THE UNIVERSITY OF CHICAGO

NOVEL METHODS FOR IN-DEPTH INVESTIGATION OF CHROMATIN STRUCTURE AND EPIGENETIC LANDMARK

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF CHEMISTRY

BY

QIANCHENG YOU

CHICAGO, ILLINOIS

JUNE 2019

For my family who has always supported me

For all my friends and colleagues who offered
their helping hands and encouragements

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I wish to express the best appreciation to my advisor, Professor Chuan He. I thank him for all the opportunities he provides. His scientific insight and guidance help me with my graduate research on these challenging and exciting topics; his patience and encouragement lead me to overcome the difficulties in scientific exploration.

I would like to thank Professor Tao Pan for his instruction on how to address the RNA related problems. I benefit a lot from his expertise in RNA biology and thank him to serve on my thesis committee. I am grateful to Professor Joseph A. Piccirilli for being my committee member again.

I would also like to thank my collaborators, Dr. Zhengqing Ouyang and Mr. Anthony Cheng from Jax Laboratory for their expertise in data analysis for the CAP-C project, Dr. Dali Han for data analysis on Nano-5hmC-Seal project.

I thank the University of Chicago genome facilities, and Berry Genomics (Beijing, China) for help with next generation sequencing.

All the work mentioned here cannot be done without other He group members and I have enjoyed all the fruitful discussion with them.

Finally, I would like to thank my family, my wife and all my friends for all their love and company. It is them who make me stronger and support me to go through all the hard time in life.

ABSTRACT

The development of next-generation sequencing has brought a comprehensive understanding of human genome to us. It has been well undertand that human genes are not consecutive but contains numerous *cis*-regulatory elements. These cis-regulatory elements was proved to be crucial for lineage-specific gene expression during development. To elucidate the gene-regulatory interactome is important not only in understanding the development of human and other species but also provide crucial guidance on studying diseases. This task is surprisingly difficult as cis-regulatory elements can regulate genes that are not their immediate neighbors. Some elements could modulate gene expression from a large genome distance. Thus, to uncover the mystery of gene regulation requires a better understanding of how chromatin fibers folded in 3D among different cell type. Moreover, epigenetic markers emerge to be proven as important indicators for gene regulation. Recent years have seen rapid progress in technologies for genome-wide analysis of 3D genome organization. However, there remains largely unknown between chromatin structure and gene regulation. In addition, lacking high resolution chromatin contacts maps also limit our understanding of mechanism of chromatin folding.To meet the requirement, two major strategies are designed and will be discussed in the thesis.

The Chemical-crosslinking Assisted Proximity Capture (CAP-C) strategy improves the chromatin contact maps at high resolution by capture more short range-chromatin contacts. Using this strategy, we also discovered an inexplicable relationship between chromatin structure and transcription and proposed that organization of genome may not be a determinant of transcription but also a consequence of its function.

The Nano-hmC-Seal strategy enable us to profile 5hmC location with limited materials. We have successifully utilize this method in studying AML and predict its clinical outcome

1 Introduction

1.1 Features of chromatin architecture and gene regulation

The ~ 2-meter-long mammalian genomic DNA is organized elegantly into compact, knot-free chromosomes in cells (*1, 2*). Thus, packaging the genome requires extensive folding to fit within the volume of a constrained cell nucleus. Protein-mediated DNA interactions are thought to influence chromatin compaction and affect transcriptional outcomes. As cells enter different cell cycle stages or differentiate according to their intended pathways of development, their chromosomes often undergo re-organization, experiencing changes in gene expression which impact cellular processes and functions (*3-6*). The structure of chromosomes is thus critical for the cell to function properly; disruption of chromosome territories often leads to the dysfunction of gene regulation or contributes to diseases (*7-10*).

Changes of gene expression during development is usually coupled with switching of promoter-enhancer interactions. Enhancers confer tissue-specific expression of target genes by recruiting sequence-specific transcription factors and chromatin remodeling complexes (*11-14*). Besides, insulator elements play an important role in gene regulation by limiting the enhancer function and preventing the spread of heterochromatin (*15*). Thus, for accurate gene expression, it is important to question how *cis*-regulatory elements are spatially stalled to prevent enhancers targeting unfavorable promoters and promote those preferred ones. And in turn, how tuning gene transcription affects the overall 3D chromatin organization.

1.2 Tools to study 3D genome organization

Recent years have seen rapid progress in technologies for genome-wide analysis of 3D genome organization in living organisms. Two major technologies have been introduced and used extensively to study the 3D chromatin structure: microscopy-based imaging tools (fluorescence in

situ hybridization (FISH)) (*16*) and chromosome conformation capture (3C)-coupled sequencing methods (*17*).

Fluorescence in situ hybridization (FISH) uses fluorescent bearing probes that bind specifically to the parts of chromatin with high degree of sequence complementary (*18*). Such DNA or RNA probes will either tagged directly with fluorophores or biotin or digitonin that can be further recognized by antibody linked fluorophores (*19*). The chromosome samples are prepared by cross-linking with para-formaldehyde and firmly attached to substrate like glasses. After denaturing and blocking repetitive DNA sequence with short DNA fragments, the designed FISH probes are applied to hybridize to their targeting regions. The results are then visualized and quantified using a microscope that is capable of exciting the dye and recording images (*20*) (Fig.1.1 right).

The use of FISH enables researchers to visualize multiple loci that are spatially in proximity in each cell, making it possible to investigate the heterogenicity of chromatin organization. In addition, live cell imaging can reveal dynamic changes in interactions among chromatin loci. However, the drawbacks of FISH are obvious as it requires pre-knowledge on possible interacting loci before designing probes. Moreover, the fosmid probes used for chromatin labeling are 40kb long in average, making it difficult to resolve those interacting loci that span less than 100Kb long (*21*). Lastly, the resolution achieved by FISH is strongly constrained by the limitation of diffraction of light sources which is ~200 nm laterally and >500 nm axially (*22*). Taken together, FISH is frequently used in validating those known long-range enhancer promoter interactions. The lack of resolution and throughput limits its application in studying chromatin structure genome-wide.

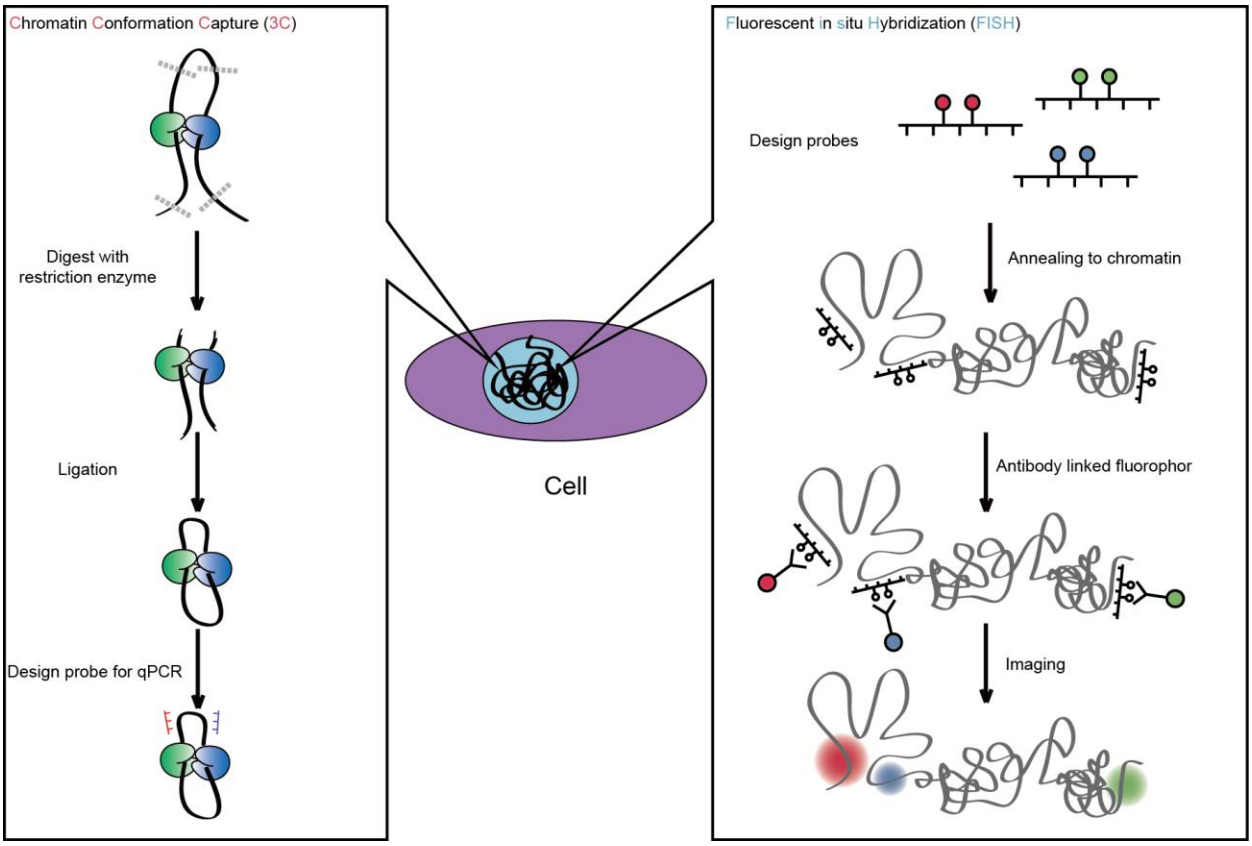


Fig.1.1 Tools to study chromatin architecture. Left: scheme of Chromatin Conformation Capture (3C); Right: scheme of Fluorescent in situ Hybridization (FISH)

Besides visualizing directly with imaging tools, 3D chromatin structure can also be inferred by pairwise contact frequency among loci. Chromatin conformation capture (3C) (*17*) and its derivatives (*23*) were introduced as an alternative strategy to study chromatin interaction. 3C is performed by first crosslinking chromatin with formaldehyde to preserve the 3D chromatin conformation, the DNA was then digested with restriction enzyme (i.e. HindIII) followed by re-ligation. If two loci are spatially in proximal at the time of crosslinking, they will be joint together known as proximity ligation, the ligated product could then be assessed by designing probes for qPCR to quantify the frequency by comparing with those non-ligated ones. Though 3C could be applied to investigate interreacting loci at all range, lacking throughput limits its broad use. Moreover, the

results acquired by 3C are often the consequence of a large population of cells, making studying the heterogenicity of chromatin structure impossible (Fig.1.1 left).

1.3 3C derived technology

In recent years, next-generation sequencing has been extensively applied to couple with all kind of 3C derived technology, resulting in proliferation of tools to map chromatin structure with various coverage and resolution. Some typical technologies that have been used extensively are introduced in below (*24*) (Fig.1.2):

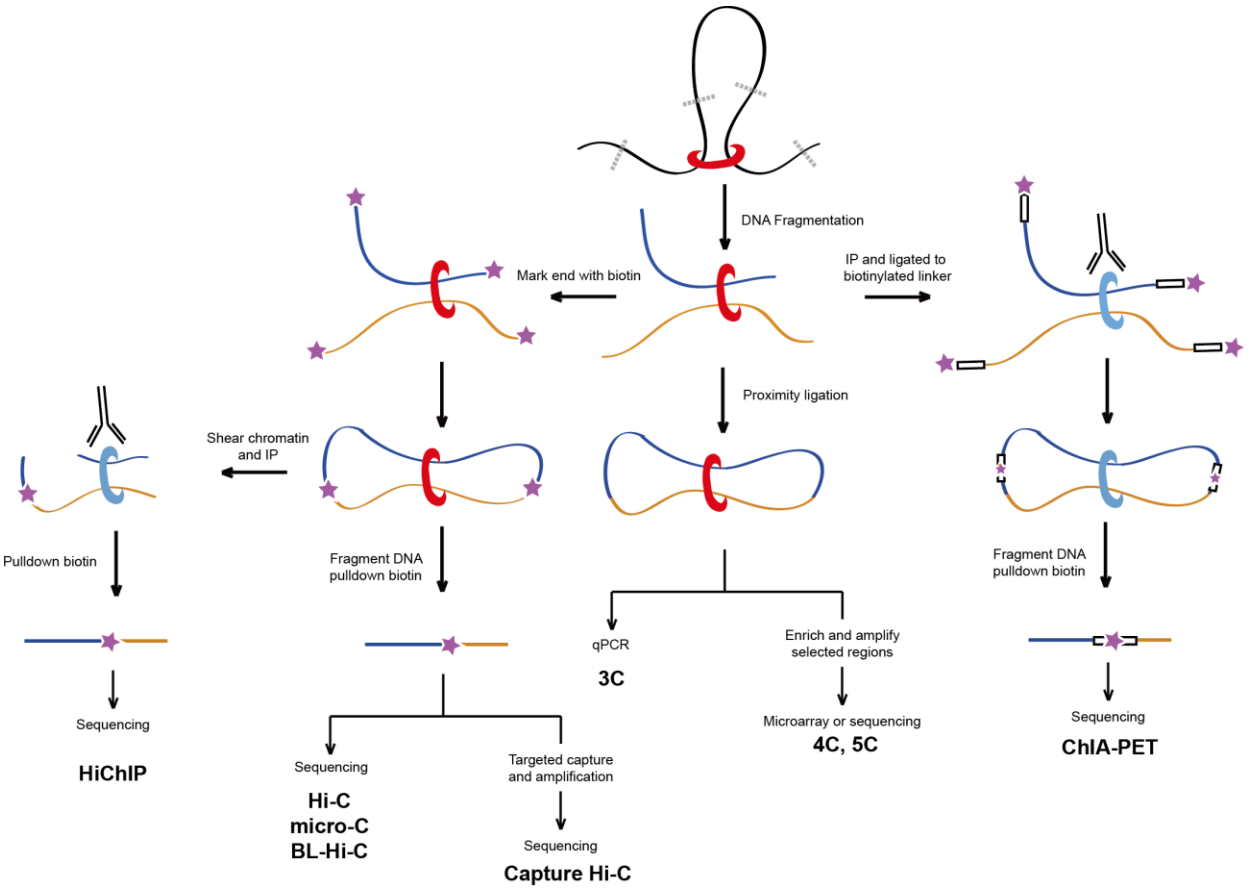


Fig.1.2 Brief overview of the C-technologies. All C-technologies share the steps of formaldehyde cross-linking, DNA fragmentation, and proximity ligation but may vary in whether to label the fragment end with biotin, the strategy to enrich the regions of interest by either antibody or designed sequence probe, and the quantification methods. The biotin label is represented by the purple star. Abbreviation: IP, immunoprecipitation.

4C: 4C (*25*) is a "one to all" technology that allows for genome-wide identification of all possible interacting patterners to one specific locus. It requires designing primers specific targeting to the interested locus (or called as view point). The primers will then by hybridized with 3C ligated products. Through PCR, all potential candidates interacting with the view point will be amplified and then quantified with high-throughput sequencing (*26*).

5C: 5C (*27*) is a "many to many" technologies, which for the first time greatly improve the genome coverage of 3C. This method takes advantage of designing primer pools that could target multiple chromatin locus. By annealing to the 3C ligated products, all interacting partners with the designed primer pools will be captured and amplified from PCR and analyzed through high-throughput sequencing (*28*) (*29*).

Hi-C: Hi-C (*30*) is a "all to all" technology, which unbiased captured all possible interactions across the genome. The major modification of Hi-C over conventional 3C is labeling the DNA fragments with biotin before proximity ligation in order to allow for the subsequent enrichment of all ligation products. There are all kinds of variants of Hi-C depending on the fragmentation method or the ligation condition, including micro-C (*31*), DNase Hi-C (*32*) and in situ Hi-C (*29*). Of note, in situ Hi-C is not only the most frequently used but also bears the highest resolution of all 3C related methods largely owing to the deep sequencing.

ChIA-PET: (*33*) Chromatin interaction analysis by paired-end tag (ChIA-PET) is a "many to many" technology which combines chromatin immunoprecipitation (ChIP) and proximity ligation and sequencing. In ChIA-PET, cross-linked chromatins are sheared by sonication and then subjected to immunoprecipitation by using antibodies against the protein of interest, followed by proximity ligation, amplification of proximity ligation products, and sequencing. Compared to Hi-C,

ChIA-PET is a cost-effective method especially suitable to investigate enhancer-promoter interactions at high resolution (*34*).

Capture-C: Capture-C (*35*) (*36*) is a "many to all" method which combines Hi-C with targeted DNA fragment capture and sequencing. Unlike 5C, Capture-C requires designing pools of DNA or RNA oligos containing biotin to specifically hybridized with multiple proximity ligation products corresponding to those interested locus (i.e. promoters). The captured DNA sequence is sequenced to reveal spatial contacts within these regions (*37*) (*38, 39*).

HiChIP: Only until recently, HiChIP, (*40*) a "many to all" technology, has been introduced as an alternative method for ChIA-PET. Unlike ChIA-PET, HiChIP is performed by in situ ligate of proximal DNA fragment within nucleus just like in situ Hi-C before immunoprecipitated of protein of interest. By doing so, it preserves more chromatin contacts of interest compared to ChIA-PET and requires less amount of cell input. In addition, by comparing with Capture-C, HiChIP doesn't require to synthesis oligo pools targeting interested locus, making it unique method to map versatile chromatin interaction by choosing different antibody against protein of interest at high resolution (*41*).

1.4 Chromatin features detected by 3C derivatives

The above described methods and their variations employ formaldehyde-mediated crosslinking followed either by in-situ enzymatic and proximity ligation or labeling with fluorescent bearing probes to infer spatial relationships between genomic loci. They have been instrumental in elucidating the principles of chromatin folding. Studies using these techniques have confirmed the existence of multiple layers of genome organization such as chromosome territories (*42*), compartments (*30*), topologically associating domains (TADs) (*43*), sub-TADs (*28*), insulated neighborhoods (*44*) and chromatin loops (*29*) (Fig.1.3).

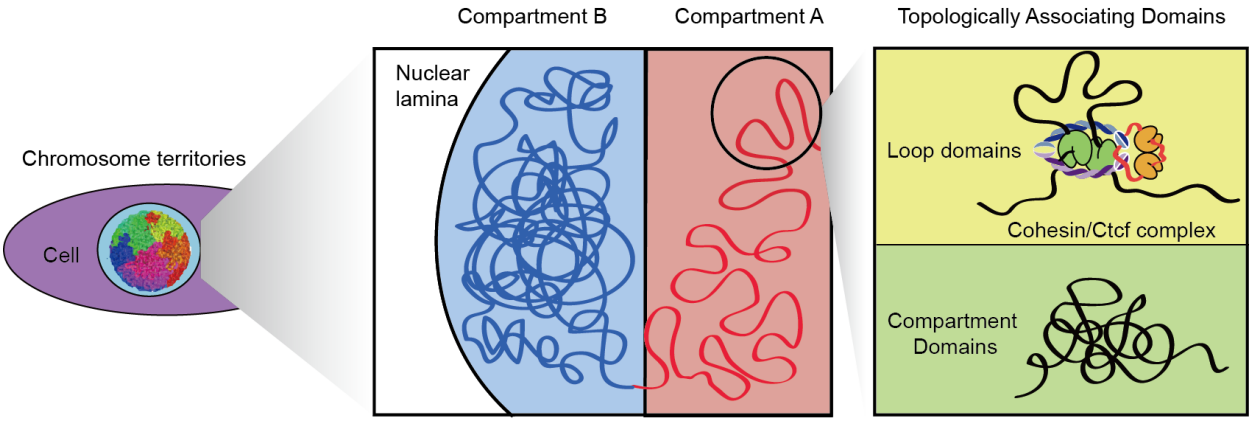


Fig.1.3 Existence of multiple layers of genome organization. Chromosome are partitioned into chromosome territories, compartments, topologically associating domains (TADs), sub-TAD (loop domain and compartment domain) and chromatin loops. CTCF is shown in green, each cohesin sub-unit is presented in different color. (Smc1 in blue; Smc3 in purple; Rad21 as a red belt; Scc3 in orange)

Chromosome territory, which could be visualized via FISH, was defined as each chromosome occupies its own position inside nucleus (*45*). Separation of chromosome into different partitions is also supported by Hi-C. Chromosome territory is not randomly distributed inside the nucleus, rather than that, Long and gene poor chromosomes tend to locate at the periphery of nucleus while those short and gene rich chromosomes will prefer to reside at the center (*46*). Moreover, chromosome territory is conserved during mitosis, which suggests an uncharacterized mechanism for mother cells to pass the information of chromosome positioning to their daughter cells (*47*). Lastly, such positioning in chromosome territory also correlates to replication timing as those early replicating loci tend to locate at the center, whereas those late replicating loci have a preference sitting at nucleus periphery (*48*, *49*).

Besides chromosome territory, contact maps generated by 3C methods after normalization display a plaid pattern when viewing in observe/expect mode, suggesting the existence of two compartments inside nucleus (*30*). Principle component analysis on the observe/expect matrix

reveals that chromosome is partitioned in A/B compartment based on the first principle component. Intriguingly, compartment A regions are early replicating and high density of gene with active transcription. They are enriched with H3K36me3 and DNaseI hyper sensitive sites (*30*). On the contrary, compartment B regions are late replicating enriched in inactive gene locus, which overlap strongly with lamina-associated domains as well as heterochromatin (*50*). Compartment A and B is also cell type specific. Human embryonic stem cells undergo over 30% compartment switching during differentiation (*51*). Thanks to high resolution achieved by deep-sequencing of in situ Hi-C, compartment A and B are further divided into 5 sub-compartments, namely A1, A2, B1, B2, B3 (*29*). And intriguingly, each sub-compartment is associate with a specific pattern of histone modifications (*52*).

When zooming in the contact matrix generated by Hi-C to achieve higher resolution using the smaller bin size (around 40Kb resolution), computational algorithms measuring the directionality of interactions in the genome identified topologically associating domains (TADs) (*43, 53*), ranging from 0.2-1.0Mb long. It corresponds to sequences that prefers to interact with themselves rather than other genomic regions. In the study of mouse embryonic stem cells (mESCs), it was reported that more than 90% of its genomes are partitioned into TADs with a median length of 800Kb. Besides, TADs are also found to be conserved among species. *Drosophila*, (*54*) *C. elegans* (*55*), zebra fish (*56*), yeast (*31, 57*) and human genomes are all organized into different size of TADs depending on their genome size. On the contrary, Hi-C contact maps generated from Arabidopsis thaliana showed no sign of TADs, indicating that genome of plant may adopt different organization principles (*58*). It is important to point out that hierarchically multiple TADs interact with each other and form compartment. Each TADs are marked specifically with either active or inactive histone modification. On the boundaries of TADs, proteins like CTCF and cohesin was found to

be enriched together with some house-keeping gene and active chromatin transcription start sites (TSS) (*43*). There exist many evidences to support that TADs are the fundamental unit of chromatin organization (*59, 60*). First of all, TADs were found to be stable across cell types and tends to remain unchanged during cell differentiation. The change of histone modification rarely changes the TADs positioning (*51*). Secondly, TADs were identified to be conserved among species, 50-70% of human and mouse TADs share similar boundaries (*51*). Thirdly, the enhancers and promoters within the same TADs tend to be coordinately weakly. Lastly, boundaries of replication domains were found to show one-to-one correspondence with TAD boundaries, suggesting that the TAD is a basic unit for replication (*61-63*).

The presence of smaller structure units within TADs were also reported by high resolution in situ Hi-C experiment, namely sub-TADs (*28*), contact domains (*29*) and insulation neighborhoods (*36*). Like TADs, sub-TADs display self-association with a decrease in contact frequency among sub-TADs boundaries. The boundaries of sub-TADs and many contact domain boundaries also enriched for CTCF/cohesin (*29*). Compared to TADs, sub-TADs seem to be less conserved among cell types but rather appeared to be more related to tissue specific gene expression (*64*). In addition, insulation neighborhoods are defined as gene regulation units with Ctcf/cohesin anchors at their boundaries, appearing to constrain the enhancer-promoter interactions within it (*65*).

According to the polymeric nature of chromatin fibers, two genomic loci will not contact with each other at high frequency via random collision (*66, 67*). In this way, chromatin looping occurs rarely without strictly regulation. Such chromatin looping always reflect the biological relevant activities like enhancer-promoter interaction during gene regulation. Given the fact that regulation elements distributed at a range of several hundreds to a few thousands base away from each other (*37, 68-72*). Kilobase resolution Hi-C contact maps are required to achieve identifying looping

contacts across genome. Visualized as a bright spot on the Hi-C contact matrix, two types of chromatin looping events have been identified (*29*). The first type is constitutive and invariant across cell types, demarcating most of the boundaries of TADs (*73, 74*). Such type of loops is often found to be mediated by CTCF/Cohesin complex. Among which, most of these loops contain convergently positioned CTCF, which explained why only a subset of CTCF will compose the TADs (*29*). On the other hand, another type of loops, mainly constructed by enhancer and enhancer associated factors as well as cohesin subunits together with mediator sub-units, are more cell type specific and might change during cell differentiation and development (*28, 74*).

1.5 Proposed mechanism for chromatin organization

To date, the smallest higher order chromatin structure identified is chromatin loops. It has thus been proposed that chromatin loops are the basic unit for chromatin architecture (*29*). Given the fact that many domain boundaries are formed by Ctcf/cohesin complex, a model called loop extrusion (Fig.1.4 left) has been proposed to explain how hierarchically loops help in formation of domain and compartment (*75-79*). Cohesin is a large protein complex containing four core sub-units: two SMC proteins (SMC1 and SMC3), an alpha-kleisin, and an orthogonal of the yeast Scc3 protein (*80*). The cohesin complex will be loaded onto the DNA by NIPBL, it will then slide and extrude along the DNA sequence until blocked by convergently oriented CTCF (*81-83*). The stopped Cohesin/CTCF complex will constrained the sequence to form domains and serve as insulator to prevent interactions with another domain. Cohesin could also be unloaded by WAPL (*84*), resulting in a dynamic on and off formation of domain along the genome and further influence the transcription.

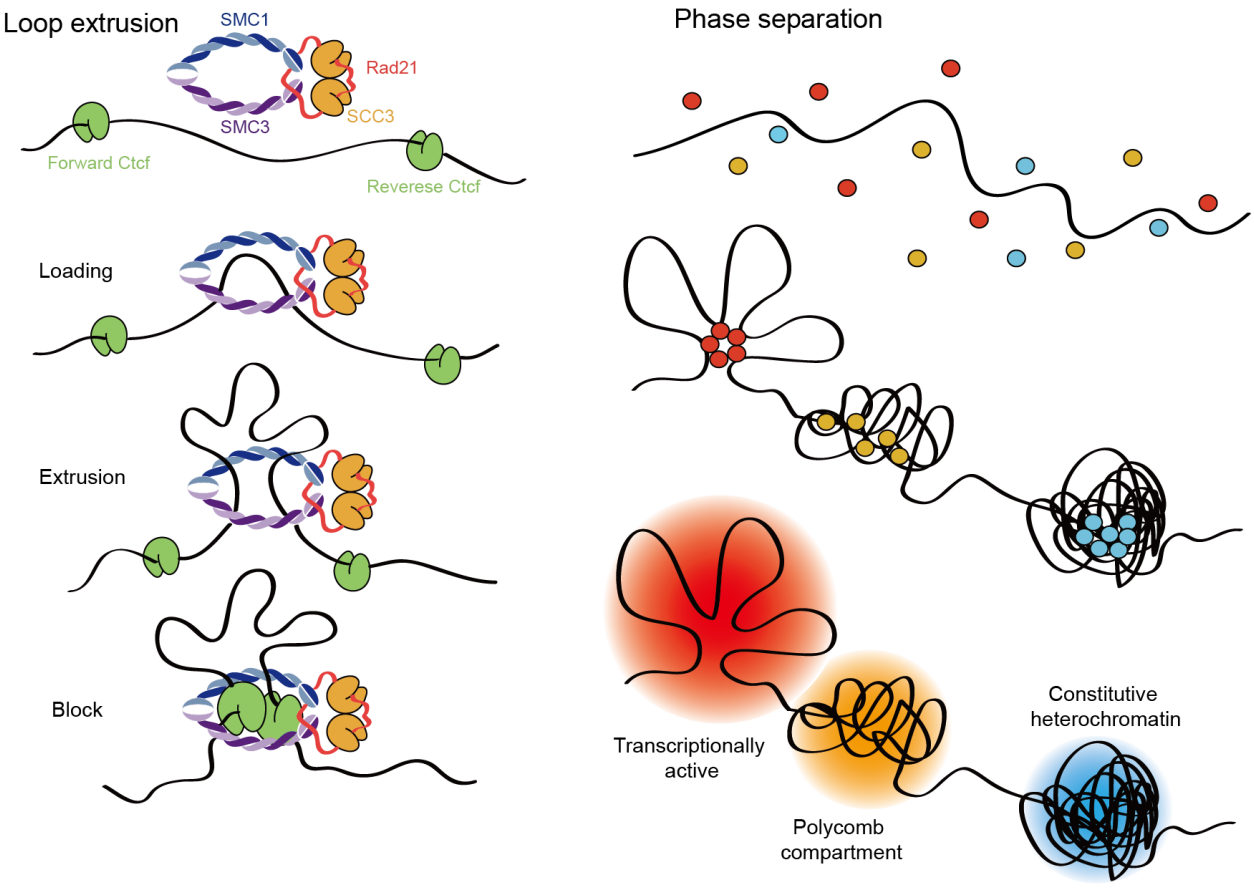


Fig.1.4 Two proposed mechanisms on chromatin organization. Left: TAD formation by loop extrusion. The Cohesin complex forms a ring structure and travels along the DNA fiber, extruding a progressively larger loop until it is stalled by bound CTCF with convergent orientation. (CTCF is shown in green, each cohesin sub-unit is presented in different color. (Smc1 in blue; Smc3 in purple; Rad21 as a red belt; Scc3 in orange)) Right: TAD formation through phase separation. Segregation of chromatin states in the nucleus may occur as a consequence of the presence of different classes of multivalent proteins that mediate class-specific interactions to create different phases, which result in droplets of distinct chromatin states within the nucleus. Red balls represent proteins and histone modifications present at genes or regulatory sequences in a transcriptionally active state, yellow balls represent histone H3 lysine 27 trimethylation (H3K27me3) and Polycomb group proteins, and blue balls represent H3K9me3 and heterochromatin protein 1α (HP1α)

Loop extrusion model indeed explained many of experimental observations from Hi-C (*85-87*), however, there still remains many contradictories as well as mysteries yet to be solved. First, there is no clear evidence to support the cohesin extrusion in vivo (*88*). The loop extrusion model speculated that cohesin subunit SMC1/3 consumes ATP and help its extruding. Indeed, when

ATP is depleted, cells with acute loss of cohesin have no ability to recover the loops. However, this might not be a direct evidence to support the idea that it is cohesin itself consumes the ATP to extrude. There are other work speculating that during transcription RNA polymerase II (RNAPII) might push cohesin ring for extruding (*89, 90*), however, calculation based on the movement speed indicates that RNAPII movement is much slow compared to cohesin extrusion. Beyond that, some simulation work has been done to support the hypothesis that transcription induced supercoiling might serve as a strong force for cohesin movement (*91, 92*). However, no direct evidence has been shown to prove that. Second, Hi-C experiments display a lot of differences between cells with acute loss of CTCF and cohesin, suggesting that there are mechanisms for domain formation other than cohesin/CTCF formed loops (*93-95*). Lastly, CTCF is a relatively small protein (~5 nm) compared to cohesin (~50nm), how such small protein could block the movement of a large protein like cohesin remains unclear (*96*). Besides, eukaryotic DNA is wrapped with histone, how cohesin slide along the DNA with so many histones around is also a mystery. Some studies have proposed cohesin walking along the chromatin but there is no clear evidence to support that (*97*).

Given the fact that compartment domains still remain after acute loss of CTCF, mechanism independent of loop extrusion exists. It was not until recently another theory has been proposed which speculate domain formation through phase separation (*98, 99*) (Fig.1.4 right). On the basis of this model, active and inactive regions of the genome may be able to interact with members of their own class, forming two different phases that preclude inter active and inactive compartmental contacts. This model was largely coming from the fact that chromatins are partitioned into A/B compartment as well as recent discoveries of phase-separated heterochromatin protein 1α (HP1α)-mediated heterochromatin droplets are formed in vitro and can be detected in vivo.

Indeed, recent studies have shown RNAPII forms phase separated droplets both in vitro and in vivo (*100*). This might suggest that transcription activation is able to promote phase separation of active chromatin. Model of phase separation would entail the constant fusion and fission of chromatin droplets, suggesting that compartmental domains are involved in dynamic interactions. Despite the above stated two mechanisms, we don't preclude other possible mechanisms on domain formation. Both loop extrusion and phase separation could only explain partial of the experimental observations. More experiments are needed to further elucidate the principle of chromatin organization.

1.6 Relationship between chromatin structure and gene regulation is complex

The above discoveries on chromatin structure highlighted the chromatin is strictly partitioned into TADs and such phenomenon is conserved across species. Each TAD may contain one or several compartment domains and loop domains while loop domains are thought to be formed through cohesin/CTCF loop extrusion. As chromatin structure is thought to be closely related to gene regulation, many studies have been conducted to indicate that insulators like cohesin and CTCF form loops/domain boudaries in order to help enhancers find their cognate promoters within the same loop/domain, while on the other hand, prevent those enhancers to target promoters that locate outside the loop/domain (*65, 73*). In this way, any inversion or deletion around the boundaries of TADs will cause dysregulation of genes or even worse, disease (*101*). This is best illustrated by studying on *EPHA4* locus (*102*) (Fig.1.5). *EPHA4* at the telomeric side resulted in brachydactyly (short digits) whereas an inversion and a duplication on the centromeric side involving part of the *EPHA4* TAD were shown to be associated with a complex form of syndactyly. Further investigations have revealed that an enhancer clusters located in the *EPHA4* TAD

that normally regulates the expression of *EPHA4* in the limb bud is activating different genes depending on the structure variation. *Paired box3* (*Pax3*) gene is activated in brachydactyly while *Wingless-type MMTV integration site family, member 6* (*Wnt6*) in the syndactyly. All this phenotype comes from the disruption of one TAD boundary so that misexpression of genes occurred due to ectopic interaction of the enhancer with the target gene (*102*).

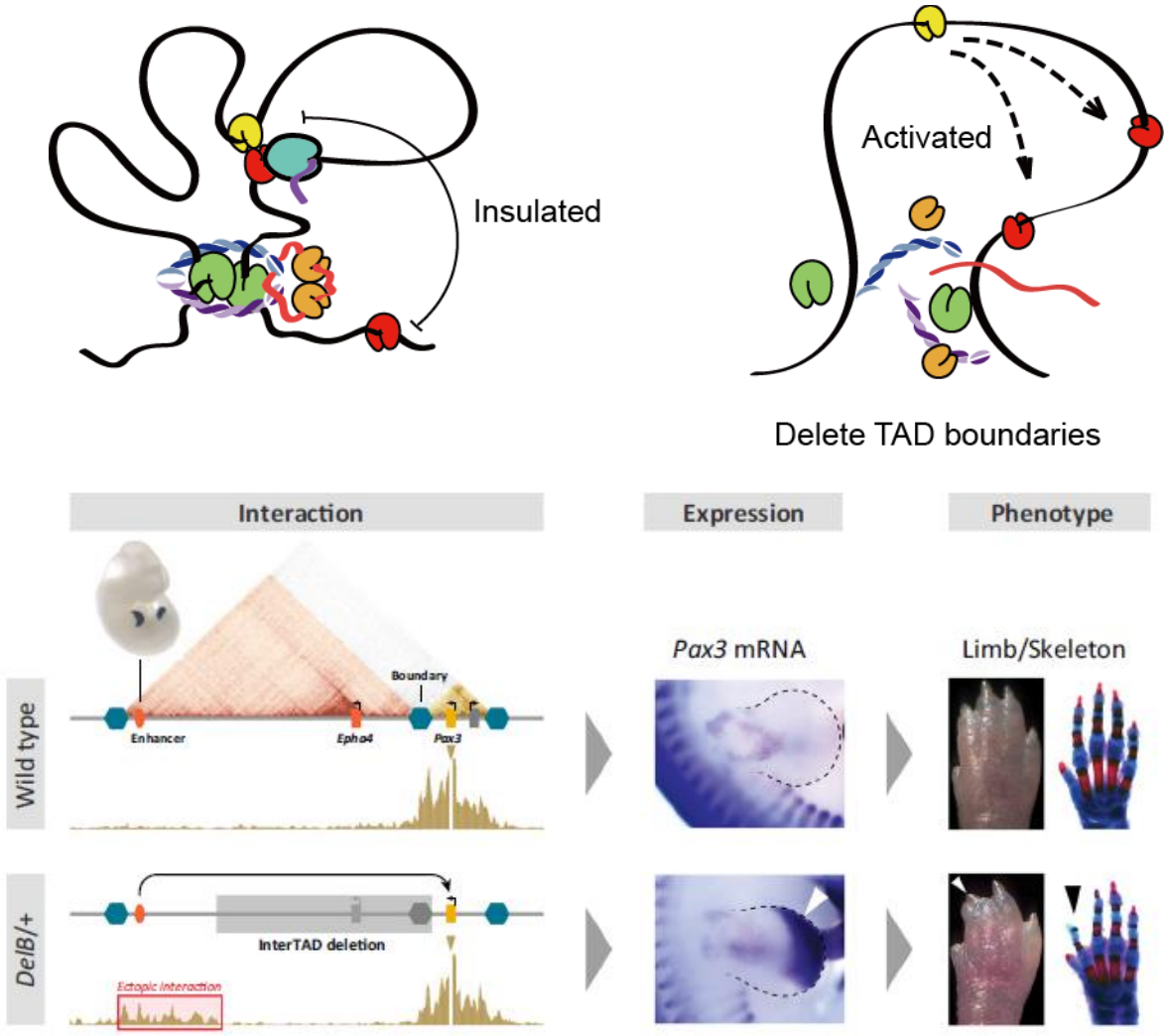


Fig.1.5 Deletion of TAD boundaries lead to disease. Top panel: Cartoons illustrate deletion of TAD boundaries will result in dysregulation of gene expression. Bottom panel: A deletion removing Epha4, parts of the Epha4 TAD, and the boundary region results in ectopic interaction of the Pax3 promoter with enhancers that originally belonged to Epha4. This ectopic interaction results in ectopic expression of Pax3 in the distal limb bud, which results in shortening of the first and second digits (arrow).

As shown above, TAD boundaries play an important role in gene regulation by limiting enhancers to target cognate promoters. CTCF, as an insulator, is often found to be located near the TAD boundaries, such observations suggests a crucial role for CTCF in gene regulation. However, the effects of CTCF removal on transcription can be quiet variable. Complete loss of CTCF is lethal during embryonic development (*103*), whereas haploinsufficiency results in intellectual disability, microcephaly and growth retardation (*104, 105*). Heterozygous CTCF-knockout mice show a high incidence of tumors, and mutation of specific CTCF binding sites correlates with various cancers in humans (*106, 107*). However, recent studies conducted on acute loss of CTCF show contradictory results as only a minority of gene expressions have been changed (*53*). Completely depletion of CTCF only result in 370 differential gene expression, while among which only 43 genes showed 5-fold change in expression. Aside from CTCF depletion, acute loss of Rad21 eliminate all CTCF loops only found 2 genes displayed 5-fold expression change. The drastic difference between phenotype in CTCF/cohesin depletion in living organism and minor changes to gene expression observed in culture cells lead to an important question: How chromatin structure affects the gene expression? More studies are required to further understand it.

On the other hand, we know little about how differential transcription changes the overall chromatin structure. Inhibition of RNAPII in the bacteria *Caulobacter crescentus* and *Bacillus subtilis* results in a great loss of compartment domain (*108*). Similar experiments conducted in mammalian cells seem to show less significant results. Inhibition of RNAPII elongation in D. melanogaster cells only result in reduction of interactions inside compartmental domains but doesn't eliminate the compartment domains (*109*). Moreover, there was minor chromatin structure changes after transcription inhibition in human cancer cell lines (*87*). All these observations indicate that the relationship between transcription and chromatin structure is complex and share

little similarities between species. One possible explanation is that the maintenance of these domains is dependent on the presence of proteins related to transcription rather than on the transcription process (*110*). As there are at least two independent mechanisms toward domain formation, it is thought that transcription might play a bigger role in compartment domain formation while loop extrusion has a more significant impact on loop domain formation. The conclusion major comes from the following aspects: 1. *D. melanogaster* cells only have few loops which largely mediated by PRC2 complex and most of its domains are compartment domains (*111*); 2. Acute loss of CTCF doesn't eliminate all the compartment domain but only CTCF mediated loops (*86*); 3. High resolution Hi-C maps show compartment domains between the start and the termination sites of transcribed genes (*109*). Taken together, more systematically studies are needed to further illustrate the causal relationship between transcription and chromatin structure.

1.7 Scope of thesis

The recent progress in understanding the chromatin structure and its related gene regulation revealed that the relationship between structure and function is much complicated than we thought. 3C related methods have provided evidence that genome is partitioned into segments in a strictly regulated manner. However, the 3C methods by far have been all rely on the formaldehyde crosslinking followed by restriction digestion, resulting in fragmentation of genome into large and uniform pieces. Consequently, the contact matrix generated by 3C methods have low resolution and full of noise. Some of the chromatin features are strongly dependent on the resolution of the current 3C technologies. Thus, new crosslinking strategies which can ideally expose all potential restriction sites, are required to ubiquitously capture proximal contacts at all length scales. The thesis will discuss some progress in a novel crosslinking strategy method development in studying large molecule interactions involving DNA-DNA and RNA-RNA.

Chapter 2 presents a Chemical-crosslinking Assisted Proximity Capture (CAP-C) for chromatin structure study. Using this strategy, we obtained a high-resolution map of local chromatin interactions in mouse embryonic stem cell, discovering two types of domains that are formed through different mechanisms, establishing a close relationship between transcription and chromatin structure.

Chapter 3 presents a Nano-5hmC-Seal for sensitive mapping of 5hmC genome-wide with limited starting materials. Using this strategy, we obtained high-quality maps of 5hmC for AML mice model and epigenetic drugs treated AML patients. We envisioned the possible utility of this method to predict the clinical outcome.

1.8 References

1. T. Sexton, G. Cavalli, The role of chromosome domains in shaping the functional genome. *Cell* **160**, 1049-1059 (2015).
2. A. Pombo, N. Dillon, Three-dimensional genome architecture: players and mechanisms. *Nat Rev Mol Cell Biol* **16**, 245-257 (2015).
3. A. G. West, P. Fraser, Remote control of gene transcription. *Human molecular genetics* **14 Spec No 1**, R101-111 (2005).
4. W. Schwarzer, F. Spitz, The architecture of gene expression: integrating dispersed cis-regulatory modules into coherent regulatory domains. *Curr Opin Genet Dev* **27**, 74-82 (2014).
5. J. H. Gibcus, J. Dekker, The hierarchy of the 3D genome. *Mol Cell* **49**, 773-782 (2013).
6. D. A. Kleinjan, V. van Heyningen, Long-Range Control of Gene Expression: Emerging Mechanisms and Disruption in Disease. *The American Journal of Human Genetics* **76**, 8-32 (2005).
7. A. L. Valton, J. Dekker, TAD disruption as oncogenic driver. *Curr Opin Genet Dev* **36**, 34-40 (2016).
8. S. Groschel *et al.*, A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell* **157**, 369-381 (2014).
9. D. G. Lupianez *et al.*, Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012-1025 (2015).
10. J. E. Bradner, D. Hnisz, R. A. Young, Transcriptional Addiction in Cancer. *Cell* **168**, 629-643 (2017).

11. T.-K. Kim, R. Shiekhattar, Architectural and functional commonalities between enhancers and promoters. *Cell* **162**, 948-959 (2015).

12. P. Deloukas *et al.*, Large-scale association analysis identifies new risk loci for coronary artery disease. *Nature genetics* **45**, 25 (2013).

13. G. Kroemer, G. Mariño, B. Levine, Autophagy and the integrated stress response. *Molecular cell* **40**, 280-293 (2010).

14. C.-T. Ong, V. G. Corces, Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics* **12**, 283 (2011).

15. M. Gaszner, G. Felsenfeld, Insulators: exploiting transcriptional and epigenetic mechanisms. *Nature Reviews Genetics* **7**, 703 (2006).

16. P. R. Langer-Safer, M. Levine, D. C. Ward, Immunological method for mapping genes on Drosophila polytene chromosomes. *Proceedings of the National Academy of Sciences* **79**, 4381-4385 (1982).

17. J. Dekker, K. Rippe, M. Dekker, N. Kleckner, Capturing Chromosome Conformation. *Science* **295**, 1306-1311 (2002).

18. T. Shimi *et al.*, The A-and B-type nuclear lamin networks: microdomains involved in chromatin organization and transcription. *Genes & development* **22**, 3409-3421 (2008).

19. A. N. Boettiger *et al.*, Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature* **529**, 418 (2016).

20. S. Wang *et al.*, Spatial organization of chromatin domains and compartments in single chromosomes. *Science*, aaf8084 (2016).

21. M. Lakadamyali, M. P. Cosma, Advanced microscopy methods for visualizing chromatin structure. *FEBS letters* **589**, 3023-3030 (2015).

22. J. Walter *et al.*, Towards many colors in FISH on 3D-preserved interphase nuclei. *Cytogenetic and genome research* **114**, 367-378 (2006).

23. A. D. Schmitt, M. Hu, B. Ren, Genome-wide mapping and analysis of chromosome architecture. *Nature reviews Molecular cell biology* **17**, 743 (2016).

24. M. Yu, B. Ren, The three-dimensional organization of mammalian genomes. *Annual review of cell and developmental biology* **33**, 265-289 (2017).

25. Z. Zhao *et al.*, Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet* **38**, 1341-1347 (2006).

26. M. Simonis *et al.*, High-resolution identification of balanced and complex chromosomal rearrangements by 4C technology. *Nature methods* **6**, 837 (2009).

27. J. Dostie *et al.*, Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome research* **16**, 1299-1309 (2006).

28. J. E. Phillips-Cremins *et al.*, Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281-1295 (2013).

29. S. S. Rao *et al.*, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).

30. E. Lieberman-Aiden *et al.*, Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326**, 289-293 (2009).

31. T. H. Hsieh *et al.*, Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* **162**, 108-119 (2015).

32. <DNase Hi-C with target capture.pdf>.

33. M. J. Fullwood *et al.*, An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462**, 58 (2009).

34. Z. Tang *et al.*, CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* **163**, 1611-1627 (2015).

35. N. H. Dryden *et al.*, Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome research*, gr. 175034.175114 (2014).

36. J. R. Hughes *et al.*, Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature genetics* **46**, 205 (2014).

37. B. M. Javierre *et al.*, Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* **167**, 1369-1384. e1319 (2016).

38. B. Mifsud *et al.*, Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature genetics* **47**, 598 (2015).

39. S. Schoenfelder *et al.*, Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nature genetics* **47**, 1179 (2015).

40. M. R. Mumbach *et al.*, HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* **13**, 919-922 (2016).

41. R. Fang *et al.*, Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell research* **26**, 1345 (2016).

42. T. Cremer, M. Cremer, Chromosome territories. *Cold Spring Harbor perspectives in biology*, a003889 (2010).

43. J. R. Dixon *et al.*, Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012).

44. J. M. Dowen *et al.*, Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**, 374-387 (2014).

45. A. Bolzer *et al.*, Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS biology* **3**, e157 (2005).

46. J. A. Croft *et al.*, Differences in the localization and morphology of chromosomes in the human nucleus. *The Journal of cell biology* **145**, 1119-1131 (1999).

47. L. A. Parada, J. J. Roix, T. Misteli, An uncertainty principle in chromosome positioning. *Trends in cell biology* **13**, 393-396 (2003).

48. F. Grasser *et al.*, Replication-timing-correlated spatial chromatin arrangements in cancer and in primate interphase nuclei. *Journal of cell science* **121**, 1876-1886 (2008).

