# Computational analysis and characterization of dysregulated chromatin interactions and RNA biology in acute myeloid leukemia

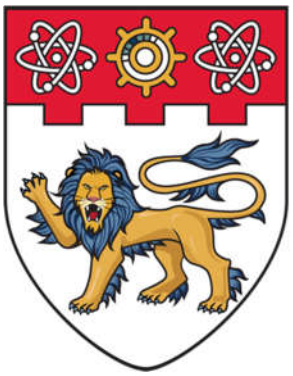
Kong, Lingshi

2021

Kong, L. (2021). Computational analysis and characterization of dysregulated chromatin interactions and RNA biology in acute myeloid leukemia. Doctoral thesis, Nanyang Technological University, Singapore. https://hdl.handle.net/10356/155629

https://hdl.handle.net/10356/155629

https://doi.org/10.32657/10356/155629



NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

# Computational Analysis and Characterization of Dysregulated Chromatin Interactions and RNA Biology in Acute Myeloid Leukemia

**KONG LINGSHI**

**SCHOOL OF BIOLOGICAL SCIENCES**

**2021**

# Computational Analysis and Characterization of Dysregulated Chromatin Interactions and RNA Biology in Acute Myeloid Leukemia

**KONG LINGSHI**

**SCHOOL OF BIOLOGICAL SCIENCES**

A thesis submitted to the Nanyang Technological University in partial fulfilment of the requirement for the degree of Doctor of Philosophy

2021

## Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research done by me except where otherwise stated in this thesis. The thesis work has not been submitted for a degree or professional qualification to any other university or institution. I declare that this thesis is written by myself and is free of plagiarism and of sufficient grammatical clarity to be examined. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.


20/8/2021

. . . . . .

Date

. . . . . .

KONG LINGSHI

## Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it of sufficient grammatical clarity to be examined. To the best of my knowledge, the thesis is free of plagiarism and the research and writing are those of the candidate's except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.



20/8/2021

. . . . . .

Date

. . . . . .

Dr. Melissa Jane Fullwood

# Authorship Attribution Statement

This thesis contains material from 1 paper is submitted and in reviewing process the following peer-reviewed journal(s) *Genomics, Proteomics & Bioinformatics* in which I am listed as an author.

**Chapter 3 is submitted and in reviewing process as:**

Wang, B., Kong, L., Babu, D., Choudhary, R., Fam, W., Tng, J. Q., Goh, Y., Liu, X., Song, F. F. & Chia, P. 2020. Three-dimensional Genome Organization Maps in Normal Haematopoietic Stem Cells and Acute Myeloid Leukemia. *bioRxiv*. (In reviewing process at *Genomics, Proteomics & Bioinformatics*). doi: https://doi.org/10.1101/2020.04.18.047738

The contributions of the co-authors are as follows:

- All the bioinformatics analyses for RNA-Seq, Hi-C excluding loop comparisons analysis, and ChIP-Seq excluding ChIP-Seq analysis on 63 published patient samples, were performed by me, including alignment, quality filtering, normalization, Hi-C clustering, TAD and loop calling, and enhancer and super enhancer calling.
- Dr. Benny Wang Zhengjie generated the CRISPR clones and performed 4C, ddPCR, qPCR, ChIP-qPCR, and growth assays.
- I prepared the 4C interaction tracks from the 4C result produced by Dr.

Benny Wang Zhengjie

- Dr. Deepak prepared the AML and Femur samples for Hi-C by Dovetail genomics, generated the AML total bone marrow Hi-C libraries and performed manual curation of the Hi-C data, and prepared the H3K27ac ChIP-Seq libraries on AML MNC samples.

- I prepared the manuscript with Dr. Benny Wang Zhengjie and Dr. Deepak Babu.

- Dr. Deepak and Ms. Winnie Fam prepared the clinical samples for RNA-Sequencing.

- Ms. Goh Yufen, Fam Wee Nih, and Ms. Tng Jiaqi collected femur and bone marrow samples and isolated CD34+ and CD33+ cells.

- Mr. Bertrand Wong Jern Han performed bioinformatics analyses to understand common and unique loops between AML and femur samples.

- Ms. Ruchi Choudhary performed bioinformatics analysis on published AML cases to identify SEs and their looping patterns in AML.

- The 4C and RNA-seq analysis were performed using the Cancer Science Institute Web Portal pipelines that were modified by Dr Omer An and Professor Henry Yang

- Clinical acute myeloid leukemia and healthy knee derived bone

marrow samples were obtained through Professor Chng Wee Joo and Professor Wilson Wang from the National University Hospital. The patient consent for these clinical samples was approved by Dr. Ming Chun Chan and coordinated by Ms. Xin Liu, Ms. Fang Fang Song, and Ms. Priscella Chia.

- Dr. Melissa Fullwood provided guidance towards the project direction and edited the manuscript.

20/8/2021

. . . . . .

Date

. . . . . .

KONG LINGSHI

# Table of Contents

# Acknowledgments

First, I would like to express my sincere gratitude to my supervisor Dr. Melissa Fullwood for her warm support and patient guidance throughout my 4-year Ph.D. study. I am very proud to be one of her students. She is a great mentor full of patience and energy to courage me to improve my skills in bioinformatics and gave a lot of significant suggestions in my understanding of 3D genome organizations.

I would like to appreciate all the lab members in the M.J.F lab for their support and continuous care, especially Dr. Benny Wang Zhengjie and Dr. Deepak, for their help and hardwork in performing experiments and collaborating with me to complete the AML paper, Ms. Kaijing Chen, Dr. Cao Fan, Dr. Yichao Cai and Mr. See Yixiang for interesting discussion in bioinformatics, and thanks to Dr. Deepak again for his work in *DNMT3A* story.

Moreover, I would like to acknowledge Dr. Zhou Qiling from Prof. Daniel Tenen's lab for assistant in K562 *DNMT3A* knockout, and my lab collaborators Prof. Wilson Wang from the National University Hospital, and Prof. Chng Wee Joo from the National University Hospital for providing the clinical samples.

Last but not least, I would like to thank my parents, family, and friends for their kind concerning, and heart-warming support throughout the journey.

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| **3C** | Chromatin Conformation Capture |
| **3D-FISH** | 3 Dimensional Fluorescence In Situ Hybridization |
| **4C** | Circularised Chromosome Conformation Capture |
| **5C** | Carbon Copy Chromosome Conformation Capture |
| **AML** | Acute Myeloid Leukemia |
| **ARID5B** | AT-Rich Interactive Domain-containing Protein 5B |
| **BRCA1** | Breast Cancer type 1 susceptibility protein |
| **CD70** | Cluster of Differentiation 70 |
| **ChIA-PET** | Chromatin Interaction Analysis by Paired-End Tag |
| **ChIP** | Chromatin Immunoprecipitation |
| **ChIP-qPCR** | Chromatin Immunoprecipitation-Quantitative Polymerase Chain Reaction |
| **ChIP-seq** | Chromatin Immunoprecipitation-Sequencing |
| **CML** | Chronic Myeloid Leukemia |
| **CNV** | Copy Number Variation |
| **CPM** | Counts of exon model Per Million mapped reads |
| **CTCF** | CCCTC-Binding Factor |
| **DMNT** | DNA Methyltransferase |
| **DOT1L** | Disruptor Of Telomeric silencing 1 |
| **EV** | Empty Vector |
| **FIRE** | Frequently Interacting Region |
| **FPKM** | Fragments Per Kilobase of exon model per Million mapped fragments |
| **GEO** | Gene Expression Omnibus |
| **GPU** | Graphics Processing Unit |

| | |
|---|---|
| **H3K27Ac** | Histone 3 Lysine 27 Acetylation |
| **H3K27Me3** | Histone 3 Lysine 27 Tri-Methylation |
| **H3K4Me3** | Histone 3 Lysine 4 Tri-Methylation |
| **Hi-C** | High-throughput Chromosome Conformation Capture |
| **HOXA9** | Homeobox A9 |
| **ICE** | Iterative Correction and Eigenvector decomposition method |
| **IDH** | Isocitrate Dehydrogenase |
| **LMA** | Ligation Mediated Amplification |
| **MACC1** | Metastasis-Associated in Colon Cancer Protein 1 |
| **MEIS1** | Myeloid Ectropic Viral Integration Site 1 |
| **MLL** | Mixed Lineage Leukemia |
| **MNC** | Mononuclear Cells |
| **mRNA** | Messenger RNA |
| **NIH** | National Institutes of Health |
| **PBS** | Phosphate-Buffered-Saline |
| **PCA** | Principal Component Analysis |
| **PHF2** | Plant Homeodomain Finger Protein 2 |
| **PLOD2** | Procollagen-Lysine,2-Oxoglutarate 5-Dioxygenase 2 |
| **RARA** | Retinoic Acid Receptor Alpha |
| **ROBO1** | Roundabout homolog 1 |
| **ROSE** | Ranking of Super Enhancers |
| **RPM** | Reads of exon model Per Million mapped reads |
| **RT-qPCR** | Reverse Transcriptase-Quantitative Polymerase Chain Reaction |
| **SE** | Super Enhancer |

| | |
|---|---|
| **SEPT9** | Septin-9 |
| **SNP** | Single Nucleotide Polymorphism |
| **SS** | Super Silencer |
| **TAD** | Topologically Associating Domain |
| **TCGA** | The Cancer Genome Atlas Program |
| **TF** | Transcription Factor |
| **TPM** | Transcripts Per kilobase of exon model per Million mapped reads |
| **TSG** | Tumor Suppressor Gene |
| **TSS** | Transcription Starting Site |
| **VC** | Vanilla Coverage method |
| **WHO** | World Health Organization |

# Summary

Cancer is a highly lethal disease. Epigenetics has been found to be influential in cancer biology. Acute Myeloid Leukemia (AML), a disease derived from the aberrant differentiation and proliferation of haematopoietic progenitor cells, has been found to have a tight connection with epigenetics.

Here we investigated how chromatin interactions, a type of epigenetic, are dysregulated in AML clinical samples and *DNMT3A* mutant myeloid leukemia. We obtained and analyzed the 3D genome organization maps through Hi-C in both AML and normal CD34+ clinical haematopoietic stem cells as well as *DNMT3A* CRISPR knockout K562 cells. Altered chromatin interactions were found in AML and *DNMT3A* CRISPR knockout K562 cells.

A Frequently Interacting Region (FIRE) in the *MEIS1* region was found to be absent in half of AML samples (4 of 8) which showed low *MEIS1* levels compared with normal samples and AML samples with the FIRE. The CRISPR excision of a CTCF binding site at the border of this FIRE led to *MEIS1* expression loss, loss of chromatin interactions between the *MEIS1* promoter with enhancers, modulation of H3K27ac levels at enhancers, and reduced cell growth.

To address the influence of *DNMT3A* mutations on chromatin interactions, clinical AML RNA-Seq from an online database was analyzed, which suggested that *DNMT3A* mutations are associated with dysregulated Topologically Associated Domain (TAD) boundaries. From Hi-C analysis, the loss of two FIREs and two loops were also observed in *DNMT3A* CRISPR knockout K562 cells, which was associated with downregulation of *PLOD2*, *MACC1,* and *ARID5B*. Further integrated analysis of CTCF and histone mark ChIP-Seq, as

well as RNA-Seq, suggested that *DNMT3A* loss led to altered histone marks, CTCF binding, chromatin interactions, and gene expression.

Taken together, our work provided a better understanding of chromatin interactions alterations and gene expression changes in AML and *DNMT3A* mutant myeloid leukemia. Our research indicates the relevance of chromatin interactions in cancer biology and suggests that drugs that modulate epigenetic, such as DNA methylation, may lead to changes in chromatin interactions. In future research, we are interested to develop therapeutic strategies for altering the dysregulated chromatin interactions seen in AML through epigenetic drugs.

# 1. Introduction

## 1.1 Cancer: An Epigenetic Disease

### 1.1.1 Epigenetics: Another Way to Influence Life

Genes control many aspects of our lives. The DNA sequences of genes play crucial aspects in coding for many functions in our bodies. However, our appearances and health are not controlled just by genetics alone but also by our environment, our behaviors, and habits. How do these factors engage in our lives without changing our DNA sequences? The study of these mysteries is called epigenetics.

Previously, the concept of epigenetics was first defined by Conrad Waddington as “the branch of biology which studies the causal interactions between genes and their products which bring the phenotype into being” (Waddington, 1942, 1968). With the rapid development of biology research, scientists have become more familiar with this area, and our current definition has changed to “heritable changes in gene functions without changing DNA sequence" (Wu & Morris, 2001). Epigenetics includes many aspects, such as 3D genome organization, chromatin remodeling, DNA methylation, histone modification, and non-coding RNA, and so on. (Handy, Castro, & Loscalzo, 2011; Portela & Esteller, 2010).

In recent years, researchers have focused more on the 3D architecture of our genome, and they found that chromatin interactions as one type of 3D genome

organization have an important place in epigenetics. We will discuss more 3D genome organization and chromatin interactions in section 1.2. DNA methylation is the process by which DNA methyltransferases transfer a methyl group to the C-5 position of the cytosine ring of DNA and further influence the transcriptional process (Jin, Li, & Robertson, 2011). “Histone modifications” refer to modifications on the histone proteins. ~146bp long DNA sequence wraps around a core of histone proteins which contains 8 histones, and this structure is called a “nucleosome”. Nucleosomes, as an example of 3D genome organization architecture, will be introduced in section 1.2.1. Histones have five major families: H1/H5, H2A, H2B, H3, and H4. Acetylation, methylation, and phosphorylation on these histones play crucial roles in regulating gene expression and chromatin interactions (Bhasin, Reinherz, & Reche, 2006).

In this thesis, we discuss Histone 3 Lysine 27 acetylation (H3K27Ac), which is regarded as a gene activating signal and usually used for calling super-enhancers (Creyghton et al., 2010), Histone 3 Lysine 27 Tri-Methylation (H3K27Me3) which function as gene repression (Barski et al., 2007) and we used it to call super silencers (Yu Zhang, Cai, Roca, Kwoh, & Fullwood, 2021); Histone 3 Lysine 4 Tri-Methylation (H3K4Me3), as an activation signal of the gene (Koch et al., 2007), which we used for calling broad H3K4Me3 domain(Cao et al., 2017; Dahl et al., 2016) in our study.

As epigenetic is heritable (Dupont, Armant, & Brenner, 2009), compared with gene alteration which may involve changes to the sequences of germline cells and the risks of changing human genome sequences, research in epigenetics might provide more conducive treatments for many diseases, both inherited and

sporadic. Epigenetics is involved in essential mechanisms for normal development and gene expression patterns in cells, especially for specific tissues. (Sharma, Kelly, & Jones, 2010), so that epigenetic research has become a hot topic today.

In this thesis, we will mainly focus on 3D genome organization and discuss its relationship with other epigenetic such as DNA methylation and histone modifications, and how they work together to influence gene activation or repression.

### 1.1.2 Cancer and Epigenetics

As epigenetic can regulate gene expression without altering gene sequences, epigenetic is connected with good health. Aberrant epigenetic is associated with disease, and cancers are one of the most fatal of all diseases in human beings (Sharma et al., 2010).

Cancer is defined as a group of diseases that contains hundreds of types. Cancer starts from abnormal cell growth and division and further leads to invasion or metastasis (spread to other parts of the body) (NIH, 2021; WHO, 2021). According to WHO, in 2020, there were a total of 10 million deaths due to cancer (Ferlay J, 2020), making it one of the leading causes of death in the world today.

By screening multiple breast cancer, colon cancer, and other cancers patients' genome profile, researchers found that perturbed epigenetic mechanisms

can be found in a variety of cancers, especially in non-germ line cancer (Jasperson, Tuohy, Neklason, & Burt, 2010; Wood et al., 2007). Many cancers arise from gene mutations, and epigenetic alterations also can have effects similar to gene mutations. For example, mutations in the *BRCA1* gene can increase the risk of breast cancer and other cancers, while hyper-DNA methylation at this gene increases predisposition to breast and other cancers. (Q. Tang, Cheng, Cao, Surowy, & Burwinkel, 2016). A similar phenomenon has been found in colorectal cancers: by checking the DNA methylation level of the *SEPT9* gene, colorectal cancer can be detected in earlier stages (Johnson et al., 2014). Taken together, studies in epigenetics indicate that understanding epigenetic mechanisms might be useful for cancer prevention, detection, and therapy (Banno et al., 2012; Novak, 2004).

## 1.2 Chromatin Interactions

### 1.2.1 3-Dimensional Genome Architecture

Humans, as a diploid species, have a total of around 6.27~6.37 Giga base pairs of DNA sequence with a total length around 205.00~208.23 cm in one cell (Piovesan et al., 2019). How could such lengths be contained in the ~10μm diameter human nucleus (H. B. Sun, Shen, & Yokota, 2000)? This question has triggered many scientists' interest in looking into the tiny world: 3-Dimensional genome architecture.

To understand how architecture looks like and how it works, scientists have

made plenty of efforts. With the development of the microscope, researchers found that the DNA double helix was folded in some specific manner, and the architecture dynamic changes with different cell stages (Kempfer & Pombo, 2020). Later when more techniques came out, like 3D-fluorescence *in situ* hybridization (3D-FISH), Electron spectroscopy imaging, High-throughput chromosome conformation capture (Hi-C), and other chromosome conformation capture-based techniques, scientists realized that genome has a non-random and highly organized manner, and it has different function structure under different scales (Kempfer & Pombo, 2020) (**Figure 1.1**).

Inside the nucleus, between the nuclear membrane and nucleolus, all the chromosomes are folded and clustered into this space to form chromosome territories (**Figure 1.1a**). These chromosomes have two types of compartments: compartment "A" which represents the active compartment and compartment "B" which is the inactivated compartment (Fortin & Hansen, 2015; Lieberman-Aiden et al., 2009) (**Figure 1.1b**). When we go to a smaller scale, those compartments can be further divided into many domains, which are called Topologically Associating Domains (TADs). The TAD provides a region inside this domain whereby DNA sequences have more chance to interact with each other (**Figure 1.1c**). CCCTC binding factors (CTCF) and cohesins help to insulate TADs from each other (Pombo & Dillon, 2015).

TADs are introduced in more detail in section 1.2.2.1, as this thesis has a large part mainly focused on TAD research. TAD is a large-scale chromatin interaction. Inside the TAD, more detailed interactions can be found. Loops are formed by the assistance of CTCF and cohesins to regulate gene expression

(**Figure 1.1d**). In a more linear format, we can find that there are nucleosomes with histone modifications that can also help to function in gene expression regulation. Considering smaller scales, we just have the DNA sequence itself left. This complicated and highly ordered structure ensures that the tiny nucleus can contain the huge amount of information that a human needs.

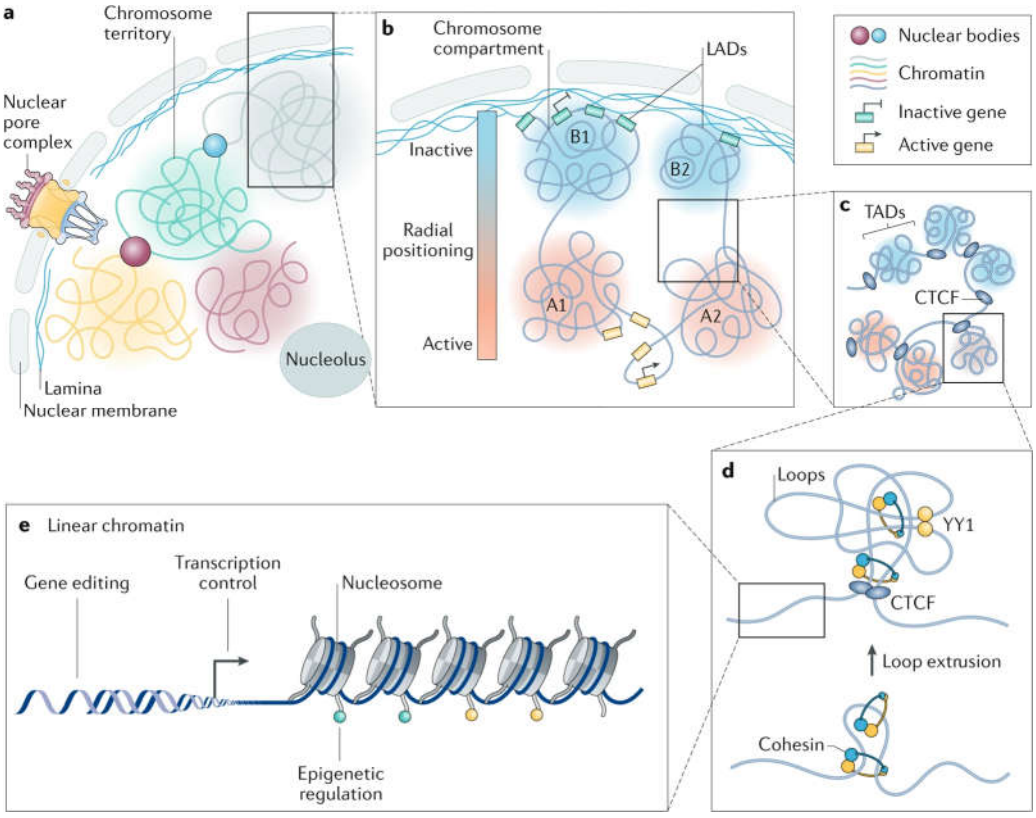



**Figure 1.1 Schematic of different scales of 3D genome organization.**

Different scales of genome organizations. From section a to e, scales decreased, and more details were shown. This figure is from (Haifeng Wang, Han, & Qi, 2021), and is reproduced with permission from Springer Nature.

### 1.2.2 Different Scales of Chromatin Interactions

Chromatin interactions start from the fundamental unit: nucleosomes, which is aproximately 146bp. In 1-100kb scales, chromatin interactions consist of mainly enhancer-promoter interactions, which contain different kinds of loops as introduced in section 1.2.1. TAD and chromosome territories are megabase size chromatin interactions. They tend to be more conserved during evolution(Rao et al., 2014). We speculate that since their sizes are large, compared with other smaller chromatin interactions, conservation would be more important because changes are likely to have a serious effect on the expression levels of many genes (**Figure 1.2**).

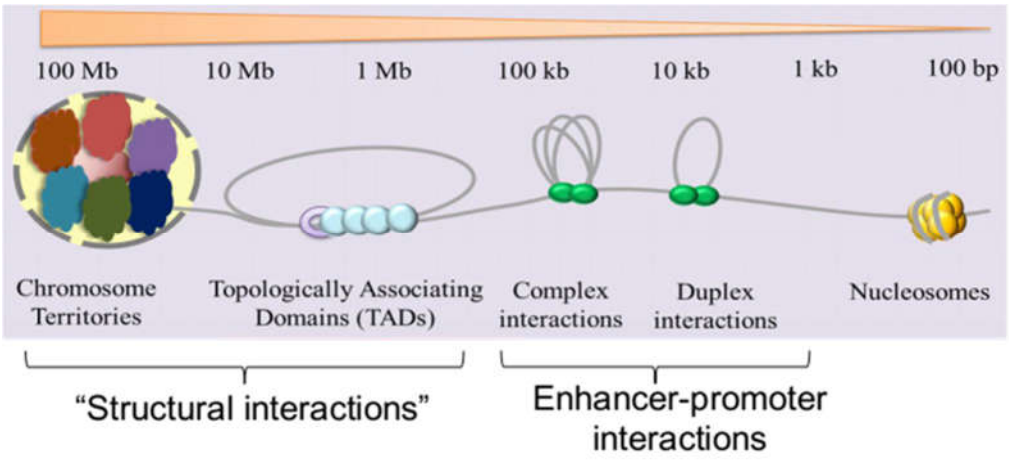



**Figure 1.2 Different scales of chromatin interactions.** From left to right is the larger scale to smaller scales of chromatin interactions. There are two types: structural interactions such as TADs and complex and duplex interactions such as enhancer-promoter interactions, enhancer-enhancer interactions, and so on. This figure is from (Babu & Fullwood, 2015) and can be reused without permission.

#### 1.2.2.1 Topologically Associating Domain (TAD)

Topologically Associating Domains (TADs) are self-interacting regions in the genome, first observed in 2009 by low-resolution Hi-C (Lieberman-Aiden et al., 2009) but did not termed as TAD, they called them some open and closed domains. In 2012, studies start to term this kind of domain as "TAD" (de Laat & Duboule, 2013; Dixon et al., 2012; Nora et al., 2012).Inside the TAD, DNA sequences have more chances to interact with each other. TADs have been discovered in many species, including Drosophila, mice, plants, fungi, and human beings (Szabo, Bantignies, & Cavalli, 2019). TAD isis considered to be conserved among different cell types and even different organisms(Dixon et al., 2015).

Once a TAD is altered, it can lead to dysregulation of gene expression of multiple genes and further lead to disease. As shown in **Figure 1.3**, gain or loss or inverse of TAD boundary might cause gene A to aberrantly interact with enhancer of gene B and influence both gene A and gene B's expression (Norton & Phillips-Cremins, 2017).

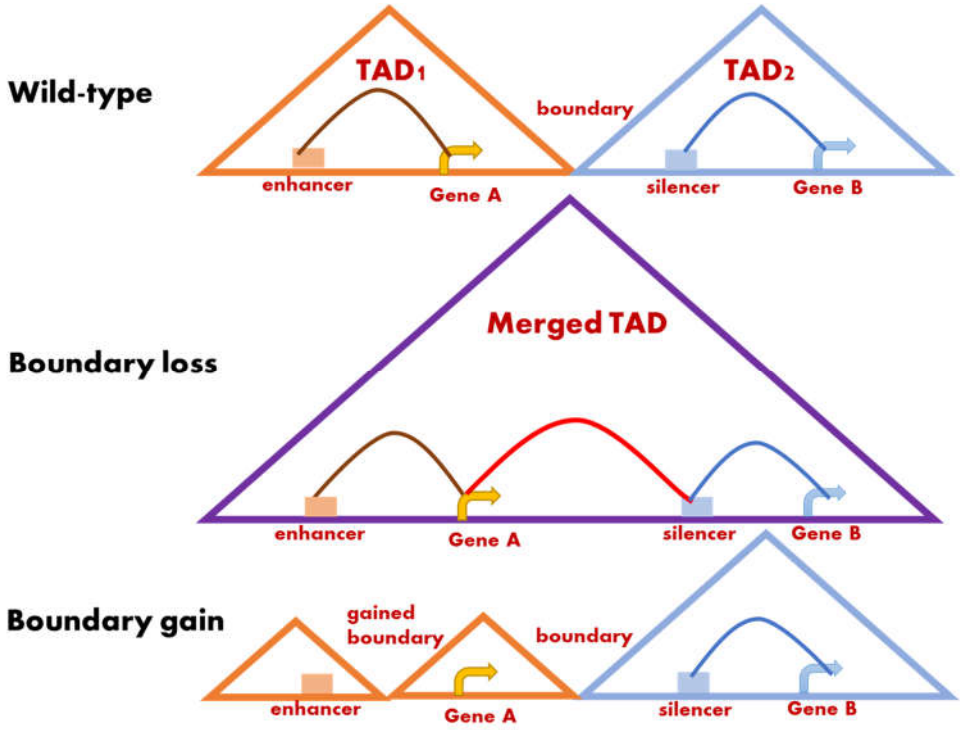



**Figure 1.3 How abnormal TAD influences gene expression.** Boundary loss and gain of TADs may lead abnormal interactions between genes and gene regulators.

### 1.2.2.2 Frequently Interacting Regions (FIRE)

Schmitt et al. first discovered and defined "Frequently Interacting Region (FIRE)" from 21 primary human tissues and cell types in 2016 (Anthony D. Schmitt, Ming Hu, Inkyung Jung, et al., 2016). They conducted Hi-C experiments in these samples and figured out that one type of chromatin interaction usually appeared in the middle of TAD with high chromatin interaction contacts. They called this type of chromatin interactions Frequently Interacting Region (FIRE)

and defined four characteristics of FIREs:

1. As a local interaction hotspot with high levels of local chromatin interactions

2. Will be depleted at TAD boundaries, usually appears in the middle of TAD

3. Associated with super-enhancers

4. Formation is partially dependent on CTCF. (Anthony D. Schmitt, Ming Hu, Inkyung Jung, et al., 2016)

They found that FIREs are tissue-specific. For instance, there was a FIRE present in two brain tissue samples around the gene *ROBO1*, which was absent in GM12878. FIREs in GM12878 are more related to the biological process of immune functions while brain-specific FIREs are more related to brain function (Anthony D. Schmitt, Ming Hu, Inkyung Jung, et al., 2016).

### 1.2.3 Techniques: 3C based "C" techniques

Many researchers have examined how chromatin is organized. In early studies, scientists used microscopes to observe the chromatin and used the fluorescence to help detect the structure of chromatin. In the 1980s, the fluorescence *in situ* hybridization (FISH) techniques came out (Langer-Safer, Levine, & Ward, 1982) and were widely used in the early study in the structur of chromatin (Garimberti & Tosi, 2010). Even though the "C" techniques are popular these days, because the "C" techniques required sequencing (except 3C),

while high throughput sequencing cost higher, FISH is still in use due to its cheap cost and easy handling as the low cost and easy handling can reduce a lot of commercial costs and labor intensity if the requirement of fineness is not high. FISH is quite suitable for a specific short region chromatin observation, but it is hard for FISH to study the whole genome chromatin structures. Along with the rapid development and spread out of high throughput sequencing techniques, the "C" techniques are taking place in studying 3D genome organizations.

The "C" methods are derived from the Chromatin Conformation Capture (3C) method, which is first developed by Job Dekker and his colleagues (Dekker, Rippe, Dekker, & Kleckner, 2002). Then Circularised Chromosome Conformation Capture (4C) was developed in 2006 by Marieke Simonis and her colleagues (Simonis et al., 2006), and further 5C (Carbon Copy Chromosome Conformation Capture) was invented in the same year of 4C by Dostie et al. (Dostie et al., 2006). Later on, in 2009, Hi-C (High-throughput Chromosome Conformation Capture) was developed to allow a whole-genome analysis of chromatin interactions (Lieberman-Aiden et al., 2009), and ChIA-PET (Chromatin Interaction Analysis by Paired-End Tag) in 2009 by Fullwood et al. (Fullwood et al., 2009), which also can explore the whole genome chromatin interactions bound by a specific factor of interest.

The "C" methods share the basic idea that chromatin interactions can be detected by cross-linking DNA and protein together, to reserve the chromatin interaction, and digest the DNA by some specific restriction enzymes, and ligate them by dilute proximity ligation. After all these procedures are done, it will lead to the formation of sequences made up of two or more genomic regions

indicating the chromatin interactions, and then we can use several PCR or sequencing methods to obtain the sequence information, which can indicate where the chromatin interaction happened (**Figure 1.4**).

The differences between 3C, 4C, and 5C were that 3C is used for confirming the expected chromatin interaction, and the specific primer needs to be designed, so 3C is “one to one”. 4C is a “one to all” chromatin interaction detection method as it involves the design of a pair of inverse primers that can amplify a circularized chromatin interaction, and the resulting PCR product can then be sequenced by next-generation sequencing. 5C relies on ligation-mediated amplification (LMA) and ligates nearby LMA so that the primer does not need to fit the known chromatin side, but the LMA instead, thus it can be used to test “many to many” interactions.

Among these techniques, Hi-C is an innovative method that can detect chromatin interaction in the whole genome, which can enable us to understand the global features of chromatin interactions. Hi-C is different from these “C”s because it does not need to design the primer, but just added biotin at the digest end, and enrich by this biotin, and then prepare the library to get the high throughput sequencing. By using paired-end sequencing, we can get sequencing pairs that can align to different locations that are not nearby, and therefore indicate that they are involved in chromatin interaction formation. Previous Hi-C experiments were usually conducted with cell lines(Burton et al., 2013), while recently scientists have used Hi-C to study cancer, for example, by examining: chromosomal rearrangement and copy number variations in anaplastic astrocytoma and glioblastomas (Harewood et al., 2017).

These methods above aim to detect chromatin interactions. However, functional chromatin interaction with specific transcription factors (TFs) and histone modifications cannot be specifically detected. Thus, the chromatin immunoprecipitation (ChIP) technique was taken into consideration. ChIP-loop, which is the method that combined ChIP and 3C together (Horike, Cai, Miyano, Cheng, & Kohwi-Shigematsu, 2005), and Chromatin Interaction Analysis by Paired-End Tag (ChIA-PET), which is similar to a combination of Hi-C and ChIP, but used ChIP but not biotin to enrich the specific signals like TF and histone marks and used the pair-end tag sequencing (Fullwood et al., 2009). Today, there are long-reads ChIA-PET and Hi-ChIP approaches that have made such analyses easier to perform (G. Li et al., 2012; Mumbach et al., 2016; Z. Tang et al., 2015).

By using these techniques above, researchers can explore chromatin interactions and interrogate the relationship between chromatin interactions and health.

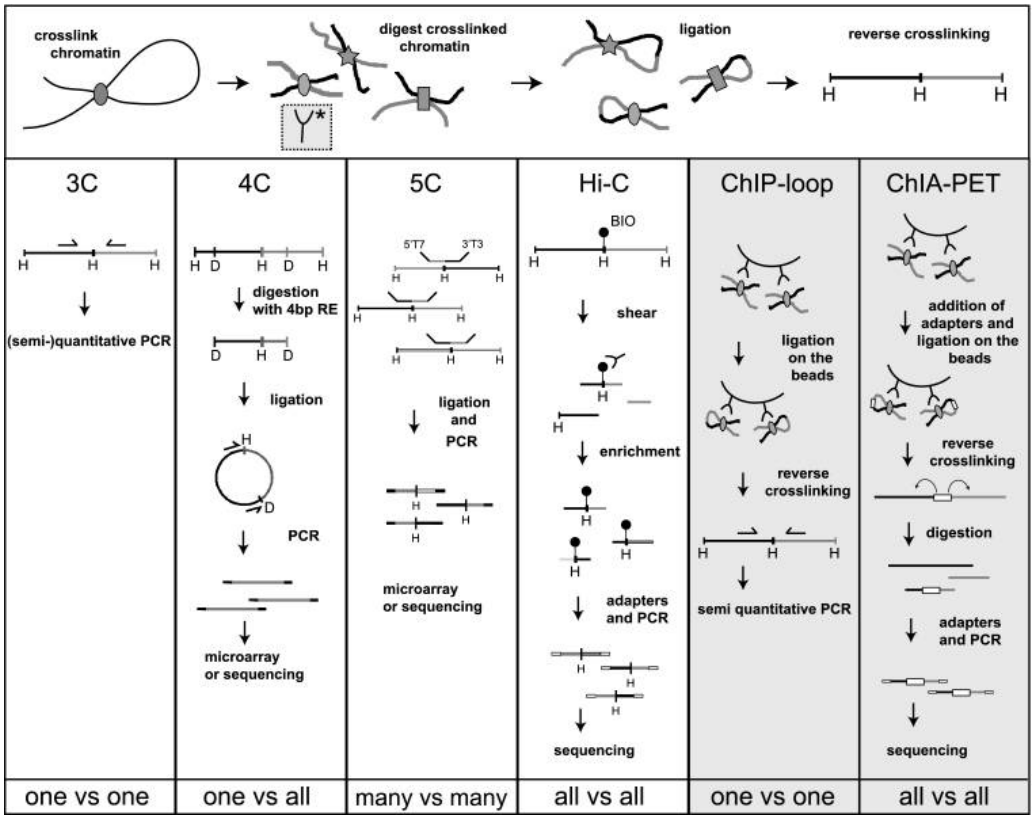



**Figure 1.4 Different Chromatin Conformation Capture Genomic Techniques.** Figure from (de Wit & de Laat, 2012) and can be reused without permission.

## 1.2.4 Algorithms for Hi-C analysis

As we introduced in section 1.2.3, Hi-C sequences multiple locations that are not near each other and then indicates whether two distant regions have chromatin interactions. By counting contact reads numbers for of every two positions, and setting each position as X and Y, we can get a 2D contact matrix. If we further analyze the matrix, we can get a profile of the whole genome chromatin interactions. That is the basic idea of Hi-C analysis. For analyzing the

matrix, the normalization method, TAD, loop, and another kind of chromatin interaction calling methods are needed.

