, Rana V, Ando-Kuri M, Karns R, Bosco G, Corces VG. 2019. Condensin II counteracts cohesin and RNA polymerase II in the establishment of 3D chromatin organization. *Cell Reports* **26**(**11**):2890–2903 DOI 10.1016/j.celrep.2019.01.116.

Rowley MJ, Nichols MH, Lyu X, Ando-Kuri M, Rivera ISM, Hermetz K, Wang P, Ruan Y, Corces VG. 2017. Evolutionarily conserved principles predict 3D chromatin organization. *Molecular Cell* **67**(**5**):837–852 DOI 10.1016/j.molcel.2017.07.022.

Schreiber J, Libbrecht M, Bilmes J, Noble WS. 2017. Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture. *bioRxiv*. 14 DOI 10.1101/103614.

Schuster M, Paliwal K. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* **45(11)**:2673–2681 DOI 10.1109/78.650093.

Schwessinger R, Gosden M, Downes D, Brown R, Telenius J, Teh YW, Lunter G, Hughes JR. 2019. DeepC: Predicting chromatin interactions using megabase scaled deep neural networks and transfer learning. *bioRxiv*. 724005 DOI 10.1101/724005.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. 2012. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148(3)**:458–472 DOI 10.1016/j.cell.2012.01.010.

Singh S, Yang Y, Poczos B, Ma J. 2019. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. *Quantitative Biology* **7(2)**:122–137 DOI 10.1007/s40484-019-0154-0.

Stadhouders R, Vidal E, Serra F, Di Stefano B, Le Dily F, Quilez J, Gomez A, Collombet S, Berenguer C, Cuartero Y, Hecht J, Filion GJ, Beato M, Marti-Renom MA, Graf T. 2018. Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nature Genetics* **50(2)**:238–249 DOI 10.1038/s41588-017-0030-7.

Trieu T, Martinez-Fundichely A, Khurana E. 2020. DeepMILO: a deep learning approach to predict the impact of non-coding sequence variants on 3D chromatin structure. *Genome Biology* **21(1)**:1–11 DOI 10.1186/s13059-019-1906-x.

Ulianov SV, Khrameeva EE, Gavrilov AA, Flyamer IM, Kos P, Mikhaleva EA, Penin AA, Logacheva MD, Imakaev MV, Chertovich A, Gelfand MS, Shevelyov YY, Razin SV. 2016. Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Research* **26(1)**:70–84 DOI 10.1101/gr.196006.115.

Wang Y, Li X, Hu H. 2014. H3K4me2 reliably defines transcription factor binding regions in different cells. *Genomics* **103(2)**:222–228 DOI 10.1016/j.ygeno.2014.02.002.

Wang Q, Sun Q, Czajkowsky DM, Shao Z. 2018. Sub-kb Hi-C in *D. melanogaster* reveals conserved characteristics of TADs between insect and mammalian cells. *Nature Communications* **9(1)**:1–8 DOI 10.1038/s41467-017-02088-w.

Waterston R, Celniker S, Snyder M, White K, Henikoff S, Karpen G. 2009. Unlocking the secrets of the genome. *Nature* **459(7249)**:927–930 DOI 10.1038/459927a.

Whalen S, Truty RM, Pollard KS. 2016. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature Genetics* **48(5)**:488–496 DOI 10.1038/ng.3539.

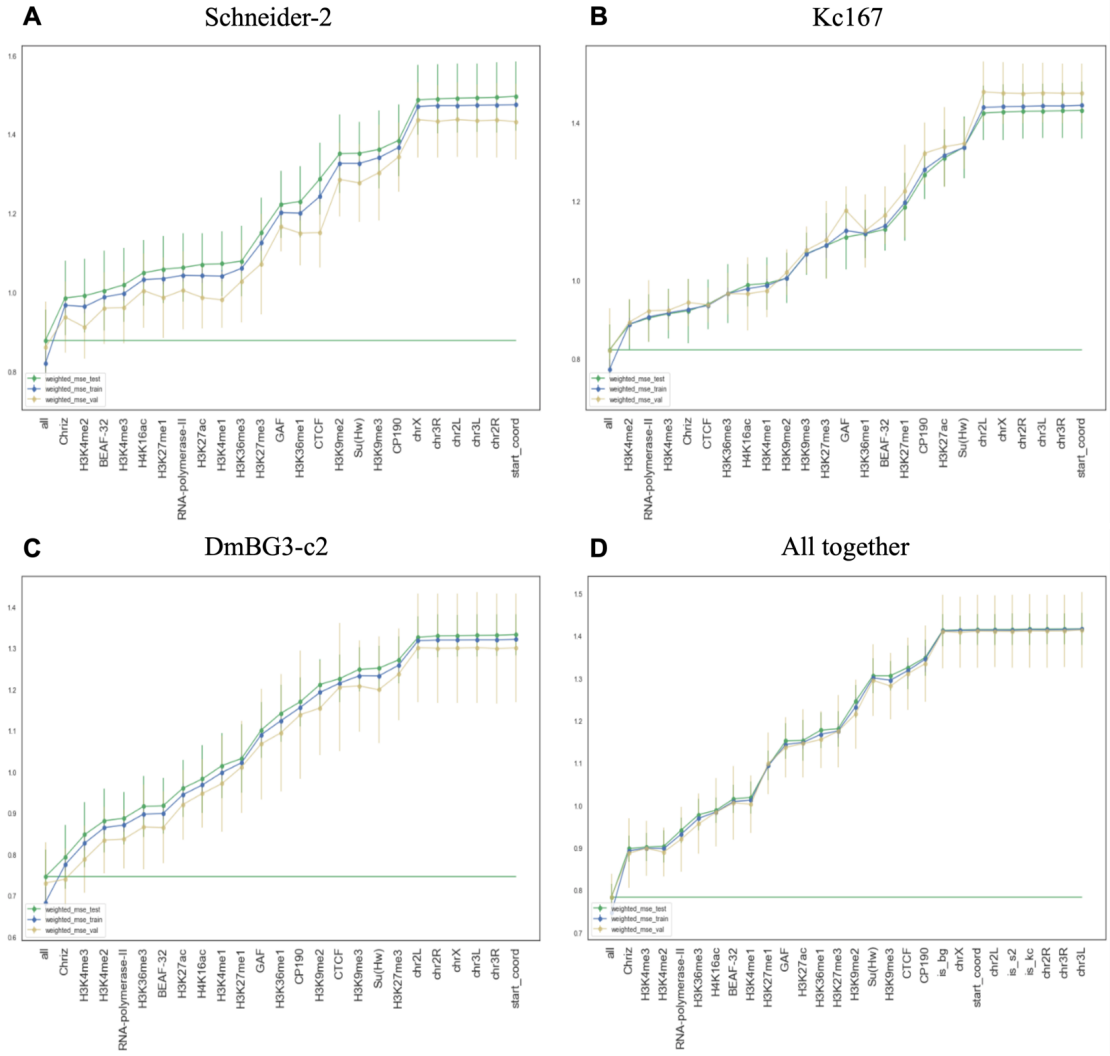
Yan X, Su X. 2009. *Linear regression analysis: theory and computing*. Singapore: World Scientific.

Yuan Y, Shi Y, Su X, Zou X, Luo Q, Feng DD, Cai W, Han Z-G. 2018. Cancer type prediction based on copy number aberration and chromatin 3D structure with convolutional neural networks. *BMC Genomics* **19(6)**:565 DOI 10.1186/s12864-018-4919-z.

Zeng W, Wang Y, Jiang R. 2020. Integrating distal and proximal information to predict gene expression via a densely connected convolutional neural network. *Bioinformatics* **36**(**2**):496–503.

Zeng W, Wu M, Jiang R. 2018. Prediction of enhancer-promoter interactions via natural language processing. *BMC Genomics* **19**(**Suppl 2**):84 DOI 10.1186/s12864-018-4459-6.

Zhimulev IF, Zykova TY, Goncharov FP, Khoroshko VA, Demakova OV, Semeshin VF, Pokholkova GV, Boldyreva LV, Demidova DS, Babenko VN, Demakov SA, Belyaeva ES. 2014. Genetic organization of interphase chromosome bands and interbands in *Drosophila melanogaster*. *PLOS ONE* **9**(**7**):1–16 DOI 10.1371/journal.pone.0101631.


A
Schneider-2
B
Kc167
C
DmBG3-c2
D
All together
weighted_mse_test
weighted_mse_train
weighted_mse_val
all
Chriz
H3K4me2
BEAF-32
H3K4me3
H4K16ac
H3K27me1
RNA-polymerase-II
H3K27ac
H3K4me1
H3K36me3
H3K27me3
GAF
H3K36me1
CTCF
H3K9me2
Su(Hw)
H3K9me3
CP190
chrX
chr3R
chr2L
chr3L
chr2R
start_coord
is_bg
is_s2
is_kc

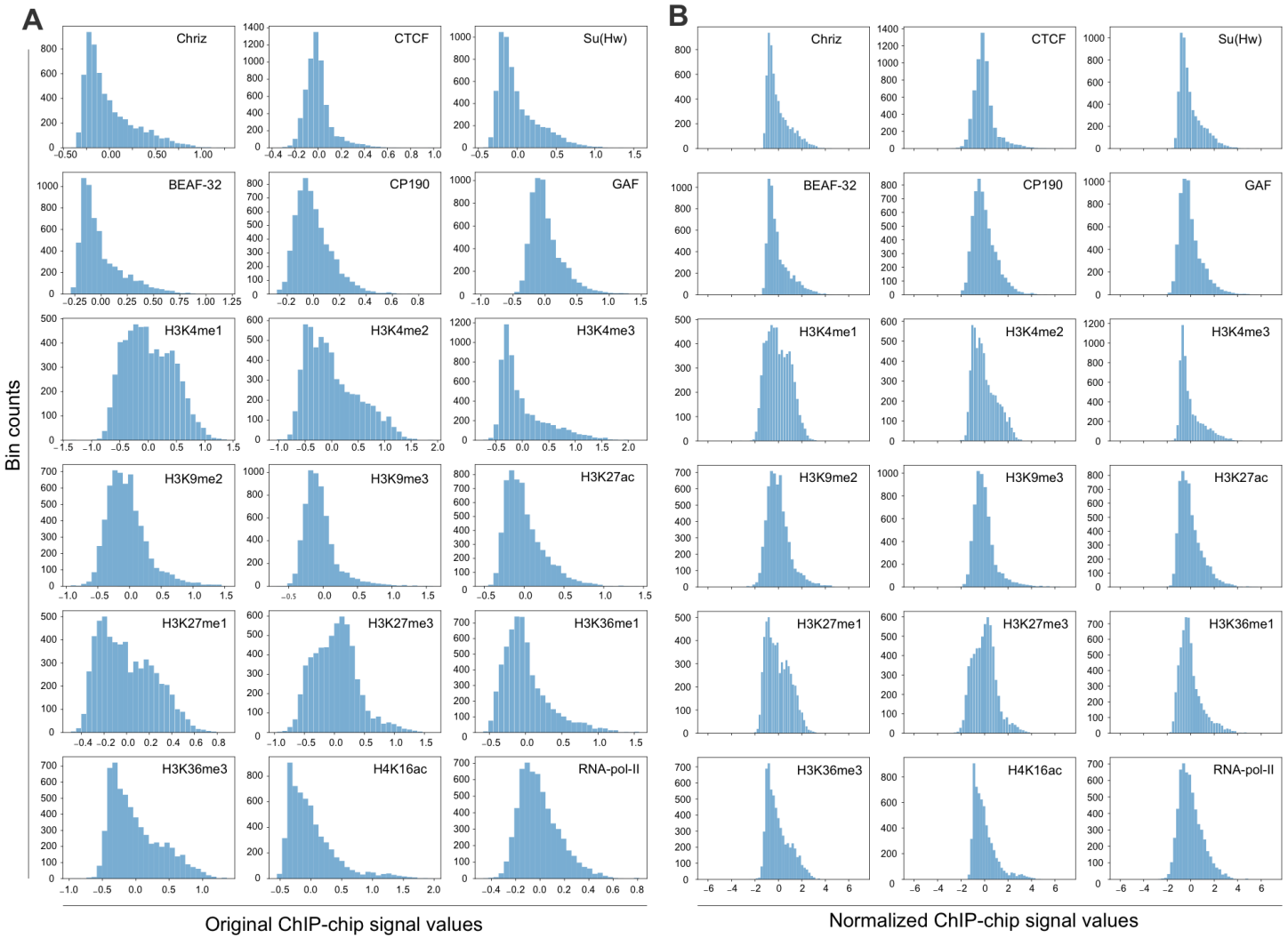

A
B
Chriz
CTCF
Su(Hw)
BEAF-32
CP190
GAF
H3K4me1
H3K4me2
H3K4me3
H3K9me2
H3K9me3
H3K27ac
H3K27me1
H3K27me3
H3K36me1
H3K36me3
H4K16ac
RNA-pol-II
Bin counts
Original ChIP-chip signal values
Normalized ChIP-chip signal values

Table 1. The modENCODE IDs of chromatin factors for three selected *Drosophila* cell lines.

	modENCODE IDs		
NAME	SCHNEIDER-2	KC167	DMBG3-C2
CHRIZ	279	277	275
CTCF	3749	3749	3671
SU(HW)	5147	3801	3717
BEAF-32	922	3745	3663
CP190	925	3748	3666
GAF	3753	3753	2651
H3K4ME1	3760	5138	2653
H3K4ME2	965	4935	2654
H3K4ME3	3761	5141	967
H3K9ME2	311	938	310
H3K9ME3	4183	3013	312
H3K27AC	3757	3757	295
H3K27ME1	3943	3942	3941
H3K27ME3	298	5136	297
H3K36ME1	3170	3003	299
H3K36ME3	303	302	301
H4K16AC	320	318	316
RNA-POLYMERASE-II	329	328	950

Chapter 6

Cumulative contact frequency of a chromatin region is an intrinsic property linked to its function

Iterative correction of population Hi-C data is a routine practice that mitigates uneven visibility of the genomic regions arising due to different coverage by restriction fragments, different PCR amplification effectiveness, chromatin accessibility to the reagents, and other experimental biases. It does so by iteratively normalizing the contact matrix by a vector of sums of contacts for each region, what we call marginal distribution of contacts or cumulative contact frequency (CCF). In other words, CCF is an aggregated number of contacts for each genomic region, or genomic bin.

However, some properties of CCF were not understood before us. In particular, we did not know what potentially important biological signal is removed with it. In this study, we deploy the correlation analysis of CCF on bulk Hi-C and demonstrate that CCF is correlated with active states of chromatin and active compartment. Thus, with iterative correction of bulk Hi-C data, we remove potentially relevant biological information instead of only technical biases. Currently, there are no other solutions to treat this fundamental problem of Hi-C data.

It is important to note that we calculate CCF correlations with other features genome-wide after the aggregation of contacts for each genomic bin. Indeed, there is another specialized application of correlation for compartments calling [Lieberman-

Aiden et al., 2009]. In compartment calling, the correlation of interaction patterns is calculated for all pairs of genomic bins. This procedure retains mainly the signal from long-range contacts; it does not allow the study of local chromatin features, such as TADs and loops. Thus, its applications are limited to compartments calling only. This approach should not be confused with our study.

Finally, I want to emphasize that in single-cell Hi-C, as opposed to bulk, we usually do not normalize raw data by CCF because the data is too sparse to obtain interpretable results after iterative correction. Thus, all critical experimental biases of CCF are present in scHi-C data.

With this study, I explored possible ways to treat the marginal distributions of contacts in individual cells in the future, the knowledge exploited in Chapter 7. Additionally, the software developed here was used in the later studies. My contribution to this work is shared with co-authors. Although they prepared the final figures, I made the preliminary calculations and independently reproduced most of the results presented here. I also wrote a substantial part of the text below.

Cumulative contact frequency of a chromatin region is an intrinsic property linked to its function

Margarita D. Samborskaia[1,*], Aleksandra Galitsyna[2,3,4,*], Ilya Pletenev[2], Anna Trofimova[2], Andrey A. Mironov[1,3], Mikhail S. Gelfand[2,3] and Ekaterina E. Khrameeva[2]

[1] Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia
[2] Center of Life Sciences, Skolkovo Institute of Science and Technology, Moscow, Russia
[3] A.A. Kharkevich Institute for Information Transmission Problems, RAS, Moscow, Russia
[4] Institute of Gene Biology, RAS, Moscow, Russia
[*] These authors contributed equally to this work.

ABSTRACT

Regulation of gene transcription is a complex process controlled by many factors, including the conformation of chromatin in the nucleus. Insights into chromatin conformation on both local and global scales can be provided by the Hi-C (high-throughput chromosomes conformation capture) method. One of the drawbacks of Hi-C analysis and interpretation is the presence of systematic biases, such as different accessibility to enzymes, amplification, and mappability of DNA regions, which all result in different visibility of the regions. Iterative correction (IC) is one of the most popular techniques developed for the elimination of these systematic biases. IC is based on the assumption that all chromatin regions have an equal number of observed contacts in Hi-C. In other words, the IC procedure is equalizing the experimental visibility approximated by the cumulative contact frequency (CCF) for all genomic regions. However, the differences in experimental visibility might be explained by biological factors such as chromatin openness, which is characteristic of distinct chromatin states. Here we show that CCF is positively correlated with active transcription. It is associated with compartment organization, since compartment A demonstrates higher CCF and gene expression levels than compartment B. Notably, this observation holds for a wide range of species, including human, mouse, and *Drosophila*. Moreover, we track the CCF state for syntenic blocks between human and mouse and conclude that active state assessed by CCF is an intrinsic property of the DNA region, which is independent of local genomic and epigenomic context. Our findings establish a missing link between Hi-C normalization procedures removing CCF from the data and poorly investigated and possibly relevant biological factors contributing to CCF.

Subjects Bioinformatics, Cell Biology, Computational Biology, Genomics, Molecular Biology
Keywords Hi-C, Chromatin, Compartments, Conformation capture

INTRODUCTION

The conformation of chromatin in the nucleus plays an important role in many cellular processes, including the regulation of gene transcription and DNA replication

Submitted 27 January 2020
Accepted 27 June 2020
Published 10 August 2020

Corresponding authors
Mikhail S. Gelfand, gelfand@iitp.ru
Ekaterina E. Khrameeva, e.khrameeva@skoltech.ru

Academic editor
Yegor Vassetzky

Additional Information and Declarations can be found on page 12

DOI **10.7717/peerj.9566**

How to cite this article Samborskaia MD, Galitsyna A, Pletenev I, Trofimova A, Mironov AA, Gelfand MS, Khrameeva EE. 2020. Cumulative contact frequency of a chromatin region is an intrinsic property linked to its function. *PeerJ* **8**:e9566 http://doi.org/10.7717/peerj.9566

(*Cremer et al., 2006*; *Sexton et al., 2007*; *Bickmore, 2013*; *Armstrong et al., 2018*; *Fulco et al., 2019*). The regulation of gene expression often involves long-range chromatin interactions between regulatory elements. Therefore, the spatial organization of chromatin could provide insight into these complex regulatory processes.

Hi-C is a method for genome-wide chromosome conformation capture, which enables the interrogation of all loci at once by combining DNA proximity ligation with high-throughput sequencing (*Lieberman-Aiden & Van Berkum, 2009*). However, data obtained by Hi-C have both technical and experiment-induced biases. Because of that, different regions of the genome may have different visibility in the experiment, yielding systematic errors in data interpretation. To correct for this bias, several approaches exist, reviewed in *Lajoie, Dekker & Kaplan (2014)* and *Schmitt, Hu and Ren (2016)*. Some recent advances in the Hi-C data analysis allow for various modifications of the correction procedure, such as probabilistic modeling (*Yaffe & Tanay, 2011*), Vanilla-Coverage (*Lieberman-Aiden & Van Berkum, 2009*), binless normalization (*Spill et al., 2019*) and other. One of the most commonly used methods for the elimination of systematic biases is the iterative correction (IC) (*Imakaev et al., 2012*). In particular, it is used as a gold standard in the Hi-C data processing package *cooler* (*Abdennur & Mirny, 2019*).

While IC is based on the assumption that all loci have equal visibility (observed number of contacts), the differences in experimental visibility may be explained not only by technical or experimental biases but also by biological factors. Notably, local chromatin conformation correlates with functional characteristics of the genome, such as individual histone modifications (*Khrameeva et al., 2012*) or their combinatorial patterns that establish certain functionality for each region, chromatin states (*Ernst et al., 2011*). Therefore, elimination of the differences in the visibility of chromatin regions could lead to the loss of a biologically meaningful signal.

Features such as TADs, enriched contacts, and compartments are usually called in normalized Hi-C interaction maps (*Forcato et al., 2017*). Experimental visibility is treated as a purely methodological artifact that is assumed not to affect the detection of these features. Other studies (*Chandradoss, Guthikonda & Kethavath, 2020*; *Beagrie et al., 2017*), have previously highlighted the importance of experimental visibility of DNA regions in Hi-C. However, to our knowledge, the relation of this genomic characteristic to expression and chromatin states has not been analyzed. Here, we establish a relation between visibility of DNA in Hi-C, assessed by the cumulative contact frequency (CCF), and chromatin states in a range of species.

METHODS

Analysis of Hi-C data

Processing of Hi-C data

We analyzed Hi-C maps for human cell lines HMEC, HUVEC, and K562, mouse cell line CH12-LX (*Rao et al., 2014*), and fruit fly (*Ulianov et al., 2015*) Schneider-2 (S2) cells (GEO database, accession numbers GSE63525 and GSE69013, respectively). We downloaded the processed Hi-C maps in the hic format from (*Rao et al., 2014*) and in the txt format

from (*Ulianov et al., 2015*). The Hi-C maps were converted to the matrix format and binned at the 1 Mb resolution. The main results were obtained for the human cell line HMEC and, where possible, for other cell lines and species to demonstrate the generalizability of our findings (see Supplemental Information).

To eliminate possible technical artifacts of Hi-C, such as single-sided reads and their subsets, mirror reads (*Galitsyna et al., 2017*), even though the coverage profile for these reads might be well-correlated with the coverage profile of double-sided reads (*Imakaev et al., 2012*), we removed the diagonal 1-Mb elements of the Hi-C maps. Additionally, for CCF calculations, we removed the secondary diagonal corresponding to regions immediately adjacent to each other and all contacts at the distance up to 5 Mb in order to remove the area of high contact frequencies that could hinder subsequent analysis. Genomic regions corresponding to rows and columns of Hi-C maps, which contained no values, were also removed from all analyses.

We calculated the cumulative contact frequency (CCF) as the sum of contact frequencies of each locus. To make CCF comparable between different cell lines and resolutions, we further report it as the percentage from the maximum CCF in the Hi-C map. We considered two types of CCF: whole-genome and inter-chromosomal (calculated for inter-chromosomal Hi-C maps). Inter-chromosomal CCF was analyzed separately to demonstrate that intra-chromosomal contacts do not drive our observations.

TAD and compartment calling

We used the *Armatus* algorithm (*Filippova et al., 2014*), as implemented in the *Lavaburst* package (accessed 01-12-18, *modularity* scoring function and *gamma* parameter 1.0 (*Abdennur, 2018*)), for TAD calling in human Hi-C maps at the 1 Mb resolution. We considered all segments smaller than three bins as interTAD regions. This allowed us to classify the genomic bins into two categories: TAD and interTAD bins. We then used these bins separately for the correlation analysis of CCF at TAD and interTAD genomic regions.

In order to identify chromatin compartments, we performed computational analysis as in *Lieberman-Aiden & Van Berkum (2009)*. For that, we normalized the whole-genome contact matrix by the expected contact frequency matrix, generated by averaging contact probabilities for loci at each genomic distance. We then calculated the Pearson correlation coefficients for each row/column pair of each element of the normalized matrix to obtain the correlation matrix. The resulting correlation matrix was then used for the principal component analysis (PCA). We used the first principal component of the resulting correlation matrix as a compartment annotation for the genome.

Notably, the first principal component for human and mouse datasets demonstrated the highest proportion of variance explained (PVE of the first component ranging from 0.60 for HMEC and HUVEC cells to 0.80 for K562 cells) and had a characteristic checkerboard pattern in accordance with previous findings (*Lieberman-Aiden & Van Berkum, 2009*; *Rao et al., 2014*). We were unable to detect compartments in the analyzed *Drosophila* dataset (PVE for the first component of S2 cells is 0.11), probably due to *Drosophila* compartments being much smaller than the selected dataset resolution (1 Mb).

Functional characteristics

We estimated the functional characteristics of genomic regions by combinatorial patterns of chromatin marks, or chromatin states, for human (*Ernst et al., 2011*), mouse (*Yue et al., 2014*), and *Drosophila* (*Kharchenko et al., 2011*). These chromatin states were originally derived from a set of ChIP-seq experiments for various chromatin factors by Hidden Markov Models, and represented distinct states with specific ChIP-seq signatures. Chromatin states are better for the assessment of functional properties of genomic regions than individual marks from two perspectives. First, they represent an integrated view of the region's expression and functional characteristics; the experimental noise of individual ChIP-seq experiments is smoothed out. Second, the analysis of chromatin states is simpler, compared to a set of marks.

We retrieved fifteen states from *Ernst et al. (2011)* for the human genome, seven states for the mouse genome (*Yue et al., 2014*), and nine states for the *Drosophila* genome (*Kharchenko et al., 2011*). The original datasets were downloaded in the format of a non-intersecting set of genomic regions, with a unique chromatin state assigned to each region. In order to match the Hi-C data uniform grid, we segmented the genome into non-overlapping 1-Mb genomic windows, or bins, starting from the first position of each chromosome. For each genomic bin, we then computed the fraction of coverage of each chromatin state. If the initial chromatin state segment spanned the bin boundary, it was split into two parts by the bin boundary and counted as contributing to both bins that it overlaps, proportionally to the resulting fragments sizes. Thus, for each bin and chromatin state, we obtained a single number from 0 to 1, reflecting the coverage of this bin by the chromatin state. Bins containing no annotation of chromatin states were removed from further analysis.

The chromatin states for the human genome from *Ernst et al. (2011)* are named by the principal function of the respective regions. We separated them into two groups by functional activity. The first group is active chromatin: Active Promoter, Weak Promoter, Inactive/poised Promoter, Strong Enhancer (2), Weak/poised Enhancer (2), Weak Transcription, Transcriptional Elongation, Transcriptional Transition. The second group is inactive chromatin: Repetitive/CNV (2), Heterochromatin.

Chromatin states for mouse from *Yue et al. (2014)* are named by the histone modifications prevalent in the corresponding state. The active marks are represented by: H3K4me3, H3K4me1/3, H3K4me1, H3K4me1+H3K36me3, and H3K36me3. Only one state, H3K27me3, represents inactive chromatin, and one state is comprised of all unmarked genomic regions.

The states for the *Drosophila* genome are called "colors" with functional load described in the original publication (*Kharchenko et al., 2011*). Based on that, we separated *Drosophila* states into two groups, active chromatin, comprized of RED (1) and MAGENTA (2) colors, and inactive/repressed chromatin, comprized of DARKGRAY (6), DARKBLUE (7), LIGHTBLUE (8), LIGHTGRAY (9).

Correlation analysis

To characterize correlation patterns, we used two approaches. First, we calculated the Pearson correlation coefficients between CCF and chromatin state proportions

in each region of the whole genome. To further validate the findings, we used Stereogene (*Stavrovskaya et al., 2017*), a tool for the genome-wide feature correlation analysis. We explored the relationship between pairs of characteristics of the genome, such as CCF, GC-content, and proportion of each chromatin state. Stereogene divides the input data into a series of fixed-length windows (adjustable parameter that was set to 10 Mb), and the independent correlation is calculated for each set. The distribution of these correlations allows one to observe the variation in the correlation coefficient across the genome and to identify regions with non-typically high positive or negative correlation. These distributions are compared against a randomized control derived from the data (*Stavrovskaya et al., 2017*), and *p*-values are calculated for the observed correlations in the real data.

Analysis of syntenic regions

Syntenic regions (size 2 Mb and larger) were obtained from the Mouse Genome Informatics database (MGI) (*Finger et al., 2011*). The regions of homology with the human genome (size 1 Mb) were established using the *LiftOver* tool (*Hinrichs, 2006*). The contact frequency for ambiguously mapped regions was split proportionally to the lengths of the mapped fragments. For this analysis, we defined large chromosomes as chromosomes 1–9, and small chromosomes as chromosomes 14–22. The Pearson correlation coefficient was calculated between the human and mouse CCF.

RESULTS

Increased CCF is associated with active transcription

As active transcription requires binding of RNA polymerase and a variety of transcription factors, increased gene expression is intuitively associated with loose packaging of chromatin and thus better accessibility to the Hi-C reaction and higher CCF. At the same time, active chromatin is involved in a larger number of interactions, including distant regulatory ones. Thus, one might expect regions with high CCF to show high gene expression levels and regions with low CCF to exhibit low gene expression. To validate this hypothesis, we constructed a whole-genome Hi-C map combined with the functional state plot showing the distributions of chromatin state proportions for each genomic region (Figs. 1A–1C). Indeed, for all analyzed human cell lines (HMEC, HUVEC, and K562), the regions with high CCF tend to be enriched in chromatin states corresponding to active transcription, while regions showing low CCF are enriched in heterochromatin and repeats (Fig. 1C, Fig. S1).

Chromosomes are known to segregate into two mutually exclusive types of chromatin, referred to as "A" and "B" compartments (*Lieberman-Aiden & Van Berkum, 2009*). Active chromatin corresponds to the A compartment, while repressed chromatin is enriched within the B compartment. Using correlation analysis of normalized Hi-C maps and PCA, we segregated the genome into two types of chromosomal regions. We observe that human compartment A has high levels of CCF in HMEC and other human cell lines (Fig. 1D, Fig. S2).

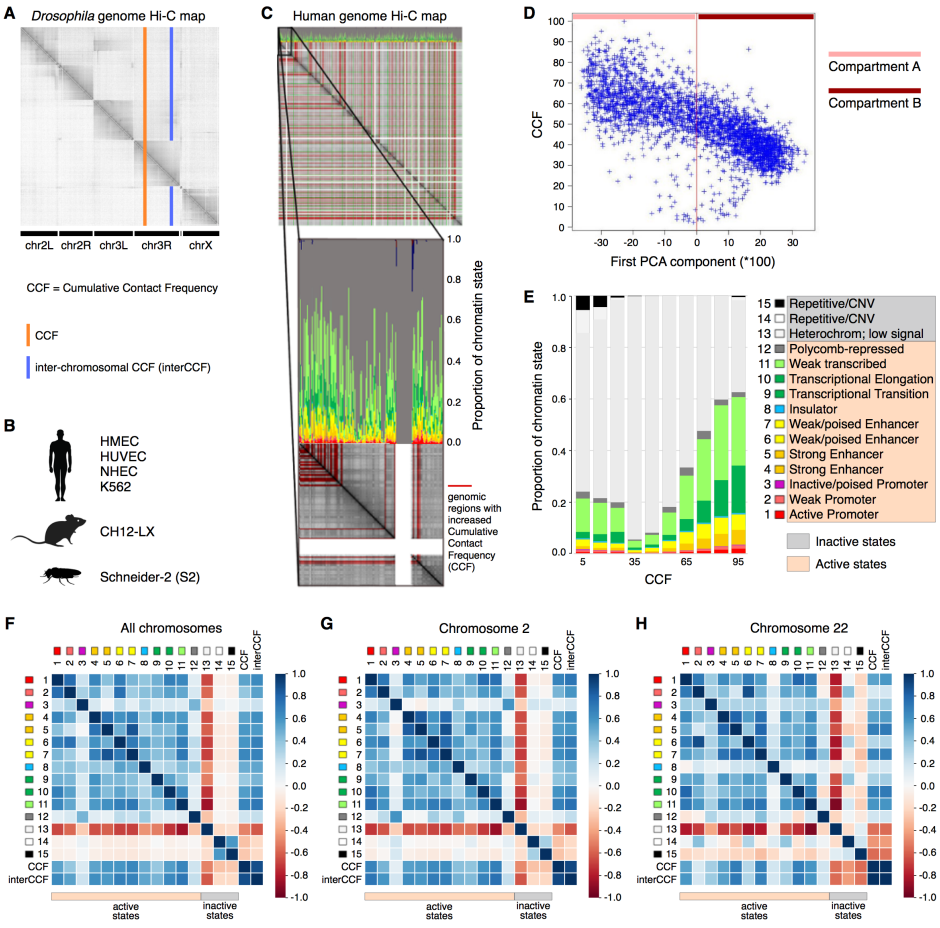


Figure 1 **Cumulative contact frequency (CCF) is positively correlated with active transcription.** (A) Schematic representation of inter-chromosomal (blue line) and total (orange line) CCF. (B) Cell lines and organisms analyzed in this study. (C) Hi-C map combined with a plot of chromatin state proportions. Red lines on the Hi-C map show regions of anomalously high CCF. Green lines separate individual chromosomes. Proportions of each chromatin state for each genomic region are displayed above the Hi-C map. An enlarged fragment of the Hi-C map for chromosome 1 is shown below. (D) Dependency of CCF on the first principal component. (E) Dependency of chromatin state proportions on CCF. (F-H) Correlation patterns between chromatin states and CCF exhibit different features for large and small chromosomes. First 15 rows in the matrix correspond to the 15 chromatin states, rows 16-17 exhibit total and inter-chromosomal CCF. Colors demonstrate the Pearson correlation coefficients. Whole-genome correlation patterns (F), correlation patterns for chromosome 2 (G) and chromosome 22 (H) are shown. Human cell line HMEC.

Full-size DOI: 10.7717/peerj.9566/fig-1

CCF is linked to active chromatin states

To get a more precise estimate of dependencies between CCF and chromatin states, we visualized the growth of chromatin state percentages at increasing CCF (Fig. 1E, HMEC cells). We observe the growth of percentages of the chromatin states corresponding to active transcription (Weak Transcription, Transcriptional Elongation, and Transcriptional

Transition, in particular) with larger CCF. This result does not depend on the Hi-C data resolution, as proved by the same analysis repeated for 1 Mb, 500 Kb, 250 Kb, 100 Kb, and 50 Kb resolutions (Fig. S3).

To further validate the result, we calculated the correlations between each of the chromatin states and CCF (Fig. 1F) enabling comparative analysis of different genomic regions. CCF is positively correlated with active chromatin state proportions in HMEC cells (correlation coefficient 0.53). The same result is obtained for other human cell lines (Fig. S1): HUVEC (correlation coefficient 0.47) and K562 (correlation coefficient 0.35). As an additional proof of concept, the homogeneity of correlations across the genome was confirmed for the cell line HMEC (Fig. S4) with the Stereogene tool (*Stavrovskaya et al., 2017*). Notably, the correlation patterns are similar for large chromosomes but different for smaller ones (Fig. 1F, Fig. S5).

To show that the dependencies between CCF and chromatin states are not specific to humans, we additionally analyzed the *Drosophila* cell line S2 and mouse cell line CH12-LX. For *Drosophila* and mouse, the chromatin state annotations (*Kharchenko et al., 2011*) differ from that in human. In particular, there are fewer chromatin states, and their functional characteristics are different. However, for *Drosophila*, we observe a positive correlation of CCF with chromatin states RED (1) and MAGENTA (2) (Fig. S6), which are representative of active chromatin with expressed genes. For mouse, we observe a positive correlation of CCF with all chromatin states but the one characterized by absence of chromatin marks (Fig. S7).

TADs and interTAD regions demonstrate different patterns for the human genome (Fig. S8). TAD CCF is correlated with active chromatin and anti-correlated with inactive chromatin, while interTAD CCF is correlated with heterochromatin and insulator chromatin states. The latter fact might be related to the interTAD insulator property. By contrast to humans, TADs and interTAD regions have only slight differences in *Drosophila* (Fig. S9), where both TAD and interTAD CCF demonstrate a positive correlation with active chromatin states and a negative correlation with inactive chromatin states.

CCF association with active chromatin is not driven by GC-content

The observed correlation between CCF and chromatin states is not necessarily direct and causative, as there might exist other genetic or epigenetic factors underlying both CCF and active chromatin state. If there is such a confounding factor, then accounting for its influence would diminish the observed correlations.

One possible type of confounders are GC-content and chromosome length. Our initial analysis demonstrates that GC-content and chromosome length are indeed both correlated with contact frequency, and the dependencies are linear or nearly linear (Fig. S10–Fig. S11). Inter-chromosomal CCF decreases with chromosome length, which indicates that small chromosomes tend to make more inter-chromosomal contacts than large chromosomes, in line with previous studies showing that small chromosomes are gene-rich and tend to interact with each other (Fig. S11A) (*Lieberman-Aiden & Van Berkum, 2009*). In particular, the correlation between the chromosome length and an average inter-chromosomal CCF is -0.42 for the cell line HMEC.

To test whether CCF is correlated with active chromatin state in the absence of these confounding factors, we performed a simple division of CCF by these factors and re-calculated the correlation plots. CCF normalized by the chromosome length or by the GC-content demonstrated the same correlation patterns as non-normalized CCF (Figs. S10B, S11B). To further validate this observation, we applied linear regression to predict CCF from the GC-content. The correlation patterns are weakened, but still the same as for non-normalized CCF (Fig. S10B). Further, normalization of CCF by the chromosome length and subsequent removal of the GC-content effect shows that, even combined, the chromosome length and GC-content cannot explain the observed correlation patterns (Figs. S10B, S11B).

CCF for different chromosomes reveals hidden variability in correlation patterns

Each chromosome has its own unique properties, which cannot be detected while considering the correlation pattern for the whole genome. Since each chromosome differs in contact frequency preferences, the correlation patterns calculated for separate chromosomes may also differ. Indeed, while the first nine chromosomes show a correlation pattern similar to that of the whole genome, smaller chromosomes exhibit individual unique correlation patterns (Figs. 1G–1H, Fig. S5).

One possibility is that we have observed a statistical artifact, caused simply by differences in the sample size, as, naturally, more fragments are considered for large chromosomes than for small ones. However, downsampling large chromosomes to the size of small chromosomes demonstrates that correlations of small chromosomes still are outliers (Figs. 2A–2B, Fig. S12). It suggests that the observed effects for small chromosomes are not due to the small sample size.

We have noticed that another important factor might be the size of centromeres, which might have different sizes hence forming different fractions of chromosomes. We have excluded the centromere regions and demonstrated that the observed correlation patterns are not related to differences in the centromere size (Fig. S13).

Notably, for individual chromosomes of the *Drosophila* genome, the correlation patterns are more similar (Fig. S6). However, a direct comparison with the results for human is impossible due to differences in chromatin state annotations between the human and *Drosophila* datasets.

Comparison of CCF in syntenic blocks between mouse and human

Small chromosomes might show unique correlation patterns due to the impact of specific evolutionary conserved regions, such as syntenic blocks. To test this hypothesis, we annotated syntenic regions and calculated CCF for them in the human and mouse datasets.

Indeed, syntenic regions of short chromosomes demonstrate correlations between contact frequency and chromatin states that are not characteristic of syntenic blocks in long chromosomes (Figs. 2A–2B, note the difference between real correlations (red line) and controls (blue bars)). Moreover, syntenic regions have similar preferences in contact frequencies between two species (Figs. 2C–2E).

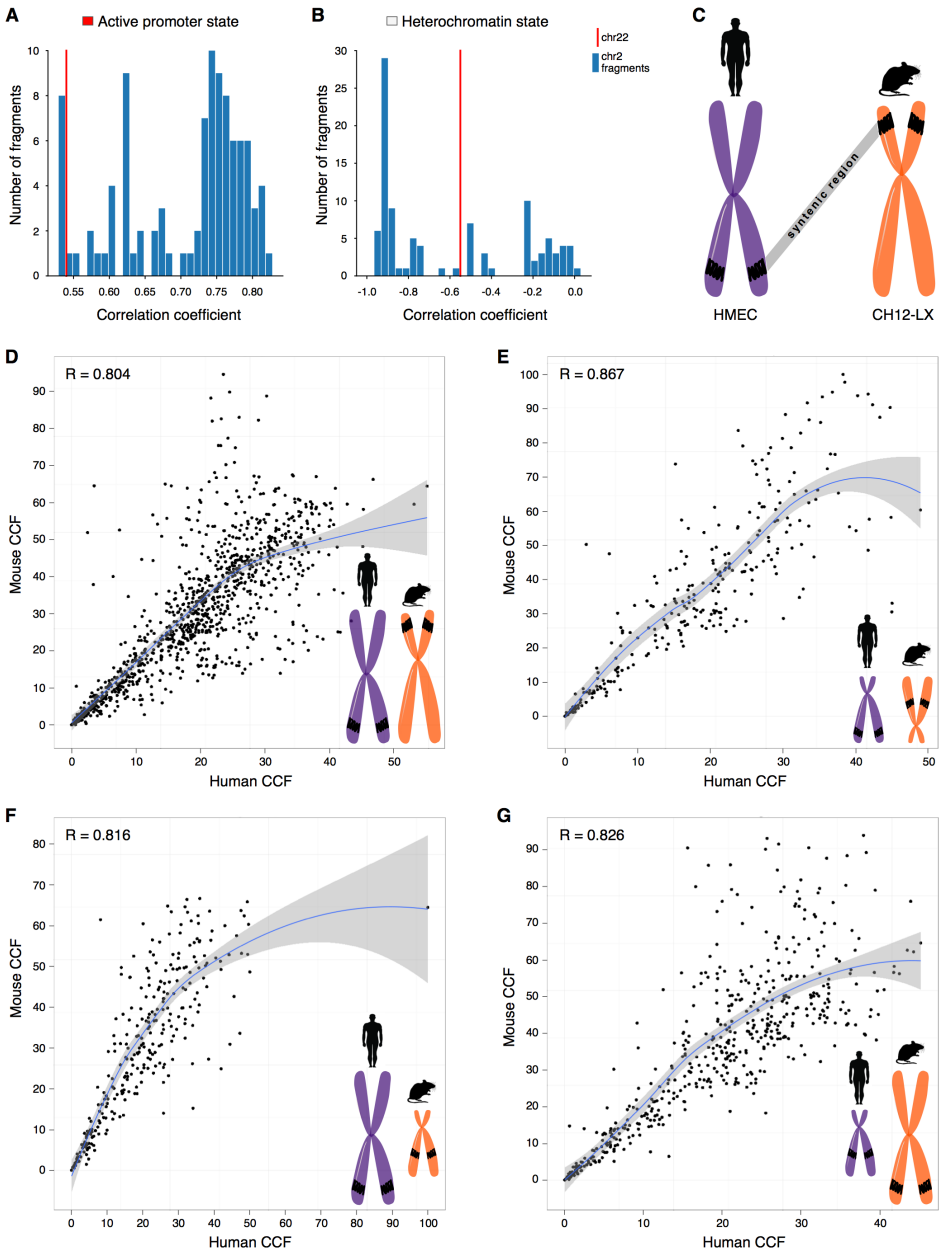


Figure 2 CCF is a position-independent inherent property of chromatin regions conserved in syntenic transitions. (A–B) Distribution of correlations between CCF and active promoter state (A) or heterochromatin state (B) for random fragments of chromosome 2 (blue bars) is compared with the real correlation for chromosome 22 (red line). Human cell line HMEC. (C) Schematic representation of a syntenic region between two chromosomes of the human and mouse genomes. Human cell line HMEC and mouse cell line CH12-LX. All syntenic regions of size 1 Mb are obtained by mapping the mouse genome to the human genome using the Liftover tool. (D–G) CCF in human versus CCF in mouse for syntenic regions in large human chromosomes and large mouse chromosomes (D), small human chromosomes and small mouse chromosomes (E), large human chromosomes and small mouse chromosomes (F), small human chromosomes and large mouse chromosomes (G). Each dot represents a syntenic region (size 1 Mb).

Full-size DOI: 10.7717/peerj.9566/fig-2

To understand how syntenic regions inherit the properties during genomic rearrangements in evolution, we identified syntenic blocks located in small chromosomes of the human genome, but in large chromosomes of the mouse genome, and visa versa. Notably, these regions exhibit similar CCF in the human and mouse genomes (Figs. 2F–2G, the correlation between the contact frequencies in the human and mouse genomes ranges from 0.82 to 0.83). Thus, the observed correlation preferences are intrinsic properties of syntenic blocks as they do not depend on the location of the region in the genome and are inherited despite evolutionary rearrangements between chromosomes (Fig. 2, Fig. S14–Fig. S15).

DISCUSSION

Data normalization is a typical step of Hi-C data processing that corrects hidden biases of the interaction signal (*Lyu, Liu & Wu, 2019*; *Calandrelli et al., 2018*; *Sauria et al., 2015*). One of the most widely used normalization methods is the iterative correction (IC), which assumes equal visibility of each genomic region in the experiment. Various features of Hi-C maps, such as TADs, enriched contacts and compartments, are called after the step of normalization. However, the equal visibility assumption might result in removal of biologically relevant information obtained from Hi-C. We sought to dissociate the technical and biological signal that is removed by IC.

Here, we introduce cumulative contact frequency (CCF) for a genomic region as the number of contacts for a region in a non-normalized Hi-C map. We then analyze CCF properties, including correlation with biologically meaningful signals such as chromatin compartments, transcriptional activity, and chromatin states.

We observe that for human cells, large CCF is predictive of active chromatin and compartment A. This result holds for multiple resolutions of the Hi-C data and several human cell types. We also have used the Stereogene approach (*Stavrovskaya et al., 2017*) to demonstrate that the correlations are reproduced for the subsets of genomic regions.

Moreover, positive correlation of CCF with active chromatin states holds for *Drosophila* and mouse, suggesting broad generalizability of our conclusions. Notably, we use human and mouse Hi-C that were mapped by *Rao et al. (2014)* and *Drosophila* Hi-C that was mapped by *Ulianov et al. (2015)* with different data processing pipelines. We find it striking that the general correlations of CCF are independent of the details of the upstream data processing, which is supportive of the biological importance of CCF. Parallel analysis of CCF properties in multiple cell types demonstrates robustness of the observed correlations, suggests a general similarity between cell types, and further supports the proposed relevance of the CCF signal.

To further separate the biologically relevant signal of CCF from possible technical artifacts, we have considered confounding factors that might affect our analysis. GC-content is a well-known source of variability in the genomic coverage for sequencing experiments, Hi-C, in particular (*Yaffe & Tanay, 2011*). We have demonstrated that CCF is predictive of active chromatin even after the removal of this confounding factor.

One of the first observations obtained using Hi-C method was the tendency of small chromosomes to interact with each other while being more active

(*Lieberman-Aiden & Van Berkum, 2009*). Thus, CCF might be different for chromosomes of different sizes. In order to control for that, we have used CCF normalized by the chromosome size and demonstrated reproducibility of the observed correlation patterns.

Surprisingly, we have observed that CCF of small and large chromosomes differs. We suggest that this difference might happen not because of the chromosome size, but because of the intrinsic properties of the regions. First, we have confirmed it by downsampling large chromosomes to the size of small ones. Second, we have compared CCF in syntenic regions between the human and mouse genomes and observed that CCF does not change after translocation between large and small chromosomes.

There are still some other possible technical confounding factors that might contribute to the CCF properties, such as the density of restriction fragments in a genomic bin, mappability of the region, chromatin openness as assessed by DNase-seq or ATAC-seq (*Yaffe & Tanay, 2011*). These factors remain out of scope of the present research.

Importantly, all these observations do not allow us to introduce a causative link between chromatin activity and CCF. We also do not account for the evolutionary history and sequence conservation of corresponding regions, which might reveal the reasons for our cross-species observations. Extensive further research is required to shed the light on these problems.

Nevertheless, our results allow to suggest that removal of CCF in the IC procedure is currently understudied. CCF contains biologically relevant information that is not affected by GC-content and chromosome size. Currently, the effect of removal of this information on calling of Hi-C features, such as TADs, enriched contacts, and compartments, has not been studied. We propose to take the Hi-C normalization step with caution and interpret Hi-C features that are robust to the removal of CCF and present in both non-normalized and normalized maps.

CONCLUSIONS

In this work, we dissociate the technical and biological signal that is removed by the iterative correction (IC), one of the most widely used methods of Hi-C data normalization. For that, we study cumulative contact frequency (CCF) defined as the number of contacts for a genomic region in a non-normalized Hi-C map. We demonstrate that CCF has significant biological properties, such as correlation with chromatin compartments, transcriptional activity, and active chromatin states. These properties are independent of GC-content and chromosome sizes. They can be generalized to a broad range of species (human, mouse, and *Drosophila*). Surprisingly, these properties are inherited and preserved between syntenic regions of human and mouse genomes. We conclude that the importance of CCF is underestimated, and it should be removed from Hi-C maps with caution.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This study was supported by the Russian Science Foundation grant 19-74-00112 to Ekaterina E. Khrameeva. The research of Aleksandra Galitsyna was supported by the Skoltech Systems Biology Fellowship. There was no additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Russian Science Foundation: 19-74-00112.
Skoltech Systems Biology Fellowship.

Competing Interests

Mikhail S. Gelfand is an Academic Editor for PeerJ.

Author Contributions

- Margarita D. Samborskaia, Aleksandra Galitsyna and Ilya Pletenev performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Anna Trofimova performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Andrey A. Mironov and Mikhail S. Gelfand conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Ekaterina E. Khrameeva conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:
We used publicly available data from GEO: GSE63525 and GSE69013.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.9566#supplemental-information.

REFERENCES

Abdennur N. 2018. Optimal domain segmentation with Lavaburst. *Available at https://github.com/nvictus/lavaburst* (accessed on 28 December 2019).

Abdennur N, Mirny LA. 2019. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36(1)**:311–316 DOI 10.1093/bioinformatics/btz540.

Armstrong RL, Penke TJ, Strah lBD, Matera, AG, McKay DJ, MacAlpine DM, Duronio RJ. 2018. Chromatin conformation and transcriptional activity are permissive regulators of DNA replication initiation in Drosophila. *Genome Research* **28(11)**:1688–1700 DOI 10.1101/gr.239913.118.

Beagrie RA, Scialdone A, Schueler M, Dorothee CA. 2017. Complex multi-enhancer contacts captured by Genome Architecture Mapping (GAM). *Nature* **543(7646)**:519–524 DOI 10.1038/nature21411.