49. T. Takizawa, K. J. Meaburn, T. Misteli, The meaning of gene positioning. *Cell* **135**, 9-13 (2008).

50. T. Ryba *et al.*, Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome research* **20**, 761-770 (2010).

51. J. R. Dixon *et al.*, Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331 (2015).

52. J. G. Van Bemmel *et al.*, A network model of the molecular organization of chromatin in Drosophila. *Molecular cell* **49**, 759-771 (2013).

53. E. P. Nora *et al.*, Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381 (2012).

54. T. Sexton *et al.*, Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458-472 (2012).

55. E. Crane *et al.*, Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**, 240 (2015).

56. C. Gómez-Marín *et al.*, Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. *Proceedings of the National Academy of Sciences* **112**, 7542-7547 (2015).

57. T. Mizuguchi, J. Barrowman, S. I. Grewal, Chromosome domain architecture and dynamic organization of the fission yeast genome. *FEBS letters* **589**, 2975-2986 (2015).

58. C. Wang *et al.*, Genome-wide analysis of local chromatin packing in Arabidopsis thaliana. *Genome research* **25**, 246-256 (2015).

59. J. Dekker, E. Heard, Structural and functional diversity of topologically associating domains. *FEBS letters* **589**, 2877-2884 (2015).

60. J. R. Dixon, D. U. Gorkin, B. Ren, Chromatin Domains: The Unit of Chromosome Organization. *Mol Cell* **62**, 668-680 (2016).

61. Y. Zhan *et al.*, Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. *Genome research*, (2017).

62. O. Symmons *et al.*, Functional and topological characteristics of mammalian regulatory domains. *Genome research*, (2014).

63. B. D. Pope *et al.*, Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**, 402 (2014).

64. S. Berlivet *et al.*, Clustering of tissue-specific sub-TADs accompanies the regulation of HoxA genes in developing limbs. *PLoS genetics* **9**, e1004018 (2013).

65. D. Hnisz, D. S. Day, R. A. Young, Insulated neighborhoods: structural and functional units of mammalian gene control. *Cell* **167**, 1188-1200 (2016).

66. H. Bornfleth, P. Edelmann, D. Zink, T. Cremer, C. Cremer, Quantitative motion analysis of subchromosomal foci in living cells using four-dimensional microscopy. *Biophysical journal* **77**, 2871-2886 (1999).

67. W. Marshall *et al.*, Interphase chromosomes undergo constrained diffusional motion in living cells. *Current Biology* **7**, 930-939 (1997).

68. F. Jin *et al.*, A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-294 (2013).

69. P. H. L. Krijger *et al.*, Cell-of-origin-specific 3D genome structure acquired during somatic cell reprogramming. *Cell Stem Cell* **18**, 597-610 (2016).

70. G. Li *et al.*, Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84-98 (2012).

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

72. S. Schoenfelder *et al.*, The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome research*, (2015).

73. J. M. Dowen *et al.*, Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**, 374-387 (2014).

74. X. Ji *et al.*, 3D chromosome regulatory landscape of human pluripotent cells. *Cell stem cell* **18**, 262-275 (2016).

75. A. L. Sanborn *et al.*, Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences* **112**, E6456-E6465 (2015).

76. G. Fudenberg *et al.*, Formation of chromosomal domains by loop extrusion. *Cell reports* **15**, 2038-2049 (2016).

77. M. H. Nichols, V. G. Corces, A CTCF code for 3D genome architecture. *Cell* **162**, 703-705 (2015).

78. K. Nasmyth, Disseminating the genome: joining, resolving, and separating sister chromatids during mitosis and meiosis. *Annual review of genetics* **35**, 673-745 (2001).

79. E. Alipour, J. F. Marko, Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic acids research* **40**, 11202-11212 (2012).

80. C. H. Haering *et al.*, Structure and stability of cohesin's Smc1-kleisin interaction. *Molecular cell* **15**, 951-964 (2004).

81. J. Stigler, G. Ö. Çamdere, D. E. Koshland, E. C. Greene, Single-molecule imaging reveals a collapsed conformational state for DNA-bound cohesin. *Cell reports* **15**, 988-998 (2016).

82. I. F. Davidson *et al.*, Rapid movement and transcriptional re‐localization of human cohesin on DNA. *The EMBO journal* **35**, 2671-2685 (2016).

83. M. Kanke, E. Tahara, T. Nishiyama, Cohesin acetylation and Wapl‐Pds5 oppositely regulate translocation of cohesin along DNA. *The EMBO journal* **35**, 2686-2698 (2016).

84. Y. Murayama, C. P. Samora, Y. Kurokawa, H. Iwasaki, F. Uhlmann, Establishment of DNA-DNA interactions by the cohesin ring. *Cell* **172**, 465-477. e415 (2018).

85. S. S. P. Rao *et al.*, Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305-320 e324 (2017).

86. E. P. Nora *et al.*, Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930-944 e922 (2017).

87. L. Vian *et al.*, The Energetics and Physiological Impact of Cohesin Extrusion. *Cell* **173**, 1165-1178. e1120 (2018).

88. M. Ocampo-Hafalla, S. Muñoz, C. P. Samora, F. Uhlmann, Evidence for cohesin sliding along budding yeast chromosomes. *Open biology* **6**, 150178 (2016).

89. C. Bausch *et al.*, Transcription alters chromosomal locations of cohesin in Saccharomyces cerevisiae. *Molecular and cellular biology* **27**, 8522-8532 (2007).

90. G. A. Busslinger *et al.*, Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl. *Nature* **544**, 503 (2017).

91. D. Racko, F. Benedetti, J. Dorier, A. Stasiak, Transcription-induced supercoiling as the driving force of chromatin loop extrusion during formation of TADs in interphase chromosomes. *Nucleic Acids Res*, (2017).

92. I. Jonkers, J. T. Lis, Getting up to speed with transcription elongation by RNA polymerase II. *Nature reviews Molecular cell biology* **16**, 167 (2015).

93. N. Kubo *et al.*, (2017).

94. A. S. Hansen, I. Pustova, C. Cattoglio, R. Tjian, X. Darzacq, CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *Elife* **6**, e25776 (2017).

95. W. Schwarzer *et al.*, Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51-56 (2017).

96. A. S. Hansen *et al.*, An RNA-binding region regulates CTCF clustering and chromatin looping. *bioRxiv*, 495432 (2018).

97. J. Gassler *et al.*, A mechanism of cohesin‐dependent loop extrusion organizes zygotic genome architecture. *The EMBO journal* **36**, 3600-3618 (2017).

98. D. Hnisz, K. Shrinivas, R. A. Young, A. K. Chakraborty, P. A. Sharp, A phase separation model for transcriptional control. *Cell* **169**, 13-23 (2017).

99. A. G. Larson *et al.*, Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. *Nature* **547**, 236 (2017).

100. H. Lu *et al.*, Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II. *Nature*, 1 (2018).

101. D. G. Lupiáñez, M. Spielmann, S. Mundlos, Breaking TADs: how alterations of chromatin domains result in disease. *Trends in Genetics* **32**, 225-237 (2016).

102. D. G. Lupiáñez *et al.*, Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012-1025 (2015).

103. J. M. Moore *et al.*, Loss of maternal CTCF is associated with peri-implantation lethality of Ctcf null embryos. *PLoS One* **7**, e34915 (2012).

104. L.-B. Wan *et al.*, Maternal depletion of CTCF reveals multiple functions during oocyte and preimplantation embryo development. *Development* **135**, 2729-2738 (2008).

105. A. Gregor *et al.*, De novo mutations in the genome organizer CTCF cause intellectual disability. *The American Journal of Human Genetics* **93**, 124-131 (2013).

106. C. J. Kemp *et al.*, CTCF haploinsufficiency destabilizes DNA methylation and predisposes to cancer. *Cell reports* **7**, 1020-1029 (2014).

107. R. Katainen *et al.*, CTCF/cohesin-binding sites are frequently mutated in cancer. *Nature genetics* **47**, 818 (2015).

108. T. B. K. Le, M. V. Imakaev, L. A. Mirny, M. T. Laub, High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome. *Science* **342**, 731-734 (2013).

109. M. J. Rowley *et al.*, Evolutionarily Conserved Principles Predict 3D Chromatin Organization. *Mol Cell* **67**, 837-852 e837 (2017).

110. M. J. Rowley, V. G. Corces, Organizational principles of 3D genome architecture. *Nature Reviews Genetics*, 1 (2018).

111. Q. Szabo *et al.*, TADs are 3D structural units of higher-order chromosome organization in Drosophila. *Science Advances* **4**, eaar8082 (2018).

2 Chemical-crosslinking Assisted Proximity Capture (CAP-C) for chromatin structure study

2.1 Introduction

In eukaryotes, DNA exists as compact, knot-free chromosomes in the nucleus. How chromosomes organize in the nucleus can influence transcription, DNA replication and other nuclear processes (*1, 2*). Chromosome conformation capture approaches (such as 3C and Hi-C) (*3-6*) have been widely used to study chromatin organization in different species and cell types. These methods and their variations employ formaldehyde-mediated crosslinking followed by *in-situ* enzymatic and proximity ligation to infer spatial relationships between genomic loci. They have been instrumental in elucidating the principles of chromatin folding. Studies using these techniques have confirmed the existence of multiple layers of genome organization such as chromosome territories, compartments (*6*), topologically associating domains (TADs) (*7*), sub-TADs (*8*), insulated neighborhoods (*9*), and chromatin loops (*10*).

Some of these chromatin features are strongly dependent on the resolution of the current 3C technologies. The sub-megabase scale chromosomal domains termed TADs (median: 880 Kb) identified in previous low-resolution Hi-C maps of mammalian cells (*7*), are in stark contrast to the contact domains (median: 185Kb) obtained from high-resolution Hi-C maps (*10*). It is still unclear whether all domains form hierarchies with nested domains that are subsequently revealed as map-resolution increases, or whether a series of small domains with irreducible length identified in a high-resolution map co-aggregate and establish a large domain in low-resolution maps. Currently, only a handful of high resolution Hi-C datasets for mammalian mouse and human genomes (*10, 11*), with map resolutions around 1 Kb, are available to address these questions. Mechanisms leading to domain formation are only just starting to be elucidated (*12-15*). Hence, there has been

a concerted effort to push past the 1 Kb resolution limit, such as by fragmenting the genome into smaller uniform units (*16-18*). However, a recurring limitation of 3C type approaches is partial digestion. All of the current 3C methods rely on formaldehyde-mediated crosslinking, which creates extensive covalent linkages of protein-protein and protein-DNA in chromatin. These crosslinks can mask certain restriction sites and prevent their full digestion. The ligation of partially digested fragments leads to an imprecise inference of their actual genomic proximity. New crosslinking strategies, which can ideally expose all potential restriction sites, are required to ubiquitously capture proximal contacts at all length scales (Fig.2.1).

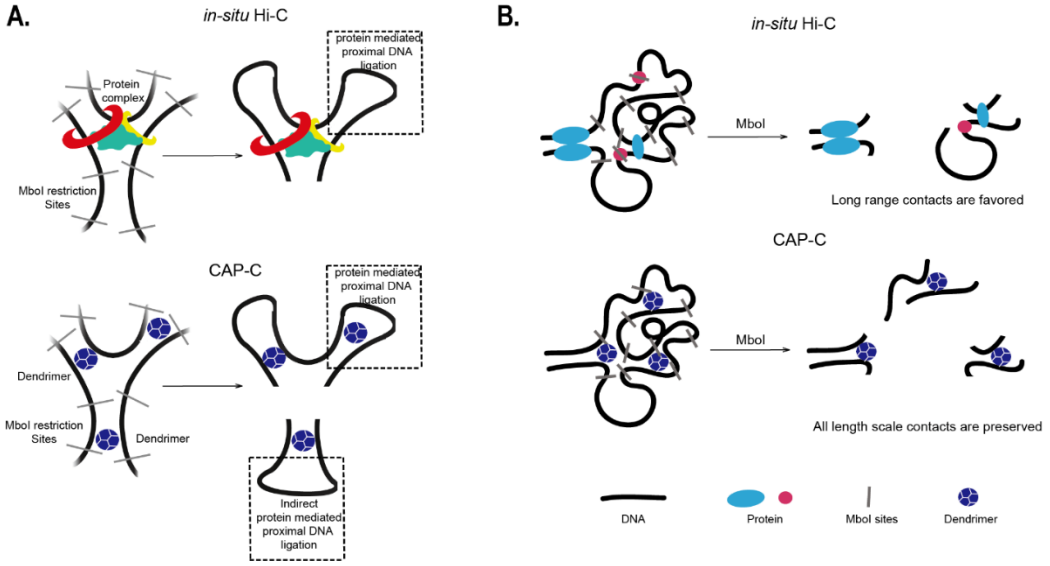


Fig. 2.1 Schematic illustration of two advantages of CAP-C over in-situ Hi-C. (A) Psoralen-functionalized PAMAM dendrimers (blue balls) not only crosslink to proximal DNA-DNA (black line) contacts mediated by direct protein-mediated interactions, but also crosslinks DNAs in close proximity without direct protein binding (indirect). Potential MboI restriction sites on genomic DNA are shown as grey lines. (**B**) In a crowded nuclear environment, binding of large protein complexes masks certain restriction sites and prevents them from being fully fragmented through restriction enzyme digestion in Hi-C procedures (Hi-C relies on protein-mediated crosslinking). Partial digestion leads to longer ligation products and reduced resolution in Hi-C procedures. The dendrimer-mediated crosslinking avoids this shortcoming. The proximal DNA loci are covalently crosslinked to the same dendrimer. After crosslinking all DNA-bound proteins are stripped away using protease treatment. All potential restriction sites are exposed and can be fully digested followed by ligation, leading to a preservation of chromatin contacts at all length scales.

Furthermore, although general principles describing the spatial conformation of mammalian chromosomes are emerging, critical gaps in our understanding of chromatin structure remain, especially regarding how domains form. An appealing model of loop extrusion was proposed based on numerous results obtained by applying 3C methods, which showed that Ctcf and cohesin loops help to bring distant DNA loci into proximity (*15*). However, this model only explains some of the observations, given that a large proportion of domains at high resolution do not form loops at their boundaries (*10*). In addition, recent studies investigating the consequences of acute cohesin loss (*19, 20*) indicated that two independent mechanisms compact chromatin: i) a cohesin-dependent loop extrusion mechanism compacts chromatin locally; ii) a cohesin-independent mechanism spatially segregates the genome into active and inactive compartments at a smaller scale than previously appreciated. These results strongly implicate an alternative mechanism that correlates the chromatin structure with transcription, which might play an important role in shaping chromatin landscape.

2.2 Result and discussion

2.2.1 Synthesize psoralen functionalized dendrimer family

To establish an approach that captures proximal chromatin contacts at all length scales, we utilized a new type of crosslinker in the form of multifunctional dendrimers (PAMAM) that bear tens of crosslinking groups on the surface of polymer spheres with diameters ranging from 3-9 nm. PAMAM dendrimers are iteratively "grown" off a central core, with a new "generation" of dendrimer being synthesized at each subsequent step. Each generation of PAMAM dendrimer has a characteristic size and can be precisely tuned to control the number of surface amine groups ranging from 16-256 amines (*21*). We used psoralen, which crosslinks to double-stranded DNA

(dsDNA) upon UV irradiation, to functionalize approximately half of the surface amine branches on generation G3, G5 and G7 PAMAM dendrimers, with the diameters of 3.6 nm, 5.4 nm, and 8.1 nm, respectively. The remaining amine branches were masked with acetyl groups, making them inert to cellular interactions (Fig.2.2).

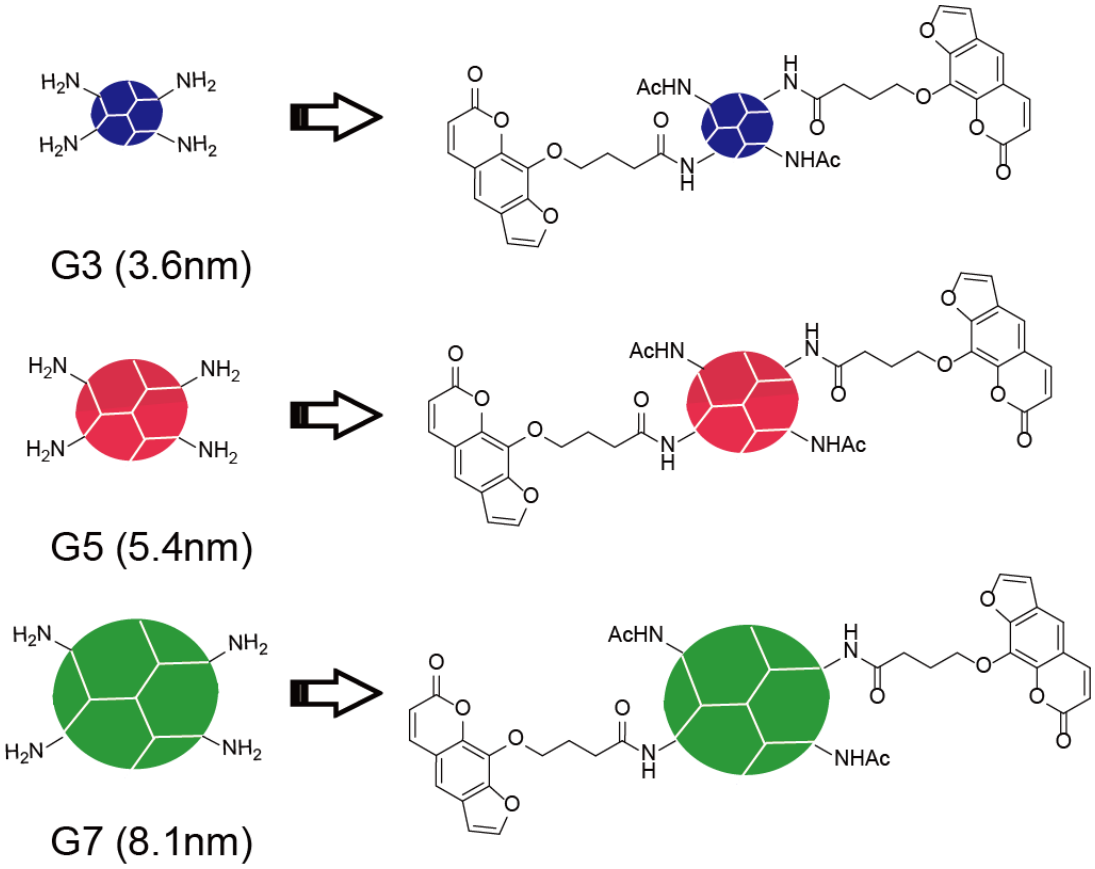


Fig. 2.2 Synthesis scheme of psoralen functionalized dendrimer family Generations of 3, 5 and 7 PAMAM dendrimers with surface amines were shown as balls in blue red and green, respectively. The diameters for each type of dendrimers are shown below. For each generation of dendrimers, half of the terminal amine branches are modified with psoralen while the rest are blocked with acetyl group.

2.2.2 Validation crosslinking between genomic DNA and psoralen functionalized dendrimer in vivo

Psoralen-functionalized dendrimers will be referred to as dendrimers throughout the theis. To confirm dendrimer functionality, we mixed different concentrations (2-25 μM) of G3 dendrimers

with 10 million of mouse embryonic stem cells (mESCs). After different time of UV irradiation, proteins and RNA were digested followed by DNA extraction and purification. We observed efficient crosslinking upon UV exposure as genomic DNA crosslinked with dendrimer will become larger so that the complex will run slower in agarose gel compared to native genomic DNA (Fig.2.3). Thus, these nanometer probes were proved to penetrate crowded chromatin environments and covalently crosslink dsDNAs, but not proteins, that are in proximity to the psoralen groups on the same dendrimer. We refer to this approach as **C**hemical-crosslinking **A**ssisted **P**roximity **C**apture (CAP-C).

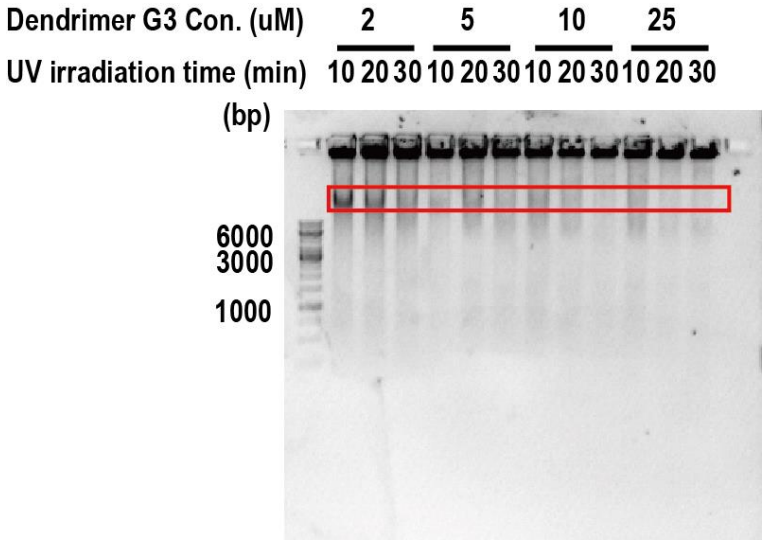


Fig. 2.3 Validation of photo-crosslinking between psoralen-modified PAMAM dendrimers and purified genomic DNA in vivo. Psoralen-functionalized dendrimer G3 was serial diluted to 25, 10, 5, 2 μM and mixed with 10 million of mouse embryonic stem cells (mESCs), respectively. Each mixture was exposed under UV irradiation 10, 20 or 30 minutes. The crosslinked complexes were analyzed on agarose gel. Red box indicates the position of genomic DNA. DNA-dendrimer complexes were characterized as a band shift to the top.

2.2.3 General scheme of CAP-C

To investigate chromatin architecture using CAP-C, we fix cells with formaldehyde to make sure the subsequent application of dendrimers does not perturb native chromosome conformation. We then diffuse dendrimers into the cell nucleus and expose these cells to UV irradiation. The formaldehyde fixing is then reversed, and DNA-bound proteins are removed with protease to expose all DNA motifs, the dendrimer-DNA complexes are subsequently purified with ethanol precipitation. The purified dendrimer-DNA complexes are then subjected to MboI restriction digestion and end-filling with biotin-bearing DNA, followed by ligation under ultra-diluted solution, biotin capture and high-throughput sequencing (Fig. 2.4). Dendrimers with a defined size were used for each CAP-C experiment.

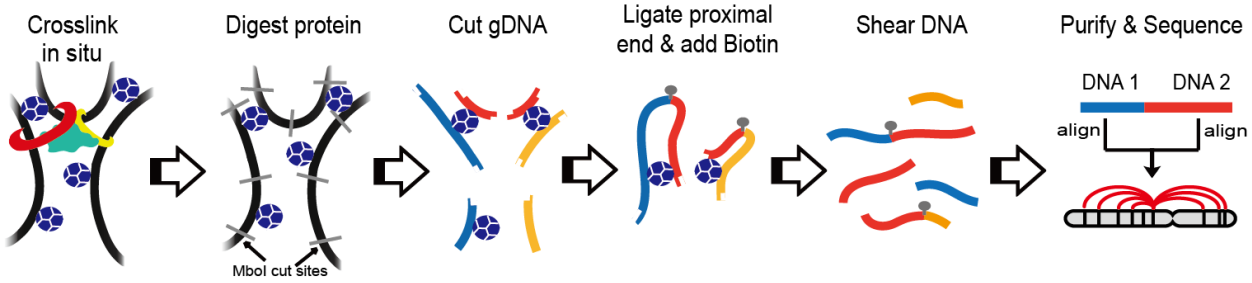


Fig. 2.4 General scheme of CAP-C. Mouse embryonic stem cells (mESs) are treated with formaldehyde to crosslink proteins (Shown in red, green, yellow) with genomic DNA (black strings). Psoralen-modified PAMAM dendrimers with fixed diameter (Shown as blue balls) are diffused into nucleus. DNA in proximity are covalently crosslinked with dendrimers under UV irradiation. Proteins are digested with protease and dendrimer-DNA complexed are purified. The purified complexes, without DNA-bound proteins, are then subjected to MboI digestion and end-filling with biotin-bearing DNA, followed by proximal ligation, biotin capture and high throughput sequencing.

2.2.4 Validation of CAP-C in capturing chromatin conformation

We validate of CAP-C strategy by conducting chromatin conformation capture in different conditions. For cells without addition of dendrimer or without UV irradiation, chromatin conformation capture was failed as we observed no sufficient chromatin interactions on contact matrix.

On the other hand, cells treated with 25uM dendrimer G3 and exposed by UV irradiation for 30 min display similar patterns on contact matrix compared to in situ Hi-C. In addition, we generated relative frequency *vs* genomic distance curve from the above 4 conditions and discovered that long-range chromatin contacts capture required both the addition of dendrimer and UV irradiation. From the curve, we also noted that CAP-C was able to capture more short-range chromatin interactions spanning a distance from 1000-20000bp compared to in situ Hi-C (Fig. 2.5).

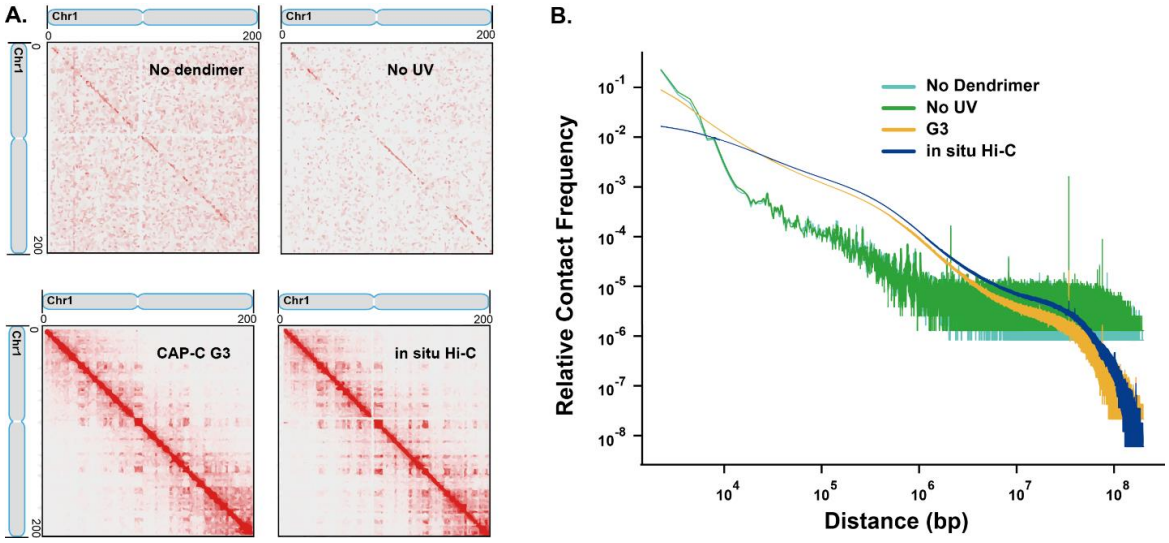


Fig. 2.5 Validation of the CAP-C method. Long-range contacts were lost without initiating the chemical-assisted UV crosslinking. (**A**) 500 Kb resolution contact maps of chromosome 1 are shown for *in situ* Hi-C under different conditions of "G3 CAP-C". *No dendrimer:* experiment was performed without introducing psoralen functionalized dendrimer; *No UV:* experiment was performed in the presence of psoralen-modified G3 dendrimer but without UV irradiation to crosslink chromatin; CAP-C G3: CAP-C was performed with addition of psoralen-modified G3 dendrimer; *in situ* Hi-C was performed as previously described. (**B**) Relative frequency *vs* genomic distance curve generated from low resolution maps show that "No dendrimer" and "No UV" conditions fail to crosslink chromatin at long-range genomic distances over 10 kB while "CAP-C G3" showed similar long-range patterns typically seen in *in situ* Hi-C.

2.2.5 CAP-C features in capturing short range chromatin contacts due to the ability of thorough chromatin digestion with restriction enzyme.

We next investigate why CAP-C features in capturing short range chromatin contacts. We performed CAP-C with or without protein removal before using restriction enzyme to fragment DNA. Proteinase-K negative maps using G3 dendrimers showed a relative contact frequency *vs* distance plot that recapitulated *in-situ* Hi-C, with significantly fewer short-range interactions than G3 CAP-C maps (with proteinase-K digestion to expose most restriction sites). (Fig. 2.6) These data strongly suggest that protein removal promotes short-length proximal chromatin contact ligation.

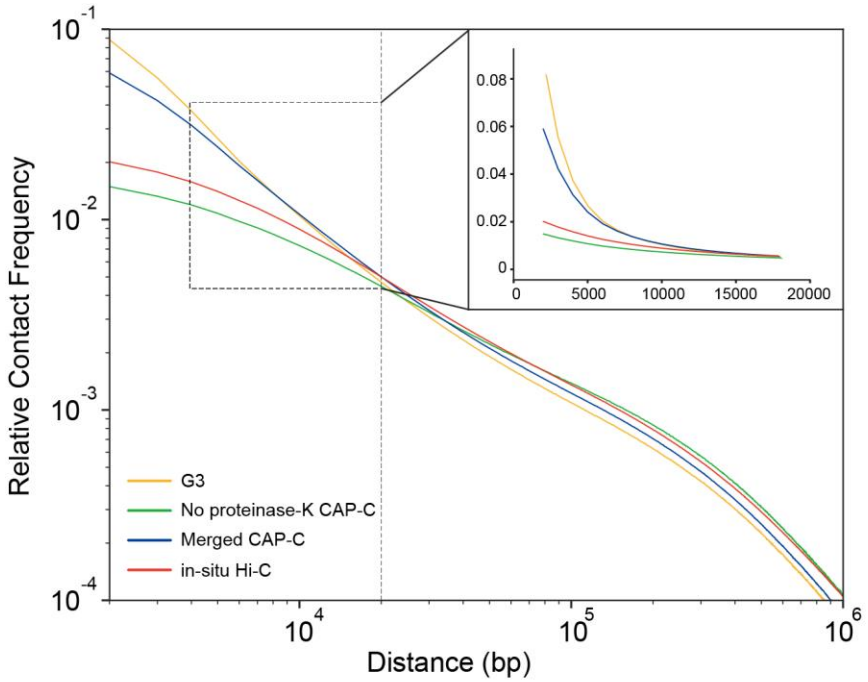


Fig. 2.6 Short length contacts could not be enriched in CAP-C without protease treatment. Relative contact frequency *vs* distance was plotted for 10 μM G3 CAP-C, 10 μM G3 CAP-C without proteinase K (no proteinase-K treated CAP-C), merged CAP-C, and *in-situ* Hi-C, respectively. We observed that in the absence of proteinase-K treatment, G3 CAP-C (*green vs yellow*) did not exhibit an enrichment of short-range contacts and was similar to the frequency-distance curve of *in-situ* Hi-C (*green vs red*), suggesting that short-range contacts could not be enriched without protease treatment.

As we hypothesized removal of protein would result in exposure of more accessible sites for restriction enzyme, we designed qPCR probes to test several MboI sites between samples that performed with CAP-C or in situ Hi-C. Overall, we found that CAP-C yielded a higher digestion efficiency over *in-situ* Hi-C (Fig.2.7).

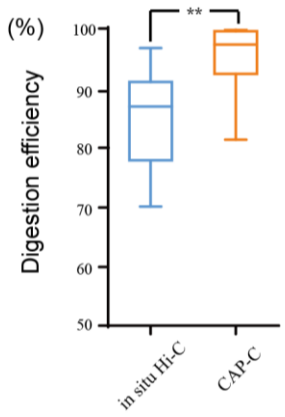


Fig. 2.7 MboI digestion efficiency between CAP-C and in-situ Hi-C. CAP-C shows higher MboI digestion efficiency over *in situ* Hi-C. Primers were designed by mapping to upstream and downstream MboI recognition sites within 200 bp. The digestion efficiency was analyzed by qPCR of MboI fragmented DNA which followed the standard protocol of CAP-C and *in situ* Hi-C.

2.2.6 Characterize deep-sequenced CAP-C contact matrix reveals the method is reproducible with high quality

Next, we turned to compare CAP-C with *in-situ* Hi-C using mESCs. We sequenced a total of 4.24 billion paired reads from six CAP-C libraries, consisting of primary and replicate libraries for each of the G3 (1.44 billion total reads), G5 (1.40 billion total reads) and G7 dendrimers (1.40 billion total reads), as well as a primary and replicate library for *in-situ* Hi-C (2.59 billion total reads). CAP-C datasets were processed employing a similar pipeline used for processing *in-situ* Hi-C libraries, followed by removal of PCR duplicates, uninformative reads, as well as reads with a low mapping quality that strongly indicate non-unique mapping (Table. S1). We also performed strand orientation analysis and removed interactions below 1Kb where read orientation is roughly

equal to +/-1% (Fig. 2.8) (*22*). Those filtered strands are the product of un-ligated DNA fragment or self-ligated DNA fragment.

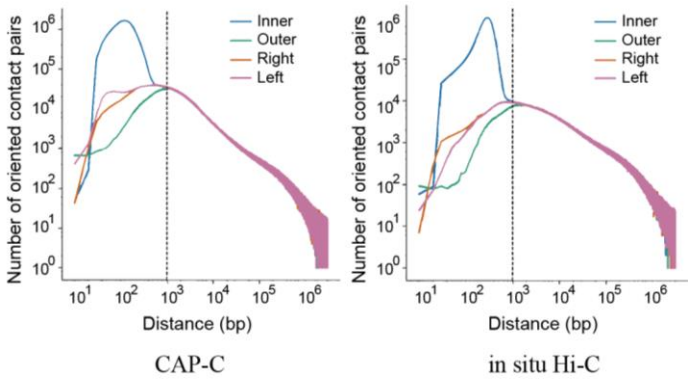


Fig. 2.8 Strand orientation analysis between CAP-C and in-situ Hi-C. Strand orientation analysis revealed that contacts above 1 Kb are legitimate ligation products in both merged CAP-C and *in-situ* Hi-C. "Inner": inward strand configuration; "Outer": outward strand configuration; "Right", "Left": same strand configuration.

The replicates for each experiment were then merged to yield a total of 732, 628, 804 and 2,093 million valid contact pairs for G3, G5, G7 and *in-situ* Hi-C libraries, respectively. Pearson's correlation coefficients between dendrimers were slightly lower than those between replicates of the experiments with same-sized dendrimers, suggesting that dendrimers of different sizes may capture different features of chromatin organization. We merged G3, G5 and G7 libraries and refer to this as the "merged CAP-C". Comparisons between primary and replicate libraries for both merged CAP-C and *in-situ* Hi-C showed that they exhibited high reproducibility at the 100 Kb (CAP-C: $R = 0.9990$ and *in-situ* HiC: $R = 0.9991$) and 25 Kb (CAP-C: $R = 0.9974$ and *in-situ* HiC: $R = 0.9959$) resolution (Fig. 2.9). Proportions of intra-chromosomal contacts were 62.1% for merged CAP-C, and 64.1% for *in-situ* Hi-C.

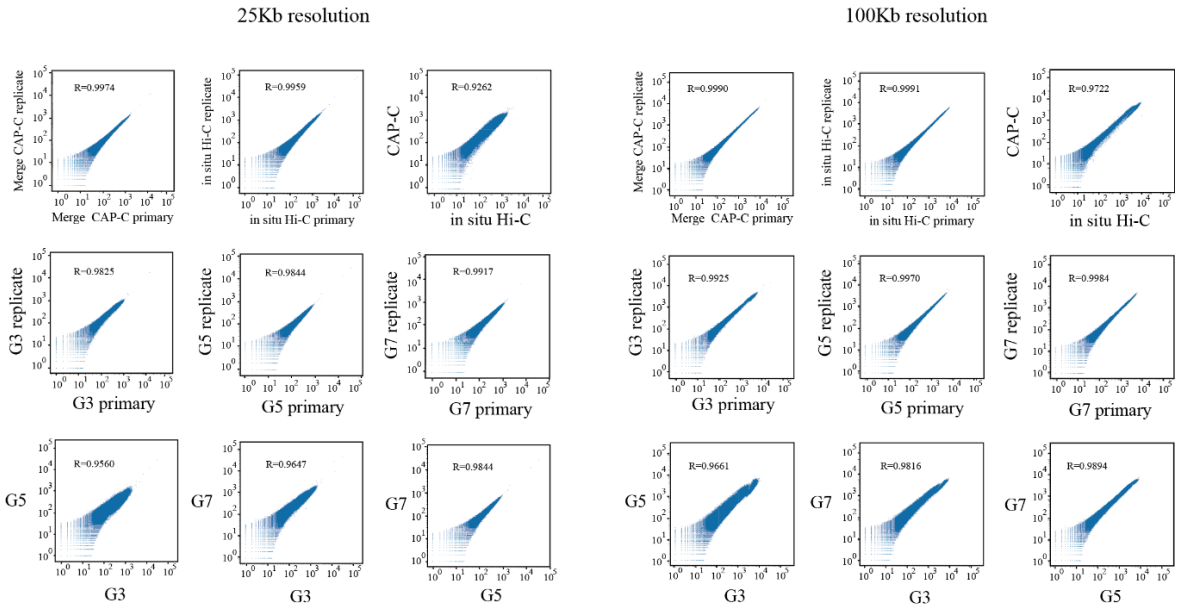


Fig. 2.9. Reproducibility of CAP-C and in-situ Hi-C experiments. Pearson's correlation analyses were performed for contact matrices binned at 25 Kb and 100 Kb resolution and visualized in log-scale. *First row*: Comparisons made between primary and biological replicate of merged CAP-C (*column 1*), *in-situ* Hi-C (*column 2*) followed by one between merged CAP-C and *in-situ* Hi-C (*column 3*). *Second row*: Comparisons made between primary and biological replicates of CAP-C G3 (*column 1*), G5 (*column 2*) and G7 (*column 3*). *Third row*: Cross comparisons between CAP-C experiments using different dendrimers.

To approximate the random ligation rate, we compared trans-interactions between mitochondria and autosomes, which are physically separate prior to crosslinking (*7*). Here, we found that even though CAP-C libraries showed 1.3-fold more mitochondria (cis+trans) interactions, there was an 8-fold enrichment of cis-interactions over *in-situ* Hi-C libraries (Fisher's Exact Test, $P < 0.0001$). Thus, CAP-C exhibits a lower random ligation rate than *in-situ* Hi-C.

2.2.7 CAP-C revealed finer local chromatin structures than in-situ Hi-C due to enrichment of short-ranged proximal chromatin contacts.

We hypothesized that different sized dendrimer crosslinkers will capture distinct spatial relationships at different length scales. Indeed, the smallest dendrimer, G3, strongly crosslinked loci

between 1 to 5 Kb in distance, whereas G5 and G7 dendrimers preferentially crosslinked loci with distances between 5 to 20 Kb. The total chromatin contacts between 1-20 Kb captured by merging all dendrimer data were 2-3 folds greater than for *in-situ* Hi-C (Fig. 2.10A). This gain in short-range interactions was offset by a relative reduction of long-range contacts after 1 Mb, however, contact maps plotted for a series of CAP-C resolutions showed similar chromatin features, such as compartments or TADs, as *in-situ* Hi-C (Fig. 2.10B).

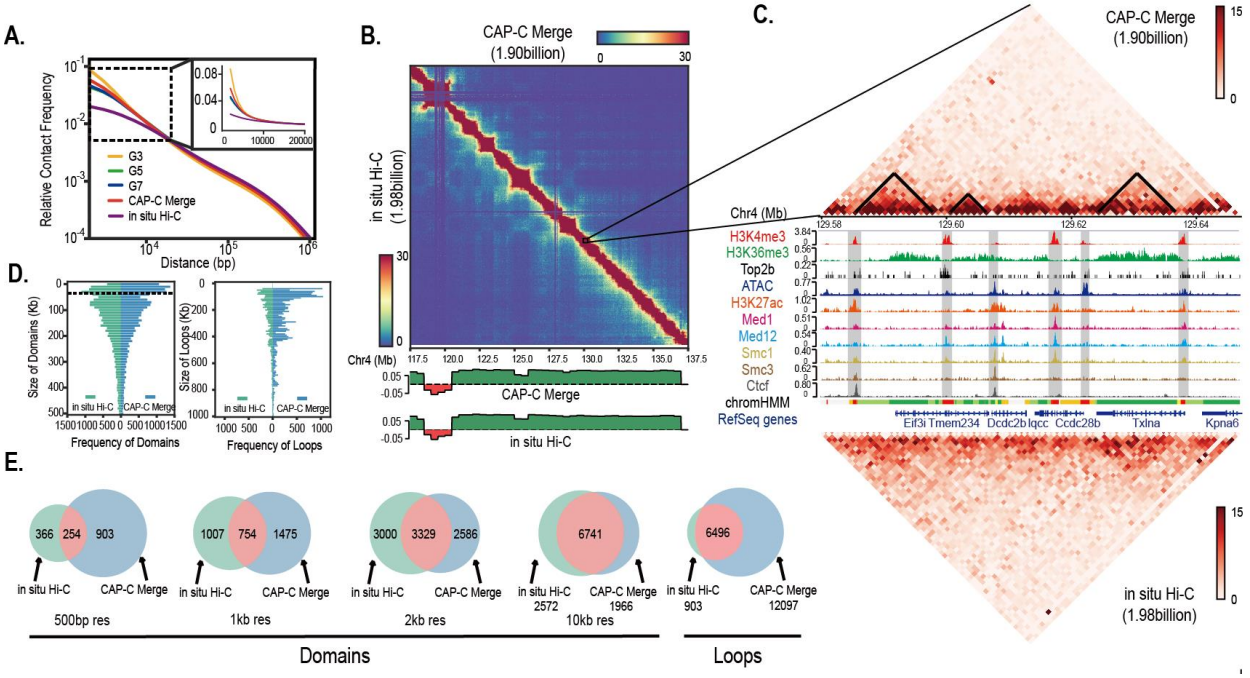


Fig. 2.10 CAP-C revealed higher resolution local chromatin structure compared to in-situ Hi-C at similar sequencing depths. (**A**) Relative contact frequency *vs* genomic distance curve shows CAP-C differentially enriched at short range (1-20kb) chromatin contacts over *in-situ* Hi-C. The zoomed-in window shows approximately 2- to 3-fold enrichment. G3, G5, G7 represent CAP-C performed with psoralen functionalized PAMAM dendrimer generation 3, 5 and 7, respectively (with the diameters of 3.6, 5.4, and 8.1 nm). CAP-C merge represents the merging of data generated from G3, G5, and G7. *In-situ* Hi-C was performed under conditions described previously. (**B**) CAP-C reproduces the same chromatin architecture identified by *in-situ* Hi-C. Top panel: 25 Kb resolution contact map of CAP-C (top right triangle) and *in-situ* Hi-C (bottom left triangle); bottom panel: CAP-C and *in-situ* Hi-C identified similar A, B compartments by PCA at 25 Kb resolution. (**C**) CAP-C reveals folding principle of local chromatin at high resolution. 70 Kb-long regions of CAP-C (top) and *in-situ* Hi-C (bottom) 1kb resolution contact matrixes are shown corresponding to Chr4: 129.58-129.65Mb. Histone modification and ChIP profiles are shown in the middle. Arrowhead detected 3 domains (Black triangle) in CAP-C enveloping Eif3i, Tmem234

Fig. 2.10 continued and Txlna, respectively. Domain boundaries are highly enriched for active promoters, mediators and enhancers. No domains were detected for *in-situ* Hi-C at this resolution. (**D** and **E**) Domains are called by Arrowhead while loops are called by HiCCUPS using high resolution map with the same sequencing depth (CAP-C: 1.90 billion; *in-situ* Hi-C: 1.98 billion). Top left: CAP-C is able to capture more small-sized domains ranging from 5-40 kb in length. Top right: CAP-C had 2-fold more loops called genome-wide. Bottom: Overlap of domains called at different resolution and loops between the two methods. CAP-C could call more domains at high resolution (500bp and 1Kb).

In contrast to higher-order chromatin structures that have been studied extensively by Hi-C, enrichment of short-range CAP-C contacts allowed us to better resolve new features of the genome at shorter length-scales. For comparison, contact maps of merged CAP-C and *in-situ* Hi-C datasets with similar depths (1.90 billion vs 1.98 billion) were plotted over a 70 Kb region (chr4:129.58-129.65Mb) encompassing 6 different genes at 1 Kb resolution. At this resolution, CAP-C maps were clearer and sharper, enabling us to visually inspect promoters, enhancers and actively-transcribed genes when placed alongside a ChromHMM track (*23*). Many of the small triangles with enhanced contact frequency close to the diagonal were observed in CAP-C, and were called as domains by using Arrowhead (*10*) at 1 Kb resolution, which were not distinguishable as domains in *in-situ* Hi-C maps with a similar sequencing depth. (Fig. 2.10C) Because contact maps consist of an ensemble of individual chromatin conformations across millions of cells, we do not assume that the genome can be partitioned into non-overlapping intervals. Thus, we called domains using Arrowhead at a series of resolutions (500 bp, 1 Kb, 2 Kb, 5 Kb and 10 Kb) and merged the call sets (nested and non-nested) into a unique but possibly overlapping set of domains. Identical domains were merged, and domains with similar boundaries were removed based on the Euclidean distance criteria of min (0.2*shortest-length, 50000). Compared to *in-situ* Hi-C maps, CAP-C maps called 1.5-fold more domains with sizes of 5 to 40 Kb, but 1.2-fold fewer domains with sizes greater than 40 Kb (Fig. 2.10D, left). Boundaries of these smaller-sized domains (5-40 Kb) called

in CAP-C tend to strongly overlap with active promoters and enhancers. In contrast, the smaller-sized domains called in *in-situ* Hi-C tend to overlap with heterochromatin. At similar sequencing depths, high-resolution peak calling using HiCCUPs (*10*) yielded more peaks (2.5-fold) with merged CAP-C contact maps than with *in-situ* Hi-C libraries. (Fig.2.10E). Proportionally, there was a 1.4-fold enrichment of peaks from CAP-C that were less than 100 Kb in size than peaks from *in-situ* Hi-C (Fisher's Exact Test, $P < 0.0001$) (Fig. 2.10, right). 87.7% (6,496) and 61.9% (7,344) of *in-situ* Hi-C and HiChIP peaks (*24*) were concordant with merged CAP-C, suggesting that peak calling was more sensitive in CAP-C than *in-situ* Hi-C.

2.2.8 Higher signal to noise ratio in CAP-C than in situ Hi-C results in more loops being identified.

Meta-analyses performed on short (100-200 Kb) and long (300-500 Kb) concordant peaks between CAP-C and *in-situ* Hi-C suggested that even though depth-normalized signal values (FPM) at the foci were similar between maps, a faster decay in mean long-range contacts between the two anchors decreases the mean lower-left background values in CAP-C. This effect significantly increases the signal-to-noise ratio and consequently increases the number of peaks called at a constant threshold (Fig. 2.11).