#### 1.2.4.1 Hi-C matrix normalization methods

All experiments, including Hi-C, come with a bias. In Hi-C, there are three major factors of systematic bias: the GC content of trimmed ligation junction, distances between restriction sites, and sequencing mappability (Yaffe & Tanay, 2011). To eliminate biases, many normalization methods have been invented. These methods can be roughly divided into two groups: explicit and implicit (Lajoie, Dekker, & Kaplan, 2015). The difference between these two approaches is that explicit methods regard systematic biases as known information from the observed data, and they design a statistical model with the consideration of biases, while implicit methods believe that the biases are unknown and are cumulated in each bin, and the biases can be calculated by the sequencing average of each bin (Schmitt, Hu, & Ren, 2016).

For the explicit approach, Yaffe & Tanay (2011) is the first group to discuss the Hi-C biases and give an explicit solution: the statistic model. Following this model, HiCNorm was developed (Hu et al., 2012), with higher computational efficiency, by improving the distribution method from Bernoulli distribution to Poisson distribution or a negative binomial distribution. (Anthony D. Schmitt, Ming Hu, & Bing Ren, 2016).

As for the implicit approach, since the idea is that the biases can be counted in each bin, and each bin's coverage should be balanced so that the solution based on this idea is aiming to balance the whole matrix by bins. Thus, this approach is

also known as the “matrix balancing” approach. Several normalization methods were developed relying on this assumption. Vanilla Coverage (VC) is one of the first methods to be published which is based on the implicit approach. Here, the observed contact frequency was divided by the sum of the respective row which is the whole-genome contact frequency of locus 1, and then divided by the sum of a respective column which is the whole-genome contact frequency of locus 2 (Lieberman-Aiden et al., 2009). At first, this method was used for normalizing the inter-chromosomal matrix (Lieberman-Aiden et al., 2009), and later, this method was used for intra-chromosomal normalization (Rao et al., 2014).

Based on the VC method, Imakaev et al. designed an optimized method called Iterative Correction and Eigenvector decomposition method (ICE) (Imakaev et al., 2012). This method iterates through the VC procedure until a normalized contact frequency convergence has been found (Anthony D. Schmitt, Ming Hu, & Bing Ren, 2016). With this optimization, the computational efficiency has been improved, but the sum of row and column is not equal to one, which seems to be not reasonable as the total contact frequency for one position should be one. Almost at the same time, Knight & Ruiz developed a fast algorithm termed “KR normalization”. This algorithm is used to normalize a symmetric matrix and its sum of row and column equals one (Knight & Ruiz, 2013).

Taken together, there are two approaches to normalize the Hi-C contact matrix to eliminate the biases, and each approach is based on different assumptions as to whether biases are known or cumulative in bins. There is no known gold standard on which way is the best. Since the biases are different

according to how you perform your experiment, it is better to try both approaches and compare the two results to select an appropriate approach.

#### 1.2.4.2 TAD calling algorithms

TAD, as described in section 1.2.2.1, is a self-regulating region. Inside one TAD, the contact frequency for each element is higher than outside of this TAD. Thus, in the Hi-C contact matrix, it appears as a square block-like pattern (**Figure 1.5**).

To identify TADs, the Hidden Markov model (HMM) method was first used by Dixon et al. in 2012 (Dixon et al., 2012). HMM is a model first mentioned by Leonard E. Baum and his collaborators in the second half of the 1960s (Baum, 1972; G. R. S. Leonard E. Baum, 1968; J. A. E. Leonard E. Baum, 1967; T. P. Leonard E. Baum, 1966; T. P. Leonard E. Baum, George Soules, Norman Weiss, 1970). It is one of the statistical Markov models and is widely used in the natural sciences such as thermodynamics, chemistry, and so on. In the late 1980s, HMMs were first used in the biological analysis of sequences (Bishop & Thompson, 1986). In applying HMMs to TADs, Dixon et al. used this model to calculate a directionality index of one bin from upstream and downstream average contact frequency to find the sharp difference between the bins, and then regarded this bin as the TAD boundary (Dixon et al., 2012).

Later in 2014, Rao et al. designed a new algorithm named Arrowhead, which used a parameter termed corner score to define TAD boundary (Rao et al.,

2014). This score is to assess whether the locus is likely to be the boundary or not. Arrowhead is reported to perform well in computation efficiency, especially when working with high-resolution matrices, but Arrowhead identifies fewer TADs compared with another algorithm (Dali & Blanchette, 2017).

Further, a variety of algorithms has been published, including HiCseg, which uses a block-wise segmentation model (Lévy-Leduc, Delattre, Mary-Huard, & Robin, 2014), Armatus, the algorithm used the multiscale dynamic program to detect TAD in multiple resolutions (Filippova, Patro, Duggal, & Kingsford, 2014), TopDom which used a diamond-like window to slide along the diagonal line and seek the local minima contact frequency position to regard as boundary (Shin et al., 2016), TADtree, which is based on the empirical distributions to determine a hierarchy of nested TADs (Weinreb & Raphael, 2016), deDoc, the graphic method using graph structure entropy, which can ensure that it is an approach that detects the global optimized structure of the genome (A. Li et al., 2018), and many other algorithms.

Even though there are plenty of methods to predict the TAD profile, TAD prediction is still a not solved problem yet (Dali & Blanchette, 2017) and there are no gold standard algorithms to identify chromatin interaction from Hi-C data yet (Forcato et al., 2017). Each algorithm has its advantages and shortages. For example, TADtree runs quite slowly and cannot analyze high-resolution data, but performs well in calculating CTCF enrichment at boundaries (Dali & Blanchette, 2017). TopDom has higher robustness across different sequencing depths and resolutions while its predicted TAD numbers are lower (Dali & Blanchette, 2017). HiCSeg has higher memory consumptions in working with high-resolution

Hi-C but has higher genome coverage of TAD (Dali & Blanchette, 2017). Also, there are no clear standard and statistical definitions of TADs which can be used to guide algorithm development. TADs are highly hierarchical, and the questions of how to define TADs, sub-TADs, the typical sizes of TAD, and what should be the exact statistical characteristics to determine the boundaries of TAD, have yet to be resolved.

Here, we chose the Arrowhead as the algorithms we used in this thesis, as other algorithms such as TopDom and HiCseg do not report overlapping TADs (Shin et al., 2016) (Lévy-Leduc et al., 2014) while Arrowhead uses definitions an overlapping list to identify TAD hierarchies, which we think is more reasonable. Since Arrowhead is reported to perform well in computation efficiency, especially when working with high-resolution matrices, as our Hi-C data is under 10kb resolution which is a high resolution, Arrowhead becomes the choice.

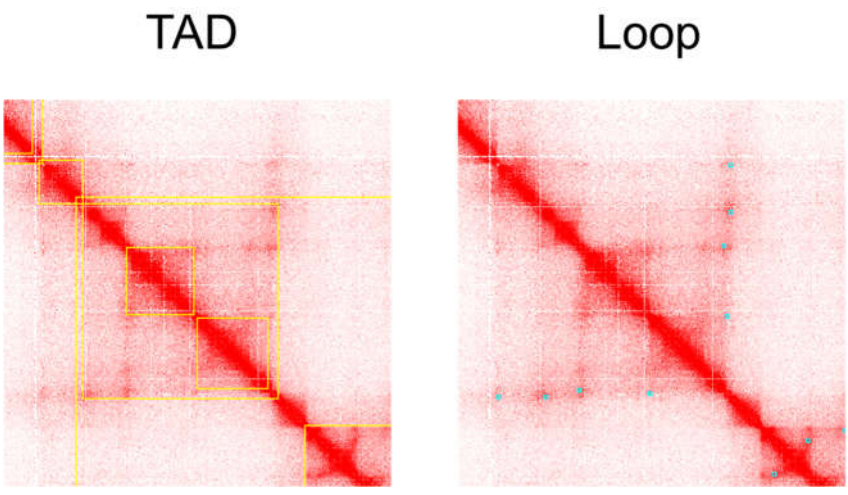



**Figure 1.5 TAD and loop patterns marked on the contact matrix heatmap.** The large yellow squares represent TADs while little blue squares represent

loops. Matrix data is adapted from GM12878 cell lines combined, primary+replicate matrix in (Rao et al., 2014), genome region: chr2:64,654,736-65,230,735, normalization method: coverage, and visualized by Juicebox (J. T. Robinson et al., 2018).

#### 1.2.4.3 Loop calling algorithms

Loop calling is also known as interaction calls because “loops” refer to chromatin interaction between two anchors. In genomic heatmaps, loops look like two small dots (**Figure 1.5**). There are fewer loop callers as compared with TAD callers. Typical methods used in the field include HOMER (Heinz et al., 2010), HiCCUPS (N. C. Durand, M. S. Shamim, et al., 2016), Fit-Hi-C (Ay, Bailey, & Noble, 2014), and diffHic (Lun & Smyth, 2015).

Homer uses the implicit normalization method to deal with matrices and uses the binomial test to detect the significant interactions by p-values, FDR, interaction read pairs, and distance (Heinz et al., 2010). HiCCUPS is a part of Juicer tools (N. C. Durand, M. S. Shamim, et al., 2016) so that its input matrix is the juicer normalized Hi-C matrix with a specific format “. hic”. It uses pixel enrichment in the nearby bottom left, donut, horizontal, vertical area to calculate the centroid of the cluster of pixels to determine the loop coordinates(Rao et al., 2014). Fit-Hi-C requires the raw matrix input and a bias file with ICE normalization, and it used a spline model with a function of distance to determine the significant chromatin interaction by FDR result. DiffHic uses a similar manner as HiCCUPS: it estimates the enrichment of neighbor areas to determine

the significant interactions, and it uses sequencing data to perform alignments.

Similar to TAD callers, there is no gold standard for loop callers either. The requirements of input files are different, and the results that are obtained from different methods from the same data are also different (Lajoie et al., 2015). In contrast to the TAD calling, loop calling does not have the problem of definition, as loops can clearly be defined: loops are significant interactions between two genomic sites. However, loop calling is even more difficult than TAD calling as it requires higher resolution because detecting two points (loop) is harder than detecting two large areas (TAD). As higher resolution data requires deeper sequencing depth, this will increase the difficulty of processing Hi-C experimental data and the hardware required to deal with larger high-throughput sequencing data, as well as the algorithms to optimize the computation efficiency.

Here in this thesis, we selected HiCCUPS for use to calculate loops, because it is reported to perform well in higher resolution (Heinz et al., 2010).

#### 1.2.4.4 The FIRE calling algorithm

As FIREs are defined by Schmitt et al. in 2016 as described in section 1.2.2.2, they also provide a package to call FIREs, FIREcaller (Crowley et al., 2021). It requires the raw matrix and then processes the data to call FIRE as shown in **Figure 1.6**. After calculation of Z-scores, they further calculated the one-sided p-values based on standard normal distribution to determine the FIREs. The bins with p-values less than 0.05 are considered to be the FIREs.

Since the FIRE is a newly defined chromatin interaction and only one algorithm has been developed for calling it, the accuracy of prediction still needs more data to be applied to test.

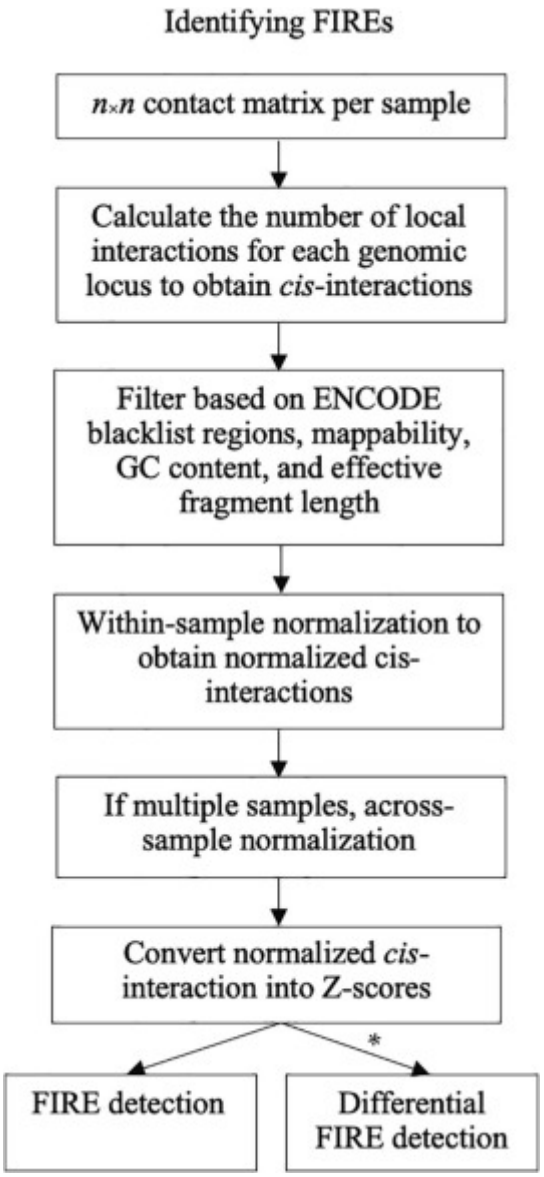



**Figure 1.6 FIREcaller flow chart.** Figure from (Crowley et al., 2021) and can be reused without permission.

### 1.2.5 Chromatin Interaction and Cancer

Recent studies suggest that aberrant chromatin interactions might cause cancer. In one study in 2016, IDH mutant glioma with a disrupted TAD caused an aberrant interaction with an enhancer and *PDGFRA* (William A. Flavahan et al., 2016). As shown in **Figure 1.7**, since the IDH is mutated, and IDH is involved in regulating DNA methylation levels in the cell, therefore in IDH1 mutated cells, the DNA methylation level will be altered. The CTCF binding site of one loop was methylated, and because CTCF cannot bind to methylated sites, the CTCF protein can no longer bind to this region (Phillips & Corces, 2009; H. Wang et al., 2012). The chromatin interaction is lost due to the absence of CTCF, allowing the oncogene inside this loop to aberrantly interact with an enhancer inside another loop, which was previously insulated by the lost loop. Ultimately, this caused improper increases in oncogene gene expression, leading to cancer. In another study, new TAD boundaries which are commonly associated with copy-number changes, are observed in the cancer genome (Achinger-Kawecka, Taberlay, & Clark, 2016).

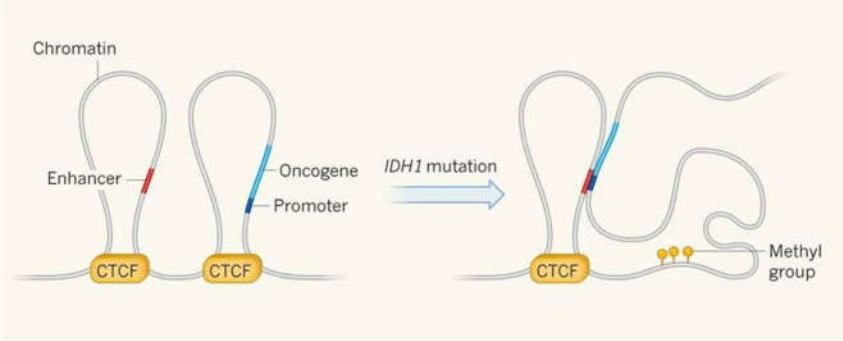


**Figure 1.7 IDH mutation caused CTCF binding site loss which further led to abnormal enhancer-promotor chromatin interactions in glioma.** This figure is from (Grimmer & Costello, 2016) and is reproduced with permission from Springer Nature.

## 1.3 Acute Myeloid Leukemia (AML)

### 1.3.1 AML: A fatal disease

Acute Myeloid Leukemia (AML) is one of the most lethal cancer types today. AML is derived from abnormal differentiation and proliferation of haematopoietic progenitor cells (including myeloid stem cells and myeloid blast) that are in the process of differentiating into myeloid cells, which further will develop abnormal red blood cells, platelets, and white blood cells (**Figure 1.8**), and causes dysregulation of the haematopoietic system. AML usually shows rapid growth of abnormal blood cells, with symptoms including tiredness, difficulty in breathing, easy bruising, and bleeding (NIH, 2020b).

Although AML has been discovered over 50 years ago, the ratio of people in remission in older patients (>60 years old) remains low (5-15%). The median survival time for older patients who cannot tolerate intensive chemotherapy is only 5-10 months (Dohner, Weisdorf, & Bloomfield, 2015). Hence, there is a critical need for targeted therapies in AML with fewer side effects that can be tolerated by the elderly.

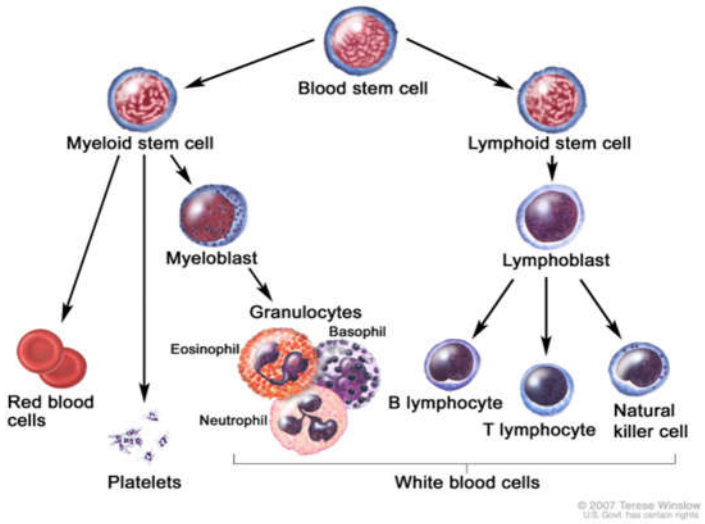



**Figure 1.8 The haematopoietic development process and cell types in different stages.** This figure is from (NIH, 2020a) and can be reused without permission.

## 1.3.2 Epigenetics and AML

The factors associated with increased risk of AML include smoking, chemical and radiation, and myelodysplastic syndrome (a group of cancer in immature blood cells), etc. (NIH, 2020b). In addition to these factors, epigenetics is also found to be important in AML. Leukemias display abnormal chromosomal characteristics (Fialkow, 1976). The Cancer Genome Atlas Research Network (TCGA) has amassed a huge amount of clinical sequencing data related to Acute Myeloid Leukemia wherein they found that 44% of mutated genes related to DNA methylation in clinical samples (Ley et al., 2013). In addition, one study indicates that DNA methylation might influence leukemic cells by altering gene

expression and phenotype rather than direct cytotoxicity to cells (Abdel-Wahab & Levine, 2013). In further research, AML was found to include many cases likely resulting from dysregulated epigenetic modulation, including DNA methylation, histone modifications, etc. (Guillamot, Cimmino, & Aifantis, 2016; Y. Sun, Chen, & Deshpande, 2018). Taken together, the evidence presented above suggests that epigenetic drugs targeted towards these epigenetic factors may be able to treat AML.

### 1.3.3 Therapeutic Ways of AML

Current treatment for AML is mainly distributed into two categories: stem cell transplantation and chemotherapy (Ferrara & Schiffer, 2013). Usually, chemotherapy is the first choice as stem cell transplantation requires careful matching between the donor and the recipient. Chemotherapy is typically given to eliminate the cancerous cells, after which stem cell transplantation is performed to restore the bone marrow ((ACS), 2020). Stem cell transplantation together with chemotherapy has a higher chance of success than chemotherapy. Chemotherapy is associated with severe side effects ((ACS), 2020), including hair loss, infection due to low levels of white blood cells, tiredness, skin and nail changes, etc. ((URMC), 2021).

As chemotherapy is intensive and has many side effects ((URMC), 2021), there is much interest in identifying additional therapies for AML. As mentioned in section 1.3.2, AML might result from aberrant epigenetic modulation, many epigenetic drugs are designed (**Table 1.1**). The targeted epigenetic factors include *DNMT*s and

*IDH*1/2, which control the DNA methylation process, and *EZH2*, MLL-complexes including *DOT1L* which influence histone modifications. Some of them are already in use and some of them are still in the trialtrial process (Wouters & Delwel, 2016). We hope more and more efficient drugs can be invented and taken into routine clinical use to help more AML patients return to heal

**Table 1.1 Current drug treatment of AML on epigenetic.** This table is adapted from (Wouters & Delwel, 2016) and is reproduced with permission from Elsevier.

| Class of epigenetic regulator | Target | Compound |
|---|---|---|
| **DNA methyltransferase** | DNMTs | Azacitidine |
| | | Decitabine |
| | | Rationally designed novel inhibitors |
| **Regulator of methylation** | IDH1, IDH2 | Inhibitors of mutant IDH1/2 |
| **Histone lysine acetyltransferase** | CREBBP (CBP) | CREBBP inhibitor |
| | EP300 (p300) | EP300 inhibitor |
| **Histone deacetylase** | HDACs | HDAC inhibitors |
| **Histone acetyl reader** | Bromodomain containing proteins (BET proteins) | BET inhibitors |
| **Histone lysine methyltransferase** | EZH2 | EZH2 inhibitors |
| | MLL-complexes | DOT1L inhibitors |
| | | Inhibitors of MLL-Menin interface |
| | | Inhibitors of MLL-LEDGF interface |
| **Histone lysine demethylase** | LSD1 | LSD1 inhibitors |
| | Jumonji family of KDMs | Small molecular inhibitors competitive for 2-oxoglutarate |
| **Histone arginine methyltransferase** | PRMTs | PRMT inhibitors |

## *1.4 DNMT3A*

### 1.4.1 *DNMT3A* Controls DNA Methylation in the Genome

The DNA Methyltransferase 3 Alpha (*DNMT3A*) is a gene that belongs to a gene family DNA methyltransferases, which is related to control of DNA methylation, both by maintaining existing DNA levels and by creating *de novo* DNA methylation (Rhee et al., 2000). As members of the *DNMT* family, *DNMT3A* and *DMNT3B* are responsible for *de novo* DNA methylation with the help of *DNMT3L* (Jia, Jurkowska, Zhang, Jeltsch, & Cheng, 2007). They are also essential for the construction of DNA methylation patterns during mammalian development (Chen, Ueda, Xie, & Li, 2002; Heyn et al., 2019; Viré et al., 2006).

As described in section 1.1.1, DNA methylation can influence gene transcriptional levels. Thus, *DNMT3A* might be crucial for gene expression maintenance. *DNMT3A* mutations may cause aberrant methylation and lead to disease.

### 1.4.2 *DNMT3A* is the Most Frequently Mutated Epigenetic Factor Gene in AML

*DNMT3A* was found to be highly mutated in AML patients (Ley et al., 2013). From **Figure 1.9**, we can conclude that *DNMT3A* is one of the most frequently mutated genes and the highest mutated epigenetic factor among AML patients.

As *DNMT3A* plays a crucial role in *de novo* DNA methylation, once it is mutated, the global DNA methylation level will be affected, which will cause cellular dysregulation. One study found that most of the genome is hypomethylated, but several repressed genes are hypermethylated upon *DNMT3A* mutation (Jeong et al., 2018). Another study found that *DNMT3A* mutations are characterized by intermediate-risk cytogenetic profiles and poor outcomes (Ley et al., 2010).

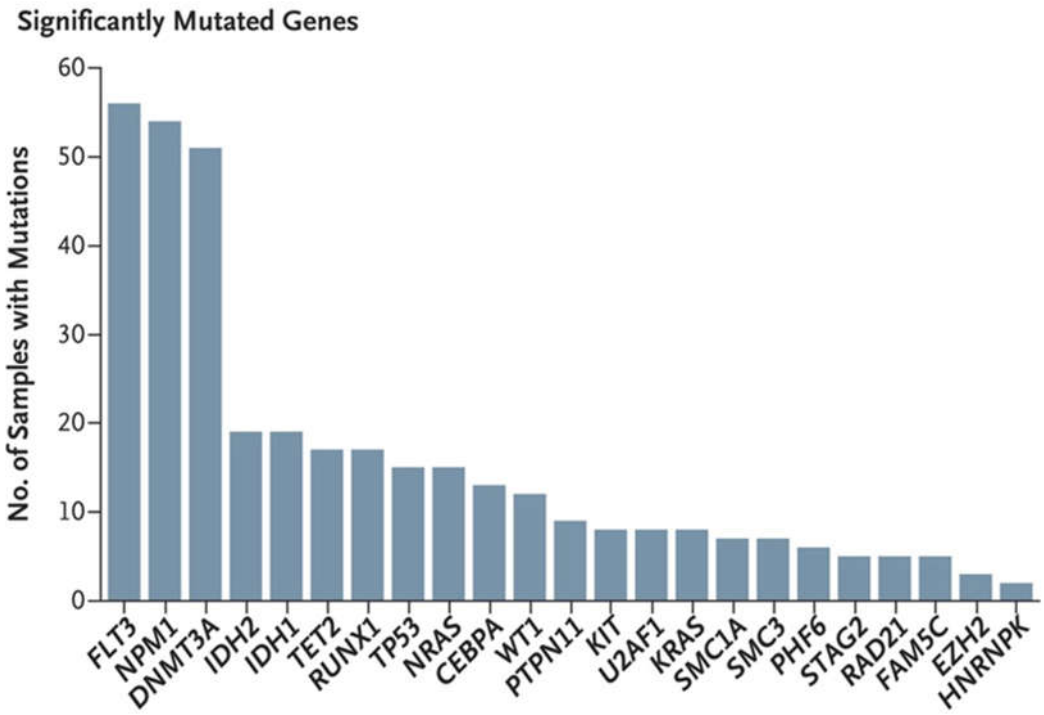



**Figure 1.9** ***DNMT3A*** **is a frequently mutated gene in AML.** The significantly mutated genes ranking by TCGA. This figure is from (Ley et al., 2013) and can be reused without permission.

### 1.4.3 *DNMT3A* Mutation Leads to Downregulation of Functional DNMT3A Protein Which Will Lead to DNA Methylation Level Change and Hematological Malignancies

The DNMT3A protein has two domains: regulatory domain and catalytic domain (**Figure 1.10**). The regulatory domain is the domain that interacts with DNMT3B and DNMT3L proteins and other regulators such as EZH2, DNA protein NP53, and so on, to regulate its function (Chaudry & Chevassut, 2017). The catalytic domain is the functional domain with the ability to methylate DNA. Inside the catalytic domain, there is a site called the R882 mutation site. This site is the R882 codon which is frequently mutated. Approximately 22% of AML and 36% cytogenetically normal AML patients carry the *DNMT3A* mutation, and of these cases, 60% have *DNMT3A* mutations at the R882 codon (Chaudry & Chevassut, 2017; Ley et al., 2010). With the R882 codon mutation, the formation of functional tetramers is impacted (Holz-Schietinger, Matje, & Reich, 2012), so the level of functional DNMT3A tetramers is reduced, which impacts the DNA methylation process, and further leads to hematological malignancies (Lin et al., 2011; O'Brien, Brewin, & Chevassut, 2014; Russler-Germain et al., 2014).

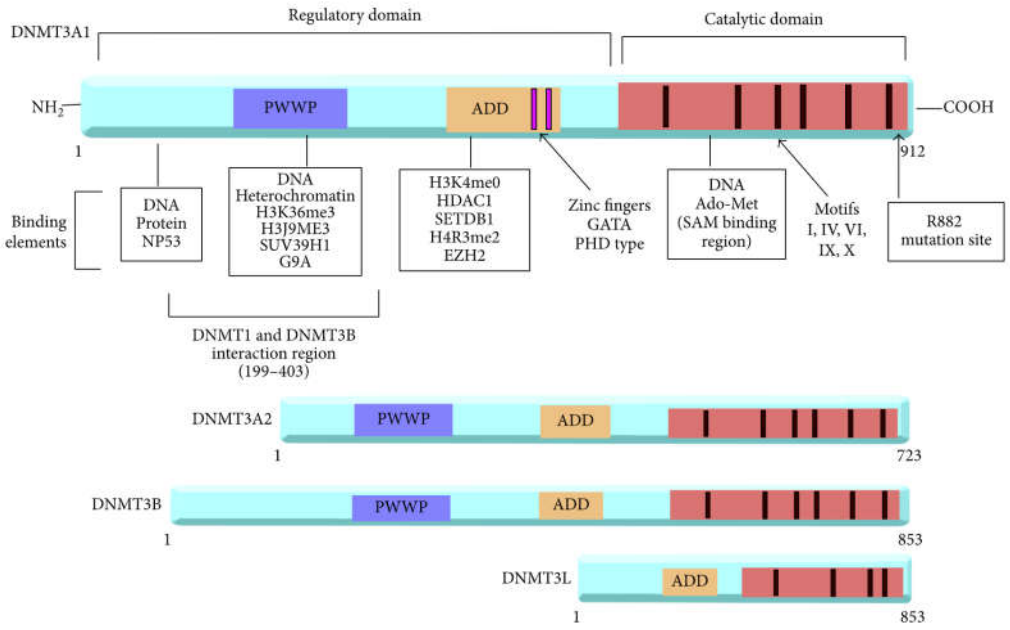



**Figure 1.10 Structure of the DNMT3A protein and its isoforms, DNMT3B and DNMT3L and their interaction regions.** This figure is from (Chaudry & Chevassut, 2017) and can be reused without permission.

# 1.5 Hypothesis and Aims

## 1.5.1 Specific Aims

In this thesis, we have the following aims:

1. To understand whether Topologically Associating Domains (TADs) and chromatin loops are dysregulated in AML compared with normal haematopoietic cells.

In this aim, we analyzed chromatin interaction information from Hi-C

analyses of clinical samples of AML compared with normal cells to see whether TADs and chromatin loops are dysregulated or not in AML.

2. To understand whether *DNMT3A* mutations lead to dysregulated TADs, and chromatin loops in AML.

As *DNMT3A* is highly mutated in AML and *DNMT3A* mutation leads to DNA methylation alteration, we inferred that *DNMT3A* might lead to TAD and chromatin loop alteration in leukemias. Thus, we designed two ways to investigate our hypothesis: first, we analyzed *DNMT3A* mutant and wild-type RNA-Seq data in the TCGA-LAML dataset (Ley et al., 2013) to see whether gene pairs correlations are changed or not which might indicate that TAD boundaries are changed. Second, we did the CRISPR knock out of *DNMT3A* in the K562 cell line to compare TAD and loop information in *DNMT3A* knock out with vector control to see whether they have some relationships or not.

### 1.5.2 Hypothesis

Corresponding to aims mentioned in 1.6.1, we have these hypotheses:

1. AMLs show many epigenetic abnormalities, which can result in altered chromatin interactions and topologically associated domains, that can lead to dysregulated transcription and cancer.

2. *DNMT3A* mutations lead to aberrant methylation in the genome, which leads to altered CTCF binding in the genome, which leads to TAD boundary

changes and chromatin loops formation, which leads to abnormal gene expression.

The reason why we have developed these hypotheses is that in previous work, Flavahan et al. 2016 showed that abnormal DNA demethylation resulting from IDH mutations in gliomas will lead to CTCF binding site loss (William A. Flavahan et al., 2016), because CTCF cannot bind to methylated DNA(Lai et al., 2010), and then leads to TADs disruption(W. A. Flavahan et al., 2016). As AML is characterized by altered epigenetic factors, we reasoned it would be likely to show dysregulated TADs and chromatin interactions compared with normal haematopoietic cells.

Next, we chose the most highly mutated epigenetic factor, *DNMT3A*, as an example, to investigate whether aberrant methylation will influence chromatin interaction and loop formation. In *DNMT3A*-mutated AML, enhancers are likely to be hypomethylated, allowing for more CTCF binding, resulting in potentially new chromatin loops, and altered Topologically Associating Domain (TAD) boundaries, while repressed genes may be hypermethylated, leading to less CTCF binding and reduced chromatin loops and TAD boundaries. These altered chromatin loops and TAD boundaries may then result in altered gene expression. Hence, we hypothesize that mutated *DNMT3A* will result in aberrant DNA methylation in AML, leading to altered chromatin interactions and TAD boundaries, leading to altered gene regulation and cancer.

The significance of our research is that if it is true that *DNMT3A* mutations in AML lead to altered chromatin loops and TAD boundaries, this may be a

vulnerability that can be exploited by epigenetic drugs, such as enhancer inhibitors such as JQ1 (Andricovich et al., 2018) that may block enhancers that make new chromatin loops or methylation inhibitors such as decitabine (Kantarjian et al., 2012) that further perturb the TAD structure of cells, potentially leading to TAD structure loss and cell death.

# 2. Materials and Methods

## 2.1 Clinical Sample Collection

### 2.1.1 Patients Sample Collection

*(Acknowledgement: My lab colleague Ms. Yufen Goh did the CD34+ selected samples collection part, and my lab colleague Ms. Winnie Fam did the total bone marrow samples collection, and my lab collaborators Prof. Wilson Wang from the National University Hospital, and Prof. Chng Wee Joo from the National University Hospital provided the clinical samples.)*

Bone marrow samples from AML patients were taken from the back of the pelvic (hip) bone. All bone marrow samples were obtained from the National University Hospital Singapore and collected according to the requirements of the Human Biomedical Research Act. Informed consent was obtained for all clinical samples used in the study. Patient clinical information see in **Table 2.1.**

## Table 2.1 Clinical information for patient samples

| | Femur47 | Femur49 | Femur50 | AML28 | AML29 | AML30 | AD796 | AD903 | AML42 | AML43 | AML44 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Gender** | Male | Female | Female | Male | Female | Female | Male | Male | Female | Male | Female |
| **Age** | 72yr | 67yr | 74yr | 52yr | 35yr | 28yr | 67yr | 47yr | 33yr | 47yr | 30yr |
| **Total MNC count** | 329 million | 567 million | 690.2 million | 56.3 million | 106 million | 98.4 million | 8 million | 38.85 million | 7 million | 4.2 million | 6 million |
| **Viability** | 88% | 93% | 88% | 91% | 72% | 93% | 24% | 36% | 75% | 90% | 96% |
| **Percentage CD34+** | 7.80% | 10.20% | 10.30% | 80.50% | 74.40% | 57.20% | n.d. | n.d. | n.d. | n.d. | n.d. |
| **Total CD34+ count** | 3.02million | 6.28 million | 6.9 million | 22 million | 22.4 million | 20.2 million | n.d. | n.d. | n.d. | n.d. | n.d. |
| **Viability** | 77% | 90% | 85% | 95% | 92% | 97% | n.d. | n.d. | n.d. | n.d. | n.d. |
| **Karyotype** | n.d. | n.d. | n.d. | Normal Karyotype | Trisomy 8 | 45,X,-X,t(8;21)(q22;q22)/46,XX | normal | 46,XY,?t(8;12;21)(q22;p13;q22),inv(9)(p11q13)[19]/46,XY,inv(9)(p11q13)[1] | 46,XX,inv(16)(p13.1q22)[20] | 49,XY,-3,+4,+5,add(5)(q11.2)x2,+8,-12,-17,idic(21)(p11.2),+der(?)t(?;3)(?;q21)ins(?;12)(?;q11q24.3),+2mar,~1dmin[19]/46,XY[1] | 46,XX,add(9)(q13)[3]/46,XX[17] |
| **FLT3** | n.d. | n.d. | n.d. | Negative | FLT3/ITD : Positive | FLT3/ITD : Positive | FLT3/ITD : Positive | FLT3/ITD : Positive | Negative | Negative | Negative |
| **NPM** | n.d. | n.d. | n.d. | Negative | Negative | Negative | Negative | Negative | Negative | Negative | Negative |
| **CEBPα** | n.d. | n.d. | n.d. | Positive | n.d. | Negative | Negative | Positive | Negative | Negative | Positive |
| **Relapse** | N.A. | N.A. | N.A. | relapse | relapse | No | No | No | No | No | No |

*"n.d." indicates "not done". "N.A." indicates "not applicable".

### 2.1.2 Sample Preparation of Mononuclear Cells (MNCs)

*(Acknowledgement: The CD34+ selected samples preparation was performed by Ms. Yufen Goh, and total bone marrow samples collection was performed by Ms. Winnie Fam, and clinical samples provided by Prof. Wilson Wang and Professor Chng Wee Joo from the National University Hospital)*

Mononuclear cells (MNCs) were isolated from AML bone marrow through a ficoll gradient (Ficoll-Paque PLUS; GE Healthcare, USA) To examine, the levels of CD34+, we" the manufacturer's instructions.

### 2.1.3 Isolation of CD34+ Haematopoietic Stem and Progenitor Cells

*(Acknowledgement: The sample preparation method was performed by Ms. Yufen Goh, and clinical samples provided by Professor Wilson Wang and Professor Chng Wee Joo from the National University Hospital)*

CD34 + cells were isolated from samples AML28, AML29, AML30, Femur47, Femur49 and Femur50. Positive selection of CD34+ cells from bone marrow samples from knee replacement operations was performed with an adapted protocol using CD34 MicroBead Kit UltraPure, human (Miltenyi Biotec, Germany). The percentage of CD34+ cells was determined by flow cytometry with a BD LSR II Flow Cytometer (BD Biosciences, Germany) at both pre-and post-isolation with cell marker (PE-conjugated anti-human CD34, clone 8G12)

after exclusion of cell debris based on scatter signals and dead cells by DAPI fluorescent stain. Data analysis was performed by FACSDiva software.