Bickmore WA. 2013. The spatial organization of the human genome. *Annual Review of Genomics and Human Genetics* **14(1)**:67–84 DOI 10.1146/annurev–genom–091212–153515.

Calandrelli R, Wu Q, Guan J, Zhong S. 2018. GITAR: an open source tool for analysis and visualization of Hi-C data. *Genomics, Proteomics and Bioinformatics* **16(5)**:365–372 DOI 10.1016/j.gpb.2018.06.006.

Chandradoss KR, Guthikonda PK, Kethavath S. 2020. Biased visibility in HiC datasets marks dynamically regulated condensed and decondensed chromatin states genome-wide. *BMC Genomics* **21(1)**:–175 DOI 10.1186/s12864-020-6580-6.

Cremer T, Cremer M, Dietzel S, Müller S, Solovei I, Fakan S. 2006. Chromosome territories—a functional nuclear landscape. *Current Opinion in Cell Biology* **18(3)**:307–316 DOI 10.1016/j.ceb.2006.04.007.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473(7345)**:43–49 DOI 10.1038/nature09906.

Filippova D, Patro R, Duggal G, Kingsford C. 2014. Identification of alternative topological domains in chromatin. *Algorithms for Molecular Biology* **9**:14 DOI 10.1186/1748–7188–9–14.

Finger JH, Smith CM, Hayamizu TF, McCright IJ, Eppig JT, Kadin JA, Richardson JE, Ringwald M. 2011. The mouse Gene Expression Database (GXD): 2011 update. *Nucleic Acids Research* **39(SUPPL. 1)**:835–841 DOI 10.1093/nar/gkq1132.

Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S. 2017. Comparison of computational methods for Hi-C data analysis. *Nature Publishing Group* **14(7)**:679–685 DOI 10.1038/nmeth.4325.

Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, Grossman SR, Anyoha R, Doughty BR, Patwardhan TA, Nguyen TH, Kane M, Perez EM, Durand NC, Lareau CA, Stamenova EK, Aiden EL, Lander ES, Engreitz JM. 2019. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics* **51(12)**:1664–1669 DOI 10.1038/s41588-019-0538-0.

Galitsyna AA, Khrameeva EE, Razin SV, Gelfand MS, Gavrilov AA. 2017. "Mirror reads" in Hi-C data. *Genomics and Computational Biology* **3(1)**:36 DOI 10.18547/gcb.2017.vol3.iss1.e36.

Hinrichs AS. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research* **34(suppl_1)**:D590–D598 DOI 10.1093/nar/gkj144.

Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. - Supplement. *Nature Methods* **9(10)**:999–1003 DOI 10.1038/nmeth.2148.

Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T, Linder-Basso D, Plachetka A, Shanower G, Tolstorukov MY, Luquette LJ, Xi R, Jung YL, Park RW, Bishop EP, Canfield TK, Sandstrom R, Thurman RE, MacAlpine DM, Stamatoyannopoulos JA, Kellis M, Elgin S. CR, Kuroda MI, Pirrotta V, Karpen GH, Park PJ. 2011. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471(7339)**:480–485 DOI 10.1038/nature09725.

Khrameeva EE, Mironov AA, Fedonin GG, Khaitovich P, Gelfand MS. 2012. Spatial proximity and similarity of the epigenetic state of genome domains. *PLOS ONE* **7(4)**:e33947 DOI 10.1371/journal.pone.0033947.

Lajoie BR, Dekker J, Kaplan N. 2014. The Hitchhiker's Guide to Hi-C analysis: practical guidelines. *Methods* **72**:65–75 DOI 10.1016/j.ymeth.2014.10.031.

Lieberman-Aiden E, Van Berkum N. 2009. Comprehensive mapping of long range interactions reveals folding principles of the human genome. *Science* **326(5950)**:289–293 DOI 10.1126/science.1181369.

Lyu H, Liu E, Wu Z. 2019. Comparison of normalization methods for Hi-C data. *BioTechniques* **68(2)**:56–64 DOI 10.2144/btn-2019-0105.

Rao S. SP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159(7)**:1665–1680 DOI 10.1016/j.cell.2014.11.021.

Sauria ME, Phillips-Cremins JE, Corces VG, Taylor J. 2015. HiFive: a tool suite for easy and efficient HiC and 5C data analysis. *Genome Biology* **16**:237 DOI 10.1186/s13059-015-0806-y.

Schmitt AD, Hu M, Ren B. 2016. Genome-wide mapping and analysis of chromosome architecture. *Nature Reviews Molecular Cell Biology* **17(12)**:743–755 DOI 10.1038/nrm.2016.104.

Sexton T, Schober H, Fraser P, Gasser SM. 2007. Gene regulation through nuclear organization. *Nature Structural and Molecular Biology* **14(11)**:1049–1055 DOI 10.1038/nsmb1324.

Spill YG, Castillo D, Vidal E, Marti-Renom MA. 2019. Binless normalization of Hi-C data provides significant interaction and difference detection independent of resolution. *Nature Communications* **10(1)**:1938 DOI 10.1038/s41467-019-09907-2.

Stavrovskaya ED, Niranjan T, Fertig EJ, Wheelan SJ, Favorov AV, Mironov AA. 2017. StereoGene: Rapid estimation of genome-wide correlation of continuous or interval feature data. *Bioinformatics* **33(20)**:3158–3165 DOI 10.1093/bioinformatics/btx379.

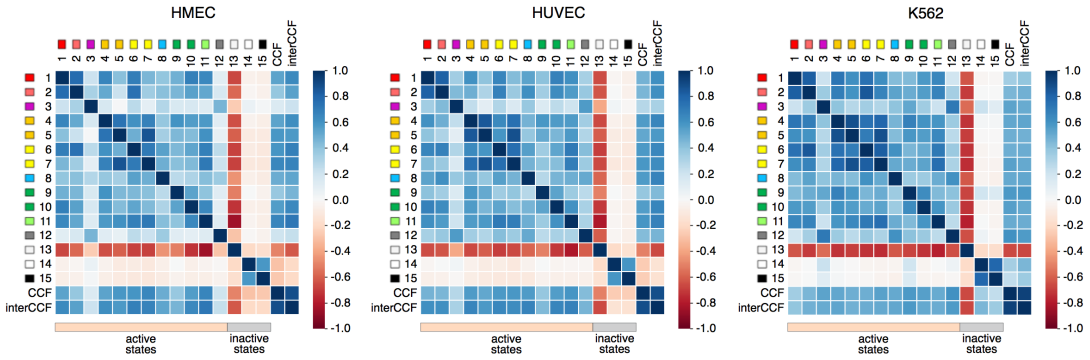
Ulianov SV, Khrameeva EE, Gavrilov AA, Flyamer IM, Kos P, Mikhaleva EA, Penin AA, Logacheva MD, Imakaev MV, Chertovich A, Gelfand MS, Shevelyov YY, Razin SV. 2015. Active chromatin and transcription play a key role in chromosome

partitioning into topologically associating domains. *Genome Research* **26(1)**:70–84 DOI 10.1101/gr.196006.115.

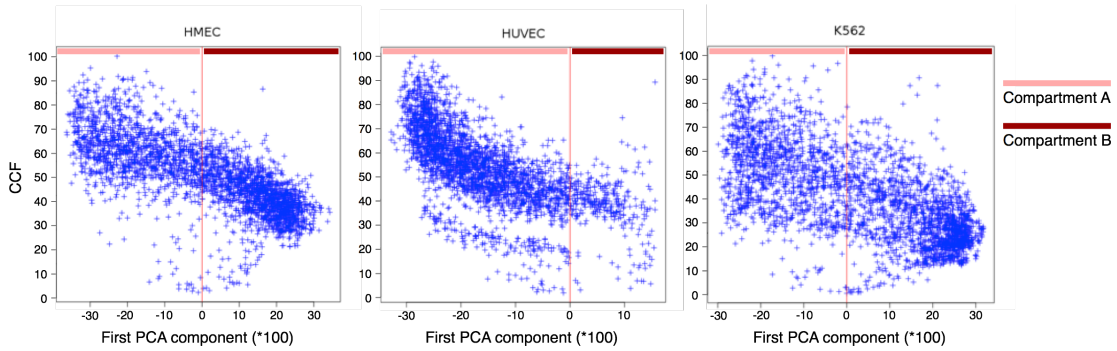
Yaffe E, Tanay A. 2011. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics* **43(11)**:1059–1065 DOI 10.1038/ng.947.

Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD, Shen Y, Pervouchine DD, Djebali S, Thurman RE, Kaul R, Rynes E, Kirilusha A, Marinov GK, Williams BA, Trout D, Amrhein H, Fisher-Aylor K, Antoshechkin I, DeSalvo G, See L.-H, Fastuca M, Drenkow J, Zaleski C, Dobin A, Prieto P, Lagarde J, Bussotti G, Tanzer A, Denas O, Li K, Bender MA, Zhang M, Byron R, Groudine MT, McCleary D, Pham L, Ye Z, Kuan S, Edsall L, Wu Y.-C, Rasmussen MD, Bansal MS, Kellis M, Keller CA, Morrissey CS, Mishra T, Jain D, Dogan N, Harris RS, Cayting P, Kawli T, Boyle AP, Euskirchen G, Kundaje A, Lin S, Lin Y, Jansen C, Malladi VS, Cline MS, Erickson DT, Kirkup VM, Learned K, Sloan CA, Rosenbloom KR, Lacerda de Sousa B, Beal K, Pignatelli M, Flicek P, Lian J, Kahveci T, Lee D, James Kent W, Ramalho Santos M, Herrero J, Notredame C, Johnson A, Vong S, Lee K, Bates D, Neri F, Diegel M, Canfield T, Sabo PJ, Wilken MS, Reh TA, Giste E, Shafer A, Kutyavin T, Haugen E, Dunn D, Reynolds AP, Neph S, Humbert R, Scott Hansen R, De Bruijn M, Selleri L, Rudensky A, Josefowicz S, Samstein R, Eichler EE, Orkin SH, Levasseur D, Papayannopoulou T, Chang K.-H, Skoultchi A, Gosh S, Disteche C, Treuting P, Wang Y, Weiss MJ, Blobel GA, Cao X, Zhong S, Wang T, Good PJ, Lowdon RF, Adams LB, Zhou X.-Q, Pazin MJ, Feingold EA, Wold B, Taylor J, Mortazavi A, Weissman SM, Stamatoyannopoulos JA, Snyder MP, Guigo R, Gingeras TR, Gilbert DM, Hardison RC, Beer MA, Ren B, Consortium T. ME. 2014. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515(7527)**:355–364 DOI 10.1038/nature13992.

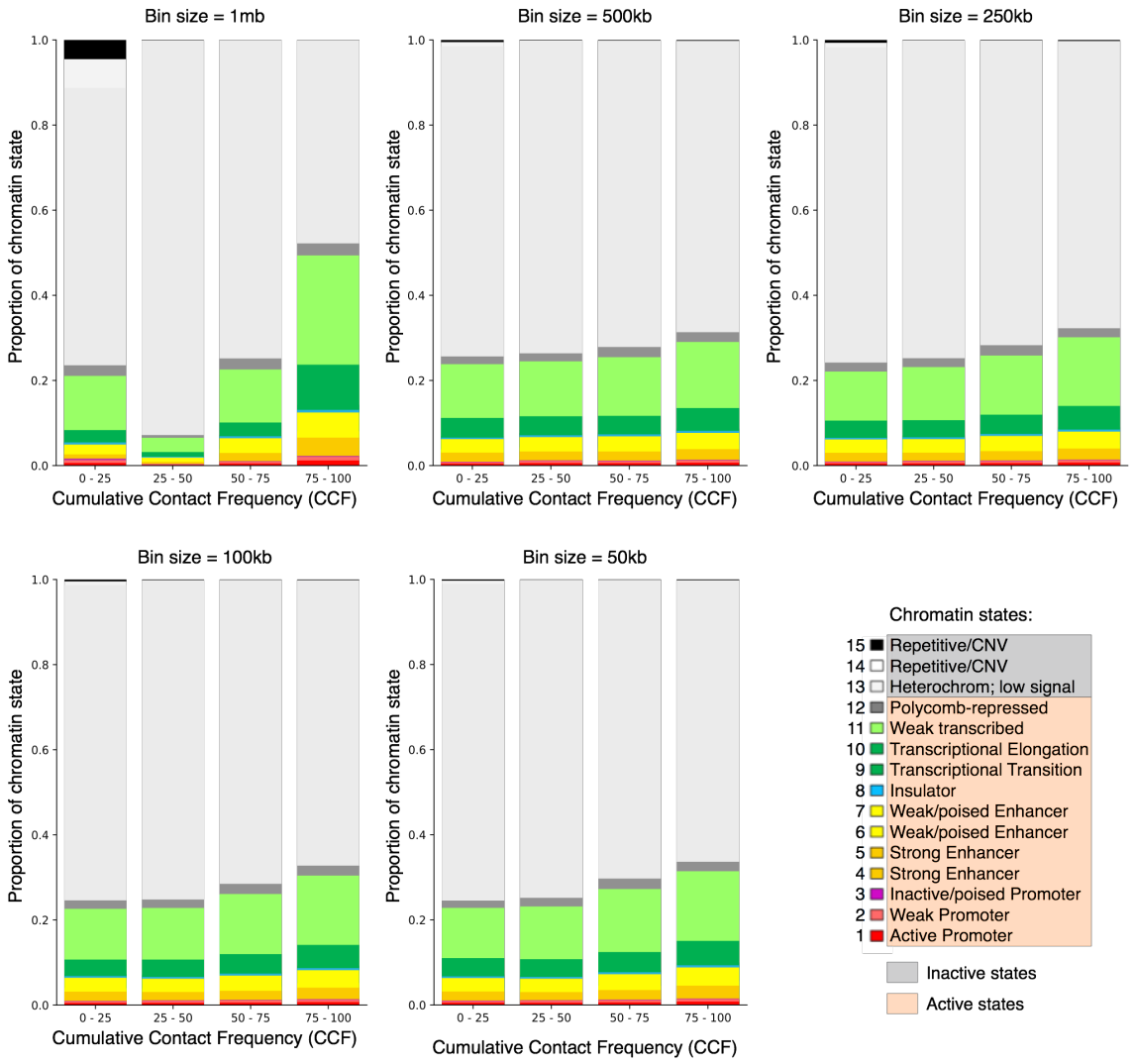
Supplementary Materials



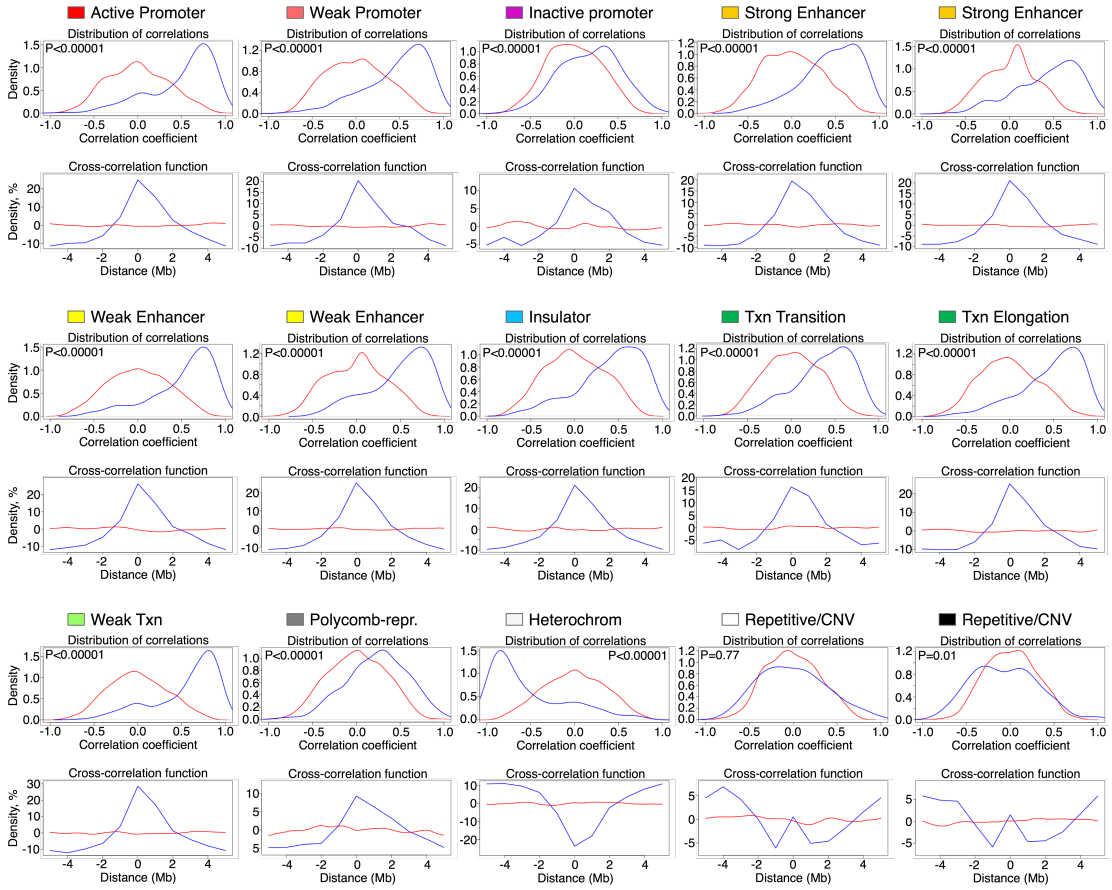
Supplemental Figure S1. Whole-genome correlation patterns for three human cell lines. Chromatin states for the human cell lines are correlated with the total CCF and inter-chromosomal CCF. First 15 rows in the matrix correspond to the 15 chromatin states, rows 16-17 exhibit whole-genome and inter-chromosomal CCF. Colors show the Pearson correlation coefficients. Note that correlation patterns are similar for whole-genome and inter-chromosomal CCF.



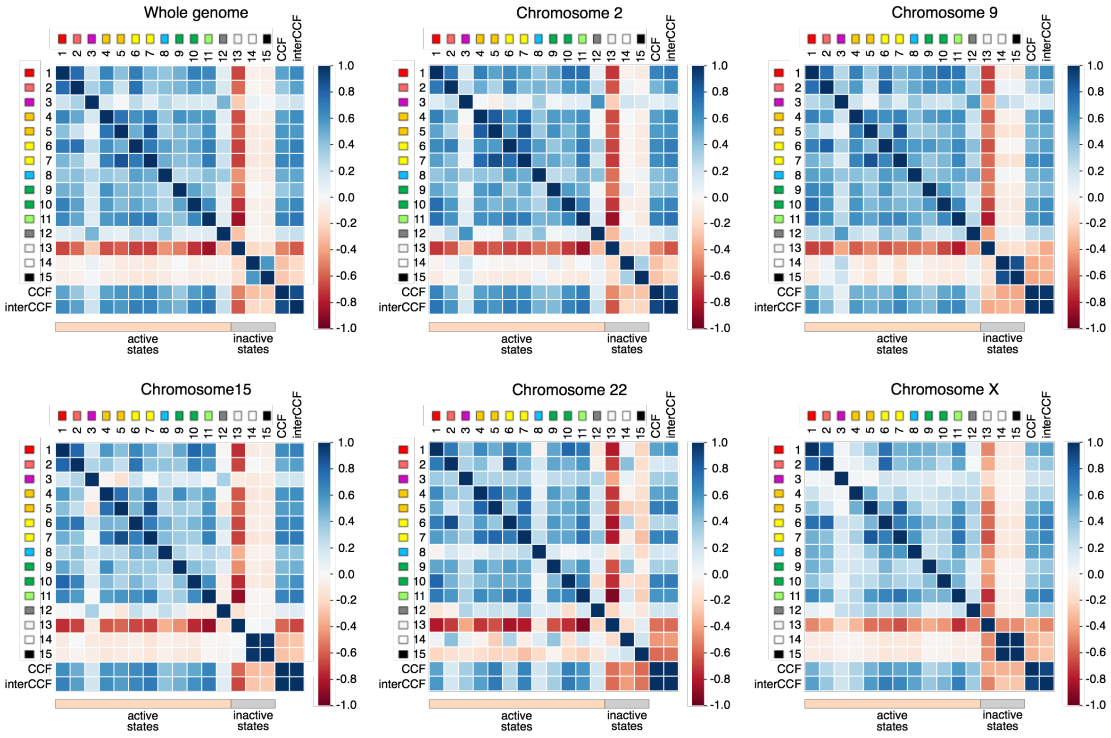
Supplemental Figure S2. Dependency of CCF on the first principal component. The first principal component was calculated using PCA analysis of the Hi-C maps at 1 Mb resolution in three human cell lines: HMEC, HUVEC, and K562.



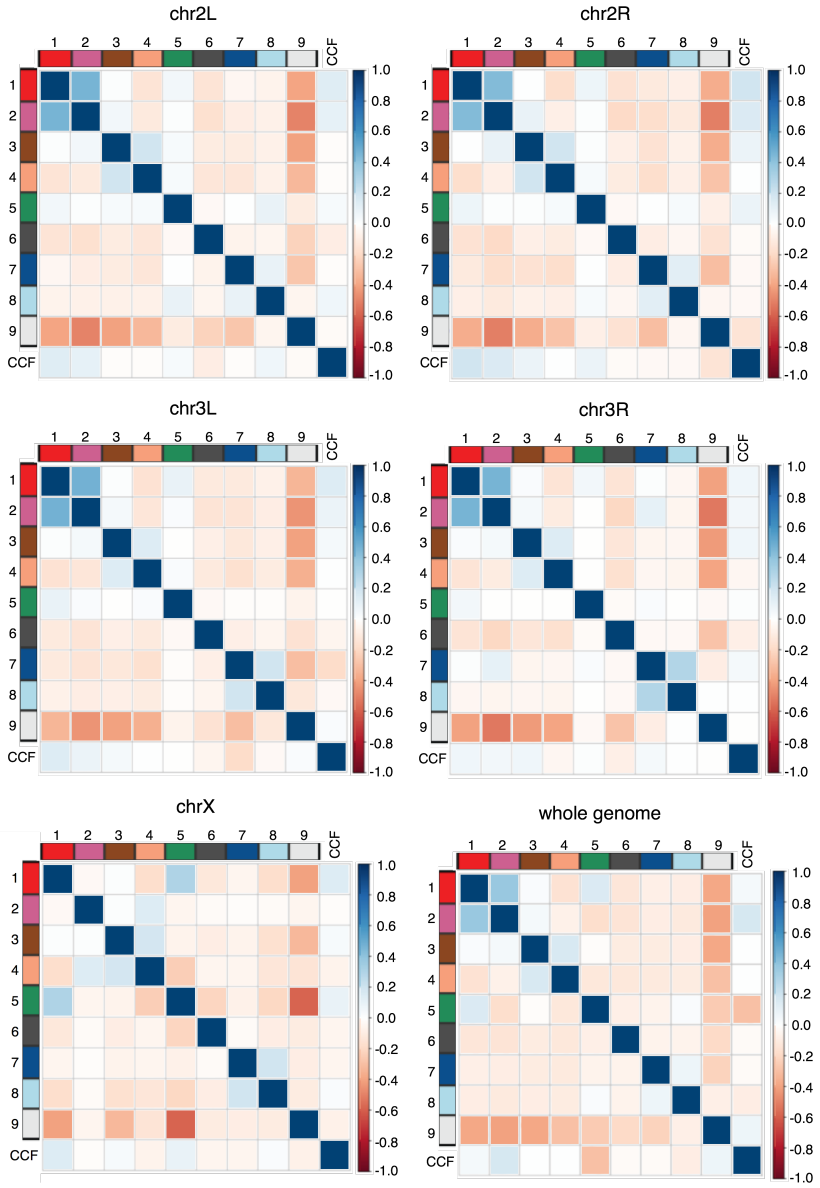
Supplemental Figure S3. Dependency of chromatin state proportions on CCF. CCF and chromatin states were calculated for five resolutions (bin sizes) – 1 Mb, 500 Kb, 250 Kb, 100 Kb and 50 Kb. Human cell line HMEC.



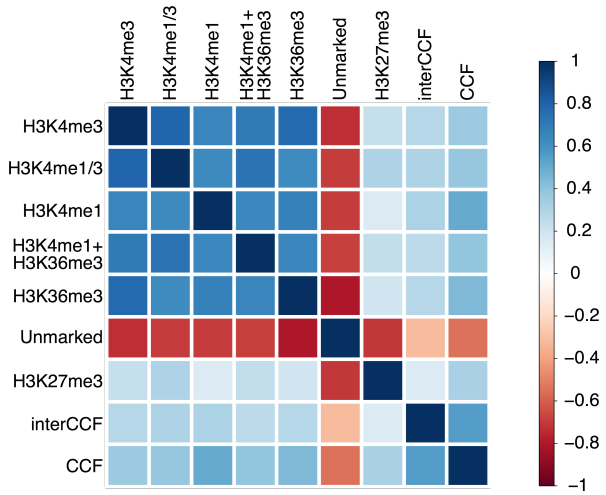
Supplemental Figure S4. Validation of correlations between chromatin states 1 to 15 and whole-genome CCF with the Stereogene tool. The distribution of real correlations (blue line) is compared with randomly selected windows (red line). Stereogene parameters except for the window size (size 10 Mb) were set to default. For each chromatin state, the top panel shows the distribution of correlations, while the bottom panel shows the cross-correlation function. Human cell line HMEC.



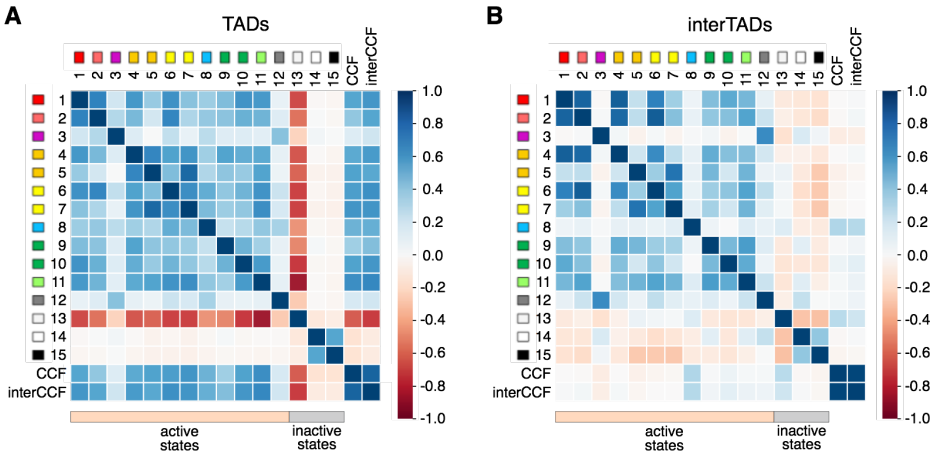
Supplemental Figure S5. Correlation patterns for different human chromosomes. Note that correlation patterns are similar for large chromosomes, but are different for small chromosomes. Human cell line HMEC.



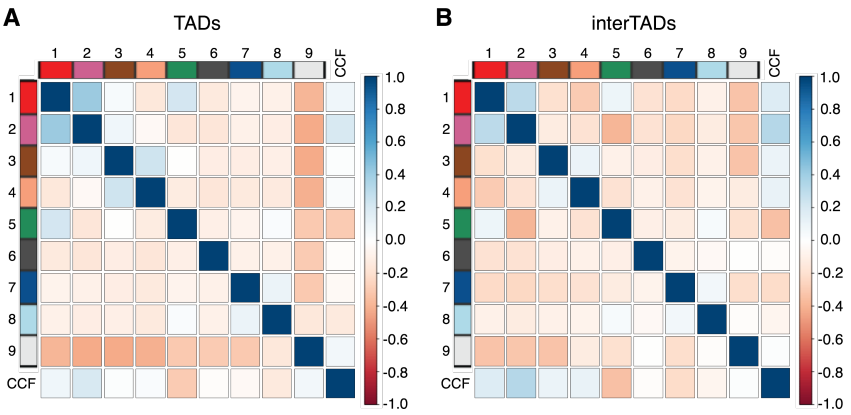
Supplemental Figure S6. Correlation patterns of whole-genome CCF and chromatin states in *Drosophila*. The correlation patterns are similar for the whole genome and individual chromosomes. *Drosophila* cell line S2-DGRC.



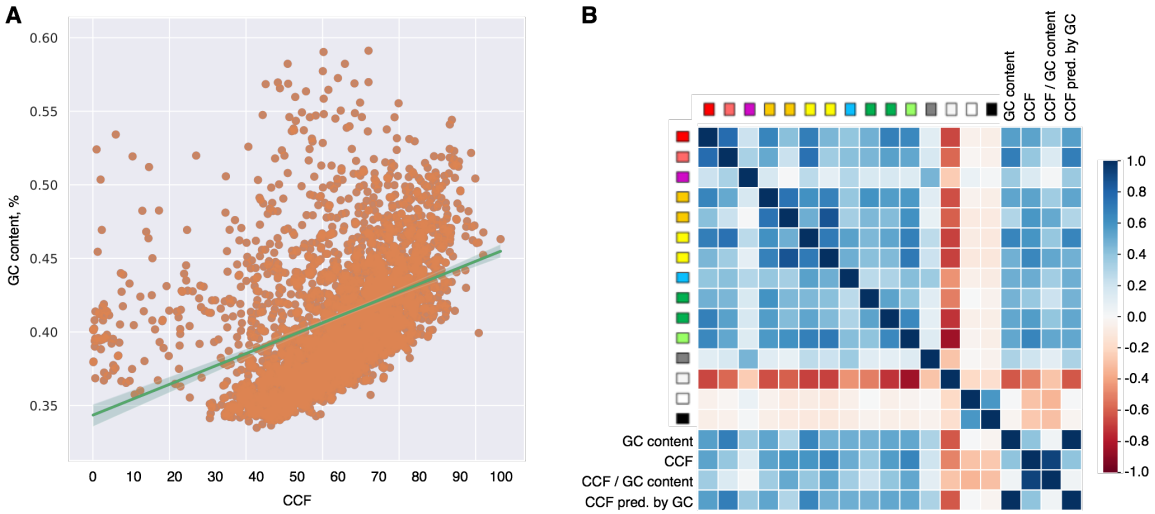
Supplemental Figure S7. Correlation patterns of whole-genome CCF, interCCF, and chromatin states in mouse. Mouse cell line CH12-LX. Resolution of the Hi-C map is 1 Mb.



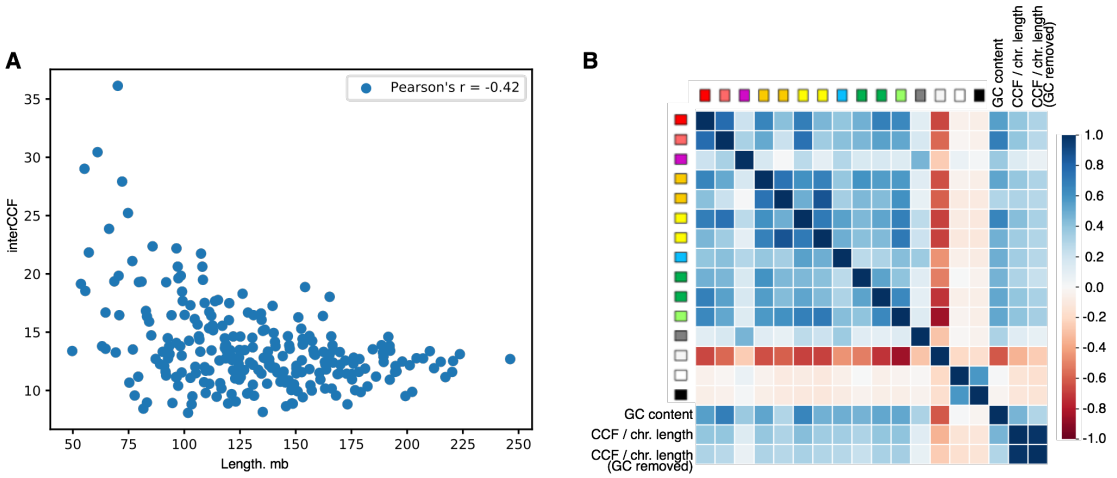
Supplemental Figure S8. Correlation patterns for TAD and interTAD regions in human. (A) Correlation patterns between two types of CCF and chromatin states for the whole-genome TADs (found using the *Armatus* algorithm, *gamma* = 1.0). (B) Correlation patterns for the whole-genome interTADs (all regions between TADs found using the *Armatus* algorithm, *gamma* = 1.0). Human cell line HMEC.



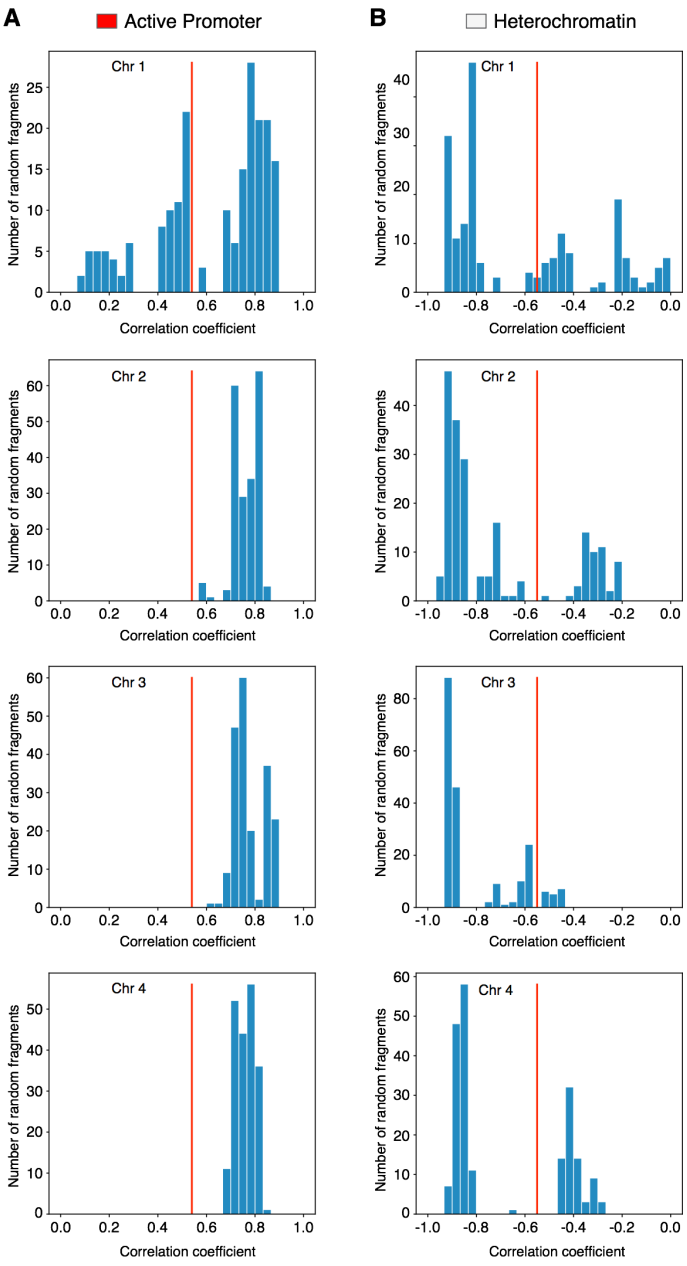
Supplemental Figure S9. Correlation patterns for TAD and interTAD regions in *Drosophila*. (A) Correlation patterns between whole-genome CCF and chromatin states (states 1 to 9) for the whole-genome TADs (found using the Armatus algorithm, *gamma* = 1.0). (B) Correlation patterns for the whole-genome interTADs (all regions between TADs found using the Armatus algorithm, *gamma* = 1.0). *Drosophila* cell line S2-DGRC.



Supplemental Figure S10. Whole-genome correlation patterns are not driven by GC-content. (A) GC-content and whole-genome CCF are highly correlated (Pearson's R=0.41). Orange circles show GC-content values in each genomic 1 Mb bin. The green line represents the linear regression, which was further used to predict CCF in B. The area around the line represents the confidence interval. (B) The Pearson correlation coefficients between the 15 chromatin states, GC-content, CCF, CCF divided by the GC-content, and CCF predicted by the GC-content using the linear regression shown in A.

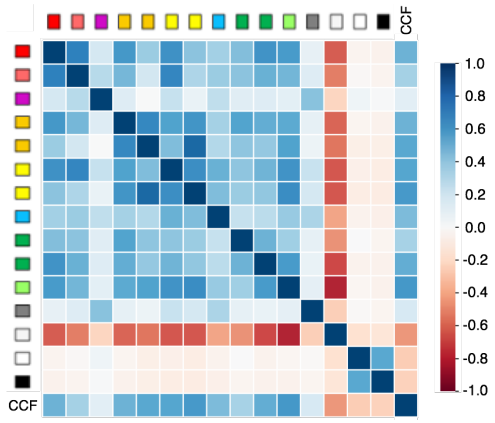


Supplemental Figure S11. Whole-genome correlation patterns are not driven by chromosome length. (A) Correlation of inter-chromosomal CCF with the chromosome length (an average length of two interacting chromosomes). The Pearson correlation coefficient is specified on the plot. (B) The Pearson correlation coefficients between the 15 chromatin states, GC-content, CCF normalized by the chromosome length, and CCF normalized by the chromosome length with subsequent removal of the GC-content dependency by additional normalization for GC-content. Human cell line HMEC.

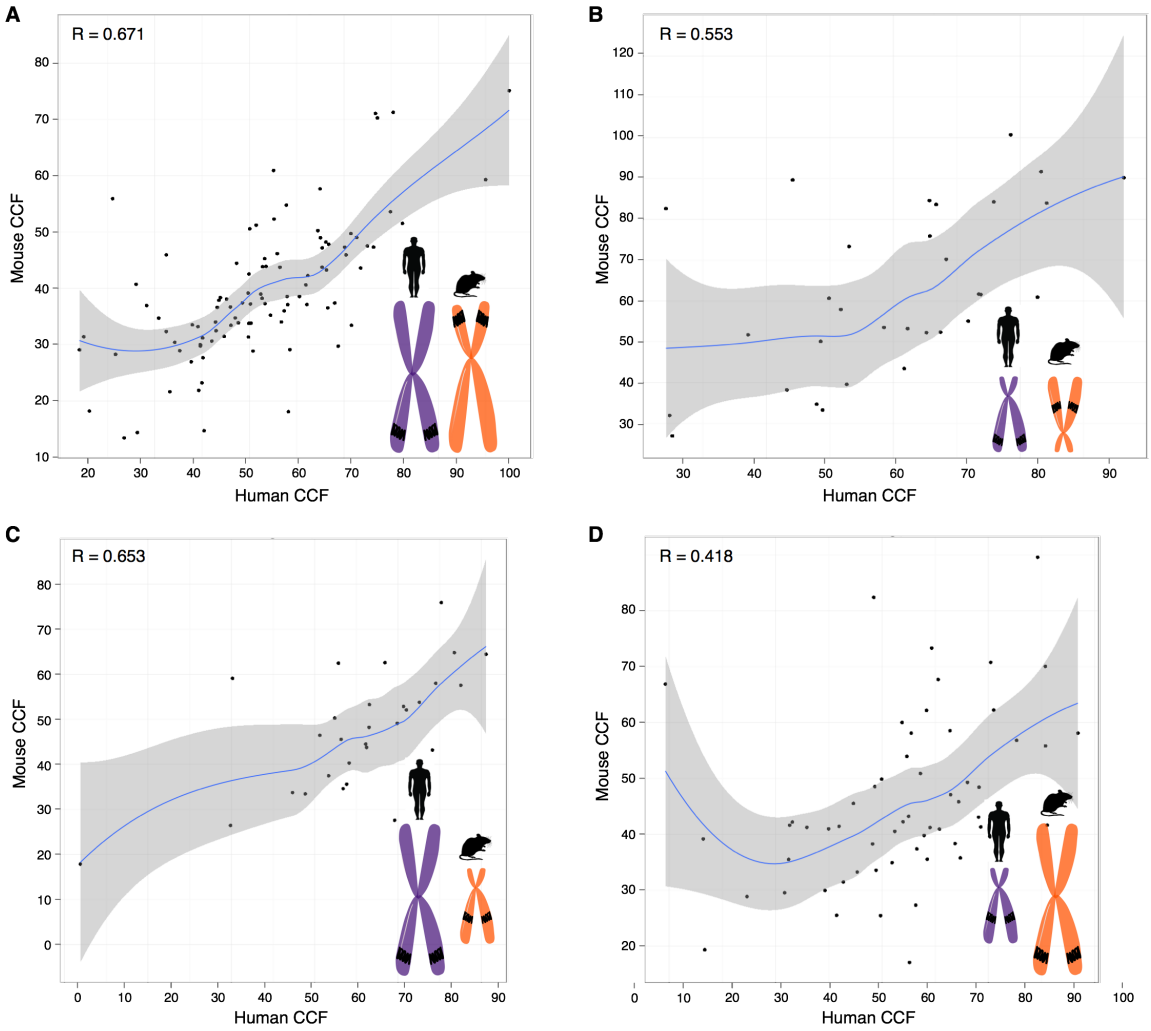


Supplemental Figure S12. Correlation for the human chromosome 22 is compared with correlations for random fragments of a large chromosome equal in length (50 Mb). (A) Distribution of correlations between inter-chromosomal CCF and active promoter chromatin state for random fragments of chromosomes 1-4 (blue histograms) compared with the real correlation for the chromosome 22 (red line). (B) Distribution of correlations between inter-chromosomal CCF and heterochromatin state for random fragments of chromosomes 1-4 (blue

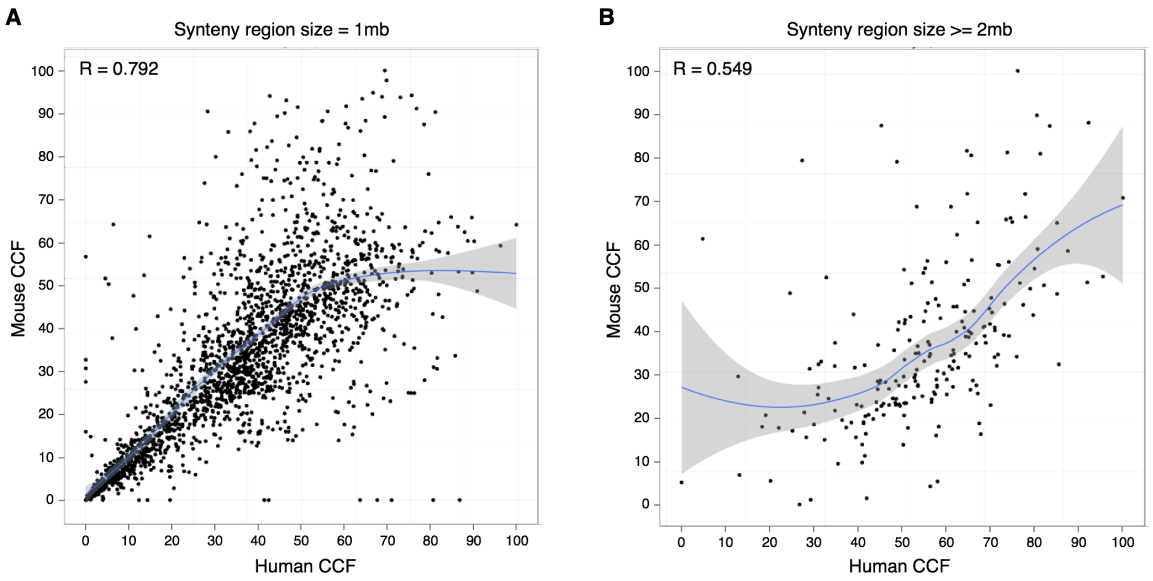
histograms) compared with the real correlation for the chromosome 22 (red line). Human cell line HMEC.



Supplemental Figure S13. Correlation patterns for the whole human genome with centromeres excluded. Correlation patterns between total CCF and chromatin states 1 to 15 for the whole genome with centromeres excluded at each chromosome. Human cell line HMEC.



Supplemental Figure S14. Contact frequency for chromatin regions is conserved in syntenic transitions between the mouse and human genomes. (A) CCF in human versus CCF in mouse for syntenic regions in large human chromosomes and large mouse chromosomes. (B) CCF in human versus CCF in mouse for syntenic regions in small human chromosomes and small mouse chromosomes. (C) CCF in human versus CCF in mouse for syntenic regions in large human chromosomes and small mouse chromosomes. (D) CCF in human versus CCF in mouse for syntenic regions in small human chromosomes and large mouse chromosomes. Each dot represents a syntenic region of size 2 Mb and larger obtained from the Mouse Genome Informatics database (MGI). Cell lines HMEC (human) and CH12-LX (mouse). Pearson's R is specified on each plot.



Supplemental Figure S15. Contact frequency for syntenic regions in the mouse and human genomes. CCF in human versus CCF in mouse is shown. (A) All syntenic regions are obtained by mapping the mouse genome to the human genome using the Liftover tool. Each dot represents a syntenic region (size 1 Mb). (B) All syntenic regions of size 2 Mb and larger obtained from the Mouse Genome Informatics database (MGI) are displayed. Each dot represents a syntenic region. Cell lines HMEC (human) and CH12-LX (mouse).

Chapter 7

Order and stochasticity in the folding of individual *Drosophila* genomes

The previous studies developed our understanding of bulk Hi-C data and the primary factors affecting chromatin structure in *Drosophila* cell population. These results guided my thought and experiments in the major project dedicated to single-cell chromatin structure analysis.

The whole project was initiated by the group of Prof. Razin, who obtained the raw data at IGB RAS and performed single-cell Hi-C (scHi-C) *Drosophila* cells. In fact, it was single-nucleus Hi-C (snHi-C), which is sometimes referred to as scHi-C, by the family name of similar experimental techniques.

When we acquired the data, four works on scHi-C existed, all of them in mammals and produced with different protocols. Thus, my role was to design the computational experiments and the tools for them from scratch. The complete description of my research path can be found in the main text of this chapter, Methods (starting from "snHi-C raw data processing and contact annotation" to "Robustness of TAD calling") and Supplementary materials (Figures 1 to 20, excluding Figure 4e-k and 7c-d).

This study had important theoretical outcomes. First of all, we discovered compartments and TADs in individual cells of *Drosophila*. Importantly, they are present when we compare to random background model that accounts for the marginal distribution of contacts, or CCF. The following surprising conclusion is the orderliness of TADs as compared to mammalian cells. TADs boundaries tend to be located

at the same positions in individual cells, which coincides with enrichment of active chromatin factors and depletion of inactive ones. Moreover, active chromatin partakes in long-range interactions, which are substantially variable between individual cells. It allowed us to conclude that stable TADs are formed of inactive chromatin, which has an increased number of short-range interactions.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-20292-z OPEN

Order and stochasticity in the folding of individual *Drosophila* genomes

Sergey V. Ulianov[1,2,16], Vlada V. Zakharova[1,2,3,16], Aleksandra A. Galitsyna[4,16], Pavel I. Kos[5,16], Kirill E. Polovnikov[4,6], Ilya M. Flyamer[7], Elena A. Mikhaleva[8], Ekaterina E. Khrameeva[4], Diego Germini[3], Mariya D. Logacheva[4], Alexey A. Gavrilov[1,9], Alexander S. Gorsky[10,11], Sergey K. Nechaev[12,13], Mikhail S. Gelfand[4,10], Yegor S. Vassetzky[3,14], Alexander V. Chertovich[5,15], Yuri Y. Shevelyov[8] & Sergey V. Razin[1,2]✉

Mammalian and *Drosophila* genomes are partitioned into topologically associating domains (TADs). Although this partitioning has been reported to be functionally relevant, it is unclear whether TADs represent true physical units located at the same genomic positions in each cell nucleus or emerge as an average of numerous alternative chromatin folding patterns in a cell population. Here, we use a single-nucleus Hi-C technique to construct high-resolution Hi-C maps in individual *Drosophila* genomes. These maps demonstrate chromatin compartmentalization at the megabase scale and partitioning of the genome into non-hierarchical TADs at the scale of 100 kb, which closely resembles the TAD profile in the bulk in situ Hi-C data. Over 40% of TAD boundaries are conserved between individual nuclei and possess a high level of active epigenetic marks. Polymer simulations demonstrate that chromatin folding is best described by the random walk model within TADs and is most suitably approximated by a crumpled globule build of Gaussian blobs at longer distances. We observe prominent cell-to-cell variability in the long-range contacts between either active genome loci or between Polycomb-bound regions, suggesting an important contribution of stochastic processes to the formation of the *Drosophila* 3D genome.

[1] Institute of Gene Biology, Russian Academy of Sciences, Moscow, Russia. [2] Faculty of Biology, M.V. Lomonosov Moscow State University, Moscow, Russia. [3] UMR9018, CNRS, Université Paris-Sud Paris-Saclay, Institut Gustave Roussy, Villejuif, France. [4] Skolkovo Institute of Science and Technology, Moscow, Russia. [5] Faculty of Physics, M.V. Lomonosov Moscow State University, Moscow, Russia. [6] Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [7] MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [8] Institute of Molecular Genetics, National Research Centre "Kurchatov Institute", Moscow, Russia. [9] Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Institute of Gene Biology, Russian Academy of Sciences, Moscow, Russia. [10] Institute for Information Transmission Problems (the Kharkevich Institute), Russian Academy of Sciences, Moscow, Russia. [11] Moscow Institute for Physics and Technology, Dolgoprudnyi, Russia. [12] Interdisciplinary Scientific Center Poncelet (CNRS UMI 2615), Moscow, Russia. [13] P.N. Lebedev Physical Institute, Russian Academy of Sciences, Moscow, Russia. [14] Koltzov Institute of Developmental Biology, Russian Academy of Sciences, Moscow, Russia. [15] Semenov Federal Research Center for Chemical Physics, Moscow, Russia. [16] These authors contributed equally: Sergey V. Ulianov, Vlada V. Zakharova, Aleksandra A. Galitsyna, Pavel I. Kos. ✉email: sergey.v.razin@usa.net

The principles of higher-order chromatin folding in the eukaryotic cell nucleus have been disclosed thanks to the development of chromosome conformation capture techniques, or C-methods[1,2]. High-throughput chromosome conformation capture (Hi-C) studies demonstrated that chromosomal territories were partitioned into partially insulated topologically associating domains (TADs)[3–5]. TADs likely coincide with functional domains of the genome[6–8], although the results concerning the role of TADs in the transcriptional control are still conflicting[6,9–12]. Analysis performed at low resolution suggested that active and repressed TADs were spatially segregated within A and B chromatin compartments[13,14]. However, high-resolution studies demonstrated that the genome was partitioned into relatively small compartmental domains bearing distinct chromatin marks and comparable in sizes with TADs[15]. In mammals, the formation of TADs by active DNA loop extrusion partially overrides the profile of compartmental domains[15,16]. Of note, TADs identified in studies of cell populations are highly hierarchical (i.e., comprising smaller subdomains, some of which are represented by DNA loops[5,17]).