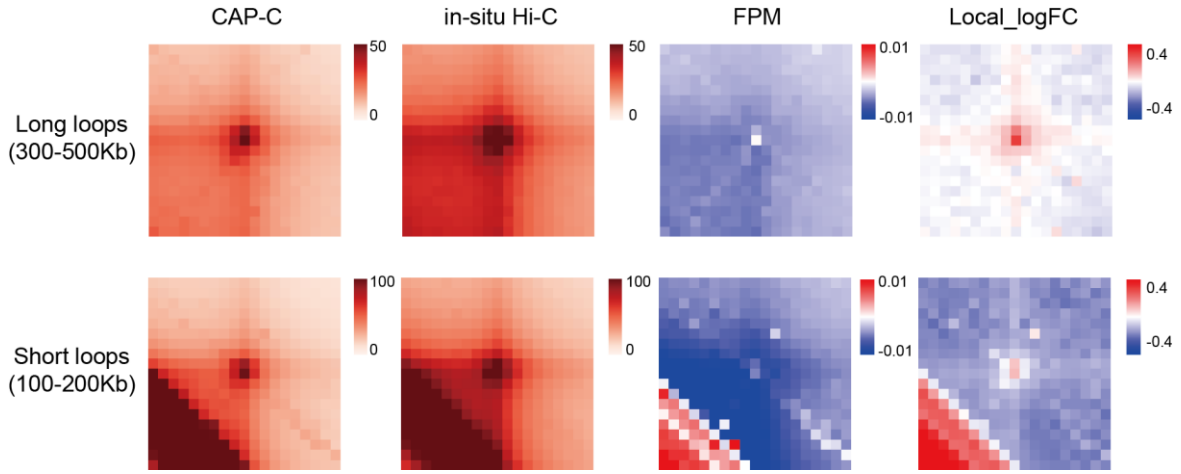


Fig. 2.11 CAP-C shows a higher signal to noise ratio around loop anchors over in-situ Hi-C. Meta-analysis of concordant peak calls was performed to explain why CAP-C shows a 3-fold increase in peak calls as well as an enrichment of peaks spanning short distances. When normalized for sequencing depth (FPM), the mean delta signal between the focal center of CAP-C and *in-situ* Hi-C maps is close to zero (*column 3*). However, normalizing the FPM value by the mean value of the local region, and comparing this local-normalized signal between maps (logFC) suggest a higher signal to background ratio around loop anchors in CAP-C than *in-situ* Hi-C (*column 4*). Peaks closer to the diagonal (*bottom row*), which are harder to call, also show higher enrichment when classified by the span of their genomic distance.

Indeed, contacts around loop anchors are expected to be low, as polymer models predict that long idealistic loops should not exhibit contacts anywhere except at the anchors where they meet (*15*, *25*), suggesting that CAP-C identifies loops better than *in-situ* Hi-C. Further, 76.3% of unique Ctcf motifs were in the convergent orientation (Fig. 2.12), which is similar to results reported previously (*10*) and validates the reliability of peaks called in CAP-C.

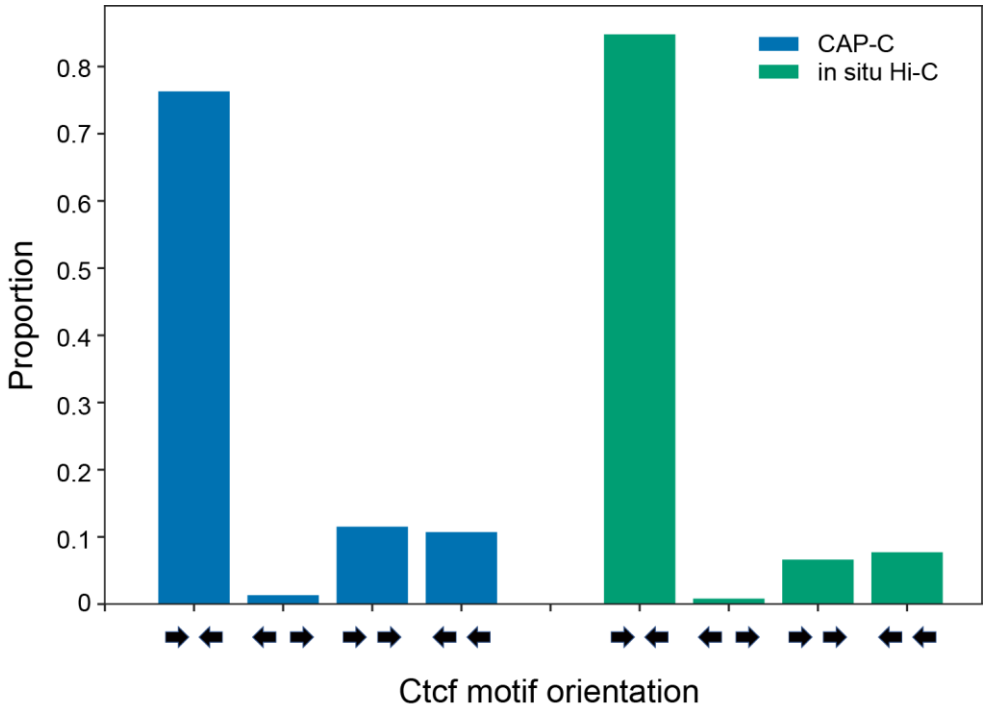


Fig. 2.12 Ctcf motif orientation analysis of loops in CAP-C and in-situ Hi-C. 76.3% and 84.8% of unique Ctcf motifs for CAP-C and *in-situ* Hi-C were in the convergent orientation, suggesting either a slightly higher false positive rate of CAP-C or that certain peaks, i.e. CTCF-cohesin independent ones, may not be conforming to the rules of the loop-extrusion model.

2.2.9 Different sized dendrimers probe different chromatin compartments

Different sized dendrimers might also access and probe distinct regions of chromatin compaction. This would be revealed by dendrimer size-dependent enrichment of interactions in distinct regions. Using principal component analysis, we determined the eigenvector with the highest eigenvalue using the pixel values of each G3, G5 and G7 contact maps and plotted a 2D map which we named as "dendrimer map" based on the eigenvector values of the 1st principal component. At multiple resolutions (500 Kb, 100 Kb, 10 Kb and 5 Kb), the 1st principal component tended to explain 90-95% of the variance instead of 50% for random contact map. Most importantly, these

"dendrimer maps" showed bifurcation similar to that of compartment intervals identified previously (*6*) (Fig. 2.13, A to D). Statistical analysis of low resolution (100 Kb) maps also yielded similar insights (see Materials and Methods). In low resolution maps, we observed differential separation along compartment intervals best represented by checkerboard patterns. Principal component loadings suggested that G5 and G7 dendrimers mostly detect interactions within open configuration chromatin, whereas G3 dendrimers identify more interactions within closed configuration. Thus, we hypothesized that the "dendrimer map" we produced could reflect A/B compartment identified by Hi-C; G5 and G7 dendrimers could capture more contacts within compartment A while G3 dendrimers identify more interactions within compartment B.

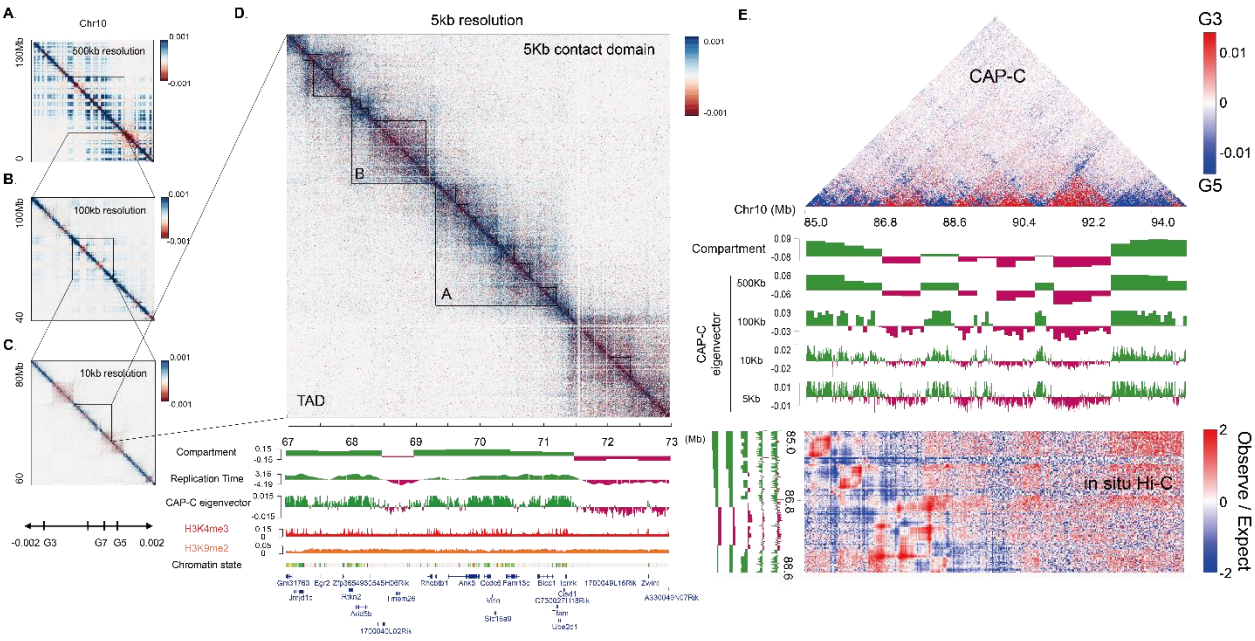


Fig. 2.13 Chromatin contacts detected by G5 and G7 dendrimers are enriched for compartment A whereas those detected by G3 dendrimers are enriched for compartment B. (**A** to **D**) The eigenvector with the highest eigenvalue (or a rescaled version commonly referred to as the proportion of explained variance ranges between 90 to 95%) calculated for a 3 by N*(N+1)/2 matrix (where N is the number of loci across a specified resolution) using principal component analysis yields a CAP-C map that shows a bifurcated separation that is similar to compartment intervals. Principal component loadings (bottom left arrow bars) suggest that G5 and G7 dendrimers contribute most to an open configuration while G3 dendrimer to a closed configuration. (A to C) In low-resolution maps, we observe checkerboard patterns, usually associated with compartments, which are also associated with the enrichment of specific dendrimers. (D) We examine a close-up (5 kb res) of the CAP-C map to reveal a fine level of compartment detail. mESC TADs (*7*) annotated in the lower-left triangle and Arrowhead contact domains annotated on the upper-right triangle reveals that the relationship between compartments intervals and domains (or domain boundaries) is complex. Eigenvector for compartment A was colored in green and compartment B in pink. (**E**) A G3 versus G5 delta map shows the one-to-one relationship with eigenvector compartments derived from (*6, 10*). CAP-C eigenvectors computed using the row-sums instead of pixels in (A to D) gives us increasing levels of high-resolution eigenvectors showing compartment intervals which are tens of kilobases in length, and missed in the low-resolution compartments calculated using the Pearson matrix (*6*). We validate our CAP-C eigenvectors by matching it with the in-situ Hi-C O/E 25 Kb resolution map to show the co-localization of compartments (in red). Eigenvector for compartment A was colored in green and compartment B in pink.

To validate above hypothesis, we produced “CAP-C eigenvector” similar to the eigenvector constructed previously in determining compartments by performing principal component analysis

on the row sums, instead of the pixels, of all three dendrimer contact maps, and arbitrarily assigned positive values to regions which are gene-rich. Indeed, our "CAP-C eigenvectors" showed good positive correlations with compartment intervals derived from the eigenvector analysis on *in-situ* Hi-C at 500 Kb resolution (Pearson's R = 0.861), and replication timing data from RepliSeq experiments in mESC (Pearson's R = 0.850) (*26*) as well as moderately negative correlation with H3K9me2 ChIP-Seq (Pearson's R = -0.329) (*27*), a histone modification mark for constitutive heterochromatin in mESC (Fig. 2.13D, bottom tracks). Moreover, we obtained "CAP-C eigenvectors" in a series of resolution and discovered smaller compartment intervals that are tens of kilobases in length, which were missed in the compartments calculated using the Pearson matrix of low resolution *in-situ* Hi-C maps (Fig. 2.13E). Given that G3 and G5 are sufficient to describe a distinction between compartment A and B, we plotted a delta map (difference in FPM) between G3 and G5 and observed the co-localization of compartment intervals with similar types (A/B) compared with 25 Kb resolution O/E maps of *in-situ* Hi-C, further proving that "CAP-C eigenvector" values are reflective of compartments.

We next inspected the "dendrimer maps" at the 5 Kb resolution to reveal additional compartment details that are missed in previous low resolution Hi-C experiments (Fig. 2.13D). When associating with chromatin states, we observed heterochromatin with negative CAP-C eigenvalue interspersed between gene-rich A compartments while smaller active open chromatin region with positive CAP-C eigenvalue were embedded in compartment B intervals. Such features were not detected in low-resolution Hi-C maps, indicating that CAP-C captures finer details of chromatin conformation. Moreover, overlapping mESC TAD annotations (Fig. 2.13D, bottom left matrix) (*7*) and high-resolution (5 Kb) contact domains (Fig. 2.13D, upper right matrix) revealed a complex relationship between compartments and domains. The smaller compartment intervals revealed in

CAP-C is further supported by recent studies using a variety of experimental techniques (*12*) and newer computational methods (*20*).

In summary, the above analyses confirmed that smaller G3 dendrimers preferentially crosslink tightly packed heterochromatin in B compartments, whereas the larger G5 and G7 dendrimers tend to capture chromatin contacts in the open and gene-rich compartments (Fig. 2.14). Thus, different sized dendrimers enrich in different regions of the genome. These fixed-size probes could be used as nanometer-scale molecular rulers to infer physical distances among different genomic loci.

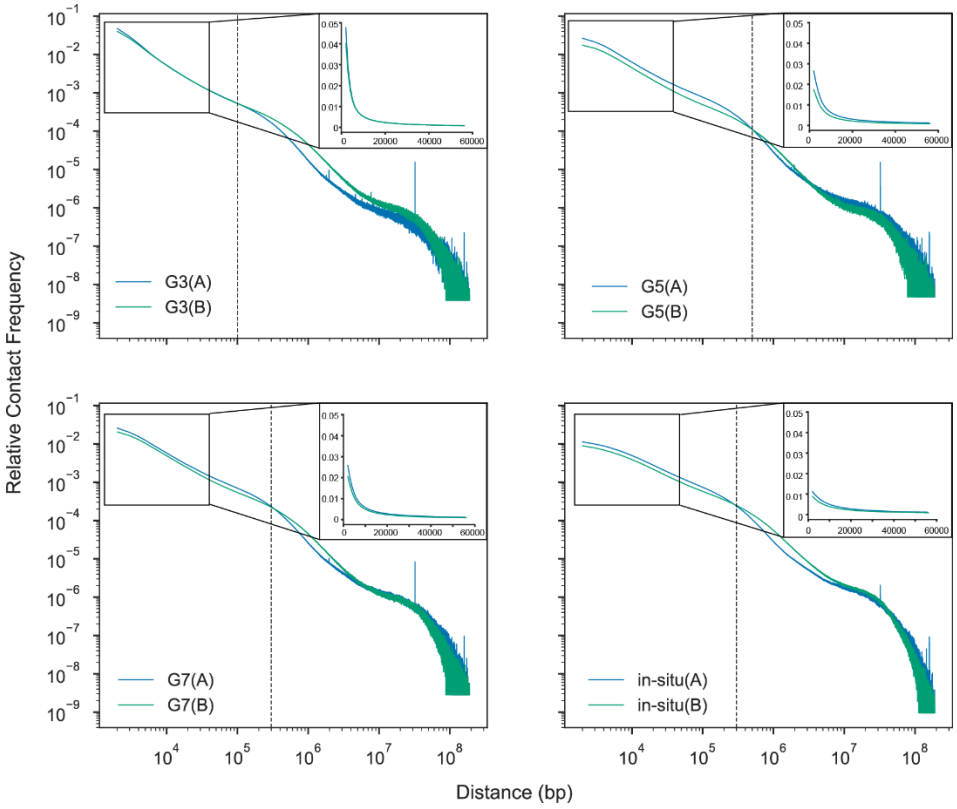


Fig. 2.14 Comparison of contact distributions in different compartments. Relative contact frequency of compartment A and B vs distance was plotted for G3; G5; G7 and *in-situ* Hi-C. As expected in Fig 2B, there was a stronger enrichment of G3 contacts in compartment B after 100 Kb in distance, and a strong enrichment of G5 and G7 contacts in compartment A below 300 Kb in distance. This suggests that large regions of compartment B are in closer contact (G3) than compartment A while smaller region in compartment A have a greater physical distance (more open) than compartment B.

2.2.10 Two types of chromatin domains with different boundary properties

Given that our dendrimer maps showed high correlation between transcription and genome segregation, we investigated how transcription affects the formation of the contact domains we discovered. Recent studies using biophysical models have proposed different mechanisms to explain the self-associating and insulating properties of chromosomal domains in prokaryotes as well as in mammals (*15, 28, 29*). In model organisms such as *C. crescentus* and *S. pombe*, which lack Ctcf, polymer models attribute transcription-induced supercoiling as the force responsible for conformational changes in the form of writhes termed plectonemes (*17, 30, 31*). Boundaries of these domains, generically termed chromosomal interacting domains (CIDs), span the transcriptional start sites of active genes. On the contrary, the detection of TADs enriched with Ctcf at its boundaries in low-resolution maps, followed by the identification of Ctcf-cohesin-mediated loops and loop-domains in high-resolution maps, suggested that loop extrusion might be responsible for chromatin organization in mammals (*7, 10, 20*). However, the loop extrusion model may not explain the self-association property in large TADs unless supercoiling is taken into account (*25*). Hence, it is not entirely clear whether chromatin loop domains form in mammals exclusively via the loop-extrusion model, or whether multiple mechanisms underlie loop domain formation. To further complicate matters, only 30% of our high-resolution contact domains show loops at the corners of loop-domains (*20*) and 65% of the same contact domains overlap Ctcf (+/-10 kB), implying that not all Ctcf-enriched boundaries form loops. In our high-resolution maps, we noticed that a substantial proportion of contact domains called at high resolution revealed boundaries starting close to the promoters of short active protein-coding genes, which either terminate at their own transcription end sites (TSS), or half-way through the gene body of another gene.

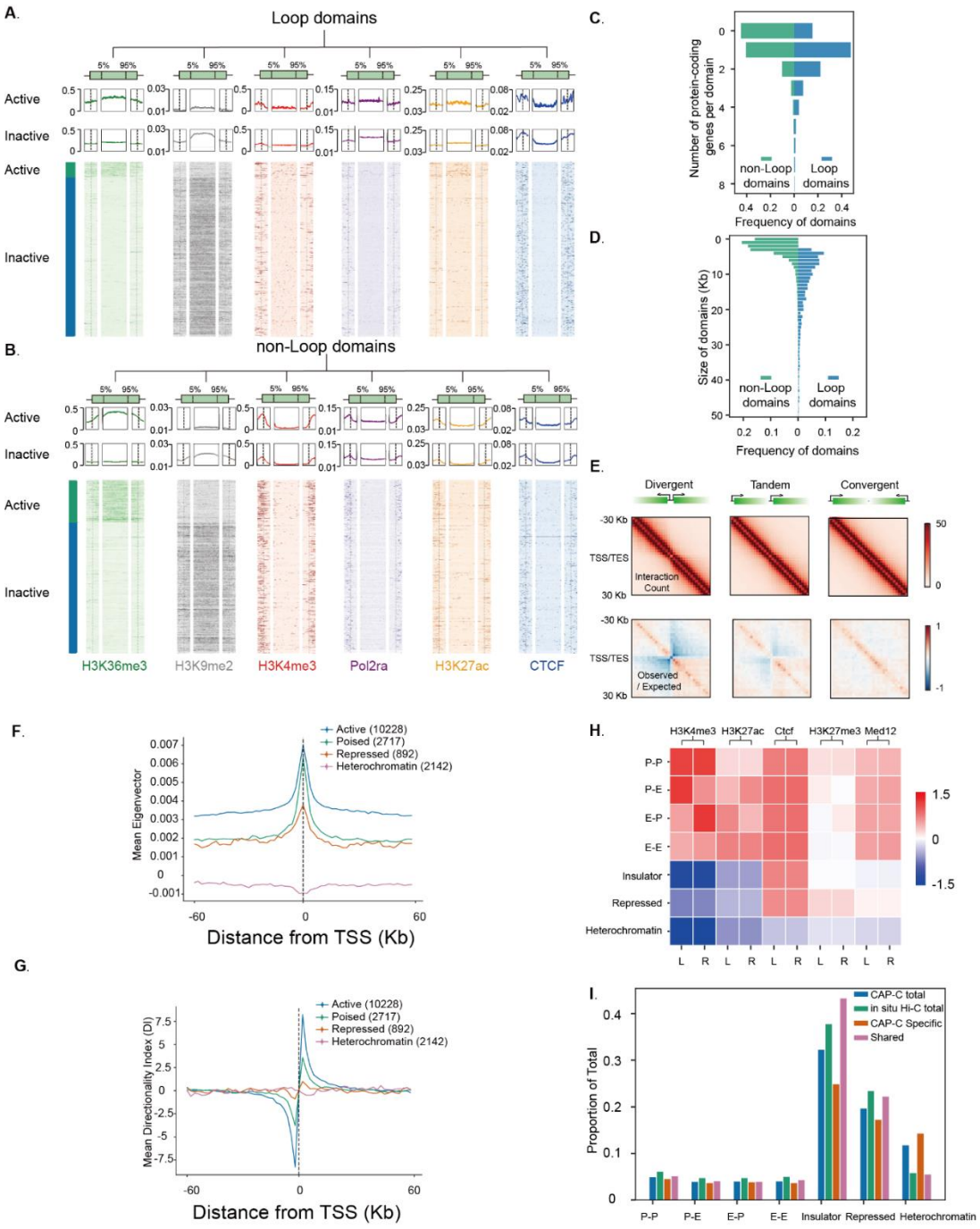


Fig. 2.15. Loop and non-loop domains show different boundary properties. (**A** and **B**) Meta-domain analysis shows two independent mechanisms responsible for domain formation. Segregating domains (called using high-resolution maps) into loop and non-loop domains indicate that boundaries of non-loop domains are enriched in RNA-polymerase binding as well as promoter and enhancer status while loop domains are enriched in CTCF binding. (**C** and **D**) Comparison of length distribution between loop domains and non-loop domains shows that non-loop domains are generally shorter and encompass a single protein-coding gene or not

Fig. 2.15 continued (i.e., domains spanning promoter-enhancer boundaries) while loop domains are longer and have one or more protein-coding genes per domains. (**E**) Plectoneme-free region causes strong separation and leads to domain formation**.** Pairs of protein-coding genes were classified by transcription directions oriented in divergent, tandem and convergent manner. Top panels show interaction counts for the three classes. Bottom panels show the same data expressed as the log2 ratio of observed interactions divided by expected interactions for a given genomic distance. Boundaries were strongest at divergent pairs. (**F** and **G**) By classifying promoters based on chromHMM states and associating them with both CAP-C eigenvector and DI calculated at 2 Kb resolution, we observed that the presence of RNA polymerase at the transcription start site and its increasing levels of elongation is strongly associated with openness and the strength of boundaries. (**H** and **I**) CAP-C captures more loops which are not functionally different from *in-situ* Hi-C. (H) Loops are classified by using histone modification and transcription factor features based on their upstream (L) and downstream (R) anchors around a +/- 5 Kb region. (I) Based on these states, the bar chart revealed that the proportion of different classes of loops are similar between CAP-C and *in-situ* HiC.

The boundaries of domains starting at active promoter regions have been previously characterized in *S. cerevisiae* and recently observed in mESCs (*11*, *17*). The associations of CAP-C loops with histone modifications and transcription factor features around their anchor points suggest that the increased loops captured in CAP-C are not artifacts but functionally similar with loops identified in *in-situ* Hi-C (Fig. 2.15, H and I). In this instance, we merged contact domains called at high-resolution (500 bp, 1 Kb, 2 Kb) into a unique set, and classified the contact domains into loop and non-loop domains, based on whether they are associated with Ctcf-mediated loops. Segregating and plotting the distributions of the number of protein-coding genes per domain and size of domains showed striking contrast between these two "arbitrary" types of domains. Non-loop domains are overwhelmingly shorter, and possess at most one protein coding gene, whereas loop domains are longer and contain 1 or more protein-coding genes (Fig. 2.15, C and D).

To study the possible mechanisms separating the two types of domains, we next overlapped domain boundaries and domain bodies with a series of histone modification marks. To account for

the long-tailed size distribution of some of these domains, and the relatively smaller peaks generally associated with histone modification marks and transcription factors, we extracted only signals +/- 2 Kb around the boundary, and signals from 5-95% around the domain body. As expected, loop domains showed stronger Ctcf and cohesin signals than non-loop domains at their boundaries. However, some of the non-loop domain boundaries are also enriched with Ctcf and cohesin binding, suggesting that not all Ctcf- and cohesion-enriched domain boundaries form loops. Conversely, non-loop domains exhibit stronger H3K4me3, H3K27ac, PolII and Top2b signals than loop domains at their boundaries (Fig. 2.15, A and B, 2.16).

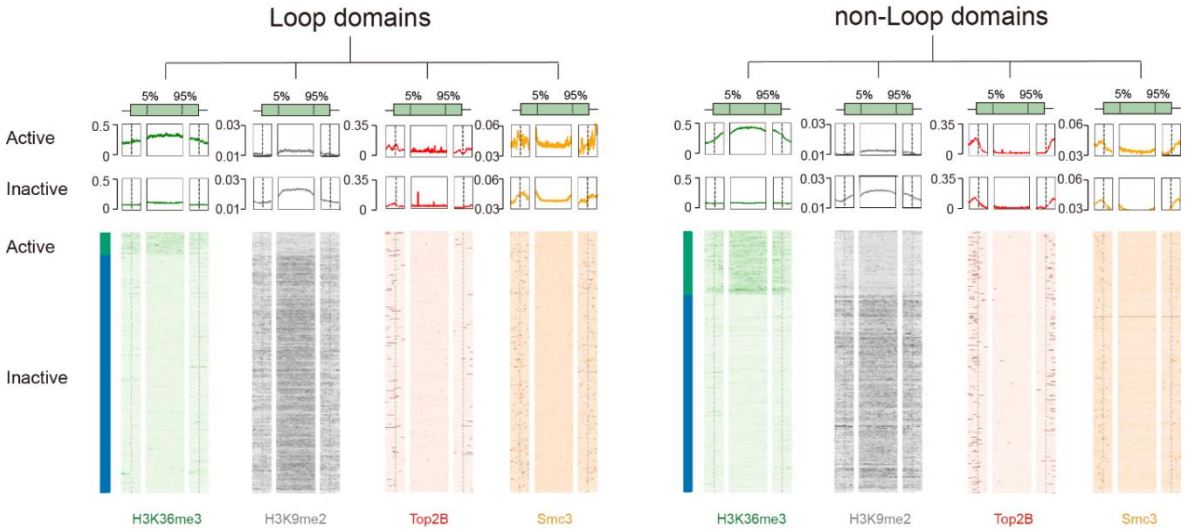


Fig. 2.16. Extended meta domain analysis. Meta-analysis of Top2b and Smc3 (cohesin) binding around boundaries (+/- 2 Kb) of loop and non-loop domains suggest that boundaries of loop domains are enriched in Smc3 binding while boundaries of non-loop domains are enriched in Top2b binding.

We calculated a 2.2-fold enrichment of active promoter marks in non-loop domains compared to loop domains (Fisher's Exact Test; $P < 0.0001$). Using K-means clustering, we classified domains by the presence of H3K36me3, an epigenetic mark for transcription elongation, in domain

bodies. Non-loop domains with low H3K36me3 signals still contain a 3.2-fold and 2.7-fold enrichment of H3K4me3 and PolII signals over loop domains with no H3K36me3 signals, respectively (Fisher's Exact Test, both $P < 0.0001$). We then asked whether the strength of domain boundaries correlate with the level of transcription. Indeed, using chromatin states information, we were able to classify TSS into four transcriptional elongation states and observed that boundaries of actively transcribed genes were highly open when plotted with mean CAP-C eigenvector values (Fig. 2.15F), and were located at the center of domain boundaries when viewed alongside mean directionality index (DI) values (Fig. 2.15G). In addition, TSS marked by poised promoters of bivalent states (PRC2-repression + PolII-H3K4me3 active promoter) and PRC2-repressed regions showed decreased levels of domain boundary formation, suggesting that transcription elongation might be associated with chromatin structure.

As loop domains were proposed to form via Ctcf-cohesin loop extrusion, the above observation led us to hypothesize that non-loop domains might be established through transcription-induced supercoiling, similar to the formation of CIDs in *S. cerevisiae* and *C. crescentus*. The twin-supercoiling domain model could predict how waves of supercoiling that propagate through diffusional pathways react when encountering each other; they either enforce or cancel each other based on the propagation direction (*32*). Consistent with this model, our mouse CAP-C maps showed similar domain formation based on the orientation of gene pairs previously shown in *S. cerevisiae* (Fig. 2.15E) (*16*).

Moreover, we observed that topoisomerase Top2b, which reduces torsional stress generated during transcription elongation (*33*), was highly associated with active promoters. Peaks called in our mESC Top2b ChIP-seq experiments showed a 2.6-fold enrichment of peaks overlapping at least 1 non-loop domain boundary versus that of loop domain boundary (Fisher's Exact Test. P-

value < 0.0001) (Fig. 2.16). The enrichment of PolII and Top2b binding on non-loop domain boundaries suggests that such chromatin domains could be formed by the action of supercoiling of duplex DNA during transcription elongation.

2.2.11 Effects of supercoiling on the structure of genes with multiple active promoters

Alternative promoter usage is a common mechanism for generating transcript complexity. Unlike alternative splicing, alternative promoter usage generates diversity across multiple cell-types by selectively positioning the pre-initiation complex at different transcription start sites (TSS) before elongation (*34*). As distances between alternative promoters can range from only tens to thousands of base pairs, these features can now be discernable by our high-resolution contact maps with enriched short-range interactions. Because multiple active promoters that occur in a single gene are in the tandem direction, we predict from the twin-domain-supercoiling model an attenuation of boundaries as positive and negative supercoils cancel each other at the active downstream promoter; this is analogous to the mean O/E contact map of gene pairs that are arranged in a tandem fashion (Fig. 2.15E). From the high-resolution CAP-C contact matrix, we observed that downstream promoters can cause insulation. Moreover, the domain boundaries correlated strongly with active promoters inside the gene body and were not well associated with Ctcf or cohein binding (Fig. 2.17A). Encouraged by this observation, we developed a scheme (see Materials and Methods) to select genes with multiple active alternative promoters (H3K4me3) and classified them into 4 possible combinations (Fig. 2.17B). As expected, we not only observed strong domain boundaries at all active alternative promoter sites, but also observed that promoters downstream of type IV genes showed weaker boundaries in the O/E contact map and reduced directionality index values, even as they are bound by PolII, and showed evidence of divergent transcription (Fig. 2.17C).

Therefore, we propose that negative supercoiling as well as the cancellation of positive and negative supercoils at the domain boundaries causes DNA to be in an unwound (low twists) or relaxed state, fulfilling the requirement of insulation, while conformational changes resulting from both positive and negative supercoiling as writhes fulfill the self-associating property (*35*).

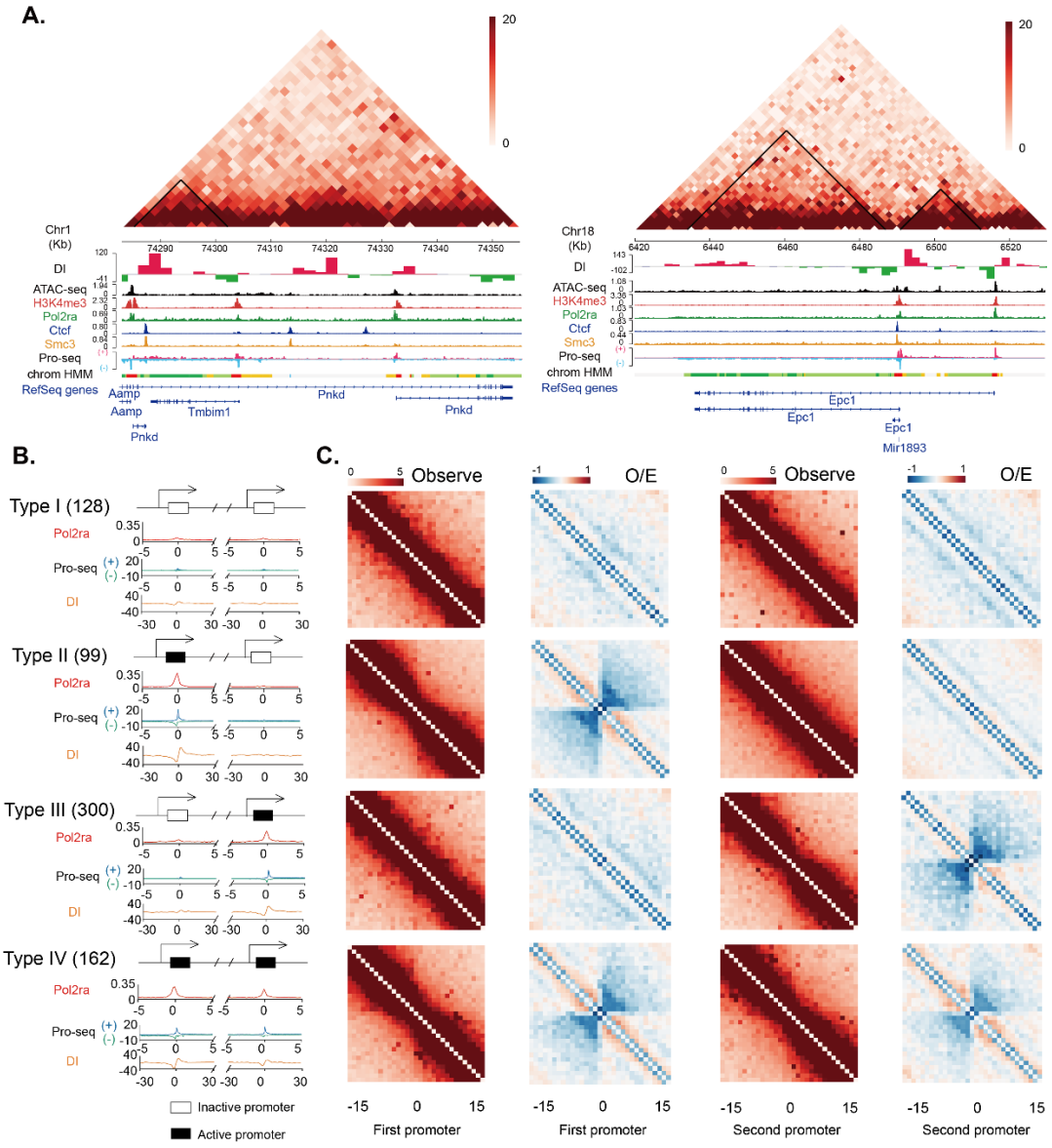


Fig. 2.17 Active multiple promoters are involved in contact domain boundary formation. (**A**) Genes are segregated into different contact domains with active promoter on the boundary. Two examples of CAP-C contact matrix (at 1 KB resolution) are shown corresponding to Chr1: 74.28-74.35Mb (Left) and Chr18: 6.42-6.53 Mb (Right). Black line depicts the domains called by Arrowhead. Direction index, histone modification and ChIP profiles are listed below each matrix.

Fig. 2.17 continued Entire Pnkd and Epc1 are separated into two domains by active promoter of Tmbim1 and Epc1 respectively. The boundaries are less correlated to Ctcf and cohesin binding. (**B** and **C**) Active alternative promoters are highly associated with strong domain boundaries. (B) Genes with alternative promoters were selected and classified into 4 different types based on the transcription state of their first and second promoters. Numbers of each type are shown inside the corresponding brackets. PolII ChIP profiles, Pro-Seq and Direction index around each promoter are shown below. (C) Interaction counts and the log2 ratio of observed interactions divided by expected interactions for a given genomic distance are shown side by side for each type.

2.2.12 Inhibition of transcription reduces supercoiling and leads to global loss of chromatin contacts

Chromatin topology highly associates with supercoiling, and supercoiling domains have been proposed and identified (*36*). These supercoiling domains were shown to partially overlap with TADs. Motivated by a relationship between transcription-induced supercoiling and domain organization, we next examined whether transcription inhibition affects chromatin architecture. We explored two different transcription elongation inhibitors, flavopiridol and α-amanitin (*37*). Reduced levels and rates of supercoiling have been observed upon transcription inhibition (*36*). Thus, we performed time-series CAP-C experiments using G5 dendrimers to crosslink mESC samples treated with 2 μM flavopiridol for 1 h and 6 h, as well as samples treated with 4 μg/ml of α-amanitin for 6 h and 12 h, respectively.

No significant differences were observed between the compartments of G5-control and inhibitor-treated G5 samples, indicating that transcription is not required to maintain compartments, and that compartmentalization may have been established much earlier during early development (Fig. 2.18).

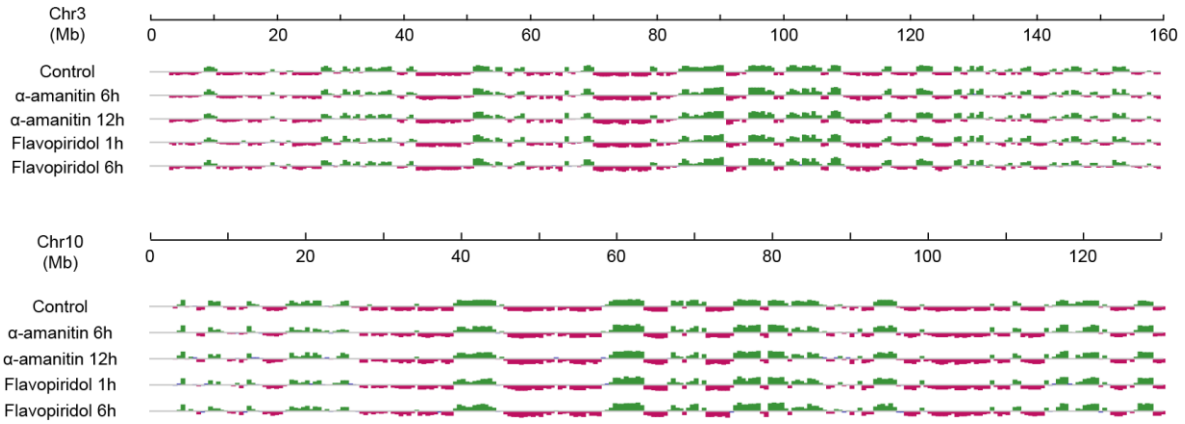


Fig. 2.18 Compartments remain unchanged upon transcription inhibition. Two examples of eigenvectors across full-length chromosomes, chr3: 0-160 Mb (top panel) and chr10: 0-130 Mb (bottom panel), before and after transcription inhibition. Compartment intervals remain unchanged for all inhibitor-treated samples. Compartment A intervals are colored green while compartment B intervals are colored red.

As expected, all 4 inhibitor-treated samples showed a profound loss in long-range interactions greater than 10 Kb in genomic-distance (Fig.2.19).

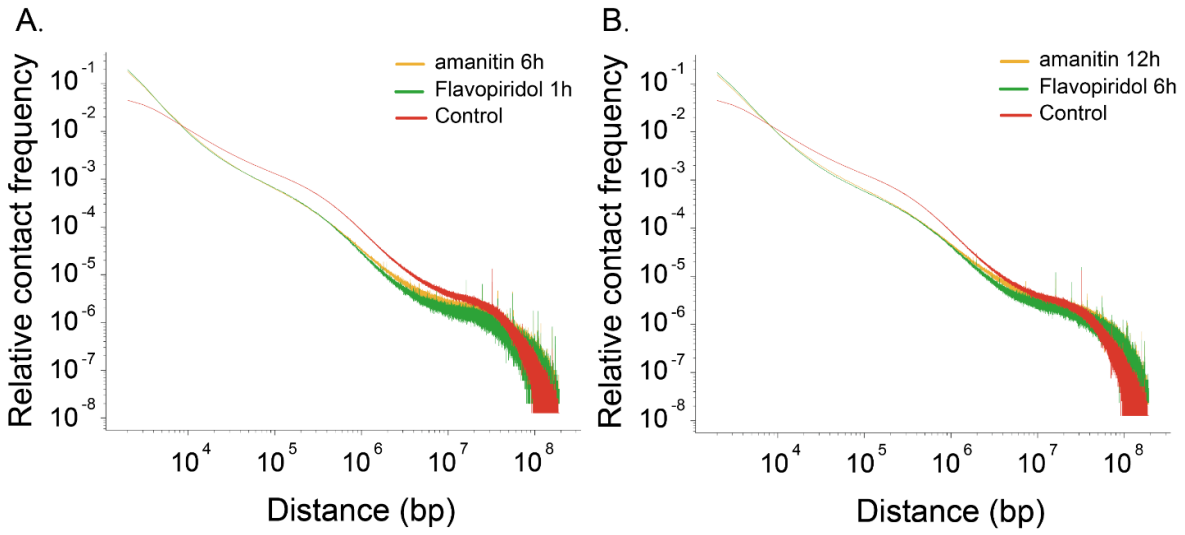


Fig. 2.19 Loss of long-range chromatin contacts are observed through transcription inactivation. Relative contact frequency *vs* distance curve generated from control and transcription inhibitor-treated samples reveal that a large portion of long-range chromatin contacts (over 10 Kb) are lost upon transcription inhibition. Flavopiridol showed a slightly strong depletion of contacts over 1 Mb.

CAP-C contact maps also displayed an extensive loss of contacts within domains for all inhibitor-treated samples. Directionality index values used to gauge domain boundaries were extensively decreased (Fig. 2.20A, 2.21), whereas the number of domains called using Arrowhead was significantly reduced (Fig. 2.20B). However, despite the disappearance of domains, more than half of peaks originally called in the control sample remained (Fig. 2.20B) after the second-time point (6 hr and 12 hr treatments for flavopiridol and α-amanitin, respectively). The signals of these peaks were, however, weakly attenuated (Fig. 2.20A). Because these inhibitor-treated and G5-control samples are low-resolution contact maps, we applied Aggregate Peak Analysis (APA) of the peaks called in the deep sequenced CAP-C dataset. A superimposed image of all signals overlapping the peak set showed enrichment of the foci in both G5-control and the inhibitor-treated samples (Fig. 2.20C, 2.22). We plotted the mean contact maps of 103-263 loop-domains (median length: 170Kb) shared between the control and inhibitor-treated samples by rescaling them into 10 bins (i.e. contact domains in G5-control overlapping loops that are shared between control and inhibitor-treated samples). Contact maps of these loop-domains showed the presence of loops at the corner of the assigned domain boundaries in the inhibitor-treated samples (Fig. 2.20D). We also quantified these maps into a contact probability vs distance-scaled bins line-plot for all loop domains, non-loop domains and randomly permutated domains. As expected, we observed a general decrease in contact probability between distance-scaled bins 2-7 for both the loop domains and non-loop domains when comparing G5-control and α-amanitin treated sample (the same trends were also seen for flavopiridol-treated samples). Increasing contact frequencies identified at corners of inhibitor-treated loop-domains (bins 8-10) indicate the presence as well as attenuation of peaks from G5-control (Fig. 2.20E). This is in stark contrast to the absence of such an increase in non-loop domains and randomly permutated domain boundaries.

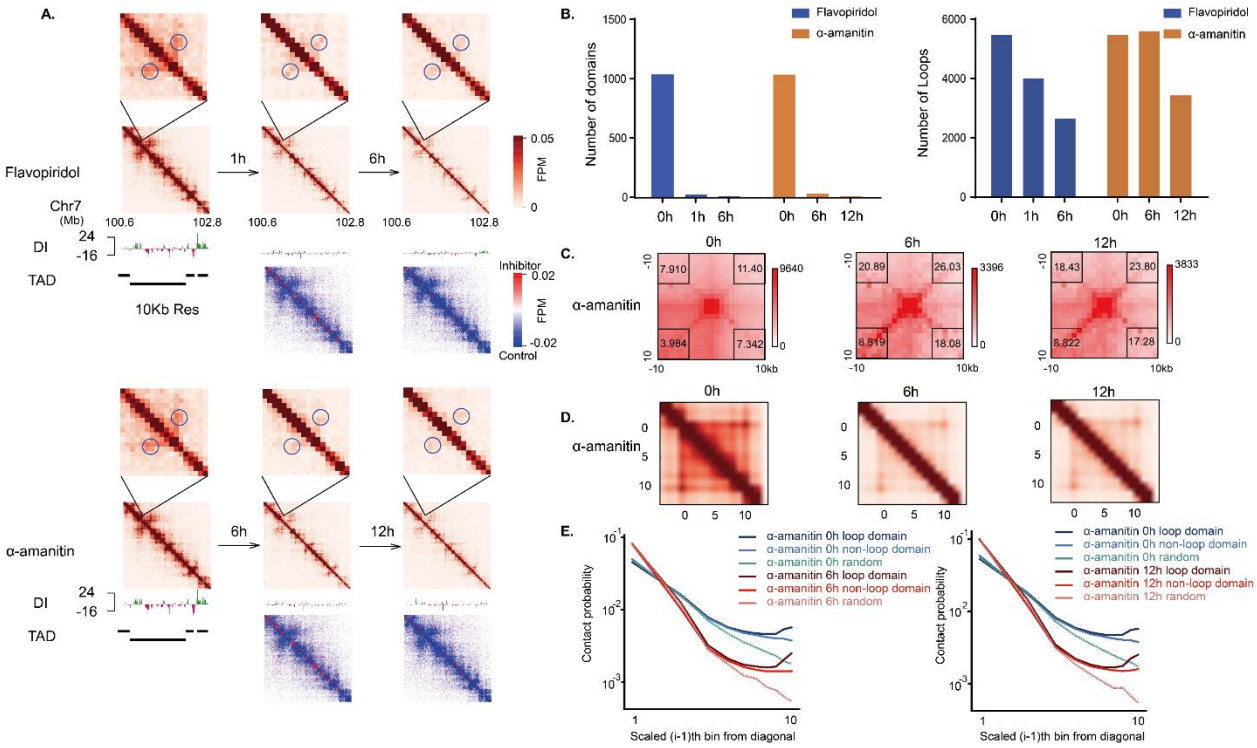


Fig 2.20 Inhibiting transcription causes widespread loss of domains and attenuates loops. (**A**) 10 kb resolution contact matrix shows a profound decrease in contacts after treatment of 2 μM flavopiridol for 1 h and 6 h as well as 4 μg/ml α-amanitin for 6 h and 12 h; directionality indices (DI) as well as inhibitor-treated vs control delta maps also indicate the weakening of boundaries and loss of intra-domain interactions respectively. A representative example of a loop shows it being present but weakened over time as the domain is lost. (**B**) Domains were called by Arrowhead and loops were called by HiCCUPS. Bar chart shows that domains are totally disappeared while loops are reduced. (**C**) Aggregate peak analysis (APA) for control, 6 h and 12 h α-amanitin treated samples indicates the presence of loops. APA was performed using the set of confident peaks called from deep-sequenced CAP-C on the lower-resolution inhibitor maps. Values greater than 1 in the bottom-left box indicate the presence of loops. APA scores (Enrichment of mean signal at peak foci over the mean signal in the lower-left corner) are also greater in α-amanitin-treated samples over control, indicating that the loss of domains during transcription inhibition is greater than the attenuation of loops. (**D**) Meta-loop analysis shows the widespread loss of domains and the presence of weakened loops. Peaks called in both control and inhibitor maps were overlapped with domains in control to yield 103-263 loop-domains with a medium length of 170 Kb for meta-analysis. Mean map of α-amanitin-treated samples, with distances rescaled into 10 bins, were shown as an example. (**E**) Contact probability of the mean maps quantifying results in (D) shows: 1) an increase of contacts at the diagonals and decrease of contacts in bins 2-7 between control and α-amanitin-treated samples; and 2) a corresponding increase in contacts of inhibitor-treated samples at the corners of all loop-domain. This is contrasted with a lack of increase at the corners of both non-loop domains and randomly permutated domains, suggesting the presence of an attenuated loop.

