2022-04-13

Detection of 3D Genome Folding at Multiple Scales

Betul Akgol-Oksuz
UMass Chan Medical School

Detection of 3D Genome Folding at Multiple Scales

A Dissertation Presented

by

Betul Akgol Oksuz

Submitted to the Faculty of the

University of Massachusetts Graduate School of Biomedical Sciences,

Worcester

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

April 13, 2022

Program in Systems Biology, Bioinformatics and Computational Biology

Detection of 3D Genome Folding at Multiple Scales

A Dissertation Presented

by

Betul Akgol Oksuz

This work was undertaken in the Graduate School of Biomedical Sciences

Interdisciplinary Graduate Program

Under the mentorship of

Job Dekker, Ph.D., Thesis Advisor

Zhiping Weng, Ph.D., Chair of Committee

Athma Pai, Ph.D., Member of Committee

Amir Mitchell, Ph.D., Member of Committee

Manuel Garber, Ph.D., Member of Committee

William Stafford Noble, Ph.D., External Member of Committee

Mary Ellen Lane, Ph.D.,

Dean of the Graduate School of Biomedical Sciences

April 13, 2022

Dedication

This thesis is dedicated to all women who were not given credit they deserved for their achievements. All women who fought hard to be seen, acknowledged and appreciated. I dedicate this thesis to you as a thank you for making it possible to do science freely today.

Acknowledgements

PhD has been a long and challenging journey. I was extremely lucky to receive a lot of support from my mentor, colleagues, family and friends.

I would like to thank my thesis advisor, Job Dekker, PhD, for his extraordinary mentorship that makes this thesis possible. Job inspired me to do science in passion and broaden my perspective interpreting biological results. He always gave me enough confidence to pursue my ideas and gave me direction to go further to the right direction. Job's mentorship is invaluable but he is also an amazing person I look up to. Last couple years were a great opportunity for me to get to know him and work with him.

My committee members always encouraged me to think about different aspects of my project and about the next move, so thank you Zhiping Weng, Amir Mitchell, Manuel Garber and Athma Pai. I would also like to thank William Stafford Noble for taking the time to attend my thesis defense and giving me amazing feedback on my thesis.

I had opportunity to work with amazing scientist in Dekker lab listed below : Nicki Fox, Liyan Yang, Ye Zhan, Johan Gibcus, Sergey Venev, Erica Hildebrand, Yu Liu (Sunny), Anne-Laure Valton, Allana Schooley, Bastiaan Dekker, Denis Lafontaine, Snehal Sambare, Davood Norouzi, George Spracklin, Marlies oomen, Kristin Abramo, Houda Belaghzal, Hakan Ozadam, Ankita Nand and Filipe Tavares-Cadete. Although it was always zoom but it was great to get to know you; Xiangru Hao, Jiangyuan Liu, Nicola Minchell. I loved our Dekker lab group lunches. Special thanks for these people who

made it possible for me to catch up with all the events and inside jokes that are happening inside the lab. Thank you Denis Lafontaine for the Hi-C 3.0 figure. Very special thanks to Anne-Laure Valton, Allana Shooley and Johan Gibcus for taking the time to edit my thesis.

At the beginning of my PhD I worked with Hakan Ozadam, who was a great mentor to me and helped me adapt to the lab quickly. Hakan is a multidisciplinary person which influenced me to think outside the box. It was great to have our mini computational group meetings to adjust to the lab and discuss papers and analysis with Sergey Venev, Ankita Nand, Hakan Ozadam and Filipe Tavares-Cadete. Later in my PhD I was mentored by Sergey Venev, who is not only a great mentor but also an amazing friend. Sergey taught me how to simplify the most complicated things by synthesizing everything step by step. Sometimes it took forever to solve a problem but at the end of the day we would know everything about the problem and all possible outcomes. I will always value the problem solving approaches I have learned from Sergey.

I will definitely miss our Monday morning meetings. Collaborating with Liyan was an amazing experience for me. Liyan is very efficient, peaceful and calm even in very stressful times. Johan always finds an angle in the analysis that I would never think of so thank you both. I will always be thankful for collaborating and discussing data and future directions with amazing scientist; Samir Abraham, Leonid Mirny, Nils Krietenstein, Feng Yue, Xiatao Wang, Jiangyuan Liu, Denis Lafontaine, Snehal Sambare and Nezar Abdennur.

To my colleagues at the Systems Biology Department who made my PhD journey more fun and exciting, thank you. I will miss our Science on tab and the happy hour after SoT.

My classmates and friends, the people who I struggled with through the Qualification exam, discussing papers and partying together, thank you for being there for me and sharing my feelings. Special thanks for Brent Horowitz, Sarah Anderson, Sneha Suresh, Serkan Sayin, Nanditha Uma Naresh, Salome funes, Debanjan Goswany, Kristopher Holloway, Sunil Guharajan and Zeynep Mirza.

I would like to thank my colleagues at NYU Langone Medical School for helping me to start my carrier in computational biology which led me to pursue to PhD.

I am very lucky to be surrounded by amazing friends. First thank you goes to my dearest friend Sureyya Karadag. We met working for an exam before college and since then our life had ups and downs but we always stayed close to make each other feel loved and special. Ezgi Guzel and Busra Uzun, thank you for being awesome friends during college and after. Our special connection makes me feel warm and happy. I have a unique friendship with Hakan Sarigul, Mehmet Ali Altunel and Cemil Hurriyetoglu. Due to distance and country borders we see each other ones or twice a year but our conversations are always sincere and fun.

I would like to thank my special friends who made New York more exciting and fun place to live; Ahu Demir, Canan Kasikara, Mehmet Emin Basbug,

Yetis Gultekin, Havva Ortabozkoyun. I miss our non-ending karaoke, dance parties and board games.

I gained exceptional friendships at UMass; Houda Belaghzal and Mercedeh Javanabkht, Anne-Laure Valton, Bastiaan Dekker and Allana Shooley thank you for being amazingly supportive when I needed you. No matter where our life takes us you will always be special for me.

I would also like to thank my amazing friends who I spend my weekends with, Ceren Gokbulut-Barut, Onur Barut, Linda Barut, Onur Hasturk, Serkan Sayin and Deniz Ozata. Unlimited barbeque parties by the chef Deniz Ozata and game nights made my PhD and Covid days fun and adventurous.

My mom and dad, I want to thank you for always supporting me pursuing my dreams. I have learned all the kindness from you which helped to succeed and be happy not only in my social life and also professionally. My beloved family; Ali, Hasan, Mehdiye, Nurettin, Husne, Sibel, Ipek, Onur, Iliya, Ali Umut, Yunus Emre, Nil and Abdullah Kemal, I am very happy to have your support along the way. It is great to have a big family because someone will always be available for you to continue 24 hour love and support. I would also like to thank my extended family Oksuz family, I love having you in my life.

My love Ozgur Oksuz, thank you for being an amazing partner, friend, colleague and more. Having you and your support makes everything fun and every problem easier to handle. Finally, my source of happiness, my son Enki Oksuz, thank you for the most beautiful smile, warm hugs and making my

everyday full of love and excitement. I cannot wait watching you growing up with all the love and support you need.

Abstract

Understanding 3D genome structure is crucial to learn how chromatin folds and how genes are regulated through the spatial organization of regulatory elements. Various technologies have been developed to investigate genome architecture. These technologies include ligation-based 3C Methodologies such as Hi-C and Micro-C, ligation-based pull-down methods like Proximity Ligation-Assisted ChIP-seq (PLAC Seq) and Paired-end tag sequencing (ChIA PET), and ligation-free methods like Split-Pool Recognition of Interactions by Tag Extension (SPRITE) and Genome Architecture Mapping (GAM). Although these technologies have provided great insight into chromatin organization, a systematic evaluation of these technologies is lacking. Among these technologies, Hi-C has been one of the most widely used methods to map genome-wide chromatin interactions for over a decade. To understand how the choice of experimental parameters determines the ability to detect and quantify the features of chromosome folding, we have first systematically evaluated two critical parameters in the Hi-C protocol: cross-linking and digestion of chromatin. We found that different protocols capture distinct 3D genome features with different efficiencies depending on the cell type (Chapter 2). Use of the updated Hi-C protocol with new parameters, which we call Hi-C 3.0, was subsequently evaluated and found to provide the best loop detection compared to all previous Hi-C protocols as well as better compartment quantification compared to Micro-C (Chapter 3). Finally, to understand how the aforementioned technologies (Hi-C, Micro-C, PLAC-Seq, ChIA-PET, SPRITE, GAM) that measure 3D organization could provide a

comprehensive understanding of the genome structure, we have performed a comparison of these technologies. We found that each of these methods captures different aspects of the chromatin folding (Chapter 4). Collectively, these studies suggest that improving the 3D methodologies and integrative analyses of these methods will reveal unprecedented details of the genome structure and function.

List of tables

List of figures

Copyright Material

retain the right to reproduce the work in my thesis dissertation. All other work presented in this thesis is currently unpublished.

Akgol Oksuz, B., Yang, L., Abraham, S. *et al.* Systematic evaluation of chromosome conformation capture assays. *Nat Methods* 18, 1046–1055 (2021). https://doi.org/10.1038/s41592-021-01248-7

Chapter I: Introduction

3D genome structure

Assembly of the ~3.3 billion nucleotides of the human genome was one of the greatest achievements of the last century, which unexpectedly raised many more questions than answers about how the genome works (Lander et al. 2001). Genomic studies have shown that 98.5% of the human genome does not encode for protein-coding genes, also referred to as non-coding regions (Schuler et al. 1996; Kellis et al. 2014). These noncoding regions are thought to help regulate the activity of the remaining 1.5% of the human genome that encodes protein-coding genes and contributes to the stability of the genome (Heintzman and Ren 2009). The Encyclopedia of DNA Elements (ENCODE) project was able to map 80% of the human genome to identify functional regulatory elements, such as enhancers (Consortium, Moore, et al. 2020; Consortium, Snyder, et al. 2020). While these regulatory elements can be located relatively far (>100 kb away) from their targets on the linear genome, they can be spatially organized in three-dimensional (3D) space to localize in close proximity to their targets (Maurano et al. 2012; Dekker et al. 2017; Dekker, Marti-Renom, and Mirny 2013; Dixon et al. 2018). GWAS studies have identified hundreds of distant gene targets that play a role in diseases (Dixon et al. 2018; Maurano et al. 2012; Boltsis et al. 2021; Lupianez et al. 2015). For example, in Amyotrophic lateral sclerosis the SYNGAP1 gene promoter interacts with a distal promoter that is 411kb away (Maurano et al. 2012). Similarly, a Breast cancer-related tumor suppressor gene TACC2 is

interacting with a promoter that is 411 kb away (Maurano et al. science.2012). Therefore, to understand the principles of gene regulation in normal and disease states, it is critical to uncover 3D genome organization at high resolution.

There has been tremendous progress in developing technologies to determine 3D genome organization at different resolutions and scales, recently reviewed in Goel et al. (Goel and Hansen 2021). All of these technologies were pioneered by chromosome conformation capture (3C) based methods, which have evolved from detecting pairwise spatial interactions (3C) to capturing all possible spatial interactions occurring in the genome at high resolution (Hi-C). 3C-based methods were derived from the fundamental 3C experiment developed in 2002 (Dekker et al. 2002) which identifies pairwise genomic interactions with Polymerase Chain Reaction (PCR). Later, more high-throughput methods were developed to capture interactions of a known region with unknown multiple loci, named Circular Chromosome Conformation Capture (4C) (Göndör, Rougier, and Ohlsson 2008; Zhao et al. 2006) and to capture interactions among multiple selected loci, called 5C (Dostie et al. 2006; van Berkum and Dekker 2009). Finally, the development of Hi-C enabled the efficient capture of genome-wide high-frequency spatial interactions at unprecedented resolution (Lieberman-Aiden et al. 2009) and provided important insights into genome organization and function. Below, I will introduce key technologies that detect various features of the genomic interactions with a particular focus on Hi-C.

- **Hi-C**

Hi-C is the most commonly used 3C-based method that identifies genome-wide contacts in an unbiased manner (Lieberman-Aiden et al. 2009) (Figure 1.1 a). The Hi-C protocol proceeds as follows: DNA-protein interactions are crosslinked in intact nuclei, followed by digestion of the crosslinked chromatin with restriction enzymes, such as HindIII or DpnII, which create 5' overhangs that are subsequently filled with biotinylated deoxyribonucleotides (Lieberman-Aiden et al. 2009; Rao et al. 2014; Belaghzal, Dekker, and Gibcus 2017; Gibcus and Dekker 2013a; Belton et al. 2012). Next, chromatin fragments are re-ligated such that only chromatin in close proximity forms a ligation product and is sonicated to give rise to fragments of interacting chromatin elements. These fragments are enriched using the biotin pulldown and identified by paired-end sequencing. The sequencing reads of these chimeric fragments are then aligned to the reference genome to identify interacting pairs generated by proximity ligation.

Analysis: Sequencing reads from the Hi-C experiment are aligned to the reference genome using bwa-mem with flag-SP or bowtie2 (Lajoie, Dekker, and Kaplan 2015; Pal, Forcato, and Ferrari 2019; Ay and Noble 2015). Hi-C reads can be parsed and pre-processed using Pairtools or SAMtools (Danecek et al. 2021) (https://github.com/open2c/pairtools). The chimeric reads that map to the same restriction fragment from both sides and any PCR duplicates are removed (Lieberman-Aiden et al. 2009; Lajoie, Dekker, and Kaplan 2015). Depending on the preference, low-quality reads can be filtered from each side of the mapped chimeric read. After filtering as required, reads

are binned into multiple resolutions to create matrices of all possible interactions for the whole genome. The interaction matrix of the whole genome needs tremendous space and memory to visualize, necessitating the development of 2D data storage solutions that enable efficient compression and access. Cool files are HDF5 containers developed to store contact matrices (Abdennur and Mirny 2020) and .hic files are compressed binary files that store contact matrices (Rao et al. 2014; Robinson et al. 2018). Multi-resolution cooler files can be created using cooler and .hic files can be created using juicertools. The matrices of Hi-C maps have inherent biases such as mappability, visibility in the genome, restriction site density,..etc (Imakaev et al. 2012; Yaffe and Tanay 2011).These biases can be corrected in a stepwise manner using a method called ICE (Iterative Correction and Eigenvector Decomposition) (Imakaev et al. 2012). This genome balancing technique corrects biases by assuming every bin in the genome should have an equal representation in the genome. Balancing provides an equal coverage profile and grants unbiased comparisons within and between Hi-C contact maps. The final Hi-C contact matrix, which includes both intra- and inter-chromosomal interactions, is used to reveal key structures that are observed in the genome such as nuclear compartments, loop domains or Topologically Associated Domains (TADs) and loops, as well as to gain insights into genome structure in various experimental settings discussed in the results section (Gibcus and Dekker 2013a; Schmitt, Hu, and Ren 2016; Lieberman-Aiden et al. 2009)

Detection and quantification of the chromatin structures in 3D

- **Compartments**

Studies using the Hi-C approach have shown that the 3D genome is organized into several layers (Lieberman-Aiden et al. 2009; Gibcus and Dekker 2013b; Dekker, Marti-Renom, and Mirny 2013) (Figure 1.2). One layer can be visualized in 2D heatmaps as a checkerboard pattern which is called compartments (Figure 1.2 a). Compartments, several megabases in size, are the largest structures that are detected by Hi-C that occur within and between chromosomes(Lieberman-Aiden et al. 2009). Compartments are identified as active A and inactive B compartments showing the segregation between euchromatin and heterochromatin, respectively. To define compartments from Hi-C data, Eigenvector Decomposition, which is a dimensionality reduction method, is used to explain the patterns and the variation of the interaction matrix. Eigenvector Decomposition is applied to the interaction matrix and generally, the eigenvector with the highest variations aligns with compartment-like structures. To determine A and B compartments, a correlation of the first eigenvector with gene coverage or GC content is used (Lajoie, Dekker, and Kaplan 2015). Gene dense or GC-rich regions correlate with positive values (A compartment) of the first eigenvector and the gene poor and GC poor regions correlate with negative values (B compartment) of the eigenvector. It is crucial to balance the interaction matrix and normalize it with its expected to remove the distance dependence effect before the Eigenvector decomposition. Finally, a one-dimensional track that represents

chromatin compartments was detected from the interaction matrix.

The mechanism that causes segregation between A (euchromatin) and B (heterochromatin) compartments is not fully understood. Recently, phase separation has been proposed to mediate such segregation (Strom et al. 2017; Erdel and Rippe 2018). Heterochromatin protein 1 alpha (HP1a), which is a key protein involved in the formation of constitutive heterochromatin together with its binding partner H3K9me2/3, has been shown to form liquid-like droplets in *Drosophila* and mammalian cells (Erdel et al. 2020; Larson et al. 2017). However, there are conflicting studies about the nature of the heterochromatin assembly being mediated by polymer-polymer or liquid-liquid phase separation (Erdel et al. 2020; Hildebrand and Dekker 2020). Another study using a combination of Hi-C, microscopy, and polymer simulations in inverted nuclei of rods in nocturnal mammals have suggested that heterochromatic regions have an attraction to interact with each other and this attraction drives the heterochromatin and euchromatin formation that is crucial for phase separation (Falk et al. 2019). This system have provided great insights into the segregation of compartments because inverted nuclei have euchromatin at the nuclear periphery and heterochromatin in the nuclear interior unlike other mammalian cells, which have euchromatin in the nuclear interior and heterochromatin at the nuclear periphery. Despite these observations, direct experimental evidence supporting a link between chromatin segregation and phase separation is lacking (McSwiggen, Hansen, et al. 2019; McSwiggen, Mir, et al. 2019). It will be important to clarify these debates and determine the relationship between Hi-C compartmentalization

and phase transition.

High coverage Hi-C experiments with deeper sequencing revealed that compartments can further be divided into subcompartments, which have shown to correlate with specific chromatin landscape. In one study, subcompartments that divide A compartment into A1 and A2 and the B compartment into B1, B2, and B3 have been identified in a human lymphoblastoid (GM12878) cell line (Rao et al. 2014). Subcompartments of A1 and A2 positively correlate with gene dense regions and active histone marks such as H3K36me3, H3K79me2, H3K27ac and H3K4me. However overall A1 has stronger correlations than A2. Subcompartments of B1 correlate with H3K27me3, B2 correlate with Lamin A/C and NADs and finally B3 correlates with Lamin A/C but not NADs. In another study, subcompartments that divide the B compartment into B0, B1, and B2 were identified in the human colorectal carcinoma cell line (HCT116). B0, B1 and B2 subcompartments have been shown to have a positive correlation with constitutive repressive chromatin marks such as H3K9me3, HP1a and HP1b (Spracklin et al. 2021). B2 has enrichment for H3K9me3, Hp1 α and HP1β but depletion of H3K27me3, whereas B1 has enrichment for H3K9me2 and H3K27me3. B0 has enrichment for H3K9me2 but lower than the B1 and enriched for H2A.Z (Spracklin et al. 2021). Conversely, A1 and A2 are depleted at the nuclear lamina and at the nucleolus associated domains (NADs) (Rao et al. 2014). A2 does not have well-defined compartments and has lower transcriptional activity compared to A1. Subcompartments might differ in different cell types. Active A compartments show faster replication

timing compared to inactive B subcompartments (Ryba et al. 2010). Taken together, these results suggest that compartments can be segregated into cell-type specific subcompartments harboring distinct chromatin features.

- **Topologically Associating Domains**

Topologically Associating Domains (TADs) are identified as domains that have high interaction frequency within square-like blocks along with the diagonal and low interaction frequency between blocks (Nora et al. 2012; Lieberman-Aiden et al. 2009; Sexton et al. 2012; Crane et al. 2015). TADs are generally smaller than compartments and are sub-megabase-sized domains. Identification of TAD boundaries is not straightforward because the majority of the time these domains have hierarchical structures lying on top of each other (Figure 1.2 b). Due to the hierarchy of the TADs there is no clear separation of TAD boundaries. To identify TAD boundaries, methods called directionality index or insulation score are commonly calculated (Dixon et al. 2012; Nora et al. 2012; Crane et al. 2015). To calculate the directionality index, average upstream and downstream interactions are determined for a given bin with a given window size. Then, the difference between them is transformed into chi-squared statistics with the values being called directionality index and the Hidden Markov Model is used to connect the boundaries with each other. To calculate the insulation score, the average interaction of squares in a sliding window along the Hi-C interaction matrix diagonal is determined and the local minimas of the insulation scores are taken to identify TAD boundaries. Insulation score is calculated for bin in the genome. The distribution of

genome-wide insulation scores is bimodal where weak and strong boundaries exist; weak boundaries being hardly detectable and strong boundaries showing sharp transitions between TADs. TADs have been proposed to form as a consequence of a dynamic loop extrusion mechanism (Dekker 2014; Mirny, Imakaev, and Abdennur 2019; Fudenberg et al. 2017; Fudenberg et al. 2016). In this model, cis-acting loop extruding factors like cohesin pulls the chromatin to form a loop-like structure. As the pulling continues the loop gets larger until it stalls by boundary elements, such as CTCF, which has footprints at the TAD boundaries. Two scenarios explain how does TADs form as a result of loop extrusion. First, loop extrusion brings genomic regions together to create loops, adding interactions in the neighboring regions as identified by TADs. CTCF blocks mixing between regions during the loop extrusion which creates insulation between TADs (Wutz et al. 2017; Nora et al. 2017; Zuin et al. 2014). Second, since loop extrusion is not a stable structure, the continued movement of the chromatin creates TADs (Hansen et al. 2018). This dynamic movement creates TAD-like structures rather than a strong dot.

TADs bring enhancers and their targets in close physical proximity, thus contributing to gene regulation (Maurano et al. 2012; Lupianez et al. 2015; Gong et al. 2021; Tang et al. 2015; Bonev and Cavalli 2016). The organization of TADs has also been shown to be important during development (Bonev et al. 2017; Pekowska et al. 2018). Mapping human and mouse distal regulatory elements showed that many of these elements, which include early developmental enhancers, locate hundreds of kilobases away from their targets (Rada-Iglesias et al. 2011). Despite being separated by long

distance in a linear genome these regulatory elements have been shown to be located in physical proximity to their targets, and reside at the same TAD (Symmons et al. 2014; Consortium, Moore, et al. 2020; Heintzman and Ren 2009; Schoenfelder and Fraser 2019). Importantly, disruption of TAD boundaries in mouse limb tissue and fibroblasts collected from patients with limb malformations resulted in misarranged interactions and aberrant rewiring of gene regulation (Lupianez et al. 2015). These results show the functional significance of TADs in regulating gene expression.

- **Chromatin Loops**

Chromatin loops are the finest detectable structures using Hi-C. Loop anchors can be enriched for certain architectural proteins, promoters, and enhancer marks, and they may provide insights into the expression levels of nearby genes (Rao et al. 2014; Nora et al. 2017; Tang et al. 2015; Ramirez et al. 2018). Loops are observed as contacts between two regions with high interaction probability compared to their neighboring regions and they look like dots in the Hi-C interaction map (Figure 1.2 c). The most common loop identification method is called HICCUPs (Rao et al. 2014). This tool first normalizes the iced Hi-C matrix to its global expected and detects pixels that have high interaction frequency. For each of these regions, the local background is calculated by comparing the high-intensity interacting pixel with its neighboring pixels, upper left, upper right, lower left, lower right, and the stripes between these corners. The interactions of the mid pixel which is a loop candidate are normalized to its donut-like background. If it is statistically

stronger than its background, it is considered as a loop. HICCUPs is considered too strict compared to other loop callers. Alternative ways of calling loops include techniques that use only global background which is the expected interaction frequency for a given distance. Some methods filter contacts using both global expected and local backgrounds and some tools use image analysis to detect dot-like structures in the Hi-C interaction heatmap (Roayaei Ardakany et al. 2020; Ay, Bailey, and Noble 2014).

SMC family proteins cohesin and condensin are identified as loop extruding factors (Fudenberg et al. 2017; Fudenberg et al. 2016; Ganji et al. 2018; Bauer et al. 2021). During interphase, cohesin has been shown to serve as a loop extruder. The Cohesin complex has two SMC proteins, SMC1 and SMC2, and two other proteins SCC1(also known as RAD21) and SCC3 (known as STAG1/SA1 and STAG2/SA2 in mammalian cells) (Nasmyth and Haering 2009). Degradation of the cohesin subunit RAD21 using the Auxin Inducible degron system caused a complete loss of loops (Rao et al. 2017). In addition, depletion of the cohesin loading factor NIPBL resulted in a complete loss of loops and TADs (Schwarzer et al. 2017). Other factors may also contribute to loop formation by regulating cohesin. For example, WAPL has shown to be essential for releasing cohesin from DNA in interphase cells, which results in less compact chromatin (Busslinger et al. 2017; Tedeschi et al. 2013). Additionally PDS5A and PDS5B have been shown to contribute chromatin compaction and cohesin localization (Wutz et al. 2017).

CTCF, a highly conserved zinc finger protein, has been shown to block

cohesin-dependent loop extrusion thereby mediating loop anchors at TAD boundaries (Nora et al. 2017). Depletion of CTCF results in reduced loops but has little effect on the global gene expression (Nora et al. 2017; Tedeschi et al. 2013). Proteins that co-occupy regions on chromatin with CTCF have also been shown to contribute to loop formation. For example, YY1, which occupies promoter and enhancer regions, mediates promoter-enhancer looping interactions (Weintraub et al. 2017). ZNF143 has also been shown to mediate CTCF-bound promoter-enhancer loops (Wen et al. 2018; Zhou et al. 2021).

CTCF is considered to mark looping interactions between active enhancers and promoters independently of cohesin enrichment. (Kubo et al. 2021; Hyle et al. 2019; Braccioli and de Wit 2019). On the other hand, not all loops contain CTCF, suggesting the existence of CTCF-independent loops (Valton et al. 2021) . A subset of loops was identified with enrichment for active promoter (H3K4me3) and active enhancer (H3K27ac) signal. We found that these loops are smaller and weaker compared to CTCF loops (Akgol Oksuz et al. 2021).

Micro-C: a variation of the Hi-C protocol that improves resolution of small scale structures

Micro-C is a variation of the Hi-C protocol developed to improve resolution and signal to noise ratio (Hsieh et al. 2015; Hsieh et al. 2016; Krietenstein et al. 2020). It was initially applied in yeast but then adapted for other cell types including mammalian cells. The general workflow of Micro-C

consists of similar steps to Hi-C. In brief, cultured cells are cross-linked with Formaldehyde (FA) and then chromatin is digested with MNase for fragmentation and with an exonuclease to remove free DNA. Chromatin fragments are then biotinylated and subjected to proximity ligation. Finally, the crosslinking is reversed and the nucleosome level DNA fragments are enriched and identified by paired-end sequencing. The Micro-C protocol was further improved by cross-linking the chromatin with additional cross-linkers such as disuccinimidyl glutarate (DSG) or ethylene glycol bis(succinimidyl succinate) (EGS), in addition to FA (Hsieh et al. 2016). These modifications improved signal to noise ratio of the interaction maps including centromere-centromere interactions (Hsieh et al. 2016). Optimization of the Micro-C protocol for mammalian cells (Human embryonic stem cells(H1-hESC) and Human Foreskin Fibroblasts clone 6, provided genome-wide interaction maps at nucleosomal level resolution and allowed identification of number of loops that have not been detected by the conventional Hi-C protocols. Micro-C has identified 18478 and 22966 more loops than Hi-C in H1-hESC and HFFc6, respectively (Krietenstein et al. 2020).

***Analysis*:** The analysis pipeline of Micro-C is identical to Hi-C. Please see the Hi-C analysis above for details.

Parameter selection for 3C-based experiments

The Hi-C protocol has evolved over the years to improve the identification of genomic interactions at high resolution (Lieberman-Aiden et al. 2009; Belton et al. 2012; Rao et al. 2014; Belaghzal, Dekker, and Gibcus

2017). Performing in situ Hi-C over diluted Hi-C was one of the most important improvements over the years (Rao et al. 2014). Fragmentation and crosslinking reagent selection were the two main parameters that have been modified for protocol optimization.

- **Fragmentation with restriction enzymes**

To fragment chromatin, the initial protocols have used restriction enzymes such as HindIII, which produces relatively large fragments of several kb (6) (Lieberman-Aiden et al. 2009). Later, 4 nucleotide cutters such as DpnII or MboI, which produce smaller fragments than 2 kb, have been used for capturing genomic interactions at around 2 kilobase resolution (Rao et al. 2014). These restriction enzymes have also been used in combination, which led to further improvements in the detection of small-scale structures such as loops. Using MNase instead of restriction enzymes in the Micro-C protocol led to fragmentation of chromatin at nucleosomal level and generation of the highest resolution maps in the detection of small-scale loop structures. These results have shown that chromatin fragmentation is an important determinant of the resolution of chromatin interaction maps.

- **Cross-linking**

To cross-link chromatin, the initial protocols have used formaldehyde reagent. However, recent studies have shown an improved signal-to-noise ratio using additional crosslinkers (DSG and EGS) that contain a longer spacer between reactive groups in various protocols measuring genomic contacts, such as Micro-C, ChIA-PET, and SPRITE (Hsieh et al. 2016;

Fullwood et al. 2009; Quinodoz et al. 2018). Thus, the selection of crosslinking chemistry is key in the detection of chromatin interactions.

Other methods that are used to measure 3D structure

Compartments, TADs and loops can be detected by other methods that measure 3D genome folding such as ligation-based ChIA-PET and PLAC-Seq, ligation-free SPRITE and GAM (Fullwood et al. 2009; Fang et al. 2016; Quinodoz et al. 2018; Beagrie et al. 2017). These methods range from predicting interactions between targeted regions (ChIA-PET and PLAC-Seq) to more unbiased detection of genome-wide contacts (SPRITE, and GAM). Below, I will introduce the experimental setup and analysis of these protocols.

- **Chromatin Interaction Analysis using Paired-End Tag sequencing (ChIA-PET)**

ChIA-PET is a method that enables the study of chromatin interactions bound by a specific protein (Fullwood et al. 2009). It incorporates chromatin immunoprecipitation (ChIP), proximity ligation, and high-throughput sequencing. The protocol starts by cross-linking the chromatin with FA followed by fragmenting chromatin by sonication. The crosslinked protein-DNA fragments are then enriched by ChIP using an antibody against the protein of interest, and subjected to proximity ligation. The resulting connected fragments can be identified by Paired-End Tag (PET) sequencing. ChIA-PET reads are then mapped to reference genomes to identify the interactions between genomic regions (Fullwood et al. 2010; Fullwood et al. 2009).

Analysis: For the quality check, DNA linkers that have unreadable barcodes were filtered and PETs are sorted such that the data have chimeric and non-chimeric PETs (Li et al. 2010). Then the chimeric PETs are aligned to the reference genome using a package called Batman, which uses Burrows-Wheeler-transform-based technique with one mismatch allowance (Li and Durbin 2009). PETs that are mapped in 1bp distance are merged and assigned to the same PET. Self-ligated PETs show the putative binding sites, which are also defined as self-circled ligations. The inter-ligated PETs provide information about the two different DNA fragments which also can be classified as intra-chromosomal inter-ligation PETs, inter-chromosomal inter-ligation PETs and different orientation ligation PETs. Different orientations show the wrong orientation of PETs and are removed from the data. PET distribution is used to decide on a cutoff to filter the self-ligated PETs. To test if the ligation products occur more frequently than a random event, a model that assumes intra-chromosomal inter-ligated PETs have an equal probability of interaction with any DNA fragments is used. This model is used to set an expected interaction frequency which is then used for comparison with the observed interaction frequency and to determine whether the difference is significant. Non-specific mappings such as translocations are flagged. The interaction frequency matrix is calculated using the number of interactions that occur within and among intra-chromosomal inter-ligation and inter-chromosomal inter-ligation products. Interactions between inter-ligated PETs are also confirmed with a model that predicts equal interaction probability between ligation products (Fullwood et al. 2010).

- **Proximity ligation-assisted ChIP-seq (PLAC-Seq)**

Similar to ChIA-PET, PLAC-Seq is used to detect and quantify chromatin contacts of genomic regions bound by specific proteins or histone marks, such as H3K4me3, H3K27ac or RNA Pol II (Fang et al. 2016). The immunoprecipitation step differs between PLAC-Seq and ChIA-PET; in PLAC-Seq immunoprecipitation is done after the proximity ligation however in ChIA-PET the immunoprecipitation was done before the proximity ligation. PLAC-Seq incorporates in situ Hi-C and ChIP protocols with high-throughput sequencing. The beginning of the protocol is similar to the Hi-C, where chromatin is fragmented with restriction enzymes and the overhangs are filled with biotin. After proximity ligation and sonication, the fragments are immunoprecipitated using antibodies targeting specific histone marks or proteins. The DNA is purified and biotin enriched regions are selected for paired-end sequencing.

Analysis: PLAC-Seq reads, unlike ChIA-PET, can be directly mapped to the reference genome without data pre-processing (Fang et al. 2016). Each of the paired-ends in PLAC-Seq data is mapped separately to the reference genome with bwa-mem in single-end mode using default settings. Then both mapped ends are paired together and MQAL>10 is used for quality filtering. Since MboI is used for digestion, reads mapped > 500bp apart of the restriction site are removed and PCR duplicates are removed using MarkDuplicates in Picard tools. Finally, the interaction matrix is created using both short-range and long-range chromatin interactions.

- **Split-Pool Recognition of Interactions by Tag Extension (SPRITE)**

SPRITE is a proximity ligation-independent method that can detect genome-wide chromatin interactions (Quinodoz et al. 2018). In addition to pairwise interactions, SPRITE can capture multiple interacting regions at the same time. Its principle relies on combinatorial barcoding with a split and pool approach. The protocol starts with crosslinking the DNA followed by splitting the crosslinked material into a 96-well plate and tagging each well with a unique barcode. The material is then combined and split again into a new 96-well plate to receive a new unique barcode. This splitting, pooling and barcoding process is repeated multiple times to uniquely tag each cluster of interacting molecules. In this case, all interacting fragments that are crosslinked to each other are expected to have the same barcodes.

ChIA-Drop is similar to the SPRITE method that captures multi-way contacts (Zheng et al. 2019). Chromatin is crosslinked and fragmented and then samples are loaded to a microfluidics device. Mixed chromatin is divided into gel-bead-in-emulsion (GEM) droplets containing unique DNA oligonucleotides and reagents for amplification and barcoding. Samples are pooled together and sequenced. Finally, barcodes are used to assign the identical GEM. ChIA-Drop data has only been generated in Drosophila to date and hence was not used in these studies (Zheng et al. 2019). By extracting pairwise interactions from SPRITE, one can compare the contact profile to Hi-C.

Analysis: The initial analysis in SPRITE focuses on only pairwise interactions**.** Bowtie 2 with default parameters is used to map the paired-end

sequences with local alignment mode with 2 bp mismatch allowance. Samtools is used to create bam files (Danecek et al. 2021). MAPQ score > 10 and MAPQ>30 thresholds are used for low-quality read filtering. RepeatMasker is used to filter PCR duplicates. Multi-mapped reads were removed except the regions generated by the ComputerGenomeMask program in the GATK with a 35 nucleotide mask. Barcode sequences are used to define SPRITE clusters. To create the 2D interaction matrix, multi-way interactions are converted to pairwise interactions as follows: within each cluster all possible pairwise interactions are created. The scale of interactions will be quadrically based on the number of reads, n. n reads that will have n-1 contacts. So each pairwise contact is normalized by n(n-1)/2 to ensure equal representation of contacts for small and large clusters. Finally, SPRITE interaction maps are normalized by Hi-Corrector (Li et al. 2015). Additionally, contacts are normalized to an expected number of contacts which are calculated by mixing human and mouse samples and determining the number of inter-species contacts. The final interaction matrix in SPRITE is comparable to other genome-wide methods such as Hi-C and Micro-C.

To analyze multiway contacts detected in SPRITE and ChIA-Drop, a newly developed method called MATCHA is used (Zhang and Ma 2020). MATCHA assumes non-overlapping genomic bins as nodes and multi-way chromatin contacts as hyperedges. Pairwise data generated from multi-way contacts pass through the Mix-n-Match autoencoder to generate hypergraphs (Wang, Herranz, and van de Weijer 2020). The pairwise contact data is used as an input to create higher-order contacts using distance weights for intra-

chromosomal interactions.

To specifically investigate the inter-chromosomal interactions, a Mix-n-Match approach is used so that the reconstruction of the interactions is made by mixing the encoder and decoder in a random manner. The encoder is created from a specific genomic bin located in a specific chromosome considering the number of bins of this chromosome and the interaction of this chromosome with another chromosome. n encoder and n decoder are both trained, where each encoder takes a vector and converts it to a hidden vector. The decoder is created using the same input size but the other way around. The interactions between chromosomes are paired randomly to determine the comparable Decoder for each chromosome.

- **Genome architecture mapping (GAM)**

Genome architecture mapping (GAM), which is a ligation independent method based on cryosectioning, measures genomic distances in the nucleus at a genome-wide scale (Beagrie et al. 2017). In this method, individual nuclei are cryosectioned with laser microdissection in multiple orientations and each slice is uniquely barcoded. Then the DNA fragments in each slice are extracted, amplified, and sequenced. Since the slices are ultrathin, the fragments identified in the same slice are considered to be in close distance. Identification of colocalized regions in multiple nuclei provides information about the position and distance between these regions. A distance matrix can be created using information from aggregated nuclear slices.

***Analysis*:** Distance information from GAM nuclei slices is converted to the

distance matrix. Sequencing reads were mapped to the reference genome using Bowtie 2 with default parameters (Langmead and Salzberg 2012). PCR duplicates and low-quality mapped reads were removed from the data. FASTQC is used to measure the per base read quality and dinucleotide repeats(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) . Next, the genome was split into 10kb to 1Mb windows, and the read coverage is calculated for each window and each nuclear profile separately. A negative binomial distribution is used to generate a threshold to determine the sequencing noise of each experiment. Nuclei that have mapped reads that are lower than the expected sequencing noise were excluded from the analysis. The number of nuclei is important in determining the resolution of the experiment. For example, 408 nuclei are sufficient to create an interaction map of 30kb resolution. Finally, the interaction frequency is determined by the number of interaction profiles in each nucleus divided by the total number of nuclear profiles.

Methods that are used to measure features of the genome

Gene activity, chromatin domains, and replication timing can be measured by various methods. Below, I will introduce the experimental setup and analysis of the commonly used methods that determine gene activity (TSA-seq), chromatin domains (DamID-seq), and replication timing (Repli-Seq) (Chen et al. 2018; Wang et al. 1999; Wu, Olson, and Yao 2016; Marchal et al. 2018; Marchal, Sima, and Gilbert 2019).

- **Tyramide Signal Amplification Sequencing**

Tyramide Signal Amplification Sequencing (TSA-Seq) maps genome-wide chromosomal distances from compartments such as nuclear speckles or Lamina to the nuclear periphery, which is used as a predictor for gene expression levels. TSA is a immunohistochemistry method developed by Wang et al. in 1999 to generate tyramide free-radicals using horseradish peroxidase (HRP) (Wang et al. 1999). Diffusion of these free-radicals results in covalent bonds with neighboring macromolecules. TSA-Seq uses TSA to estimate genome-wide cytological distances (Chen et al. 2018). The TSA reaction is followed by reversed FA cross-linking, isolating DNA, and pulling down biotinylated DNA for high-throughput sequencing. In this chapter, I focused on TSA-Seq data targeting the proteins below: Nuclear Speckle protein (SON), Nuclear Lamina protein (LMNB1), Nucleolar protein (NIFK), Centromere protein (CENB1) and RNA Polymerase I subunit E (POLR1E) (https://data.4dnucleome.org/).

Analysis: Genome-wide cytological distances measured by TSA-Seq can be generated as follows. Single-end sequenced reads are mapped to their reference genome using Bowtie 2 with default parameters excluding chromosome Y (Langmead and Salzberg 2012). SAMtools is used to remove PCR duplicates (Danecek et al. 2021). Mapped reads are normalized such that reads in 20 kb window size are normalized to the total number of reads in that sample and then divided by their corresponding input which is similarly normalized. The log2 of the fold-change track is used for further analysis.

- **DNA adenine methyltransferase identification sequencing**

DNA adenine methyltransferase identification (DamID-Seq) is used to detect genome-wide protein-DNA interactions as an alternative to ChIP-Seq experiments (Vogel, Peric-Hupkes, and van Steensel 2007; Wu, Olson, and Yao 2016). *Escherichia coli* DNA methyltransferase (Dam) protein is fused to a chromatin-binding protein of interest in cells to make preferential methylation of nearby adenines. Since the methylation of adenines does not occur naturally, it is used to tag DNA proximal to the protein of interest, which can then be specifically amplified by PCR. Recent DamID protocols, DamID-Seq incorporate high-throughput sequencing to obtain genome-wide information.

Analysis: DamID-Seq data was mapped exactly like the TSA-Seq data mentioned above.

- **Replication Timing Sequencing**

Replication Timing Sequencing (Repli-Seq) was developed to measure the replication timing of different regions on the genome (Marchal et al. 2018). Newly synthesized DNA strands are labeled by 5-bromo-2-deoxyuridine (BrdU) which is incorporated into the replicated DNA in place of thymidine. Labeled cells are sorted based on DNA content into G1, S1, S2, S3, S4, and G2. BrdU labeled DNA strands are extracted using anti-BrdU monoclonal antibodies and the isolated BrdU fractions are sequenced. Mapping the sequences of BrdU-labeled nascent DNA replication strands on the genome enables inferences about DNA replication timing.

Analysis: Data analysis begins with quality checks and mapping the data to the reference genome. Then the ratio of early replication timing to late replication timing is calculated. The coverage of the log 2 ratio is smoothed and used for visualization and downstream analysis.

Integration of Hi-C, Micro-C, ChIA-PET, PLAC-Seq, SPRITE, GAM, TSA-Seq, DamID-Seq and Repli-Seq

Structures detected in 3D methods highly correlate with signals detected from TSA-Seq (gene activity), DamID-Seq (chromatin domains) and Repli-Seq (replication timing) (Boninsegna et al. 2022; Vouzas and Gilbert 2021; Wang et al. 2021; Marchal, Sima, and Gilbert 2019; Zheng et al. 2018). Strong signal of TSA-Seq speckle association is an indication of active euchromatic region and lower signal of TSA-Seq is an indicator of heterochromatin, which is reversed in TSA-Seq LaminB1 and most DamID experiments. Repli-Seq signals also show that the active regions have faster replication than the inactive heterochromatic regions. TSA-Seq, DamID and Repli-Seq correlate with the compartmentalization signal detected in 3D experiments. EigenVector Decomposition is used to detect the compartment signal in 3D experiments so that the positive values of the first eigenvector corresponds to active euchromatin and negative values correspond to inactive heterochromatic regions.

Measuring the reproducibility of 3D data

Reproducibility of an experiment is one of the most important measures that show the data is real and of good quality. Below, I briefly explain the main

methods developed to measure the 3D data reproducibility: Pearson correlation, HiCRep, GenomeDISCO, Hi-C-Spector and QuASAR-Rep (Yang et al. 2017; Ursu et al. 2018; Yan et al. 2017; Yardimci et al. 2019; Sauria and Taylor 2017).

Pearson correlation is the first approach used to measure the correlation between two Hi-C contact matrices generated using different enzymes (Lieberman-Aiden et al. 2009).

HiCRep uses the stratum adjusted correlation coefficient (SCC) to quantify the similarity between two matrices (Yang et al. 2017). HiCRep starts with smoothing the data matrix and takes the distance-dependence effect into account using weighted Pearson correlations.

GenomeDISCO starts with smoothing the contact maps using random walks, which are then converted to networks (Ursu et al. 2018). Nodes represent the genomic loci and the edges represent the interaction frequency between two loci. The smoothed maps are compared. Noise and distance decay are considered for each contact map and systematic simulations are used to quantify the methods.

Hi-C Spector starts with calculating the Laplacian matrix of each dataset followed by calculation and comparison of the eigenvector decompositions of these matrices (Yan et al. 2017).

QuASAR-Rep assesses the reproducibility considering multiple resolutions of transformed matrices (Sauria and Taylor 2017). Matrix transformation is done using local correlations of matrices weighted by the contact frequency of each

matrix.

HiCRep and GenomeDISCO quantify chromosome-scale reproducibility while the other described methods measure genome-wide correlations. Although these methods have been commonly used, they do not perform similarly. To compare the performance of these methods controlling noise and sparsity, one study has used real and simulated datasets such as Pseudo replicates, biological replicates, and non-replicates (Yardimci et al. 2019). Hi-C Spector was found to be the strongest and the Pearson correlation to be the weakest measure for Hi-C data reproducibility. Although Hi-C Spector performs the best finding high reproducibility between replicates and classification of noisy datasets, HiCRep, GenomeDISCO and QuASAR-Rep have a powerful performance separating pseudo replicates, biological replicates, and non-replicates. Hi-C Spector produces the strongest results seperating i) the simulated data sets with different noise injection levels, ii) biological replicates, pseudo replicates and non-replicate, iii) all aforementioned replicates with different coverage levels, iv) reproducibility in different cell types.

Structure detection in 3D data

- **Compartment detection**

Compartments are the largest structures detected in 3D genome organization. To identify these compartments Eigenvector decomposition is used but for subcompartment identification unsupervised clustering algorithms like Hidden Markov Model and k-means clustering are generally used.

Eigenvector decomposition has been used on the correlation matrix of Hi-C to define A and B compartments (Lieberman-Aiden et al. 2009). The correlation matrix is computed with Pearson correlation between ith row and jth column of a matrix M. The matrix is normalized to its expected (obs/exp) values to remove the distance-decay effect before performing the eigenvector decomposition. To compute A and B compartments, an automated method called Cooltools (https://github.com/open2c/cooltools) is used. Cooltools reads multi-resolution matrix and uses eigenvector decomposition on obs/exp, which is a strong method to identify compartments in a matrix with a low (a few million) number of interactions.

Detailed investigation of compartments using a Hidden Markov Model has revealed subcompartments (Rao et al. 2014). Using a high resolution deep (~4.9 billion contacts) Hi-C data an inter-chromosomal matrix was constructed with odd chromosomes being in the rows and even chromosomes being in the columns. Z-score was applied to the odd chromosomes of this matrix after filtering rows and columns that had low coverage. Z-scored matrix is used for unsupervised Gaussian hidden Markov model clustering algorithm (GaussianHMM). The z-score was independently applied to even chromosomes and GaussianHMM clustering was applied to this matrix as well. Finally, the authors found that each cluster in the even chromosomes preferentially interacts with one cluster of odd chromosomes. They identified 5 clusters/subcompartments (A1, A2, B1, B2 and B3) that have preferential interactions. Identification of these clusters was determined by correlating them with A and B compartments.

Another approach to identify subcompartments is Subcompartment iNference, which uses Imputed Probabilistic ExpRessions (SNIPER) based on de-noising autoencoder and a multilayer perceptron classifier using around 500 million interactions (Xiong and Ma 2019). SNIPER creates a similar inter-chromosomal interaction matrix as the aforementioned study as an input for an autoencoder, which outputs a dense inter-chromosomal contact matrix and features of the sparse matrix as latent variables in low-dimensional (Rao et al. 2014). Latent variables lower the dimensionality of the inter-chromosomal matrix and inputs 5 subcompartment categories into the classifier that are identified based on GM12878. The training was done for both odd and even numbered chromosomes. Multi-layer perceptron network is used to predict subcompartment annotations. Since SNIPER was trained using a lymphoblastoid cell line (GM12878), the prediction may not hold true for very different cell types.

Another study identified subcompartments by using 9 eigenvectors that have the most variation (Spracklin et al. 2021). These eigenvectors highly correlate with signal tracks from speckles, replication timing, GC content, Lamin B1 and Protect Seq. This subcompartment prediction approach does not require deep datasets and shows a great correlation with genome function.

- **TAD detection**

TADs form as a consequence of loop extrusion mechanisms and look like rectangles in the chromatin interaction map with enhanced interactions

within a TAD and depleted interactions between TADs. Unlike compartments, TADs have a hierarchical structure, which makes TAD identification more difficult than the compartments. These embedded TAD structures led to the development of over thirty methods to identify TAD-like structures. These methods are based on multiple features, which include averaging interaction frequency of the Hi-C heatmap across the diagonal with a given sliding window, calculating average upstream and downstream interactions of a given bin and transforming these interactions into chi-squared statistics, graph theory-based methods (Nora et al. 2012; Dixon et al. 2012; Norton et al. 2018). Arrowhead algorithm that uses the global expected to find the TAD boundaries, and graph theory-based algorithms to find TADs and subTADs (Rao et al. 2014).

Multiple tools have been developed to calculate insulation score and directionality index. Cooltools is used to calculate insulation score and FAN-C is used to find both insulation score and directionality index (Kruse, Hug, and Vaquerizas 2020). Juicer is used to compute TADs using the arrowhead algorithm and 3DNetMod is a method developed to define TADs and subTADs using a graph based algorithm (Robinson et al. 2018; Norton et al. 2018). Over thirty tools developed for TADs comparison using Hi-C and Hi-C-like datasets have been compared in a comprehensive study (Liu et al. 2022). These methods showed differences in the number of detected TADs, TAD size, detection of subTADs, and the detection of the same TADs. Some methods use histone modifications to identify TAD boundaries.

