MOLECULAR AND REGULATORY PROFILING OF MURINE TRANSCRIPTION FACTOR CTCFL IN NON-GERM CELLS AND MALE GERMLINE STEM CELLS

MAISARAH BINTI AB SAMAD

UNIVERSITI SAINS MALAYSIA

2022

MOLECULAR AND REGULATORY PROFILING OF MURINE TRANSCRIPTION FACTOR CTCFL IN NON-GERM CELLS AND MALE GERMLINE STEM CELLS

by

MAISARAH BINTI AB SAMAD

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

April 2022

ACKNOWLEDGEMENT

All praises and gratitude are to Allah SWT, the Lord to whom every single creature in heaven and the earth belongs. Alhamdulillah for giving us the opportunity and help us endlessly in finishing the thesis. I would like to express my gratitude to Professor Dr Shaharum Shamsuddin, my principal supervisor, for the chance he gave me to participate in the USM-RIKEN Interdisciplinary Collaboration for Advanced Sciences (URICAS). I am also very grateful for his advice, support, and constructive comment throughout the journey. My deepest gratitude goes to Professor Kuniya Abe, my field supervisor from RIKEN BioResource Research Center (RIKEN BRC), Tsukuba, Japan, for the opportunity to work towards this thesis and his support over the last four years of my PhD. I would also like to thank all of my colleagues in the Technology and Development Team for Mammalian Genome Dynamics in RIKEN BRC (Dr Suzuki Shinnosuke, Dr Ura Hiroki, Dr Sugimoto Michihiko, Dr Tada Yuhki, Mrs Koga Yumiko, Mr Cho Dooseon and Mrs Kusayama Miwako) for their advice and generous assistance throughout my stay in RIKEN BRC. I would like to thank RIKEN for awarding me with the International Program Associate fellowship, which has allowed me to carry out this research in RIKEN BRC, Tsukuba. Japan. I would also like to thank the Public Service Department of Malaysia (JPA) for awarding me the Program Pelajar Cemerlang fellowship and School of Health Sciences, USM, to award me the Graduate Research Assistance scheme. Sincere gratitude also to my co-supervisor, Dr Daruliza Kernain Mohd Azman, for the continuous support received. Finally, I would like to thank my mother, family and friends for their patience, understanding, encouragement and support throughout everything.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF SYMBOLS

%	Percentage
1×	One time dilution
10×	Ten-time dilution
$\times 10^4$	Ten raised to a power of four
bp	Base pair(s)
kb	Kilobase pair
µg	Microgram
µl	Microlitre
µM	Micromolar
M	Molar/molarity
mL	Millilitre
mm	Millimetre
mM	Millimolar
ng	Nanogram
nM	Nanomolar
^oC	Degree Celsius
pM	Picomolar
pmol	Picomol
rpm	Revolutions per minute
V	Volt
w/v	Mass per volume

LIST OF ABBREVIATIONS

(DMSO)	DMSO treated cells
(DOX)	Doxycycline treated cells, (3×FLAG-CTCFL expressed cells)
3’ UTR	3’ Untranslated regions
BSA	Bovine serum albumin
cDNA	Complementary deoxyribonucleic acid
CDS	Coding sequence
ChIP	Chromatin immunoprecipitation
ChIP-seq	Chromatin immunoprecipitation and sequencing
cRNA	Complementary ribonucleic acid
CTA	Cancer testis antigen
CTCF	CCCTC-binding factor
Ctcf	CCCTC-binding factor (gene)
CTCFL	CCCTC-binding factor like
Ctcfl	CCCTC-binding factor-like (gene)
CTCF&CTCFL	DNA sequences targeted by CTCF and CTCFL
CTCF-only	DNA sequences targeted by CTCF only
CTCFL-only	DNA sequences targeted by CTCFL only
DE gene	Differentially expressed gene
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DOX	Doxycycline
dsDNA	Double-stranded DNA
DsiRNA	Dicer-substrate short interfering RNAs
E	Embryonic day
E	Efficiency
EGFP	Enhanced Green Fluorescent Protein
ESCs	Embryonic stem cells
ESC-CTCFL	ESC-CTCFL cell line
ESC-CTCFL(DMSO)	Control ESC-CTCFL cell line (repressed CTCFL, DMSO treated cells)
ESC-CTCFL(DOX)	ESC-CTCFL cell line with 3×FLAG-CTCFL expression (doxycycline-treated cells)

FBS	Fetal bovine serum
FC	Fold-change
FDR	False discovery rate
GO	Gene ontology
GSCs	Germline stem cells
GSC(Neg)	Control GSCs (without *Ctcfl* knockdown)
GSC-*Ctcfl*(kd)	*Ctcfl* knockdown GSCs
HF	High fidelity
H3K4me3	Histone-3-lysine-4-trimethylation
H3K27me3	Histone-3-lysine-27-trimethylation
JK1	JK1 testicular stromal cell line
JK1-CTCFL	JK1-CTCFL cell line
JK1-CTCFL(DMSO)	Control JK1-CTCFL cell line (repressed CTCFL, DMSO treated cells)
JK1-CTCFL(DOX)	JK1-CTCFL cell line with 3×FLAG-CTCFL expression (doxycycline-treated cells)
kb	Kilo base pairs
kd	Knockdown
kDa	Kilo Dalton
KEGG	Kyoto Encyclopedia of Genes and Genomes
KSR	Knockout serum replacement
MEFs	Mouse embryonic fibroblasts
mRNA	Messenger RNA
n	Number/count
n	Number of replicates
neg	Negative control
NGS	Next-generation sequencing
P	Postnatal day
p	Passage number
PAGE	Polyacrylamide gel electrophoresis
PBS	Phosphate buffered saline
PCA	Principle component analysis
PCR	Polymerase chain reaction
PRC2	Polycomb repressor complex 2
qPCR	Quantitative polymerase chain reaction
R	Replicate

Rep	Replicate
RIN	RNA Integrity Number
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RNAi	RNA interference
RT-qPCR	Reverse transcription-quantitative polymerase chain reaction
SSCs	Spermatogonial stem cells
SUZ12	Suppressor of zeste 12 protein
TNF	Tumour necrosis factor
TSS	Transcriptional start sites
ZF	Zinc finger

LIST OF APPENDICES

PEMPROFILAN MOLEKUL DAN PENGAWALATURAN TERHADAP FAKTOR TRANSKRIPSI TIKUS CTCFL DALAM SEL BUKAN GERMA DAN SEL STEM GERMA JANTAN