If the sample contained <20% CD34+, positive selection of CD34+ was performed according to the manufacturer's instructions using CD34 MicroBead Kit UltraPure, human.

If the sample contained 20 to 50% CD34+, positive selection of CD34+ was performed with double the volume of microbeads, relative to the sample volume. 10µl PBS, 0.5% FBS, 2mM EDTA was added to every 108 mononuclear cells. An equal volume of FcR blocking solution (Miltenyi Biotec, Germany) and twice the amount of CD34+ ultrapure beads were added to the sample.

If the sample contained >50% CD34+, a positive selection of CD34+ was performed with triple the volume of microbeads, relative to the sample volume. 10µl PBS, 0.5% FBS, 2mM EDTA is added to every 108 mononuclear cells. An equal volume of FcR blocking and thrice the amount of CD34+ ultrapure beads was added to the sample.

Magnetic separation with the autoMACS Pro Separator (Miltenyi Biotec, Germany) was carried out using the program, Posselds. CD34+ cells were collected as the positive fraction.

### 2.1.4 Flow Cytometry Analysis

*(Acknowledgement: This method was performed by Ms. Yufen Goh, and the*

*clinical samples were provided by Professor Wilson Wang and Professor Chng Wee Joo from the National University Hospital)*

Samples AML28, AML29, AML30, Femur47, Femur49, and Femur50 were processed by this step. To check the percentage of CD34+ cells in the mononuclear cells population of all clinical samples and purity of the CD34+ cells after MACs separation, 100-250k cells were subjected to flow cytometry analysis with a BD FACSAria II. To examine the levels of CD34+, we stained the cells with phycoerythrin (PE)-conjugated anti-CD34 (BD Biosciences, 348057). To further confirm if the CD34+ cells obtained were primitive haematopoietic precursors or myeloid progenitors, the cells were also co-immunostained with allophycocyanin (APC)-conjugated CD33 (BD Bioscience, 555626) and fluorescein isothiocyanate (FITC)-conjugated CD45 (BD Bioscience, 555485).

# 2.2 K562 CRISPR Knock Out

## 2.2.1 K562 *MEIS1* Region CTCF Knock Out

### 2.2.1.1 CRISPR-Cas9 Plasmid Cloning

*(Acknowledgement: This experiment was conducted by Dr. Benny Wang Zhengjie.)*

CRISPR-Cas9 excision was performed with the All-in-One vector system as described previously (Sakuma, Nishikawa, Kume, Chayama, & Yamamoto, 2014). Two sgRNA oligonucleotides (1$^{st}$ base) targeting two regions of interest

(chr2:66,802,343-66,802,362 & chr2:66,803,315-66,803,334) were designed using the Benchling web interface (Benchling [Biology Software]. (2019). Retrieved from https://benchling.com) and annealed (**Table2.2**). The annealed region was ligated into the pX330A-Cas9-2A-GFP or pX330S-Cas9-2A-GFP vector plasmid by Golden Gate Assembly(Sakuma et al., 2014). The Golden Gate Assembly was performed with the pX330A-Cas9-2A-GFP and pX330S-Cas9-2A-GFP plasmids containing each targeted cut side to yield a single fused pX330A-Cas9-2A-GFP plasmid containing two cut sides. Generated all-in-one plasmids were confirmed for the successful insertion of the sgRNAs by Sanger sequencing (1$^{st}$ base) using the CRISPR-step2-F and CRISPR–step-2-R primers (**Table 2.3**).

**Table 2.2 CRISPR-Cas9 Excision Primers (5' to 3')**

| | |
|---|---|
| **Region One** | AAGCCAAAAAACGTGCCTTG |
| **Region Two** | TGCCCCGAGAGGAAATCCAG |

**Table 2.3 CRISPR-Cas9 Sanger Sequence Primers (5' to 3')**

| | |
|---|---|
| **CRISPR-step2-F** | GCCTTTTGCTGGCCTTTTGCTC |
| **CRISPR-step2-R** | CGGGCCATTTACCGTAAGTTATGTAACG |

#### 2.2.1.2 Transfection of K562 Cells

*(Acknowledgement: This experiment was conducted by Dr. Benny Wang Zhengjie.)*

Transfection of K562 cells was performed with the Neon Transfection System (ThermoFisher). Briefly, 5µg of pX330A-Cas9-2A-GFP with two cut sites was added to 1 million K562 cells and electroporated. The K562 cells were transfected with the plasmids containing sgRNAs of the designed cut sites as well as with control plasmids without the sgRNAs in different bio replicates. Transfected cells were kept at 37ºC with 10% fetal bovine serum and 1% of penicillin-streptomycin RPMI1640 media for 48 hours before being Fluorescence-activated cell sorting (BD FACSAria Flow Cytometer) sorted for GFP fluorescing positive clones. Each positive clone was sorted into a single well of a 96 well-plate and cultured until a visible cell pellet could be observed.

#### 2.2.1.3 Genotyping of CRISPR Clones

*(Acknowledgement: This part was conducted by Dr. Benny Wang Zhengjie.)*

Cell pellets of the positive CRISPR clones were resuspended and passaged into one well of a 6 well plate containing 10% Fetal Bovine Serum and 1% of penicillin-streptomycin RPMI1640 media and further cultured for another five days. One million cells were subsequently harvested from each clone and their genomic DNA was extracted (Wizard SV Genomic DNA purification system, Promega). Genotyping of each clone was done with specific region-specific

internal and flanking primers (**Table 2.4**).

**Table 2.4 Genotyping Primers (5' to 3')**

| | |
|---|---|
| **Flanking Forward** | CTGCAATTCATCCGCTGCTC |
| **Flanking Reverse** | TCCCAGGCTCCTGTAGTCTC |
| **Internal Forward** | CGACTCGGTAGGAAACGGAG |
| **Internal Reverse** | CACACAGCAACTAACCCCGA |

#### 2.2.1.4 Growth curve assay

*(Acknowledgement: This part was conducted by Dr. Benny Wang Zhengjie.)*

10 000 cells/well were seeded in 96 well plates and measured for cell growth at 0, 24, 48, 72 hours using the CellTiterGlo assay kit (Promega, G7571). Luminescence was measured on a Tecan plate reader.

### 2.2.2 K562 *DNMT3A* Knock Out

*(Acknowledgement: This experiment was conducted by my lab collaborator Dr. Qiling Zhou from Prof. Daniel Tenen's lab in the Cancer Science Institute of Singapore. This experiment resulted in DNMT3A CRISPR knockout clone 1. In*

*addition, my lab colleague Dr. Deepak Babu repeated this experiment, to obtain DNMT3A CRISPR knockout clone 2)*

We used CRISPR/Cas 9 to target exon 7 in *DNMT3A* to partially knock out a part of DNA sequence ~100bp and checked by Sanger sequencing and protein electrophoresis to ensure *DNMT3A* has been knocked out. The guide RNA is AGCATCGGACCCCACGGGCT (5' to 3').

K562 cells were transfected with Cas9 (mCherry+) and gRNA lentivirus (GFP+) separately. After adding Dox to induce gRNA expression for 5-7 days, mCherry and GFP double-positive single cells were sorted into 96-well plates. gDNA was extracted from each cell colony for Sanger sequencing to check whether *DNMT3A* homozygous KO was successfully introduced. For the control cells, the *DNMT3A* gRNA lentivirus was replaced by empty vector lentivirus, and mCherry and GFP double positive cells were used as control cells after Sanger sequencing to confirm the genotype.

## 2.3 Hi-C Experiments and Analyses

### 2.3.1 Hi-C Libraries Preparation for CD34+ Selected AML and Femur Samples

*(Acknowledgement: This experiment was performed by my lab colleague Dr. Deepak Babu with Dovetail Biosciences).*

Sample AML28, AML29, AML30, Femur47, Femur49 were prepared using

this method. Libraries were prepared with Dovetail™ Hi-C Kit in a similar manner as described previously (Lieberman-Aiden et al., 2009). Index primer 6 (GCCAAT) and 12 (CTTGTA) included in the Hi-C kit were used. Libraries were dissolved in TE buffer. Hi-C libraries were sequenced on a high throughput Illumina sequencer HiSeq 4000.

### 2.3.2 Hi-C Libraries Preparation for Total Bone Marrow AML Samples and K562 *DNMT3A* Knock Out and Vector Control Cells

*(Acknowledgement: This experiment was performed by my lab colleague Dr Deepak Babu with Arima-Hi-C Kit)*

Hi-C analyses using the Arima Hi-C kit were performed on AML42, AML43, AML44, K562 *DNMT3A* KO, and K562 vector control (Vec_Con). Libraries were prepared by Arima-Hi-C Kit(A. D. Schmitt et al., 2016), which uses a proprietary enzyme mixture called "Arima" to digest the DNA fragments. The cut sites of this combination are G^ANTC and ^GATC. Hi-C libraries were sequenced on the high throughput Illumina sequencer NovaSeq 6000.

### 2.3.3 Hi-C Data Process

#### 2.3.3.1 Alignment and Contact Matrix Construction

Hg38 p2 is used as a genome reference, which is the same reference as the TCGA-LAML project dataset used (Ley et al., 2013), because I will use the data of the TCGA-LAML dataset in this thesis, and I want to keep all the data under the same genome reference. The TCGA-LAML dataset is a dataset of Acute Myeloid Leukemia clinical samples. I utilized Juicer (version 1.5) (Neva C Durand et al., 2016) software produced by Aiden lab from fasta format to hic format, which can be used to visualize heatmap in Juicebox (version 1.2.3) (N. C. Durand, J. T. Robinson, et al., 2016; J. T. Robinson et al., 2018). All heatmaps generated in Juicebox are under 10kb resolution and normalized by coverage. MAPQ <30 reads were filtered out for further analysis.

#### 2.3.3.2 Principal Component Analysis of AML Hi-C data

To figure out whether AML samples are distinctly different from normal samples, I performed a Principal Component Analysis (PCA) based clustering using the python package provided by scHiCluster (Jingtian Zhou et al., 2019). I chose this package is because it can calculate multiple principal component value while other packages only concern PC1 in Hi-C data as PC1 is important for A/B compartment analysis. But in this analysis, I want to see how PC1 and PC2 differentiate different samples. I first used the “dump” tool from Juicer (Version 1.5) (Neva C Durand et al., 2016) to extract 1Mb resolution sparse format

matrices (KR normalized) for each sample and each chromosome, then converted them to n*n rows of sparse format matrices and made them into inputs of the scHiCluster to get principal component matrices for each sample. I only used the first and second principal components (PC1 and PC2) to draw the clustering plots.

#### 2.3.3.3 Topologically Associating Domain and Chromatin Loop Calling

TAD calling for this thesis was processed by Arrowhead which is software compiled within Juicer tools (version 1.5) (Neva C Durand et al., 2016). 10kb resolution was used as our contact matrix can reach this resolution because I tested several samples and 10kb works the best. KR normalization (described in section 1.2.4.1) was used for all AML samples, individual Femur samples, K562 CRISPR cells, and vector control cells. Only three Femurs combined Hi-C data which is used for TCGA data analysis (see in section 2.7) used VC normalization to call TADs as KR normalization matrix is absent in some chromosomes of such a large data.

Chromatin loops were called by HiCCUPS (described in section 1.2.4.3) which is also complied in Juicer tools (version 1.5) (Neva C Durand et al., 2016). KR as the normalization method to call under a total of three resolutions: 5kb, 10kb, and 25kb, and finally merged to get a final list of loops.All the TAD and loop lists were re-organized by homemade scripts and I used the UCSC tool bedToBigBed (Kent, Zweig, Barber, Hinrichs, & Karolchik, 2010) to make the bigbed file for visualization on the UCSC genome browser (Kent et al., 2002).

### 2.3.3.4 K562 *DNMT3A* Knock Out cells TAD and Loop Comparison

TAD comparison was done by a homemade script that used the similarity ratio to detect the same or different TAD (**Figure 2.1**). Here, the total length between head and tail means the "tail" coordinate minus the "head" coordinate.

$$Similarity\ ratio = \frac{Overlap\ region\ length}{Total\ length\ between\ head\ and\ tail}$$

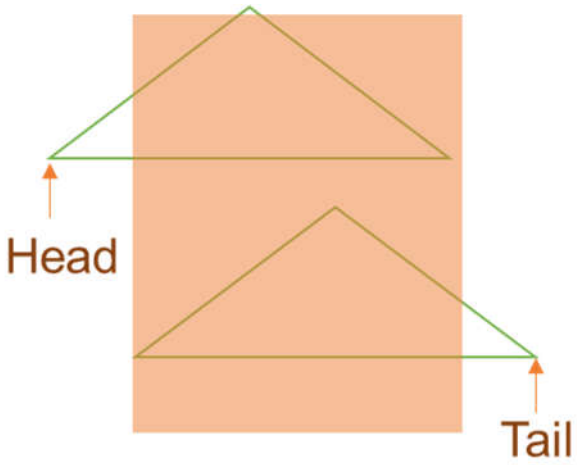



**Figure 2.1 How to calculate the similarity ratio of two overlapping TADs**

I considered two TADs with a similarity ratio of no less than 90% to be the same TADs. If multiple TADs overlapped with one TAD were found, I will choose the highest similarity ratio one. Genes with the transcription starting site inside each TAD will be regarded as common/specific TAD-associated genes according to TADs they belong to.

Loop comparison is also done by a homemade script by my colleague Mr. Bertrand Wong Jern Han, which used the idea that two loops share at least 1bp common region for every two anchors (**Figure 2.2**). Loops with only one anchor overlapped or no anchor overlapped were regarded as specific loops. Genes with the transcription starting site inside the region start from upstream 15kb to downstream 15kb of two anchors were regarded as common/specific loop associated genes according to loops they belong to.

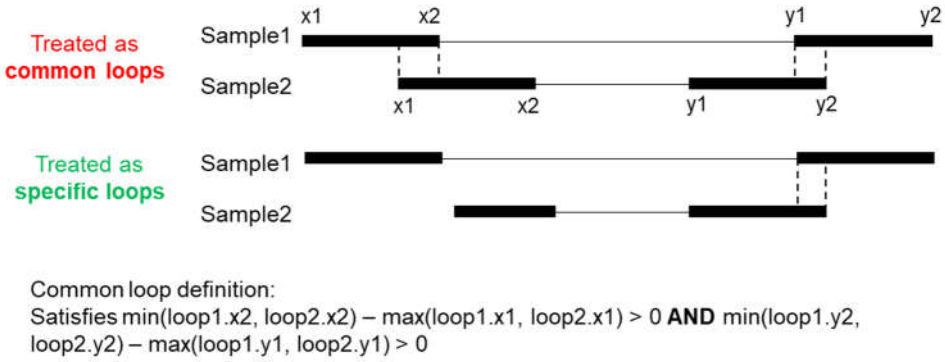



**Figure 2.2 How to define common/specific loops** *(Note: This figure is produced by Mr. Bertrand Wong Jern Han, an undergraduate of NTU SBS who did an internship in Dr. Melissa Fullwood's lab)*

#### 2.3.3.5 CD34+ AML and Femur Samples Identification of Genes and Enriched Gene Sets Associated with Common/Specific Loops

*(Acknowledgement: This part is conducted by Mr. Bertrand Wong Jern Han, an undergraduate of NTU SBS who did an internship in the lab of Dr. Melissa Fullwood)*

Loops with both anchors overlapping by at least a single base pair were regarded as being shared between the AML and femur sample. To identify loops that were specific to either AML or femur samples, these loops were excluded. Subsequently, from the remaining class-specific loops, to identify loops that were common to all three AML or all three femur samples, all samples within a class were compared in a pairwise manner and overlapping loops that occur in all pairwise comparisons were considered common. Genes within 15kb of the highlighted loops were identified using the bedtools window function. Cancer-linked information for these genes was obtained from the COSMIC Cancer Gene Census (Tate et al., 2019). to which we also added manually curated information about specific genes known to be important in AML such as *MEIS1*.

#### 2.3.3.6 Insulation Score Calling

Insulation scores of K562 *DNMT3A* knockout and vector control Hi-C data were calculated by the HiCExplorer (Version 3.4.2) (Ramírez et al., 2018; Wolff et al., 2018; Wolff et al., 2020). This software used a diamond-like a window slide along the diagonal line of the Hi-C matrix to calculate the mean z-score to get the TAD separate score, which in this thesis is called as insulation score. This score can be simply described as the likelihood of a region to be a TAD. With a more negative score, the region is more likely to be a boundary, while with a more positive score, the region is more likely to be a TAD.

#### 2.3.3.7 Copy Number Variation (CNV) and Translocation Analyses

HiNT (version 2.2.7) is used for both CNV and translocation analyses as it is a software which can do both CNV and translocation analysis with a Hi-C processed file input. HiNT requires software BICseq2, in this thesis,version v0.7.3 of BICseq2 is used. For CNV, 50kb resolution is used. Juicer produced inter_30.hic file is the input file. For translocation, $p<0.05$ is te significant translocation cutoff.

## 2.4 RNA Experiments and Analyses

### 2.4.1 Total RNA Isolation and Sequencing

*(Acknowledgement: CD34+ AML and CD34+ Femur samples RNA isolation performed by Ms. Yufen Goh; Total bone marrow RNA isolation was performed by Ms. Winnie Fam and Dr. Deepak Babu; DNMT3A knockout K562 and wild-type K562 RNA isolation performed by Dr. Deepak Babu; Sequencing runs were done by the Genome Institute of Singapore.)*

About 5 million mononuclear cells (MNCs) and 2.5 million CD34+ cells were set aside, and the DNA/RNA were extracted using AllPrep DNA/RNA/miRNA universal kit (Qiagen, Germany) according to manufacturer's instructions (Qiagen, Germany). This is a dual-extraction kit, which permits the extraction of RNA followed by DNA from the same input tissue. The extracted nucleic acids were assessed using a NanoDrop spectrophotometer (ThermoFisher Scientific, USA). The assessment of RNA integrity was done on all samples

using the RNA 6000 Nano kit on the Agilent 2100 Bioanalyzer System (Agilent, USA). Sequencing was performed through the ribosomal rRNA-depleted RNA-Seq approach (Illumina), either 2x150-151 or 2x101 bases.

### 2.4.2 Reverse Transcription (RT) and Quantitative Polymerase Chain Reaction (qPCR)

*(Acknowledgement: RT-qPCR experiments in K562 CTCF knock out samples were conducted by Dr. Benny Wang Zhengjie, DNMT3A RT-qPCR experiments in K562 DNMT3A knock out samples were conducted by Dr. Deepak Babu, RT-qPCR experiments of other genes in K562 DNMT3A knock out samples were conducted by Ms. Judy Shao, a Ph.D. student at the Cancer Science Institute who did a Ph.D. rotation in Dr. Melissa Fullwood's lab)*

Following RNA extraction (section 2.4.1), reverse transcription of RNA into cDNA was performed using the qScript cDNA Supermix (Quantabio). Quantitative polymerase chain reaction (qPCR) was performed with the GoTaq qPCR Mastermix (Promega) and QuantStudio 5 Real-Time PCR (Applied Biosystems). GAPDH was selected as the endogenous control for qPCR.

### 2.4.3 Droplet Digital Polymerase Chain Reaction (ddPCR)

*(Acknowledgement: This experiment was conducted by Dr. Benny Wang Zhengjie.)*

Following cDNA preparation (section 2.4.1 and 2.4.2), ddPCR experiments on cDNA were performed with the EvaGreen Mastermix (Biorad) and the QX200 Droplet Digital PCR system. Post analyses were done with the Quantasoft Analysis Pro Softward (Biorad).

### 2.4.4 RNA-Seq Analyses

All RNA-Seq alignments were done by STAR (v2.7.3a) (Dobin et al., 2013). The reference genome was hg38 p2. I normalized the raw data reads numbers input by their read length and unique mapping ratio. For AML28, AML29, AML30, CD34 normal sample1, and CD34 normal sample2 as they were sequenced by different lengths (either 2x150-151 or 2x101 bases), i randomly selected reads after we calculated the normalized read numbers, to make sure each sample has the same genome coverage (7× of the genome). For other samples, 2x151bp were used with similar sequencing depth, so that no reads selection was applied. UCSC genome browser tracks for RNA-Seq were also prepared by STAR (v2.7.3a) (Dobin et al., 2013). I normalized the signals by Reads of exon model per Million mapped reads (RPM) and then converted the signals to bigwig format by bedGraphToBigWig tool from UCSC (Kent et al., 2010). Transcripts Per Kilobase of exon model per Million mapped reads (TPM) were also calculated by the homemade script to indicate the exact observed RNA signals for each AML and Femur sample. The formula is below:

$$TPM = \frac{\frac{Ni}{Li} * 10^6}{sum\left(\frac{N1}{L1} + \frac{N2}{L2} + \cdots + \frac{Nn}{Ln}\right)}$$

N is the reads number mapped to the exon i, L is the length of the exon i.

Differential gene expression analysis was done by edgeR (version 4.1) (M. D. Robinson, McCarthy, & Smyth, 2010).

## 2.5 Chromatin Immunoprecipitation -Quantitative Polymerase Chain Reaction (ChIP-qPCR), ChIP-Seq Experiments and Analyses

### 2.5.1 Chromatin Immunoprecipitation -Quantitative Polymerase Chain Reaction (ChIP-qPCR) and ChIP-Seq Experiments

*(Acknowledgement: The ChIP-qPCR for AML samples were done by Dr. Benny Wang Zhengjie; ChIP-Seq for all samples were done by Dr. Deepak Babu)*

Cells were crosslinked with 1% formaldehyde (Thermo Scientific) for 15 minutes and quenched with glycine for 5 minutes at room temperature. Following this, the crosslinked cells were lysed with 1% SDS lysis buffer supplemented with protease inhibitor (Roche). Lysed cells were sonicated at 25 cycles with the Bioruptor Pico (Diagenode) and subsequently added to antibody-conjugated A/G beads (Invitrogen) and rotated overnight at $4^0$C. Anti-HOXA9 (Sigma-

HPA061982) and Anti-MEIS1 (abcam-ab19867) antibodies were used. The incubated beads were then washed in the following order: thrice with 0.1% SDS lysis buffer, once with high salt wash buffer, once with lithium chloride wash buffer, and once with Tris-EDTA buffer. The beads were eluted in ChIP elution buffer before treatment with RNase A (Qiagen) and Proteinase K (Ambion) at $37^0$C for 4 hours. The ChIP DNA was cleaned up with the QIAquick PCR purification kit (Qiagen).

The ChIP DNA was then used for performing ChIP-qPCR or ChIP-Seq. All ChIP-qPCR experiments were performed with four biological replicates of cells. In the case of ChIP-Seq, 2x150 bases were sequenced on a high throughput Illumina sequencer.

## 2.5.2 ChIP-Seq Analyses

### 2.5.2.1 Published AML Patient Samples H3K27Ac ChIP-Seq Analysis

*(Acknowledgement: This part is done by my lab colleague Ms. Ruchi Choudhary, a Ph.D. student in Nanyang Technological University School of Biological Sciences)*

Super-enhancers (SEs) were called from previously published H3K27ac ChIP-Seq data from 63 AML patient samples and 2 FACS-purified haematopoietic stem and progenitor cell (HSPC) samples (McKeown et al., 2017).

H3K27ac ChIP-seq sequences were aligned to the human genome using Bowtie2 (Langmead & Salzberg, 2012) with the default parameters. PCR duplicates were removed using 'samtools markdup'. Blacklisted regions that fall within the ENCODE consensus were removed using 'bedtools intersect'. After sorting and indexing sequences with 'samtools', narrow peaks were called using MACS2 (version 2.1.2) (Yong Zhang et al., 2008). Enhancer peaks within a 4 kb distance were stitched together and identified as SEs based on the ChIP-seq signal using a custom script similar to the ROSE package as previously described (Cao et al., 2017). The alignment was performed with genome reference hg19 and then the super-enhancer bed tracks were lifted over from hg19 to hg38 by the UCSC LiftOver tool (Navarro Gonzalez et al., 2021).

#### 2.5.2.2 THP-1 and K562 H3K27Ac ChIP-Seq Analysis

THP-1 single-end ChIP-Seq data was downloaded from (Mohaghegh et al., 2019). I chose SRR8329547 and SRR8329548 as two biological replicates for H3K27Ac ChIP-Seq data, and SRR8329549 and SRR8329550 as two biological replicates for total input background ChIP-Seq data.

K562 single-end ChIP-Seq data is from the project, Histone Modifications by ChIP-seq from ENCODE/Broad Institute (Consortium, 2012). I chose SRR227385 and SRR227386 as two biological replicates for H3K27Ac ChIP-Seq data, and SRR227650 for total input background ChIP-Seq data.

The basic analysis pipeline followed the pipeline described in section

2.5.2.3.

### 2.5.2.3 AML Samples and K562 *DNMT3A* Knock Out and Vector Control Cells H3K27Ac ChIP-Seq Analysis

First, I aligned all sequencing data with Bowtie2 (Version 2.2.5) (Langmead & Salzberg, 2012) with default settings by genome reference hg38 p2, then narrow peaks were called by MACS2 (version 2.2.7.1) (Yong Zhang et al., 2008) using the "-q 0.05 --keep-dup auto" settings to filter the q value less than 0.05 and remove duplicates. Super-enhancers and enhancers were called by a custom script similar to the ROSE package as previously described (Cao et al., 2017) by using a 4kb stitch distance.

For K562 *DNMT3A* Knock Out and Vector Control Cells H3K27Ac ChIP-Seq only, I called the peaks, enhancers, and super-enhancers comparison between KO and Vec_Con following the method which is described in detail in section 2.5.2.5.

### 2.5.2.4 K562 *DNMT3A* Knock Out and Vector Control Cells CTCF, H3K27Me3, and H3K4Me3 ChIP-Seq Analysis

First, I aligned all sequencing data with Bowtie2 (Version 2.2.5) (Langmead & Salzberg, 2012) with default settings by genome reference hg38 p2.

For CTCF ChIP-Seq, narrow peaks were called by MACS2 (version

2.2.7.1) (Yong Zhang et al., 2008) using the "-q 0.05 --keep-dup auto" settings to filter the q value less than 0.05 and remove duplicates. Then I just compare the CTCF peaks between KO and Vec_Con which are described in detail in section 2.5.2.5.

For H3K27Me3 ChIP-Seq, broad peaks were called by MACS2 (version 2.2.7.1) (Yong Zhang et al., 2008) using the "-q 0.05 --broad-cutoff 0.05 --keep-dup auto" settings to filter the q value less than 0.05 and remove duplicates. "BroadPeak" file is used for calling silencers and H3K27me3-rich regions (MRRs), and "gappedPeak" was used as the peaks list. Silencers and H3K27me3-rich regions (MRRs) were called in a similar manner of enhancers and super-enhancers called by a custom script similar to the ROSE package but which used a 4kb stitch distance, as previously described (Cao et al., 2017). This follows the method and ideas previously described in (Y. Zhang et al., 2021). Peaks, silencers, super-silencers were compared between KO and Vec_Con following the method described in detail in section 2.5.2.5.

For H3K4Me3 ChIP-Seq, broad peaks were called by MACS2 (version 2.2.7.1) (Yong Zhang et al., 2008) using the "-q 0.05 --broad-cutoff 0.05 --keep-dup auto" settings to filter the q value less than 0.05 and remove duplicates. BroadPeak is used for calling broad H3K4Me3 domains, and gappedPeak was used as the peaks list. Broad H3K4Me3 domains were selected by identifying the top 5% by the ranking of peak sizes as previously described (Cao et al., 2017; Dahl et al., 2016). Peaks and broad H3K4Me3 domains were compared between KO and Vec_Con following the method described in detail in section 2.5.2.5.

#### 2.5.2.5 Peaks, Enhancers, Super-Enhancers, Silencers, Super-Silencers and Broad H3K4Me3 Domains Comparison Between K562 *DNMT3A* Knock Out and Vector Control Cells.

These comparisons were all done by BEDtools (version 2.29.2) (Quinlan & Hall, 2010). As we have two replicates for both KO and Vec_Con, I first used the "merge" tool of BEDtools to merge the two replicates' lists into one, then I compared KO and Vec_Con lists by "intersect" tool of BEDtools, to identify the regions with at least 1bp overlap which were regarded as "common" regions. The regions without any overlap were regarded as "specific " regions.

## 2.6 Circular Chromosome Conformation Capture (4C) Experiments and Analyses

*(Acknowledgement: 4C-Seq experiments were conducted by Dr. Benny Wang Zhengjie, alignment, and r3cseq analysis was done by Dr. Benny Wang Zhengjie on CSI NGS Portal set up by Dr. Omer An and Dr. Henry Yang at the Cancer Science Institute (An et al., 2020))*

4C-seq was performed as previously described with some modifications (Splinter, de Wit, van de Werken, Klous, & de Laat, 2012). In brief, $4 \times 10^7$ cells were harvested and crosslinked with 1% formaldehyde for 10 min at room temperature with rotation. The crosslinking was quenched by glycine for 5 min at room temperature with rotation. Following SDS and Triton X-100

permeabilization, nuclei were digested with HindIII-HF (NEB) overnight. Following proximity ligation, reverse cross-linking, and DNA purification, the circular DNA was digested with DpnII (NEB) 37 °C overnight and circularized.

The 4C-seq library was generated by performing nested inverse PCR using Phusion DNA polymerase (Thermo Scientific) with the primers. 10% of the 1st PCR product was used for the 2nd PCR. The 4C-seq library was purified by 4–20% gradient TBE PAGE gel (ThermoFisher Scientific) and the smear band regions including the expected sizes were excised. The library was recovered by incubating the crushed gel slice with 200 uL TE buffer overnight at 37 °C and the DNA in the supernatant was ethanol precipitated in presence of GlycoBlue (ThermoFisher Scientific). The multiplex 4C-seq library was pooled in equal molar ratio and sequenced on MiSeq (Illumina) with 1X150 bp. 500,000–1,000,000 reads were produced for each library. BWA-Mem (0.4.17-r1188) was used to map to the human genome hg19. The mapped 4C-seq data was analyzed by r3CSeq (1.30.0). All the resulting bed tracks for 4C were first analyzed with human genome reference hg19 and subsequently, we used the UCSC LifeOver tool (Navarro Gonzalez et al., 2021) to lift over the data hg38.

Further, the file after being lifted over was converted to a biginteract format by the bedToBigBed tool from UCSC (Kent et al., 2010).

## 2.7 Gene Correlation Analysis for TCGA-LAML Data

### 2.7.1 Assignment of Genes into Different Pairs

Our computational pipelines followed Flavahan, W. A., et al., 2016 with several modifications (William A. Flavahan et al., 2016). As shown in the flow chart (**Figure 2.4**), first, I collected a gene list from hg38 p2 annotation files, and only chose the annotated element which marked as "gene" in the annotation file. Flavahan, W. A., et al., 2016 used GM12878 and IMR90 TAD lists which from (Rao et al., 2014), while I used our Femur47, Femur 49, and Femur50 combined Hi-C matrix to call a haematopoietic stem cell-specific TAD list as the reference. The detailed process can be found in section 2.3.3.3. Next, I began to assign these genes to their corresponding TAD by checking whether the transcription starting site is located inside the TAD or not. Genes that cannot be assigned to a domain and genes that can be assigned to multiple domains, but which could not be unambiguously assigned to an inner-most domain were discarded before I divided genes into pairs (**Figure 2.3 & Figure 2.4**).

After assigning genes into their respective TADs, I divided them into pairs by distance. For example, using 500 kb as the cut-off for the distance criteria, if the distance between two genes, we treated the genes as follows: (1)If these two genes were in the same domain, then this gene pair was be regarded as a "same domain pair". (2)If these two genes belonged to different domains, then the pair was defined as a "cross boundary pair". I used two different distance criteria in this thesis: 1000kb for detecting dysregulated boundaries in *DNMT3A* mutant AML samples, and 500kb for the same domain and cross boundary correlation

comparison analysis (**Figure 4.1**), which is similar to what Flavahan, W. A., et al., 2016 have done.

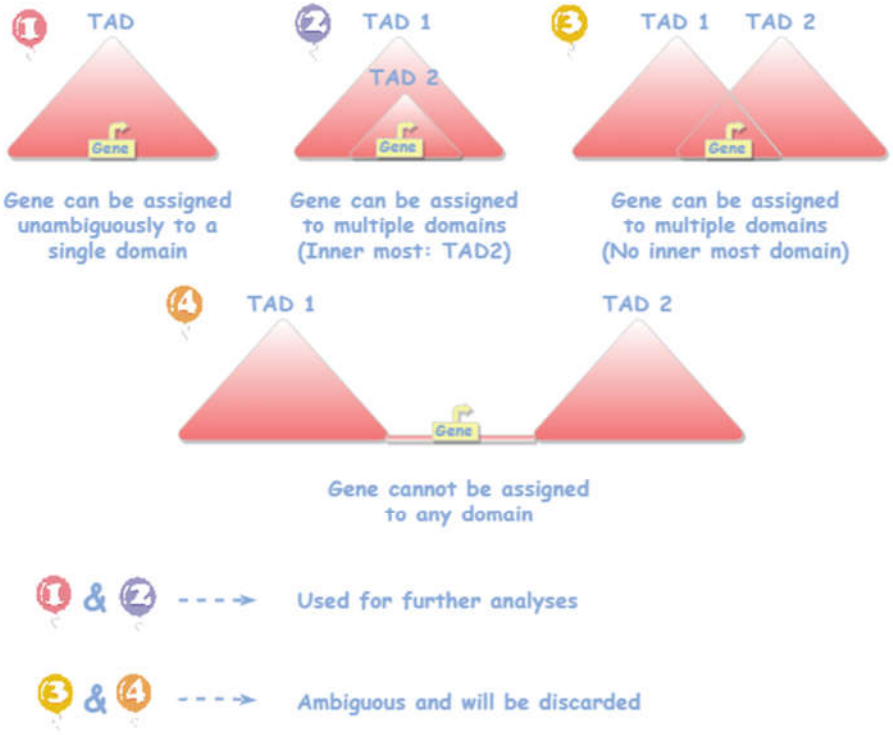



**Figure 2.3 How to assign a gene into a domain**

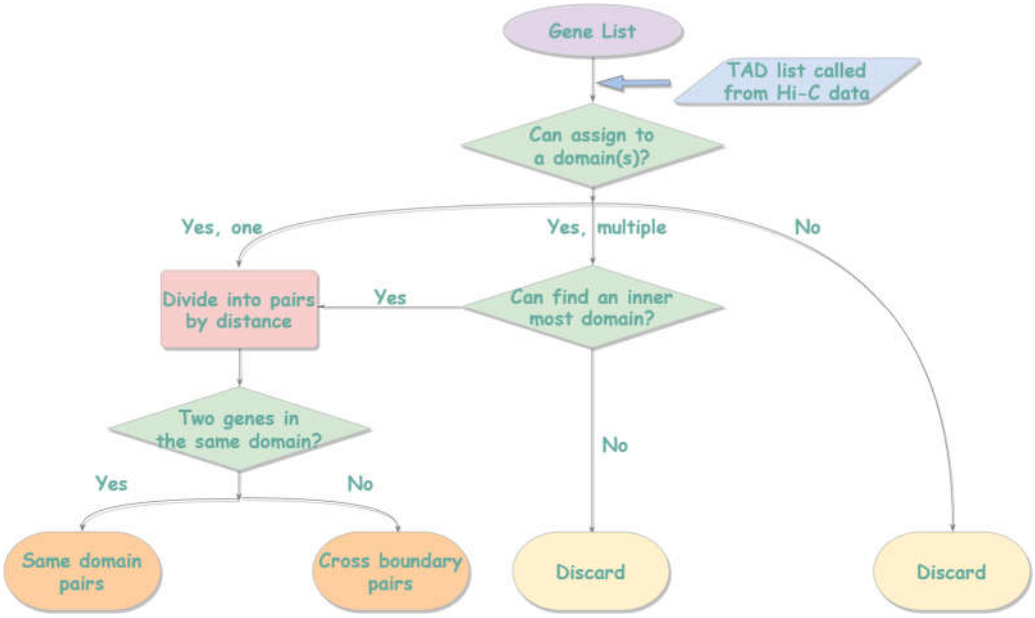



**Figure 2.4 Flow chart for assignment of genes into different pairs**

## 2.7.2 Gene Correlation Analysis and Dysregulated Boundaries Identification.

TCGA AML project TCGA-LAML RNA-Seq gene expression files (Tier 3, htseq FPKM (Fragments Per Kilobase of exon model per Million mapped fragments) files) were downloaded and used (total 151 clinical samples, 124 *DNMT3A* wild type cases, and 27 *DNMT3A* mutant cases (Ley et al., 2013), hg38p2 as a default reference genome) to calculate gene correlations. FPKM was then converted to TPM by a homemade script following the formula:

$$TPM = e^{\log(FPKM) - \log(sum(FPKM)) + \log(1e6)}$$

First, 124 *DNMT3A* wild-type cases were used for comparing same domain

pairs and cross boundary pairs correlation changes against increases in distance between gene pairs. Gene pairs under 500kb distance of transcription starting site were used to calculate the correlation. Gene correlations were calculated by Pearson correlation using cor() function in Python (version 3.6.4), I generated a line plot illustrating correlation trend differences between "same domain pairs" and "cross boundary pairs" against increasing distance using R (version 3.3.1). Lines were smoothed by weighted linear least squares (LOESS, span=0.1).