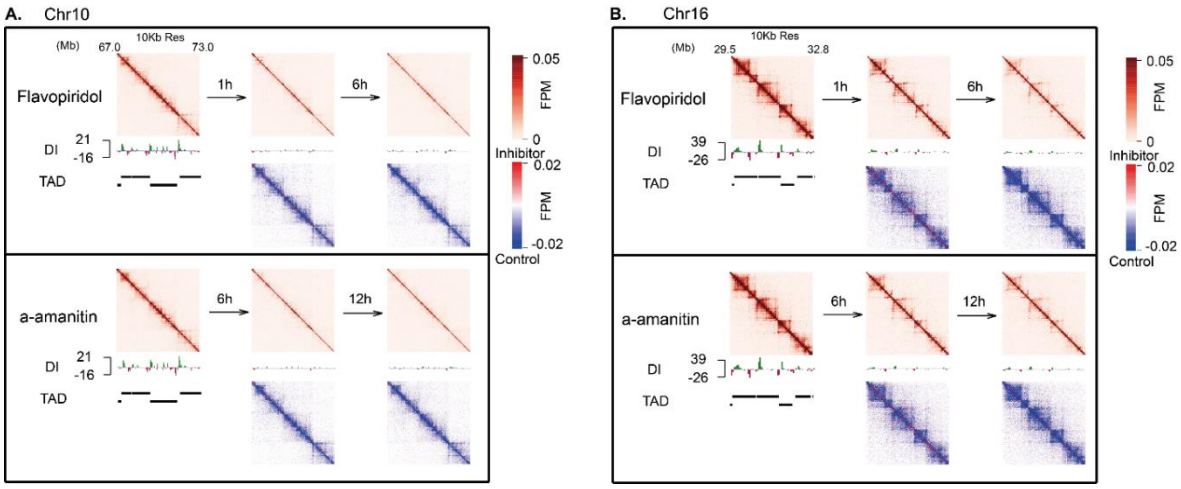


Fig. 2.21 Transcription inhibition causes widespread loss of domains. 10 Kb resolution contact maps shown here for chr10: 67.0-73.0Mb (left panel) and chr16: 29.5-32.8Mb (right panel). Reduction in directionality indices (DI) between inhibitor-treated vs control indicate the weakening of boundaries. Similarly, delta maps between inhibitor-treated vs control maps indicate the loss of intra-domain interactions.

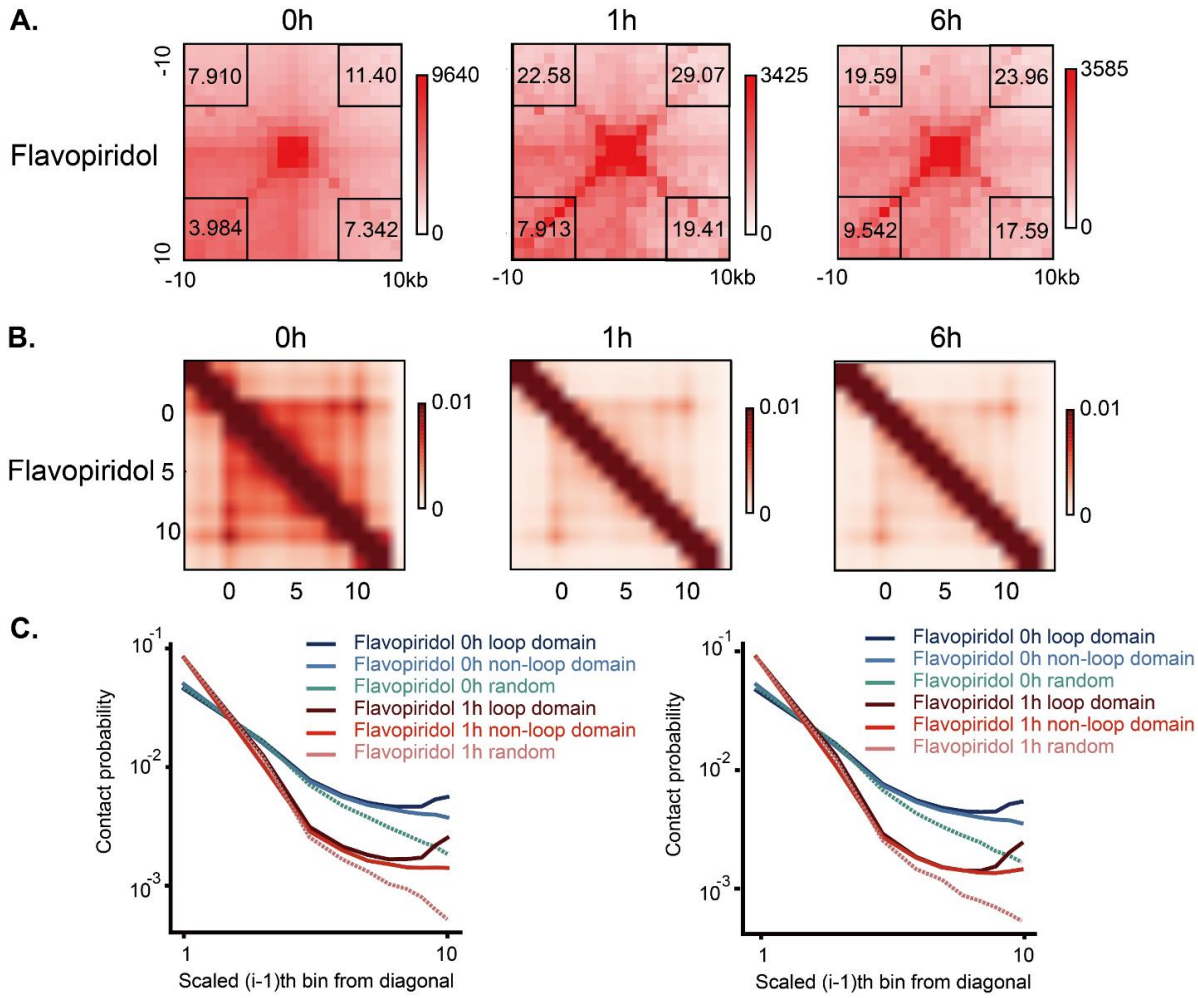


Fig. 2.22 Loops are attenuated but preserved in flavopiridol-treated samples. (**A**) APA analysis was performed on low-resolution maps using loops called in the high-resolution CAP-C map. Aggregated signals showed the presence of loops after flavopiridol treatment. (**B**) Similar meta-analyses of loops were performed (See Fig. 5) for flavopiridol-treated samples. (**C**) Mean maps were quantified as contact probability vs scaled distance and showed that loops are preserved but attenuated (bin 8 to 10).

Therefore, we conclude that domain formations are dependent on transcription-induced supercoiling. Blocking transcription elongation abrogated both loop and non-loop domains; however, loops were attenuated but largely retained. These observations support the critical role of transcription-induced supercoiling in the formation of non-loop domains, but also suggest that transcription-induced supercoiling and loop extrusion likely work synergistically to shape the overall

chromatin architecture as the formation of loop domains also appear to be dependent on transcription. Taken together, we propose that positive and negative supercoiling generated during transcription elongation are responsible for the intra-domain contact interactions observed in our experiments.

2.2.13 Different sizes of dendrimer show different binding preference around transcription start sites (TSS)

Next, we question whether the fact that those small size dendrimer G3 preferred to be enriched at close compartment while large size dendrimer G5 and G7 enriched at open compartment comes from the ability of dendrimer to probe different size of the chromatin conformation or if dendrimer has its own preference on binding to specific DNA regions. In order to solve that puzzle we synthesized dendrimer with psoralen as well as biotin on the branch. We mix different size of the synthesized dendrimer with mESCs and initiating the CAP-C crosslinking. The crosslinked DNA-dendrimer complex is then purified and fragmented by sonication. The sheared DNA fragments were further subjected to library construction and high-throughput sequencing. We call this experiment as 1D-dendrimer capture (Fig. 2.23).

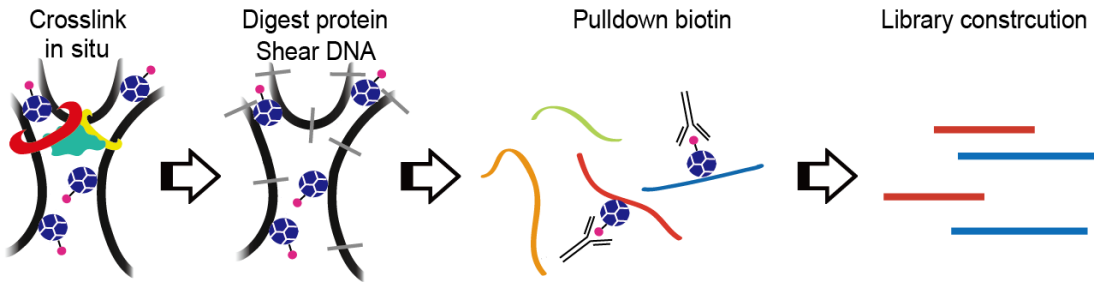


Fig. 2.23 General scheme of 1D dendrimer capture experiment. Dendrimer G1, G3, G5 and G7 are modified with psoralen as well as biotin (shown as pink ball) on the branch. Each size of the dendrimer is crosslinked with mESCs with UV irradiation. The DNA/dendrimer complex is purified and fragment with sonication. The DNA fragments are further subjected to library construction and next generation sequencing.

First of all, we mapped all the DNA sequence that have been pulldown with biotinylated dendrimer. The peaks called in 1D dendrimer capture experiment by using different sizes of dendrimer showed an evenly distributed profile. The result here suggests that difference of absolute numbers of psoralen on each type of dendrimer showed no sequence preference for dendrimer to crosslink with DNA. In addition, such observation also supports the idea that different size of the dendrimer has preference on different chromatin compartment comes from the versatile of dendrimer size fitting different chromatin conformation instead of bias on psoralen in crosslinking specific DNA sequence. Next, we treated the cells with RNAPII elongation inhibitor, flavopiridol, for 1h and performed the 1D dendrimer capture experiment. Again, the DNA sequence captured by different size of the dendrimer is still evenly distributed along the genome (Fig. 2.24).

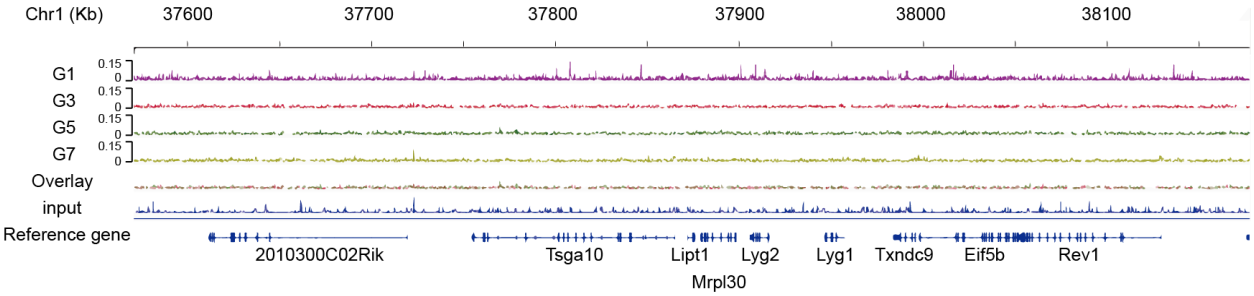


Fig. 2.24 DNA sequences captured by different size of dendrimer are evenly distributed across genome. Raw reads captured by dendrimer G1, G3, G5 and G7 are aligned to the mouse genome and shown on IGV. Overlay of the 4 track and the input are shown below. (Selected genomic region: chromosome 1: 37570769-38179982bp)

However, when normalizing the counts by sequencing depth around the transcription start sites (TSS), we found that aside from dendrimer G1, all other 3 dendrimers displayed increasing of enrichment around TSS followed by the order of the increment of dendrimer generation. However, cells treated with flavopiridol showed a decrease of binding intensity for dendrimer G5 and G7. On the contrary, dendrimer G1 starts to enrich at the TSS compared to WT and dendrimer G3

seems to gain most favorable fitting around the TSS. As we know the DNA sequence keep invariable, changes of binding preference around TSS by different dendrimer is not the result of bias for psoralen crosslink to specific DNA sequence but comes from the ability of dendrimer to fit at different chromatin conformation (Fig. 2.25).

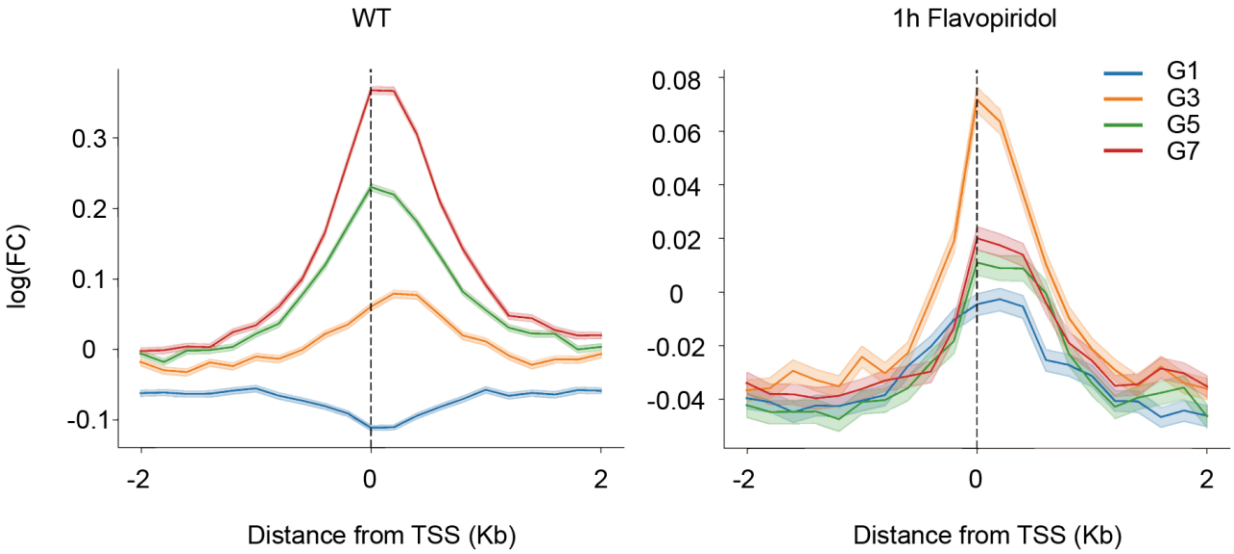


Fig.2.25 Dendrimer is able to probe the chromatin conformation change around TSS. Each mapped reads captured by different size of the dendrimer are normalized by the depth of sequencing and centered at TSS +/- 2Kb region. Left: wild type cells (WT); right: cells treated with flavopiridol for 1h (1h flavopiridol).

To test the relationship of transcription and the binding preference of dendrimer around corresponding TSS, we first classified (TSS) of wild type cells by their transcription strength using Pro-seq data into 10 percentiles. 90th percentile shows the highest nascent gene expression while the 0th percentile exhibits lowest. Then, we normalized the counts by sequencing depth (FPM) and plotted +/- 2Kb around each type of TSS. Intriguingly, for the wild type (WT) mESCs, we found that aside from dendrimer G1, all other 3 dendrimers showed slightly enriched at TSS. Besides, more significant enrichment was observed if the TSS showed higher nascent gene expression. However, G1 on the contrary, displayed a depleted binding preference at TSS, especially for TSS with higher nascent gene expression. Next, we treated the cells with RNAPII elongation inhibitor,

flavopiridol, for 1h and performed the 1D dendrimer capture experiment again. Similarly, we discovered that G5 and G7 dendrimer showed decrease binding preference around TSS while dendrimer G1 and G3 start to enrich at TSS. It is important to noted that degree of transcription level was thought to be close related to the chromatin accessibility. Higher transcription is always correlated to more open chromatin with more accessibility. On the contrary, inhibition of RNAPII elongation leads to a decrease of the amount of RNAPII around TSS as well as the transcription machinery. In this way, the corresponding chromatin region around TSS become relatively close and less accessible for large dendrimer but more favorable for small dendrimer compared to wild type. Taken together, we concluded that the phenomenon we observed here suggests that different sizes of dendrimer are able to probe the openness conformation around TSS (Fig. 2.26).

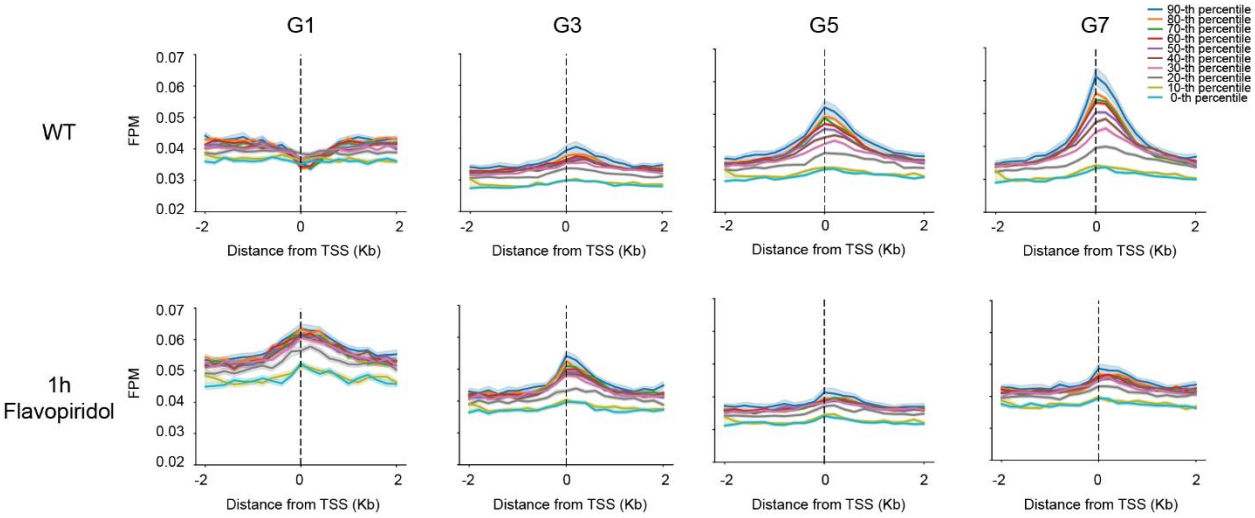


Fig.2.26 Probe the openness of transcription starting sites (TSS) by biotinylated psoralen functionalized dendrimers. TSS are classified into 10 groups based on their transcription state. Each line represents the mean of the percentile of nascent gene expression at its TSS using Pro-seq signal. (90^{th} percentile shows the highest nascent gene expression while 0^{th} exhibits the lowest) Normalized sequencing counts (FPM) are plotted +/- 2Kb around TSS for each percentile. (First row: experiment was performed on wild type cells; Second row: experiment was performed on cells treated with flavopiridol for 1h)

2.2.14 Characterize condensin and YY1 as potential player for non-loop domain boundary formation.

As we have demonstrated the existence of two type of domains, namely loop domain and non-loop domain. Loop domains are thought to form through cohesin/CTCF mediated loop extrusion and upon depletion of cohesin, all loop domains disappear. Recovery of cohesin by removing auxin treatment showed the re-appearance of loop domain. However, there is little or no evidence in demonstrating the mechanism of non-loop domain formation until recently, people found two other proteins, condensin and YY1, could be served as potential players contributing in chromatin structure formation. First thing first, unlike CTCF, YY1 could form loops bringing distal enhancers and promoters together and form insulated neighborhoods. Insulated neighborhoods are small in size compared to TADs and is thought to be closely related to gene regulation. As we found non-loop domain boundaries are enriched for enhancer and promoter, we thought YY1 could be the reason for non-loop domain formation. Secondly, single molecule experiment done in vitro suggests that condensin is able to consume ATP and translocate on the DNA sequence. And it is not until recently people have discovered that one condensin sub-units NCAPH2 is enriched around active promoter sites. Based on the above discoveries, we are questioning whether these two proteins might play an important role in non-loop domain formation.

We first did a meta-domain analysis using NCAPH2 ChIP-seq data and discovered that NCAPH2, as a sub-unit of condensin, indeed enriched at non-loop domain boundaries but not the loop domain boundaries. Such observation motivates us to check whether NCAPH2 will form loop around non-loop domain. We did HiChIP using NCAPH2 and YY1 antibody followed by proximity ligation. The preliminary results showed no or little enrichment when compared HiChIP versus in situ Hi-C, which serve as a non-enrichment input. We then try to check the domain boundary

insulation score in each HiChIP experiment and compare them with in situ Hi-C. The insulation score in both HiChIP experiment showed 2-3 folds enrichment compared to in situ HiC at the loci where NCAPH2 or YY1 showed strong binding preference based on the ChIP-seq results. Moreover, we also compare our YY1 HiChIP data with published YY1 ChIA-PET by looking at the insulation score. To our delight, the enrichment of insulation score toward YY1 binding sites share similar pattern (Fig. 2.27).

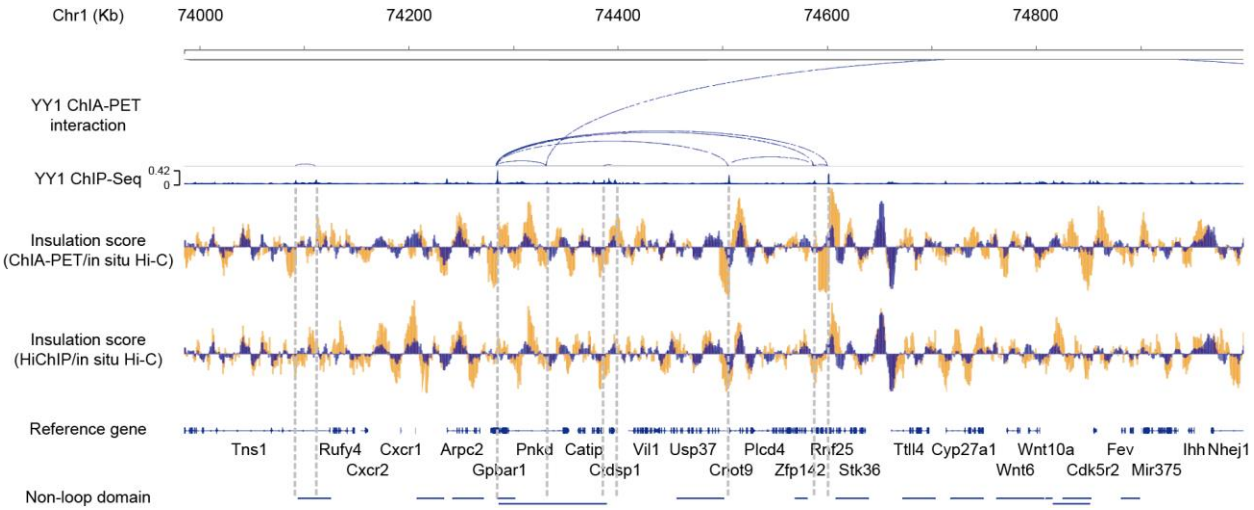


Fig.2.27 Validation of YY1 HiChIP by comparing with published YY1 ChIA-PET. Insulation score is calculated by dividing ChIA-PET or HiChIP YY1 directional index value (DI value, shown in yellow) by in situ Hi-C DI value (shown in blue). Non-loop domains from 500bp, 1Kb, 2Kb resolution are merged and listed below. (Selected genomic region: chromosome 1: 73985252-75001652bp)

As we validate our YY1 HiChIP data, we then compare the significant YY1 interactions with our non-loop domain boundaries. However, though some of the YY1 interactions do span from one non-loop domain coordinator to another, most of the YY1 loops locate inside one non-loop domain or even span across the boundaries of two domains, suggesting though YY1 could mediate enhancer and promoter interactions, is not the cause of non-loop domain formation. We also turn to check NCAPH2, though NCAPH2 HiChIP showed interesting enrichment of insulation score

at some of the active promoter sites, most sites display less enrichment on insulation score compared to YY1. Moreover, most of the NCAPH2 significant interactions also span across two non-loop domains, suggesting NCAPH2 is also not crucial for non-loop domain formation. However, as the active promoter features in non-loop domain, we hypothesized that is it plausible that RNAPII loops might correlate better with our non-loop domain. Indeed, when comparing our non-loop domain boundaries with significant interactions mediated by RNAPII, we found that PolII loops won't span over two non-loop domain boundaries and all of the loops tend to locate inside one non-loop domain. The dynamic of RNAPII elongation within the non-loop domain also support the fact that non-loop domain boundaries contain no significant static interactions like loop domains. In this way, we demonstrate that RNAPII most likely be the cause of non-loop domain formation while YY1 and condensin might play other role on chromatin architecture (Fig. 2.28).

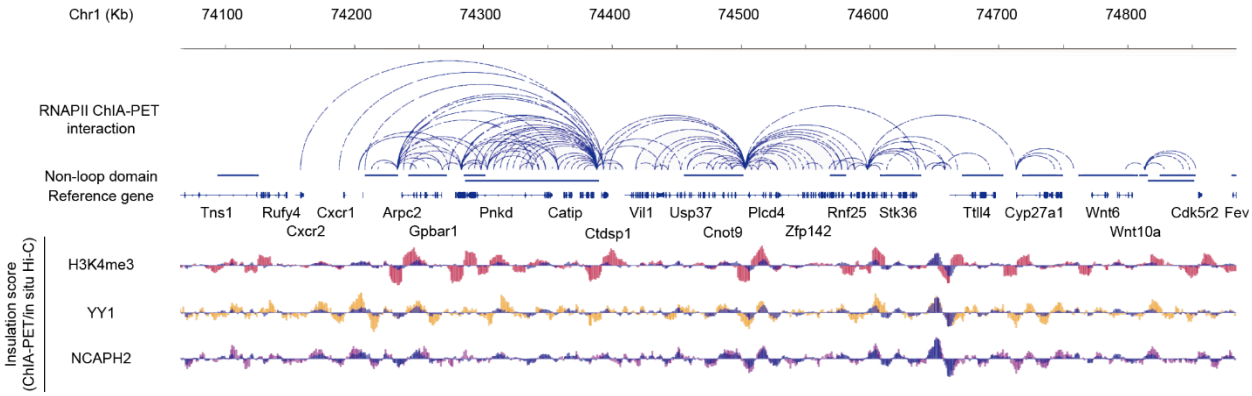


Fig.2.28 RNAPII mediated interactions showed better correlation with non-loop domain boundaries. Insulation score is calculated by dividing HiChIP H3K4me3 (shown in pink); YY1 (shown in yellow); NCAPH2 (shown in purple) by in situ Hi-C DI value (shown in blue). Non-loop domains from 500bp, 1Kb, 2Kb resolution are merged and listed below. (Selected genomic region: chromosome 1: 74064606-74887135bp)

2.2.15 Improve CAP-C by introducing a bridge linker for better proximal ligation

Though CAP-C showed advantages toward enrichment for short-range chromatin contacts as well as being able to probe chromatin conformation compaction, it still relies on formaldehyde crosslinking and restriction enzyme digestion. Furthermore, from an intuitive perspective it is not

clear that the differential diameters between dendrimers (3.6, 5.4, and 8.1 nm) would produce different features, as local contacts (generated by the crosslinking with a near zero-length linker formaldehyde) will get labeled with any of these distances. Thus, the argument that other method relies on formaldehyde fixation, but this psoralen method does not is incorrect, as the interactions that are "frozen in" to be probed with psoralen are crosslinked with formaldehyde, and thus this is the background of distances the CAP-C method probes. Moreover, though we digest away all the DNA binding protein and exposed all potential restriction enzyme digestion sites, the genomic DNA is still cut into non-evenly distributed fragments, causing potential bias on illustrating proximal chromatin interactions at all length. Thus, it is crucial to improve the current CAP-C strategy without using formaldehyde fixation. Besides, changing DNA fragmentation approach from restriction enzyme to another enzyme which could cut DNA into evenly distributed fragments such as DNaseI or MNase is also important. Lastly, in situ Hi-C and original CAP-C protocol all relies on blunt end ligation. The ligation efficiency is low and random ligation could occur, leading to high background on Hi-C contact matrix. Here we describe a modified version of CAP-C. Despite functionalized psoralen on the dendrimer, an azide linker is also added to the dendrimer tips. The azide linker provide dendrimer ability to further react with a designed alkyne modified sequence through click chemistry. The designed sequence bears a biotin and one thymine protruding out at both 5' end. The biotin helps to provide a handle for downstream enrichment by streptavidin beads while the 5' protruding T will complement 3' A tailing of the sequence captured by dendrimer. In this way, the proximal ligation switch from blunt end ligation to sticky end ligation. Such modification was designed to promote the ligation efficiency and reduce the random ligation. Attaching biotin directly on the dendrimer also give us a chance to fragment DNA with shorter endonuclease such as MNase.

2.2.16 General scheme of modified CAP-C

To test the feasibility of modified CAP-C, we first fixed mESCs with formaldehyde. We then diffuse azide and psoralen functionalized dendrimers into the cell nucleus and expose these cells to UV irradiation. The formaldehyde fixing is then reversed, and DNA-bound proteins are removed with protease to expose all DNA motifs, the dendrimer-DNA complexes are subsequently purified with ethanol precipitation. The purified dendrimer-DNA complexes are then subjected to MNase digestion followed by end polishing and A tailing. After that, DNA-dendrimer is purified again with ethanol precipitation to remove the excessed enzyme. Bridge linker containing biotin and is attached to dendrimer through click chemistry. Excess bridge linker is purified away by XP beads selection. The DNA-dendrimer complex is then ultra-diluted in ligation buffer and sticky proximal end is joint together by overnight ligation. The ligated products will then be pulled out with strep-tavidin beads followed by library construction and next generation sequencing (Fig. 2.29).

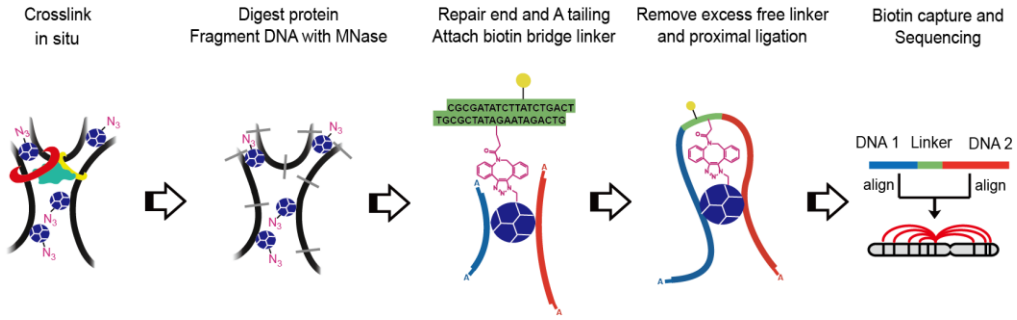


Fig.2.29 Scheme of modified CAP-C. Mouse embryonic stem cells (mESs) are treated with formaldehyde to crosslink proteins (Shown in red, green, yellow) with genomic DNA (black strings). Azide and psoralen-modified PAMAM dendrimers with fixed diameter (Shown as blue balls) are diffused into nucleus. DNA in proximity are covalently crosslinked with dendrimers under UV irradiation. Proteins are digested with protease and dendrimer-DNA complexed are purified. The purified complexes, without DNA-bound proteins, are then subjected to MNase digestion, end filling and A tailing. Biotin, DBCO bifunctional linker is attached to dendrimer through click chemistry. Excess unreacted linkers are removed by size selection. The dendrimer/linker/DNA complex is then ligated with ultra-dilution followed by biotin capture and high throughput sequencing.

2.2.17 Modified CAP-C capture more short-range contacts genome-wide.

Utilizing MNase leads to fragmentation of genome into evenly smaller pieces compared to restriction enzyme. Relative frequency of chromatin contacts for short range (from 1Kb to 10Kb) showed 30% increase compared to original CAP-C and 60% more compared to in situ Hi-C. This allows us to visualizing contact matrix with an even higher resolution down to 500bp. When compared to the highest mESC chromatin contact matrix, modified CAP-C map at 500bp has lower background and sharper domain boundaries. (Fig. 2.30) In addition, shorter fragmentation also leads us to get a finer structure of mitochondria genome. Previous restriction enzyme-based methods have limited cutting sites, resulting in capturing limited interactions for analysis. Here, for the first time, we obtained mitochondria genome structure with high resolution. The preliminary results suggest that mitochondria in mESCs are also partitioned into several domains. Each domain seems to correlate one specific mitochondria gene. As such feature is also observed in bacterial genome. We envisioned that mitochondria folds its genome through similar ways as compared in bacteria (Fig. 2.31).

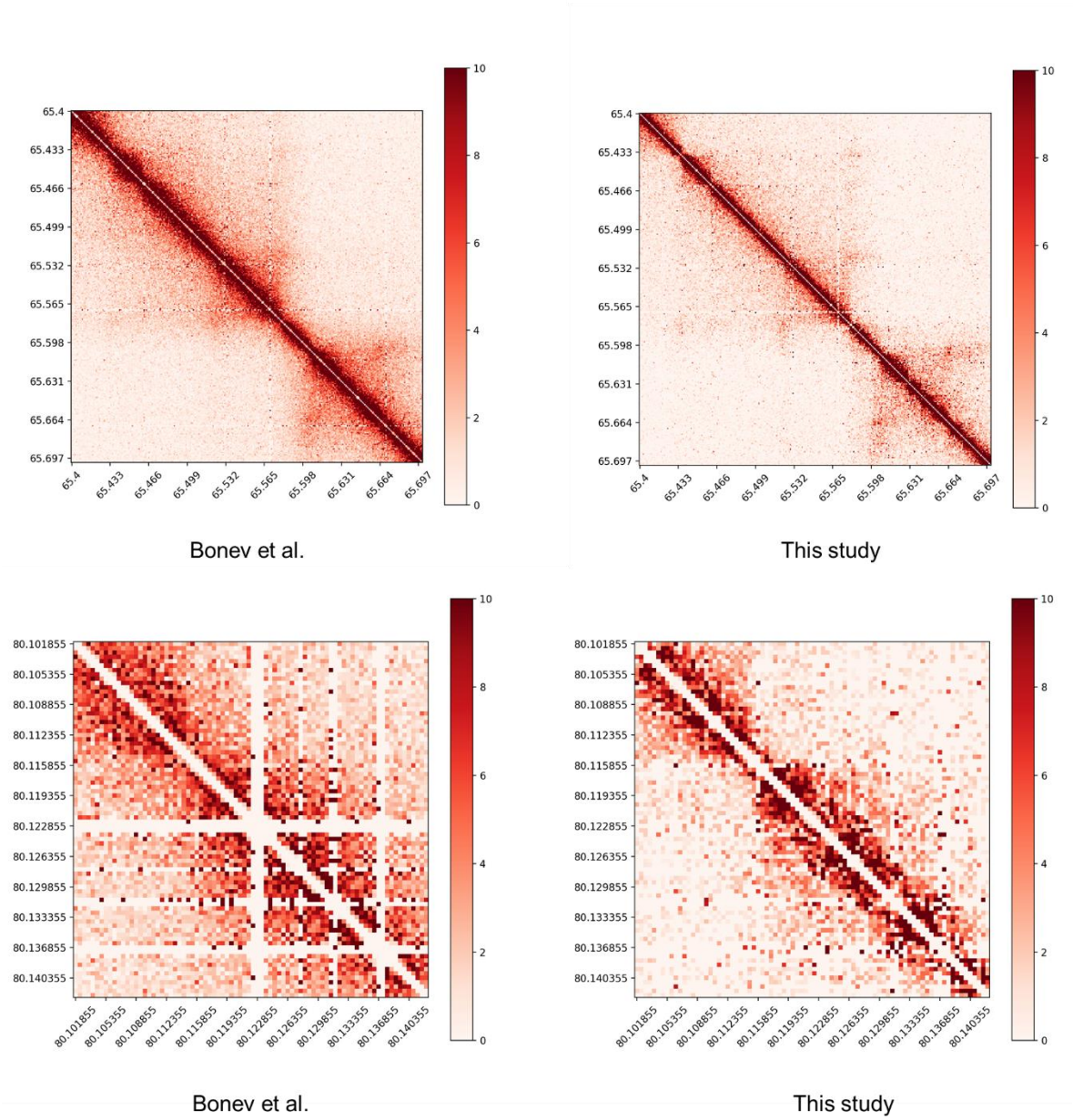


Fig.2.30 Modified CAP-C shows clean background and clear chromatin feature at high resolution. Two examples are shown for contact matrix at 500bp resolution. (Left: in-situ Hi-C; right: modified CAP-C)

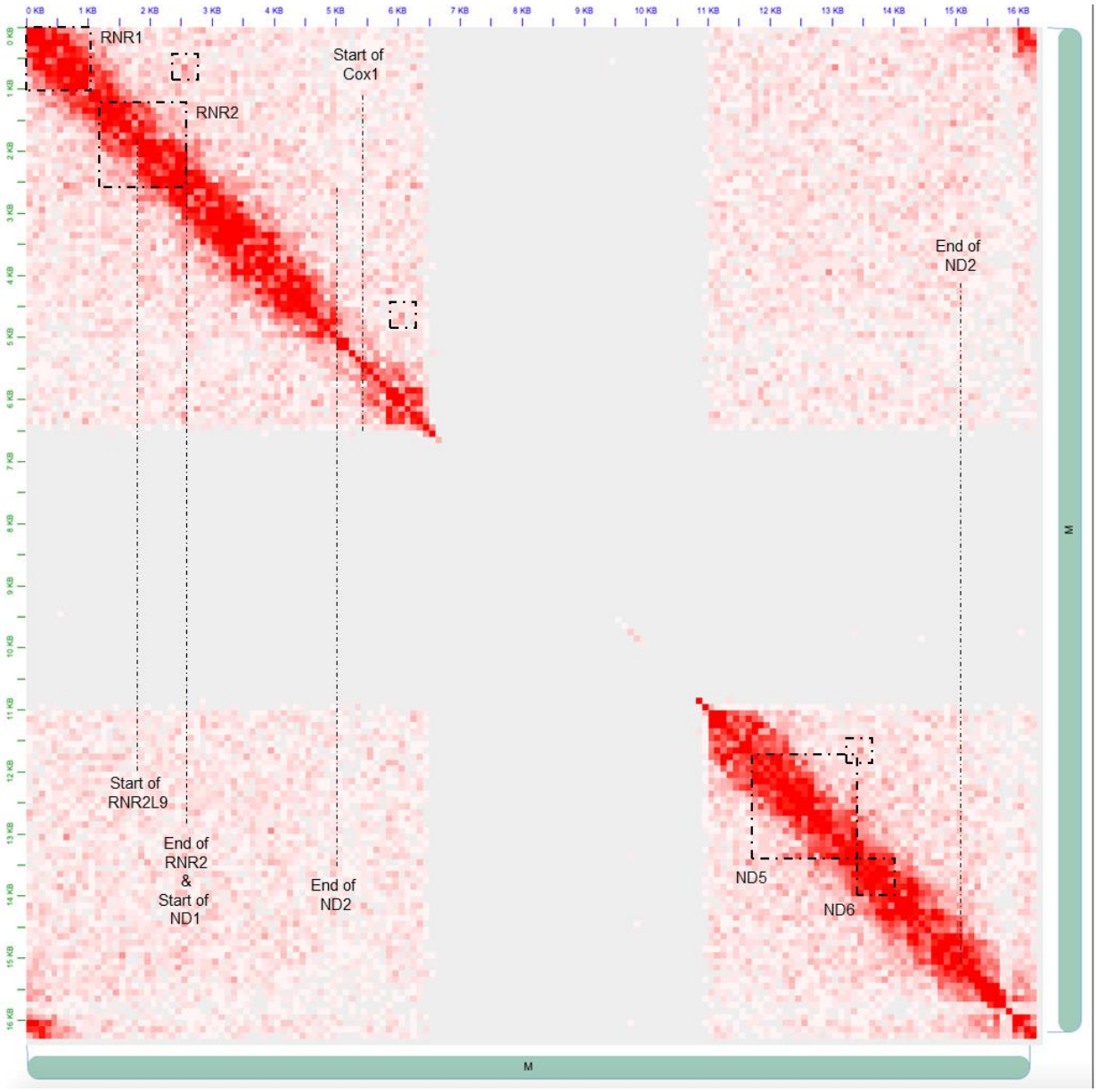


Fig.2.31 Modified CAP-C reveals mitochondria genome structure at high resolution. Dash lines depict start or end of gene position. Domains are highlight in the box.

We then compared loops and domains called by different methods and discovered a universal more loops being captured by modified CAP-C. Domains, on the contrary, CAP-C capture more smaller domains ranging from 5Kb to 25Kb while in situ Hi-C detect slightly more larger domains. The overall domains captured by these two methods are generally the same. We then questioned

whether smaller domains specifically captured by CAP-C are positive and functional related structure. Indeed, compared with in situ Hi-C we observed higher correlation between boundaries of smaller domains captured by CAP-C and active histone modification, CTCF and Smc1. We discovered that the smallest domain is around 5.5Kb and the smallest loop is 35Kb (Fig. 2.32).

Fig.2.32 Modified CAP-C reveals functional related chromatin feature at high resolution. Left panel: loops are called by HiCCUPS and grouped by size; Middle panel: domains are called by Arrowhead and grouped by size; Right panel: Enrichment folds of given proteins around domain boundaries are listed accordingly to domain size.

2.2.18 Acute deletion CTCF system.

Genetic perturbation is a powerful tool to analyze the function of proteins in vivo. (CRISPR)/CRISPR-associated (Cas) system-based gene-editing technology has revolutionized the generation of gene knockouts in mammalian cells. However, some of the essential genes like

CTCF is so important that constitutive knockout will be lethal. Thus, conditional depletion of encoding genes like CTCF will be useful for its function study. Conditional depletion can be achieved by fusing a destabilizing domain (so-called degron) that can be conditionally controlled. Here we use the auxin-inducible degron (AID) technology by transplanting a plant-specific degradation pathway controlled by a phytohormone, auxin, into mouse embryonic stem cells (mESCs). In mESCs expressing the auxin perceptive F-box protein TIR1, which forms a functional SCF (Skp1–Cullin–F-box) ubiquitin ligase, CTCF fused with an AID tag derived from the IAA17 protein of *Arabidopsis thaliana* can be induced for rapid degradation by the addition of auxin to the culture medium. In this way, we could transiently deplete CTCF and study its function on chromatin architecture.

2.2.19 Investigation the effect on genome architecture by inhibition of transcription or acute depletion of CTCF.

In order to further characterize loop domain vs non-loop domain, we questioned whether these two domains are formed and regulated through different mechanisms. As we hypothesized non-loop domain has a higher correlation to active histone modification and transcription while loop domain is closely related to CTCF/cohesin mediated loop, we tried to inhibit transcription and acute deplete CTCF in our CTCF-AID mESCs system and compare the two conditions with wild type and investigate the effects on chromatin structure. Besides, we also combine the above two treatment to see whether transcription and CTCF might work together on genome architecture.

For sufficient inhibition of transcription, we treated CTCF-AID mESCs with 1 μM of flavopiridol for 6h. For acute depletion of CTCF, we treated CTCF-AID mESCs with 500 μM auxin

for 48h. We also tried to combine the two treatment together on this cell line. Surprisingly, we found all 3 treatments lead to decrease of chromatin interactions in different degree. Among which, KO CTCF while inhibition transcription at the same time has the most significant effect on reducing the chromatin interactions genome-wide. Compare to that, KO CTCF alone showed comparable decrease but to a less degree while inhibition of transcription alone has the least impact on reduction of chromatin interactions. We then compare our KO CTCF results with the published KO CTCF data and identified similar decrease pattern when compared with wild type. In addition, we also performed CTCF, Rad21 ChIP-Seq on these 4 conditions. Both KO CTCF and KO CTCF + PolII inhibition lead to a global loss of CTCF peaks (Fig. 2.33). Besides, we found inhibition PolII reduce part of the CTCF level, which might come from the decrease of RNA when inhibition of PolII elongation. Taken together, these observations suggest the successful KO of CTCF in our CTCF-AID system (Fig. 2.34).

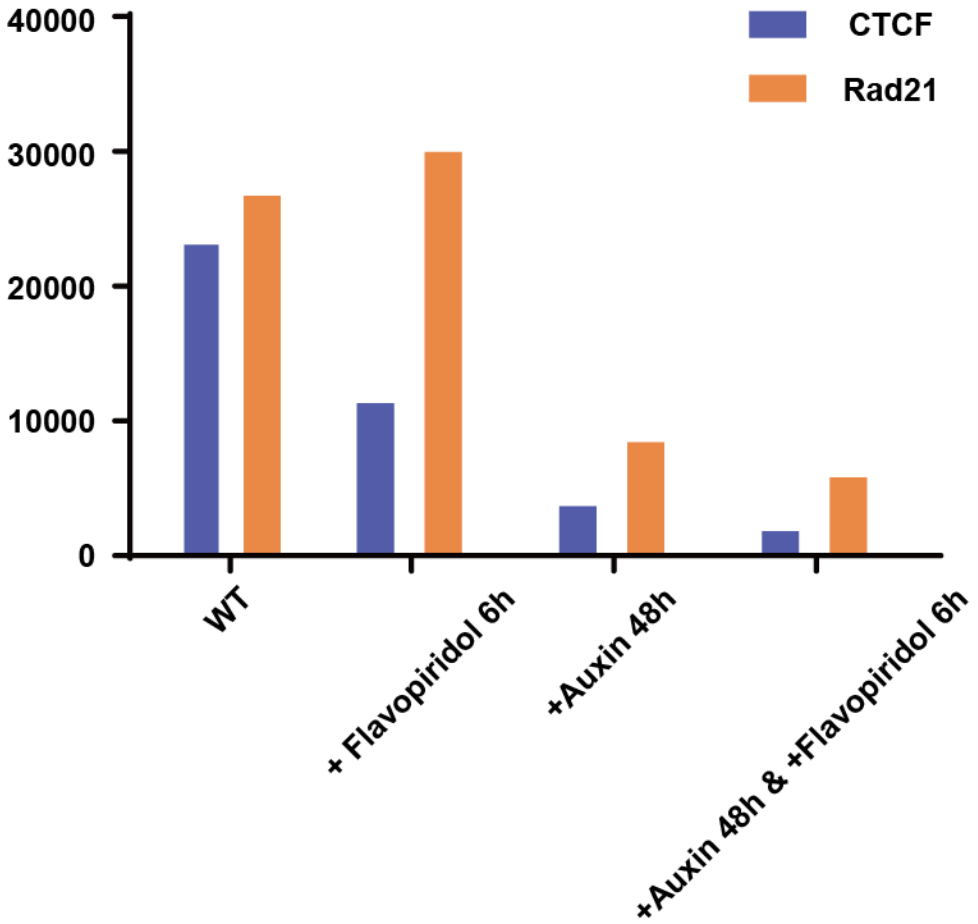


Fig.2.33 CTCF and Rad21 ChIP-seq. CTCF and Rad21 peaks with high confidence are called from ChIP-seq for sample of wild type (WT); RNAPII inhibition (+ Flavopiridol 6h); KO CTCF (+Auxin 48h); KO CTCF +RNAPII inhibition (+Auxin 48h & + Flavopiridol 6h).