ABSTRAK

Faktor pengikat CCCTC khusus testis (CTCFL) adalah faktor transkripsi yang diekspresi dalam sel germa jantan dan penting untuk penghasilan sperma. CTCFL ialah paralog kepada faktor pengikat CCCTC (CTCF), faktor transkripsi dan protein pengatur-bina kromatin. Mereka berkongsi domain pengikat DNA jejari-zink yang serupa tetapi mempunyai dua domain hujung yang berbeza. CTCFL tidak ekspres dalam tisu somatik; walau bagaimanapun, ekspresi yang abnormal dalam sel bukan germa dikaitkan dengan kanser. Fungsi pengawalseliaan CTCFL dalam penghasilan sperma dan pertumbuhan tumor masih tidak jelas. Kajian ini bertujuan untuk menjelaskan profil pengawalseliaan CTCFL murin dalam sel bukan germa dan sel germa jantan. Ekspresi ektopik CTCFL bertanda FLAG (3×FLAG-CTCFL) daktifkan dalam sel bukan germa, iaitu sel induk embrio tikus (ESCs) dan sel stroma testis JK1 (JK1) selama 24 Jam. Untuk mengetahui peranan CTCFL dalam sel germa, ekspresi Ctcfl endogen dalam sel stem germanium jantan (GSCs) telah dikurangkan melalui gangguan RNA selama 48 jam. Perubahan global dalam transkrip telah dianalisis menggunakan microarray dan penjujukan RNA. Analisis anotasi fungsi mendapati gen dan proses perkembangan telah diganggu apabila CTCFL diekspresi dalam ESC, sebagai ESC-CTCFL(DOX) dan sel JK1, sebagai JK1-CTCFL(DOX). Pengurangan ekspresi Ctcfl dalam GSCs menindas ekspresi gen yang terlibat dalam peraturan kematian sel dan proses selular dalam spermatogonia. Mod pengawalseliaan CTCFL telah disimpulkan berdasarkan profil ikatan protein-DNA dalam genom ESC-CTCFL(DOX). Analisis awal in silico menganggarkan

perkaitan utama antara gen-gen yang mempunyai pertambahan ekspresi dalam ESC-CTCFL(DOX) dengan pengayaan CTCFL, CTCF, dan komponen-komponen Polycomb Repressor Complex 2 (PRC2), termasuk Suppressor of zeste 12 protein (SUZ12). Pengayaan CTCFL, CTCF, dan SUZ12 terhadap genom dalam ESC-CTCFL(DOX) telah dianalisis menggunakan analisis imunoprekipitasi kromatin dan penjujukan (ChIP-seq). Hasilnya mengesahkan pertindihan pengikatan CTCFL dengan tapak pengayaan CTCF dan SUZ12 yang mendominasi tapak promotor pengatur perkembangan sel. Pengikatan bersama CTCF dan CTCFL di tapak CTCF&CTCFL boleh memacu pengaktifan gen germa dalam sel bukan germa. Pengayaan SUZ12 telah dikurangkan di tapak CTCF&CTCFL berhampiran promotor gen perkembangan yang ditambah ekpresinya, menyimpulkan perubahan fungsi penindasan PRC2 oleh CTCFL. Ringkasnya, CTCFL mungkin memainkan peranan dalam peraturan perkembangan dengan mengubah suai proses dalam sel bukan germa atau dengan mengawal diferensiasi spermatogonia. Disregulasi aktiviti CTCF dan PRC2 boleh menjadi mod pengawalseliaan CTCFL dalam pertumbuhan barah dan memerlukan penyiasatan lanjut.

MOLECULAR AND REGULATORY PROFILING OF MURINE TRANSCRIPTION FACTOR CTCFL IN NON-GERM CELLS AND MALE GERMLINE STEM CELLS

ABSTRACT

Testis-specific CCCTC-binding factor-like (CTCFL) is a transcription factor expressed in male germ cells and essential for spermatogenesis. CTCFL is the paralog of CCCTC-binding factor (CTCF), a transcription factor and chromatin architectural protein. They share a similar zinc-finger DNA binding domain but different end termini. CTCFL is repressed in somatic tissues; however, the aberrant expression in the non-germ cells is associated with cancer. The regulatory functions of CTCFL in spermatogenesis and tumourigenesis are still unclear. This study aimed to elucidate the regulatory profile of murine CTCFL in non-germ cells and male germ cells. To investigate the transcriptional effects of CTCFL aberrant expression in non-germ cells, ectopic FLAG-tagged CTCFL (3×FLAG-CTCFL) was expressed in the pluripotent mouse embryonic stem cells (ESCs) and JK1 testicular stromal cell line (JK1) for 24 hours. To discover the roles of CTCFL in the germ cells, the expression of endogenous *Ctcfl* in male germline stem cells (GSCs) was reduced by RNA interference for 48 hours. Global changes in the transcriptome were measured by microarray and RNA sequencing, respectively. The functional annotation analysis observed altered developmental genes and processes in CTCFL expressing ESCs, ESC-CTCFL(DOX), and JK1 cells, JK1-CTCFL(DOX). *Ctcfl* knockdown in GSCs repressed the genes involved in cell death regulation and cellular processes in spermatogonia. The regulatory mode of CTCFL was inferred based on the genome-wide protein-DNA binding profile in ESC-CTCFL(DOX). Preliminary *in silico* analysis highlighted the association of the upregulated genes in ESC-CTCFL(DOX)

with the enrichment of CTCFL, CTCF, and Polycomb Repressor Complex 2 (PRC2) components, including Suppressor of zeste 12 protein (SUZ12). The enrichment of CTCFL, CTCF, and SUZ12 to the genome in ESC-CTCFL(DOX) were analysed by chromatin immunoprecipitation and sequencing (ChIP-seq). The results validated the intersection of CTCFL binding with CTCF and SUZ12 enrichment sites that dominated the promoter of developmental regulators. CTCF and CTCFL co-binding at CTCF&CTCFL sites could drive the activation of germline genes in the non-germ cells. SUZ12 enrichment was reduced at the CTCF&CTCFL sites near the promoter of upregulated developmental genes, inferring alteration of PRC2 repression by CTCFL. In summary, CTCFL may play a role in developmental regulation by modifying the processes in non-germ cells and regulating spermatogonia's differentiation. CTCF and PRC2 activity dysregulation could be the CTCFL regulatory mode in tumourigenesis and warrant further investigations.

CHAPTER 1

INTRODUCTION

1.1 Background of Study

Transcription factors are one of the main components in the transcriptional machinery regulating gene transcription into mRNA. In turn, the gene expression programs convey cell-fate decisions for cellular activity and development. Transcription factors act by binding to specific DNA motifs within gene regulatory elements (Stadhouders *et al.*, 2019). The sequence-specific transcription factors may become the activator or repressor that regulate the general transcription factor complex that controls the gene transcription (Suter, 2020). Regulation of gene expression by transcription factors interplays with chromatin regulators to trigger the local alteration of chromatin structure. Transcription factors and chromatin regulators confer epigenomic and transcriptomic levels to govern the circuitry of the gene regulatory network (Wilson and Filipp, 2018).

