FUNCTIONAL CHARACTERIZATION OF THE INTERACTIONS BETWEEN CTCF TRUNCATED TRANSCRIPTIONAL FACTOR WITH Y-BOX BINDING PROTEIN-1 AND THE CTD OF POL II

DARULIZA KERNAIN BINTI MOHD AZMAN

UNNIVERSITI SAINS MALAYSIA

2016

FUNCTIONAL CHARACTERIZATION OF THE INTERACTIONS BETWEEN CTCF TRUNCATED TRANSCRIPTIONAL FACTOR WITH Y-BOX BINDING PROTEIN-1 AND THE CTD OF POL II

by

DARULIZA KERNAIN BINTI MOHD AZMAN

Thesis submitted in fulfilment of the requirement

for the degree of

Doctor of Philosophy

September 2016

ACKNOWLEDGMENTS

Thanks to the Almighty Allah for giving me strength and ability to understand, learn and complete this thesis. I would like to express my immeasurable appreciations and deepest gratitude to my Supervisor, Prof. Shaharum Shamsuddin who believed in me and let me grew in his lab to who I am today. I am also deeply thankful to Prof. Elena Klenova, who supervised me during the 3 months attachment in Essex University, UK. I am also very thankful to my Co-Supervisors Associate Prof. Dr See Too Wei Chun and Dr. Tan Suat Cheng for their invaluable advices and guidance throughout my PhD journey.

This thesis would not be possible without the love and support of family members especially my beloved parents Mohd Azman Zulkernain and Kazrina Abdullah for allowing me to realize my own potential and believe in my dreams. I am much thankful for the love and supports given by my siblings during the ups and downs of my life. I would also like to thank all the lab members, lab technologists, scientific officers for their help throughout my stay both in USM Malaysia and Essex University, UK. Lastly I would like to thank Ministry of Higher Education Malaysia and USM for the financial assistance ASTS (Academic Staff Training Scheme) and Universiti Sains Malaysia-Research University grant (1001/PPSK/813074) without with my dream of accomplishing this thesis would not be possible

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

AGRP	Agouti Related Neuropeptide
AMKL	megakaryoblastic leukaemia
Amp	Ampicillin
AP-FRET	Acceptor Photobleaching Fö rster Resonance Energy Transfer
APP	Amyloid Protein Precursor
APS	Ammonium persiflage
Bp	Base pairs
BSA	Bovine serum albumin
BORIS	Brother of the regulator of imprinting sites
$CaCl_2$	Calcium chloride
CDH1	Cadherin 1, Type 1
CDH13	Cadherin 13
CDKs	Cyclin-dependent Kinases
cDNA	Complementary DNA
CKII	Casein kinase II
CSD	Cold Shock Domain
CTCF	CCCTC sequence binding factor
CTD	C-terminal domain
CTS	CTCF target sequences
DAPI	4',6-diamidino-2-phenylindole dilactate

DMD	Differentially methylated region
DMSO	Dimethylsulfoxide
DNase	Deoxyribonuclease
DNMT1	DNA (cytosine-5)-methyltransferase 1
ECL	Enhanced chemiluminescence immunodetection
EDTA	Ethylenediaminetetraacetic acid
EGFP	Enhanced green fluorescence protein
FBS	Fetal bovine serum
FITC	Fluorescin isothiocyanate
FOG	Friend of GATA-1
GNAO1	Guanine Nucleotide Binding Protein (G Protein)
GTF	General Transcription Factor
HDAC	Histone deacetylase
HRP	Horseradish peroxidase
ICR	Imprinting control region
IHC	Immunohistochemistry
IF	Immunofluorescence
Inr	Initiator
kDa	Kilodalton
KIFC3	Kinesin Family Member C3
LB	Luria Bertani
LOH	Loss of heterozygosity
LOI	Loss of imprinting
LS Pol II	Largest subunit of RNA Pol II

MC1R	Melanocortin 1 Receptor (Alpha Melanocyte Stimulating Hormone
mRNA	Messenger RNA
mTFP	monomeric Teal Fluorescent Protein
MW	Molecular Weight
NaB	Sodium butyrate
NFATC3	Nuclear Factor Of Activated T-Cells, Cytoplasmic, Calcineurin-Dependent 3
NLS	Nuclear localisation signal
ORF	Open reading frame
RNA	Ribonucleic acid
RPM	Revolutions per minute
RT	Room temperature
PAR	Poly(ADP-ribose)
PARG	Poly(ADP-ribose)glycohydrolase
PARP	Poly(ADP-ribose)polymerase
PBS	phosphate buffered saline
PCNA	proliferating cell nuclear antigen
PLCG2	Phospholipase C, Gamma 2 (Phosphatidylinositol-Specific)
RGBM	Recurrent Glioblastoma Multiforme
SDS	Sodium Dodecyl Sulfate
SDS-PAGE	Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis
SRO	small region of overlap
SUMO	Small ubiquitin-like modifier
SRO	smallest region of overlap
TBP	TATA binding protein