Partitioning of the genome into TADs is relatively stable across cell types of the same species[3,4]. The recent data suggest that mammalian TADs are formed by active DNA loop extrusion[18,19]. The boundaries of mammalian TADs frequently contain convergent binding sites for the insulator protein CTCF that are thought to block the progression of loop extrusion[19–21]. Contribution of DNA loop extrusion in the assembly of *Drosophila* TADs has not been demonstrated yet[22]; thus, *Drosophila* TADs might represent pure compartmental domains[23]. Large TADs in the *Drosophila* genome are mostly inactive and are separated by transcribed regions characterized by the presence of a set of active histone marks, including hyperacetylated histones[5,24]. Some insulator/architectural proteins are also overrepresented in *Drosophila* TAD boundaries[24–26], but their contribution to the formation of these boundaries has not been directly tested. The results of computer simulations suggest that *Drosophila* TADs are assembled by the condensation of nucleosomes of inactive chromatin[24].

The current view of genome folding is based on the population Hi-C data that present integrated interaction maps of millions of individual cells. It is not clear, however, whether and to what extent the 3D genome organization in individual cells differs from this population average. Even the existence of TADs in individual cells may be questioned. Indeed, the DNA loop extrusion model considers TADs as a population average representing a superimposition of various extruded DNA loops in individual cells[18]. Heterogeneity in patterns of epigenetic modifications and transcriptomes in single cells of the same population was shown by different single-cell techniques, such as single-cell RNA-seq[27], ATAC-seq[28], and DNA-methylation analysis[29]. Studies performed using FISH demonstrated that the relative positions of individual genomic loci varied significantly in individual cells[30]. The first single-cell Hi-C study captured a low number of unique contacts per individual cell[31] and allowed only the demonstration of a significant variability of DNA path at the level of a chromosome territory. Improved single-cell Hi-C protocols[32,33] allowed to achieve single-cell Hi-C maps with a resolution of up to 40 kb per individual cell[32,34] and investigate local and global chromatin spatial variability in mammalian cells, driven by various factors, including cell cycle progression[33]. Of note, TAD profiles directly annotated in individual cells demonstrated prominent variability in individual mouse cells[32]. The possible contribution of stochastic fluctuations of captured contacts in sparse single-cell Hi-C matrices into this apparent variability was not analyzed[32]. More comprehensive observations were made when super-resolution microscopy (Hi-M, 3D-SIM) coupled with high-throughput hybridization was used to analyze chromatin folding in individual cells at a kilobase-scale resolution. These studies demonstrated chromosome partitioning into TADs in individual mammalian cells and confirmed a trend for colocalization of CTCF and cohesin at TAD boundaries, although the positions of boundaries again demonstrated significant cell-to-cell variability[35]. Condensed chromatin domains coinciding with population TADs were also observed in *Drosophila* cells[36,37]. In accordance with previous observations made in cell population Hi-C studies[24], the obtained results suggested that partitioning of the *Drosophila* genome into TADs was driven by the stochastic contacts of chromosome regions with similar epigenetic states at different folding levels[38].

Although studies performed using FISH and multiplex hybridization allowed to construct chromatin interaction maps with a very high resolution[35], they cannot provide genome-wide information. Here, we present single-nucleus Hi-C (snHi-C) maps of individual *Drosophila* cells with a 10-kb resolution. These maps allow direct annotation of TADs that appear to be non-hierarchical and are remarkably reproducible between individual cells. TAD boundaries conserved in different cells of the population bear a high level of active chromatin marks supporting the idea that active chromatin might be among determinants of TAD boundaries in *Drosophila*[24].

Results

High-resolution single-nucleus Hi-C reveals distinct TADs in *Drosophila* genome. To investigate the nature of TADs in single cells and to characterize individual cell variability in *Drosophila* 3D genome organization, we performed single-nucleus Hi-C (snHi-C)[32] (Fig. 1a) in 88 asynchronously growing *Drosophila* male Dm-BG3c2 (BG3) cells (Supplementary Fig. 1a) in parallel with the bulk BG3 in situ Hi-C analysis and obtained 2–5 million paired-end reads per single-cell library (for the data processing workflow, see Supplementary Fig. 1b). To select the libraries for deep sequencing, we subsampled the snHi-C data to estimate the expected number of unique contacts that could be extracted from the data (Supplementary Fig. 2a; also see "Methods"). Twenty libraries were additionally sequenced with 16.7–36.5 million paired-end reads, and we extracted 8032–107,823 unique contacts per cell (Supplementary Table 1). We developed a custom *pairtools*-based approach termed ORBITA (One Read-Based Interaction Annotation) (Fig. 1b) to eliminate artificial contacts generated by spontaneous template switches of the Phi29 DNA-polymerase[39,40] (Fig. 1c, d) during the whole-genome amplification (WGA) step (see "Methods"). In contrast to the *hiclib*[32,41] (see "Methods") annotations showing up to 20 contacts per restriction fragment (RF) in a single nucleus, ORBITA detects one or two unique contacts per RF (Fig. 1d, Supplementary Fig. 2b, c). We tested ORBITA by analyzing previously published snHi-C data from murine oocytes[32] and found that ORBITA allowed us to filter out artificial junctions in this dataset (Supplementary Fig. 3a). Notably, *hiclib* and ORBITA detect a similar number of contacts per RF in single-cell Hi-C data obtained without the usage of Phi29 DNA-polymerase[33] (Supplementary Fig. 3b). Thus, ORBITA efficiently filters out artificial Phi29 DNA-polymerase-produced DNA chimeras from snHi-C libraries.

We then constructed snHi-C maps with a resolution of up to 10 kb (Fig. 1e). In single nuclei, the dependence of the contact probability on the genomic distance, $P_c(s)$, has a shape comparable to that observed in the bulk BG3 in situ Hi-C regardless of the number of captured contacts (Fig. 1f), indicating that the key steps of the snHi-C protocol such as fixation, DNA fragmentation, and in situ ligation were performed successfully. To estimate the overall quality of the snHi-C libraries, we first

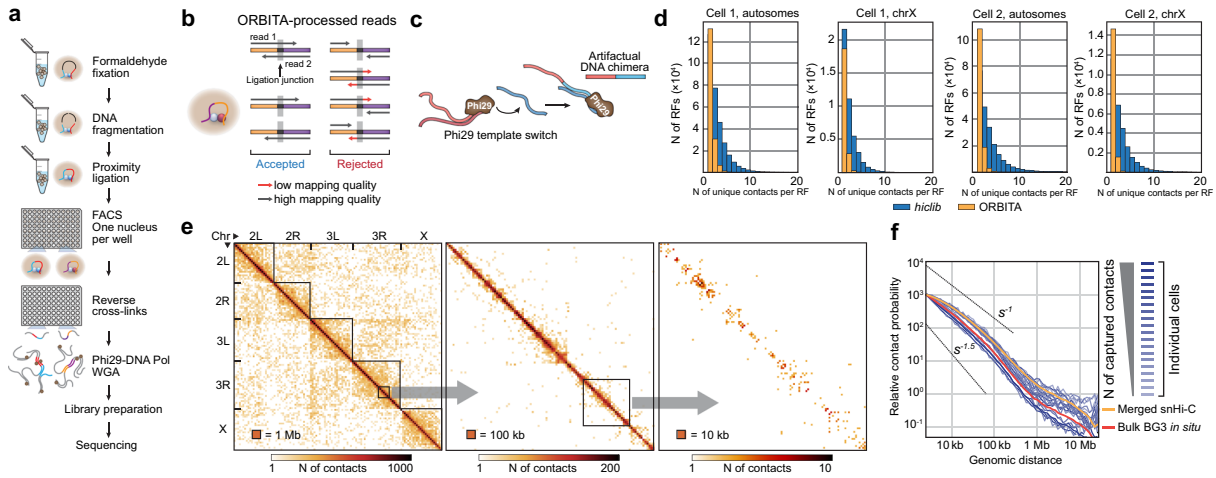


Fig. 1 ORBITA-processed Hi-C data from single *Drosophila* nuclei. a Single-nucleus Hi-C protocol scheme (see "Methods" for details). **b** Workflow of ORBITA function for detection of unique Hi-C contacts. ORBITA processes only chimeric reads with good mapping quality containing ligation junction marked by the cleavage site for restriction enzyme used for the snHi-C map construction. **c** Scheme of an artefactual DNA chimera formation by Phi29-DNA-polymerase. **d** Number of unique contacts per restriction fragment (RF) captured by ORBITA (orange) and *hiclib* (blue) for autosomes and the X chromosome. BG3 is a diploid male cell line; accordingly, in a single nucleus, each RF from autosomes and the X chromosome could establish no more than four and two unique contacts, respectively. Cell 1, autosomes: $n = 148{,}415$ and 159,060 for ORBITA and *hiclib*, respectively; ChrX: $n = 22{,}016$ and 26,674 for ORBITA and *hiclib*, respectively. Cell 2: autosomes: $n = 113{,}988$ and 119,066 for ORBITA and *hiclib*, respectively; ChrX: $n = 16{,}384$ and 19,429 for ORBITA and *hiclib*, respectively. **e** Visualization of a single-nucleus Hi-C map at 1-Mb, 100-kb, and 10-kb resolution for the cell with 107,823 captured unique contacts. **f** Dependence of the contact probability $P_c(s)$ on the genomic distance s for single nuclei (shades of blue reflect the number of unique contacts captured in individual nuclei), merged snHi-C data (orange), and bulk in situ BG3 Hi-C data (red). Black lines show slopes for $P_c(s) = s^{-1.5}$ and $P_c(s) = s^{-1}$.

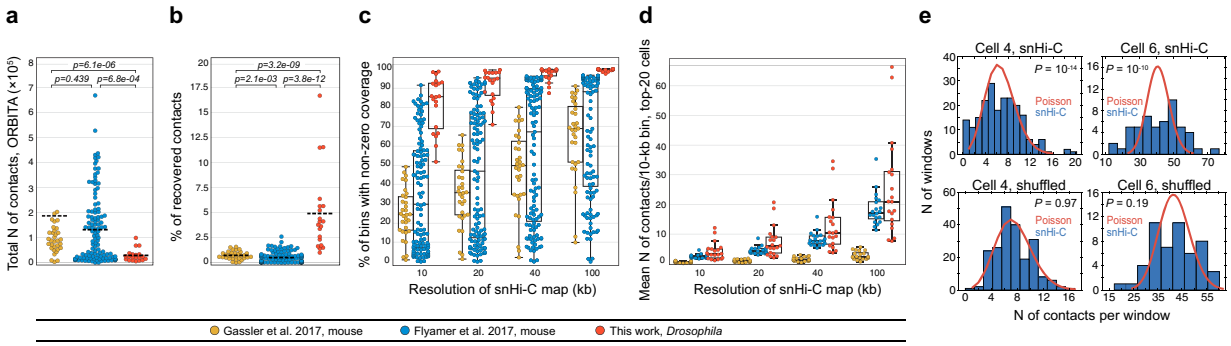


Fig. 2 snHi-C datasets in *Drosophila* represent a major portion of the genome and are not random matrices. a Number of ORBITA-captured contacts per individual nuclei obtained for *Drosophila* in the current work, compared with mouse oocytes from Flyamer et al.[32] and G2 zygotes pronuclei from Gassler et al.[34]. $^{**}p < 0.01$ using the Mann–Whitney two-sided test. $n = 20$, 120, and 32 nuclei for *Drosophila* in the current work, mouse oocytes from Flyamer et al.[32] and G2 zygotes pronuclei from Gassler et al.[34], respectively (the same is true for (**b**) and (**c**)). **b** Percentage of recovered contacts out of the total possible for *Drosophila* in the current work, compared with mouse oocytes from Flyamer et al.[32] and G2 zygotes pronuclei from Gassler et al.[34]. *P*-values are calculated using the Mann–Whitney two-sided test. **c** Percentage of bins with non-zero coverage for autosomes and sex chromosome of *Drosophila*, murine oocytes, and G2 zygote pronuclei. Boxplots represent the median, interquartile range, maximum and minimum. **d** Mean number of contacts per 10-kb genomic bin in top-20 cells in the current work, compared with mouse oocytes from Flyamer et al.[32] and G2 zygotes pronuclei from Gassler et al.[34]. Boxplots represent the median, interquartile range, maximum and minimum. **e** Distributions of the number of contacts in windows of fixed size (100 kb for the Cell 4, and 400 kb for the Cell 6; chr2R) in snHi-C data and shuffled maps for two individual cells (blue bars). The red curve shows the Poisson distribution expected for an entirely random matrix with the same number of contacts. *P*-values were estimated by the goodness of fit test. $n = 211$ and 52 windows for the cell 4 and for the cell 6, respectively.

calculated the number of captured contacts per cell. On average, we extracted 33,291 unique contacts from individual nuclei that represented 5% of the theoretical maximum number of contacts and corresponded to four contacts per 10-kb genomic bin (see "Methods"); in the best cell, 17% of contacts were recovered (Fig. 2a, b, Supplementary Table 1). Relying on the number of captured contacts, we then estimated the proportion of the genome available for the downstream analysis. At 10-kb resolution, ~82% of the genome on average was covered with contacts in each individual cell, and 67% of genomic bins established more than 1 contact (Fig. 2c). Notably, in the previously published mouse snHi-C datasets, ~0.6% of theoretically possible contacts were detected on average (Fig. 2b). Because the top-20 mouse snHi-C libraries from Flyamer et al.[32] demonstrated a comparable genome coverage with contacts and a number of contacts per 10-kb genomic bin (Fig. 2d), we could

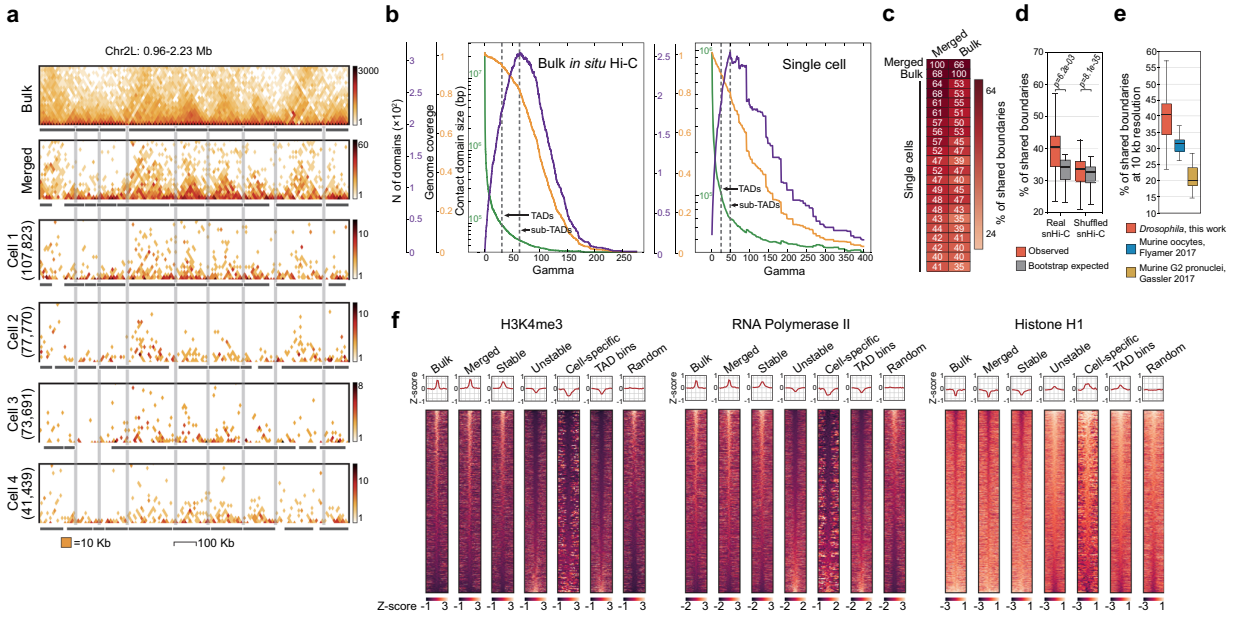


Fig. 3 Stable TAD boundaries are defined by high level of active epigenetic marks. a Example of a genomic region on Chromosome 2L with a high similarity of TAD profiles (black rectangles) in individual cells and bulk BG3 in situ Hi-C data. Number of unique captured contacts is shown in brackets. Positions of TAD boundaries identified in bulk BG3 in situ Hi-C data (top panel) are highlighted with gray lines. Here and below, TADs are identified using lavaburst software. **b** Dependence of the contact domain (CD) size (green), genome coverage by CDs (orange), and number of identified CDs (violet) on the γ value in bulk (left) and single-cell (right) BG3 Hi-C data. γ values selected for the calling of sub-TADs (γ_{max}) and TADs ($\gamma_{max}/2$) are marked with vertical gray lines. **c** Percentage of TAD boundaries shared between single cells, bulk BG3 in situ Hi-C, and merged snHi-C data. **d** Percentage of shared boundaries in real snHi-C, shuffled control maps, and bootstrap expected. Boxplots represent the median, interquartile range, maximum and minimum. $^{**}p < 0.01$ using the Mann–Whitney two-sided test. $n = 380$ comparisons between individual cells. **e** Percentage of shared boundaries in real snHi-C for *Drosophila*, murine oocytes from Flyamer et al.[32] and G2 zygote pronuclei from Gassler et al.[34]. Boxplots represent the median, interquartile range, maximum and minimum. $n = 380$ comparisons between individual cells. **f** Heatmaps of active (H3K4me3, RNA Polymerase II) and inactive (H1 histone) chromatin marks centered at single-cell TAD boundaries from different groups (±100 kb). Bulk—conventional BG3 in situ Hi-C; merged—aggregated snHi-C data from all individual cells; stable and unstable—boundaries found in more and in less than 50% of cells, respectively; cell-specific—boundaries identified in any one individual cell; TAD bins—genomic bins from TAD interior; random—randomly selected genomic bins.

directly compare the *Drosophila* and mouse snHi-C maps (see below). Next, to verify that these sparse snHi-C matrices were not generated by random fluctuations of captured contacts, we calculated the distributions of the contact numbers in sliding non-intersecting windows of different fixed sizes. In contrast to the shuffled maps, these distributions in the original data are distinct from the Poisson shape typical for random matrices (Fig. 2e, see "Methods" and Supplementary Fig. 4). We conclude that the snHi-C maps obtained here are of acceptable quality and indeed reflect specific patterns of spatial contacts in chromatin.

Visual inspection of snHi-C maps revealed distinct 50–200 kb contact domains that closely recapitulated the TAD profile in the bulk BG3 in situ Hi-C data (Fig. 3a). To call TADs in snHi-C data systematically, we used the lavaburst Python package with the modularity scoring function[32]. For each nucleus, we performed TAD segmentation in snHi-C maps of 10-kb resolution at a broad range of the gamma (γ) master parameter values (Fig. 3b, see "Methods" and Supplementary Fig. 5). Of note, the majority of the identified boundaries were resistant to the data down-sampling, indicating that these boundaries did not result from fluctuations of captured contacts in sparse snHi-C matrices (Supplementary Fig. 6). In individual nuclei, we identified 554–1402 TADs with a median size of 60 kb covering 40–76% of the genome at the γ value corresponding to the maximal number of TADs called (γ_{max}). At 10–20 kb resolution, the median size of *Drosophila* TADs was previously estimated as 100–150 kb[5,24,25]. To obtain a robust TAD profile, we used $\gamma_{max}/2$ corresponding to TADs with a median size equal to that for TADs identified in the *Drosophila* cell population according to the previously published data[24]. At $\gamma_{max}/2$, we identified 510–1175 TADs with a median size ~90 kb covering up to 89% of the genome in best snHi-C matrices (Supplementary Fig. 5).

To additionally validate the single-cell TAD segmentations, we utilized a modification of the recently published[42] spectral clustering method based on the non-backtracking random walks (NBT; see "Methods"). The non-backtracking operator is used to resolve communities in sufficiently sparse networks[42,43], thus providing a useful tool for TAD annotation in single-cell Hi-C matrices. The method performs dimensionality reduction of the network using the leading eigenvectors of the non-backtracking operator, which has a distinctive disc-shape complex spectrum with a number of isolated eigenvalues on the real axis (Supplementary Fig. 7d). The resulting average size of the detected TADs was 110 kb, closely matching the typical TAD size in the population-averaged data and in the single-cell modularity-derived segmentations. The mean number of detected TADs per cell (855 and 920 for the NBT and modularity, respectively) and single-cell TAD segmentations were remarkably similar between the two methods (Supplementary Fig. 7a) and demonstrated the same epigenetic properties (Supplementary Fig. 7c, see below). Moreover, the modularity-derived TAD boundaries were robust to the data resolution changes. On average, 84.8% of modularity-derived boundaries at the 20-kb resolution and 78.6% of boundaries at the 40-kb resolution have a matching boundary

at the 10-kb resolution. This is significantly higher than the 43 and 58% expected at random, respectively. Taken together, these results indicate that TAD profiles are robust and, thus, acceptable for the downstream analysis.

TADs are largely conserved in individual *Drosophila* nuclei, and stable TAD boundaries are enriched with active chromatin. We found that TADs tended to occupy similar positions in different cells regardless of the number of captured contacts (Fig. 3a, Supplementary Fig. 8). On average, 46.6% of population-identified TAD boundaries were present in each of the single cells analyzed (Fig. 3c), and 39.5% of boundaries were shared between different cells in pairwise comparisons (Supplementary Fig. 8). This is significantly higher than the percentage of shared boundaries for shuffled control maps (32.9%) and the percentage expected at random (33.1%, Fig. 3d). Notably, 44% of NBT-identified single-cell TAD boundaries were conserved in pairwise cell-to-cell comparisons (Supplementary Fig. 7b), supporting the results obtained in the analysis of modularity-derived TAD boundary profiles. In individual mammalian cells, TADs frequently overpassed the boundaries identified in the cell population, arguing for a substantial degree of stochasticity in genome folding[32,35,44]. We used the ORBITA algorithm to reanalyze previously published snHi-C data from murine oocytes[32] and G2 zygote pronuclei[34] and found that 31.2 and 21% of boundaries were shared on average between any two cells, respectively (Fig. 3e, Supplementary Fig. 9). This result is reproduced at 40-kb resolution and persists for a broad range of snHi-C datasets' quality (Supplementary Fig. 10). We conclude that, in *Drosophila*, TADs have more stable boundaries as compared to mammals. This corroborates recent observations of the Cavalli lab[37] and may reflect the differential impact of loop extrusion[18,19,34] and internucleosomal contacts[24] on TAD formation[16,23].

Population TADs in *Drosophila* identified at 10–20 kb resolution mostly correspond to inactive chromatin, whereas their boundaries and inter-TAD regions correlate with highly acetylated active chromatin[24,45]. These are further partitioned into much smaller domains with the size of about 9 kb[25] and, thus, unavailable for the analysis at the resolution of our Hi-C maps. To examine the properties of TAD boundaries at the single-cell level, we divided all TAD boundaries from snHi-C data into three groups according to the proportion of cells where these boundaries were present and analyzed them separately (number of boundaries of each type and distances between neighboring boundaries within each type are shown in Supplementary Fig. 13). The boundaries present in a large fraction of cells (more than 50% of cells) defined here as "stable" overlapped 73% of conserved boundaries between BG3 and Kc167 cell lines[46] and had high levels of active chromatin marks (RNA polymerase II, H3K4me3; Fig. 3f, Supplementary Figs. 11, 12). They were also slightly enriched in some architectural proteins associated with active promoters (BEAF-32, Chriz, CTCF, and GAF; Supplementary Fig. 11, 12). In contrast, boundaries identified in less than 50% of cells and defined here as "unstable" (as well as boundaries identified in just one cell termed cell-specific boundaries) were remarkably depleted of acetylated histones and features of transcriptionally active chromatin while being enriched in histone H1 and other proteins of repressed chromatin similarly to the internal TAD bins (Fig. 3f, Supplementary Fig. 11, 12). The epigenetic profiles of "unstable" boundaries may be due to the fact that actual profiles of active chromatin in individual cells differ from the bulk epigenetic profiles used in our analysis. However, it may also reflect a certain degree of stochasticity in chromatin fiber folding into contact domains[35]. Taking into consideration the fact that active chromatin regions mostly colocalize with stable boundaries, one would expect the "unstable" boundaries tend to be located in the inactive parts of the chromosome.

TADs in individual *Drosophila* cells are not hierarchical. *Drosophila* TADs are hierarchical in cell population-based Hi-C maps[45,47]. It is, however, not clear whether the hierarchy exists in individual cells or emerges in the bulk BG3 in situ Hi-C maps as a result of averaging of alternative chromatin configurations over a number of individual cells. To test this proposal, we focused on two TAD segmentations: at $\gamma_{max}/2$ (TADs) and γ_{max} (smaller domains referred to as sub-TADs located inside TADs, Fig. 4a). We analyzed only the haploid X chromosome to avoid combined folding patterns of diploid somatic chromosomes. We assumed that if TADs in individual nuclei are truly hierarchical, then sub-TADs belonging to the same TAD should be demarcated with well-defined boundaries arising from specific folding of the chromatin. To determine whether this is the case, we tested the resistance of sub-TAD boundaries to the data downsampling (two-fold depletion of total number of contacts in the snHi-C maps). In contrast to relatively stable TAD boundaries, sub-TAD boundaries showed a two-fold reduction in the probability of detection in downsampled datasets (Fig. 4b). Moreover, we found that profiles of sub-TADs were highly different in individual nuclei: only approximately 20% of sub-TAD boundaries in individual cells were shared in pairwise comparisons, similar to the shuffled controls (Supplementary Fig. 14). Hence, a potential hierarchy of TAD structure in single cells appears to reflect local Hi-C signal fluctuations. The hierarchical structure of TADs observed in bulk *Drosophila* Hi-C data[45,48], thus, likely results from the superposition of multiple alternative chromatin folding patterns present in individual nuclei; this is also supported by the visual inspection of snHi-C maps (Fig. 4c).

A-compartment in individual *Drosophila* nuclei. In animal cells, TADs of the same epigenetic type interact with each other across large genomic distances, forming compartments that spatially segregate active and inactive genomic loci in the nuclear space[13]. Similarly to *Drosophila* embryo[5], S2[49], and Kc167 cells[50], we observed an increased long-range interaction frequency within the A-compartment in the bulk BG3 in situ Hi-C data (Fig. 4d–f; Supplementary Fig. 15). Supporting this observation, we also found increased long-range interactions between genomic regions of the X chromosome activated by male-specific-lethal (MSL) complex binding[51] (Fig. 4h) in both BG3 in situ Hi-C data and the merged cell. In contrast, we observed a weak enrichment of long-range interactions between Polycomb-repressed regions[52,53] bound by dRING (Fig. 4i)[54] and nearly no enrichment for B-compartment regions (Fig. 4d, e, g).

We could not directly detect compartments in individual nuclei due to the sparsity of the maps, but we observed a substantial enrichment of contacts in the A-compartment after averaging contacts in each individual nucleus across the population-based compartment mask (Fig. 4d, Supplementary Fig. 15). Compartmentalization might, thus, be a genuine feature of chromatin folding of *Drosophila* individual nuclei. The presence of extensive long-range contacts between the active genome regions in individual chromosomes is also supported by the contact probability $P_c(s)$ plotted for active and inactive genomic bins separately: $P_c(s)$ between active genome regions has a gentler slope outside TADs, indicating that active, but not inactive chromatin forms spatial contacts across large genomic distances (Fig. 4e). These results suggest that active and inactive genome loci are spatially segregated in individual *Drosophila* nuclei; active regions establish long-distance contacts, possibly at transcription factories and nuclear speckles[55–58].

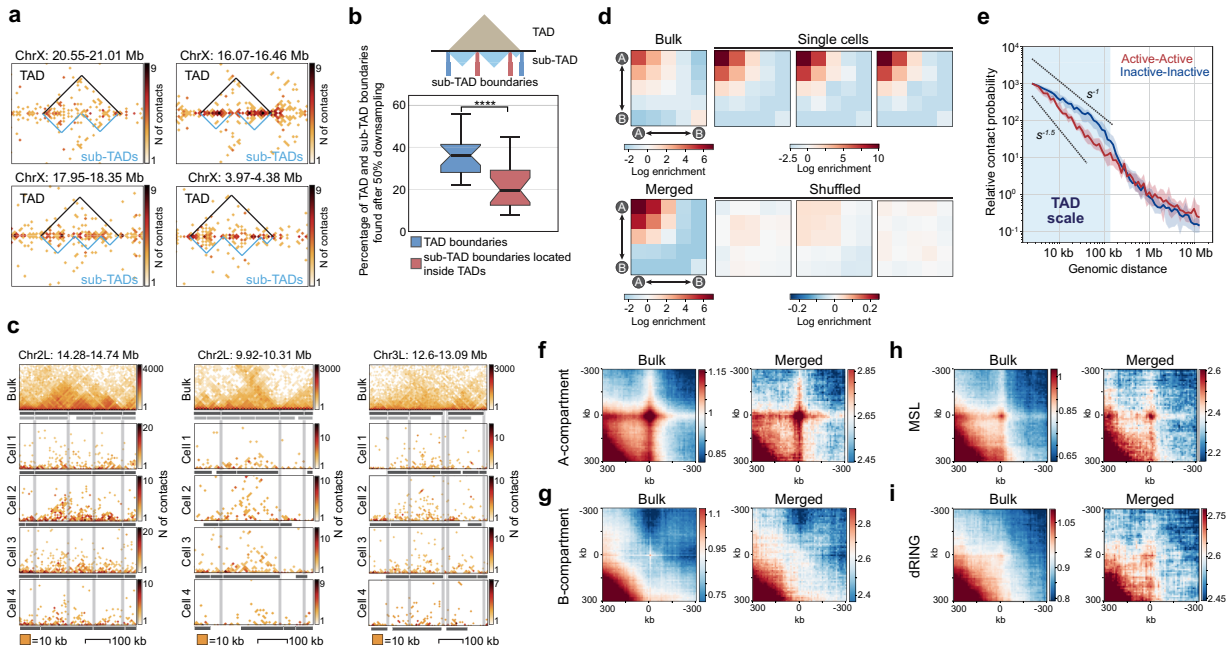


Fig. 4 Chromatin in individual *Drosophila* cells is compartmentalized and lacks folding hierarchy at the level of TADs. a Examples of TAD (black triangles) and sub-TAD (light blue triangles) positions in the haploid X chromosome in individual nuclei with 77,770 unique contacts. **b** Percentage of TAD and sub-TAD boundaries per cell (excluding TAD boundaries for the same cells) found as sub-TAD boundaries in the snHi-C maps after 50% downsampling. Downsampling was performed 10 times. At the top: TAD boundaries are highlighted with blue lines, sub-TAD boundaries located inside TADs are highlighted with red lines. Boxplots represent the median, interquartile range, maximum and minimum. ****p-value < 0.0001 using the Mann–Whitney two-sided test. $n = 20$ cells. **c** Genomic regions with alternative chromatin folding patterns in individual cells. Positions of sub-TADs and TADs identified in bulk BG3 in situ Hi-C data (top panels) are highlighted with light gray and dark gray rectangles, respectively. Positions of TAD boundaries in bulk BG3 in situ Hi-C data are shown with vertical light gray lines. **d** Heatmaps showing $\log_2$ values of contact enrichment between genomic regions belonging to putative A- (negative PC1 values) and B- (positive PC1 values) compartments (saddle plot). PC1 profile is constructed using the bulk BG3 in situ Hi-C data. **e** Contact probability $P_c(s)$ between transcriptionally active (red) and inactive (blue) genomic bins in the snHi-C data. The light blue shading shows the genomic distances corresponding to the average TAD size in single nuclei. **f** Average plot of long-range interactions between top 1000 regions of A compartment annotated by bulk Hi-C data (in bulk Hi-C and merged snHi-C). **g** Average plot of long-range interactions between top 1000 regions of B compartment annotated by bulk Hi-C data (in bulk Hi-C and merged snHi-C). **h** Average plot of interactions between top 500 regions enriched in MSL (in bulk Hi-C and merged snHi-C) on chromosome X. **i** Average plot of interactions between top 500 regions enriched in dRING (in bulk Hi-C and merged snHi-C).

Modeling of DNA fiber folding within X-chromosome by constrained polymer collapse. We next applied dissipative particle dynamics (DPD) polymer simulations[59] to reconstruct the 3D structures of haploid X chromosomes (Supplementary Fig. 16a) in individual cells using the snHi-C data (Fig. 5a, Supplementary Fig. 16b). The chromatin fiber path in these structures is strictly determined by the pattern of contacts derived from the snHi-C experiments and, thus, reflects the actual folding of the X chromosome in living cells[60]. As revealed by TAD annotation, the DPD simulations successfully reproduced chromatin fiber folding even at short and middle genomic distances because TAD positions along the X chromosome were largely preserved between the models and the original snHi-C data (Fig. 5a, Supplementary Figs. 17, 19a, b; also see "Methods"). Moreover, the simulations correctly reproduced chromatin folding at the scale of the whole chromosome with a well-defined A-compartment (Fig. 5a, Supplementary Fig. 18). Additionally, to validate the simulation results using an alternative approach, we performed multicolor in situ fluorescence hybridization (FISH) with two intra-TAD probes and one probe located outside the selected TAD. The distributions of inter-probe spatial distances extracted from the X chromosome model closely resemble those of the FISH analysis (Supplementary Fig. 19c). Taken together, these observations confirm the validity of our approach.

The snHi-C maps show remarkable cell-to-cell variability in the distribution of captured contacts (Figs. 3a, 4c); therefore, we performed a pairwise comparison of 3D models of the X chromosome in individual cells using the coefficient of the difference at a broad range of genomic distances (Fig. 5b; see "Methods"). The higher the value of the coefficient, the higher the difference between the distance matrices obtained from the models. We have found that chromatin fiber conformation was strikingly different between individual models (red curve, Fig. 5b) in comparison to different configurations (at different time points) of each particular model (blue curve, Fig. 5b), showing the prominent cell specificity in the organization of the X chromosome territory (CT). Notably, shuffling of contacts (see "Methods") in the snHi-C data prior to simulations significantly decreased the variability in the chromatin fiber conformation at long distances (gray curve, Fig. 5b). Despite cell-to-cell differences in the overall 3D shape of a particular TAD (Fig. 5c, Supplementary Fig. 19d), the variability of the chromatin fiber conformation was substantially lower at short ranges (within TADs) as compared to long-range distances (Fig. 5b). This difference could be due to an increased flexibility in chromatin folding arising from larger genomic distances. In addition, the curve of the coefficient of difference between individual models reached the plateau outside TADs (Fig. 5b), suggesting that the

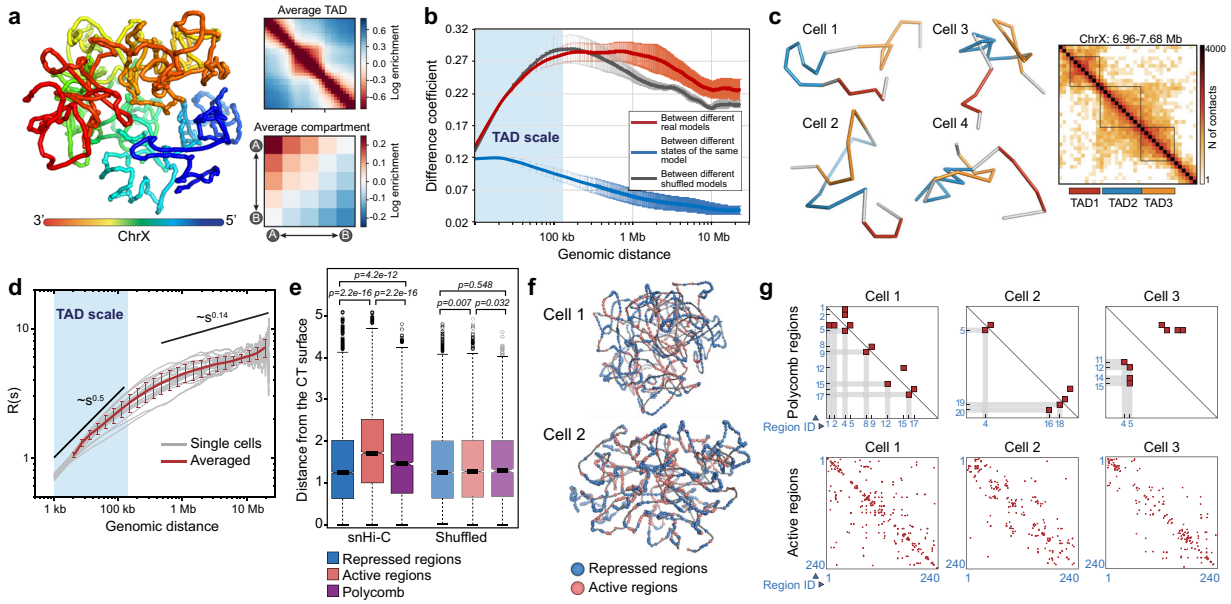


Fig. 5 3D folding of the haploid X chromosome. a Left panel: 3D structure of the haploid X chromosome from an individual nucleus derived from snHi-C data by the DPD polymer simulations. Right panel: averaging of contacts in the 3D model over TAD positions in the corresponding snHi-C data (top) and compartment (bottom) positions annotated in the bulk BG3 in situ Hi-C data. Source data are provided as a Source data file. **b** Coefficient of difference over a broad range of genomic distances. The central curves represent average values. Error bars show standard deviation (SD) for 20 independent model realizations, $n = 2242$ distinct ranges of genomic distances used for the curve construction. **c** Single-nucleus 3D structures of a genomic region covered by three TADs (left). Right, bulk BG3 in situ Hi-C map of this region. TAD positions are shown by colored rectangles below the map and by black squares on the map. **d** Dependence of the Euclidean spatial inter-particle distance R (see "Methods") on the genomic distance s between these particles along the chromatin fiber. Black lines show the slopes characteristic for the random walk behavior ($s^{0.5}$) and the crumpled globule build-up from Gaussian blobs ($s^{0.14}$). Error bars show standard deviation (SD) for 20 independent model realizations, $n = 20$. **e** Spatial distance from the surface of a chromosome territory (CT) to transcriptionally active ($n = 8966$) and inactive ($n = 17{,}103$; according to RNA-seq from ref. [24]) regions, and Polycomb-bound domains ($n = 2160$; according to the 9-state chromatin type annotation[54]). Boxplots show data aggregated from all individual models analyzed. Boxplots represent the median, interquartile range, maximum and minimum. *P*-values are calculated using the Mann–Whitney two-sided test. **f** Examples of simulated ChrX 3D structures demonstrating the preferential location of transcriptionally inactive regions (blue particles) at the surface of the CT. **g** Heatmap showing cell-to-cell variability in interactions detected between Polycomb domains (upper panels) or between transcriptionally active regions (bottom panels) in individual cells. Red rectangles denote detected interactions. The total number of Polycomb and active regions identified in the X chromosome are 20 and 240, respectively (see "Methods"). Only interacting domains are numbered for Polycomb domains; interacting active regions are not numbered due to their multiplicity.

variability of chromatin folding inside and outside TADs was governed by different rules. Due to the fact that TADs in *Drosophila* (at the 10–20 kb resolution of the Hi-C maps) are largely composed of inactive chromatin, we propose that the chromatin fiber conformation within TADs is mostly determined by interactions between adjacent non-acetylated nucleosomes. In contrast, at large genomic distances, TADs interact with each other in a stochastic manner, imposing the spherical form of the CT that is observed in all model structures (Fig. 5a, Supplementary Figs. 16, 20). In line with this hypothesis, the dependence of spatial distance R between any two particles on the genomic distance s revealed two distinct modes of polymer folding (Fig. 5d). At the scale of ~100 kb (e.g., inside TADs), the chromatin fiber demonstrated a random walk behavior ($s^{0.5}$) similar to the chromatin of budding yeast. At larger distances, R(s) had a scaling similar to a crumpled globule build of Gaussian blobs ($s^{0.14}$)[61]. Thus, chromatin folding within TADs and at the scale of the whole CT could be driven by different molecular mechanisms.

Analysis of the radial distributions of transcriptionally active, inactive, and Polycomb-bound genome regions in our models demonstrated that active chromatin tended to be located in the CT interior, whereas inactive regions were located near the CT surface (Fig. 5e, f); this can be driven by interactions with the nuclear lamina[62]. Formation of nuclear microcompartments such as Polycomb bodies[63] represents another factor determining the large-scale spatial structure of the X chromosome territory. We analyzed patterns of interactions between individual Polycomb-occupied regions in the 3D models. To this aim, each of such regions was assigned a consecutive number according to their positions along the chromosome. The examples of 2D maps demonstrating regions residing in a spatial proximity in each cell are presented in Fig. 5g (upper panels). We found that Polycomb-occupied regions interacted with each other in a cell-specific manner and, moreover, such contacts occurred between loci regardless of the genomic distances between them (Fig. 5g, upper panels). Using a similar approach, we constructed 2D interaction maps of active genomic regions (Fig. 5g, bottom panels). Active genome regions also interacted with each other across large genomic distances in a cell-specific manner (Fig. 5g, bottom panels). We propose that these two types of long-range interactions: stochastic assembly of Polycomb bodies and transcription-related microcompartments (factories[64]), underlie the cell-specific conformation of the chromatin fiber within CTs in *Drosophila*.

Discussion

Folding of interphase chromatin in eukaryotes is driven by multiple mechanisms operating at different genome scales and generating distinct types of the 3D genome features[16,20]. In

mammalian cells, cohesin-mediated chromatin fiber extrusion mainly impacts the genome topology at the scale of ~100–1000 kb by producing loops, resulting in the formation of TADs[18,19] and establishing enhancer-promoter communication[65]. Chromatin loop formation by the loop extrusion complex (LEC) in mammalian cells is a substantially deterministic process due to the preferential positioning of loop anchors encoded in DNA by CTCF binding sites (CBS). The cohesin-CTCF molecular tandem modulates folding of intrinsically disordered chromatin fiber[16,23]. On the other hand, association of active and repressed gene loci in chromatin compartments[13,14], and formation of Polycomb and transcription-related nuclear bodies[66,67] in both mammalian and *Drosophila* cells shape the 3D genome at the scale of the whole chromosome. These associations appear to be stochastic: a particular Polycomb-bound or transcriptionally active region in individual cells interacts with different partners located across a wide range of genomic distances[68].

Here, we applied the single-nucleus Hi-C to probe the 3D genome in individual *Drosophila* cells at a relatively high resolution that was not achieved previously in single-cell Hi-C studies. Based on our observations, we suggest that, in *Drosophila*, both deterministic and stochastic forces govern the chromatin spatial organization (Fig. 6a).

We found that the entire individual *Drosophila* genomes were partitioned into TADs; this observation supports the results of recent super-resolution microscopy studies[37]. TAD profiles are highly similar between individual *Drosophila* cells and demonstrate lower cell-to-cell variability as compared to mammalian TADs. According to our model[24], large inactive TADs in *Drosophila* are assembled by multiple transient electrostatic interactions between non-acetylated nucleosomes in transcriptionally silent genome regions. Conversely, TAD boundaries and inter-TAD regions at the 10-kb resolution of Hi-C maps in *Drosophila* were found to be formed by transcriptionally active chromatin. This result may explain why TADs in individual cells occupy virtually the same genomic positions (Fig. 6b). Gene expression profile is a characteristic feature of a particular cell type, and, thus, should be relatively stable in individual cells within the population. In agreement with this, we demonstrated that invariant TAD boundaries present in a major portion of individual cells were highly enriched in active chromatin marks. Moreover, stable boundaries were also largely conserved in other cell types (see "Results" and ref. [46]), possibly due to the fact that TAD boundaries were frequently formed at the position of housekeeping genes.

In contrast to stable TAD boundaries, the boundaries that demonstrate cell-to-cell variability bear silent chromatin. Some cell-specific TAD boundaries may originate at various positions due to a putative size limit of large inactive TADs or other restrictions in chromatin fiber folding. Indeed, it appears that the assembly of randomly distributed TAD-sized self-interacting domains is an intrinsic property of chromatin fiber folding[35]. In mammals, the positioning of these domains is modulated by cohesin-mediated DNA loop extrusion[35], whereas in *Drosophila*, it may be modulated by segregation of chromatin domains bearing distinct epigenetic marks[16,23]. Even if cell-specific and unstable TAD boundaries are distributed in a random fashion, they should be depleted in active chromatin marks because active chromatin regions are mainly occupied by stable TAD boundaries. We also cannot exclude that variable boundaries and the TAD boundary shifts are caused by local variations in gene expression and active chromatin profiles in individual cells that we cannot assess simultaneously with constructing snHi-C maps.

Our results are also compatible with an alternative mechanism of TAD formation. Given that the above-mentioned cohesin-driven loop extrusion is evolutionarily conserved from bacteria to mammals[69], it is compelling to assume that extrusion works in *Drosophila* as well. Despite the presence of all potential components of LEC (cohesin, its loading and releasing factors), TAD boundaries in *Drosophila* are not significantly enriched with CTCF[24,25] and do not form CTCF-enriched interactions or TAD corner peaks. These observations suggest that the binding sites of CTCF or other distinct proteins do not constitute barrier elements for the *Drosophila* LEC even if these proteins are enriched in TAD boundaries; this may be due to some other properties of a genomic region. For example, stably bound cohesins were proposed to act as the barriers for cohesin extrusion in yeast[70].

Active transcription interferes with DNA loop extrusion[71,72]. Because TAD boundaries in *Drosophila* are highly transcribed, we propose that open chromatin with actively transcribing polymerase and/or a high density of chromatin remodeling complexes could serve as a barrier for the *Drosophila* LEC. Contrary to the strictly positioned and short CBSs in mammals, active loci flanking *Drosophila* TADs represent relatively extended regions up to several dozens of kb in length. Probabilistic termination of LEC at varying points within such regions in different cells of the population could explain the absence of canonical loop signals and the presence of strong compartment-like interactions between active regions flanking a TAD (Fig. 6c). This model also provides a potential explanation for the relatively high stability of TAD positioning in individual *Drosophila* cells in comparison to mammals. A relative permeability of CBSs in mammalian cells allows LEC to proceed through thousands of kilobases and to produce large contact domains[17]. Extended active regions acting as "blurry" barrier elements where LEC termination occurs at multiple points, should stop the LEC more efficiently, making the TAD pattern more stable and pronounced.

Taken together, the order in the *Drosophila* chromatin 3D organization is manifested in a TAD profile that is relatively stable between individual cells and likely dictated by the distribution of active genes along the genome. On the other hand, our molecular simulations of individual haploid X chromosomes indicate a prominent stochasticity in both the form of individual TADs and the overall folding of the entire chromosome territory. According to our data, the active A-compartment is easily detectable in individual cells, and the profiles of interaction between individual active regions are highly variable between individual cells. Notably, this also holds true for Polycomb-occupied loci that are known to shape chromatin fiber in living cells[48].