CCCTC-binding factor (CTCF) is a transcription factor and an architectural protein for chromatin organisation (Phillips and Corces, 2009; Xiang and Corces, 2020). CTCF binds to its cognate DNA sites highly distributed across the genome in all cell types (Chen *et al.*, 2012a; Maurano *et al.*, 2015). CTCF plays diverse roles in transcription regulation and mediating higher-order chromatin organisation (Braccioli and de Wit, 2019; Herold *et al.*, 2012; Phillips and Corces, 2009; Wu *et al.*, 2020). Polycomb repressor complex (PRC2) is a chromatin regulator that controls the gene repression by catalysing the methylation of lysine 27 on Histone H3 (H3K27) (Glancy *et al.*, 2021). Both PRC2 and CTCF are essential regulators for cellular development (Arzate-Mejia *et al.*, 2018; Deevy and Bracken, 2019). They also cooperate in regulating the expression of imprinted genes (Li *et al.*, 2008). Both

are essential for the appropriate expression of developmental genes in pluripotent stem cells (Dowen *et al.*, 2014; Xu *et al.*, 2014). CTCF and PRC2 are often mutated, causing dysregulation of gene expression and driving to cancer (Debaugny and Skok, 2020; Piunti and Shilatifard, 2021).

CTCFL (CCCTC-binding factor-like) is a zinc-finger transcription factor expressed selectively in male germ cells (Loukinov *et al.*, 2002). CTCFL is the paralog of CTCF with restrictive expression in male germ cells (Klenova *et al.*, 2002; Loukinov *et al.*, 2002; Sleutels *et al.*, 2012). Hence, *Ctcfl* is a germline gene as its expression is essential for spermatogenesis. CTCFL regulates the expression of essential germline genes, which are *Gal3st1* and *Prss50* (Kosaka-Suzuki *et al.*, 2011; Suzuki *et al.*, 2010). Depletion of CTCFL results in defective spermatogenesis and sub-sterility in mice (Sleutels *et al.*, 2012).

Nevertheless, CTCFL is also a cancer-testis antigen (CTA). CTA genes have a restrictive expression in normal testis, but its derepression in the non-testicular tissues is associated with malignancies (Gibbs and Whitehurst, 2018; Kalejs and Erenpreisa, 2005). CTCFL aberrant expression beyond the male germ cells is linked with the invasive phenotype of cancer (Debaugny and Skok, 2020; Debruyne *et al.*, 2019; Janssen *et al.*, 2020; Klenova *et al.*, 2002; Soltanian and Dehghani, 2018). The reactivation of CTAs beyond the male germ cells leads to the hijacking of cell development processes by supporting the survivability of cancer cells (Gordeeva, 2018). As a transcription factor, the derepression of CTCFL in cancer may contribute to the alteration of transcription profile which drives the malignant genome reprogramming and cell transformation (Debaugny and Skok, 2020; Wang *et al.*, 2011). CTCFL is also a potential candidate for immunotherapeutic against cancer and metastasis (Loukinov, 2018).

1.2 Problem Statement

CTCFL aberrant expression in the non-testicular tissues can be detrimental as the protein is a transcription factor and a CTA, which may drive the soma-to-germline expression program for malignant reprogramming (Wang *et al.*, 2011). In addition, CTCFL can activate the expression of other CTAs and germline genes in cancer, although the derepression is speculated by several studies (Debaugny and Skok, 2020; Hillman *et al.*, 2019; Woloszynska-Read *et al.*, 2011).

A study by Pugacheva *et al.* (2015) discovered the regulatory mode of CTCFL that is common between germline and cancer cells. CTCFL can bind to the CTCF binding sites near the promoter of germline genes, forming a heterodimer with CTCF and activating the aberrant gene expression in the somatic cells. This abnormal regulatory action may predispose them to cancer. In parallel, induction of aberrant CTCFL expression by Sati *et al.* (2015) during mouse embryogenesis resulted in poor growth development. They observed that the dead pups had growth retardation and malformations in multiple somatic organs, particularly the eyes, brain, and vascular system. It is peculiar that a testis-specific transcription factor could cause grievous alterations in developing non-testicular tissues. This finding provides a clue about the dysregulations by CTCFL in somatic development by altering the Transforming Growth Beta (TGF-β) signalling pathway. Still, establishing a link between CTCFL activation of the germline programme and cancer is difficult due to the lack of study on its role in germ cells.

Most transcription factors do not independently participate in gene transcriptional regulation (Will and Helms, 2014). Besides binding to their target DNA binding sites, they also interact with other regulatory proteins such as mediators and chromatin modifiers (Nakagawa *et al.*, 2018). Albeit being discovered

for 20 years, less is known about the functional binding sites of CTCFL and its cooperation with other regulators. CTCFL binding overlaps a subset of CTCF target sites as paralogous proteins due to the similar ZF DNA binding domain (Pugacheva *et al.*, 2015). The different end termini and configuration of ZF domains modulate the selection of DNA binding and interaction with protein partners (Bergmaier *et al.*, 2018; Nishana *et al.*, 2020). Less is known about the downstream regulations, primarily when CTCF and CTCFL co-express in the tumour and male germ cells. They may interact with similar DNA sequences and other regulatory proteins but with unique downstream regulations.

1.3 Rationale of Study

Germline genes, including CTAs, involve a wide range of processes such as genomic maintenance, transcriptional regulation and meiosis in spermatogenesis, which could connect to their aberrant activation derepression outcomes with neoplastic behaviours (Gibbs and Whitehurst, 2018; Whitehurst, 2014). CTCFL participates in transcriptional regulation in tumourigenesis and spermatogenesis. The implications of CTCFL “out of context” expression may associate with the derepression of germline genes, as well as dysregulation of somatic development (Pugacheva *et al.*, 2015; Sati *et al.*, 2015). Still, there is a gap of understanding of how CTCFL functions. CTCFL regulation and its role in male germ cells require further elucidation.

The current study was conducted to investigate CTCFL regulation in both expression settings for a comprehensive understanding, based on a hypothesis that the implications of CTCFL aberrant expression in non-germ cells could be linked with its functional roles in male germ cells. In addition, this study also explored the

mechanism of transcriptional regulation of CTCFL as a transcription factor that may link CTCFL function in spermatogenesis and cancer.

1.4 Research Objectives

The main objective of this study was to elucidate the regulatory profile of murine CTCFL in non-germ cells and male germ cells. The following specific objectives were targeted in this study.

(1) To investigate the transcriptional effects of CTCFL aberrant expression in non-germ cells by introducing the ectopic CTCFL (Chapter 4).

(2) To discover the roles of CTCFL in male germ cells by knocking down the endogenous *Ctcfl* transcript expression (Chapter 5).

(3) To infer the regulatory mode of CTCFL as a transcription factor based on the genome-wide protein-DNA binding profile (Chapter 6).

The first and second objectives focussed on the transcriptional regulation of CTCFL in the non-germ cells and male germ cells, respectively. For the first objective, this work utilised the recombinant construct, 3×FLAG-CTCFL, for the ectopic CTCFL expression in the non-germ cells. Ectopic expression is the expression of a gene in a cell type, developmental stage, or condition where it should not be expressed (Prelich, 2012). The ectopic CTCFL expression was conducted in pluripotent embryonic stem cells (ESCs) and JK1 testicular stromal cell lines, in which the protein is repressed. Tumour or cancer cell lines were opted out of this study as the altered genetic and epigenetic events in are confounding factors may conceal the direct effects of CTCFL aberrant expression. The ectopic CTCFL was tagged FLAG-tag epitope to allow protein probing by anti-FLAG monoclonal antibody for protein expression analysis and chromatin immunoprecipitation. The

implications of CTCFL aberrant expression in the ESCs and JK1 cells was analysed by microarray to measure the global gene expression changes.