TEMED	N,N,N’,N’,-tetramethylehylenediamine
TRE	Thyroid hormone responsive element
TSS	Transcription start site
UTR	Untranslated region
WB	Western Blot

PENCIRIAN FUNGSI INTERAKSI ANTARA FAKTOR TRANSKRIPSI CTCF TERPANGKAS DENGAN PROTEIN PENGIKAT KOTAK Y-1 DAN TERMINAL C DOMAIN POL II

ABSTRAK

CTCF adalah faktor transkripsi jejari 11-Zn, yang terlibat dalam pengawalan transkripsi, penebat, kawalan cetakan, dan pentakaktifan kromosom X. Dua daripada protein yang dapat berinteraksi dengan CTCF adalah YB-1 dan RNA Polymerase II. Protein yang pertama ialah faktor transkripsi YB-1 yang terdapat dalam kalangan Y-kotak, dan protein yang kedua ialah Subunit Besar RNA polimerase II (LS Pol II) yang merupakan enzim utama untuk transkripsi. Berdasarkan kajian-kajian terdahulu, interaksi antara CTCF dan YB-1 berlaku pada ZF dan CSD domain manakala interaksi antara CTCF dan RNA Poll II berlaku pada C domain. Objektif kajian ini adalah untuk mengenalpasti interaksi anatara CTCF dan YB-1 di dalam titisan sel RGBM dan mengenalpasti motif pada CTCF C domain yang terlibat dalam interaksi dengan RNA Pol II. Hasil cerakin co-IP menunjukkan CTCF dapat membentuk kompleks dengan YB-1 dalam keseluruhan lisis sel RGBM. Interaksi antara kedua-dua protein tersebut turut digambarkan melalui cerakin *in vitro pull-down*. Berdasarkan hasil kajian, CTCF-ZF adalah satu-satunya

domain yang berjaya diikat dengan YB-1 CSD, manakala domain yang selebihnya tidak menunjukkan sebarang interaksi. Selain itu, kajian ini juga berjaya menentukan interaksi antara CTCF dan LS RNA Pol II. Kajian terdahulu telah melaporkan interaksi antara kedua-dua protein tersebut berlaku di kawasan terminal C. Dua tapak dalam domain terminal C CTCF telah menunjukkan kepentingan dengan pengikatan Pol II CTD. Kajian ini turut menganalisis tujuh mutan varian CTCF C untuk mengenal pasti kesesuaian untuk diikat dengan Poll II CTD. Hasil co-IP menunjukkan interaksi positif antara semua mutan varian CTCF C dengan LS Poll II dalam titisan sel RGBM. Interaksi ini digambarkan menggunakan cerakin pull-down. Hasil cerakin *pull-down* menunjukkan motif EEEE merupakan kawasan penting bagi interaksi dengan Poll II CTD. Akhirnya, kesan CTCF C mutan varian pada kematian sel oleh apoptosis menunjukkan CTCF Complete Mutant (CM) memberi peratusan tertinggi dalam menyebabkan apoptosis dalam sel. Sebagai kesimpulan, CTCF ZF domain telah dikenalpasti sebagai domain yang dapat berinteraksi dengan YB-1 didalam sel RGBM. Selain itu, motif penting di dalam CTCF C domain telah Berjaya dikenalpasti di dalam kajian ini dimana EEEE motif merupakan motif yang penting bagi interaksi dengan RNA Pol II. Akhirnya, kajiian lanjutan diperlukan bagi menganalpasti implikasi motif in dalam interaksi antara CTCF and RNA Poll II.

FUNCTIONAL CHARACTERIZATION OF THE INTERACTIONS BETWEEN CTCF TRUNCATED TRANSCRIPTIONAL FACTOR WITH Y-BOX BINDING PROTEIN-1 AND THE CTD OF POL II

ABSTRACT

CCCTC sequence binding factor (CTCF) involved in the regulation of transcription, insulator function, control of imprinting and the X-chromosome inactivation. Two of the CTCF protein interacting partners were the transcriptional factor YB-1, a member of the Y-box binding protein 1 and the second protein was the large subunit of RNA polymerase II (LS Pol II) the principal enzyme for transcription. Previous studies have shown the interaction between CTCF and YB-1 occurred at the Zinc finger (ZF) and Cold Shock Domain (CSD) respectively whereas the interaction between CTCF and RNA Pol II occurred at the C-terminal domains (CTD) for both proteins. The objectives of this study are to determine the interaction between CTCF and YB-1 in Recurrent Glioblastoma Multiforme (RGBM) cell line and to identify the motif within CTCF CTD that is critical for binding with RNA Pol II CTD. CTCF truncated proteins were produced and used to map the interaction to YB-1 in RGBM cell line. The interaction of these truncated proteins was mapped using co-immunoprecipitation (Co-IP) and pull-down assay and the functional significant of the interaction was characterized via mammalian two hybrid system. On the other hand, CTCF and RNA Poll II interactions