Although these highly variable long-range interactions of active regions and Polycomb-occupied loci are closely related to the shape of chromosome territory (CT), the cause-and-effect relationships between them and the stochastic nature of the cell-specific chromatin chain path are currently unclear. The main question to be answered by future studies is whether these interactions are fully stochastic or at least partially specific. The possible molecular mechanisms that may provide specific communication between remote genomic loci separated by up to megabases of DNA are not known. In a scenario of the absence of any specificity, the pattern of contacts inside A-compartment and within Polycomb bodies in a particular cell is established by stochastic fluctuations of the large-scale chromatin fiber folding. In this case, the large-scale chromatin fiber folding dictates the cell-specific location of Polycomb-enriched and active chromatin regions in the 3D nuclear space. The formation of Polycomb bodies and transcription-related chromatin hubs is achieved by confined diffusion of these regions and might be further stabilized by specific protein-protein interactions and liquid-liquid phase separation[73]. This mechanism allows to sort through alternative configurations of the 3D genome and to transiently stabilize those that are functionally relevant under specific conditions. A balance

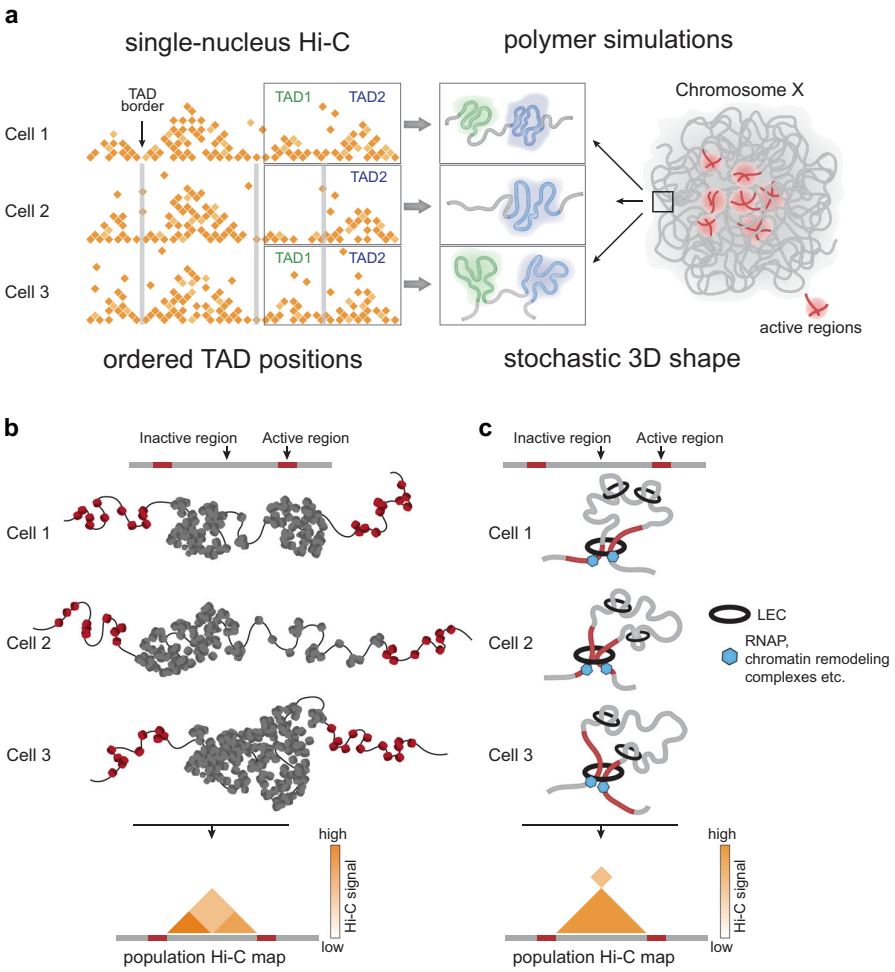


Fig. 6 Order and stochasticity in the *Drosophila* 3D genome. a Schematic representation of the ordered and stochastic components in the *Drosophila* genome folding. Positions of TAD boundaries are largely conservative between individual cells and determined by active chromatin. Chromatin fiber path within a particular TAD and within the whole chromosome territory is largely stochastic and demonstrate prominent cell-to-cell variability. **b** Determined positions of active regions along the *Drosophila* genome define TAD boundaries persistent in individual cells. Inactive region is folded into chromatin globule due to interactions between non-acetylated "sticky" nucleosomes. This region adopts different configurations in individual cells (and at different time points in a particular cell). In a cell 1, it is folded into two globules separated with stochastically formed fuzzy boundary. In a cell 2, one part of the region is compact (left) and the other part (right) is transiently decondensed. In a cell 3, the entire region forms one densely packed globule. Averaging of these configuration results in a TAD containing two sub-TADs in a population-based Hi-C map. Note, that the hierarchical structure of the TAD emerging in a population Hi-C map reflects different configuration of the region in individual cells. We note that the absence of any structure at inactive TAD borders denotes ambiguity of folding of these regions with snHi-C, but not the absence of this structure. **c** Extended active regions serving as barrier elements for potential loop extrusion complex (LEC) in *Drosophila* cells. It has been previously shown that transcription might interfere with loop extrusion[71,72]. Since stable TAD boundaries in *Drosophila* are enriched with transcribed genes, we propose that extended regions of active chromatin but not binding sites of architectural proteins represent barrier elements for LEC in *Drosophila* cells. In this scenario, LEC is looping out a TAD and terminates within flanking active regions colliding with RNA-polymerases, large chromatin-remodeling complexes and other components of active chromatin. In different individual cells, termination occurs accidentally at different points within these regions. In a population-based Hi-C map that results in a compartment-like signal but not in a conventional pointed loop observed in mammalian cells where CTCF binding sites serve as barrier elements for LEC.

between the order and the stochasticity appears to be an intrinsic property of nuclear organization that enables rapid adaptation to changing environmental conditions.

Methods

Cell culture. *Drosophila melanogaster* ML-DmBG3-c2 cell line (Drosophila Genomics Resource Center) was grown at 25 °C in a mixture (1:1 v/v) of Shields and Sang M3 insect medium (Sigma) and Schneider's Drosophila Medium (Gibco) supplemented with 10% heat-inactivated fetal bovine serum (FBS, Gibco), 50 units/ml penicillin, and 50 µg/ml streptomycin.

Single-nucleus Hi-C library preparation. We modified the previously published single-nucleus Hi-C protocol[32] as follows: 5–10 million cells were fixed in 1× phosphate-buffered solution (PBS) with 2% formaldehyde for 10 min with occasional mixing. The reaction was stopped by the addition of 2 M glycine to give a final concentration of 125 mM. Cells were centrifuged (1000 × *g*, 10 min., 4 °C), resuspended in 50 µl of 1× PBS, snap-frozen in liquid nitrogen, and stored at −80 °C. Defrozen cells were lysed in 1.5 ml isotonic buffer (50 mM Tris-HCl pH 8.0, 150 mM NaCl, 0.5% (v/v) NP-40 substitute (Fluka), 1% (v/v) Triton-X100 (Sigma), 1× Halt™ Protease Inhibitor Cocktail (Thermo Scientific) on ice for 15 min. Cells were centrifuged at 2500 × *g* for 5 min, resuspended in 100 µl of 1× DpnII buffer (NEB), and pelleted again. The pellet was resuspended in 200 µl of 0.3% SDS in 1.1× DpnII buffer and incubated at 37 °C for 1 h. Then, 330 µl of 1.1× DpnII buffer and 53 µl of 20%

Triton X-100 (Sigma) were added, and the suspension was incubated at 37 °C for 1 h. Next, 600 U of DpnII enzyme (NEB) were added, and the chromatin was digested overnight (14–16 h) at 37 °C with shaking (1400 rpm). On the following day, 200 U of DpnII enzyme were added, and the cells were incubated for an additional 2 h. DpnII was then inactivated by incubation at 65 °C for 20 min. Nuclei were centrifuged at 3000 × *g* for 5 min, resuspended in 100 µl of 1× T4 DNA ligase buffer (Fermentas), and pelleted again. The pellet was resuspended in 400 µl of 1× T4 DNA ligase buffer, and 75 U of T4 DNA ligase (Fermentas) were added. Chromatin fragments were ligated at 16 °C for 6 h. Next, the nuclei were centrifuged at 5000 × *g* for 5 min, resuspended in 100 µl of sterile 1× PBS, stained with Hoechst, and single nuclei were isolated into wells of a standard 96-well PCR plate (Thermo Fisher) using FACS (BD FACSAriaTMIII). Each well contained 3 µl of sample buffer from the Illustra GenomiPhi v2 DNA amplification kit (GE Healthcare). Sample buffer drops containing isolated nuclei were covered by 5 µl of mineral oil (Thermo Fisher) and incubated at 65 °C for 3 h to reverse formaldehyde cross-links. Total DNA was amplified according to a previously published protocol[74]. The amplification was considered successful if the sample contained ≥1 µg DNA. The DNA was then dissolved in 500 µl of sonication buffer (50 mM Tris-HCl (pH 8.0), 10 mM EDTA, 0.1% SDS) and sheared to a size of ~100–1,000 bp using a VirSonic 100 (VerTis). The samples were concentrated (and simultaneously purified) using AMICON Ultra Centrifugal Filter Units to a total volume of about 50 µl. The DNA ends were repaired by adding 62.5 µl MQ water, 14 µl of 10× T4 DNA ligase reaction buffer (Fermentas), 3.5 µl of 10 mM dNTP mix (Fermentas), 5 µl of 3 U/µl T4 DNA polymerase (NEB), 5 µl of 10 U/µl T4 polynucleotide kinase (NEB), 1 µl of 5 U/µl Klenow DNA polymerase (NEB), and then incubating at 20 °C for 30 min. The DNA was purified with Agencourt AMPure XP beads and eluted with 50 µl of 10 mM Tris-HCl (pH 8.0). To perform an A-tailing reaction, the DNA samples were supplemented with 6 µl 10× NEBuffer 2, 1.2 µl of 10 mM dATP, 1 µl of MQ water, and 3.6 µl of 5 U/µl Klenow (exo-) (NEB). The reactions were carried out for 30 min at 37 °C in a PCR machine, and the enzyme was then heat-inactivated by incubation at 65 °C for 20 min. The DNA was purified using Agencourt AMPure XP beads and eluted with 100 µl of 10 mM Tris-HCl (pH 8.0). Adapter ligation was performed at 22 °C for 2.5 h in the following mixture: 41.5 µl MQ water, 5 µl 10× T4 DNA ligase reaction buffer (Fermentas), 2.5 µl of Illumina TruSeq adapters, and 1 µl of 5 U/µl T4 DNA ligase (Fermentas). Test PCR reactions containing 4 µl of the ligation mixture were performed to determine the optimal number of PCR cycles required to generate sufficient PCR products for sequencing. The PCR reactions were performed using KAPA High Fidelity DNA Polymerase (KAPA) and Illumina PE1.0 and PE2.0 PCR primers (10 pmol each). The temperature profile was 5 min at 98 °C, followed by 6, 9, 12, 15, and 18 cycles of 20 s at 98 °C, 15 s at 65 °C, and 20 s at 72 °C. The PCR reactions were separated on a 2% agarose gel containing ethidium bromide, and the number of PCR cycles necessary to obtain a sufficient amount of DNA was determined based on the visual inspection of gels (typically 12–15 cycles). Four preparative PCR reactions were performed for each sample. The PCR mixtures were combined, and the products were separated on a 1.8% agarose gel. 200–600 bp DNA fragments were excised from the gel and purified with a QIAGEN Gel Extraction Kit.

Bulk BG3 in situ Hi-C library preparation. Bulk BG3 in situ Hi-C libraries were prepared as described previously[24] with minor modifications. The first steps of the protocol (from fixation to DpnII enzyme inactivation) were completely identical to the corresponding steps in the single-cell Hi-C library preparation procedure described above. After DpnII inactivation, the nuclei were harvested for 10 min at 5000 × *g*, washed with 100 µl of 1× NEBuffer 2, and resuspended in 50 µl of 1× NEBuffer 2. Cohesive DNA ends were biotinylated by the addition of 7.6 µl of the biotin fill-in mixture prepared in 1× NEBuffer 2 (0.025 mM dATP (Thermo Scientific), 0.025 mM dGTP (Thermo Scientific), 0.025 mM dTTP (Thermo Scientific), 0.025 mM biotin-14-dCTP (Invitrogen), and 0.8 U/µl Klenow enzyme (NEB)). The samples were incubated at 37 °C for 75 min with shaking (1400 rpm). Nuclei were centrifuged at 3000 × g for 5 min, resuspended in 100 µl of 1× T4 DNA ligase buffer (Fermentas), and pelleted again. The pellet was resuspended in 400 µl of 1× T4 DNA ligase buffer, and 75 U of T4 DNA ligase (Fermentas) were added. Chromatin fragments were ligated at 20 °C for 6 h. The cross-links were reversed by overnight incubation at 65 °C in the presence of proteinase K (100 µg/ml). After cross-link reversal, the DNA was purified by single phenol-chloroform extraction followed by ethanol precipitation with 20 µg/ml glycogen (Thermo Scientific) as the co-precipitator. After precipitation, the pellets were dissolved in 100 µl 10 mM Tris-HCl pH 8.0. To remove residual RNA, samples were treated with 50 µg of RNase A (Thermo Scientific) for 45 min at 37 °C. To remove residual salts and DTT, the DNA was additionally purified using Agencourt AMPure XP beads (Beckman Coulter). Biotinylated nucleotides from the non-ligated DNA ends were removed by incubating the Hi-C libraries (2 µg) in the presence of 6 U of T4 DNA polymerase (NEB) in NEBuffer 2 supplied with 0.025 mM dATP and 0.025 mM dGTP at 20 °C for 4 h. Next, the DNA was purified using Agencourt AMPure XP beads. The DNA was then dissolved in 500 µl of sonication buffer (50 mM Tris-HCl (pH 8.0), 10 mM EDTA, 0.1% SDS) and sheared to a size of approximately 100–1000 bp using a VirSonic 100 (VerTis). The samples were concentrated (and simultaneously purified) using AMICON Ultra Centrifugal Filter Units to a total volume of approximately 50 µl. The DNA ends were repaired by adding 62.5 µl MQ water, 14 µl of 10× T4 DNA ligase reaction buffer (Fermentas), 3.5 µl of 10 mM dNTP mix (Fermentas), 5 µl of 3 U/µl T4 DNA polymerase (NEB), 5 µl of 10 U/µl T4 polynucleotide kinase (NEB), 1 µl of 5 U/µl Klenow DNA polymerase (NEB), and then incubating at 20 °C for 30 min. The DNA was purified with Agencourt AMPure XP beads and eluted with 50 µl of 10 mM Tris-HCl (pH 8.0). To perform an A-tailing reaction, the DNA samples were supplemented with 6 µl 10× NEBuffer 2, 1.2 µl of 10 mM dATP, 1 µl of MQ water, and 3.6 µl of 5 U/µl Klenow (exo−) (NEB). The reactions were carried out for 30 min at 37 °C in a PCR machine, and the enzyme was then heat-inactivated by incubation at 65 °C for 20 min. The DNA was purified using Agencourt AMPure XP beads and eluted with 100 µl of 10 mM Tris-HCl (pH 8.0). Biotin pulldown of the ligation junctions was performed as described previously, with minor modifications. Briefly, 4 µl of MyOne Dynabeads Streptavidin C1 (Invitrogen) beads were used to capture the biotinylated DNA, and the volumes of all buffers were decreased by 4-fold. The washed beads with captured ligation junctions were resuspended in 50 µl of adapter ligation mixture comprising 41.5 µl MQ water, 5 µl 10× T4 DNA ligase reaction buffer (Fermentas), 2.5 µl of Illumina TruSeq adapters, and 1 µl of 5 U/µl T4 DNA ligase (Fermentas). Adapter ligation was performed at 22 °C for 2.5 h, and the beads were sequentially washed twice with 100 µl of TWB (5 mM Tris-HCl (pH 8.0), 0.5 mM EDTA, 1 M NaCl, 0.05% Tween-20), once with 100 µl of 1× binding buffer (10 mM Tris-HCl (pH 8.0), 1 mM EDTA, 2 M NaCl), once with 100 µl of CWB (10 mM Tris-HCl (pH 8.0) and 50 mM NaCl), and then resuspended in 20 µl of MQ water. Test PCR reactions containing 4 µl of the streptavidin-bound Hi-C library were performed to determine the optimal number of PCR cycles required to generate sufficient PCR products for sequencing. The PCR reactions were performed using KAPA High Fidelity DNA Polymerase (KAPA) and Illumina PE1.0 and PE2.0 PCR primers (10 pmol each). The temperature profile was 5 min at 98 °C, followed by 6, 9, 12, 15, and 18 cycles of 20 s at 98 °C, 15 s at 65 °C, and 20 s at 72 °C. The PCR reactions were separated on a 2% agarose gel containing ethidium bromide, and the number of PCR cycles necessary to obtain a sufficient amount of DNA was determined based on the visual inspection of gels (typically 12–15 cycles). Four preparative PCR reactions were performed for each sample. The PCR mixtures were combined, and the products were separated on a 1.8% agarose gel. 200–600 bp DNA fragments were excised from the gel and purified with a QIAGEN Gel Extraction Kit. Two biological replicates were performed.

snHi-C raw data processing and contact annotation. The whole-genome amplification step of snHi-C uses the Phi29 DNA polymerase, which is known to produce chimeric DNA molecules by randomly switching the DNA template[40]. DNA molecules created by the template switch were further amplified during the snHi-C protocol and resulted in chimeric reads. Notably, in theory, template switches can be detected by the presence of two consecutive parts of the same read that map to different genomic locations and do not align immediately next to the restriction sites at the DNA breakpoint. This situation is different from the standard Hi-C, where each read pair is considered to be a true contact pair regardless of the DNA breakpoint presence and annotation. Standard Hi-C processing tools, such as *hiclib*[32,41], *Juicer*[75], and *HiCExplorer*[26], typically rely on mapping of both reads in a Hi-C pair and do not account for the presence of chimeric parts in a single side of paired-end sequencing. We devised a more accurate approach for processing of snHi-C data that annotates each DNA breakpoint observed in each single-end read, and selects the contacts that do not represent possible template switches of Phi29 polymerase. Thus, we developed a custom approach for snHi-C data processing termed ORBITA (One Read-Based Interaction Annotation), as described below.

Reads mapping. As the first step of the approach, FASTQ files with paired-end sequencing data are mapped to *Drosophila* reference genome dm3 using Burrows-Wheeler Aligner (BWA-MEM, console version 0.7.17-r1188)[76] with default parameters. Notably, this mapping procedure allows independent alignment of chimeric parts of both forward and reverse reads. This step results in BAM files with paired-end mapping information.

Annotated pairs retrieval. In the next step, the BAM files are parsed with an adapted version of *pairtools* (https://github.com/mirnylab/pairtools) with our newly implemented option ORBITA. Among many other utilities for Hi-C data processing, we selected *pairtools* from the Mirny lab as the basis of our approach, due to the convenience and modular structure of its code. This version of the tool can be accessed at the GitHub repository https://github.com/agalitsyna/pairtools.

ORBITA treats each read in the BAM file independently, regardless of whether it is forward or reverse. Reads that are uniquely mapped to a single location of the genome are marked as type P, meaning that they are part of a standard Hi-C Pair with no DNA breakpoint evidence. Reads that contain precisely two successive regions uniquely mapped to different genomic locations (MAPQ > 1) are selected for further DNA breakpoint annotation. ORBITA takes the genome restriction annotation (provided as a BED file with DpnII restriction fragments positions, produced by *cooler digest*[77]) and compares each breakpoint against the list of restriction sites. For each 3′-end of the right chimeric part and 5′-end of the left chimeric part (in other words, ligated ends), both upstream and downstream restriction sites are annotated, and the distance to the closest one is calculated. If both ends are located sufficiently close (<10 bp) to any restriction site in the genome, ORBITA considers them as a true ligation junction of restricted fragments

in the snHi-C proximity ligation step. These cases are marked as J type (ligation Junction), with the evidence of traversing the ligation junction of DpnII restriction fragments. If at least one ligated end of the chimeric read was not mapped to the restriction site, ORBITA marks it as H (template switch, or Hopping of Phi29 DNA polymerase). To simplify the ORBITA approach, we omit the cases with more complicated scenarios of read mapping, when three or more uniquely mapped chimeric parts of a single-end read were present. If the read contains multiple mapped chimeric parts, it is discarded. ORBITA produces the resulting PAIRS file with annotation of JJ pairs (with the evidence of the ligation) that are accepted for further processing. If not explicitly mentioned, the generic names "pair" or "contact" are used for snHi-C contacts with the evidence of the ligation junction.

Amplification duplicates removal. In the next step, we performed a correction for amplified duplicates of snHi-C contacts. Standard Hi-C uses amplification by the Illumina PCR protocol with primers that are ligated to the ends of sheared DNA[17]. Thus, two independent Hi-C pairs can be PCR duplicates if their mapping positions coincide (e.g., see *hiclib*). However, the amplification in snHi-C[32] is followed by sonication, resulting in random breaks of ligated DNA fragments. Hence, coinciding mapping positions cannot be used as a criterion of PCR duplication. Notably, we cannot distinguish the amplified pair contacting restriction fragments from the contacts of the same regions in the homologous chromosomes. Thus, we removed all multiple copies of restriction fragment pairs and retained unique contacts for each combinatorial pair of restriction fragments.

Fragment filtration. In the next step, we used restriction fragment filtration to reduce the possible contribution of copy number variation, read misalignment, and Phi29 DNA polymerase template switch that had not been removed by the ORBITA filter.

In theory, each restriction fragment of DNA has two ends and is present twice in the diploid nucleus of ML-DmBG3-c2 *Drosophila* cells; thus, we expect the upper limit of four unique contacts per restriction fragment if no unannotated genomic rearrangements, mismappings, or template switches occurred. For each restriction fragment, we calculated the observed number of contacts and removed fragments that had more than four contacts.

Before contact filtration by this rule, we compared the number of restriction fragments with more than four unique contacts according to ORBITA and one previous approach, *hiclib* for Flyamer et al. 2017. We obtained datasets for mouse nuclei from Flyamer et al. 2017 and Nagano et al. 2017 and mapped with the *hiclib* and ORBITA pipelines. We found a significant reduction in the number of unique contacts per fragment for snHi-C from Phi29 DNA polymerase datasets (Flyamer et al. 2017, present work], but not for scHi-C without Phi29 DNA polymerase (Nagano et al. 2017) (Supplementary Figs. 2, 3). Thus, we conclude that ORBITA is an effective approach to reduce the number of snHi-C artefactual contacts arising from random template switches of Phi29 DNA polymerase.

Cell selection by raw data subsampling. We obtained filtered contacts for 88 individual nuclei after the initial round of sequencing. Before the second round of sequencing, we assessed the robustness of the number of unique contacts by subsampling of raw datasets (Supplementary Fig. 2a). For each library, we created a uniform grid of sequencing depth (from 0 to the resulting number of reads with the step of 100,000 reads). We then randomly selected X reads from the full library and calculated the number of unique contacts (as described above) for each number from the grid X. We repeated this procedure ten times and plotted the mean number of unique contacts for each sequencing depth from the grid.

We proposed that there are a significant number of cells containing PCR duplicates and that the number of contacts increases slowly depending on the sequencing depth due to the poor efficiency of the snHi-C protocol. Further sequencing of these cells would result in a relatively small improvement of the detectable number of unique contacts. The number of contacts for other cells increases more rapidly with the number of reads but reaches a plateau once the maximum number of unique contacts is achieved. Thus, additional sequencing of these cells might result in reading duplicated contacts.

For other cells, the number of contacts grew slowly with sequencing depth (Supplementary Fig. 2a). However, for all these cells, the number of unique contacts gradually increased with no plateau signature. We selected the cells displaying the best growth of the number of contacts, indicative of the good quality of the dataset. The top 20 cells by the number of unique contacts were subjected to an additional round of sequencing. The same mapping and parsing pipeline was used for these datasets. Technical replicates (initial and additional rounds of snHi-C libraries sequencing) were merged at the annotated PAIRS file stage.

snHi-C interaction map construction. The resulting pair data were binned at 1 kb, 10 kb. 20-kb, 40-kb, and 100-kb resolutions with *cooler* version 0.8.5[77] and stored in the COOL format. We constructed the merged dataset by summing all snHi-C maps. To exclude self-interacting genomic bins and possible contribution of dangling ends, self-circles[41], and mirror reads[78], we removed the first diagonal in both single cells and the merged maps. The *HiGlass* server was used for data visualization[79]. 10-kb resolution was used throughout the paper if another resolution is not specified.

Bulk BG3 in situ Hi-C raw data processing. For bulk BG3 in situ Hi-C (two biological replicates), reads were mapped to *Drosophila* reference genome dm3 with Burrows-Wheeler Aligner (BWA-MEM, console version 0.7.17-r1188)[76] with default parameters. For consistency with the snHi-C analysis, the resulting BAM files were parsed with *pairtools* v0.3.0, (https://github.com/mirnylab/pairtools) using default parameters. The resulting files were sorted by the *pairtools* module "sort"; replicates were merged by the *pairtools* module "merge" and duplicates were removed, allowing one mismatch between possible duplicates (*pairtools* dedup with --max-mismatch 1 and —mark-dups options). The resulting PAIRS file was binned with *cooler*[77] at the same resolutions as the single-cell datasets. To remove the contribution of possible Hi-C technical artifacts, such as backward ligation, dangling ends, self-circles[41], and mirror reads[78], the first two diagonals of Hi-C maps were removed. As the last step of bulk Hi-C processing, the maps were iteratively corrected for the removal of coverage bias[41] by the *cooler* balance tool with default parameters[77].

For the reproducibility control, both replicates were converted to interaction maps independently by the above pipeline. The resulting maps demonstrated a correlation of 0.9–0.95 as estimated by the *HiCRep* stratum-adjusted correlation coefficient for intrachromosomal maps smoothed with one-bin offset and genomic distance up to 300 kb at 20 kb resolution[80].

snHi-C background model construction. We sought to create a background model for snHi-C that can be used as a control for the subsequent analysis of intrachromosomal snHi-C interaction maps. For that, we considered two major factors contributing to the intrachromosomal contact frequency in the genomic region: the contact probability for a particular genomic distance $P_c(s)$[13], and region visibility[81].

For bulk BG3 in situ Hi-C, the $P_c(s)$ is assessed by the mean number of contacts for a certain genomic distance[13]. However, the same procedure cannot be readily used for snHi-C due to data sparsity and missing data. Thus, to calculate $P_c(s)$ for a snHi-C dataset, we counted the number of contacts for a certain genomic distance and normalized by the number of genomic bins that had contact in at least one snHi-C experiment at any distance. Notably, we use the same procedure for the visualization of snHi-C $P_c(s)$ dependence on the genomic distance s (Fig. 1f and Fig. 4e); the genomic distance step size was set to 1 kb. For snHi-C background models, we used $P_c(s)$ genomic distance step size 10 kb.

We assessed the region visibility in snHi-C by the marginal distribution of the number of contacts for the region $marg_i$ (in other words, the total number of observed intrachromosomal contacts for a genomic region) using maps at a 10-kb resolution.

For each snHi-C map, we calculated $P_c(s)$ and the marginal distribution of contacts and shuffled the positions of the contacts for each chromosome, so that the marginal distribution was preserved, and $P_c(s)$ was at least approximated (Supplementary Fig. 4a–d). Note that for 3D modeling, we used more crude shuffling without saving the marginal distribution of contacts.

Assessment of percentage of recovered contacts. To compare snHi-C datasets across species (Fig. 2a–c), we assessed the percentage of recovered contacts out of all possible contacts per nuclei.

First, we determined the theoretical size of the pool of restriction fragments for the nucleus of each species and cell type. For *Drosophila*, we used a diploid male cell line. Thus, the total number of restriction fragments was ~600,000, composed of the double amount of fragments in autosomes (2 × 265,167, as assessed by the dm3 in silico digestion) in addition to the number of fragments on chromosome X (64,108). For mice, Flyamer et al. (2017) analyzed oocytes with four copies of the genome, resulting in a total of 4 × 6,407,802 ~ 25,600,000 fragments. Gassler et al. (2017) analyzed G2 zygotes pronuclei with two copies of the genome, resulting in a total of 2 × 6,407,802 ~ 12,800,000 fragments (we did not distinguish between the maternal and paternal pronuclei because the contribution of chromosome X is not as significant for the mouse genome).

We next assessed the upper limit of the total number of possible contacts per single nucleus, which is achieved when each restriction fragment formed two contacts with the ends of any other restriction fragments from the pool. Because the valency of each fragment is two, the theoretical upper limit is equal to the number of restriction fragments.

We then divided the total number of observed contacts (recovered by ORBITA) by the upper bound of the possible number of contacts, and we recovered up to ~16% of the total number of possible contacts for *Drosophila* (see Fig. 2b); this number is approximately 2.6% for the best mouse dataset. The mean percentage of recovered contacts is 4.9% for our dataset and <1% for Flyamer et al. (2017) and Gassler et al. (2017).

However, this assessment of the percentage of recovered contacts is not exact for several reasons: (1) we did not perform sorting prior to snHi-C to isolate G1 cells; hence, some regions of the genome might have an increased copy number in S or G2 cells; (2) some regions of the genome might be affected by deletions and copy number variations that were not accounted for in our analysis. However, even in the worst-case scenario, if we imagine that all *Drosophila* cells are in the G2 phase of the cell cycle, we recovered at least 8% of all possible contacts for the best cells in our analysis, which is still a substantial improvement compared to recovery for the best cells from mammalian studies.

TAD calling in snHi-C and bulk BG3 in situ Hi-C data. We used Hi-C map segmentation with *lavaburst* (v0.2.0) (https://github.com/nvictus/lavaburst) with the modularity scoring function for TAD calling in Hi-C maps at 10-kb resolution[32]. All TAD segments smaller or equal to 3 bins (30 kb) were considered to be inter-TADs[24]. *lavaburst* has a gamma (γ) parameter controlling the size and the number of resulting TADs. We varied g from 0 to 375 with a step of 0.1 for *Drosophila* datasets. The range and the step were selected to guarantee the comprehensive coverage of both extremes (a small amount of unusually large TADs and a large amount of smallest possible TADs). We observed a sharp decrease in median TAD size and an increase in the number of TADs with the γ increase (Fig. 3b, Supplementary Fig. 5). After reaching the peak, the number of TADs starts to drop because many segments fall beyond the minimal allowed TAD size. For large γ, both the number of TADs and mean TAD size reach a plateau at low levels. We considered the point of the maximum number of TADs (γ_{max}) as the most informative segmentation reachable by the algorithm for a particular dataset. The mean TAD size is ~70 kb on average between cells compared to the expected 120 kb size of *Drosophila* TADs[24]. Thus, we considered this level to be the sub-TADs. To guarantee a uniform γ selection procedure for all the cells, we arbitrarily selected $\gamma_{max}/2$ to obtain a resulting TAD segmentation (mean TAD size ~90 kb).

For the other resolutions of snHi-C maps, the same protocol of TAD calling was applied, except the inter-TAD size threshold was set to 60 kb (3 bins at 20 kb) for 20 kb and 120 kb (3 bins of 40 kb) for 40 kb.

Robustness of TAD calling. To assess TAD calling robustness and filter out potentially artifactual TAD boundaries, we performed TAD calling on snHi-C maps with random subsampling of the contacts as a control. For each cell, we performed ten iterations of independent subsampling of contacts leaving 95%, 90%, … 5% of the initial number of unique contacts per dataset. For each subsampling, we performed the TAD calling in the same manner as for the full dataset. We then assumed the bins found as TAD boundaries in the full snHi-C maps with no subsampling to be positives and inner TAD bins to be negatives. Based on this definition, we calculated both false positive rates (FPR) and false negative rates (FNR) for each cell and all subsampling levels. As expected, FNR gradually decreased with the percentage of remaining contacts. FPR reached a maxima at 10–30% subsampling level and then gradually decreased (Supplementary Fig. 6a, b).

We then defined a TAD boundary support for a given subsampling level (X%). TAD boundary support is calculated for each genomic bin as the number of subsampling iterations with the number of contacts equal to or larger than X%, where the bin was annotated as the TAD boundary (allowing a one-bin offset). We used TAD boundary support as a predictor of observed TAD boundaries in each cell (with no subsampling of the snHi-C dataset). We plotted receiver operating characteristic (ROC) curves for each X = (95%, 90%, … 5%) and calculated the ROC area under the curve (AUC) for each case (Supplementary Fig. 6c). Based on the largest ROC AUC, we selected the best subsampling level predictive of boundaries, X = 90% ROC AUC 0.9969 (Supplementary Fig. 6c). We then chose the TAD boundary support threshold by optimizing the accuracy. We obtained an accuracy of 0.9765 for the final criteria that the TAD boundary support is larger than 45% for (90%..95%) subsampling levels.

We refined the boundaries based on these final criteria and observed only a mild decrease in the number of boundaries per cell (Supplementary Fig. 6d). Thus, we conclude that the TAD calling procedure is robust to subsampling. We used the non-refined boundaries set in the paper if not stated otherwise.

For the refined boundaries set, we allowed a 10-kb offset for each boundary and assessed the number of cells in which each genomic bin was annotated as a boundary. We then defined the stable boundaries as bins that were annotated as boundaries in more than or equal to 50% of cells (> = 7), and unstable boundaries as the bins annotated as boundaries in less than 50% of cells (<7).

We compared stable boundaries with boundaries conserved between Kc167 and BG3 cells[46]. For that, we obtained TAD positions from[46], mapped them to the dm3 genome with liftover, and coarse-grained the coordinates to 10-kb bins. We then allowed the 10-kb offset and counted the boundaries that overlapped with stable boundaries obtained in the single-cell analysis.

Segmentation comparison. We introduced two types of similarity scores for TAD/sub-TAD segmentation comparison:

(1) the percentage of shared boundaries, where we fixed the first segmentation and compared it with the second segmentation. Each TAD boundary bin of the second segmentation was allowed to include two of its closest neighbors at a 10 kb distance (one bin offset). The number of shared boundaries between two segmentations was calculated as a simple intersection of sets. The percentage was calculated by division by the total number of bins annotated as TAD boundaries in the first segmentation.
(2) Jaccard index for TAD bins, where the bins inside a TAD (excluding the boundaries) were considered. The shared TAD bins between two segmentations were calculated and divided by the total number of bins annotated as TADs in both segmentations.

To assess the significance of obtained similarity score of TADs, we randomized the locations of TAD boundaries preserving the distributions of TAD and inter-TAD sizes and the number of TADs/inter-TADs per chromosome. Each randomization was performed 1000 times; the distribution of scores was approximated by Gaussian distribution; *p*-values were inferred from these backgrounds. The same procedure was used for sub-TADs.

Non-backtracking approach for annotation of TADs in single cells contact maps. The chromatin network, constructed on the basis of the single-cell Hi-C data, can be classified as sparse (i.e., the number of actual contacts per bin in a single-cell contact matrix (adjacency matrix of the network) is much less than the matrix size N). The sparsity of the data significantly complicates the community detection problem in single cells. It is known that upon dilution of the network, there is a fundamental resolution threshold for all community detection methods[82]. Furthermore, traditional operators (adjacency, Laplacian, modularity) fail far above this resolution limit (i.e., their leading eigenvectors become uncorrelated with the true community structure above the threshold)[43]. That is explained by the emergence of tree-like subgraphs (hubs) overlapping with true clusters in the isolated part of the spectrum for these operators. Localization on the hubs, but not on true communities in the network, is a drawback of all conventional spectral methods in the sparse regime.

To overcome the sparsity issue and to make spectral methods useful in the sparse regime, Krzakala et al.[43] proposed to construct the transfer-matrix of non-backtracking random walks (NBT) on a directed network. The NBT operator B is defined on the edges $i \to j$, $k \to l$ as follows:

$$B_{i\to j,k\to l} = \delta_{il}(1 - \delta_{jk}) \quad (1)$$

By construction, NBT walks cannot revisit the same node on the subsequent step and, thus, they do not concentrate on hubs. It has been shown that the non-backtracking operator is able to resolve the community structure in a sparse stochastic block model up to the theoretical resolution limit. In recently published paper[42], we have proposed the neutralized towards the expected contact probability NBT operator for the sake of a large-scale splitting of a sparse polymer network into two compartments.

Here, we are interested in the small-scale clustering into TADs, for which the conventional NBT operator is appropriate. To eliminate the compartmental signal from the data, we first cleansed all chromosome contact matrices starting from the diagonal, corresponding to 1 Mb separation distance (100th diagonal in the 10-kb resolution). To respect the polymeric nature of the contact matrices, we have filled all empty cells on the leading sub-diagonals with 1. Then, the NBT spectra of all single-cell contact matrices were computed. The majority of eigenvalues of the non-Hermitian NBT operator are located inside the disc in a complex plane, and some number of isolated eigenvalues with large amplitudes lie on the real axis. The edge of the isolated part of the spectrum was defined as the real part of the largest in absolute value eigenvalue with a non-zero imaginary part. All eigenvalues λ_i such that $R_e(\lambda_i) > r_c$ are isolated, and the corresponding eigenvectors correlate with annotation into the TADs. The position of the spectral edge, determined by the procedure above, has been found to be very close to the edge of the disk for the stochastic block model $r_c = \sqrt{d^{-1}\left\langle \frac{d}{d-1} \right\rangle}$, where d is the vector of degrees[83]. The typical number of the isolated eigenvalues was around 100 for dense contact matrices and somewhat less for sparser ones. The leading eigenvectors define the coordinates $u_j^{(i)}, j = 1, 2, \ldots, N$ of the nodes (bins) of the network in the space of reduced dimension $k << N$. At the second step, the clustering of the data was performed using the spherical k-means method, realized in the Python library *spherecluster*[84]. The number of isolated eigenvalues establishes a lower bound on the new space dimension k to be used for the clustering algorithm, since the respective leading eigenvectors are linearly independent. The dimension of the space k establishes a lower bound on the number of clusters because the leading eigenvectors are linearly independent. To take into account the hierarchical organization of TADs, we have communicated to the spherical k-means the number of clusters somewhat larger than the lower bound. Although the final splitting was found to be not particularly sensitive to this number, we have chosen to split the network into $2.5*k$ clusters in order to obtain the same mean amount of TADs per chromosome as with the modularity method (171 TADs).

The annotations produced by the spherical k-means on the single-cell Hi-C matrices were contiguous (i.e., the clusters were sequence respective, thus resembling TADs). The clusters (i) of size less than 30 kb and (ii) with amount of contacts equal to $2(l-1)$ (i.e., with no contacts other than on the sub-diagonals) were excluded from the set as the inter-TADs regions. The ultimate median size of the TADs across all single cells obtained by this algorithm was 110 kb (from 60 kb to 260 kb), and the mean chromosome coverage was 82% (from 57 to 93%). The same analyses of shuffled contact maps have revealed a similar number, size, and coverage of the domains, formed purely due to fluctuations. The boundaries of the NBT TADs in single cells were significantly conserved from cell to cell: the mean pairwise fraction of matched boundaries was 44% for all the cells and 59% for the five densest ones (for the shuffled cells with preservation of stickiness and scaling, see the MSS model; the mean pairwise fraction was 38 and 50% for the five densest cells).

Regarding the comparison of TAD boundaries with the modularity approach, the mean fraction of conserved modularity boundaries is somewhat less – 42% for all pairs of cells in the analyses and 52% for the five densest cells, whereas the number of TADs per chromosome is the same in the two methods (171). Between

the two methods, the mean number of matched boundaries for the corresponding cells is 61%.

Compartment annotation in snHi-C and bulk BG3 in situ Hi-C. For compartment annotation in bulk BG3 in situ Hi-C, we used eigenvector decomposition of cis-interactions maps for each chromosome, as implemented in *cooltools* call-compartments tool version 0.2.0 (https://github.com/mirnylab/cooltools). We then reversed the sign of eigenvalues based on GC content (positive values corresponding to an A compartment with larger GC content)[26]. We next carried out a saddle plot analysis for each snHi-C dataset based on bulk BG3 in situ Hi-C compartment annotation[32]. For this procedure, the bins in raw scHi-C maps were reordered by ascending first eigenvector values and averaged to 5 × 5 saddle plots[32].

Epigenetic analysis of TAD boundaries. For the functional annotation of TAD boundaries, we downloaded modENCODE normalized array files[85]: total RNA of ML-DmBG3-c2 cell line assessed by RNA tiling array (modENCODE id 713) and the ChIP-chip for MOF (id 3041), BEAF-32 (id 921), Chriz (275), CP190 (924), CTCF (3280), dmTopo-II (5058), GAF (2651), H1 (3299), HP1a (2666), HP1b (3016), HP1c (942), HP2 (3026), HP4 (4185), ISWI (3030), JIL-1 (3035), mod (mdg4) (324), MRG15 (3045), NURF301 (5063), Pc (325), RNA-polymerase-II (950), Su(Hw) (951), Su(var)3-7 (2671), Su(var)3-9 (952), WDS (5148), H3 (3302), H3K27ac (295), H3K27me3 (297), H3K36me1 (299), H3K36me3 (301), H3K4me1 (2653), H3K4me3 (967), H3K9me2 (310), H3K9me3 (312), H4K16ac (316). For RNA-Seq coverage, we used the data from ref. [24]. The files were binned at 10-kb resolution by summation.

We plotted the ChIP-chip signal around different types of boundaries with *pybbi* utility (https://github.com/nvictus/pybbi.git) based on UCSC tools[86] and constructed six sets of boundaries: boundaries found in the bulk in situ Hi-C, boundaries found in the merged snHi-C dataset, boundaries present in > = 50% of cells (> = 7 cells, stable boundaries), boundaries present in <50% of cells (<7 cells, unstable boundaries), boundaries present in just one single cell, and random boundaries. To obtain randomized boundaries, we shuffled bulk in situ Hi-C boundaries across the *Drosophila* genome, preserving the number of boundaries per chromosome. We also used the bins from the inner parts of TADs as a control for the epigenetic analysis.

Functional annotation of distant contacts. The 10-kb genomic bins were separated into four groups based on chromatin states for BG3 from Kharchenko et al.[54]: active chromatin (>0.5 of RED and MAGENTA color), inactive chromatin (>0.5 LIGHT GRAY), Polycomb chromatin (>0.5 DARK GRAY), and unannotated (all the rest) for functional annotation of distant contacts. The thresholds for functional enrichment of particular types of chromatin were selected in order to guarantee the selection of the regions with the most prominent properties of active/inactive/Polycomb chromatin.

The 10-kb genomic bins were split into five groups based on the average expression from two RNA-seq replicates in BG3 cells[24] (0 expression, 38.1–40%, 40–60%, 60–80%, top 20% expression) for expression activity annotation. We were not able to split the data using an even grid of percentiles (e.g., 0–20%, 20–40%) because ~38% of all genomic bins had zero expression in both replicates. The same functional annotation was used later for polymer model coloring.

Average loop. For the construction of an average loop of A-compartment regions (Fig. 4f) and B compartment regions (Fig. 4g), MSL complex (Fig. 4h) and Polycomb (Fig. 4i), we selected the top 1000 genomic regions with the highest abundance of the corresponding genomic annotations as potential looping positions. A and B compartments were assessed by a *cis*-derived eigenvector of the bulk BG3 Hi-C data. MSL ChIP-Seq was obtained from Ramirez et al.[51], GEO ID GSE58821). dRING binding data were obtained from modENCODE as a ChIP-chip normalized array file (ID 927[54]). We considered the pairs of potential looping positions corresponding to intrachromosomal interactions, at the genomic distances of more than 600 kb, separated by up to 50 other looping positions. The snipping of Hi-C square 600-kb windows, centered on the corresponding looping positions, was done with cooltools (https://github.com/mirnylab/cooltools/tree/master/cooltools). The aggregation was performed by summation. $\log_{10}$ values were plotted as heatmaps.

Assessment of folding hierarchy of TADs. To assess the folding hierarchy at the level of TADs, we used the assumption that the successive sub-TADs that form the same TAD will have more interactions in the observed real snHi-C maps than in the control maps described in the section "snHi-C background model" of these Methods. We calculated the number of contacts directly from snHi-C maps and the control maps. Only sequential sub-TADs falling into the same TAD were considered. The distribution of the number of contacts in the windows between sequential sub-TADs was calculated. We compared the distributions of the number of contacts between sub-TADs falling into the same TAD for real snHi-C maps and the control maps. For each cell, we used either TAD/sub-TAD annotations from the corresponding snHi-C map or TAD/sub-TAD annotation from bulk in situ Hi-C.

Marginal scaling (MS) and marginal scaling and stickiness (MSS) models. We carried out the statistical analysis of the single-cell Hi-C maps to provide statistical arguments supporting the premise that the clustering observed in snHi-C contact matrices "is not random". For this, we used two different models of a polymer network based on Erdos-Renyi graphs, where bins of the contact map resemble graph vertices, and contacts between bins are graph edges[87] (Supplementary Fig. 4a):

(a) In the MS model, we require the probability of contact between nodes to respect the contact probability of the experimental contact map, i.e. $P(s) = P_c(|i-j|)$. Decay of the contact probability originates from the intrinsic linear connectivity of the chromatin nodes; therefore, it is an important ingredient for studying fluctuations in a polymer network. The probability of the link between i and j in the random graph I, $j = 1, 2\ldots, N$ is, thus, defined as follows:

$$p_{ij} = \frac{P_c(|i-j|)}{\sum_{s=1}^{N-1}(N-s)P_c(s)}N_c \quad (2)$$

where the normalization factor in the denominator guarantees that the mean number of links in the graph equals N_c (i.e., the number of experimentally observed links in each single cell). To obtain the average scaling, we merge all contacts from the available single cells and compute the average $P_c(s)$. Given the probability p_{ij} by Eq. 2, we randomly generate adjacency matrices that have a homogenous distribution of contacts along the diagonals and do not respect local peculiarities of the bins, such as insulation score, acetylation, and protein affinity. Nevertheless, some non-homogeneity (clustering) of contacts still emerges as a result of stochasticity in each realization of this graph (Supplementary Fig. 4e).

(b) the MSS model introduces probabilistic non-homogeneity along the diagonals of the adjacency matrices through definition of the "stickiness" of bins, or. Specifically, under "stickiness", we understand a non-selective affinity k_i of a bin i to other bins; the probability that the bin i forms a link with any other bin in the polymer graph is proportional to its stickiness. Thus, the clusters of contacts close to the main diagonal of contact matrices form as a result of different "stickiness" of bins in the MSS model. Stickiness might effectively emerge as a result of a particular distribution of "sticky" proteins, such as PcG proteins known to mediate bridging interactions between nucleosomes and to participate in stabilization of the repressed chromatin state.

Assuming that the stickiness is distributed independently of the polymer scaling $P_c(|i-j|)$, we use the following expression for the probability of the link, p_{ij}, in the MSS model:

$$p_{ij} = \frac{k_i k_j P_c(|i-j|)}{\sum_{i<j} k_i k_j P_c(|i-j|)}N_c \quad (3)$$

To derive the values of stickiness, we calculated the coverage at each bin in the merged contact map $\tilde{k}_i$, which stands for the average number of contacts at a particular bin. Due to the polymer scaling, the rates of contacts along each row (column) vary. Thus, $\tilde{k}_i$ is not equal to stickiness, $\tilde{k}_i \neq k_i$. To determine the stickiness values k_i, one should correlate the experimental coverage $\tilde{k}_i$ with the theoretical mean number of contacts per bin, according to Eq. 3:

$$\tilde{k}_i = \sum\nolimits_j p_{ij} = k_i \alpha_i \quad (4)$$

where is "activity" of surrounding bins, measured for the i-th bin:

$$\alpha_i = \frac{1}{Z}\sum\nolimits_j k_j P_c(|i-j|),\ Z = \frac{1}{N_c}\sum\nolimits_{i<j} k_i k_j P_c(|i-j|) \quad (5)$$

Equation 3 sets a system of N non-linear equations that cannot be solved analytically. To determine the stickiness values, we implement the numerical method of iterative approximations. Namely, we start with:

$$k_i^{(0)} = \tilde{k}_i,\ \alpha_i^{(0)} = \alpha_i(\tilde{k}_i) \quad (6)$$

and recalculate $k_i^{(1)}$ using Eqs. (4, 5) at the second step. After several recursive steps, we find good convergence of the stickiness and activity to their limiting values k_i^{∞} and α_i^{∞}. In particular, the derived values of the stickiness provide a good estimate for the averaged theoretical coverage $\tilde{k}_i$ as compared to the experimental coverage; see Supplementary Fig. 4f, g. Therefore, the derived null-model of single-cell maps reproduces, on average, the observed coverage of contacts of each bin by means of the individual stickiness assignment. We would like to point out the difference between the limiting values of the stickiness and $\tilde{k}_i$, used as a starting approximation in the iterative procedure; Supplementary Fig. 4h. This difference is a result of the non-homogeneous redistribution of contacts at each particular row in accordance with the marginal polymeric scaling $P_c(|i-j|)$.

Number of contacts in windows. The MS and MSS models introduced above demonstrate apparent clustering of generated contacts close to the main diagonal in realizations of adjacency matrices. In the MS model, this is purely due to fluctuations: the mean weight of the link $w_{ij} = p_s$ depends only on the genomic

distance between the bins $s = |i - j|$ in the respective Poisson version of the weighted network. In contrast, in the MSS model, the non-homogeneity of bin sicknesses allows for a deterministic non-homogeneous distribution of contacts along the main diagonal.

To statistically compare the clustering of contacts generated by the two models with the clustering in experimental single cell Hi-C maps, we studied distributions of the number of contacts in certain "windows" of different sizes. The inspected windows are isosceles triangles with the base located on the main diagonal and having the angle with the congruent sides. These windows look like TADs but, in contrast to the latter, have a fixed size throughout the genome.

At a given window size W, we sampled the number of contacts falling in the defined windows in each snHi-C map. We compared the samples originating from 100 random MS-generated maps and 100 random MSS-generated maps with derived limiting values of stickiness (see the previous section for discussion of the models).

Note that in the theoretical models (MS and MSS), all contacts are statistically independent: in both models, the number of contacts falling in a window of size can be interpreted as a number of "successes" occurring independently in a certain fixed interval. In the MS model, the "success" rate is constant along each diagonal; thus, for rather sparse MS maps (i.e. sufficiently small rates), one would expect the observed contacts in the windows to follow the Poisson distribution. In the MSS maps, the stickiness distributions introduce non-homogeneity to "success" rates along the diagonals; however, as our analyses suggest, the random MSS maps exhibit much more satisfactory Poisson statistics than their original experimental counterparts; Supplementary Fig. 4j, k.

Deviations from the Poisson statistics of the snHi-C contact maps are evaluated by the p-value of the χ^2 goodness of fit test (Supplementary Fig. 4k). The heatmaps of the common logarithm of p-values for the top-10 single cells and the corresponding MS and MSS maps are presented in Supplementary Fig. 4j. The random maps (the second and third rows) demonstrate reasonably even distributions of the p-values across distinct single cells that rarely enter below the significance level $\alpha = 10^{-5}$. Several atypically low p-values correspond either to the most dense single cells and small window sizes (upper-left corner), for which the sparse Poisson limit is violated, or to a quite uneven distribution of stickiness for a given chromosome. Notably, the snHi-C maps demonstrate remarkable deviations from the Poisson statistics for small window size $W < 40$ bins (<400 kb). As can be seen from the heatmaps (Supplementary Fig. 4j) the χ^2 test rejects the null hypothesis at the significance level $\alpha = 10^{-5}$ for most of the single cells at small scales. Therefore, the probability that the experimental contact maps are described by the Poisson statistics is significantly low (α).

To understand the source of inconsistency between the experimental and Poisson distributions, we plotted the histograms of the number of contacts along with their best Poisson-fit for $W = 10$ (Supplementary Fig. 4k, left) and $W = 40$ (Supplementary Fig. 4k, right). The presence of large-scale heavy tails and low-scale shoulders in the experimental histograms results in the rejection of the null hypothesis.

Finally, the samples corresponding to larger windows are notably better described by the Poisson distribution, exhibiting a level of p-values similar to the random maps. The crossover $W_0 \approx 40$ (400 kb) corresponds to the scale of 3–4 typical TADs; this implies that the positioning of the contacts inside a single TAD is sufficiently correlated. Correlations between the contacts of different pairs of loci can originate from a specific non-ideal folding of chromatin (e.g., fractal globule) or be a signature of active processes (e.g., loop extrusion) operating at the scale of one TAD. Larger window sizes accumulate contacts from different TADs, whereas most of the inter-TADs contacts are much less correlated. As a result, we see reasonable Poisson statistics of the number of contacts from larger windows with $W > W_0$. Taken together, we conclude that correlations in contacts is a structural feature of experimental single cell maps and that clusters (TADs) identified in the maps cannot be reduced to random fluctuations imposed by the white noise or imperfections of the experimental setup.