Next, I calculated the delta correlation using the 500kb as the distance criteria. I chose 500 kb instead of 180kb that Flavahan, W. A., et al., 2016 used as 180kb is too small for cross boundary pairs (Very few cross boundary pairs were found in our results). Gene correlations were calculated by Pearson correlation using the cor() function in Python (version 3.6.4). Delta correlation equals correlations in *DNMT3A* mutant samples minus *DNMT3A* wildtype samples. Significances were calculated using fisher-z-transformation:

$$(1)\ z = \frac{1}{2}\ln(\frac{1+\rho}{1-\rho})$$

$$(2)\ Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

$$(3)\ P = 2(1 - pnorm(Z))$$

ρ in formula (1) is the gene pair correlations. After calculation each z value using formula (1) for *DNMT3A* mutant and *DNMT3A* wildtype, calculate Z using formula (2) as a variance of two z value. Then using pnorm functions in formula (3) is the normal distribution function in Python(version 3.6.4) to get the

significance P. (pnorm might be different across different programming languages).Volcano plots were generated by script written in R(version 3.3.1)

Dysregulated boundaries detection used 1000kb as the gene pairs distance and follow the flow chart in **Figure 2.5**. Femur Hi-C combined matrices was used to call TADs, and the TAD list was used to define the same domain and cross boundary gene pairs. For two TADs, the boundary between them was regarded as dysregulated if at least 1 same domain gene pair correlation is decreased and at least 1 cross boundary gene pair correlation is increased.

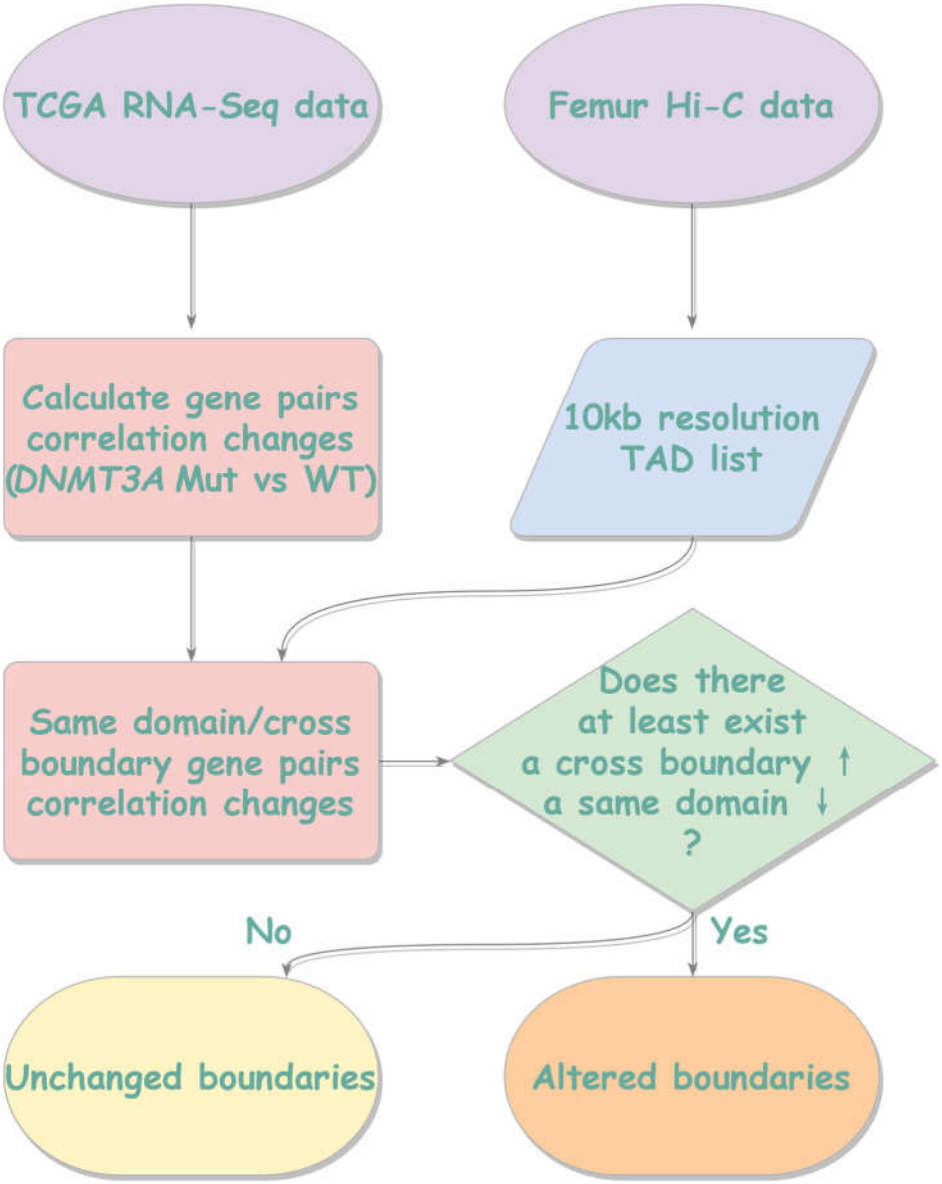



**Figure 2.5 The flow chart of how to detect the altered boundaries by integrated analysis of gene correlation calculation and Hi-C analysis of Femur samples**

## 2.8 Data Availability

All the Chapter 3 data can be viewed in GEO accession GSE149381 in referee private access. To review GEO accession GSE149381: Go to https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE149381 Enter token kzexsyyuljyhvuv into the box.

All Chapter 3 UCSC tracks can be viewed in:

http://genome.ucsc.edu/s/lincy/AML

All Chapter 4 UCSC tracks can be viewed in:

http://genome.ucsc.edu/s/lincy/DNMT3A

All homemade scripts can be viewed in:

https://github.com/lingshi951129/thesis-code

# 3. The Three-Dimensional Chromatin Interaction Landscapes of Acute Myeloid Leukemia are Altered Compared with Normal Haematopoietic Stem Cells

AML is derived from abnormal differentiation and proliferation of haematopoietic progenitor cells, which are marked by CD34 protein (Bonnet & Dick, 1997; Shlush et al., 2014), along with several aberrant activations and up-regulation of some fusion oncogenic proteins (Kundu et al., 2002; Matsuo et al., 2018; Y. Wang, Wu, Liu, & Jin, 2017). As AML shows dysregulated epigenetic (discussed in section 1.3.2), we asked how chromatin interactions, a form of epigenetic, are altered in AML. Altered 3D genome architecture has been observed in Acute Lymphocytic Leukemia and leukemia cell lines in previous studies (Kloetgen et al., 2020; Y. Li et al., 2018; Yang et al., 2019). However, how chromatin interactions differ between AML and normal haematopoietic stem cells are poorly understood.

Here, we performed Hi-C analyses in AML and normal haematopoietic clinical samples to figure out how chromatin interaction impacts AML. To understand the potential impact of dysregulated chromatin interactions on transcription, we performed integrated RNA-Seq in several AML samples.

Furthermore, for several AML samples, we also obtained H3K27ac ChIP-Seq to characterize super-enhancers in the AML samples.

## 3.1 Chromatin Interaction Alterations were Observed in CD34+ Acute Myeloid Leukemia Samples Compared with CD34+ Normal Haematopoietic Clinical Samples at Key Oncogenes.

First, with the help of my lab collaborator Prof. Wilson Wang, and Prof. Chng Wee Joo from National University Hospital who provided the clinical samples, my colleague Ms. Yufen Goh collected and prepared the CD34+ sorted AML and CD34+ sorted normal femur samples for further Hi-C experiments (CD34+ AML samples: AML28, AML29, and AML30; CD34+ normal femur samples: Femur47, Femur49, and Femur50). In this part, I did all the bioinformatics analysis for RNA-Seq and Hi-C except loop comparison analysis.

The reason why we sorted the clinical samples with CD34+ is due to the heterogeneity of AML (Horibata et al., 2019). As AML is a disease derived from abnormal differentiation and proliferation of haematopoietic progenitor cells (which are marked by CD34+), isolation of CD34+ AML cells leads to the enrichment of a more homogeneous group of primitive AML cells (Bonnet & Dick, 1997; Shlush et al., 2014). My colleague Dr. Deepak Babu performed the Hi-C experiments with Dovetail Biosciences. Billions of sequencing reads were applied, and 347-596 million Hi-C contact reads were used for further analysis.

927-1,682 of TADs and 4,733-24,294 loops were called from the Hi-C data (**Table 3.1**).

**Table 3.1 Hi-C statistics for CD34+ sorted AML and Femur clinical samples**

| | Total Sequenced Reads | Hi-C Contacts | #TAD | #loop |
|---|---|---|---|---|
| **AML28** | 1,290,109,443 | 535,816,129(41.53%) | 1,682 | 24,394 |
| **AML29** | 1,245,251,169 | 402,803,175(32.35%) | 1,108 | 19,541 |
| **AML30** | 1,387,562,168 | 347,777,411(25.06%) | 1,296 | 10,733 |
| **Femur47** | 1,337,973,222 | 596,400,948(44.57%) | 1,153 | 4,733 |
| **Femur49** | 1,156,332,234 | 396,674,984(34.30%) | 927 | 13,107 |
| **Femur50** | 1,165,296,399 | 481,662,160(41.33%) | 1,234 | 13,795 |

Then, Hi-C clustering analysis was applied in both CD34+ AML samples and CD34+ normal femur samples. By using the Hi-C contact matrix under 1Mb resolution, we calculated the principal component 1 and principal component 2 and drew the PCA plot. We can see that AML and Femur samples were separately clustered (**Figure 3.1A**), which indicates that CD34+ AML and CD34+ normal femur clinical samples indeed have variations in chromatin interactions.

In the next step, Mr. Bertrand Wong Jern Han, an undergraduate of NTU SBS who did an internship in Dr. Melissa Fullwood's lab compared the loop lists from my analysis between CD34+ AML samples and CD34+ normal femur samples. Interestingly, he found most loops in femur samples tend to be common, but AML tends to have more specific loops (**Figure 4.1B**). When he overlapped these AML associated loops with the oncogene list provided by the COSMIC

Cancer Gene Census (Tate et al., 2019), some interesting oncogenes were found, including *RAD21* which was found to be significantly mutated in AML (**Figure 1.9**) (Ley et al., 2013), *RARA* which was involved in *PML-RARA* fusion that commonly found in AML (Ley et al., 2013), MYC which was found to influence chromatin interactions (Kieffer-Kwon et al., 2017), and *MEIS1* and *HOXA9* which were found to be overexpressed in over half of AML cases (Andreeff et al., 2008; C. Collins et al., 2014; C. T. Collins & Hess, 2016; Gao, Sun, Liu, Zhang, & Ma, 2016) and presented poor prognosis (C. T. Collins & Hess, 2016; Mohr et al., 2017; Yuqing Sun et al., 2018).

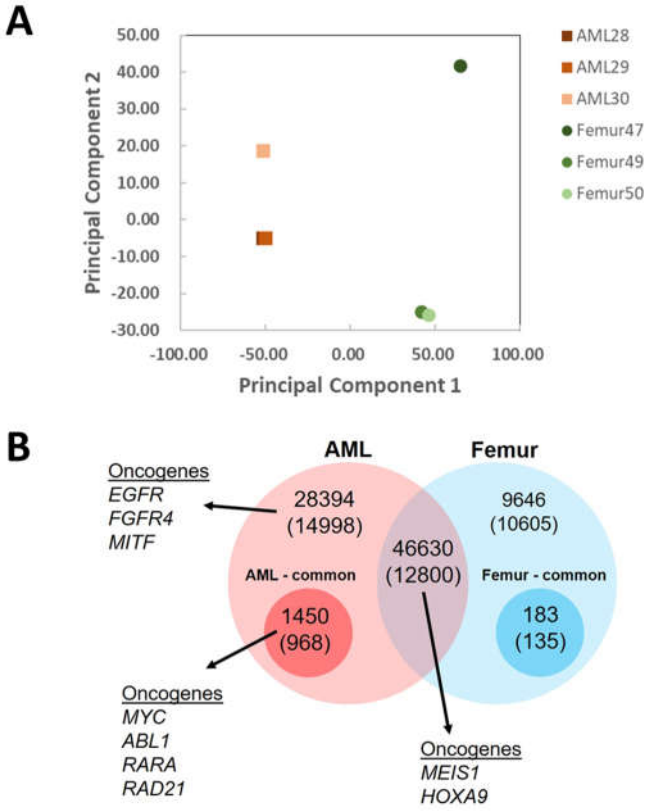



**Figure 3.1 Principal Component Analysis (PCA) and loop comparisons**

**results indicate chromatin interaction alterations in CD34+ sorted Acute Myeloid Leukemia clinical samples compared with CD34+ sorted normal haematopoietic stem cells**. **A.** Principal Component Analysis of CD34+ sorted AML and CD34+ Femur clinical samples. (Using Hi-C contact matrix of all chromosomes under 1Mb resolution, KR normalized) **B.** Loop comparison analysis revealed that oncogenes are associated with chromatin interactions in AML, numbers above brakes are loop numbers that are specific or common, and numbers inside brackets are associated gene numbers. Typical oncogenes found in these associated genes were marked out. *(Note: My lab colleague Ms. Yufen Goh did the sample collection and preparation part, and my lab collaborators Prof. Wilson Wang, and Prof. Chng Wee Joo from National University Hospital provided the clinical samples. The Hi-C experiment was conducted by my colleague Dr. Deepak Babu. Analysis and figure in part B were produced by Mr. Bertrand Wong Jern Han, an undergraduate of NTU SBS who did an internship in Dr. Melissa Fullwood's lab.)*

## 3.2 Dysregulation of a Frequently Interacting Region (FIRE) in *MEIS1* region was Heterogeneously Present in CD34+ AML Clinical Samples

We then explored these oncogenes deeper, especially the genes which were reported to be related to AML. In these genes, one gene termed *MEIS1*, in full, Myeloid Ecotropic Viral Integration Site 1 Homolog, is a gene of the Homeobox family, which is crucial for normal development. When we looked at the *MEIS1*

region, we found an interesting chromatin interaction that was heterogeneously present in CD34+ AML samples, but always present in the CD34+ normal femur samples (**Figure 3.2A**). Zoomed in heatmaps in this chromatin interaction region showed that AML28 and AML30 lost this chromatin interaction and AML29 retained the chromatin interaction (**Figure 3.2B**). My colleague Dr. Benny Wang Zhengjie then conducted the *MEIS1* ddPCR in AML28, AML29, and AML30, and compared these results with ddPCR results in several femur samples. Interestingly, AML28 and AML30 which lost this chromatin interaction showed the absence of *MEIS1* gene expression. By contrast, AML29 showed an extremely high MEIS1 copy per μL (over 400) compared with normal femurs which showed *MEIS1* expression in 5 of 6 patients, but at a much lower level - less than 150 (**Figure 4.5 C & D**). RNA-Seq for AML28, AML29 and AML30 (the experiment was conducted by Dr. Deepak Babu, and I analyzed the data) also presented similar results (**Figure 4.5E**).

As this chromatin interaction fulfilled all the features of Frequently Interacting Region (FIRE) (See in section 1.2.2.2) (present a local interaction hotspot with high levels of local chromatin interactions; in the middle of TAD; associated with super-enhancers (will confirm later); formation is partially dependent on CTCF (will confirm later)), we will call this chromatin interaction "*MEIS1* FIRE" in this thesis.

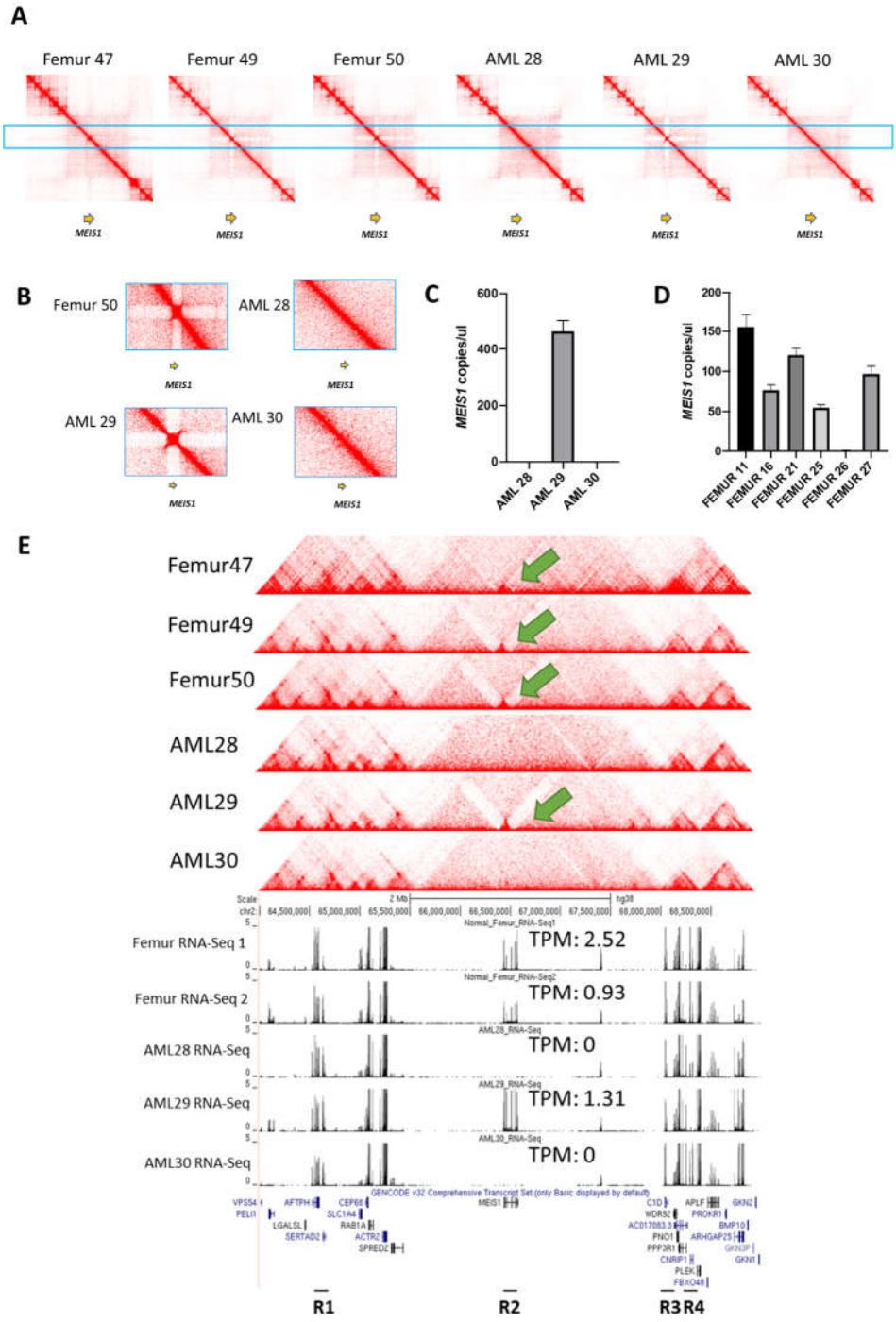



**Figure 3.2 A FIRE at *MEIS1* is heterogeneously present in AML clinical samples and the absence of the FIRE is associated with a lack of *MEIS1* gene expression. A.** Different *MEIS1* region FIRE profiles were observed in heatmaps

of CD34+ AML and Femur clinical samples (genomic region: chr2: 64,000,000-69,000,000, visualized by Juicebox (J. T. Robinson et al., 2018), color setting number: 5, normalization: coverage). **B.** Zoomed in heatmaps of *MEIS1* FIRE in Femur50, AML28, AML29, and AML30. **C.** *MEIS1* ddPCR of CD34+ AML clinical samples: AML28, AML29, and AML30. **D.** *MEIS1* ddPCR of CD34+ normal femur samples. **E.** RNA-Seq results visualized by UCSC genome browser tracks (Kent et al., 2002) integrated with Hi-C heatmaps in *MEIS1* region (genomic region: chr2: 64,000,000-69,000,000). *(Note: My lab colleague Ms. Yufen Goh did the sample collection and preparation part, and my lab collaborators Prof. Wilson Wang, and Prof. Chng Wee Joo from National University Hospital provided the clinical samples. Hi-C and RNA-Seq experiments were conducted by my colleague Dr. Deepak Babu. ddPCR experiment was conducted by my colleague Dr. Benny Wang Zhengjie.*)

## 3.3 Four Enhancer Regions Around the *MEIS1* FIRE Identified from 63 AML Patients were Involved in Chromatin Interactions with *MEIS1* in THP-1 Cells

Next, we tried to identify what kinds of elements are involved in chromatin interactions in this region, which might have resulted in *MEIS1* expression changes. As AML29, which contains the *MEIS1* FIRE, has an extremely high expression level of *MEIS1*, and a previous publication indicated that *MEIS1* expression was regulated by distal enhancers (Q. f. Wang et al., 2014), we asked

whether enhancers might be associated with the chromatin interactions. My colleague Ms. Ruchi Choudhary, a Ph.D. student in Nanyang Technological University School of Biological Sciences, has analyzed H3K27Ac ChIP-Seq data from published 63 AML patients and 2 CD34+ normal samples to identify the super-enhancers in this region (McKeown et al., 2017). The AML cell line THP-1 Hi-C from Phanstiel et al., 2017 indicated that THP-1 also shows this FIRE (**Figure 3.6A & B**) (Phanstiel et al., 2017). Thus, Dr. Benny Wang Zhengjie conducted 4C at the *MEIS1* Transcription Start Site (TSS) to detect the interacting regions of *MEIS1* in THP-1, and I helped him to make the interaction track. Integrated results of both 4C and super-enhancers analysis showed that in THP-1, *MEIS1* interacts with a few enhancer regions shown in most AML patients (**Figure 3.3A**). THP-1 H3K27Ac published data also presented these enhancers except region R2 (**Figure 3.3A**) (Mohaghegh et al., 2019). Thus, we concluded that *MEIS1* can interact with four enhancer regions: R1 (57 in 63 cases), R2 (38 in 63 cases), R3 (30 in 63 cases), and R4 (52 in 63 cases).

Additionally, Ms. Ruchi Choudhary also analyzed the H3K27Ac enrichment in these patients and normal samples in these four regions. This analysis revealed that even the normal samples also present some of the enhancers in R1, R2, R3, and R4 (**Figure 3.3A**), the H3K27Ac enrichment is lower than most of the AML samples (**Figure 3.3B**). This discovery indicates that normal haematopoietic cells might have weaker enhancer intensities in these enhancer regions. This might be the explanation of why AML29 expresses higher *MEIS1* than normal femurs.

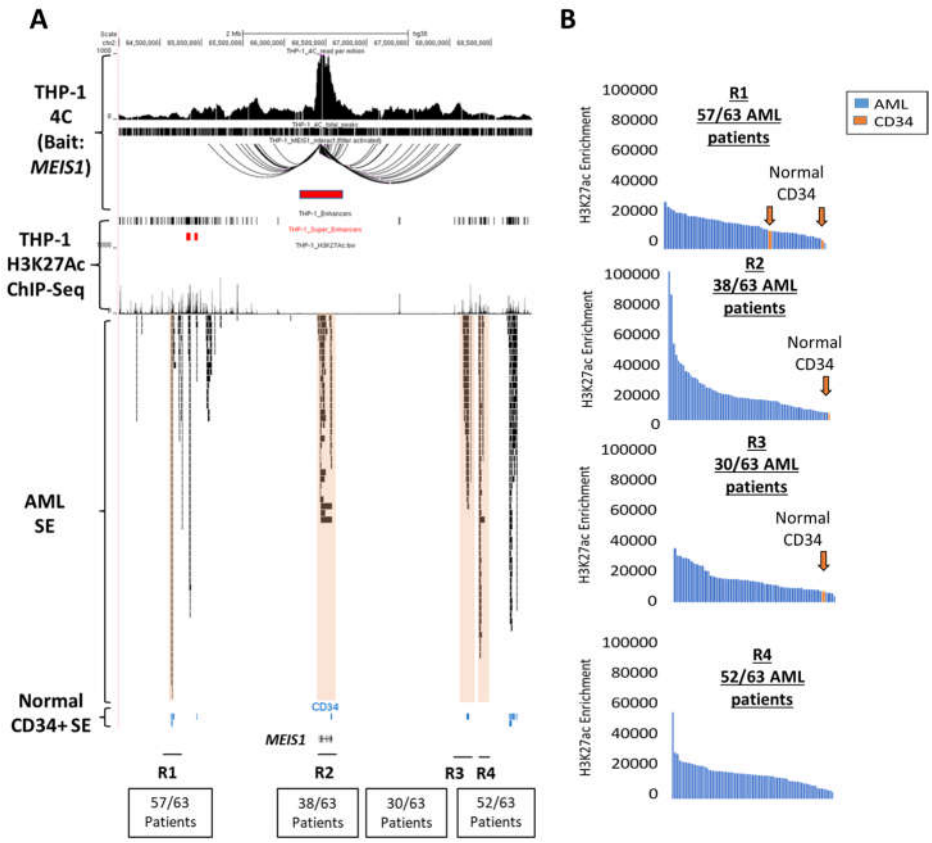



**Figure 3.3** ***MEIS1*** **region Super Enhancers (SE) profile in 63 AML clinical samples indicates four regions of enhancers interact with** ***MEIS1*****. A.** THP-1 *MEIS1* 4C and published H3K27Ac data (Mohaghegh et al., 2019) analysis integrated with super-enhancers from published 63 AML patient samples and 2 normal CD34+ clinical samples (McKeown et al., 2017). Tracks visualized in UCSC genome browser (Kent et al., 2002) (genomic region: chr2: 64,000,000-69,000,000). **B.** H3K27Ac enrichment of 63 AML patient samples and 2 normal CD34+ clinical samples in each enhancer region. Normal CD34+ samples were indicated by orange color and arrows. *(Note: MEIS1 4C experiment in THP-1 cells was conducted by my colleague Dr. Benny Wang Zhengjie. 63 AML patient SE analysis was done by my colleague Ms. Ruchi Choudhary, a PhD student in Nanyang Technological University School of Biological Sciences.)*

## 3.4 Integrated Hi-C, RNA-Seq and H3K27Ac ChIP-Seq Analyses in Total Bone Marrow AML Clinical Samples Indicates that the FIRE can bring together *MEIS1* and Enhancers

To investigate whether the enhancer regions present in the MEIS1 region observed in THP-1 and published AML patient samples are also present in AML samples with and without the *MEIS1* FIRE, we further performed integrated analyses of Hi-C, RNA-Seq, and H3K27Ac ChIP-Seq in a new batch of AML clinical samples. These samples were provided by my lab collaborator Prof. Chng Wee Joo from National University Hospital, collected and prepared by my lab colleague Ms. Winnie Fam for further experiments.

Total bone marrow was used this time instead of CD34+ sorted cells, as we could only obtain limited amounts of each clinical sample, and we planned to perform a variety of sequencing on these samples. AD796 and AD903 are frozen total bone marrow AML samples, and AML42, AML43, and AML44 are fresh total bone marrow AML samples. Hi-C, ChIP-Seq, and RNA-Seq experiments were conducted by my colleague Dr. Deepak Babu. For the Hi-C experiment, this time we used the Arima Hi-C Kit instead of Dovetail services. The Hi-C statistics are shown in **Table 3.2**. In this part, I did all the bioinformatics analysis of RNA-Seq, Hi-C, and ChIP-Seq.

Hi-C clustering analysis by PCA was also applied on total bone marrow AML samples, along with CD34+ sorted AML and femur samples (**Figure 3.4A**). We can see clearly that femurs and AMLs separately cluster together.

Interestingly, this time we figured out that frozen total bone marrow samples AD796 and AD903 were separated from other AML samples by PC2. Similarly, Femur47 separated with other femurs by PC2 (**Figure 3.4A**). We also discovered that Femur47, AD796, and AD903 all presented higher inter-chromosomal ratios compared with other samples in the same batches (**Table 3.3**). Since they also have lower numbers of called TADs and loops compared with other samples (**Table 3.1 & 3.2**), we guess the high ratio of inter-chromosomal interactions might be influencedinfluen the TAD and loop detections.

**Table 3.2 Hi-C statistics for frozen and fresh total bone marrow AML clinical samples**

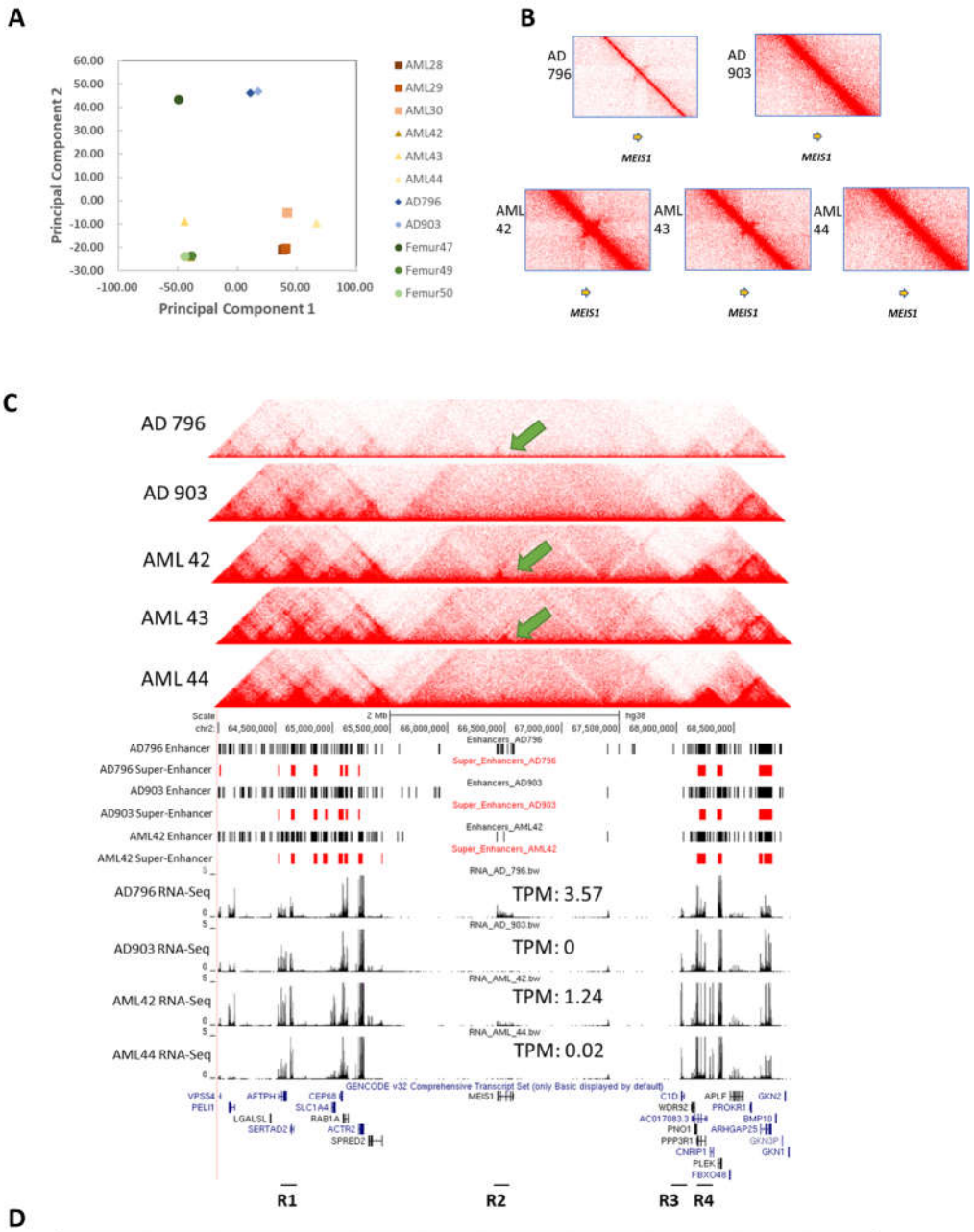
| | Total Sequenced Reads | Hi-C Contacts | #TAD | #loop |
|---|---|---|---|---|
| **AD796** | 884,279,692 | 556,385,368(62.92%) | 139 | 191 |
| **AD903** | 893,319,065 | 533,741,101(59.75%) | 1,778 | 719 |
| **AML42** | 844,725,953 | 520,623,955(61.63%) | 3,216 | 4,663 |
| **AML43** | 907,371,475 | 533,849,323(58.83%) | 2,580 | 3,305 |
| **AML44** | 894,400,600 | 550,344,133(61.53%) | 3,958 | 4,525 |

**Table 3.3 Hi-C quality statistics of all clinical samples**

| | Inter-chromosomal | Intra-chromosomal |
|---|---|---|
| **AML28** | 39,629,090(3.07% ) | 496,187,039(38.46%) |
| **AML29** | 30,706,047(2.47%) | 372,097,128(29.88%) |
| **AML30** | 31,659,727(2.28%) | 316,117,684(22.78%) |
| **Femur47** | 105,328,806(7.87%) | 491,072,142(36.70%) |
| **Femur49** | 31,044,622(2.68%) | 365,630,362(31.62%) |
| **Femur50** | 42,950,254(3.69%) | 438,711,906(37.65%) |
| | | |
| **AD796** | 318,071,761(35.97%) | 238,313,607(26.95%) |
| **AD903** | 198,409,825(22.21%) | 335,331,276(37.54%) |
| **AML42** | 139,260,217(16.49%) | 381,363,738(45.15%) |
| **AML43** | 175,204,085(19.31%) | 358,645,238(39.53%) |
| **AML44** | 160,151,828(17.91%) | 390,192,305(43.63%) |

Further, we examined the *MEIS1* FIRE using Juicebox heatmaps. AD796, AML42 and AML43 showed the FIRE while AD903 and AML44 did not (**Figure 3.4B & C**). Enhancers and super-enhancers called from H3K27Ac ChIP-Seq in AD796, AD903, and AML42 indicated that regardless of the presence of the FIRE, enhancer region R1, R2, and R4 were always present in AML samples. Enhancer R2 was present sometimes, and this discovery matches what we observed in 63 published patient samples (**Figure 3.3A**). *MEIS1* observed in RNA-Seq results also expressed in a similar manner we observed in CD34+ sorted clinical samples (**Figure 3.2**), that expression of *MEIS1* correlated with FIRE.