ESCs are ideal for investigating mammalian transcriptional regulation (Rao, 2012). An exhaustive range of datasets for ESCs allows simultaneous understanding and construction of hypotheses behind the multiple layers of gene expression regulation in the cell system. Furthermore, as ESCs can differentiate into all cell lineages (the germ layers and germline cells), the cells are heavily utilised for investigating developmental process and lineage specification. Several studies have used ESCs or pluripotent stem cells for investigating the roles of CTCFL (Bergmaier *et al.*, 2018; Nishana *et al.*, 2020; Sati *et al.*, 2015; Sleutels *et al.*, 2012). Meanwhile, JK1 cells are immortalised cells originating from CD34+ enriched testicular stromal cells, particularly the peritubular myoid cells (Kim *et al.*, 2008; Seandel *et al.*, 2007). This stromal cell line can be utilised as feeder cells that support the expansion of adult spermatogonial stem cells (Kim *et al.*, 2008). Thus, JK1 cells may resemble the testicular somatic cell type, closely connected with the growth of male germ cells.

Ctcfl expression knockdown was conducted in male germline stem cells (GSCs) for the second objective. GSCs are the *in vitro* cell model representing undifferentiated spermatogonial stem cells (SSCs). GSCs were established from the SSCs of neonatal mice and continuously proliferated as undifferentiated spermatogonia (Kanatsu-Shinohara *et al.*, 2003a; Kanatsu-Shinohara *et al.*, 2003b). CTCFL has the most outstanding expression level in mitotic spermatogonia compared to other germ cell stages (Rivero-Hinojosa *et al.*, 2021). Hence, GSCs express the endogenous CTCFL and can analyse the physiological role of this transcription factor in male germ cells. The knockdown approach or gene silencing was chosen to reduce the gene expression aberrantly and elucidate the direct roles of

CTCFL in the cells. The biggest concern of using an RNAi-based system for knockdown studies is the off-target effect. This study used an improved RNAi system, the Dicer-substrate short interfering RNAs (DsiRNA), to target the *Ctcfl* at specific sites. DsiRNA has improved performance than the canonical small interfering RNA (siRNA) in the RNAi pathway by minimising the off-target effect (Snead *et al.*, 2013). RNA sequencing (RNA-seq) analysed the transcriptome profile of *Ctcfl* knockdown GSCs, and data analysis computed the differential expressed genes.

Lastly, the genome-wide CTCFL enrichment was investigated to infer the regulatory mode or the mechanism of CTCFL regulation. ChIP-seq profiled the CTCFL binding sites, and further analysis retrieved the possible direct targets and associated downstream biological processes. Genomic enrichment of its paralog, CTCF and a core subunit of PRC2, SUZ12, were also assessed. An integrative analysis was conducted to assess the association between the transcriptional regulation and the annotated protein-DNA binding profile. The analysis revealed the possible mechanism of CTCFL-dependent gene regulation in non-germ cells and germ cells. Figure 1.1 illustrates the research design in this study based on the three specific objectives.

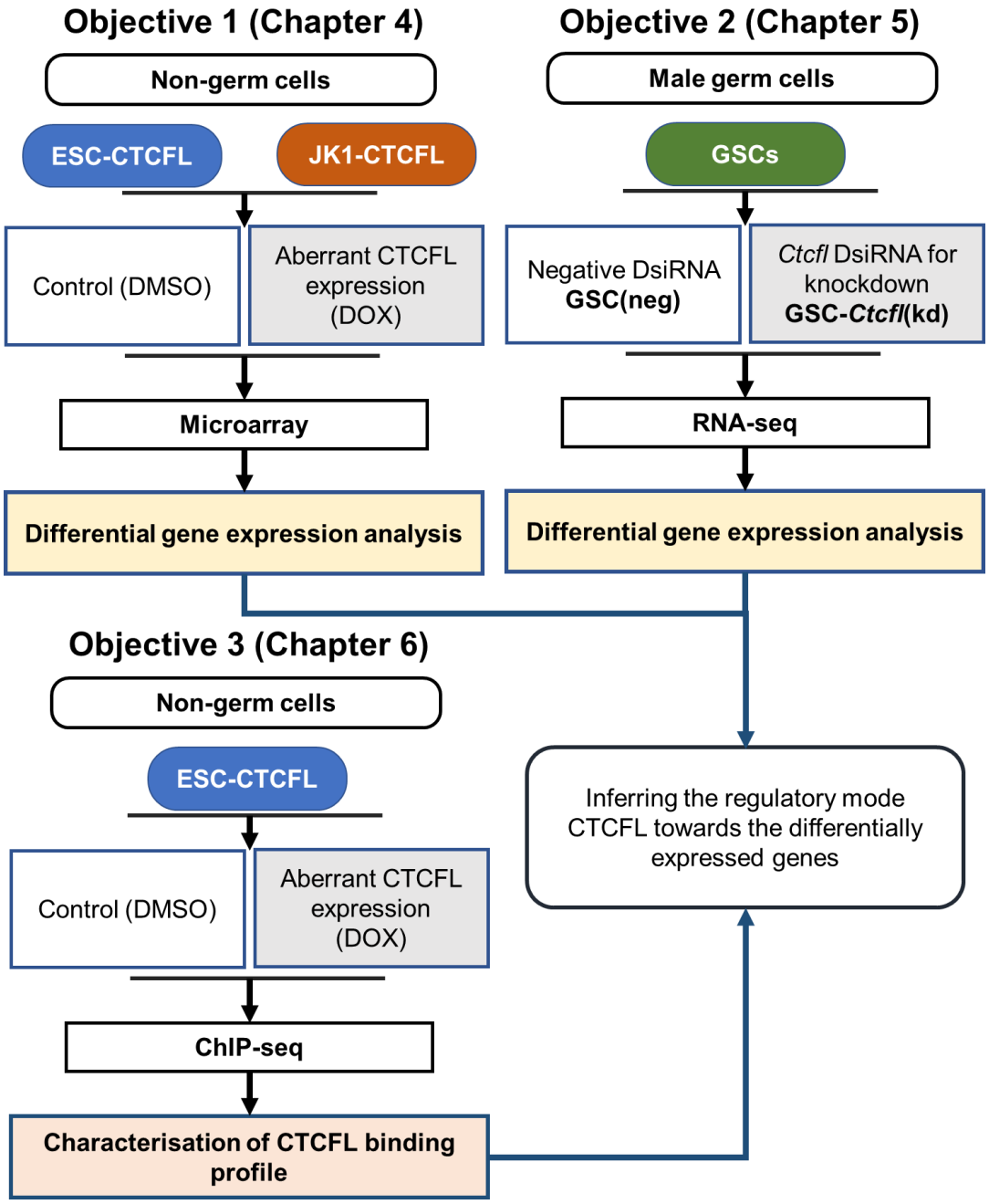


Figure 1.1 The overall research design conducted in this study was based on three specific objectives.