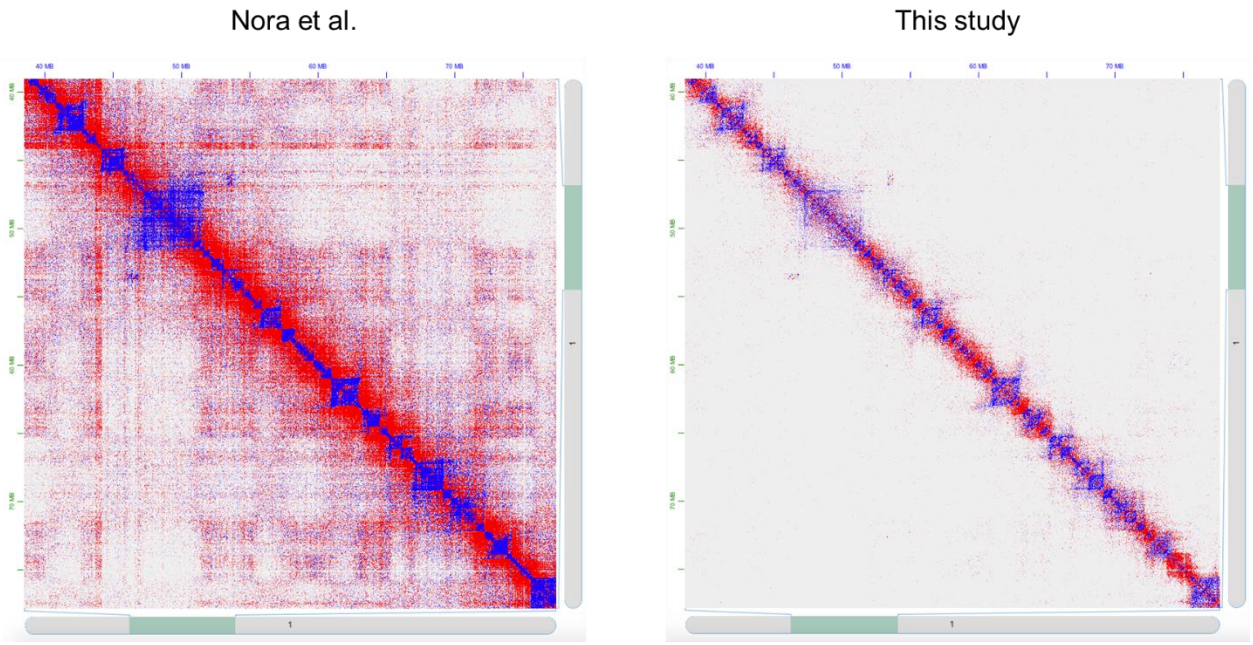


Fig.2.34 KO CTCF showed similar interaction loss on contact matrix compared to published results. KO CTCF / Wild type contact matrix is viewed in Juicebox. Regions in red indicates enrichment of interactions in KO CTCF while regions in blue shows enrichment of contacts in wild type.

We then tried to investigate the above 3 conditions on reducing chromatin contacts in details. First of all, we investigate domain boundary by looking at the differential of insulation score as well as DI value. The results here suggest KO CTCF leads to a significant loss on boundary strength, inhibition of PolII also decrease the domain boundary to a relatively less extent. Moreover, by association of CTCF, Rad21 expression level at the domain boundaries we could see a dramatic decrease for both two proteins once deletion of CTCF, however, inhibition of PolII only exhibits a slightly decrease for CTCF and Rad21 at the domain boundaries. Secondly, upon inhibition of transcription, about one quarter of domains disappear at 10 or 5Kb resolution, KO CTCF gives a more dramatic impact on reducing domain numbers to about half remaining at 10 or 5 Kb resolution. (Fig. 2.35) As for loops, same trend has been observed as half of the loops disappeared for PolII inhibition samples. However, loops are more affected in CTCF KO samples as only 10% loops remain after depletion of CTCF (Fig. 2.36). To make sure the

drop on loop numbers are not false positive, we did APA analysis on loop peaks, which showed a slightly decrease on APA score for PolII inhibition samples. On the other hand, APA score drops dramatically in those samples without CTCF. This result again proves that the presence of CTCF is more important for loop while transcription only has a less significant effect on stabilizing the loop (Fig. 2.37).

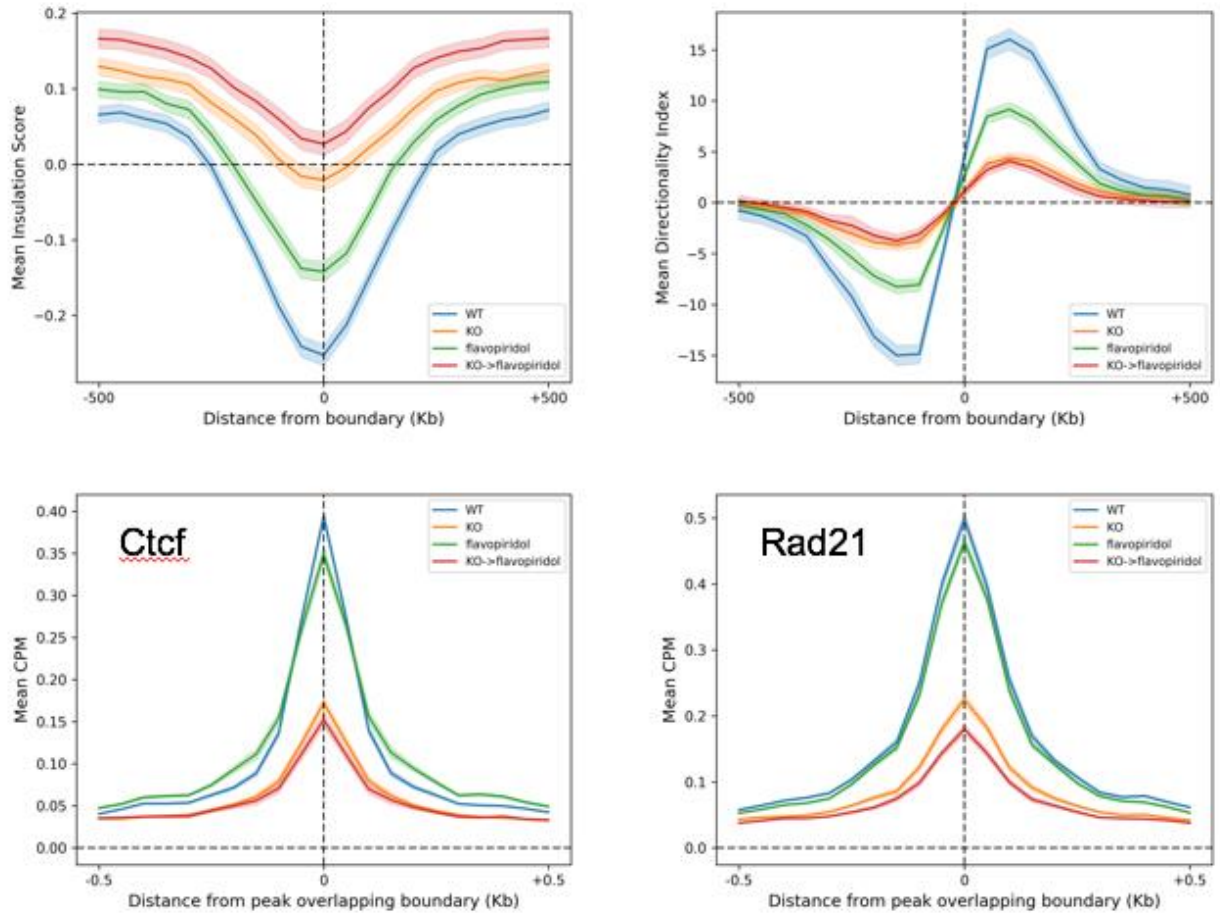


Fig.2.35 KO CTCF and RNAPII inhibition leads to different extent of domain boundary loss. Normalized insulation score and directional index value (DI) are centered around domain boundaries (+/- 500bp); CTCF and Rad21 expression level is normalized and centered around domain boundaries (+/- 500bp) for wild type (WT); RNAPII inhibition (flavopiridol); KO CTCF (KO); KO CTCF +RNAPII inhibition (KO-flavopiridol).

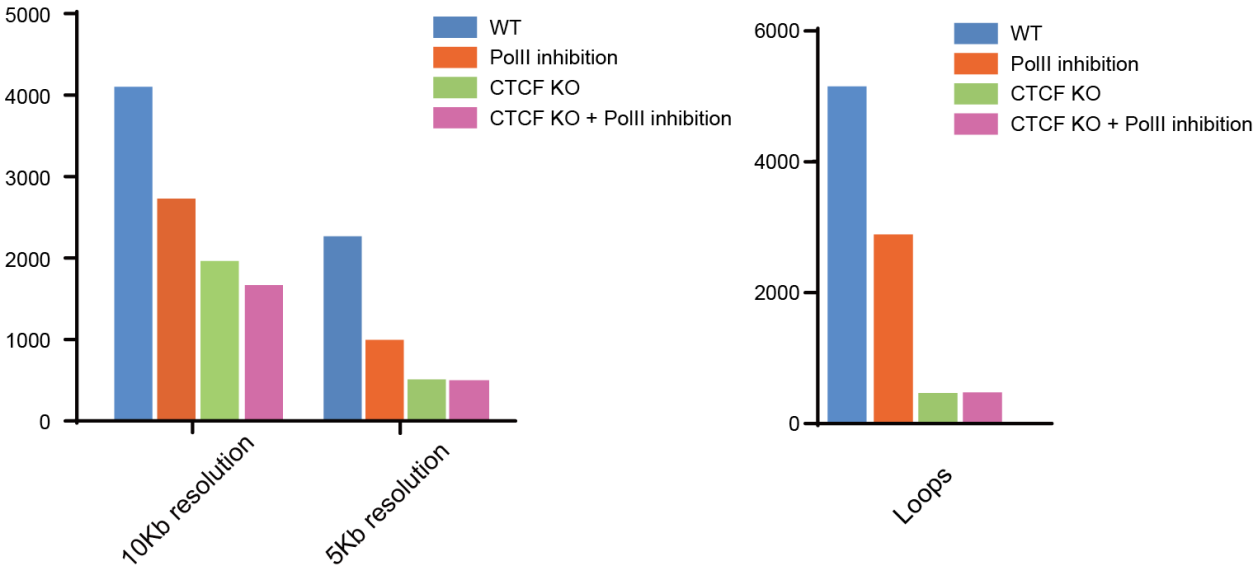


Fig.2.36 KO CTCF and RNAPII inhibition leads to different extent of domain and loop loss. Number of domains are called for 10Kb and 5Kb resolution with Arrowhead. High confidence loops are called with HiCCUPS.

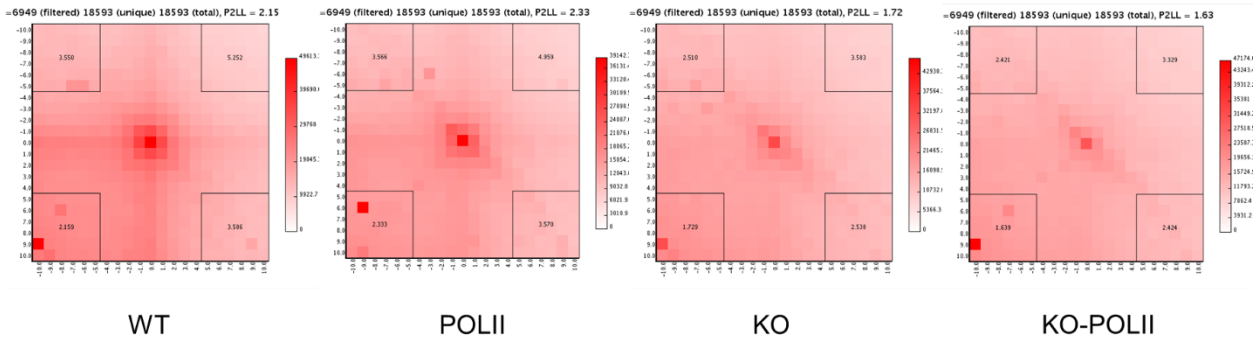


Fig.2.37 Loops are more affected by KO CTCF. APA analysis was performed on contact maps using loops called in the high-resolution CAP-C map. Aggregated signals showed the significant decrease of loops after KO CTCF while inhibition of PolII leads to less significant reduction.

As there are still loops decrease in PolII inhibition samples, we asked a question if it has anything to do with the drop on CTCF levels we see on ChIP-seq. As it is known that loop loss is result from CTCF depletion in KO samples, correlation of the loss of CTCF level with loss of loops showed a deviated curve. However, similar correlation curve showed no such deviation

pattern in PolII inhibition curve, suggesting that the loss of loops in PolII inhibition samples are not result from the decrease level of CTCF level (Fig. 2.38 and 2.39).

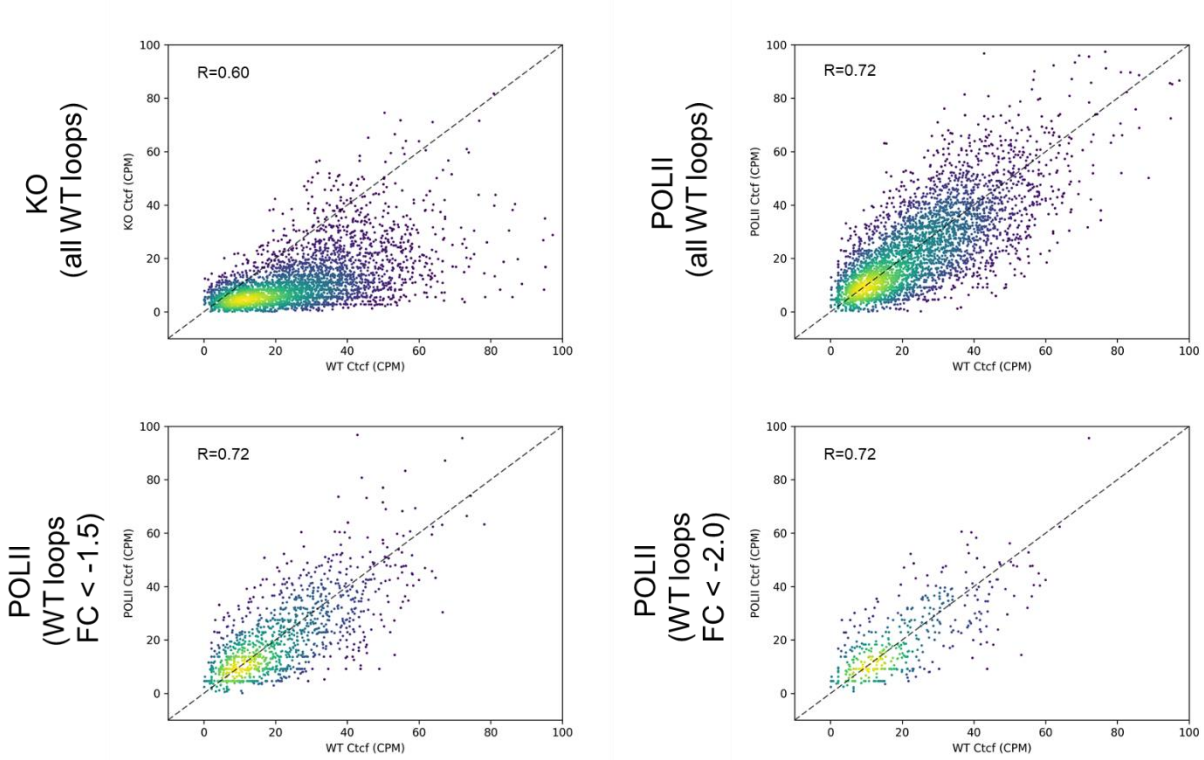


Fig.2.38 Decrease of loops are not result from reduction of CTCF around loop anchors by inhibition of transcription. CTCF expression level around loop anchors are associated with normalized signals around the same location on contact matrix. FC=Sample/wild type (WT)

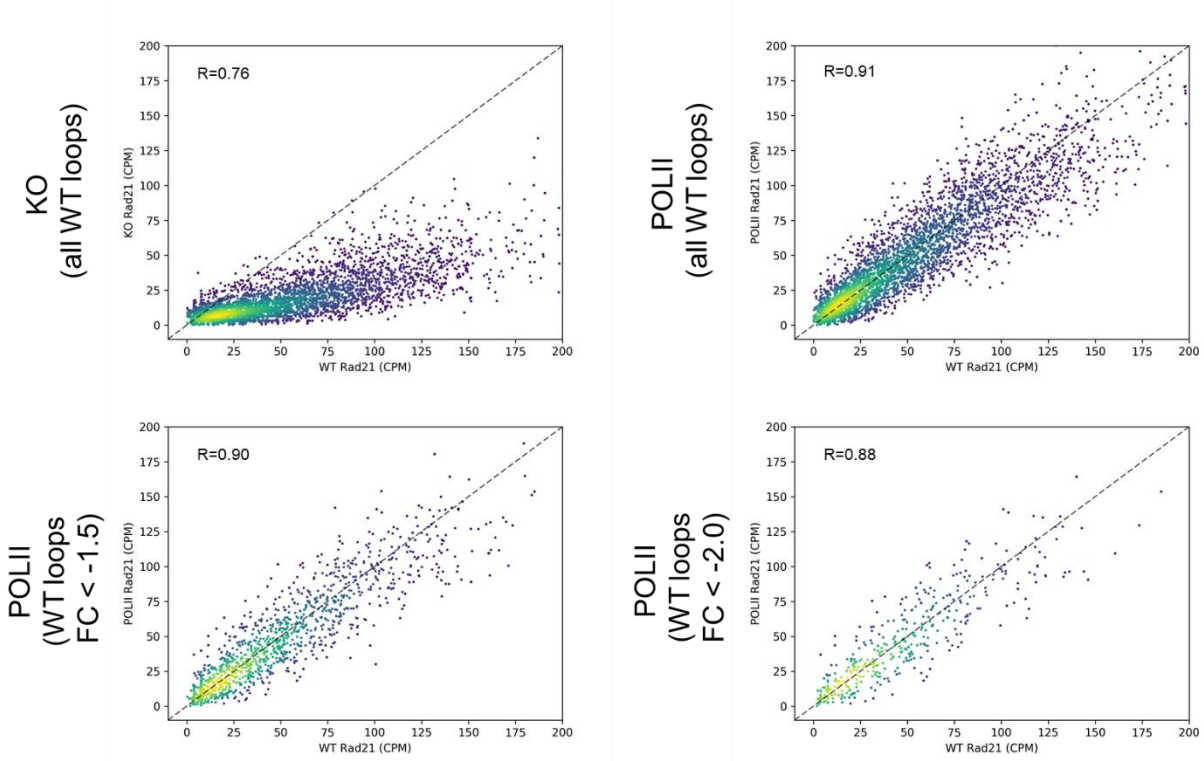


Fig.2.39 Decrease of loops are not result from reduction of cohesin around loop anchors by inhibition of transcription. Rad21 expression level around loop anchors are associated with normalized signals around the same location on contact matrix. FC=Sample/wild type (WT)

The results above seem to show that PolII inhibition and CTCF KO each contribute to chromatin change in a unique way. It encourages us to dig deeper by checking the global effect of PolII and CTCF on intra domain interactions and strength of loops. First, as for loops, we classify loops into two types based on their location in A/B compartment and discovered that CTCF plays a more important role on loop formation as KO CTCF leads to a much greater loss when compared to WT. PolII inhibition, however, did show some effect on decreasing the loop strength but to a less degree. Moreover, combine the two treatment gives us an even more loop decrease in both compartments. Next, we turned to check how PolII inhibition and KO CTCF might work on domain. 10Kb and 5Kb resolution domains are divided into two groups based on their location in A/B compartment. Intra-domain interactions are summarized for each condition

and compared to WT. For both resolution, inhibition of PolII has a more profound impacts on intra-domain interactions drop in active compartments, on the contrary, KO CTCF results a more significant loss for intra-domain interactions decrease in inactive compartments. This result is highly correlated to previous classification of domain boundary features for loop domain and non-loop domain. As non-loop domain boundaries are demarcated more with active promoters (Fig. 2.40). We next classify loop domain and non-loop domain and investigate the effect of PolII inhibition and KO CTCF on intra-domain interactions. In agreement with previous hypothesis, KO CTCF has little effect on non-loop domain as only for loop-domains in compartment B see a moderate intra-domain decrease. However, PolII inhibition have a more significant effects on intra-domain interaction loss in both compartments. As loop domain, the trend goes opposite as CTCF KO shows a profound decrease in intra-domain interactions in both compartments, especially for those domains in inactive compartment compared to PolII inhibition. Moreover. combined the two treatment give us an even more decrease in intra-domain interactions for all domains genome-wide. Taken together, these observations suggest that transcription and CTCF play a mutually exclusive role on domain formation but different in specific region. Transcription plays a more important role on active chromatin region while CTCF is more crucial on inactive chromatin region. In addition, CTCF is more important for loop formation while transcription seems to play a role in stabilizing loops based on our observations (Fig. 2.41).

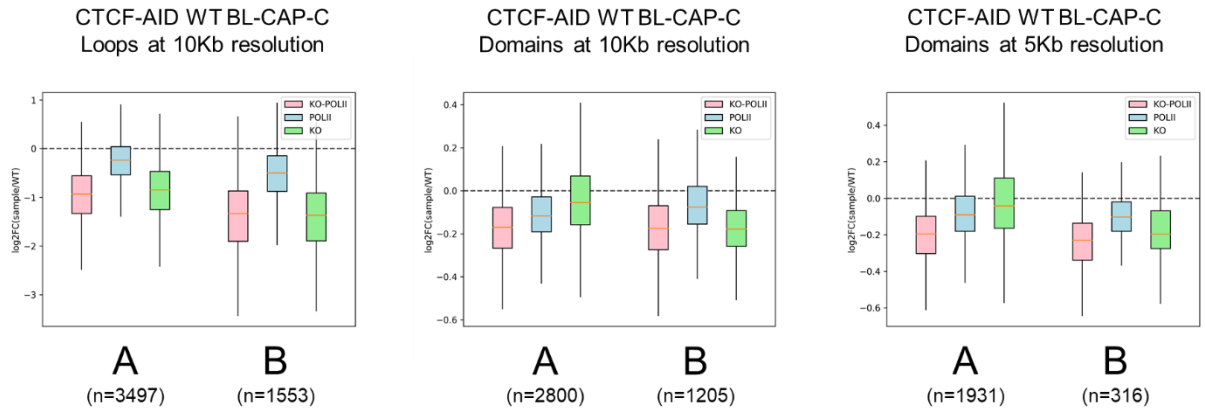


Fig.2.40 Transcription and CTCF each contribute to chromatin in different way. Normalized contacts are summed and divided into A/B compartment for 10kb loops. Intra-domain interactions are summed and divided into A/B compartment for domains at 10Kb or 5Kb resolution. N is the total number of loops or domains.

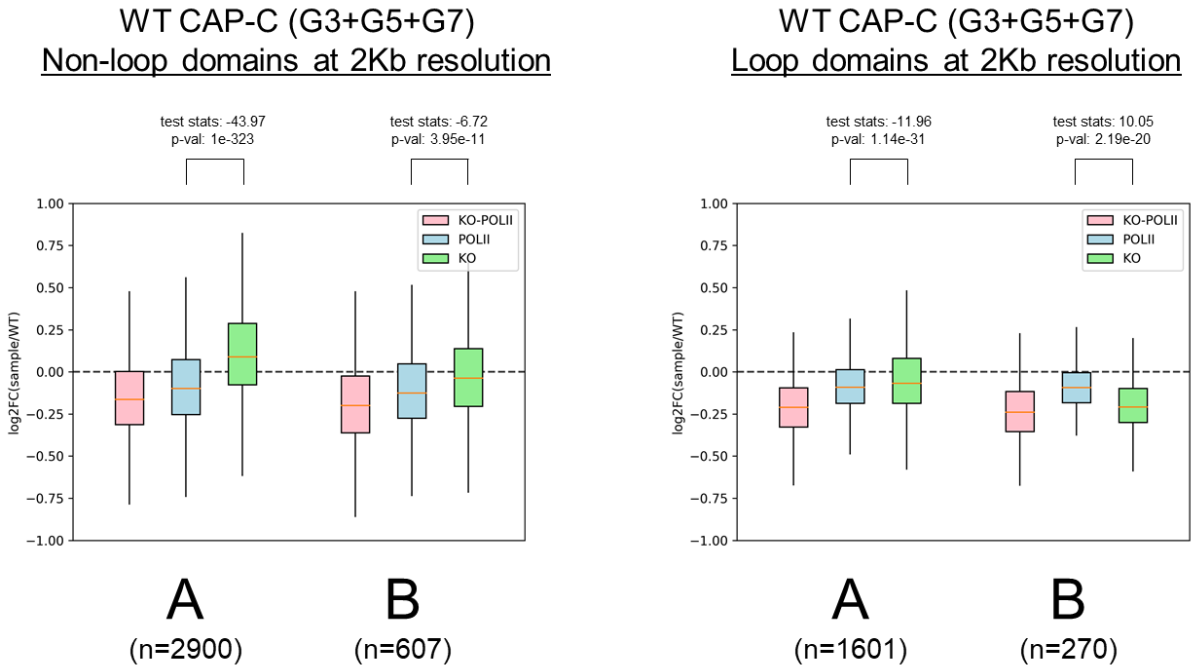


Fig.2.41 Transcription contributes more in Non-loop domains while CTCF is more responsible for loop domain. Intra-domain interactions are summed and divided into A/B compartment for loop domain and non-loop domain at 2Kb resolution. N is the total number of loops or domains.

We then turn to study another chromatin feature, stripe, and check how CTCF and transcription might work on its formation. Visual inspection of ultra-deep CAP-C maps with Juicebox

also showed a previous identified feature, stripe, where a single locus forms frequent contacts with a contiguous genomic interval. Stripes frequently appeared along the edges of domains with abundant cohesin loading and ranged from a few to hundreds of Kb. The stripe has been found to highly associate with transcription as 79% of stripe domains were enriched for active enhancers, including conventional and super enhancers (SEs) but repelled for poised enhancers. In addition, investigation of stripes between mESCs and B cells reveals different stripe distribution pattern, suggesting stripe is not conserve between cell types. Moreover, deletion of stripe anchors weakens the functional interactions between SE and promoters, leading to a reduction of corresponding gene expression. Lastly, depletion of ATP leads to decrease of loops as well as stripes (Fig.2.42).

As stripe is highly associate with transcription, we questioned whether inhibition of transcription will in turn affect stripe formation. As expected, upon KO CTCF, a previously characterized stripe anchor, eliminates most of the stripe domains. Intriguingly, inhibition of transcription also leads to a significant decrease of stripe compared to WT. We also found there is no significant change on CTCF level around the stripe anchor that has been weakened by inhibition of transcription. As stripe was thought to be the consequence of loop extrusion by cohesin and CTCF, loop extrusion is thought to be independent from transcription. Here, we demonstrate that CTCF mediated stripes, though remained after completely transcription shut down, are significantly weakened. The transcription could promote stripe formation or even loop extrusion independent of changing CTCF or cohesin level. We hypothesize that stripe is highly correlated with transcription. On the one hand, deplete stripe anchors like CTCF will weaken enhancer-promoter interactions to repress gene transcription. On the other hand, active transcription within the stripe promote the dynamic of loop extrusion and strengthen stripe formation in a feedback manner.

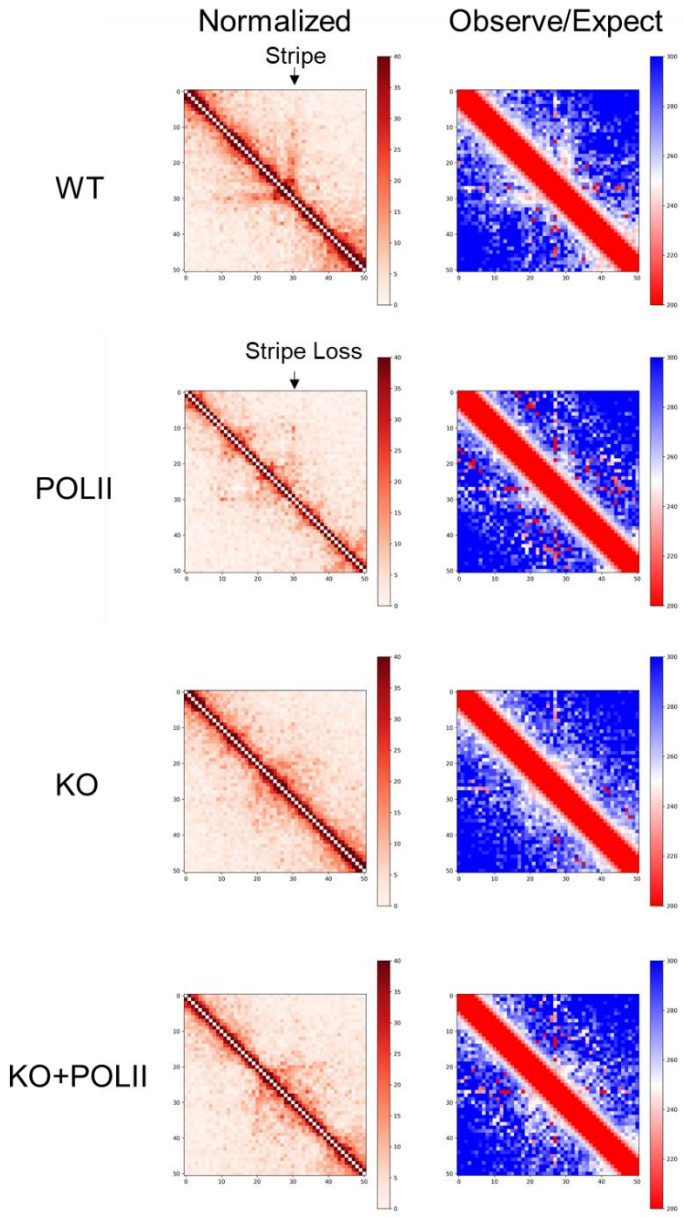


Fig.2.42 Transcription inhibition leads to stripe loss. Interaction counts and the log2 ratio of observed interactions divided by expected interactions for a given genomic distance are shown side by side for each treatment.

In addition to loop loss in transcription inhibition samples, we also discover a decent amount of loop appear in both PolII inhibition sample as well as PolII inhibition plus KO CTCF sample. However, there is no sign for new loop emerge in KO CTCF at the same location. By comparing with ChIP-Seq data we found no sign of CTCF or cohesin on the new loop anchors, suggesting

other proteins might mediate those loops formation. Although with no clear evidence, we hypothesized it to be YY1 mediated loops. Further exploration is required to study the mechanism of formation for these loops (Fig.2.43).

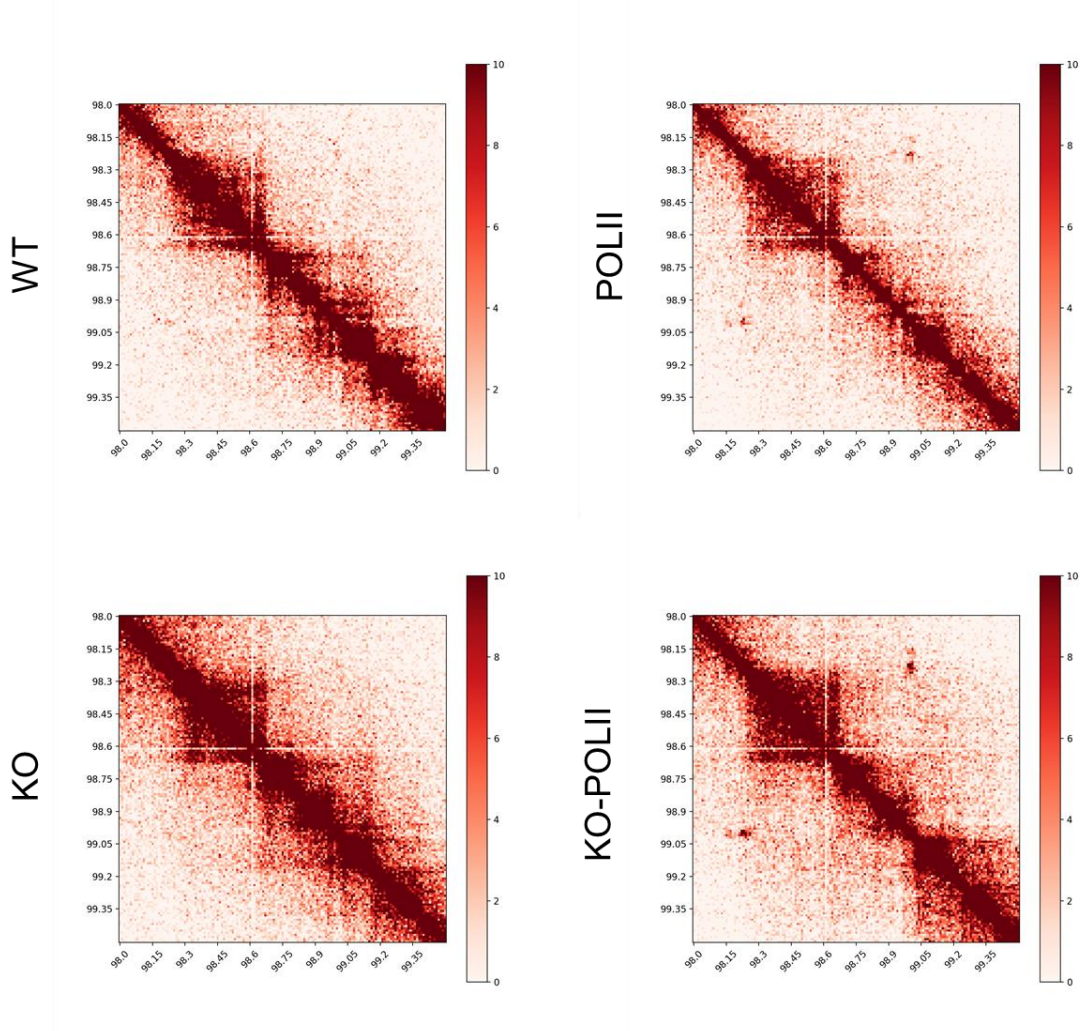


Fig.2.43 New loops emerge in transcription inhibitor treated samples.

2.3 Experiment section

2.3.1 Construction of plasmids for CTCF-AID mouse embryonic cell

The CRISPR/Cas9 plasmid was assembled using the Multiplex CRISPR/Cas9 Assembly System (*38*) kit (Addgene kit #1000000055). Oligonucleotides for three gRNA templates were synthesized, annealed and introduced into the corresponding intermediate vectors. The first gRNA

matches the genome sequence 23 bp upstream of the stop codon of mouse CTCF. The oligonucleotides with sequences (5'- CACCGTGATCCTCAGCATGATGGAC-3') and (5'-AAAC-GTCCATCATGCTGAGGATCAC-3') were annealed. The other two gRNAs direct in vivo linearization of the donor vector: the first pair of oligonucleotides are (5'-CACCGCTGAG-GATCATCTCAGGGGC -3') and (5'- AAACGCCCCTGAGATGATCCTCAGC -3'); the second pair are (5'- CACCGATGCTGGGGCCTTGCTGGC-3') and (5'-AAACGCCAG-CAAGGCCCCAGCATC-3'). The three gRNA-expressing cassettes were incorporated into one single plasmid using Golden Gate assembly. The donor vector was constructed using PCR and Gibson Assembly Cloning kit (New England Biolabs). The insert cassette includes sequences that codes for a 5GA linker, the auxin-induced degron (AID), a T2A peptide and the neomycin resistant marker, and is flanked by 24-bp homology arms to integrate into the CTCF locus. The left and right arms have sequences CCTGAGATGATCCTCAGCATGATG and GAC-CGGTGATGCTGGGGCCTTGCT, respectively. The AID coding sequence was amplified from pcDNA5-H2B-AID-EYFP (*39*) (Addgene plasmid #47329) and the T2A-neomycin was amplified from pAC95-pmax-dCas9VP160-2A-neo (*40*) (Addgene plasmid #48227). The sequence for the 5GA linker was included in one of the primers. The original donor backbone was a gift from Dr. Ken-ichi T. Suzuki from Hiroshima University, Hiroshima, Japan. The lentiviral vector for expressing TIR1 was constructed using PCR and Gibson Assembly Cloning kit (New England Biolabs). The backbone was modified from lentiCRISPR v2 (*41*) (a gift from Feng Zhang, Addgene plasmid #52961) and the TIR1-9myc fragment was amplified from pBabe TIR1-9myc (*39*) (a gift from Don Cleveland, Addgene plasmid #47328). The expressing cassette includes a puromycin resistant marker followed by sequences that code for P2A peptide and TIR1-9myc protein. The gene expression is driven by EFS promoter in the original lentiCRISPR v2.

2.3.2 General mouse embryonic cell culture

F123 mouse embryonic stem cells were grown on gamma-irradiated mouse embryonic fibroblast cells under standard conditions (85% High Glucose DMEM, 15% Knockout Serum Replacement, 0.1 mM non-essential amino acids, 0.1 mM β-mercaptoethanol, 1 mM Glutamine, LIF 500U/mL, +P/S). Before harvesting for CAP-C and *in-situ* Hi-C, F123 mESCs were passaged onto feeder free 0.2% gelatin coated plates for at least 2 passages to rid the culture of feeder cells.

2.3.3 Transfection and establishment of CTCF-AID knock-in clones

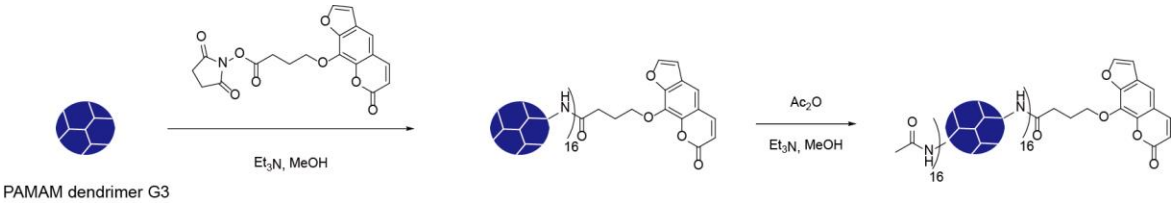
The cells were passaged once on 0.1% gelatin-coated feeder-free plates before transfection. The cells were transfected using the Mouse ES Cell Nucleofector Kit (Lonza) and Amaxa Nucleofector (Lonza) with 10 µg of the CRISPR plasmid and 5 µg of the donor plasmid following the manufacturer's instructions. After transfection, the cells were plated on drug-resistant MEFs (GlobalStem). Two days after transfection, drug selection was started by addition of 160 µg/ml G418 (Geneticin, Gibco) to the medium. Drug-resistant colonies were isolated and the clones with AID knock-in on both alleles were found by performing PCR of the genomic DNA using primers specific to sequences flanking the 3' end of the CTCF coding sequence (AAATGTTAAAGTGGAGGCCTGTGAG and AAGATTTGGGCCGTTTAAACACAGC). The sequence at the CTCF-AID junction on both alleles were checked by sequencing of allele-specific PCR products, which were generated by using either a CTCF-129-specific (CTGACTTGGGCATCACTGCTG) or a CTCF-Cast-specific (GTTTTGTTTCTGTTGACTTAGGCATCACTGTTA) forward primer and a reverse primer in the AID coding se-

quence (GAGGTTTGGCTGGATCTTTAGGACA). The expression of CTCF-AID fusion protein was confirmed by observing the difference in the molecular weight compared to the control cells by Western blot with anti-CTCF antibody (Millipore, 07-729).We produced the lentivirus for expressing TIR1-9myc using Lenti-X Packaging Single Shots system (Clontech) and infected the CTCF-AID knock-in mouse ES cells following the manufacturer's instructions. After infection, the cells were selected by culturing with 1 μg/ml puromycin. Drugresistant colonies were isolated and expression of TIR1-9myc was confirmed by Western blot using anti-Myc antibody (Santa Cruz, sc-40). Clones expressing high level of TIR1-9myc were used for the experiments.The CTCF-AID knock-in mouse ES cells expressing TIR1-9myc were passaged on 0.1% gelatincoated plates without MEFs. We added 1 ul 500 mM auxin (Abcam, ab146403) per 1 ml medium to deplete CTCF, and changed medium with auxin every 24 hours. Cells were harvested 48 hours after stating auxin treatment.

2.3.4 Synthesis of psoralen functionalized PAMAM dendrimer

PAMAM dendrimer G3, G5 and G7 (Sigma Aldrich) were dissolved in 2 mL methanol respectively. Half molar ratio of SPB (Thermo Fisher) was then added followed by 5 μl Et_3N. Each reaction mixture was stirred at r.t. for overnight before addition of 100 μl Ac_2O. The reaction was continuing for another 12 h. The reaction was quenched by addition of 3 mL ddH_2O. The modified PAMAM dendrimer was purified by centrifugation with 3 K Amicon Ultra Centrifugal Filter Unit (Millipore).

Below shows a synthetic scheme using PAMAM dendrimer G3 as an example:



PAMAM dendrimer G3

2.3.5 CAP-C

2.3.5.1 Fixing cells in situ

Grow five million cells under recommended culture conditions. Detach adherent cells by centrifugation at 300 x G for 5 min. Resuspend cells in fresh medium at 1 million cells per 1 ml medium. Add 16% formaldehyde solution to a final concentration of 1%, v/v. Incubate at r.t. for 5 min on rotating rocker.Add 2.5 M glycine solution to a final concentration of 0.2 M to quench the reaction. Incubate at r.t. for 5 min on rotating rocker. Centrifuge for 5 min at 300 x G at 4 ˚C. Discard supernatant. Resuspend cells in 1 ml of cold 1X PBS and spin for 5 min at 300 x G at 4 ˚C. Discard supernatant and flash-freeze cell pellets in liquid nitrogen (can be stored in -80 ˚C for up to a year).

2.3.5.2 UV crosslinking cells with dendrimers

Combine 250 µl of ice-cold lysis buffer (10 mM Tris-HCl, pH 8.0, 10 mM NaCl, 0.2% Ige-pal CA630) with 50 µl of protease inhibitors (Sigma, P8340). Add to formaldehyde fixed pellet of cells. Incubate cell suspension on ice for 20 min. Centrifuge at 2500xG for 5 min. Discard the supernatant. Wash pelleted nuclei once with 500 µl of ice-cold Hi-C lysis buffer. Centrifuge and discard the supernatant. Resuspend the cell pellet in 1 ml 10 uM dendrimer in methanol. Incubate at 4 ˚C on a rocker with rotation. Photo crosslink the nuclei by irradiating under 365 nm UV for 30 min. Centrifuge for 5 min at 2500 x G at 4 ˚C. Discard supernatant. Wash pelleted nuclei

twice with 500 µl of ice-cold Hi-C lysis buffer. Centrifuge and discard the supernatant. Resuspend the pellet in proteinase K buffer (420 µl Hi-C lysis buffer, 50 µl 10% SDS, 30 µl 20 mg/ml proteinase K) Incubate at 65 ˚C for O/N on a thermomixer at 800 rpm.

2.3.5.3 Purify UV crosslinked DNA-Dendrimer complexes:

Extract the DNA with 500 µl phenol:chloroform. Centrifuge at max for 10 min at r.t. Transfer the upper layer to a new tube. Add 800 µl EtOH and 50 µl 3 M NaOAc (pH 5.5). Incubate at -80 ˚C for 3-4 h. Centrifuge at max for 15 min at 4 ˚C. Discard the supernatant. Wash the pellet twice with 500 µl 70% EtOH. Centrifuge at max for 5 min at 4 ˚C. Discard the supernatant.

Resuspend the DNA pellet in 50 µl water.

2.3.5.4 Digest DNA-Dendrimer complexes with MboI:

50 µl of DNA-Dendrimer complex

20 µl of 10 X NEBuffer 2

20 µl of 5U/ µl MboI (NEB, R0147)

110 µl of ddH_2O

Incubate at 37˚C for O/N on a thermomixer at 800 rpm. On the next day, inactivate MboI by incubating at 65 ˚C for 20 min on a thermomixer at 800 rpm.

2.3.5.5 Marking DNA ends with Biotin

200 µl of above DNA-Dendrimer complex

37.5 µl of 0.4 mM biotin-14-dATP (Life Technologies, 19524-016)

1.5 µl of 10 mM dCTP

1.5 µl of 10 mM dGTP

1.5 µl of 10 mM dTTP

10 µl of 5 U/µl DNA Polymerase I, Large (Klenow) Fragment (NEB, M0210)

Incubate at 37 ˚C for 1 h on a rotating rocker. Inactivate the Klenow by incubating at 65 ˚C for 30 min on a thermomixer at 800 rpm.

2.3.5.6 Proximity Ligation in the ultra-diluted solution

6.2 ml of water

500 µl of 10X NEB T4 DNA ligase buffer (NEB, B0202)

300 µl of above DNA-Dendrimer complexes

12 µl of 10mg/ml Bovine Serum Albumin (100X BSA)

20 µl of 400 U/ µl T4 DNA Ligase (NEB, M0202)

Incubate at 16˚C for 8 h on a rotating rocker.

2.3.5.7 Purify DNA-Dendrimer complexes

Extract the DNA with 7 ml phenol:chloroform. Centrifuge at 2000 G for 10 min at r.t. Transfer the upper layer to a new tube. Add 17.5 ml EtOH and 700 µl 3 M NaOAc (pH 5.5). Incubate at -80 ºC for O/N. Centrifuge at 10000 G for 20 min at 4 ˚C. Discard the supernatant. Resuspend the pellet in 300 µl water.

2.3.5.8 Shear DNA-Dendrimer complexes

Instrument: Diagenode Bioruptor Pico

Volume of Library: 100 µl in a 0.65 ml Diagenode tube

Program: 30 s on; 30 s off; 8 cycles.

2.3.5.9 Library construction

Perform all the following steps in low-bind tubes. Prepare for biotin pull-down by washing 150 µl of 10 mg/ml Dynabeads MyOne Streptavidin C1 beads (Life technologies) with 400 µl of 1X Tween Washing Buffer (1X TWB: 5 mM Tris-HCl (pH 7.5); 0.5 mM EDTA; 1 M NaCl; 0.05% Tween 20). Separate on a magnet and discard the solution. Resuspend the beads in 300 µl of 2X Binding Buffer (2X BB: 10 mM Tris-HCl (pH 7.5); 1 mM EDTA; 2 M NaCl) and add to the reaction. Incubate at room temperature for 15 min with rotation to bind biotinylated DNA to the streptavidin beads. Separate on a magnet and discard the solution. Wash the beads by adding 600 µl of 1X TWB and transferring the mixture to a new tube. Heat the tubes on a Thermomixer at 55 °C for 2 min with mixing. Reclaim the beads using a magnet. Discard supernatant. Repeat wash. Resuspend beads in 100 ul 1X NEB T4 DNA ligase buffer (NEB, B0202) and transfer to a new tube. Reclaim beads and discard the buffer. To repair ends of sheared DNA and remove biotin from unligated ends, resuspend beads in 100 µl of master mix: 88 µl of 1X NEB T4 DNA ligase buffer with 10 mM ATP S33, 2 µl of 25mM dNTP mix, 5 µl of 10 U/µl NEB T4 PNK (NEB, M0201), 4 µl of 3 U/µl NEB T4 DNA polymerase I (NEB, M0203), 1 µl of 5 U/µl NEB DNA polymerase I, Large (Klenow) Fragment (NEB, M0210) Incubate at room temperature for 30 min. Separate on a magnet and discard the solution. Wash the beads by adding 600 µl of 1X TWB and transferring the mixture to a new tube. Heat the tubes on a Thermomixer at 55°C for 2 min with mixing. Reclaim the beads using a magnet. Discard supernatant. Repeat wash. Resuspend beads in 100 µl 1X NEBuffer 2 and transfer to a new tube. Reclaim beads and discard the buffer. Resuspend beads in 100 µl of dATP attachment master mix: 90 µl of 1X NEBuffer 2, 5 µl of 10 mM dATP, 5 µl of 5 U/µl NEB Klenow exo minus (NEB, M0212). Incubate at 37°C for 30 min. Separate on a magnet and discard the solution. Wash the beads by adding 600 µl of 1X

TWB and transferring the mixture to a new tube. Heat the tubes on a Thermomixer at 55°C for 2 min with mixing. Reclaim the beads using a magnet. Discard supernatant. Repeat wash. Resuspend beads in 100 µl 1X Quick ligation reaction buffer (NEB, B6058) and transfer to a new tube. Reclaim beads and discard the buffer. Resuspend in 50 µl of 1X NEB Quick ligation reaction buffer. Add 2 µl of NEB DNA Quick ligase (NEB, M2200). Add 3 µl of Illumina indexed adapter. (Nextflex) Record the sample-index combination. Mix thoroughly. Incubate at room temperature for 15 min. Separate on a magnet and discard the solution. Wash the beads by adding 600 µl of 1X TWB and transferring the mixture to a new tube. Heat the tubes on a Thermomixer at 55°C for 2 min with mixing. Reclaim the beads using a magnet. Remove supernatant. Repeat wash. Wash 3 times with 100 µl water. Reclaim the beads with 50 µl water. Incubate at 98 ˚C for 10min to elute the DNA from the beads. Transfer the supernatant to a 8-well PCR tube.