were characterized by identifying two sites within CTCF C-terminal domain important for the interaction and seven candidates CTCF C mutant variants named 1A, 1B, 1A1B, 2A, 2B, 2A2B and CM deficient for binding with Poll II CTD were produced. The interaction of these mutants to Poll II CTD was characterized via Co-immunoprecipitation (Co-IP) and pull-down assay, while the ability of these mutants to induce apoptosis in recurrent glioblastoma multiforme (RGBM) cell line was also analysed using Caspase 3/7 Glo assay. From the results obtained for CTCF and YB-1 interactions, the co-immunoprecipitation (co-IP) showed that CTCF was able to form a complex with YB-1 in the RGBM total cell lysate. From pull down assay results, CTCF-ZF was the only domain binds with YB-1 Cold Shock Domain (CSD) and the rest of domains fail to interact. As for CTCF and RNA Pol II interactions, the Co-IP results shown positive interaction for CTCF wild-type and all the mutant variants in RGBM cell line. The interaction was further characterized using pull-down assay and EEEE motif was identified as a critical motif for binding to Pol II CTD. The effects of CTCF C mutant variants on cell death by apoptosis assay showed that CTCF complete mutant (CM) induced highest apoptosis percentage in the cell as compared to the wild-type CTCF. As conclusion, CTCF ZF domain was found to interact with YB-1 CSD in RGBM cell line and the important motif in CTCF C terminal region critical for Poll II CTD binding was successfully identified. The EEEE motif was found to be a critical motif for binding; hence further study is needed to determine the significant of these motifs in CTCF and RNA Poll II interactions.

CHAPTER 1

INTRODUCTION

1.1 GENERAL CHARACTERISTICS OF CTCF TRANSCRIPTION FACTOR

CTCF is a multifunctional transcription factor that ubiquitously expressed in many species. It was categorized as a zinc finger transcriptional factor and it was reported to bind to the multiple target sites using different combinations of its 11 zinc finger domains (Filippova *et al.*, 1996). It is known as 11 zinc fingers transcriptional factor and highly conserved from *Drosophila melanogaster* to humans (Ladomery and Dellaire, 2002).

CTCF was originally identified as a protein that binds to the three repeats of the core sequence of CCCTC spaced ranging from 12-13 bp intervals within the chicken *c-Myc* promoter and was named as CCCTC Binding Factor, or CTCF. However, subsequent studies had shown CTCF was able to bind to various sequences in the human, mouse and avian *c-Myc* promoters suggesting its ability to have multiple roles in the biological functions (Lobanenkov *et al.*, 1991).

CTCF was reported to be a multifunctional protein with the ability to bind to the genomic DNA sites ~40,000 within the mouse embryonic stem cell (Chen *et al.*, 2008). It is functions include promoter activation or repression, hormone-responsive gene silencing, regulation of cell growth and apoptosis. In addition, CTCF was also reported to involve in the chromatin barrier, insulation and genomic imprinting (Ohlsson *et al.*, 2001).

1.1.1 The *CTCF* gene

CTCF is a single copy gene and human *CTCF* is localized at the chromosome region 16q22.1 (Figure 1.1) which is often associated with genomic alterations in human tumours including breast and prostate (Filippova *et al.*, 1998). The *CTCF* cDNA was first isolated, cloned and sequenced from chicken by Klenova *et al*. (1993) and the open reading frame (ORF) of *CTCF* gene is highly conserved between different species to which it was reported there was 93 % similarity between avian and human.

1.1.2 The structure of the CTCF protein

The CTCF protein is composed of three main domains which are N, Zinc Finger (ZF) and C (Figure 1.2) (Filippova *et al.*, 1996). The ZF domain of CTCF was reported to contain ten fingers of C_2H_2 type and one finger of C_2HC type (Fox *et al.*, 1999). The third motif within the C-terminal region of CTCF contains two repeats of the PXXP-signature characteristic of SH3-domain binding proteins. The protein was reported to consist of two isoforms which are CTCF-130 and CTCF-180.

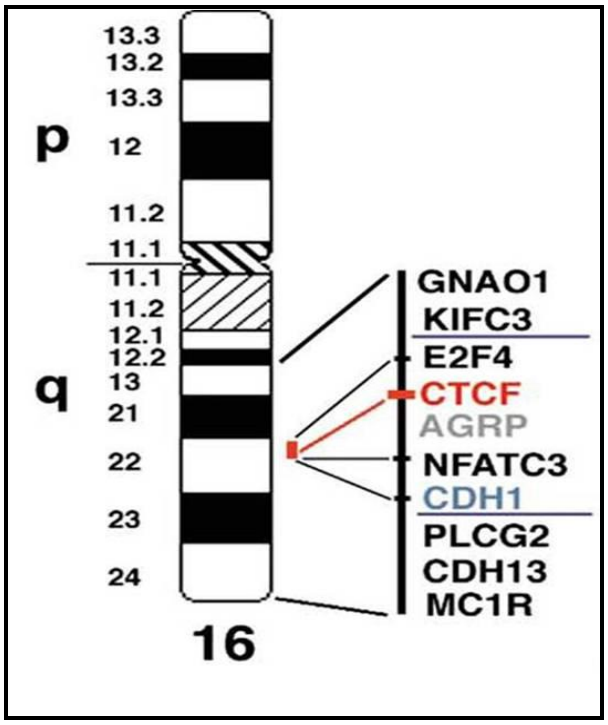

Figure 1. 1 Human CTCF gene localized at the chromosome region 16q22.1 (Klenova *et al.*, 2002). The abbreviation of each gene is described as follow:

GNAO1 : Guanine Nucleotide Binding Protein (G Protein)

KIFC3 : Kinesin Family Member C3

E2F4 : E2F Transcription Factor 4

CTCF : CCCTC-Binding Factor (Zinc Finger Protein)

AGRP : Agouti Related Neuropeptide

NFATC3 : Nuclear Factor Of Activated T-Cells, Cytoplasmic, Calcineurin-Dependent 3

CDH1 : Cadherin 1, Type 1

PLCG2 : Phospholipase C, Gamma 2 (Phosphatidylinositol-Specific)

CDH13 : Cadherin 13

MC1R : Melanocortin 1 Receptor (Alpha Melanocyte Stimulating Hormone Receptor)

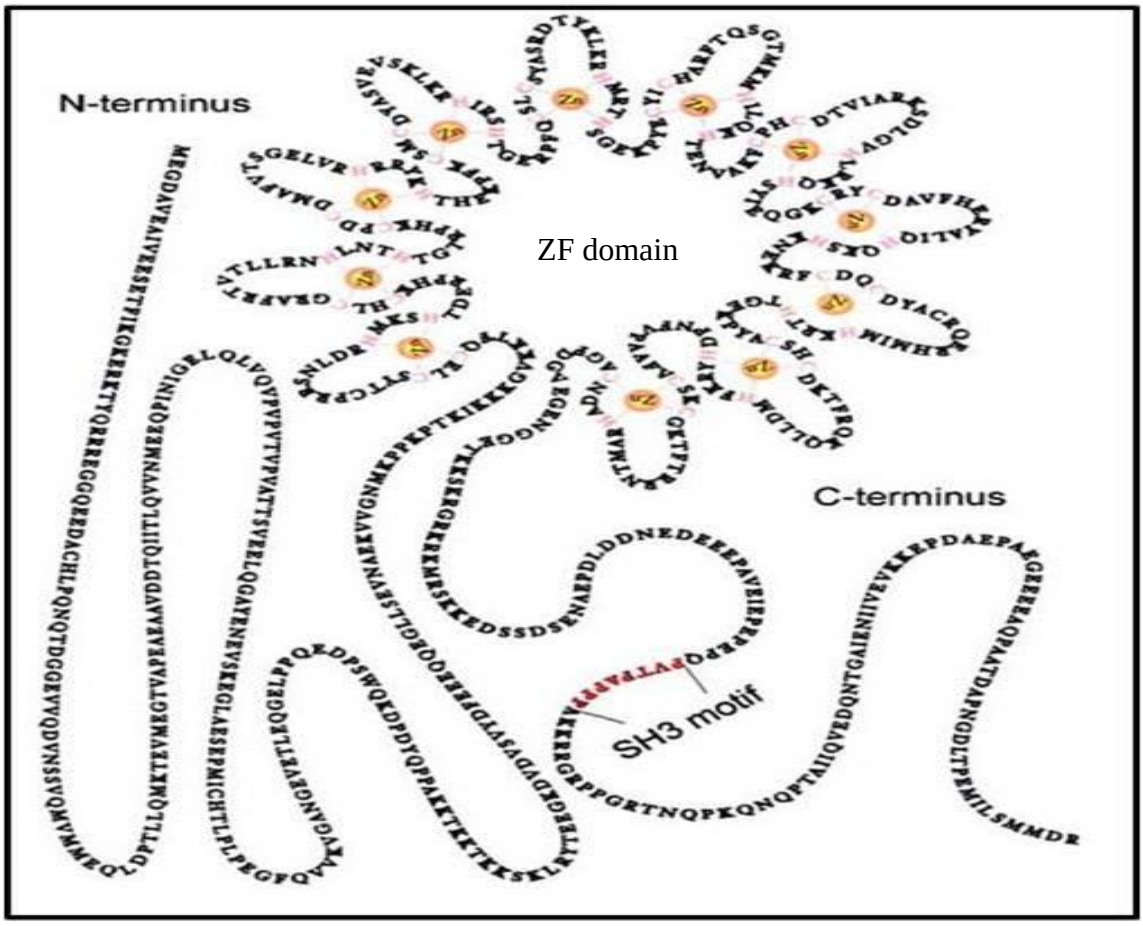

Figure 1. 2 The structure of the CTCF protein. It consists of three main domains which are N, ZF and C (Filippova *et al.*, 1996). It consists of PXXP motif which is a characteristic of SH3 binding domain.