With these results above, we summarized all the observations into a table (**Figure 3.4D**). From this table, we can find that half of AML samples (AML28, AML30, AD903, and AML44 in total 8 AML samples) lost this *MEIS1* FIRE, along with nearly no expression of *MEIS1*. As the enhancer regions R1, R3, and R4 are present in AML samples with and without the FIRE, and interactions between *MEIS1* and these enhancer regions were found in both THP-1 cells and AML42 (**Figure 3.3A & Figure 3.5**), we next explored how the FIRE affects these chromatin interactions, enhancers, and transcription of *MEIS1* in next section.




| | Normal Femur1 | Normal Femur2 | AML 28 | AML 29 | AML30 | AD 796 | AD 903 | AML 42 | AML 43 | AML 44 |
|---|---|---|---|---|---|---|---|---|---|---|
| **FIRE presence?** | Yes | Yes | No | Yes | Very weak | Yes | No | Yes | Yes | Very weak |
| ***MEIS1* RNA level? (TPM)** | High (2.52) | Medium (0.93) | None (0) | Medium (1.31) | None (0) | Very high (3.57) | None (0) | Medium (1.24) | Not done | Low (0.02) |
| **Enhancers?** | Not done | Not done | Not done | Not done | Not done | Enhancers at R1, 2, 3, 4 | Enhancers at R1, 3, 4 only | Enhancers at R1, 2, 3, 4 | Not done | Not done |

**Figure 3.4 Hi-C, ChIP-Seq, and RNA-Seq integrated analysis on total bone**

**marrow AML clinical samples in the *MEIS1* region. A.** Principal Component Analysis of CD34+ sorted AML, CD34+ Femur, and total bone marrow AML clinical samples. (Using Hi-C contact matrix of all chromosomes under 1Mb resolution, KR normalized). **B.** Zoomed in heatmaps of all total bone marrow AML clinical samples in *MEIS1* FIRE region (Visualized by Juicebox (J. T. Robinson et al., 2018), color setting number: 8, normalization: coverage). **C.** *MEIS1* region heatmaps integrated with enhancer and super-enhancers called from ChIP-Seq and RNA-Seq of total bone marrow AML clinical samples visualized in UCSC genome browser (Kent et al., 2002) (genomic region: chr2: 64,000,000-9-69,000,000). **D.** Summary table of integrated analyses for all AML and Femur clinical samples. *(Note: My lab colleague Ms. Winnie Fam did the sample collection and preparation part, and my lab collaborators Prof. Chng Wee Joo from National University Hospital provided the clinical samples. Hi-C, ChIP-Seq, and RNA-Seq experiments were conducted by my colleague Dr. Deepak Babu.)*

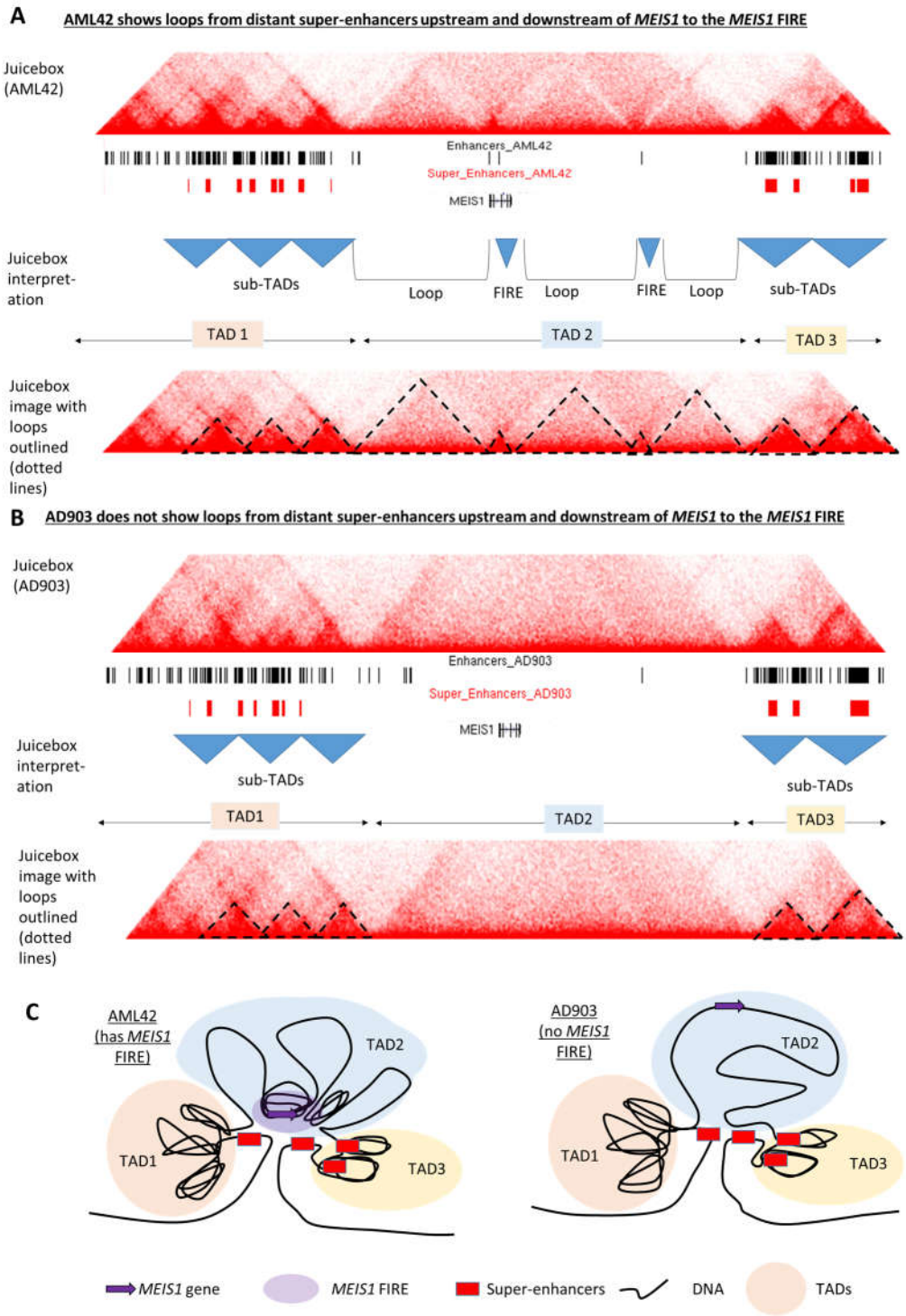



**Figure 3.5 Chromatin interactions indicated by heatmaps in AML42 and AD903 suggest that FIRE is essential for maintaining *MEIS1* chromatin interactions with enhancer regions. A.** The chromatin interactions annotated on

the heatmap of AML42 overlapped with AML42 enhancer and super-enhancer tracks. **B.** The chromatin interactions annotated on the heatmap of AD903 overlapped with AD903 enhancer and super-enhancer tracks. (Two heatmaps visualized by Juicebox (J. T. Robinson et al., 2018), genomic region: chr2: 64,000,000-9-69,000,000, color setting number: 8, normalization: coverage. All enhancer and super-enhancer tracks visualized on UCSC genome browser with the same genomic region) **C.** Schematics of chromatin interactions in AML42 and AD903 inferred from heatmaps. *(Note: This figure was produced by my supervisor Dr. Melissa Jane Fullwood.)*

To address tha the FIRE is a true chromatin interaction but not a structure variation, and the expression level change is not due to copy number variation (CNV), we did the TAD and loop calling, translocation and CNV analyses on these clinical samples. **Figure 3.6A** and **Figure 3.7A is** aligned TAD and loop calling tracks with heatmaps. Even though these calling method seems not perform perfectly on clinical samples, for example, *MEIS1* FIRE is not always been called through TAD and loop calling, but if we can combine two methods results, we can find that TAD or loop appeared in AML29, AML42, AML43 and all Femur samples. AD796 has no TAD and loop been called but this might because AD796 is a frozen sample which might have poor quality. Interestingly, AML30 has a loop been called while we cannot observe any interacion in MEIS1 region on the heatmap. This might be a false positive or a weak loop which cannot influence *MEIS1* expression. **Table 3.4** is a summary of top 4 ranked translocations in these clinical samples. No *MEIS1* region translocation is found.

Taken together with these results, we can say that *MEIS1* FIRE might be the true chromatin interactions that heterogeneously appeared in AML. CNV analyses in Figure 3.6B and Figure 3.7B also indicate no *MEIS1* region CNV is found. This result let us make sure that *MEIS1* expression alterations are not due to CNV.

**Table 3.4 Top 4 significant translocations of clinical samples**

| | Femur47 | Femur49 | Femur50 | AML28 | AML29 | AML30 | AD796 | AD903 | AML42 | AML43 | AML44 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | N.A. | chr19:22200000-22400000 with chrX:81500000-81700000 | chr1:188000000-188200000 with chr17:71600000-71800000 | chr4:49100000-49300000 with chr17:74100000-74300000 | chr4:34600000-34800000 with chr19:30000000-30200000 | chr2:41900000-42100000 with chr15:32700000-32900000 | N.A. | chr8:136600000-136800000 with chr12:34200000-34400000 | Chr8:2000000-2200000 with chr17:44400000-44600000 | chr3:27300000-27500000 with chr5:74800000-75000000 | chr15:29900000-30100000 with chrX:132600000-132800000 |
| **2** | N.A. | chr7:142400000-142600000 with chr12:25800000-26000000 | chr8:60400000-60600000 with chr17:71600000-71800000 | chr1:29200000-29400000 with chr22:49700000-49900000 | chr14:92200000-92400000 with chr17:35900000-36100000 | chr1:21580000-216000000 with chr15:29300000-29500000 | N.A. | chr8:92000000-92200000 with chr21:34800000-35000000 | chr6:168100000-168300000 with chr17:71600000-71800000 | chr3:33100000-33300000 with chr12:27800000-28000000 | chr5:165600000-165800000 with chr22:35700000-35900000 |
| **3** | N.A. | chr3:61300000-61500000 with chr22:11200000-11400000 | chr13:113500000-113700000 with chr19:19200000-19400000 | chr3:66600000-66800000 with chr22:23300000-23500000 | chr1:194600000-194800000 with chr17:71600000-71800000 | N.A. | N.A. | N.A. | chr10:45300000-45500000 with chr17:34500000-34700000 | chr5:45800000-46000000 with chr17:65200000-65400000 | chr15:29900000-30100000 with chr17:65200000-65400000 |
| **4** | N.A. | N.A. | chr16:52700000-52900000 with chr17:71600000-71800000 | chr8:21000000-23000000 with chr17:72000000-72200000 | chr1:181400000-181600000 with chr22:11000000-11200000 | N.A. | N.A. | N.A. | chr8:700000-900000 with chr15:63000000-63200000 | chr3:24700000-24900000 with chr17:20700000-20900000 | chr5:69100000-69300000 with chr19:27700000-27900000 |

*"N.A." indicates "not applicable".

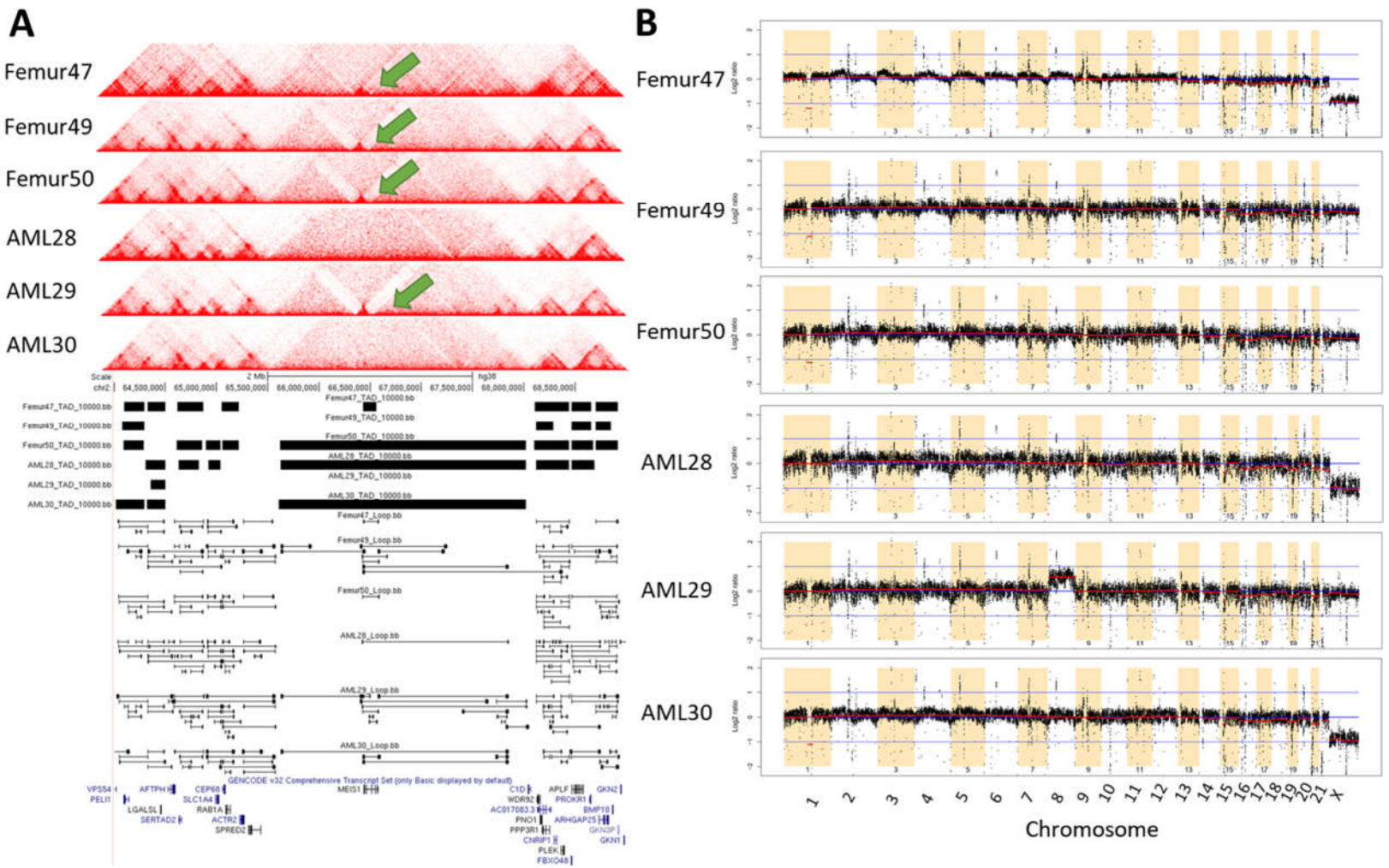

A
Femur47
Femur49
Femur50
AML28
AML29
AML30
B
Femur47
Femur49
Femur50
AML28
AML29
AML30
Chromosome

**Figure 3.6 No CNV is found in MEIS1 regions, and TAD and loop calling results indicate different chromatin interactions appeared in CD34+ selected clinical samples.A**.TAD and loop calling tracks visualized in UCSC genome browser (Kent et al., 2002) (genomic region: chr2: 64,000,000-9-69,000,000). **B.** CNV analysis show no copy number variation in MEIS1 region in CD34+ selected clinical samples. *(Note: My lab colleague Ms. Winnie Fam did the sample collection and preparation part, and my lab collaborators Prof. Chng Wee Joo from National University Hospital provided the clinical samples. Hi-C experiment experiments were conducted by my colleague Dr. Deepak Babu.)*

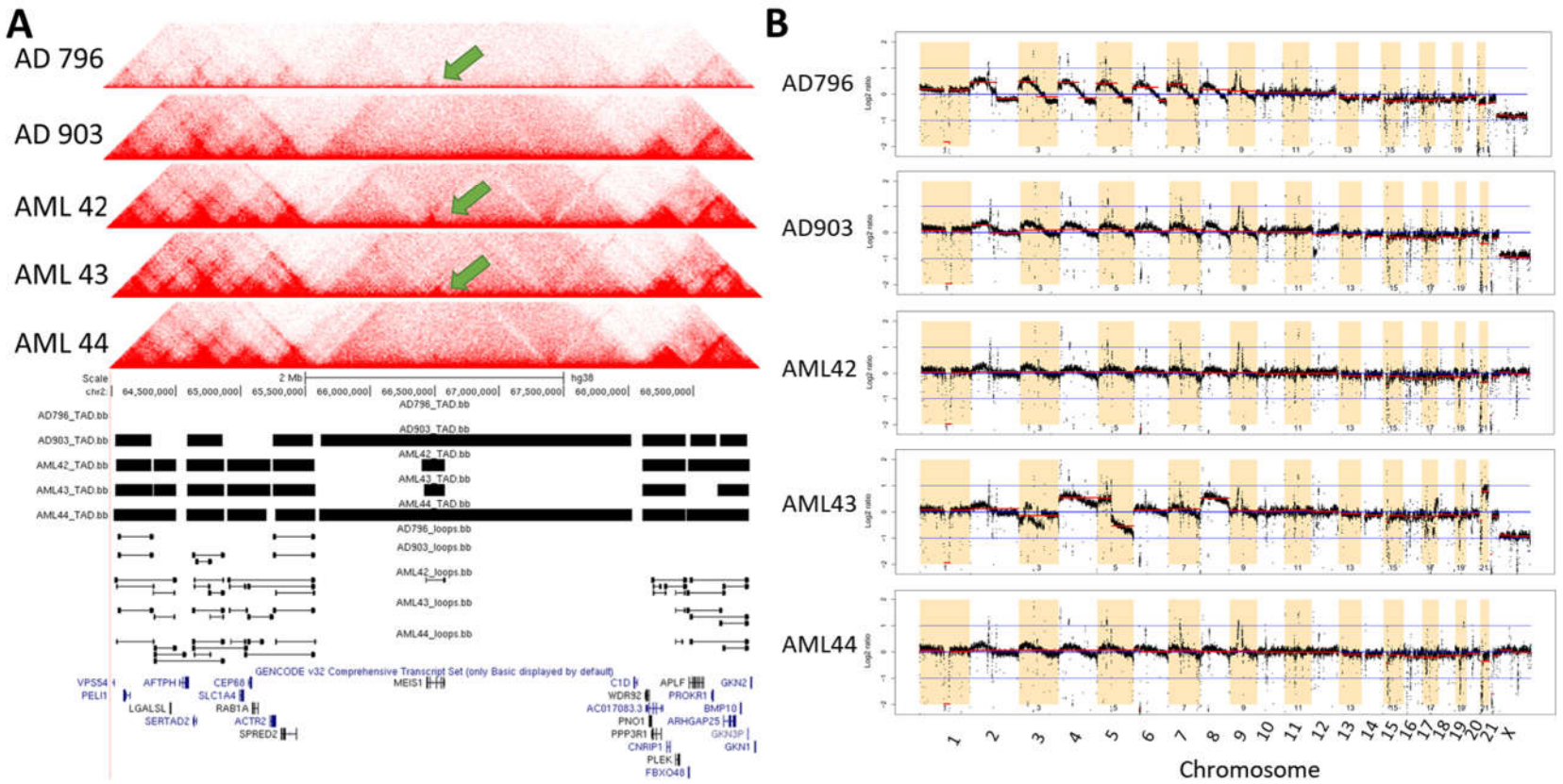



**Figure 3.7 No CNV is found in MEIS1 regions, and TAD and loop calling results indicate different chromatin interactions appeared in total bone marrow AML clinical samples.A**.TAD and loop calling tracks visualized in UCSC genome browser (Kent et

al., 2002) (genomic region: chr2: 64,000,000-9-69,000,000). **B.** CNV analysis show no copy number variation in MEIS1 region in total bone marrow AML clinical samples. *(Note: My lab colleague Ms. Winnie Fam did the sample collection and preparation part, and my lab collaborators Prof. Chng Wee Joo from National University Hospital provided the clinical samples. Hi-C experiment experiments were conducted by my colleague Dr. Deepak Babu.)*

## 3.5 CTCF Binding Site CRISPR Excision of *MEIS1* FIRE Border Indicates the FIRE is Essential for Maintaining Chromatin Interactions Between *MEIS1* and Enhancers in Myeloid Leukemia

### 3.5.1 THP-1 and K562 Can be Used as a Model to Study the *MEIS1* FIRE.

To study how the FIRE influences chromatin interactions between *MEIS1* and enhancer regions R1, R3, and R4, we planned to perform CRISPR excision of a CTCF site in *MEIS1* FIRE in a cell line. Myeloid leukemia cell lines such as THP-1, K562, and HL-60 were taken into considerations. We examined published Hi-C heatmaps in THP-1 (Phanstiel et al., 2017), K562, and GM12878 (Rao et al., 2014). GM12878 is a human lymphoblastoid cell line, which was used here as a representative sample without FIRE. Heatmaps indicated that both THP-1 and K562 have the FIRE, and the FIRE in THP-1 is weaker than K562 (**Figure 3.8A & B**). We could not find HL-60 published Hi-C data, so the HL-60 heatmap is not placed here. Dr. Benny Wang Zhengjie then conducted the *MEIS1* ddPCR in these cell lines. As we expected, K562 has the highest *MEIS1* expression while THP-1 has a lower expression, and GM12878 and HL-60 show no expression (**Figure 3.8C**). Taken together, we inferred that K562 might be a good model to help us study the *MEIS1* FIRE.

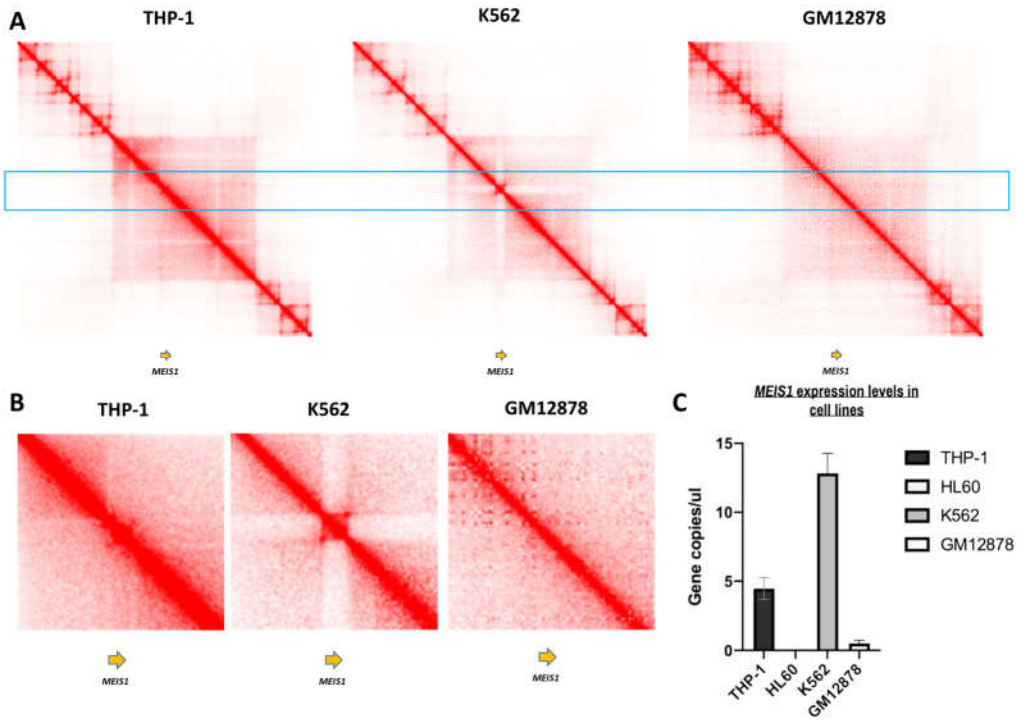



**Figure 3.8 THP-1 and K562 show *MEIS1* FIRE and MEIS1 expression. A.** *MEIS1* region heatmaps from published Hi-C data of THP-1 (Phanstiel et al., 2017), K562 and GM12878 (Rao et al., 2014) (genomic region: chr2:64,227,134-69,227,132 in human genome reference hg19. Visualized by Juicebox (J. T. Robinson et al., 2018), color setting number: 150 for THP-1, 40 for K562, and 50 for GM12878, normalization: coverage).**B.** Zoomed in *MEIS1* FIRE region heatmap for THP-1, K562, and GM12878. **C.** *MEIS1* ddPCR of THP-1, HL-60, K562, and GM12878. *(Note: THP-1 monocytes data, K562 combined data, and GM12878 combined data were selected from corresponding papers. ddPCR experiment in this figure was conducted by my colleague Dr. Benny Wang Zhengjie.)*

### 3.5.2 Reduced Chromatin Interactions Between *MEIS1* and Enhancer Regions were Observed in K562 CRISPR Excised Cells.

After we chose K562 as the study model, Dr. Benny Wang Zhengjie designed the CRISPR excision site at the CTCF binding site which acts as a border at the right side of FIRE (**Figure 3.9**), and applied the CRISPR knock out the operation. After excision, a 4C experiment was also conducted by him. I helped him to make the 4C interaction track. The K562 ChIP-Seq was also analyzed by me.

K562 published H3K27Ac ChIP-Seq (Consortium, 2012) was analyzed to ensure that enhancer regions R1, R2, R3, and R4 were present in K562 (**Figure 3.9**). Compared with the empty vector which presented a similar interaction profile of THP-1 *MEIS1* 4C, CTCF binding site knock out cells showed a sharp decrease in chromatin interactions reaching out of the FIRE region with enhancers (**Figure 3.9**).

Thus, we can conclude that the CTCF binding site at the border of the *MEIS1* FIRE is important in maintaining the chromatin interactions between *MEIS1* and enhancer regions such as R1, R3, and R4.

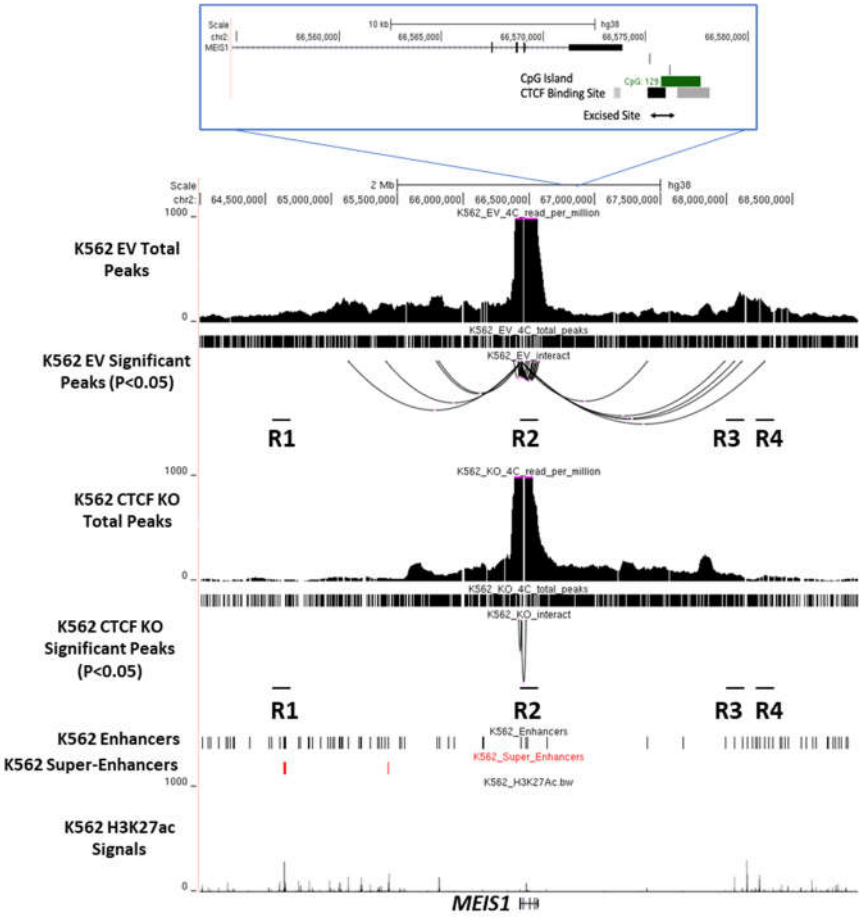



**Figure 3.9 CTCF knockout at *MEIS1* FIRE region in K562 cells reduced chromatin interactions between *MEIS1* and enhancer regions.** From top to bottom: Zoomed in CRISPR region of CTCF binding site; 4C signal track, total peak track and significant interactions track in empty vector cells; 4C signal track, total peak track, and significant interactions track in knock out cells; Published K562 H3K27Ac data (Consortium, 2012) analyses of enhancers, super-enhancers, and signal tracks. All tracks were visualized with UCSC genome browser (Kent et al., 2002) (genomic region: chr2: 64,000,000-9-69,000,000). *(Note: K562 CTCF binding cite knock out and 4C experiments in this figure were conducted by my colleague Dr. Benny Wang Zhengjie)*

### 3.5.3 Multiple Cellular Alterations were Induced by the Absence of *MEIS1* CTCF Binding Site in K562

After we observed the chromatin interactions loss in CRISPR excision of *MEIS1* FIRE applied K562 cells, we tried to investigate what patterns were influenced by these alterations. Dr. Benny Wang Zhengjie further conducted ChIP-qPCR of *MEIS1* and H3K27Ac at enhancer regions R1, R2, and R3, and *MEIS1* promoter region. The reason why we examined the MEIS1 protein binding here is that we were very curious about whether MEIS1 will regulate itself by binding to the sequence. The results showed that MEIS1 protein shows a significantly decreased binding in the R1 and *MEIS1* promoter region (**Figure 3.10A**). These results suggested that MEIS1 protein might bind to enhancer and promoter regions of itself to regulate its expression, and FIRE loss let MEIS1 protein decrease which reduced such binding amount. H3K27Ac signals also decreased in R1, R3, and *MEIS1* promoters, which suggested that enhancer ability might be weakened after loss of the CTCF binding site at the border of the *MEIS1* FIRE.

RT-qPCR of *MEIS1* and *MYC*, which was previously found to be a downstream target of *MEIS1* in Zebrafish (Bessa et al., 2008), was also done by Dr. Benny Wang Zhengjie. As expected, once FIRE loss, the *MEIS1* expression decreased in all three KO clones, especially in C1 and C3, no *MEIS1* is detected. *MYC* as the possible downstream target also decreased significantly (**Figure 3.10B**).

Cell viability assay was also applied by Dr. Benny Wang Zhengjie to figure out whether cell growth and death were altered in knock-out cells or not. By checking every 24 hours, KO clones showed significantly lower cell viability after 48 hours compared with empty vector clones, which suggested that the CTCF binding site at the *MEIS1* FIRE border is also responsible for maintaining cell growth and health (**Figure 3.10C**)

Taken together, we summarized a schematic of what kinds of cellular changes were caused after CRISPR excision of a CTCF binding site at the border of the FIRE (**Figure 3.10D**). After excision, chromatin loops between *MEIS1* promotor and enhancers were perturbed as well as the H3K27ac levels at the enhancers and the binding of MEIS1 protein to the promoter and enhancers. *MEIS1* expression was greatly decreased to almost no expression at all and MYC was also downregulated. Cell growth was slowed down.

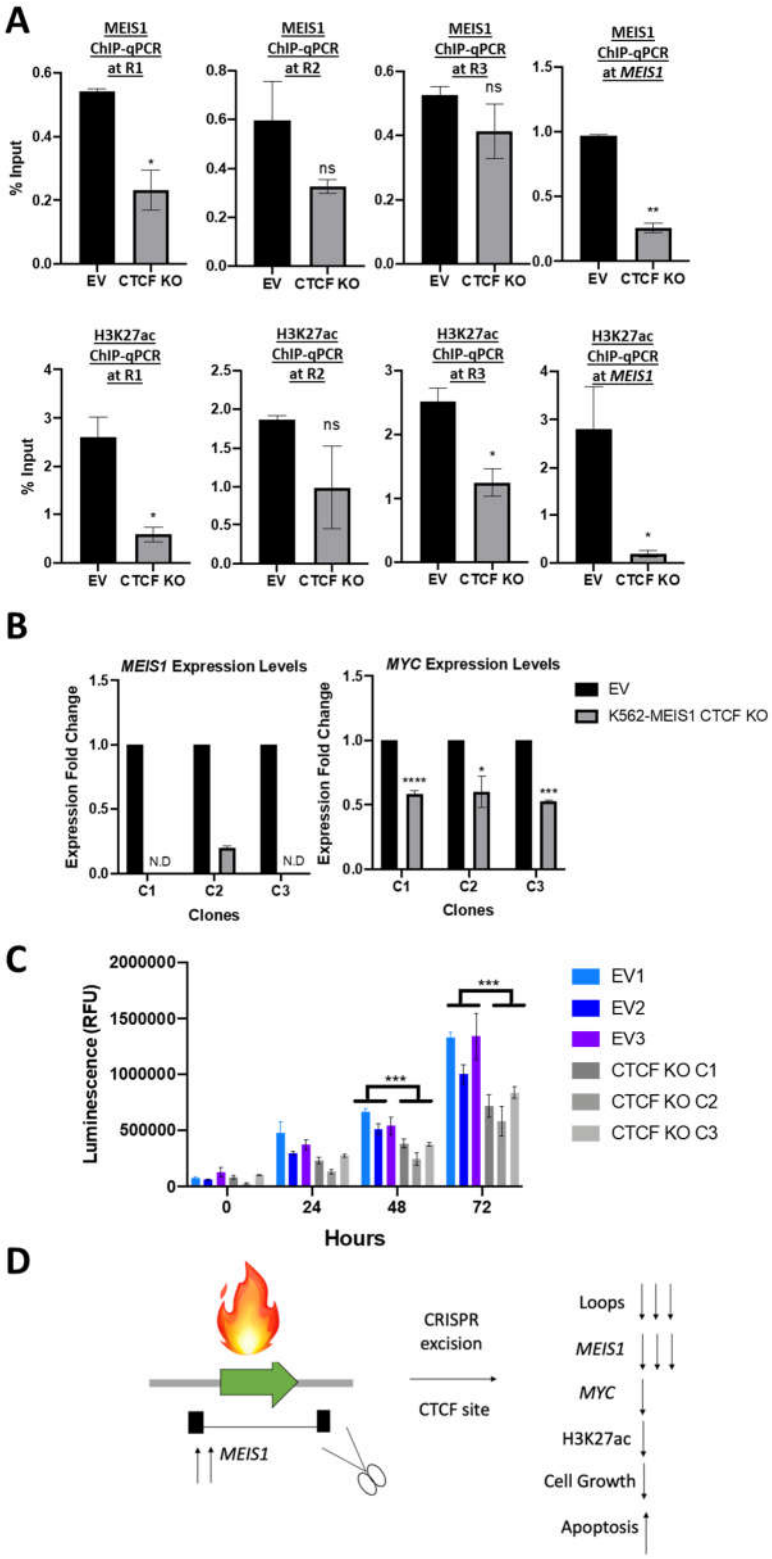



**Figure 3.10 CRISPR excision of CTCF binding site of *MEIS1* FIRE region**

**led to multiple cellular changes. A.** *MEIS1* and H3K27Ac ChIP-qPCR in enhancer regions R1, R2, and R3 and *MEIS1* region (** - P< 0.01, *- P < 0.05, ns- not significant). **B.** RT-qPCR of *MEIS1* and MYC in 3 clones of CTCF binding cite knock out cells (****P<0.0001, *** - P< 0.001 ** - P< 0.01, *- P < 0.05, N.D.- not detected). **C.** Cell viability assays in 3 days on both empty vectors and knock out 3 clones for each. **D.** Schematic summary about the cellular changes after CRISPR excision. A "fire" symbol represents *MEIS1* FIRE. More arrows mean more fold changes of alterations. *(Note: ChIP-qPCR, ddPCR, and cell viability assay were conducted by my colleague Dr. Benny Wang Zhengjie. Figure in part D was produced by my supervisor Dr. Melissa Jane Fullwood.)*

## 3.6 Summary

In summary, we found that chromatin interaction landscapes might change in AML clinical samples compared with normal haematopoietic stem cells, and more altered loops are associated with oncogenes (**Figure 3.1**). Research in the oncogene *MEIS1* region surprisingly led us to conclude that a specific chromatin interaction termed Frequently Interacting Region (FIRE) was heterogeneously present in AML clinical samples, which is stable in normal femur samples. The absence of a CTCF binding site at the border of this FIRE will subsequently cause *MEIS1* downregulation, and *MEIS1* downstream target *MYC* downregulation by destroying the chromatin interactions between *MEIS1* promoter and four enhancer regions: R1, R2, R3, and R4 and weakening the H3K27ac binding levels at these regions. We speculate the following schematic

of the mechanism of *MEIS1* FIRE functioning: Two different subtypes of *MEIS1* FIRE might exist in AML, with or without the FIRE. With the existence of the FIRE, the MEIS1 promoter can interact with enhancers as normal cells, but the ability of enhancers was strengthened so that *MEIS1* increased, and more MEIS1 protein will bind to this region. While in the subtype without FIRE, no interactions can be formed without the help of the FIRE, and *MEIS1* will decrease as well as the MEIS1 protein binding (**Figure 3.11**).

There are still mysteries left in this proposed mechanism. For example, what factors initiate carcinogenesis? What alterations lead to the different subtypes? How does MEIS1 protein binding regulate its expression? Why is the enhancer ability of *MEIS1* in AML stronger than normal? Further investigations need to be performed to examine these questions.