- **Loop Detection**

Chromatin loops are the finest structures detected using 3D methods and look like a dot in the interaction heatmap. Detection of these dots requires a correction for global and local backgrounds. A variety of methods have been developed to detect dot-like structures. Some methods focused on detecting enriched interactions regardless of their shape using the global background.

HICCUPS, SIP, Mustache and Pikachu are some of the methods that are used to detect chromatin loops correcting for local and global background (Rao et al. 2014; Rowley et al. 2020; Roayaei Ardakany et al. 2020; Salameh et al. 2020). HICCUPS detects the chromatin loops by first normalizing the observed matrix by the global expected to identify potential enrichments and then each of these enrichments were compared to their neighbor pixels to show the local enrichment and dot-like structures. Cooltools uses the same strategy to identify loops. Cooltools is written in python and easier to implement compared to HICCUPS which is written in Java and has many dependencies.

SIP uses image processing algorithms to identify chromatin loops (Rowley et al. 2020). Images are adjusted and corrected using gaussian blur, contrast enhancement, White Top-Hat and minimum-maximum filters to pass the initial screening. Then loop candidates are filtered based on their enrichment and position in the genome. SIP detects more loops than HICCUPS with lower false positive rate.

Peakachu uses supervised Machine Learning algorithms to find

chromatin loops (Salameh et al. 2020). Peakachu is constructed using positive and negative training sets. The positive set is created using loop-like enrichments extracted from ChIA-PET/HiChIP, Capture Hi-C and HiFISH and the negative set is created using random interactions that have similar genomic distance as loops and interactions extracted from longer distances. Hyperparameter search is used to find the best random forest model that segregates positive and negative loop sets. Peakachu has advantages over other methods by detecting loops in high resolution and with low sequencing depth.

The Mustache algorithm uses a 2D Gaussian approach to convolve an interaction map and produce a scale-space image representation (Roayaei Ardakany et al. 2020). Then it subtracts the images from each other to find the local maximum enrichment in each image and compared to its local background. Mustache provides reproducible results compared to HICCUPS and SIP and is able to call loops in 1kb resolution.

In addition to these methods, Fit-HiC and HiC-DC detect any enrichment in the interaction map regardless of the local background (Ay, Bailey, and Noble 2014; Carty et al. 2017). Both Fit-HiC and HiC-DC use statistical models based on the global background to identify enriched pixels in 3C-based methods. The advantages of Fit-HiC and HiC-DC is that they provide a more global approach for structure identification regardless of the shape of the structures. They can detect stripe-like structures that are observed in interaction maps and occur as a consequence of cohesin stalling.

Such stripe-like structures show high interaction frequency in an adjoining genomic interval with a great amount of cohesin loading. Juicebox is one of the methods that can be used to identify stripes (Robinson et al. 2018).

Understanding 3D genome organization is crucial to unveil complicated relationships between genes and their distal enhancers, DNA folding patterns and relationships between different regions of the chromatin. Hi-C 3.0 was updated to increase the potential of capturing multi-layer structures in 3D. But many other methods are still available to unveil information about 3D structures that are not captured by Hi-C. Every method that measures the 3D genome captures some aspects of chromatin organization. Such methods include Micro-C, PLAC Seq, ChIA PET, SPRITE and GAM. Analyzing these methods together and combining information extracted from these methods give a global view of the chromatin organization. This study showed the basic differences between 3D methods and how method selection is crucial in understanding biological problems.

a

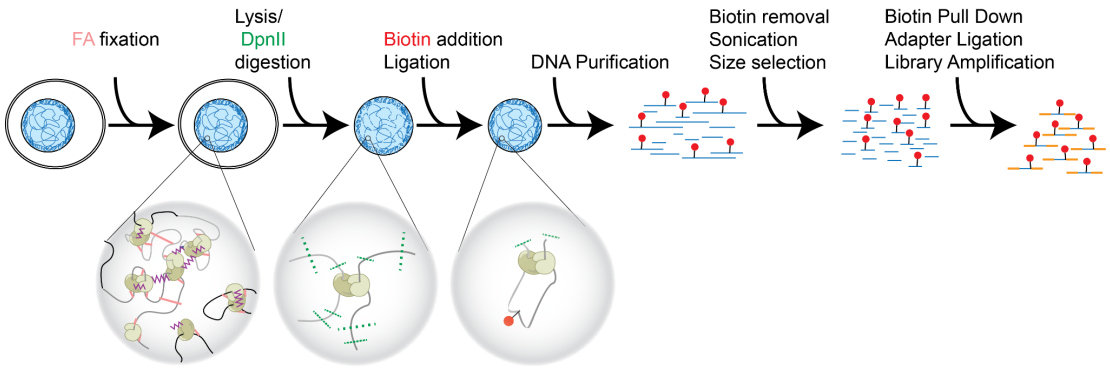

Figure 1.1: Hi-C Protocol

a. An illustration of the steps in Hi-C 2.0 Protocol

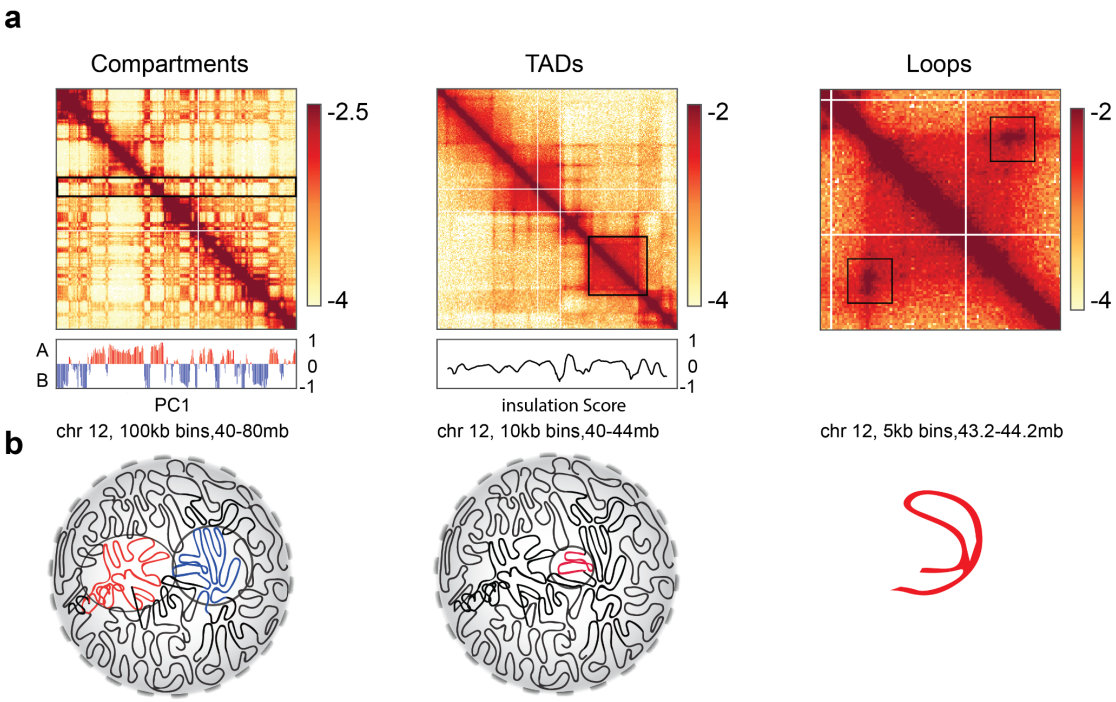

Figure 1.2: Multi-layer chromatin folding

a. Interaction maps of a Hi-C experiment showing compartments, TADs and chromatin loops.

b. Observed formation of compartments, TADs and loops in the nucleus

Chapter II: Systematic evaluation of chromosome conformation capture assays

Preface

This research chapter is published work in Nature Methods by

Betul Akgol Oksuz, Liyan Yang, Sameer Abraham, Sergey V. Venev, Nils Krietenstein,Krishna Mohan Parsi, Hakan Ozadam, Marlies E. Oomen, Ankita Nand, Hui Mao, Ryan M. J. Genga, Rene Maehr, Oliver J. Rando, Leonid A. Mirny, Johan H. Gibcus and Job Dekker. The publication is entitled "Systematic evaluation of chromosome conformation capture assays". PMID: 34480151 PMCID: PMC8446342 DOI: 10.1038/s41592-021-01248-7

Summary

Chromosome conformation capture (3C) assays are used to map chromatin interactions genome-wide. Chromatin interaction maps provide insights into the spatial organization of chromosomes and the mechanisms by which they fold. A number of 3C protocols such as Hi-C and Micro-C are now widely used and these differ in key experimental parameters including cross-linking chemistry and chromatin fragmentation strategy. To understand how the choice of experimental protocol determines the ability to detect and quantify aspects of chromosome folding, we have performed a systematic evaluation of experimental parameters of 3C-based protocols. We identified optimal protocol variants for either loop detection (fine fragmentation and use of combinations of formaldehyde and disuccinimidyl glutarate or Ethylene glycol bis(succinimidylsuccinate) cross-linking) or compartment detection (large

fragments). We used this information to develop a greatly improved Hi-C protocol that can detect both loops and compartments relatively effectively. In addition to providing benchmarked protocols, this work produced ultra-deep chromatin interaction maps using Micro-C, conventional Hi-C and improved Hi-C for key cell lines used by the 4D Nucleome project.

Introduction

Chromosome conformation capture (3C)-based assays (Dekker et al. 2002) have become widely used to generate genome-wide chromatin interaction maps (Denker and de Laat 2016). Analysis of chromatin interaction maps has led to the detection of several features of the folded genome. Such features include precise looping interactions (0.1-1 Mb scale) between pairs of specific sites that appear as local dots in interaction maps. Many of such dots represent loops formed by cohesin-mediated loop extrusion that is stalled at convergent CTCF sites (Rao et al. 2014; Kagey et al. 2010; Fudenberg et al. 2016). Loop extrusion will also produce other features in interaction maps including stripe-like patterns anchored at specific sites that block loop extrusion, and the effective depletion of interaction across such blocking sites leading to domain boundaries (insulation). At the Mb scale interaction maps of many organisms including mammals display checkerboard patterns that represent the spatial compartmentalization of two main types of chromatin: active and open A-type chromatin domains and inactive and more closed B-type chromatin domains (Lieberman-Aiden et al. 2009).

There have been significant improvements to the 3C assay since its

development to reduce signal-to-noise ratio. Several parameters have been tested to improve Hi-C, including performing crosslinking, digestion, and ligation in intact nuclei as well as modifications on the crosslinking chemistry and fragmentation strategies (Rao et al. 2014; Nagano et al. 2015). Crosslinking and fragmentation parameters have been shown to be critical for 3C-based protocols. Formaldehyde (FA) is the most commonly used crosslinking reagent in 3C-based assays. Efficiency of FA crosslinking has been improved over the years by various optimizations focusing on FA concentration, crosslinking time, serum presence in the cell media before crosslinking, freshness of the FA before crosslinking and the way FA is added to the cells (Naumova et al. 2012; Naumova et al. 2013; Belton et al. 2012; Golloshi, Sanders, and McCord 2018). In addition to FA, using extra crosslinking reagents also improved the crosslinking efficiency. For example, additional usage of homobifunctional N-hydroxysuccinimide (NHS) ester crosslinking reagent (EGS) improved protein-DNA crosslinking in a ChIA-PET assay, which is a 3C-based technique that measures chromatin interactions of regions bound by a specific protein (Fullwood et al. 2009). Using EGS and another NHS esther crosslinker, disuccinimidyl glutarate (DSG), also led to an improvement of short and long-range signal at nucleosome-free regions in Micro-C assay, which is a derivation of Hi-C that uses micrococcal nuclease (MNase) to fragment chromatin (Hsieh et al. 2016; Krietenstein et al. 2020). To fragment chromatin various restriction enzymes, MNase and DNase have been tested (Krietenstein et al. 2020; Ramani et al. 2016). For example, to increase the resolution of Hi-C, chromatin is digested into smaller fragments

using 4-base cutters such as MboI and DpnII instead of 6-base cutters EcoRI NcoI, HindIII (Rao et al. 2014; Belaghzal, Dekker, and Gibcus 2017; Belton et al. 2012; Lieberman-Aiden et al. 2009). These modifications have significantly improved the resolution and signal-to-noise ratio in Hi-C experiments. In spite of these improvements, most conformation capture techniques still cross-link chromatin as developed in the original 3C protocol, i.e. with 1% formaldehyde and uses several commonly used restriction enzymes to fragment chromatin (Dekker et al. 2002).

Specific features of genome folding, such as compartments TADs and loops result from different mechanisms and exist at different length scales (Ganji et al. 2018; Nora et al. 2012; Dixon et al. 2012; Lieberman-Aiden et al. 2009; Mirny, Imakaev, and Abdennur 2019). The current Hi-C protocol is limited in detecting all of these structures at high resolution. Evidences suggest that optimization of experimental parameters such as crosslinking chemistry and fragmentation, may improve detection of specific features of genomic structures (Krietenstein et al. 2020; Rao et al. 2014; Dixon et al. 2015; Naumova et al. 2013; Gibcus et al. 2018; Abramo et al. 2019; Golloshi, Sanders, and McCord 2018). Thus, we reasoned that it might be possible to optimize a single protocol that can capture genomic interaction at all scales, which would improve our understanding of genome structure and function.

It is critical to ascertain how key parameters of these 3C-based methods quantitatively influence the detection of chromatin interaction frequencies and the detection of different chromosome folding features that

range from local looping between small cis elements to global compartmentalization of Mb-sized domains. Here we systematically assessed how different cross-linking and fragmentation methods yield quantitatively different chromatin interaction maps.

Results

We set out to explore how two key parameters of 3C-based protocols, cross-linking and chromatin fragmentation, determine the ability to quantitatively detect chromatin compartment domains and loops. We selected three cross-linking chemistries widely used to cross-link chromatin: 1) 1% formaldehyde (FA) , conventional for most 3C-based protocols, 2) 1% FA followed by an incubation with 3 mM disuccinimidyl glutarate (FA+DSG protocol), and 3) 1% FA followed by an incubation with 3 mM Ethylene glycol bis(succinimidylsuccinate) (FA+EGS protocol) (Figure 2.1 a). We selected 4 different nucleases for chromatin fragmentation: MNase, DdeI, DpnII and HindIII, which fragment chromatin in sizes ranging from single nucleosomes to multiple kilobases. Combined, the 3 cross-linking and 4 fragmentation strategies yield a matrix of 12 distinct 3C-based protocols (Figure 2.1 b). To determine how performance of these protocols varies for different states of chromatin we applied this matrix of protocols to multiple cell types and cell cycle stages. We analyzed 4 different cell types: pluripotent H1-hESCs (12 Protocols), differentiated endoderm (DE) cells derived from H1-hESCs (12 protocols), fully differentiated Human Foreskin Fibroblasts (HFF (12 Protocols), and a clonal derivate HFFc6) and HeLa S3 (9 Protocols) cells.

Furthermore, we analyzed two cell cycle stages: G1 (9 Protocols) and mitotic (9 Protocols) HeLa S3 cells (Figure 1). Each interaction library was then sequenced each on a single lane of a HiSeq4000 instrument, producing ~150-200 million uniquely mapping read pairs for each experiment (Table 1).

We first assessed the size range of the chromatin fragments produced after digestion by the twelve protocols for HFF cells (see Methods). Digestion with HindIII resulted in 5-20 kb DNA fragments; DpnII and DdeI produced fragments of 0.5-5kb; and MNase digested up to the level of mononucleosomes (~150 bp) (Figure 2.2 a). For protocols using MNase we included a size selection step and therefore all interactions in those datasets involve pairs of nucleosome-sized fragments. Different cross-linking chemistries did not affect the size ranges produced by the different nucleases much, though DSG cross-linking lowered digestion efficiency slightly (Figure 2.2 b).

- **All tested 3C-based protocols can differentiate between cell states**

We first assessed similarity between the 63 datasets by global and pairwise correlations using HiCRep and hierarchical clustering (Figure 2.2 c) (Abdennur and Mirny 2020; Imakaev et al. 2012). We found that the datasets are highly correlated and cluster primarily by cell type and state and then by cell type similarity, as for example H1-ESC and ESC derived DE cells cluster together; and the most distinct cluster is formed by mitotic HeLa cells. Although data from all protocols is highly correlated, MNase protocols show slightly lower correlations with Hi-C experiments (Figure 2.2 d-g).

- **Extra cross-linking yields more intra-chromosomal interactions in all cell states**

Given that chromosomes occupy individual territories, intra-chromosomal (cis) interactions are more frequent than inter-chromosomal (trans) interactions (Yang et al. 2017). The ratio of the number of interactions found in cis and trans is commonly used as an indicator of Hi-C library quality given that inter-chromosomal interactions are a mixture of true chromatin interactions and interactions that are the result of random ligations (see below) (Yang et al. 2017; Yardimci et al. 2019). For all enzymes and cell types, we found that the addition of DSG or EGS to FA cross-linking decreased the percentage of trans interactions (Figure 2.3 a (HFF) ; Fig.2.4 a (H1-hESC, DE, Hela S3)).

With respect to intra-chromosomal interactions, we noticed two distinct patterns. First, digestion into smaller fragments resulted in a relative increase in short range interactions. MNase digestion resulted in a higher number of interactions between loci separated by less than 10 kb, whereas digestion with either DdeI, DpnII or HindIII resulted in a relatively larger number of interactions between loci separated by more than 10 kb (Figure 2.3 a,b (HFF), Figure 2.4 a,b (DE, H1-hESC, HelaS3)). Second, $P(s)$ plots showed that the addition of either DSG or EGS resulted in a steeper decay in interaction frequency as a function of genomic distance for all fragmentation protocols. Moreover, for a given chromatin fragmentation level, additional cross-linking with DSG or EGS reduced the percentage of trans interactions, as shown for HFF and all other cell types and cell stages studied (Figure 2.3 c, d and

Figure 2.4 c). Addition of DSG or EGS could have reduced fragment mobility and formation of spurious ligations, resulting in a steeper slope of the *P(s)*. We note a difference in slopes for data obtained with different cell types and cell cycle stages, which could reflect state-dependent differences in chromatin compaction.

Random ligation events between uncrosslinked, freely diffusing fragments lead to noise in 3C-based experiments. Such noise will mostly be seen in trans interactions, or very long range cis interactions where true signal is low. To estimate the rate of random ligations, we compared trans interaction frequencies for each protocol to interactions between mitochondrial and nuclear genomes, as these interactions can only result from random ligations (Figure 2.4 d). We could not use this noise metric for experiments using MNase because MNase completely degrades the mitochondrial genome. Interestingly we observed that random ligations between genomic and mitochondrial DNA were the lowest when chromatin was fragmented with HindIII, and generally higher when chromatin was fragmented in smaller segments with DpnII or DdeI. Additional DSG or EGS cross-linking reduced random ligation in experiments using DpnII or DdeI, but did not further reduce random ligations when HindIII was used.

Reduced noise in experiments using HindIII or in experiments using DpnII with additional cross-linkers is readily visible in chromatin interaction maps: for these protocols we observed a general decrease in trans contacts, while uncovering a stronger inter-chromosomal compartmental pattern (Figure

2.3 e). We conclude that noise as a result of random ligation can be reduced by either using 6 bp-cutters (HindIII) or, when using more frequent cutters (DpnII, DdeI), by using additional DSG or EGS cross-linking. The reduced noise improves trans compartment detection and possibly long range cis interactions.

- **Quantitative detection of compartmentalization is enhanced by long fragments and extra cross-linkers**

Visual inspection of interaction matrices (binned at 100 kb resolution) suggested that the contrast between the domains that make up the A and B compartments can vary between protocols. For instance, for HFF cells cross-linked with only FA, interaction matrices obtained with MNase digestion displayed a relatively weak compartment pattern, whereas those obtained with HindIII digestion showed much stronger patterns (Fig 2.5 a).

To investigate compartmentalization and determine the positions of A and B compartments, we used eigenvector decomposition (Lieberman-Aiden et al. 2009; Lajoie, Dekker, and Kaplan 2015) for all cell states except for mitotic cells that do not display compartmentalized chromosomes (Schmitt, Hu, and Ren 2016). Correlation between compartment profiles of all experiments showed that the greatest difference in profiles can be attributed to cell type (Figure 2.6 a). Within each cell type, positions of compartment domains obtained with different protocols were highly similar (Spearman correlation >0.8; Figure 2.6 a).

Compartment strength analysis using saddle plots revealed three important trends (Lajoie, Dekker, and Kaplan 2015; Naumova et al. 2013; Nora et al. 2017). To generate saddle plots the contact matrix of Hi-C or Micro-C was sorted based on the eigenvector values from lowest to highest (B to A). Sorted maps were then corrected for their expected contact frequencies. The strongest B-B contacts are located in the upper left corner and the strongest A-A contacts are located in the lower right corner. The upper right and lower left are B-A and A-B, respectively. The 20% of the strongest A-A and B-B interactions were normalized to the strongest 20% of A-B and B-A interactions to generate a single value for compartment strength.

First, protocols that generate larger fragments (e.g. using HindIII; Figure 2.5 b, c) and protocols that include additional DSG or EGS cross-linking produced quantitatively stronger compartment patterns (Figure 2.5 c; Figure 2.6 b-e) for all 4 cell types. Second, different cell types differed in compartment strength: HFF cells displayed the strongest compartment pattern, while H1-hESCs displayed weak compartments regardless of the protocol used. This could be related to differences in chromatin state and/or cell cycle distribution between the cell types. Third, compartment strength was much stronger in cis than in trans. Furthermore, some protocols, including the conventional Hi-C protocol (cross-linked with FA and digestion with DpnII) and MNase-based protocols (Micro-C, regardless of cross-linking protocol) did not detect enrichment of B-B interactions in trans (Figure 2.6). Such preferred B-B interactions were detected only when Hi-C was performed with HindIII (Figure 2.6 d, e)). Additionally, trans preferential A-A interactions were more frequent than trans

preferential B-B interactions for all protocols and cell types. In summary, compartment strength was stronger in both cis and trans, when protocols produce larger fragments or employ additional cross-linking.

- **Chromatin loops are better detected in experiments with finer fragmentation and additional DSG cross-linking**

Of all structural Hi-C features, the detection of loops depends the most on sequencing depth. We applied 1) conventional Hi-C using FA and DpnII digestion (FA-DpnII); 2) Hi-C using DSG in addition to FA cross-linking and DpnII digestion (FA+DSG-DpnII); and 3) the standard Micro-C protocol (FA+DSG-MNase) to two cell types, H1-hESC and HFFc6, and sequenced the resulting libraries to a depth of 2.4-3.9 billion valid interactions (two biological replicates combined). HFFc6 cells are a subclone of HFF cells and are used by the 4D Nucleome Consortium (Dekker et al. 2017). Interaction maps of data obtained from these “deep” datasets showed quantitative differences in interactions for both H1-hESC and HFFc6 (Figure 2.7 a,b). As compared to conventional Hi-C (FA-DpnII), the use of additional cross-linking with DSG and finer fragmentation produced contact maps with more contrast and more pronounced focal enrichment of specific looping contacts. We used a reimplementation of the HICCUPS approach to identify looping interactions that appear as dots (Rao et al. 2014; Krietenstein et al. 2020) (see Methods).

First, we compared the number of loops detected in individual and merged biological replicates for the deeply sequenced protocols. We observed that 1) the number of loops detected with protocols that cross-link

chromatin with only FA was more sensitive to sequencing depth compared to the number of loops detected with protocols that cross-linked with FA+DSG; 2) loops were more consistent between replicates for datasets obtained with protocols that cross-link chromatin with FA+DSG compared to those detected in FA-only crosslinked replicates (Figure 2.7 c, d). We used the lists of loops that were detected in merged replicates for our analyses. In H1-hESCs we detected 3,951 loops with the FA-DpnII protocol, 12,396 loops with the FA+DSG-DpnII protocol, and 22,507 loops with the FA+DSG-MNase protocol (Figure 2.9 a). For HFFc6 these numbers were 13,867, 22,934 and 36,988 respectively (Figure 2.8 a). To investigate the properties of detected loops, we compared loops that were called in individual or multiple protocols (Figure 2.8 a). While a large fraction of loops was detected by all three protocols, we found that the protocols with additional cross-linking (FA+DSG-DpnII) and finer fragmentation (FA+DSG-MNase) detected a large set of additional loops (Figure 2.8 a).

When we aggregated interaction data for the various subsets of loops detected with one or multiple protocols, we observed a focal increase in interaction frequency for all subsets of loops for all datasets; even for data obtained with protocols where that subset of loops was not detected as significantly enriched (Figure 2.8 b for HFFc6, Figure 2.9 b for H1-ESC). For instance, loops only detected with the FA+DSG-MNase protocol were also visible in aggregated data obtained with the FA-DpnII protocol. Quantifying the strength of the different subsets of loops detected by one or multiple protocols, we found that loops detected by all three protocols were the

strongest, while loops detected only by the FA+DSG-MNase protocol were relatively weak.

We then defined a consensus set of loops that were detected in all three deep datasets and then used this set to analyze the data obtained with the matrix of 12 protocols described in Figure 2.1 that differ in cross-linking and fragmentation strategies. We observed a gradual increase in average loop strength with decreasing fragment size and after addition of DSG or EGS (Figure 2.9 d, e).

To explore this in another way, we quantified the strengths of each loop in the sets of consensus and union loops and found that the majority of loops were strengthened by additional DSG crosslinking (Figure 2.8 c left panel). Further, loop strengths increased by digestion with MNase as compared to DpnII (Figure 2.8 c, middle). Loops were strongest when both additional cross-linkers were used and chromatin was fragmented with MNase (Figure 2.8 c, right plot). A similar trend is also observed in H1-hESC cells (Figure. 2.9 c).

We conclude that the use of additional cross-linkers and enzymes that fragment chromatin in smaller fragments independently contribute to the loop strength and the number of loops that are detectable.

A looping interaction is defined by a pair of frequently interacting loci, anchors. When each anchor is involved in only 1 looping interaction, the number anchors will be twice the number of loops. In contrast, when anchors engage in multiple looping interactions with other anchors, the number of anchors will

be smaller than twofold the number of loops (Fudenberg et al. 2016). To examine the relationship between loops, we compared the number of anchors as a function of the number of loops detected in deep sequenced datasets for HFFc6 cells (Figure 2.10 a). We found that this relation is proportional for the FA-DpnII experiment with a ratio of two, but disproportionate for experiments with improved loop detection (FA-DSG-DpnII and FA-DSG-MNase). This suggests that many of the newly identified loops involved anchors that were also detected with FA-DpnII (Figure 2.10 a, Figure 2.11 a). In other words, many of additionally detected loops are arranged along stripes emanating from the same anchors.

To further investigate this we directly determined the number of loops that a given anchor is engaged in as detected by different protocols and then calculated the difference between them. For each anchor, we subtracted the number of loops detected by the FA-DpnII protocol from the number of loops detected using the FA+DSG-DpnII or the FA-DSG-MNase protocol. We found that using extra cross-linkers as well as finer fragmentation increased the number of detectable loops for most anchors (Figure 2.10 b, c, Figure 2.11 b, c). We conclude that protocols that use additional cross-linkers and finer fragmentation detect more loops in two ways: first, more loops are detected per anchor, and second, additional looping anchors are detected.

We split loop anchors into two categories: 1) anchors detected with more than 1 protocol and 2) anchors detected with only 1 protocol. We observed that anchors detected with at least 2 protocols were engaged in multiple loops

(loop “valency” >1). In contrast, anchors that were detected with only 1 protocol mostly had a loop valency of 1 (Figure 2.11 d, e). Interestingly, for H1-ESCs the majority of additional loops detected with FA-DSG-MNase protocol (62%) involve two anchors not detected with other protocols. In comparison for HFFc6 this was only 21% indicating that most new loops shared at least one anchor with loops detected with other protocols.

We investigated factor binding (CTCF and cohesin (SMC1), YY1 and RNA polII) and chromatin state (H3K4Me3, H3K27Ac) at the two categories of loop anchors. We used publicly available datasets (Dekker et al. 2017; Janssens et al. 2018) and new data generated using a variety of techniques (Cut&Run, Cut&Tag, ChIP Seq and ATAC-Seq for this analysis (Zhang et al. 2020; Skene and Henikoff 2017; Kaya-Okur et al. 2019; Buenrostro et al. 2015). An example region is shown in Figure 2.10 d. Some loop anchors were detected with all protocols and in the example shown these correspond to sites bound by CTCF and cohesin (cyan squares). Other loop anchors were only detected with the FA-DSG-MNase protocol (black squares). In this example these do not correspond to sites bound by CTCF and cohesin, but were enriched in H3K27Ac and H3K4Me3. Possibly, the ability of different protocols to detect various loop anchors is related to factor binding and chromatin state. To investigate this genome-wide we aggregated CTCF, SMC1, YY1, RNA PolII binding data and histone modification data (H3K4me3 and H3K27ac) at loop anchors detected with all protocols or with only FA-DSG-MNase (Figure 2.10 e). Interestingly, in HFFc6 cells we found that FA+DSG-MNase-specific loop anchors were less enriched for CTCF and SMC1 but more enriched for

H3K4me3 and H3K27ac compared to the loop anchors that were detected by all three protocols which were more enriched for CTCF and SMC1 but less enriched for H3K4me3 and H3K27ac (Figure 2.10 e, Figure 2.11 f).

Next, we examined the predicted cis regulatory elements that are located at shared loop anchors across the three deep datasets and at loop anchors detected only with the FA+DSG-MNase protocol. We used candidate cis-Regulatory element (cCREs) predictions (predictions were made for H1-hESC and HFFc6) from ENCODE which uses DNase hypersensitive regions, CTCF, H3K4me3 and H3K27ac ChIP seq data for predicting and annotating cCREs with and without CTCF binding sites (Buenrostro et al. 2013). We found that the majority of the shared anchors had cCREs but only a small part of these cCREs were predicted promoter or enhancer elements without CTCF site (5.2% for HFFc6, 9.8% for H1-ESC; Figure 2.10 f, Figure 2.11 g). In contrast, around half of the FA-DSG-MNase-specific anchors had predicted cCREs and for this subset the number of predicted promoter or enhancer elements without CTCF site is higher compared to loop anchors detected with all protocols (21% for HFFc6, 30% for H1-ESC; Figure 2.10 f, Figure 2.11 g). FA+DSG-DpnII-specific loop anchors show similar enrichments as FA-DSG-MNase-specific anchors.

Finally, we compared the chromatin organization at CTCF-enriched loop anchors with respect to the orientation of the CTCF binding motif. Remarkably, using Cut&Tag or Cut&Run data we found an asymmetric distribution of signal for all factors (Figure 2.10 g), including CTCF (Cut&Tag

data). Both CTCF and cohesin signals are skewed towards the inside of the loop. We noted that the Cut&Tag data was generated with an antibody against the N-terminus of CTCF (Figure 2.10 g). We also analyzed Cut&Run data that was generated with an antibody directed against the C-terminus of CTCF (Figure 2.11 h) and observed signal enrichment skewed at CTCF sites towards the outside of the loop. These observations are consistent with the orientation of CTCF binding to its motif and interactions between the N-terminus of CTCF with cohesin on the inside of the loop (Consortium, Moore, et al. 2020). The stronger enrichment of H3K4me3 and H3K27ac on the inside of the loop is intriguing, but the mechanism of this asymmetry is unknown.

- **Insulation quantification is robust to experimental variations**

Next we investigated chromatin insulation, i.e. the reduced interaction probability across domain boundaries (Li et al. 2020; Hou et al. 2012; Dixon et al. 2012). Loop anchors often form domain boundaries as they represent sites at which cohesin-mediated loop extrusion is blocked. We used the previously described insulation metric that quantifies the frequency of interactions across any genomic locus within a set window size (Nora et al. 2012). We quantified insulation for each 10 kb bin by aggregating interactions across each bin over a 200 kb window. Local minima in this metric represent positions of insulation, i.e. domain boundaries). The local depth of the minimum is a measure for the strength of the boundary. By identifying local minima in the insulation profiles we obtained a set of boundary positions genome-wide.

First, we compared the boundary strength as detected with the deep

datasets obtained with the FA-DpnII, FA+DSG-DpnII and FA+DSG-MNase protocols in HFFc6. We observed that the distribution of the boundary strengths was bimodal: for each dataset we identify a relatively large set of very weak boundaries, and a smaller set of strong boundaries (Figure 2.12 a). Insulation at the weak boundaries was very small, and possibly noise (Figure 2.12 d). Focussing on the strong boundaries, we aggregated insulation profiles at 1) loop anchors detected with each of the three deep datasets (Figure 2.12 b (left)), 2) strong boundaries (Figure 2.12 b (middle)) and 3) loop anchors that are at strong boundaries (Figure 2.12 b (right)). We found that insulation at these elements was very similar for each of the three deep datasets, indicating that the different protocols performed comparably in quantitative detection of strong insulation sites. In general, insulation at strong boundaries was stronger than at loop anchors, possibly because of our stringent threshold for boundary detection.

Second, we investigated whether insulation strength depends on sequencing depth. We compared two biological replicates, one with ~150 M interactions (matrix data, Figure 2.12 c (left)) and the other with 2.5 billion interactions (deep data, Figure 2.12 a) for data obtained with the FA-DpnII, FA+DSG-DpnII and FA+DSG-MNase protocols. Deeper sequencing reduced the relative number of weak boundaries, suggesting these were due to noise (Figure 2.12 a, c (left)). The majority (>85%) of strong boundaries are detected in both deep data and the less deeply sequenced data obtained with the matrix of 12 protocols, and the insulation scores of these shared strong boundaries were highly correlated across all datasets (r >0.80) (Figure 2.12 c

(right)).

Third, we investigated the number and the strength of the boundaries detected using data obtained with the matrix of 12 protocols for HFF, H1-hESC, DE cells and the 9 protocols for HelaS3. Insulation strength at boundaries detected with each protocol was very similar (Figure 2.12 d (right)). We observed the same results for H1-hESC (Figure 2.12 e-h).

We found a positive correlation between boundary strength and the number of protocols that detected that boundary (Figure 2.12 i, j). Focusing on the set of boundaries that were detected by at least half of the protocols we then investigated how insulation varied for data obtained with the matrix of 12 protocols. We found that insulation strength was very similar for data obtained with all protocols (Figure 2.12 k). Similarly, we detected only minor variations in insulation when insulation was aggregated at the set of loop anchors detected by all three deep datasets using data obtained with the matrix of 12 protocols. In summary insulation detection and quantification was robust to variations in protocol (Figure 2.12 l).

Discussion

We present a systematic evaluation of experimental parameters of 3C-based protocols that determine the ability to detect and quantify aspects of chromosome folding. Fragmentation level and cross-linking chemistry determined the ability to detect chromatin loops or compartmentalization in different ways. Loop detection was improved when chromatin was cross-linked with additional (DSG) cross-linking and cut into small fragments (Figure

2.13 a,b for HFFc6 and Figure 2.13 d,e for H1-hESC). Loops detected with such protocols were more enriched for cis elements like enhancers and promoters as compared to sets of loops detected with conventional Hi-C. However, this comes at a cost of a reduced ability to quantitatively detect compartmentalization in cis and in trans (Figure 2.13 c for HFFc6 and Figure 2.13 f for H1-hESC). Quantification of compartmentalization improved with longer fragments such as those produced with DpnII in conventional Hi-C. Compartment strength improved further with additional cross-linkers or when chromatin was digested in even longer fragments, e.g. using HindIII (Figure 2.13 g for 12 protocols performed in HFF).

Fragmentation level and cross-linking chemistry determine assay performance by affecting the level of noise due to random ligation events in datasets (Yang et al. 2017). We find that smaller fragments result in more random ligation events. Possibly the number of cross-links per fragment is low for small fragments, leading to a higher mobility and increased random ligations during the assay. Random ligation events diminish when additional cross-linking is used or when chromatin is fragmented into larger fragments. This results in a decrease in inter-chromosomal interactions and steeper $P(s)$ plots. Improved signal-to-noise ratios allowed better detection of loops, compartments, and more bona fide inter-chromosomal interactions.

Detection of compartmentalization strength is improved when protocols are used that produce relatively long fragments and include additional cross-linking. Possibly, compartmental interactions are more difficult to capture than

looping interactions that are closely held together by cohesin complexes. Recently, we found that interfaces between compartment domains appear relatively unmixed (Tavares-Cadete et al. 2020). Longer fragments or extra cross-linkers may be required to more efficiently capture contacts across these interfaces. Interestingly, cell-type-specific differences in strength of compartmentalization are only observed with some protocols. Conventional Hi-C (FA+DpnII) suggests that compartmentalization strength is quite similar in H1-ESCs, Hela S3, DE cells, and HFF. However, when Hi-C is performed with additional cross-linkers and/or with restriction enzymes that produce longer fragments, HFF and Hela S3 have stronger compartmentalization, while compartmentalization strength for H1-ESCs and DE are unaffected. This suggests that quantitative differences in cell type-specific chromosome organization can be missed or underestimated depending on the 3C-based protocol.

Hi-C protocols with additional cross-linkers allowed detection of many thousands more loops than conventional Hi-C (FA+DpnII), while also detecting strong compartmentalization. Depending on the objective of their study, investigators may choose different protocols: Micro-C for loop detection, or Hi-C for detection of compartmentalization. In contrast to restriction enzyme digestion, MNase digestion requires optimization of conditions, which may not always be possible when cell numbers are low, or samples are rare. Micro-C is not applicable for organisms without nucleosomes.

The very deeply sequenced Hi-C, Micro-C and Hi-C 3.0 datasets we produced for H1-ESCs and HFFc6 cells will be useful resources for the chromosome folding community given that these cell lines are widely used for method benchmarking and analysis by the 4D nucleome project (Dekker et al. 2017). Further, the comprehensive collection of chromatin interaction data generated with the matrix of twelve 3C-based protocol variants for each cell line can also be a valuable resource for benchmarking computational methods for data analysis given their different cross-linking distances and chemistry, fragment lengths and noise levels.

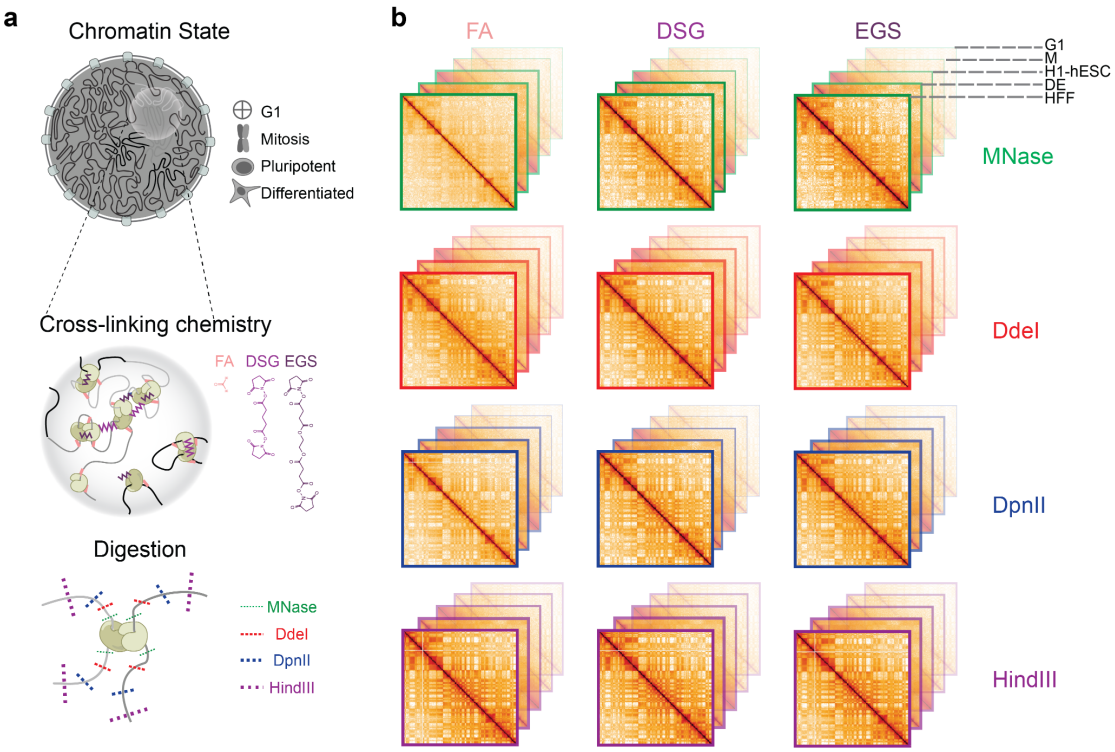

Figure 2.1: Outline of the experimental design.

a. Experimental design for conformation capture on cells with the indicated chromatin states (left), using various cross-linking chemistries (middle) and digestion methods (right).

b. Representation of interaction maps generated by combinations of experimental conditions depicted in panel a.

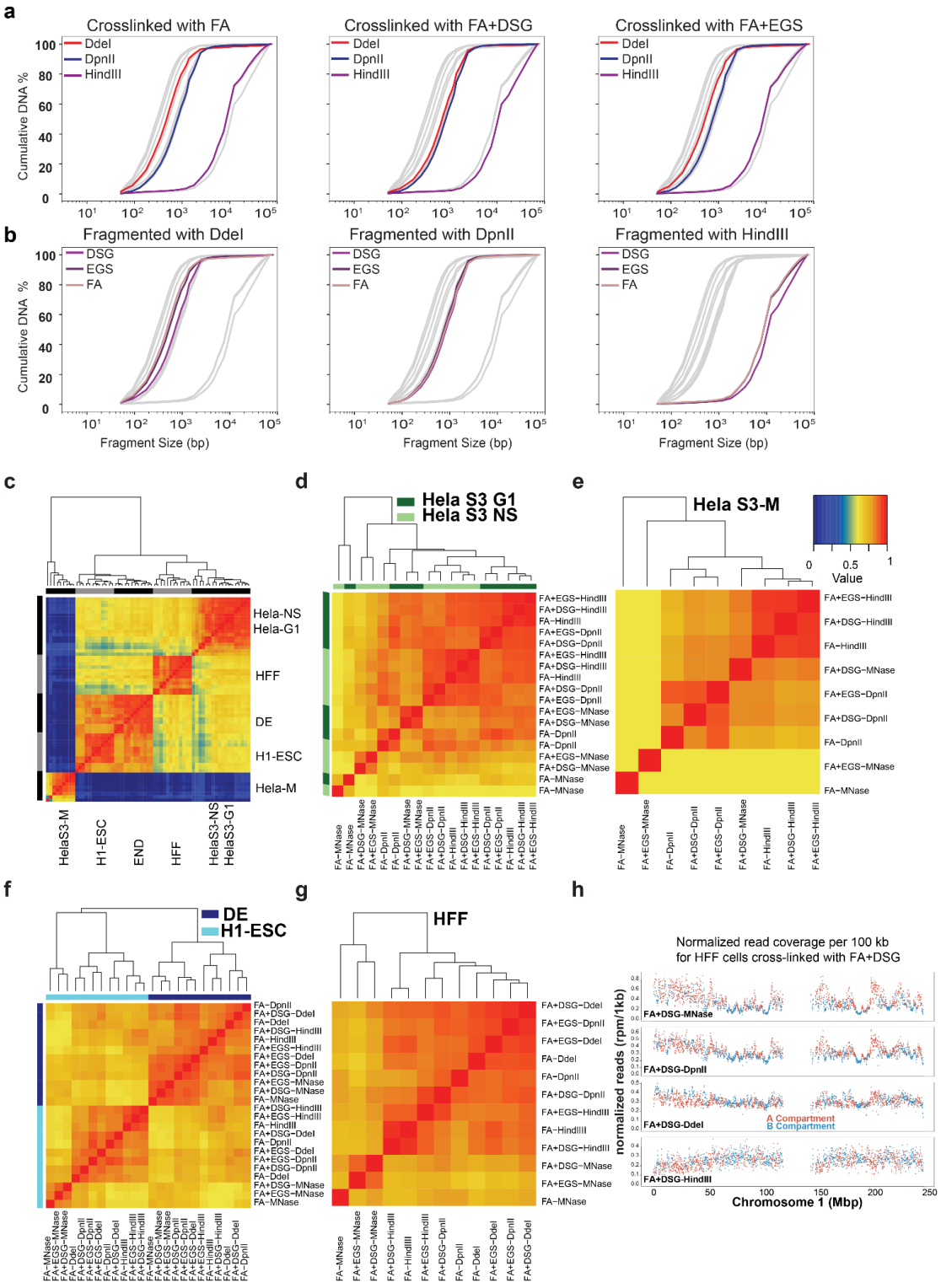
a
Crosslinked with FA
Crosslinked with FA+DSG
Crosslinked with FA+EGS
DdeI
DpnII
HindIII
Cumulative DNA %
b
Fragmented with DdeI
Fragmented with DpnII
Fragmented with HindIII
DSG
EGS
FA
Fragment Size (bp)
c
Hela-NS
Hela-G1
HFF
DE
H1-ESC
Hela-M
HelaS3-M
H1-ESC
END
HFF
HelaS3-NS
HelaS3-G1
d
Hela S3 G1
Hela S3 NS
e
Hela S3-M
Value
f
DE
H1-ESC
g
HFF
h
Normalized read coverage per 100 kb
for HFF cells cross-linked with FA+DSG
normalized reads (rpm/1kb)
FA+DSG-MNase
FA+DSG-DpnII
FA+DSG-DdeI
FA+DSG-HindIII
A Compartment
B Compartment
Chromosome 1 (Mbp)

Figure 2.2: DNA fragmentation and hierarchical clustering of distance corrected correlation (HiCRep)

a,b. Cumulative distribution of the lengths of fragmented DNA obtained from fragment analyzer data in HFF cells stratified for different cross-linkers (a) and restriction enzymes (b). Gray lines indicate all datasets, colored lines indicate data obtained with the indicated nuclease/cross-linkers.

c-g. Hierarchical clustering of HiCRep correlations for: all protocols comparing cell states (c), synchronized Hela S3 G1 cells (dark green) and non-synchronized Hela S3 cells (light green) (d), synchronized Hela S3 mitotic cells (e), H1-hESC and H1-hESC derived DE cells (f), 12 protocols applied to HFF cells (g). One color key is indicated for all of the heatmaps.

h. Genome coverage of data generated using MNase, DdeI, DpnII and HindIII. The read density was normalized to reads per million, separated by the coverage in A and B compartments (Methods).

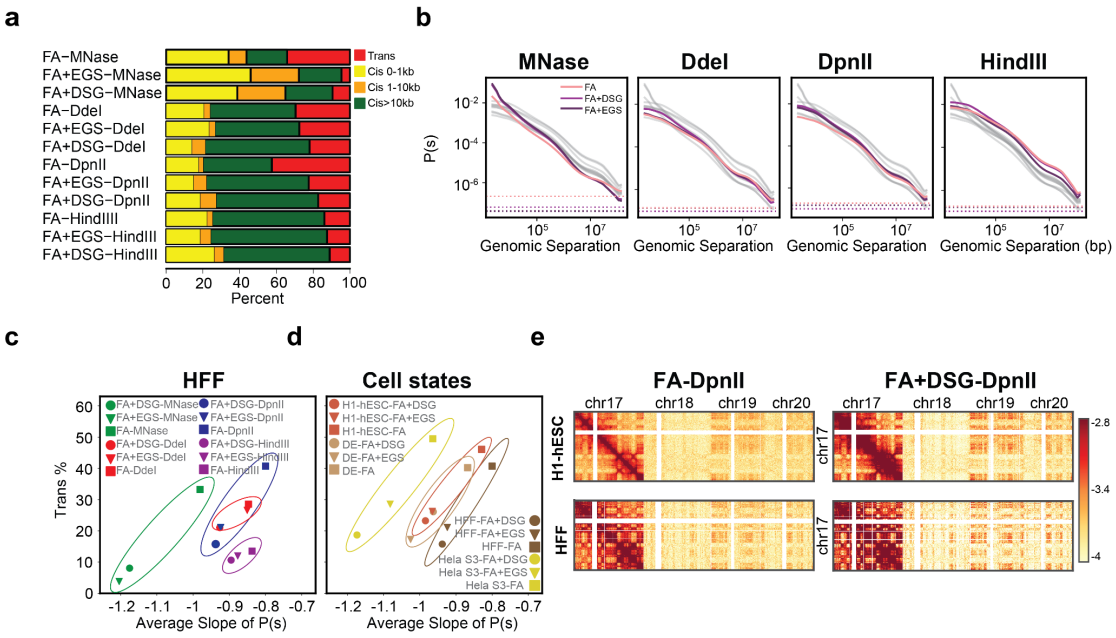

Figure 2.3: Distance dependent interaction frequency and the number of inter-chromosomal interactions change across protocols that use various enzyme and cross-linker combinations.