Fluorescence in situ hybridization. The cells were harvested overnight on poly-l-lysine coated coverslips placed in culture flasks. The cells were fixed in 4% paraformaldehyde for 10 min, permeabilized in 0.5% Triton X-100, washed in PBS, dehydrated in ethanol series, air-dried, stored at room temperature for 2 days, and then frozen at −80 °C. Probes were prepared from fosmids by labeling with fluorophore-conjugated dUTPs using nick-translation. Approximately 150 ng of each probe was used in hybridization. Denaturation was performed at 80 °C for 30 min in 70% formamide (pH 7.5), 2× SSC. Hybridization of probes was done for 24 h in 50% formamide, 2× SSC, 10% dextran sulfate, 1% Tween 20. Washing steps were performed in 2× SSC at 45 °C followed by 0.1× SSC at 60 °C and 4× SSC, 0.1% Triton X-100. For imaging, cells were counterstained with DAPI, and epifluorescent images were acquired using a microscope setup comprising a Zeiss Axiovert 200 fluorescence microscope (Carl Zeiss UK, Cambridge, UK), X-Cite ExFo 120 Mercury Halide (Exfo X-cite 120, Excelitas Technologies) fluorescent source with liquid light guide and 10-position excitation, neutral density, and emission filter wheels (Sutter Instrument, Novato, CA), ASI PZ2000 3-axis XYZ stage with integrated piezo Z-drive (Applied Scientific Instrumentation, Eugene, OR), Retiga R1 CCD camera (Qimaging, Surrey, BC, Canada). The filter wheels were populated with a #89903 ET BV421/BV480/AF488/AF568/AF647 quinta set (Chroma Technology Corp., Rockingham, VT). Image capture was performed using Micromanager 1.4 (https://open-imaging.com/). Hardware control and image capture were carried out using μManager[88]. Images were deconvolved using Nikon NIS-Elements. Measurements were taken using Imaris.

Polymer simulations. Simulation of 3D chromatin fiber enabled substantiation of assumptions about factors that play key roles in chromatin organization and to obtain important information about its packaging. We focused on the static properties of the system and did not consider its dynamic properties.

Modeling pipeline, general description of the procedure. Many methods are currently used to perform computer modeling of polymers. Due to the actual size and complexity of the chromatin, the all- or united-atom model cannot be used to simulate spatial scales of interest. The dissipative particle dynamics (DPD) technique was used because it enables modeling of the physical properties of polymer systems[59]. DPD is a coarse-grain method of molecular dynamics. Newton's equations are solved numerically for each particle in the system for every time step. The total force consists of conservative, dissipative, random, and elastic forces.

Conservative force is described by a soft potential within the sphere with cutting radius $R_c = 1.0$. The soft potential has no singularity at the zero point (Supplementary Fig. 21a). It is possible to use a large time step in the Velocity Verlet integration scheme, in contrast to classical molecular dynamics (CMD) with the Lennard-Jones potential. The typical time step in CMD is 20 times smaller than in DPD. The solvent is taken into account explicitly; it is necessary for the DPD thermostat to work[89,90]. The temperature control of the system is ensured by a balance of dissipative and random forces that conserve the momentum. The elastic force simulates the presence of a bond between beads. An ensemble of NVT (number of particles, volume, temperature) is used. A detailed description of the simulation method can be found elsewhere[91]. We used our own implementation of DPD that is 2D parallelized and lightweight[92].

In all simulations, the following parameters were used: $a_{pp} = a_{ss} = 25.0$, $a_{ps} = 26.63$ (soft potential repulsion coefficient), in terms of Flory-Huggins' theory $\chi = 0.5 = 0.306*(a_{ps} - a_{pp})$, where a_{pp}—repulsion coefficient between polymer and polymer beads, a_{ss}—between solvent and solvent beads, a_{ps}—between polymer and solvent beads; $l_0 = 0.5$ (undeformed bond length), $k = 40$ (bond stiffness), $dt = 0.04$ (integration timestep), $\sigma = 3$ (number density), simulation box size $22 \times 22 \times 22$ DPD a.u.

With these parameters, the polymer chain (or chromatin fiber) is able to self-intersect but still has an effective excluded volume. At $\chi = 0.5$, the single polymer chain in a dilute solution has a Gaussian conformation (i.e. it corresponds to a simple random walk).

Each simulation was organized as follows:

The polymer chain is generated as a random walk within the cubic cell with the size of 10 DPD units. Adjacent solvent particles are included into the simulation cell with the size of 22 DPD units until the number density $\sigma = 3$. Additional bonds between beads are added according to the snHi-C contact matrix. If i-th and j-th beads have a contact, an additional harmonic bond between i-th and j-th beads is added to the system if $|i - j| > 1$. We define contact as an event when the distance between two beads (i, j) meets criterion $D_{ij} < R_{cut} = 0.7$ Such R_{cut} value corresponds to the average bond length. We count all the contacts in the system. So, in a system any bead can have more than 1 contact. Additional bonds could be overstretched; therefore, the system is equilibrated over 106 steps. The simulation time is two orders of magnitude higher than the necessary equilibration time (Supplementary Fig. 21b); hence, there are no doubts regarding the system equilibrium. According to our calculations, the equilibration time is ~20k steps. The equilibrated system contained overstretched bonds, which were removed one by one until the maximum length became less than the threshold $l_{max} < 1.5$ DPD a.u. (Supplementary Fig. 21c, Supplementary Table 2). Backbone bonds were not removed, because they represented reliable information. The system was equilibrated for 20k steps after each bond removal.

Values of the single-cell Hi-C matrix elements could vary because the restriction fragment is smaller than the selected resolution (10 kb). Data regarding the exact number of contacts between two fragments were not used. Therefore, the contact matrix was considered to be binary. Only the X chromosome was simulated because it is haploid. The X chromosome corresponds to the polymer chain consisting of 2242 beads at 10 kb resolution. Every single chain bead represents 50 nucleosomes. Our model does not consider the shape of a 10-kb region or any other internal properties.

Control simulations were organized in the same manner, but the contacts were shuffled. Shuffling was performed while maintaining the number of contacts at each genomic distance. We also performed simulations with shuffling on the long genomic distances only and sampling the contacts from two cells (Supplementary Table 3). The second case shows that reconstruction of the 3D conformation from diploid chromosomes is meaningless in comparison with haploid chromosomes.

Coefficient of the difference. To compare two 3D structures, corresponding distance matrices were calculated. Orientation of the chain in 3D space did not affect the elements of distance matrices. The Coefficient of the difference is introduced as $K = M_{asym}/M_{sym}$, where $M_{asym} = ||D–D'||/2$ and $M_{sym} = ||D–D'||/2$,

where D and D'—distance matrices. ||Matrix||—is the Euclidean distance ($d = \sqrt{a_{11}^2 + a_{12}^2 + .. + a_{21}^2 + \ldots}$, $a_{\#\#}$—matrix element). To avoid the contribution of thermal fluctuations, each distance matrix was averaged over 100 conformations with an output rate of 10k steps.

To demonstrate the independence of the final result on the initial conformation, we repeated the calculation of the system ten times with the maximal number of contacts. For each repeat, we created a new independent initial conformation, but we kept the same set of additional bonds. The initial conformation does not affect the final result in the simulation protocol.

Visualization of epigenetic states. The visualization was performed using the pymol software v. 2.3.2 (https://pymol.org/2/). 1D epigenetic data were added to the structure as a bead type and represented with a corresponding color. Analysis of different epigenetic states was performed via Python scripts (https://github.com/polly-code/DPD_withRemovingBonds). Before the visualization, some of the conformations were smoothed by averaging coordinates within the window of 15 beads along the chain. This approach ensured that thermal fluctuations were avoided (Supplementary Figs. 16, 21).

Radial distances and center of mass. We calculated the surface of the chromosome territory as a convex hull. The distance to the surface was evaluated as the minimal distance from the particle to the surface, and then the distance arrays were averaged.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw and processed snHi-C and bulk BG3 in situ Hi-C data are available in the GEO NCBI under accession number "GSE131811". List of publicly available GEO sources used in this study: "GSE122603" (Hi-C for Kc167 and BG3 cell lines for comparison of stable TAD boundaries), "GSE58821" (MSL; ChIP-seq), "GSE69013" (RNA-Seq). List of publicly available modENCODE data sources used in this study: total RNA of ML-DmBG3-c2 cell line assessed by RNA tiling array (modENCODE id 713) and the ChIP-chip for MOF (id 3041), BEAF-32 (id 921), Chriz (id 275), CP190 (id 924), CTCF (id 3280), dmTopo-II (id 5058), GAF (id 2651), H1 (id 3299), HP1a (id 2666), HP1b (id 3016), HP1c (id 942), HP2 (id 3026), HP4 (id 4185), ISWI (id 3030), JIL-1 (id 3035), mod(mdg4) (id 324), MRG15 (id 3045), NURF301 (id 5063), Pc (id 325), RNA-polymerase-II (id 950), Su(Hw) (id 951), Su(var)3-7 (id 2671), Su(var)3-9 (id 952), WDS (id 5148), H3 (id 3302), H3K27ac (id 295), H3K27me3 (id 297), H3K36me1 (id 299), H3K36me3 (id 301), H3K4me1 (id 2653), H3K4me3 (id 967), H3K9me2 (id 310), H3K9me3 (id 312), H4K16ac (id 316). dRING binding data were obtained from modENCODE as a ChIP-chip normalized array file (id 927). All other relevant data supporting the key findings of this study are available within the article and its Supplementary Information files or from the corresponding author upon reasonable request. A reporting summary for this Article is available as a Supplementary Information file. Source data are provided with this paper.

Code availability

The data processing pipeline is available at https://github.com/agalitsyna/sc_dros. The modeling pipeline is available at https://github.com/polly-code/DPD_withRemovingBonds.

Received: 13 February 2020; Accepted: 23 November 2020;
Published online: 04 January 2021

References

1. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
2. Kim, T. H. & Dekker, J. 3C-based chromatin interaction analyses. *Cold Spring Harbor protoc.* https://doi.org/10.1101/pdb.top097832 (2018).
3. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
4. Nora, E. P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
5. Sexton, T. et al. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148**, 458–472 (2012).
6. Lupianez, D. G. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
7. Symmons, O. et al. Functional and topological characteristics of mammalian regulatory domains. *Genome Res.* **24**, 390–400 (2014).
8. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin domains: The Unit of Chromosome Organization. *Mol. CeLL* **62**, 668–680 (2016).
9. Franke, M. et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016).
10. Akdemir, K. C. et al. Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nat. Genet.* **52**, 294–305 (2020).
11. Schwarzer, W. et al. Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51–56 (2017).
12. Rao, S. S. P. et al. Cohesin loss eliminates all loop domains. *Cell* **171**, 305–320 e324 (2017).
13. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
14. Hildebrand, E. M. & Dekker, J. Mechanisms and Functions of Chromosome Compartmentalization. *Trends Biochem Sci.* **45**, 385–396 (2020).
15. Drucker, J. L. & King, D. H. Management of viral infections in AIDS patients. *Infection* **15**, S32–S33 (1987).
16. Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N. & Mirny, L. A. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proc. Natl Acad. Sci. USA* **115**, E6697–E6706 (2018).
17. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
18. Fudenberg, G. et al. Formation of chromosomal domains by loop extrusion. *Cell Rep.* **15**, 2038–2049 (2016).
19. Sanborn, A. L. et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl Acad. Sci. USA* **112**, E6456–E6465 (2015).
20. Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nat. Rev. Genet.* **19**, 789–800 (2018).
21. Wutz, G. et al. Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J.* **36**, 3573–3599 (2017).
22. Matthews, N. E. & White, R. Chromatin architecture in the fly: living without CTCF/cohesin loop extrusion?: Alternating chromatin states provide a basis for domain architecture in *Drosophila*. *BioEssays* **41**, e1900048 (2019).
23. Rowley, M. J. et al. Evolutionarily conserved principles predict 3D chromatin organization. *Mol. Cell* **67**, 837–852 e837 (2017).
24. Ulianov, S. V. et al. Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Res.* **26**, 70–84 (2016).
25. Wang, Q., Sun, Q., Czajkowsky, D. M. & Shao, Z. Sub-kb Hi-C in D. melanogaster reveals conserved characteristics of TADs between insect and mammalian cells. *Nat. Commun.* **9**, 188 (2018).
26. Ramirez, F. et al. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.* **9**, 189 (2018).
27. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58**, 610–620 (2015).
28. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
29. Clark, S. J., Lee, H. J., Smallwood, S. A., Kelsey, G. & Reik, W. Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome Biol.* **17**, 72 (2016).
30. Fraser, J., Williamson, I., Bickmore, W. A. & Dostie, J. An Overview of Genome Organization and How We Got There: from FISH to Hi-C. *Microbiol Mol. Biol. Rev.* **79**, 347–372 (2015).
31. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013).
32. Flyamer, I. M. et al. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* **544**, 110–114 (2017).
33. Nagano, T. et al. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* **547**, 61–67 (2017).
34. Gassler, J. et al. A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture. *EMBO J.* **36**, 3600–3618 (2017).
35. Bintu, B. et al. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* https://doi.org/10.1126/science.aau1783 (2018).
36. Cardozo Gizzi, A. M. et al. Microscopy-based chromosome conformation capture enables simultaneous visualization of genome organization and transcription in intact organisms. *Mol. Cell* **74**, 212–222 e215 (2019).
37. Szabo, Q. et al. TADs are 3D structural units of higher-order chromosome organization in *Drosophila*. *Sci. Adv.* **4**, eaar8082 (2018).
38. Cattoni, D. I. et al. Single-cell absolute contact probability detection reveals chromosomes are organized by multiple low-frequency yet specific interactions. *Nat. Commun.* **8**, 1753 (2017).
39. Murthy, V., Meijer, W. J., Blanco, L. & Salas, M. DNA polymerase template switching at specific sites on the phi29 genome causes the in vivo accumulation of subgenomic phi29 DNA molecules. *Mol. Microbiol.* **29**, 787–798 (1998).
40. Lasken, R. S. & Stockwell, T. B. Mechanism of chimera formation during the multiple displacement amplification reaction. *BMC Biotechnol.* **7**, 19 (2007).

41. Imakaev, M. et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012).
42. Polovnikov, K., Gorsky, A., Nechaev, S., Razin, S. V. & Ulianov, S. V. Non-backtracking walks reveal compartments in sparse chromatin interaction networks. *Sci. Rep.* https://doi.org/10.1038/s41598-020-68182-0 (2020).
43. Krzakala, F. et al. Spectral redemption in clustering sparse networks. *Proc. Natl Acad. Sci. USA* **110**, 20935–20940 (2013).
44. Hansen, A. S., Cattoglio, C., Darzacq, X. & Tjian, R. Recent evidence that TADs and chromatin loops are dynamic structures. *Nucleus* **9**, 20–32 (2018).
45. Luzhin, A. V. et al. Quantitative differences in TAD border strength underly the TAD hierarchy in *Drosophila* chromosomes. *J. Cell Biochem.* **120**, 4494–4503 (2019).
46. Chathoth, K. T. & Zabet, N. R. Chromatin architecture reorganization during neuronal cell differentiation in *Drosophila* genome. *Genome Res.* **29**, 613–625 (2019).
47. Wang, X. T., Cui, W. & Peng, C. HiTAD: detecting the structural and functional hierarchies of topologically associating domains from chromatin interactions. *Nucleic Acids Res.* **45**, e163 (2017).
48. Schwartz, Y. B. & Cavalli, G. Three-dimensional genome organization and function in drosophila. *Genetics* **205**, 5–24 (2017).
49. Ulianov, S. V. et al. Nuclear lamina integrity is required for proper spatial organization of chromatin in Drosophila. *Nat. Commun.* **10**, 1176 (2019).
50. Rowley, M. J. et al. Condensin II counteracts cohesin and RNA polymerase II in the establishment of 3D chromatin organization. *Cell Rep.* **26**, 2890–2903 e2893 (2019).
51. Ramirez, F. et al. High-affinity sites form an interaction network to facilitate spreading of the MSL complex across the X chromosome in *Drosophila*. *Mol. Cell* **60**, 146–162 (2015).
52. Eagen, K. P., Aiden, E. L. & Kornberg, R. D. Polycomb-mediated chromatin loops revealed by a subkilobase-resolution chromatin interaction map. *Proc. Natl Acad. Sci. USA* **114**, 8764–8769 (2017).
53. Ogiyama, Y., Schuettengruber, B., Papadopoulos, G. L., Chang, J. M. & Cavalli, G. Polycomb-dependent chromatin looping contributes to gene silencing during *Drosophila* development. *Mol. CeLL* **71**, 73–88 e75 (2018).
54. Kharchenko, P. V. et al. Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature* **471**, 480–485 (2011).
55. Osborne, C. S. et al. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat. Genet.* **36**, 1065–1071 (2004).
56. Iborra, F. J., Pombo, A., Jackson, D. A. & Cook, P. R. Active RNA polymerases are localized within discrete transcription "factories' in human nuclei. *J. Cell Sci.* **109**, 1427–1436 (1996).
57. Quinodoz, S. A. et al. Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* **174**, 744–757 e724 (2018).
58. Chen, Y. et al. Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *J. Cell Biol.* **217**, 4025–4048 (2018).
59. Español, P. & Warren, P. B. Perspective: dissipative particle dynamics. The. *J. Chem. Phys.* **146**, 150901 (2017).
60. Stevens, T. J. et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**, 59–64 (2017).
61. Chertovich, A. & Kos, P. Crumpled globule formation during collapse of a long flexible and semiflexible polymer in poor solvent. *J. Chem. Phys.* **141**, 134903 (2014).
62. Shevelyov, Y. Y. & Ulianov, S. V. The nuclear lamina as an organizer of chromosome architecture. *Cells* https://doi.org/10.3390/cells8020136 (2019).
63. Pirrotta, V. & Li, H. B. A view of nuclear Polycomb bodies. *Curr. Opin. Genet Dev.* **22**, 101–109 (2012).
64. Razin, S. V. et al. Transcription factories in the context of the nuclear and genome organization. *Nucleic Acids Res.* **39**, 9085–9092 (2011).
65. Robson, M. I., Ringel, A. R. & Mundlos, S. Regulatory landscaping: how enhancer-promoter communication is sculpted in 3D. *Mol. CeLL* **74**, 1110–1122 (2019).
66. Loubiere, V., Martinez, A. M. & Cavalli, G. Cell fate and developmental regulation dynamics by polycomb proteins and 3D genome architecture. *BioEssays* **41**, e1800222 (2019).
67. Cook, P. R. & Marenduzzo, D. Transcription-driven genome organization: a model for chromosome structure and the regulation of gene expression tested through simulations. *Nucleic Acids Res* **46**, 9895–9906 (2018).
68. Rhodes, J. D. P. et al. Cohesin disrupts polycomb-dependent chromosome interactions in embryonic stem cells. *Cell Rep.* **30**, 820–835 e810 (2020).
69. Banigan, E. J. & Mirny, L. A. Loop extrusion: theory meets single-molecule experiments. *Curr. Opin. Cell Biol.* **64**, 124–138 (2020).
70. Costantino, L., Hsieh, T.-H. S., Lamothe, R., Darzacq, X. & Koshland, D. Cohesin residency determines chromatin loop patterns. *eLife* **9**, e59889 (2020).
71. Brandao, H. B. et al. RNA polymerases as moving barriers to condensin loop extrusion. *Proc. Natl Acad. Sci. USA* **116**, 20489–20499 (2019).
72. Davidson, I. F. et al. Rapid movement and transcriptional re-localization of human cohesin on DNA. *EMBO J.* **35**, 2671–2685 (2016).
73. Yoshizawa, T., Nozawa, R. S., Jia, T. Z., Saio, T. & Mori, E. Biological phase separation: cell biology meets biophysics. *Biophysical Rev.* **12**, 519–539 (2020).
74. Kumar, G., Garnova, E., Reagin, M. & Vidali, A. Improved multiple displacement amplification with phi29 DNA polymerase for genotyping of single human cells. *Biotechniques* **44**, 879–890 (2008).
75. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
76. Li, H. & Durbin, R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
77. Abdennur, N. & Mirny, L. A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36**, 311–316 (2020).
78. Gavrilov, A. A., Gelfand, M. S., Razin, S. V., Khrameeva, E. E. & Galitsyna, A. A. "Mirror reads" in Hi-C data. *Genomics Comput. Biol.* **3**, 36 (2017).
79. Kerpedjiev, P. et al. HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol.* **19**, 125 (2018).
80. Yang, T. et al. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res.* **27**, 1939–1949 (2017).
81. Chandradoss, K. R. et al. Biased visibility in Hi-C datasets marks dynamically regulated condensed and decondensed chromatin states genome-wide. *BMC Genomics* **21**, 175 (2020).
82. Decelle, A., Krzakala, F., Moore, C. & Zdeborova, L. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E* **84**, 066106 (2011).
83. Newman, M. E. J. Spectral methods for community detection and graph partitioning. *Phys. Rev.* https://doi.org/10.1103/PhysRevE.88.042822 (2013).
84. Banerjee, A., Dhillon, I. S., Ghosh, J. & Sra, S. Clustering on the unit hypersphere using von Mises-Fisher distributions. *J. Mach. Learn Res* **6**, 1345–1382 (2005).
85. Celniker, S. E. et al. Unlocking the secrets of the genome. *Nature* **459**, 927–930 (2009).
86. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207 (2010).
87. Anderson, G. W., Guionnet, A. & Zeitouni, O. *An Introduction to Random Matrices* (Cambridge University Press, 2010).
88. Edelstein, A. D. et al. Advanced methods of microscope control using muManager software. *J. Biol. Methods* https://doi.org/10.14440/jbm.2014.36 (2014).
89. Hoogerbrugge, P. J. & Koelman, J. M. V. A. Simulating microscopic hydrodynamic phenomena with dissipative particle dynamics. *Europhys. Lett.* **19**, 155–160 (1992).
90. Koelman, J. M. V. A. & Hoogerbrugge, P. J. Dynamic simulations of hard-sphere suspensions under steady shear. *Europhys. Lett.* **21**, 363–368 (1993).
91. Groot, R. D. & Warren, P. B. Dissipative particle dynamics: bridging the gap between atomistic and mesoscopic simulation. *J. Chem. Phys.* **107**, 4423–4435 (1997).
92. Gavrilov, A. A., Chertovich, A. V., Khalatur, P. G. & Khokhlov, A. R. Effect of nanotube size on the mechanical properties of elastomeric composites. *Soft Matter.* **9**, 4067 (2013).

Acknowledgements

This work was supported by Russian Science Foundation (RSF) grant #19-14-00016 to S.V.R. Bioinformatics analysis of the data was supported by RSF grant #19-74-00112 to E.E.K and Russian Foundation for Support of Fundamental Science (RFBR) grant #18-29-13013 to S.K.N. A.A.Gal. was supported by RFBR grant #19-34-90136. The research is carried out using the equipment of the shared research facilities of HPC computing resources at Lomonosov Moscow State University and the Makarich HPC cluster provided by the Faculty of Bioengineering and Bioinformatics. The research of P.I.K. is supported partly by RFBR grant #18-29-13041 and by Skoltech Systems Biology Fellowship. The research of A.V.C. is supported by RFBR grant #18-29-13041. S.V.U. and S.V.R. were supported by the Interdisciplinary Scientific and Educational School of Moscow University «Molecular Technologies of the Living Systems and Synthetic Biology». We thank the Center for Precision Genome Editing and Genetic Technologies for Biomedicine, IGB RAS, and IGB RAS facilities supported by the Ministry of Science and Higher Education of the Russian Federation for providing research equipment.

Author contributions

S.V.R., S.V.U., and I.M.F. conceived the project; D.G. performed cell sorting; V.V.Z. and Y.S.V. prepared snHi-C and bulk BG3 in situ Hi-C libraries; A.A.Gal., K.E.P., E.E.K., S.V.U., A.A.Gav., A.S.G., S.K.N., and M.S.G. analyzed snHi-C, bulk BG3 in situ Hi-C, and publicly available data; P.I.K. and A.V.C. performed polymer simulations; I.M.F. performed FISH; E.A.M. and Y.Y.S. maintained cell cultures; M.D.L. performed sequencing of snHi-C and bulk BG3 in situ Hi-C libraries; S.V.U., V.V.Z., Y.S.V., A.A.Gal., E.E.K., and S.V.R. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-20292-z.

Correspondence and requests for materials should be addressed to S.V.R.

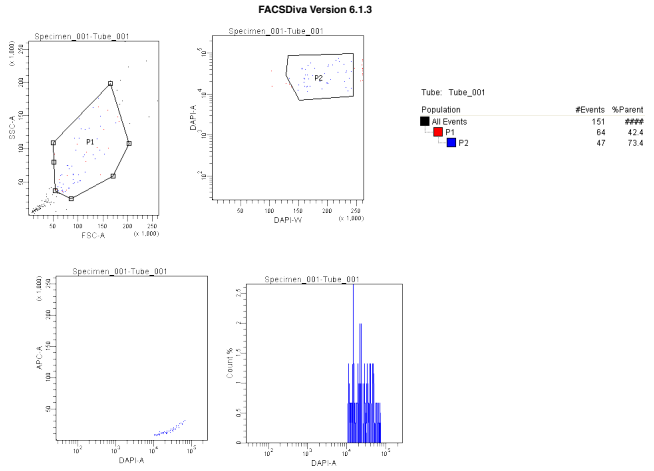
Peer review information *Nature Communications* thanks Nicolae Radu Zabet and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

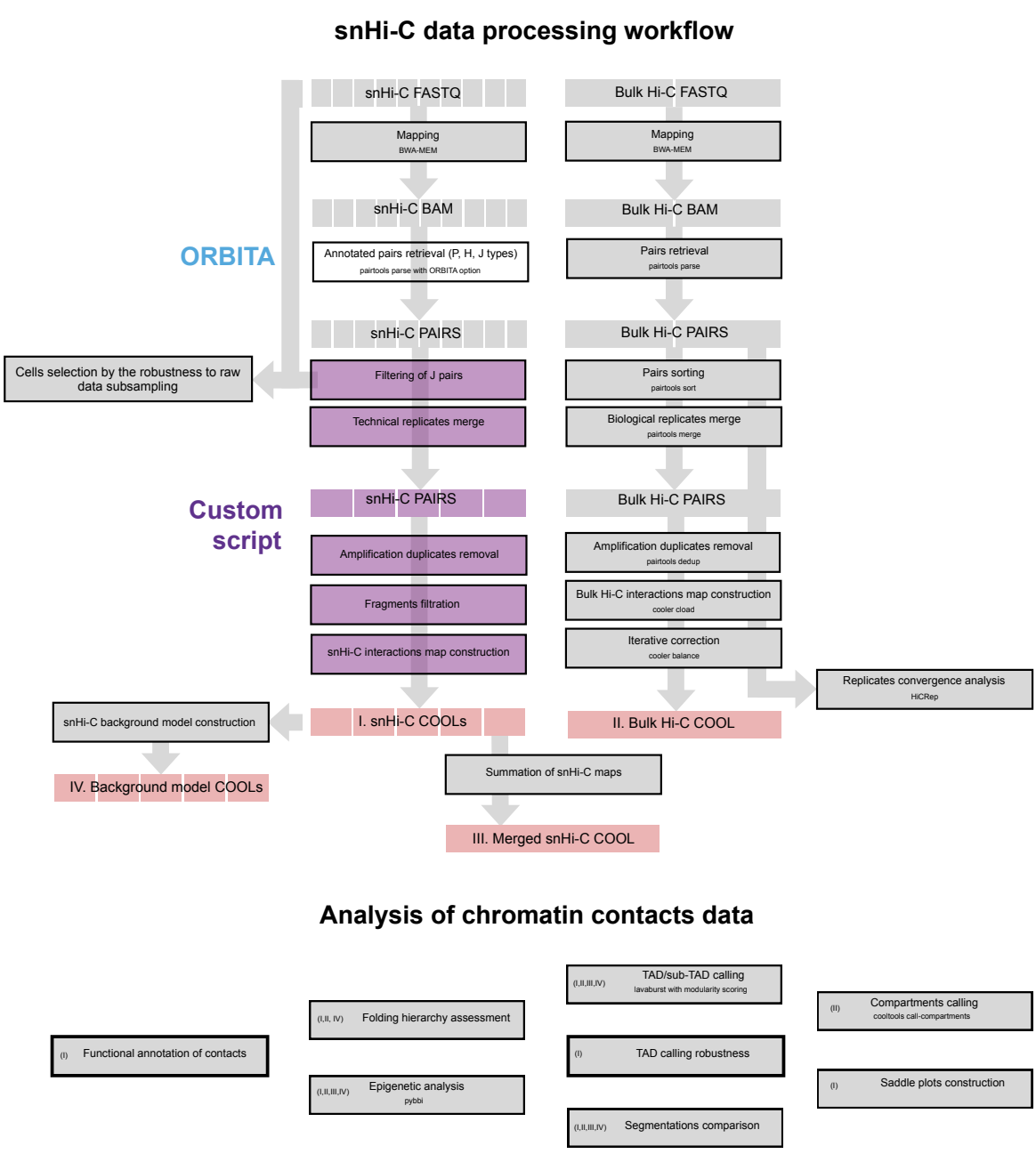
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Figure 1

a

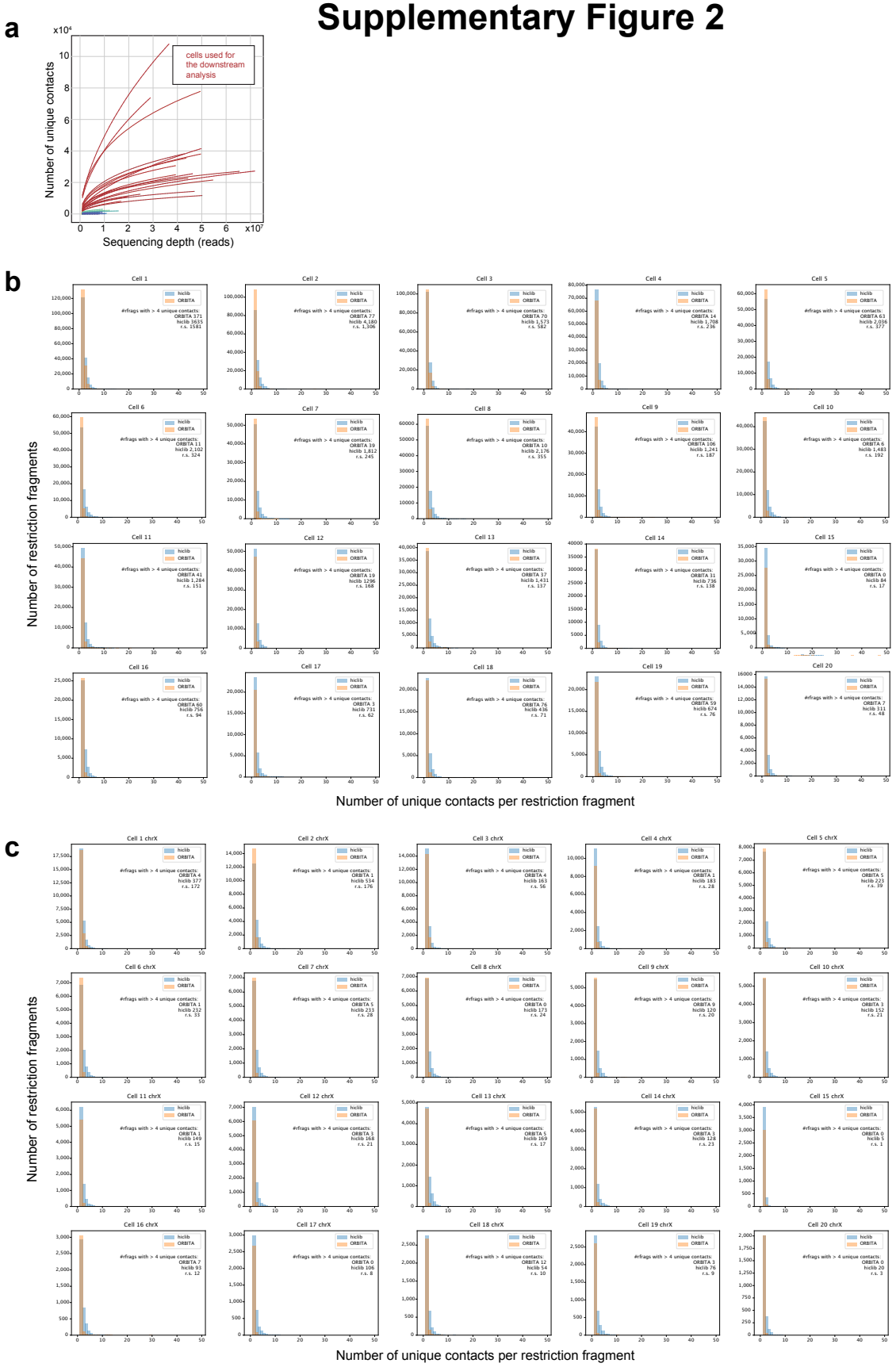


b



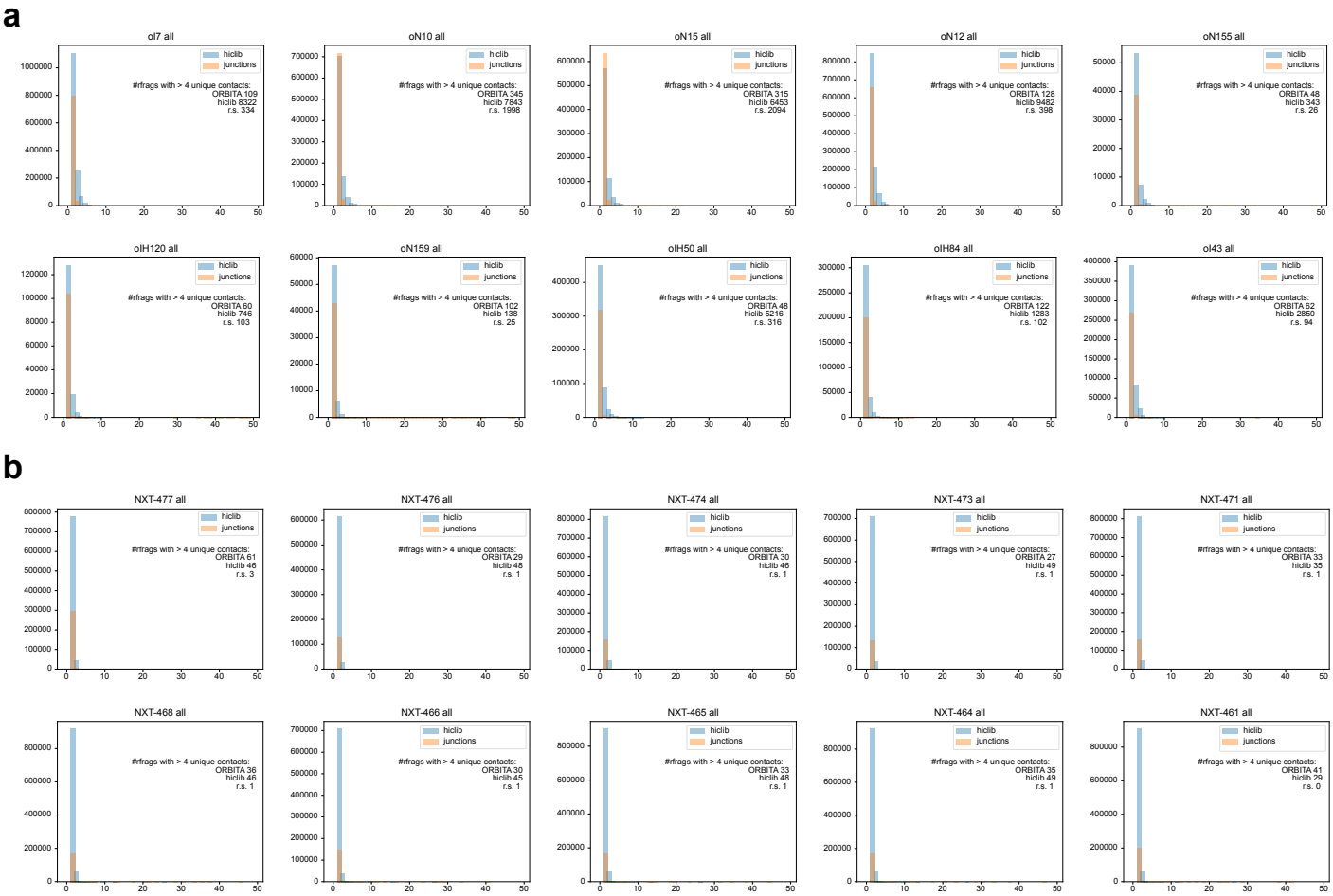
Supplementary Figure 1.
(a) Example of individual nuclei isolation. FACS was performed using DAPI staining. P1 zone contains individual nuclei, and zone P2 contains high-confidence signals. Nuclei were harvested from zone P2. **(b)** Workflow of snHi-C data analysis.

Supplementary Figure 2



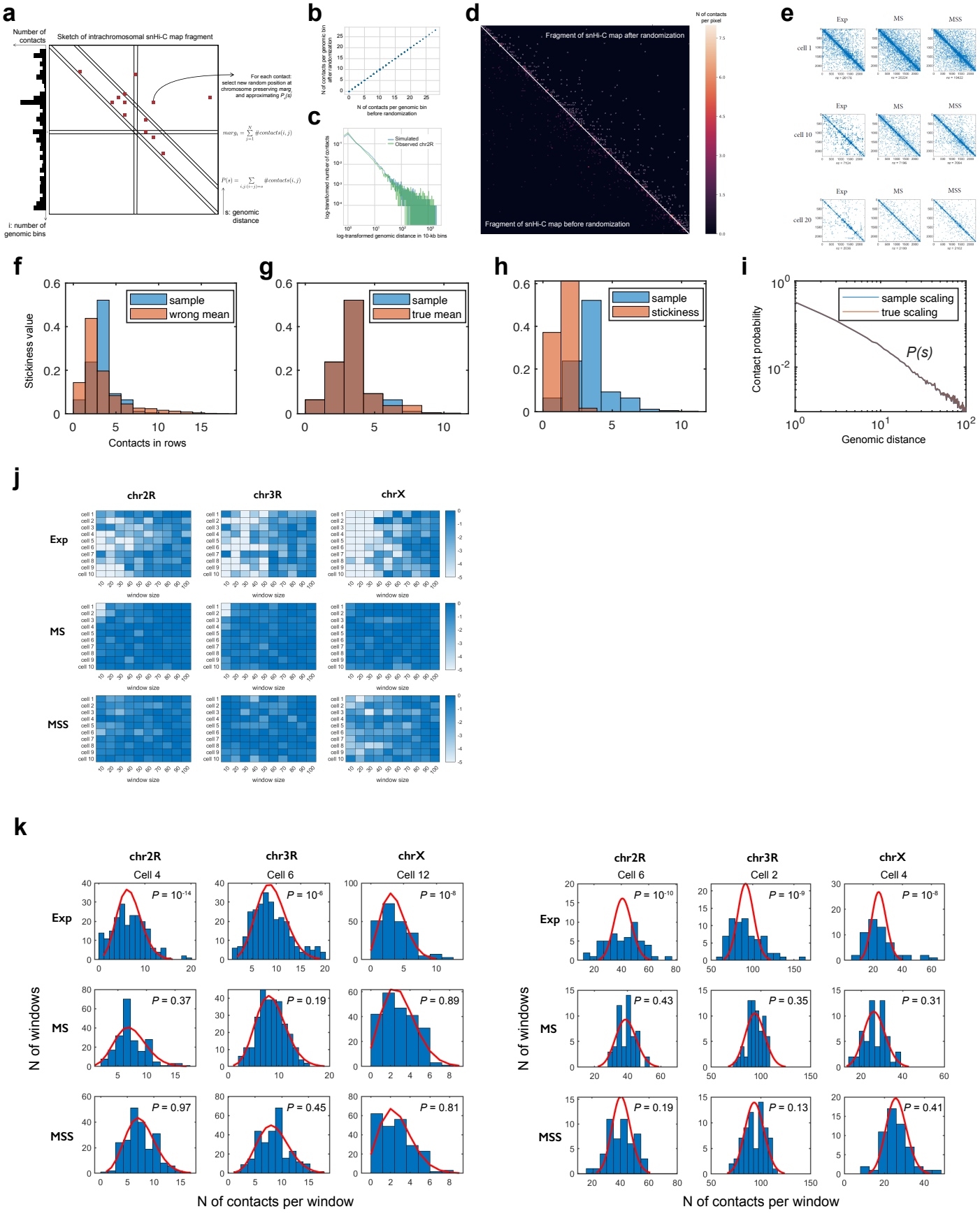
Supplementary Figure 2. Quality control of *Drosophila* snHi-C datasets.
(a) Dependence of number of unique contacts on library size in downsampling analysis (see Methods). The best 20 single cells (red) were selected and then used for additional sequencing. **(b)** Number of unique contacts per restriction fragment (RF) captured by ORBITA (orange) and *hiclib* (blue) for 20 best BG3 snHi-C datasets. Number of restriction fragments with more than four contacts is shown in every plot. Note that the number of contacts called by *hiclib* is larger than the number of contacts called by ORBITA. BG3 is a diploid male cell line; accordingly, in a single nucleus, each RF from autosomes and the X chromosome could establish no more than four and two unique contacts, respectively (see Online Methods for details); r.s., results of downsampling control, averaged over 10 repeat. **(c)** Number of unique contacts per restriction fragment (RF) captured by ORBITA (orange) and *hiclib* (blue) for chromosome X. BG3 cells have only one X chromosome; thus, only 2 unique contacts are possible for a single restriction fragment, corresponding to 2 ends of linear DNA fragment after restriction. r.s., as above.

Supplementary Figure 3



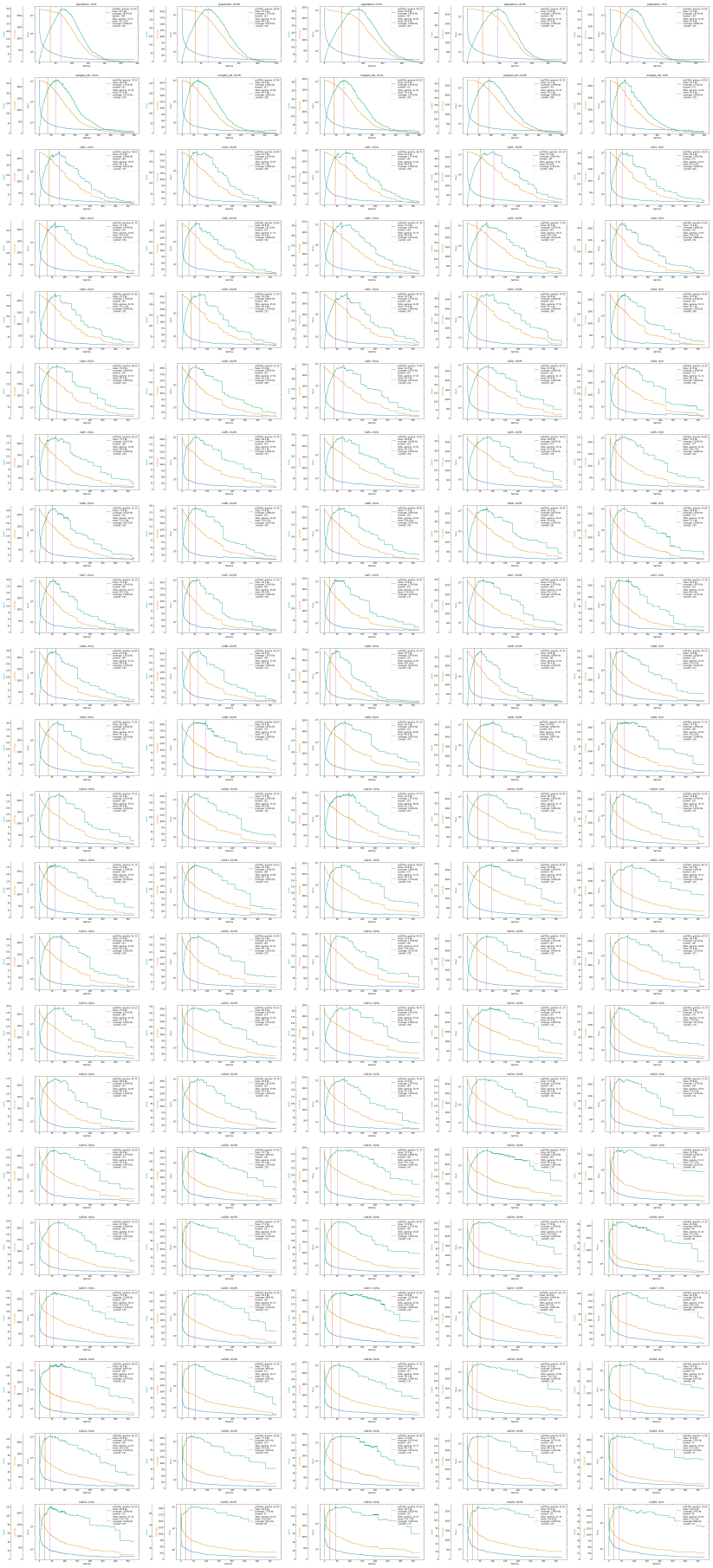
Supplementary Figure 3. Quality control for published mouse snHi-C using ORBITA.
(a) Number of unique contacts per restriction fragment (RF) captured by ORBITA (orange) and *hiclib* (blue) for ten cells from Flyamer et al. (2017)[32]. The cells are named according to the Supplementary Data from Flyamer et al. (2017)[32]. o – oocyte, N – non-surrounded nucleolus, H – Hoechst stain, I – Intermediate, r.s., results of downsampling control, averaged over 10 repeats. (b) Number of unique contacts per RF captured by ORBITA (orange) and *hiclib* (blue) for ten cells from Nagano et al. (2017)[33]. The cells are named according to Supplementary Data from Nagano et al. (2017)[33]; r.s., as above.

Supplementary Figure 4



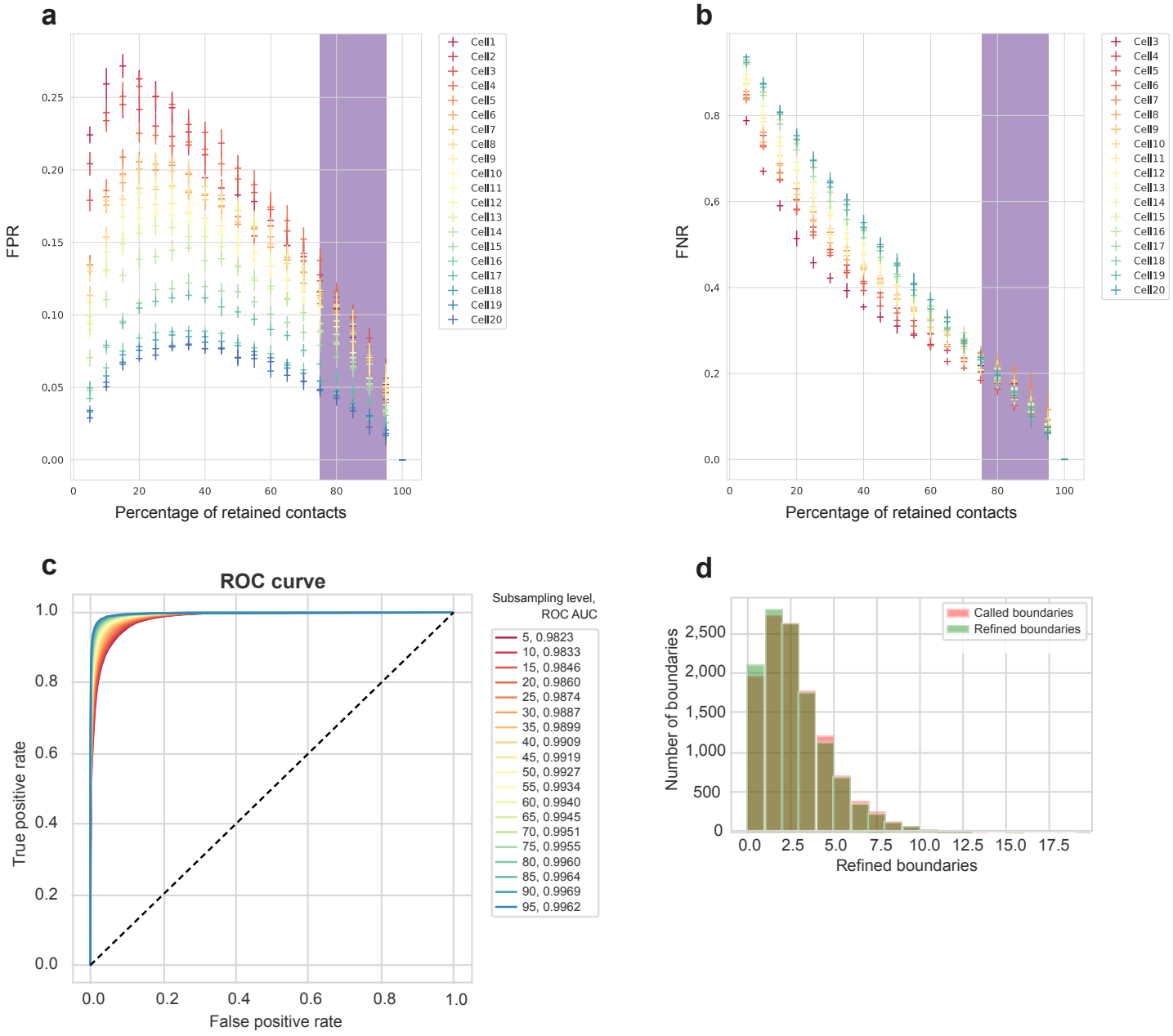
Supplementary Figure 4. snHi-C maps do not follow the rules of random distribution of contacts.
(a) Background model of snHi-C interactions (MSS; Marginal Scaling with Stickiness Model). **(b)** Scatter plot of the initial number of contacts per genomic bin in snHi-C map and after randomization for chr2R of Cell 1 with 107,823 unique contacts. **(c)** Contact probability $P_c(s)$ for chr2R of Cell 1 before (green) and after (blue) randomization. **(d)** Cell 1 snHi-C interactions map for a region of chr2R (lower triangle) and randomized background control (upper triangle) (see Methods). Note the presence of contact clusters at the diagonal both in original and reshuffled data. **(e)** Examples of experimental (Exp) single-cell Hi-C maps with those simulated using the MSS and MS models. **(f-h)** Derivation of the stickiness values (Y axis) given the coverage of bins (numbers of contacts in rows, X axis) obtained by iterative approximations for the MSS model and chr2L (merged snHi-C data were used). n = 2,302 bins. **(f)** Histograms of observed coverage from merged snHi-C map (blue) and of theoretical values (brown) calculated with (red) at the first step of the iterative procedure; wrong mean – computed with wrong stickiness. **(g)** The same histogram as in (f) after a series of iterative corrections of the stickiness values that led to convergence towards the limiting values. The resulting distribution of the coverage (red) reproduces the experimental values; true mean – computed with true stickiness, which is the outcome of the iterative procedure. **(h)** Distributions of the experimental coverage (blue) and of the limiting stickiness (red) are significantly different. Notably, the stickiness values have lower variance than the experimental coverage because the latter incorporate fluctuations of the contact probability. **(i)** Initial and limiting scaling probability functions (see Methods) remain unchanged after the iterative approach. **(j)** Heatmaps of log10 of *p*-values for the test for the top-10 cells sorted to their contact densities. Clustering of contacts at the scale of TADs cannot be explained by the random models at the significance level. **(k)** Experimental, MS, and MSS distributions of the number of contacts in windows of the size bins (100 kb) (left) and bins (400 kb) (right) displaced at the main diagonal and their best Poisson distribution (in red). *P*-values are calculated in Chi-Square Goodness of Fit Test. Left: n = 211, 279 and 224 windows for 2R, 3R and X, respectively. Right: n = 52, 69 and 56 windows for 2R, 3R and X, respectively.