The first isoform, CTCF-130 cDNA was first cloned from the chicken genome with an open reading frame (ORF) of 727 amino acids corresponding to a polypeptide chain of 82 kDa (Klenova *et al.*, 1993). However in SDS-PAGE, this isoform migrates as a 130 kDa protein (CTCF-130) and this could be due to the modifications of its N and the C-terminal domains (Klenova *et al.*, 1997). The second isoform of CTCF was discovered by the same group much later and it was reported to migrate in SDS-PAGE at 180 kDa. The increase in molecular weight could be due to some post-translational modification such as poly (ADP-ribosyl)ation (PARylation) that takes place in the cell.

CTCF carries a unique feature to which it recognizes diverse DNA target sites through a combination of its 11 ZFs domain. This means each particular target sites utilizes combination of different ZFs for instance in chicken, CTCF utilized ZFs 2 to 7 to bind to the c-*Myc* promoter while in human, CTCF utilizes ZFs 4 to 11 to bind to c-*Myc* promoter (Ohlsson *et al.*, 2001).

1.2 POST-TRANSLATIONAL MODIFICATIONS OF CTCF

The post-translational modification that takes place in the protein was reported to modulate their functions and affect their localization during the cell cycle. The post-translational modification process was reported to take place in CTCF which includes phosphorylation, SUMOylation and Poly(ADP-ribosyl)ation or also known as PARylation.

These processes were found to be important in the regulation of CTCF functions to which specific phosphorylation by Casein kinase II (CK2) affects its function in the transcription regulation while Poly (ADP-ribosyl)ation modification affects its activity in the insulation function (MacPherson *et al.,* 2009).

1.2.1 Phosphorylation

The involvement of CTCF phosphorylation in the regulation of differentiation was proposed earlier with the study conducted on the myeloid cell differentiation using four leukaemia cell lines: K562, HL60, U937 and THP1 (Delgado *et al.*, 1999). The study suggested that phosphorylation of CTCF may be responsible for the regulation of gene-specific lineage expression and may also be responsible for the activation of selected programmes in the biological functions (Delgado *et al.*, 1999). A Study conducted by Klenova *et al*. (2001) showed that CTCF was able to be phosphorylated by protein kinase CK2 both *in vivo* and *in vitro* particularly in highly conserved sequences which were found in the C-terminal region. This group suggested that CTCF phosphorylation could be an important event in the cell for the modulation of CTCF function in regards to the activation of *c-Myc* promoter sequence.

1.2.2 SUMOylation

Sumoylation is a post-translational modification process, which involves an addition of SUMOs (small ubiquitin-like modifiers) to the protein of interest. This process can affect the protein's structure and subcellular localization in the cell.

Previously reported study had shown that CTCF can be SUMOylated both in vivo and in vitro (MacPherson *et al.*, 2009). Two sites within CTCF that were reported to be SUMOylated are the N and the C-terminal domains. This group has suggested that SUMOylation does not change the DNA binding properties of CTCF though it enhances the CTCF repressive functions on the *c-Myc* P2 promoter. In addition, another study conducted in 2010 revealed that SUMOylation of CTCF enhances the repression and enhancer blocking functions by reducing the transactivation and chromatin opening (Kitchen and Schoenherr, 2010). Furthermore, this process was reported to involve in the regulation of nuclear architecture, particularly the organisation of chromatin loops (MacPherson and Sadowski, 2010). Moreover, a recent study revealed that de-SUMOylation of CTCF is linked to hypoxic stress although the mechanism of this process is still not well understood (Wang *et al.*, 2012).

1.2.3 Poly (ADP-ribosyl)ation (PARylation)

Poly (ADP) ribose polymerases (PARPs) is a type of protein mainly involved in the cellular processes involving DNA repair and programmed cell death (Beneke and Burkle, 2007). The earlier study had reported that CTCF was co-purified with PARP-1 from the nuclear fractions of HeLa cell lysates (Yusufzai *et al.*, 2004). However, the constitutive levels of this protein in the un-stimulated cells are usually low (D'Amours *et al.*, 1999; Farrar *et al.*, 2010) though in the presence of mitogenic stimuli or DNA damage the levels of poly-ADP-ribose may incline up to 500 times (Gagne *et al.*, 2006). Meanwhile, the researcher has conducted a study on the relation between PARylation with CTCF and reported that PARylation was essential for the

CTCF insulator function at the maternal ICR of the *H19*/*Igf2* locus (Yu *et al.*, 2004). Based on these findings, a model was proposed to which it was stated that PARylation of CTCF could be responsible in the formation of higher chromatin order structures at the maternal allele of the *H19/Igf2* locus (Klenova and Ohlsson, 2005).