Figures a, b and c created using 12 protocols that are performed in HFF cells.

a. The number of valid pairs in each of the 12 protocols categorized by genomic distance.

b. Distance dependent contact probability as detected by the set of 12 protocols split by used nucleases. Gray lines indicate all datasets, colored lines indicate data obtained with the indicated nuclease.

c. The relationship between the percentage of trans interactions and the average slope of the distance dependent contact probability separated by cross-linker and enzyme combinations. Oval lines group datasets obtained with the same nuclease.

d. The relationship between percentage of trans interactions and average slope of the distance dependent contact probability separated by cell

type. Only experiments in which chromatin cross-linked with FA, FA+DSG or FA+EGS and digested with DpnII are shown.

e. Interaction map (log transformed) of chromosome 17 with chromosomes 17, 18, 19 and 20 for FA or FA+DSG cross-linking and DpnII digestion, in H1-hESC and HFF. Total trans interactions for FA-DpnII protocols in H1-hESC: 47.7%, HFF: 42.5% and for FA+DSG-DpnII protocols in H1-hESC: 25% and HFF: 17.3%.

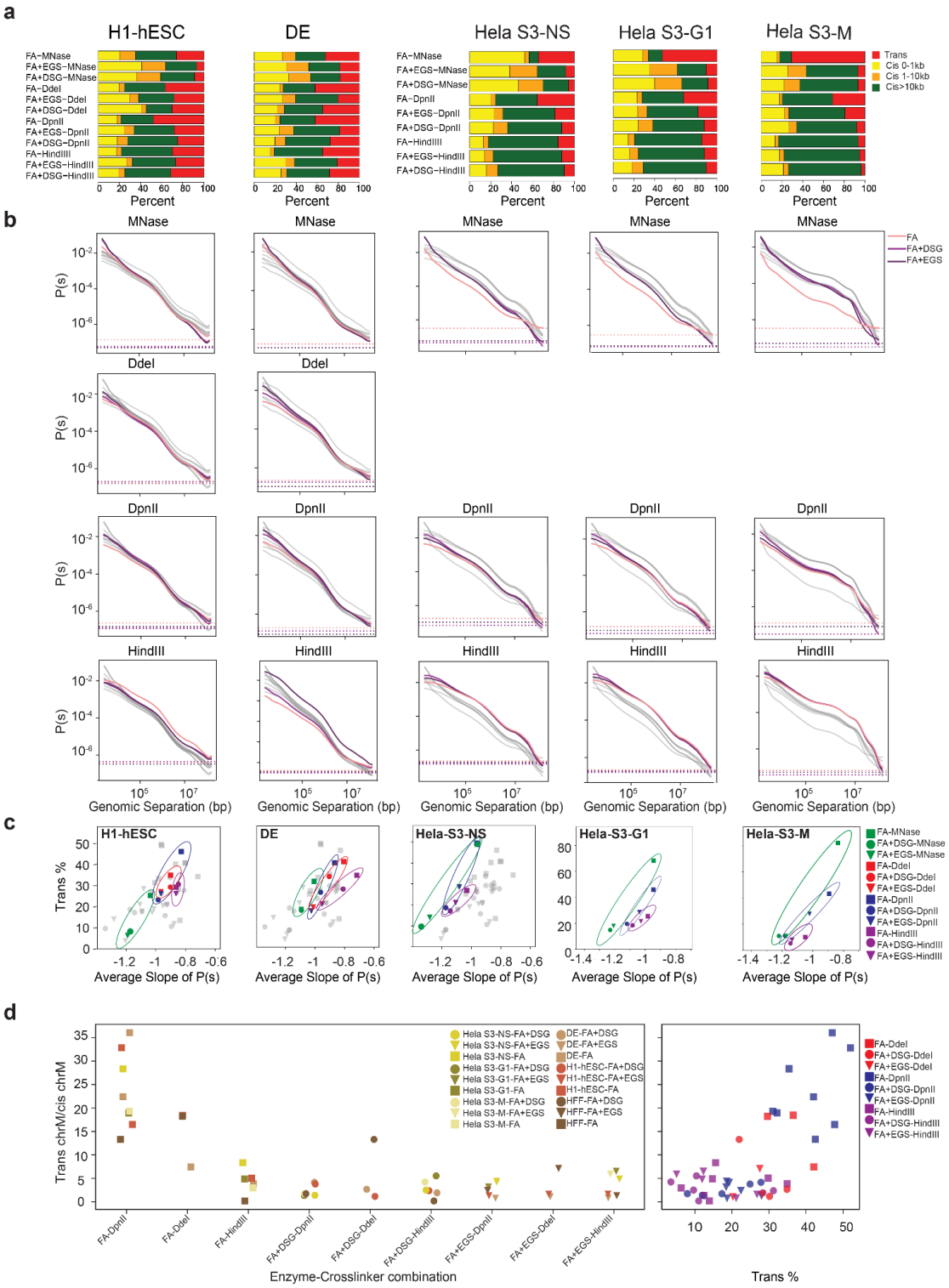
a
H1-hESC
DE
Hela S3-NS
Hela S3-G1
Hela S3-M
FA−MNase
FA+EGS−MNase
FA+DSG−MNase
FA−DdeI
FA+EGS−DdeI
FA+DSG−DdeI
FA−DpnII
FA+EGS−DpnII
FA+DSG−DpnII
FA−HindIII
FA+EGS−HindIII
FA+DSG−HindIII
Percent
Trans
Cis 0-1kb
Cis 1-10kb
Cis>10kb
b
MNase
DdeI
DpnII
HindIII
P(s)
Genomic Separation (bp)
FA
FA+DSG
FA+EGS
c
H1-hESC
DE
Hela-S3-NS
Hela-S3-G1
Hela-S3-M
Trans %
Average Slope of P(s)
FA-MNase
FA+DSG-MNase
FA+EGS-MNase
FA-DdeI
FA+DSG-DdeI
FA+EGS-DdeI
FA-DpnII
FA+DSG-DpnII
FA+EGS-DpnII
FA-HindIII
FA+DSG-HindIII
FA+EGS-HindIII
d
Trans chrM/cis chrM
Hela S3-NS-FA+DSG
Hela S3-NS-FA+EGS
Hela S3-NS-FA
Hela S3-G1-FA+DSG
Hela S3-G1-FA+EGS
Hela S3-G1-FA
Hela S3-M-FA+DSG
Hela S3-M-FA+EGS
Hela S3-M-FA
DE-FA+DSG
DE-FA+EGS
DE-FA
H1-hESC-FA+DSG
H1-hESC-FA+EGS
H1-hESC-FA
HFF-FA+DSG
HFF-FA+EGS
HFF-FA
FA-DpnII
FA-DdeI
FA-HindIII
FA+DSG-DpnII
FA+DSG-DdeI
FA+DSG-HindIII
FA+EGS-DpnII
FA+EGS-DdeI
FA+EGS-HindIII
Enzyme-Crosslinker combination
Trans %

Figure 2.4: Distance dependent interaction frequency and the number of inter-chromosomal interactions change across protocols that use various enzyme and cross-linker combinations

a. The number of valid pairs in each of the 12 protocols applied to H1-hESC, DE, Hela S3-NS, Hela S3-G1 and Hela S3-M cells partitioned by genomic distances.

b. Distance dependent contact probability of 12 protocols ordered as in (a), partitioned by fragmenting nucleases used (gray lines indicate all datasets, colored lines indicate datasets generated with the nucleases indicated for each plot).

c. The relationship between the trans percent and the average slope of the distance dependent contact probability for the 12 protocols ordered as in Figure 2.4

d. Quantification of protocol introduced noise as defined by inter-mitochondrial interactions (chrM with chr1-22), normalized by intra-mitochondrial (chrM with chrM) interactions.

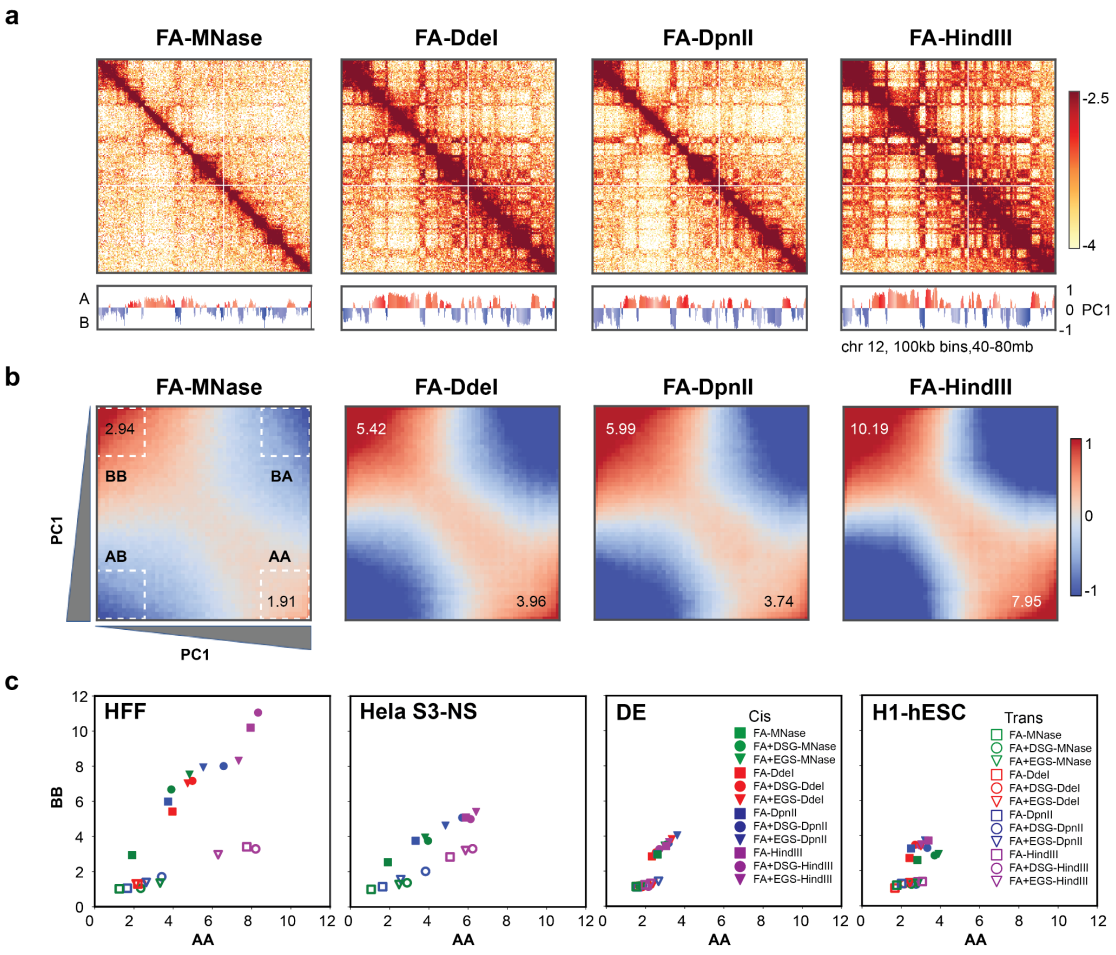

Figure 2.5: Cross-linking the chromatin with DSG or EGS and digesting it with HindIII strengthen compartment signals.

a. Interaction maps (log transformed) for HFF cells obtained with protocols where the chromatin is cross-linked with FA only and digested with MNase, DdeI, DpnII and HindIII, respectively. Values of the first eigenvector for all genomic regions are displayed below the figure.

b. Saddle plots of the genome-wide interaction maps of the data shown in Figure 2.5 a . A-A and B-B compartment signals in cis get stronger with increasing fragment size.

c. Quantification of the compartment strength using saddle plots of cis and trans interactions for 12 protocols applied to HFF cells, 9 protocols

to Hela S3 NS, 12 protocols to DE, 12 protocols to H1-hESC. Y-axis represents the quantification for the strongest 20% of B-B and x-axis represents the quantification of the strongest 20% of A-A interactions divided by the sum of corresponding A-B interactions.

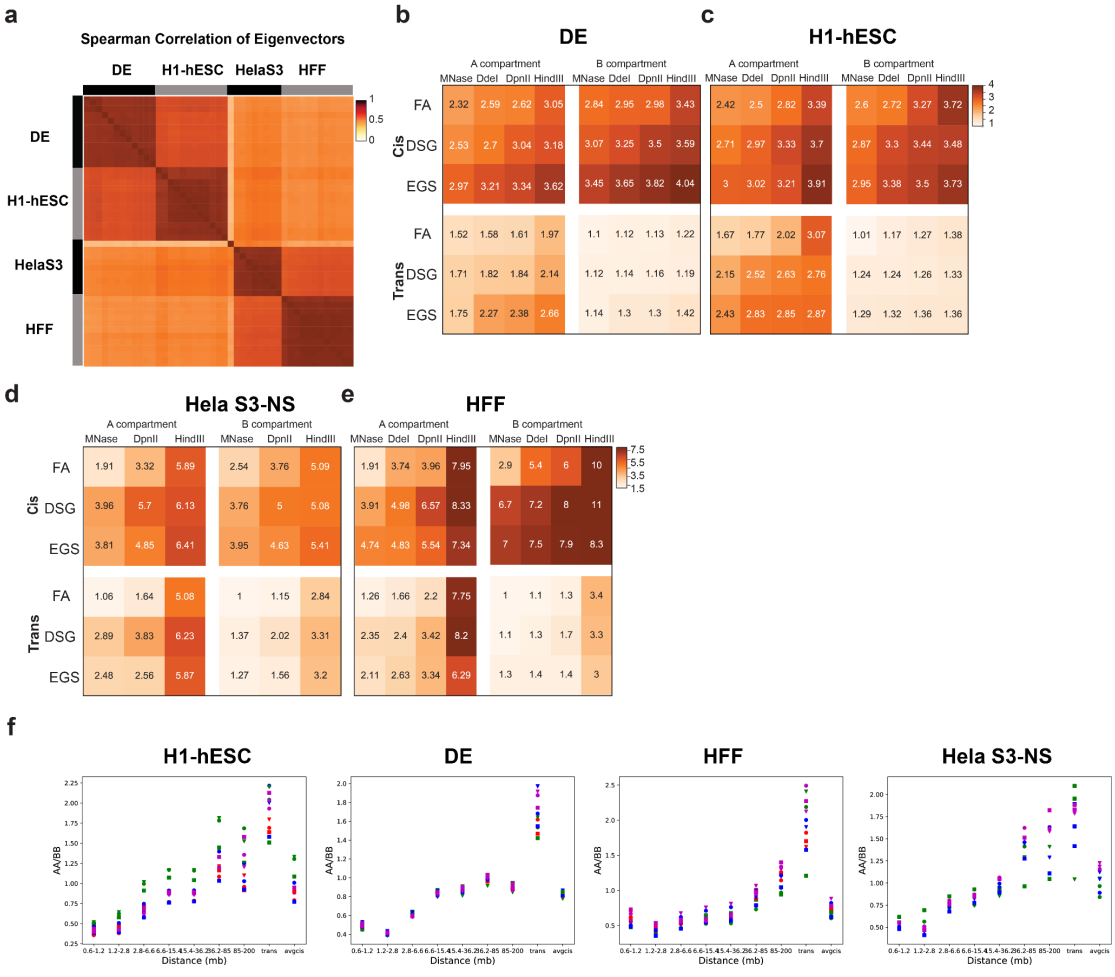

Figure 2.6: Compartment identity is robust for protocol variation but the strength of the compartments differs between protocols

a. Hierarchical clustering of Spearman correlations of Eigenvectors (E1) for 63 protocols. Clustering shows strong correlations between compartments from data obtained with varying protocols applied to the

same cell types and weaker correlations for data obtained with the same protocols applied to different cell types.

b-e. A-A and B-B compartment strength of saddle plots for fixation versus enzyme stratified by cell state: DE (b), H1-hESC (c), Hela S3-NS (d), HFF (e). For each cell type, saddle plot quantification was done for cis and trans reads separately.

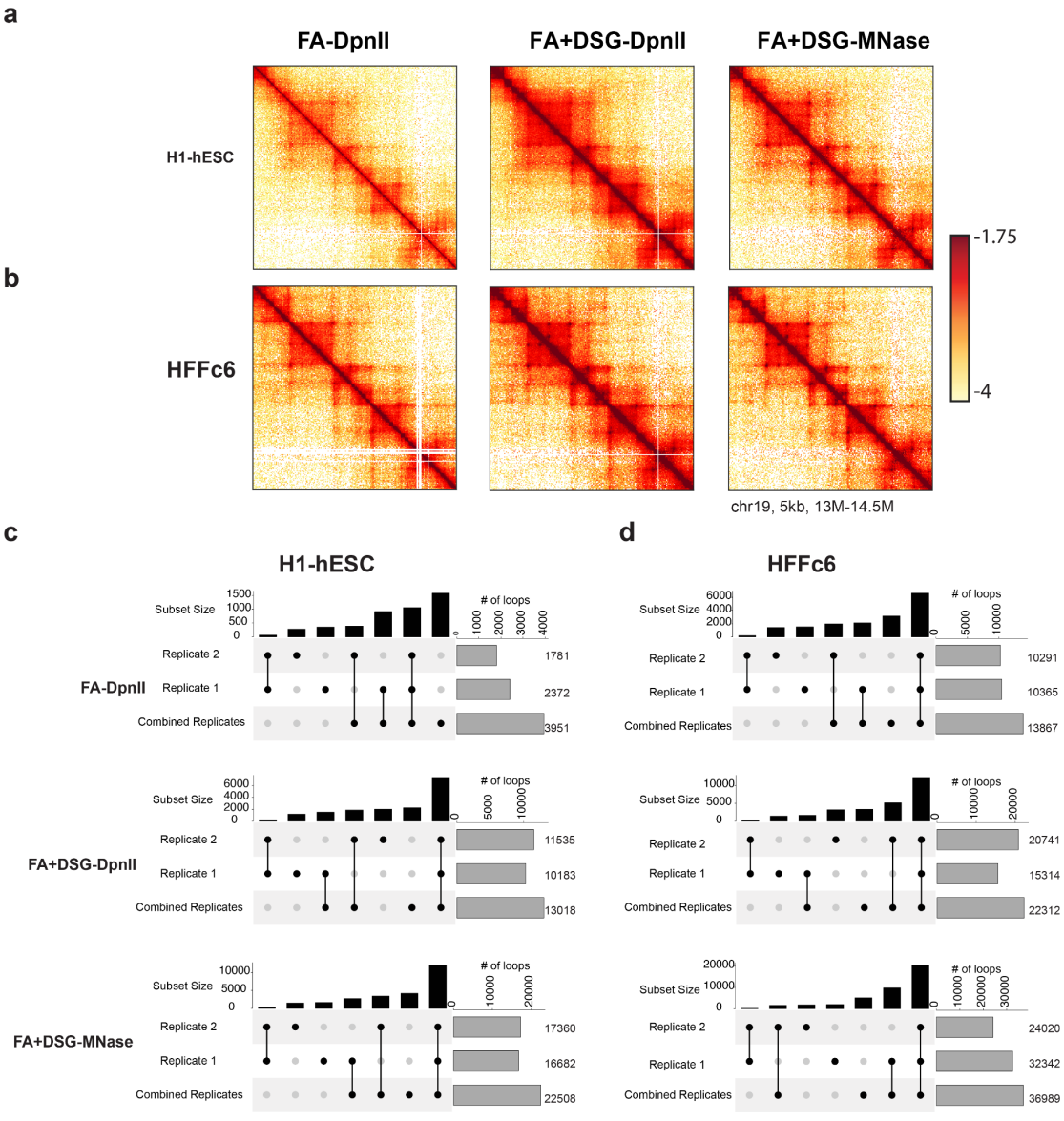

Figure 2.7: Chromatin loops are more consistent between replicates that are cross-link the chromatin with FA+DSG

a. Interaction heatmaps (log transformed) of experiments for H1-ESC cells obtained from the following cross-linker-enzyme combinations (from left to right): FA-DpnII, FA+DSG-DpnII and FA+DSG-MNase.

b. Interaction heatmaps (log transformed) of experiments for HFFc6 cells obtained from the following cross-linker-enzyme combinations (from left to right): FA-DpnII, FA+DSG-DpnII and FA+DSG-MNase.

c. Upset plots of loops detected with different replicates for H1-hESC show: 1) total number of loops detected in Replicate 1, Replicate 2 and merged replicates on the right side (gray bars), 2) number of loops detected in the one, two or three experiments shown in black bars. Loops found with only one or multiple experiments are highlighted and connected with black dots. Here Upset plots investigate the consistency of loops between each of the replicates and combined replicates for FA-DpnII, FA+DSG-DpnII and FA+DSG-MNase in H1-hESC.

d. Upset plots (as explained in Figure 2.7 c) of loops detected with different replicates for HFFc6 cells.

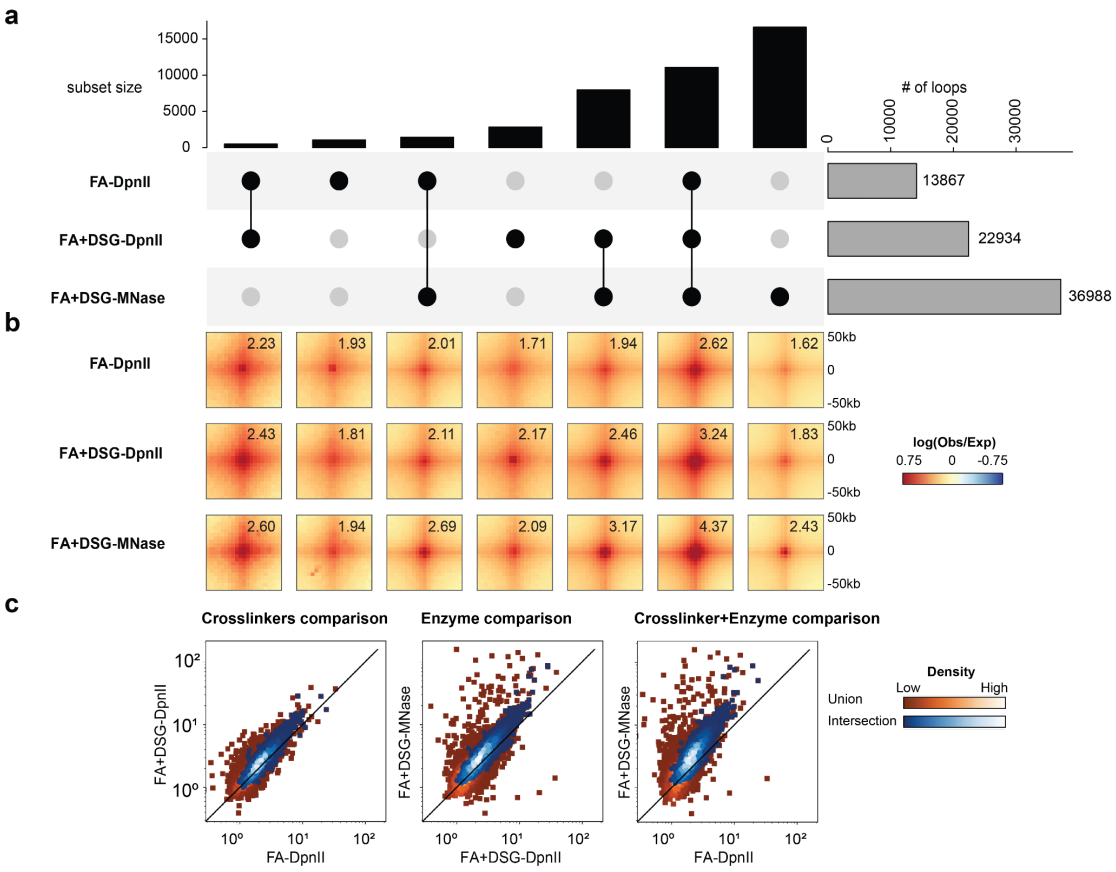

Figure 2.8: Chromatin loops are better detected in experiments with fine fragmentation and DSG cross-linking

a. Upset plot of loops detected in protocols performed using HFFc6 shows the 1) total number of loops detected in FA-DpnII, FA+DSG-DpnII and FA+DSG-MNase on the right side (gray bars), 2) number of loops detected in one, two or three of these protocols shown in black bars. Loops found with only one or multiple protocols are highlighted and connected with black dots.

b. The pileups of the loops in HFFc6 for every set of loops shown in Figure 2.8 a. Numbers in each pileup represent the signal enrichment at the loop compared to local background. See methods for quantification of loop strength.

c. Scatter plots for the relative strengths of individual loops between pairs of protocols for HFFc6 cells. The pairs are (from left to right): FA-DpnII v/s FA+DSG-DpnII (different crosslinking with the same enzyme), FA+DSG-DpnII v/s FA+DSG-MNase (different enzymes with same crosslinking) and FA-DpnII v/s FA+DSG-MNase (different enzymes and different crosslinking). Loop strengths were calculated the same as in panel b but for individual looping interactions. Scatter plots display two sets of looping interactions - the union of all locations from the three protocols (red squares) and interaction of all locations from the three protocols (blue circles). The color scale represents the density of loop interactions present at those strengths.

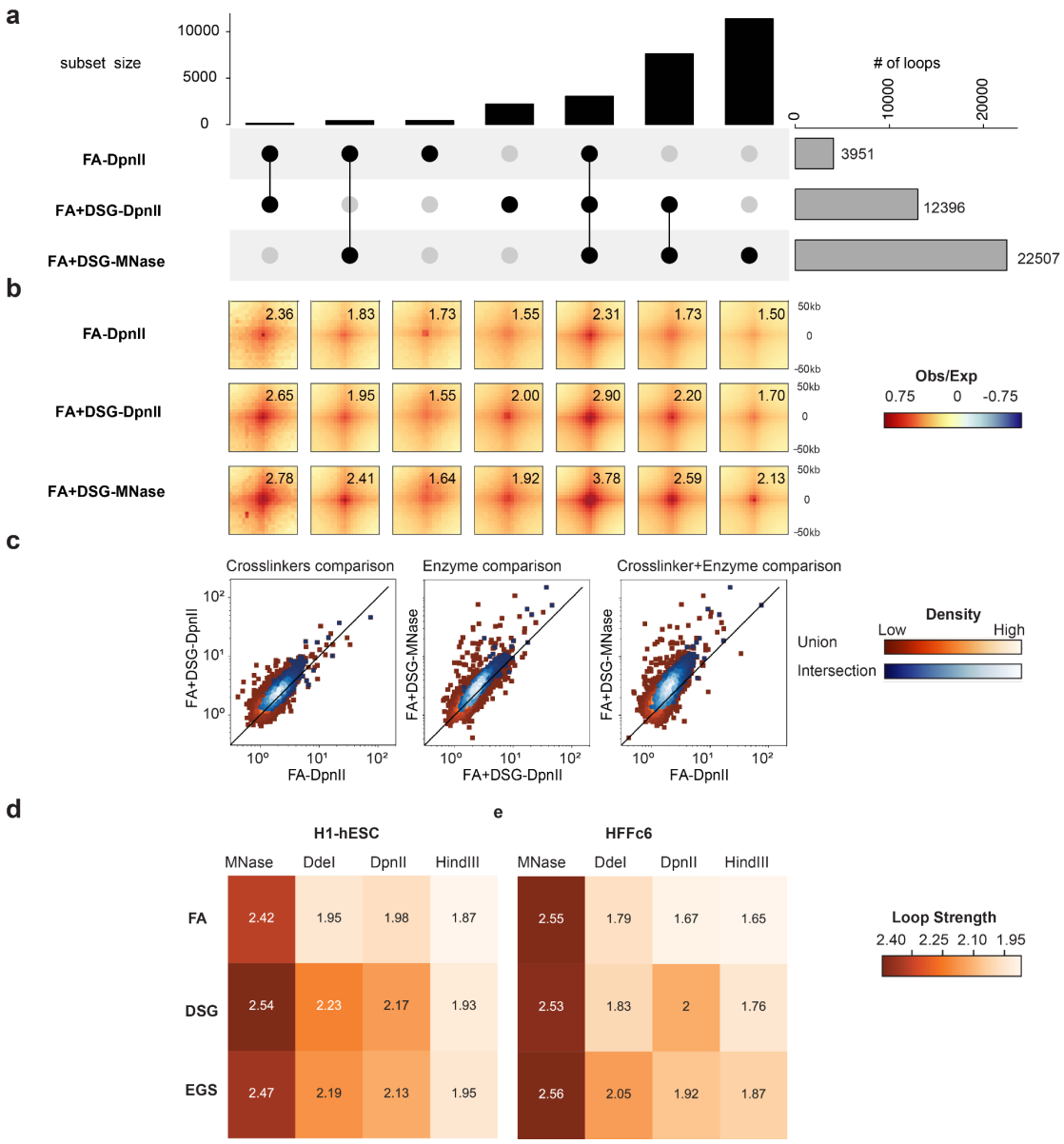

Figure 2.9: Chromatin loops are better detected in experiments with fine fragmentation and DSG cross-linking

a. Upset plot of loops detected in protocols performed in H1-hESC showing 1) total number of loops detected in FA-DpnII, FA+DSG-DpnII and FA+DSG-MNase (gray bars on the right), 2) number of loops detected in one, two or three of these protocols shown as black bars. The combinations highlighted with connected black dots.

b. Pileups of the loops detected in all combinations for H1-hESC protocols shown in Figure 2.9 a. Quantification of the loop strength was done taking an observed/expected interaction matrix (50x50 kb) around the loop and then normalizing the loop intensity to its local background (see methods).

d. Scatter plots with relative strengths of individual loops between pairs of protocols for H1-hESC; from left to right: FA-DpnII v/s FA+DSG-DpnII (different cross-linking with the same enzyme), FA+DSG-DpnII v/s FA+DSG-MNase (different enzymes with same cross-linking) and FA-DpnII v/s FA+DSG-MNase (different enzymes and different cross-linking). Loop strengths were calculated as in panel b but for individual looping interactions. Scatter plots were drawn for two sets of looping interactions (1) the union of all locations from the three protocols (red squares) and (2) interaction of all locations from the three protocols (blue circles). Color scale represents the density of looping interactions.

d,e. Quantification of aggregated loop strengths from the matrix of 12 protocols described in Fig 2.1 a for H1-hESC cells (d), and HFFc6 cell (e). Pileups represent looping interactions detected across all three deep protocols (FA-DpnII, FA+DSG-DpnII and FA+DSG-MNase) in each cell type.

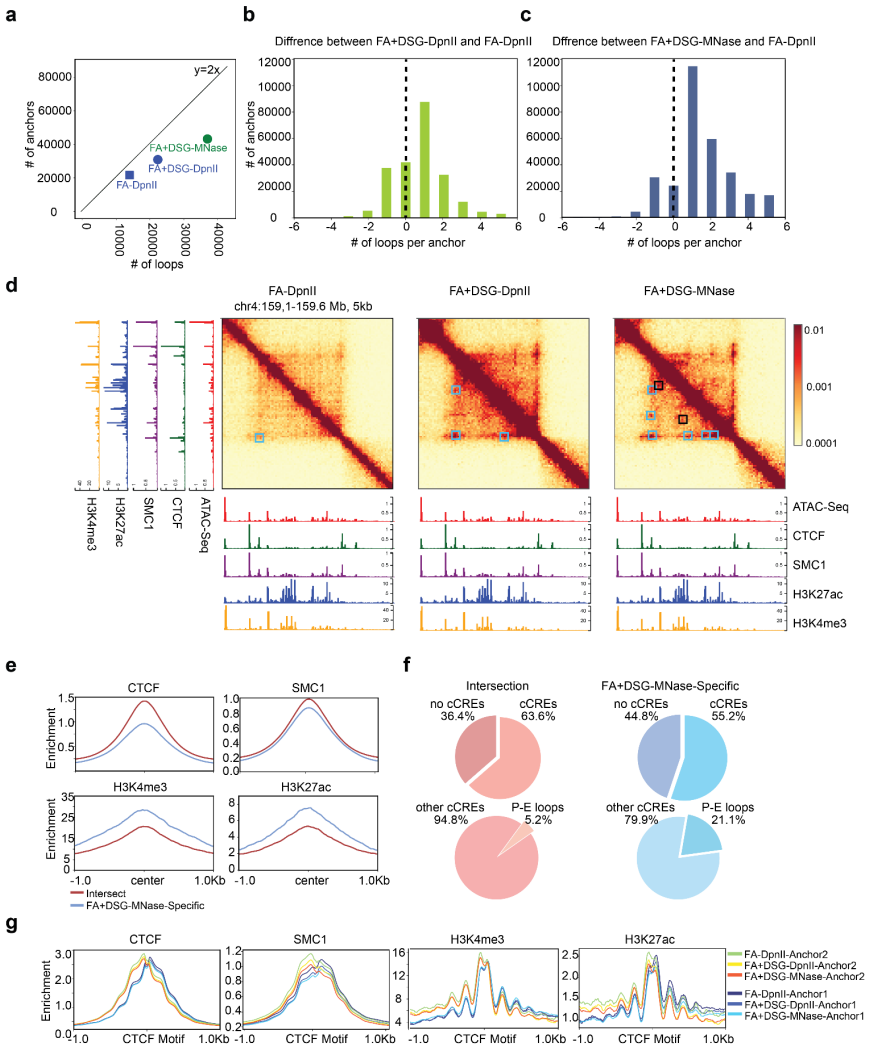

Figure 2.10: Characterization of interactions and chromatin features of loop anchors detected with different protocols.

a. The number of loops detected in HFFc6 (x-axis) plotted against the number of loop anchors (y-axis). For y=2x depicts the expected relationship when each anchor is engaged in only 1 loop.

b,c. The number of FA-DpnII loops subtracted from the number of FA+DSG-DpnII (b) or FA+DSG-MNase (c) loops detected at the same

anchors; the union list of the plotted protocols was used.

d. Interaction maps (linear scale) for 3 protocols applied to HFFc6 cells and CUT$Run/CUT&Tag data for CTCF, SMC1, H3K4me3 and H3K27ac. Cyan squares highlight the loop anchors detected with all three protocols. Black squares indicate loop anchors detected with FA+DSG-MNase only.

e. CTCF, SMC1, H3K4me3 and H3K27ac enrichments at loop anchors detected by all protocols (union) or FA+DSG-MNase alone in HFFc6. Open chromatin regions within anchor coordinates were used to center average enrichments.

f. Candidate Cis Regulatory elements (cCREs) detected in common and FA+DSG-MNase specific loop anchors from Figure 5e (top) and stratified percentage of Promoter-Enhancer cCREs without CTCF enrichment (bottom).

g. Enrichment of CTCF, SMC1, H3K4me3 and H3K27ac separated between the left (Anchor1) and right (Anchor 2) anchor for loop anchors detected in HFFc6 using FA-DpnII, FA+DSG-DpnII or FA+DSG-MNase.

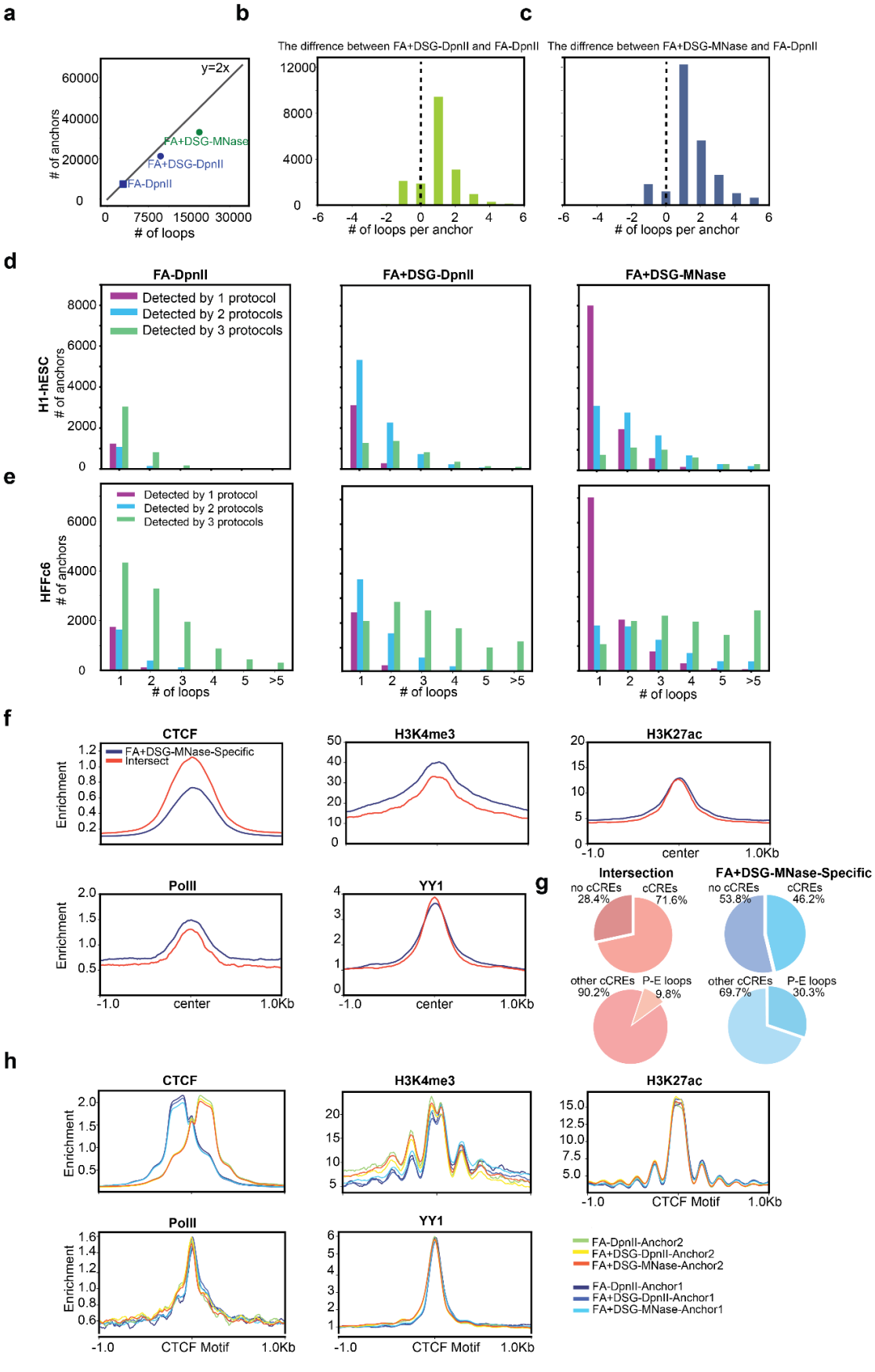
a
b
c
y=2x
FA+DSG-MNase
FA+DSG-DpnII
FA-DpnII
# of anchors
# of loops
The diffrence between FA+DSG-DpnII and FA-DpnII
The diffrence between FA+DSG-MNase and FA-DpnII
# of loops per anchor
d
FA-DpnII
FA+DSG-DpnII
FA+DSG-MNase
Detected by 1 protocol
Detected by 2 protocols
Detected by 3 protocols
H1-hESC
# of anchors
e
HFFc6
# of loops
f
CTCF
FA+DSG-MNase-Specific
Intersect
Enrichment
H3K4me3
H3K27ac
PolII
YY1
-1.0
center
1.0Kb
g
Intersection
FA+DSG-MNase-Specific
no cCREs
cCREs
28.4%
71.6%
53.8%
46.2%
other cCREs
P-E loops
90.2%
9.8%
69.7%
30.3%
h
CTCF Motif
FA-DpnII-Anchor2
FA+DSG-DpnII-Anchor2
FA+DSG-MNase-Anchor2
FA-DpnII-Anchor1
FA+DSG-DpnII-Anchor1
FA+DSG-MNase-Anchor1

Figure 2.11: Characterization of interactions and chromatin features of loop anchors detected with different protocols.

a. The number of loops detected in H1-hESC (x-axis) plotted against the number of loop anchors (y-axis). y=2x line shows the expected relationship between loops and loop anchors when each anchor is engaged in only 1 loop.

b,c. In H1-hESC, the number of FA-DpnII loops subtracted from the number of FA+DSG-DpnII loops (b) or the number of FA+DSG-MNase loops (c) detected at the same anchors. The union of loop calls from both of the plotted protocols was used here.

d. Histograms of valencies of loop anchors detected in H1-hESC. Each panel represents loop anchors found in the given protocol (from left to right, FA-DpnII, FA+DSG-DpnII, FA+DSG-MNase). For each protocol, the anchors were further stratified into three categories. The leftmost panel (FA-DpnII) is used as a guiding example. The categories are: anchors detected by 1 protocol (FA-DpnII), anchors detected in 2 protocols (FA-DpnII and either FA+DSG-DpnII or FA+DSG-MNase) and anchors detected in all 3 protocols (FA-DpnII and FA+DSG-DpnII and FA+DSG-MNase).

e. The same plots as shown in Figure 2.11 d but generated using HFFc6 cells.

f. The comparison of CTCF, H3K4me3, H3K27ac, YY1 and PolII enrichments at loop anchors centered at open chromatin regions. Open chromatin regions (as quantified by ATAC Seq) that are located within the anchor coordinates are used to center the average enrichments. Anchors detected by all protocols and

FA+DSG-MNase-specific anchors in H1-hESC.

g. Top row: candidate Cis Regulatory elements (cCREs) in common (left) and FA+DSG-MNase specific loop anchors (right), as specified in Figure 2.11 f detected in H1-hESC. Bottom row: the percentage of these cCREs for Promoter-Enhancer elements without CTCF enrichment.

h. The enrichment of CTCF, H3K4me3, H3K27ac, YY1 and PolII in left (Anchor1) and right (Anchor 2) anchors centered by CTCF sites. Loop anchors present in FA-DpnII, FA+DSG-DpnII or FA+DSG-MNase applied to H1-hESC.

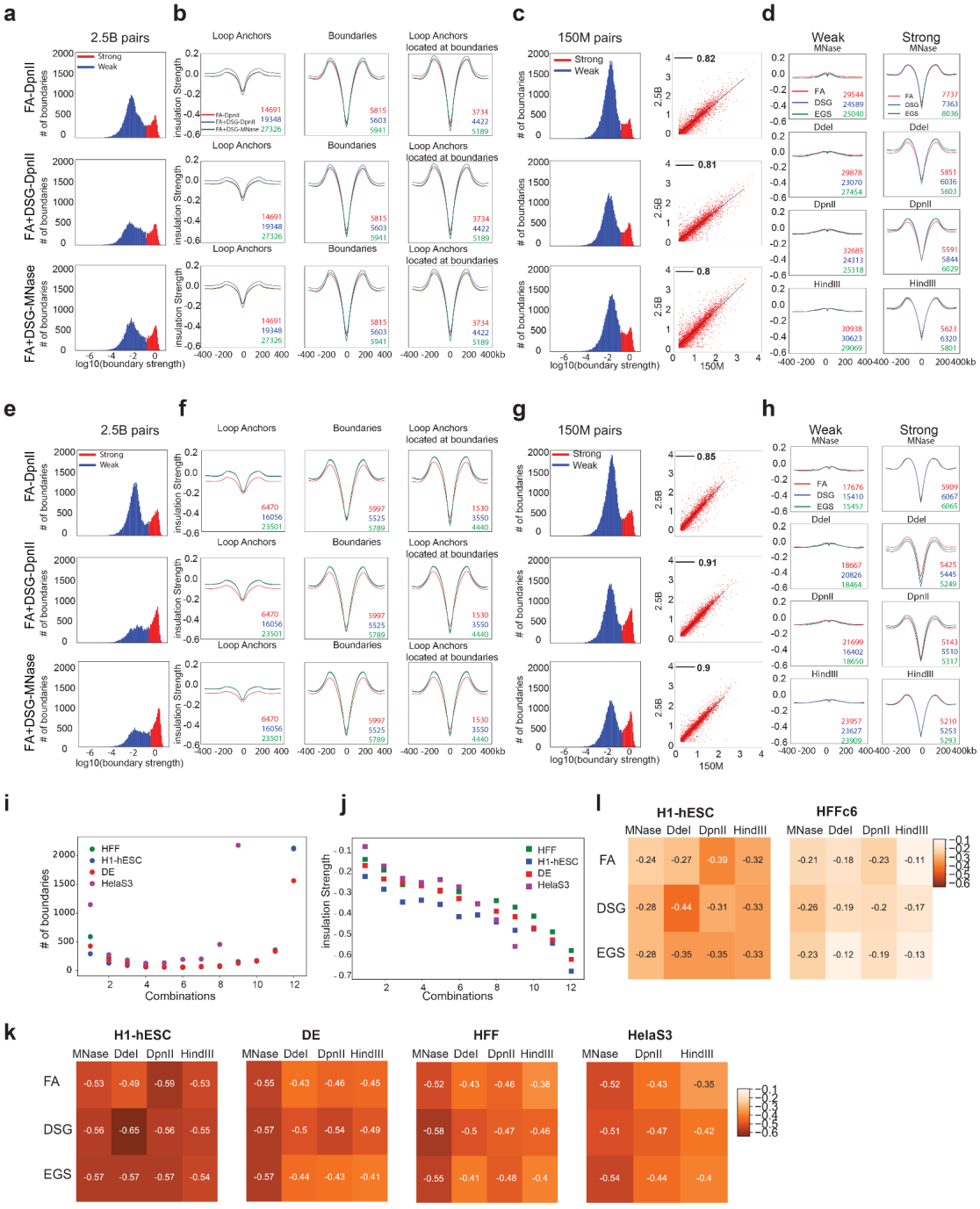

Figure 2.12: Insulation boundaries show modest differences across experimental variations

a. Boundary strength distribution (log scale) for HFFc6 deep data (~2.5 B valid pairs) from FA-DpnII (top) , FA+DSG-DpnII (middle) and

FA+DSG-MNase (bottom). Boundaries are classified as weak (blue) or strong (red) based on boundary strength (see methods).

b. Pileups in FA-DpnII (top row), FA+DSG-DpnII (middle row) and FA+DSG-MNase (bottom row) for aggregate insulation scores at loop anchors (left), strong insulation boundaries (middle) and loop anchors colocalizing with strong insulation boundaries (right) as detected in deeply sequenced libraries: FA-DpnII (red), FA+DSG-DpnII (blue) and FA+DSG-MNase (green).

c. The effect of sequencing depth on boundary strength. Left panel shows the boundary strength distribution of matrix data (Fig. 2.1a, ~150 M valid pairs) for FA-DpnII (top) , FA+DSG-DpnII (middle) and FA+DSG-MNase (bottom) applied to HFFc6 cells. An excess of weak boundaries is observed when comparing to the equivalent deeply sequenced library as shown in Figure 2.12 a. Focusing on the strong boundaries, the right panel shows a strong correlation between deep and matrix data for FA-DpnII, FA+DSG-DpnII and FA+DSG-MNase.

d. Aggregate insulation profile of the boundaries obtained from matrix data for HFFc6 stratified by cross-linkers and nucleases. Boundaries further separated based on their insulation strength as weak and strong insulation.

e-h. H1-hESC data displayed like Figure 2.12 a-d.

i. Distributions of the number of boundaries (y-axis) stratified by the number of protocols in which a given boundary was detected (x-axis). The number of protocols varies between 1 to 12 (Figure 2.1 a).

j. Insulation strength of the boundaries stratified in the same manner as Figure 2.12 i.

k. Mean insulation strength from boundaries detected in at least half of the protocols for various cross-linkers and enzyme combinations of H1-hESC, DE, HFF and Hela S3-NS (see methods).

l. Mean insulation strength of loop anchors that are detected in all three deep protocols (FA-DpnII, FA+DSG-DpnII and FA+DSG-MNase) for both HFFc6 and H1-hESC, averaged for 12 protocols of H1-hESC and HFF.