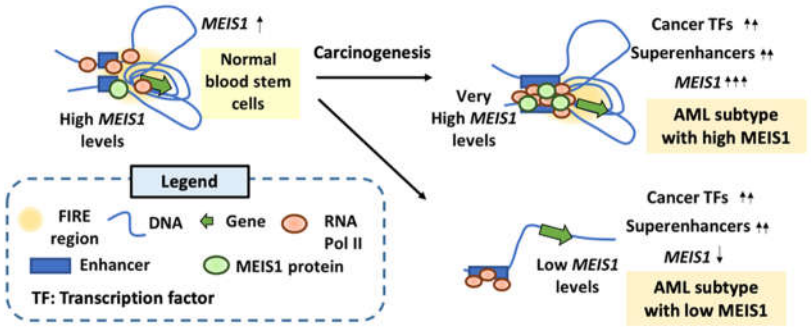



**Figure 3.11 Proposed schematic of the mechanisms of how *MEIS1* FIRE influences *MEIS1* expression and chromatin interactions, as well as other cellular changes.** We hypothesized that there will be two subtypes of *MEIS1* FIRE in AML cells, which were indicated in this figure *(Note: This figure was produced by my supervisor Dr. Melissa Jane Fullwood.)*

# 4. *DNMT3A* Loss Leads to Altered Chromatin Interactions and Epigenetic Landscapes in Myeloid Leukemia

## 4.1 *DNMT3A* Mutation Might Lead to Dysregulation of TAD Boundaries in Clinical AML Samples.

The DNA Methyltransferase 3 Alpha (*DNMT3A*) is a gene responsible for maintaining and modulating DNA methylation. As described in section 1.4.2, *DNMT3A* is the most frequently mutated epigenetic factor gene in AML (Ley et al., 2013). AML patients with *DNMT3A* mutations usually show poor outcomes (Hou et al., 2012; Ley et al., 2010). Within this group of patients, over half of them were affected by the mutation in the R882 codon (37 patients in 62 DNMT3A mutated patients) (Ley et al., 2010), which tends to form non-functional tetramers. In other words, functional DNMT3A protein is lost. Thus, we asked, what is the influence upon chromatin interactions and epigenetic if *DNMT3A* is mutated or lost? To investigate this question, we first looked at the RNA-Seq data provided by the TCGA-LAML project, from The Cancer Genome Atlas Program (TCGA), which contains 201 cases in total, with 27 *DNMT3A* mutant patient cases (Ley et al., 2013).

Inspired by William A. Flavahan et al., 2016, which found that certain *IDH* mutant gliomas have dysregulated TAD boundaries due to CTCF binding site loss (William A. Flavahan et al., 2016), we inferred that *DNMT3A* might have a

similar influence. We reasoned that as *IDH* is also DNA methylation-related gene, dysregulation of another DNA methylation gene such as *DNMT3A* might also lead to altered TAD boundaries. With the DNA methylation level changes, we anticipate that some of the methylation-sensitive CTCF binding sites will change in terms of CTCF occupancy because CTCF is anticorrelated with DNA methylation (Phillips & Corces, 2009; H. Wang et al., 2012). This loss of CTCF occupancy at methylation-sensitive CTCF binding sites will further alter the chromatin interactions as CTCF is an important protein information of chromatin interactions (Hansen, Pustova, Cattoglio, Tjian, & Darzacq, 2017; Phillips & Corces, 2009; Rao et al., 2014).

Thus, we followed the gene correlation method for investigating altered TAD boundaries which were described in William A. Flavahan et al., 2016 (William A. Flavahan et al., 2016), by using the TCGA-LAML RNA-Seq data (Ley et al., 2013), to investigate whether *DNMT3A* changes tend to lead to changes in TADs in AML. First, we divided the RNA-Seq data which have a total of 201 cases into two groups: *DNMT3A* wild type (127 cases) and *DNMT3A* mutant (27 cases). Then we merged the Hi-C contact matrix of 3 Femur samples mentioned in chapter 3, to get a high-resolution Hi-C matrix of normal AML patient samples, and called TADs from this matrix under 10kb resolution, using VC as the normalization.

We further used this TAD list as a reference to divide genes into different pairs and defined whether they are the same domain pairs or cross boundary pairs. If two genes belonged to the same TAD, they are regarded as “same domain pairs”, while if they belonged to different TADs, they are regarded as

"cross boundary pairs" (Detailed methods can be found in section 2.7). We calculated the gene correlations from RNA-Seq data for both same domain and cross boundary gene pairs, to get the correlation changes of different gene pairs in *DNMT3A* mutant and wild types cases.

The correlation analysis results in *DNMT3A* wild-type AML cancers revealed that the gene correlation of same domain pairs is always higher than cross boundary pairs with the increase of gene pair distances (**Figure 4.1A**). This is expected because TADs are self-interacting genomic regions. This observation suggests that if TADs are not dysregulated, the same domain correlation will keep a higher correlation than a cross boundary.

Next, when we investigated the alterations of gene pairs correlations across *DNMT3A* mutant and wild type, we calculated the delta correlation by using the correlation in *DNMT3A* mutant minus wild type. If the delta correlation of one gene pair is less than zero, this observation indicates that these gene pair have a decreased correlation in mutant, and *vice versa*. In **Figure 4.1B**, we can observe more cross boundaries in the positive delta correlation part, which indicates that cross boundary pair (orange dots) correlations tend to increase in *DNMT3A* mutant, while same domain pairs (blue dots) tend to show decreased correlation. This scenario supports our hypothesis that some of the TAD boundaries might be dysregulated in *DNMT3A* mutant cases.

We then identified all possible dysregulated boundaries by analyzing the changes in the gene pairs correlation between *DNMT3A* cases and unmutated AMLs. If at least one same domain gene pairs in two TADs shows a decreased

correlation, and at the same time, cross boundary gene pairs which belong to these two TADs are found at least one support an increased correlation, then the boundary in between them will be regarded as probably dysregulated (Total number: 2364).

Examining these possible dysregulated boundaries, we found chromosome 1 has the highest proportion of dysregulated boundaries, and chromosome 21 has the lowest number of dysregulated boundaries, which correlates with the length of each chromosome – chromosome 1 is longer than chromosome 21 (**Figure 4.1C**). However, we also found that several chromosomes such as chromosome 4, chromosome 13, chromosome 16, and chromosome 18 have fewer results compared with even shorter chromosomes (e.g., chromosome 20). This might indicate that these chromosomes have fewer gene expressional level changes or more conserved chromatin interactions. We note that this investigation is correlative, and further investigations need to be conducted to check whether there are dysregulated regions or not.

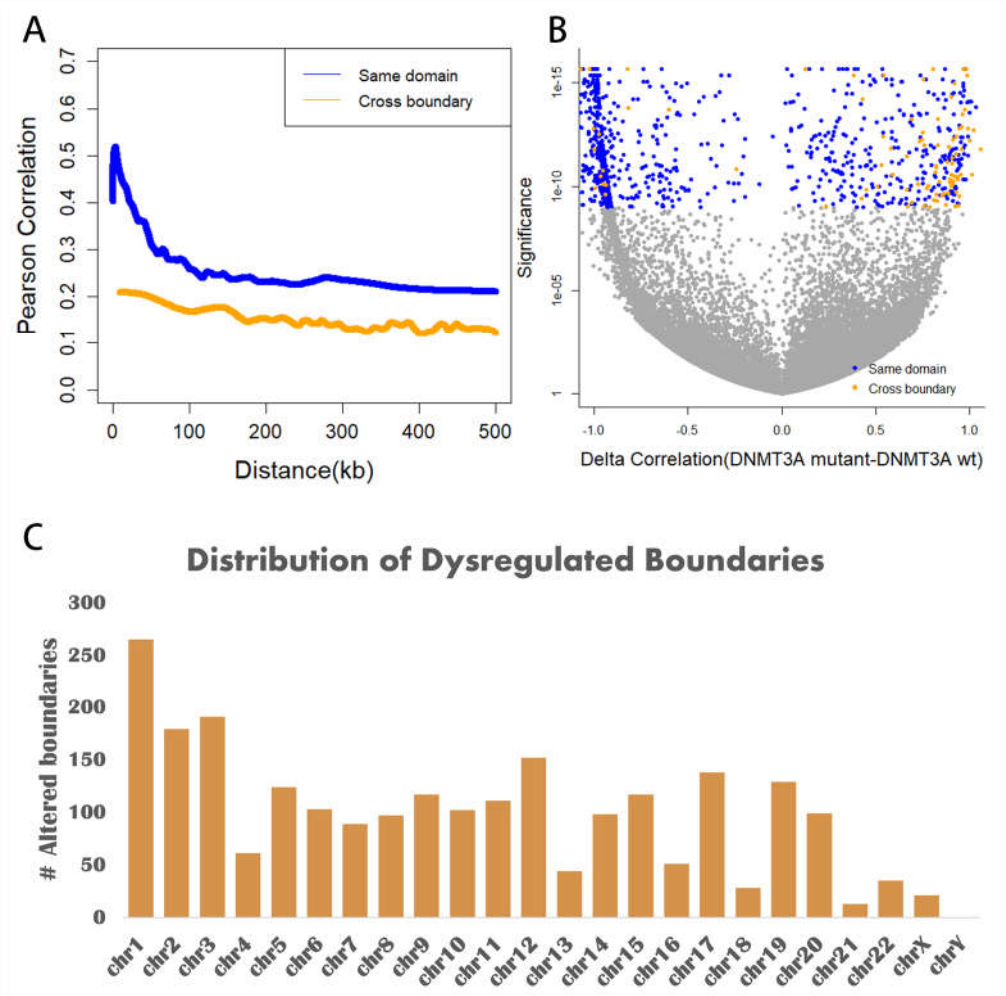



**Figure 4.1 Correlation analysis on TCGA-LAML dataset indicated boundaries might be altered due to *DNMT3A* mutation.A.** Correlation changes in *DNMT3A* wild type cases as a function of the distance of same domain pairs (blue) and cross boundary pairs (orange). **B.** Correlation changes between *DNMT3A* mutant and *DNMT3A* wild type cases in same domain pairs (blue) and cross boundary pairs (orange). **C.** Distribution in different chromosomes of possible altered boundaries detected after correlation analysis.

## 4.2 Altered Chromatin Interaction and Other Epigenetic and Transcriptional Profile Have been Observed in *DNMT3A* CRISPR Knock Out Cells.

Since around 60% of *DNMT3A* mutation cases involve the R882 codon and this kind of mutation leads to loss of function in *DNMT3A*, we tried to mimic this kind of mutation by using CRISPR to knock out *DNMT3A* in K562 cells and lead to loss of DNMT3A protein. To check that our reasoning that *DNMT3A* mutations tend to lead to reduced function of DNMT3A and reduced gene expression is correct, I analyzed the transcriptional level of *DNMT3A* in *DNMT3A* mutant AML cases and found that the transcriptional level is decreased compared with wild type cases (**Figure 4.2B**). This gives us more confidence that the knockout of *DNMT3A* can mimic the real *DNMT3A* mutations in myeloid leukemia cells.

First, my lab collaborator Dr. Qiling Zhou from Prof. Daniel Tenen's lab in the Cancer Science Institute of Singapore designed the guide RNA to target the exon 7 of *DNMT3A*. With the CRISPR experiment applied, one clone of *DNMT3A* knock-out cells with a deletion of 11bp by checking with Sanger sequencing in the target region are successfully generated (**Figure 4.2A**). My lab colleague Dr. Deepak Babu used this clone to prepare two replicates of Hi-C, RNA-Seq, and ChIP-Seq experiments. He also repeated this knock-out an experiment to get a clone 2 with 1 base pair of inserted "T" by checking with Sanger sequencing (**Figure 4.2A**). This clone 2 is only used in the RT-qPCR of *DNMT3A* and *PLOD2*, and we have not performed any additional Hi-C, ChIP-Seq, RNA-Seq studies on this clone yet. Thus, in this thesis, we will call knock

out clone 1 as KO, and empty vector as Vec_Con, without specifying the KO clone 1, if clone 2 is not present.

In the RT-qPCR experiment of *DNMT3A*, which was conducted by Dr. Deepak Babu, we can see a clear and significant loss of *DNMT3A* transcription in both clones (**Figure 4.2C**).

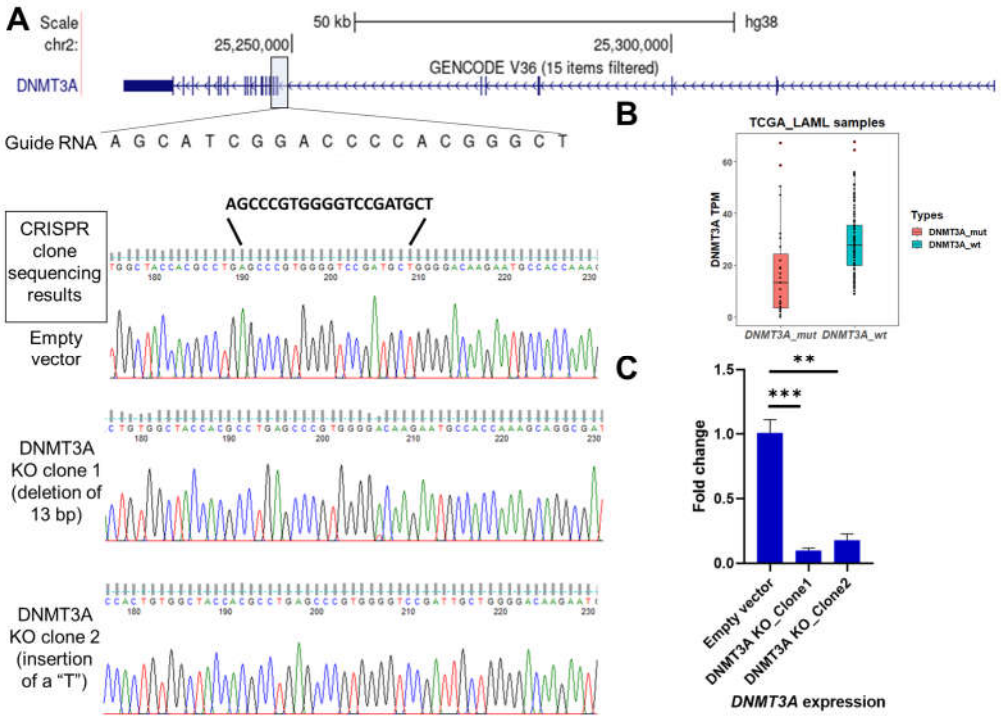



**Figure 4.2 CRISPR *DNMT3A* knockout in K562 cells. A.** Sanger sequencing for *DNMT3A* knock-out clones 1 and 2, and empty vector control. **B.** *DNMT3A* transcription level in TCGA-LAML *DNMT3A* mutant and wild-type cases. **C.** RT-qPCR shows *DNMT3A* down-regulation in *DNMT3A* knock out clone versus the empty vector *(Note: CRISPR experiment of KO clone1 and Empty Vector was done by my lab collaborator Dr. Qiling Zhou from Prof. Daniel Tenen's lab in*

*the Cancer Science Institute of Singapore, and my lab colleague Dr. Deepak Babu repeated the CRISPR experiment, to obtain DNMT3A CRISPR knockout clone 2. DNMT3A RT-qPCR experiment was also conducted by Dr. Deepak Babu).*

After we knocked out the *DNMT3A* in K562 cells, we further investigated the consequences of *DNMT3A* loss on chromatin interactions, transcription, and epigenetic. Hi-C, RNA-Seq, and ChIP-Seq sequencing was conducted by my colleague Dr. Deepak Babu, and I did the further bioinformatics analyses.

TADs and loops were predicted for both KO and Vec_Con (Detailed methods can be seen in section 2.3.3.3). To figure out whether chromatin interactions are dysregulated or not, we compared the TADs and loops between KO and Vec_Con. Similarity ratio (see in section 2.3.3.4) over 90% TADs in KO and Vec_Con were regarded as common TADs, and TADs with less than 90% similarity were considered to be specific TADs. Loops with both two anchors that overlapped in KO and Vec_Con are common loops and others were considered to be specific loops. By these comparisons, a greater proportion of loop alterations (more than half) were found in KO compared with TAD alterations (around 30%) (**Figure 4.3A**). This difference can be observed clearly in our Venn plot (**Figure 4.3B**) as TADs have more overlapped proportions. Even though this result might have false positives and false negatives due to the limitations of TAD and loop calling algorithms (see in section 1.2.4.2 and section 1.2.4.3) and further manual curation needs to be done, we can conclude from this observation that TADs tend to be more conserved compare with loops, and chromatin interactions have a high chance to be altered in *DNMT3A* KO cells.

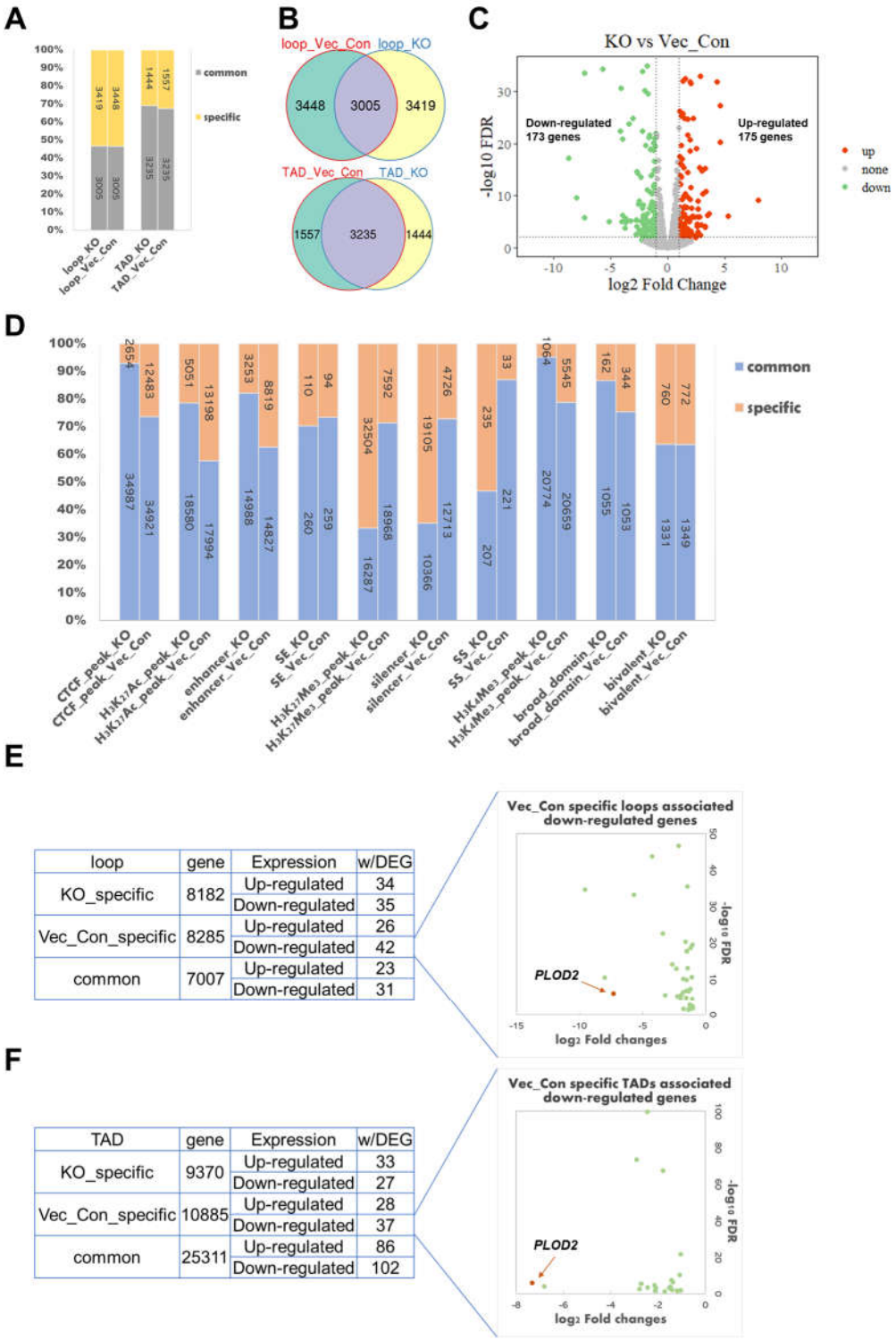
Next, we tried to understand what is the impact of *DNMT3A* loss on other types of epigenetic, such as histone modifications, as well as gene expression levels. We applied CTCF, H3K27Ac, H3K27Me3, and H3K4Me3 ChIP-Seq and RNA-Seq to study KO and Vec_Con. Previous work on DNA methylation has shown that DNA methylation levels will influence the epigenetic landscapes (Gu et al., 2018), as well as CTCF binding (Phillips & Corces, 2009; H. Wang et al., 2012).

In our results, ChIP-Seq for CTCF, H3K27Ac, H3K27Me3, and H3K4Me3 showed an altered profile in KO cells (**Figure 4.3D**). Enhancers and super-enhancers were called from H3K27Ac signals, silencers, and H3K27Me3-rich regions (MRRs), which we also called "super silencers", were called from H3K27Me3 signals, and broad H3K4Me3 domain was defined as top 5% size of H3K4Me3 peaks as previously described (Cao et al., 2017; Dahl et al., 2016). An altered profile of super-enhancers, super silencers, and broad H3K4Me3 domains were observed in KO. Each of these categories had gains (KO specific) and losses (Vec_Con specific).

As the altered epigenetic was observed, we were curious whether there were alterations in gene expression level profile. Through analyzing the RNA-Seq data of KO and Vec_Con, and filtering with FDR<0.05 and log2 fold change >1 and <-1, 173 down-regulated genes and 175 up-regulated genes were found (**Figure 4.3C**), which indicates that *DNMT3A* knock out has an impact upon the transcription profile.

Different epigenetic alterations function to affect cells in different ways. For example, H3K27Ac or H3K4Me3 are associated with gene activation, hence the gain or loss of H3K27Ac and H3K4Me3 could further control the corresponding gene expressions. H3K27Me3, as a mark of silencers, can also control gene expression. Chromatin interactions are another form of epigenetic marks which we found to have changed in KO cells and given that chromatin interactions have been associated with control of gene expression (Peng et al., 2019), (**Figure 4.3A**), we then asked how the changed chromatin interactions are associated with altered gene expression.

To investigate this question, we further counted the genes where their transcription starting site (TSS) is near altered loops/ TADs by using the criteria that TSSs located inside altered TAD are the altered (KO/Vec_Con specific) TAD associated genes, and TSSs inside a flanking region of +15kb and -15kb of two anchors of altered loops are sorted into altered KO/Vec_Con specific) loops associated genes. Then we overlapped the altered loop/TAD associated genes with differentially expressed genes and performed the manual curation to identify examples of genes to investigate in more detail. One interesting gene named Procollagen-Lysine,2-Oxoglutarate 5-Dioxygenase 2 (*PLOD2*) was found to be both involved in altered loops and TADs, and extremely down-regulated. (**Figure 4.3E & F**). We then prepared an integrated map for this gene region to figure out what influenced this gene expression, which includes a Hi-C Juicebox image to show the TAD and loop structures, and a UCSC genome browser screenshot to visualize the CTCF, H3K27Ac, H3K27Me3, and H3K4Me3 ChIP-Seq as well as RNA-Seq patterns in this region.




| loop | gene | Expression | w/DEG |
|---|---|---|---|
| KO_specific | 8182 | Up-regulated | 34 |
| | | Down-regulated | 35 |
| Vec_Con_specific | 8285 | Up-regulated | 26 |
| | | Down-regulated | 42 |
| common | 7007 | Up-regulated | 23 |
| | | Down-regulated | 31 |

| TAD | gene | Expression | w/DEG |
|---|---|---|---|
| KO_specific | 9370 | Up-regulated | 33 |
| | | Down-regulated | 27 |
| Vec_Con_specific | 10885 | Up-regulated | 28 |
| | | Down-regulated | 37 |
| common | 25311 | Up-regulated | 86 |
| | | Down-regulated | 102 |

**Figure 4.3 Chromatin interaction and other epigenetic alterations found in *DNMT3A* knock-out cells. A.** Common and specific TADs and loops in KO and

Vec_Con. **B.** Venn plots for common and specific TADs and loops in KO and Vec_Con. **C.** Volcano plot for up and down-regulated genes in KO compared with Vec_Con **D.** CTCF peaks and histone marks generally presented a proportion of alteration in KO versus Vec_Con. **E.** KO and Vec_Con common/specific loops associated genes can find several up/down-regulated. *PLOD2* (orange dot) can be found in Vec_Con specific loops associated genes and it is conspicuously downregulated in KO cells. **F.** KO and Vec_Con common/specific TADs associated genes can find several up/down-regulated. *PLOD2* (orange dot) also can be found in Vec_Con specific TADs associated genes and it is also conspicuously downregulated in KO cells. *(Note: CRISPR knock-out experiment in this figure was conducted by my lab collaborator Dr. Qiling Zhou from Prof. Daniel Tenen's lab. Hi-C, ChIP-Seq, and RNA-Seq experiments in this figure were performed by my colleague Dr Deepak Babu)*

## 4.3 *DNMT3A* Loss Leads to Alterations in FIREs, CTCF binding, Histone Modifications, and Expression of *PLOD2* and *MACC1*.

*PLOD2* codes for the lysyl hydroxylasesis protein LH2, and is responsible for catalyzing the hydroxylation of collagen lysyl residues (Qi & Xu, 2018). Mutations in *PLOD2* can cause Bruck Syndrome, a disease that consists of bone fragility and congenital joint contractures (Gistelinck et al., 2016), but there are also studies that show that overexpressed *PLOD2* also be detected in many types of cancers, such as glioma, cervical and liver cancer (Gjaltema, de Rond, Rots, &

Bank, 2015), suggesting that *PLOD2* may have an impact on cancer biology. However, we note that decreased *PLOD2* expression in myeloid leukemia has not been reported and the interplay of how it can decrease is not yet known. To check what influenced *PLOD2* expression in the case of *DNMT3A* KO, we integrated all results to generate an integrated map.

First, we checked the Hi-C heatmap of the *PLOD2* region, and we observed that a specific pattern of a small square TAD-like structure in the Vec_Con heatmap disappeared in KO (**Figure 4.4A**). This type of pattern is similar to the pattern in Chapter 3, the *MEIS1* FIRE. In addition to the heatmap, when we looked at the insulation score at this region, the *PLOD2* region also showed a peak in Vec_Con but not KO, which indicates that a sub-TAD was lost in KO. The same pattern could be observed in the TAD and loop list: Vec_Con showed both a TAD and a loop that did not show up in KO. As so many pieces of evidence suggest that there is loss of chromatin interactions in this region, and since this region also fulfills the FIRE definition (mentioned in section 1.2.2.2), we would consider this region to be an example of a FIRE loss in *DNMT3A* KO cells.

Along with this FIRE loss, CTCF binding loss, H3K27Ac loss, and H3K27Me3 gained and decreased expression of *PLOD2* were also found (**Figure 4.4D**). My colleague Dr. Deepak Babu further confirmed the significant downregulation of *PLOD2* in both two clones of KO by RT-qPCR (**Figure 4.4B**).

Taken together of these observations, a possible assumption of why *PLOD2* is down-regulated is that the methylation might be altered at the CTCF binding

region due to *DNMT3A* knockout, and the CTCF binding loss caused the FIRE loss. This FIRE is responsible for the interactions between *PLOD2* TSS and the left side enhancer. Loss of this FIRE, as well as the enhancer, caused *PLOD2* downregulated. At the same time, a gained H3K27Me3 silencer around the TSS region together with an unchanged H3K4Me3 peak creates a gained bivalent region around *PLOD2* TSS, which might further decrease the *PLOD2* expression as this region used to be active because only the H3K4Me3 peak existed in the vector control cells.

Interestingly, in clinical samples from the TCGA-LAML project, *PLOD2* seems to be usually low expressed in *DNMT3A* mutated cases, while *DNMT3A* wild-type cases show a few numbers of outliers with extremely high *PLOD2* expression (**Figure 4.4C**). This might be an indication that *DNMT3A* loss can inhibit *PLOD2* overexpression by destroying the FIRE at *PLOD2*. In further work, it would be interesting to explore what is the functional relevance and significance of *PLOD2* downregulation in *DNMT3A* mutated AML.

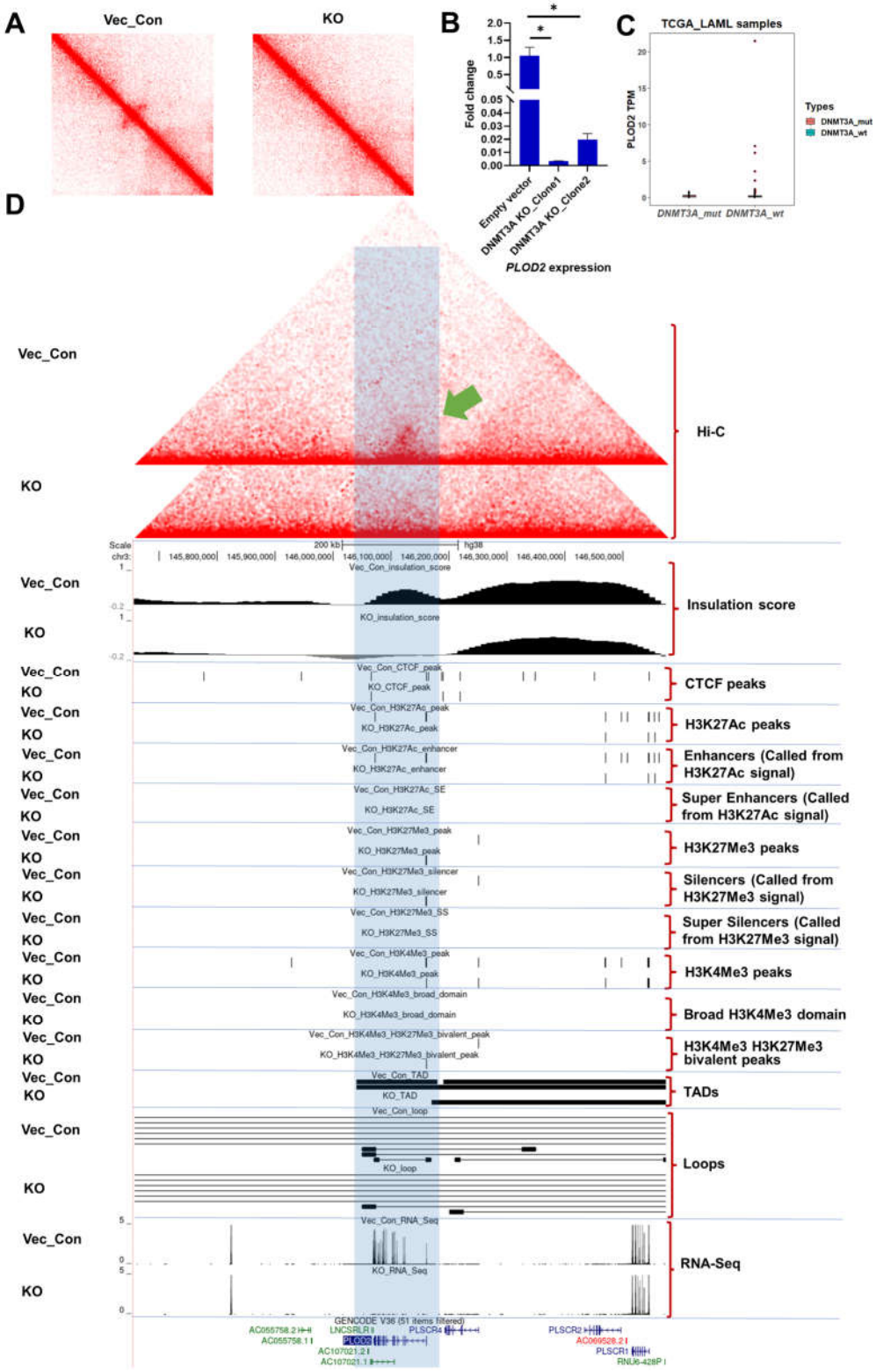



**Figure 4.4** ***PLOD2*** **region integrated analyses indicate a CTCF binding loss**

**caused boundary loss in KO cells. A.** Hi-C heatmaps in *PLOD2* region in Vec_Con and KO cells revealed a FIRE loss (genomic region: chr3:145,656,587-146,574,036, visualized by Juicebox (J. T. Robinson et al., 2018), coverage normalization is used, 8 as the color number setting). **B.** RT-qPCR results showed an *PLOD2* downregulation in all 2 KO clones. **C.** TCGA-LAML *DNMT3A* mutated patient samples always show a low expression of *PLOD2*, while *DNMT3A* wild type cases show a few outliers of extremely high *PLOD2* expression. **D.** Integrated map with tracks visualized in UCSC genome browser (Kent et al., 2002) in *PLOD2* region (genomic region: chr3:145,656,587-146,574,036). *(Note: CRISPR knock out, Hi-C, RT-qPCR, ChIP-Seq and RNA-Seq experiments in this figure were performed by my colleague Dr. Deepak Babu)*

Next, we asked whether more FIREs can be found to be dysregulated in KO. After a round of manual curation by my colleagues Ms. Judy Xiaoman Shao, a Ph.D. student who did her Ph.D. rotation in Dr. Melissa Fullwood's Lab, and Dr. Deepak Babu, another interesting region was found around the Metastasis-Associated in Colon Cancer Protein 1 (*MACC1*) gene.

*MACC1* is a gene involved in cell growth and hepatocyte growth factor pathway and acts as a prognostic indicator in colon cancer metastasis (Ge, Meng, Zhou, Zhang, & Ding, 2015; Stein et al., 2009). The link between *MACC1* and AML is still unknown. Only one paper mentioned that *MACC1* expression might

cause lymphatic metastasis in colorectal cancer (Z. Zhang, Jia, Wang, Du, & Zhong, 2021).

We can observe that *MACC1* is expressed, albeit at a low level, in the UCSC RNA-Seq track of Vec_Con, while there is nearly no expression in KO (**Figure 4.5D**). We note that *MACC1* can be found in Vec_Con specific TAD genes, but not in the differential expression list as it is quite lowly expressed so it was filtered out during the analysis. To confirm this down-regulation, Ms. Judy Xiaoman Shao, conducted an RT-qPCR experiment in KO clone1, which showed a significant decrease compared with Vec_Con (**Figure 4.5B**). We will further repeat this experiment in KO clone 2 as well in the future. In clinical samples, *DNMT3A* wild-type cases also show a higher expression of *MACC1* compared with *DNMT3A* mutant (**Figure 4.5C**). The functional significance of *MACC1* down-regulation upon *DNMT3A* loss needs further elucidation in the future.

The FIRE in the *MACC1* region is absent in KO cells which can be observed in heatmaps (**Figure 4.5A**), insulation score as well as TAD tracks (**Figure 4.5D**). At the same time, lost CTCF peaks in the FIRE region, and a common H3K4Me3 peak at the right side of FIRE was found. Based on these observations, we speculate that *DNMT3A* loss changes the methylation level, and further leads to CTCF loss. This FIRE can assist the interaction of TSS of *MACC1* and H3K4Me3 region, but in KO samples, the FIRE disappeared, and *MACC1* can no longer interact with the active signal, and then the expression level goes down. With *DNMT3A* knocked out, an H3K4Me3 peak is also lost right near the *MACC1* TSS region, this might also be another explanation for the reduced gene expression at *MACC1*.