1.5 Thesis Organisation

The thesis is organised into chapters that cover the three specific objectives. Chapter 2 describes the biological background of transcription factors and chromatin regulators by introducing CTCF and PRC2 as the respective example. The literature review highlights the CTCFL as the protein of interest in this study. An overview of genome-wide analysis methods, including microarray, RNA-seq and ChIP-seq, was described later in this chapter.

Chapter 3 provides the essential research methods used to achieve the objectives of this study. The thesis highlighted the detailed protocol for the mammalian cell culture, quantitative polymerase chain reaction (qPCR), protein expression analysis, microarray and library preparation for RNA-seq and ChIP-seq.

Chapters 4, 5 and 6 present the experiments and findings for the three specific objectives, respectively. In Chapter 4, the aim was to investigate the implications of CTCFL aberrant expression in ESCs and JK1 cells. Transcriptomic changes were analysed using microarray to discover the differentially expressed (DE) genes. This chapter highlights the effects of CTCFL expression toward the potential dysregulation of cell development in the stem cells and somatic cells.

In Chapter 5, the aim was to elucidate the roles of CTCFL in GSCs. The expression of *Ctcfl* in GSCs was aberrantly reduced via gene knockdown followed by RNA-seq. Transcriptomic analysis exhibited alteration of gene expression, which involved cellular processes such as cell death and signalling pathways. This chapter suggests the potential role of CTCFL in the male germ cells during the early stage of spermatogenesis.

In Chapter 6, the aim was to investigate the regulatory mode of CTCFL based on its genome-wide binding profile. The enrichment of CTCFL and CTCF and a core subunit of PRC2; SUZ12 were characterised by ChIP-seq. CTCFL binding was predominant at the gene promoters that CTCF and SUZ12 occupied. Furthermore, the analysis discovered the probable alteration of the roles of CTCF and PRC2 in response to CTCFL enrichment to the overlapping binding sites.

Finally, Chapter 7 concludes the thesis by summarising the findings of this study and mapping out possible areas for future research.

CHAPTER 2

LITERATURE REVIEW

2.1 An Overview of Transcription Factors and Histone Modifications

2.1.1 Chromatin organisation

DNA stores the hereditary genetic information of an organism. DNA is a two-stranded helix in which each strand consists of four different nucleotides called adenine (A), guanine (G), cytosine (C) and thymine (T). DNA exists within the nucleus and would stretch up to two meters for human DNA. DNA is folded in a highly compacted structure known as chromatin in eukaryotic cells, including nucleic acids and histone proteins. The protein-DNA complex is called the nucleosomes in which 147 base pairs (bp) of DNA is wrapped approximately 1.65 to 1.7 times around a core protein octamer composed of two units of each of the histone proteins, H2A, H2B, H3, and H4 (Luger and Hansen, 2005; Luger *et al.*, 1997; Richmond and Davey, 2003).

Chromatin structure divides into three hierarchical states based on compaction: primary, secondary, and tertiary (Figure 2.1). The primary state consists of core chromatin that appears as beads-on-a-string, forming an eleven nm chromatin fibre. Next, a range of 20 to 50 bp of linker DNA associated with histone H1 separates the nucleosomes (Oudet *et al.*, 1975). The secondary state involves condensing the eleven nm fibre further to form a 30 nm chromatin fibre through nucleosomal interactions, but the formation remains controversial (Felsenfeld and McGhee, 1986; Maeshima *et al.*, 2010). Finally, in the tertiary level of compaction, the chromatin is folded, resulting in interchromosomal and intrachromosomal interactions of the 30 and 11 nm fibres that form the higher-order structures (Bian and Belmont, 2012; Rattner and Hamkalo, 1978).

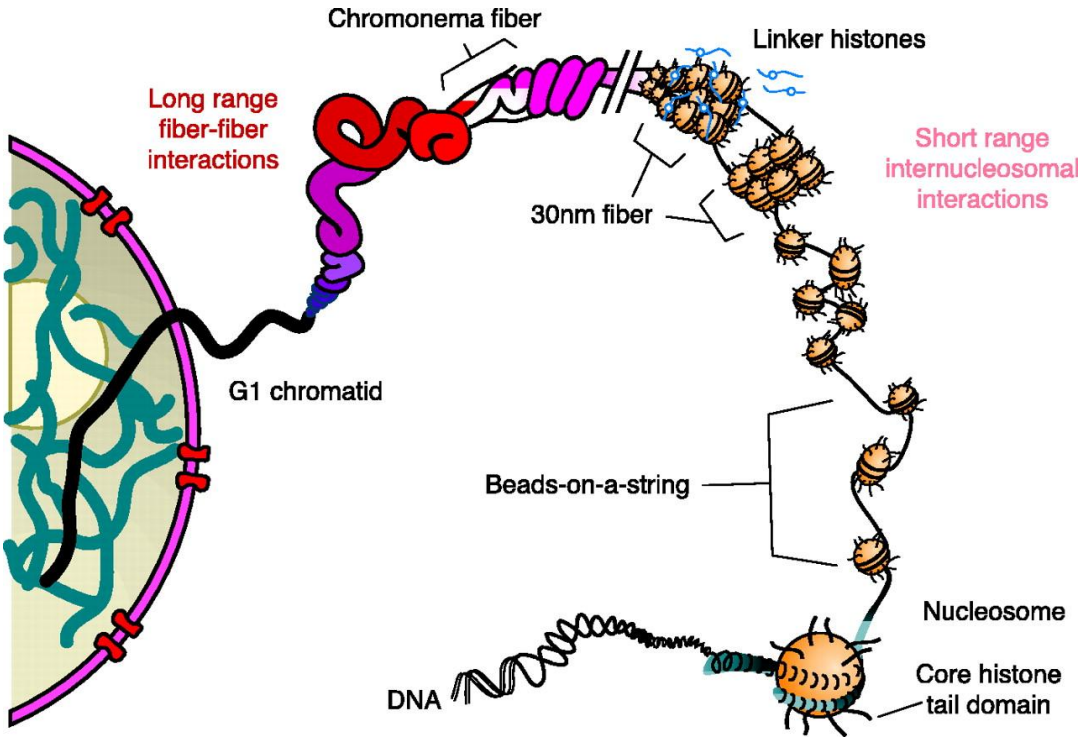


Figure 2.1 Levels of chromatin organisation. DNA and its associated proteins are called chromatin. The DNA is tightly wrapped around a histone octamer to form the nucleosome. These nucleosomes are separated by linker histones and compacted in a 30 nm chromatin fibre. Chromatin is further compacted into higher-order structures. The figure was adapted from Botchkarev *et al.* (2012)

Chromatin is functionally divided into two general states, euchromatin and heterochromatin, distributed throughout the genome. Euchromatin is less condensed, gene-rich, and often associated with active gene expression (Dillon and Festenstein, 2002). Heterochromatin is highly condensed, generally gene-poor and lacks much active gene expression. The transcriptionally inactive chromatin can be in two varieties: permanently silenced or constitutive heterochromatin and facultative heterochromatin. Constitutive chromatin localises extensively near centromeres, telomeres, and gene desert regions. Facultative chromatin is the condensed euchromatic region where the compaction is developmentally regulated (Gilbert *et al.,* 2003). The convoluted chromatin changes through histone modifications and accessibility to regulatory proteins are linked with the coordinated regulation of gene expression for cellular functions such as transcription, DNA replication and DNA repair (Blakey and Litt, 2015; Parmar and Padinhateeri, 2020).