PCR amplify 12 cycles with following conditions:

98 ˚C 30 s ;98 ˚C 15 s ;60 ˚C 30 s; 72 ˚C 30 s; Repeat 12 cycles; 72 ˚C 1 min.

Purify the libraries with 0.9X Ampure beads. Elute with 30 µl water. Check the ligation efficiency by aliquote 8 µl DNA libraries and adding 1 µl 10X CutSmart buffer, 1 µl BspdI. Incubate at 37 ºC for 1 h. Run a 2% agarose gel with digested libraries and original libraries side by side. A clear shift-down to small size should be observed with BspdI digested libraries.

2.3.6 *In-situ* Hi-C

In-situ Hi-C experiments were performed as previously described using the MboI restriction enzyme (Rao et al., 2014). For fixation, trypsinized mES cells were counted and resuspended with adjusted volume (1 million cells per mL) of fresh media. Formaldehyde was added to 1% final, and samples were incubated for 10 minutes at room temperature. Then, 25 µl per mL of

2.5M Glycine was added followed by a 5-minute incubation at room temperature and then a 15-minute incubation on ice. Aliquots were spun down at 3,500 x g for 15 minutes, washed with cold 1 x PBS, frozen on dry ice, and stored at -80˚C. The crosslinked pellets were thawed on ice, and were incubated with 200ul of lysis buffer (10 mM Tris-HCl pH 8.0, 10 mM NaCl, 0.2% Igepal CA630, 33 µL Protease Inhibitor (Sigma, P8340)) on ice for 15 min, washed with 300 µL cold lysis buffer, and then incubated in 50uL of 0.5% SDS for 10min at 62°C. After heating, 170 µL of 1.47% Triton X-100 was added and incubated for 15min at 37˚C. To digest chromatin 100U MboI and 25uL of 10X NEBuffer2 were added followed by overnight incubation at 37˚C with agitation at 700rpm on a thermomixer. After incubation, MboI was inactivated by heating at 62°C for 20 minutes. Digestion efficiency was confirmed by performing agarose gel electrophoresis of the samples. The digested ends were filled and labeled with biotin by adding 37.5uL of 0.4mM biotin-14-dATP (Life Tech), 1.5 µL of 10mM dCTP, 10mM dTTP, 10mM dGTP, and 8uL of 5U/ul Klenow (New England Biolabs) and incubating at 23°C for 60 minutes with shaking at 500 rpm on a thermomixer. Then the samples were mixed with 1x T4 DNA ligase buffer (New England Biolabs), 0.83% Trition X100, 0.1 mg/mL BSA, 2000U T4 DNA Ligase (New England Biolabs, M0202), and incubated for at 23°C for 4 hours with shaking at 300rpm on a thermomixer to ligate the ends. After the ligation reaction, samples were spun and pellets were resuspended in 550uL 10 mM Tris-HCl, pH 8.0. To digest the proteins and to reverse the cross-links, 50 µL of 20mg/mL Proteinase K (New England Biolabs) and 57 µL of 10% SDS were mixed with the samples, and incubated at 55°C for 30 minutes, and then 67 µL of 5M NaCl were added followed by overnight incubation at 68°C. After cooling the samples for 10 minutes at room temperature, 0.8X Ampure (Beckman-Coulter) purification was performed and samples

were eluted in 100 μL 10 mM Tris-HCl, pH 8.0. Next, the samples were sonicated to mean fragment length of 400 bp using Covaris M220 with the following parameters: 70 seconds duration at 10.0% duty factor, 50.0 peak power, 200 cycles per burst. To collect 200-600 bp size of fragmented DNA, two rounds of Ampure (Beckman-Coulter) beads purification was performed and the samples were eluted in 300 μL 10 mM Tris-HCl, pH 8.0. The DNA labeled with biotin was purified using Dynabeads My One T1 Streptavidin beads (Invitrogen). 100 μL of 10 mg/mL Dynabeads My One T1 Streptavidin beads was washed with 400 μL of 1x Tween Wash Buffer (5 mM Tris-HCl pH 7.5, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween-20), and resuspended in 300 μL of 2x Binding Buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA, 2 M NaCl). The beads were transferred to the sample tube, incubated for 15 minutes at room temperature, and the supernatant was removed from the beads using a magnetic rack. Then the beads were washed twice by adding 600 μL of 1x Tween Wash Buffer, heating on a thermomixer for 2 minutes at 55˚C with mixing, and removing the supernatant using a magnetic rack. Then the beads were equilibrated once in 100 uL 1x NEB T4 DNA ligase buffer (New England Biolabs) followed by removal of the supernatant using a magnetic rack. To repair the fragmented ends and remove biotin from unligated ends, the beads were resuspended in 100uL of the following: 88 μL 1X NEB T4 DNA ligase buffer (New England Biolabs, B0202), 2 μL of 25mM dNTP mix, 5 μL of 10 U/μL T4 PNK (New England Biolabs), 4 μL of 3 U/μL NEB T4 DNA Polymerase (New England Biolabs), 1 μL of 5U/μL Klenow (New England Biolabs). The beads were incubated for 30 minutes at room temperature, followed by removal of the supernatant using a magnetic rack. The beads were washed twice by adding 600 μL of 1x Tween Wash Buffer, heating on a thermomixer for 2 minutes at 55˚C with mixing, and removing the supernatant using a magnetic rack. To add dA-tail, the beads were resuspended in 100uL of the following: 90 μL of 1X NEB Buffer2, 5 μL of

10mM dATP, and 5 μL of 5U/ul Klenow (exo-) (New England Biolabs). The beads were incubated for 30 minutes at 37˚C, followed by removal of the supernatant using a magnetic rack. The beads were washed twice by adding 600 μL of 1x Tween Wash Buffer, heating on a thermomixer for 2 minutes at 55˚C with mixing, and removing the supernatant using a magnetic rack. Following the washes, the beads were equilibrated once in 100 μL 1x NEB Quick Ligation Reaction Buffer (New England Biolabs) and the supernatants were removed using a magnetic rack. Then the beads were resuspended again in 50 μL 1x NEB Quick Ligation Reaction Buffer. To ligate adapters, 2 μL of NEB DNA Quick Ligase (New England Biolabs) and 3 μL of Illumina Indexed adapter were added to the beads and incubated for 15 minutes at room temperature. The supernatant was removed using a magnetic rack and the beads were washed twice by adding 600 μL of 1x Tween Wash Buffer, heating on a thermomixer for 2 minutes at 55˚C with mixing, and removing the supernatant using a magnetic rack. Then the beads were resuspended once in 100 μL 10 mM Tris-HCl, pH 8.0, followed by removal of the supernatant and resuspension again in 50 μL 10 mM Tris-HCl, pH 8.0. After deciding an optimal PCR cycle number using KAPA DNA Quantification kit (Kapa Biosystems), 8-9 cycles of PCR amplification was performed with the following: 10 μL Fusion HF Buffer (New England Biolabs), 3.125 μL 10uM TruSeq Primer 1, 3.125 μL 10uM TruSeq Primer 2, 1 μL 10mM dNTPs, 0.5 μL Fusion HotStartII, 20.75 μL ddH20, 11.5 μL Bead-bound HiC library. Then, PCR products underwent final purification using AMPure beads (Beckman-Coulter) and were eluted in 30 μL 10 mM Tris-HCl, pH 8.0. Libraries were sequenced on Illumina HiSeq 4000.

2.3.7 Modified CAP-C

2.3.7.1 Crosslink cells with dendrimers

For modified CAP-C with formaldehyde crosslinking:

Grow five million cells under recommended culture conditions. Detach adherent cells by centrifugation at 300 x G for 5 min. Resuspend cells in fresh medium at 1 million cells per 1 ml medium. Add 16% formaldehyde solution to a final concentration of 1%, v/v. Incubate at r.t. for 5 min on rotating rocker. Add 2.5 M glycine solution to a final concentration of 0.2 M to quench the reaction. Incubate at r.t. for 5 min on rotating rocker. Centrifuge for 5 min at 300 x G at 4 ˚C. Discard supernatant. Resuspend cells in 1 ml of cold 1X PBS and spin for 5 min at 300 x G at 4 ˚C. Discard supernatant and flash-freeze cell pellets in liquid nitrogen (can be stored in -80 ˚C for up to a year). Combine 250 µl of ice-cold lysis buffer (10 mM Tris-HCl, pH 8.0, 10 mM NaCl, 0.2% Igepal CA630) with 50 µl of protease inhibitors (Sigma, P8340). Add to formaldehyde fixed pellet of cells. Incubate cell suspension on ice for 20 min. Centrifuge at 2500xG for 5 min. Discard the supernatant. Wash pelleted nuclei once with 500 µl of ice-cold Hi-C lysis buffer. Centrifuge and discard the supernatant. Resuspend the cell pellet in 1 ml 50 uM dendrimer in methanol. Incubate at 4 ˚C on a rocker with rotation. Photo crosslink the nuclei by irradiating under 365 nm UV for 30 min. Centrifuge for 5 min at 2500 x G at 4 ˚C. Discard supernatant. Wash pelleted nuclei twice with 500 µl of ice-cold Hi-C lysis buffer. Centrifuge and discard the supernatant. Resuspend the pellet in proteinase K buffer (420 µl Hi-C lysis buffer, 50 µl 10% SDS, 30 µl 20 mg/ml proteinase K) Incubate at 65 ˚C for O/N on a thermomixer at 800 rpm.

For modified CAP-C without formaldehyde crosslinking:

Grow five million cells under recommended culture conditions. Detach adherent cells by centrifugation at 300 x G for 5 min. Combine 250 µl of ice-cold nucleus lysis buffer (10mM Tris, pH7.5, 10mM NaCl, 3mM MgCl2, 0.5% NP-40, 0.15mM spermine, 0.5mM spermidine) with 50 µl of protease inhibitors (Sigma, P8340). Add to pellet of cells. Incubate cell suspension on ice for 5 min. Centrifuge at 500xG for 5 min. Discard the supernatant. Wash pelleted nuclei

once with 500 µl of resuspension buffer (10mM Tris-HCl pH7.4, 15mM NaCl, 60mM KCl, 0.15mM spermine, 0.5mM spermidine). Centrifuge at 500xG for 5 min and discard the supernatant. Resuspend the cell pellet in 1 ml 50 uM dendrimer in methanol. Incubate at 4 ˚C on a rocker with rotation for 10min. Photo crosslink the nuclei by irradiating under 365 nm UV for 30 min. Centrifuge for 5 min at 2500 x G at 4 ˚C. Discard supernatant. Wash pelleted nuclei twice with 500 µl of resuspension buffer. Centrifuge and discard the supernatant. Resuspend the pellet in proteinase K buffer (420 µl Hi-C resuspension buffer, 50 µl 10% SDS, 30 µl 20 mg/ml proteinase K) Incubate at 65 ˚C for O/N on a thermomixer at 800 rpm.

2.3.7.2 Purify UV crosslinked DNA-Dendrimer complexes:

Extract the DNA with 500 µl phenol:chloroform. Centrifuge at max for 10 min at r.t. Transfer the upper layer to a new tube. Add 800 µl EtOH and 50 µl 3 M NaOAc (pH 5.5). Incubate at -80 ˚C for 1 h. Centrifuge at max for 15 min at 4 ˚C. Discard the supernatant. Wash the pellet twice with 500 µl 70% EtOH. Centrifuge at max for 5 min at 4 ˚C. Discard the supernatant.

Resuspend the DNA pellet in 100 µl MNase digestion buffer (10mM Tris-HCl pH7.4, 15mM NaCl, 60mM KCl, 1mM $CaCl_2$, 0.15mM spermine, 0.5mM spermidine).

2.3.7.3 Digest DNA-Dendrimer complexes with MNase:

100 µl of DNA-Dendrimer complex

1 unit of MNase

Incubate at 37˚C for 5 min then stop the reaction by adding 150µl of Stop Buffer. (20mM EDTA, 20mM EGTA, 0.4% SDS) Incubate the mixture at 65°C for 30min. Purify DNA with ethanol precipitation by adding 800 µl EtOH and 50 µl 3 M NaOAc (pH 5.5). Incubate at -80 ˚C

for 1 h. Centrifuge at max for 15 min at 4 ˚C. Discard the supernatant. Wash the pellet twice with 500 µl 70% EtOH. Centrifuge at max for 5 min at 4 ˚C. Discard the supernatant. Resuspend the DNA pellet in 100 µl H_2O.

2.3.7.4 Repair end and add "A"

Repair DNA ends and add "A" using the KAPA Hyper plus kit.

100 µl of above DNA-Dendrimer complex

28 µl of ER&AT buffer mix

12 µl of ER&AT enzyme mix

Incubate at 20˚C for 30 min then 65˚C for 30 min. Purify DNA with ethanol precipitation by adding 500 µl EtOH and 20 µl 3 M NaOAc (pH 5.5). Incubate at -80 ˚C for 1 h. Centrifuge at max for 15 min at 4 ˚C. Discard the supernatant. Wash the pellet twice with 500 µl 70% EtOH. Centrifuge at max for 5 min at 4 ˚C. Discard the supernatant. Resuspend the DNA pellet in 100 µl H_2O.

2.3.7.5 Attach biotin linker to the dendrimer

100 µl of above DNA-Dendrimer complex

2 µl of 100 µM biotin linker

Incubate at 37˚C for 2 h on a thermomixer at 800 rpm. Excess of biotin linkers are removed by XP beads size selection. DNA is eluted with 100 µl of H_2O.

2.3.7.6 Immobilize dendrimer-DNA complex on streptavidin beads

Prepare for biotin pull-down by washing 20 µl of 10 mg/ml Dynabeads MyOne Streptavidin C1 beads (Life technologies) with 400 µl of 1X Tween Washing Buffer (1X TWB: 5 mM Tris-

HCl (pH 7.5); 0.5 mM EDTA; 1 M NaCl; 0.05% Tween 20). Separate on a magnet and discard the solution. Resuspend the beads in 100 µl of 2X Binding Buffer (2X BB: 10 mM Tris-HCl (pH 7.5); 1 mM EDTA; 2 M NaCl) and add to the reaction. Incubate at room temperature for 15 min with rotation to bind biotinylated DNA to the streptavidin beads. Separate on a magnet and discard the solution. Wash the beads by adding 600 µl of 1X TWB and transferring the mixture to a new tube. Heat the tubes on a Thermomixer at 55 °C for 2 min with mixing. Reclaim the beads using a magnet. Discard supernatant. Repeat wash.

2.3.7.7 Proximity Ligation in the ultra-diluted solution

4 ml of water

500 µl of 10X NEB T4 DNA ligase buffer (NEB, B0202)

1 ml of above DNA-Dendrimer complexes

20 µl of 400 U/ µl T4 DNA Ligase (NEB, M0202)

Incubate at 16˚C for overnight on a rotating rocker. Separate on a magnet and discard the solution.

2.3.7.8 Purify DNA-Dendrimer complexes and library construction

Wash the beads by adding 600 µl of 1X TWB and transferring the mixture to a new tube. Heat the tubes on a Thermomixer at 55 °C for 2 min with mixing. Reclaim the beads using a magnet. Discard supernatant. Repeat wash. Perform all the following steps in low-bind tubes. Resuspend beads in 100 ul 1X NEB T4 DNA ligase buffer (NEB, B0202) and transfer to a new tube. Reclaim beads and discard the buffer. To repair ends of sheared DNA and remove biotin from unligated ends, resuspend beads in 100 µl of master mix: 88 µl of 1X NEB T4 DNA ligase buffer with 10 mM ATP S33, 2 µl of 25mM dNTP mix, 5 µl of 10 U/µl NEB T4 PNK (NEB, M0201), 4 µl of 3 U/µl NEB T4 DNA polymerase I (NEB, M0203), 1 µl of 5 U/µl NEB DNA

polymerase I, Large (Klenow) Fragment (NEB, M0210) Incubate at room temperature for 30 min. Separate on a magnet and discard the solution. Wash the beads by adding 600 µl of 1X TWB and transferring the mixture to a new tube. Heat the tubes on a Thermomixer at 55°C for 2 min with mixing. Reclaim the beads using a magnet. Discard supernatant. Repeat wash. Resuspend beads in 100 µl 1X NEBuffer 2 and transfer to a new tube. Reclaim beads and discard the buffer. Resuspend beads in 100 µl of dATP attachment master mix: 90 µl of 1X NEBuffer 2, 5 µl of 10 mM dATP, 5 µl of 5 U/µl NEB Klenow exo minus (NEB, M0212). Incubate at 37°C for 30 min. Separate on a magnet and discard the solution. Wash the beads by adding 600 µl of 1X TWB and transferring the mixture to a new tube. Heat the tubes on a Thermomixer at 55°C for 2 min with mixing. Reclaim the beads using a magnet. Discard supernatant. Repeat wash. Resuspend beads in 100 µl 1X Quick ligation reaction buffer (NEB, B6058) and transfer to a new tube. Reclaim beads and discard the buffer. Resuspend in 50 µl of 1X NEB Quick ligation reaction buffer. Add 2 µl of NEB DNA Quick ligase (NEB, M2200). Add 3 µl of Illumina indexed adapter. (Nextflex) Record the sample-index combination. Mix thoroughly. Incubate at room temperature for 15 min. Separate on a magnet and discard the solution. Wash the beads by adding 600 µl of 1X TWB and transferring the mixture to a new tube. Heat the tubes on a Thermomixer at 55°C for 2 min with mixing. Reclaim the beads using a magnet. Remove supernatant. Repeat wash. Wash 3 times with 100 µl water. Reclaim the beads with 50 µl water. Incubate at 98 ˚C for 10min to elute the DNA from the beads. Transfer the supernatant to an 8-well PCR tube.

PCR amplify 12 cycles with following conditions:

98 ˚C 30 s ;98 ˚C 15 s ;60 ˚C 30 s; 72 ˚C 30 s; Repeat 12 cycles; 72 ˚C 1 min.

Purify the libraries with 0.9X Ampure beads. Elute with 30 µl water. Check the ligation efficiency by aliquote 8 µl DNA libraries and adding 1 µl 10X CutSmart buffer, 1 µl BspdI. Incubate at 37 ºC for 1 h. Run a 2% agarose gel with digested libraries and original libraries side by side. A clear shift-down to small size should be observed with EcoRV digested libraries.

2.3.7.9 ChIP-Seq

ChIP-seq experiments was performed as described in ENCODE experiments protocols (*42*) with minor modifications. Cells were crosslinked with 1% formaldehyde for 10 min and quenched with 200 mM glycine. 5.0 million mESCs were used for each ChIP sample. Shearing of chromatin was performed using Diagenode Bioruptor Pico for sonication with following parameters: 30 s on; 30 s off; 30 cycles at 4°C. The concentration of fragmented DNA was diluted to 0.1 µg/µl with 1xTE. For immunoprecipitation, we used 20 µL anti-mouse IgA Dynabeads (Life Technologies) and wash them with cold BSA/PBS (0.5 mg / mL bovine serum albumin in 1x phosphate buffered saline) for 3 times. After washing, 3 µL anti-Top2B (Santa Cruz) with 147 µL cold BSA/PBS were added to the beads and incubated on a rotating platform at 4°C for 2 hours. After incubation, beads were washed with150 mL cold BSA/PBS for 3 times, and mixed with 100 µL Binding Buffer (1% Triton X-100, 0.1% Sodium Deoxycholate, 1x complete protease inhibitor (Roche)) plus 100 µL 0.2 µg/µl chromatin followed by overnight incubation on a rotating platform at 4°C. Beads were collected on a magnetic rack and washed 5 times with 50 mM Hepes pH 8.0, 1% NP-40, 1 mM EDTA, 0.70% Sodium Deoxycholate, 0.5 M LiCl, 1x complete protease inhibitor (Roche) and washed once with 150 µL cold 1x TE. After removing the TE, 150 µL ChIP elution buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA, 1% SDS) was added, and samples were incubated at 65°C for 4 h at 800 rpm on a thermomixer. The beads were dis-

carded using a magnetic rack and the samples were further incubated at 65°C overnight to reverse crosslinks. For input samples, 20 μL of chromatin was added to 130 μL ChIP elution buffer and incubated at 65°C overnight with the other samples. The input samples were processed in parallel with the ChIP samples from here on. RNase A was added to each sample and incubated at 37°C for 1 hour, and Proteinase K was added and incubated at 55°C for 1 hour. The samples were extracted with phenol: chloroform: isoamyl alcohol (25:24:1) using Phase Lock tube (5 Prime). Then the samples were precipitated with ethanol and resuspended in 50 μL10 mM Tris (pH 8.0). The KAPA Hyper plus Kit (KAPA) was used to for preparing Illumina sequencing libraries. Libraries were sequenced on HiSeq4000 single end for 50 bp.

2.3.8 HiChIP

Pellet detached adherent or suspension cells and resuspend in freshly made 1% formaldehyde (methanol free) at a volume of 1 mL of formaldehyde for every one million cells. Incubate cells at room temperature for 10 minutes with rotation. Add glycine to a final concentration of 125 mM to quench the formaldehyde, and then incubate at room temperature for 5 minutes with rotation. Pellet cells, wash in PBS, pellet again, and then store in -80°C or proceed into the HiChIP protocol. Resuspend up to 15 million crosslinked cells in 500 μL of ice-cold Hi-C Lysis Buffer and rotate at 4 for 30 minutes. For cell amounts greater than 15 million, split the pellet in half for contact generation and then recombine for the sonication. Spin down at 2500 rcf for 5 minutes and discard the supernatant. Wash pelleted nuclei once with 500 μL of ice-cold Hi-C Lysis Buffer. Remove the supernatant and resuspend pellet in 100 μL of 0.5% SDS. Incubate at 62°C for 10 minutes and then add 285 μL of H_2O and 50 μL of 10% Triton X-100 to quench the SDS. Mix well and incubate at 37°C for 15 minutes. Add 50 μL of 10X NEB Buffer 2 and 375 U

of MboI restriction enzyme (NEB, R0147), and digest chromatin for 2 hours at 37°C with rotation. For lower starting material, less restriction enzyme was used: 15 μL was used for 10-15 million cells, 8 μL for 5 million cells, and 4 μL for 1 million cells. Heat inactivate MboI at 62°C for 20 min. To fill in the restriction fragment overhangs and mark the DNA ends with biotin, add 52 μL of fill-in master mix: Fill-in Master Mix 1 Reaction 0.4 mM biotin-dATP (Thermo 19524016) 37.5 μL 10 mM dCTP 1.5 μL 10 mM dGTP 1.5 μL 10 mM dTTP 1.5 μL 5U/μL DNA Polymerase I, Large (Klenow) Fragment (NEB, M0210) 10 μL. Mix and incubate at 37°C for 1 hour with rotation. Add 948 μL of ligation master mix: Ligation Master Mix (10X NEB T4 DNA ligase buffer with 10 mM ATP (NEB, B0202) 10% Triton X-100 125 μL 50 mg/mL BSA 3 μL 400 U/μL T4 DNA Ligase (NEB, M0202) 10 μL Water) Incubate at room temperature for 4 hours with rotation. Pellet nuclei at 2500 rcf for 5 minutes and remove supernatant. Bring pellet up to 880 μL in Nuclear Lysis Buffer. Aliquot 300 μL samples in 1.5ml Bioruptor tube each and sonicate chromatin on Diagenode Bioruptor Pico instrument with 30 s on; 30 s off for 30 cycles. Clarify sample for 15 minutes at 16100 rcf at 4°C. Add 2X volume of ChIP Dilution Buffer. Wash 60 μL of Protein A beads for every 10 million cells in ChIP Dilution Buffer. Amounts of beads (for preclearing and capture) and antibody should be adjusted linearly for different amounts of cell starting material. Resuspend Protein A beads in 50 μL of Dilution Buffer per tube (100 μL per HiChIP), add to sample and rotate at 4°C for 1 hour. Put samples on magnet and transfer supernatant into new tubes. Add 7.5 ug of antibody (YY1 or NCAPH2) for every 10million cells and incubate at 4°C overnight with rotation. Wash 60 μL of Protein A beads for every 10m cells in ChIP Dilution Buffer. Resuspend Protein A beads in 50 μL of Dilution Buffer (100 μL per HiChIP), add to sample and rotate at 4°C for 2 hours. Wash beads three times each

with Low Salt Wash Buffer, High Salt Wash Buffer, and LiCl Wash Buffer. Washing was performed at room temperature on a magnet by adding 500 µL of a wash buffer, swishing the beads back and forth twice by moving the sample relative to the magnet, and then removing the supernatant. Resuspend ChIP sample beads in 100 µL of DNA Elution Buffer (make fresh). Incubate at RT for 10 minutes with rotation, followed by 3 minutes at 37°C with shaking. For ChIP samples, place on magnet and remove supernatant to a fresh tube. Add another 100 µL of DNA Elution Buffer to ChIP samples and repeat incubations. Remove ChIP samples supernatant again to the new tube. There should now be 200 µL of ChIP sample. Add 10 µL of Proteinase K to each sample and incubate at 55°C for 45 minutes with shaking. Increase temperature to 67°C and incubate for at least 1.5 hours with shaking. Use Zymo DNA Clean & Concentrator to purify the samples and elute in 10 µL of water. Quantify post-ChIP DNA to estimate the amount of Tn5 needed to generate libraries at the correct size distribution. Prepare for biotin pull-down by washing 5 µL of Streptavidin C-1 beads with Tween Wash Buffer. Resuspend the beads in 10 µL of 2X Biotin Binding Buffer and add to the samples. Incubate at room temperature for 15 minutes with rotation. Separate on a magnet and discard the supernatant. Wash the beads twice by adding 500 µL of Tween Wash Buffer and incubating at 55°C for 2 minutes shaking. Wash the beads in 100 µL of 1X (from 2X) TD Buffer. Resuspend beads in 25 µL of 2X TD Buffer, the appropriate amount of Tn5 for your material amount (2.5 µL for 50 ng of post-ChIP DNA), and water to 50 µL. Adjust Tn5 amount linearly for different amounts of post-ChIP DNA, with a maximum amount of 4 µL of Tn5. Incubate at 55°C with interval shaking for 10 minutes. Place samples on magnet and remove supernatant. Add 50 mM EDTA to samples and incubate at 50°C for 30 minutes, then quickly place on magnet and remove supernatant. Wash samples twice with 50 mM EDTA at 50°C for 3 minutes, removing quickly on magnet. Wash samples twice in Tween

Wash Buffer at 55°C for 2 minutes, removing quickly on magnet. Wash samples in 10 mM Tris. Resuspend beads in 50 μL of PCR master mix (Phusion HF 2X 25 μL Nextera Adapter1.1 (Universal) 12.5 uM 1 μL Nextera Adapter2.x (Barcoded) 12.5 uM 1 μL Water 23 μL) Run the following PCR program: First run 5 cycles on a regular PCR and then remove from beads. Add 0.25X SYBR green and then run on a qPCR and pull out samples at the beginning of exponential amplification. Run reactions on a PCR and estimate cycle number based on the amount of material from the post-ChIP Qubit (greater than 50 ng was ran in five cycles, while approximately 50 ng was running in six, 25 ng was running in seven, 12.5 ng was ran in eight, etc.). PCR Program: 72°C 5 minutes; 98°C 1 minute; Cycle 98°C 15 seconds 63°C 30 seconds 72°C 1 minute. Place libraries on a magnet and elute into new tubes. After PCR place libraries on a magnet and elute into new tubes. Then add 25 μL of Ampure XP beads and keep the supernatant to capture fragments less than 700 bp. Transfer supernatant to a new tube and add 15 μL of fresh beads to capture fragments greater than 300 bp. Finally elute libraries in 10 μL of water. Sequence the libraries on Hi-Seq4000 with PE 50bp.

2.3.9 1D-dendrimer capture experiment

Pellet detached adherent or suspension cells and resuspend in freshly made 1% formaldehyde (methanol free) at a volume of 1 mL of formaldehyde for every one million cells. Incubate cells at room temperature for 10 minutes with rotation. Add glycine to a final concentration of 125 mM to quench the formaldehyde, and then incubate at room temperature for 5 minutes with rotation. Pellet cells, wash in PBS, pellet again. Resuspend cell pellet in 500 μL lysis buffer (10 mM Tris-HCl pH 8.0, 10 mM NaCl, 0.2% Igepal CA630) In cubate on ice for 15min, pellet cells. Add 50 μM each generation of biotinylated dendrimer functionalized with psoralen into cell pellet. Incubate at 4°C for 10min. Crosllink the cells with UV365 for 30min. Pellet cells and wash

twice with lysis buffer (10 mM Tris-HCl pH 8.0, 10 mM NaCl, 0.2% Igepal CA630) Resuspend the pellet in proteinase K buffer (420 µl Hi-C lysis buffer, 50 µl 10% SDS, 30 µl 20 mg/ml proteinase K) Incubate at 65 ˚C for O/N on a thermomixer at 800 rpm. Extract the DNA with 500 µl phenol:chloroform. Centrifuge at max for 10 min at r.t. Transfer the upper layer to a new tube. Add 800 µl EtOH and 50 µl 3 M NaOAc (pH 5.5). Incubate at -80 ˚C for 3-4 h. Centrifuge at max for 15 min at 4 ˚C. Discard the supernatant. Wash the pellet twice with 500 µl 70% EtOH. Centrifuge at max for 5 min at 4 ˚C. Discard the supernatant. Resuspend the DNA pellet in 100 µl water. Prepare for biotin pull-down by washing 20 µl of 10 mg/ml Dynabeads MyOne Streptavidin C1 beads (Life technologies) with 400 µl of 1X Tween Washing Buffer (1X TWB: 5 mM Tris-HCl (pH 7.5); 0.5 mM EDTA; 1 M NaCl; 0.05% Tween 20). Separate on a magnet and discard the solution. Resuspend the beads in 100 µl of 2X Binding Buffer (2X BB: 10 mM Tris-HCl (pH 7.5); 1 mM EDTA; 2 M NaCl) and add to the reaction. Incubate at room temperature for 15 min with rotation to bind biotinylated DNA to the streptavidin beads. Separate on a magnet and discard the solution. Wash the beads by adding 600 µl of 1X TWB and transferring the mixture to a new tube. Heat the tubes on a Thermomixer at 55 °C for 2 min with mixing. Reclaim the beads using a magnet. Discard supernatant. Repeat wash. Resuspend beads in 100 ul 1X NEB T4 DNA ligase buffer (NEB, B0202) and transfer to a new tube. Reclaim beads and discard the buffer. To repair ends of sheared DNA and remove biotin from unligated ends, resuspend beads in 100 µl of master mix: 88 µl of 1X NEB T4 DNA ligase buffer with 10 mM ATP S33, 2 µl of 25mM dNTP mix, 5 µl of 10 U/µl NEB T4 PNK (NEB, M0201), 4 µl of 3 U/µl NEB T4 DNA polymerase I (NEB, M0203), 1 µl of 5 U/µl NEB DNA polymerase I, Large (Klenow) Fragment (NEB, M0210) Incubate at room temperature for 30 min. Separate on a magnet and discard the solution. Wash the beads by adding 600 µl of 1X TWB and transferring the mixture to a new

tube. Heat the tubes on a Thermomixer at 55°C for 2 min with mixing. Reclaim the beads using a magnet. Discard supernatant. Repeat wash. Resuspend beads in 100 µl 1X NEBuffer 2 and transfer to a new tube. Reclaim beads and discard the buffer. Resuspend beads in 100 µl of dATP attachment master mix: 90 µl of 1X NEBuffer 2, 5 µl of 10 mM dATP, 5 µl of 5 U/µl NEB Klenow exo minus (NEB, M0212). Incubate at 37°C for 30 min. Separate on a magnet and discard the solution. Wash the beads by adding 600 µl of 1X TWB and transferring the mixture to a new tube. Heat the tubes on a Thermomixer at 55°C for 2 min with mixing. Reclaim the beads using a magnet. Discard supernatant. Repeat wash. Resuspend beads in 100 µl 1X Quick ligation reaction buffer (NEB, B6058) and transfer to a new tube. Reclaim beads and discard the buffer. Resuspend in 50 µl of 1X NEB Quick ligation reaction buffer. Add 2 µl of NEB DNA Quick ligase (NEB, M2200). Add 3 µl of Illumina indexed adapter. (Nextflex) Record the sample-index combination. Mix thoroughly. Incubate at room temperature for 15 min. Separate on a magnet and discard the solution. Wash the beads by adding 600 µl of 1X TWB and transferring the mixture to a new tube. Heat the tubes on a Thermomixer at 55°C for 2 min with mixing. Reclaim the beads using a magnet. Remove supernatant. Repeat wash. Wash 3 times with 100 µl water. Reclaim the beads with 50 µl water. Incubate at 98 ˚C for 10min to elute the DNA from the beads. Transfer the supernatant to a 8-well PCR tube.

PCR amplify 12 cycles with following conditions:

98 ˚C 30 s ;98 ˚C 15 s ;60 ˚C 30 s; 72 ˚C 30 s; Repeat 12 cycles; 72 ˚C 1 min.

Purify the libraries with 0.9X Ampure beads. Elute with 30 µl water.

2.3.10 ChIP-seq data analysis

Top2b, CTCF, Rad21 ChIP-seq and control datasets was aligned to the mm10 reference genome using BWA-MEM (*43*). PCR duplicates were removed using Picard Tools version 2.2.2 (http://broadinstitute.github.io/picard/). Peak calling was performed using MACS2 (Zhang et al.) with both treated and control BAM files and the following parameters: -q 0.05 –f BAM –g mm. For visualization of all internally and externally generated ChIP-seq as tracks, we aligned each dataset to the mm10 reference genome, and used bedtools version 2.27 (*44*) to create bedGraph files of depth-normalized RPM or FPM values (depending on whether it is a paired-end or single-end dataset). bedGraph files were then converted to bigWig files for its indexing and fast query retrievals using UCSC bedGraphToBigWig (https://genome.ucsc.edu/util.html).

2.3.11 Alignment and pre-processing of CAP-C and *in-situ* Hi-C datasets

All CAP-C and *in-situ* Hi-C libraries were sequenced on the Illumina NextSeq500 or HiSeq2500. Paired-end FASTQ formatted files were aligned to the mm10 reference genome using BWA-MEM (*3*). Each read end was aligned independently before combining them into a single BAM file. Singletons and chimeric reads were set aside. PCR duplicates were removed using Picard Tools. Separately, an MboI restriction sites BED file was prepared containing the positions of all possible GATC cut sites using HiCPro (*45*). We performed an extra filtering step to remove all contact distances less than 1Kb in length to ensure that 1) all strand orientations were equally represented as a function of genomic distance and 2) because they were highly inconsistent across all replicates sequenced, and severely skewed the relative contact frequencies of longer range interactions. Only valid, and filtered contact pairs were used for all downstream analysis. A pre-formatted text file was prepared before converting it into a .hic format file with

the following resolutions (0.5 Kb, 1 Kb, 2 Kb, 5 Kb, 10 Kb, 25 Kb, 40 Kb, 50 Kb, 100 Kb, 250 Kb, 500 Kb, 1 Mb and 2.5 Mb). Finally, a MAPQ>=1 and MAPQ>=30 .hic file was then generate for each sample.

To generate a CAP-C merge file, all G3, G5 and G7 primary and replicated libraries in the pre-formatted text file were concatenated and merged-sorted before using Juicer to convert it into a .hic formatted file. All contact matrices visually represented as contact maps were VC-normalized with Juicer (*46*).

2.3.12 CAP-C eigenvectors and other statistical analyses

Because we had prepared CAP-C libraries using 3 different dendrimers, a standard differential analysis could not be sufficiently tested with large number of pixels and limited replicates at high-resolution. Instead, we depth-normalized all libraries and performed a principal component analysis using a 3 by N*(N+1)/2 matrix where N is the number of loci based on the specified resolution. To calculate CAP-C eigenvectors, we performed principal component analysis on a 3 * N matrix by using the depth-normalized row sum values of all 3 contact matrices. The first principal component or the eigenvector with the highest eigenvalue and its loadings were extracted. Regions with high gene densities were arbitrarily assigned as positive values.

To compare the difference of enrichment between groups of dendrimers, we deployed a multi-step statistical test to determine whether the dendrimer with the highest relative contact frequency in the bin is different from the relative contact frequency of the other two groups of dendrimers. For sufficient power, we performed the statistical analysis of all contact matrices at low-resolution (100Kb). First, we performed a one-way ANOVA (analysis of variance), and if there is a significant difference between groups ($F_{stat} < F_{critical}$) for the appropriate significance,

we followed up with post-hoc analysis using a linear contrast. Rather than use it as a decision to reject the null, the calculated p-value is used to intensify the color of the bin to reflect the variability among biological replicates. Because the objective of the analysis is to compare between dendrimers, we scale contact frequencies (and contact probabilities) by the total number of intra-chromosomal valid contact pairs, and do not perform implicit normalization such as KR or ICE matrix balancing approaches because we do not assume that the row (or column) sums should be equal, i.e. the row sums are bias for the enrichment of regions in which the dendrimers target for.

Euclidean Distance Criteria:

Determining the overlap or concordance of 2D features is based on the criteria proposed by (*1*)

$$\min(0.2 * length, 50Kb) <= \sqrt{(x_1 - x_2)^2 + (y_1 + y_2)^2}$$

2.3.13 Compartments, domains, loops calling and external validation

Compartments were called at 500Kb resolution using the eigenvector module of Juicer.

Contact domains were called using the Arrowhead module of Juicer. To utilize the increase in short-range contacts in CAP-C, we sensitively call more short-range contact domains by making changes to the Blockbuster.java file to toggle the minimum width parameter. We set this to zero. A customized Juicer jar file was recompiled based on these changes. We called contact domains at 500bp, 1Kb, 2Kb, 5Kb and 10Kb resolutions, and merged all nested and non-nested contact domains into a unique call set. Out of all domains that overlapped (based on the Euclidean distance criteria), the contact domain with the highest corner score was retained.

Peaks in the deeply sequenced CAP-C and in-situ Hi-C libraries were called using HiCCUPs with the following high-resolution parameters (hiccups –m 1024 –r 5000,10000 –k KR –f .1,.1 –p 4,2 –i 7,5 –t 0.02,1.5,1.75,2 –d 20000,20000), while lower-resolution G5 inhibitor-treated libraries were called with the following medium-resolution parameters (hiccups –m 1024 –r 5000,10000,25000 –k KR –f .1,.1,.1 –p 4,2,1 –i 7,5,3 –t 0.02,1.5,1.75,2 –d 20000,20000,50000). We validated our peaks calls by comparing peak calls to mESC Smc1 HiChIP (*24*) and mESC cohesin ChIA-PET (*47*). Comparisons between peaks or domains in a call set were determined to be concordant based on the Euclidean distance criteria. For peaks, we took the mid-point as the coordinate for each anchor. Loop domains were determined by overlapping peaks with domains under the same criteria. All remaining contact domains were classified as non-loop domains.

2.3.14 Meta-analysis of domain boundaries

To investigate the properties of loop versus non-loop domains, we first merged all nested and non-nested contact domains called at the 500bp, 1Kb and 2Kb resolution into a unique call set, and segregate them into high-resolution loop versus non-loop domains based on whether they overlapped a peak or not. For all ChIP-Seq datasets of histone modification marks and transcription factor peaks, we extracted only signals +/- 2 Kb around the boundary, and signals at 5% to 95% around the domain body using pyBigWig (*48*).

To build mean contact maps of contact domains with varying sizes, boundaries of each contact domain were first extended both upstream and downstream by 30%, and contact frequencies recomputed into 32 bins (6 upstream + 20 for the body + 6 downstream). The rescaled contact probability was then computed for each contact map, and a mean contact map visualized based on the array of rescaled contact matrices. We then performed meta-analysis on loop-domains,

non-loop domains, and domains whose boundaries overlaps a promoter and enhancer state. At the same time, we also visualized cohesin ChIA-PET interactions (*8*) overlapping promoter and enhancers. Promoter-enhancer interactions is unlikely to have the same characteristics as CTCF-cohesin loops we see in loop-domains.

Contact domains overlapping at least 50% of the length of the gene is considered to overlap a single gene.

2.3.15 Directionality Index and TADs

Directionality indices were calculated based on the following accorded to (*7*) for quantifying strength of domain boundaries (i.e the upstream vs downstream bias) in a one-dimensional linear fashion.

$$DI = \left(\frac{D-U}{|D-U|}\right)\left(\frac{(U-E)^2}{E} + \frac{(D-E)^2}{E}\right)$$

Where

- U is the total number of contacts pairs for all R-th resolution bins between the locus to the locus R*40 bp upstream
- D is the total number of contact pairs for all R-th resolution bins between the locus to the locus R*40 bp downstream
- E is the expected number of contact pairs for each locus

2.3.16 Classification of loops, domains and TSS

Loops were classified as promoter-promoter (P-P), promoter-enhancer (P-E), enhancer-promoter (E-P), enhancer-enhancer (E-E), insulator, repressed or heterochromatin based on the following histone modification marks: H3K4me3, H3K27ac, Ctcf and H3K27me3. Because loop resolution (5Kb) was still significantly lower than the peak resolution (200bp) of histone modification and transcription factors, we devised the following a priori criteria to classify loops. A loop anchor was defined as overlapping a promoter region if the mean H3K4me3 mark was more enriched than the global H3K4me3 average. A loop anchor was defined as overlapping an enhancer region if the mean H3K27ac mark was more enriched than the global H3K27ac average. A loop anchor was defined as overlapping a repressed region if the mean H3K27me3 mark was more enriched than the global H3K27me3 average. Loop anchors were defined as overlapping insulator regions if the loop anchors were not classified as overlapping promoter, enhancer, or repressed regions but whose mean CTCF signals were greater than the global CTCF average. Loop anchors with mean signals of H3K4me3, H3K27ac, CTCF and H3K27me3 lesser than the global averages were classified as overlapping heterochromatin regions. Med12 was used to independently validate the classification of these anchors, as Med12 is known to be present at promoter and enhancer regions, but not at CTCF-cohesin mediated insulator loops. Classification of loops are not mutually exclusive, and each loop may have multiple assignments.

Because resolution of contact domains and TSS were higher than loops, we simply overlapped the upstream and downstream boundaries with chromatin state track and classified them. For contact domains, we were only interested in promoter-enhancer domains. For each gene, we classified their activity status based on the chromatin states around the TSS. We classified them into 4 categories: active promoter, poised promoter, repressed and heterochromatin states.

2.3.17 Analyses involving domain boundary formation based on the orientation of gene pairs

Ensembl gene annotations (v87) of the mm10 reference genome was used to determine the longest transcript of a gene. We classified the orientation of each consecutive gene pair as divergent (reverse-forward strand), convergent (forward-reverse strand), and tandem (forward-forward, and reverse-reverse strand). We plot average contact maps of VC-normalized counts and O/E values by centering the TSS or TES of the second gene in the gene pair.

2.3.18 Alternative Promoter Analyses

RefSeq annotated transcripts were used to discover alternative promoters by selecting genes with different transcription start sites (TSS) sites (1,046 genes or 4.2% of all RefSeq genes had more than one TSS). Because certain genes have more than 2 known and unique transcription start sites, we simplified the analysis by selecting genes with only two unique TSS which are at least 5Kb apart. Genes that were less than 10Kb at their longest were filtered away (689 genes remained). We classified these genes with multiple alternative promoters into 4 different classes based on the arrangement in which their TSS overlapped an active promoter ChromHMM state. 128 (18.6%) did not have either TSS overlap any active promoter states. 99 (14.4%) had their upstream TSS, and not their downstream TSS, overlap an active promoter region. 300 (43.5%) had their downstream TSS, but not their upstream TSS, overlap an active promoter region. 162 (23.5%) had both TSS overlapped active promoters. Based on these classifications, we plotted VC-normalized mean contact maps and O/E maps to determine the strength of their domain

boundaries. We also overlapped these regions with PolII ChIP-seq, PRO-seq (determine presence and strength of divergent transcription) and DI (strength of domain boundary) values. Contact maps and mean signal values of these genes were all adjusted to the forward orientation.

2.3.19 Analyses of inhibitor experiments

Primary and replicate inhibitor maps at each individual time-point were found to be reproducible and merged. Because the merged maps are low-resolution quality and have a map-resolution of 25Kb, HiCCUP Peaks were called using medium resolution parameters while domains were called using the unmodified Arrowhead at 5Kb, 10Kb and 25Kb resolutions. Aggregate peak analyses (APA) was performed using the APA module of Juicer. APA extracts out a 200Kb by 200Kb contact matrices at 10Kb resolution around all peaks and superimposes the images to calculate the overall enrichment of the foci over the local background values. To examine the effect of loops on inhibitors in low-resolution maps, the APA P2LL scores (ranging from 3.85 to 23.93) in all inhibitor-treated maps were greater than 1, implying the presence of loops. What is striking is the increase in enrichment over the local background in all inhibitor-treated maps because the decrease in intra-domain contacts decreased faster than peak signals at the foci.

Mean maps were visualized by first extending the boundaries of domains by 30% in both directions, and re-scaling the contact matrices (in 10 Kb resolution) into (16 x 17)/2 bins (Length of each bin represents 10% of the domain length). Random domain boundaries were permuted using bedtools shuffle based on the size distribution of all domains from control.

2.4 Discussion and future perspective:

CAP-C represents a new method for studying chromatin architecture. CAP-C utilizes a multifunctional dendrimer platform instead of DNA-bound proteins to crosslink DNA, achieving informative spatial chromatin organization at higher resolution than *in situ* Hi-C. The high resolution achieved with CAP-C is not dependent on the sequencing depth but stems from its ability to preserve abundant informative short-range (1-20 Kb) chromatin contacts.

RNAPII is a powerful rotational motor whose activity has a significant impact on the topology of DNA (*49*). Using CAP-C, we discovered a role for transcription-induced supercoiling in delineating chromatin domains. The twin-domain supercoiling model predicts that during transcription, DNA regions ahead of the polymerase incur positive supercoiling (over-wound DNA) while those behind the polymerase incur negative supercoiling (under-wound DNA) (*50*). At the same time, conformational changes existing in the form of writhes in eukaryotes and plectonemes in prokaryotes start to form in both directions. Plectonemes (or writhe-like structure) can propagate quickly, and almost always faster than the rate of elongation (*51*). Biophysical models have also suggested that plectonemes can be propagated after elongation of just 5 bp (*52, 53*), explaining why poised promoters of bivalent states have also been observed to be insulated at its boundaries. Recent studies sought to map and measure the distribution of negative supercoiling across the genome using microarrays (*36, 54*) or next generation sequencing (*55*). These methods use psoralen derivatives as a probe to measure under-wound DNA by intercalating base-pairs of double-stranded DNA and crosslinking these regions with UV irradiation. In one study, low-resolution (~10 Kb) mapping of chromosome 11 in human retinal cells identified large-scale domains with varying degrees of

supercoiling (*36*), strongly resembling high-resolution compartment intervals (*10*) and compartment domains (*12*). In addition, high-resolution mapping (150 bp to 1 Kb) showed general enrichment of negative torsional stress detected as unwound DNA at the promoter region of actively transcribed genes (*54, 56*).

Our study here draws an inexplicable link between transcription-induced supercoiling (torsional stress) and higher-order structures like TADs, offering experimental evidence that transcription-induced supercoiling does cause chromatin conformation changes. Locally, self-associated domains are a manifestation of writhes that result from unconstrained supercoiling generated from RNAPII rotational torsion which helps to increase the contact possibilities between promoter and gene body. On the broader level, several proximal gene domains with distal regulatory element co-aggregate to form domains that further interact to build insulated neighborhood and TADs, while the supercoiling force generated within TADs might push the cohesin handcuff to form loops on the boundaries (*57*). Our findings raise the possibility that chromatin domains might form through transcription-induced supercoiling and this mechanism may be common among species from prokaryotes to eukaryotes. In species like mouse and human, that have acquired insulator proteins such as Ctcf and cohesin during evolution, loop domains might further insulate regulatory elements like enhancers with their targeted promoters to elevate their communication.