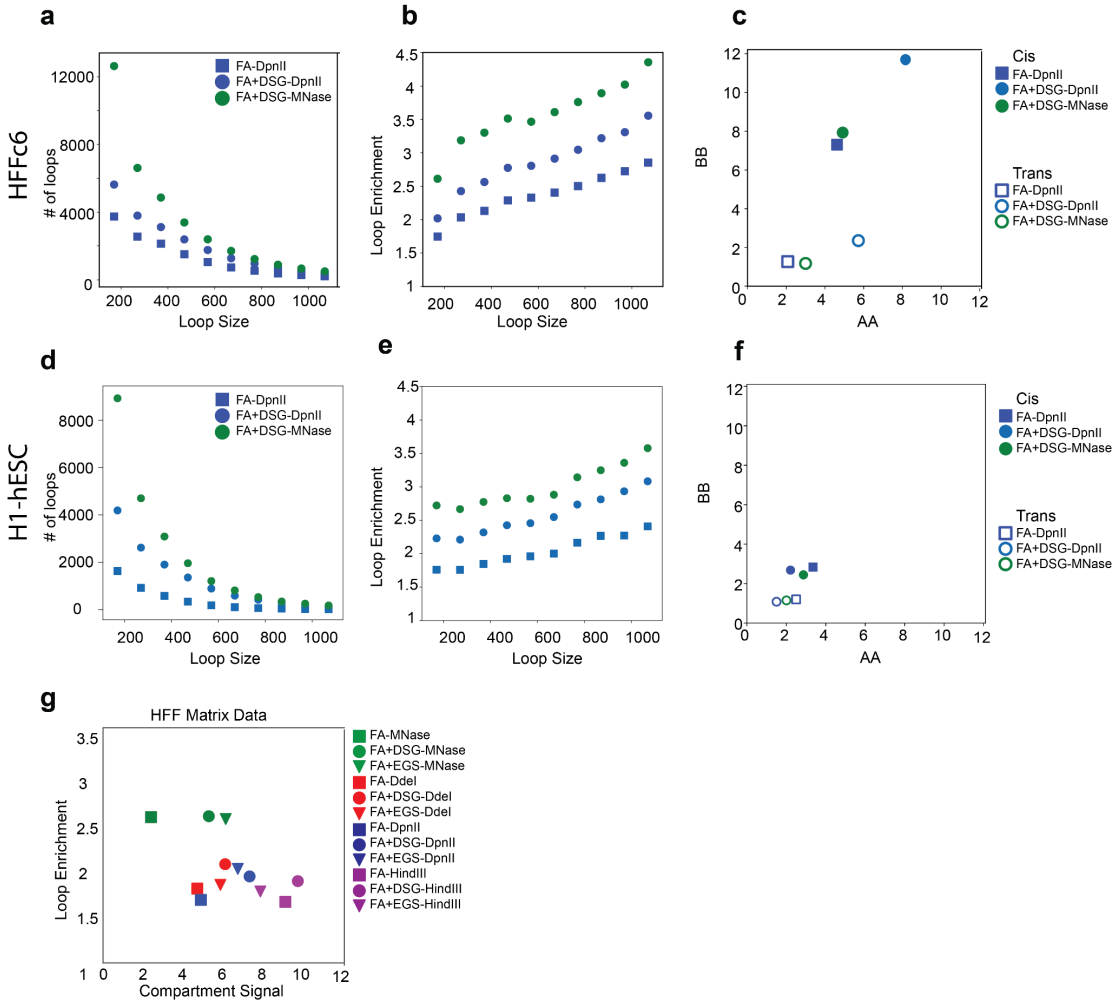

Figure 2.13 : Loop detectability and strength increase when the chromatin is digested with two restriction enzymes while preserving strong compartment signal.

a. The number of loops detected within 100kb intervals (loop size) starting at 70kb in HFFc6. Bin intervals: 70-170 kb, 170-270 kb, 270-370,.....,970-1070 kb.

b. Loop strength of 1,000 loops sampled from 100 kb intervals (Figure 2.13 a) in HFFc6. When less than 1,000 loops were available, loop strengths for available loops were used.

c. A-A (x-axis) and B-B (y-axis) compartment strengths in cis and trans derived from saddle plot analysis of HFFc6.

d.f. Figure 2.13 a-c generated for H1-hESC.

g. Compartment strength obtained from the matrix of 12 protocols applied to HFF (x-axis)

compared to loop enrichment for the set of 10,000 loops sampled from the deep data using interaction data obtained from the same matrix of 12 protocols (y-axis).

Table 1 : The list of 3D methods used in Chapter II, Matrix Data

Experiment_name	Experiment_type	Short_experiment_name	Biological_rep	Technical_rep	Cell_Type	Cell_cycle_stage	Enzyme	Crosslinker	total_non_duplicated_valid_pairs	number_of_cis_valid_apirs	cis_ratio	!Sample_description = Experiment 4DN accession
U54-END-DSG-DdeI-20161031-R1-T1	Hi-C	END-FA+DSG-DdeI	1	1	H1-Derived Endoderm(DE)	Non-Synchronized	DdeI	FA+DSG	205924091	134138539	65.13979901	4DNEXGQTF6GX
U54-END-DSG-DpnII-20190711-R2-T1	Hi-C	END-FA+DSG-DpnII	1	1	H1-Derived Endoderm(DE)	Non-Synchronized	DpnII	FA+DSG	210572905	151943203	72.15705316	4DNEXLSE8I9F
U54-END-DSG-HindIII-20161206-R1-T1	Hi-C	END-FA+DSG-HindIII	1	1	H1-Derived Endoderm(DE)	Non-Synchronized	HindIII	FA+DSG	229896147	164696856	71.6396765	4DNEXML4UEI5

U54-END-DSG-MNase-20170508-R1-T1	Micro-C	END-FA+DSG-Mnase	1	1	H1-Derived Endoderm(DE)	Non-Synchronized	MNase	FA+DSG	189326455	154461467	81.58472465	4DNEXYOKS2Q9
U54-END-EGS-DdeI-20161219-R1-T1	Hi-C	END-FA+EGS-DdeI	1	1	H1-Derived Endoderm(DE)	Non-Synchronized	DdeI	FA+EGS	163609737	130469461	79.74431314	4DNEXW481RNF
U54-END-EGS-DpnII-20170119-R2-T1	Hi-C	END-FA+EGS-DpnII	1	1	H1-Derived Endoderm(DE)	Non-Synchronized	DpnII	FA+EGS	186302452	151497006	81.31777353	4DNEXZX5SXU3
U54-END-EGS-HindIII-20161219-R1-T1	Hi-C	END-FA+EGS-HindIII	1	1	H1-Derived Endoderm(DE)	Non-Synchronized	HindIII	FA+EGS	174413205	137515283	78.84453646	4DNEXYNR65K2
U54-END-EGS-MNase-20170508-R1-T1	Micro-C	END-FA+EGS-Mnase	1	1	H1-Derived Endoderm(DE)	Non-Synchronized	MNase	FA+EGS	179727994	146231786	81.3628321	4DNEXBNTOKDI
U54-END-FA-DdeI-20161118-R1-T1	Hi-C	END-FA-DdeI	1	1	H1-Derived Endoderm(DE)	Non-Synchronized	DdeI	FA	186429934	107921572	57.88854273	4DNEX9DK757I
U54-END-FA-DpnII-20170119-R2-T1	Hi-C	END-FA-DpnII	1	1	H1-Derived Endoderm(DE)	Non-Synchronized	DpnII	FA	178823010	103638144	57.95570939	4DNEX281COZ4
U54-END-FA-HindIII-20160311-R1-T1	Hi-C	END-FA-HindIII	1	1	H1-Derived Endoderm(DE)	Non-Synchronized	HindIII	FA	211056400	137365408	65.08469205	4DNEXGQDKD1Y
U54-END-FA-MNase-20170508-R1-T1	Micro-C	END-FA-Mnase	1	1	H1-Derived Endoderm(DE)	Non-Synchronized	MNase	FA	138646531	94018758	67.81183584	4DNEXVCM43FD
U54-ESC-DSG-DdeI-20161014-R1-T1	**Hi-C**	**H1-hESC-FA+DSG-DdeI**	**1**	**1**	**Human Embryonic Stem Cells (H1-hESC)**	Non-Synchronized	**DdeI**	**FA+DSG**	**159048920**	**111063786**	**69.82995295**	**4DNEXGYGE5BH**
U54-ESC-DSG-DpnII-20160722-R1-T1	**Hi-C**	**H1-hESC-FA+DSG-DpnII**	**1**	**1**	**Human Embryonic Stem Cells (H1-hESC)**	Non-Synchronized	**DpnII**	**FA+DSG**	**152477256**	**114504941**	**75.09640717**	**4DNEXPENVNQD**
U54-ESC-DSG-HindIII-20161206-R1-T1	**Hi-C**	**H1-hESC-FA+DSG-HindIII**	**1**	**1**	**Human Embryonic Stem Cells (H1-hESC)**	Non-Synchronized	**HindIII**	**FA+DSG**	**240405140**	**165348847**	**68.77924781**	**4DNEXSZZEHXT**
U54-ESC-DSG-MNase-20170508-R2-T1	**Micro-C**	**H1-hESC-FA+DSG-Mnase**	**1**	**1**	**Human Embryonic Stem Cells (H1-hESC)**	Non-Synchronized	**MNase**	**FA+DSG**	**191222360**	**174297800**	**91.14927773**	**4DNEX8M8ALDF**
U54-ESC-EGS-DdeI-20161118-R1-T1	**Hi-C**	**H1-hESC-FA+EGS-DdeI**	**1**	**1**	**Human Embryonic Stem Cells (H1-hESC)**	Non-Synchronized	**DdeI**	**FA+EGS**	**219023043**	**157259639**	**71.80049955**	**4DNEXYTD5A9M**
U54-ESC-EGS-DpnII-20170119-R2-T1	**Hi-C**	**H1-hESC-FA+EGS-DpnII**	**1**	**1**	**Human Embryonic Stem Cells (H1-hESC)**	Non-Synchronized	**DpnII**	**FA+EGS**	**171785294**	**123692616**	**72.00419379**	**4DNEXW32FN59**
U54-ESC-EGS-HindIII-20161206-R1-T1	**Hi-C**	**H1-hESC-FA+EGS-HindIII**	**1**	**1**	**Human Embryonic Stem Cells (H1-hESC)**	Non-Synchronized	**HindIII**	**FA+EGS**	**227891397**	**166889610**	**73.23207993**	**4DNEX83C4KR6**
U54-ESC-EGS-MNase-20170508-R1-T1	**Micro-C**	**H1-hESC-FA+EGS-Mnase**	**1**	**1**	**Human Embryonic Stem Cells (H1-hESC)**	Non-Synchronized	**MNase**	**FA+EGS**	**210749954**	**196351305**	**93.16789934**	**4DNEX4O48PL9**
U54-ESC-FA-DdeI-20190711-R2-T1	**Hi-C**	**H1-hESC-FA-DdeI**	**1**	**1**	**Human Embryonic Stem Cells (H1-hESC)**	Non-Synchronized	**DdeI**	**FA**	**178046987**	**113002759**	**63.46794231**	**4DNEXRAUHVDK**
U54-ESC-FA-DpnII-20170119-R2-T1	**Hi-C**	**H1-hESC-FA-DpnII**	**1**	**1**	**Human Embryonic Stem Cells (H1-hESC)**	Non-Synchronized	**DpnII**	**FA**	**179273461**	**93706733**	**52.27027608**	**4DNEXSUMVBKJ**
U54-ESC-FA-HindIII-20160311-R1-T1	**Hi-C**	**H1-hESC-FA-HindIII**	**1**	**1**	**Human Embryonic Stem Cells (H1-hESC)**	Non-Synchronized	**HindIII**	**FA**	**184165420**	**129344818**	**70.23295579**	**4DNEXP6V97PN**
U54-ESC-FA-MNase-	**Micro-C**	**H1-hESC-FA-Mnase**	**1**	**1**	**Human Embryonic**	Non-Synchronized	**MNase**	**FA**	**144239839**	**106636709**	**73.9301359**	**4DNEXB5YWPOO**

20170508-R1-T1					Stem Cells (H1-hESC)							
U54-HFF-plate-DSG-DdeI-20160812-R1-T1	Hi-C	HFF-FA+DSG-DdeI	1	1	Human Foreskin Fibroblast(HFF)	Non-Synchronized	DdeI	FA+DSG	214335421	167274633	78.04339209	4DNEX4A7XMOY
U54-HFF-plate-DSG-DpnII-20170119-R2-T1	Hi-C	HFF-FA+DSG-DpnII	1	1	Human Foreskin Fibroblast(HFF)	Non-Synchronized	DpnII	FA+DSG	181557851	150141539	82.69625256	4DNEX982NDRF
U54-HFF-plate-DSG-HindIII-20160226-R1-T1	Hi-C	HFF-FA+DSG-HindIII	1	1	Human Foreskin Fibroblast(HFF)	Non-Synchronized	HindIII	FA+DSG	143226547	127420770	88.96449204	4DNEXP44ZRAL
U54-HFF-plate-DSG-MNase-20190509-R2-T1	Micro-C	HFF-FA+DSG-Mnase	1	1	Human Foreskin Fibroblast(HFF)	Non-Synchronized	MNase	FA+DSG	191818051	173678959	90.54359488	4DNEXPBOFI77
U54-HFF-plate-EGS-DdeI-20161031-R1-T1	Hi-C	HFF-FA+EGS-DdeI	1	1	Human Foreskin Fibroblast(HFF)	Non-Synchronized	DdeI	FA+EGS	207979631	150711489	72.46454293	4DNEXFU1OR7Q
U54-HFF-plate-EGS-DpnII-20160902-R1-T1	Hi-C	HFF-FA+EGS-DpnII	1	1	Human Foreskin Fibroblast(HFF)	Non-Synchronized	DpnII	FA+EGS	204078240	158303618	77.57006234	4DNEXDX613HH
U54-HFF-plate-EGS-HindIII-20190718-R2-T1	Hi-C	HFF-FA+EGS-HindIII	1	1	Human Foreskin Fibroblast(HFF)	Non-Synchronized	HindIII	FA+EGS	209839776	183636276	87.51261534	4DNEXOEZYCRL
U54-HFF-plate-EGS-MNase-20190509-R2-T1	Micro-C	HFF-FA+EGS-Mnase	1	1	Human Foreskin Fibroblast(HFF)	Non-Synchronized	MNase	FA+EGS	185701606	177114562	95.37589136	4DNEXSRJNTWI
U54-HFF-plate-FA-DdeI-20170119-R2-T1	Hi-C	HFF-FA-DdeI	1	1	Human Foreskin Fibroblast(HFF)	Non-Synchronized	DdeI	FA	157229847	110693012	70.40203505	4DNEXXEAFS42
U54-HFF-plate-FA-DpnII-20180904-R1-T1	Hi-C	HFF-FA-DpnII	1	1	Human Foreskin Fibroblast(HFF)	Non-Synchronized	DpnII	FA	180559079	103803806	57.49021682	4DNEXL1W2O7X
U54-HFF-plate-FA-HindIIII-20160226-R2-T1	Hi-C	HFF-FA-HindIII	1	1	Human Foreskin Fibroblast(HFF)	Non-Synchronized	HindIII	FA	202269739	174118277	86.08221767	4DNEXQ6NUVKG
U54-HFF-plate-FA-MNase-20190509-R2-T1	Micro-C	HFF-FA-Mnase	1	1	Human Foreskin Fibroblast(HFF)	Non-Synchronized	MNase	FA	155081327	102025719	65.78852591	4DNEXA2GHZRV
U54-HelaS3-NS-DSG-DpnII-20180709-R1-T1	Hi-C	HelaS3-NS-FA+DSG-DpnII	1	1	Hela S3	Non-Synchronized	DpnII	FA+DSG	152488378	134193834	88.00266339	4DNEXNEX6XA5
U54-HelaS3-NS-DSG-HindIII-20180730-R1-T1	Hi-C	HelaS3-NS-FA+DSG-HindIII	1	1	Hela S3	Non-Synchronized	HindIII	FA+DSG	220380646	199140824	90.36221084	4DNEX21BBVGH
U54-HelaS3-NS-DSG-MNase-08072018-R1-T1	Micro-C	HelaS3-NS-FA+DSG-MNase	1	1	Hela S3	Non-Synchronized	MNase	FA+DSG	232516616	219712667	94.49331871	4DNEXDCH6ZSC
U54-HelaS3-NS-EGS-DpnII-20180709-R1-T1	Hi-C	HelaS3-NS-FA+EGS-DpnII	1	1	Hela S3	Non-Synchronized	DpnII	FA+EGS	124247392	101150749	81.41076233	4DNEX4QN19KR
U54-HelaS3-NS-EGS-HindIII-20180730-R1-T1	Hi-C	HelaS3-NS-FA+EGS-HindIII	1	1	Hela S3	Non-Synchronized	HindIII	FA+EGS	212106796	186751411	88.04593465	4DNEXPQPFGFI
U54-HelaS3-NS-EGS-MNase-08072018-R1-T1	Micro-C	HelaS3-NS-FA+EGS-MNase	1	1	Hela S3	Non-Synchronized	MNase	FA+EGS	222343091	204124873	91.80625855	4DNEXDGA81X4
U54-HelaS3-NS-FA-DpnII-20180709-R1-T1	Hi-C	HelaS3-NS-FA-DpnII	1	1	Hela S3	Non-Synchronized	DpnII	FA	160722183	103742779	64.54789069	4DNEXQZN872V
U54-HelaS3-NS-FA-HindIII-20180730-R1-T1	Hi-C	HelaS3-NS-FA-HindIII	1	1	Hela S3	Non-Synchronized	HindIII	FA	207604208	175183648	84.38347647	4DNEXJLC7L9V
U54-HelaS3-NS-FA-MNase-08072018-R1-T1	Micro-C	HelaS3-NS-FA-MNase	1	1	Hela S3	Non-Synchronized	MNase	FA	76650152	50355346	65.69503737	4DNEXAL2NDFQ

U54-HelaS3-G1-DSG-DpnII-20180709-R1-T1	Hi-C	HelaS3-G1-FA+DSG-DpnII	1	1	Hela S3	Synchronized-G1	DpnII	FA+DSG	137615088	120438296	87.51823492	4DNEXQGIFW97
U54-HelaS3-G1-DSG-HindIII-20180730-R1-T1	Hi-C	HelaS3-G1-FA+DSG-HindIII	1	1	Hela S3	Synchronized-G1	HindIII	FA+DSG	224808021	201397983	89.58665358	4DNEX5PGT9DS
U54-HelaS3-G1-DSG-MNase-08072018-R1-T1	Micro-C	HelaS3-G1-FA+DSG-MNase	1	1	Hela S3	Synchronized-G1	MNase	FA+DSG	238609768	217832255	91.29226218	4DNEX9TYL3XW
U54-HelaS3-G1-EGS-DpnII-20180709-R1-T1	Hi-C	HelaS3-G1-FA+EGS-DpnII	1	1	Hela S3	Synchronized-G1	DpnII	FA+EGS	150034641	122376379	81.56541595	4DNEXHEVORF1
U54-HelaS3-G1-EGS-HindIII-20180730-R1-T1	Hi-C	HelaS3-G1-FA+EGS-HindIII	1	1	Hela S3	Synchronized-G1	HindIII	FA+EGS	219847694	192790542	87.69277425	4DNEXDN6IM45
U54-HelaS3-G1-EGS-MNase-08072018-R1-T1	Micro-C	HelaS3-G1-FA+EGS-MNase	1	1	Hela S3	Synchronized-G1	MNase	FA+EGS	233861989	209045642	89.38846492	4DNEX3Q5WM53
U54-HelaS3-G1-FA-DpnII-20180709-R1-T1	Hi-C	HelaS3-G1-FA-DpnII	1	1	Hela S3	Synchronized-G1	DpnII	FA	127692395	86554611	67.78368516	4DNEXMF677YU
U54-HelaS3-G1-FA-HindIII-20180730-R1-T1	Hi-C	HelaS3-G1-FA-HindIII	1	1	Hela S3	Synchronized-G1	HindIII	FA	226779995	193431514	85.29478713	4DNEX8V59T74
U54-HelaS3-G1-FA-MNase-08072018-R1-T1	Micro-C	HelaS3-G1-FA-Mnase	1	1	Hela S3	Synchronized-G1	MNase	FA	79643500	37612874	47.22654579	4DNEXD4NIABB
U54-HelaS3-M-DSG-DpnII-20180709-R1-T1	Hi-C	HelaS3-M-FA+DSG-DpnII	1	1	Hela S3	Synchronized-Mitosis	DpnII	FA+DSG	125076028	115102454	92.02599078	4DNEX6RNZ4I9
U54-HelaS3-M-DSG-HindIII-20180730-R1-T1	Hi-C	HelaS3-M-FA+DSG-HindIII	1	1	Hela S3	Synchronized-Mitosis	HindIII	FA+DSG	216873203	209261829	96.49040366	4DNEXVH3GE23
U54-HelaS3-M-DSG-MNase-08072018-R1-T1	Micro-C	HelaS3-M-FA+DSG-MNase	1	1	Hela S3	Synchronized-Mitosis	MNase	FA+DSG	250327507	233434833	93.25177077	4DNEXXV9HQZQ
U54-HelaS3-M-EGS-DpnII-20180709-R1-T1	Hi-C	HelaS3-M-FA+EGS-DpnII	1	1	Hela S3	Synchronized-Mitosis	DpnII	FA+EGS	112195020	90585992	80.73976189	4DNEXW59M74B
U54-HelaS3-M-EGS-HindIII-20180730-R1-T1	Hi-C	HelaS3-M-FA+EGS-HindIII	1	1	Hela S3	Synchronized-Mitosis	HindIII	FA+EGS	214538340	203691547	94.94412374	4DNEX286UZDL
U54-HelaS3-M-EGS-MNase-08072018-R1-T1	Micro-C	HelaS3-M-FA+EGS-MNase	1	1	Hela S3	Synchronized-Mitosis	MNase	FA+EGS	235502044	219350822	93.14179116	4DNEXWBODDF1
U54-HelaS3-M-FA-DpnII-20180709-R1-T1	Hi-C	HelaS3-M-FA-DpnII	1	1	Hela S3	Synchronized-Mitosis	DpnII	FA	109633817	75578240	68.93697772	4DNEXNVLEW4T
U54-HelaS3-M-FA-HindIII-20180730-R1-T1	Hi-C	HelaS3-M-FA-HindIII	1	1	Hela S3	Synchronized-Mitosis	HindIII	FA	202441463	189675265	93.69388177	4DNEXR3L8DZL
U54-HelaS3-M-FA-MNase-08072018-R1-T1	Micro-C	HelaS3-M-FA-MNase	1	1	Hela S3	Synchronized-Mitosis	MNase	FA	177300242	52080538	29.37420582	4DNEXSQB9QVT

Table 2 : The list of 3D methods used in Chapter II, Deep Data

U54-ESC4DN-DSG-DpnII-20190530-R1-T1	Hi-C	H1-hESC-FA+DSG-DpnII	1	1	Human Embryonic Stem Cells (H1-hESC)	Non-Synchronized	DpnII	FA+DSG	1719156895	1384796845	80.55092872	4DNEXNQBY5YS
U54-ESC4DN-DSG-DpnII-20190530-R2-T1	Hi-C	H1-hESC-FA+DSG-DpnII	2	1	Human Embryonic Stem Cells (H1-hESC)	Non-Synchronized	DpnII	FA+DSG	1677562310	1350603279	80.50987263	4DNEXUBKKN5P
U54-ESC4DN-FA-DpnII-2017524-R1-T1	Hi-C	H1-hESC-FA-DpnII	1	1	Human Embryonic Stem Cells (H1-hESC)	Non-Synchronized	DpnII	FA	1362318048	660952512	48.51675517	4DNEXENKINA2

U54-ESC4DN-FA-DpnII-2017524-R1-T2	Hi-C	H1-hESC-FA-DpnII	2	1	Human Embryonic Stem Cells (H1-hESC)	Non-Synchronized	DpnII	FA	1246833143	597363946	47.91049623	4DNEXZJZ5EBZ
U54-ESC4DN-FA-DSG-MNase-R1-T1	Micro-C	H1-hESC-FA+DSG-MNase	1	1	Human Embryonic Stem Cells (H1-hESC)	Non-Synchronized	MNase	FA+DSG	499643562	449673411	89.99884021	4DNEXQMEU2O4
U54-ESC4DN-FA-DSG-MNase-R1-T2	Micro-C	H1-hESC-FA+DSG-MNase	1	2	Human Embryonic Stem Cells (H1-hESC)	Non-Synchronized	MNase	FA+DSG	205146303	176530863	86.05120366	4DNEX24VTARO
U54-ESC4DN-FA-DSG-MNase-R1-T3	Micro-C	H1-hESC-FA+DSG-MNase	1	3	Human Embryonic Stem Cells (H1-hESC)	Non-Synchronized	MNase	FA+DSG	299729273	248779306	83.00133768	4DNEX9JAVMKE
U54-ESC4DN-FA-DSG-MNase-R1-T4	Micro-C	H1-hESC-FA+DSG-MNase	1	4	Human Embryonic Stem Cells (H1-hESC)	Non-Synchronized	MNase	FA+DSG	473301507	385885358	81.53055765	4DNEXWHALJ1K
U54-ESC4DN-FA-DSG-MNase-R2-T1	Micro-C	H1-hESC-FA+DSG-MNase	2	1	Human Embryonic Stem Cells (H1-hESC)	Non-Synchronized	MNase	FA+DSG	226763614	188421398	83.09154836	4DNEXQQM6EMK
U54-ESC4DN-FA-DSG-MNase-R2-T2	Micro-C	H1-hESC-FA+DSG-MNase	2	2	Human Embryonic Stem Cells (H1-hESC)	Non-Synchronized	MNase	FA+DSG	220396074	182809153	82.94573931	4DNEXACM6M51
U54-ESC4DN-FA-DSG-MNase-R2-T3	Micro-C	H1-hESC-FA+DSG-MNase	2	3	Human Embryonic Stem Cells (H1-hESC)	Non-Synchronized	MNase	FA+DSG	435493545	386194490	88.67972773	4DNEX2WPDBN7
U54-ESC4DN-FA-DSG-MNase-R2-T4	Micro-C	H1-hESC-FA+DSG-MNase	2	4	Human Embryonic Stem Cells (H1-hESC)	Non-Synchronized	MNase	FA+DSG	449531807	398203541	88.58183888	4DNEXCL5AQ32
U54-ESC4DN-FA-DSG-MNase-R2-T5	Micro-C	H1-hESC-FA+DSG-MNase	2	5	Human Embryonic Stem Cells (H1-hESC)	Non-Synchronized	MNase	FA+DSG	421247032	377494216	89.61350166	4DNEXJPSM64Q
U54-HFFc6-DSG-DdeI-20180319-R1-T1	Hi-C	HFFc6-FA+DSG-DdeI	1	1	Human Foreskin Fibroblast clone 6 (HFFc6)	Non-Synchronized	DdeI	FA+DSG	1381215653	1260258773	91.24272305	4DNEX95OQ954
U54-HFFc6-DSG-DdeI-20181023-R2-T1	Hi-C	HFFc6-FA+DSG-DdeI	2	1	Human Foreskin Fibroblast clone 6 (HFFc6)	Non-Synchronized	DdeI	FA+DSG	1353058162	1248663160	92.28451482	4DNEX2SUQP87
U54-HFFc6-DSG-DdeI-DpnII-20190711-R1-T1	Hi-C	HFFc6-FA+DSG-DdeI-DpnII	1	1	Human Foreskin Fibroblast clone 6 (HFFc6)	Non-Synchronized	DdeI-DpnII	FA+DSG	1715480034	1539578927	89.74624574	4DNEXDL7KBH2
U54-HFFc6-DSG-DdeI-DpnII-20191219-R3-T1	Hi-C	HFFc6-FA+DSG-DdeI-DpnII	2	1	Human Foreskin Fibroblast clone 6 (HFFc6)	Non-Synchronized	DdeI-DpnII	FA+DSG	1563512452	1390580534	88.9395241	4DNEXSFVDRQD
U54-HFFc6-DSG-DpnII-20180319-R1-T1	Hi-C	HFFc6-FA+DSG-DpnII	1	1	Human Foreskin Fibroblast clone 6 (HFFc6)	Non-Synchronized	DpnII	FA+DSG	1340569555	1190949842	88.83909362	4DNEXBO6JSO9
U54-HFFc6-DSG-DpnII-20190102-R2-T1	Hi-C	HFFc6-FA+DSG-DpnII	2	1	Human Foreskin Fibroblast clone 6 (HFFc6)	Non-Synchronized	DpnII	FA+DSG	1496036040	1355447143	90.60257285	4DNEXQBYXQKH
U54-HFFc6-p17-FA-DpnII-20170327	Hi-C	HFFc6-FA-DpnII	1	1	Human Foreskin Fibroblast clone 6 (HFFc6)	Non-Synchronized	DpnII	FA	1549715647	1041819394	67.22648739	4DNEX7POCO84
U54-HFFc6-p22-FA-DpnII-20170327	Hi-C	HFFc6-FA-DpnII	2	1	Human Foreskin Fibroblast clone 6 (HFFc6)	Non-Synchronized	DpnII	FA	1468505740	1050752790	71.55251501	4DNEXRAEERUF
U54-HFFc64DN-FA-DSG-MNase-R1-T1	Micro-C	HFFc6-FA+DSG-MNase	1	1	Human Foreskin Fibroblast clone 6 (HFFc6)	Non-Synchronized	MNase	FA+DSG	502076577	468995532	93.41115549	4DNEXC1TYVLD
U54-HFFc64DN-FA-DSG-MNase-R1-T2	Micro-C	HFFc6-FA+DSG-MNase	1	2	Human Foreskin Fibroblast clone 6 (HFFc6)	Non-Synchronized	MNase	FA+DSG	523271269	488273007	93.31164081	4DNEX1MDLILR
U54-HFFc64DN-FA-DSG-MNase-R1-T3	Micro-C	HFFc6-FA+DSG-MNase	1	3	Human Foreskin Fibroblast clone 6 (HFFc6)	Non-Synchronized	MNase	FA+DSG	498218862	468920452	94.11936957	4DNEXRJYVOZJ
U54-HFFc64DN-FA-DSG-MNase-R1-T4	Micro-C	HFFc6-FA+DSG-MNase	1	4	Human Foreskin Fibroblast clone 6 (HFFc6)	Non-Synchronized	MNase	FA+DSG	514269682	483369664	93.99147586	4DNEXONC5BJ1
U54-HFFc64DN-FA-DSG-MNase-R1-T5	Micro-C	HFFc6-FA+DSG-MNase	1	5	Human Foreskin Fibroblast clone 6 (HFFc6)	Non-Synchronized	MNase	FA+DSG	504813661	470582890	93.21912744	4DNEX9TF73VI
U54-HFFc64DN-FA-DSG-MNase-R1-T6	Micro-C	HFFc6-FA+DSG-MNase	1	6	Human Foreskin Fibroblast clone 6 (HFFc6)	Non-Synchronized	MNase	FA+DSG	518371787	487007507	93.94946238	4DNEX5F27GL2
U54-HFFc64DN-FA-DSG-MNase-R2-T1	Micro-C	HFFc6-FA+DSG-MNase	2	1	Human Foreskin Fibroblast clone 6 (HFFc6)	Non-Synchronized	MNase	FA+DSG	189730532	180050178	94.89784069	4DNEX3W33RA7
U54-HFFc64DN-FA-DSG-MNase-R2-T2	Micro-C	HFFc6-FA+DSG-MNase	2	2	Human Foreskin Fibroblast clone 6 (HFFc6)	Non-Synchronized	MNase	FA+DSG	179019084	169939401	94.92809214	4DNEX2RPZBFM
U54-HFFc64DN-FA-DSG-MNase-R2-T3	Micro-C	HFFc6-FA+DSG-MNase	2	3	Human Foreskin Fibroblast clone 6 (HFFc6)	Non-Synchronized	MNase	FA+DSG	172219324	162093947	94.1206499	4DNEX5X6Y4QY

U54-HFFc64DN-FA-DSG-MNase-R2-T4	Micro-C	HFFc6-FA+DSG-MNase	2	4	Human Foreskin Fibroblast clone 6 (HFFc6)	Non-Synchronized	MNase	FA+DSG	136819555	128058436	93.59658859	4DNEXF6J2UBH
U54-HFFc64DN-FA-DSG-MNase-R2-T5	Micro-C	HFFc6-FA+DSG-MNase	2	5	Human Foreskin Fibroblast clone 6 (HFFc6)	Non-Synchronized	MNase	FA+DSG	438117029	411984245	94.03520469	4DNEXNCO8M9N
U54-HFFc64DN-FA-DSG-MNase-R2-T6	Micro-C	HFFc6-FA+DSG-MNase	2	6	Human Foreskin Fibroblast clone 6 (HFFc6)	Non-Synchronized	MNase	FA+DSG	397291449	376088185	94.66304547	4DNEXXXWOYOB

Table 3: The list of 1D methods used in Chapter II

Experiment_name	Experiment_type	Biological_rep	Technical_rep	Total_number_of_reads	Uniquely_mapped_reads	Cell_Type	!Sample_description = Experiment 4DN accession
HFF1_CTCF_1_2020JUL28_PE40_chip	Cut&Tag	1	1	56207456	52976144	Human Foreskin Fibroblast clone 6 (HFFc6)	4DNEXHYA4NOW
HFF1_CTCF_2_2020JUL28_PE40_chip	Cut&Tag	2	1	64136278	61681502	Human Foreskin Fibroblast clone 6 (HFFc6)	4DNEXAG8IFJU
HFF1_Smc1_1_2020JUL28_PE40_chip	Cut&Tag	1	1	61033294	57812682	Human Foreskin Fibroblast clone 6 (HFFc6)	4DNEXLAHVC6T
HFF1_Smc1_2_2020JUL28_PE40_chip	Cut&Tag	2	1	39619704	37139568	Human Foreskin Fibroblast clone 6 (HFFc6)	4DNEXMHZPZVI
CTCF-H1-hESC-R1	Cut&Run	1	1	23607294	21556708	Human Emryonic Stem Cells (H1-hESC)	4DNEXIHB6H1I
CTCF-H1-hESC-R2	Cut&Run	2	1	21375302	19466596	Human Emryonic Stem Cells (H1-hESC)	4DNEXCZB2WY3
CTCF-H1-hESC-R3	Cut&Run	3	1	12400636	11489860	Human Emryonic Stem Cells (H1-hESC)	4DNEXJ5UFKWU
CTCF-H1-hESC-R4	Cut&Run	4	1	10062780	9406374	Human Emryonic Stem Cells (H1-hESC)	4DNEXR1FX8LV
CTCF-H1-hESC-R5	Cut&Run	5	1	19640118	17947694	Human Emryonic Stem Cells (H1-hESC)	4DNEX65LZI8G
CTCF-H1-hESC-R6	Cut&Run	6	1	17746118	15911678	Human Emryonic Stem Cells (H1-hESC)	4DNEXUS23EIB
H3K4me3 - H1-hESC-R1	Cut&Run	1	1	23925878	18164810	Human Emryonic Stem Cells (H1-hESC)	4DNEX3V6Q5XT
H3K4me3 - H1-hESC-R2	Cut&Run	2	1	16194056	13465432	Human Emryonic Stem Cells (H1-hESC)	4DNEXNYV5G5T
H3K4me3 - H1-hESC-R3	Cut&Run	3	1	11783986	10413744	Human Emryonic Stem Cells (H1-hESC)	4DNEXU6CZ816
H3K4me3 - H1-hESC-R4	Cut&Run	4	1	27026796	22450414	Human Emryonic Stem Cells (H1-hESC)	4DNEX8N5MVDM
H3K4me3 - H1-hESC-R5	Cut&Run	5	1	28823924	23427374	Human Emryonic Stem Cells (H1-hESC)	4DNEXEVVM5HF
H3K27ac - H1-hESC-R1	Cut&Run	1	1	38834218	34785352	Human Emryonic Stem Cells (H1-hESC)	4DNEXSPSRIYS
H3K27ac - H1-hESC-R2	Cut&Run	2	1	33571412	30143014	Human Emryonic Stem Cells (H1-hESC)	4DNEXAX88165
H3K27ac - H1-hESC-R3	Cut&Run	3	1	48609358	44808318	Human Emryonic Stem Cells (H1-hESC)	4DNEXVYA898L
H3K27ac - H1-hESC-R4	Cut&Run	4	1	24897206	22186198	Human Emryonic Stem Cells (H1-hESC)	4DNEX81NIWVM
H3K27ac - H1-hESC-R5	Cut&Run	5	1	26258098	22660174	Human Emryonic Stem Cells (H1-hESC)	4DNEX5EQJ2P2
H3K4me3 - HFFc6-R1	Cut&Run	1	1	18958168	16936068	Human Foreskin Fibroblast clone 6 (HFFc6)	4DNEX67Q84RB
H3K4me3 - HFFc6-R2	Cut&Run	2	1	15898068	14331352	Human Foreskin Fibroblast clone 6 (HFFc6)	4DNEXRL2LUNJ
H3K4me3 - HFFc6-R3	Cut&Run	3	1	21000500	18703244	Human Foreskin Fibroblast clone 6 (HFFc6)	4DNEXQ8TTJJ5
H3K4me3 - HFFc6-R4	Cut&Run	4	1	17015302	14894516	Human Foreskin Fibroblast clone 6 (HFFc6)	4DNEX62TLS34

H3K27ac - HFFc6-R1	Cut&Run	1	1	18551090	13098528	Human Foreskin Fibroblast clone 6 (HFFc6)	4DNEXEYS34V6
H3K27ac - HFFc6-R2	Cut&Run	2	1	16743986	12338000	Human Foreskin Fibroblast clone 6 (HFFc6)	4DNEX4PTP92V
H3K27ac - HFFc6-R3	Cut&Run	3	1	14007368	12512660	Human Foreskin Fibroblast clone 6 (HFFc6)	4DNEXC666QIV
H3K27ac - HFFc6-R4	Cut&Run	4	1	16002260	14142742	Human Foreskin Fibroblast clone 6 (HFFc6)	4DNEXTSVULI3
H1-hESC-R1	ATAC-Seq	1	1	133703774	132816798	Human Emryonic Stem Cells (H1-hESC)	4DNEXKMOP6RJ
H1-hESC-R2	ATAC-Seq	2	1	138530478	137501566	Human Emryonic Stem Cells (H1-hESC)	4DNEX9J9YFK3
U54-HFFc6-NS-AT-R1	ATAC-Seq	1	1,2,3,4	702534390	681091376	Human Foreskin Fibroblast clone 6 (HFFc6)	4DNEXF6RQQ27, 4DNEXQUCGLW9, 4DNEXJGIUGRH, 4DNEXBYHX8E4
U54-HFFc6-NS-AT-R2	ATAC-Seq	2	1,2,3,4	708825898	692621990	Human Foreskin Fibroblast clone 6 (HFFc6)	4DNEX97A5KCP, 4DNEX6C42Q5B, 4DNEXN7N1U82, 4DNEXMHGYBOJ

Table 4: Experimental Protocols used in Chapter IV

Methods	Target	Required sequencing depth	Pros	Cons
Hi-C	Genome-wide	~2 Billions	Stong compartments	Limited resolution
Micro-C	Genome-wide	~2 Billions	Strong loop detection	Weak compartmes
SPRITE	Genome-wide	~2 Billions	Multi-contacts, Stong compartments	Limited resolution
GAM	Genome-wide	~1.1 Billions	Real genomic distance	Limited resolution
ChIA-PET	CTCF/PolII	~500 Millions	Strong loops	Weak compartmes
PLAC-Seq	H3K4me3	~500 Millions	Strong loops	Weak compartmes

Chapter III: An improved Hi-C Protocol

Preface

The chapter III has unpublished data as well as data published in two papers.

The protocol of Hi-C 3.0 is published by Denis L. Lafontaine, Liyan Yang, Job Dekker, and Johan H. Gibcus. The protocol is entitled as : Hi- C 3.0: Improved Protocol for Genome- Wide Chromosome Conformation Capture. PMID: 34286910 PMCID: PMC8362010 DOI: 10.1002/cpz1.198

Some of the analysis of Hi-C 3.0 is published by

Betul Akgol Oksuz, Liyan Yang, Sameer Abraham, Sergey V. Venev, Nils Krietenstein,Krishna Mohan Parsi, Hakan Ozadam, Marlies E. Oomen, Ankita Nand, Hui Mao, Ryan M. J. Genga, Rene Maehr, Oliver J. Rando, Leonid A. Mirny, Johan H. Gibcus and Job Dekker. The publication is entitled "Systematic evaluation of chromosome conformation capture assays". PMID: 34480151 PMCID: PMC8446342 DOI: 10.1038/s41592-021-01248-7

Summary

Chromatin conformation capture (3C)-based methods have greatly improved detecting genome-wide interactions at different scales partly determined by the fragmentation level of the genome. Micro-C, with nucleosome level fragmentation, has the greatest resolution to detect small-scale structures such as looping interactions compared to methods that measure chromatin contacts genome-wide. However, it has a weak performance detecting large-scale structures, like compartmental domains. It has been a challenge to

develop one protocol that best detects both small and large-scale structures. Here, we developed Hi-C 3.0, an updated version of Hi-C, which improves the detection of genomic interactions at all scales. Hi-C 3.0 uses two crosslinking agents and two restriction enzymes, which allow capturing more genomic interactions at higher resolution compared to existing Hi-C protocols. Hi-C 3.0 eliminates the need of using separate protocols to identify both small and large genomic structures and provides a quantitative comparison of genomic interactions across multiple scales in normal and perturbed conditions.

Introduction

Hi-C is the most widely used method for measuring three-dimensional (3D) genomic interactions at genome-wide level. Genomic interactions in 3D play a crucial role in gene regulation and function (Lupianez et al. 2015; Melo et al. 2020; Nora et al. 2012; Uhler and Shivashankar 2017; Valton et al. 2021; Zheng and Xie 2019; Zuin et al. 2014). In brief, the Hi-C protocol relies on cross-linking, fragmenting and ligation of chromatin in spatial proximity. After digestion, fragments are labeled with biotin, which is then used to pull-down the ligation products to enrich them for identification by high-throughput sequencing. In the Hi-C protocol, crosslinking the chromatin and fragmenting it by restriction enzyme are the two critical steps that determine the chromatin contacts. Chapter II has shown how additional crosslinking to FA greatly improves the detection and quantification of all genomic structures and how fragmentation level determines the detectability of these structures over short and long distances. In this chapter, I will focus on combining different

restriction enzymes that would affect the detection of genome structures.

Historically, many different restriction enzymes have been used in Hi-C protocol, which contributed to the generation of genomic interaction maps at different resolutions (Lieberman-Aiden et al. 2009; Rao et al. 2014; Belaghzal, Dekker, and Gibcus 2017). Fragment length and sequencing depth have been shown to be critical in determining the resolution of a Hi-C experiment (Rao et al. 2014). Initial Hi-C experiments used 6 base pair cutters such as HindIII and NcoI to fragment the chromatin (Lieberman-Aiden et al. 2009). Later, Rao et al. improved the Hi-C protocol two ways; by ligating the chromatin in intact nuclei, i.e. *in situ,* and by digesting the chromatin with a 4-base pair (4-bp) cutter instead of 6. This study improved the resolution of Hi-C by increased fragmentation frequency and deeper sequencing. Consistently, in chapter II, we showed that fragmenting the genome into smaller pieces either by using MNase fragmentation or 4-bp cutters improves the detection of small-scale structures, such as chromatin loops. Insights from the literature and chapter II led us to develop a new Hi-C protocol called Hi-C 3.0, which combines two frequently cutting restriction enzymes: DdeI (CTNAG) and DpnII (GATC) to produce short fragments that are on average ~1.5kb in size and two crosslinking agents: disuccinimidyl glutarate (DSG) and formaldehyde to capture frequently interacting genomic regions (Lafontaine et al. 2021). Hi-C 3.0 showed improved detection of genomic interactions at small scales (loops, Topologically Associating Domains (TADs)) while maintaining the detection and quantification of large-scale structures (compartments) (Akgol Oksuz et al. 2021).

Results

To test the performance of Hi-C 3.0 we compared it to previously developed Hi-C protocols and Micro-C by quantifying the detectability of both small and large scale structures. All compared protocols used cross-linking with FA and DSG. Compared protocols' chromatin is digested with MNase, DdeI, DpnII or both DdeI and DpnII. We compared the interaction frequency as a function of distance, detectability, and strength of small and large scale structures, inter-chromosomal interactions, and the importance of sequencing depth in the detection of small and large scale structures. We found that compared to other Hi-C protocols, the finer fragments in Hi-C 3.0 led to a stronger detection of small-scale structures such as chromatin loops. Although Micro-C is superior at detecting loops, Hi-C 3.0 provides a more powerful detection of large-scale structures like compartments.

- **Fragment size in Hi-C 3.0**

We compared the fragment size distribution of Hi-C 3.0, performed using both DdeI and DpnII, to Hi-C performed with either DdeI or DpnII. Using two restriction enzymes to digest the chromatin produces finer fragmentation compared to the use of one enzyme at a time. Therefore, Hi-C 3.0 has the finest fragmentation compared to all other Hi-C experiments. Smaller fragments increase the complexity and the resolution of the experiments by increasing the probability of capturing possible pairwise interactions.

Before sequencing, we used a fragment analyzer to examine the fragment size distribution of a Hi-C 3.0 library, digested with DdeI (CTNAG) and DpnII

(GATC. We observed that two-enzyme digestion resulted in smaller fragment sizes and more uniform fragmentation compared to one enzyme digestion (Figure 3.1a). Fragmenting the genome uniformly is important for getting contact information from all regions in the genome.

Because Hi-C 3.0 uses two crosslinkers, we wanted to confirm that using extra crosslinkers does not influence fragment size distributions. Indeed, we did not observe significant differences in the fragment size distributions with additional crosslinkers (Figure 3.1a). As expected, in silico digested chromatin gave smaller fragment sizes compared to the experimentally digested chromatin, confirming that the digestion efficiency is not 100 % (Figure 3.1b upper panel, dotted lines). The median fragment size for experiments that digest the chromatin with DdeI is smaller than experiments that digest the chromatin with DpnII. Expectedly, experiments that digested chromatin with both DdeI and DpnII have the smallest fragments (Figure 3.1b lower panel). We observed similar fragment lengths with and without additional cross-linkers. These results confirm that using the two restriction enzymes generated the desired small fragments and that the extra cross-linkers does not affect the nature of the experiment.

Next, we performed paired-end sequencing and tested the relationship between fragment size and the distribution of sequenced reads located in the corresponding fragment. We sampled 1M reads and quantified the number of reads associated with a given fragment. We did not observe a correlation between the number of reads and their corresponding fragment size which

indicates that there is no bias toward specific fragment length in these experiments (Figure 3.1c).

- **Short distance contact probability in Hi-C 3.0**

To investigate the performance of Hi-C 3.0 in detecting short-range genomic interactions, we first compared it to existing Hi-C and Micro-C protocols. These protocols have been used to generate contact maps at different resolutions, where Hi-C is biased in detection towards large (e.g. compartments) and Micro-C is biased towards fine (e.g. chromatin loops) structures, respectively (Akgol Oksuz et al. 2021; Krietenstein et al. 2020). Comparison of contact maps generated from existing Hi-C protocols (Hi-C 2.5, two crosslinkers and one enzyme), Hi-C 3.0 and Micro-C protocols at 100 kb binned (long) or 2 kb (short) binned resolutions showed that Hi-C 3.0 performs similar to the existing Hi-C protocols at long-distance and to Micro-C at short-distance. (Figure 3.2a and Figure 3.2b). Distance corrected HicREP correlation showed that Micro-C highly correlates with all existing Hi-C protocols. Yet, it correlates best with Hi-C 3.0 for all chromosomes (Figure 3.2c). These results indicate that Hi-C 3.0 has improved detection of short-range contacts without compromising the detection of long-range interactions.

Due to chromatin nature and chromosome territories, 3C-based methods detect more interactions in cis and fewer interactions in trans. High percent trans interactions is an indication of random ligation. To further support the performance of the Hi-C 3.0 protocol, we investigated the cis percent of each library as a function of genomic distance. We observed that

almost 50% of Micro-C reads are located in the first 1kb region and only ~20% of the reads are > 40kb. Overall, the existing Hi-C experiments have fewer reads in 1kb distance compared to Micro-C. Reads < 1kb are mostly products of artifacts such as dangling ends and self circles in Hi-C experiments. Hi-C 3.0 has more reads in the first 1kb compared to other Hi-C experiments and more reads > 40kb compared to Micro-C (Figure 3.2d). We found that Hi-C 3.0 improved the distance-dependent contact probability (Figure 3.2e). Compared to data obtained by single DdeI or DpnII digests, the number of contacts increased for loci separated by less than 10 kb, making the results from this protocol more similar to results obtained with protocols using MNase digestion. Yet, longer distance contacts resembled data obtained with protocols using single restriction enzymes more than data obtained with protocols using MNase. These results supported the improved performance of the Hi-C protocol in the detection of short-range contacts without a loss in the detection of the long-range contacts. (Figure 3.2e).

- **Compartment quantification in Hi-C 3.0**

To identify the compartments in Hi-C 3.0 and Micro-C, we used eigenvector decomposition, which is a dimensionality reduction method that can segregate euchromatic (A) and heterochromatic (B) compartments. The first eigenvector which generally detects the compartment signal, showed a high correlation between Micro-C and Hi-C (Figure 3.3.a). To quantify the strength of compartments detected by these protocols, we used saddle plots (Methods). Consistent with previous studies (chapter II), all Hi-C protocols

showed an overall stronger compartment signal in both *cis* and *trans* compared to Micro-C (Figure 3.3.b) (Akgol Oksuz et al. 2021). Among the Hi-C protocols, which showed similar compartment strength in cis, Hi-C 3.0 showed a modest reduction in the compartment signal in trans (Figure 3.3.c).

To determine if the compartment strength is affected by the genomic distance, we quantified compartment strength detected in Hi-C 3.0 and Micro-C at genomic distances between 6 Mb and 250 Mb. At all tested genomic distances, Hi-C 3.0 showed stronger compartmentalization compared to Micro-C (Figure 3.3.d left for A-A and middle panel for B-B).

Next, we tested the strength of the interactions between non-preferential A and B compartments which might indicate the noise level of an experiment. To compare the noise levels for genomic interactions between Micro-C and Hi-C protocols, we quantified interaction between A (active) and B (inactive) compartments. Hi-C showed a weaker signal for non-specific A-B interactions compared to Micro-C. As an additional analysis for noise, we compared the interaction frequency between genomic and mitochondrial DNA of all three Hi-C experiments and found that Hi-C 3.0 has a similar mitochondrial signal as other Hi-C protocols (Figure 3.3.e). These results suggest that Hi-C 3.0 does not lead to increased random ligations and noise levels compared to other Hi-C protocols and provides a better signal-to-noise ratio compared to Micro-C.