2.1.2 Transcription factors

Gene transcriptional regulation is central to tissue-specific gene expression and stimulus-mediated gene activity in development and cellular differentiation (Holmberg and Perlmann, 2012; Hu and Gallo, 2010; Molina *et al.*, 2013; Sonawane *et al.*, 2017). Transcription factors are regulatory proteins bound to short specific DNA sequences or motifs and form a protein complex system to control the gene transcription (Lambert *et al.*, 2018; Latchman, 2010). Hence, transcription factors cooperate with DNA and coregulators, determining the transcription factor function and transcriptional activity (Cheatle Jarvela and Hinman, 2015). The protein-DNA and protein-protein interactions are the key parameters determining the three functional features of transcription factors: local chromatin access, self-sustained

remodelling, and sporadic occupancy/selective gene activation (Francois *et al.*, 2020).

Specific transcription factor binding motif localises to *cis*-regulatory elements that include the promoters, enhancers (or silencers) and insulators for precise control of gene expression (Inukai *et al.*, 2017). The promoter is a stretch of upstream DNA sequence where transcription is initiated, and this region includes the TSS to which the basal transcriptional machinery is enriched. The promoters can be tissue-specific or broadly scattered in many tissues. The tissue-specific promoters are primarily governed by transcriptional control, while the broad class is more dependent on the chromatin accessibility and epigenetic marks (Lenhard *et al.*, 2012; Mora *et al.*, 2016).

The distal *cis*-regulatory elements such as enhancers, silencers and insulators are localised far from the TSS. Enhancers and silencers are short motifs containing binding sites for transcription factors, but they have the opposite regulation (Kolovos *et al.*, 2012). The enhancers function in amplifying transcriptional regulation while the silencers suppress the gene expression. These elements possess a bifunctional role (enhancers or silencers) depending on the specific enrichment of transcription factors and epigenetic marks (Bandara *et al.*, 2021; Gisselbrecht *et al.*, 2020). Meanwhile, insulators are boundary elements that demarcate the active domains and restrict the promoters' promiscuous interaction between enhancers or silencers (Gaszner and Felsenfeld, 2006). The interactions between the promoters and the distal regulatory elements determine the transcriptional activity in housekeeping and cell or tissue-specific regulations for development and disease (Ko *et al.*, 2017; Zabidi and Stark, 2016).

The transcription factor binding specificity and functional features are modulated by multiple simultaneous protein-protein interactions (PPIs) with other transcription factors and transcriptional coregulators (Ahsendorf *et al.*, 2017; Inukai *et al.*, 2017). Most transcription factors operate as homo- or heterodimers and exhibit cooperative binding to execute multiple regulatory roles (Funnell and Crossley, 2012; Morgunova and Taipale, 2017). Transcriptional coregulators are also required to activate or suppress chromatin-dependent transcriptional signalling after transcription factors bind to cis-regulatory DNA sequences. Transcriptional coregulators (coactivators and corepressors) are diverse functional classes of chromatin-associated proteins including, but not limited to, chromatin modifiers (writers and erasers), chromatin remodellers, chromatin readers, and other scaffolding proteins (Bishop *et al.*, 2019). Transcription factors and coregulators assemble diverse regulatory complexes at the *cis*-regulatory elements. Dysregulation of these assemblies has severe implications for cell homeostasis and often leads to disease development.

2.1.3 Histone modifications

Chromatin regulators conduct the dynamic modification of chromatin architecture known as chromatin remodelling for the accessibility of transcription machinery to the regulatory elements (Morgan and Shilatifard, 2020). Remodelling occurs via covalent histone modifications by enzymes such as histone acetyltransferases (HATs), histone deacetyltransferases (HDACs) and histone methyltransferases (HMTs) (Wang *et al.*, 2007). Histone modifications include acetylation, methylation, and ubiquitination. Other modifications such as crotonylation, 5-hydroxylation, sumoylation, ADP-ribosylation, and glycosylation also may occur at the histone polypeptides and their residues (Rothbart and Strahl,

2014). Each modification plays a crucial role in mediating the functional chromatin states for the proper organisation and efficient execution of nuclear processes (Adkins *et al.*, 2013).

The best-studied modifications are methylated and acetylated lysine residues on histone H3. They are deposited on particular areas of the genome and associated with distinct states of gene transcription. Histone acetylation is generally correlated with active transcription, histone methylation associates with a variety of transcriptional states, depending on the particular lysine position that is methylated. (Gates *et al.*, 2017). For example, histone-3-lysine-4-mono-, di- and trimethylation (H3K4me1/2/3), histone-3-lysine-36 mono- and tri-methylation (H3K36me1/3), histone-3-lysine-27-acetylation (H3K27ac) and histone-3-lysine-9-acetylation (H3K9ac) are located in actively transcribed regions and accumulated near the euchromatin. On the other hand, the heterochromatic regions are often marked by histone-3-lysine 9 (H3K9), histone-3-lysine-27 (H3K27) and histone-4-lysine-20 methylation (H4K20me) and lack of histone acetylation. Table 2.1 lists the genomic localisation of notable histone H3 modifications associated with transcription states of chromatin (Gates *et al.*, 2017; Zhao and Garcia, 2015). Another transcription state is the poised state of chromatin by the bivalency of histone marks. Bivalent chromatin is the genomic feature co-occupied by active H3K4me3 and repressive H3K27me3, mainly at the promoters (Bernstein *et al.*, 2006; Blanco *et al.*, 2020). Bivalency regulates gene expression for cell fate and tissue lineage commitment in development (Harikumar and Meshorer, 2015; Jeon and Tucker-Kellogg, 2020).

Table 2.1 Transcriptional role and localisation of histone modifications (Gates *et al*., 2017; Zhao and Garcia, 2015).