CAP-C offers several distinct advantages over conventional 3C-based methods. For chromatin packed in a highly crowded environment, DNA-bound proteins block the accessibility of DNA motifs for efficient restriction digestion and subsequent ligation in conventional 3C. These proteins are stripped away in CAP-C before restriction enzyme digestion, thus exposing all potential restriction sites to favor ligation of proximal contacts at all length scales. Unlike conventional 3C, CAP-C can also reveal DNA-DNA interactions that are not mediated by protein complexes. The

association of proximal DNA contacts within the same dendrimer molecule could facilitate derivation of loci-specific interactomes in the future, by enrichment of DNA bait without ligation.

PAMAM dendrimers allow precise control of the spherical polymer size, with different sized dendrimers serving as "molecular rulers" that fit chromatin conformations of various densities and potentially "measuring" the physical distances between two genomic loci. Different sized dendrimers offer an opportunity to discern open and closed chromatin at high resolution. Small dendrimers such as G3 favor tightly compacted, closed chromatin regions, whereas open chromatin regions are packed loosely and enrich for large dendrimers. Future work will explore synthesizing even larger dendrimer platforms to probe interactions at large scale to investigate potential communications between chromosome territories.

Lastly, the CAP strategy is not limited to studying chromatin structure via proximity ligation and high throughput sequencing. The crosslinked DNA-dendrimer complexes, which preserve intact chromatin structure, could be purified and coupled with other downstream methods (*58*) such as electronic microscopy or fluorescent microscopy to directly visualize native chromatin structure at high resolution. In addition, the surface exposed amines could be functionalized with crosslinking groups for RNA and protein, allowing broad application of the strategy to study all potential interactions among large biomolecules.

2.5 References

1. T. Sexton, G. Cavalli, The role of chromosome domains in shaping the functional genome. *Cell* **160**, 1049-1059 (2015).
2. A. Pombo, N. Dillon, Three-dimensional genome architecture: players and mechanisms. *Nat Rev Mol Cell Biol* **16**, 245-257 (2015).
3. J. Dekker, K. Rippe, M. Dekker, N. Kleckner, Capturing Chromosome Conformation. *Science* **295**, 1306-1311 (2002).

4. Z. Zhao *et al.*, Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet* **38**, 1341-1347 (2006).

5. J. Dostie *et al.*, Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome research* **16**, 1299-1309 (2006).

6. E. Lieberman-Aiden *et al.*, Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326**, 289-293 (2009).

7. J. R. Dixon *et al.*, Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012).

8. J. E. Phillips-Cremins *et al.*, Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281-1295 (2013).

9. J. M. Dowen *et al.*, Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**, 374-387 (2014).

10. S. S. Rao *et al.*, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).

11. B. Bonev *et al.*, Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557-572.e524.

12. M. J. Rowley *et al.*, Evolutionarily Conserved Principles Predict 3D Chromatin Organization. *Mol Cell* **67**, 837-852 e837 (2017).

13. C. B. Hug, A. G. Grimaldi, K. Kruse, J. M. Vaquerizas, Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription. *Cell* **169**, 216-228 e219 (2017).

14. B. Bonev *et al.*, Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557-572 e524 (2017).

15. G. Fudenberg *et al.*, Formation of Chromosomal Domains by Loop Extrusion. *Cell reports* **15**, 2038-2049 (2016).

16. T. S. Hsieh, G. Fudenberg, A. Goloborodko, O. J. Rando, Micro-C XL: assaying chromosome conformation from the nucleosome to the entire genome. *Nat Methods* **13**, 1009-1011 (2016).

17. T. H. Hsieh *et al.*, Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* **162**, 108-119 (2015).

18. W. Ma *et al.*, Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nat Methods* **12**, 71-78 (2015).

19. W. Schwarzer *et al.*, Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51-56 (2017).

20. S. S. P. Rao *et al.*, Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305-320 e324 (2017).

21. D. Astruc, E. Boisselier, C. Ornelas, Dendrimers Designed for Functions: From Physical, Photophysical, and Supramolecular Properties to Applications in Sensing, Catalysis, Molecular Electronics, Photonics, and Nanomedicine. *Chemical Reviews* **110**, 1857-1959 (2010).

22. F. Jin *et al.*, A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-294 (2013).

23. J. Ernst, M. Kellis, Chromatin-state discovery and genome annotation with ChromHMM. *Nature Protocols* **12**, 2478 (2017).

24. M. R. Mumbach *et al.*, HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* **13**, 919-922 (2016).

25. F. Benedetti, J. Dorier, Y. Burnier, A. Stasiak, Models that include supercoiling of topological domains reproduce several known features of interphase chromosomes. *Nucleic Acids Res* **42**, 2848-2855 (2014).

26. I. Hiratani *et al.*, Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS biology* **6**, e245 (2008).

27. N. Liu *et al.*, Recognition of H3K9 methylation by GLP is required for efficient establishment of H3K9 methylation, rapid target gene repression, and mouse viability. *Genes & development* **29**, 379-393 (2015).

28. T. B. Le, M. T. Laub, Transcription rate and transcript length drive formation of chromosomal interaction domain boundaries. *The EMBO journal* **35**, 1582-1595 (2016).

29. F. Benedetti, D. Racko, J. Dorier, Y. Burnier, A. Stasiak, Transcription-induced supercoiling explains formation of self-interacting chromatin domains in S. pombe. *Nucleic Acids Research* **45**, 9850-9859 (2017).

30. T. B. K. Le, M. V. Imakaev, L. A. Mirny, M. T. Laub, High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome. *Science* **342**, 731-734 (2013).

31. T. Mizuguchi *et al.*, Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe. *Nature* **516**, 432-435 (2014).

32. M. Dunaway, E. A. Ostrander, Local domains of supercoiling activate a eukaryotic promoter in vivo. *Nature* **361**, 746 (1993).

33. L. Uuskula-Reimand *et al.*, Topoisomerase II beta interacts with cohesin and CTCF at topological domain borders. *Genome Biol* **17**, 182 (2016).

34. T. A. Ayoubi, W. J. Van De Ven, Regulation of gene expression by alternative promoters. *The FASEB Journal* **10**, 453-460 (1996).

35. J. R. Dixon, D. U. Gorkin, B. Ren, Chromatin Domains: The Unit of Chromosome Organization. *Mol Cell* **62**, 668-680 (2016).

36. C. Naughton *et al.*, Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nat Struct Mol Biol* **20**, 387-395 (2013).

37. O. Bensaude, Inhibiting eukaryotic transcription: Which compound to choose? How to evaluate its activity? *Transcription* **2**, 103-108 (2011).

38. T. Sakuma, A. Nishikawa, S. Kume, K. Chayama, T. Yamamoto, Multiplex genome engineering in human cells using all-in-one CRISPR/Cas9 vector system. *Scientific reports* **4**, 5400 (2014).

39. A. J. Holland, D. Fachinetti, J. S. Han, D. W. Cleveland, Inducible, reversible system for the rapid and complete degradation of proteins in mammalian cells. *Proceedings of the National Academy of Sciences* **109**, E3350-E3357 (2012).

40. A. W. Cheng *et al.*, Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. *Cell research* **23**, 1163 (2013).

41. N. E. Sanjana, O. Shalem, F. Zhang, Improved vectors and genome-wide libraries for CRISPR screening. *Nature methods* **11**, 783 (2014).

42. S. G. Landt *et al.*, ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research* **22**, 1813-1831 (2012).

43. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*, (2013).

44. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).

45. N. Servant *et al.*, HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome biology* **16**, 259 (2015).

46. N. C. Durand *et al.*, Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell systems* **3**, 95-98 (2016).

47. J. M. Dowen *et al.*, Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**, 374-387 (2014).

48. F. Ramírez *et al.*, deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic acids research* **44**, W160-W165 (2016).

49. J. Ma, M. D. Wang, DNA supercoiling during transcription. *Biophysical Reviews* **8**, 75-87 (2016).

50. H.-Y. Wu, S. Shyy, J. C. Wang, L. F. Liu, Transcription generates positively and negatively supercoiled domains in the template. *Cell* **53**, 433-440.

51. M. T. J. van Loenhout, M. V. de Grunt, C. Dekker, Dynamics of DNA Supercoils. *Science* **338**, 94-97 (2012).

52. C. Bécavin, M. Barbi, J.-M. Victor, A. Lesne, Transcription within Condensed Chromatin: Steric Hindrance Facilitates Elongation. *Biophysical Journal* **98**, 824-833 (2010).

53. S. S. Teves, S. Henikoff, DNA torsion as a feedback mediator of transcription and chromatin dynamics. *Nucleus* **5**, 211-218 (2014).

54. F. Kouzine *et al.*, Transcription dependent dynamic supercoiling is a short-range genomic force. *Nature structural & molecular biology* **20**, 396-403 (2013).

55. S. S. Teves, S. Henikoff, Transcription-generated torsional stress destabilizes nucleosomes. *Nat Struct Mol Biol* **21**, 88-94 (2014).

56. S. S. Teves, C. M. Weber, S. Henikoff, Transcribing through the nucleosome. *Trends in biochemical sciences* **39**, 577-586.

57. D. Racko, F. Benedetti, J. Dorier, A. Stasiak, Transcription-induced supercoiling as the driving force of chromatin loop extrusion during formation of TADs in interphase chromosomes. *Nucleic Acids Research*, gkx1123-gkx1123 (2017).

58. R. A. Beagrie *et al.*, Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**, 519-524 (2017).

3 A highly sensitive and robust method for genome-wide 5hmC profiling and its application on acute myeloid leukemia (AML) study

3.1 Introduction

5-hydroxymethylcytosine (5hmC), the oxidative product of 5-methylcytosine (5mC) catalyzed by ten-eleven translocation (TET) enzymes, is found in various mammalian tissues and cell types. Emerging evidence indicates that 5hmC is not only an intermediate of DNA demethylation, but also acts as a potential epigenetic mediator, which modulates a spectrum of biological processes and human diseases. (*1-4*) The recent development of high-throughput, genome-wide sequencing technologies has enabled genome-wide mapping of 5hmC in mammalian systems. (*5-11*) While applications of these methods have provided key information about the distribution of 5hmC and its functional insights, the need for a large amount of cells to obtain sufficient genomic DNA starting material for 5hmC localization precludes their use with rare cell populations including normal and malignant stem cells, homogeneous neuronal cells, and clinical isolates including needle biopsies, circulating tumor cells, and cell-free DNA. Therefore, new approaches are needed to allow for the detection of 5hmC in rare cell populations. Here, we developed a sensitive and robust 5hmC sequencing approach which allows genome-wide profiling of 5hmC based on a previously invented selective chemical labeling. (*8*) Using a limited amount of genomic DNA that can be readily isolated from 1,000 cells (nano-hmCSeal). To demonstrate the advantage and utility of this strategy, we have applied this approach to compare 5hmC profiles between hematopoietic stem cell (HSC) and progenitor cell populations. We found that 5hmC is enriched in the gene body of highly expressed genes and the level of 5hmC positively correlates

with histone modifications that mark active transcription. Moreover, we observed that the differentiation of murine HSCs to progenitor cells is strongly associated with dynamic alterations in 5hmC patterns with lineage-specific enhancers marked by pronounced 5hmC peaks. We further applied this technology to profile leukemia stem cells from a murine model of Tet2-mutant acute myeloid leukemia (AML) and obtained high-quality maps of 5hmC in tumor-initiating cells.

3.2 Result and discussion

3.2.1 General Scheme of Nano-hmC-Seal

We modified an engineered Tn5 transposase-based library construction strategy, (*12*-*14*) which allows fragmentation of genomic DNA into 300–500-bp fragments and appends sequencing compatible adaptors in a single tagmentation step. (*12*) Next, to enrich 5hmC-containing DNA fragments, we took advantage of the selective 5hmC chemical labeling (hmC-Seal) strategy, which has been previously developed for efficient, unbiased, and genome-wide labeling and covalent pull down of 5hmC. (*8*, *9*) Specifically, sequencing adaptors are incorporated through the transposase-catalyzed DNA tagmentation. (*12*)*2* The T4 bacteriophage enzyme β-glucosyltransferase (βGT) is then employed to transfer an engineered glucose moiety containing an azide-group to 5hmC in duplex DNA, yielding β-6-azide-glucosyl-5-hydroxymethyl-cytosine (N_3-5-gmC), as reported previously. (*8*)*2* Then a biotin tag is installed onto the azide group by using Huisgen cycloaddition (click) chemistry. Finally, 5hmC-containing DNA fragments with biotin tags are efficiently captured from the random DNA fragments pool by avidin beads. A regular PCR amplification reaction generates the library, which is then subjected to high-throughput sequencing (Fig.3.1A).

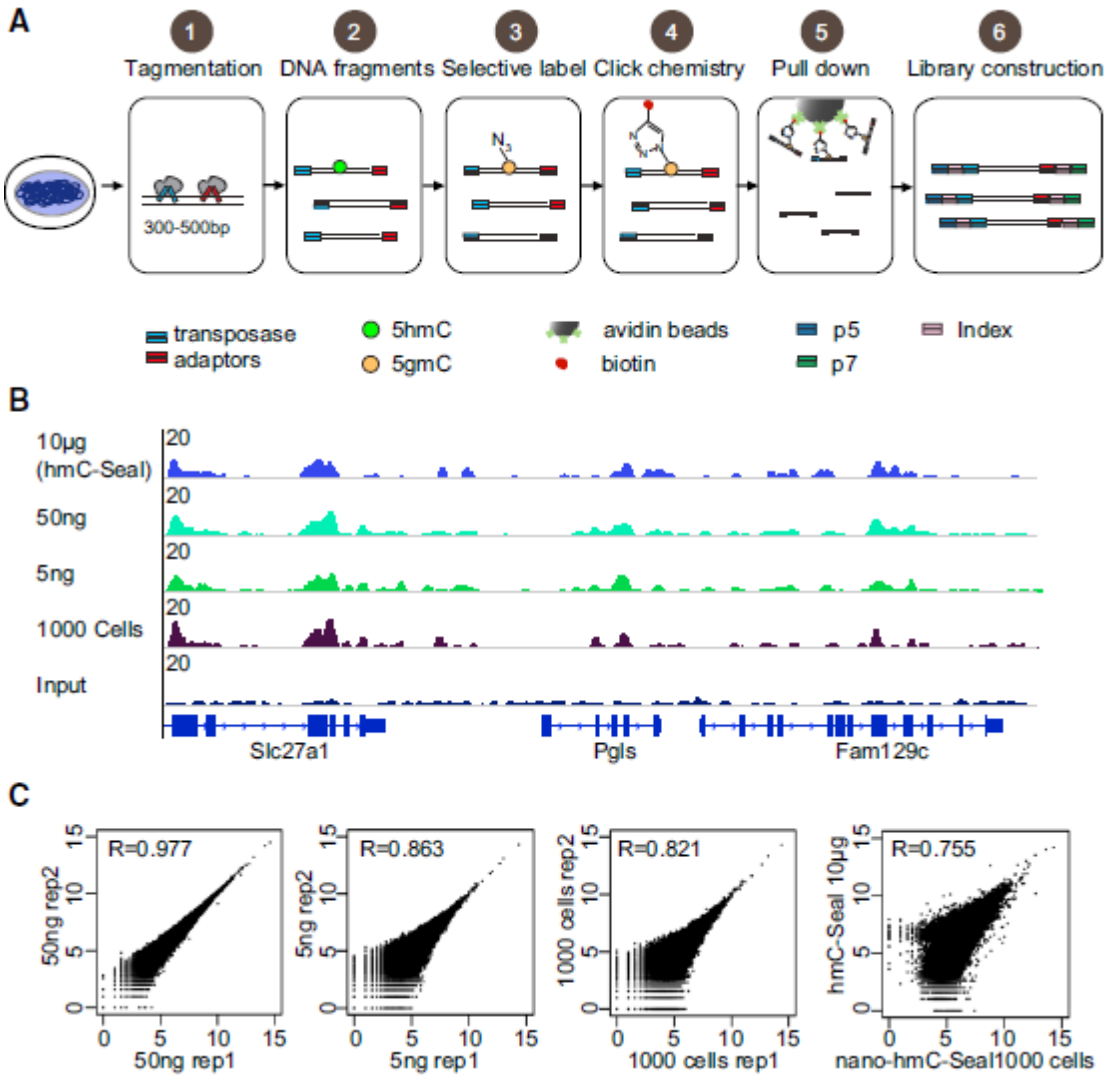


Fig. 3.1 Nano-hmC-Seal to Generate Genome-wide 5hmC Maps from Ultra-Low DNA Starting Materials (**A**) Schematic overview of the nano-hmC-Seal approach. (**B**) Genome browser views of 5hmC signals detected in a 30 kb region from libraries generated with 5 ng–10 mg of starting genomic DNA from mESCs. In blue, the 5hmC profile obtained using regular hmC-Seal with 10 mg genome DNA (top). Approximately 5–6 ng of genomic DNA could be isolated from 1,000 mESCs (1,000 cells). (**C**) Scatterplots showing correlation between nano-hmC-Seal replicates with Pearson correlation (r) displayed. Each dot represents a 5hmC enriched peak. The read counts were transformed to log2 base. From left to right: correlation between replicate libraries prepared from 50 ng and 5 ng mESC genomic DNA, and genomic DNA isolated from 1,000 mESCs, respectively, and between libraries using 1,000 cells with regular hmC-Seal using 10 mg genomic DNA.

3.2.2 Generation of Nano-hmC-Seal Libraries with Ultra-Low Starting Material

We tested the nano-hmC-Seal approach with 50 ng and 5 ng genomic DNA isolated from mouse embryonic stem cells (mESCs) or directly starting from 1,000 mESCs in replicates. We found that nano-hmC-Seal enrichment profiles are similar to results obtained from the regular hmC-Seal profiling starting with 10 mg mESC DNA (Fig. 3.1B and Fig. 3.2A– C). The 5hmC profiles were highly reproducible between replicates (R = 0.979 for 50 ng, 0.863 for 5 ng, and 0.821 for 1,000-cell samples) (Fig. 3.1C). The pairwise correlation of 5hmC signals in 2,000-bp tiling regions across the genome is 0.98 (Pearson correlation coefficient) when comparing the two replicates using 50 ng DNA, while regular hmC-Seal libraries showed a correlation of 0.88 (Fig. 3.2C). The 5hmC signals from 5 ng and 1,000-cell libraries correlated well with those obtained from the regular hmC-Seal libraries (Pearson's r ranging from 0.76 to 0.82) (Fig. 3.2C), albeit with slightly lower levels of correlations likely due to reduced input materials used (Fig. 3.1C Fig. 3.2C). Furthermore, the fractions of reads in high density 5hmC clusters were calculated as a measure of specific enrichment. All of the libraries generated by nano-hmC-Seal had similar 5hmC enrichment levels compared to the regular hmC-Seal libraries (Fig. 3.2D). We next employed the PreSeq (*15*) package to extrapolate and estimate library complexity (Fig. 3.2E). The results showed that libraries using 50 ng DNA have similar complexity as regular hmC-Seal libraries. Although libraries built from 5 ng or 1,000 cells displayed lower complexity, deeper sequencing of these libraries could still generate enough unique reads for downstream analysis. The 5hmC peaks detected from libraries using 50 ng DNA have similar number and quality compared to peaks obtained from regular hmC-Seal libraries; the quality starts to decrease with libraries employing 5 ng DNA or from 1,000 cells, but still yields a large number of 5hmC-enriched peaks for analysis (Fig. 3.2F). To compare our data sets with the ‘‘gold-standard’’ base-

resolution 5hmC maps generated by using TAB-seq, (*11*) 5hmC sites detected by TAB-seq were divided into three sets with low (10%–25%), medium (25%–40%), and high (>40%) 5hmC percentage. 5hmC signals obtained using nano-hmC-Seal around these sites positively correlate with the 5hmC abundance determined by TAB-seq (Fig. 3.2G). In addition, a strong enrichment of nano-hmC-Seal signals was also observed at the 5hmC-containging CpG islands (CGIs) detected by oxRRBS (Fig. 3.2H). (*5*) We therefore conclude that nano-hmC-Seal provides a reliable approach to the genome wide detection of 5hmC using limited genomic DNA materials.

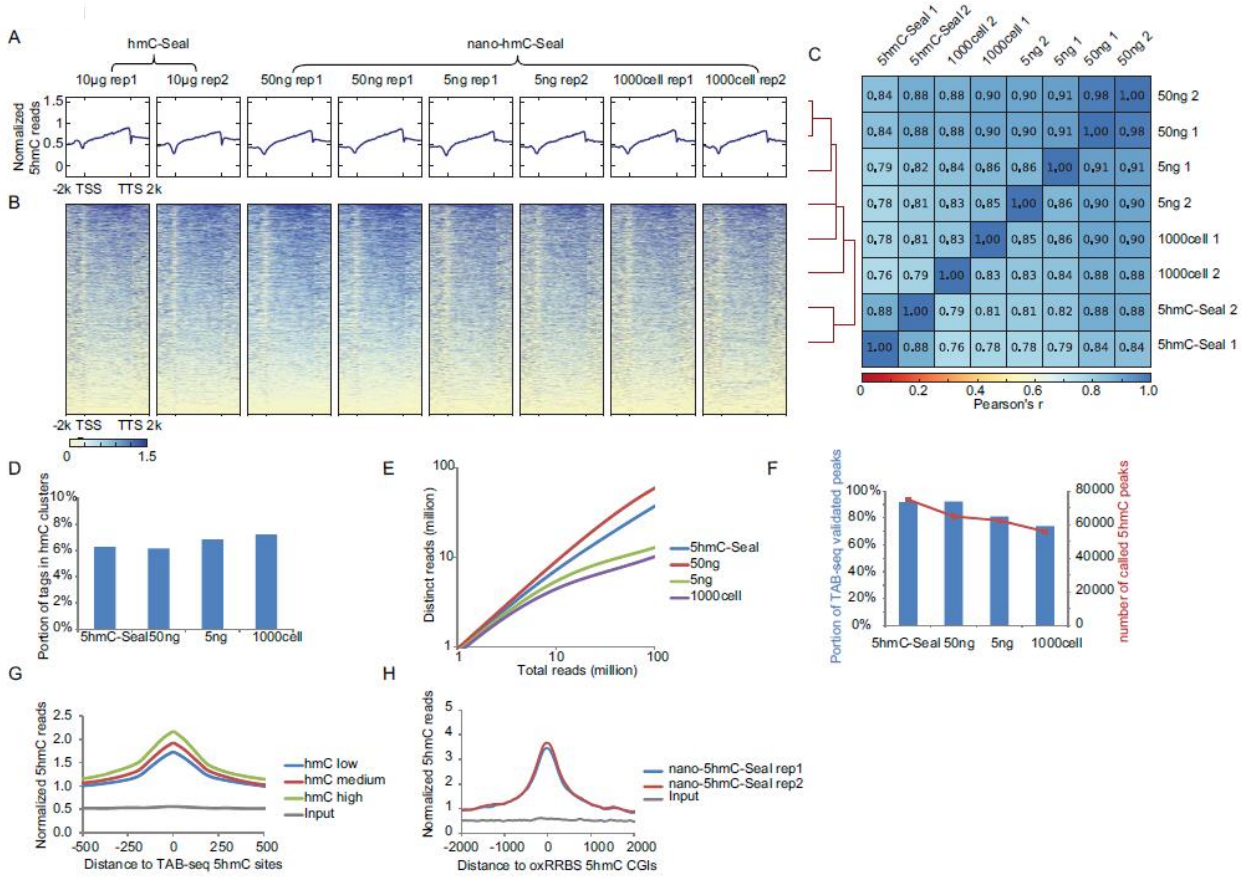


Fig. 3.2 Global comparison of conventional hmC-Seal and nano-hmC-Seal sequencing data. (A and B) Average profiles (A) and heatmap (B) across gene regions ±2,000 bp for nano-hmC-Seal libraries. Each row represents a gene, ordered by the mean value signals. Regular hmC-Seal libraries were generated from 10 μg genomic DNA and used as references. (C) Genome-wide correlations (2,000bp tiling windows) of sequencing results obtained using conventional hmC-Seal and nano-5hmC-Seal (with 50 ng and 5 ng DNA as well as genomic DNA isolated from 1,000 cells). (D) Fraction of reads located in 5hmC high-density clusters for libraries constructed using different amounts of input DNA. (E) Preseq library complexity curves for different libraries. (F) The number of high confident 5hmC-enriched peaks (right axis) called from different libraries and the portion of peaks validated by TAB-seq 5hmC sites (left axis). The 5hmC-enriched peaks were considered as validated if TAB-seq detected 5hmC sites reside in the area of 1000 bp surrounding the peak center. (G) The distribution of nano-hmC-Seal (50ng) signals at 5hmC sites detected by TAB-seq. 5hmC sites were further divided into low (10 - 25%), medium (25 - 40%), high (40% and above) subgroups (5,000 sites were randomly selected for each subgroup) (H) The distribution of nano-hmC-Seal (50ng) signals at 562 5hmC-containing CGIs detected by ox-RRBS method.

3.2.3 Nano-hmC-Seal Reveals Dynamic Hydroxy Methylation Localization at Enhancer Sites during Early Hematopoietic Differentiation

Having validated nano-hmC-Seal with limited starting genomic DNA, we next applied this approach to the study of hematopoiesis, a process in which epigenetic regulation plays a central role (*16*, *17*). In order to profile the dynamics of 5hmC during hematopoietic cell differentiation (Fig. 3.3A), we performed nano-hmC-Seal on cells isolated from mouse bone marrow by fluorescence-activated cell sorting (FACS). We analyzed flow sorted HSCs (LSK), common myeloid progenitors (CMP), granulocyte-macrophage progenitors (GMP), and megakaryocyte erythroid progenitors (MEP), with approximately 5,000 cells isolated from each purified population. Each cell type was sequenced in triplicate to monitor biological variability. These libraries yielded a total of 226.62 million reads, of which, 92.5% could be aligned to the mouse reference genome. A total of 323,854 5hmC peaks in the genome were identified.

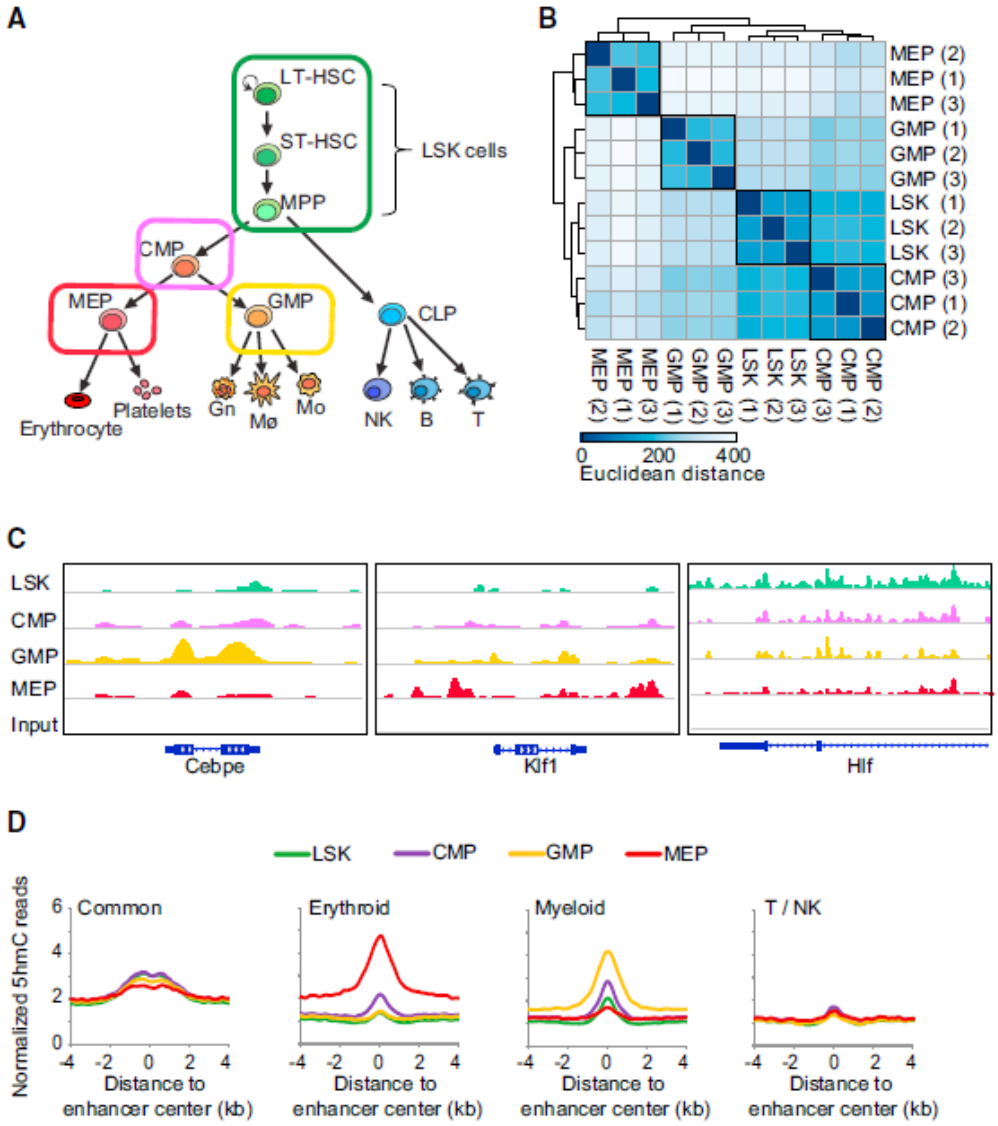


Fig. 3.3 Nano-hmC-Seal Provides Dynamic 5hmC Profiles during Early Hematopoiesis
(A) Schematic of the hematopoietic differentiation stages. The cell types investigated in this study are outlined. The cell surface phenotypes were LSK (lin$^-$ Sca$^+$ cKit$^+$), MPP (lin$^-$ Sca$^+$ cKit$^+$ CD48$^+$ CD150$^-$), CMP (lin$^-$ Sca$^-$ cKit$^+$ CD34$^+$ CD16/32$^-$), MEP (lin$^-$ Sca$^-$ cKit$^+$ CD34$^-$ CD16/32$^-$), and GMP (lin$^-$ Sca$^-$ cKit$^+$ CD34$^+$ CD16/32$^+$). (B) Relationship of 5hmC profiles in four different types of hematopoiesis stem and progenitor cells. Hierarchical clustering applied to the matrix of sample-to-sample distance based on rlog-transformed read counts in 304,069 detected 5hmCen-riched peaks is shown. (C) Representative examples of 5hmC profile in several loci (from left to right: *Cebpe*, *Klf1*, and *Hlf*). (D) Distribution of 5hmC signals at lineage-specific enhancers (from left to right: common enhancers, erythroid enhancer, myeloid enhancers, and T/NK cells enhancers). The lineage-specific enhancers were determined based on K-means analysis of H3K4me1 signals. The genomic locations of enhancers were previously defined.

The replicates displayed high similarity of 5hmC signal density in these peaks, thus demonstrating high reproducibility of our method (Fig. 3.3B). Of note, the 5hmC pattern of LSK bears more resemblance to the 5hmC pattern of CMP, and the 5hmC pattern of GMP cells are closer to that of CMP than LSK, whereas the 5hmC pattern of MEP is distinct from the other three lineages (Fig. 3.3B), indicating that the transition from CMP to MEP is associated with significant changes in 5hmC localization. Furthermore, at the differentially methylated regions (DMRs), which distinguish between GMP and MEPs obtained using RRBS (*17*), the 5hmC density showed an inverse comparative pattern in comparison with 5mC changes (Fig. 3.4A–C). Our results represent comprehensive genome-wide 5hmC maps obtained during in vivo differentiation of HSCs to committed progenitors.

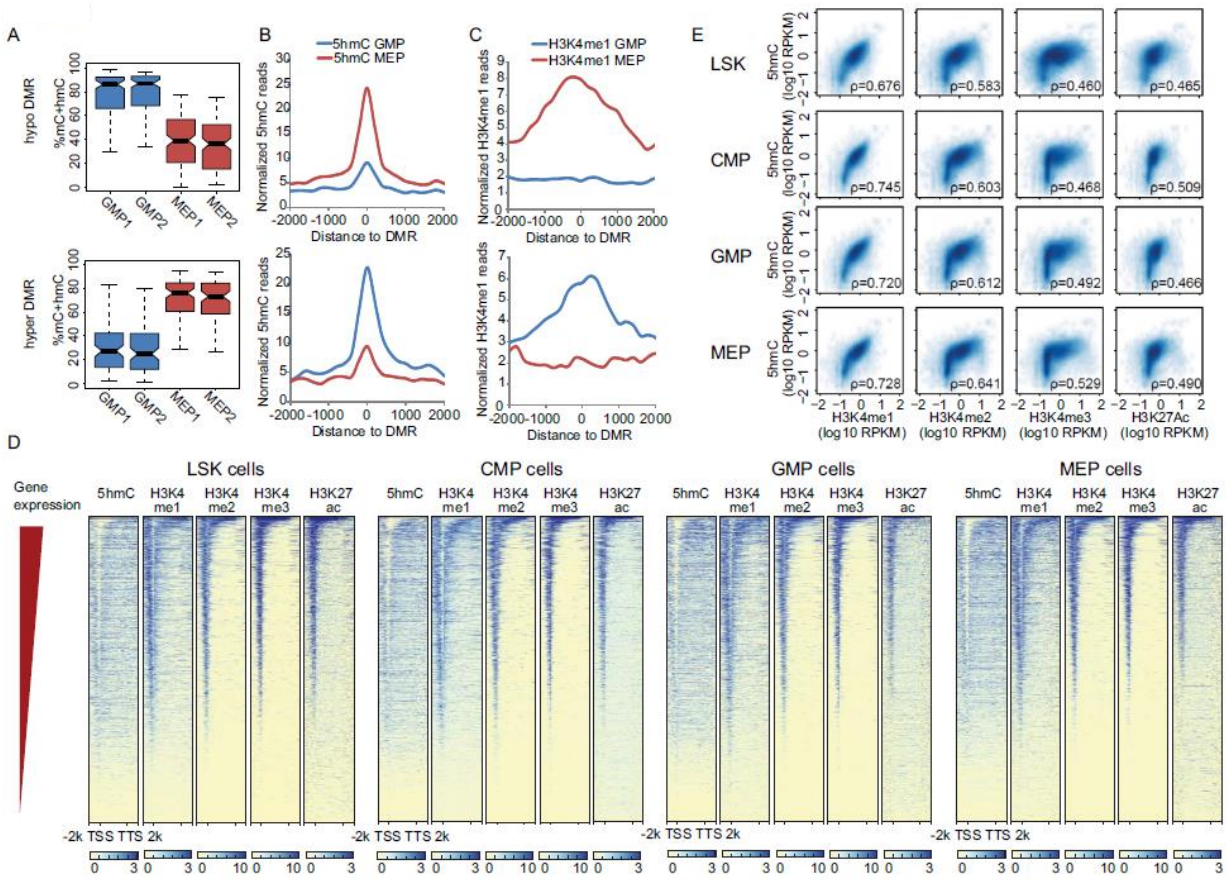


Fig. 3.4 5hmC levels correlate with DNA methylation and with histone marks in HSC and progenitor cells. (A) Boxplot to show the level of DNA modification (5mC+5hmC) at hypo (upper) and hyper (lower) DMRs. (B-C) The distribution of 5hmC (B) and H3K4me1 (C) signals at hypo (upper) and hyper (lower) DMRs. (D) Heatmap displaying the reads density distribution of 5hmC and indicated histone modifications in all annotated genes ordered by decreasing expression in LSK, CMP, GMP cells and MEP cells. Each row represents a gene. Due to lack of LSK histone modification data in published datasets, LSK hmC data was compared with histone modification data obtained from MPP cells. (E) Scatter plot displaying the correlation of 5hmC with H3K4me1, H3K4me2, H3K4me3 and H3K27ac in LSK, CMP, GMP cells and MEP cells. Each dot represents a gene. Due to lack of LSK histone modification data in published dataset, LSK hmC data was compared with histone modification data obtained from MPP cells. All values are represented as log10 RPKM. The spearman rank correlation coefficient is shown (ρ) in each comparison.

We next examined differential 5hmC localizations at genes encoding master transcriptional factors known to be expressed or silenced during HSC differentiation (*18, 19*)*2*. For example, 5hmC was observed at the highest level across the gene body of the Cebpe gene in GMP; Cebpe

encodes a bZIP transcription factor responsible for the lineage determination of GMP cells (*19, 20*) (Fig. 3.3C). In contrast, Klf1, a gene encoding a transcription factor critical for erythropoiesis, showed an increased 5hmC deposition at genic-proximal regions in MEP. Moreover, we observed a stepwise loss of 5hmC across the intragenic regions of Hlf, a gene essential for maintaining HSC function (*21*). To assess the regulatory role of 5hmC during differentiation, we compared 5hmC levels for each gene with previously reported chromatin immunoprecipitation sequencing (ChIP-seq) signals of histone modifications (*19*). The distributions of 5hmC and histone modifications were plotted ±2 kb around annotated genes and sorted based on their expression levels (Fig. 3.4D). As expected, 5hmC is enriched in the gene-body of highly expressed genes. In all cell types, there is a positive correlation of gene-body 5hmC with histone modifications that mark gene activation, especially for H3K4me1 (Fig. 3.4D-E).

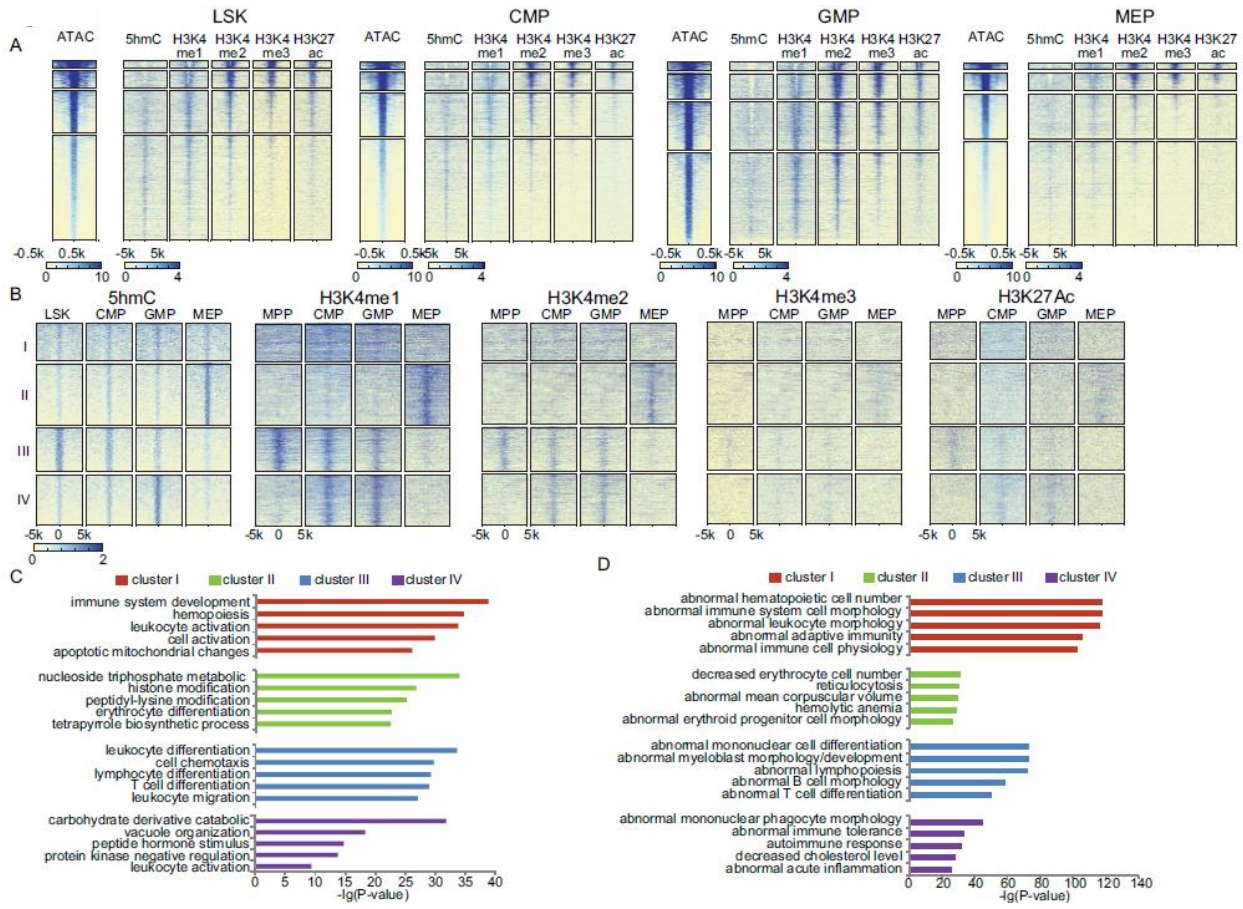


Fig. 3.5 The distribution of 5hmC and histone modifications at selected genomic regions. (A) Heatmap displaying read densities of ATAC-seq, 5hmC and histone modifications around the ATAC-seq signal-enriched peaks. ATAC-seq peaks were divided into four clusters by k-means clustering and ranked according to decreased ATAC signal values. Due to lack of LSK histone modification data in the published dataset, LSK ATAC-seq and hmC data were compared with histone modification data obtained from MPP cells. (B) Heatmap displaying read densities of 5hmC, H3K4me1, H3K4me2, H3K4me3 and H3K27Ac around DhMRs across differentiation stages. Clusters were generated by k-means clustering of 5hmC signals. Histone modification datasets were arranged to match the order of 5hmC heatmap. (C-D) Functional annotation of DhMRs in each cluster was performed using GREAT. The top over-represented categories belonging to Gene Ontology biological process (C) and Mouse Genome Informatics phenotype ontology (D) are shown. The x axis values correspond to the log10-transformed binomial P-values.

We next probed the relationship between 5hmC distribution and chromatin accessibility detected by assay for transposase-accessible chromatin (ATAC)-seq (*13, 19*). We divided ATAC-seq peaks into four groups based on signal intensity and found that 5hmC is depleted in the center of ATAC-seq peaks with high chromatin accessibility but enriched in ATAC-seq peaks with

lower signal intensity in all cell stages (Fig. 3.5A). The distribution of H3K4me1 peaks showed similar patterns with depletion in the center of ATAC-seq peaks, but not for other active histone makers (Fig. 3.5A). These observations are consistent with previous findings that 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC), the further oxidized intermediates in active demethylation, mark open chromatin sites (*22*), whereas 5hmC tends to mark less active or ‘‘poised’’ chromatin elements (*7, 11, 22*) and enriches around, but not on the transcription factor binding sites (*11*). Next, we studied the potential association between changes in 5hmC sites and in histone modifications during differentiation (Fig. 3.5B). We identified 21,791 regions that display differential 5hmC peaks across different differentiation stages. Clustering of these regions by their dynamic 5hmC profiles revealed four clusters. 5hmC sites in cluster I do not show noticeable trends of 5hmC changes, nor correlation patterns to reported histone modification differences. Gene set enrichment analysis with the Genomic Regions Enrichment of Annotations Tool (GREAT) (*23*) showed that genes associated to these 5hmC sites had enriched for genes with a role in immune system development and related to abnormalities in hematopoietic cell number and immune cell physiology (Fig. 3.5C). H3K4me1 and H3K4me2 signals show similar correlation patterns with respect to differential 5hmC in clusters II–IV (Fig. 3.5B). However, cluster III comprises peaks with the highest 5hmC level in LSK cells, which gradually reduces during differentiation. GREAT annotation of these peaks showed association with genes highly enriched for leukocyte differentiation. Cluster II comprises sites with increased 5hmC signals during differentiation from CMP to MEP cells, and cluster IV contains sites with gradually increasing 5hmC signals upon differentiation from LSK to GMP cells, but with slightly reduced 5hmC signals in MEP cells. Cluster II sites are linked to genes essential for erythrocyte differentiation and

cluster IV sites are related to genes whose knockout phenotypes exhibit defects in immune tolerance and acute inflammation (Fig. 3.5C). Both clusters contain regions that are pre-marked with relatively low levels of 5hmC in the stem cell stage; however, the 5hmC level increases upon differentiation to GMP or MEP cells, which is accompanied by the appearances of H3K4me1 and H3K4me2 marks that were not observed at the stem cell stage. These observations may suggest that 5hmC precede H3K4me1/2 at these regions in the chromatin activation process during HSC differentiation. In general gene-body 5hmC changes correlate with histone mark changes. Previous studies have shown that enhancer dynamics during hematopoiesis are critical for the differential access of transcriptional factors that drive lineage specification (*19*). Furthermore, studies in mESCs have suggested that the establishment and maintenance of enhancer regions primarily depend on TET-mediated active demethylation, and that the activity of these enhancers correlates with the abundance of 5hmC (*24, 25*). Analysis of genome-wide data obtained by nano-hmC-Seal allowed us to observe that 85.6% (41,450/48,415) of predicated enhancers (*19*), defined by enriched H3K4me1/2 histone modification signatures, are co-localized with 5hmC peaks in hematopoietic cells. We therefore hypothesized that dynamic changes in lineage specific enhancers during hematopoiesis (particularly those marked by H3K4me1) are associated with corresponding changes in 5hmC. Indeed, 5hmC signatures at common enhancers are remarkably similar across all cell types, whereas 5hmC signatures were gradually established in the myeloid specific enhancers during myeloid transition from LSK to GMP cells, but not in MEP cells (Fig. 3.3D). Moreover, 5hmC is highly enriched in erythroid specific enhancers in MEP cells, indicating a strong correlation of 5hmC distribution with lineage commitment. Further analyses verified that 5hmC signals detected in these cells are not enriched at enhancers specific for T or NK cells (Fig. 3.3D).

3.2.4 Nano-hmC-Seal Analysis of Murine Leukemia Stem Cells with *Tet2* Loss and *Flt3^{ITD}* Mutation

After mapping 5hmC dynamics during normal hematopoiesis, we next chose to focus our attention on AML, a disease marked by recurrent mutations in epigenetic regulators. We therefore applied nano-hmC-Seal to dissect how *Tet2* loss, combined with other known AML disease alleles, could potentially dysregulate 5hmC and contribute to leukemogenesis. To this end, we generated global 5hmC maps of MPP and GMP isolated from wild-type mice and T2F3 mice (a murine AML model harboring *Tet2* and *Flt3^{ITD}* mutations;(*26*)). Of note, previous studies have shown that MPP, but not GMP, from *Tet2* and *Flt3^{ITD}* mice have leukemia stem cell (LSC) potential with the ability to serially transplant in vivo (*26*). Unsupervised hierarchical clustering of a 5hmC-enriched region demonstrated clear separation of leukemic T2F3 samples from wild-type (WT) samples in both cell types (Fig. 3.6A). To pinpoint specific loci that display differential 5hmC profiles between leukemic samples and WT samples, we proceeded to identify and characterize differentially hydroxymethylated regions (DhMRs). A total of 9,204 DhMRs were found in MPP and 5,008 DhMRs in GMP (Fig. 3.6B-C). We compared DhMRs in MPP with those identified in GMP and found that the majority (>80%) of these DhMRs are not shared between the two cell types, suggesting distinct 5hmC deposition and maintenance in these two cell types isolated from this AML model. Regions with 5hmC-loss and 5hmC-gain in MPP and GMP were examined for enrichment in lineage specific enhancers. Regions with increased 5hmC in MPP were found preferentially overlapping with the myeloid specific enhancers, while losses of 5hmC in MPP were enriched for the erythroid-related enhancers (Fig. 3.6D). These results suggest that the alterations of 5hmC deposition may favor an aberrant differentiation toward myeloid cells at the expense of erythroid cells, a conclusion consistent with previous observations of

increased GMP and reduced MEPs in T2F3 mice (*26*). To better understand the regulatory sequence codes that are associated with 5hmC changes, we performed de novo motif analysis within these DhMR regions. The most enriched recognition motifs in loci with reduced 5hmC matched the binding sites of known regulators implicated in AML pathogenesis, including GATA and MYB (Fig. 3.6E). In contrast, we detected a known PU.1 motif (AGAGGAAGTG, $p = 1X10^{-106}$) in the regions that gained 5hmC (Fig. 3.7B), supporting the notion that reciprocal antagonism between PU.1 and the GATA family is critical for the myeloid lineage commitment in normal and malignant hematopoiesis (*27*). Interestingly, we detected a PU.1:IRF composite motif in the regions that gained 5hmC (Fig. 3.6E).