Supplementary Figure 5



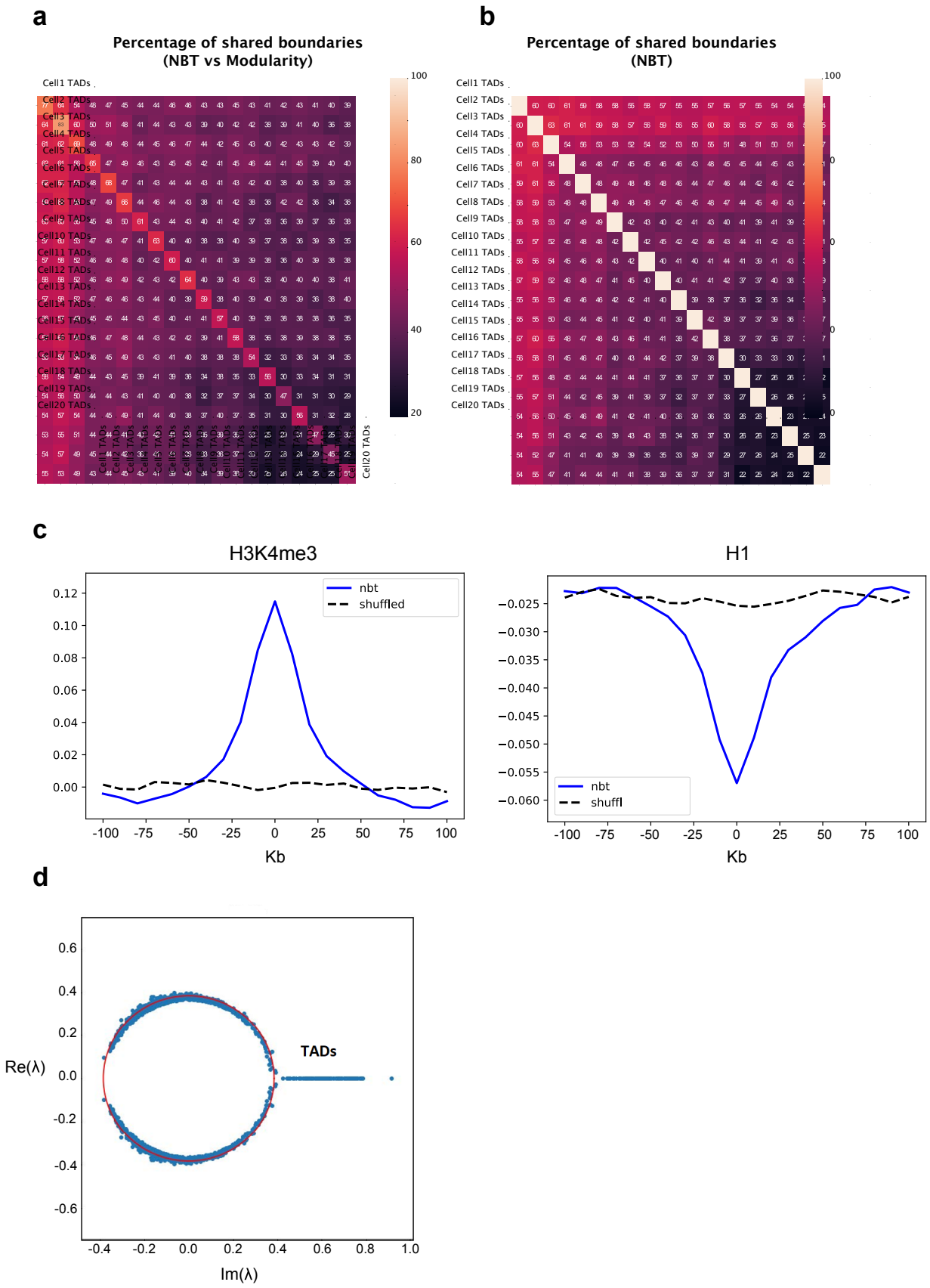
Supplementary Figure 5. Selection of TAD calling parameters.
Dependence of mean TAD size, genome coverage with TADs, and number of called TADs on the γ parameter value for all analyzed chromosomes for bulk in situ Hi-C data, merged dataset, and all single cells. Iterative correction of the maps was used prior to TAD calling for the bulk *in situ* Hi-C dataset. Count – number of TADs identified; Coverage – coverage of the genome with the TADs identified; mean – mean TAD size. In the inset: number – number of TADs and sub-TADs identified.

Supplementary Figure 6



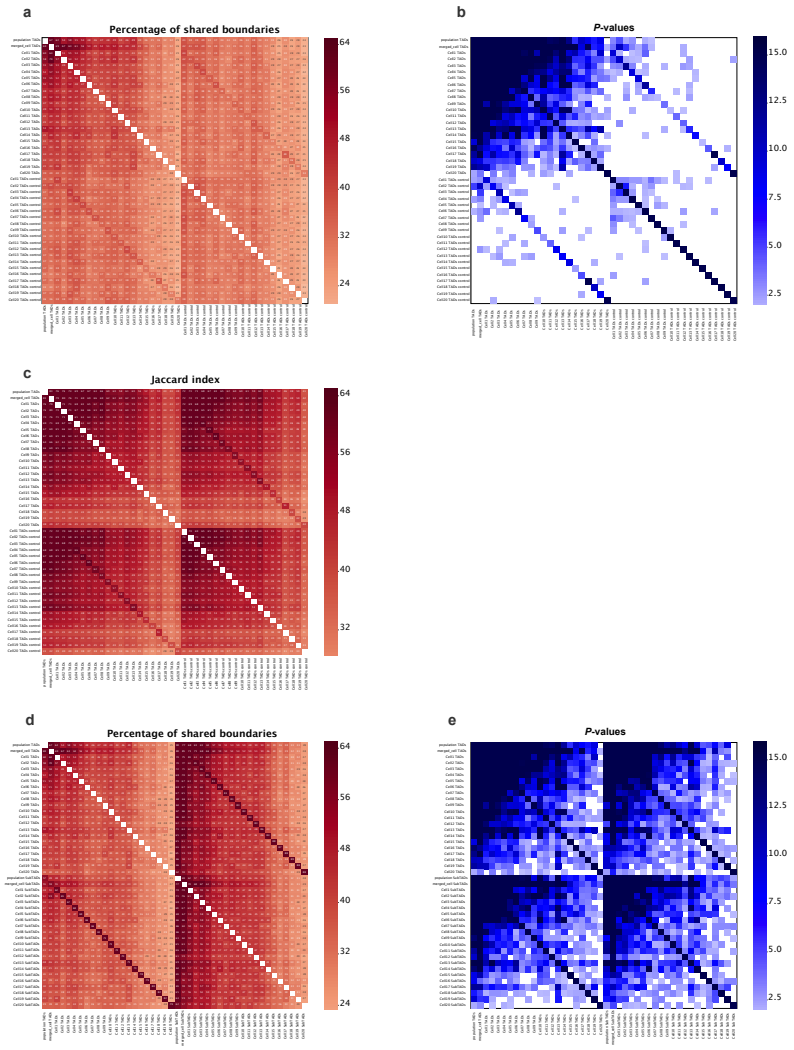
Supplementary Figure 6. TAD calling robustness to sampling procedure.
(a-b) Mean ratio of overrepresented boundaries (false positives) out of true non-boundary genomic bins (false positive rate (FPR)) (a) or underrepresented boundaries (False Negatives) out of true boundaries (false negative rate (FNR)) (b) for TAD calling on subsampled snHi-C maps (see Methods). The snHi-C maps were independently subsampled by 5% levels from the initial contacts of each dataset. Ten subsampling iterations per dataset and subsampling level were performed. Violet rectangle shows the diapason of data downsampling used for testing of TAD boundary robustness to sampling procedure.
(c) True TAD boundaries were predicted based on the collective support by different subsampling levels and iterations (see Methods). ROC curves for different threshold subsampling levels are shown. Subsampling level 95 corresponds to 95% of initial contacts per dataset; level 90 corresponds to collective support from 95% and 90% subsampling iterations etc. Level 90 was selected as the threshold. The following final criteria were selected: collective support is smaller than 45% at 90–95% subsampling levels. The resulting accuracy is 0.9765.
(d) Distribution of number of cells in which the boundary is present. Out of 9,942 initially called boundaries across all the cells and chromosomes, 9,788 are confirmed by subsampling analysis.

Supplementary Figure 7



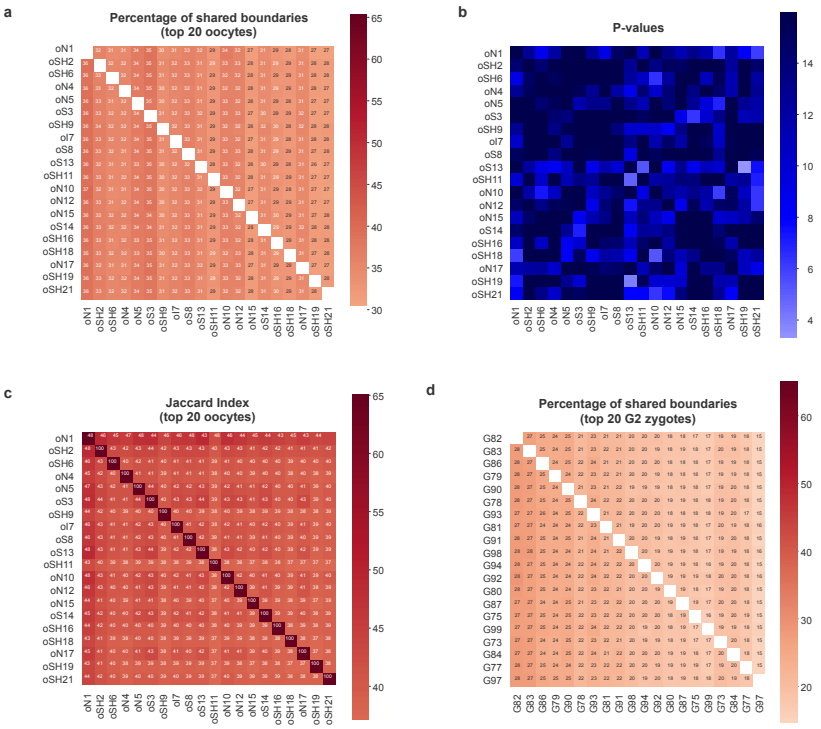
Supplementary Figure 7. NBT as an alternative approach for identification of TAD boundaries.
(a) Percentage of TAD boundaries shared between NBT- and modularity-derived TAD segmentations in individual cells. The mean percentage of shared boundaries is 61%. **(b)** Percentage of TAD boundaries shared between single cells for the NBT TAD calling procedure. The mean percentage of shared boundaries is 42%. **(c)** Epigenetic profiles around the NBT-identified TAD boundaries. **(d)** Spectrum of the non-backtracking operator for cell 3, chr3L. The corresponding constraining disk of the radius r_c for the stochastic block model is shown by red.

Supplementary Figure 8



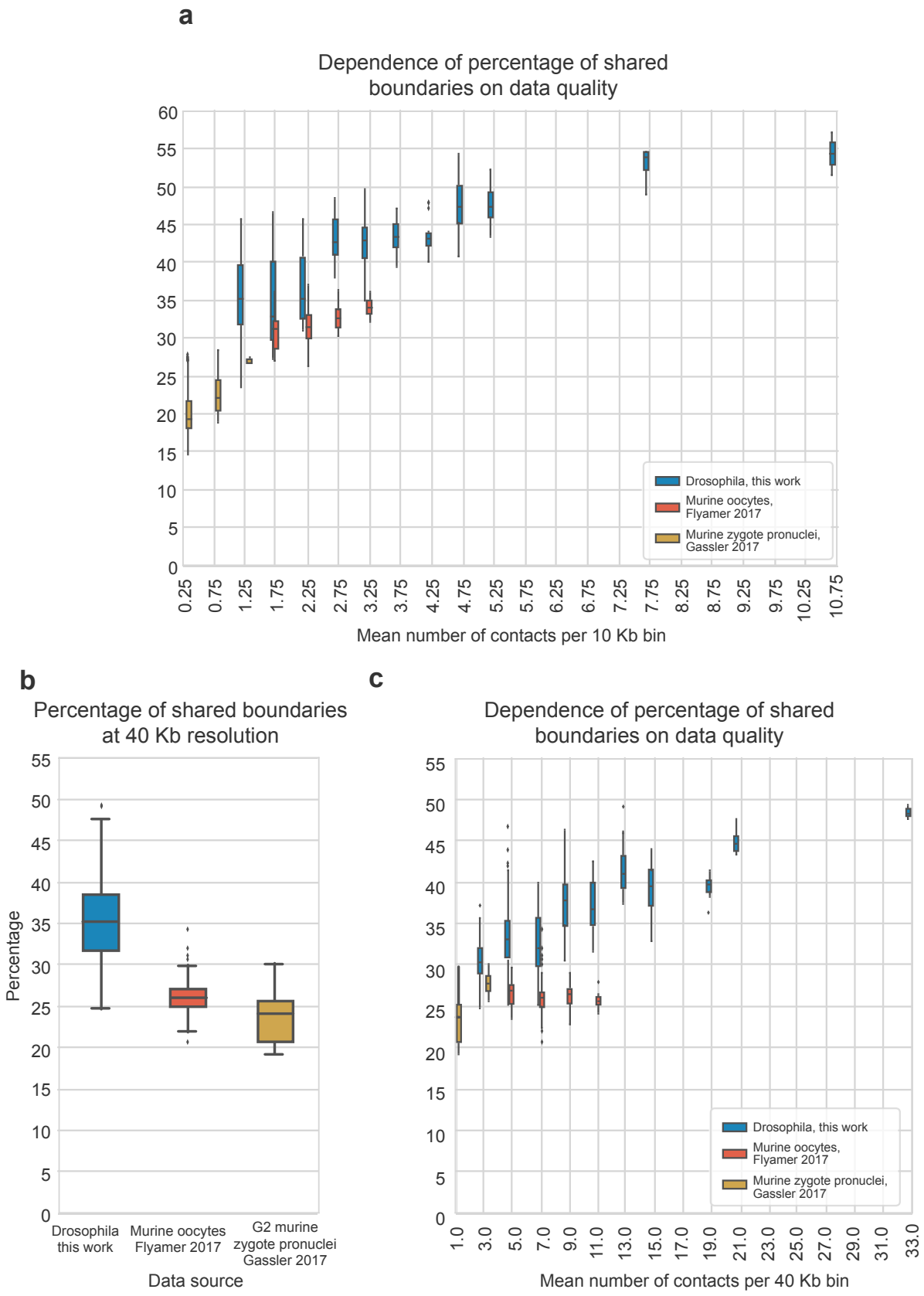
Supplementary Figure 8. TAD boundaries are shared between individual *Drosophila* cells.
(a) Percentage of TAD boundaries shared between *Drosophila* BG3 single cells, bulk BG3 *in situ* Hi-C, and merged snHi-C data, pairwise comparisons. The mean percentage of shared boundaries between individual cells is 39.45%. On average, 46.6% of population-identified TAD boundaries are present in the single cells. In control maps with shuffled contacts (lower part of the plot) preserving marginal distributions and scaling, only 34.95% of boundaries are shared with population boundaries, and 32.47% of boundaries are shared between pairs of shuffled maps (see Methods). **(b)** P-values of permutation tests for the TAD boundaries from (a). Permutation tests were performed 1,000 times (see Methods). -log10 values are shown. **(c)** Jaccard index of shared TAD regions between *Drosophila* BG3 single cells, bulk BG3 *in situ* Hi-C, and merged snHi-C data, pairwise comparisons (see Methods). **(d)** Percentage of TAD boundaries shared between *Drosophila* BG3 single cells, bulk BG3 *in situ* Hi-C, and merged snHi-C data, TAD and sub-TAD, pairwise comparisons. **(e)** *P*-values of permutation tests for the TAD and sub-TAD boundaries from (d). Log10 values are shown.

Supplementary Figure 9



Supplementary Figure 9. TAD boundaries are less stable between mouse cells.
(a) Percentage of TAD boundaries shared between murine oocytes single cells[32] (top-20 cells based on the number of contacts), pairwise comparisons. The mean percentage of shared boundaries is 31.2%. **(b)** *P*-values of permutation tests for the TAD boundaries from (a). Permutation tests were performed 1,000 times. -log10 values are shown. **(c)** Jaccard index of shared TAD regions between mouse single oocytes[34], pairwise comparisons. TADs were called with the procedure similar to the TAD calling in *Drosophila*, as described in Methods, varying gamma value from 0 to 375 during step 1. **(d)** Percentage of TAD boundaries shared between murine G2 zygotes pronuclei (top-20 pronuclei based on the number of contacts), pairwise comparisons. The mean percentage of shared boundaries is 21%.

Supplementary Figure 10



Supplementary Figure 10. Different stability of TAD boundaries between mouse and *Drosophila* is robust to the data resolution selection and data quality.
(a) Percentage of shared boundaries for different levels of data quality, as assessed by mean number of contacts per 10-kb genomic bin per dataset. Boxplots represent the median, interquartile range, maximum and minimum. Total n = 380 cell-to-cell pairwise comparisons for each of three types of boxplots in the analysis. Related to Fig. 3e. **(b)** Percentage of shared boundaries between all pairs of cells of *Drosophila* in this work, top-20 oocytes from Flyamer et al. (2017)[32] and top-20 G2 zygote pronuclei[34] at 40 kb. Boxplots represent the median, interquartile range, maximum and minimum. Total n = 380 cell-to-cell pairwise comparisons for each of three types of boxplots in the analysis. **(c)** Percentage of shared boundaries for different levels of data quality, as assessed by mean number of contacts per 40-kb genomic bin per dataset. Boxplots represent the median, interquartile range, maximum and minimum. Total n = 380 cell-to-cell pairwise comparisons for each of three types of boxplots in the analysis. Related to Supplementary Fig. 10b.

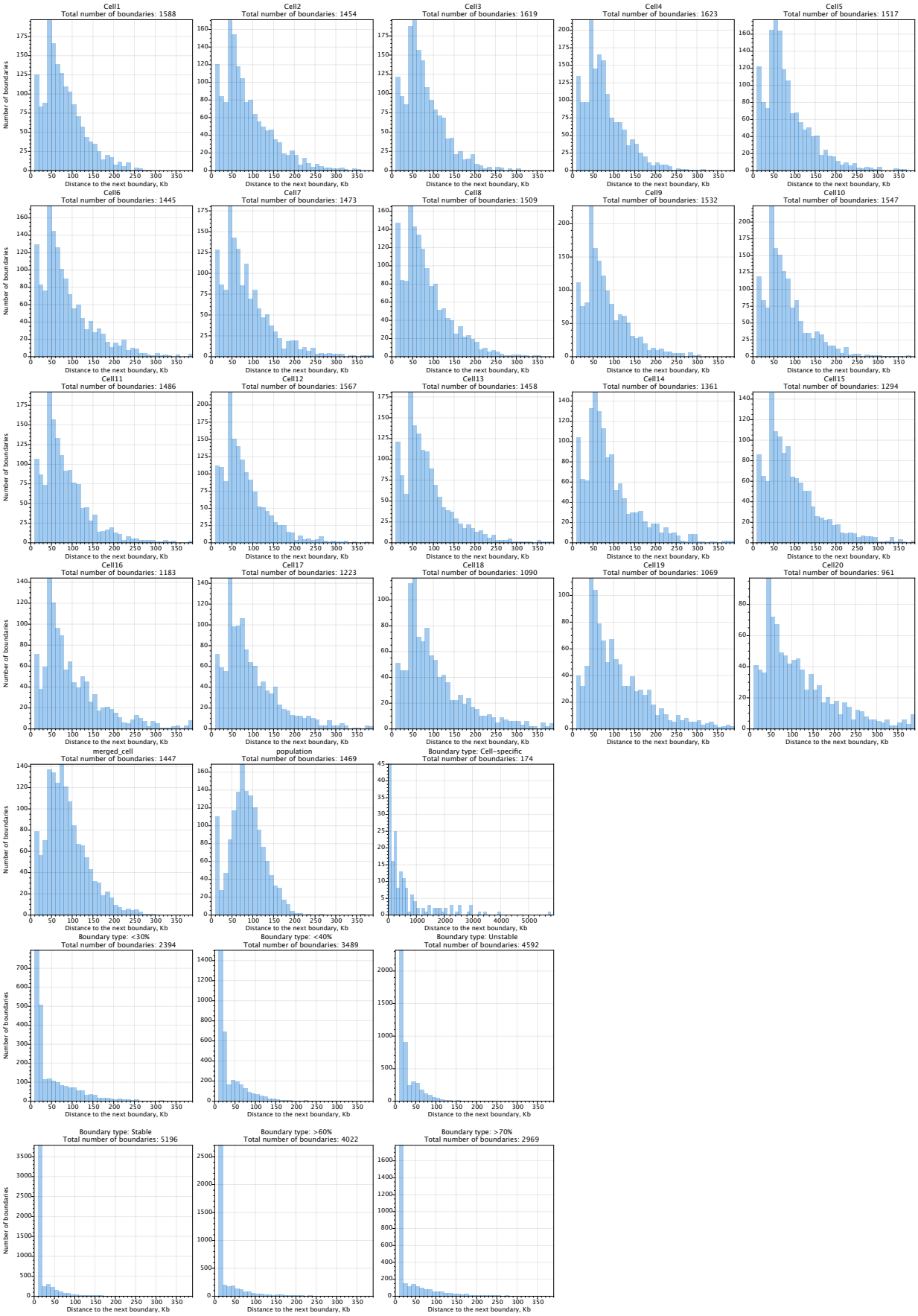
Supplementary Figure 11

Supplementary Figure 11. Epigenetic properties of different types of TAD boundaries.
Heatmaps with z-score (upper panel) of selected chromatin marks centered at single-cell TAD boundaries from different groups (+/- 100 kb). Bulk – conventional BG3 *in situ* Hi-C; merged – aggregated snHi-C data from all individual cells; stable – boundaries found in more than 50% of cells; unstable – boundaries found in less than 50% of cells; cell-specific – boundaries identified in any one individual cell; TAD bins – genomic bins from TAD interior; random – randomly selected genomic bins. TAD boundaries are refined by the subsampling robustness protocol (see Methods).

Supplementary Figure 12

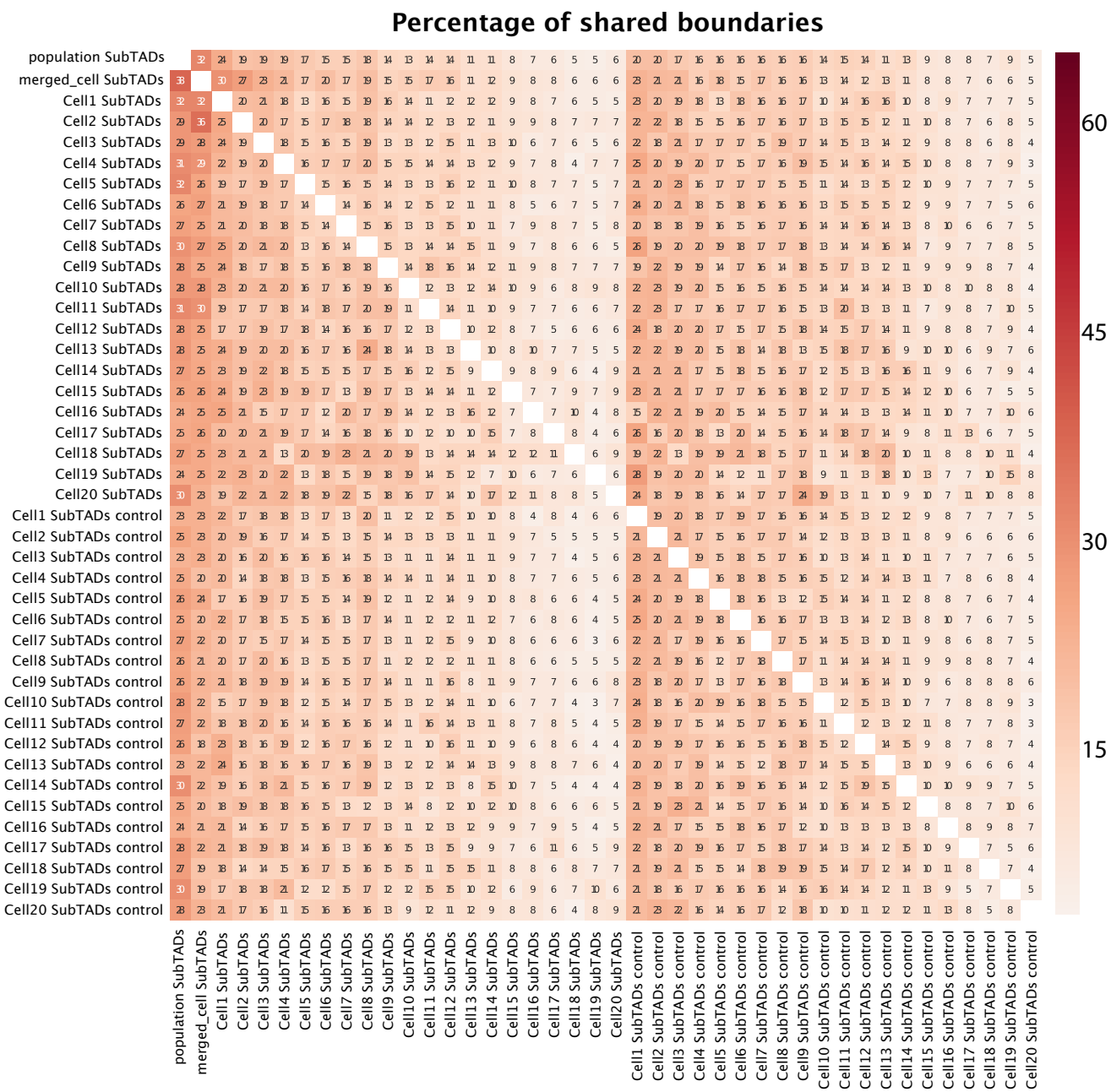
Supplementary Figure 12. Epigenetic properties of different types of TAD boundaries (related to Supplementary Figure 11; stable – boundaries found in more than 60% of cells; unstable – boundaries found in less than 40% of cells).

Supplementary Figure 13



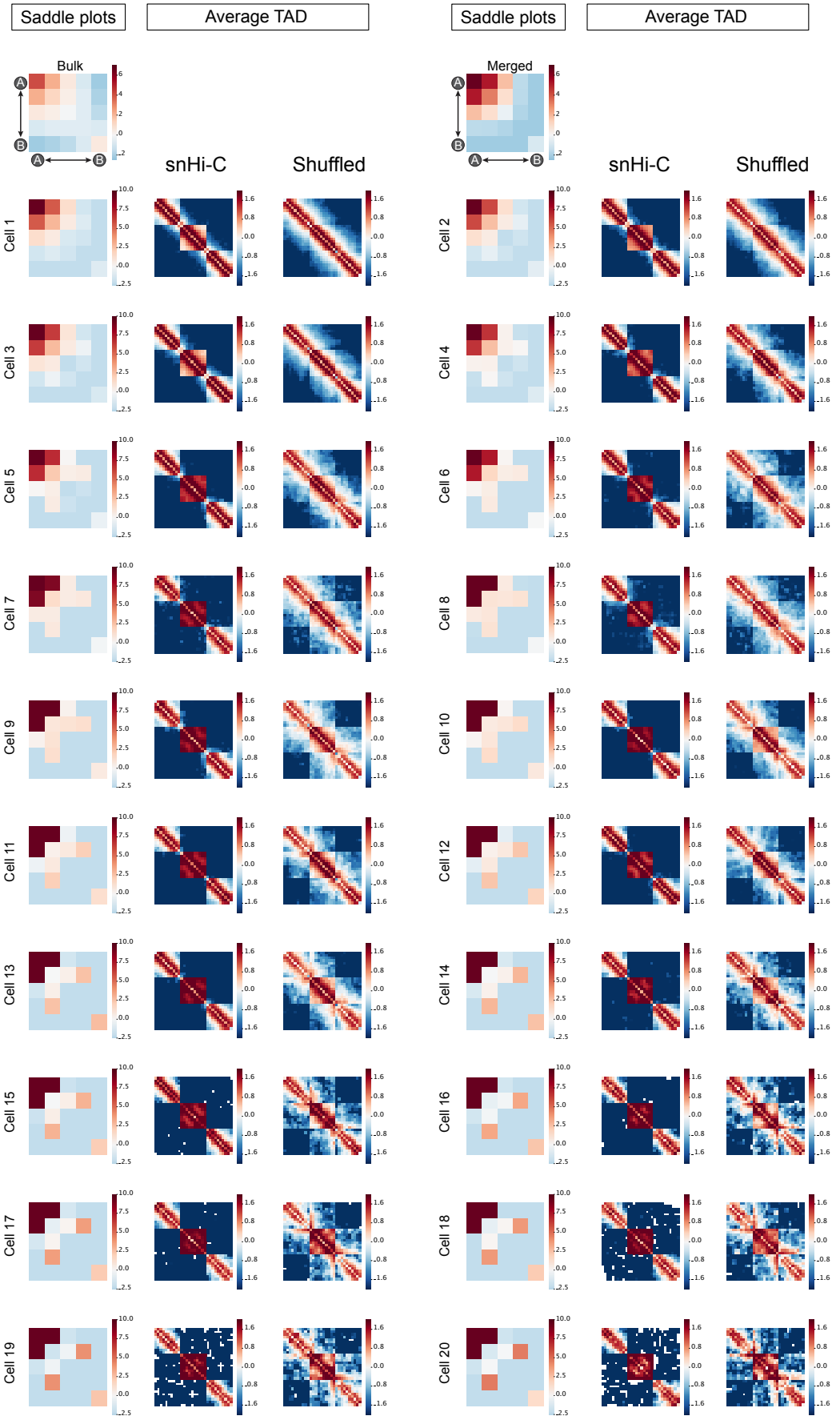
Supplementary Figure 13. Distributions of distances between boundaries of different types and the number of boundaries. Rows 1-4 represent the boundaries from individual cells. Rows 5-7 represent the boundaries from the bulk Hi-C data, merged datasets, alongside cell-specific boundaries, unstable and stable boundaries. In all the cases except cell-specific boundaries, the distributions are demonstrated up to 400 kb.

Supplementary Figure 14



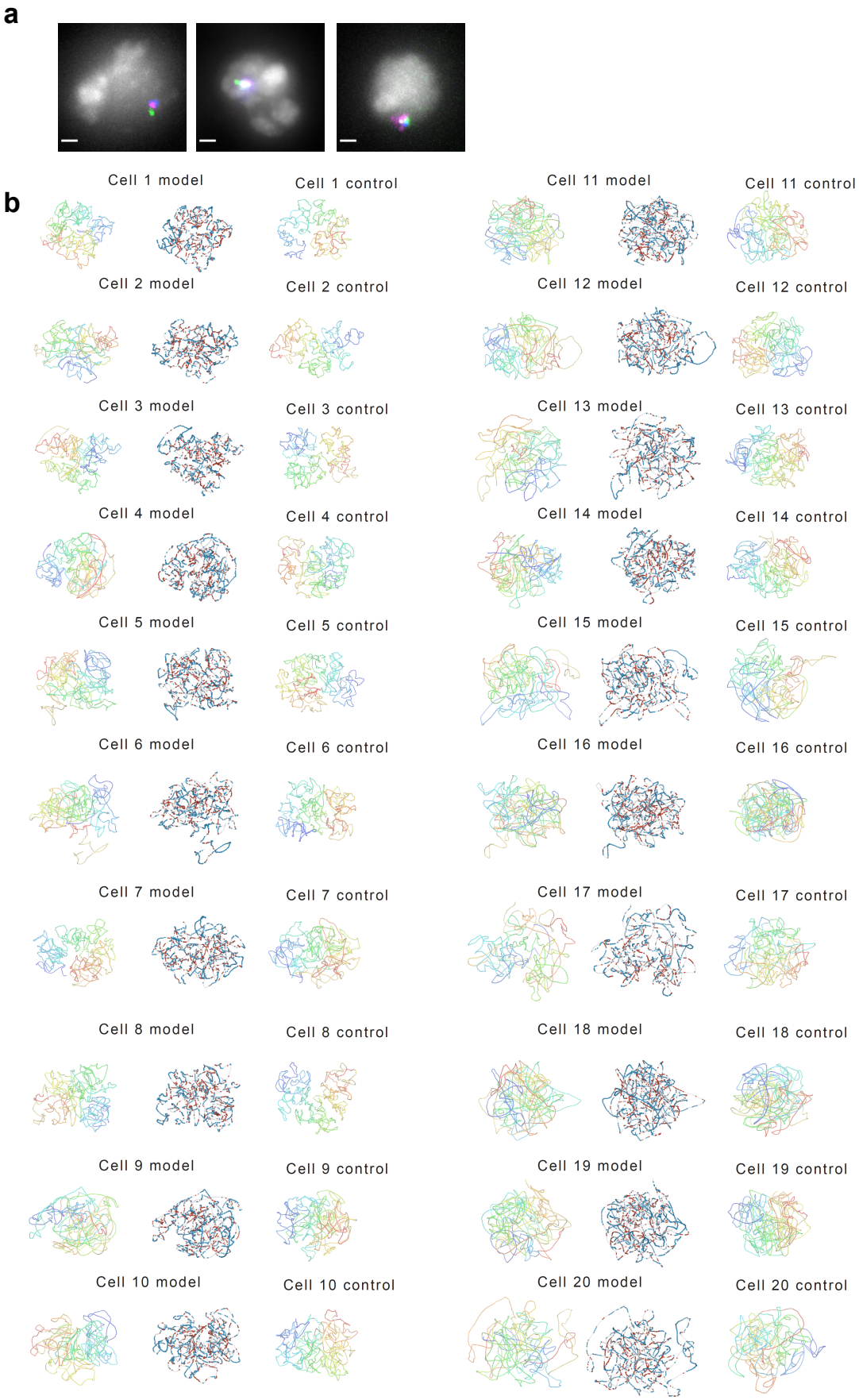
Supplementary Figure 14. Sub-TAD boundary profiles are not conserved between individual cells. Percentage of sub-TAD boundaries excluding TAD boundaries shared between *Drosophila* BG3 single cells, bulk BG3 *in situ* Hi-C, merged snHi-C data, and control shuffled datasets, pairwise comparisons. *P*-values of permutation tests for the sub-TAD boundaries were performed 1,000 times and are not shown. All of them are <0.01.

Supplementary Figure 15



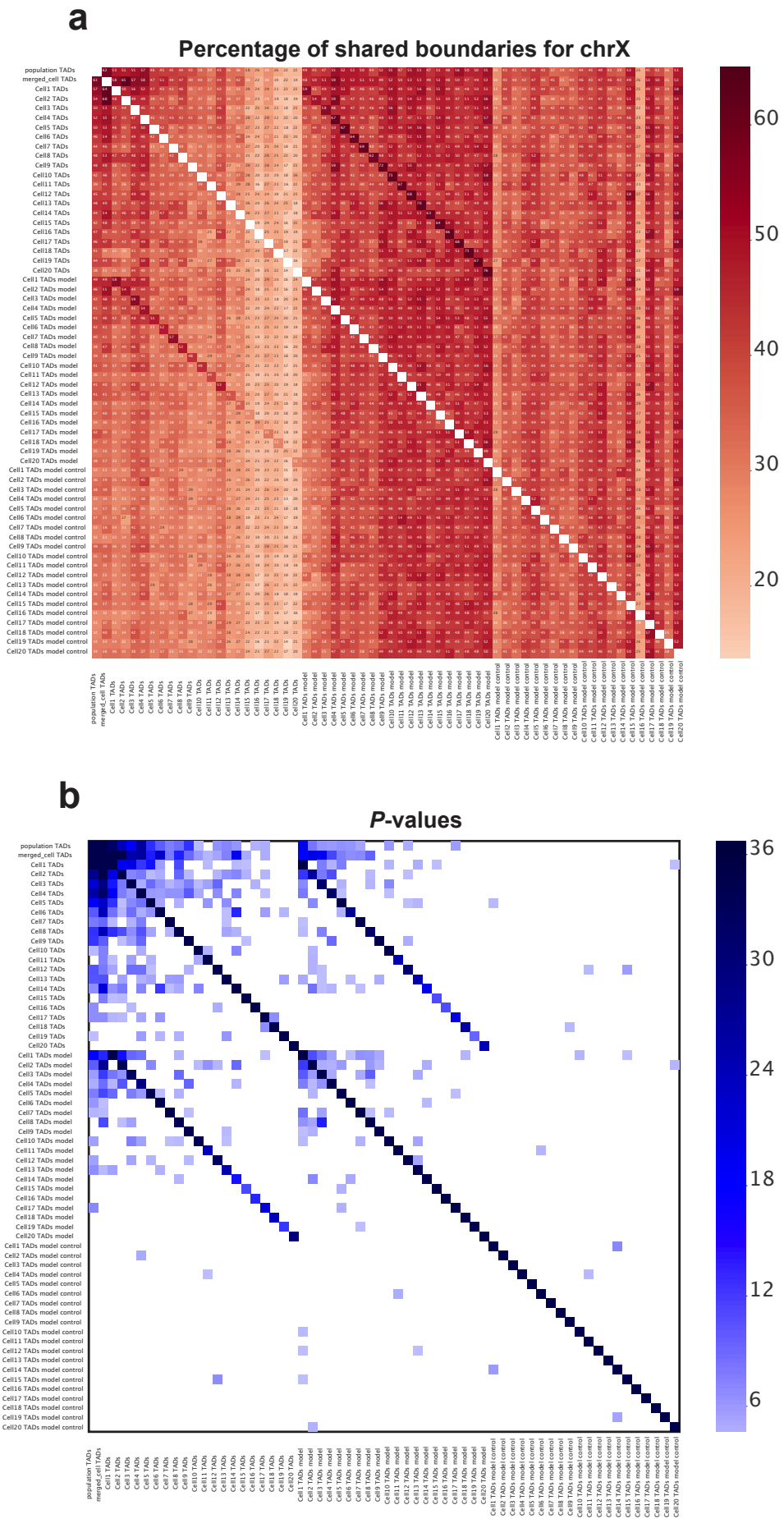
Supplementary Figure 15. Saddle plots and average TAD in individual cells.
Top: saddle plots for bulk Hi-C and merged contacts from all snHi-C datasets are shown on top. Below: saddle plot, average TAD plot, and average TAD plot for shuffled controls are displayed for each cell. Shuffled controls were obtained by the procedure described in Supplementary Fig. 4a-d. For saddle plots of bulk and merged data, log2 of observed over expected of iteratively corrected maps was used. For individual cells and saddle plot, log2 of observed over expected of pulled raw maps is shown. For individual cells and average TAD plots, log2 of pulled raw maps normalized by the mean number of contacts in a sliced window around TAD is provided.

Supplementary Figure 16



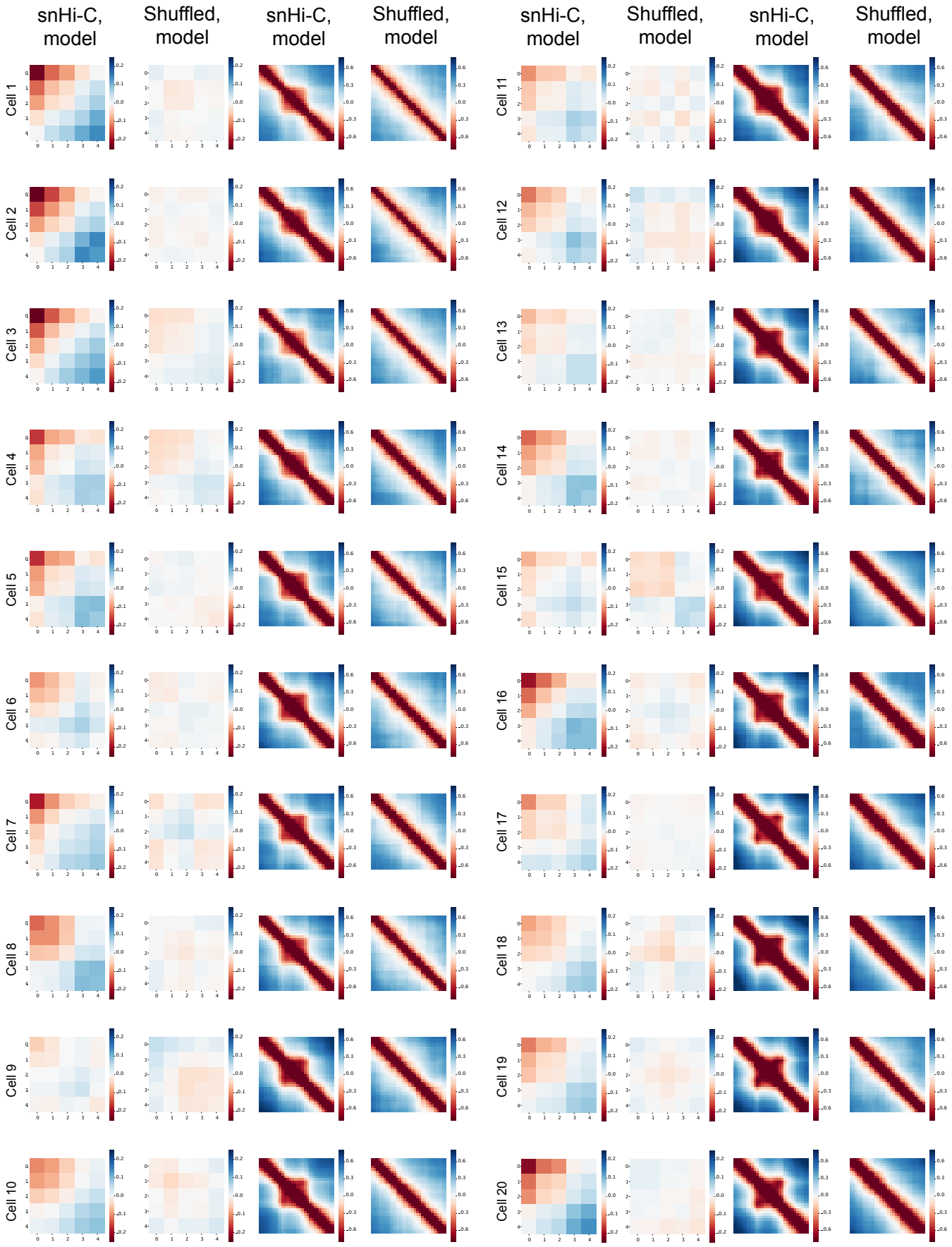
Supplementary Figure 16. Modeling of 3D folding of individual chromosomes.
(a) Examples of FISH images (related to Supplementary Fig. 19c) demonstrating that ChrX is actually haploid in the used batch of the BG3 cell line. Scale bar = 1 μm. **(b)** 3D models of individual X chromosomes obtained from real snHi-C contacts and from shuffled control snHi-C maps. The models are shown as a stick model; positions of each genomic bin were averaged in the sliding window of size 15 centered at each bin. The models obtained from real snHi-C contacts are colored by a rainbow approach and by chromatin colors (as described in Methods, red – active, dark grey – inactive). Source data are provided as a Source Data file.

Supplementary Figure 17



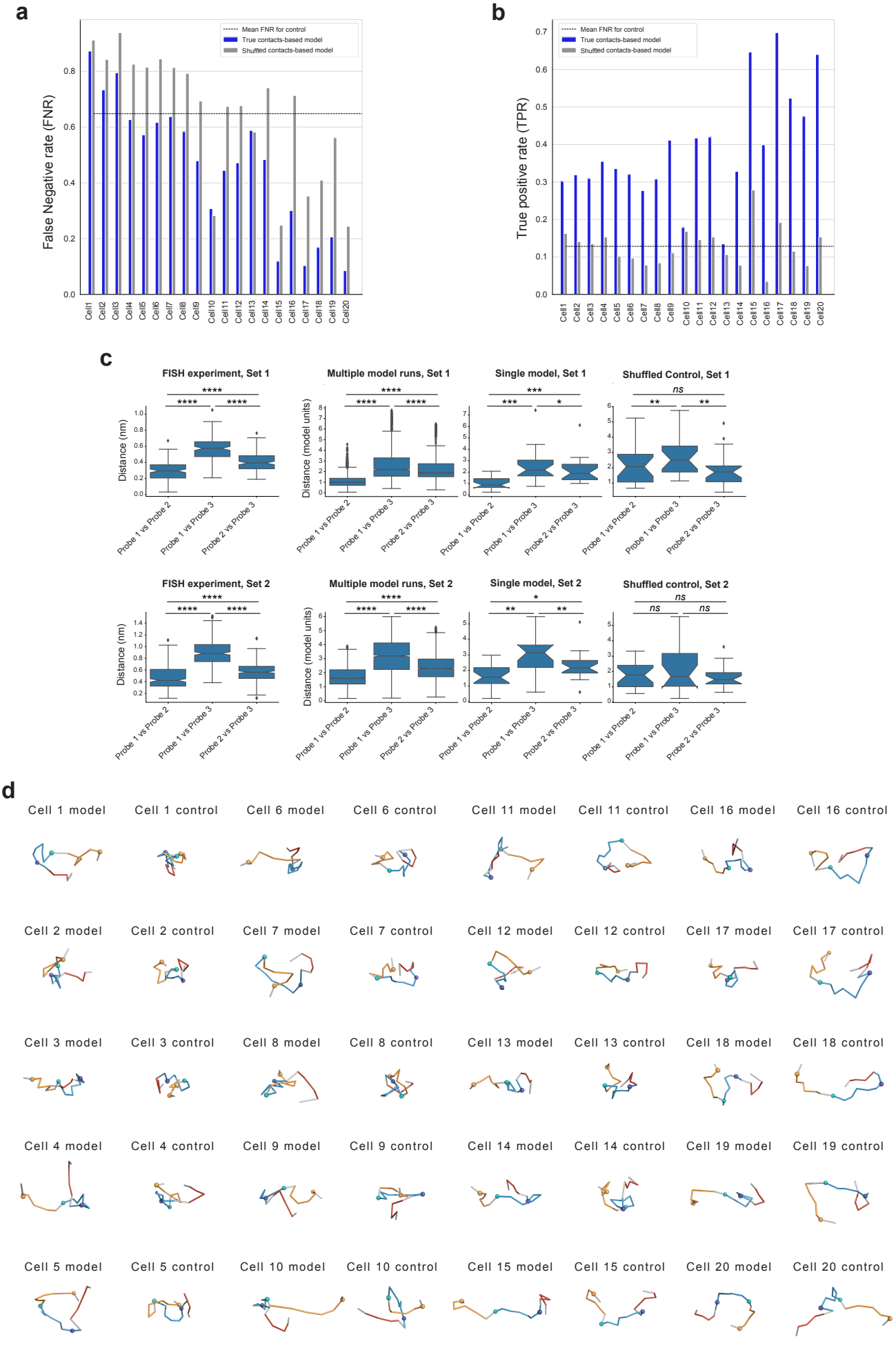
Supplementary Figure 17. TAD boundaries are reproduced in single-chromosome models.
(a) Percentage of TAD boundaries shared between individual cells, 3D models, and 3D models derived from shuffled snHi-C data (control). **(b)** *P*-values of permutation tests for TAD boundaries for (a).