1.3 THE BIOLOGICAL FUNCTIONS OF CTCF

The capacity of CTCF protein binds to various DNA sequences, recruiting numerous protein partners thus suggesting its functional complexity in the biological system. Among the regulatory roles play by CTCF are promoter activation or repression, chromatin insulator protein, genomic imprinting and genomic organization (Klenova *et al*., 2001). The details of each regulatory roles played by CTCF are described in the following pages.

1.3.1 CTCF as a transcriptional activator or repressor

CTCF may act as a positive regulators of human *amyloid β-protein precursor*, *APP* gene (Vostrov and Quitschke, 1997). It regulates the over-expression of this protein which is known to be a contributing factor in the neurodegenerative disorders such as Alzheimer's disease and Down's syndrome (Mann *et al.*, 1990). It was reported that the CTCF N-terminal domain was responsible for the activation of the *APP* gene promoter (Vostrov *et al.*, 2002). Besides transcriptional activator, CTCF may also act as a transcriptional repressor. Previously reported study has shown that CTCF represses mouse, chicken and human *c-Myc* genes (Filippova *et al.*, 1996; Lobanenkov *et al.*, 1990). The repression process carried out by CTCF was essential in the regulation of

cellular functions such as cell cycle progression, differentiation and apoptosis (Marcu *et al.*, 1992; Packham and Cleveland, 1995).

1.3.2 CTCF functions as a chromatin insulator protein

The genome of higher eukaryotes consists of active and inactive genes separated by DNA sequences called chromatin insulators. The enhancer-blocking activities of CTCF were first described with the discovery of CTCF binding to 5'HS4 DNase I hypersensitive site (HS) of the insulator sequences upstream to the chicken β-*globin* locus (Bell *et al.*, 1999). In humans, CTCF was shown to bind to two sites within the 5' and 3' flanking regions of the β-*globin* locus (Farrell *et al.*, 2002). Studies of the spatial chromosomal organization in β-globin locus revealed that CTCF mediates cell type specific interactions in which β-globin domain folds into a globule with the LCR and the active globin genes on the periphery (Junier *et al.*, 2012).

1.3.3 CTCF and genomic imprinting

The genetic information of eukaryotes is stored in two copies and in the mammalian systems, hundreds of genes exhibit parental imprinting. The gene expression in the biological system is in a selective manner to which and it can be attributed either from the mother or father. Genomic imprinting relies on the parental-specific methylation of DNA elements which control the differential expression of maternal or paternal alleles. Imprinted genes are often organized in the clusters such as *insulin-like growth factor 2 (Igf2/H19)* genes. The *Igf2/H19* genes were first identified as imprinted genes (Ferguson-Smith *et al.*, 1991).

Both genes are located 80 kb apart in the genome and the monoallelic expression of *Igf2/H19* locus is regulated by an imprinting control region (ICR) located 2 kb to 4.4 kb relative to the transcription start site of *H19*. It was proposed that the insulator factor is present at ICR in order to block the enhancer function at the maternal *Igf2* gene.

Studies have shown that *CTCF* involves in the differential methylation of *Igf2/H19* locus (Bell and Felsenfeld, 2000; Hark *et al.*, 2000; Kanduri *et al.*, 2000). The methylation process by *CTCF* occurred when the maternally inherited chromosome is unmethylated and CTCF binds to this site, resulting in the blocking of *Igf2* gene activation by the enhancer. On the paternal chromosome, the ICR is methylated, thus preventing *CTCF* binding. As a result, the enhancer can bind to the *Igf2* promoter and initiate its transcription of *H19* gene.

1.4 CTCF AS A CANDIDATE TUMOUR SUPPRESSOR

The tumor suppressor genes encode proteins that suppress the tumor formation. Tumor suppressors have diverse functions and are involved in various cellular processes which include cell cycle control, detection and repair of DNA damage, protein degradation, mitogenic signalling, cell specifications, tumor angiogenesis and others (Sherr, 2004). Numerous studies have demonstrated that genetic and epigenetic changes of tumor suppressor genes could lead to functional abnormalities often causing cancer development (Gronbaek *et al.*, 2007; Sadikovic *et al.*, 2008).

CTCF is reported to be the candidate of tumor suppressor gene because of few factors such as the loss of *CTCF* heterozygosity locus in different malignancies, the association of specific ZF mutations in some cancers, the growth suppressive features of *CTCF* and the ability of *CTCF* to maintain epigenetic marks of growth- related genes via its insulator functions.

The previous report on the cytogenetic studies at the 16q22.1 locus demonstrated that deletions at this locus are often associated with various cancers including breast (Driouch *et al.*, 1997) and prostate (Latil *et al.*, 1997). These two studies demonstrated the loss of heterozygosity (LOH) in the regions where tumor suppressor genes were located and defined the smallest region of overlap (SRO) for commonly deleted genetic material from the 16q chromosome. Additionally, the re-arrangement of the exons within the ZF domain associated with sporadic breast and prostate tumors were observed, further suggesting *CTCF* involvement as a tumor suppressor gene (Filippova *et al.*, 1998).