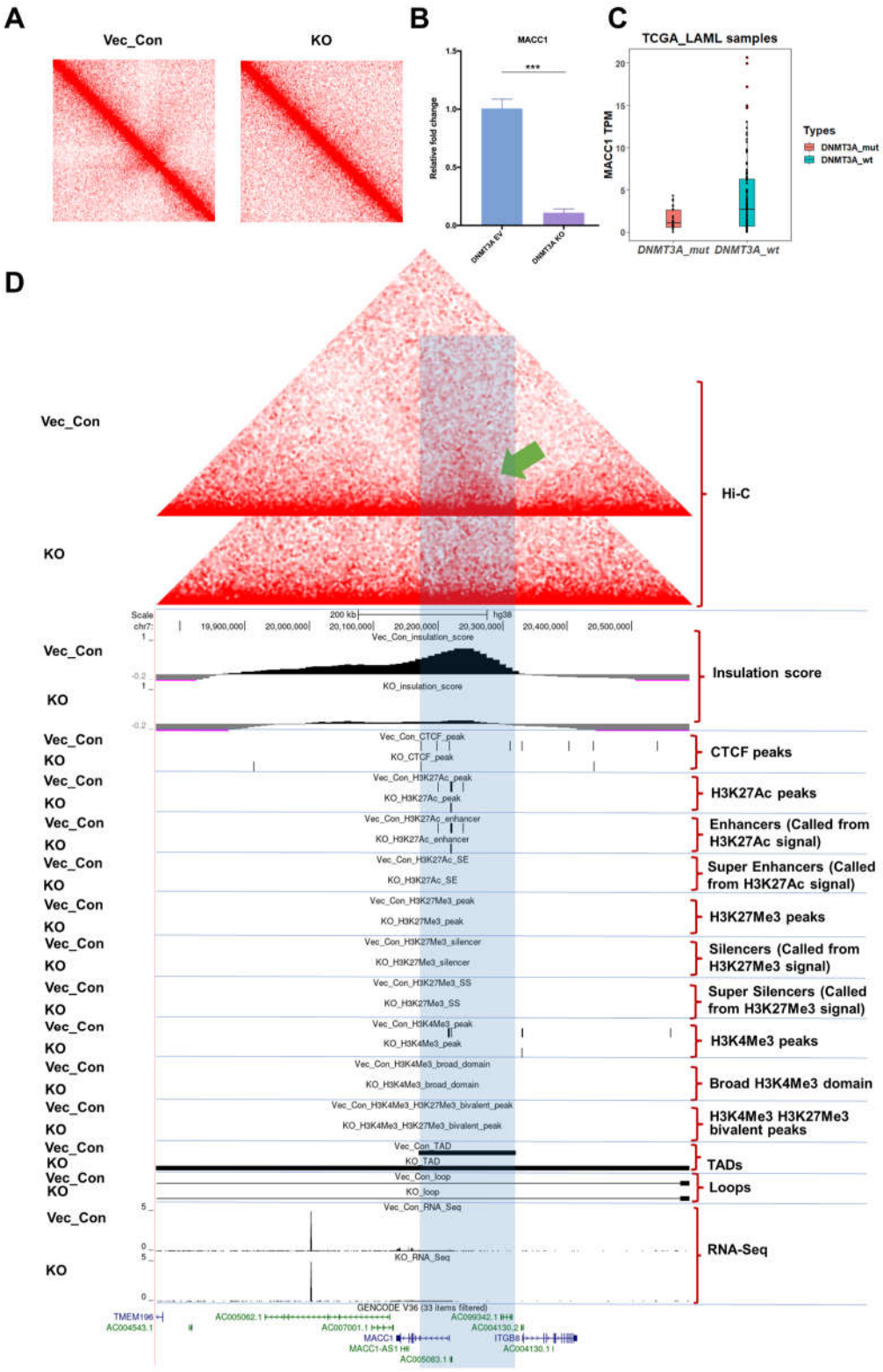



**Figure 4.5** ***MACC1*** **region integrated analyses indicate a CTCF binding**

**loss caused boundary loss in KO cells. A.** Hi-C heatmaps in *MACC1* region in Vec_Con and KO cells revealed a FIRE loss (genomic region: chr7:19,762,370-20,589,669, visualized by Juicebox (J. T. Robinson et al., 2018), coverage normalization is used, 8 as the color number setting). **B.** RT-qPCR results showed an *MACC1* downregulation in the KO1 clone. **C.** TCGA-LAML *DNMT3A* mutated patient cases show a lower expression of *MACC2* compared with *DNMT3A* wild-type cases. **D.** Integrated map with tracks visualized in UCSC genome browser (Kent et al., 2002) in *MACC1* region (genomic region: chr7:19,762,370-20,589,669). *(Note: RT-qPCR experiments in this figure were conducted by Ms. Judy Shao, a Ph.D. student at the Cancer Science Institute who did a Ph.D. rotation in in Dr. Melissa Fullwood's lab. CRISPR knock out, Hi-C, ChIP-Seq, and RNA-Seq experiments in this figure were performed by my colleague Dr Deepak Babu)*

## *4.4 DNMT3A* Loss also Leads to Alterations in Chromatin Loops, CTCF Bindings, Histone Modifications, and Expression of *ARID5B*.

As we have observed two FIREs altered in the KO clone, we attempted to figure out whether chromatin loops alterations can be observed. As our loop and TAD comparison shown in **Figure 4.3A & B**, there are more loop alterations in KO cells compared with TADs. During the manual curation of Ms. Judy Xiaoman Shao and Dr. Deepak Babu, a candidate with loss of loops was found in AT-Rich Interactive Domain-Containing Protein 5B (*ARID5B*) region.

*ARID5B* is one of the genes in the AT-rich interaction domain (ARID) family, a family of DNA binding protein, which will modulate chromatin structure (Gregory, Kortschak, Kalionis, & Saint, 1996; Herrscher et al., 1995). *ARID5B* plays an important role in transcription modulation by recruiting PHF2 in the target gene region (P. Wang et al., 2020). A Single Nucleotide Polymorphism (SNP) in *ARID5B* was reported to be influential in Acute Lymphoblastic Leukemia (ALL) (Reyes-León et al., 2019; Tao et al., 2019), childhood leukemia (Emerenciano et al., 2014), and male promyelocytic leukemia (J. Zhou et al., 2019). Downregulated *ARID5B* is associated with leukemia relapse (P. Wang et al., 2020). Taken together, alterations in *ARID5B* might be important for myeloid leukemia.

We observed the loss of two chromatin loops loss with CTCF loss in the *ARID5B* region (**Figure 4.6A & D**), and loss of H3K27Ac enhancers and H3K4Me3 peaks and broad domains, as well as gain of H3K27Me3 silencers at the TSS region which formed a bivalent region (**Figure 4.6D**). *ARID5B* has also listed in Vec_Con specific loops associated genes but has been filtered out of the differential expressed gene list due to low expression. Dr. Deepak Babu also designed the RT-qPCR and confirmed the downregulation of *ARID5B* in both KO clone 1 and clone 2 (**Figure 4.6B**). *ARID5B* is also downregulated in clinical *DNMT3A* mutated cases compared with *DNMT3A* wild-type cases (**Figure 4.6C**).

We suggest that CTCF loss caused the loss of two loops and these two loops facilitated the interaction of the TSS of *ARID5B* with H3K27Ac enhancers and H3K4Me3 broad domains in Vec_Con to maintain high expression. Loss of active signals of histones and gain of silencer signal could lead to the decrease of

*ARID5B* gene expression levels.

As *ARID5B* can also modulate other gene expression levels (P. Wang et al., 2020), and it can influence chromatin structure (Gregory et al., 1996; Herrscher et al., 1995), we suggest that *ARID5B* might be a gene that was directly affected by *DNMT3A* loss which further influence the landscape of 3D genome architecture and expression profile in myeloid leukemia, thus leading to indirect effects of *DNMT3A* loss.

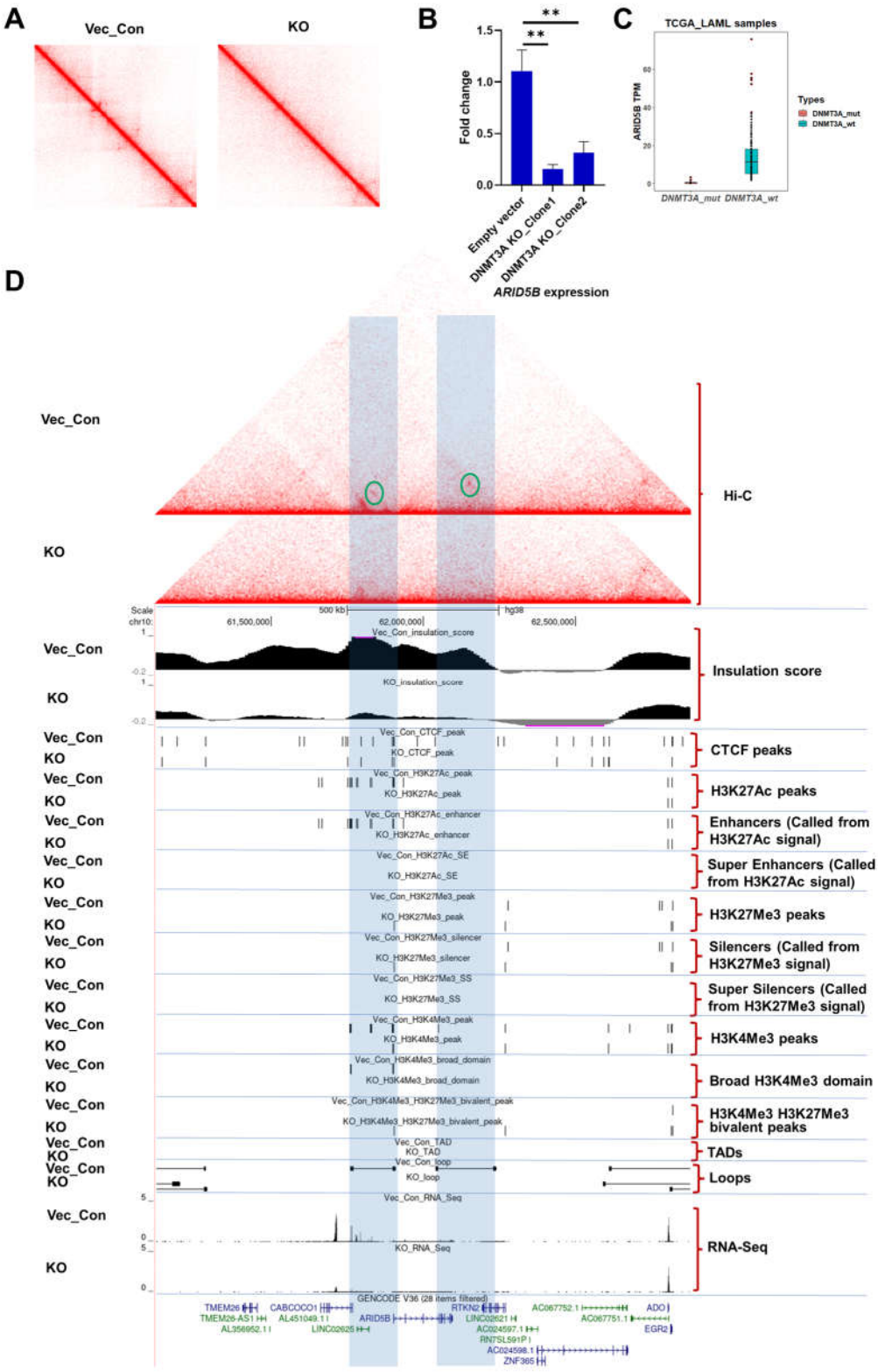



**Figure 4.6** ***ARID5B*** **region integrated analyses indicate two chromatin**

**loops loss caused three gene expression level changes in KO cells. A.** Hi-C heatmaps in *ARID5B* region in Vec_Con and KO cells revealed two chromatin loops loss (genomic region: chr10:61,120,715-62,877,928, visualized by Juicebox (J. T. Robinson et al., 2018), coverage normalization is used, 8 as the color number setting). **B.** RT-qPCR results showed an *ARID5B* downregulation in the KO1 clone. **C.** TCGA-LAML *DNMT3A* mutated patient cases show a lower expression of *ARID5B* compared with *DNMT3A* wild-type cases. **D.** Integrated map with tracks visualized in UCSC genome browser (Kent et al., 2002) in *ARID5B* region (genomic region: chr10:61,120,715-62,877,928). *(Note: RT-qPCR, CRISPR knock out, Hi-C, ChIP-Seq and RNA-Seq experiments in this figure were performed by my colleague Dr Deepak Babu)*

## 4.5 Summary

In summary, we found two FIREs lost in the *PLOD2* and *MACC1* region and two loops lost in the *ARID5B* region in *DNMT3A* knock-out clone, as well as other altered epigenetic changes such as CTCF binding and histone marks. All these alterations might act together to lead to lower gene expression levels of these three genes. *ARID5B* might also lead to cause further alterations in gene expression and chromatin structure modulation in myeloid leukemia, as *ARID5B* is an epigenetic regulator. This, and other mechanisms, may contribute to the widespread dysregulated gene expression seen in the *DNMT3A* knock-out clone. The dysregulated gene expression and altered epigenetic factors in *DNMT3A*

might be reasons why *DNMT3A* mutated cases in AML usually show a poorer outcome as compared with wild-type AML.

# 5. Conclusions and Future Directions

## 5.1 Conclusions

In this thesis, we asked two questions: (1) whether Topologically Associating Domains (TADs) and chromatin loops are dysregulated in AML compared with normal haematopoietic cells, and (2) whether *DNMT3A* mutations lead to dysregulated TADs and chromatin loops in AML.

To answer the first question, we examined the 3D genome architecture of AML clinical samples by Hi-C experiments compared with normal haematopoietic stem cells and figured out that there are differences in chromatin interactions between AML and normal haematopoietic stem cells, and many altered chromatin interactions are associated with oncogenes. Going further, we found the heterogeneous presence of *MEIS1* FIRE as an example of altered chromatin interactions that correlated with *MEIS1* expression. We also used a K562 model to perform CRISPR investigation of the *MEIS1* FIRE region by removing a CTCF binding site at the FIRE border to investigate the mechanisms behind the *MEIS1* FIRE in modulating the *MEIS1* expression. We proposed a mechanism that two subtypes of *MEIS1* FIRE might exist in AML patients and FIRE can facilitate the *MEIS1* promoter interaction with enhancers, which further influence the *MEIS1* expression, MYC expression, enhancer strength, and cell growth.

For the second question, we first analyzed the AML clinical RNA-Seq data in the TCGA online database, to compare the *DNMT3A* mutant and wild type and

figure out the differences in gene pairs correlations. By dividing gene pairs into same domain pairs and cross boundary pairs and checking the gene correlations, we observed the features of dysregulated boundaries. Then we used the K562 cell line as a model to mimic the *DNMT3A* mutation by CRISPR knockout of *DNMT3A*. Through an integrated analysis of Hi-C, ChIP-Seq of CTCF, H3K27Ac, H3K27Me3, and H3K4Me3 and RNA-Seq in *DNMT3A* CRISPR knock out K562 cells compared with vector control cells, we identified three examples of dysregulated chromatin interactions: *PLOD2*, *MACC1,* and *ARID5B*. Two FIRE regions and two loops were lost in these three examples, along with gene expression level change and alterations of epigenetic landscapes.

Through the research work completed above, we concluded these key points:

1. Chromatin interactions including FIRE, loops, and TADs tend to be altered in AML compared with normal stem cells, and these altered loops are associated with oncogenes.
2. *MEIS1* FIRE is heterogeneously present in AML patients.
3. A CTCF binding site at the *MEIS1* FIRE is essential for facilitating the chromatin interaction between *MEIS1* promotor with enhancers and modulation of enhancer intensities.
4. Loss of a CTCF binding site at the *MEIS1* FIRE will induce many cellular changes such as gene expressional changes and cell growth changes.
5. *DNMT3A* is required for the maintenance of TAD boundaries.
6. *DNMT3A* knockout in K562, the myeloid leukemia cell line, will lead

to chromatin interaction changes, altered epigenetic landscape, and gene expression changes.

7. *PLOD2*, *MACC1,* and *ARID5B* regions were affected by *DNMT3A* loss, and showed altered chromatin interactions, CTCF binding, histone modifications, an gene expression levels. *MACC1* and *ARID5B* were also downregulated in *DNMT3A* mutant AML patient samples compared with *DNMT3A* wild-type patients.

## 5.2 Discussion

In this thesis, we investigated the 3D genome organization in AML clinical samples compared with normal samples and in the *DNMT3A* CRISPR knock-out K562 cells versus vector control cells. From this work, we found out that AML samples, and *DNMT3A* loss of function, lead to chromatin interaction alterations. In this discussion, I will address the technical limitations we faced, unsolved problems in our research, interesting phenomena we found, and possible therapeutic strategies based on our findings.

### 5.2.1 Limitations of Hi-C Technique and Analysis

In this thesis, we used Hi-C to study chromatin interactions. The reason we choose Hi-C is that the Hi-C technique can detect the chromatin interactions in the whole genome, and we wished to investigate the complete 3D genome organization maps in our research questions. Compared with other “C” techniques, such as 4C and 5C, Hi-C indeed was the best choice for our works to

detect possible dysregulated chromatin interactions. However, it still has some limitations.

First is its high cost, both in terms of experimental cost and analysis cost. To detect the whole genome chromatin interactions, Hi-C requires a very deep sequencing depth, especially in the clinical samples. In our case, 300,000,000 of Hi-C contact reads were the minimum requirement if we want to reach the 10kb resolution. To obtain this Hi-C contact number, at least 2 × 150bp 1,000,000,000 reads are required as the input for clinical samples due to the duplicates and mapping quality filtering. In other word, ~100 × sequencing depth for each base is required (300 × 1,000,000,000 base pairs, divided by the genome length 3,088,269,832 (hg38.p2), equals to 97.14). Cell line Hi-C is less technically challenging, but ~30 × sequencing depth is also needed.

Such deep sequencing will generate large data files, which will take up storage space for hard disks ~1Tb per sample to store raw data and metadata. To analyze these data, the hardware requirement is also crucial. Taking the software Juicer as an example, the ideal settings are >= 4 cores (min 1 core) and >= 64 GB RAM (min 16 GB RAM). When calling loops by HiCCUPS, Graphics Processing Unit (GPU) is also required (N. C. Durand, M. S. Shamim, et al., 2016). If the researcher only has a workstation or personal computer with limited hardware settings, the analysis of high-resolution Hi-C data could be a tough challenge. Even with a good server, it takes days or weeks per sample for a single analysis step. For the current research, Hi-C is still an expensive and irreplaceable tool for studying 3D genome organization. However, in the future, we hope that with

further technical optimization, the costs of Hi-C can be reduced.

The second limitation is the limitations in the algorithms. Even though there are a variety of algorithms to call different scales of chromatin interactions such as TAD and loops, the unsolved problem for these algorithms is the accuracy. False-positive and false negative predictions make it difficult to get accurate comparisons between samples. Manual curation for these false positives and false negatives is highly labor-intensive as usually thousands of results will be reported. Moreover, as TADs are hierarchical structures, there is no gold standard for TAD definition in computational terms (Dali & Blanchette, 2017).

Some of the TAD calling algorithms do not report overlapping TADs (e.g., TopDom and HiCseg) (Shin et al., 2016) (Lévy-Leduc et al., 2014) while some of the algorithms use an overlapping list to identify TAD hierarchies (e.g., arrowhead and TADtree) (Rao et al., 2014) (Weinreb & Raphael, 2016). The questions of "what is the statistical definition for TADs and sub-TADs?", "What are the typical sizes of TADs" and "What should be the exact statistical characteristics to determine the boundaries of TAD?", have yet to be resolved.

### 5.2.2 Difficulties in Clinical Research in AML

In this thesis, we carried out the clinical sample study comparing AML versus normal, and we faced some difficulties in the clinical research.

The first ddifficulty is the heterogeneity of AML. AML is a highly heterogeneous disease with many different subtypes (Horibata et al., 2019; Ley et

al., 2013; S. Li, Mason, & Melnick, 2016; Swaminathan et al., 2018), making genetic and epigenetic research difficult. In our work, we figured out that the *MEIS1* FIRE is heterogeneously present in AML patients. However, due to the heterogeneity of AML, to understand this phenomenon better, we would require more clinical samples, and here we faced the second problem - the difficulty to acquire samples.

Clinical samples are always rare for research due to many reasons. For example, the patients might not be willing to provide samples for study, or the sample quality might not be suitable for sequencing, or the storage of samples cannot perform well, etc. The samples collected from patients usually have a limited cell amount, which makes it difficult to carry out integrated analysis in many aspects. In our case, the Hi-C, ChIP-Seq, and RNA-Seq integrated analysis could only be carried out in total bone marrow samples without CD34+ sorting due to the limited number of cells. Also, despite these efforts, we were still left with several samples that lacked RNA-Seq or ChIP-Seq.

In addition, due to the heterogeneity of AML, and the limited sample numbers, it is difficult to research specific mutations. For example, for our study, when we planned to study *DNMT3A* mutation, we failed to acquire a *DNMT3A* mutated clinical sample, even though we obtained a total of eight AML clinical samples for testing. Thus, we could only mimic the mutation in the K562 cell line.

Clinical research is important for therapeutic advances, and we hope that with further improvements in terms of applying integrated epigenetic and RNA-

Seq methods to clinical samples, more of these limitations can be overcome.

### 5.2.3 Limitations of Different Cell Lines as the Study Model of *MEIS1* FIRE loss as well as *DNMT3A* loss

When we tried to seek out an alternative way to study the *MEIS1* FIRE and *DNMT3A* mutation, we decided to use cell lines for our analyses due to the limitations of clinical research. We considered leukemia cell lines including HL-60, THP-1, and K562.

HL-60 is an Acute promyelocytic leukemia cell line, but unfortunately, it does not appear likely to have the *MEIS1* FIRE (**Figure 3.5**). Thus, it is not suitable for the *MEIS1* FIRE CRISPR study. Although THP-1 has a weak *MEIS1* FIRE, it is not amenable for CRISPR studies, because THP-1 cells cannot tolerate growth in single-cell colonies and need the proximity of other cells to grow. HL-60 also has this problem. According to several tries by my colleagues and other researchers in Cancer Science Institute, THP-1 and HL-60 usually die quickly after CRISPR transfection and single-cell sorting. Thus, we finally chose K562.

However, K562 is a Chronic Myelogenous Leukemia (CML) cell line. Even though K562 is a leukemia cell line and also of the myeloid lineage, we do not know whether K562 will reflect the same effects in AML or not. Thus, further studies need to be applied to verify the discoveries in our works.

### 5.2.4 Unsolved Problems in *MEIS1* FIRE Region

For the *MEIS1* FIRE region, we gave a proposed mechanism on how the FIRE modulates *MEIS1* expression and figured out that two subtypes (with and without FIRE) might exist in AML. But there are still some questions that need to be answered.

First, why has the FIRE disappeared in some of the AML patients? Based on the current understanding that CTCF is essential for maintaining chromatin interactions (Hansen et al., 2017; Phillips & Corces, 2009; Rao et al., 2014), and the finding that CTCF is anticorrelated with DNA methylation (Phillips & Corces, 2009; H. Wang et al., 2012), one possible answer might be the DNA methylation level change caused CTCF loss of FIRE. To confirm this answer, bisulfite sequencing is needed to test the DNA methylation levels, as well as CTCF ChIP-Seq to investigate whether CTCF binding is lost.

Another question is, why is there stronger H3K27Ac enhancer intensity in AML? Maybe some epigenetic factors which control H3K27Ac levels have been altered in AML. We need to explore the changes in epigenetic factors through RNA-Seq. Changes in DNA methylation level may also be the reason, as H3K27Ac is found to be strongly anticorrelated with DNA methylation (Kundaje et al., 2015). Again, to test this assumption, bisulfite sequencing is needed.

### 5.2.5 *MEIS1* and *HOXA9* Co-expression

*MEIS1* and *HOXA9* co-expression has been widely studied for years in leukemia, and aberrant co-expression will induce AML (Thorsteinsdottir, Kroon,

Jerome, Blasi, & Sauvageau, 2001). During our research on *MEIS1* FIRE, I have also engaged in the bioinformatics analyses of research in the relations between *MEIS1* FIRE and *HOXA9*. Since the discovery is not related to the thesis hypothesis, and most of the work has done by my colleague Dr. Benny Wang Zhengjie, I did not put this section into the thesis. But it is still an interesting topic to discuss (B. Wang et al., 2021).

We found that a super-enhancer heterogeneously presented in AML patients which interacted with *HOXA9* promoter by a chromatin loop. ChIP-qPCR experiments in THP-1 show that MEIS1 protein can bind to *MEIS1* promoter, *HOXA9* promoter, *MYC* promoter, and this super-enhancer region. HOXA9 protein can bind to these regions as well. This might suggest that *MEIS1* and *HOXA9* co-express by binding to each promoter and enhancer to regulate their expressions, which forms a positive feedback loop. Once one of them is overexpressed due to the appearance of super-enhancers, their expression will be highly elevated in myeloid leukemia. Loss of *MEIS1* FIRE will largely reduce both of their expressions. We have proposed a schematic of mechanisms in **Figure 5.1**.

In this part, the remaining question is: why is the super-enhancer heterogeneously present in *HOXA9*? We inferred that super-enhancer loss may be because of the *MEIS1* FIRE loss, but more investigations should be carried out to confirm our assumptions and to explain why the *MEIS1* FIRE can control the super-enhancer.

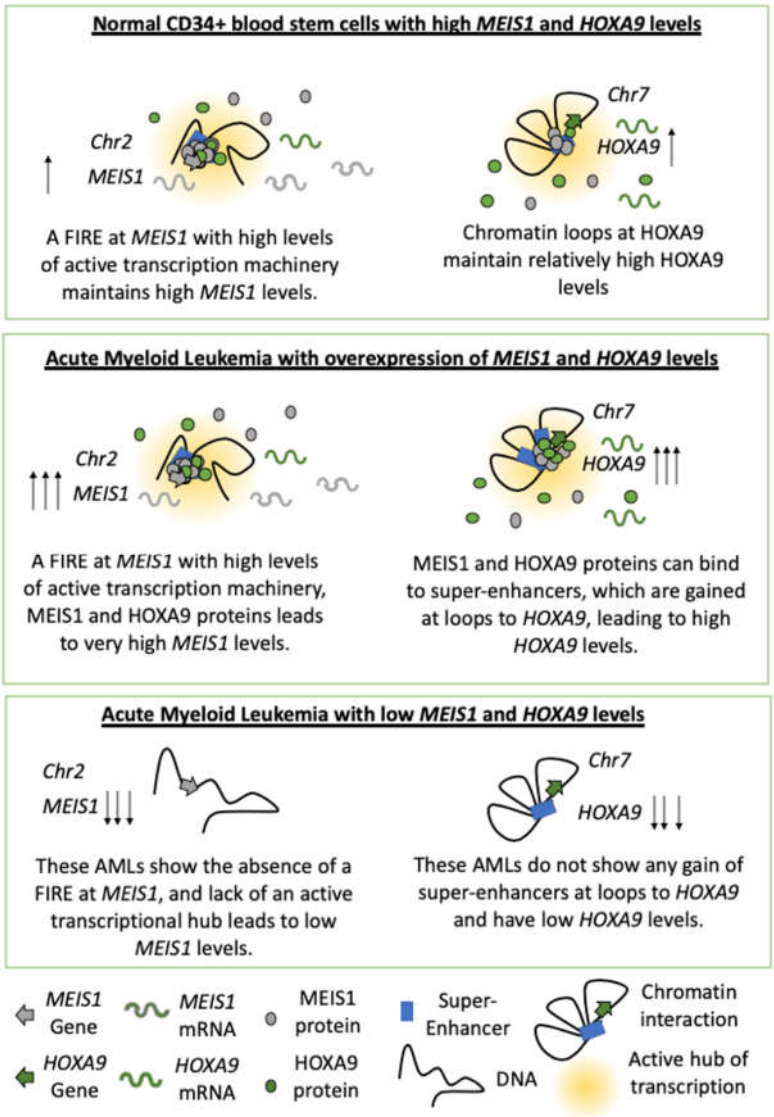



**Figure 5.1 Proposed schematic of the mechanisms leading to the heterogeneous expression of *MEIS1* and *HOXA9* in different sub-types of AML** *(Note: This figure was produced by my supervisor Dr. Melissa Jane Fullwood)*

### 5.2.6 The Relationships Between *DNMT3A*, Histone Marks and CTCF Binding

We have observed altered histone modification and CTCF binding profile in *DNMT3A* knock-out K562 cells, raising several questions. For instance, how does *DNMT3A* influence the histone marks? Previous works found that *DNMT3A* and *TET1* can help the depositions of H3K27Me3 in bivalent regions (Gu et al., 2018), but in our case, when *DNMT3A* was knocked out, we found that an H3K27Me3 is gained to form a bivalent region in the *PLOD2* region (**Figure 4.4**). One study showed that CTCF binding can remove H3K27Me3 (Weth et al., 2014), which might be the reason why the H3K27Me3 mark was gained in this region as the CTCF binding was lost there.

And for the CTCF binding, the three altered chromatin interaction regions all presented CTCF binding loss. *DNMT3A* knockout leads to mainly reduced DNA methylation, which as we inferred, should result in more CTCF binding. The possible explanation for this might be that *DNMT3A* does not function by itself only and other DNA methylases - *DNMT1*, *DNMT3B*, *DNMT3L* also play an important role in DNA methylation. The question of how these epigenetic factors change or have compensatory roles after *DNMT3A* knockout remains to be explored.

### *5.2.7* The Role of *PLOD2* and *ARID5B* in Leukemia

During our K562 *DNMT3A* study, we found three interesting regions:

*PLOD2*, *MACC1,* and *ARID5B*. These three genes are all found to be related to cancer previously (Emerenciano et al., 2014; Ge et al., 2015; Gjaltema et al., 2015; Reyes-León et al., 2019; Stein et al., 2009), while interestingly we found *PLOD2* overexpressed in glioma, cervical and liver cancer, which indicate overexpression of *PLOD2* might facilitate cancer formation and/or progression.

However, in our study, *PLOD2* showed overexpression in a few *DNMT3A* wild-type clinical samples, while the level of *PLOD2* in *DNMT3A* mutant remains low. We do not know what is the impact of down regulation of *PLOD2* on leukemia progression, but if *DNMT3A* mutation can control the *PLOD2* expression by influencing the *PLOD2* FIRE, and if further work indicates that downregulation of *PLOD2* is therapeutically valuable for leukemia patients, then this consequence of the *DNMT3A* mutation might be beneficial for leukemia patients. If we can study more on this mechanism, new drugs can be designed to mimic *DNMT3A* influences on *PLOD2* FIRE to treat *PLOD2* overexpressed cancers. Alternatively, existing drugs could be repurposed. For example, decitabine and azacytidine are drugs used to treat Acute Myeloid Leukemia patients and these drugs work by inhibiting DNA methyltransferases (Kantarjian et al., 2012).

Another interesting gene is the *ARID5B*, like *DNMT3A*, *ARID5B* can also modulate other gene expression levels (P. Wang et al., 2020), and it can influence chromatin structure (Gregory et al., 1996; Herrscher et al., 1995), we inferred that *ARID5B* may be another epigenetic factor under the control of *DNMT3A*, which will cause further indirect effect of *DNMT3A* loss. This might be a reason why *DNMT3A* mutant AML usually shows a poor outcome. To address this assumption, more analysis between *DNMT3A*, *ARID5B* and other gene

expressions needs to be done.

### 5.2.8 Possible Therapeutic Strategies Suggested from Our Works

Through our works in AML clinical samples and *DNMT3A* knock-out K562 cells, several possible therapeutic ways can be suggested. First, as we conclude that chromatin interactions tend to be altered in AML and *DNMT3A* mutated AML, it would be useful to develop strategies to target chromatin interactions.

In our *DNMT3A* knock-out cells, we found the histone modification profile changed. In addition to using DNA methylation drugs to target AML(Contieri, Duarte, & Lazarini, 2020), our discovery may suggest that drugs that target histone modifications might also be useful. (give some examples of drugs that target histone modifications)

Last, as we inferred that *ARID5B* might be controlled by *DNMT3A* and may cause further indirect effect of *DNMT3A* loss if this assumption can be confirmed, drugs targeting *ARID5B* might also be another treatment strategy.

## 5.3 Future Directions

### 5.3.1 Our Future Plans

We hypothesized that *DNMT3A* might control the chromatin interactions by controlling the DNA methylation level. We also thought that CTCF binding

alterations caused by DNA methylation change will be the factor of dysregulated chromatin interaction. However, we lack a DNA methylation profile of *DNMT3A* loss samples. Thus, in future work, bisulfite sequencing will be used to fill in this piece of the puzzle.

Another work that needs to be improved is the ChIP-Seq analysis. We noticed that some of the ChIP-Seq with an unbalanced sequencing depth (e.g., CTCF and H3K27Me3), and we will try to balance them to get a more accurate comparison between KO and Vec_Con.

We are also planning to establish some software from our used scripts in our study to help other researchers to reduce the bioinformatics work. The initial plan about this includes: (1) Establish a software of gene correlation analysis to indicate boundary dysregulation, and (2) Establish a software of loop comparison, which can figure out specific loops associated genes and further overlap with differential expressed genes and enhancer/silencer altered genes.

In the future, we are also interested to investigate epigenetic drugs in Acute Myeloid Leukemia for their impact on chromatin interactions, for example, DOT1L inhibitors. DOT1Li treated THP-1 cells have already been sequenced for Hi-C and RNA-Seq, and further analysis will be applied to seek out the impact of DOT1Li in myeloid leukemia.

### 5.3.2 Suggestions for Future Research

As we discussed in section 5.2.1, there are plenty of algorithms that can predict the chromatin interactions, but the poor accuracy, the ambiguous

statistical standard of TAD, and the hardware requirement hinder chromatin interaction analyses. We hope that in the future, we will have a better understanding of 3D genome organization, and a better sense of how a TAD should be statistically termed. With these improvements, we can optimize the algorithms of TAD/loop call.

In terms of clinical research, we hope that more resources of a clinical database of Hi-C can be established. We now have the TCGA database as a cancer clinical sample data resource, which includes whole genome sequencing and RNA profile, but no epigenetic database including Hi-C data and ChIP-Seq of cancer clinical samples has been established. By establishing such a database in the future, cancer clinical research will become easier.

We noticed that current drugs for cancer treatment are mainly influenced by methylations, histone marks, or targeting a specific gene. We suggest that in the future, we can try to develop drugs that target chromatin interactions. In addition, we can investigate how existing drugs affect chromatin interactions.

We also have some suggestions based on the problems we found in our study. As we found that our ChIP-Seq seems to have an unbalanced sequencing depth, but if the ChIP enriched signal differs a lot in the original sample, it is hard to clarify what kind of reason may have caused this unbalance. Thus, we suggest using the spike-in strategy (Egan et al., 2016) when quantificational comparisons between groups are needed.

## 5.4 Overview

Taken together, our work provided a better understanding of chromatin interactions alterations and gene expression changes in AML and *DNMT3A* mutant myeloid leukemia. Our research indicates the relevance of chromatin interactions in cancer biology, and suggests that drugs that modulate epigenetic, such as DNA methylation, may lead to changes in chromatin interactions. In future research, we are interested to develop therapeutic strategies for altering the dysregulated chromatin interactions seen in AML through epigenetic drugs.

# Bibliography


(ACS), A. C. S. (2020, September 3, 2020). Typical Treatment of Acute Myeloid Leukemia (Except APL). Retrieved from https://www.cancer.org/cancer/acute-myeloid-leukemia/treating/typical-treatment-of-aml.html

(URMC), U. o. R. M. C. (2021). Acute Myeloid Leukemia (AML): Chemotherapy Retrieved from https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=34&contentid=BAMLT3

Abdel-Wahab, O., & Levine, R. L. (2013). Mutations in epigenetic modifiers in the pathogenesis and therapy of acute myeloid leukemia. *Blood, 121*(18), 3563-3572. doi:10.1182/blood-2013-01-451781

Achinger-Kawecka, J., Taberlay, P. C., & Clark, S. J. (2016). Alterations in Three-Dimensional Organization of the Cancer Genome and Epigenome. *Cold Spring Harb Symp Quant Biol, 81*, 41-51. doi:10.1101/sqb.2016.81.031013

An, O., Tan, K.-T., Li, Y., Li, J., Wu, C.-S., Zhang, B., . . . Yang, H. (2020). CSI NGS Portal: An Online Platform for Automated NGS Data Analysis and Sharing. *International Journal of Molecular Sciences, 21*(11). doi:10.3390/ijms21113828

Andreeff, M., Ruvolo, V., Gadgil, S., Zeng, C., Coombes, K., Chen, W., . . . Drabkin, H. (2008). HOX expression patterns identify a common signature for favorable AML. *Leukemia, 22*(11), 2041-2047.

Andricovich, J., Perkail, S., Kai, Y., Casasanta, N., Peng, W., & Tzatsos, A. (2018). Loss of KDM6A Activates Super-Enhancers to Induce Gender-Specific Squamous-like Pancreatic Cancer and Confers Sensitivity to BET Inhibitors. *Cancer Cell, 33*(3), 512-526.e518. doi:10.1016/j.ccell.2018.02.003

Ay, F., Bailey, T. L., & Noble, W. S. (2014). Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res, 24*(6), 999-1011. doi:10.1101/gr.160374.113

Babu, D., & Fullwood, M. J. (2015). 3D genome organization in health and disease: emerging opportunities in cancer translational medicine. *Nucleus, 6*(5), 382-393. doi:10.1080/19491034.2015.1106676

Banno, K., Kisu, I., Yanokura, M., Tsuji, K., Masuda, K., Ueki, A., . . . Aoki, D. (2012). Epimutation and cancer: a new carcinogenic mechanism of Lynch syndrome (Review). *Int J Oncol, 41*(3), 793-797. doi:10.3892/ijo.2012.1528

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

Baum, L. E. (1972). An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of a Markov Process. *Inequalities, 3*, 1-8.