Supplementary Figure 18



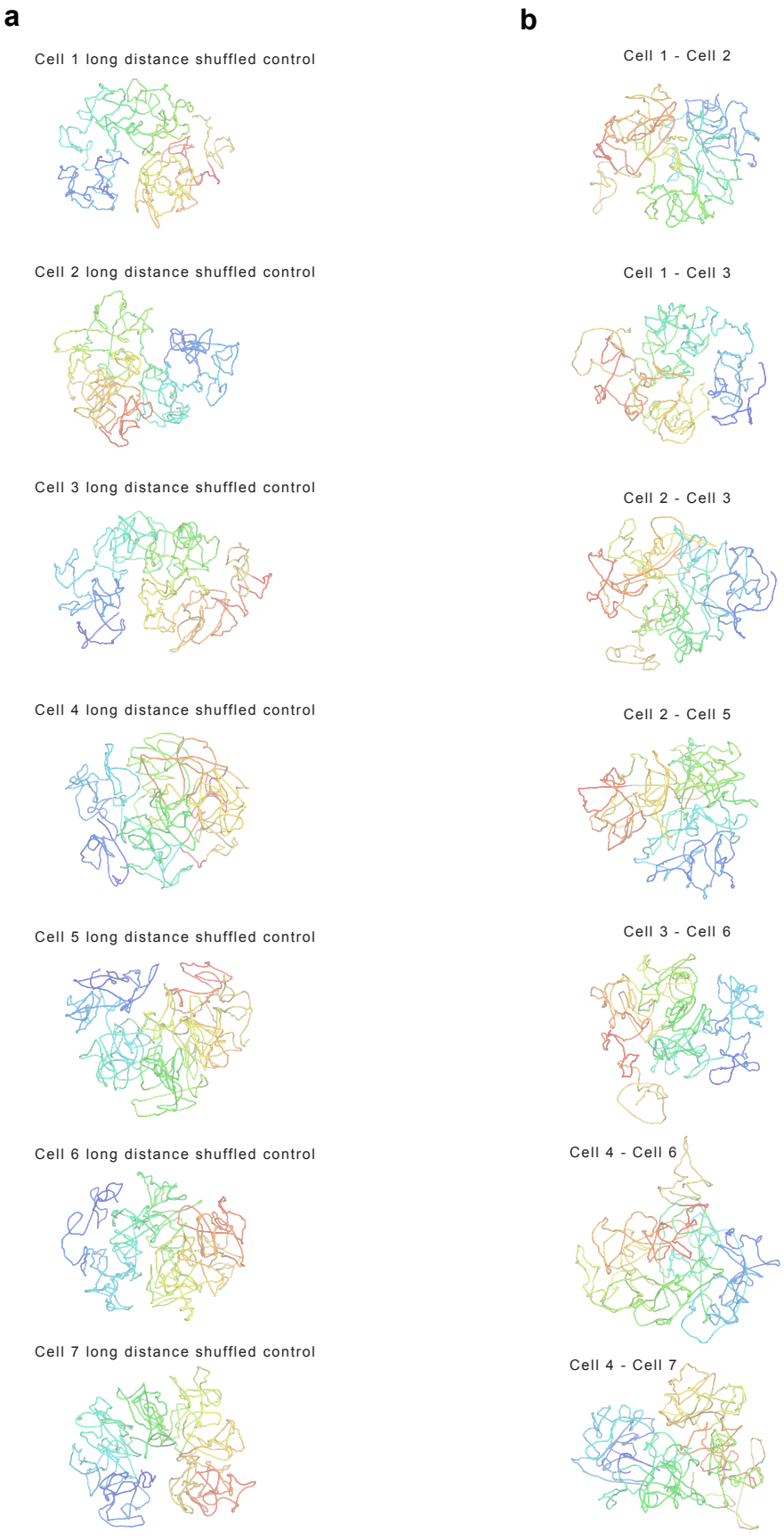
Supplementary Figure 18. Saddle plots and average TADs in individual models.
Four images in a row represent a single model of a chromosome X. For each model, there are saddle plots for real snHi-C modeling, saddle plots for shuffled controls, average TADs for snHi-C, and average TADs for control models. log2 values of observed over expected for average distance matrices were used; thus, smaller values represent a closer distance of corresponding bins of genome in 3D space.

Supplementary Figure 19



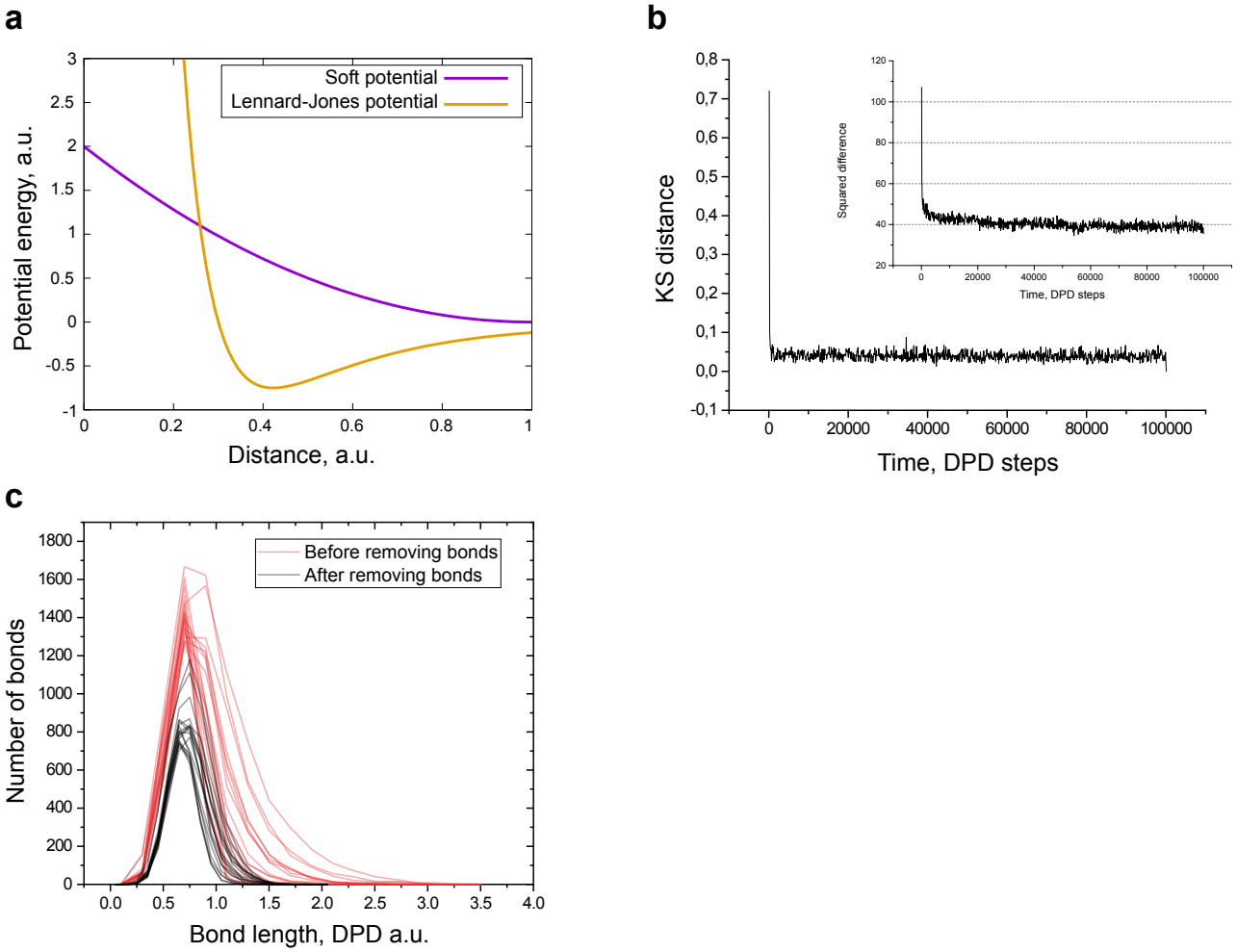
Supplementary Figure 19. Models recapitulate biologically meaningful contacts in single cells.
(a) False negative rate (FNR) of contacts obtained from models (real contacts and control are compared). Here, the set of true contacts for each cell is defined as the list of genomic bin pairs (10-kb resolution) with at least 1 interaction in the snHi-C map at a distance > 20 kb. The set of predicted contacts in the model is defined as the list of pairs of polymer monomers closer than 0.7 units based on average distance matrices. False negatives for the cell are defined as the number of pairs of bins that interact based on snHi-C but do not interact in the models by the criteria above. FNR is smaller for all the models based on real contacts except for Cells 10 and 13. Dotted line – mean FNR for the control; blue – true contacts-based model; grey – shuffled contacts-based model. **(b)** True positive rate (TPR) of contacts obtained from models. TPR is higher for all the models based on real contacts and is even higher than the mean TPR for the control shuffled contacts-based models. True positives for the cell are defined as the number of pairs of bins that interact based on snHi-C and also interact in the models by the criteria above. **(c)** Comparison of spatial distances between FISH probes in the DPD-simulated models (three right plots) and in situ FISH (left). Multiple model runs are 100 independent runs of DPD simulations. Single model and control are distance measurements for the models represented in Supplementary Fig. 16. Two independent sets of FISH probes were selected. ****$p < 0.0001$, ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, ns – non-significant difference in the two-sided Wilcoxon test. Set of probes 1: n = 132 independent measurements; Set of probes 2: n = 122 independent measurements.
Set 1:
Probe-1 = chrX:3,871,158..3,892,065 bp
Probe-2 = chrX:3,960,041..3,983,074 bp
Probe-3 = chrX:4,054,120..4,075,361;
Set 2:
Probe-1 = chrX:17,644,479..17,663,154 bp
Probe-2 = chrX:17,704,670..17,725,745 bp
Probe-3 = chrX:17,764,735..17,783,789.
Source data are provided as a Source Data file.
(d) Visualization of FISH probe regions (set 2) in models and conformation comparison with models obtained from real contacts and the shuffled control. Yellow, light-blue, and red stick fragments represent three TADs; three spheres represent the corresponding FISH probes.

Supplementary Figure 20



Supplementary Figure 20. 3D modeling controls.
(a) Models obtained for snHi-C interactions with shuffled contacts at a distance > 200 kb (long-distance shuffled control).
(b) Models obtained for snHi-C pairs of cells that were artificially merged and subsampled to the number of contacts of one of the cells. The number of contacts of the first cell in the pair was selected as a reference.

Supplementary Figure 21



Supplementary Figure 21. DPD polymer simulations.
(a) Qualitative comparison of soft potential from DPD and Lennard-Jones potential (see Online Methods) from classical MD demonstrates a significant difference in behavior near zero. **(b)** Main graph is the dependency of Kolmogorov-Smirnov distance between distance matrices of final conformation and conformation at other timepoints. In the inset, there is the Euclidean norm of difference between two distance matrices: the final one and at other timepoints. Such tests revealed the fast relaxation of spatial conformation for the proposed reconstruction method. **(c)** Distribution of bonds' length before and after removing overstretched bonds. Despite the threshold for additional bonds being l<1.5, the length of the backbone bonds could be higher than 1.5. After removing additional bonds, heavy tails of the distributions "before" are not present.

Supplementary Table 1

Sequencing and mapping statistics for snHi-C datasets

Cell	Raw reads	Mapped reads	P (typical Hi-C) pairs	H (hops) pairs	J (junctions supported by restriction site) pairs	Removed duplicates	Number of unique contacts	Contacts from restriction fragments with > 4 contacts	Contacts from restriction fragments with > 4 contacts on chrX	Maximum number of contacts per restriction fragments	Maximum number of contacts per restriction fragments on chrX	Number of restriction fragments with > 4 contacts	Number of chr X restriction fragments with > 4 contacts	Final unique contacts	Final unique contacts on chrX
Cell1	44148251	26639477	23543936	507058	2588483	5322544	349051	1905	18	11	5	371	4	107823	11898
Cell2	68632913	39168691	33757160	1059441	4352090	7703393	160751	421	4	25	5	77	1	77770	8174
Cell3	34187865	19797650	17446338	374445	1976867	3071733	295915	368	11	8	6	70	4	73691	8408
Cell4	58947730	22197761	19741636	508544	1947581	6938793	193845	72	6	11	6	14	1	41439	4867
Cell5	50651907	31979756	26885818	892647	4201291	4047611	70297	314	7	13	11	63	5	38174	3917
Cell6	49705351	16513692	13628560	661470	2223662	3629266	94498	58	5	6	5	11	1	35420	3628
Cell7	43834863	13472471	10757036	489079	2226356	3858726	86946	194	13	10	8	39	5	30620	3348
Cell8	55049855	11567500	9733256	438663	1395581	2461501	145606	56	0	8	4	10	0	38019	3275
Cell9	78541387	65267034	55840373	2020682	7405979	7281553	50349	666	22	24	13	106	9	27059	2732
Cell10	41639146	8557235	7157042	376074	1024119	2137327	67414	41	15	16	5	6	3	24882	2585
Cell11	59065312	34086673	30192109	740740	3153824	10265078	129935	229	1	20	5	41	1	25483	2558
Cell12	97406463	41318464	33589103	5018090	2711271	6957870	132613	155	18	37	10	19	3	27215	2988
Cell13	52867991	38130313	33034704	1502780	3592829	8718835	118149	201	9	10	10	37	5	22714	2344
Cell14	68676167	18782199	12614794	2894410	3272995	4745700	78900	163	2	11	11	31	3	21541	2046
Cell15	15004439	755908	661306	18173	76429	141119	58575	0	0	4	3	0	0	14919	1448
Cell16	58244415	22185479	16514468	872321	4798690	7708534	69613	388	7	29	29	60	7	14325	1168
Cell17	22231710	17614914	15932015	320181	1362718	2931803	47619	16	0	6	3	3	0	11087	1282
Cell18	38703945	22557412	10679605	597650	11280157	15133641	42358	453	10	29	7	76	12	12534	929
Cell19	59894379	34471518	28505516	602827	5363175	10805432	58616	280	1	12	10	59	3	11688	1002
Cell20	24336293	6568388	3720850	61977	2785561	4022939	24282	37	0	8	2	7	0	8032	765

Supplementary Table 2

Number of removed bonds. Green means number of the removed contacts in the shuffled models greater than in experimental data, red is the opposite.

Cell ID	Experimental data	Shuffled data	Difference between shuffled and experimental data	Shuffled on the long genomic distances only
1	1075	64	-1011	1271
2	578	712	134	654
3	640	1285	645	1257
4	171	268	97	236
5	224	438	214	
6	190	486	296	482
7	208	503	295	451
8	230	552	322	471
9	59	128	69	
10	20	61	41	
11	38	214	176	
12	0	0	0	
13	13	248	235	
14	85	175	90	
15	8	166	158	
16	0	0	0	
17	0	2	2	
18	5	82	77	
19	2	25	23	
20	0	0	0	

Supplementary Table 3

Number of removed contacts if contacts sampled from two cells. Green means that number of the removed contacts in the model higher than in case of each cell of the mixture.

Cell IDs	Number of removed bonds
Cell4 - Cell8	353
Cell4 - Cell7	299
Cell1 - Cell2	1146
Cell1 - Cell3	1238
Cell2-Cell7	703
Cell2-Cell3	703
Cell3-Cell6	893

Chapter 8

Single-cell Hi-C data analysis: safety in numbers

Single-cell Hi-C (scHi-C) is a rapidly developing technology that allows for unraveling the variability of the chromatin structure between individual cells. While relatively new, it has already improved our understanding of chromatin transformations during the cell cycle, embryogenesis, and brain development. With scHi-C protocols becoming accessible to more labs, specialized computational tools for analyzing the generated data are being developed. However, while various approaches exist, there is little guidance on what computational methods should be applied to solve specific problems arising in the scHi-C data analysis.

In this review, I provide a comprehensive summary of the existing approaches to analyzing diverse scHi-C data types and characterize limitations arising due to the complex and sparse nature of the data. This is a guide through all aspects of scHi-C data analysis, from the read processing to the aggregation analysis, chromatin feature calling, embeddings, and 3D modeling. It reveals the theoretical limit of the number of possible contacts, sources of experimental artifacts, and computational approaches to mitigate them.

I introduce an important concept of ideal scHi-C experiment an ideal scHi-C with 100% recovery of contacts, and demonstrate that it still cannot generate more than 2.4 interactions per 1 Kb of the haploid mouse genome, which is two orders of magnitude lower than bulk Hi-C. Thus, even with ideal scHi-C the contact matrices will remain sparse, with compartments and TADs represented as individual contacts.

This explains why specialized approaches are needed to analyze scHi-C data.

Despite there are approaches to obtain haplotype-specific contacts from scHi-C data, majority of the scHi-C studies do not distinguish homologous chromosomes nor sister chromatids. Current solutions to this problem involve specialized experimental techniques, such as cell sorting, working with highly heterozygous cells and analyzing sex chromosomes present in a single copy (as done in Chapter 7). Haplotype-resolved contact maps of individual cells are one of future directions of development of scHi-C field.

Finally, in this chapter I highlight future improvements and share our outlook on possible developments of this powerful technique.

Briefings in Bioinformatics, 00(0), 2021, 1–13

https://doi.org/10.1093/bib/bbab316
Review

Single-cell Hi-C data analysis: safety in numbers

Aleksandra A. Galitsyna and Mikhail S. Gelfand

Corresponding author: Mikhail S. Gelfand, Skolkovo Institute of Science and Technology, Skolkovo, Russia. Tel: +7 (495) 280-14-81; E-mail: M.Gelfand@skoltech.ru

Abstract

Over the past decade, genome-wide assays for chromatin interactions in single cells have enabled the study of individual nuclei at unprecedented resolution and throughput. Current chromosome conformation capture techniques survey contacts for up to tens of thousands of individual cells, improving our understanding of genome function in 3D. However, these methods recover a small fraction of all contacts in single cells, requiring specialised processing of sparse interactome data. In this review, we highlight recent advances in methods for the interpretation of single-cell genomic contacts. After discussing the strengths and limitations of these methods, we outline frontiers for future development in this rapidly moving field.

Key words: single cell; chromatin; single-cell Hi-C; conformation capture; single-cell sequencing

Introduction

Detecting specific DNA positioning in single cells was first proposed over half a century ago [44, 72]. Deriving statistically reliable general patterns of chromatin folding in single cells, however, has been challenging [5]. Improvements towards this goal have included: increasing the number of analysed cells, studying more loci (up to the complete genome), reducing the size of the interacting regions and improving discriminative power for detection of contacts at a broader scale of spatial distances. There are two main approaches: *microscopy based* and *capture based*. These two types of methods, despite their limitations, provide complementary views on the chromatin structure of single cells [92].

Targeted microscopy approaches measure spatial distances between genomic regions in individual cells using labelled probes. These typically involve complicated probe design, which can be overcome with a new *in situ sequencing* technique [73] but remains challenging to implement. With any microscopy approach, trade-offs have to be considered: which cells are analysed (fixed or living), number of targeted regions, time dynamics and resolution of obtained images. For an extended discussion, we refer the reader to recent reviews [5, 8].

Chromosome conformation capture uses crosslinking, digestion and proximity ligation to detect genomic regions located close to each other in 3D space. It was originally designed for inputs of millions of cells and had higher statistical power than microscopy [23]. An explosion of conformation-based techniques, including the high-throughput sequencing-based Hi-C [64], has paved the way for new discoveries expanding our general understanding of DNA folding in eukaryotic cells [34], bacterial cells [19] and even viruses [9]. For eukaryotes, these patterns include *topologically associating domains (TADs)*, promoter-enhancer and architectural *loops* and *compartments* (reviewed in-depth by [6, 21, 22, 84]).

A long-standing impediment to our interpretation and understanding of structure formation principles is that chromatin features in individual cells are not equivalent to the average features in a population of cells [31]. To address this problem, the first *single-cell chromosome conformation capture assay (scHi-C)* reduced the scale of the traditional Hi-C protocol to one cell per reaction tube [68]. Then, scHi-C was extended by

Aleksandra Galitsyna is a PhD student in Prof. Gelfand's group at Skolkovo Institute of Science and Technology. Her scientific agenda includes various topic in the broad field of chromatin research focused on biology and bioinformatics of single cells.
Mikhail Gelfand is a bioinformatician, professor at the Center of Life Sciences and vice president for biomedical research at the Skolkovo Institute of Science and Technology, Moscow, Russia. His scientific interests include a broad range of problems, ranging from bacterial genomics and evolution to transcriptomics, splicing and mRNA editing in eukaryotes, with chromatin structural analysis being one of them.
Submitted: 31 May 2021; **Received (in revised form):** 09 July 2021

introduction of sorting into multi-well plates and tagmentation followed by polymerase chain reaction (PCR) [69]. A similar approach, *single-nucleus Hi-C (snHi-C)* substituted traditional PCR with whole-genome amplification and cut out the biotin fill-in step. This came, however, at the cost of larger sequencing volumes and data processing [28, 32]. *Diploid chromatin conformation capture (Dip-C)* has adapted tagmentation-based strategies [86, 87], simplifying the experimental protocol [85]. Single-cell combinatorial indexed Hi-C (sciHi-C) is yet another powerful technique based on several rounds of combinatorial barcoding of diluted samples without isolation of individual cells [49, 76]. scHi-C can be combined with other assays to investigate the methylome, such as *Methyl-3C* and *sn-m3C-seq* [55, 57]. For the sake of simplicity, we will refer to all the family of methods a scHi-C throughout this review.

Alongside scHi-C, there is a growing family of many-body interaction capture methods, including *MC-3C* [88], *PORE-C* [91], *Nano-C* [13]. These methods recover up to several dozens of pairwise contacts from individual cells but cannot yet compete with scHi-C in genome-wide searches for architectural features. *Single-cell SPRITE* is a ligation-free method that generates 30 times more contacts but captures interacting complexes instead of pairs [2].

The main challenge of analysing scHi-C data is extreme data sparsity. On average, up to 700 000 interactions are captured in any given cell (for mouse [55]). Thus, the power of scHi-C manifests itself when data for multiple cells are available. Firstly, it makes the detection of chromatin patterns of individual cells statistically reliable. Twenty cells may already be sufficient to assess the presence of TADs, compartments and loops at the level of individual cells of *Drosophila* [93]. Secondly, multiple cells may be clustered into groups of similar types and pooled *in silico*. Such pseudo-bulk Hi-C of scHi-C-guided groups is a better alternative to bulk Hi-C, where the contacts formed in different cell types are indistinguishable [69, 85, 87]. To analyse such data, one needs specialised tools and computational pipelines, which are currently designed *ad hoc* and are rarely re-used or cross-tested. Here, we describe the diversity of recent scHi-C studies and summarise computational approaches to single-cell interactome data (for a recent review of similar topics, see [102]).

Overview of single-cell Hi-C techniques

Like traditional bulk Hi-C, single-cell Hi-C includes chromatin crosslinking, cells permeabilisation, DNA digestion, proximity ligation and library preparation. A crucial step of scHi-C, however, is either *isolation* or *barcoding* of individual cells. To separate contacts from each nucleus, a typical approach is to isolate cells or nuclei into individual reaction mixtures and perform subsequent steps separately. The isolation can be done following crosslinking of cells [28, 82], after ligation [85–87] or right before de-crosslinking [16, 68, 69]. Technically, this is performed by manual placement of each nucleus into a single tube [28, 32] or fluorescence-activated cell/nucleus sorting (FACS/FANS) into individual wells of a plate [16, 82, 87]. Right before or during sorting, optional steps can be included, such as imaging [52, 82] or bisulfite conversion [55, 57]. Isolation-free technique *single-cell combinatorial indexed Hi-C (sciHi-C)* involves several rounds of combinatorial barcoding of the diluted cells [49, 76]. Isolation-free sciHi-C requires demultiplexing as one of the first data processing steps, while the isolation approach may [69] or may not include this step. A more comprehensive overview of the scHi-C experimental technique can be found [92], but we will highlight aspects of different protocols that are particularly relevant for data processing (Figure 1).

The initial step of the scHi-C protocol is to crosslink cells with formaldehyde, resulting in the fixation of DNA-DNA interactions. Next, cell membranes are lysed to guarantee the delivery of reagents into the nucleus. Then, DNA is digested by a restriction enzyme such as DpnII that cuts at the four-letter palindromic motif GATC (Figure 1A). This produces free ends of restriction fragments, which are then ligated either directly [28, 32, 85–87], after biotin fill-in [68, 69, 82] or after ligation of a biotinylated bridge adaptor [49, 76] (Figure 1B). Ligation junctions containing biotin-labelled nucleotides are pulled down using streptavidin. This pulldown is omitted in some scHi-C variants because it results in a loss of meaningful contacts [28, 32, 85–87]. Regardless of the ligation procedure, properly formed junctions are expected to contain specific sequences (restriction sites with or without a bridge, Figure 1), which can be used to computationally select real contacts [93]. The final step of scHi-C is to extract DNA and prepare it for sequencing. Multiple library preparation strategies were probed with scHi-C (Figure 1C), including whole-genome amplification (Illustra WGA in [28], META WGA in [87]), tagmentation followed by PCR [69], digestion with a restriction enzyme followed by primers ligation and PCR [68], barcoding and PCR [76] or PCR with random primers [57]. While tagmentation and restriction enzyme digestion generate fixed-point cuts in the DNA resulting in simple rules for computational deduplication of the pairs with coinciding mapping positions, this is not the case for whole-genome amplification and PCR with random primers, for which other deduplication schemes should be used. Finally, amplified DNA is purified and sequenced in the paired-end mode.

Data processing workflow

The data processing workflow (Figure 2A and B) consists of general steps shared with typical Hi-C: optional pre-processing of reads (trimming, demultiplexing, etc.), read mapping, optional restriction fragment assignment, filtration of contacts and deduplication and binning with generation of single-cell Hi-C maps. The cells are typically filtered by the quality and/or the number of contacts.

Mapping of reads

As with any other conformation capture, scHi-C generates chimeric DNA molecules (Figure 2C), making the mapping of these discontinuous reads to multiple genomic locations non-trivial [51]. Standard mappers, such as BOWTIE2 [54], cannot reliably map such reads. There are four main approaches to treat scHi-C chimaeras, three of them transferred from traditional bulk Hi-C: split read alignment, iterative mapping and read clipping. The fourth approach is one-read-based mapping (ORBITA), a special case of the split read alignment [93], which attempts to find only those contact pairs that are directly ligated (Figure 2c). In the *split read alignment* strategy, specialised mappers like BWA MEM [58] detect multiple sequential alignments in each read. Of these, only the representative alignments are retained (typically, the alignments at 5′-end). Some studies use the information about 3′-end alignments to specify the endpoints of contacting fragments [86]. *Iterative mapping* is a method of analysing chimeric reads initially used for traditional Hi-C [42] and adapted for single-cells [28, 32]: short 5′ sequences of increasing size are iteratively selected on both forward and reverse reads until the mapping of the pair

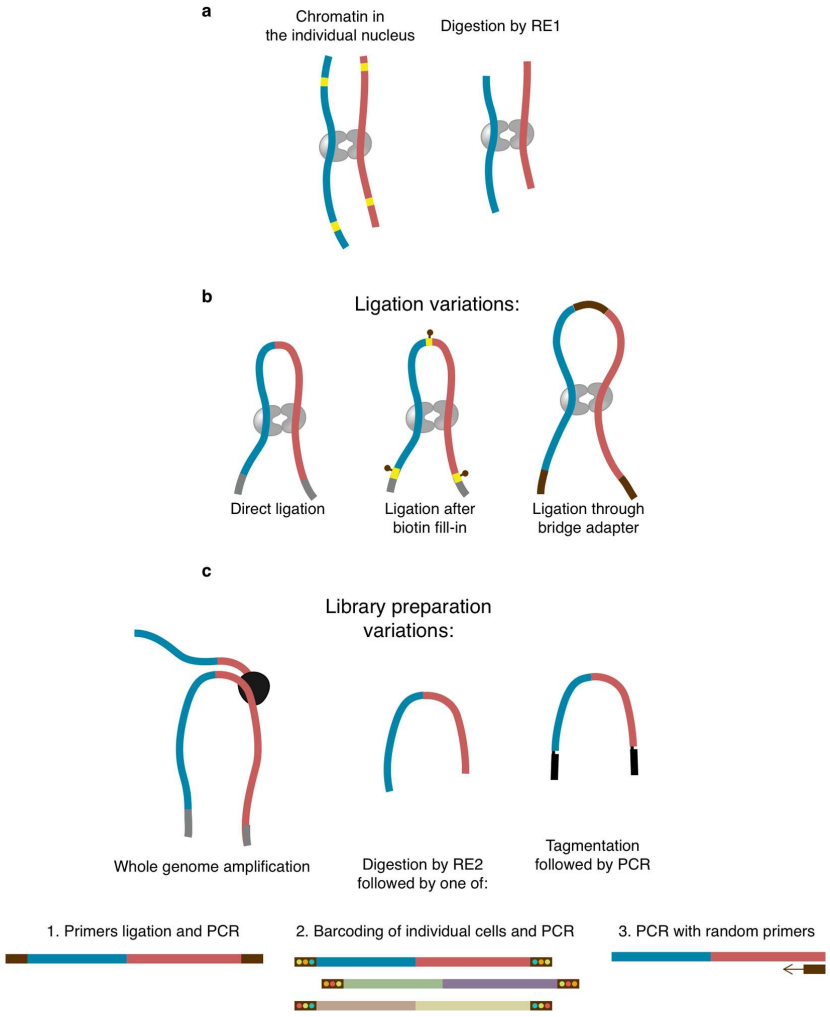


Figure 1. Overview of variations in scHi-C protocols relevant for data processing. **A**. Cross-linking and digestion, used in any scHi-C. **B**. Variations of the ligation step. **C**. Variations of the library preparation. RE1 and RE2 denote restriction enzymes selected for corresponding stages.

(or coverage of the full read length) is achieved [51]. In *read clipping*, reads are scanned for the restriction site [69, 82] or bridge adapter [76], and all the 3′ sequences after the match are removed. Two resulting paired sequences (one for forward and one for reverse read) are mapped independently and form a contact pair if the mapping was successful. However, only *one-read-based interactions annotation* utilises the information on chimeric parts to guarantees that the observed pair is a direct ligation junction of DNA fragments (Figure 2c). This approach reduces erroneous contacts in scHi-C data [93].

Another problem during scHi-C read mapping is genetic variation. Some regions of the genome of the studied cells differ from the reference hampering the mappability. Moreover, the cells are not guaranteed to descend from a single clone [86] and may have intrinsic variation, such as single-nucleotide polymorphisms (SNPs). Thus, some studies [82] ignore genomic locations with SNPs and prohibit mapping mismatches. On the other hand, SNPs can be a powerful source of information to help distinguish haplotype alleles [16, 69, 86] and impute the contacts of the maternal and paternal chromosomes [86].

Filtering of contacts

After mapping, the scHi-C maps are vastly populated with *amplification duplicates*, *contacts of promiscuous genomic regions* and *artifactual contacts*, which can be detected and filtered out.

Amplification duplicates are identical or nearly identical copies of the same contact pairs generated during library preparation. Depending on the experimental protocol, the scHi-C duplicates do not necessarily have the same mapping positions in the genome. Whole-genome amplification and PCR with random primers produce DNA fragments that may originate at random locations close to actual ligation position. Thus, if a group of contact pairs has the same restriction fragments [76] or their termini [69, 93], these contacts are likely to have been duplicated and should be merged into a single contact. Alternatively, contacts of the same 500 bp-bins [28] or contacts located closer than 1 kb [86] may be merged directly [28] or iteratively [86].

The genome coverage in conformation capture is affected by multiple factors, including replication, DNA accessibility, GC-content and active chromatin state [42, 78, 97]. In bulk Hi-C,

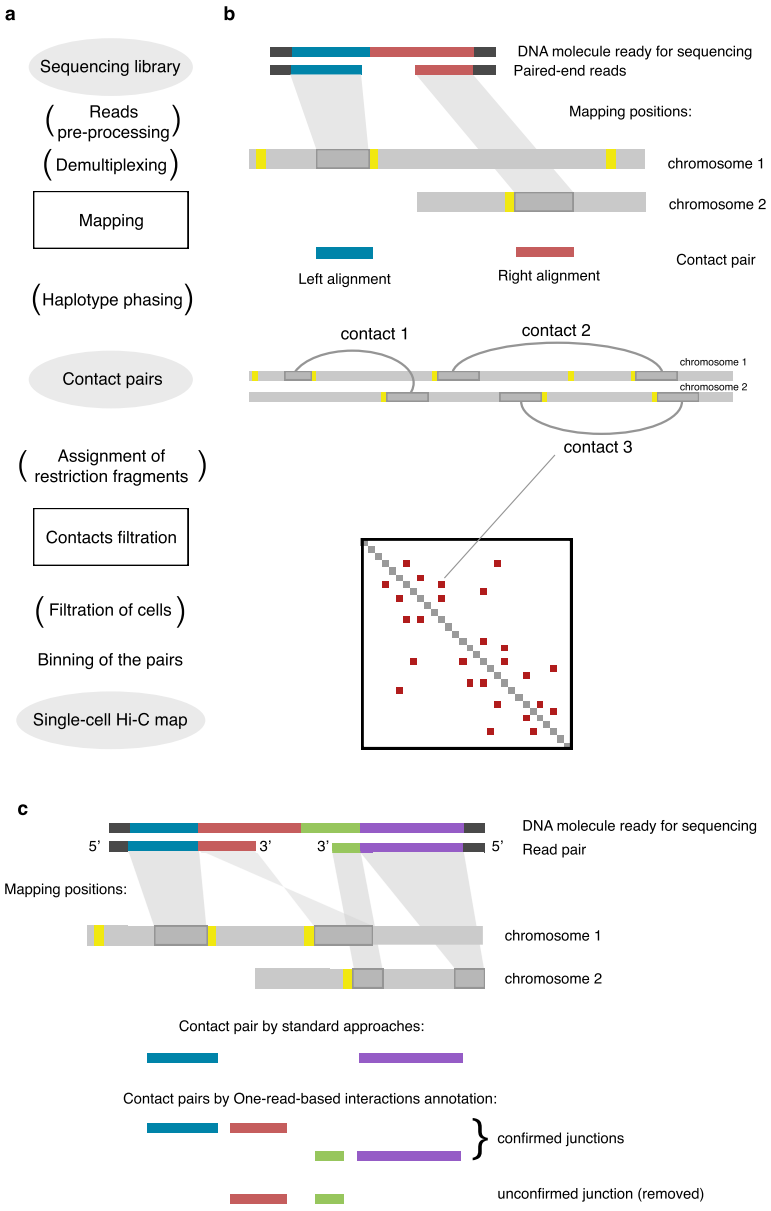


Figure 2. Outline of single-cell Hi-C data processing. The steps in brackets are optional, depending on the scHi-C protocol and the pipeline specifics.

this is mitigated by correction, such as iterative balancing [42]. However, due to data sparsity, this step is not recommended for scHi-C (although proposed as intermediary step of quality assessment [40]), and little research has been devoted to scHi-C correction alternatives [59, 66]. In the absence of data correction, scHi-C may bear intrinsic biases, such as larger numbers of contacts formed by active regions [93] and early replication domains [69]. Larger numbers of contacts have been suggested for regions with genomic rearrangements [69], e.g. Stevens *et al.* [82] detected trisomy by the increased number of contacts for the whole chromosome. As a partial remedy, one can remove *contacts of promiscuous genomic regions* [68, 69, 93], e.g. 1 Kb regions that have more than ten contacts in a given cell [86].

Artifactual contacts are random contacts happening at various stages of scHi-C sample preparation and data processing, typically not representative of the real 3D conformation of chromatin and impairing downstream analysis. First of all, properly formed and mapped pairs should be located close to the restriction sites. scHi-C protocols using Phi29 phage polymerase can generate switch templates during WGA that are devoid of this feature and should be discarded [93]. The original scHi-C protocol generates a number of *spurious ligations*, likely represented by the pairs supported by a single read [68]. Frequent artefacts are *sequencing pairing mismatches*, having a global rate of 0.1% for Illumina [69, 76], as assessed by admixture of phiX174 DNA to mouse cells [69]. Stevens *et al.* [82] suggested a general scheme for filtering a broad

range of scHi-C artefacts, which is based on the assumption that the regions in close spatial proximity have the neighbouring genomic regions located nearby, also forming a contact. Thus, if the contact is isolated (e.g. is not supported by neighbours within 2 Mb distance [82]), it is likely to represent an artefact and should be removed [82, 86].

Filtering of cells

Data from some cells should entirely be discarded due to the failure of the protocol in those cells. Multiple criteria to identify such problematic cells were proposed: *robustness to downsampling* [93], *fraction of read-pairs sequenced only once* [68] and *fraction of non-digested DNA* [69]. The most commonly used criterion is *cell coverage*, that is, the total number of detected contacts per cell [69]. For example, the cell coverage in sciHi-C follows the bimodal distribution, with low-coverage cells likely representing in-solution DNA noise [76]. Yet, another popular criterion is cumulative contacts properties, such as *cis-to-trans ratio* [68, 69, 76], defined as the ratio of the intrachromosomal contacts to the interchromosomal contacts. Typically, interchromosomal contacts in the chromatin occur with a lower probability than intrachromosomal ones, a phenomenon called *chromosome territoriality* [17]. Artifactual contacts are less likely to depend on the 3D distance between corresponding genomic positions and, thus, a deviating cis-to-trans ratio for a cell might signify excessive spurious ligation. Similar assumptions are used to filtrate the cells by *distance decay properties* of contacts [69] and *cross-species ligation frequency* [69, 76]. Another notion guiding the choice of high-quality cells is that scHi-C contacts tend to be found in clusters. Based on this observation, *GiniQC* measures the level of unevenness of inter-chromosomal scHi-C maps [40].

Data structure

The scHi-C contact data are typically represented as a matrix, similar to the standard Hi-C [68]. Each cell in this matrix corresponds to a pair of genomic bins, and the value in a cell is the absolute number of interactions between these bins. A set of experiments is stored as a set of matrices, while specialised file formats exist to store matrices for a number of cells, such as scool [96]. Hypergraphs [99] and 'topics' [49] are representations for a set of cells used for specialised applications, such as prediction of contacts using machine learning [99] and data decomposition [49]. For special applications, scHi-C can be represented as a vector, for example, when scRNA-Seq methods are transferred to 2D data [37]. The 3D model is a popular representation, although it requires substantial preprocessing of the data and is not necessarily back-convertible to the set of initial contacts [68, 82, 86].

Graph representation

Graph representation [10, 100] is a popular representation that can be used to *upper bound for the number of pairwise contacts in scHi-C maps* [93] (Figure 3A). This upper bound can be defined for scHi-C but not bulk because a single cell with defined DNA content is used in the experiment. It depends on the number of restriction fragments that can potentially form contacts, which in each cell depends on the restriction site frequency, the organism's genome size and the number of DNA copies in a particular cell type. For example, a single copy of the mouse genome mm10 [15] contains 6.6 million DpnII restriction sites (Figure 3B). In theory, if both ends of each restriction fragment were ligated to the ends of other restriction fragments and all ends are ligated, then the fragments form a circle graph. Thus, the number of contacts that could be detected would equal to the number of restriction fragments (Figure 3A). If two copies of the mouse genome are present (in a diploid cell), the number of possible contacts will be around 13.2 million. This number may be higher for cells during mitosis, S or G2 phase of the cell cycle, when the genomic content, and hence the number of restriction fragments, is completely or partially doubled. Although non-realistic to achieve in the working conditions of scHi-C, this number can serve as a theoretical upper bound to the possible number of pairwise contacts in a single nucleus. Notably, the largest number of contacts per cell obtained to date for mammals [57, 85] is already larger than the theoretical limit for the haploid genome of *Drosophila melanogaster* (Figure 3C), suggesting that the complete recovery of contacts of small genomes is possible with scHi-C.

The upper bound estimate can serve as a normalisation factor for contacts recovery in scHi-C studies (Figure 3D). The best standard scHi-C [93] has 17% contacts recovery and the joint assay with methylation, sn-m3C-seq, almost reaches 25% [55]. It is important to note that for an ideal scHi-C with 100% recovery, we still cannot expect more than 2.4 interactions per 1 Kb of the genome (for haploid mm10 genome). This number is two orders of magnitude lower than bulk Hi-C (around 1700 contacts per 1Kb or genome in neural progenitor cells [7]). Thus, even if the theoretical limit is reached, scHi-C remains profoundly sparse and specialised software is required for its downstream analysis.

Data analysis

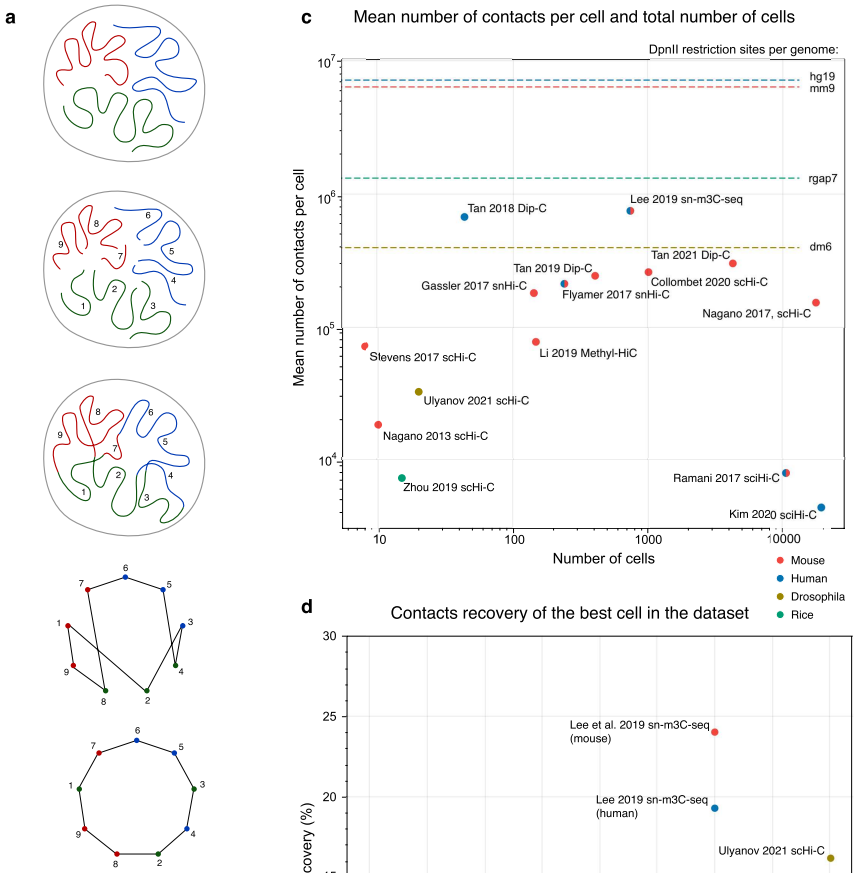
There are two general approaches to the scHi-C data analysis, depending on the solution to the problem of low statistical power of scHi-C data sparsity. In the first one, every single cell is processed independently. It includes building its 3D model, data imputation, aggregation analysis and features calling. In the second approach, single-cell maps are analysed together, then grouped and pooled to produce pseudo-bulk Hi-C maps.

Structure reconstruction

A typical approach for the 3D structure reconstruction is to build a *beads-on-string model* restrained by molecular dynamics with simulated annealing [68]. Each bead corresponds to a genomic bin of a given size (ranging from 10 Kb [93] to 1 Mb [69]), while each bond is either a polymer backbone or an observed scHi-C contact. The simulation starts from a random initial conformation, where the beads involved in observed scHi-C interactions might be overstretched. The beads connected by bonds are attracted to each other, forcing a rearrangement of the structure so that connected beads are located in close spatial proximity. Some bonds do not balance and remain overstretched; thus, they can be removed [82, 93] as potential experimental artefacts [53]. Other proposed solutions include Bayesian inference [11], recurrence plots [39] and lattice models [103]. All these methods remain data driven and do not account for the actual mechanisms of chromatin structure formation [43].

Imputation of missing data

Due to contacts sparsity, applications of bulk Hi-C analysis tools to scHi-C are restricted [59]. To mitigate this effect, imputation techniques bring the numbers of scHi-C contacts closer to bulk [102]. Zhou *et al.* [100] populate the map with contacts



b

Genome	Number of DpnII restriction sites
Human (hg38)	7199481
Human (hg19)	7227576
Mouse (mm10)	6684538
Mouse (mm9)	6415222
Drosophila (dm3)	332951
Rice (rgap7)	1325057

Figure 3. **A**. Illustrative upper bound estimation of the possible number of pairwise contacts per single cell. The theoretical genome has nine restriction fragments that form a circle graph after ideal ligation (nodes are restriction fragments with the valency of 2, edges denote ligation of their ends). **B**. Total numbers of DpnII restriction sites for the single copies of popular genomes. **C**. Descriptive statistics of published scHi-C studies. The lines represent the upper bounds for the possible number of contacts per single cell from (B). Colour indicate species. **D**. The best cells for some of the published scHi-C datasets as a function of the publication time. For C and D, we use the numbers reported in the supplementary materials of the original studies, when possible. For each study, we indicate the first author and the names of scHi-C techniques self-reported by the authors. For [49] and [76], the mean is calculated based on the median count per dataset. For [86], we used the cleaned contacts after removal of damaged cells. For [55], the calculated mean is based on the numbers reported for 741 cells in the supplementary table.

generated by a random walk, making the scHi-C graph closer to a complete clique. Stevens *et al.* [82] and Ulianov *et al.* [93] use the maps imputed by polymer models. Notably, both TADs and compartments can also be readily assessed from model-imputed maps [82, 93], with TADs similar to those in original scHi-C data [93]. As a substantial breakthrough in scHi-C data imputation, inter-cellular patterns of contacts can be accounted for by the hypergraph neural network [99]. Some studies test the technical possibility to transfer dropout imputation algorithms for single-cell RNA-Seq, although lacking theoretical support [37].

Contacts aggregation and features calling

Two approaches have been suggested to study TADs, loops and compartments in scHi-C maps, aggregation analysis and *features calling* (Figure 5). During *aggregation*, the statistics of contacts is accumulated over predefined regions of the genome (e.g. CTCF

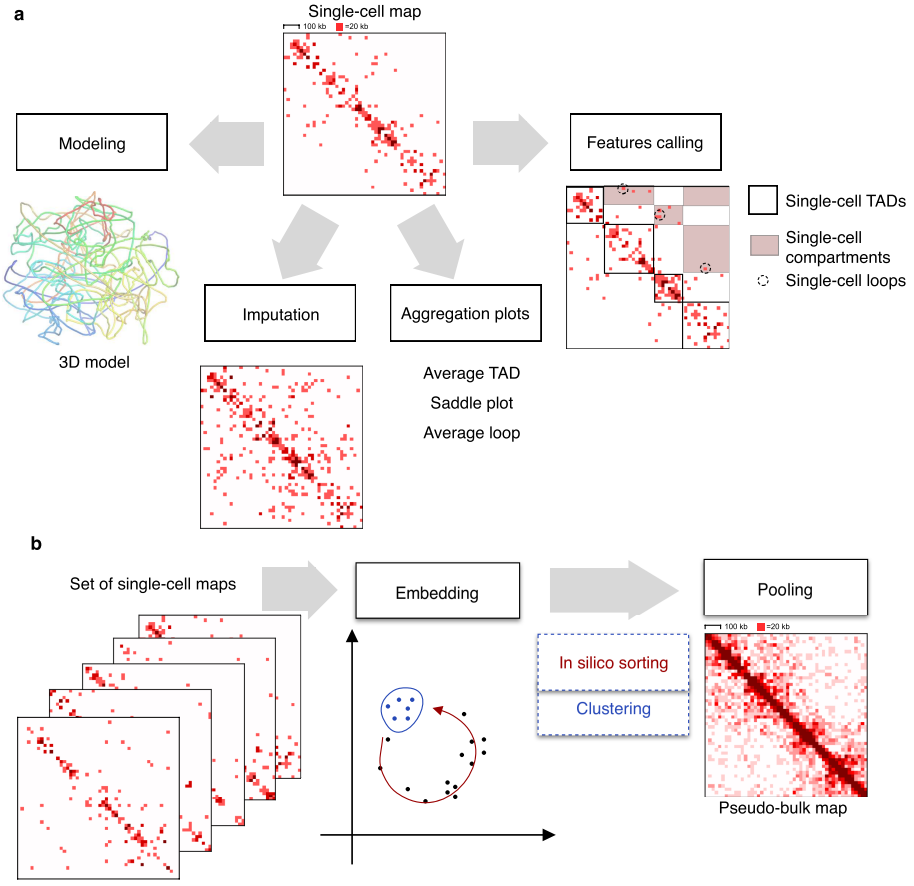


Figure 4. Approaches to studying a single scHi-C map (**A**) and a set of scHi-C maps (**B**). Single-cell Hi-C maps from [28] for the region chr1:9000000-1000000.

binding positions to assess loops; bulk TADs or bulk compartments). Aggregation confirms the presence of these chromatin features in individual cells [32], and there is specialised software for this purpose [29]. With *features calling*, the positions of individual loops [82], TADs [28, 60, 75, 93] and compartments [75, 86] are found directly in the scHi-C map, demanding high-quality scHi-C maps and providing insight into variability between individual cells. For example, the positions of TADs in individual cells demonstrated higher stability of TAD boundaries between individual cells of *Drosophila* than between mouse oocytes [93].

scHi-C embedding

scHi-C data are multidimensional ($\sim N^2$ contacts measurements for N genomic regions) and can be projected into a space of lower dimension for visualisation, clustering and sorting. Typical visualisation is a scatter plot where each dot is a cell, and the axes correspond to some characteristics of the cells. The values on the axes can be derived from some additional measurement, such as the levels of the DNA replication marker geminin and DNA content from FACS [69] or the level of DNA methylation [55, 57]. Alternatively, the axes can represent some explicitly calculated interpretable characteristic of the scHi-C maps, such as the total number of contacts, the cis-to-trans ratio [16, 69] and the percentage of local/mitotic contacts [16, 69]. Tan *et al.* [86] characterise the 3D models instead, plotting the strength of the Rabl configuration, the centrality of telomeres, the number of interchromosomal neighbours, the average CpG content of the neighbours and the probability of cell-type-specific loops.