Analysis of the DNA sequences of the *CTCF* gene from different tumors identified missense mutations in breast, prostate and Wilms' tumours. These mutations selectively affected the ability of *CTCF* to bind to the CTS in the genes (Filippova *et al.*, 2002). Review that is more recent reported on *CTCF* harbouring 13 different mutations analyzed in 825 clinical patients with primary breast cancer (Cancer Genome Atlas, 2012).

The previous study reported, the existence of acquired *CTCF* mutations in patients with transient abnormal myelopoiesis, mostly affecting infants with Down syndrome, that leads to acute megakaryoblastic leukaemia (AMKL) (Yoshida *et al.*, 2013). However, some studies suggested that mutations in the *CTCF* gene do not always account for cancer development but rather abnormal *CTCF* function and epigenetic defects could possibly contribute more to tumor development (Recillas-Targa *et al.*, 2011).

CTCF binds to two CTSs upstream of the *BRCA1* gene simultaneously with Sp1, which creates a protective barrier for DNA methylation leading to epigenetic silencing of these genes (Butcher *et al.*, 2004). A similar mechanism of action is observed at the *Rb* gene promoter, where it protects the promoter against epigenetic silencing (De La Rosa-Velazquez *et al.*, 2007). The additional mechanism was also proposed by Phillips and Corces (Phillips and Corces, 2009) for the same gene loci where it was suggested that CTCF could interact with a transcriptional factor such as Kaiso to form alternative loops, thus protecting this locus via 3D organization.

1.5 CTCF PROTEIN PARTNERS

CTCF has been shown to interact with a number of proteins, via N, ZF and C-terminal domains that modulate CTCF-dependent molecular processes (Figure 1.3) (Ohlsson *et al.*, 2010). One of the CTCF's unique features to perform diverse functions could be attributed to the combinatorial usage of different ZFs to bind to DNA sequences and interacts with proteins.