Histone modification	**Transcriptional role**	**Location of enrichment**
H3K4me1	Activation	Enhancers
H3K4me3	Activation	Promoters
H3K27ac	Activation	Enhancers, promoters
H3K9Ac	Activation	Enhancers, promoters
H3K36me3	Activation	Gene bodies
H3K79me3	Activation	Gene bodies
H3K27me3	Repression	Promoters, gene-rich regions
H3K9me3	Repression	Satellite repeats, telomeres, pericentromeres

Histone-modifying enzymes are chromatin regulators that mediate histone modifications that frequently operate as large multiprotein complexes (DesJarlais and Tummino, 2016). The histone-modifying enzymes are grouped into writers, erasers, and readers (Biswas and Rao, 2018; Hojfeldt *et al.*, 2013). The 'writers' are enzymes that set the post-translational modification to the residues on histones. The covalently modified histone tails are recognised via specific domains of a group of proteins referred to as 'readers'. The enzymes that remove the post-translational modifications are called 'erasers'. These enzymes and protein domains modify and read the specific amino acids on the histones and form the basis of the dynamic epigenetic regulation of gene expression (Figure 2.2).

Among the writers are the Polycomb group (PcG) and Trithorax group (TrxG) proteins which play a fundamental role in regulating histone modifications connected with gene repression and activation, respectively (Blanco *et al.*, 2020; Kuroda *et al.*, 2020). The TrxG complexes display cooperativity with PcG complexes through their opposing roles in regulating gene expression and development (Kadoch *et al.*, 2016; Kuroda *et al.*, 2020). PRC2 is a PcG complex that maintains the repression of gene expression by depositing H3K27me3 at the gene regulatory elements (Gaydos *et al.*, 2014). Meanwhile, the TrxG complex deposits active H3K4me3 to oppose the PcG silencing (Tie *et al.*, 2014). Both complexes are also responsible for establishing the bivalent domains for poised chromatin and may act as the master switch that coordinates the transcriptional programming (Kuroda *et al.*, 2020).

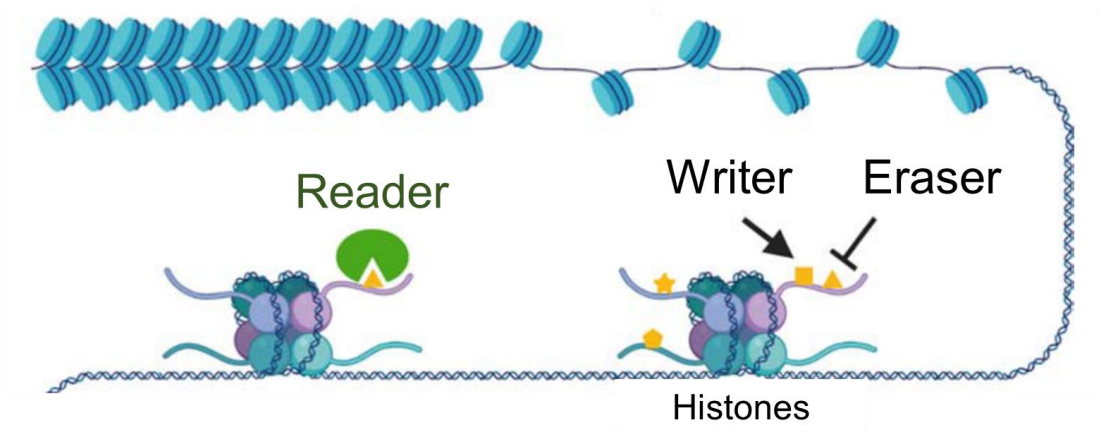


Figure 2.2 The regulators of histone modifications. Histone post-translational modifications are added or removed by specific enzymes (‘writers’ and ‘erasers’, respectively) and recognised by their binding proteins (‘readers’). The image was adapted from Cao and Yan (2020).

2.2 Transcription Factor: CCCTC-binding Factor (CTCF)

2.2.1 Background

CTCF is a highly conserved zinc-finger DNA-binding protein initially discovered as a transcriptional repressor of c-myc (Filippova *et al.*, 1996; Klenova *et al.*, 1993). CTCF plays multiple roles in transcriptional activation and repression, regulation of imprinted loci and X chromosome inactivation (XCI) and organisation of higher-order chromatin by chromosomal looping and nuclear tethering (Braccioli and de Wit, 2019; Herold *et al.*, 2012; Phillips and Corces, 2009; Wu *et al.*, 2020). CTCF expresses in all stages of development across most metazoan tissues (Phillips and Corces, 2009). Notably, CTCF regulates the lineage specification during cell/tissue-type specific gene expression by controlling the transcriptional regulation and chromatin organisation (Arzate-Mejia *et al.*, 2018; Braccioli and de Wit, 2019).

CTCF comprises an N-terminal domain, a central DNA binding domain with 11 C2H2 ZFs and a C-terminal domain (Klenova *et al.*, 1993). The plethora of CTCF roles is contributed by the three domains that determine the binding site preference and regulation of chromosome organisation (Nishana *et al.*, 2020; Ong and Corces, 2014; Pugacheva *et al.*, 2020). The dynamic combination of ZFs binds to different DNA sequences and interacts with various protein factors (Hashimoto *et al.*, 2017; Nakahashi *et al.*, 2013). In addition, CTCF may interact with other proteins, RNA and susceptible to post-translational modifications that could affect interactions with DNA or other proteins, as reviewed elsewhere (Arzate-Mejia *et al.*, 2018; Braccioli and de Wit, 2019; Wu *et al.*, 2020; Zlatanova and Caiafa, 2009).

2.2.2 CTCF target sites (CTSes)

CTCF is a sequence-specific transcription factor that regulates 3D genome organisation and critical aspects of gene expression such as transcription activation

and repression, RNA splicing, and enhancer/promoter insulation. (Ong and Corces, 2014). CTCF target sites (CTSes) are highly conserved and widely spread across the mammalian genome. A vast number of more than 80,000 CTSes may be found in the mammalian genome (Chen *et al.*, 2012a; Maurano *et al.*, 2015). CTCF binding is predominant to the distal regions and gene bodies than the TSS (Handoko *et al.*, 2011; Holwerda and de Laat, 2013). Still, CTCF enrichment near TSS is crucial for forming transcription or enhancer hubs through long-distance chromatin interactions that bridge distal enhancers with target promoters (Guo *et al.*, 2012; Kubo *et al.*, 2021).