- **Insulation strength in Hi-C 3.0**

  In chapter II, we showed that the position of TADs boundaries and their

strength don't differ between 3C-based protocols. Here, we tested if the TAD boundaries change in Hi-C 3.0 compared to other Hi-C protocols and Micro-C. We use the insulation score to quantify the TAD boundaries. The insulation score is calculated as follows: the average interaction frequency is calculated for a given window size moving on the diagonal of the interaction map. The depth of a local minimum displays the strength of the TAD boundary. We observed that insulation scores in 10kb bins with 50 kb sliding window size are highly correlated between Micro-C and Hi-C experiments (Figure 3.4 a). However, the correlation between Hi-C 3.0 and Micro-C is higher than between Hi-C 3.0 and other Hi-C protocols, suggesting that the finer fragmented Hi-C experiment becomes more similar to Micro-C. The pileup of the insulation scores, centered on TAD boundaries, showed that Micro-C has the strongest insulation strength followed by Hi-C 3.0 (Figure 3.4 b).

- **Loop detection in Hi-C 3.0**

We investigated the detection and strength of loops in Micro-C and Hi-C protocols. We used Cooltools that adopted the approach and parameters from HICCUPS to identify chromatin loops (https://github.com/open2c/cooltools) (Rao et al. 2014). Cooltools is a multi-functional python package developed to analyze interaction matrices and quantify all the structures detected in these interaction matrices. HICCUPS uses global and local donut-like filtering for loop detection. Fragmentation level determines the detectability and the strength of chromatin loops hence Hi-C 3.0 uses two restriction enzymes to shorten the fragment size. Next, we

investigated how the finer fragmentation of Hi-C 3.0 changes the detectability and the strength of chromatin loops. We observed that the number of detected loops greatly increased in Hi-C 3.0 compared to all other Hi-C experiments. Hi-C 3.0 detects 28,921 loops, Hi-C performed with DdeI detects 22,219 loops and Hi-C performed with DpnII detects 22,312 loops. Hi-C 3.0 detects more loops for all genomic distances compared to other Hi-C methods and the strength of these loops is higher in Hi-C 3.0 (Figure 3.5 a, left). Although Hi-C 3.0 improved the detection of loops compared to other Hi-C protocols, Micro-C detects the highest number of loops which is 36,989.

To examine the performance of methods detecting loops in different genomic distances, we investigated the number and the strength of the loops separated by specific genomic distances. In smaller distances, Micro-C detects almost two-fold loops compared to Hi-C experiments but this number decreases as the genomic separation between loops increases (Figure 3.5 a, left). Micro-C has the strongest loop strength for all genomic distances (Figure 3.5 a, right).

Next, we investigated the effect of sequencing depth on loop detection. We combined the Hi-C data obtained with either DdeI or DpnII digestion, which resulted in ~4.8B valid pairs. The Hi-C 3.0 dataset consisted of ~2.8 B valid pairs. We observed that the combined datasets detected 26,919 loops whereas Hi-C 3.0 detected 28,921 loops. Even though the combined DdeI and DpnII data sets had an almost twofold amount of reads, our analysis found more loops in the Hi-C 3.0 data set. This result indicates that for loop

detection, fragmentation is more important than sequencing depth. Finer fragmentation is the key for improving the loop detection here. Loop size distribution of Micro-C and Hi-C protocols showed that as expected the finer fragmentation in Micro-C led to the detection of smaller loops (Figure 3.5 b).

Additionally, we observed that the majority of the loops that are detected in Hi-C 3.0 are also detected by Micro-C (Figure 3.5 c). Loops that are detected by both Hi-C 3.0 and Micro-C are the strongest. The enrichment of the loops that are detected by only Hi-C 3.0 or only Micro-C are weaker (Figure 3.5 d).

Furthermore, we investigated the features of the loops by checking the underlying mechanisms of loops detected in Hi-C 3.0 and Micro-C. DNA looping can be ascribed to promoter-enhancer (P-E) interactions, enrichment for active promoters, active enhancers, CTCF, and cohesin. Krietenstein et al. have shown that Micro-C detects more promoter-enhancer (P-E) loops compared to Hi-C (Krietenstein et al. 2020). We examined the enrichment of CTCF, SMC1, H3K4me3, and H3K27ac at Hi-C 3.0 specific loop anchors, Micro-C specific loop anchors and loop anchors that are shared between these two protocols. We observed that common loop anchors are more enriched for CTCF and cohesin (SMC1) and less enriched for active promoter mark H3K4me3 and enhancer mark H3K27ac. On the other hand protocol-specific loops (Hi-C 3.0 specific and Micro-C specific) are less enriched for CTCF and SMC1 and more enriched for H3K4me3 and H3K27ac (Figure 3.5 e). Using cis-regulatory elements that are predicted by SCREEN (Consortium, Moore, et al. 2020), we investigated the proportion of promoter-enhancer

loops and CTCF loops in Hi-C 3.0 and Micro-C. Examining these cis-regulatory elements enriched at loop anchors showed that CTCF-independent, loops which are the protocol-specific loops, were more likely to be promoter enhancer loops. Since promoter-enhancer loops are closer in the spatial distance, they are better detected with protocols that have smaller fragments. Micro-C and Hi-C 3.0, both with shorter average fragment length than conventional Hi-C, were better at detecting P-E loops.

Interchromosomal interactions

In interphase, chromosomes mostly occupy their own volume in the nucleus without intermingling. Due to the nature of these chromosome territories, proximity ligation detects more interactions within chromosomes (cis) and fewer interactions between chromosomes (trans). We tested if varying Hi-C protocols and Micro-C differ in detecting inter-chromosomal interactions. Compartmental domains detected in both cis and trans. We tested if detection of compartment differs between cis and trans. We computed the number and the strength of A and B compartments in cis and trans for various protocols and h1-hESC and HFFc6 cells. First, we counted the number of A and B compartments detected in cis and trans. In H1-hESC we compared the number of A and B compartments detected in each chromosome using all Hi-C protocols and Micro-C and found that smaller chromosomes consistently have more A compartments than B. in both cis and trans, Micro-C identifies more A compartments in bigger chromosomes compared to Hi-C (Figure 3.6a). Second, in HFFc6 cells, Hi-C identifies more

B compartments than Micro-C and consistently Micro-C identifies more A compartments than Hi-C (Figure 3.6b). Third, in H1-hESC, more A compartments can be found on smaller chromosomes whereas in HFFc6, more A compartments were detected on larger chromosomes. H1-hESC and HFFc6 are different in detecting A and B compartments (Figure 3.6 c, d-f).

In chapter II, we showed that standard Hi-C protocol and Micro-C have a weak performance detecting trans contacts. Detection of trans contacts was improved in Hi-C after the addition of extra DSG crosslinkers. Overall, we observed stronger trans interactions for all Hi-C experiments compared to Micro-C.

To follow up, we investigated if the interactions between specific individual chromosomes differed between Hi-C and Micro-C. We quantified the average compartment strength for A-A, B-B and A-B interactions in trans for both Hi-C and Micro-C. First, we observed that in trans, A compartments were stronger than B compartments. As expected, small chromosomes were interacting more than the big chromosomes. Second, Micro-C captures stronger A-A interactions between all pairwise chromosomes compared to Hi-C. A-B interactions were slightly higher in Micro-C compared to Hi-C. However, Hi-C protocols capture stronger B-B interactions across all pairwise chromosomes. Third, H1-hESC and HFFc6 have different strengths of A-A, B-B, and A-B interactions.

In summary, we did not find a relationship between the number of A and B compartments and the strength of these compartments detected by

different methods.

Sequencing depth and genome structure detection in 3C-based methods

The resolution of Hi-C or Micro-C experiments depends on two parameters, fragment size, and sequencing depth. Since we already discussed the effect of the fragment size in the results section, here we tested the effect of sequencing depth on feature detection. We tested how compartment strength and loop detection changed at various sequencing depths. We down sampled and compared deep datasets derived from H1-hESC and HFFc6 from a minimum of 200 Million reads to a maximum of 2 Billion reads. We then quantified and compared compartments and loops from these 10 experiments. First, we found that compartment identifications were comparable for all read depths (Spearman correlation > 0.9) (Figure 3.7 a, b, c, d). Second, we observed that the compartment strength (for A and B) is similar for all read depths (Figure 3.7 c, d). Third, more loops were detected as the number of reads increased (Figure 3.7 g, h). Loop detection did not improve after a certain read depth in conventional Hi-C , which suggests that for conventional Hi-C, saturation was reached at a lower read depth than Hi-C 3.0 and Micro-C. Importantly, at all read-depths, the number of loops detected increased with finer fragmentation and additional cross-linking. To summarize, a higher sequencing depth significantly improved the detection and quantification of chromatin loops without affecting the detection and the quantification of chromatin compartments.

Discussion

In chapter II, we have shown that different protocols could be used to detect small and large-scale structures in 3C-based methods. Here, we tested the effect of double digestion with DdeI and DpnII after cross-linking with FA+DSG (FA+DSG-DdeI+DpnII, referred to as “Hi-C 3.0”) (Lafontaine et al. 2021). We observed that using two enzymes shortened the fragment size compared to individual enzyme digestion. We found that the Hi-C 3.0 protocol affected the distance-dependent contact probability. Compared with data obtained by single DdeI or DpnII digests, in the data obtained with the Hi-C 3.0 protocol increased contacts for loci separated by less than 10 kb, making the results from this protocol more similar to results obtained with protocols using MNase digestion. However, longer distance contacts more closely resembled data obtained with protocols using single restriction enzymes than data obtained with protocols using MNase. Combined, this protocol improved the short-range signal without loss of the long-range signal.

Loop strength increased in Hi-C 3.0 compared with data obtained with protocols that use a single restriction enzyme. We found ~6,000 more looping interactions than with either single DpnII or single DdeI digestion. Furthermore, the average enrichment of contacts at these looping interactions was also higher for data obtained with the double digestion protocol. Nonetheless, the MNase library remained superior in detecting loops, both in number and in contact strength. Importantly, when we investigated compartmental interactions we found that smaller restriction fragments did not result in a loss of quantitative detection of preferential compartmental

interactions. In summary, the FA+DSG-DdeI+DpnII double-digest protocol allowed for the efficient detection of both loops and compartments in a single protocol.

Hi-C 3.0, as a single protocol, can be used to understand the genome-wide contacts in a cost-efficient way. Additionally, using specific protocols to detect chromatin structures at specific length scales brings additional variation comparing different samples. Using one protocol to understand genome structures simplifies the interpretation of the experiments.

a

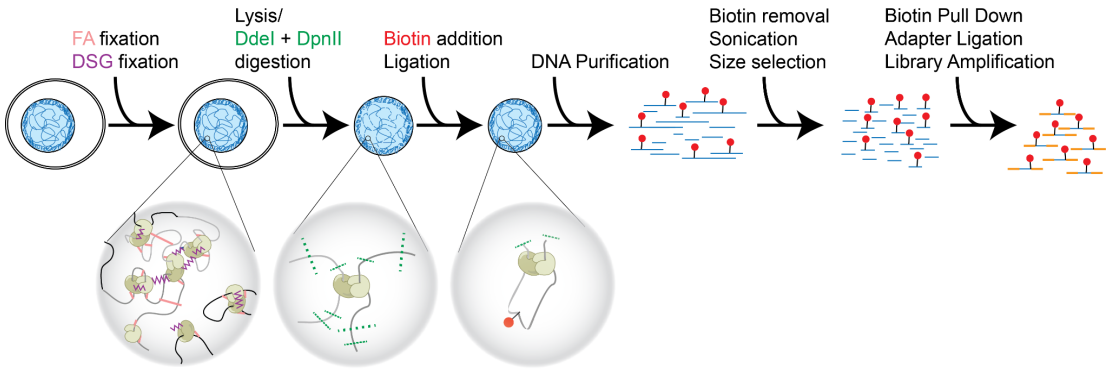

Figure 3.1: An illustration of the step in Hi-C 3.0 protocol

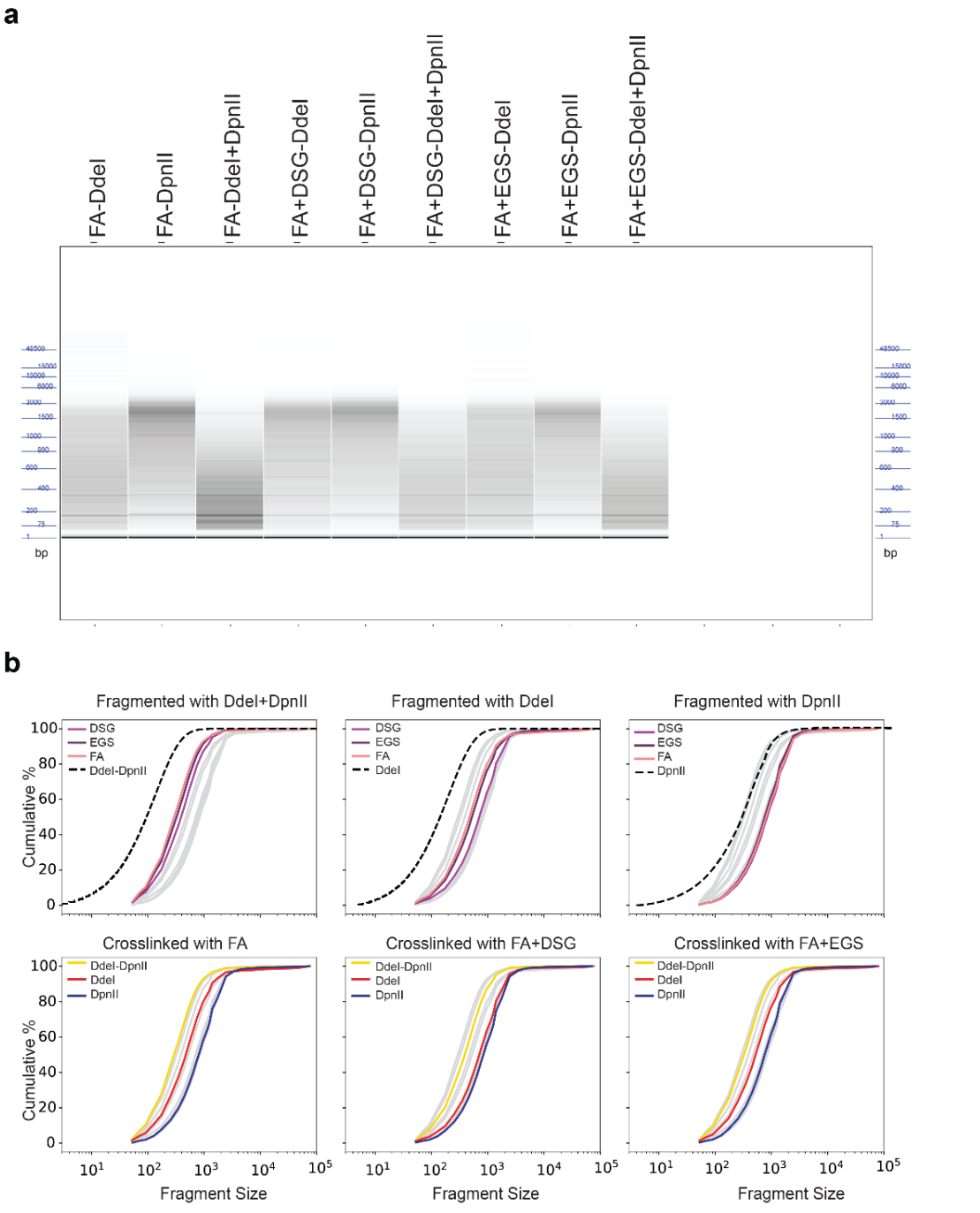
a
FA-DdeI
FA-DpnII
FA-DdeI+DpnII
FA+DSG-DdeI
FA+DSG-DpnII
FA+DSG-DdeI+DpnII
FA+EGS-DdeI
FA+EGS-DpnII
FA+EGS-DdeI+DpnII
bp
b
Fragmented with DdeI+DpnII
Fragmented with DdeI
Fragmented with DpnII
DSG
EGS
FA
DdeI-DpnII
DdeI
DpnII
Cumulative %
Crosslinked with FA
Crosslinked with FA+DSG
Crosslinked with FA+EGS
Fragment Size
c
Hi-C 2.5 DdeI
Hi-C 2.5 DpnII
Hi-C 3.0 DdeI-DpnII
log 10 # of reads
log 10 Fragment Size
Fragments density

Figure 3.2: Hi-C that uses two enzymes for digesting the chromatin shortens the fragment size

a. Fragment size distributions from Fragment Analyzer for specified protocols.

b. Cumulative distributions of fragmented DNA in HFF cells stratified for cross-linking agents (top row) or restriction enzymes (bottom row). Dashed lines in each of the panels represent expected fragment size distribution from *in silico* digestion of hg38 for enzymes indicated. Gray lines represent all data from all other enzymes (columns).

c. Shows the expected fragment sizes and y-axis represents the number of reads that are assigned to these specific fragments. Panel shows the graphs for libraries that are deeply sequenced.The color indicates the local density of fragments.

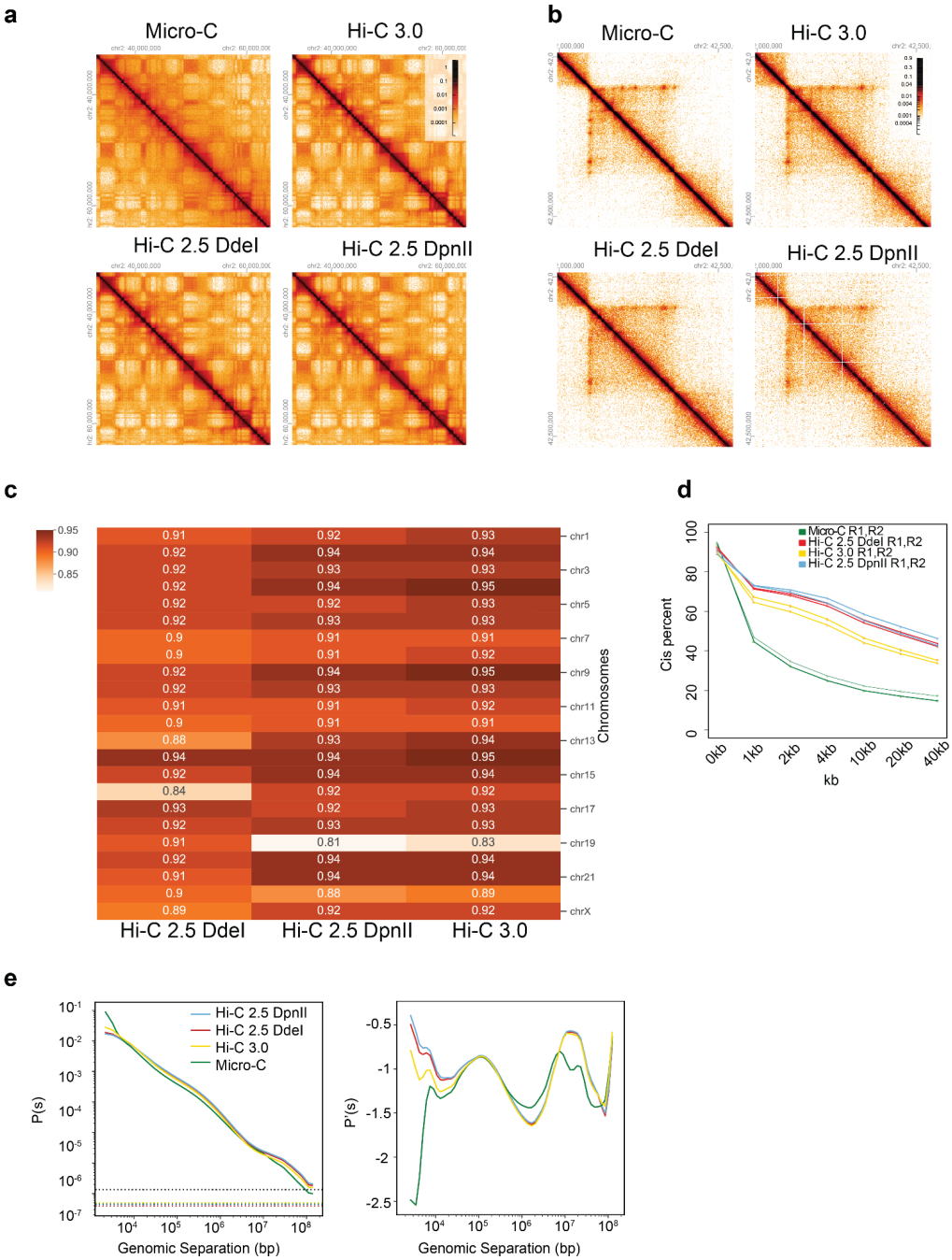

Figure 3.3: Loop detectability and strength increase when the chromatin is digested with two restriction enzymes while preserving strong compartment signal.

a. Interaction maps (log transformed) of Micro-C, Hi-C- 3.0, Hi-C 2.5 DdeI, Hi-C 2.5 DpnII. Chromosome 2, 35mb-65mb, 50kb bins

b. Interaction maps (log transformed) of Micro-C, Hi-C 3.0, Hi-C 2.5 DdeI, Hi-C 2.5 DpnII. Chromosome 2, 42mb-42.6mb, 2kb bins

c. HicREP correlation of Micro-C with Hi-C 2.5 DdeI, Hi-C 2.5 DpnII and Hi-C 3.0 for all chromosomes.

d. Graph shows the cis percent for Micro-C, Hi-C 3.0, Hi-C 2.5 DdeI, Hi-C 2.5 DpnII calculated for different genomic distances. ~40% of cis interactions are between 0-1kb in Micro-C experiments.

e. *P*(*s*) plot showing distance dependent contact probability of interactions detected with 4 protocols applied to HFFc6 cells. Dashed lines show the percentage of trans interactions for each dataset left), Derivative of the *P*(s) plots (right) shown in panel e.

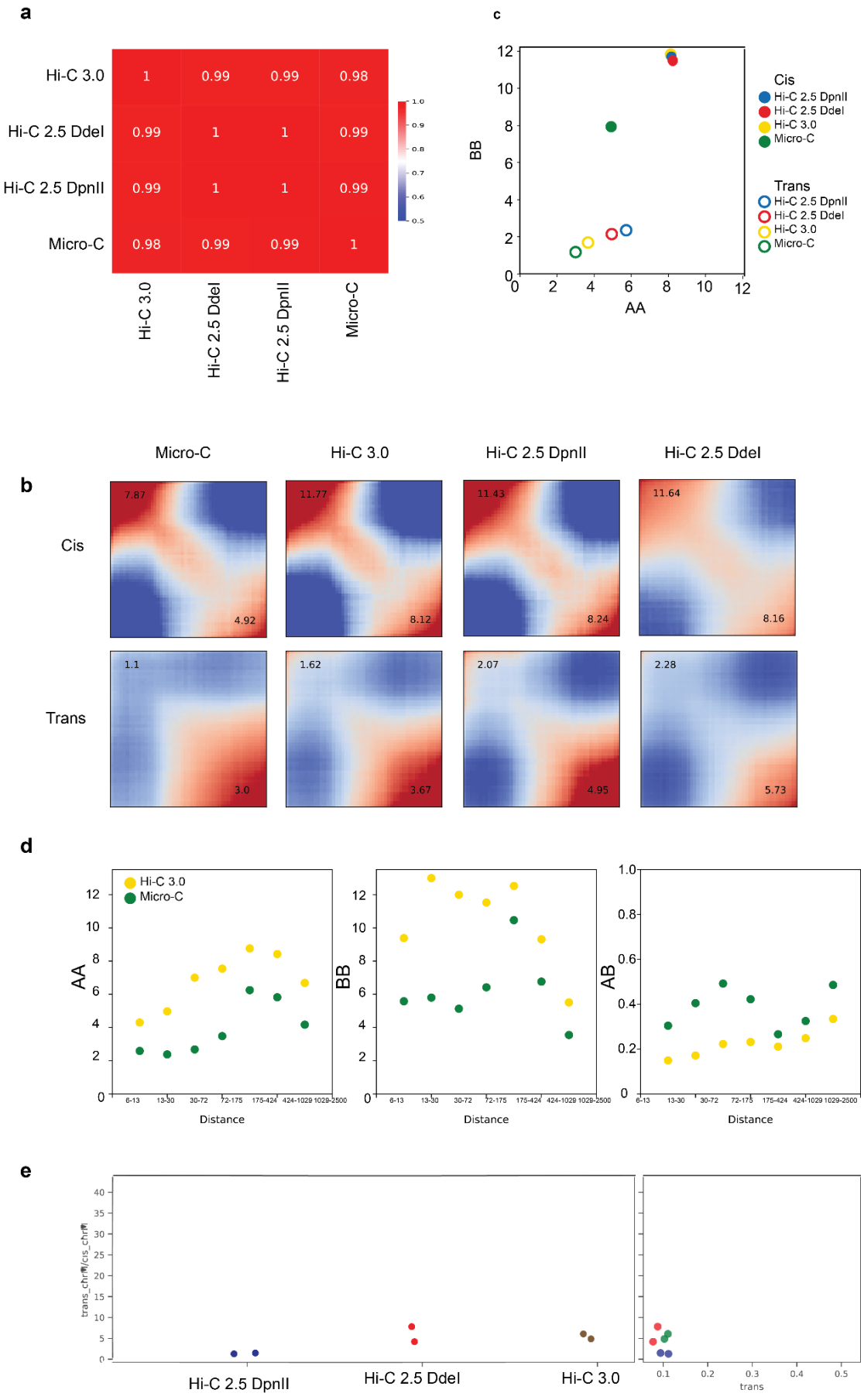
a
Hi-C 3.0
Hi-C 2.5 DdeI
Hi-C 2.5 DpnII
Micro-C
1
0.99
0.99
0.98
0.99
1
1
0.99
0.99
1
1
0.99
0.98
0.99
0.99
1
Hi-C 3.0
Hi-C 2.5 DdeI
Hi-C 2.5 DpnII
Micro-C
c
BB
AA
Cis
Hi-C 2.5 DpnII
Hi-C 2.5 DdeI
Hi-C 3.0
Micro-C
Trans
Hi-C 2.5 DpnII
Hi-C 2.5 DdeI
Hi-C 3.0
Micro-C
b
Micro-C
Hi-C 3.0
Hi-C 2.5 DpnII
Hi-C 2.5 DdeI
Cis
Trans
7.87
4.92
11.77
8.12
11.43
8.24
11.64
8.16
1.1
3.0
1.62
3.67
2.07
4.95
2.28
5.73
d
Hi-C 3.0
Micro-C
AA
BB
AB
Distance
6-13
13-30
30-72
72-175
175-424
424-1029
1029-2500
e
trans_chrM/cis_chrM
Hi-C 2.5 DpnII
Hi-C 2.5 DdeI
Hi-C 3.0
trans

Figure 3.4: Compartments are stronger in Hi-C experiments compared to Micro-C

a. Spearman correlation of the first eigenvectors called using Eigenvector decompositions on matrices generated from Micro-C, Hi-C 3.0, Hi-C 2.5 DdeI, Hi-C 2.5 DpnII. The first eigenvector represents the compartment signal.

b. Saddle plots of the genome-wide interaction maps of the data shown in Figure 3.3 a . A-A (bottom right corner) and B-B (upper left corner) compartment signals are stronger in Hi-C experiments compared to Micro-C.

c. Quantification of the compartment strength using saddle plots of cis and trans interactions A-A (x-axis) and B-B (y-axis).

d. Quantification of the A-A compartment strength using saddle plots of cis for Hi-C 3.0 and Micro-C calculated for multiple genomic distances (left). Quantification of the B-B compartment strength using saddle plots of cis for Hi-C 3.0 and Micro-C calculated for multiple genomic distances (middle). Quantification of the A-B compartment strength using saddle plots of cis for Hi-C 3.0 and Micro-C calculated for multiple genomic distances (right).

e. Quantification of protocol introduced noise as defined by inter-mitochondrial interactions (chrM with chr1-22), normalized by intra-mitochondrial (chrM with chrM) interactions for three Hi-C variants.

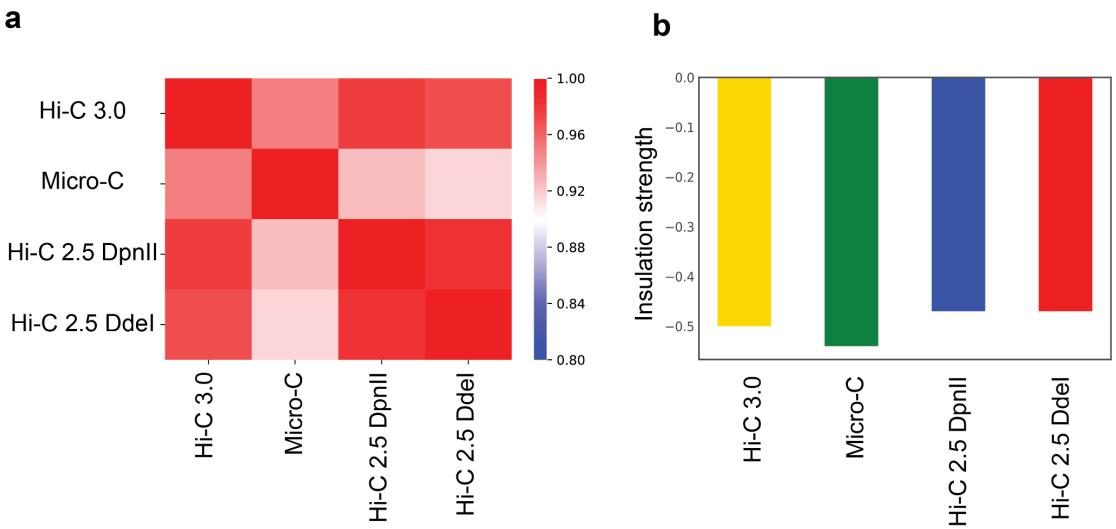

Figure 3.5: Micro-C has the strongest TAD boundaries

a. Pearson correlation of genome wide insulation scores calculated for Micro-C, Hi-C 3.0, Hi-C 2.5 DdeI, Hi-C 2.5 DpnII.

b. The insulation score of TAD boundaries for all four protocols mentioned in Figure 3.4 a. Mean insulation score of all four protocols set as a threshold to define the TAD boundaries.

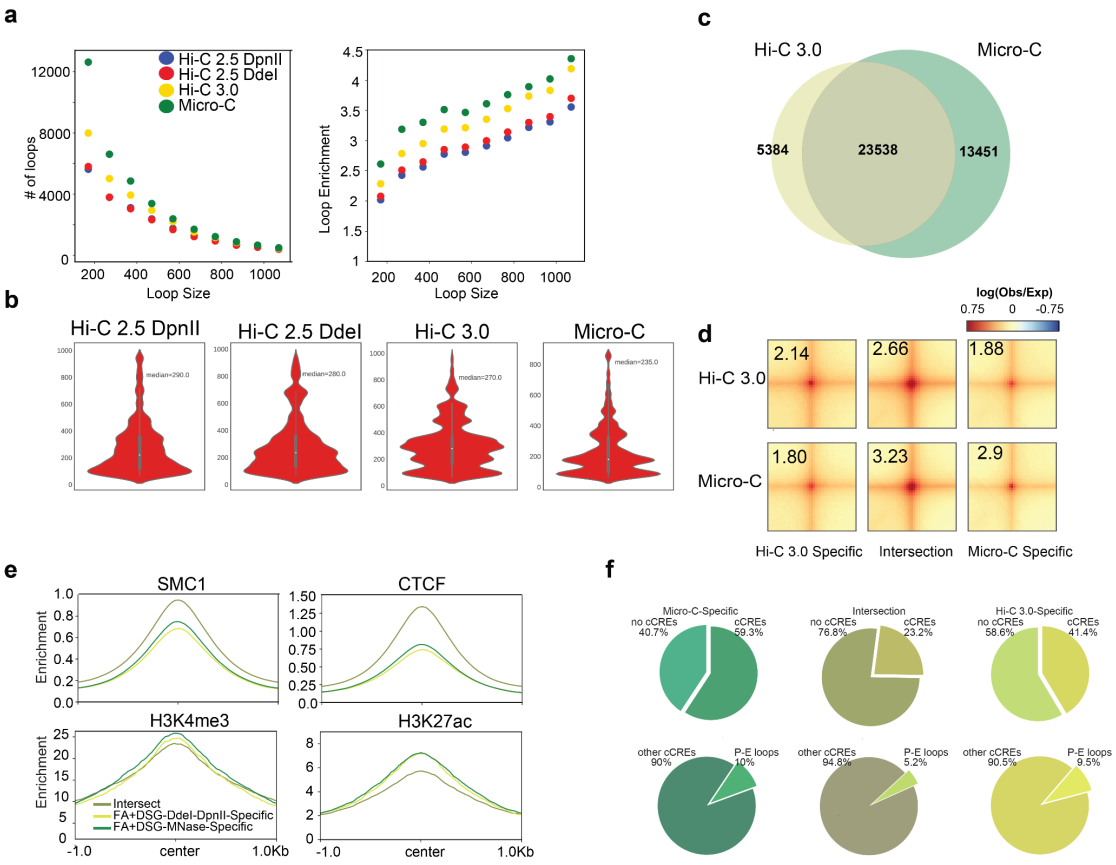

Figure 3.6: Characterization of interactions and chromatin features of loop anchors detected with Hi-C 3.0 and Micro-C.

a. The number of loops detected within 100kb intervals (loop size) starting at 70kb detected using Micro-C, Hi-C 3.0, Hi-C 2.5 with DdeI, Hi-C 2.5 with DpnII (left). Bin intervals: 70-170 kb, 170-270 kb, 270-370,.....,970-1070 kb. Loop strength of 1,000 loops sampled from 100 kb intervals. When less than 1,000 loops were available, loop strengths for available loops were used (right).

b. Loop size distribution (genomic separation of anchors that create the chromatin loop) of loops detected in Micro-C, Hi-C 3.0, Hi-C 2.5 DdeI, Hi-C 2.5 DpnII.

c. Venn diagram shows the overlap between the number of loops detected with Hi-C 3.0 and Micro-C.

d. Loop pileups of 3 loop lists described in Figure 3.6 c. Common loops that are detected in both Hi-C 3.0 and Micro-C are stronger than the protocol specific loops that are detected in only Hi-C 3.0 or only Micro-C.

e. Comparison of CTCF, SMC1, H3K4me3 and H3K27ac enrichments at loop anchors centered at open chromatin regions. Open chromatin regions (ATAC Seq) located within the anchor coordinates were used to center the average enrichments. Anchors were separated into sets detected by Hi-C 3.0, Micro-C or both.

f. Percentage of cCREs and promoter-enhancer elements located at loop anchors specific to Hi-C 3.0, Micro-C or shared between them.

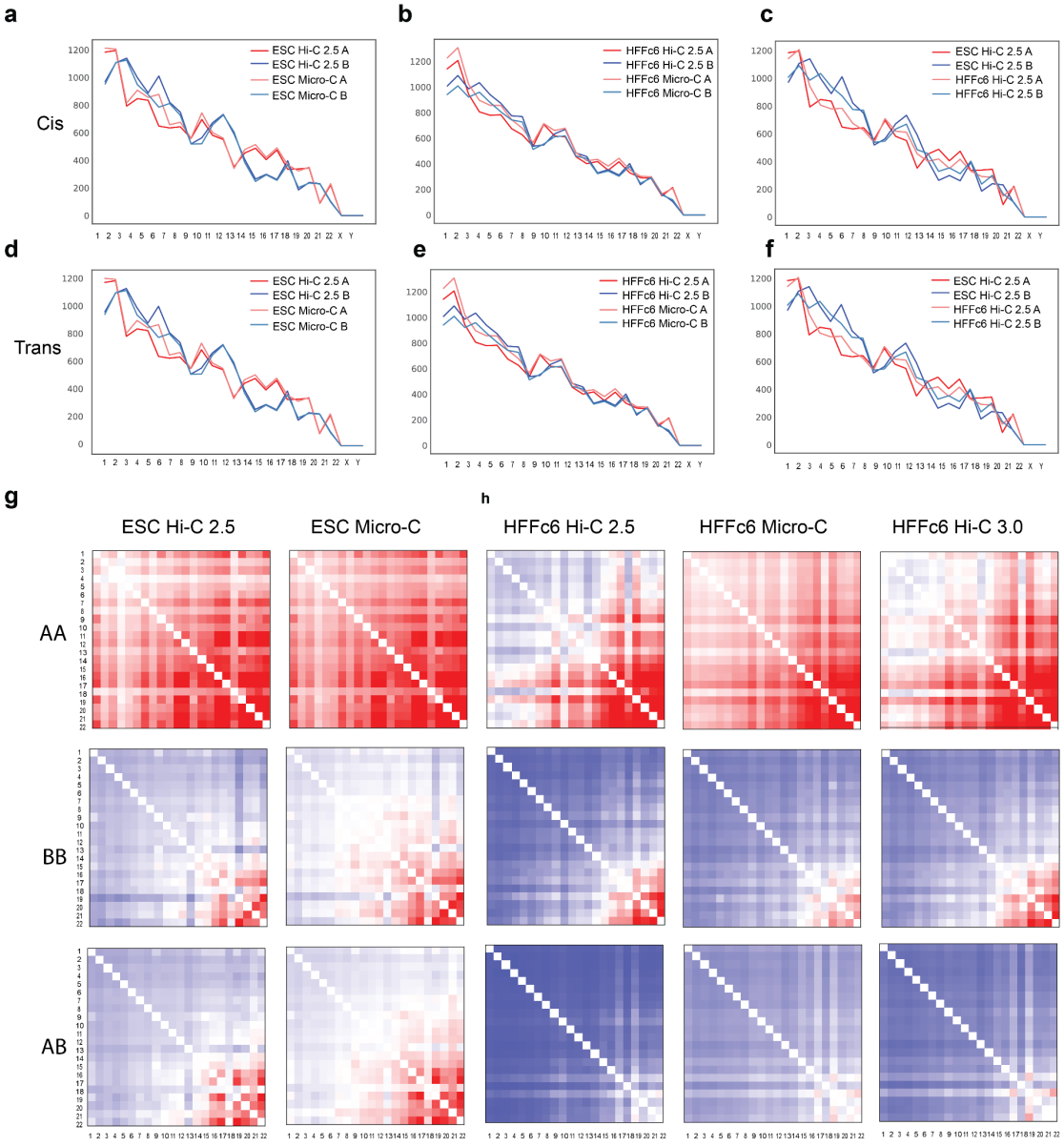

Figure 3.7: Detection of trans contacts vary between protocols and cell types

a-c. The number of A and B compartments in cis; Hi-C 2.5 compared to Micro-C for H1-hESC (a) and HFFc6 (b), Hi-C 2.5 comparing cell types; H1-hESC and HFFc6 (c)

d-f. The number of A and B compartments in trans Hi-C 2.5 compared to

Micro-C for H1-hESC (a) and HFFc6 (e), Hi-C 2.5 comparing cell types; H1-hESC and HFFc6 (f)

g. Average inter-chromosomal interactions quantified in A-A, B-B and A-B compartments in Hi-C 2.5 and Micro-C in H1-hESC

h. Average inter-chromosomal interactions quantified in A-A, B-B and A-B compartments in Hi-C 2.5, Hi-C 3.0 and Micro-C in HFFc6

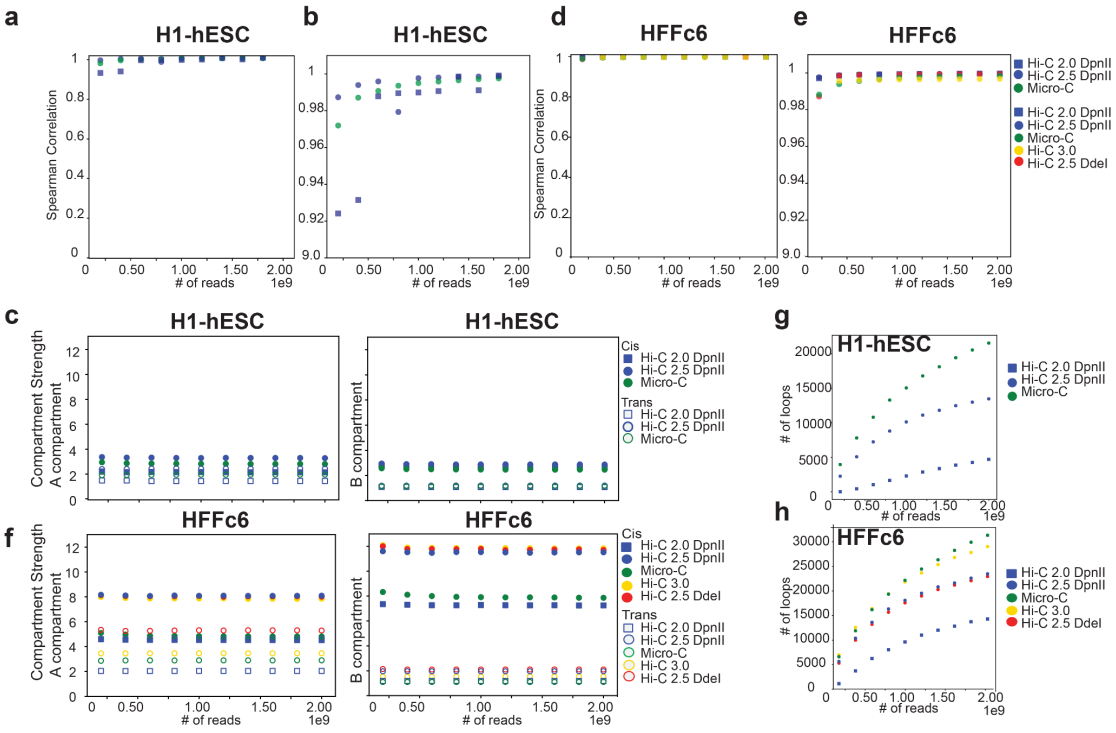

Figure 3.8: Compartmentalization is read depth independent; however, detection of chromatin loops is dependent on the read depth.

a. Spearman correlation of the eigenvectors for different sequencing depths in H1-hESC. Each point represents one sampled experiment. X-axis shows the sequencing depth (200M reads-2B reads) and y- axis shows the correlation of the eigenvectors for each depth with the eigenvector of the experiment with 2 Billion reads. The right plot shows the.

b. Zoomed correlations of H1-hESC from Figure 3.8 a

c. Compartment strength of A compartment for experiments with different read depths quantified in cis and trans for H1-hESC (left). Compartment strength of B compartment for experiments with different read depths quantified in cis and trans for H1-hESC (right).

d-f Analysis that is shown in a-c repeated for experiments performed in HFFc6 cells

g. # of loops detected in experiments with different read depths in H1-hESC.

h. # of loops detected in experiments with different read depths in HFFc6.

Chapter IV: Comparison of methods that measure genome folding

Preface

Chapter IV has unpublished data. The manuscript of this chapter is in preparation with 4DN Joint analysis group members.

Summary

For over a decade, a variety of methods have been developed to study genome organization. Each of these methods captures some information about the genome, ranging from locus-specific to genome-wide chromatin interaction profiles. However, none of them gives the entire picture of the chromosome folding in the finest possible resolution at all scales. To better understand the complete picture of chromosome folding, data gathered with these methods should be analyzed in a comparative and integrative manner. Here, we performed an integrative analysis of ligation-based and ligation-free methods (Hi-C, Micro-C, SPRITE, GAM, PLAC-Seq, ChIA-PET) measuring chromosome folding at local and genome-wide scales in commonly used cellular model systems: H1-hESC and HFFc6. These analyses revealed details of the genome folding that were not detected by individual protocols.

Introduction

Unveiling 3D genome organization is crucial to understanding gene expression and the relationship between regulatory elements and their targets (Boltsis et al. 2021; Busslinger et al. 2017; Creyghton et al. 2010; Engreitz,

Ollikainen, and Guttman 2016; Kubo et al. 2021; Rada-Iglesias et al. 2011; Uhler and Shivashankar 2017; van Steensel and Furlong 2019; Zheng and Xie 2019). 3C and its derivatives have greatly improved our understanding of the 3D genome (Goel and Hansen 2021). 3C-based methods capture contacts genome-wide or in a targeted manner and rely on the proper cross-linking and ligation of the fragments that are in close proximity (Belaghzal, Dekker, and Gibcus 2017; Krietenstein et al. 2020; Fullwood et al. 2009; Fang et al. 2016). Ligation-free methods have also been developed to detect genome-wide interactions (Quinodoz et al. 2018; Beagrie et al. 2017). Each of these methods capture an aspect of chromatin folding with different sensitivity and resolutions, however, none of these methods provide a complete picture of chromosome folding.

Each of the methods measuring 3D genome organization has several inherent limitations and advantages that depend on the experimental approach and parameters. Hi-C and Micro-C, which are ligation-based methods that capture genome-wide contacts, are dependent on the key experimental parameters, such as crosslinking and fragmentation. Efficient crosslinking is important for keeping the chromatin intact for proximity ligation. While proper fragmentation is crucial to get a uniformly fragmented genome that will enable gathering information from all regions of the genome. Hi-C protocol is limited by the restriction sites of an enzyme used for fragmenting the genome. For instance, if a genomic region does not have restriction sites the contacts will not be captured in this region. An ideal restriction enzyme should have uniformly distributed and closely spaced restriction sites all over

the genome. To overcome some limitations in fragmenting the genome, multiple restriction enzymes have been used to increase the probability of getting a uniformly fragmented genome. (See Hi-C 3.0 in Chapter III) . Hi-C 3.0 provided approximately 2kb resolution. Hi-C variant uses MNase for fragmentation, which produces nucleosome level resolution (~150 bp). However, Micro-C has a bias toward nucleosome-free active regions and has limited capacity in detecting long-range interactions.

PLAC-Seq and ChIA-PET are also ligation-based methods that capture genome-wide contacts of specific protein-bound chromatin regions (Fang et al. 2016; Fullwood et al. 2009). They have been used to detect long-range interactions between active promoters, active enhancers and regulatory elements. Similar to 3C-based protocols, PLAC-Seq and ChIA-PET methods include chromatin cross-linking and fragmentation steps. They also include a chromatin pull-down step to enrich for targeted regions using antibodies against the targeted protein. The performance of these methods depends on the specificity of the antibodies used. Because these methods are region-specific, they do not capture the genome-wide contacts in an entirety, while they capture the contacts of targeted regions at high resolution.

SPRITE is a ligation-free method that captures genome-wide contacts and is dependent on the split-pool barcoding strategy (Quinodoz et al. 2018). The protocol requires extensive optimizations for efficient barcoding. SPRITE can capture multiway genomic contacts by detecting clusters of chromatin that share the same barcodes because they travel together during the split-pool

step. The SPRITE cluster sizes can range from two to up to thousands of DNA molecules. Fragments that are identified to be in the same cluster are assumed to interact with each other. Considering the small clusters this assumption is correct. However, this assumption might be misleading for bigger clusters since it is unclear what is the distance between the fragments within the same cluster. In this chapter, I will investigate the features of various cluster sizes detected by SPRITE. Such features include the quantification of chromatin compartments and loops.

GAM is another ligation-free method that captures genome-wide contacts by measuring the distance between genomic regions within the nucleus (Beagrie et al. 2017). It is based on random cryosectioning of the crosslinked nucleus and the resulting sections are then barcoded and sequenced. GAM uses different cross-linking techniques compared to other genome-wide techniques, it uses 4% and 8% paraformaldehyde to crosslink the cells. All other techniques use FA with DSG or EGS for crosslinking the chromatin. If two genomic regions are observed to be in multiple nuclei sections at the same time they are expected to be in close distance to each other. Information from multiple nuclei sections is combined to predict the distance between genomic regions. The resolution of structures detected in GAM is limited by the number of nuclei that are sectioned for the experiment. For example, to generate a genomic interaction map at 30 kb resolution, ~407 nuclei have been used (Beagrie et al. 2017).

The information from all these methods described above can be

visualized by contact matrices. Pre-processed files of Hi-C, Micro-C, ChIA-PET, PLAC-Seq, SPRITE, and GAM are downloaded from the 4DN Data portal for the analysis in this chapter. This allows diminishing the variations that might occur due to different aligners and using variant tools for data processing. Briefly, all datasets are mapped using BWA-mem and processed using pairtools as described in 4DN Data Portal pipelines (https://data.4dnucleome.org/). Cooler is used to create multi-resolution cooler files for downstream analysis (Abdennur and Mirny 2020).

The relationship between genome organization and gene activity is not clear. Spatial organization of the genome correlates with chromatin landscape, gene expression, and replication timing (Boninsegna et al. 2022; Yildirim et al. 2022; Rao et al. 2014; Zheng and Xie 2019). To determine chromatin landscape, several methods including DamID have been developed. DamID is used to quantify protein-DNA interactions using antibodies targeting the protein of interest. The position of a gene relative to nuclear speckles (high transcriptional activity) and nuclear lamina (low transcriptional activity) can be used as a predictor of gene expression levels. A method called TSA-Seq has been developed to measure the distance between genes and nuclear speckles or nuclear lamina. TSA-Seq, a cytological ruler, and DamID mapping provide information for gene activation and gene silencing (Wang et al. 2021; Chen et al. 2018; Vogel, Peric-Hupkes, and van Steensel 2007). Dividing cells follow a program called replication timing that determines duplication speed of their DNA in S phase. Replication timing is a conserved process between eucaryotes, stable within the cell type

and could be important for conducting genomic alterations for specific parts of the genome (Marchal, Sima, and Gilbert 2019; Pope et al. 2014). To determine the replication timing, a method called Repli-Seq, has been developed. These methods are used to extract information about the function of the genome and how it correlates with genome organization.