Bessa, J., Tavares, M. J., Santos, J., Kikuta, H., Laplante, M., Becker, T. S., . . . Casares, F. (2008). meis1 regulates cyclin D1 and c-myc expression, and controls the proliferation of the multipotent cells in the early developing zebrafish eye. *Development, 135*(5), 799-803.

Bhasin, M., Reinherz, E. L., & Reche, P. A. (2006). Recognition and classification of histones using support vector machine. *J Comput Biol, 13*(1), 102-112. doi:10.1089/cmb.2006.13.102

Bishop, M. J., & Thompson, E. A. (1986). Maximum likelihood alignment of DNA sequences. *J Mol Biol, 190*(2), 159-165. doi:10.1016/0022-2836(86)90289-5

Bonnet, D., & Dick, J. E. (1997). Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat Med, 3*(7), 730-737. doi:10.1038/nm0797-730

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., & Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol, 31*(12), 1119-1125. doi:10.1038/nbt.2727

Cao, F., Fang, Y., Tan, H. K., Goh, Y., Choy, J. Y. H., Koh, B. T. H., . . . Fullwood, M. J. (2017). Super-Enhancers and Broad H3K4me3 Domains Form Complex Gene Regulatory Circuits Involving Chromatin Interactions. *Scientific Reports, 7*(1), 2186. doi:10.1038/s41598-017-02257-3

Chaudry, S. F., & Chevassut, T. J. T. (2017). Epigenetic Guardian: A Review of the DNA Methyltransferase DNMT3A in Acute Myeloid Leukaemia and Clonal Haematopoiesis. *BioMed Research International, 2017*, 5473197. doi:10.1155/2017/5473197

Chen, T., Ueda, Y., Xie, S., & Li, E. (2002). A novel Dnmt3a isoform produced from an alternative promoter localizes to euchromatin and its expression correlates with active de novo methylation. *J Biol Chem, 277*(41), 38746-38754. doi:10.1074/jbc.M205312200

Collins, C., Wang, J., Miao, H., Bronstein, J., Nawer, H., Xu, T., . . . Hess, J. L. (2014). C/EBPα is an essential collaborator in Hoxa9/Meis1-mediated leukemogenesis. *Proceedings of the National Academy of Sciences, 111*(27), 9899-9904.

Collins, C. T., & Hess, J. L. (2016). Deregulation of the HOXA9/MEIS1 axis in acute leukemia. *Current opinion in hematology, 23*(4), 354.

Consortium, E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57-74. doi:10.1038/nature11247

Contieri, B., Duarte, B. K. L., & Lazarini, M. (2020). Updates on DNA methylation modifiers in acute myeloid leukemia. *Ann Hematol, 99*(4), 693-701. doi:10.1007/s00277-020-03938-2

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., . . . Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America, 107*(50), 21931-21936. doi:10.1073/pnas.1016071107

Crowley, C., Yang, Y., Qiu, Y., Hu, B., Abnousi, A., Lipiński, J., . . . Li, Y. (2021). FIREcaller: Detecting frequently interacting regions from Hi-C data. *Computational and Structural Biotechnology Journal, 19*, 355-362. doi:https://doi.org/10.1016/j.csbj.2020.12.026

Dahl, J. A., Jung, I., Aanes, H., Greggains, G. D., Manaf, A., Lerdrup, M., . . . Klungland, A. (2016). Broad histone H3K4me3 domains in mouse oocytes modulate maternal-to-zygotic transition. *Nature, 537*(7621), 548-552. doi:10.1038/nature19360

Dali, R., & Blanchette, M. (2017). A critical assessment of topologically associating domain prediction tools. *Nucleic acids research, 45*(6), 2994-3005. doi:10.1093/nar/gkx145

de Laat, W., & Duboule, D. (2013). Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature, 502*(7472), 499-506. doi:10.1038/nature12753

de Wit, E., & de Laat, W. (2012). A decade of 3C technologies: insights into nuclear organization. *Genes Dev, 26*(1), 11-24. doi:10.1101/gad.179804.111

Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *Science, 295*(5558), 1306-1311. doi:10.1126/science.1067799

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., . . . Ren, B. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature, 518*(7539), 331-336. doi:10.1038/nature14222

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., . . . Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature, 485*(7398), 376-380. doi:10.1038/nature11082

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15-21.

Dohner, H., Weisdorf, D. J., & Bloomfield, C. D. (2015). Acute Myeloid Leukemia. *N Engl J Med, 373*(12), 1136-1152. doi:10.1056/NEJMra1406184

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., . . . Dekker, J. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res, 16*(10), 1299-1309. doi:10.1101/gr.5571506

Dupont, C., Armant, D. R., & Brenner, C. A. (2009). Epigenetics: definition, mechanisms and clinical perspective. *Semin Reprod Med, 27*(5), 351-357. doi:10.1055/s-0029-1237423

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., & Aiden, E. L. (2016). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst, 3*(1), 99-101. doi:10.1016/j.cels.2015.07.012

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell systems, 3*(1), 95-98.

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst, 3*(1), 95-98. doi:10.1016/j.cels.2016.07.002

Egan, B., Yuan, C. C., Craske, M. L., Labhart, P., Guler, G. D., Arnott, D., . . . Trojer, P. (2016). An Alternative Approach to ChIP-Seq Normalization Enables Detection of Genome-Wide Changes in Histone H3 Lysine 27 Trimethylation upon EZH2 Inhibition. *PLoS One, 11*(11), e0166438. doi:10.1371/journal.pone.0166438

Emerenciano, M., Barbosa, T. C., Lopes, B. A., Blunck, C. B., Faro, A., Andrade, C., . . . The Brazilian Collaborative Study Group of Infant Acute, L. (2014). ARID5B polymorphism confers an increased risk to acquire specific MLL rearrangements in early childhood leukemia. *BMC Cancer, 14*(1), 127. doi:10.1186/1471-2407-14-127

Ferlay J, E. M., Lam F, Colombet M, Mery L, Piñeros M, et al. . (2020). Global Cancer Observatory: Cancer Today. Lyon: International Agency for Research on Cancer. Retrieved from https://gco.iarc.fr/today

Ferrara, F., & Schiffer, C. A. (2013). Acute myeloid leukaemia in adults. *Lancet, 381*(9865), 484-495. doi:10.1016/s0140-6736(12)61727-9

Fialkow, P. J. (1976). Clonal origin of human tumors. *Biochim Biophys Acta, 458*(3), 283-321.

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

Flavahan, W. A., Drier, Y., Liau, B. B., Gillespie, S. M., Venteicher, A. S., Stemmer-Rachamimov, A. O., . . . Bernstein, B. E. (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature, 529*(7584), 110-114. doi:10.1038/nature16490

Flavahan, W. A., Drier, Y., Liau, B. B., Gillespie, S. M., Venteicher, A. S., Stemmer-Rachamimov, A. O., . . . Bernstein, B. E. (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature, 529*(7584), 110-114. doi:10.1038/nature16490

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

Fortin, J.-P., & Hansen, K. D. (2015). Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome biology, 16*(1), 180-180. doi:10.1186/s13059-015-0741-y

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., . . . Ruan, Y. (2009). An oestrogen-receptor-α-bound human chromatin interactome. *Nature, 462*(7269), 58-64. doi:10.1038/nature08497

Gao, L., Sun, J., Liu, F., Zhang, H., & Ma, Y. (2016). Higher expression levels of the HOXA9 gene, closely associated with MLL-PTD and EZH2 mutations, predict inferior outcome in acute myeloid leukemia. *OncoTargets and therapy, 9*, 711.

Garimberti, E., & Tosi, S. (2010). Fluorescence in situ hybridization (FISH), basic principles and methodology. *Methods Mol Biol, 659*, 3-20. doi:10.1007/978-1-60761-789-1_1

Ge, Y., Meng, X., Zhou, Y., Zhang, J., & Ding, Y. (2015). Positive MACC1 expression correlates with invasive behaviors and postoperative liver metastasis in colon cancer. *International journal of clinical and experimental medicine, 8*(1), 1094-1100.

Gistelinck, C., Witten, P. E., Huysseune, A., Symoens, S., Malfait, F., Larionova, D., . . . Coucke, P. J. (2016). Loss of Type I Collagen Telopeptide Lysyl Hydroxylation Causes Musculoskeletal Abnormalities in a Zebrafish Model of Bruck Syndrome. *Journal of Bone and Mineral Research, 31*(11), 1930-1942. doi:https://doi.org/10.1002/jbmr.2977

Gjaltema, R. A. F., de Rond, S., Rots, M. G., & Bank, R. A. (2015). Procollagen Lysyl Hydroxylase 2 Expression Is Regulated by an Alternative Downstream Transforming Growth Factor β-1 Activation Mechanism. *J Biol Chem, 290*(47), 28465-28476. doi:10.1074/jbc.M114.634311

Gregory, S. L., Kortschak, R. D., Kalionis, B., & Saint, R. (1996). Characterization of the dead ringer gene identifies a novel, highly conserved family of sequence-specific DNA-binding proteins. *Mol Cell Biol, 16*(3), 792-799. doi:10.1128/mcb.16.3.792

Grimmer, M. R., & Costello, J. F. (2016). Cancer: Oncogene brought into the loop. *Nature, 529*(7584), 34-35. doi:10.1038/nature16330

Gu, T., Lin, X., Cullen, S. M., Luo, M., Jeong, M., Estecio, M., . . . Goodell, M. A. (2018). DNMT3A and TET1 cooperate to regulate promoter epigenetic landscapes in mouse embryonic stem cells. *Genome biology, 19*(1), 88. doi:10.1186/s13059-018-1464-7

Guillamot, M., Cimmino, L., & Aifantis, I. (2016). The Impact of DNA Methylation in Hematopoietic Malignancies. *Trends in cancer, 2*(2), 70-83. doi:10.1016/j.trecan.2015.12.006

Handy, D. E., Castro, R., & Loscalzo, J. (2011). Epigenetic modifications: basic mechanisms and role in cardiovascular disease. *Circulation, 123*(19), 2145-2156. doi:10.1161/circulationaha.110.956839

Hansen, A. S., Pustova, I., Cattoglio, C., Tjian, R., & Darzacq, X. (2017). CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *eLife, 6*, e25776. doi:10.7554/eLife.25776

Harewood, L., Kishore, K., Eldridge, M. D., Wingett, S., Pearson, D., Schoenfelder, S., . . . Fraser, P. (2017). Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *18*(1), 125. doi:10.1186/s13059-017-1253-8

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., . . . Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell, 38*(4), 576-589. doi:10.1016/j.molcel.2010.05.004

Herrscher, R. F., Kaplan, M. H., Lelsz, D. L., Das, C., Scheuermann, R., & Tucker, P. W. (1995). The immunoglobulin heavy-chain matrix-associating regions are bound by Bright: a B cell-specific trans-activator that describes a new DNA-binding protein family. *Genes Dev, 9*(24), 3067-3082. doi:10.1101/gad.9.24.3067

Heyn, P., Logan, C. V., Fluteau, A., Challis, R. C., Auchynnikava, T., Martin, C. A., . . . Jackson, A. P. (2019). Gain-of-function DNMT3A mutations cause microcephalic dwarfism and hypermethylation of Polycomb-regulated regions. *Nat Genet, 51*(1), 96-105. doi:10.1038/s41588-018-0274-x

Holz-Schietinger, C., Matje, D. M., & Reich, N. O. (2012). Mutations in DNA methyltransferase (DNMT3A) observed in acute myeloid leukemia patients disrupt processive methylation. *J Biol Chem, 287*(37), 30941-30951. doi:10.1074/jbc.M112.366625

Horibata, S., Gui, G., Lack, J., DeStefano, C. B., Gottesman, M. M., & Hourigan, C. S. (2019). Heterogeneity in refractory acute myeloid leukemia. *Proceedings of the National Academy of Sciences, 116*(21), 10494-10503. doi:10.1073/pnas.1902375116

Horike, S., Cai, S., Miyano, M., Cheng, J. F., & Kohwi-Shigematsu, T. (2005). Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome. *Nat Genet, 37*(1), 31-40. doi:10.1038/ng1491

Hou, H. A., Kuo, Y. Y., Liu, C. Y., Chou, W. C., Lee, M. C., Chen, C. Y., . . . Tien, H. F. (2012). DNMT3A mutations in acute myeloid leukemia: stability during disease evolution and clinical implications. *Blood, 119*(2), 559-568. doi:10.1182/blood-2011-07-369934

Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B., & Liu, J. S. (2012). HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics, 28*(23), 3131-3133. doi:10.1093/bioinformatics/bts570

Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., . . . Mirny, L. A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods, 9*(10), 999-1003. doi:10.1038/nmeth.2148

Jasperson, K. W., Tuohy, T. M., Neklason, D. W., & Burt, R. W. (2010). Hereditary and familial colon cancer. *Gastroenterology, 138*(6), 2044-2058. doi:10.1053/j.gastro.2010.01.054

Jeong, M., Park, H. J., Celik, H., Ostrander, E. L., Reyes, J. M., Guzman, A., . . . Challen, G. A. (2018). Loss of Dnmt3a Immortalizes Hematopoietic Stem Cells In Vivo. *Cell Rep, 23*(1), 1-10. doi:10.1016/j.celrep.2018.03.025

Jia, D., Jurkowska, R. Z., Zhang, X., Jeltsch, A., & Cheng, X. (2007). Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature, 449*(7159), 248-251. doi:10.1038/nature06146

Jin, B., Li, Y., & Robertson, K. D. (2011). DNA methylation: superior or subordinate in the epigenetic hierarchy? *Genes & cancer, 2*(6), 607-617. doi:10.1177/1947601910393957

Johnson, D. A., Barclay, R. L., Mergener, K., Weiss, G., König, T., Beck, J., & Potter, N. T. (2014). Plasma Septin9 versus fecal immunochemical testing for colorectal cancer screening: a prospective multicenter study. *PLoS One, 9*(6), e98238. doi:10.1371/journal.pone.0098238

Kantarjian, H. M., Thomas, X. G., Dmoszynska, A., Wierzbowska, A., Mazur, G., Mayer, J., . . . Arthur, C. (2012). Multicenter, randomized, open-label, phase III trial of decitabine versus patient choice, with physician advice, of either supportive care or low-dose cytarabine for the treatment of older patients with newly diagnosed acute myeloid leukemia. *J Clin Oncol, 30*(21), 2670-2677. doi:10.1200/jco.2011.38.9429

Kempfer, R., & Pombo, A. (2020). Methods for mapping 3D chromosome architecture. *Nature Reviews Genetics, 21*(4), 207-226. doi:10.1038/s41576-019-0195-2

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Res, 12*(6), 996-1006. doi:10.1101/gr.229102

Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., & Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics, 26*(17), 2204-2207. doi:10.1093/bioinformatics/btq351

Kieffer-Kwon, K. R., Nimura, K., Rao, S. S. P., Xu, J., Jung, S., Pekowska, A., . . . Casellas, R. (2017). Myc Regulates Chromatin Decompaction and Nuclear Architecture during B Cell Activation. *Molecular cell, 67*(4), 566-578.e510. doi:10.1016/j.molcel.2017.07.013

Kloetgen, A., Thandapani, P., Ntziachristos, P., Ghebrechristos, Y., Nomikou, S., Lazaris, C., . . . Tsirigos, A. (2020). Three-dimensional chromatin landscapes in T cell acute lymphoblastic leukemia. *Nat Genet, 52*(4), 388-400. doi:10.1038/s41588-020-0602-9

Knight, P. A., & Ruiz, D. (2013). A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis, 33*(3), 1029-1047. doi:10.1093/imanum/drs019

Koch, C. M., Andrews, R. M., Flicek, P., Dillon, S. C., Karaöz, U., Clelland, G. K., . . . Dunham, I. (2007). The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res, 17*(6), 691-707. doi:10.1101/gr.5704207

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., . . . Principal, i. (2015). Integrative analysis of 111 reference human epigenomes. *Nature, 518*(7539), 317-330. doi:10.1038/nature14248

Kundu, M., Chen, A., Anderson, S., Kirby, M., Xu, L., Castilla, L. H., . . . Liu, P. P. (2002). Role of Cbfb in hematopoiesis and perturbations resulting from expression of the leukemogenic fusion gene Cbfb-MYH11. *Blood, 100*(7), 2449-2456. doi:10.1182/blood-2002-04-1064

Lai, A. Y., Fatemi, M., Dhasarathy, A., Malone, C., Sobol, S. E., Geigerman, C., . . . Wade, P. A. (2010). DNA methylation prevents CTCF-mediated silencing of the oncogene BCL6 in B cell lymphomas. *J Exp Med, 207*(9), 1939-1950. doi:10.1084/jem.20100204

Lajoie, B. R., Dekker, J., & Kaplan, N. (2015). The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods (San Diego, Calif.), 72*, 65-75. doi:10.1016/j.ymeth.2014.10.031

Langer-Safer, P. R., Levine, M., & Ward, D. C. (1982). Immunological method for mapping genes on Drosophila polytene chromosomes. *Proceedings of the National Academy of Sciences of the United States of America, 79*(14), 4381-4385. doi:10.1073/pnas.79.14.4381

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*(4), 357-359. doi:10.1038/nmeth.1923

Leonard E. Baum, G. R. S. (1968). Growth transformations for functions on manifolds. *Pacific Journal of Mathematics, 27*(2), 211-227. doi:10.2140/pjm.1968.27.211

Leonard E. Baum, J. A. E. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society, 73*(3), 360. doi:10.1090/S0002-9904-1967-11751-8

Leonard E. Baum, T. P. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics., 37*(6), 1554-1563. doi:10.1214/aoms/1177699147

Leonard E. Baum, T. P., George Soules, Norman Weiss. (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics., 41*(1), 164-171. doi:10.1214/aoms/1177697196

Lévy-Leduc, C., Delattre, M., Mary-Huard, T., & Robin, S. (2014). Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics, 30*(17), i386-392. doi:10.1093/bioinformatics/btu443

Ley, T. J., Ding, L., Walter, M. J., McLellan, M. D., Lamprecht, T., Larson, D. E., . . . Wilson, R. K. (2010). DNMT3A mutations in acute myeloid leukemia. *N Engl J Med, 363*(25), 2424-2433. doi:10.1056/NEJMoa1005143

Ley, T. J., Miller, C., Ding, L., Raphael, B. J., Mungall, A. J., Robertson, A., . . . Eley, G. (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med, 368*(22), 2059-2074. doi:10.1056/NEJMoa1301689

Li, A., Yin, X., Xu, B., Wang, D., Han, J., Wei, Y., . . . Zhang, Z. (2018). Decoding topologically associating domains with ultra-low resolution Hi-C data by graph structural entropy. *Nature Communications, 9*(1), 3265. doi:10.1038/s41467-018-05691-7

Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., . . . Ruan, Y. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell, 148*(1-2), 84-98. doi:10.1016/j.cell.2011.12.014

Li, S., Mason, C. E., & Melnick, A. (2016). Genetic and epigenetic heterogeneity in acute myeloid leukemia. *Curr Opin Genet Dev, 36*, 100-106. doi:10.1016/j.gde.2016.03.011

Li, Y., He, Y., Liang, Z., Wang, Y., Chen, F., Djekidel, M. N., . . . Chen, Y. (2018). Alterations of specific chromatin conformation affect ATRA-induced leukemia cell differentiation. *Cell Death & Disease, 9*(2), 200. doi:10.1038/s41419-017-0173-6

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., . . . Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science, 326*(5950), 289-293. doi:10.1126/science.1181369

Lin, J., Yao, D. M., Qian, J., Chen, Q., Qian, W., Li, Y., . . . Xu, W. R. (2011). Recurrent DNMT3A R882 mutations in Chinese patients with acute myeloid leukemia and myelodysplastic syndrome. *PLoS One, 6*(10), e26906. doi:10.1371/journal.pone.0026906

Lun, A. T., & Smyth, G. K. (2015). diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics, 16*, 258. doi:10.1186/s12859-015-0683-0

Matsuo, H., Iijima-Yamashita, Y., Yamada, M., Deguchi, T., Kiyokawa, N., Shimada, A., . . . Horibe, K. (2018). Monitoring of fusion gene transcripts to predict relapse in pediatric acute myeloid leukemia. *Pediatr Int, 60*(1), 41-46. doi:10.1111/ped.13440

McKeown, M. R., Corces, M. R., Eaton, M. L., Fiore, C., Lee, E., Lopez, J. T., . . . Koenig, J. L. (2017). Superenhancer analysis defines novel epigenomic subtypes of non-APL AML, including an RARα dependency targetable by SY-1425, a potent and selective RARα agonist. *Cancer discovery, 7*(10), 1136-1153.

Mohaghegh, N., Bray, D., Keenan, J., Penvose, A., Andrilenas, K. K., Ramlall, V., & Siggers, T. (2019). NextPBM: a platform to study cell-specific transcription factor binding and cooperativity. *Nucleic acids research, 47*(6), e31. doi:10.1093/nar/gkz020

Mohr, S., Doebele, C., Comoglio, F., Berg, T., Beck, J., Bohnenberger, H., . . . Wachter, A. (2017). Hoxa9 and Meis1 cooperatively induce addiction to

Syk signaling by suppressing miR-146a in acute myeloid leukemia. *Cancer cell, 31*(4), 549-562. e511.
Mumbach, M. R., Rubin, A. J., Flynn, R. A., Dai, C., Khavari, P. A., Greenleaf, W. J., & Chang, H. Y. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nature methods, 13*(11), 919-922. doi:10.1038/nmeth.3999
Navarro Gonzalez, J., Zweig, A. S., Speir, M. L., Schmelter, D., Rosenbloom, K. R., Raney, B. J., . . . Kent, W. J. (2021). The UCSC Genome Browser database: 2021 update. *Nucleic acids research, 49*(D1), D1046-d1057. doi:10.1093/nar/gkaa1070
NIH. (2020a, March 11, 2020). Adult Acute Lymphoblastic Leukemia Treatment (PDQ®)–Patient Version. Retrieved from https://www.cancer.gov/types/leukemia/patient/adult-all-treatment-pdq
NIH. (2020b, March 6, 2020). Adult Acute Myeloid Leukemia Treatment (PDQ®)–Patient Version. Retrieved from https://www.cancer.gov/types/leukemia/patient/adult-aml-treatment-pdq#section/all
NIH. (2021, May 5, 2021). What is Cancer. Retrieved from https://www.cancer.gov/about-cancer/understanding/what-is-cancer
Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., . . . Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature, 485*(7398), 381-385. doi:10.1038/nature11049
Norton, H. K., & Phillips-Cremins, J. E. (2017). Crossed wires: 3D genome misfolding in human disease. *Journal of Cell Biology, 216*(11), 3441-3452. doi:10.1083/jcb.201611001
Novak, K. (2004). Epigenetics changes in cancer cells. *MedGenMed, 6*(4), 17.
O'Brien, E. C., Brewin, J., & Chevassut, T. (2014). DNMT3A: the DioNysian MonsTer of acute myeloid leukaemia. *Therapeutic advances in hematology, 5*(6), 187-196. doi:10.1177/2040620714554538
Peng, Y., Xiong, D., Zhao, L., Ouyang, W., Wang, S., Sun, J., . . . Li, X. (2019). Chromatin interaction maps reveal genetic regulation for quantitative traits in maize. *Nature Communications, 10*(1), 2632. doi:10.1038/s41467-019-10602-5
Phanstiel, D. H., Van Bortle, K., Spacek, D., Hess, G. T., Shamim, M. S., Machol, I., . . . Snyder, M. P. (2017). Static and Dynamic DNA Loops form AP-1-Bound Activation Hubs during Macrophage Development. *Mol Cell, 67*(6), 1037-1048 e1036. doi:10.1016/j.molcel.2017.08.006
Phillips, J. E., & Corces, V. G. (2009). CTCF: Master Weaver of the Genome. *Cell, 137*(7), 1194-1211. doi:https://doi.org/10.1016/j.cell.2009.06.001
Piovesan, A., Pelleri, M. C., Antonaros, F., Strippoli, P., Caracausi, M., & Vitale, L. (2019). On the length, weight and GC content of the human genome. *BMC research notes, 12*(1), 106-106. doi:10.1186/s13104-019-4137-z

Pombo, A., & Dillon, N. (2015). Three-dimensional genome architecture: players and mechanisms. *Nat Rev Mol Cell Biol, 16*(4), 245-257. doi:10.1038/nrm3965

Portela, A., & Esteller, M. (2010). Epigenetic modifications and human disease. *Nat Biotechnol, 28*(10), 1057-1068. doi:10.1038/nbt.1685

Qi, Y., & Xu, R. (2018). Roles of PLODs in Collagen Synthesis and Cancer Progression. *Frontiers in Cell and Developmental Biology, 6*(66). doi:10.3389/fcell.2018.00066

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), 841-842. doi:10.1093/bioinformatics/btq033

Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K. C., Grüning, B. A., Villaveces, J., . . . Manke, T. (2018). High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature Communications, 9*(1), 189. doi:10.1038/s41467-017-02525-w

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., . . . Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell, 159*(7), 1665-1680. doi:10.1016/j.cell.2014.11.021

Reyes-León, A., Ramírez-Martínez, M., Fernández-García, D., Amaro-Muñoz, D., Velázquez-Aragón, J. A., Salas-Labadía, C., . . . Pérez-Vera, P. (2019). Variants in ARID5B gene are associated with the development of acute lymphoblastic leukemia in Mexican children. *Ann Hematol, 98*(10), 2379-2388. doi:10.1007/s00277-019-03730-x

Rhee, I., Jair, K. W., Yen, R. W., Lengauer, C., Herman, J. G., Kinzler, K. W., . . . Schuebel, K. E. (2000). CpG methylation is maintained in human cancer cells lacking DNMT1. *Nature, 404*(6781), 1003-1007. doi:10.1038/35010000

Robinson, J. T., Turner, D., Durand, N. C., Thorvaldsdóttir, H., Mesirov, J. P., & Aiden, E. L. (2018). Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. *Cell Syst, 6*(2), 256-258.e251. doi:10.1016/j.cels.2018.01.001

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26*(1), 139-140. doi:10.1093/bioinformatics/btp616

Russler-Germain, D. A., Spencer, D. H., Young, M. A., Lamprecht, T. L., Miller, C. A., Fulton, R., . . . Ley, T. J. (2014). The R882H DNMT3A mutation associated with AML dominantly inhibits wild-type DNMT3A by blocking its ability to form active tetramers. *Cancer cell, 25*(4), 442-454. doi:10.1016/j.ccr.2014.02.010

Sakuma, T., Nishikawa, A., Kume, S., Chayama, K., & Yamamoto, T. (2014). Multiplex genome engineering in human cells using all-in-one CRISPR/Cas9 vector system. *Scientific reports, 4*(1), 1-6.

Schmitt, A. D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C. L., . . . Ren, B. (2016). A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell reports, 17*(8), 2042-2059. doi:10.1016/j.celrep.2016.10.061

Schmitt, A. D., Hu, M., & Ren, B. (2016). Genome-wide mapping and analysis of chromosome architecture. *Nat Rev Mol Cell Biol, 17*(12), 743-755. doi:10.1038/nrm.2016.104

Sharma, S., Kelly, T. K., & Jones, P. A. (2010). Epigenetics in cancer. *Carcinogenesis, 31*(1), 27-36. doi:10.1093/carcin/bgp220

Shin, H., Shi, Y., Dai, C., Tjong, H., Gong, K., Alber, F., & Zhou, X. J. (2016). TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic acids research, 44*(7), e70-e70. doi:10.1093/nar/gkv1505

Shlush, L. I., Zandi, S., Mitchell, A., Chen, W. C., Brandwein, J. M., Gupta, V., . . . Dick, J. E. (2014). Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature, 506*(7488), 328-333. doi:10.1038/nature13038

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., . . . de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nat Genet, 38*(11), 1348-1354. doi:10.1038/ng1896

Splinter, E., de Wit, E., van de Werken, H. J., Klous, P., & de Laat, W. (2012). Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. *Methods, 58*(3), 221-230.

Stein, U., Walther, W., Arlt, F., Schwabe, H., Smith, J., Fichtner, I., . . . Schlag, P. M. (2009). MACC1, a newly identified key regulator of HGF-MET signaling, predicts colon cancer metastasis. *Nat Med, 15*(1), 59-67. doi:10.1038/nm.1889

Sun, H. B., Shen, J., & Yokota, H. (2000). Size-Dependent Positioning of Human Chromosomes in Interphase Nuclei. *Biophysical Journal, 79*(1), 184-190. doi:https://doi.org/10.1016/S0006-3495(00)76282-5

Sun, Y., Chen, B.-R., & Deshpande, A. (2018). Epigenetic Regulators in the Development, Maintenance, and Therapeutic Targeting of Acute Myeloid Leukemia. *Frontiers in oncology, 8*, 41-41. doi:10.3389/fonc.2018.00041

Sun, Y., Zhou, B., Mao, F., Xu, J., Miao, H., Zou, Z., . . . Koche, R. (2018). HOXA9 reprograms the enhancer landscape to promote leukemogenesis. *Cancer cell, 34*(4), 643-658. e645.

Swaminathan, M., Morita, K., Yuanqing, Y., Wang, F., Burks, J. K., Gumbs, C., . . . Takahashi, K. (2018). Clinical Heterogeneity of AML Is Associated with Mutational Heterogeneity. *Blood, 132*(Supplement 1), 5240-5240. doi:10.1182/blood-2018-99-117287

Szabo, Q., Bantignies, F., & Cavalli, G. (2019). Principles of genome folding into topologically associating domains. *Science advances, 5*(4), eaaw1668-eaaw1668. doi:10.1126/sciadv.aaw1668

Tang, Q., Cheng, J., Cao, X., Surowy, H., & Burwinkel, B. (2016). Blood-based DNA methylation as biomarker for breast cancer: a systematic review. *Clin Epigenetics, 8*, 115. doi:10.1186/s13148-016-0282-6

Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., Szalaj, P., . . . Ruan, Y. (2015). CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell, 163*(7), 1611-1627. doi:10.1016/j.cell.2015.11.024

Tao, R., Liu, Y. J., Liu, L. F., Li, W., Zhao, Y., Li, H. M., . . . Zhao, Z. Y. (2019). Genetic polymorphisms of ARID5B rs7089424 and rs10994982 are associated with B-lineage ALL susceptibility in Chinese pediatric population. *J Chin Med Assoc, 82*(7), 562-567. doi:10.1097/jcma.0000000000000038

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., . . . Dawson, E. (2019). COSMIC: the catalogue of somatic mutations in cancer. *Nucleic acids research, 47*(D1), D941-D947.

Thorsteinsdottir, U., Kroon, E., Jerome, L., Blasi, F., & Sauvageau, G. (2001). Defining roles for HOX and MEIS1 genes in induction of acute myeloid leukemia. *Mol Cell Biol, 21*(1), 224-234. doi:10.1128/mcb.21.1.224-234.2001

Viré, E., Brenner, C., Deplus, R., Blanchon, L., Fraga, M., Didelot, C., . . . Fuks, F. (2006). The Polycomb group protein EZH2 directly controls DNA methylation. *Nature, 439*(7078), 871-874. doi:10.1038/nature04431

Waddington, C. H. (1942). The epigenotype. *Endeavour, 1*, 18-20.

Waddington, C. H. (1968). Towards a Theoretical Biology. *Nature, 218*(5141), 525-527. doi:10.1038/218525a0

Wang, B., Kong, L., Babu, D., Choudhary, R., Fam, W., Tng, J. Q., . . . Fullwood, M. J. (2021). Three-dimensional Genome Organization Maps in Normal Haematopoietic Stem Cells and Acute Myeloid Leukemia. *bioRxiv*, 2020.2004.2018.047738. doi:10.1101/2020.04.18.047738

Wang, H., Han, M., & Qi, L. S. (2021). Engineering 3D genome organization. *Nature Reviews Genetics, 22*(6), 343-360. doi:10.1038/s41576-020-00325-5

Wang, H., Maurano, M. T., Qu, H., Varley, K. E., Gertz, J., Pauli, F., . . . Stamatoyannopoulos, J. A. (2012). Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res, 22*(9), 1680-1688. doi:10.1101/gr.136101.111

Wang, P., Deng, Y., Yan, X., Zhu, J., Yin, Y., Shu, Y., . . . Lu, X. (2020). The Role of ARID5B in Acute Lymphoblastic Leukemia and Beyond. *Front Genet, 11*, 598. doi:10.3389/fgene.2020.00598

Wang, Q. f., Li, Y. j., Dong, J. f., Li, B., Kaberlein, J. J., Zhang, L., . . . Thirman, M. J. (2014). Regulation of MEIS1 by distal enhancer

elements in acute leukemia. *Leukemia, 28*(1), 138-146. doi:10.1038/leu.2013.260
Wang, Y., Wu, N., Liu, D., & Jin, Y. (2017). Recurrent Fusion Genes in Leukemia: An Attractive Target for Diagnosis and Treatment. *Curr Genomics, 18*(5), 378-384. doi:10.2174/1389202918666170329110349
Weinreb, C., & Raphael, B. J. (2016). Identification of hierarchical chromatin domains. *Bioinformatics, 32*(11), 1601-1609. doi:10.1093/bioinformatics/btv485
Weth, O., Paprotka, C., Günther, K., Schulte, A., Baierl, M., Leers, J., . . . Renkawitz, R. (2014). CTCF induces histone variant incorporation, erases the H3K27me3 histone mark and opens chromatin. *Nucleic acids research, 42*(19), 11941-11951. doi:10.1093/nar/gku937
WHO. (2021, 3 March 2021). Cancer. Retrieved from https://www.who.int/en/news-room/fact-sheets/detail/cancer
Wolff, J., Bhardwaj, V., Nothjunge, S., Richard, G., Renschler, G., Gilsbach, R., . . . Grüning, B. A. (2018). Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization. *Nucleic acids research, 46*(W1), W11-W16. doi:10.1093/nar/gky504
Wolff, J., Rabbani, L., Gilsbach, R., Richard, G., Manke, T., Backofen, R., & Grüning, B. A. (2020). Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic acids research, 48*(W1), W177-W184. doi:10.1093/nar/gkaa220
Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjöblom, T., Leary, R. J., . . . Vogelstein, B. (2007). The genomic landscapes of human breast and colorectal cancers. *Science, 318*(5853), 1108-1113. doi:10.1126/science.1145720
Wouters, B. J., & Delwel, R. (2016). Epigenetics and approaches to targeted epigenetic therapy in acute myeloid leukemia. *Blood, 127*(1), 42-52. doi:https://doi.org/10.1182/blood-2015-07-604512
Wu, C., & Morris, J. R. (2001). Genes, genetics, and epigenetics: a correspondence. *Science, 293*(5532), 1103-1105. doi:10.1126/science.293.5532.1103
Yaffe, E., & Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet, 43*(11), 1059-1065. doi:10.1038/ng.947
Yang, M., Vesterlund, M., Siavelis, I., Moura-Castro, L. H., Castor, A., Fioretos, T., . . . Paulsson, K. (2019). Proteogenomics and Hi-C reveal transcriptional dysregulation in high hyperdiploid childhood acute lymphoblastic leukemia. *Nature Communications, 10*(1), 1519. doi:10.1038/s41467-019-09469-3
Zhang, Y., Cai, Y., Roca, X., Kwoh, C. K., & Fullwood, M. J. (2021). Chromatin loop anchors predict transcript and exon usage. *Briefings in Bioinformatics*. doi:10.1093/bib/bbab254

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., . . . Liu, X. S. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biology, 9*(9), R137. doi:10.1186/gb-2008-9-9-r137

Zhang, Z., Jia, H., Wang, Y., Du, B., & Zhong, J. (2021). Association of MACC1 expression with lymphatic metastasis in colorectal cancer: A nested case-control study. *PLoS One, 16*(8), e0255489. doi:10.1371/journal.pone.0255489

Zhou, J., Gou, H., Zhang, L., Wang, X., Ye, Y., Lu, X., & Ying, B. (2019). ARID5B Genetic Polymorphisms Contribute to the Susceptibility and Prognosis of Male Acute Promyelocytic Leukemia. *DNA Cell Biol, 38*(11), 1374-1386. doi:10.1089/dna.2019.4926

Zhou, J., Ma, J., Chen, Y., Cheng, C., Bao, B., Peng, J., . . . Ecker, J. R. (2019). Robust single-cell Hi-C clustering by convolution-and random-walk–based imputation. *Proceedings of the National Academy of Sciences, 116*(28), 14011-14018.