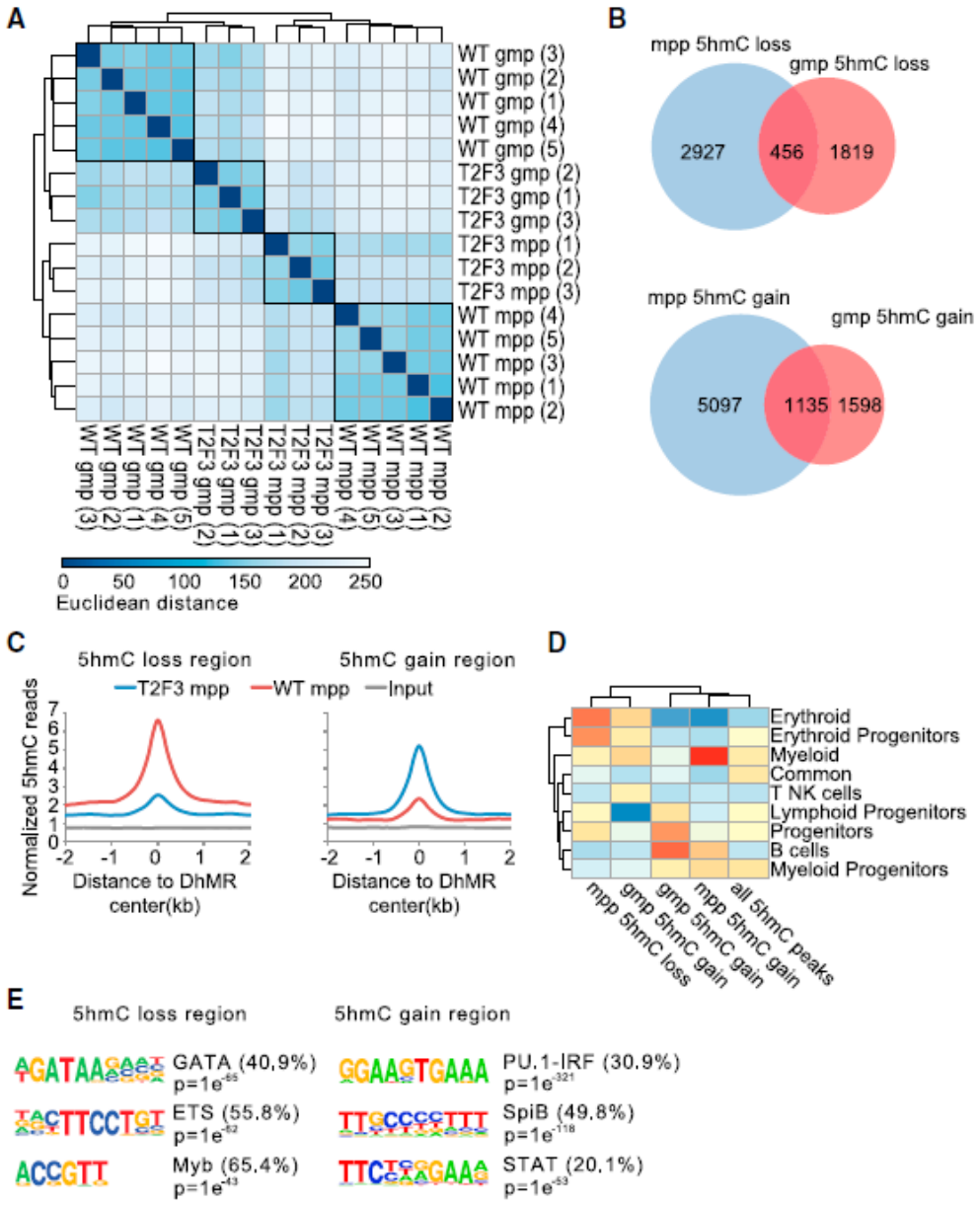


Fig. 3.6 Nano-hmC-Seal Reveals 5hmC Redistribution in a Murine AML Model. (A) Comparison of 5hmC profiles in bone marrow MPP and GMP cells from WT and *Tet2*$^{-/-}$; *Flt3*^{ITD} mice. Hierarchical clustering applied to the matrix of sample-to-sample distance based on rlogtransformed read counts in 272,087 detected 5hmC-enriched peaks is shown. (B) Venn diagram showing the overlap of detected DhMRs between MPP and GMP cells. (C) Distribution of 5hmC signals at DhMRs (left: 5hmC loss and right: 5hmC gain) in MPP cells. (D) Relative enrichment of the genomic overlap between DhMRs and lineage-specific enhancers. Only DhMRs with at least 2-fold 5hmC change, adjust $p < 0.1$, and mean of normalized counts >20 were used for analysis. All 5hmC peaks were used as control set. (E) De novo motif analysis by HOMER at DhMRs (left: 5hmC loss and right: 5hmC gain) in MPP cells.

We also observed an accumulation of 5hmC at the edge of DNA methylation canyons, which is a recently reported epigenetic feature of AML driven by mutations in epigenetic regulators (*28*). Although no significant variations of 5hmC were noticed within the canyon, a decrease of 5hmC outside the canyon boundary region was observed in AML samples compared to the WT (Fig. 3.7C), suggesting that the disruption of 5hmC distribution in the leukemia model may affect the maintenance of methylation canyons. To further confirm the functional influence of the observed 5hmC alternation, we performed RNA-seq of mRNA obtained from the same cells. As expected, the alternation of 5hmC positively correlated with gene expression (Fig. 3.8A). In MPP, we identified 366 differentially expressed genes that were associated with significant 5hmC changes. We performed gene ontology analysis (GO) and found that these genes are enriched in functions such as cell adhesion, inflammatory response, and regulation of immune effector process (Fig. 3.8B-C). Consistent with the motif analysis, we also confirmed the downregulation of Gata1/Gata2 and upregulation of Irf8 (Fig. 3.8D). Together, these results highlight the utility of nano-hmC-Seal as a sensitive tool to map 5hmC distribution and dynamic methylation/demethylation in rare cell populations.

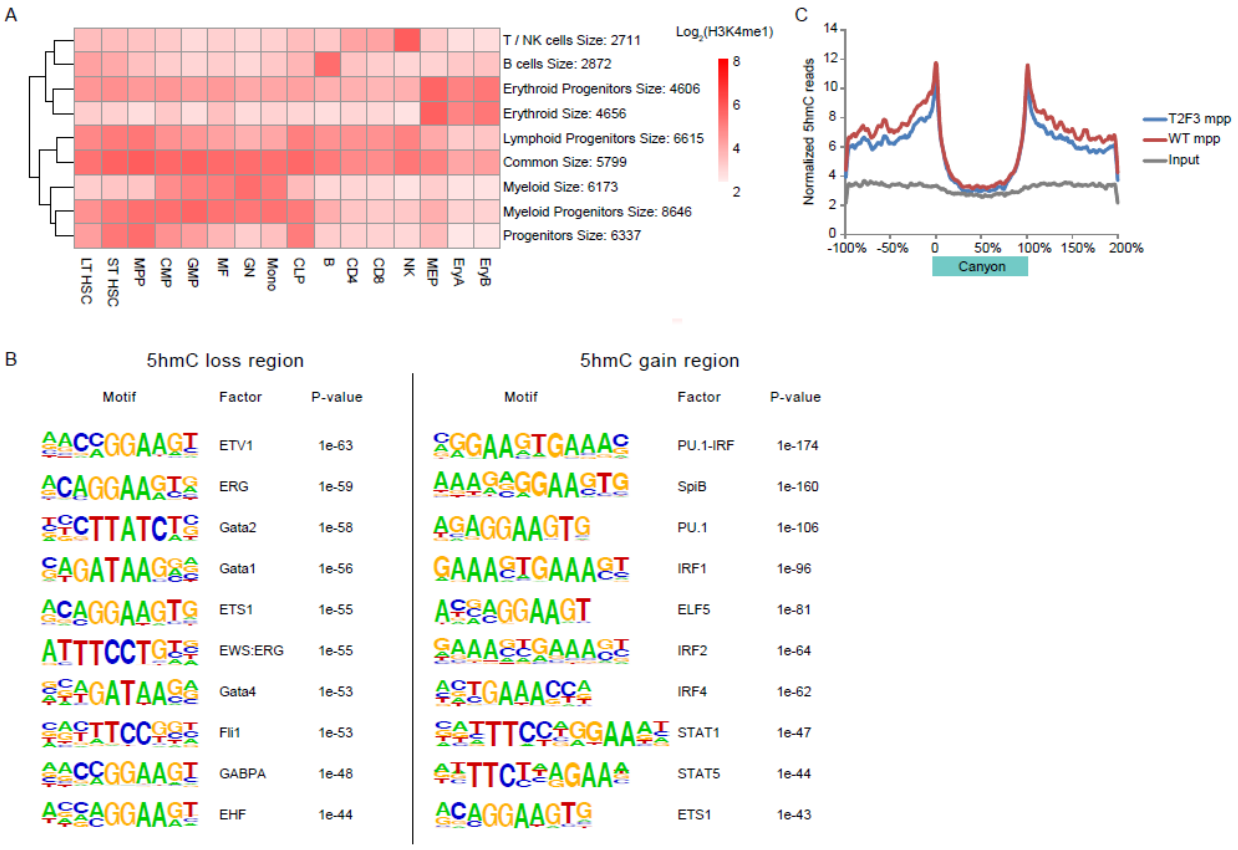


Fig. 3.7 The relationship between 5hmC and functional regulatory elements in WT or AML model mice. (A) The activity of lineage specific enhancers during hematopoiesis. Heatmap showing lineage-specific hematopoiesis enhancers with k-means cluster analysis of H3K4me1 signals (K=9). The genomic location of hematopoietic enhancers and H3K4me1 ChIP-Seq dataset were obtained from a previously published study (Lara-Astiaso et al., 2014). K-mean cluster analysis is performed by R package "pheatmap". (B) Top enriched known transcription factor binding motifs detected at DhMRs (left: 5hmC loss; right: 5hmC gain) in MPP cells. Motif information was obtained from the Homer motif database. (C) Normalized distribution profiles of 5hmC in MPP cells across DNA methylation canyon regions detected in hematopoietic stem cells. The genomic location of DNA canyon was obtained from a previously published study (Jeong et al., 2014).

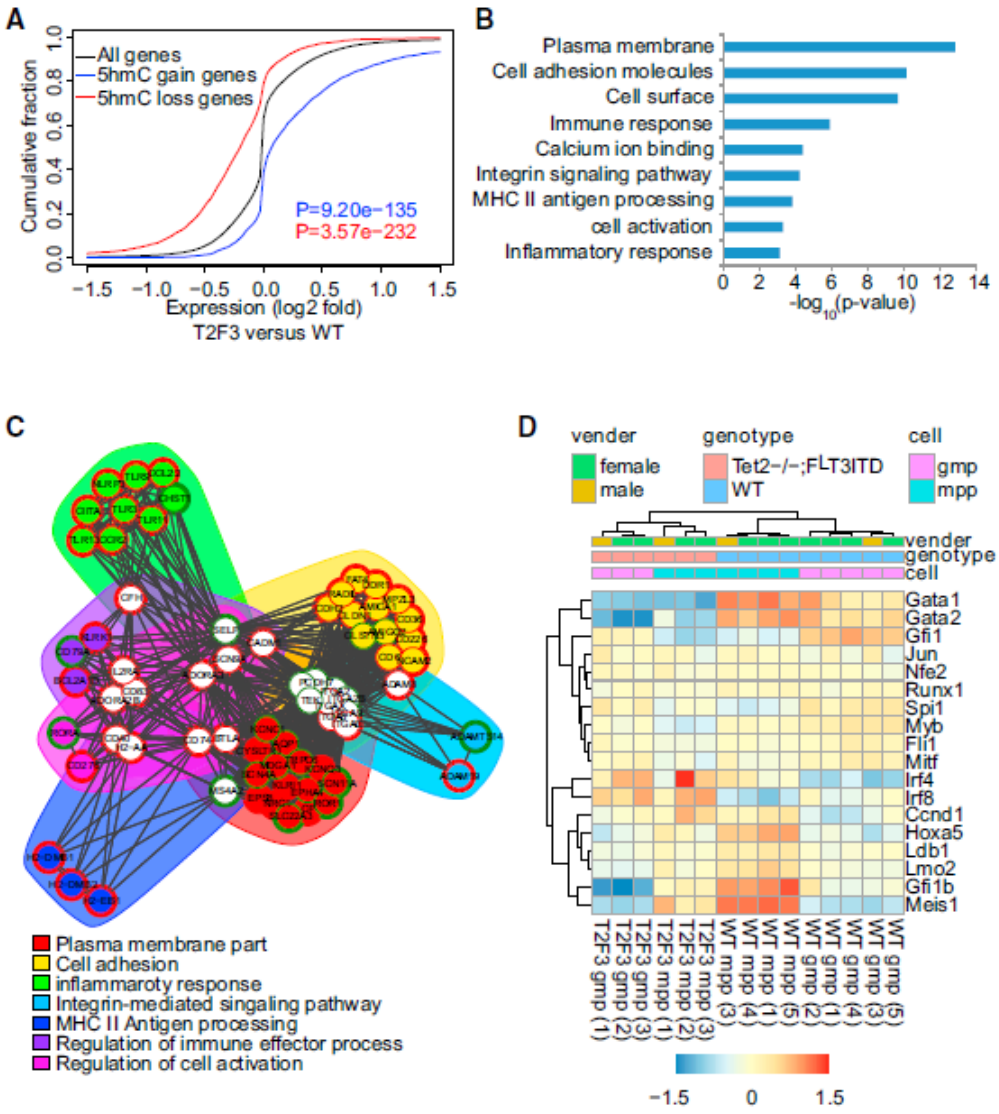


Fig. 3.8 Alternations of 5hmC in Gene Body Correlate with Gene Expression Changes in AML Model. (A) Cumulative distribution of 5hmC gain genes (blue) and 5hmC loss genes (red) correlate with expression changes in MPP cells from *Tet2*$^{-/-}$*; Flt3*^{ITD} (T2F3) mice versus WT mice. All genes were used as control. p values were calculated versus all genes (two-sided Wilcoxon rank-sum test). (B) GO enrichment analysis of 366 differentially expressed genes (|fold change| >2 and adjust p < 0.1) associated with significant 5hmC changes (adjust p < 0.1) in MPP cells. (C) Functional gene networks derived from GO enrichment analysis. The border color of the nodes denotes the upregulation (red) or downregulation (green) of genes in MPP cells from *Tet2*$^{-/-}$*; Flt3*^{ITD} versus WT mice. The genes in the same terms were linked to each other. The genes within the same clusters are surrounded by a common background color. (D) Heatmap of RNA expression to compare gene expression of hematopoietic transcriptional regulators in different samples. The genes and samples were clustered by Euclidean distance using centered rlog-transformed expression counts.

3.2.5 Apply Nano-hmC-Seal on prognosis of clinical outcome for AML patients with DAC treatment.

As Nano-hmC-Seal is proved to be a sensitive tool to map 5hmC distribution and dynamic methylation/demethylation in AML mice model, we then turned to apply this method to probe differential 5hmC in normal and pathologic disease states with limited samples and provide a robust platform to assess 5hmC localizations in tumor models and clinical samples in order to identify target loci and biomarkers with biologic and therapeutic relevance. We took bone marrow (BM) and peripheral blood (PB) from AML patients who received mito+hiDAC for treatment. As mito and hiDAC are both FDA approved drugs for potential AML cure by changing the distribution of epigenetic markers, we designed to utilize Nano-hmC-Seal to map the differential 5hmC regions between patient samples before and after treatment. In addition, we also noticed that the clinical outcome is different among patients after mito+hiDAC treatment. The post-diagnosis displayed three outcomes: CR (complete remission), CRi (partially remission) and TF (treatment failure). Based on that, studying differential of 5hmC distribution in above 3 classifiers could potentially help determining the prognosis outcome of mito+hiDAC treatment. Thus, it would benefit patients who are sensitive to the drug and at the same time, avoid tragic for potential treatment failure for patients who are resist to the drug.

Firstly, as a pilot study, we obtained DNA of bone marrow (BM) and peripheral blood (PB) from 85 AML patients who received mito+hiDAC treatment and performed Nano-hmC-Seal. Based on the 5hmC profiling data, we selected 1,050 differential gene among 48 AML samples (Bone marrow and before mito+hiDAC treatment). Via Hierarchical Clustering analysis, we found genes in cluster II shows distinct patterns in subgroups divided by diagnosis or treatment response (Fig. 3.9A-B). In addition, we carried out the PCA analysis for genes in cluster II and

found that CR (complete remission) samples showed distinct signatures and could be distinguished from TF (treatment failure) samples (Fig. 3.9C). To explore the function of 5hmC dysregulated genes, we performed GO term analysis for the differential 5hmC genes between TF and CR patients. We found that the leukemia related pathways such as lymphocyte activation and cytokine signaling were highly enriched in the up-regulated genes of CR patients (Fig. 3.9D).

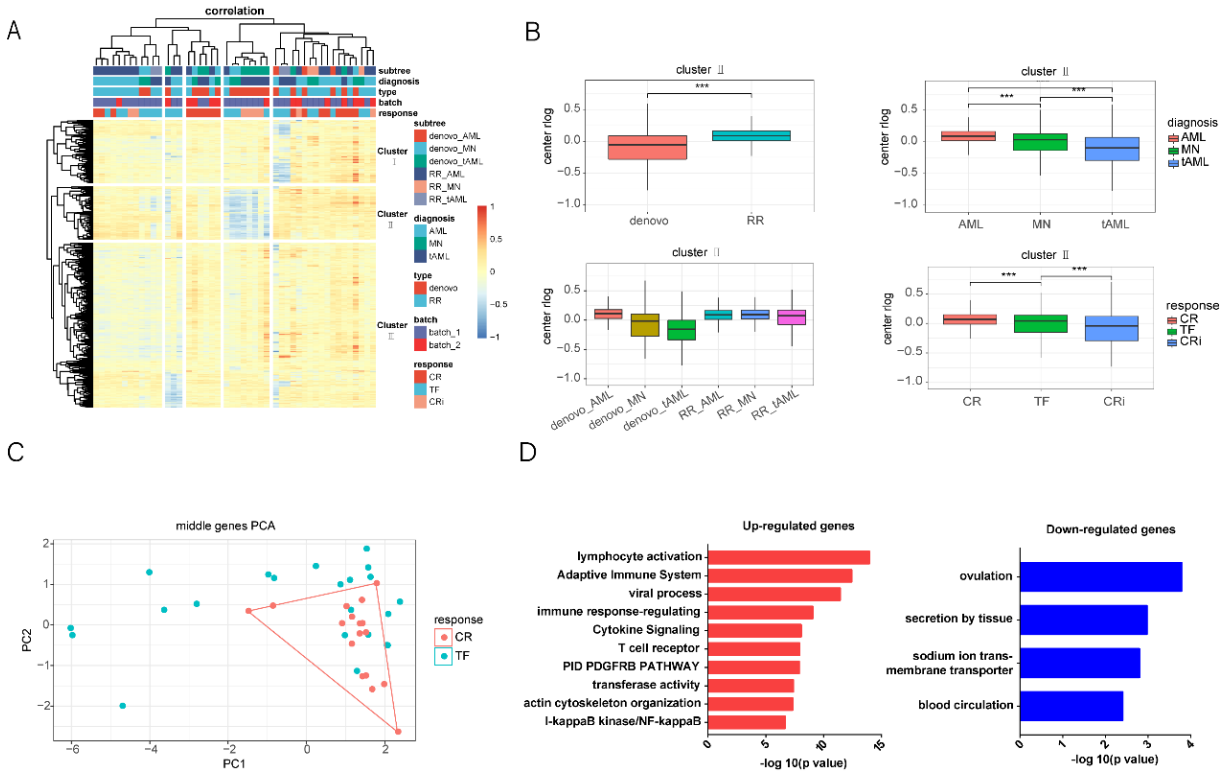


Fig. 3.9 Differential 5hmC genes in AML patients. (A) Clustering result of 1,050 differential 5hmC genes in 48 bone marrow samples of AML patients. Three differential genes clusters (I, II, III) were detected. (B) Boxplots to show normalized 5hmC levels of genes in cluster II. Samples were divided into subgroups according to the clinical information. (C) PCA plot of 5hmC levels of genes in the middle cluster from CR and TF samples. (D)GO enrichment analysis of 199 genes in the middle cluster.

Next, we turned to utilize 5hmC characteristics to predict the response to treatment in AML patients. We divided 85 samples from batch 1 into training group (68 samples) and validation groups (17 samples). 40 additional samples from batch 2 were used as an independent testing set.

By using XGBoost algorithm, the prediction performance achieved AUC = 0.97 for validation sets and 0.93 for independent test set (Fig. 3.10A). In contrast, models trained with random labeled samples showed a poor performance (AUC = 0.48). By applying our model, patients with TF response have significant lower predicted probability compared to CR and CRi patients (Fig. 3.10B), while the RR patients (Relapsed/Refractory) have lower probability compared to denovo patients (Fig. 3.10C). Noticeably, our results show that samples obtained from peripheral blood (PB) were similar with bone marrow (BM) samples (Fig. 3.10C), indicating that our prognosis method could be successfully applied to PB samples. In addition, top genes were selected according to the contribution in the model (Fig. 3.10D, E). The performance of classifier is stable with only five genes which show highest contribution (Fig. 3.10F).

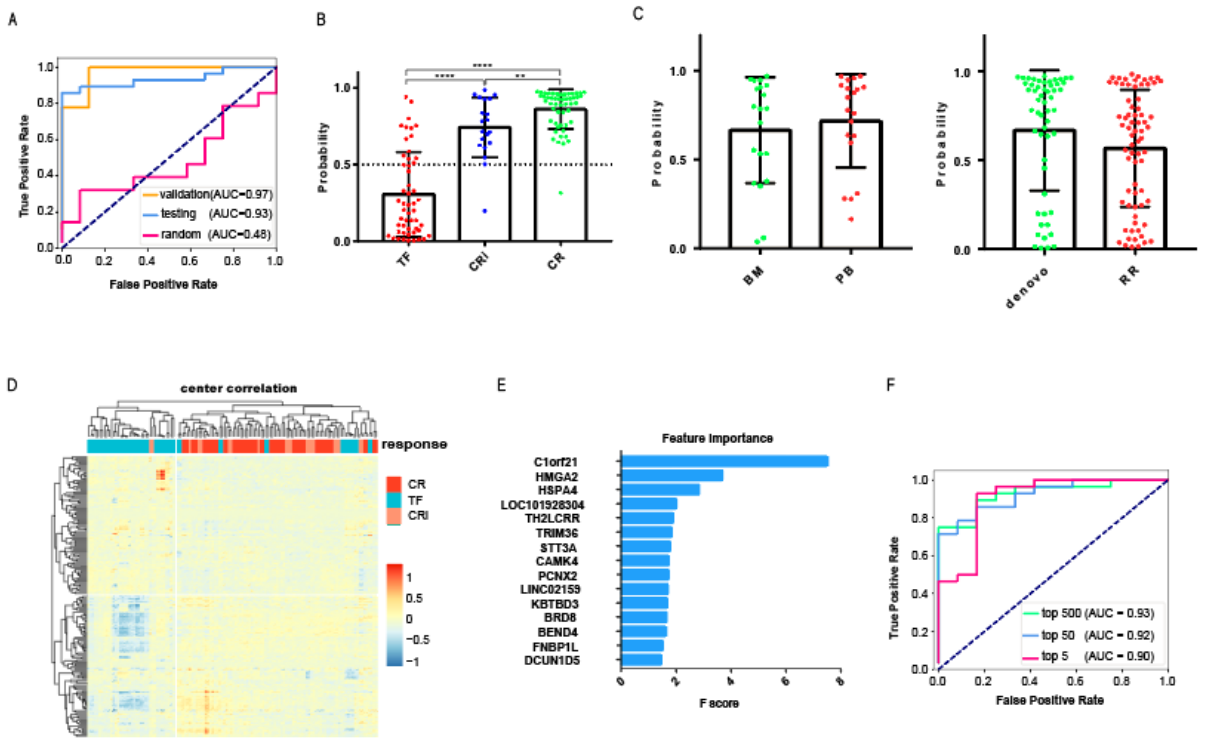


Fig. 3.10 Prediction of treatment response for AML patients. (A) Performance of 5hmC classifier to predict AML treatment response. A prognosis model with 5hmC data of AML patients was trained by using XGBoost algorithm. 68 and 17 samples were used as training and validation set from batch 1, 40 samples from batch 2 were used as an independent testing set. ROC curve was used to evaluate the performance of our model. (B)The predicted CR (complete remission) probability from 5hmC classifier shows a significant difference between CR, CRi (complete remission with incomplete blood count recovery) and TF (treatment failure) patients. (C)The probability was similar in bone marrow (BM) and peripheral blood (PB) samples. The RR samples (relapsed/refractory) show lower probability compared to denovo AML samples. (D)The top 400 important genes in model training were selected to plot the heatmap of 5hmC in 125 AML samples. CR and TF samples were clearly separated. (E) Bar plot showing the F score of top 15 important genes based on their contribution (feature importance) in model training. (F) Classifier performance is stable upon reducing the number of contributive genes.

Lastly, we compared the prognosis performance between classifier based on 5hmC profiling and RNA-Seq. We performed RNA-Seq in 59 AML samples. The performance of RNA classifier is similar to 5hmC classifier (RNA AUC=0.91, 5hmC AUC=0.87). However, if we build combination classifier based on both 5hmC-profiling and RNA-Seq, the model showed best performance (AUC=0.98) (Fig. 3.11A). 54.76% of contributed features are obtained from 5hmC-

profiling data. In addition, 72 genes were both detected in RNA and 5hmC classifier (Fig. 3.11A), including HMGA2, KCTD1, ARHGAP32 and EBF1 (Fig. 3.11B). We further obtained published RNA-Seq and clinical information of AML patients from TCGA database. In this dataset, patients with abnormal expression of these genes also showed poor prognosis (Fig. 3.11D).

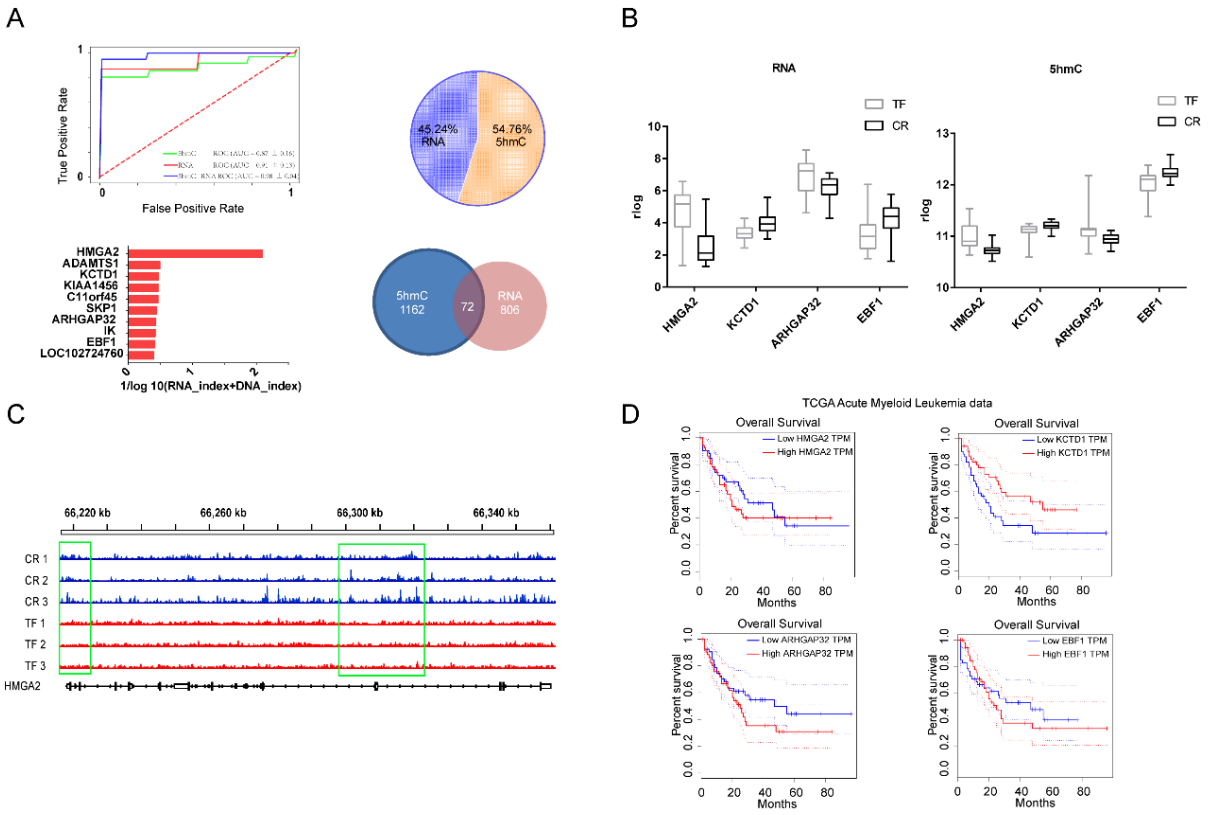


Fig. 3.11 Prognosis classifier between 5hmC and RNA-seq. (A) RNA-Seq was performed in 59 samples. The Classifiers were trained by 5hmC features only, RNA features only and the combination of 5hmC/RNA features. The combination classifier achieves the best performance. The pie chart shows the proportion of 5hmC and RNA features in the combination model. 1,234 5hmC features and 878 RNA features were detected as contributive genes in the classifier separately. 72 genes were detected in both classifiers and top 10 genes were showed. (B) Both RNA and 5hmC features of these 4 genes contribute significantly to the combination model. The 5hmC and RNA level of these genes are significantly different Between TF and CR patients. (C) The normalized 5hmC values of HMGA2 between CR and TF samples. (D) The survival curves based on TCGA Acute Myeloid Leukemia data. The samples were divided into two groups based on gene expression level.

3.3 Experiment section

3.3.1 Nano-hmC-Seal Protocol

Genomic DNA is extracted from cells using Quick-gDNA MicroPrep (Zymo) according to the manufacture's instruction. Tagmentation reactions were performed in a 50 ml solution with 13 tagmentation buffer from Nextera DNA Sample Preparation Kit (FC-121-1031). The input DNA ranged from 5 ng to 50 ng. Reactions were performed and purified according to the manufacturer's instruction. In brief, the fragmentation reaction was performed in 50 ml solutions containing 25 ml 23 tagmentation buffer, gDNA, and 0.5-5 3l Tagmentase, at 55 ˚C for 5 min. The fragmented DNA was eluted into 17.5 ml ddH_2O. The glucosylation reactions were performed in a 20 ml solution containing 50 mM HEPES buffer (pH 8.0), 25 mM $MgCl_2$, purified DNA, 100 mM N_3-UDP-Glc, and 1 mM βGT, at 37 ˚C for 1 hr. After the reaction, 2 ml DBCO-PEG4-biotin (Click Chemistry Tools, 20 mM stock in DMSO) was added to the reaction mixture, and the mixture was incubated at 37 ˚C for 2 hr. Next, the modified DNA was purified by a Micro Bio-Spin 30 Column (Bio-Rad). The purified DNA was incubated with 5 ml C1 Streptavidin beads (Life Technologies) in 2X buffer (1X buffer: 5 mM Tris [pH 7.5], 0.5 mM EDTA, and 1 M NaCl) for 15 min according to the manufacture's instruction. The beads were subsequently undergone six 5 min washes with 1X buffer. All binding and washing were done at room temperature with gentle rotation. The captured DNA fragments were amplified with 12–17 cycles of PCR amplification using the enzyme mix supplied in the Nextera kit. The PCR products were purified using AMPure XP beads according to the manufacture's instruction. Separate input libraries were made by direct PCR from fragmented DNA without labeling and capture according

to the manufacture's instruction in the Nextera kit. DNA concentration of each library was measured with a Qubit fluorometer (Life Technologies). Sequencing was performed on the HiSeq instrument.

3.3.2 Cell Culture

mESCs were cultured in feeder-free gelatin-coated plates in Dulbecco's modified Eagle medium (DMEM) (Invitrogen Cat. No. 11995) supplemented with 15% fetal bovine serum (FBS) (Gibco), 2 mM L-glutamine (Gibco), 0.1 mM 2-mercaptoethanol (Sigma), 13 nonessential amino acids (Gibco), 1,000 units/ml LIF (Millipore Cat. No. ESG1107), 13 pen/strep (Gibco), 3 mM CHIR99021 (Stemgent), and 1 mM PD0325901 (Stemgent). Half of the medium was changed every day for 5 days before harvesting mEBs by sedimentation.

3.3.3 Isolation of Hematopoietic Progenitor Cells

Mouse strains, antibody staining, and FACS protocol were as previously described (Shih et al., 2015). Briefly, leg and arm bones were isolated from mice and bone marrow was isolated by centrifugation at 6,000 rcf for 1 min. Cells were then lysed in ACK lysis buffer. Stem cell enrichment was performed using the Progenitor Cell Enrichment Kit (STEMCELL Technologies). Antibodies used for flow cytometry were as follows: (anti-mouse) Gr1 (Ly6G), B220 (RA3-682), CD34 (RAM34), CD16/32 (93) Sca-1 (D7), cKit (2B8), Mac-1 (CD11b) (M1/70), NK1.1 (PK136), Ter119 (Ter119,553673), CD3 (145- 2C11), CD150 (TC-15-12F2.2), and CD48 (HM48-1) from BioLegend and eBioscience. The ‘‘lineage cocktail’’ included CD3, Gr-1, Mac-1 (CD11b), NK1.1, B220, and Terr-119. FACS was performed on a BD FACSAria sorter. DNA and RNA were isolated using the AllPrep DNA/RNA Mini Kit from QIAGEN.

All animal procedures were conducted in accordance with the Guidelines for the Care and Use of Laboratory Animals and were approved by the Institutional Animal Care and Use Committees at Memorial Sloan Kettering Cancer Center.

3.3.4 RNA-Seq and Analysis

RNA-seq libraries were constructed by TrueSeq Stranded mRNA Sample Preparation Kit (Illumina). The reaction was performed according to the manufacture's instruction. Sequencing reads were aligned to mm9 genome by STAR (*29*). Count matrices were generated by summarizeOverlaps functionality from R package GenomicAlignments (*30*). Differential expression analysis was performed by R package DESeq2 (*31*).

3.3.5 Data Processing and Analysis

Illumina reads were post-processed and aligned to the mouse mm9 assembly using the bowtie program with default parameters. To visualize sequencing signals in the genome browser, we generated bedGraph files with HOMER (*32*). deepTools (*33*) software was used to plot the heatmap of signal distribution, perform K-means clustering, and calculate genome-wide correlations (2,000-bp tiling regions). The identification of 5hmC-enriched regions (peaks) in each sample was performed using MACS (*34*). Peaks that were detected by all replicates were considered as high confident peaks. Peaks from all samples were combined into one unified catalog for each study separately by ''mergePeak'' functionality from HOMER (*32*) *2*. Potential library complexity was determined by using the PreSeq (*15*) package with extrapolate function. To show the correlation between Nano-hmC-Seal and standard hmC-Seal in mESC (or between replicates), tag counts in merged high-confident peaks were calculated by HOMER. The log2 transform values

were then used to generate scatterplots and calculate the Pearson correlation. The data set of standard hmC-Seal using bulk mESCs was obtained from a previously published study (*35*). The fractions of reads in high density 5hmC clusters were calculated as a measure of specific enrichment. High density 5hmC cluster were regions that contained at least five TAB-Seq detected 5hmCs, each within 200 bp of each other. To detect 5hmC-containing CGIs from an ox-RRBS data set, R package methylKit (*36*) was used for analysis.

To assess overall difference between samples, we calculated the Euclidean distance based on rlog-transformed 5hmC signals and visualized the distance in a heatmap figure by using an R package ‘‘pheatmap’’. To detect the genotype-specific DhMR or genes, the Bioconductor DESeq2 packages was used for analysis. Functional annotation of DhMRs was obtained with GREAT (*23*). To detect DMR from an RRBS data set, R package methylKit (*36*) was used for analysis. The HOMER software was used to perform de novo motif analysis at ±500bp around DhMR. GO term analyses were performed by DAVID (*37*) and visualized as functional gene network by FGNet (*38*).

3.3.6 Definition of Lineage Specific Enhancer Subgroups

The genomic locations of 48,415 enhancers identified across 16 stages of hematopoietic differentiation were previously defined (*19*). To be consistent with previous studies, these enhancers were further grouped into nine major clusters based on log-transformed H3K4me1 level using K-means cluster, as described before (*19*).

3.4 Discussion and future perspective

Several methods have been developed to map 5hmC genome wide. Bisulfite-based approaches such as TAB-seq (*11*) and oxBS-seq (*5*) provide the most comprehensive and quantitative information because they are able to detect 5hmC at single-base resolution. However, these methods can be prohibitively expensive due to the requirement for high sequencing coverage. The oxidation and bisulfite treatments also lead to substantial degradation of genomic DNA, which limits their utility to investigate systems with limited input materials. By contrast, enrichment-based profiling methods require lower sequencing costs and can be routinely applied to 5hmC studies. However, antibody-based enrichment still requires microgram levels of input genomic DNA and can introduce sequence biases. In addition, it cannot be used to profile biological processes or clinical samples where target cells or input DNA exist in rare quantities and often are difficult-toobtain. We show here that nano-hmC-Seal, an approach based on selective chemical labeling and capture, provides a highly sensitive and robust method to profile 5hmC with a few nanograms of genomic DNA (~1,000 cells). We have demonstrated the utility of this method in the study of HSC differentiation from mouse models. Importantly, we show that the covalent labeling and pull-down technology could be widely applied to profile 5hmC in sorted homogenous neurons and stem cells, clinical biopsy samples, circulating tumor cells, and cell-free DNA. We anticipate expansion of this approach into the study of dynamics of 5mC with limited input DNA materials when combined with Tet-mediated covalent labeling of 5mC (TAmC-seq) (*39*). Because 5hmC serves as a potential indicator of active DNA demethylation, a crucial mechanism commonly associated with transcriptional activation, the distribution patterns of 5hmC not only provide a global view of gene activation, but also uncover insight into the use of lineage/disease-specific enhancers and promoters. As such, differential 5hmC patterns reflect

the underlying gene expression patterns and the binding activities of transcription factors at enhancer regions, especially for pioneer factors. Our studies using nano-hmC-Seal in normal HSC and progenitor cells and in AML LSCs allowed us to define multiple specific loci with dynamic changes in 5hmC patterns that are enriched in enhancers and gene bodies with a known or putative role in normal and malignant hematopoiesis and which strongly correlate with differential gene expression. Therefore, integrated states of 5hmC at enhancer and gene-body regions in patients can potentially unravel the molecular mechanisms that drive transformation from normal stem/progenitor cells to different malignancies. Our results reveal new insights of 5hmC dynamics during HSC differentiation: (1) gene-associated 5hmC is positively correlated with active histone modifications, in particular with H3K4me1/2 changes. The genes proximal to these regions are annotated to blood cell functions including leukocyte/erythrocyte differentiation and immunological disorders and (2) 5hmC peaks tend to locate in genomic regions that are marked by less strong ATAC-seq signal intensity and are depleted in the center of ATAC-seq peaks that may mark high chromatin accessibility. We identified regions with gain and loss of the Tet2-deficient LSCs, indicating that the inactive mutation of Tet2 does not simply lead to loss of 5hmC, but instead drives a redistribution of this modification (*40*). In addition, we identified PU.1 motifs, a key factor required for the generation of myeloid and lymphoid cells (*27*), as significantly enriched at sites with increased 5hmC. PU.1 is known to drive different normal and pathogenic hematopoietic functions by co-binding with various transcription factors including c-Jun, C/EBPA, GATA1/2, and IRF4/8 (*41*). The identification of a significant PU.1: IRF ($p = 1 \times 10^{-321}$) motif discovered by de novo motif analysis in these loci implies a specific PU.1 regulatory mechanism associated with active demethylation and 5hmC distribution. It will be of interest to

investigate whether the PU.1/IRF complex has a functional role in leukemogenesis. These observations also open an interesting question about how the binding of PU.1, which was reported to physically interact with TET2 (*42*), contributes to 5hmC redistribution induced by inactivation of TET2. One possible explanation is that PU.1 might also interact with other Tet proteins, thus leading to a unique genomic binding pattern in the absence of TET2. Our 5hmC profiling further showed that regions with loss of 5hmC in AML stem cells exhibit significant enrichment of binding motifs for GATA factors, which are known to be mutated and altered by translocations and differential expression in AML (*43*). This result underscores our previous observation that LSCs have reduced expression of *Gata2* in the same AML mice models, and that restoration of *Gata2* expression abrogates leukemogenesis in *Tet2/Flt3*- mutant AML (*26*). Together, these findings suggest a multi-layered epigenetic regulation of GATA factors by repressing transcription initiation and silencing of downstream binding sites. These studies demonstrate the utility of nano-hmC-Seal or similar approaches based on covalent 5hmC labeling to probe differential 5hmC in normal and pathologic disease states with limited samples and provide a robust platform to assess 5hmC localizations in tumor models and clinical samples in order to identify target loci and biomarkers with biologic and therapeutic relevance.

3.5 References

1. J. A. Hackett *et al.*, Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine. *Science* **339**, 448-452 (2013).
2. X. Lu, B. S. Zhao, C. He, TET family proteins: oxidation activity, interacting molecules, and functions in diseases. *Chemical reviews* **115**, 2225-2239 (2015).
3. W. Sun, L. Zang, Q. Shu, X. Li, From development to diseases: the role of 5hmC in brain. *Genomics* **104**, 347-351 (2014).
4. S. L. Topalian *et al.*, Safety, activity, and immune correlates of anti–PD-1 antibody in cancer. *New England Journal of Medicine* **366**, 2443-2454 (2012).

5. M. J. Booth *et al.*, Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934-937 (2012).

6. L. Cui, T. H. Chung, D. Tan, X. Sun, X.-Y. Jia, JBP1-seq: a fast and efficient method for genome-wide profiling of 5hmC. *Genomics* **104**, 368-375 (2014).

7. W. A. Pastor *et al.*, Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473**, 394 (2011).

8. C.-X. Song *et al.*, Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nature biotechnology* **29**, 68 (2010).

9. C.-X. Song, C. Yi, C. He, Mapping recently identified nucleotide variants in the genome and transcriptome. *Nature biotechnology* **30**, 1107 (2012).

10. Z. Sun *et al.*, A sensitive approach to map genome-wide 5-hydroxymethylcytosine and 5-formylcytosine at single-base resolution. *Molecular cell* **57**, 750-761 (2015).

11. M. Yu *et al.*, Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368-1380 (2012).

12. A. Adey *et al.*, Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome biology* **11**, R119 (2010).

13. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**, 1213 (2013).

14. C. Schmidl, A. F. Rendeiro, N. C. Sheffield, C. Bock, ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nature methods* **12**, 963 (2015).

15. T. Daley, A. D. Smith, Predicting the molecular complexity of sequencing libraries. *Nature methods* **10**, 325 (2013).

16. K. Rice, I. Hormaeche, J. Licht, Epigenetic regulation of normal and malignant hematopoiesis. *Oncogene* **26**, 6697 (2007).

17. A. H. Shih, O. Abdel-Wahab, J. P. Patel, R. L. Levine, The role of mutations in epigenetic regulators in myeloid malignancies. *Nature reviews Cancer* **12**, 599 (2012).

18. V. Moignard *et al.*, Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nature cell biology* **15**, 363 (2013).

19. D. Lara-Astiaso *et al.*, Chromatin state dynamics during blood formation. *science* **345**, 943-949 (2014).

20. J. Lekstrom-Himes, K. G. Xanthopoulos, CCAAT/enhancer binding protein ε is critical for effective neutrophil-mediated response to inflammatory challenge. *Blood* **93**, 3096-3105 (1999).

21. R. Gazit *et al.*, Transcriptome analysis identifies regulators of hematopoietic stem and progenitor cells. *Stem cell reports* **1**, 266-280 (2013).

22. X. Lu *et al.*, Base-resolution maps of 5-formylcytosine and 5-carboxylcytosine reveal genome-wide DNA demethylation dynamics. *Cell research* **25**, 386 (2015).

23. C. Y. McLean *et al.*, GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28**, 495 (2010).

24. G. C. Hon *et al.*, 5mC oxidation by Tet2 modulates enhancer activity and timing of transcriptome reprogramming during differentiation. *Molecular cell* **56**, 286-297 (2014).
25. M. B. Stadler *et al.*, DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, (2011).
26. A. H. Shih *et al.*, Mutational cooperativity linked to combinatorial epigenetic gain of function in acute myeloid leukemia. *Cancer cell* **27**, 502-515 (2015).
27. J. C. Walsh *et al.*, Cooperative and antagonistic interplay between PU. 1 and GATA-2 in the specification of myeloid cell fates. *Immunity* **17**, 665-676 (2002).
28. M. Jeong *et al.*, Large conserved domains of low DNA methylation maintained by Dnmt3a. *Nature genetics* **46**, 17 (2014).
29. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
30. M. Lawrence *et al.*, Software for computing and annotating genomic ranges. *PLoS computational biology* **9**, e1003118 (2013).
31. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550 (2014).
32. S. Heinz *et al.*, Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38**, 576-589 (2010).
33. F. Ramírez, F. Dündar, S. Diehl, B. A. Grüning, T. Manke, deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic acids research* **42**, W187-W191 (2014).
34. J. Feng, T. Liu, B. Qin, Y. Zhang, X. S. Liu, Identifying ChIP-seq enrichment using MACS. *Nature protocols* **7**, 1728 (2012).
35. C.-X. Song *et al.*, Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **153**, 678-691 (2013).
36. A. Akalin *et al.*, methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome biology* **13**, R87 (2012).
37. D. W. Huang, B. T. Sherman, R. A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* **37**, 1-13 (2008).
38. S. Aibar, C. Fontanillo, C. Droste, J. De Las Rivas, Functional Gene Networks: R/Bioc package to generate and analyse gene networks derived from functional enrichment and clustering. *Bioinformatics* **31**, 1686-1688 (2015).
39. L. Zhang *et al.*, Tet-mediated covalent labelling of 5-methylcytosine for its genome-wide detection and sequencing. *Nature communications* **4**, 1517 (2013).
40. J. Madzo *et al.*, Hydroxymethylation at gene regulatory regions directs stem/early progenitor cell commitment during erythropoiesis. *Cell reports* **6**, 231-244 (2014).
41. P. Gupta, G. U. Gurudutta, D. Saluja, R. P. Tripathi, PU. 1 and partners: regulation of haematopoietic stem cell fate in normal and malignant haematopoiesis. *Journal of cellular and molecular medicine* **13**, 4349-4363 (2009).
42. L. de la Rica *et al.*, PU. 1 target genes undergo Tet2-coupled demethylation and DNMT3b-mediated methylation in monocyte-to-osteoclast differentiation. *Genome biology* **14**, R99 (2013).

43. S. Gröschel *et al.*, A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell* **157**, 369-381 (2014).