In this chapter, we have integrated contact information from the methods explained above and quantified the chromatin structures to gain a broad understanding of genome folding. Additionally, we used functional assays such as TSA-Seq, DamID-Seq and Repli-Seq to investigate the correlation between genome structure, chromatin landscape, and replication timing.

Results

- **Distance-dependent interaction frequency of all methods.**

To better understand the complete picture of the chromosome folding, we have used Hi-C, Micro-C, ChIA-PET, PLAC-seq, SPRITE and GAM data that were generated in two commonly used cell lines: H1-hESC and HFFc6.

To compare genomic interactions obtained in these methods at different scales, we generated heatmaps (Figure 4.1 a). Visual inspection of the heatmaps showed that short-range structures are more pronounced in the pull-down methods (lower panel, PLAC-Seq and ChIA-PET) compared to unbiased genome-wide methods (upper panel, Hi-C, Micro-C, SPRITE, and GAM). However, long-range structures were more obvious in the unbiased

genome-wide methods, particularly in Hi-C and SPRITE.

The number of sequencing reads obtained from these experiments were considerably different, which is expected considering the differences in these methods (Figure 4.1 b). Before starting data analysis, we first wanted to check the quality of these datasets. One aspect of determining data quality is to compare the percentage of intra-chromosomal (cis) versus inter-chromosomal (trans) interactions. The expectation is that cis interactions should be higher than the trans interactions due to arrangements of chromosomes into territories (Cremer and Cremer 2010). All of these methods showed around 85% cis contacts, which confirms the good quality of these data (Figure 4.1 c).

Distance-dependent interaction probability (P(s)) graph shows that these methods differ in both short and longer genomic distances. Micro-C showed similar interaction probability in small distances to pull-down methods in HFFc6 (Figure 4.1 d). While, Hi-C, SPRITE, and GAM, showed significantly lower interactions in small distances compared to Micro-C and pull-down methods

Next, using HiCRep we correlated the interaction profiles of these methods. HiCRep calculates the Pearson correlation of specific genomic distances of the methods and takes the weighted average of these correlations to calculate the stratum-adjusted correlation coefficient. Stratum-adjusted correlations of these methods showed that the variation between methods does not exceed the cell-type-specific variations because the cluster

formation was similar in different cell types regardless of the methods (Figure 4.1 e). SPRITE and GAM don't cluster with their corresponding cell types. SPRITE experiments performed in H1-hESC and HFFc6 cluster together. Based on these results we concluded that the SPRITE experiments might have some bias that leads to clustering two cell types together. Additionally, we observed that the GAM experiment produces the most distinct interaction map (HicRep correlation < 0.5). For a given cell type observing strong correlations between methods (assures the reproducibility of the genomic interactions. It also suggests that these methods are comparable. Next, we compared specific structures detected in these methods.

- **Compartment detection/strength.**

Genomes are segregated into active A and inactive B compartments which correlate with euchromatin and heterochromatin, respectively. To investigate the detectability of A and B compartments in all methods, we identified compartments in Hi-C, Micro-C, ChIA-PET, PLAC-seq, and SPRITE in HFFc6 and H1-hESC cell lines using eigenvector decomposition. Although, eigenvector decomposition initially developed for Hi-C experiments, we observed strong correlation between the first eigenvectors of different methods (Spearman correlation coefficient > 0.73) (Figure 4.2 a). In other words overall detection of the compartments was similar in all methods (Figure 4.2 a). SPRITE eigenvector showed a slightly lower correlation compared to other methods indicating that SPRITE might be detecting a set of compartments that are unique to the method. To quantify the strength of the

detected A and B compartments, we used eigenvectors, which provided a number of observations explained below (Figure 4.2 b): First, HFFc6 cell line showed stronger compartments than H1-hESC cell line. For a given method the compartment strength of HFFc6 is around three fold higher than the compartment strength of H1-hESC. Second, the strongest compartmentalization was observed in SPRITE and Hi-C methods. Hi-C detected quantitatively similar A and B compartments, while SPRITE detected stronger B compartments and weaker A compartments (Figure 4.2 b). Third, the strength of the compartments showed significant differences among methods. These differences were more obvious in the HFFc6 cell line that has stronger compartments as compared to H1-hESC.

Compartment detection in SPRITE showed a lower correlation (0.73) with other methods indicating that these compartments may have some differences. We indeed found some interesting examples of compartments, which were detected as a whole B compartment in SPRITE but showed alternating A and B compartment patterns in other methods (Figure 4.2 c). This result raised the possibility that SPRITE may detect compartments as large blocks, which would reduce the number of detected compartments. To test this, we plotted the number of compartments detected in each of these protocols and found that SPRITE has the lowest number of compartments in HFFc6 but not for H1-hESC (Figure 4.2 d). To determine the size of these compartments, we sampled the same number of A and B compartments from each method and plotted their cumulative size distribution. SPRITE showed larger compartments in HFFc6 but not in H1-hESC (Figure 4.2 e). The

compartments are weak in H1-hESC but strong and better defined in HFFc6 which indicate that HFFc6 highlights more differences between protocols than H1-hESC. Furthermore, we investigated the genomic regions that are identified differentially between methods. These genomic regions are assigned to different compartments in at least one of the methods. We found that ChIA-PET for RNA Pol II and SPRITE identify distinct sets of compartments in both HFFc6 and H1-hESC (Figure 4.2 f). These sets of new compartments might be special interactions that are harder to detect by other methods. Alternatively, they may be artifacts caused by the experimental technique.

- **Preferential interactions detected by 3D Methods**

Multiple studies have shown that the spatial organization of the genome correlates with chromatin landscape, gene expression, and replication timing (Boninsegna et al. 2022; Yildirim et al. 2022; Marchal, Sima, and Gilbert 2019; Wang et al. 2021) .To investigate these correlations with the 3D methods described above, we analyzed DamID of LMNB1, TSA-Seq for predicting gene expression levels, and Repli-Seq for determining replication timing. The TSA-Seq data is targeting SON, Laminin subunit beta 1 (LMNB1) , Nuclear factor NF-kappa-B p105 subunit (NFK1), Centromere protein (CENB1), RNA Polymerase I Subunit E (POLR1E). For correlation analysis, we performed pairwise Spearman correlation between these assays and the compartment signal detected by the 3D methods. Overall, compartments showed correlation below 0.2 for both HFFc6 and H1-hESC (Figure 4.3 a,b).

The signal for DamID, TSA-Seq is locus-specific which explains the lower correlation for the whole genome. Additionally, using two time points (early and late) for Repli-Seq might not be informative enough to cover the whole genome information. Correlation of specific regions in the genome could be used for more accurate correlation coefficients. An example of such region is shown in Figure 4.3 c. To investigate the relationship between these regions and 3D methods we extracted the genomic regions that have the strongest 20% signal from these methods. Then we quantified the preferential interactions between these regions using 3D methods. We observed that preferential interactions are more diverse in HFFc6 compared to H1-hESC (Figure 4.3 d, e). We don't observe differences between methods in H1-hESC, however, the differences become more obvious in HFFc6, stronger in Hi-C and SPRITE compared to all other methods (Figure 4.3 d, e). Hi-C and SPRITE capture preferential interactions better than other methods, especially the interactions between speckles that are measured by SON TSA-Seq. In summary, the detection of preferential genomic contacts is method dependent. It's crucial to select the right method to detect desired interaction in the genome.

- **Insulation Strength**

Topologically associating domains (TADs) are mid-megabase structures that have been suggested to play a fundamental role in gene regulation (Nora et al. 2012; Dixon et al. 2012). The boundary of a TAD has been shown to restrict interactions between cis-regulatory elements and

genes outside of the TAD, thereby contributing to the regulation of gene expression (Crane et al. 2015; Hyle et al. 2019; Lupianez et al. 2015; Melo et al. 2020). To measure the boundaries of TADs, we examined the insulation scores of all 3D methods.

First, we calculated the Pearson correlation of the genome-wide insulation scores of all methods. All methods are highly correlated except SPRITE (> 0.75 for H1-hESC and > 0.8 for HFFc6) (Figure 4.4 a,b). SPRITE has lower correlations (0.35-0.54 in H1-hESC, 0.47-0.71 for HFFc6.) than other methods in both H1-hESC and HFFc6 (Figure 4.4 a,b). An example of insulation score plotted below interaction maps in Figure 4.4 c. Tracks of insulation scores look similar between methods but less pointy in SPRITE.

Second, we investigated the distribution of insulation scores and calculated their average (Figure 4.4 a,b). We considered the insulation scores below the average as weak (blue), the insulation scores above the average as strong (red). Although insulation scores showed some variations in different 3D methods, SPRITE did not clearly distinguish the strong and weak boundaries (Figure 4.4 d, e). Then, we quantified the insulation strength of possible TAD boundaries by selecting the strongest insulation scores and found that Micro-C has the strongest insulation followed by ChIA-PET protocols in H1-hESC. However, ChIA-PET protocols have stronger insulation strength in HFFc6. Finally, SPRITE has the weakest insulation in both cell types (Figure 4.4 f,g).

All 3D methods show similar performance detecting and quantifying TAD boundaries. It seems like SPRITE shows a strong quantification

detecting large scale structures like compartments but the quantifications are weaker for smaller-scale structures such as TADs.

- **Chromatin Loops**

Chromatin loops are the finest structures detected in 3D-based methods. Detectability and the strength of the loops are crucial to understand the genome-wide high-resolution chromatin interactions such as promoter-enhancer contacts. Detection of such promoter-enhancer loops is important to understand the gene expression. 3D-based methods have been improved to better detect these promoter-enhancer loops. We identified loops in Hi-C, Micro-C, ChIA-PET and PLAC-Seq. SPRITE data did not have the sequencing depth that is required for loop identification therefore we did not call loops in SPRITE data. ChIA-PET PolII and Micro-C detect the most loops in both H1-hESC and HFFc6 (Figure 4.5 a,b). Investigating the genomic separation of these loops we found that ChIA-PET PolII detects the smallest loops compared to other methods (Figure 4.5 c,d).

To compare the strength of the loops, we took the union of loops detected in all methods. We piled up and averaged the union loop set to quantify their strength in all these methods and we found that ChIA-PET PolII and Micro-C have the highest loop strengths (Figure 4.5 e,f). It is very appealing to see a genome-wide method, Micro-C, with a great resolution detecting thousands of loops and yet these loops are stronger than all other methods. Next, we examined the combination of loops detected using different protocols for H1-hESC. Extracting the top 31 combinations we

observed that i) loops detected in pull-down methods are the smallest loops, ii) Micro-C detects both small and large loops iII) Loops that are detected by multiple methods are stronger than the loops that are detected by one or two methods (Figure 4.5 g,h). In summary, the union set of all these methods provides a great reference into cell-type-specific looping interactions including promoter-enhancer contacts. This reference set can be used to compare the performance of methods at loop detection.

- **SPRITE clusters show differences in quantifying genomic structures**

SPRITE is unique compared to other methods that are considered in this study by capturing not only pair-wise but also higher-order contacts in different chromatin clusters. The number of possible multi-way interactions increases with larger clusters. However, since multi-way contacts are captured only by SPRITE, for comparisons the information about multi-way contacts is not used here. Instead, all possible pairwise interactions are created from each cluster separately and then these interactions are combined. In small clusters it is possible that all fragments are interacting, however, in bigger clusters, it is difficult to predict the interactions between fragments. Fragments in bigger clusters don’t have equal interaction probability. These differences between small and big clusters led us to investigate cluster-specific features in SPRITE. We quantified and compared the structures such as compartments, loops detected by SPRITE clusters with sizes 2-10, 11-100, 101-1000, 1001-10000 fragments. We found that i) the

number of reads as well as the cis percent decrease as the clusters get larger in SPRITE (Figure 4.6 a,b). ii) The interaction frequencies differ for different genomic distances such that smaller clusters have more short-range interactions whereas larger clusters have more long-range interactions (Figure 4.6 c). Also, larger clusters detect more inter-chromosomal interactions.

We sought to quantify the compartments and loops for different SPRITE clusters to determine cluster size effect on structure detection. For compartment identification, we used the eigenvector computed from combined clusters. Quantification of compartment strength showed that clusters with size 2-10 are more powerful detecting compartments (Figure 4.6 e). It is important to point that the compartment detection is independent of the read count (Figure 4.6 d). Since SPRITE data did not have the resolution for loop detection, we used the union loop sets that are used in Figure 4.5 to quantify the loop strengths in different SPRITE clusters. Similar to compartment analysis, cluster sizes of 2-10 are better at detecting chromatin loops in both H1-hESC and HFFc6 (Figure 4.6 f, g). Compartment and loop quantification of the combined clusters show a slightly lower signal than the clusters that have 2-10 fragments.

Due to lack of bioinformatics tools, SPRITE data is not analyzed with its whole potential. Since information detected by different SPRITE clusters is different it would be crucial to understand this difference and select specific clusters for specific comparisons. Larger clusters of SPRITE seem to collect information

distinctly compared to other methods. However, further investigation is needed to characterize these large clusters.

Discussion

Comprehensive comparison of Hi-C, Micro-C, ChIA-PET, PLAC-seq, and SPRITE showed that genome-wide methods are good at detecting larger chromatin structures such as compartments, and pull-down methods are good at detecting smaller scale structures such as loops. Unlike other genome-wide methods, Micro-C performs the best at detecting high-resolution chromatin loops.

Hi-C, Micro-C, SPRITE and GAM are advantageous over pull-down methods by capturing contacts genome-wide and in an unbiased manner. The unbiased detection of structures allows performing a comparative analysis of the interaction frequencies for the whole genome. On the other hand, pull-down methods target specific regions in the genome which provide enhanced interactions of these regions. This bias toward specific regions makes it challenging to compute if these interactions are statistically significantly enriched. Additionally, genome-wide methods require a high sequencing depth hence higher cost to reach a specific resolution, whereas pull-down methods can reach the desired resolution with less sequencing depth. This makes pull-down methods cost-efficient compared to genome-wide methods.

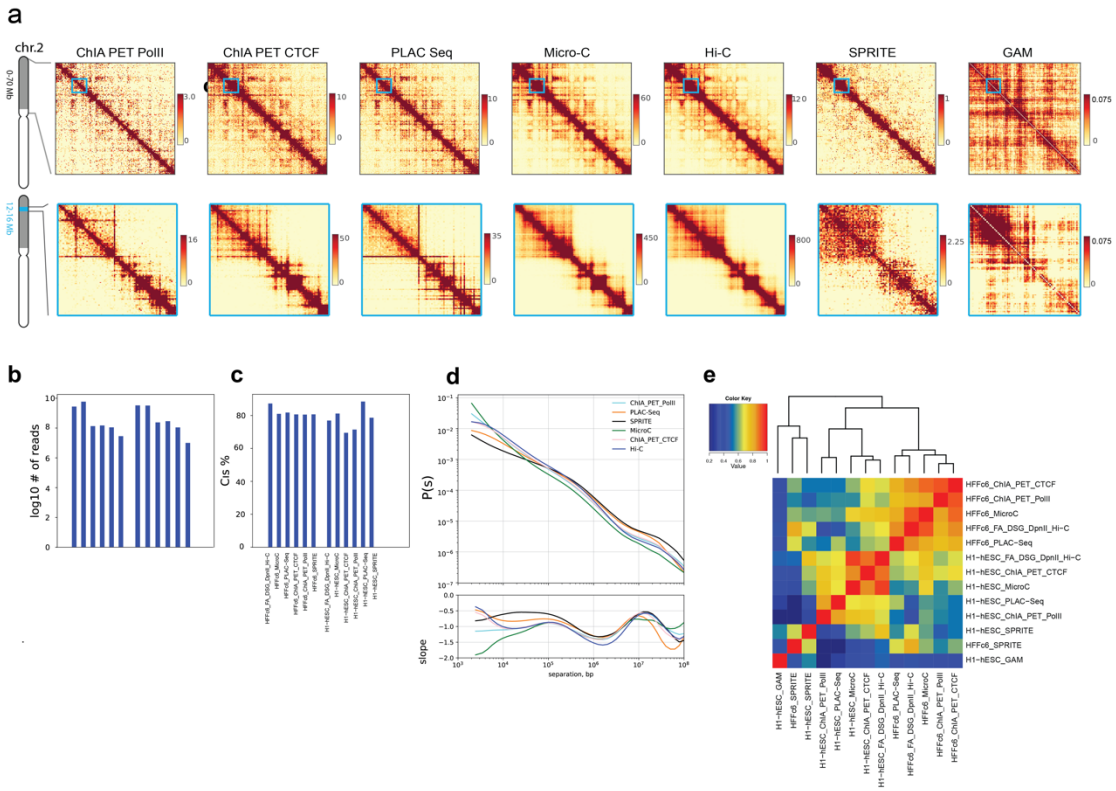

Figure 4.1: Methods Overview

a. Heatmaps of contact maps generated using Hi-C, Micro-C, ChIA PET, PLAC Seq, SPRITE and GAM (upper heatmaps 100kb bins, chr 2, 0-70mb, lower heatmaps 25 kb bins, chr2, 12-16mb)

b. The number of valid pairs for Hi-C, Micro-C, ChIA PET, PLAC Seq, SPRITE

c. The % of cis contacts in each method specified in Figure 4.1 b

d. *P*(*s*) plot showing distance dependent contact probability of interactions detected with all protocols applied to HFFc6 cells (top). Derivative of the *P*(s) plots shown in panel d (bottom).

e. HiCRep correlation of Hi-C, Micro-C, ChIA-PET, PLAC-Seq, SPRITE and GAM

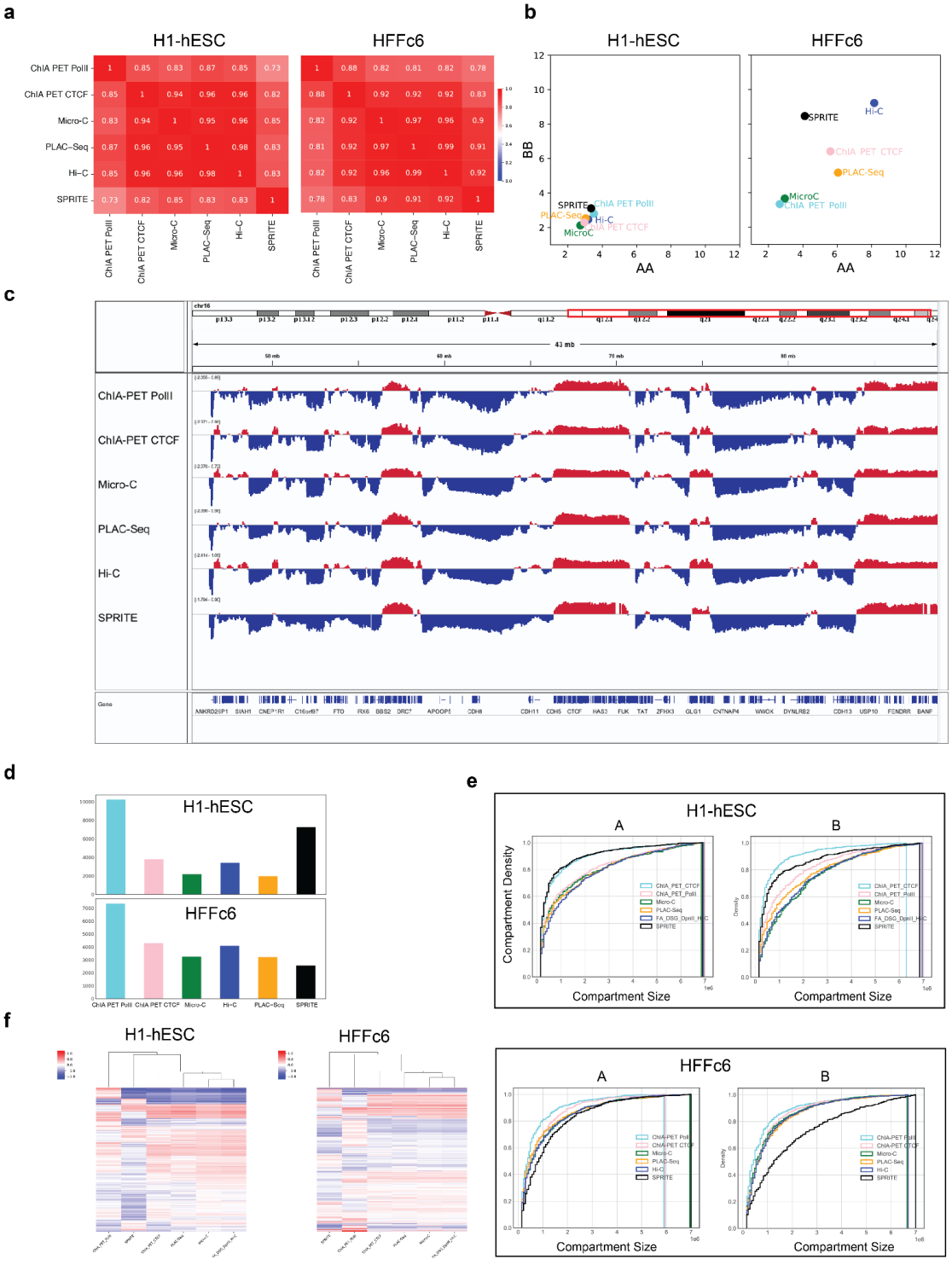

Figure 4.2: Compartments are better detected in Hi-C and SPRITE

a. Spearman correlation of the first eigenvectors identified using

Hi-C, Micro-C, ChIA PET, PLAC Seq, SPRITE in H1-hESC and HFFc6

b. Quantification of the A-A and B-B compartment strength using saddle plots of cis for Hi-C, Micro-C, ChIA PET, PLAC Seq, SPRITE in H1-hESC and HFFc6

c. An example of IGV track of the first eigenvectors showing a large B compartment in SPRITE that is detected differently from other methods.

d. The number of compartments detected in Hi-C, Micro-C, ChIA PET, PLAC Seq, SPRITE in H1-hESC and HFFc6

e. Cumulative plot of compartments sizes plotted for A and B compartments separately.

f. Hierarchical clustering of genomic regions are assigned to a different compartment in at least one of the methods

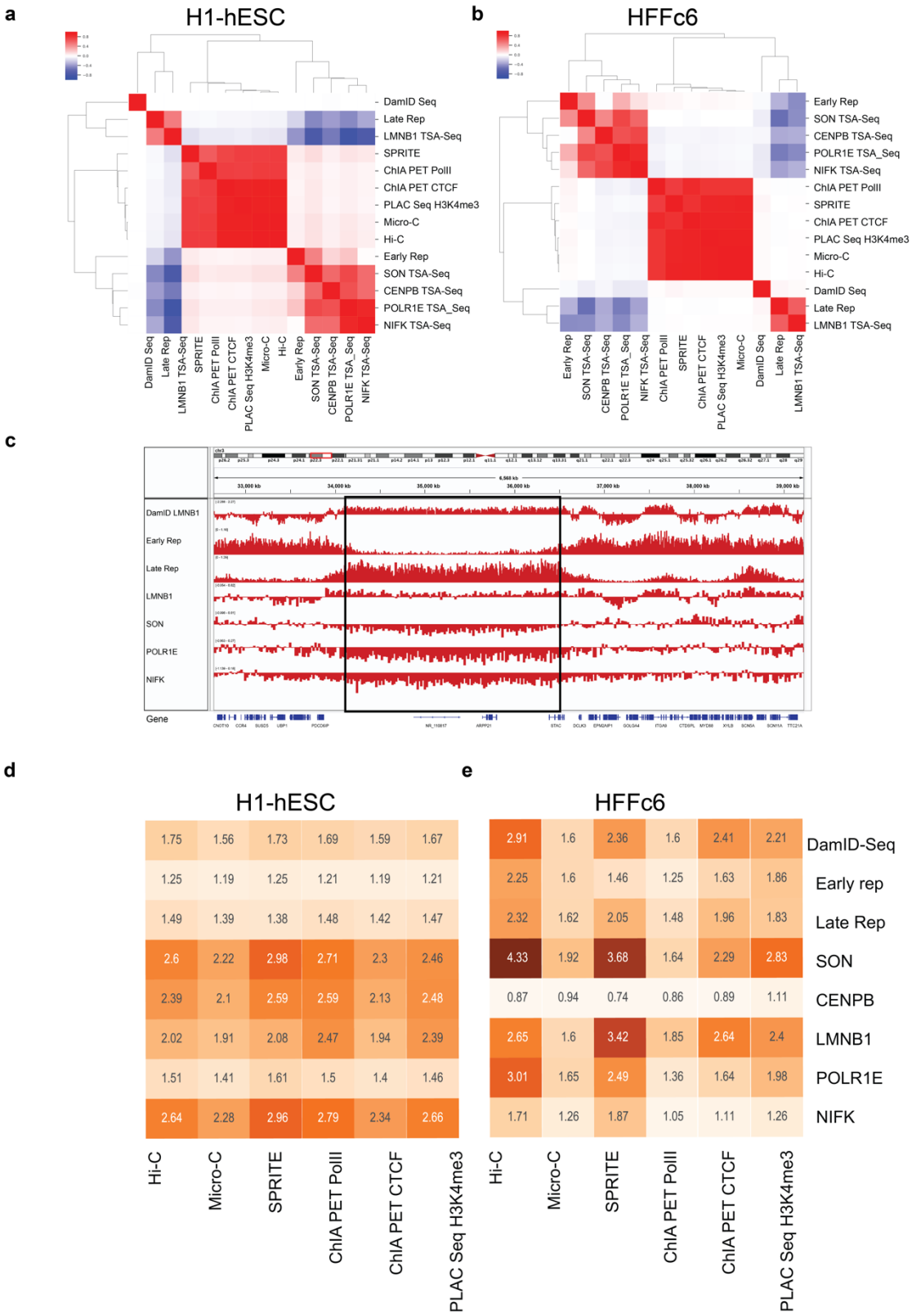
a
H1-hESC
DamID Seq
Late Rep
LMNB1 TSA-Seq
SPRITE
ChIA PET PolII
ChIA PET CTCF
PLAC Seq H3K4me3
Micro-C
Hi-C
Early Rep
SON TSA-Seq
CENPB TSA-Seq
POLR1E TSA_Seq
NIFK TSA-Seq
b
HFFc6
Early Rep
SON TSA-Seq
CENPB TSA-Seq
POLR1E TSA_Seq
NIFK TSA-Seq
ChIA PET PolII
SPRITE
ChIA PET CTCF
PLAC Seq H3K4me3
Micro-C
Hi-C
DamID Seq
Late Rep
LMNB1 TSA-Seq
c
DamID LMNB1
Early Rep
Late Rep
LMNB1
SON
POLR1E
NIFK
Gene
d
H1-hESC
1.75 1.56 1.73 1.69 1.59 1.67
1.25 1.19 1.25 1.21 1.19 1.21
1.49 1.39 1.38 1.48 1.42 1.47
2.6 2.22 2.98 2.71 2.3 2.46
2.39 2.1 2.59 2.59 2.13 2.48
2.02 1.91 2.08 2.47 1.94 2.39
1.51 1.41 1.61 1.5 1.4 1.46
2.64 2.28 2.96 2.79 2.34 2.66
Hi-C
Micro-C
SPRITE
ChIA PET PolII
ChIA PET CTCF
PLAC Seq H3K4me3
e
HFFc6
2.91 1.6 2.36 1.6 2.41 2.21 DamID-Seq
2.25 1.6 1.46 1.25 1.63 1.86 Early rep
2.32 1.62 2.05 1.48 1.96 1.83 Late Rep
4.33 1.92 3.68 1.64 2.29 2.83 SON
0.87 0.94 0.74 0.86 0.89 1.11 CENPB
2.65 1.6 3.42 1.85 2.64 2.4 LMNB1
3.01 1.65 2.49 1.36 1.64 1.98 POLR1E
1.71 1.26 1.87 1.05 1.11 1.26 NIFK
Hi-C
Micro-C
SPRITE
ChIA PET PolII
ChIA PET CTCF
PLAC Seq H3K4me3

Figure 4.3: Preferential Interactions differ between methods and cell types

a.b. Spearman correlation of eigenvector detected using Hi-C, Micro-C, ChIA PET, PLAC Seq, SPRITE, TAS Seq with different targets, DamID and Replication Timing in H1-hESC (a) and HFFc6 (b)

c. 1D tracks show enrichment for DamID Seq, TSA-Seq and Replication timing.

d.e. Preferential interactions quantified in Hi-C, Micro-C, ChIA PET, PLAC Seq, SPRITE using DamID Seq, Early and Late replication timing and TSA Seq targeting SON, LMNB1, POLR1E, CENB1, NIFK in H1-hESC (d) and HFFc6 (e)

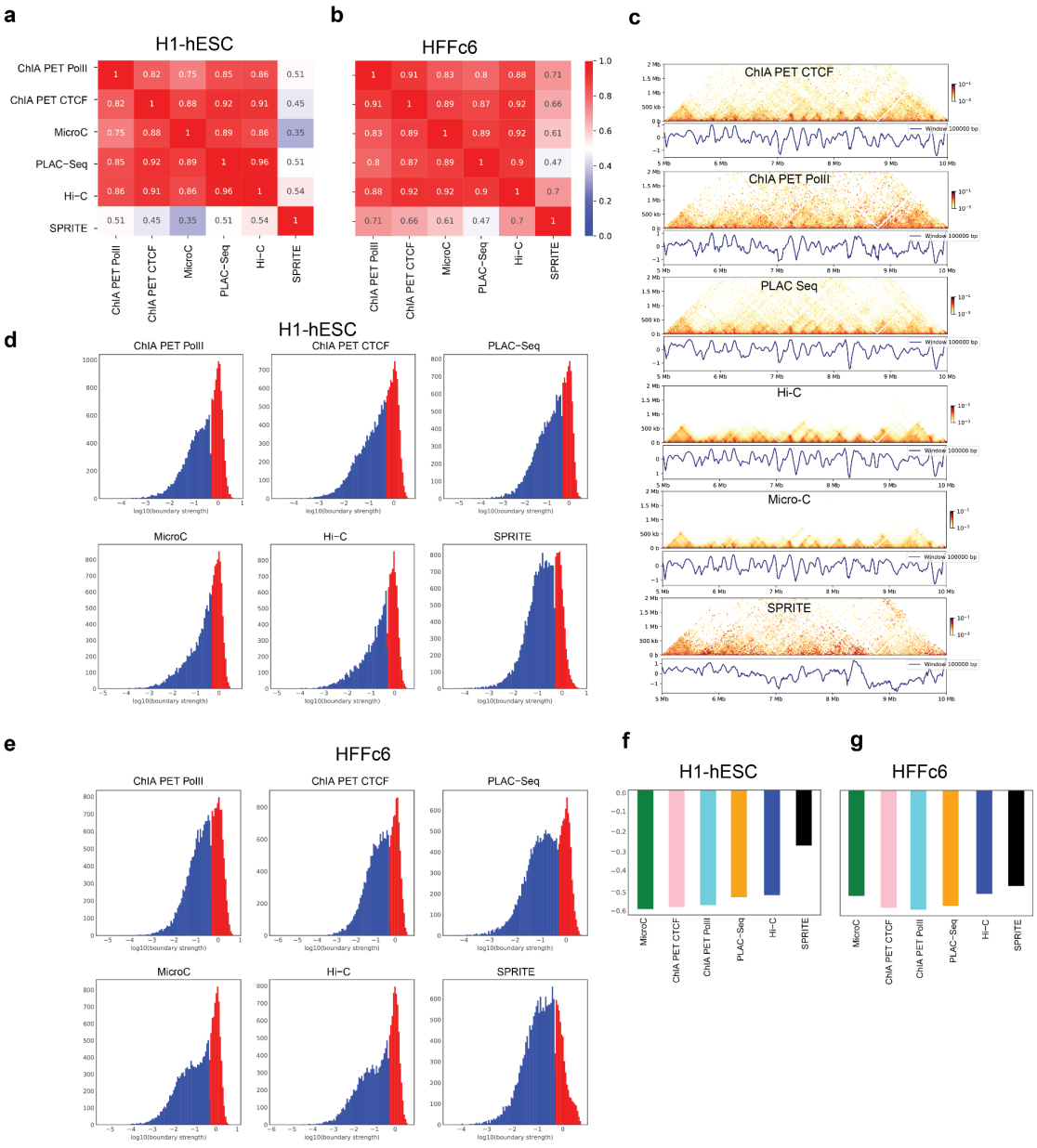

Figure 4.4: TAD boundaries are consistent between methods except SPRITE

b. Pearson Correlation of insulation score for H1-hESC (a) and HFFc6 (b).

c. Heatmaps generated from ChIA PET, PLAC Seq, Hi-C, Micro-C and SPRITE with their insulation score

d-e. Distribution of insulation scores for H1-hESC (d) and HFFc6 (e). Values larger than the mean are plotted in red and values smaller than the mean are plotted in blue.

f.g Quantification of average insulation strength of the strong insulation scores (Figure 4.4 d, e red) for H1-hESC (f) and HFFc6 (g).

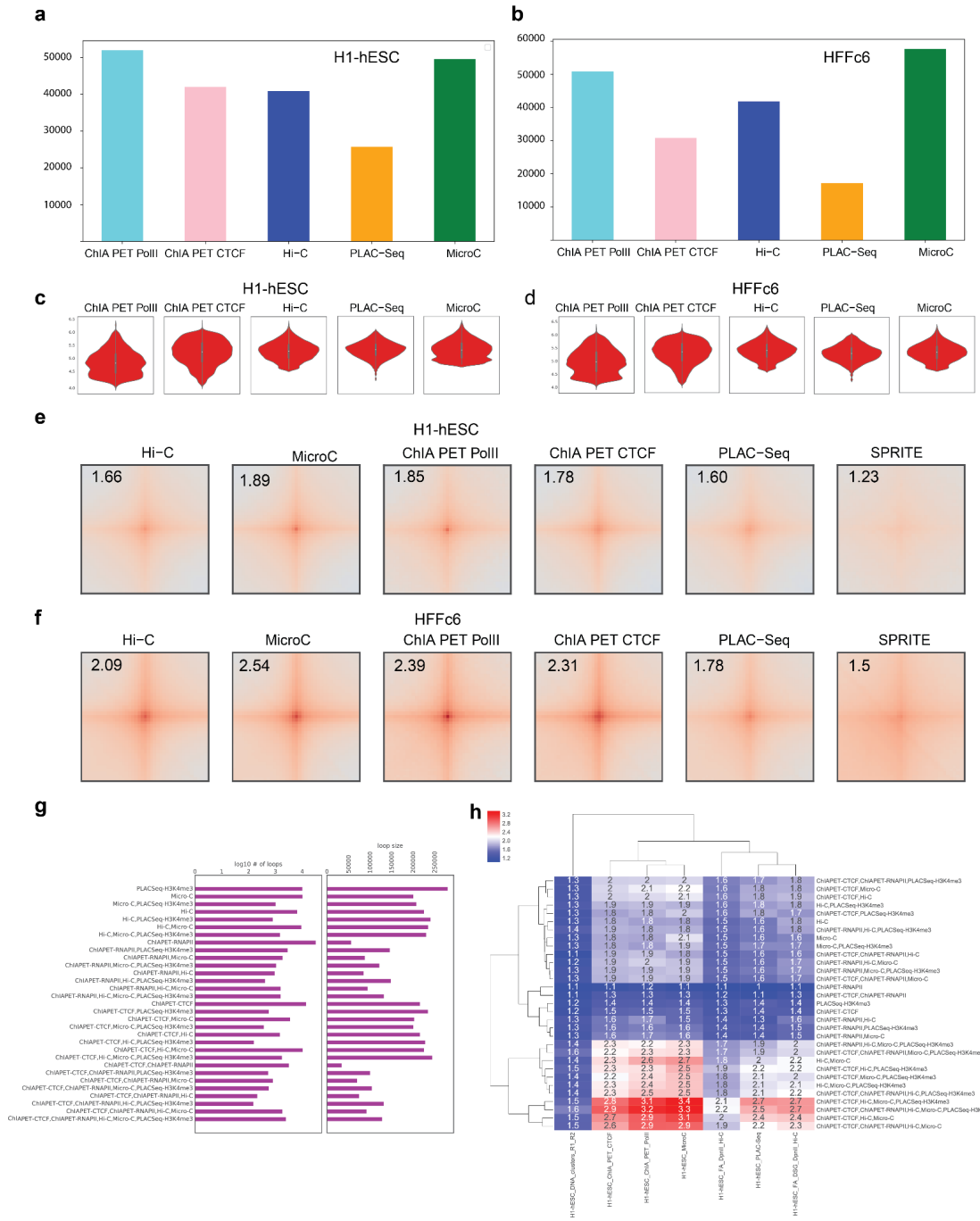

Figure 4.5: ChIA-PET PolII and Micro-C detects the most loops

a-b. The number of loops detected in ChIA PET, PLAC Seq, Hi-C, Micro-C in H1-hESC (a) and HFFc6 (b)

c-d. Loop size distribution (genomic separation of anchors that create the

chromatin loop) of loops detected in ChIA PET, PLAC Seq, Hi-C, Micro-C in H1-hESC (c), in HFFc6 (d).

e-f. Loop pileups of union loop lists created using loop sets from Figure 4.5 a for H1-hESC (e) and b for HFFc6 (f).

g. Top 31 loop combinations to of loops (left) and loop sizes (right) for H1-hESC

h. The heatmap of hierarchical clustering of loop strengths for the top 31 loop sets mentioned in Figure 4.5 g

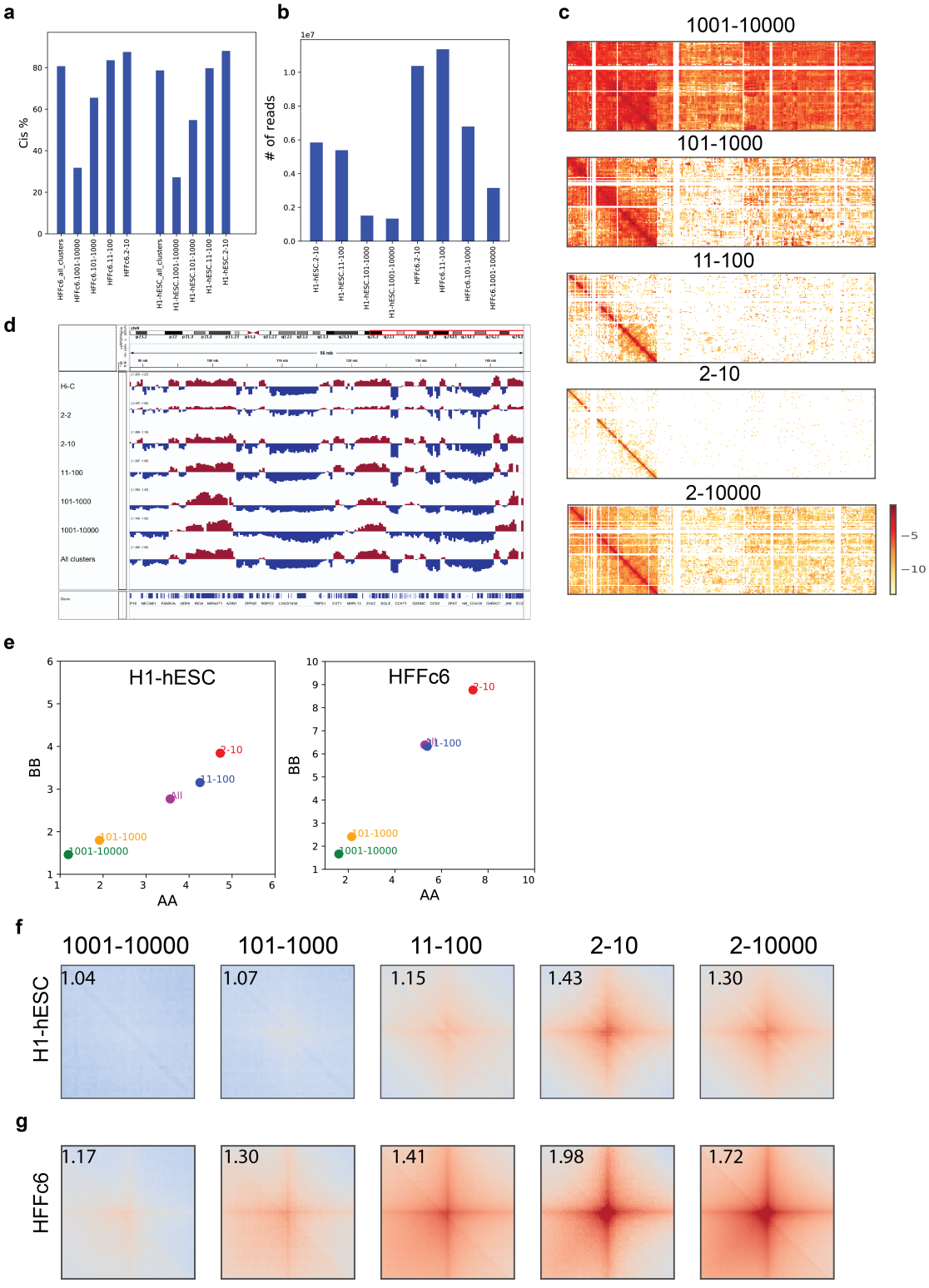

Figure 4.6: SPRITE clusters detect different chromatin features

a. The number of fragments in each SPRITE cluster. Cluster sizes are

formed using 2-10, 11-100, 101-1000, 1001-10000 fragments.

b. The % of cis contacts in each cluster

c. Heatmaps of chr 17 interacting with chromosome 17,18 and 19 for interaction map created using different cluster sizes.

d. Compartment detection of all SPRITE clusters compared to Hi-C.

e. Compartment strength of various clusters for H1-hESC (d), HFFc6 (e).

f.g Loop pileups for various cluster sizes for H1-hESC (f) and HFFc6 (g). Union loops sets described in Figure 4.5 are used for pileups.

Chapter V: Discussion

Why do we need to precisely map the genome structure?

Human DNA is about 2 meters long and fits into a 10 μm nucleus (Bickmore and van Steensel 2013) which requires extensive packaging through chromatin folding. This folding inside the nucleus is not random and is organized in a hierarchical manner. Chromosomes occupy individual volumes inside the nucleus called chromosome territories (Cremer and Cremer 2010). 3D methods have not only allowed the detection of chromosome territories but also led to the identification of layered organization inside each chromosome from compartmental domains to Topologically associated domains (TADs) and chromatin loops (Gibcus and Dekker 2013).

The organization of chromatin appears to have a great correlation with gene expression, replication timing, and distinct chromatin domains, At a global scale, A and B compartments detected with 3D methods correlate with active euchromatin and inactive heterochromatin respectively. At a smaller scale, the expression of genes within a TAD tends to be regulated similarly (Kagey et al. 2010; Schwarzer et al. 2017). This co-regulation may be important to achieve cell-type-specific gene expression during development (Bonev et al. 2017; Dixon et al. 2015). Genes are often regulated by an enhancer that is located within the same TAD where more frequent contacts occur as compared to other TADs (Symmons et al. 2014; Consortium, Moore, et al. 2020; Heintzman and Ren 2009). The frequency of enhancers and promoters interactions through chromatin looping can contribute to the

regulation of gene expression (Fudenberg et al. 2016; Ay, Bailey, and Noble 2014; Dekker and Mirny 2016; Hnisz et al. 2016; Valton and Dekker 2016). The organization of chromatin also correlates with replication timing and chromatin domains marked by specific histone modifications. Thus, understanding the organization of chromatin is crucial to unveil how genome activity is regulated and will provide insights into the regulation of replication timing and formation of chromatin domains.

In this study, we have investigated genome folding in detail and its relationship to genome activity by i) extensively evaluating the experimental parameters that determine interaction matrix in the most commonly used 3D method; Hi-C, ii) integrating 3D methods to have a complete picture of the genome folding and resolve method-specific features.

In chapter II, we have evaluated how two parameters; cross-linking chemistry and fragmentation level influence the interaction maps in 3C-based methods. We found that using multiple cross-linkers improves the signal-to-noise ratio hence the detection of small and large scale structures. Finer fragmentation also improved the detection of small-scale structures. These results led us to develop a new Hi-C method Hi-C 3.0, which is capable of detecting both small and large scale structures at relatively high resolution compared to other 3-C based genome-wide capture assays (Chapter III). The improved Hi-C 3.0 method will allow us to resolve the role of proteins involved in various scales of genome folding and will facilitate new discoveries related to genome organization.

In chapter IV, we compared 3D methods to capture different aspects of genome folding. We have also used TSA-Seq, DamID and Repli-Seq assays to investigate the performance of these methods to correlate with genome activity, chromatin landscape, and replication timing. We found that 3D methods are biased in detecting specific loop sets that are defined by ChromHMM as poised promoter, insulator, transcriptional transition, active promoter-strong enhancer, active promoter, and transcriptional elongation (Ernst and Kellis 2012). Integrative analysis of these 3D methods allows us to capture a more complete picture of genome folding and provide insights into the formation of specific loop sets.

Taken together, these results showed that improving 3D methods and combining information gained from these methods helps develop new approaches to unveil the relationship between chromatin folding and genome activity.

The relationship between genome folding and disease

Structures detected by 3D methods provide explanations for the action of regulatory elements and their targets. Such actions play a crucial role in diseases such as sickle cell anemia, β-thalassemia, human lymphomas, Glioma and limb malformations (Deng et al. 2014; Roix et al. 2003; Flavahan et al. 2016; Lupianez et al. 2015).

The proximity of a distal enhancer called the Locus control region (LCR) to β-globin genes is dynamic and differs between cell types (Deng et al. 2014; Deng et al. 2012). The interaction between LCR and globin genes

determines the level of hemoglobin production and the disruption of these interactions leads to sickle cell anemia and β-thalassemia (Wilber et al. 2010; Deng et al. 2014). The looping interactions between LCR and globin genes can be leveraged to develop potential therapies for these diseases.

Genomic regions that are in close spatial proximity have a higher probability for translocations. For instance, normal B cells have MYC, BCL, and immunoglobulin loci close in spatial distance. These regions are often translocated in B-cell lymphomas and contribute to the disease progression (Roix et al. 2003). These results point to the contributions of genome organization in the acquisition of genomic aberrations in diseases.

Gain of function isocitrate dehydrogenase (IDH) mutants are used to identify gliomas (Pirozzi and Yan 2021). IDH mutants are DNA hypermethylated and this methylation causes changes in CTCF binding sites, which then alter TAD boundaries that are located around these IDH mutants. This results in continuous expression of PDGRFA, a well-known glioma oncogene (Flavahan et al. 2016).

Rearrangement of a TAD boundary can also result in limb malformations. Disruption of TAD boundaries around WNT6/IHH/EPHA4/PAX3 loci via deletions, insertions, and inversions led to aberrant gene expression and contributed to Brachydactyly, F-syndrome, and Polydactyly diseases (Lupianez et al. 2015).

The examples above illustrate the application of 3D genome organization in the diagnosis and prognosis of diseases. Controlling gene

expression by altering the chromatin interactions could provide new opportunities for future therapies.

Measuring genome folding

Over the years the number of methods that measure 3D genome organization increased. Each of these methods measures different aspects of chromosome folding. Some of them quantify genome-wide interactions (Hi-C, Micro-C, SPRITE, GAM) whereas others specialize to capture locus-specific interactions (ChIA-PET, PLAC-Seq) (Belaghzal, Dekker, and Gibcus 2017; Krietenstein et al. 2020; Quinodoz et al. 2018; Beagrie et al. 2017; Fang et al. 2016; Fullwood et al. 2009). Below, the pros and cons of these methods are discussed.

Genome-wide methods require deeper sequencing to detect genomic loops, whereas locus-specific methods can reach the same resolution with fewer reads. Reaching the desired resolution with fewer reads makes locus-specific methods cost-efficient. However, finding the significantly enriched interactions is not straightforward for locus-specific methods since they are enriched and biased toward particular sets of interactions, and the expected interaction frequency of these regions is not known. On the other hand, genome-wide methods are unbiased and it is easier to compare different genomic regions so as to detect the enriched loci.

Hi-C and Micro-C rely on proximity ligation; they use the same cross-linkers but different strategy to fragment chromatin. Quantifying the strength of union loops detected by all methods showed that Micro-C has the strongest

loop signal indicating that Micro-C performs the best in detecting chromatin loops compared to both genome-wide (Hi-C, SPRITE, and GAM) and locus-specific methods (PLAC-Seq and ChIA-PET). Micro-C targets nucleosome-free regions which makes it biased toward active regions. Hence more interactions are detected in active regions compared to inactive regions in Micro-C. The newly developed Hi-C 3.0 performs nearly as Micro-C in detecting small-scale looping interactions and does not show a bias towards active regions, because uniform fragmentation is achieved in Hi-C 3.0 using multiple restriction enzymes.