Finally, the axes might not readily correspond to any known biological characteristics—scHi-C maps can be transformed and subjected to *dimensionality reduction* by the *principal component analysis (PCA)* or other techniques (see Table 1 for comparison). For example, Ramani *et al.* [76] apply PCA to matrices of interchromosomal interactions and find that the first component explains a large fraction of the variance (52.1%) and strongly correlates with the coverage. The combination of the first and second (1.07% explained variance) components distinguishes cell types. Nagano *et al.* [69] observe the cell cycle-dependent embedding of scHi-C by calculating the pairwise symmetric Kullback–Liebler divergence on vectors of distance decays and subsequent spectral embedding. Collombet *et al.* [16] apply *uniform manifold approximation and projection (UMAP)* to vectors of TAD contact profiles; Li *et al.* [60] perform PCA on pairwise similarities of TAD profiles; Tan *et al.* [87] calculate the compartment score profiles for each cell, take 20 principal components and then visualise it with *t-distributed stochastic neighbour embedding (t-SNE)*. One of the most generalised approaches is HiCRep [100], which calculates a similarity matrix between each pair of individual cells, taking the *stratum-adjusted correlation coefficient (SCC)* measure of similarity. HiCRep with *subsequent multidimensional scaling (MDS)* has proved to be one of the best approaches to study embedded

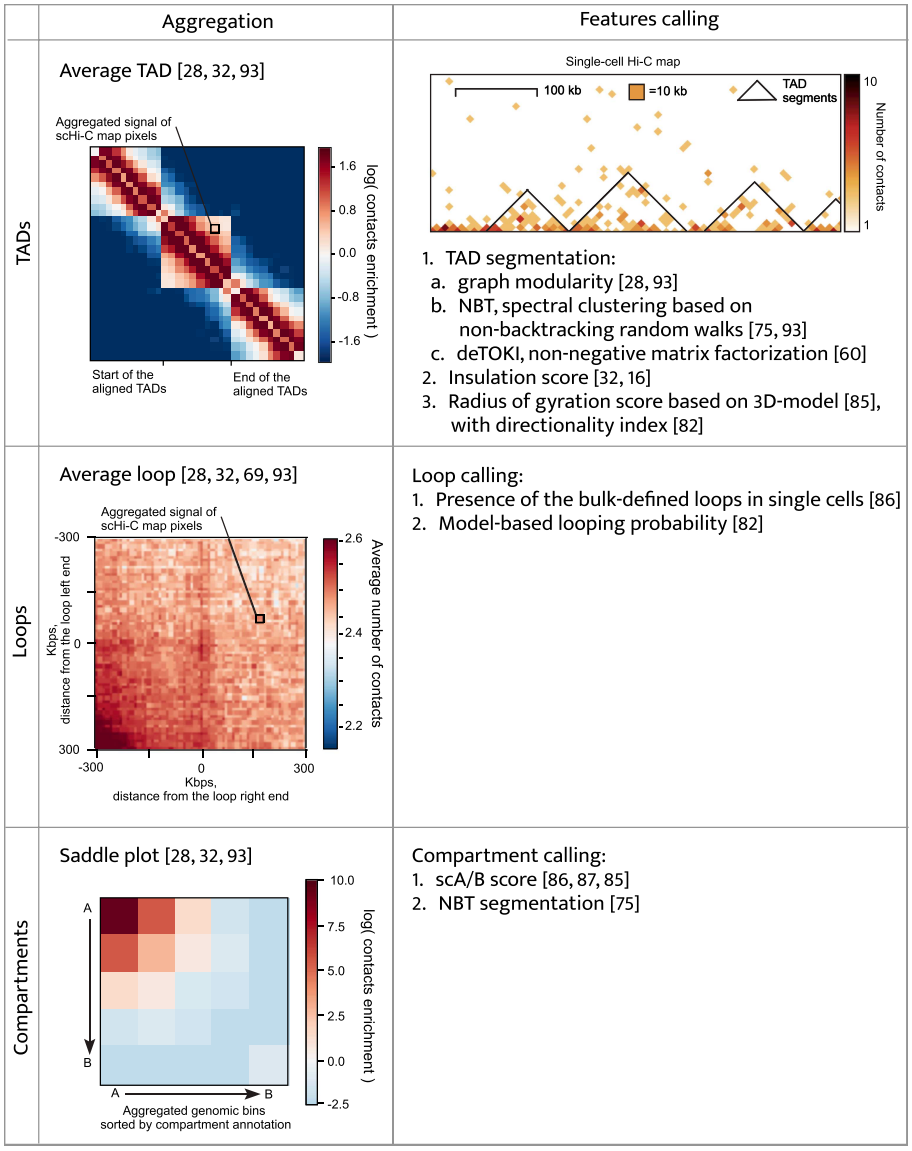


Figure 5. Comparison of aggregation of contacts and features calling for TADs, loops and compartments. All the examples are for the *Drosophila* scHi-C map of Cell 1 from [93]. Average TAD and saddle plot are for bulk TADs and compartments, while average loop is for the top 1000 regions with the highest content of RED chromatin state from [48].

scHi-C datasets [65]. In this approach, Zhou *et al.* [100] propose to impute potential dropouts before the embedding to increase the cluster separation. The imputation was further supplemented it with scRNA-Seq dropout correction methods [37] (but see the discussion above).

An alternative, scHiCExplorer [96], implements an approximate nearest neighbour method with a local sensitive hash function, MinHash. Finally, some approaches suggest using the co-occurrence of contacts in individual cells to base the embedding on meaningful single-cell patterns. For example, Kim *et al.* [49] applied *latent Dirichlet allocation* to factorise the scHi-C dataset into a set of documents, words and topics, and Zhang *et al.* [99] used a *hyper-graph neural network*. In all these studies, the axes created by *in silico* approaches are rarely interpreted, and it might be of interest to correlate them with various scHi-C characteristics such as the contact coverage, distance decay, strength of TADs, loops and compartments.

A more exotic approach is to describe scHi-C space in terms of topological data analysis [10]. Finally, joint assays of the methylome and interactome [55, 57] allow for independent embeddings of scHi-C and single-cell methylation patterns and subsequent comparison of resulting embeddings.

To date, no comprehensive studies on embedding all existing scHi-C datasets have been published. Moreover, there have been no attempts to embed datasets originating from different species, although scHi-C data for human [28, 49, 76, 86], mouse

Table 1. Summary of major scHi-C embedding techniques

Family of embedding methods	Linearity	Primary reference	Special scHi-C pre-processing	Special measure of similarity/difference between cells	Explicit usage of contacts co-occurrence patterns
PCA	Linear	[100]	Raw binned matrix	–	No
		[76]	Interchromosomal interactions profile	–	No
		[60]	TAD profile	–	No
		[87]	Compartment score profile	–	No
t-SNE	Non-linear	[85]	20 PCs of compartment score profiles	–	No
Spectral embedding	Non-linear	[69]	Distance decays	Symmetric KL	No
MDS	Non-linear	[65]	Distance decay	Jensen–Shannon divergence	No
		[100]	scHi-C binned matrix after smoothing and random-walk imputation	SCC	No
UMAP	Non-linear	[16]	TAD contact profiles	–	No
		[49]	Cell-topic matrix after LDA	–	Yes
		[99]	Hypergraph embedding	–	Yes

MDS indicates multidimensional scaling; SCC, stratum-adjusted correlation coefficient; UMAP, uniform manifold approximation and projection.

[16, 68, 69, 76, 82, 85, 87], *Drosophila* [93] and rice [101] are available. This might identify species-specific patterns in genomic interactions and their variability.

While both linear and non-linear embeddings of scHi-C have been proposed, advanced *manifold learning* techniques are yet to be developed for scHi-C, analogous to the outbreak of embedding methods for single-cell RNA-Seq data (reviewed in [67]). At that, multiple, diverse formalisations of scHi-C as matrices, graphs and vectors allow for a broad field of embedding techniques to be studied on these datasets.

In silico sorting, clustering and pooling

Based on the position in the embedding space, scHi-C data can be *in silico* sorted [69] or clustered [85]. Nagano *et al.* [69] observed the ordering of the cells by the position in the cell cycle, while Tan *et al.* [85] derived subtypes of mouse brain cells using k-means. Collombet *et al.* [16] relied on outliers in the embedding space to filter out cells undergoing mitosis and retain only interphase embryonic cells.

Specialised approaches, including the ones based on machine learning, have been designed for scHi-C data clustering. Typically, these applications require embedding (see below). The quality of clustering is tested on datasets with known ground truth (e.g. types of pronuclei in the mouse zygote [28] or types of cells forming the dataset [76]). Each cluster, or group of cells, is assigned with a particular cell type and the quality is usually assessed by normalised mutual information [62] or adjusted rand score [62, 100].

The resulting groups of cells can be pooled by simple summation of single-cell Hi-C maps, resulting in *ensemble*, or *pseudo-bulk*, Hi-C and analysed as typical bulk Hi-C [16, 69, 85]. Pseudo-bulk scHi-C maps are a powerful technique for detection of cell-type specific differences in the chromatin architecture. For example, pseudo-bulk mitotic cells lack the TAD and compartment structure [69], while subtypes of brain cells have differences in regions of cell-type specific genes [85].

The long-studied field is the reverse of the pooling, namely deconvolution of bulk interaction maps into a set on single cells [46]. Such approaches aim to construct a population of genome structures with a total set of genomic interactions approximating (or equal to) a set observed in a population of nuclei. Several advanced techniques including machine learning have been suggested, such as *maximum likelihood* [89], *Bayesian inference* [12], *fractal Monte Carlo weight enrichment with Bayesian deconvolution* [74], *Monte Carlo with bag of little bootstraps* for the generation of bootstrap structures [83] and, most recently, stochastic embedding [36]. However, these approaches are limited by the number of models that approximate bulk datasets (up to several tens of thousands), although around 5–10 million structures contribute to the typical bulk Hi-C map. Nevertheless, it might be interesting to demonstrate the reversibility of the pooling of a low number of single-cell maps by applying some of these methods to pseudo-bulk datasets. Guarnera *et al.* [36] assessed the variability of polymers after deconvolution, which might be interesting to compare with results obtained from embeddings of real scHi-C.

Design of scHi-C controls

Due to the complex nature of scHi-C data, a good practice is to design scHi-C controls to validate the hypotheses. These include *sampled*, *shuffled* or *de novo generated randomised scHi-C maps*, which typically have the same number of contacts as real cells. *Sampled maps* are populated by contacts randomly selected from

bulk [93] or ensemble [68, 76] datasets. However, it creates maps less sparse and heterogeneous than real scHi-C maps [100]. Thus, an effective number of sampled contacts can be increased or additional artificial noise can be introduced [100]. *Shuffled maps* are single-cell maps with randomly permuted pairs of contacts [68]. This procedure retains coverage by contacts but removes any information on the spatial structure, including distance decay. Sampling and shuffling can be combined together: bulk Hi-C maps first randomised, preserving the coverage and distance decay, and then sampled [69]. *De novo generative models* do not rely directly on the observed contact maps while preserving the meaningful properties of scHi-C maps. For example, thresholding the distance between genomic regions in polymer models [93] produces control maps with meaningful distance decays. A more advanced alternative, stepwise generation of single-cell Hi-C-like maps, preserves both distance decay and observed coverage by contacts [93].

Controls like this allow differentiating the technical and biological properties of the single-cell contact maps for features calling (such as TADs) and aggregation analysis [93]. They provided the baseline for assessing the general quality of modelling by the number of violated constraints [68]. Further, they demonstrated that scHi-C maps are non-random [82] and chromatin features of the modelled cells are similar to that of the real cells [69, 82, 93]. Yet another important observation is that real scHi-C data are more variable and sparse than bulk subsamples [100]. Although randomised scHi-C control is a powerful method, it is sporadically used in scHi-C studies. This will improve with the development of specialised tools for this task and the emergence of theoretical studies on the statistical properties of single-cell contacts.

Outlook and challenges

Single-cell Hi-C is a young and rapidly developing field in chromatin biology. Due to its extreme data sparsity and complicated experimental protocol, the quality of the datasets has been a limiting factor. However, with the emergence of simplified and cheaper protocols [76, 86], we anticipate continued growth of both coverage of scHi-C and number of cells analysed, leading to improved data resolution and statistical reliability of the biological results. This will also stimulate the development of new data processing and analysis methods. However, as we demonstrated here, scHi-C data have a natural upper bound for the possible number of recovered single-cell interactions; thus, data sparsity will remain a challenge for the field.

Despite the substantial efforts to work with sparse data, the computational analysis of scHi-C has not reached maturity yet. For example, a recent re-analysis of datasets from three studies demonstrated that inappropriate contacts mapping may result in the accumulation of experimental artefacts and overestimation of the number of recovered contacts [93]. However, if the data from multiple studies were processed uniformly, it demonstrated that TAD boundaries in *Drosophila* are more conserved than in mouse. Similar comparative analysis of scHi-C results will further shed light on reproducible chromatin features in individual cells in an unbiased way.

Machine learning has a growing impact on our understanding of biological systems (reviewed in [27, 63]) and 3D genomics [4, 30, 77, 79, 94, 95]. For single-cell chromatin research, imputation and embedding are already driven by neural networks [99] and other advanced machine learning methods will emerge. Importantly, features calling from single-cell data will be improved.

Next, an important direction is improving structural reconstruction approaches. To date, scHi-C structure reconstruction does not account for a specific mechanism of structure formation. Alternative *de novo* modelling assumes the particular mechanism but does not incorporate scHi-C contacts [28, 30]. These approaches can be, in theory, united to open intriguing perspectives. For example, can we simulate loop extrusion [31] that will produce the contact maps similar to those observed in scHi-C? Can we infer the cohesin loading sites in individual cells based on observed contacts? Finally, can we differentiate the cohesin-dependent contacts in single cells from compartmental ones [32] and study them independently?

These challenges are not the only ones that will require computational solutions. An important direction will be the design of new assays, as well as tools for their data processing. For example, currently, restriction enzymes digest chromatin into relatively large restriction fragments, which dictates the strict upper bound for the total number of pairwise contacts recoverable from a single cell. If micrococcal nuclease is used instead, it will allow for up to 15 million contacts of individual nucleosomes in the haploid human genome [1], increasing the theoretical upper bound at least twice.

Joint assays, other than Methyl-3C and sc-me3C, will unravel the interplay of chromatin architecture with other cellular mechanisms. For example, measuring single-cell lamina-associating domains (LADs) alongside scHi-C will shed light on the lamina association of individual TADs. Indeed, bulk TADs do not entirely correspond to either bulk [91] and single-cell LADs [50]. However, it is possible that single-cell TADs are elementary units of interaction with lamina if there is a one-to-one correspondence between TADs and LADs observed in the same cell. Next, measuring chromatin openness and/or transcriptional activity will accelerate the research on interplay and causality between regulation, chromatin folding and gene expression [24]. On the computational side, having more than one type of measurement in single cells is a unique opportunity to develop joint embedding [56] methods, which use both interaction graphs and single-cell features to create meaningful low-dimensionality representation. Also, having several types of measurements will help to develop and benchmark standard scHi-C embedding techniques.

Single-cell RNA-DNA contacts will help distinguish RNA-mediated interactions from the rest and depict the single-cell pattern of regulatory RNA functioning. However, the resolution of bulk RNA-DNA interaction capture techniques is relatively low [3, 33, 61, 81], which will remain a major impediment for single-cell RNA-DNA interactions as well.

Currently, scHi-C requires vast sequencing with relatively low meaningful output (e.g. Ramani *et al.* [76] sequenced over 170 mln reads per dataset on average, only 11% of them resulting in unique contacts). However, studying biological mechanisms of chromatin compaction and regulation frequently requires engineering and targeting of individual regions of the genome limited in size. Thus, it might be beneficial to develop single-cell Hi-C with enrichment for targets. Target enrichment for a genomic region is already well developed for bulk chromosome capture approaches [20, 25, 35]. Adaptation of these approaches for the single-cell level will allow for specific enrichment of single-cell interactions of regulatory regions that might undergo the specific architectural changes in a cell population.

As both wet-lab and computational scHi-C methods improve, it will lead to breakthroughs in understanding biological systems currently restricted by bulk Hi-C. For example, chromatin transitions during mouse embryogenesis were studied by low-input Hi-C [26, 47], which accommodates the limited number

of embryos available but does not distinguish individual cells. Starting from the zygote and up until the gastrulation (stage E7.5), chromatin features gradually emerge. At stage E7.5, the embryo has approximately 15000 cells, some differentiated into progenitors of diverse tissues and organs [90]. Their variability can be recovered only by scHi-C. Indeed, scHi-C demonstrated cell- and allele-specific patterns of chromosomes folding in mouse embryos but only up to a much earlier stage of 64 cells [16, 28, 32]. Given the fact that existing scHi-C assay several tens of thousands of cells [49], a whole-embryo single-cell chromatin structure study is a realistic short-term goal. This opens an intriguing perspective to answer fundamental questions about chromatin dynamics in development. What paths do chromosomes follow in individual nuclei during tissue differentiation and organogenesis? Can we track the lineages of cells based on their chromatin, as we do for single-cell transcription [80]? Finally, what are the rules governing chromatin transitions in individual cells during the development of other species studied by bulk Hi-C, including human [14], *Xenopus tropicalis* [71], Medaka fish [70], *Danio rerio* [45] and *Drosophila melanogaster* [41]?

Next, scHi-C will uncover the diversity of chromatin architecture within cancer cell, contributing to the clonal analysis of solid and liquid tumours currently done with genomic and transcriptomic methods. Finally, single-cell atlases of chromatin architecture for cell types of different organs will expand our knowledge on chromatin structural diversity. Their proper association with single-cell atlases of transcription [38] and chromatin openness [18, 98] will unravel the interplay between epigenetics, chromatin structure and gene expression.

Key Points

- Single-cell Hi-C is a powerful and rapidly developing technology to study chromatin architecture, with computational analysis playing a crucial role in extracting biological meaning from its sparse readouts.
- The number of scHi-C pairwise genomic contacts is limited by the number of genomic fragments in the nucleus requiring special approaches for sparse interactome data analysis, including structure reconstruction, imputation of interactions, aggregation of contacts and feature calling for a single map and embedding, sorting, clustering and pooling for a set of maps.
- We anticipate improvements in scHi-C data quality and computational analysis to lead to the expansion of scHi-C applications, eventually resulting in breakthroughs in our understanding of cell function comparable with those achieved by scRNA-seq and scATAC-seq.

Author contributions statement

A.A.G. wrote the manuscript and analysed the data. M.S.G. conceived the idea, wrote the manuscript and supervised the work.

Funding

This study was supported by grants from Russian Foundation for Basic Research (19-34-90136 to A.G. and 18-29-13011 to M.S.G.).

References

1. Alberts B, Johnson A, Lewis J, *et al.*. Molecular biology of the cell 5th edition. *Garland Science* 2008.
2. Arrastia MV, Jachowicz JW, Ollikainen N, *et al.* A single-cell method to map higher-order 3D genome organization in thousands of individual cells reveals structural heterogeneity in mouse ES cells. *bioRxiv* 2020. Preprint biorxiv:2020.08.11.242081.
3. Bell JC, Jukam D, Teran NA, *et al.* Chromatin-associated RNA sequencing (chAR-seq) maps genome-wide RNA-to-DNA contacts. *Elife* 2018;**7**:e27024.
4. Belokopytova PS, Nuriddinov MA, Mozheiko EA, *et al.* Quantitative prediction of enhancer–promoter interactions. *Genome Res* 2020;**30**(1):72–84.
5. Boettiger A, Murphy S. Advances in chromatin imaging at kilobase-scale resolution. *Trends Genet* 2020;**36**(4):273–87.
6. Bonev B, Cavalli G. Organization and function of the 3D genome. *Nat Rev Genet* 2016;**17**(11):661–78.
7. Bonev B, Cohen NM, Szabo Q, *et al.* Multiscale 3D genome rewiring during mouse neural development. *Cell* 2017;**171**(3):557–72.
8. Brandao HB, Gabriele M, Hansen AS. Tracking and interpreting long-range chromatin interactions with super-resolution live-cell imaging. *Curr Opin Cell Biol* 2021;**70**: 18–26.
9. Campbell M, Watanabe T, Nakano K, *et al.* KSHV episomes reveal dynamic chromatin loop formation with domain-specific gene regulation. *Nat Commun* 2018;**9**(1):1–14.
10. Carriere M, Rabadan R. Topological data analysis of single-cell Hi-C contact maps. *Abel Symp* 2020;**15**:147–62.
11. Carstens S, Nilges M, Habeck M. Inferential structure determination of chromosomes from single-cell Hi-C data. *PLoS Comput Biol* 2016;**12**(12):e1005292.
12. Carstens S, Nilges M, Habeck M. Bayesian inference of chromatin structure ensembles from population-averaged contact data. *Proc Natl Acad Sci U S A* 2020;**117**(14): 7824–30.
13. Chang L-H, Ghosh S, Papale A, *et al.* A complex CTCF binding code defines TAD boundary structure and function. *bioRxiv* 2021. Preprint biorxiv:2021.04.15.440007.
14. Chen X, Ke Y, Wu K, *et al.* Key role for CTCF in establishing chromatin structure in human embryos. *Nature* 2019;**576**(7786):306–10.
15. Church DM, Schneider VA, Graves T, *et al.* Modernizing reference genome assemblies. *PLoS Biol* 2011;**9**(7):1–5.
16. Collombet S, Ranisavljevic N, Nagano T, *et al.* Parental-to-embryo switch of chromosome organization in early embryogenesis. *Nature* 2020;**580**(7801):142–6.
17. Cremer T, Cremer M. Chromosome territories. *Cold Spring Harb Perspect Biol* 2010;**2**(3):1–22.
18. Cusanovich DA, Hill AJ, Aghamirzaie D, *et al.* A single-cell atlas of *in vivo* mammalian chromatin accessibility. *Cell* 2018;**174**(5):1309–24.
19. Dame RT, Rashid FZM, Grainger DC. Chromosome organization in bacteria: mechanistic insights into genome structure and function. *Nat Rev Genet* 2020;**21**(4):227–42.
20. Davies JOJ, Telenius JM, McGowan SJ, *et al.* Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nat Methods* 2015;**13**(1):74–80.
21. de Wit E. TADs as the caller calls them. *J Mol Biol* 2020; **432**(3):638–42.
22. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting

chromatin interaction data. *Nat Rev Genet* 2013;**14**(6): 390–403.
23. Dekker J, Rippe K, Dekker M, *et al*. Capturing chromosome conformation. *Science* 2002;**295**(5558):1306–11.
24. Delaneau O, Zazhytska M, Borel C, *et al*. Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science* 2019;**364**(6439):1044–5.
25. Dostie J, Richmond TA, Arnaout RA, *et al*. Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* 2006;**16**(10):1299–309.
26. Du Z, Zheng H, Huang B, *et al*. Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature* 2017;**547**(7662):232–5.
27. Eraslan G, Avsec Z, Gagneur J, *et al*. Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet* 2019;**20**(7):389–403.
28. Flyamer IM, Gassler J, Imakaev M, *et al*. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* 2017;**544**(7648):110–4.
29. Flyamer IM, Illingworth RS, Bickmore WA. Coolpup.py: versatile pile-up analysis of Hi-C data. *Bioinformatics* 2020;**36**(10):2980–5.
30. Fudenberg G, Kelley DR, Pollard KS. Predicting 3D genome folding from DNA sequence with Akita. *Nat Methods* 2020;**17**(11):1111–7.
31. Fudenberg G, Imakaev M, Lu C, *et al*. Formation of chromosomal domains by loop extrusion. *Cell Rep* 2016;**15**(9): 2038–49.
32. Gassler J, Brandao HB, Imakaev M, *et al*. A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture. *EMBO J* 2017;**36**(24):3600–18.
33. Gavrilov AA, Zharikova AA, Galitsyna AA, *et al*. Studying RNA-DNA interactome by Red-C identifies noncoding RNAs associated with various chromatin types and reveals transcription dynamics. *Nucleic Acids Res* 2020;**48**(12): 6699–714.
34. Goel VY, Hansen AS. The macro and micro of chromosome conformation capture. *Wiley Interdiscip Rev Dev Biol* 2020;e395. https://pubmed.ncbi.nlm.nih.gov/32987449/.
35. Golov AK, Ulianov SV, Luzhin AV, *et al*. C-TALE, a new cost-effective method for targeted enrichment of Hi-C/3C-seq libraries. *Methods* 2020;**170**(2019):48–60.
36. Guarnera E, Tan ZW, Berezovsky IN. Three-dimensional chromatin ensemble reconstruction via stochastic embedding. *Structure* 2021;1–13.
37. Han C, Xie Q, Lin S. Are dropout imputation methods for scRNA-seq effective for scHi-C data? *Brief Bioinform* 2020; **22**:1–12.
38. He S, Wang L-H, Liu Y, *et al*. Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs. *Genome Biol* 2020;**21**(1):1–34.
39. Hirata Y, Oda A, Ohta K, *et al*. Three-dimensional reconstruction of single-cell chromosome structure using recurrence plots. *Sci Rep* 2016;**6**:3–8.
40. Horton CA, Alver BH, Park PJ. GiniQC: a measure for quantifying noise in single-cell Hi-C data. *Bioinformatics* 2020;**36**(9):2902–4.
41. Hug CB, Grimaldi AG, Kruse K, *et al*. Chromatin architecture emerges during zygotic genome activation independent of transcription article chromatin architecture emerges during zygotic genome activation independent of transcription. *Cell* 2017;**169**(2):216–28.
42. Imakaev M, Fudenberg G, McCord RP, *et al*. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* 2012;**9**(10): 999–1003.
43. Imakaev MV, Fudenberg G, Mirny LA. Modeling chromosomes: beyond pretty pictures. *FEBS Lett* 2015;**589**(20): 3031–6.
44. John HA, Birnstiel ML, Jones KW. RNA-DNA hybrids at the cytological level. *Nature* 1969;**223**(5206):582–7.
45. Kaaij LJT, van der Weide RH, Ketting RF, *et al*. Systemic loss and gain of chromatin architecture throughout zebrafish development. *Cell Rep* 2018;**24**(1):1–10.
46. Kalhor R, Tjong H, Jayathilaka N, *et al*. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol* 2012;**30**(1):90–8.
47. Ke Y, Xu Y, Chen X, *et al*. 3D chromatin structures of mature gametes and structural reprogramming during mammalian embryogenesis. *Cell* 2017;**170**(2):367–81.
48. Kharchenko PV, Alekseyenko AA, Schwartz YB, *et al*. Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature* 2011;**471**(7339):480–6.
49. Kim HJ, Yardimci GG, Bonora G, *et al*. Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell Hi-C data. *PLoS Comput Biol* 2020;**16**(9):1–19.
50. Kind J, Pagie L, De Vries SS, *et al*. Genome-wide maps of nuclear lamina interactions in single human cells. *Cell* 2015;**163**(1):134–47.
51. Lajoie BR, Dekker J, Kaplan N. The hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods* 2015;**72**:65–75.
52. Lando D, Basu S, Stevens TJ, *et al*. Combining fluorescence imaging with Hi-C to study 3D genome architecture of the same single cell. *Nat Protoc* 2018;**13**(5):1034–61.
53. Lando D, Stevens TJ, Basu S, *et al*. Calculation of 3D genome structures for comparison of chromosome conformation capture experiments with microscopy: an evaluation of single-cell Hi-C protocols. *Nucleus* 2018;**9**(1):190–201.
54. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**(4):357–9.
55. Lee DS, Luo C, Zhou J, *et al*. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nat Methods* 2019;**16**(10):999–1006.
56. Lerique S, Abitbol JL, Karsai M. Joint embedding of structure and features via graph convolutional networks. *Appl Netw Sci* 2020;**5**(1):1–24.
57. Li G, Liu Y, Zhang Y, *et al*. Joint profiling of DNA methylation and chromatin architecture in single cells. *Nat Methods* 2019;**16**(10):991–3.
58. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013. Preprint arXiv:1303.3997.
59. Li X, An Z, Zhang Z. Comparison of computational methods for 3D genome analysis at single-cell Hi-C level. *Methods* 2020;**181–182**:52–61.
60. Li X, Zhang Z. DeTOKI identifies and characterizes the dynamics of chromatin topologically associating domains in a single cell. *bioRxiv* 2021. Preprint biorxiv: 2021.02.23.432401.
61. Li X, Zhou B, Chen L, *et al*. GRID-seq reveals the global RNA-chromatin interactome. *Nat Biotechnol* 2017;**35**(10):940–50.
62. Li X, Feng F, Hongxi P, *et al*. A computational toolbox for analyzing single-cell Hi-C data. *PLoS Comput Biol* 2021;**17**(5):e1008978.
63. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet* 2015;**16**(6):321–32.
64. Lieberman-Aiden E, Van Berkum NL, Williams L, *et al*. Comprehensive mapping of long-range interactions

reveals folding principles of the human genome. *Science* 2009;**326**(5950):289–93.
65. Liu J, Lin D, Yardlmcl GG, *et al*. Unsupervised embedding of single-cell Hi-C data. *Bioinformatics* 2018;**34**(13):i96–i104.
66. Liu T, Zheng W. ScHiCNorm: a software package to eliminate systematic biases in single-cell Hi-C data. *Bioinformatics* 2018;**34**(6):1046–7.
67. Moon KR, Stanley JS, Burkhardt D, *et al*. Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Curr Opin Syst Biol* 2018;**7**:36–46.
68. Nagano T, Lubling Y, Stevens TJ, *et al*. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 2013;**502**(7469):59–64.
69. Nagano T, Lubling Y, Várnai C, *et al*. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* 2017;**547**(7661):61–7.
70. Nakamura R, Motai Y, Kumagai M, *et al*. CTCF looping is established during gastrulation in medaka embryos. *Genome Res* 2021;**31**(6):968–80.
71. Niu L, Shen W, Shi Z, *et al*. Systematic chromatin architecture analysis in xenopus tropicalis reveals conserved three-dimensional folding principles of vertebrate genomes. *bioRxiv* 2020. Preprint biorxiv:2020.04.02.021378.
72. Pardue ML, Gall JG. Molecular hybridization of radioactive DNA to the DNA of cytological preparations. *Proc Natl Acad Sci U S A* 1969;**64**(2):600–4.
73. Payne AC, Chiang ZD, Reginato PL, *et al*. *In situ* genome sequencing resolves DNA sequence and structure in intact biological samples. *Science* 2021;**371**(6532): eaay3446.
74. Perez-Rathke A, Sun Q, Wang B, *et al*. Chromatix: computing the functional landscape of many-body chromatin interactions in transcriptionally active loci from deconvolved single cells. *Genome Biol* 2020;**21**(1):1–17.
75. Polovnikov K, Gorsky A, Nechaev S, *et al*. Non-backtracking walks reveal compartments in sparse chromatin interaction networks. *Scientific Reports* 2020;**10**(1):1–1.
76. Ramani V, Deng X, Qiu R, *et al*. Massively multiplex single-cell Hi-C. *Nat Methods* 2017;**14**(3):263–6.
77. Rozenwald MB, Galitsyna AA, Sapunov GV, *et al*. A machine learning framework for the prediction of chromatin folding in *Drosophila* using epigenetic features. *PeerJ Comput Sci* 2020;**6**:2–21.
78. Samborskaia MD, Galitsyna A, Pletenev I, *et al*. Cumulative contact frequency of a chromatin region is an intrinsic property linked to its function. *PeerJ* 2020;**8**:1–15.
79. Schwessinger R, Gosden M, Downes D, *et al*. DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nat Methods* 2020;**17**(11):1118–24.
80. Soldatov R, Kaucka M, Kastriti ME, *et al*. Spatiotemporal structure of cell fate decisions in murine neural crest. *Science* 2019;**364**(6444):eaas9536.
81. Sridhar B, Rivas-Astroza M, Nguyen TC, *et al*. Systematic mapping of RNA-chromatin interactions *in vivo*. *Curr Biol* 2017;**27**(4):602–9.
82. Stevens TJ, Lando D, Basu S, *et al*. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* 2017;**544**(7648):59–64.
83. Sun Q, Perez-Rathke A, Czajkowsky DM, *et al*. High-resolution single-cell 3D-models of chromatin ensembles during *Drosophila* embryogenesis. *Nat Commun* 2021;**12**(1): 1–12.
84. Szabo Q, Bantignies F, Cavalli G. Principles of genome folding into topologically associating domains. *Sci Adv* 2019;**5**(4):eaaw1668.
85. Tan L, Ma W, Wu H, *et al*. Changes in genome architecture and transcriptional dynamics progress independently of sensory experience during post-natal brain development. *Cell* 2021;**184**(3):741–58.
86. Tan L, Xing D, Chang CH, *et al*. Three-dimensional genome structures of single diploid human cells. *Science* 2018;**361**(6405):924–8.
87. Tan L, Xing D, Daley N, *et al*. Three-dimensional genome structures of single sensory neurons in mouse visual and olfactory systems. *Nat Struct Mol Biol* 2019;**26**(4): 297–307.
88. Tavares-Cadete F, Norouzi D, Dekker B, *et al*. Multi-contact 3C reveals that the human genome during interphase is largely not entangled. *Nat Struct Mol Biol* 2020;**27**(12): 1105–14.
89. Tjong H, Li W, Kalhor R, *et al*. Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proc Natl Acad Sci U S A* 2016;**113**(12): E1663–72.
90. Tzouanacou E, Wegener A, Wymeersch FJ, *et al*. Redefining the progression of lineage segregations during mammalian embryogenesis by clonal analysis. *Dev Cell* 2009;**17**(3): 365–76.
91. Ulahannan N, Pendleton M, Deshpande A, *et al*. Nanopore sequencing of DNA concatemers reveals higher-order features of chromatin structure. *bioRxiv* 2019. Preprint biorxiv:833590.
92. Ulianov SV, Tachibana-Konwalski K, Razin SV. Single-cell Hi-C bridges microscopy and genome-wide sequencing approaches to study 3D chromatin organization. *Bioessays* 2017;**39**(10):1–8.
93. Ulianov SV, Zakharova VV, Galitsyna AA, *et al*. Order and stochasticity in the folding of individual drosophila genomes. *Nat Commun* 2021;**12**(1):1–17.
94. Vanhaeren T, Divina F, García-Torres M, *et al*. A comparative study of supervised machine learning algorithms for the prediction of long-range chromatin interactions. *Genes* 2020;**11**(9):1–17.
95. Whalen S, Truty RM, Pollard KS. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet* 2016;**48**(5):488–96.
96. Wolff J, Abdennur N, Backofen R, *et al*. Scool: a new data storage format for single-cell Hi-C data. *Bioinformatics* 2020;**37**(9):1337.
97. Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* 2011;**43**(11): 1059–65.
98. Zhang K, Hocker JD, Miller M, *et al*. A cell atlas of chromatin accessibility across 25 adult human tissues. *bioRxiv* 2021. Preprint biorxiv:2021.02.17.431699.
99. Zhang R, Zhou T, Ma J. Multiscale and integrative single-cell Hi-C analysis with Higashi. *bioRxiv* 2020. Preprint biorxiv:2020.12.13.422537.
100. Zhou J, Ma J, Chen Y, *et al*. Robust single-cell Hi-C clustering by convolution- and random-walk-based imputation. *Proc Natl Acad Sci U S A* 2019;**116**(28):14011–8.
101. Zhou S, Jiang W, Zhao Y, *et al*. Single-cell three-dimensional genome structures of rice gametes and unicellular zygotes. *Nat Plants* 2019;**5**(8):795–800.
102. Zhou T, Zhang R, Ma J. The 3D genome structure of single cells. *Annu Rev Biomed Data Sci* 2021;**4**:21–41.
103. Zhu H, Zheng W. SCL: a lattice-based approach to infer 3D chromosome structures from single-cell Hi-C data. *Bioinformatics* 2019;**35**(20):3981–8.

Chapter 9

Conclusion

- Hi-C-based readouts can be affected by intrinsic properties of genomic regions. In particular, the cumulative contact frequency (CCF) is associated with chromatin active state.
- Binding of insulator factor Chriz (Chromator) and histone modification H3K4me3 are the best predicting factors of TAD prominence in *Drosophila* based on population Hi-C, suggesting the importance of both histone-based and insulator-based mechanisms of structure formation in *Drosophila*.
- Lamina binding is not crucial for the formation of TADs and compartments in *Drosophila*, although its disruption causes redistribution of contacts.
- Single-cell Hi-C is a powerful and promising technology to study chromatin architecture, with computational analysis playing a crucial role in extracting biological meaning from its sparse readouts.
- Cell-to-cell variability in long-range contacts between active genomic regions in *Drosophila* is prominent, while the local scale of domains is highly conserved between individual cells. Stochastic processes significantly contribute to the formation of *Drosophila* 3D genome with two possible models ("sticky" nucleosomes and loop extrusion with barriers) explaining this outcome.

Bibliography

Kristin Abramo, Anne-Laure Valton, Sergey V Venev, Hakan Ozadam, A Nicole Fox, and Job Dekker. A chromosome folding intermediate at the condensin-to-cohesin transition during telophase. *Nature cell biology*, 21(11):1393–1402, 2019.

Chiara Anania and Darío G Lupiáñez. Order and disorder: abnormal 3D chromatin organization in human disease. *Briefings in Functional Genomics*, 19(2):128–138, 2020.

Edward J Banigan, Aafke A van den Berg, Hugo B Brandão, John F Marko, and Leonid A Mirny. Chromosome organization by one-sided and two-sided loop extrusion. *Elife*, 9:e53558, 2020.

Polina S Belokopytova, Miroslav A Nuriddinov, Evgeniy A Mozheiko, Daniil Fishman, and Veniamin Fishman. Quantitative prediction of enhancer–promoter interactions. *Genome Research*, 30(1):72–84, 2020.

Hugo B Brandão, Payel Paul, Aafke A van den Berg, David Z Rudner, Xindan Wang, and Leonid A Mirny. RNA polymerases as moving barriers to condensin loop extrusion. *Proceedings of the National Academy of Sciences*, 116(41):20489–20499, 2019.

by Open Chromosome Collective. distiller, pipeline for Hi-C data processing. URL https://github.com/open2c/distiller-nf.

Keerthi T Chathoth and Nicolae Radu Zabet. Chromatin architecture reorganization during neuronal cell differentiation in Drosophila genome. *Genome Research*, 29 (4):613–625, 2019.

Neville Cobbe and Margarete MS Heck. The evolution of SMC proteins: phylogenetic analysis and structural implications. *Molecular biology and evolution*, 21(2): 332–347, 2004.

UniProt Consortium. Uniprot: a hub for protein information. *Nucleic acids research*, 43(D1):D204–D212, 2015.

Lorenzo Costantino, Tsung-Han S Hsieh, Rebecca Lamothe, Xavier Darzacq, and Douglas Koshland. Cohesin residency determines chromatin loop patterns. *Elife*, 9:e59889, 2020.

Emily Crane, Qian Bian, Rachel Patton McCord, Bryan R Lajoie, Bayly S Wheeler, Edward J Ralston, Satoru Uzawa, Job Dekker, and Barbara J Meyer. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, 523(7559):240–244, 2015.

Job Dekker and Edith Heard. Structural and functional diversity of topologically associating domains. *FEBS letters*, 589(20):2877–2884, 2015.

Zhenhai Du, Hui Zheng, Yumiko K Kawamura, Ke Zhang, Johanna Gassler, Sean Powell, Qianhua Xu, Zili Lin, Kai Xu, Qian Zhou, et al. Polycomb group proteins regulate chromatin architecture in mouse oocytes and early embryos. *Molecular Cell*, 77(4):825–839, 2020.

Kyle P Eagen, Erez Lieberman Aiden, and Roger D Kornberg. Polycomb-mediated chromatin loops revealed by a subkilobase-resolution chromatin interaction map. *Proceedings of the National Academy of Sciences*, 114(33):8764–8769, 2017.

Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.

Fabian Erdel, Anne Rademacher, Rifka Vlijm, Jana Tunnermann, Lukas Frank, Robin Weinmann, Elisabeth Schweigert, Klaus Yserentant, Johan Hummert, Caroline Bauer, et al. Mouse heterochromatin adopts digital compaction states without showing hallmarks of HP1-driven liquid-liquid phase separation. *Molecular Cell*, 2020.

Guillaume J Filion, Joke G van Bemmel, Ulrich Braunschweig, Wendy Talhout, Jop Kind, Lucas D Ward, Wim Brugman, Inês J de Castro, Ron M Kerkhoven, Harmen J Bussemaker, et al. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell*, 143(2):212–224, 2010.

Ilya M Flyamer, Johanna Gassler, Maxim Imakaev, Hugo B Brandao, Sergey V Ulianov, Nezar Abdennur, Sergey V Razin, Leonid A Mirny, and Kikue Tachibana-Konwalski. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature*, 544(7648):110–114, 2017.

Geoff Fudenberg, David R. Kelley, and Katherine S. Pollard. Predicting 3D genome folding from DNA sequence with Akita. *Nature Methods*, 17(11):1111–1117, 2020.

Geoffrey Fudenberg, Maxim Imakaev, Carolyn Lu, Anton Goloborodko, Nezar Abdennur, and Leonid A. Mirny. Formation of chromosomal domains by loop extrusion. *Cell Reports*, 15(9):2038–2049, 2016.

Geoffrey Fudenberg, Nezar Abdennur, Maxim Imakaev, Anton Goloborodko, and Leonid A Mirny. Emerging evidence of chromosome folding by loop extrusion. In *Cold Spring Harbor symposia on quantitative biology*, volume 82, pages 45–55. Cold Spring Harbor Laboratory Press, 2017.

Johanna Gassler, Hugo B Brandao, Maxim Imakaev, Ilya M Flyamer, Sabrina Ladstatter, Wendy A Bickmore, Jan-Michael Peters, Leonid A Mirny, and Kikue

Tachibana. A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture. *The EMBO Journal*, 36(24):3600–3618, 2017.

Viraat Y Goel and Anders S Hansen. The macro and micro of chromosome conformation capture. *Wiley Interdisciplinary Reviews: Developmental Biology*, page e395, 2020.

Yosef Gruenbaum and Roland Foisner. Lamins: nuclear intermediate filament proteins with fundamental functions in nuclear mechanics and genome regulation. *Annual Review of Biochemistry*, 84:131–164, 2015.

Peter Heger, Birger Marin, Marek Bartkuhn, Einhard Schierenberg, and Thomas Wiehe. The chromatin insulator CTCF and the emergence of metazoan diversity. *Proceedings of the National Academy of Sciences*, 109(43):17507–17512, 2012.

Eimear E Holohan, Camilla Kwong, Boris Adryan, Marek Bartkuhn, Martin Herold, Rainer Renkawitz, Steven Russell, and Robert White. CTCF genomic binding sites in Drosophila and the organisation of the bithorax complex. *PLoS Genet*, 3 (7):e112, 2007.

Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R Lajoie, Job Dekker, and Leonid A Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods*, 9(10):999–1003, 2012.

Peter Kerpedjiev, Nezar Abdennur, Fritz Lekschas, Chuck McCallum, Kasper Dinkla, Hendrik Strobelt, Jacob M Luber, Scott B Ouellette, Alaleh Azhir, Nikhil Kumar, et al. HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome biology*, 19(1):1–12, 2018.

Peter V Kharchenko, Artyom A Alekseyenko, Yuri B Schwartz, Aki Minoda, Nicole C. Riddle, Jason Ernst, Peter J Sabo, Erica Larschan, Andrey A Gorchakov, Tingting Gu, Daniela Linder-Basso, Annette Plachetka, Gregory Shanower, Michael Y. Tolstorukov, Lovelace J Luquette, Ruibin Xi, Youngsook L Jung, Richard W. Park, Eric P Bishop, Theresa K. Canfield, Richard Sandstrom, Robert E. Thurman, David M. MacAlpine, John A. Stamatoyannopoulos, Manolis Kellis, Sarah C.R. Elgin, Mitzi I. Kuroda, Vincenzo Pirrotta, Gary H. Karpen, and Peter J. Park. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature*, 471(7339):480–486, 2011.

David Lando, Tim J. Stevens, Srinjan Basu, and Ernest D. Laue. Calculation of 3D genome structures for comparison of chromosome conformation capture experiments with microscopy: An evaluation of single-cell Hi-C protocols. *Nucleus*, 9 (1):190–201, 2018.

Yan Li, Judith HI Haarhuis, Ángela Sedeño Cacciatore, Roel Oldenkamp, Marjon S van Ruiten, Laureen Willems, Hans Teunissen, Kyle W Muir, Elzo de Wit, Benjamin D Rowland, et al. The structural basis for cohesin–CTCF-anchored loops. *Nature*, 578(7795):472–476, 2020.

Erez Lieberman-Aiden, Nynke L. Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

Nicholas E Matthews and Rob White. Chromatin architecture in the fly: Living without CTCF/cohesin loop extrusion? *BioEssays*, 41(9):1900048, 2019.

Leonid A Mirny, Maxim Imakaev, and Nezar Abdennur. Two major mechanisms of chromosome organization. *Current opinion in cell biology*, 58:142–152, 2019.

Charlotte Moretti, Isabelle Stévant, and Yad Ghavi-Helm. 3D genome organisation in Drosophila. *Briefings in Functional Genomics*, 19(2):92–100, 2020.

Takashi Nagano, Yaniv Lubling, Tim J. Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D. Laue, Amos Tanay, and Peter Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, 2013.

Takashi Nagano, Yaniv Lubling, Csilla Várnai, Carmel Dudley, Wing Leung, Yael Baran, and Netta Mendelson Cohen. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature Publishing Group*, 547(7661):61–67, 2017.

Natalia Naumova, Maxim Imakaev, Geoffrey Fudenberg, Ye Zhan, Bryan R Lajoie, Leonid A Mirny, and Job Dekker. Organization of the mitotic chromosome. *Science*, 342(6161):948–953, 2013.

Elphège P Nora, Anton Goloborodko, Anne-Laure Valton, Johan H Gibcus, Alec Uebersohn, Nezar Abdennur, Job Dekker, Leonid A Mirny, and Benoit G Bruneau. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell*, 169(5):930–944, 2017.

Johannes Nuebler, Geoffrey Fudenberg, Maxim Imakaev, Nezar Abdennur, and Leonid A Mirny. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proceedings of the National Academy of Sciences*, 115 (29):E6697–E6706, 2018.

Vijay Ramani, Xinxian Deng, Ruolan Qiu, Kevin L. Gunderson, Frank J. Steemers, Christine M. Disteche, William S. Noble, Zhijun Duan, and Jay Shendure. Massively multiplex single-cell Hi-C. *Nature Methods*, 14(3):263–266, 2017.

Fidel Ramírez, Vivek Bhardwaj, Laura Arrigoni, Kin Chung Lam, Björn A Grüning, José Villaveces, Bianca Habermann, Asifa Akhtar, and Thomas Manke. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature Communications*, 9(1):1–15, 2018.

Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

Suhas SP Rao, Su-Chen Huang, Brian Glenn St Hilaire, Jesse M Engreitz, Elizabeth M Perez, Kyong-Rim Kieffer-Kwon, Adrian L Sanborn, Sarah E Johnstone, Gavin D Bascom, Ivan D Bochkov, et al. Cohesin loss eliminates all loop domains. *Cell*, 171(2):305–320, 2017.

M Jordan Rowley, Michael H Nichols, Xiaowen Lyu, Masami Ando-Kuri, I Sarahi M Rivera, Karen Hermetz, Ping Wang, Yijun Ruan, and Victor G Corces. Evolutionarily conserved principles predict 3D chromatin organization. *Molecular Cell*, 67(5):837–852, 2017.

Michal B Rozenwald, **Aleksandra A Galitsyna**, Grigory V Sapunov, Ekaterina E Khrameeva, and Mikhail S Gelfand. A machine learning framework for the prediction of chromatin folding in Drosophila using epigenetic features. *PeerJ Computer Science*, 6:e307, 2020.

Matteo Vietri Rudan, Christopher Barrington, Stephen Henderson, Christina Ernst, Duncan T Odom, Amos Tanay, and Suzana Hadjur. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Reports*, 10(8):1297–1309, 2015.

Margarita D Samborskaia*, **Aleksandra Galitsyna***, Ilya Pletenev, Anna Trofimova, Andrey A Mironov, Mikhail S Gelfand, and Ekaterina E Khrameeva. Cumulative contact frequency of a chromatin region is an intrinsic property linked to its function. *PeerJ*, 8:e9566, 2020.

Tom Sexton, Eitan Yaffe, Ephraim Kenigsberg, Frédéric Bantignies, Benjamin Leblanc, Michael Hoichman, Hugues Parrinello, Amos Tanay, and Giacomo Cavalli. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, 148(3):458–472, 2012.

Quentin Szabo, Daniel Jost, Jia-Ming Chang, Diego I Cattoni, Giorgio L Papadopoulos, Boyan Bonev, Tom Sexton, Julian Gurgo, Caroline Jacquier, Marcelo Nollmann, et al. TADs are 3D structural units of higher-order chromosome organization in Drosophila. *Science Advances*, 4(2):eaar8082, 2018.

Longzhi Tan, Dong Xing, Chi Han Chang, Heng Li, and X. Sunney Xie. Three-dimensional genome structures of single diploid human cells. *Science*, 361(6405): 924–928, 2018.

Aleksandra A Galitsyna and Mikhail S Gelfand. Single-cell Hi-C data analysis: safety in numbers. *Briefings in Bioinformatics*, 08 2021. bbab316.

Sergey V Ulianov, Ekaterina E Khrameeva, Alexey A Gavrilov, Ilya M Flyamer, Pavel Kos, Elena A Mikhaleva, Aleksey A Penin, Maria D Logacheva, Maxim V Imakaev, Alexander Chertovich, et al. Active chromatin and transcription play

a key role in chromosome partitioning into topologically associating domains. *Genome Research*, 26(1):70–84, 2016.

Sergey V Ulianov, Kikue Tachibana-Konwalski, and Sergey V Razin. Single-cell Hi-C bridges microscopy and genome-wide sequencing approaches to study 3D chromatin organization. *BioEssays*, 39(10):1–8, 2017. ISSN 15211878.

Sergey V Ulianov, Semen A Doronin, Ekaterina E Khrameeva, Pavel I Kos, Artem V Luzhin, Sergei S Starikov, **Aleksandra A Galitsyna**, Valentina V Nenasheva, Artem A Ilyin, Ilya M Flyamer, et al. Nuclear lamina integrity is required for proper spatial organization of chromatin in Drosophila. *Nature Communications*, 10(1):1–11, 2019.

Sergey V Ulianov*, Vlada V Zakharova*, **Aleksandra A Galitsyna***, Pavel I Kos*, Kirill E Polovnikov, Ilya M Flyamer, Elena A Mikhaleva, Ekaterina E Khrameeva, Diego Germini, Mariya D Logacheva, et al. Order and stochasticity in the folding of individual Drosophila genomes. *Nature Communications*, 12(1):1–17, 2021.

Qi Wang, Qiu Sun, Daniel M Czajkowsky, and Zhifeng Shao. Sub-kb Hi-C in D. melanogaster reveals conserved characteristics of TADs between insect and mammalian cells. *Nature Communications*, 9(1):1–8, 2018.

Andrew M Waterhouse, James B Procter, David MA Martin, Michèle Clamp, and Geoffrey J Barton. Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191, 2009.

Stanislau Yatskevich, James Rhodes, and Kim Nasmyth. Organization of chromosomal DNA by SMC complexes. *Annual Review of Genetics*, 53:445–482, 2019.

Tianming Zhou, Ruochi Zhang, and Jian Ma. The 3D genome structure of single cells. *Annual Review of Biomedical Data Science*, 4, 2021.