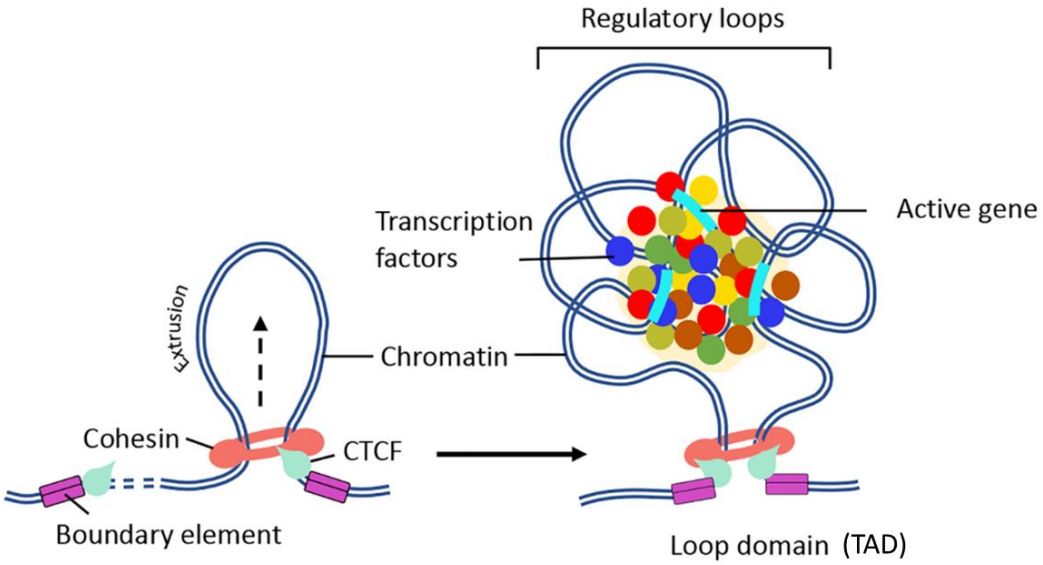
CTCF forms a complex with cohesin, a key component of chromatin and binds to the DNA to establish the DNA loop domains (Hansen *et al.*, 2017; Lee and Iyer, 2012) (Figure 2.3A). The loops facilitate the interactions between promoter and enhancer elements. A loop domain, known as topologically associated domains (TAD), is formed by multiple loops of intervening DNA between two convergent CTCF sites. TADs form the basis of chromatin regulatory segmentation that demarcate the gene regulatory elements (Merkenschlager and Nora, 2016; Xiang and Corces, 2020).

CTCF exhibits dynamic binding to the cognate genomic sites, recognised by diverse combinations of its 11 ZFs as a “multivalent protein” (Xu *et al.*, 2018a). The systematic mapping of CTCF ZFs to specific chromatin sites enables CTCF to execute diverse functions in different contexts and cell types. CTCF core motif consists of about 20 bp, in which the CTCF ZF domain employs ZFs three to seven to bind to the 15 bp consensus sequence. Meanwhile, the ZFs nine to eleven modulate CTCF-binding stability by interacting with the second CTCF motif located at 21 to 22 bp upstream to the core motif (Guo *et al.*, 2015; Hashimoto *et al.*, 2017;

Nakahashi *et al.*, 2013; Persikov and Singh, 2014; Xie *et al.*, 2007; Xu *et al.*, 2018a). The core sequences can be partial palindromic, and the directionality is determined by the second motif (Guo *et al.*, 2015; Wu *et al.*, 2020; Xie *et al.*, 2007). The CTCF binding motifs are displayed in Figure 2.3B, adapted from an extensive review by Wu *et al.* (2020).

CTSes elements can exist as single non-clustered and clustered CTCF sites. CTCF clustered sites represent a group of at least two CTSes separated by less than ten kb on the genome. The vast majority of clustered CTSes colocalise with cohesin and enrich near transcription start sites to stabilise the chromatin contacts (Kentepozidou *et al.*, 2020). Meanwhile, the single CTSes function as the insulators that restrict the interactions between promoters and distal enhancers (Jia *et al.*, 2020).

A.



B.

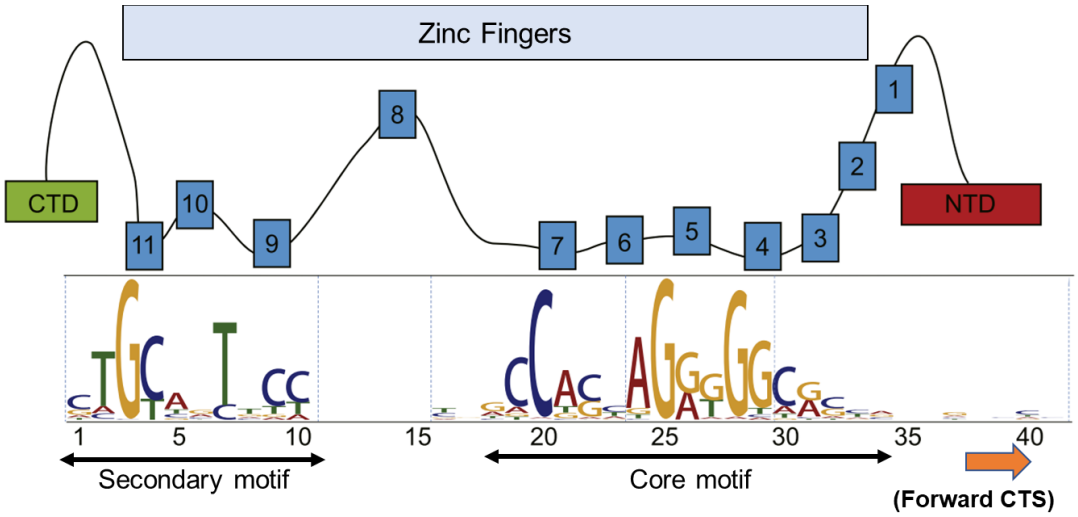


Figure 2.3 The formation of chromatin loops and CTCF binding motifs. **A.** CTCF and cohesin generate chromatin loops through loop extrusion and form TAD. The image was adapted from Pongubala and Murre (2021). **B.** The combination of CTCF ZFs binds to the core and secondary binding motifs. The image was adapted from Wu *et al.* (2020).

.

2.2.3 CTCF in the development and cancer

The CTCF-mediated interactions are also closely related to developmental gene regulation (Allahyar *et al.*, 2018; Merkenschlager and Nora, 2016). CTCF is required for embryonic development since CTCF-null embryos cannot implant (Moore *et al.*, 2012). CTCF mediates chromatin looping in the pluripotent genome and is essential to determine the stem cells' fate (Rao *et al.*, 2014). As an architectural protein, CTCF controls the expression of pluripotency genes and lineage-specifying genes by organising the chromatin state (Balakrishnan *et al.*, 2012; Dowen *et al.*, 2014; Handoko *et al.*, 2011; Nora *et al.*, 2017; Pekowska *et al.*, 2018). The pluripotency genes such as *Nanog* and *Oct4* are localised within the CTCF-anchored super-enhancer domain. The domains insulate the super-enhancers from interacting with lineage-specific master transcription factors and maintain the stem cell state (Dowen *et al.*, 2014; Justice *et al.*, 2020; Whyte *et al.*, 2013).

CTCF contributes to the lineage specification and development in multiple tissues, including brain, cardiovascular, limb, muscle, retina, immune cells, and gametes (Arzate-Mejia *et al.*, 2018). The chromatin has a flexible structure with weak organisation and high accessibility in the pluripotent stem cells (Schlesinger and Meshorer, 2019). Gain and enhancement of chromatin loops, especially at CTCF binding sites, mark the exit from pluripotency to differentiation (Pekowska *et al.*, 2018). The chromatin becomes more compartmentalised as the developmental program activates and progresses to more lineage-specific regulation (Zheng and Xie, 2019).

CTCF binding at the promoters mediates the promoter-promoter and promoter-enhancer interactions for active gene expression during cellular differentiation (Kubo *et al.*, 2021). An architectural protein, YY1 interacts with