SPRITE uses a split-pool approach and does not rely on proximity ligation. It has a unique capacity for detecting multi-way contacts within a given chromatin cluster. Fragments located in the same cluster are assumed to have a physical interaction in SPRITE. That assumption might be for small clusters, but as the cluster size increases, it becomes challenging to predict the distance and interaction frequency between two fragments that are located in the same cluster.

GAM measures the genomic distance between fragments that are co-localized in cryo-sectioned nuclei. Slice orientation and the number of slices in nuclei are the two deterministic factors for measuring the genomic distance. Slicing the nuclei from multiple orientations is required to have enough representation of the whole chromatin. Due to the nature of GAM, it is more likely to detect interactions that are enriched in the center of the nucleus and less likely to detect the interactions that are enriched in the nucleus periphery.

The number of cryo-sectioned nuclei is another factor that determines the distances in GAM. As the number of nuclei increases, it is more likely to detect less represented interactions and increase the resolution. GAM differs from other methods by measuring the distance, not the genomic interactions.

PLAC-Seq and ChIA-PET capture locus-specific interactions. PLAC-seq has some advantages over ChIA-PET has superiority over ChIA-PET in terms of using in terms of requiring 20 fold fewer cells, being 100 times more cost-effective, producing more unique read pairs, having a smaller PCR duplication rate and producing more intra-chromosomal interactions. Additionally, PLAC-seq has more specificity and sensitivity in detecting chromatin loops than ChIA-PET.

Genome folding with two mechanisms - loop extrusion and compartmentalization

Mechanisms that form chromatin loops and compartments have been proposed to be different (Nora et al. 2017; Rao et al. 2017; Schwarzer et al. 2017). Loops form by loop extrusion mechanism where cohesin pulls the chromatin until it hits boundary element, CTCF, to create a loop (Fudenberg et al. 2017; Mirny, Imakaev, and Abdennur 2019). Similarly, TADs also form by loop extrusion mechanism where the TAD boundaries are defined by the CTCF binding sites. On the other hand, chromatin compartments are predicted to form by phase separation or chromatin compaction by heterochromatin (Hildebrand and Dekker 2020).

Multiple studies have investigated the interplay between chromatin loops and

compartments by depletion or deletion of CTCF and/or cohesion (Rao et al. 2017; Nora et al. 2017; Schwarzer et al. 2017). These studies found that chromatin loops and TADs completely lost or weakened the dependence of the experiment while compartments are not changed or enhanced. One such study knocked down cohesin loading factor NIPBL using liver-specific tamoxifen-inducible Cre driver (Schwarzer et al. 2017). Depletion of NIPBL led to complete loss of chromatin loops and TADs but enhanced signal for compartments. This study showed that the loop extrusion and compartmentalization are independent. However, the analysis in this paper was done genome-wide without considering the drastic differences between scales of compartment and loop interactions. Because loop extrusion exists in smaller scales than compartments, we predicted that there is an interplay between loops and compartments in specific genomic distances where loops and compartments co-exist. To investigate this interplay, we used the published data and quantified the compartment strength for different genomic distances in wild-type mice, Tamoxifen control (TAM) and NIPBL deleted mice (Schwarzer et al. 2017). In smaller genomic distances where loops and compartments co-exist, NIPBL depletion led to stronger compartments. This effect was more prominent for A compartment, where most of the loop extrusion occurs (Figure 5.1 a). The change in compartment strength due to NIPBL depletion is higher in smaller genomic distances compared to large genomic distances. As a sanity check, we used Hi-C and Micro-C data (3 experimental conditions for H1-hESC and 3 experimental conditions for HFFc6 generated in Chapter II) to confirm that the number of loops detected

in A compartments is higher than the number of loops detected in B compartments, especially in small genomic distances (Figure 5.1 b). These results suggest that loop extrusion and compartmentalization may not be completely independent, and there may be an interplay between loop and compartment formation in small genomic distances. Understanding the loop extrusion, compartmentalization and interplay between them could be important to identify the forces that drive the chromatin to form loops and compartments. The newly developed Hi-C 3.0 could provide an advantage in investigating this interplay at various genomic scales, as Hi-C 3.0 could capture both large (compartments) and small (loops) genomic scales due to its low signal-to-noise ratio compared to other methods (conventional Hi-C and Micro-C).

CTCF and cohesin in regulating genome organization

CTCF and cohesin play a fundamental role in genome organization. However, deletion or depletion of CTCF and cohesin has minimal changes on global gene expression. CTCF is involved in multiple functions including genome organization, regulation of cell type-specific genes as a result of distal promoter-enhancer and promoter-promoter interactions, RNA splicing and RNA processing (Kubo et al. 2021; Hyle et al. 2019; Braccioli and de Wit 2019; Shukla et al. 2011; Ruiz-Velasco et al. 2017; Valton et al. 2021). Furthermore, some studies have reported that different CTCF sites have distinct functions (Khoury et al. 2020; Luan et al. 2021). Various 3C-based methods have been used to measure and classify CTCF interactions in the

aforementioned studies which makes it harder to have a complete classification of CTCF sites. Furthermore, in absence of CTCF and cohesin release factor Wings apart-like (Wapl), cohesin accumulates in 3' of active genes indicating that cohesin forms domain boundaries that are CTCF-independent (Busslinger et al. 2017; Valton et al. 2021). A deeper understanding of the CTCF and cohesin function requires genome-wide high-resolution 3D maps. The newly developed Hi-C 3.0 protocol could be used to generate high-resolution contact maps for better categorization of the CTCF and cohesin binding sites based on their functions. Additionally, Hi-C 3.0 could reveal new functions of CTCF and cohesin.

Future directions

Outstanding research has been done to unveil 3D genome organization and its functionality over the years. However, the field is still lacking answers for some of the main questions. Major discoveries in biology have often been led by developing tools/technologies to address critical questions. The genome organization field has been developing tools to map genome organization at the highest possible resolutions. In addition to these efforts, integration of imaging technologies, physics-based modeling, and machine learning approaches will be necessary to get a better understanding of genome organization.

It has been proposed that chromatin compartments form independent of chromatin loops (Rao et al. 2017; Schwarzer et al. 2017; Nora et al. 2017). However, the complete picture of the interplay between compartments and

loops is lacking. There are several questions that need to be answered to understand the formation of compartments and loops: How does a genomic region decide to form a compartment but not a loop? How do loop extrusion and chromatin compartmentalization play a role in cell fate? What are the implications of the interplay between loops and compartments in diseases? These questions could only be answered by further improving the detection of genome structures and integrating various approaches from multiple fields, such as super-resolution microscopy, physics-based modeling, and machine learning.

Chromatin loops have mostly been studied under steady-state conditions. However, growing evidence suggests that loops are highly dynamic and heterogeneous across cell populations. It would be important to understand the mechanisms that regulate the mobility of chromatin loops and how this mobility regulates expression of genes. Advancement in imaging technologies and developing new tools to track chromatin loops in live cells will provide insights into the formation and function of these dynamic loops.

Integrating population-based assays with single cell assays would be informative to better understand the chromatin organization. In population-based assays capturing a chromatin loop depends on two criteria; the strength of the loop and the number of cells these loops exist. For example, if a loop is weak and formed in only <5% of the cells, it would be very difficult to capture. Single cell assays could help to determine the number of cells that contain a specific loop, which informs about the heterogeneity in a cell

population. On the other hand, population-based assays are helpful in detecting weak loops because many weak loops will create a stronger signal for detection. Combining information from the two types of assays would be useful to learn about cell to cell variation and the strength of structures.

Trans interactions : Measuring trans interactions emerge to understand the positions of chromosomes inside the nucleus. Translocations occur within and between chromosomes that are close in spatial distance. Therefore chromosome positions play an important role in understanding translocations in cancer and developing therapeutic approaches for drug development. Additionally, it is important to know the relationship between a chromosome position and its activity. Chromosomes located in the middle of the nucleus interact more and tend to be more active than the chromosomes located at the periphery. Chromosome positions differ between cells hence detection of trans interactions is challenging. Due to proximity ligation 3C based methods are limited to detecting trans interactions. Improving current methods or developing new methods is required to understand the interactions between chromosomes and the functional outcome of these interactions.

Analysis tool development: Tool development is needed to analyze, visualize and integrate diverse datasets.

3D methods and chromatin binding assays have different resolutions. For example, current loop calling tools detect loops that are ~5kb. New tools are needed to detect chromatin loops in higher resolution than 5kb. Detecting loops in high resolution (<=1kb) would help a better integration of protein-

chromatin binding assays because these assays have ~500bp resolution. Additionally tools are needed to integrate 3D methods with binding assays.

Another important tool development is needed to integrate 3D methods with high resolution imaging data to track single molecules in the genome.

Finally, developing predictive tools (machine-learning algorithms) and models for annotating coding and non-coding variants in the genome would be useful for genome-wide chromatin structure.

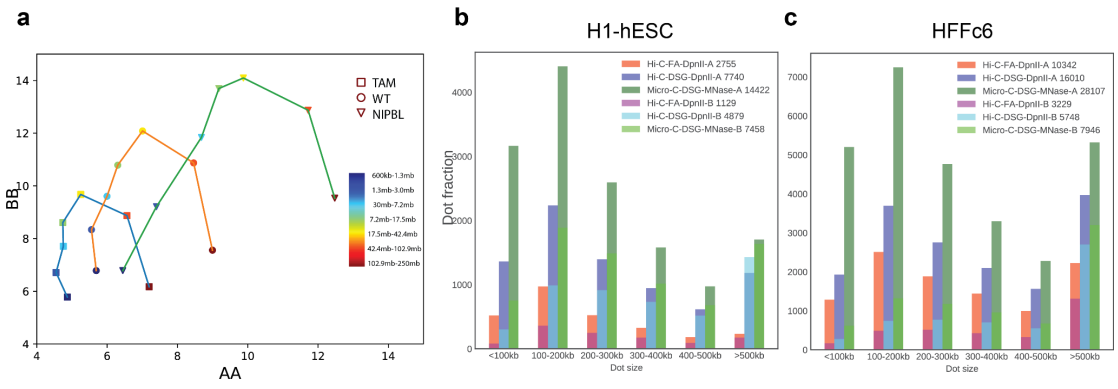

Figure 5.1: Interplay between compartments and loops

a. Compartment strength of A and B compartments is plotted for different genomic distances for wild type mice, Tamoxifen control (TAM) and NIPBL deleted mice.

b. # of loops detected in A and B compartments for H1-hESC and HFFc6.

Materials and Methods

cLIMS: A Laboratory Information Management System for C-Data

cLims is a web-based lab information management system tailored for chromosome conformation capture experiments. It can be used to organize, store and export metadata of various experiment types such as HiC, 5C, ATAC-Seq, etc. The metadata organization is compatible with 4DN DCIC standards and data to cLIMS can be used to export 4DN DCIC and GEO

systems with one click.

For the matrix project, we had increasing levels of detail in metadata, growing number of experiments, long time periods between data creation and submission and many people working on the same data sets, hence cLIMS helped us to keep this information properly maintained. The details included cell line, assay, treatments, sequencing, contributor's information. This will also help us in reproducibility of experiments.

cLIMS has been developed using the Django web framework on the back-end and HTML5 and Javascript libraries on the front-end. It is running on PostgreSQL database and Apache web server and can be hosted on major Linux distributions.

Cell line culture and fixation

- **HFFc6**

HFFc6 was cultured according to 4DN SOP (https://data.4dnucleome.org/biosources/4DNSRC6ZVYVP/). Cells were grown at 37°C under 5% CO2 in 75cm^2 flasks containing Dulbeco's Modified Eagle Medium (DMEM), supplemented with 20%, heat-inactivated Fetal Bovine Serum (FBS). For sub-culture, cells were rinsed with 1x DPBS and detached using 0.05% trypsin at 37 °C for 2-3 minutes. Cells were typically split every 2-3 days at a 1:4 ratio and harvested while sub-confluent, ensuring they would not overgrow.

- **H1-hESC**

Human Embryonic stem cells (H1 – WiCell, WA01, lot # WB35186) were cultured in mTeSR1 media (StemCell Technologies, 85850) under feeder-free conditions on Matrigel H1-hESC-qualified matrix (Corning, 354277, lot # 6011002) coated plates at 37°C and 5% CO_2. H1 cells were daily fed with fresh mTeSR1 media and passaged every 4-5 days using ReLeSR reagent (StemCell Technologies, 05872). Cells were dissociated into single cells with TrypLE Express (Thermo Fisher, 12604013).

- **Fixation protocol**

Final harvest of 5 million HFFc6 and H1-hESC cells was performed after washing twice with Hank's Buffered Salt Solution (HBSS) before cross-linking in HBSS with 1% Formaldehyde for 10 minutes at room temperature. Formaldehyde was quenched with glycine (128 mM final concentration) at room temperature for 5 minutes and on ice for an additional 15 minutes. Cells were washed twice with DPBS before pelleting and flash freezing with liquid nitrogen into 5 million aliquots. Alternatively, formaldehyde fixed cells were centrifuged at 800x*g* and subjected to additional cross-linking with either 3mM Disuccinimidyl glutarate (DSG) or Ethylene glycol bis (succinimidylsuccinate) (EGS), freshly prepared and diluted from a 300mM stock in DMSO, for 40 minutes at room temperature. DSG and EGS cross-linked cells were both quenched with 0.4M glycine for 5 minutes and washed twice with DPBS, supplemented with 0.5% Bovine Serum Albumin, before flash freezing with liquid nitrogen into 5 million aliquots.

Hi-C protocol

Chromosome conformation capture was performed as described previously and we refer to Belaghzal et al. (Belaghzal, Dekker, and Gibcus 2017) for a step-by-step version similar to this protocol. Briefly, $5x10^6$cross-linked cells were lysed for 15 minutes in ice cold lysis buffer (10 mM Tris-HCl pH8.0, 10 mM NaCl, 0.2% Igepal CA-630) in the presence of Halt protease inhibitors (Thermo Fisher, 78429). Then, the cells were disrupted by homogenization with pestle A for 2x 30 strokes. An aliquot if 8 µL was taken to later assess chromatin integrity. Remaining chromatin was solubilized in 0.1% SDS at 65°C for 10 minutes, quenched by 1% Triton X-100 (Sigma, 93443) and digested with 400 units of either HindIII (R0104 in NEBuffer 2.1), DdeI (R0175, in NEBuffer 3.1) or DpnII (R0543 in NEBuffer 3.1) for 16 hours at 37°C. Samples were incubated at 65℃ for 20 minutes to inactivate the restriction enzyme after which 10 µL was set aside to assess digestion efficiency. Fill-in of digested overhangs by DNA polymerase I, large Klenow fragment (NEB, M0210) in the presence of 250 nM biotin-14-dCTP (HindIII; Thermo Fisher, 19518018) or biotin-14-dATP (DdeI, DpnII; Thermo Fisher, 19524016) for 4 hours at 23°C was performed prior to ligation with 50 µL T4 DNA ligase (Thermo Fisher, 15224090) for 4 hours at 16°C in a total volume of 1.2 mL. Cross-links of ligated chromatin were reversed at 65°C overnight by 2 separate 50 µL additions of 10 mg/mL proteinase K (Fisher, BP1750I-400). DNA was isolated by adding 2.6 mL saturated phenol pH 8.0: chloroform (1:1) to 1.3 mL of sample. The mixture was vortexed and spun

down in phase-lock tubes (Quiagen, 129065) before standard precipitation with 100% ethanol in the presence of 1/10 vol/vol of 3 M sodium acetate pH 5.2. DNA cleanup and desalting was performed using an AMICON Ultra Centrifuge filter, following manufacturer's instructions (EMD Millipore, UFC5030BK). RNA was removed by incubation with 1 µL of 1 mg/mL RNAase A for 30 minutes at 37°C in a total of 100 µL TLE (10 mM Tris-HCl, 0.1 mM EDTA in milliQ) and quantified on a 0.7% agarose gel. Biotin was removed at 20°C for 4 hours in a 50 µL reaction for every 5 µg of DNA using 15 units of T4 DNA polymerase (NEB, M0203L) and 25 nM dATP and 25nM dGTP in NEBuffer 3.1 (no dTTP and dCTP). Polymerase was inactivated for 20 mins at 75°C and placed at 4°C. Volume was brought up to 130 µL and DNA was sheared for 3 minutes using a Covaris sonicator (E220 evolution: Duty Cycle 10%, Intensity 5, Cycles per Burst 200 or M220: Peak Incident Power 50W, Duty Cycle 20%, Cycles per Burst 200) and size selected with Agencourt AMPure® XP (Beckman Coulter, A63881) to obtain 150 - 350 basepair fragments, validated by DNA gel electrophoresis. DNA was repaired by adding a cocktail of 20 µL of the end-repair mix [3.5X NEB ligation buffer (NEB, B0202S), 17.5 mM dNTP mix, 7.5 units of T4 DNA polymerase (NEB, M0203L), 25 units of T4 polynucleotide kinase (NEB, M0201S), 2.5 units Klenow polymerase Polymerase I (NEB, M0210L)] to 50µL DNA solution at 20°C for 30 minutes, followed by a 20 minute incubation at 75°C to inactivate Klenow polymerase. For every library, at least 10 µL of streptavidin coated

Dynabeads™ MyOne™ Streptavidin C1 (Thermo Fisher, 65001) in LoBind tubes (Eppendorf, 022431021) were prepared by washing the beads twice with Tris Wash Buffer [5 mM Tris-HCl pH8.0, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween20] and resuspending in 400 µL of 2X Binding Buffer [10 mM Tris-HCl pH 8, 1 mM EDTA, 2 M NaCl]. The washed beads were added to the biotinylated DNA (brought to 400 µL TLE) and incubated for 15 minutes at room temperature under rotation. Thereafter, the beads were first washed with 1x Binding Buffer and then with TLE before final elution in 41µL TLE on a magnetic stand. Then, a 9 µL A-tailing mix consisting of 5 µL 10x NEBuffer 2.1, 1 µL of 10 mM dATP and 15 units of Klenow 3' → 5' exo- (NEB M0212L) was added to blunted ends and incubated at 37°C for 30 minutes, followed by inactivation for 20 minutes at 65°C. Beads were reclaimed, washed with 1x ligation buffer (from 5x T4 DNA ligase buffer, Thermo Fisher, 46300-018) and Illumina paired-end adapters were added by ligation with T4 DNA ligase (Thermo Fisher, 15224090) for 2 hours at room temperature. To determine the minimal number of PCR cycles needed to generate a Hi-C library, a PCR titration was performed prior to the production PCR (using Illumina primers PE1.0 and PE2.0). Primers were separated from the final library by size selection with AMpure XP (1:1 ratio) prior to 50 bp paired-end sequencing on

an Illumina HiSeq 4000 sequencer (Thermo Fisher).

For each deep library repeat, we generated 4 Hi-C libraries in parallel (20x 10^6 cells total) and sequenced each of the generated libraries on a single lane of an Illumina HiSeq 4000 flow cell.

Micro-C-XL protocol

The Micro-C XL protocol was adopted from Hsieh et al. and Krietenstein et al. (7, 8). Frozen cells were resuspended in 200 µl cold 1x PBS (10 mM Na_2HPO_4/KH_2PO_4, pH 7.4, 137 mM NaCl, 2.7 mM KCl,) per 1 mio cells and split into 1 mio cells aliquots. Note, 1x BSA (NEB,#B9000S) was added to PBS prior resuspension and wash to reduce stickiness of HFFc6 cells to the tub walls. After 20 min incubation on ice, cells were collected by centrifugation (5000x g, 5 min), washed with 500 µl buffer MB#1 (10 mM Tris-HCl, pH 7.5, 50 mM NaCl, 5 mM $MgCl_2$, 1 mM $CaCl_2$, 0.2% NP-40, 1x Roche cOmplete EDTA-free (Roche diagnostics, 04693132001)), collected by centrifugation (5000x g, 5 min), and resuspended in 200 µl MB#1. Chromatin was fragmented with MNase for 10 min at 37°C. MNase concentrations were chosen to yield mostly mono-nucleosomal fragments, as tested in prior digestion tests, typically 5-20 U MNase (Wortington Biochem, LS004798). The digestion was stopped by addition of 0.5 M EGTA (Bioworld, #405200081) to a 1.5 mM final concentration and incubation at 65°C for 10 min. Chromatin aliquots were pooled for further processing. Here, the equivalent of 2.5 mio

cells input yielded the best results, more than 5 mio cell-equivalent per aliquot is not recommended. The chromatin was collected by centrifugation (5000x g, 5 min), washed with 500 µl 1x NEBuffer 2.1 (NEB, #B7202S), collected by centrifugation (5000x g, 5 min), and resuspended in 45 µl NEBuffer 2.1. DNA ends were dephosphorylated by addition of 5 µl rSAP (NEB, #M0203) and incubation at 37°C for 45 min. The reaction was stopped by incubation at 65°C for 5 min. 5' overhangs were generated by 3' resection. Here, 40 µl pre-mix (5 µl 10x NEBuffer 2.1, 2 µl 100 mM ATP (Thermo Fisher, #R0441), 3 µl 100 mM DTT, 30 µl H2O) and 8 µl Large Klenow Fragment (NEB, #M0210L) and 2 µl T4 PNK (NEB,#M0201L) were added to the sample in respective order. The reaction was incubated at 37°C for 15 min. The DNA overhangs were filled with biotinylated nucleotides by addition of 100 µl pre-mix (25 µl 0.4 mM Biotin-dATP (Invitrogen, #19524016), 25 µl 0.4 mM Biotin-dCTP (Invitrogen, #19518018), 2 µl 10 mM dGTP and 10 mM dTTP (stock solutions: NEB, #N0446), 10 µl 10x T4 DNA Ligase Reaction Buffer (NEB #B0202S), 0.5 µl 200x BSA (NEB, #B9000S), 38.5 µl H2O) and incubation at 25°C for 45 min. The reaction was stopped by addition of 12 µl 0.5 M EDTA (Invitrogen, #15575038) and incubation at 65°C for 20 min. The chromatin was collected by centrifugation (10000x g), washed in 500 µl 1x Ligase Reaction Buffer, and collected by centrifugation (10000x g). The chromatin pellet was resuspended in 2500 µl ligation reaction buffer (1x NEB Ligase buffer, 1x NEB BSA, 12500 U NEB T4 Ligase (NEB, #M0202L)) and incubated rotating at RT for 2.5-3 h.

After proximity ligation, the chromatin was collected, resuspended in 200 µl 1x NEBuffer 1 (NEB, #B7001S) and 200 U NEB Exonuclease III (NEB, #M0206S), and incubated for 5 min at 37°C to remove biotin from unligated ends. For deproteination and reverse crosslinking, 25 µl ProteinaseK (25mg/ml in TE with 50% glycerol) and 25 µl 10% SDS (Invitrogen, #15553-035) were added and the sample was incubated at 65°C o/n. The DNA was first phenol/chloroform purified and second purified with DNA Clean & Concentrator Kit (Zymo, #D4013). The 300 bp sized MicroC library was purified via 1.5% agarose gel electrophoresis and extracted with Zymoclean Gel DNA Recovery Kit (ZymoResearch, #D4002) with a final elution volume of 50 µl. 5 µl Dynabeads^{TM}MyOne^{TM}Streptavidin C1 beads (Invitrogen, #65001) were washed twice with 300 µl 1x TBW (5 mM Tris-HCl, pH7.5, 0.5 mM EDTA, 1 M NaCl) and suspended in 150 µl 2x TBW (10 mM Tris-HCl, pH 7.5, 1 mM EDTA, 2 M NaCl). 100 µl H2O and150 µl washed Streptavidin beads in 2x TBW were added to the sample and incubated rotation at RT for 20 min. The beads were washed twice with 300 µl 1x TBW and resuspended in 50 µl TE buffer. Sequencing libraries were prepared with NEBNext®UltraII DNA Library Prep Kit for Illumina® (NEB, #E7645) according to protocol, except for the DNA purification and size selection prior PCR. Here, adaptor-ligated DNA was pulled-down via still attached Streptavidin beads and washed twice with 300 µl 1x TBW and once with 0.1x TE. Finally, the beads were resuspended in 20 µl 0.1x TE. PCR amplification, sample indexing, and DNA purification

after PCR was performed with according to (NEB, #E7645) using NEBNext®Multiplex Oligos for Illumina®. The samples were sequenced on an Illumina HiSeq 4000 on 50 base pair paired end mode.

Size range of chromatin fragments produced after digestion

Cells were cross-linked, lysed and digested as with the Hi-C protocol (see above). Then, cross-links were reversed and DNA was isolated as in Hi-C, but without ligation and biotin incorporation. DNA was loaded on an Advanced Analytical Fragment Analyzer (Agilent) for size range analysis and data was analyzed with PROsize3 software (Agilent). PROsize3 traces were exported separately as 4x8 bins (32 total) ranging from 40-500; 500-1300; 1300-8000 and 8000-100000 basepairs. Size ranges of potential restriction sites (hg38) were identified with cooltools genome digest (https://cooltools.readthedocs.io/en/latest/cli.html?highlight=enzyme#cooltools-genome-digest).

Cut&Tag protocol

Samples were processed as previously described in Kaya-Okur et.al. (22) , with few modifications. Briefly, approximately 100K cells per sample were permeabilized in the wash buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM Spermidine, 1× Protease inhibitor cocktail) and then cells were coupled with activated concanavalin A-coated magnetic beads for 10 min at RT. Pelleted beads were resuspended in antibody buffer (Mix 8 µL 0.5 M

EDTA and 6.7 µL 30% BSA with 2 mL Dig-wash buffer) with 1:100 dilution of SMC1 (Bethyl, cat# A300-055A) or CTCF antibody (Active motif, cat # 61311) and incubated overnight at 4 °C on a rotator. The next day, the pelleted bead complex was incubated with 1: 50 dilution of secondary antibody (guinea pig α-rabbit antibody, cat. # ABIN101961) in Dig-Wash buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM Spermidine, 1× Protease inhibitor cocktail, 0.05% Digitonin) and incubated at RT for 30 min on rotator. After two washes in Dig-Wash buffer, 1:250 diluted pAG-Tn5 adapter complex in Dig-300 buffer (20 mM HEPES pH 7.5, 300 mM NaCl, 0.5 mM Spermidine, 1× Protease inhibitor cocktail, 0.05% Digitonin) were added to bead complex and incubated at RT for 1 hr. After two washes in Dig-300 buffer, beads were resuspended in 300 µL of Tagmentation buffer (20 mM HEPES pH 7.5, 300 mM NaCl, 0.5 mM Spermidine, 1× Protease inhibitor cocktail, 0.05% Digitonin,10 mM MgCl2) and incubated at 37 °C for 1 h 45 min. Samples were subjected to Proteinase K treatment and extracted tagmented DNA using Phenol:Chloroform:Isoamyl Alcohol (25:24:1). In preparation for Illumina sequencing, 21 µL DNA was mixed with 2 µL of a universal i5, 2 µL of a uniquely barcoded i7 primer, and 25 µL of NEBNext HiFi 2× PCR Master mix. The sample was placed in a thermocycler with a heated lid using the following cycling conditions: 72 °C for 5 min; 98 °C for 30 s; 14 cycles of 98 °C for 10 s and 63 °C for 30 s; final extension at 72 °C for 1 min and hold at 4 °C. Post-PCR clean-up was performed by adding 1.1× volume of Ampure XP beads and incubated for 15 min at RT, washed twice gently in 80% ethanol, and eluted in 30 µL 10 mM Tris pH 8.0. Final library samples were paired-end

sequenced on Nextseq500.

Cut&Run Protocol

Cut&Run raw data (fastq files) of H1-hESC are downloaded from Janssens et al. 2018 and raw files of HFFc6 are generated by Steve Henikoff Lab using Skene et al. 2017 protocol (Skene and Henikoff 2017).

ATAC Seq Protocol

We have followed a published protocol to perform H1-hESC ATAC Seq experiments. The protocol details are described in Genga et al. 2019 (Genga et al. 2019).

ATAC-seq experiments on HFFc6 cells were performed following previously published protocol (Buenrostro et al. 2015). Briefly, 50,000 cells per experiment were washed and lysed using a lysis buffer (0.1% NP-40, 10 mM Tris-HCl (pH 7.4), 10 mM NaCl and 3 mM MgCl2). Lysed cells were then transposed using the Nextera DNA library prep kit (Illumina #FC-121-1030) for 30 min at 37C, immediately followed by DNA collection using Qiagen MinElute columns (Qiagen #28004). Appropriate cycle numbers for amplification were determined for each sample individually using qPCR. Finally, primers were removed using AMpure XP beads (Beckman Coulter #A63881) prior to 2x50bp paired-end sequencing.

Data analysis

- **Chromosome capture data processing**

Distiller (https://github.com/mirnylab/distiller-nf) pipeline is used to process Hi-C and Micro-C datasets. First, sequencing reads were mapped to hg38 using bwa mem with flags-SP. Second, mapped reads were parsed and classified using the pairtools package (https://github.com/mirnylab/pairtools) to get 4DN-compliant pairs files. PCR and/optical duplicates removed by matching the positions of aligned reads with 2bp flexibility. Next, pairs were filtered using mapping quality scores (MAPQ > 30) on each side of aligned chimeric read, binned into multiple resolutions and low coverage bins are removed. Finally multiresolution cooler files were created using the cooler package (33)(https://github.com/mirnylab/cooler.git). We normalized contact matrices using the iterative correction procedure from Imakaev et al. 2012 (Imakaev et al. 2012). Interaction heatmaps were created using the “cooler show“ command from the cooler package.

- **Hicrep correlations**

We used HiCRep to do distance corrected correlations of the various protocols and cell states. Correlation is calculated in two steps. First, interaction maps are stratified by genomic distances and the correlation coefficients are calculated for each distance separately. Second, the reproducibility is determined by a novel stratum-adjusted correlation coefficient statistic (SCC) by aggregating stratum-specific correlation coefficients using a weighted average. We correlated the individual

chromosomes between protocols and averaged the correlations across all chromosomes.

- **Cis and Trans Ratio**

Trans percent is calculated by dividing the total interactions between chromosomes with the sum of interactions within and between chromosomes (trans/cis+trans). Distance seperated cis interactions are calculated by dividing total interactions within specified distance of the chromosomes by the sum of interactions within and between chromosomes (cis of specific distance/cis+trans). Pairtools provides statistics for the numbers of interactions captured within and between chromosomes.

- **P(s) Plots**

P(s) plots describe the decay of the average probability of contact between two regions on a chromosome as a function of the genomic separation between them.

As per best practices, scalings are typically computed for each chromosomal arm of the genome before being aggregated. In order to obtain the extent of each chromosomal arm, the sizes of the chromosomes and the positions of their associated centromeres must be obtained. The sizes of the chromosome were obtained using the *fetch_chromsizes* function that is found in the bioframe library (https://github.com/open2c/bioframe/blob/master/bioframe/io/resources.py#L61) and the starts and ends of the centromere were obtained from bioframe

using *fetch_centromeres*(https://github.com/open2c/bioframe/blob/master/bioframe/io/resources.py#L109). The results of these two functions were combined to create a single list containing the extents of each chromosomal arm of the Human hg38 genome. For all libraries except those made from Hela S3 cells, all chromosome arms were used in the scaling calculation. For Hela libraries we excluded the chromosomes with translocations and used only chromosomes 4, 14, 17, 18, 20, and 21.

We used the *diagsum* function from the cooltools library (https://github.com/open2c/cooltools/blob/master/cooltools/expected.py#L541) to calculate scaling. This function takes in a cooler, extracts the table of non-zero read counts across the genome (known as the pixel table) and calculates the sum of read counts based on its distance from the main diagonal. It also simultaneously calculates the total number of possible counts obtainable at a given distance (called valid pairs) based on masking of region due to balancing and other use provided criteria. Additionally, this function also has the ability of transforming the read-counts obtained from the pixel table before aggregating the result. This is done by passing the appropriate use defined function to the “transforms” parameter of diagsum.

To obtain the scaling plots shown in the manuscript, for each library, the diagsum function was applied on the 1kb cooler associated with the library. 1kb is the recommended resolution to calculate scalings as it allows us to observe variations at the finest scales. Along with the cooler, the

chromosomal arms extents were also provided using the regions argument. A transform (named “balanced”) was also applied to the data to convert raw read-counts to balanced read-counts. This was done by multiplying the count value with the associated row and column weights obtained from balancing the cooler.

The resulting output is a single table with 4 relevant columns: 1) “region” which describes what chromosome arm a specific row was obtained from; 2) “diag” which refers to the genomic separation at which the data was aggregated; 3) “balanced.sum” which is the sum of read-counts for that given region and genomic separation after they were transformed by the “balanced” transform and 4) “n_valid” the number of possible valid pairs at a given distance (as described earlier). The individual column values were aggregated over the different arms and then further aggregated into logarithmically spaced bins of genomic separation. Finally, the “balanced.sum” column was divided by the “n_valid” column to create the “balanced.avg” column that is a measure of the average number of contacts across the genomic for a given genomic separation. The curves shown in the main text are the “balanced.avg” values plotted as a function of “diag” for the different libraries.

In addition to the interaction decay within a chromosome, interaction between different chromosomes can also be quantified. This is done using the “*blocksum_asymm*” function in cooltools (https://github.com/open2c/cooltools/blob/master/cooltools/expected.py#L820) which uses a very similar methodology. Two sets of regions are provided to

blocksum_asymm and “balanced.sum” and “n_valid” is calculated for every pair of regions (entire chromosomes in this case). Since the interactions are between two chromosomes there is no notion of genomic separation between two regions. The “balanced.avg” is calculated in the same manner as above and the mean of this value is visualized as horizontal dashed lines in the main text figures.

- **Average slope of scaling**

In order to magnify small variations between the different libraries, we calculated “derivative curves” from the scaling curves. Derivative curves represent the rate of change of scaling curves as observed on a log-log scale. These are computed by taking the log of scaling data (both x and y), calculating the finite difference measure of the slope and smoothing that value with a gaussian kernel. The smoothing function used is *gaussian_filter1d* from the scipy library (with a spread of 1). The smooth finite difference values can be plotted as a function of distance. Alternatively, the average value of this derivative is calculated and correlated with other features.

- **Genome Coverage Analysis**

For genome wide coverage analysis, the mapped read pairs were split into two individual files and the read coverage at respective bins (genome-wide at 100 kb bins) were computed with bedtools coverage (v2.29.2) function (Quinlan and Hall 2010). The read density was normalized to reads per million to compare between samples with different total read counts and subsequently by reads per 1 kb to compare between annotations with

different bin sizes. The compartment associations were extracted from HindIII compartment calls using the respective cell types.

- **Compartment Analysis**

We assessed compartments using eigenvector decomposition on observed-over-expected contact maps at 100kb resolution separated for each chromosomal arm using the cooltools package derived scripts. Eigenvector that has the strongest correlation with gene density is selected, then A and B compartments were assigned based on the gene density profiles such that A compartment has high gene density and B compartment has low gene density profile . Spearman correlation was used to correlate the eigenvectors of different experiments performed with various protocols and cell states. Saddle plots were generated as follows: the interaction matrix of an experiment was sorted based on the eigenvector values from lowest to highest (B to A). Sorted maps were then normalized for their expected interaction frequencies; the upper left corner of the interaction matrix represents the strongest B-B interactions, lower right represents strongest A-A interactions, upper right and lower left are B-A and A-B respectively. To quantify saddle plots we took the strongest 20% of BB and strongest 20% of AA interactions and normalized them by the sum of AB and BA (top(AA)/(AB+BA) and top(BB)/(AB+BA)). Saddle quantifications were used to create the scatter plots in figure 3c and heatmaps in supplemental figure 3 that compare A and B compartments for all cell types. Both scatter plots and heatmaps in figure 3 and Supplemental figure 3 were created using the Matplotlib package from Python.

- **Identification of chromatin loops**

The cooltools call-dots function (https://github.com/open2c/cooltools/blob/master/cooltools/cli/call_dots.py),a reimplementation of HICCUPS was used to detect the chromatin loops that are reflected as dots in the interaction matrix. We used the following parameters to call the loops: fdr=0.1, diag_width=10000000, tile_size = 5000000, --max-nans-tolerated 4. We called dots in deep data at both 5kb and 10 kb resolutions, using MAPQ> 30 pairs and before merging the results using the criteria mentioned in Rao et al. 2014 (Rao et al. 2014). Briefly, to merge 5kb and 10 kb loop calls, both the reproducible 5kb calls and unique 10 kb calls were kept. Unique 5kb calls were kept if the genomic separation of the region was <100kb or if the dots were particularly strong (i.e.more than 100 raw interactions per 5kb pixel). More detailed explanations for dot calling can be found in Rao et al. 2014 and Krietenstein et al. 2020.

- **Comparison of loops detected in different protocols**

Bedtools intersect was re-implemented to overlap 2D loops between protocols. Since loop calls are fundamentally 2 dimensional data, they needed to be processed for use with bedtools (which operate on 1d data) (Quinlan and Hall 2010).

Each loop call consists of 6 coordinates: chrom1, start1, end1, chrom2, start2, end2. Since chrom1 is always the same as chrom2 for loop calls, we ignored these two columns and reduced our space to4 coordinates. Furthermore, to account for errors in the positioning of loop during the loop calling, we

introduced the following margin of error around the called region (typically 10kb):

pos1 = (start1 + end1)/2; start1 = (pos1 - 5kb); end1 = (pos1 + 5kb)

pos2 = (start1 + end1)/2; start2 = (pos1 - 5kb); end2 = (pos2 + 5kb)

In order to overlap two lists, we performed 2 separate 1D overlaps with bedtools and then merged the results. To this end, every entry on each list is given a unique "loop ID." Using *bedtools overlap* on each dimension of the loop list, we obtained a pair of loop IDs (one from each list) that were used to track which pairs of dots overlapped along both dimensions. Thus only pairs of dots with overlaps in both dimensions are merged and outputted.

- **Upset Plots**

Upset plots were created for overlapping loops using the following R package: https://cran.r-project.org/web/packages/UpSetR/vignettes/basic.usage.html.

- **Quantification of chromatin loops**

We created the loop pileups using notebooks from the *hic-data-analysis-bootcamp* notebook (https://github.com/hms-dbmi/hic-data-analysis-bootcamp/blob/master/notebooks/06_analysis_cooltools-snipping-pileups.ipynb). The pileups were done at 5kb resolution and with a 50kb extension on each side of the loop. To quantify the loop strength, first, we created an interaction matrix of 50x50 kb, centered around the loop. Then, we

calculated the intensity of the loop by dividing the average of a 3x3 square in the middle of the interaction matrix by the average of its neighboring pixels; upper left, upper middle, upper right, lower left and right middle. See the image below:

This quantification of loop enrichment using its local background was also done to identify the loops. These quantifications are shown in figure 4b-4c, supplemental figure 5b-e.

- **Anchor Analysis**

We concatenated the genomic positions of the left and the right anchors for each loop to create a 1D anchor list for each deep dataset (FA-DpnII, FA+DSG-DpnII, FA+DSG-MNase) derived from both H1-hESC and HFFc6 cell lines.

We used *BEDtools merge* with "--c 1 -o count " parameters to remove redundant anchors (based on their genomic position) and to find the number of merged anchors in each genomic location. The number of merged anchors in a given genomic locus reflected loop valency at this anchor. Using *BEDtools multiinter*(https://bedtools.readthedocs.io/en/latest/content/overview.html) we identified the anchors that were shared in 1,2 or 3 protocols (Figure 5a-5b-5c and Supplemental Figure 6a-e).

- **Cut&Run, Cut&Tag and ChIP Seq Analysis**

Cut&Run data (HFFc6 H3K4me3, HFFc6 H3K27ac, H1-hESC CTCF, H1-

hESC H3K4me3, H1-hESC H3K27ac) was generated in the lab of Steve Henikoff and can be found on the 4DN Data Portal (https://data.4dnucleome.org/). Cut&Tag (HFFc6 CTCF, HFFc6 SMC1) data was generated in the lab of René Maehr at UMass Medical school. Finally, ChIP Seq data was downloaded from ENCODE. We processed raw fastq files for Cut&Run and Cut&Tag data and downloaded already processed bigwig and peak lists for ChIP Seq data. We mapped and processed the fastq files using *nf-core ATAC Seq* (35) pipelines. BWA was used for mapping the fastq files to the hg38 reference genome; MACS2 (with default parameters) was used to find the enriched peaks and *BEDtools intersect* was subsequently used to identify the loop anchors from these enriched peaks.

We found the intersected anchors between the three protocols (FA-DpnII, FA+DSG-DpnII, FA+DSG-MNase) and the FA+DSG-MNase specific anchors using *bedtools intersect*. We extracted the open chromatin (ATAC Seq peaks) regions located at these anchors and then aggregated the average signal enrichments of CTCF, SMC1, H3K4me3, H3K27ac, YY1 and RNA PolII. *Deeptools* was used to create the enrichment profiles in Figure 5e and Supplemental Figure 6f (36). We downloaded the lists of candidate Cis Regulatory Elements (cCREs) for H1-hESC and HFFc6 from ENCODE (24) and overlapped these cCREs with the intersected anchor list and the FA+DSG-MNase anchor list, again using *BEDtools intersect*. Finally separated them based on the cCRE categories.

To compare the anchor specific enrichments, we used the loop lists of FA-

DpnII, FA+DSG-DpnII and FA+DSG-MNase. We identified enriched convergent CTCF sites located at these loop anchors and compared the enrichments of CTCF, SMC1, H3K4me3, H3K27ac, YY1 and RNA PolII per anchor. To obtain convergent CTCF sites, we selected Anchor 1 (left anchor) to overlap with CTCF sites that had a "+" orientation and a CTCF peak and Anchor 2 (right anchor) to overlap with CTCF sites that had a "-" orientation. We plotted convergent CTCF sites located at Anchor 1 and Anchor 2 for FA-DpnII, FA+DSG-DpnII and FA+DSG-MNase in both HFFc6 and H1-hESC.

For HFFc6, we used Cut&Tag data generated with an antibody against the N-terminus of CTCF. For H1-hESC cells, we used Cut&Run data generated with an antibody against the C-terminus of CTCF. Since CTCF motifs are known to locate at the N-terminus of the CTCF protein, the orientation of the CTCF enrichments differed between the data sets from Cut&Tag and Cut&Run.

- **Insulation Score**

We calculated diamond insulation scores using cooltools (https://github.com/open2c/cooltools/blob/master/cooltools/cli/diamond_insulation.py) as implemented from Crane et al. We defined the insulation and boundary strengths of each 10 kb bin by detecting the local minima of 10 kb binned data with a 200kb window size. We used cooltools's *diamond-insulation* function with these parameters: " --ignore-diags 2, --window-pixels 20". We separated weak and strong boundaries using the mean insulation score of each protocol (i.e.: weak boundaries < mean < strong boundaries). Since diamond insulation pipelines cannot differentiate between compartment

boundaries and insulation boundaries we manually removed the compartment boundaries before any further analysis. Therefore the depth in local minima here is a result of strong insulation strength not a compartment switch. Next, we aggregated the insulation strength of the deep datasets at loop anchors, strong boundaries and loop anchors located at the strong boundaries using scripts from the *hic-data-analysis-bootcamp* notebook (https://github.com/hms-dbmi/hic-data-analysis-bootcamp/blob/master/notebooks/06_analysis_cooltools-snipping-pileups.ipynb). For both deep and matrix data we used only strong boundaries for further analysis since they reflected the true boundaries across protocols. Since the position of insulation boundaries was often offset by one or two bins between protocols, we extended the boundary bin by 10 kb on each side (30 kb total) in each protocol. We then used *bedtools multiinter*(https://bedtools.readthedocs.io/en/latest/content/overview.html) to count the boundaries that were found in one or more protocols within the cell type. We defined our stringent boundary list as the boundaries that were shared in at least 50% of the matrix protocols within each cell type and used these boundary lists for further comparisons. In heatmaps, we used the average insulation strength of these boundaries per protocol. To create the heatmaps we used the loop anchors that were shared between the 3 protocols that were deeply sequenced: FA+DpnII, FA+DSG-DpnII and FA+DSG-MNase in both H1-hESC and HFFc6.

- **Loop quantification for specific genomic separations**

To quantify the loop strengths for HFFc6 deep datasets described (FA-DpnII, FA+DSG-DpnII, FA+DSG-MNase), first, we seperated the loops based on their genomic separations into 100 kb bins, starting from 70kb (i.e. 70-170-kb, 170-270kb,...970-1070 kb), because 70kb was the smallest detectable loop size and then plotted the number of loops detected in each distance interval (Fig. 6d). Since the number of detected loops in these genomic separations was different for each library, we sampled 1,000 loops for each distance from FA+DSG-DdeI-DpnII to quantify loop enrichments of the 5 libraries. If the number of loops at a specified distance was smaller than 1000 we use the entire loop set at this distance.

Finally, we sampled 2,000 loops from each HFFc6 deep dataset, (FA-DpnII, FA+DSG-DpnII, FA+DSG-DdeI, FA+DSG-DdeI-DpnII, FA+DSG-MNase), combined them and then quantified the loop strength of the total 10,000 loops in these deep datasets and in matrix datasets described in Fig. 2.1 a. Loop enrichments were quantified as described in the “Quantification of chromatin loops” section.

- **Determining the # of reads as a function of fragment size**

10 million reads sampled from 3 experimental conditions; Hi-C 2.5 DdeI, Hi-C 2.5 DpnII and Hi-C 3.0. Then these reads were mapped to the fragments they belong to. In figure we plot the number of interactions mapped to each fragment vs the size of the fragment.

- **Loop quantification for specific genomic separations**

To quantify the loop strengths for HFFc6 deep datasets described in Fig 6d (FA-DpnII, FA+DSG-DpnII, FA+DSG-DdeI, FA+DSG-DdeI-DpnII, FA+DSG-MNase), first, we seperated the loops based on their genomic separations into 100 kb bins, starting from 70kb (i.e. 70-170-kb, 170-270kb,...970-1070 kb), because 70kb was the smallest detectable loop size and then plotted the number of loops detected in each distance interval. Since the number of detected loops in these genomic separations was different for each library, we sampled 1,000 loops for each distance from FA+DSG-DdeI-DpnII to quantify loop enrichments of the 5 libraries. If the number of loops at a specified distance was smaller than 1000 we use the entire loop set at this distance.

Finally, we sampled 2,000 loops from each HFFc6 deep dataset, (FA-DpnII, FA+DSG-DpnII, FA+DSG-DdeI, FA+DSG-DdeI-DpnII, FA+DSG-MNase), combined them and then quantified the loop strength of the total 10,000 loops in these deep datasets and in matrix datasets described in Fig. 2.1 a. Loop enrichments were quantified as described in the “Quantification of chromatin loops” section.

- **Sampling Experiment**

We combined two biological replicates for the deep datasets obtained with each of the protocols. We then sampled 10 experiments with different numbers of interactions (valid pairs): 200 Million reads, 400 M, ...1800M, 2B reads. For each sample we then called and quantified compartment strength, and loops exactly as described above.

- **Visualization of methods**

We used cooler package to create interaction heatmaps of all 3D methods. The color scale of each method was determined by the 10th and 90th percentile of the data distribution.

- **Processing TSA-Seq, DamID and Replication Timing data**

TSA-Seq, DamID and Replication timing bedGraph files were downloaded from 4DN Data Portal. These files were binned to 50kb bins for the comparison with 3D methods. Basically the signal in 50kb bins is summed during binning.

- **Quantification of preferential interactions**

Interaction heatmaps were sorted using TSA-Seq, DamID and Replication timing datasets. The strongest 20% signal is quantified and normalized to the non-preferential interactions. Non-preferential interactions are defined as the interactions that occur between strong TSA-Seq regions and weak TSA-Seq regions.

- **Processing SPRITE data**

SPRITE data for H1-hESC and HFFc6 downloaded from 4DN Data portal. Fragments were classified based on their barcode combinations such that fragments that have the same barcode combination locate at the same cluster. All possible combinations of pairs of fragments within the cluster are created. Cooler is used to convert pairs of fragments to interaction matrices. Then the methods specified above were used to quantify compartments and

TADs. Since the resolution of SPRITE not high enough to call loops; union loop lists that are created in Figure 4.5 a and 4.5 b are used to create loop pileups.

Bibliography

Abdennur, Nezar, and Leonid A. Mirny. 2020. 'Cooler: scalable storage for Hi-C data and other genomically labeled arrays', *Bioinformatics*, 36: 311-16.

Abramo, K., A. L. Valton, S. V. Venev, H. Ozadam, A. N. Fox, and J. Dekker. 2019. 'A chromosome folding intermediate at the condensin-to-cohesin transition during telophase', *Nat Cell Biol*, 21: 1393-402.

Akgol Oksuz, B., L. Yang, S. Abraham, S. V. Venev, N. Krietenstein, K. M. Parsi, H. Ozadam, M. E. Oomen, A. Nand, H. Mao, R. M. J. Genga, R. Maehr, O. J. Rando, L. A. Mirny, J. H. Gibcus, and J. Dekker. 2021. 'Systematic evaluation of chromosome conformation capture assays', *Nat Methods*, 18: 1046-55.

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

Ay, Ferhat, Timothy L. Bailey, and William Stafford Noble. 2014. 'Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts', *Genome research*, 24: 999-1011.

Bauer, B. W., I. F. Davidson, D. Canena, G. Wutz, W. Tang, G. Litos, S. Horn, P. Hinterdorfer, and J. M. Peters. 2021. 'Cohesin mediates DNA loop extrusion by a "swing and clamp" mechanism', *Cell*, 184: 5448-64 e22.

Beagrie, R. A., A. Scialdone, M. Schueler, D. C. Kraemer, M. Chotalia, S. Q. Xie, M. Barbieri, I. de Santiago, L. M. Lavitas, M. R. Branco, J. Fraser, J. Dostie, L. Game, N. Dillon, P. A. Edwards, M. Nicodemi, and A. Pombo. 2017. 'Complex multi-enhancer contacts captured by genome architecture mapping', *Nature*, 543: 519-24.

Belaghzal, H., J. Dekker, and J. H. Gibcus. 2017. 'Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation', *Methods*, 123: 56-65.

Belton, J. M., R. P. McCord, J. H. Gibcus, N. Naumova, Y. Zhan, and J. Dekker. 2012. 'Hi-C: a comprehensive technique to capture the conformation of genomes', *Methods*, 58: 268-76.

Bickmore, W. A., and B. van Steensel. 2013. 'Genome architecture: domain organization of interphase chromosomes', *Cell*, 152: 1270-84.

Boltsis, I., F. Grosveld, G. Giraud, and P. Kolovos. 2021. 'Chromatin Conformation in Development and Disease', *Front Cell Dev Biol*, 9: 723859.

Bonev, B., and G. Cavalli. 2016. 'Organization and function of the 3D genome', *Nat Rev Genet*, 17: 661-78.

Bonev, B., N. Mendelson Cohen, Q. Szabo, L. Fritsch, G. L. Papadopoulos, Y. Lubling, X. Xu, X. Lv, J. P. Hugnot, A. Tanay, and G. Cavalli. 2017. 'Multiscale 3D Genome Rewiring during Mouse Neural Development', *Cell*, 171: 557-72 e24.

Boninsegna, L., A. Yildirim, Y. Zhan, and F. Alber. 2022. 'Integrative approaches in genome structure analysis', *Structure*, 30: 24-36.
Braccioli, L., and E. de Wit. 2019. 'CTCF: a Swiss-army knife for genome organization and transcription regulation', *Essays Biochem*, 63: 157-65.
Buenrostro, J. D., P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf. 2013. 'Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position', *Nat Methods*, 10: 1213-8.
Buenrostro, J. D., B. Wu, H. Y. Chang, and W. J. Greenleaf. 2015. 'ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide', *Curr Protoc Mol Biol*, 109: 21 29 1-21 29 9.
Busslinger, G. A., R. R. Stocsits, P. van der Lelij, E. Axelsson, A. Tedeschi, N. Galjart, and J. M. Peters. 2017. 'Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl', *Nature*, 544: 503-07.
Carty, M., L. Zamparo, M. Sahin, A. González, R. Pelossof, O. Elemento, and C. S. Leslie. 2017. 'An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data', *Nat Commun*, 8: 15454.
Chen, Y., Y. Zhang, Y. Wang, L. Zhang, E. K. Brinkman, S. A. Adam, R. Goldman, B. van Steensel, J. Ma, and A. S. Belmont. 2018. 'Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler', *J Cell Biol*, 217: 4025-48.
Consortium, Encode Project, J. E. Moore, M. J. Purcaro, H. E. Pratt, C. B. Epstein, N. Shoresh, J. Adrian, T. Kawli, C. A. Davis, A. Dobin, R. Kaul, J. Halow, E. L. Van Nostrand, P. Freese, D. U. Gorkin, Y. Shen, Y. He, M. Mackiewicz, F. Pauli-Behn, B. A. Williams, A. Mortazavi, C. A. Keller, X. O. Zhang, S. I. Elhajjajy, J. Huey, D. E. Dickel, V. Snetkova, X. Wei, X. Wang, J. C. Rivera-Mulia, J. Rozowsky, J. Zhang, S. B. Chhetri, J. Zhang, A. Victorsen, K. P. White, A. Visel, G. W. Yeo, C. B. Burge, E. Lecuyer, D. M. Gilbert, J. Dekker, J. Rinn, E. M. Mendenhall, J. R. Ecker, M. Kellis, R. J. Klein, W. S. Noble, A. Kundaje, R. Guigo, P. J. Farnham, J. M. Cherry, R. M. Myers, B. Ren, B. R. Graveley, M. B. Gerstein, L. A. Pennacchio, M. P. Snyder, B. E. Bernstein, B. Wold, R. C. Hardison, T. R. Gingeras, J. A. Stamatoyannopoulos, and Z. Weng. 2020. 'Expanded encyclopaedias of DNA elements in the human and mouse genomes', *Nature*, 583: 699-710.
Consortium, Encode Project, M. P. Snyder, T. R. Gingeras, J. E. Moore, Z. Weng, M. B. Gerstein, B. Ren, R. C. Hardison, J. A. Stamatoyannopoulos, B. R. Graveley, E. A. Feingold, M. J. Pazin, M. Pagan, D. A. Gilchrist, B. C. Hitz, J. M. Cherry, B. E. Bernstein, E. M. Mendenhall, D. R. Zerbino, A. Frankish, P. Flicek, and R. M. Myers. 2020. 'Perspectives on ENCODE', *Nature*, 583: 693-98.
Crane, E., Q. Bian, R. P. McCord, B. R. Lajoie, B. S. Wheeler, E. J. Ralston, S. Uzawa, J. Dekker, and B. J. Meyer. 2015. 'Condensin-driven remodelling of X chromosome topology during dosage compensation', *Nature*, 523: 240-4.
Cremer, T., and M. Cremer. 2010. 'Chromosome territories', *Cold Spring Harb Perspect Biol*, 2: a003889.
Creyghton, M. P., A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young, and R. Jaenisch. 2010. 'Histone H3K27ac separates active from poised enhancers and predicts developmental state', *Proc Natl Acad Sci U S A*, 107: 21931-6.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li. 2021. 'Twelve years of SAMtools and BCFtools', *Gigascience*, 10.

Dekker, J. 2014. 'Two ways to fold the genome during the cell cycle: insights obtained with chromosome conformation capture', *Epigenetics Chromatin*, 7: 25.

Dekker, J., A. S. Belmont, M. Guttman, V. O. Leshyk, J. T. Lis, S. Lomvardas, L. A. Mirny, C. C. O'Shea, P. J. Park, B. Ren, J. C. R. Politz, J. Shendure, S. Zhong, and D. Nucleome Network. 2017. 'The 4D nucleome project', *Nature*, 549: 219-26.

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

Dekker, J., and L. Mirny. 2016. 'The 3D Genome as Moderator of Chromosomal Communication', *Cell*, 164: 1110-21.

Dekker, Job, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. 2002. 'Capturing Chromosome Conformation', *Science*, 295: 1306-11.

Deng, W., J. Lee, H. Wang, J. Miller, A. Reik, P. D. Gregory, A. Dean, and G. A. Blobel. 2012. 'Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor', *Cell*, 149: 1233-44.

Deng, W., J. W. Rupon, I. Krivega, L. Breda, I. Motta, K. S. Jahn, A. Reik, P. D. Gregory, S. Rivella, A. Dean, and G. A. Blobel. 2014. 'Reactivation of developmentally silenced globin genes by forced chromatin looping', *Cell*, 158: 849-60.

Denker, A., and W. de Laat. 2016. 'The second decade of 3C technologies: detailed insights into nuclear organization', *Genes Dev*, 30: 1357-82.

Dixon, J. R., I. Jung, S. Selvaraj, Y. Shen, J. E. Antosiewicz-Bourget, A. Y. Lee, Z. Ye, A. Kim, N. Rajagopal, W. Xie, Y. Diao, J. Liang, H. Zhao, V. V. Lobanenkov, J. R. Ecker, J. A. Thomson, and B. Ren. 2015. 'Chromatin architecture reorganization during stem cell differentiation', *Nature*, 518: 331-6.

Dixon, J. R., S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. 2012. 'Topological domains in mammalian genomes identified by analysis of chromatin interactions', *Nature*, 485: 376-80.

Dixon, J. R., J. Xu, V. Dileep, Y. Zhan, F. Song, V. T. Le, G. G. Yardimci, A. Chakraborty, D. V. Bann, Y. Wang, R. Clark, L. Zhang, H. Yang, T. Liu, S. Iyyanki, L. An, C. Pool, T. Sasaki, J. C. Rivera-Mulia, H. Ozadam, B. R. Lajoie, R. Kaul, M. Buckley, K. Lee, M. Diegel, D. Pezic, C. Ernst, S. Hadjur, D. T. Odom, J. A. Stamatoyannopoulos, J. R. Broach, R. C. Hardison, F. Ay, W. S. Noble, J. Dekker, D. M. Gilbert, and F. Yue. 2018. 'Integrative detection and analysis of structural variation in cancer genomes', *Nat Genet*, 50: 1388-98.

Dostie, J., T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green, and J. Dekker. 2006. 'Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements', *Genome Res*, 16: 1299-309.

Engreitz, J. M., N. Ollikainen, and M. Guttman. 2016. 'Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression', *Nat Rev Mol Cell Biol*, 17: 756-70.

Erdel, F., A. Rademacher, R. Vlijm, J. Tunnermann, L. Frank, R. Weinmann, E. Schweigert, K. Yserentant, J. Hummert, C. Bauer, S. Schumacher, A. Al

Alwash, C. Normand, D. P. Herten, J. Engelhardt, and K. Rippe. 2020. 'Mouse Heterochromatin Adopts Digital Compaction States without Showing Hallmarks of HP1-Driven Liquid-Liquid Phase Separation', *Mol Cell*, 78: 236-49 e7.

Erdel, F., and K. Rippe. 2018. 'Formation of Chromatin Subcompartments by Phase Separation', *Biophys J*, 114: 2262-70.

Ernst, J., and M. Kellis. 2012. 'ChromHMM: automating chromatin-state discovery and characterization', *Nat Methods*, 9: 215-6.

Falk, M., Y. Feodorova, N. Naumova, M. Imakaev, B. R. Lajoie, H. Leonhardt, B. Joffe, J. Dekker, G. Fudenberg, I. Solovei, and L. A. Mirny. 2019. 'Heterochromatin drives compartmentalization of inverted and conventional nuclei', *Nature*, 570: 395-99.

Fang, R., M. Yu, G. Li, S. Chee, T. Liu, A. D. Schmitt, and B. Ren. 2016. 'Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq', *Cell Res*, 26: 1345-48.

Flavahan, W. A., Y. Drier, B. B. Liau, S. M. Gillespie, A. S. Venteicher, A. O. Stemmer-Rachamimov, M. L. Suva, and B. E. Bernstein. 2016. 'Insulator dysfunction and oncogene activation in IDH mutant gliomas', *Nature*, 529: 110-4.

Fudenberg, G., N. Abdennur, M. Imakaev, A. Goloborodko, and L. A. Mirny. 2017. 'Emerging Evidence of Chromosome Folding by Loop Extrusion', *Cold Spring Harb Symp Quant Biol*, 82: 45-55.

Fudenberg, G., M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur, and L. A. Mirny. 2016. 'Formation of Chromosomal Domains by Loop Extrusion', *Cell Rep*, 15: 2038-49.

Fullwood, M. J., Y. Han, C. L. Wei, X. Ruan, and Y. Ruan. 2010. 'Chromatin interaction analysis using paired-end tag sequencing', *Curr Protoc Mol Biol*, Chapter 21: Unit 21 15 1-25.

Fullwood, M. J., M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Chew, P. Y. Huang, W. J. Welboren, Y. Han, H. S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K. V. Desai, J. S. Thomsen, Y. K. Lee, R. K. Karuturi, T. Herve, G. Bourque, H. G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W. K. Sung, E. T. Liu, C. L. Wei, E. Cheung, and Y. Ruan. 2009. 'An oestrogen-receptor-alpha-bound human chromatin interactome', *Nature*, 462: 58-64.

Ganji, M., I. A. Shaltiel, S. Bisht, E. Kim, A. Kalichava, C. H. Haering, and C. Dekker. 2018. 'Real-time imaging of DNA loop extrusion by condensin', *Science*, 360: 102-05.

Genga, R. M. J., E. M. Kernfeld, K. M. Parsi, T. J. Parsons, M. J. Ziller, and R. Maehr. 2019. 'Single-Cell RNA-Sequencing-Based CRISPRi Screening Resolves Molecular Drivers of Early Human Endoderm Development', *Cell Rep*, 27: 708-18.e10.

Gibcus, J. H., and J. Dekker. 2013a. 'The hierarchy of the 3D genome', *Mol Cell*, 49: 773-82.

Gibcus, J. H., K. Samejima, A. Goloborodko, I. Samejima, N. Naumova, J. Nuebler, M. T. Kanemaki, L. Xie, J. R. Paulson, W. C. Earnshaw, L. A. Mirny, and J. Dekker. 2018. 'A pathway for mitotic chromosome formation', *Science*, 359.

Gibcus, Johan H., and Job Dekker. 2013b. 'The hierarchy of the 3D genome', *Molecular Cell*, 49: 773-82.

Goel, V. Y., and A. S. Hansen. 2021. 'The macro and micro of chromosome conformation capture', *Wiley Interdiscip Rev Dev Biol*, 10: e395.

Golloshi, R., J. T. Sanders, and R. P. McCord. 2018. 'Iteratively improving Hi-C experiments one step at a time', *Methods*, 142: 47-58.
Göndör, A., C. Rougier, and R. Ohlsson. 2008. 'High-resolution circular chromosome conformation capture assay', *Nat Protoc*, 3: 303-13.
Gong, H., Y. Yang, S. Zhang, M. Li, and X. Zhang. 2021. 'Application of Hi-C and other omics data analysis in human cancer and cell differentiation research', *Comput Struct Biotechnol J*, 19: 2070-83.
Hansen, A. S., C. Cattoglio, X. Darzacq, and R. Tjian. 2018. 'Recent evidence that TADs and chromatin loops are dynamic structures', *Nucleus*, 9: 20-32.
Heintzman, N. D., and B. Ren. 2009. 'Finding distal regulatory elements in the human genome', *Curr Opin Genet Dev*, 19: 541-9.
Hildebrand, E. M., and J. Dekker. 2020. 'Mechanisms and Functions of Chromosome Compartmentalization', *Trends Biochem Sci*, 45: 385-96.
Hnisz, D., A. S. Weintraub, D. S. Day, A. L. Valton, R. O. Bak, C. H. Li, J. Goldmann, B. R. Lajoie, Z. P. Fan, A. A. Sigova, J. Reddy, D. Borges-Rivera, T. I. Lee, R. Jaenisch, M. H. Porteus, J. Dekker, and R. A. Young. 2016. 'Activation of proto-oncogenes by disruption of chromosome neighborhoods', *Science*, 351: 1454-58.
Hou, C., L. Li, Z. S. Qin, and V. G. Corces. 2012. 'Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains', *Mol Cell*, 48: 471-84.
Hsieh, T. H., A. Weiner, B. Lajoie, J. Dekker, N. Friedman, and O. J. Rando. 2015. 'Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C', *Cell*, 162: 108-19.
Hsieh, T. S., G. Fudenberg, A. Goloborodko, and O. J. Rando. 2016. 'Micro-C XL: assaying chromosome conformation from the nucleosome to the entire genome', *Nat Methods*, 13: 1009-11.
Hyle, J., Y. Zhang, S. Wright, B. Xu, Y. Shao, J. Easton, L. Tian, R. Feng, P. Xu, and C. Li. 2019. 'Acute depletion of CTCF directly affects MYC regulation through loss of enhancer-promoter looping', *Nucleic Acids Res*, 47: 6699-713.
Imakaev, M., G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny. 2012. 'Iterative correction of Hi-C data reveals hallmarks of chromosome organization', *Nat Methods*, 9: 999-1003.
Janssens, D. H., S. J. Wu, J. F. Sarthy, M. P. Meers, C. H. Myers, J. M. Olson, K. Ahmad, and S. Henikoff. 2018. 'Automated in situ chromatin profiling efficiently resolves cell types and gene regulatory programs', *Epigenetics Chromatin*, 11: 74.
Kagey, Michael H., Jamie J. Newman, Steve Bilodeau, Ye Zhan, David A. Orlando, Nynke L. van Berkum, Christopher C. Ebmeier, Jesse Goossens, Peter B. Rahl, Stuart S. Levine, Dylan J. Taatjes, Job Dekker, and Richard A. Young. 2010. 'Mediator and cohesin connect gene expression and chromatin architecture', *Nature*, 467: 430-35.
Kaya-Okur, Hatice S., Steven J. Wu, Christine A. Codomo, Erica S. Pledger, Terri D. Bryson, Jorja G. Henikoff, Kami Ahmad, and Steven Henikoff. 2019. 'CUT&Tag for efficient epigenomic profiling of small samples and single cells', *Nature Communications*, 10: 1930.
Kellis, M., B. Wold, M. P. Snyder, B. E. Bernstein, A. Kundaje, G. K. Marinov, L. D. Ward, E. Birney, G. E. Crawford, J. Dekker, I. Dunham, L. L. Elnitski, P. J. Farnham, E. A. Feingold, M. Gerstein, M. C. Giddings, D. M. Gilbert, T. R. Gingeras, E. D. Green, R. Guigo, T. Hubbard, J. Kent, J. D. Lieb, R. M. Myers, M. J. Pazin, B. Ren, J. A. Stamatoyannopoulos, Z. Weng, K. P. White,

and R. C. Hardison. 2014. 'Defining functional DNA elements in the human genome', *Proc Natl Acad Sci U S A*, 111: 6131-8.

Khoury, A., J. Achinger-Kawecka, S. A. Bert, G. C. Smith, H. J. French, P. L. Luu, T. J. Peters, Q. Du, A. J. Parry, F. Valdes-Mora, P. C. Taberlay, C. Stirzaker, A. L. Statham, and S. J. Clark. 2020. 'Constitutively bound CTCF sites maintain 3D chromatin architecture and long-range epigenetically regulated domains', *Nat Commun*, 11: 54.

Krietenstein, N., S. Abraham, S. V. Venev, N. Abdennur, J. Gibcus, T. S. Hsieh, K. M. Parsi, L. Yang, R. Maehr, L. A. Mirny, J. Dekker, and O. J. Rando. 2020. 'Ultrastructural Details of Mammalian Chromosome Architecture', *Mol Cell*, 78: 554-65 e7.

Kruse, Kai, Clemens B. Hug, and Juan M. Vaquerizas. 2020. 'FAN-C: a feature-rich framework for the analysis and visualisation of chromosome conformation capture data', *Genome Biology*, 21: 303.

Kubo, N., H. Ishii, X. Xiong, S. Bianco, F. Meitinger, R. Hu, J. D. Hocker, M. Conte, D. Gorkin, M. Yu, B. Li, J. R. Dixon, M. Hu, M. Nicodemi, H. Zhao, and B. Ren. 2021. 'Promoter-proximal CTCF binding promotes distal enhancer-dependent gene activation', *Nat Struct Mol Biol*, 28: 152-61.

Lafontaine, D. L., L. Yang, J. Dekker, and J. H. Gibcus. 2021. 'Hi-C 3.0: Improved Protocol for Genome-Wide Chromosome Conformation Capture', *Curr Protoc*, 1: e198.

Lajoie, B. R., J. Dekker, and N. Kaplan. 2015. 'The Hitchhiker's guide to Hi-C analysis: practical guidelines', *Methods*, 72: 65-75.

Lander, Eric S., Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie LeVine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P. Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd, Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C. Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Robert H. Waterston, Richard K. Wilson, LaDeana W. Hillier, John D. McPherson, Marco A. Marra, Elaine R. Mardis, Lucinda A. Fulton, Asif T. Chinwalla, Kymberlie H. Pepin, Warren R. Gish, Stephanie L. Chissoe, Michael C. Wendl, Kim D. Delehaunty, Tracie L. Miner, Andrew Delehaunty, Jason B. Kramer, Lisa L. Cook, Robert S. Fulton, Douglas L. Johnson, Patrick J. Minx, Sandra W. Clifton, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan-Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward Uberbacher, Marvin Frazier, Richard A. Gibbs, Donna M. Muzny, Steven E. Scherer, John B. Bouck, Erica J. Sodergren, Kim C. Worley, Catherine M. Rives, James H. Gorrell, Michael L. Metzker, Susan L. Naylor, Raju S. Kucherlapati, David L. Nelson, George M. Weinstock, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, Jean Weissenbach, Roland Heilig, William Saurin, Francois Artiguenave, Philippe Brottier, Thomas Bruls,

Eric Pelletier, Catherine Robert, Patrick Wincker, André Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, Douglas R. Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Hong Mei Lee, JoAnn Dubois, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Ronald W. Davis, Nancy A. Federspiel, A. Pia Abola, Michael J. Proctor, Bruce A. Roe, Feng Chen, Huaqin Pan, Juliane Ramser, Hans Lehrach, Richard Reinhardt, W. Richard McCombie, Melissa de la Bastide, Neilay Dedhia, Helmut Blöcker, Klaus Hornischer, Gabriele Nordsiek, Richa Agarwala, L. Aravind, Jeffrey A. Bailey, Alex Bateman, Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G. Brown, Christopher B. Burge, Lorenzo Cerutti, Hsiu-Chuan Chen, Deanna Church, Michele Clamp, Richard R. Copley, Tobias Doerks, Sean R. Eddy, Evan E. Eichler, Terrence S. Furey, James Galagan, James G. R. Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L. Steven Johnson, Thomas A. Jones, Simon Kasif, Arek Kaspryzk, Scot Kennedy, W. James Kent, Paul Kitts, Eugene V. Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M. Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V. Moran, Nicola Mulder, Victor J. Pollara, Chris P. Ponting, Greg Schuler, Jörg Schultz, Guy Slater, Arian F. A. Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I. Wolf, Kenneth H. Wolfe, Shiaw-Pyng Yang, Ru-Fang Yeh, Francis Collins, Mark S. Guyer, Jane Peterson, Adam Felsenfeld, Kris A. Wetterstrand, Richard M. Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R. Cox, Maynard V. Olson, Rajinder Kaul, Christopher Raymond, Nobuyoshi Shimizu, Kazuhiko Kawasaki, Shinsei Minoshima, Glen A. Evans, Maria Athanasiou, Roger Schultz, Aristides Patrinos, Michael J. Morgan, Consortium International Human Genome Sequencing, Center for Genome Research Whitehead Institute for Biomedical Research, Centre The Sanger, Center Washington University Genome Sequencing, Us Doe Joint Genome Institute, Center Baylor College of Medicine Human Genome Sequencing, Riken Genomic Sciences Center, Genoscope, U. M. R. Cnrs, Institute of Molecular Biotechnology Department of Genome Analysis, G. T. C. Sequencing Center, Center Beijing Genomics Institute/Human Genome, The Institute for Systems Biology Multimegabase Sequencing Center, Center Stanford Genome Technology, Technology University of Oklahoma's Advanced Center for Genome, Genetics Max Planck Institute for Molecular, Lita Annenberg Hazen Genome Center Cold Spring Harbor Laboratory, G. BF—German Research Centre for Biotechnology, Group *Genome Analysis, U. S. National Institutes of Health Scientific management: National Human Genome Research Institute, Center Stanford Human Genome, Center University of Washington Genome, Keio University School of Medicine Department of Molecular Biology, Dallas University of Texas Southwestern Medical Center at, U. S. Department of Energy Office of Science, and Trust The Wellcome. 2001. 'Initial sequencing and analysis of the human genome', *Nature*, 409: 860-921.

Langmead, Ben, and Steven L. Salzberg. 2012. 'Fast gapped-read alignment with Bowtie 2', *Nature Methods*, 9: 357-59.

Larson, A. G., D. Elnatan, M. M. Keenen, M. J. Trnka, J. B. Johnston, A. L. Burlingame, D. A. Agard, S. Redding, and G. J. Narlikar. 2017. 'Liquid droplet formation by HP1alpha suggests a role for phase separation in heterochromatin', *Nature*, 547: 236-40.

Li, Guoliang, Melissa J. Fullwood, Han Xu, Fabianus Hendriyan Mulawadi, Stoyan Velkov, Vinsensius Vega, Pramila Nuwantha Ariyaratne, Yusoff Bin Mohamed, Hong-Sain Ooi, Chandana Tennakoon, Chia-Lin Wei, Yijun Ruan, and Wing-Kin Sung. 2010. 'ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing', *Genome Biology*, 11: R22.

Li, H., and R. Durbin. 2009. 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*, 25: 1754-60.

Li, W., K. Gong, Q. Li, F. Alber, and X. J. Zhou. 2015. 'Hi-Corrector: a fast, scalable and memory-efficient package for normalizing large-scale Hi-C data', *Bioinformatics*, 31: 960-2.

Li, Y., J. H. I. Haarhuis, A. Sedeno Cacciatore, R. Oldenkamp, M. S. van Ruiten, L. Willems, H. Teunissen, K. W. Muir, E. de Wit, B. D. Rowland, and D. Panne. 2020. 'The structural basis for cohesin-CTCF-anchored loops', *Nature*, 578: 472-76.

Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. 2009. 'Comprehensive mapping of long-range interactions reveals folding principles of the human genome', *Science*, 326: 289-93.

Liu, K., H. Li, Y. Li, J. Wang, and J. Wang. 2022. 'A comparison of topologically associating domain callers based on Hi-C data', *IEEE/ACM Trans Comput Biol Bioinform*, Pp.

Luan, J., G. Xiang, P. A. Gómez-García, J. M. Tome, Z. Zhang, M. W. Vermunt, H. Zhang, A. Huang, C. A. Keller, B. M. Giardine, Y. Zhang, Y. Lan, J. T. Lis, M. Lakadamyali, R. C. Hardison, and G. A. Blobel. 2021. 'Distinct properties and functions of CTCF revealed by a rapidly inducible degron system', *Cell Rep*, 34: 108783.

Lupianez, D. G., K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel, and S. Mundlos. 2015. 'Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions', *Cell*, 161: 1012-25.

Marchal, C., T. Sasaki, D. Vera, K. Wilson, J. Sima, J. C. Rivera-Mulia, C. Trevilla-Garcia, C. Nogues, E. Nafie, and D. M. Gilbert. 2018. 'Genome-wide analysis of replication timing by next-generation sequencing with E/L Repli-seq', *Nat Protoc*, 13: 819-39.

Marchal, C., J. Sima, and D. M. Gilbert. 2019. 'Control of DNA replication timing in the 3D genome', *Nat Rev Mol Cell Biol*, 20: 721-37.

Maurano, M. T., R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, and J. A. Stamatoyannopoulos. 2012. 'Systematic localization of common disease-associated variation in regulatory DNA', *Science*, 337: 1190-5.

McSwiggen, D. T., A. S. Hansen, S. S. Teves, H. Marie-Nelly, Y. Hao, A. B. Heckert, K. K. Umemoto, C. Dugast-Darzacq, R. Tjian, and X. Darzacq. 2019. 'Evidence for DNA-mediated nuclear compartmentalization distinct from phase separation', *Elife*, 8.

McSwiggen, D. T., M. Mir, X. Darzacq, and R. Tjian. 2019. 'Evaluating phase separation in live cells: diagnosis, caveats, and functional consequences', *Genes Dev*, 33: 1619-34.
Melo, U. S., R. Schopflin, R. Acuna-Hidalgo, M. A. Mensah, B. Fischer-Zirnsak, M. Holtgrewe, M. K. Klever, S. Turkmen, V. Heinrich, I. D. Pluym, E. Matoso, S. Bernardo de Sousa, P. Louro, W. Hulsemann, M. Cohen, A. Dufke, A. Latos-Bielenska, M. Vingron, V. Kalscheuer, F. Quintero-Rivera, M. Spielmann, and S. Mundlos. 2020. 'Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases', *Am J Hum Genet*, 106: 872-84.
Mirny, L. A., M. Imakaev, and N. Abdennur. 2019. 'Two major mechanisms of chromosome organization', *Curr Opin Cell Biol*, 58: 142-52.
Nagano, Takashi, Csilla Várnai, Stefan Schoenfelder, Biola-Maria Javierre, Steven W. Wingett, and Peter Fraser. 2015. 'Comparison of Hi-C results using in-solution versus in-nucleus ligation', *Genome Biology*, 16: 175.
Nasmyth, K., and C. H. Haering. 2009. 'Cohesin: its roles and mechanisms', *Annu Rev Genet*, 43: 525-58.
Naumova, N., M. Imakaev, G. Fudenberg, Y. Zhan, B. R. Lajoie, L. A. Mirny, and J. Dekker. 2013. 'Organization of the mitotic chromosome', *Science*, 342: 948-53.
Naumova, N., E. M. Smith, Y. Zhan, and J. Dekker. 2012. 'Analysis of long-range chromatin interactions using Chromosome Conformation Capture', *Methods*, 58: 192-203.
Nora, E. P., A. Goloborodko, A. L. Valton, J. H. Gibcus, A. Uebersohn, N. Abdennur, J. Dekker, L. A. Mirny, and B. G. Bruneau. 2017. 'Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization', *Cell*, 169: 930-44 e22.
Nora, E. P., B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Bluthgen, J. Dekker, and E. Heard. 2012. 'Spatial partitioning of the regulatory landscape of the X-inactivation centre', *Nature*, 485: 381-5.
Norton, Heidi K., Daniel J. Emerson, Harvey Huang, Jesi Kim, Katelyn R. Titus, Shi Gu, Danielle S. Bassett, and Jennifer E. Phillips-Cremins. 2018. 'Detecting hierarchical genome folding with network modularity', *Nature Methods*, 15: 119-22.
Pal, K., M. Forcato, and F. Ferrari. 2019. 'Hi-C analysis: from data generation to integration', *Biophys Rev*, 11: 67-78.
Pekowska, A., B. Klaus, W. Xiang, J. Severino, N. Daigle, F. A. Klein, M. Oles, R. Casellas, J. Ellenberg, L. M. Steinmetz, P. Bertone, and W. Huber. 2018. 'Gain of CTCF-Anchored Chromatin Loops Marks the Exit from Naive Pluripotency', *Cell Syst*, 7: 482-95 e10.
Pirozzi, C. J., and H. Yan. 2021. 'The implications of IDH mutations for cancer development and therapy', *Nat Rev Clin Oncol*, 18: 645-61.
Pope, B. D., T. Ryba, V. Dileep, F. Yue, W. Wu, O. Denas, D. L. Vera, Y. Wang, R. S. Hansen, T. K. Canfield, R. E. Thurman, Y. Cheng, G. Gulsoy, J. H. Dennis, M. P. Snyder, J. A. Stamatoyannopoulos, J. Taylor, R. C. Hardison, T. Kahveci, B. Ren, and D. M. Gilbert. 2014. 'Topologically associating domains are stable units of replication-timing regulation', *Nature*, 515: 402-5.
Quinlan, A. R., and I. M. Hall. 2010. 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26: 841-2.
Quinodoz, S. A., N. Ollikainen, B. Tabak, A. Palla, J. M. Schmidt, E. Detmar, M. M. Lai, A. A. Shishkin, P. Bhat, Y. Takei, V. Trinh, E. Aznauryan, P. Russell, C. Cheng, M. Jovanovic, A. Chow, L. Cai, P. McDonel, M. Garber, and M.

Guttman. 2018. 'Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus', *Cell*, 174: 744-57.e24.
Rada-Iglesias, A., R. Bajpai, T. Swigut, S. A. Brugmann, R. A. Flynn, and J. Wysocka. 2011. 'A unique chromatin signature uncovers early developmental enhancers in humans', *Nature*, 470: 279-83.
Ramani, Vijay, Darren A. Cusanovich, Ronald J. Hause, Wenxiu Ma, Ruolan Qiu, Xinxian Deng, C. Anthony Blau, Christine M. Disteche, William S. Noble, Jay Shendure, and Zhijun Duan. 2016. 'Mapping 3D genome architecture through in situ DNase Hi-C', *Nature Protocols*, 11: 2104-21.
Ramirez, F., V. Bhardwaj, L. Arrigoni, K. C. Lam, B. A. Gruning, J. Villaveces, B. Habermann, A. Akhtar, and T. Manke. 2018. 'High-resolution TADs reveal DNA sequences underlying genome organization in flies', *Nat Commun*, 9: 189.
Rao, S. S., M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden. 2014. 'A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping', *Cell*, 159: 1665-80.
Rao, S. S. P., S. C. Huang, B. Glenn St Hilaire, J. M. Engreitz, E. M. Perez, K. R. Kieffer-Kwon, A. L. Sanborn, S. E. Johnstone, G. D. Bascom, I. D. Bochkov, X. Huang, M. S. Shamim, J. Shin, D. Turner, Z. Ye, A. D. Omer, J. T. Robinson, T. Schlick, B. E. Bernstein, R. Casellas, E. S. Lander, and E. L. Aiden. 2017. 'Cohesin Loss Eliminates All Loop Domains', *Cell*, 171: 305-20 e24.
Roayaei Ardakany, Abbas, Halil Tuvan Gezer, Stefano Lonardi, and Ferhat Ay. 2020. 'Mustache: multi-scale detection of chromatin loops from Hi-C and Micro-C maps using scale-space representation', *Genome Biology*, 21: 256.
Robinson, J. T., D. Turner, N. C. Durand, H. Thorvaldsdottir, J. P. Mesirov, and E. L. Aiden. 2018. 'Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data', *Cell Syst*, 6: 256-58 e1.
Roix, J. J., P. G. McQueen, P. J. Munson, L. A. Parada, and T. Misteli. 2003. 'Spatial proximity of translocation-prone gene loci in human lymphomas', *Nat Genet*, 34: 287-91.
Rowley, M. J., A. Poulet, M. H. Nichols, B. J. Bixler, A. L. Sanborn, E. A. Brouhard, K. Hermetz, H. Linsenbaum, G. Csankovszki, E. Lieberman Aiden, and V. G. Corces. 2020. 'Analysis of Hi-C data using SIP effectively identifies loops in organisms from C. elegans to mammals', *Genome Res*, 30: 447-58.
Ruiz-Velasco, M., M. Kumar, M. C. Lai, P. Bhat, A. B. Solis-Pinson, A. Reyes, S. Kleinsorg, K. M. Noh, T. J. Gibson, and J. B. Zaugg. 2017. 'CTCF-Mediated Chromatin Loops between Promoter and Gene Body Regulate Alternative Splicing across Individuals', *Cell Syst*, 5: 628-37 e6.
Ryba, T., I. Hiratani, J. Lu, M. Itoh, M. Kulik, J. Zhang, T. C. Schulz, A. J. Robins, S. Dalton, and D. M. Gilbert. 2010. 'Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types', *Genome Res*, 20: 761-70.
Salameh, T. J., X. Wang, F. Song, B. Zhang, S. M. Wright, C. Khunsriraksakul, Y. Ruan, and F. Yue. 2020. 'A supervised learning framework for chromatin loop detection in genome-wide contact maps', *Nat Commun*, 11: 3428.
Sauria, Michael E. G., and James Taylor. 2017. 'QuASAR: Quality Assessment of Spatial Arrangement Reproducibility in Hi-C Data', *bioRxiv*: 204438.
Schmitt, A. D., M. Hu, and B. Ren. 2016. 'Genome-wide mapping and analysis of chromosome architecture', *Nat Rev Mol Cell Biol*, 17: 743-55.

Schoenfelder, S., and P. Fraser. 2019. 'Long-range enhancer-promoter contacts in gene expression control', *Nat Rev Genet*, 20: 437-55.

Schuler, G. D., M. S. Boguski, E. A. Stewart, L. D. Stein, G. Gyapay, K. Rice, R. E. White, P. Rodriguez-Tomé, A. Aggarwal, E. Bajorek, S. Bentolila, B. B. Birren, A. Butler, A. B. Castle, N. Chiannilkulchai, A. Chu, C. Clee, S. Cowles, P. J. Day, T. Dibling, N. Drouot, I. Dunham, S. Duprat, C. East, C. Edwards, J. B. Fan, N. Fang, C. Fizames, C. Garrett, L. Green, D. Hadley, M. Harris, P. Harrison, S. Brady, A. Hicks, E. Holloway, L. Hui, S. Hussain, C. Louis-Dit-Sully, J. Ma, A. MacGilvery, C. Mader, A. Maratukulam, T. C. Matise, K. B. McKusick, J. Morissette, A. Mungall, D. Muselet, H. C. Nusbaum, D. C. Page, A. Peck, S. Perkins, M. Piercy, F. Qin, J. Quackenbush, S. Ranby, T. Reif, S. Rozen, C. Sanders, X. She, J. Silva, D. K. Slonim, C. Soderlund, W. L. Sun, P. Tabar, T. Thangarajah, N. Vega-Czarny, D. Vollrath, S. Voyticky, T. Wilmer, X. Wu, M. D. Adams, C. Auffray, N. A. Walter, R. Brandon, A. Dehejia, P. N. Goodfellow, R. Houlgatte, J. R. Hudson, Jr., S. E. Ide, K. R. Iorio, W. Y. Lee, N. Seki, T. Nagase, K. Ishikawa, N. Nomura, C. Phillips, M. H. Polymeropoulos, M. Sandusky, K. Schmitt, R. Berry, K. Swanson, R. Torres, J. C. Venter, J. M. Sikela, J. S. Beckmann, J. Weissenbach, R. M. Myers, D. R. Cox, M. R. James, D. Bentley, P. Deloukas, E. S. Lander, and T. J. Hudson. 1996. 'A gene map of the human genome', *Science*, 274: 540-6.

Schwarzer, W., N. Abdennur, A. Goloborodko, A. Pekowska, G. Fudenberg, Y. Loe-Mie, N. A. Fonseca, W. Huber, C. H. Haering, L. Mirny, and F. Spitz. 2017. 'Two independent modes of chromatin organization revealed by cohesin removal', *Nature*, 551: 51-56.

Sexton, Tom, Eitan Yaffe, Ephraim Kenigsberg, Frédéric Bantignies, Benjamin Leblanc, Michael Hoichman, Hugues Parrinello, Amos Tanay, and Giacomo Cavalli. 2012. 'Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome', *Cell*, 148: 458-72.

Shukla, S., E. Kavak, M. Gregory, M. Imashimizu, B. Shutinoski, M. Kashlev, P. Oberdoerffer, R. Sandberg, and S. Oberdoerffer. 2011. 'CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing', *Nature*, 479: 74-9.

Skene, P. J., and S. Henikoff. 2017. 'An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites', *Elife*, 6.

Spracklin, George, Nezar Abdennur, Maxim Imakaev, Neil Chowdhury, Sriharsa Pradhan, Leonid Mirny, and Job Dekker. 2021. 'Heterochromatin diversity modulates genome compartmentalization and loop extrusion barriers', *bioRxiv*: 2021.08.05.455340.

Strom, A. R., A. V. Emelyanov, M. Mir, D. V. Fyodorov, X. Darzacq, and G. H. Karpen. 2017. 'Phase separation drives heterochromatin domain formation', *Nature*, 547: 241-45.

Symmons, O., V. V. Uslu, T. Tsujimura, S. Ruf, S. Nassari, W. Schwarzer, L. Ettwiller, and F. Spitz. 2014. 'Functional and topological characteristics of mammalian regulatory domains', *Genome Res*, 24: 390-400.

Tang, Z., O. J. Luo, X. Li, M. Zheng, J. J. Zhu, P. Szalaj, P. Trzaskoma, A. Magalska, J. Wlodarczyk, B. Ruszczycki, P. Michalski, E. Piecuch, P. Wang, D. Wang, S. Z. Tian, M. Penrad-Mobayed, L. M. Sachs, X. Ruan, C. L. Wei, E. T. Liu, G. M. Wilczynski, D. Plewczynski, G. Li, and Y. Ruan. 2015. 'CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription', *Cell*, 163: 1611-27.

Tavares-Cadete, F., D. Norouzi, B. Dekker, Y. Liu, and J. Dekker. 2020. 'Multi-contact 3C reveals that the human genome during interphase is largely not entangled', *Nat Struct Mol Biol*.

Tedeschi, A., G. Wutz, S. Huet, M. Jaritz, A. Wuensche, E. Schirghuber, I. F. Davidson, W. Tang, D. A. Cisneros, V. Bhaskara, T. Nishiyama, A. Vaziri, A. Wutz, J. Ellenberg, and J. M. Peters. 2013. 'Wapl is an essential regulator of chromatin structure and chromosome segregation', *Nature*, 501: 564-8.

Uhler, C., and G. V. Shivashankar. 2017. 'Regulation of genome organization and gene expression by nuclear mechanotransduction', *Nat Rev Mol Cell Biol*, 18: 717-27.

Ursu, O., N. Boley, M. Taranova, Y. X. R. Wang, G. G. Yardimci, W. Stafford Noble, and A. Kundaje. 2018. 'GenomeDISCO: a concordance score for chromosome conformation capture experiments using random walks on contact map graphs', *Bioinformatics*, 34: 2701-07.

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

Valton, Anne-Laure, Sergey V. Venev, Barbara Mair, Eraj Khokhar, Amy H. Y. Tong, Matej Usaj, Katherine S. K. Chan, Athma A. Pai, Jason Moffat, and Job Dekker. 2021. 'A cohesin traffic pattern genetically linked to gene regulation', *bioRxiv*: 2021.07.29.454218.

van Berkum, N. L., and J. Dekker. 2009. 'Determining spatial chromatin organization of large genomic regions using 5C technology', *Methods Mol Biol*, 567: 189-213.

van Steensel, B., and E. E. M. Furlong. 2019. 'The role of transcription in shaping the spatial organization of the genome', *Nat Rev Mol Cell Biol*, 20: 327-37.

Vogel, M. J., D. Peric-Hupkes, and B. van Steensel. 2007. 'Detection of in vivo protein-DNA interactions using DamID in mammalian cells', *Nat Protoc*, 2: 1467-78.

Vouzas, A. E., and D. M. Gilbert. 2021. 'Mammalian DNA Replication Timing', *Cold Spring Harb Perspect Biol*, 13.

Wang, G., C. L. Achim, R. L. Hamilton, C. A. Wiley, and V. Soontornniyomkij. 1999. 'Tyramide signal amplification method in multiple-label immunofluorescence confocal microscopy', *Methods*, 18: 459-64.

Wang, Y., Y. Zhang, R. Zhang, T. van Schaik, L. Zhang, T. Sasaki, D. Peric-Hupkes, Y. Chen, D. M. Gilbert, B. van Steensel, A. S. Belmont, and J. Ma. 2021. 'SPIN reveals genome-wide landscape of nuclear compartmentalization', *Genome Biol*, 22: 36.

Wang, Yaxing, Luis Herranz, and Joost van de Weijer. 2020. 'Mix and Match Networks: Cross-Modal Alignment for Zero-Pair Image-to-Image Translation', *International Journal of Computer Vision*, 128: 2849-72.

Weintraub, A. S., C. H. Li, A. V. Zamudio, A. A. Sigova, N. M. Hannett, D. S. Day, B. J. Abraham, M. A. Cohen, B. Nabet, D. L. Buckley, Y. E. Guo, D. Hnisz, R. Jaenisch, J. E. Bradner, N. S. Gray, and R. A. Young. 2017. 'YY1 Is a Structural Regulator of Enhancer-Promoter Loops', *Cell*, 171: 1573-88 e28.

Wen, Z., Z. T. Huang, R. Zhang, and C. Peng. 2018. 'ZNF143 is a regulator of chromatin loop', *Cell Biol Toxicol*, 34: 471-78.

Wilber, A., U. Tschulena, P. W. Hargrove, Y. S. Kim, D. A. Persons, C. F. Barbas, 3rd, and A. W. Nienhuis. 2010. 'A zinc-finger transcriptional activator designed to interact with the gamma-globin gene promoters enhances fetal hemoglobin production in primary human adult erythroblasts', *Blood*, 115: 3033-41.

Wu, F., B. G. Olson, and J. Yao. 2016. 'DamID-seq: Genome-wide Mapping of Protein-DNA Interactions by High Throughput Sequencing of Adenine-methylated DNA Fragments', *J Vis Exp*: e53620.

Wutz, G., C. Varnai, K. Nagasaka, D. A. Cisneros, R. R. Stocsits, W. Tang, S. Schoenfelder, G. Jessberger, M. Muhar, M. J. Hossain, N. Walther, B. Koch, M. Kueblbeck, J. Ellenberg, J. Zuber, P. Fraser, and J. M. Peters. 2017. 'Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins', *EMBO J*, 36: 3573-99.

Xiong, K., and J. Ma. 2019. 'Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions', *Nat Commun*, 10: 5069.

Yaffe, E., and A. Tanay. 2011. 'Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture', *Nat Genet*, 43: 1059-65.

Yan, K. K., G. G. Yardimci, C. Yan, W. S. Noble, and M. Gerstein. 2017. 'HiC-spector: a matrix library for spectral and reproducibility analysis of Hi-C contact maps', *Bioinformatics*, 33: 2199-201.

Yang, T., F. Zhang, G. G. Yardımcı, F. Song, R. C. Hardison, W. S. Noble, F. Yue, and Q. Li. 2017. 'HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient', *Genome Res*, 27: 1939-49.

Yardimci, G. G., H. Ozadam, M. E. G. Sauria, O. Ursu, K. K. Yan, T. Yang, A. Chakraborty, A. Kaul, B. R. Lajoie, F. Song, Y. Zhan, F. Ay, M. Gerstein, A. Kundaje, Q. Li, J. Taylor, F. Yue, J. Dekker, and W. S. Noble. 2019. 'Measuring the reproducibility and quality of Hi-C data', *Genome Biol*, 20: 57.

Yildirim, Asli, Nan Hua, Lorenzo Boninsegna, Guido Polles, Ke Gong, Shengli Hao, Wenyuan Li, Xianghong Jasmine Zhou, and Frank Alber. 2022. 'Population-based structure modeling reveals key roles of nuclear microenviroment in gene functions', *bioRxiv*: 2021.07.11.451976.

Zhang, Jing, Donghoon Lee, Vineet Dhiman, Peng Jiang, Jie Xu, Patrick McGillivray, Hongbo Yang, Jason Liu, William Meyerson, Declan Clarke, Mengting Gu, Shantao Li, Shaoke Lou, Jinrui Xu, Lucas Lochovsky, Matthew Ung, Lijia Ma, Shan Yu, Qin Cao, Arif Harmanci, Koon-Kiu Yan, Anurag Sethi, Gamze Gürsoy, Michael Rutenberg Schoenberg, Joel Rozowsky, Jonathan Warrell, Prashant Emani, Yucheng T. Yang, Timur Galeev, Xiangmeng Kong, Shuang Liu, Xiaotong Li, Jayanth Krishnan, Yanlin Feng, Juan Carlos Rivera-Mulia, Jessica Adrian, James R. Broach, Michael Bolt, Jennifer Moran, Dominic Fitzgerald, Vishnu Dileep, Tingting Liu, Shenglin Mei, Takayo Sasaki, Claudia Trevilla-Garcia, Su Wang, Yanli Wang, Chongzhi Zang, Daifeng Wang, Robert J. Klein, Michael Snyder, David M. Gilbert, Kevin Yip, Chao Cheng, Feng Yue, X. Shirley Liu, Kevin P. White, and Mark Gerstein. 2020. 'An integrative ENCODE resource for cancer genomics', *Nature Communications*, 11: 3696.

Zhang, R., and J. Ma. 2020. 'MATCHA: Probing multi-way chromatin interaction with hypergraph representation learning', *Cell Syst*, 10: 397-407 e5.

Zhao, Z., G. Tavoosidana, M. Sjolinder, A. Gondor, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K. S. Sandhu, U. Singh, V. Pant, V. Tiwari, S. Kurukuti, and R. Ohlsson. 2006. 'Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions', *Nat Genet*, 38: 1341-7.

Zheng, H., and W. Xie. 2019. 'The role of 3D genome organization in development and cell differentiation', *Nat Rev Mol Cell Biol*, 20: 535-50.

Zheng, M., S. Z. Tian, D. Capurso, M. Kim, R. Maurya, B. Lee, E. Piecuch, L. Gong, J. J. Zhu, Z. Li, C. H. Wong, C. Y. Ngan, P. Wang, X. Ruan, C. L. Wei, and Y. Ruan. 2019. 'Multiplex chromatin interactions with single-molecule precision', *Nature*, 566: 558-62.

Zheng, X., J. Hu, S. Yue, L. Kristiani, M. Kim, M. Sauria, J. Taylor, Y. Kim, and Y. Zheng. 2018. 'Lamins Organize the Global Three-Dimensional Genome from the Nuclear Periphery', *Mol Cell*, 71: 802-15 e7.

Zhou, Qiling, Miao Yu, Roberto Tirado-Magallanes, Bin Li, Lingshi Kong, Mingrui Guo, Zi Hui Tan, Sanghoon Lee, Li Chai, Akihiko Numata, Touati Benoukraf, Melissa Jane Fullwood, Motomi Osato, Bing Ren, and Daniel G. Tenen. 2021. 'ZNF143 mediates CTCF-bound promoter–enhancer loops required for murine hematopoietic stem and progenitor cell function', *Nature Communications*, 12: 43.

Zuin, J., J. R. Dixon, M. I. van der Reijden, Z. Ye, P. Kolovos, R. W. Brouwer, M. P. van de Corput, H. J. van de Werken, T. A. Knoch, IJcken W. F. van, F. G. Grosveld, B. Ren, and K. S. Wendt. 2014. 'Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells', *Proc Natl Acad Sci U S A*, 111: 996-1001.