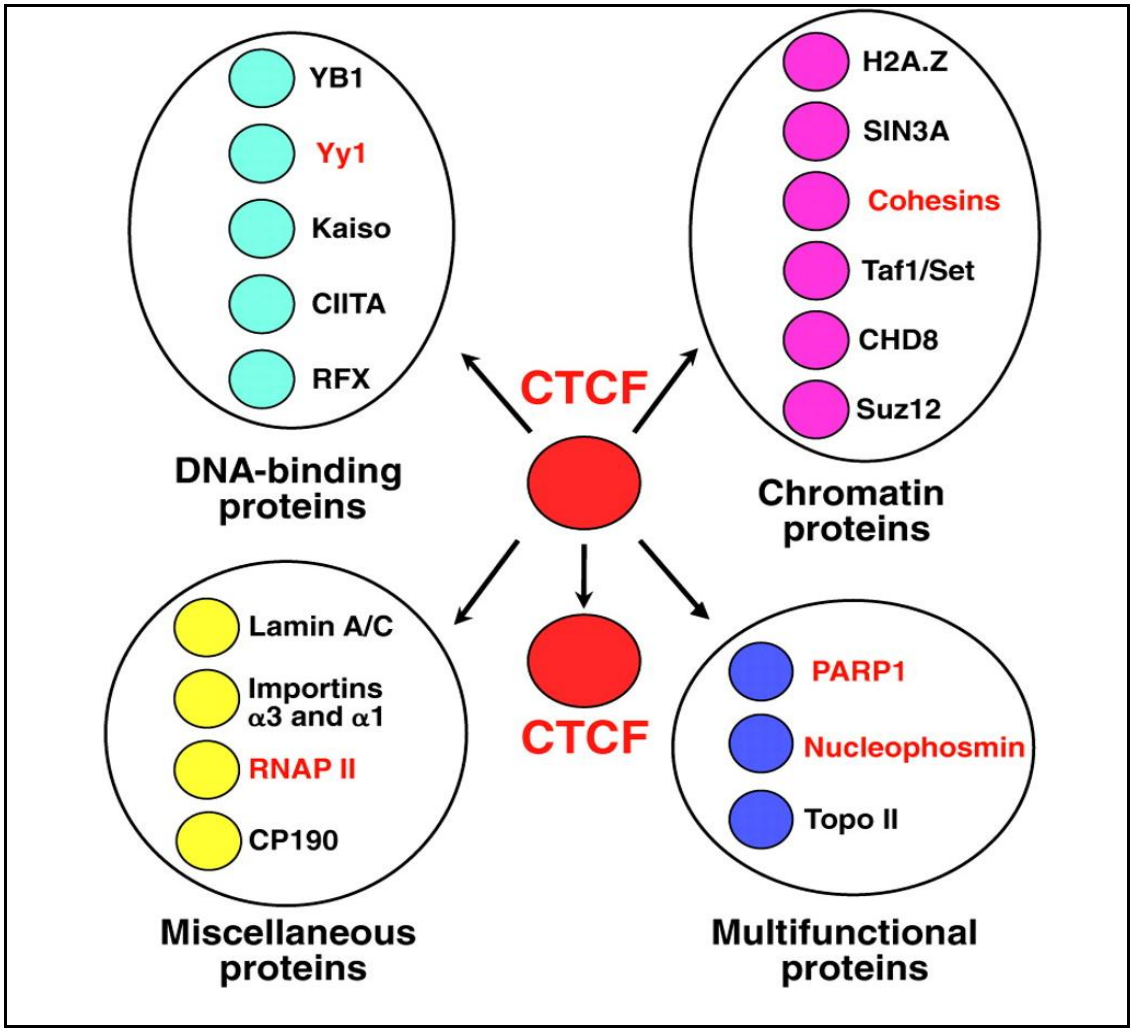

Figure 1. 3 Structure of CTCF with its known interacting partners (Ohlsson *et al.*, 2010).

The previous reports have shown that, CTCF insulator function is regulated by a few protein partners such as transcription factor Kaiso (Defossez *et al.*, 2005), chromatin helicase protein CHD8 (Ishihara *et al.*, 2006), nucleophosmin/B23 (Yusufzai *et al.*, 2004), and cohesin (Wendt *et al.*, 2008). CTCF was reported to form a complex with YY1 which this complex transactivate *Tsix* gene (Donohoe *et al.*, 2007). In addition, CTCF interaction with the multifunctional Y-box DNA/RNA binding factor, YB1 (Chernukhin *et al.*, 2000), and the co-repressor of histone deacetylases, Sin3A (Lutz *et al.*, 2000), resulted in the negative regulation of the *c-Myc* promoter. Other proteins reported to associate with CTCF include proliferating cell nuclear antigen, PCNA and breast cancer associated tumour suppressor protein BRCA-1 (Irminger-Finger *et al.*, 1999).

It has been reported that CTCF can homodimerize via ZF and C-terminal domains. Homodimerization of CTCF is said to involve in the 3D organization of the genome via the interaction between distant DNA regions (Pant *et al.*, 2004). Among nuclear architectural proteins CTCF interacts with are cohesin (Parelho *et al.*, 2008), nucleophosmin/B23, topoisomerase II (Yusufzai *et al.*, 2004) and components of nuclear lamina (Handoko *et al.*, 2011).

1.5.1 RNA Polymerase II

In prokaryotes, one DNA-dependent RNA polymerase is sufficient to transcribe all genes into the variety of RNA molecules needed for the cell's functioning (Egloff and Murphy, 2008).

However, that is not true for eukaryotes. Transcription of specific gene types is shared among three different RNA Polymerases. Polymerase I (Pol I) transcribes genes encoding 18S, 28S, and 5.8S rRNA; whereas Polymerase III (Pol III) transcribes a range of short genes including those encoding tRNA and 5S RNA (Egloff and Murphy, 2008). Polymerase II (Pol II) is responsible for transcription of all protein-coding genes and also small nuclear RNAs. All three polymerases have the largest catalytic subunits which share homology between each other and with the largest subunit of bacterial polymerase (Smale and Kadonaga, 2003).

1.5.1.1 CTCF interacts with RNA Pol II

CTCF is a multifunctional protein with protein partners specific for each function it performs (Zlatanova and Caiafa, 2009). CTCF was reported to interact with the Large Subunit (LS) Pol II through the series of assays, in which the interaction was found in holo-Pol II complex and not in the TFIIH complex. The previous study had reported that, the interaction of CTCF occurred at the C-terminal domain of LS Pol II (Chernukhin *et al.*, 2007).

1.5.1.2 General characteristic features of RNA Pol II enzyme

RNA polymerase II is a central enzyme in gene expression. It was classified into two forms which are “holo” polymerase and “core” polymerase (Figure 1.4). The “holo” enzyme of RNA Pol II consists of the “core” RNA Pol II and other complexes implicated in transcription, splicing, polyadenylation and DNA repair.

These complexes include, but are not limited, to the general transcription factors (GTFs), the core Srb-mediator complex, the Srb10 cyclin-dependent kinase (CDK) complex, and the Swi-Snf complex (Myer and Young, 1998). The molecular weight of the "holo" enzyme is over 1MDa (Jenkins and Spencer, 2001).

The "core" polymerase is a complex enzyme composed of multiple subunits (1-12 subunits, from 6 to 600 kDa) with a molecular mass of 500 kDa (Myer and Young, 1998). Genes which are encoding the subunits of the "core" enzyme are *RPB1-RPB12.* The two largest Pol II subunits, Rpb1 (220 kDa) and Rbp2 (150 kDa), are the most highly conserved and are homologous to the β and β' subunits of the bacterial polymerase. Additionally, three yeast polymerases were reported to have several subunits in common which are Rpb5, Rpb6, Rpb8, Rpb10 and Rpb12 (Sugaya *et al.*, 2001).

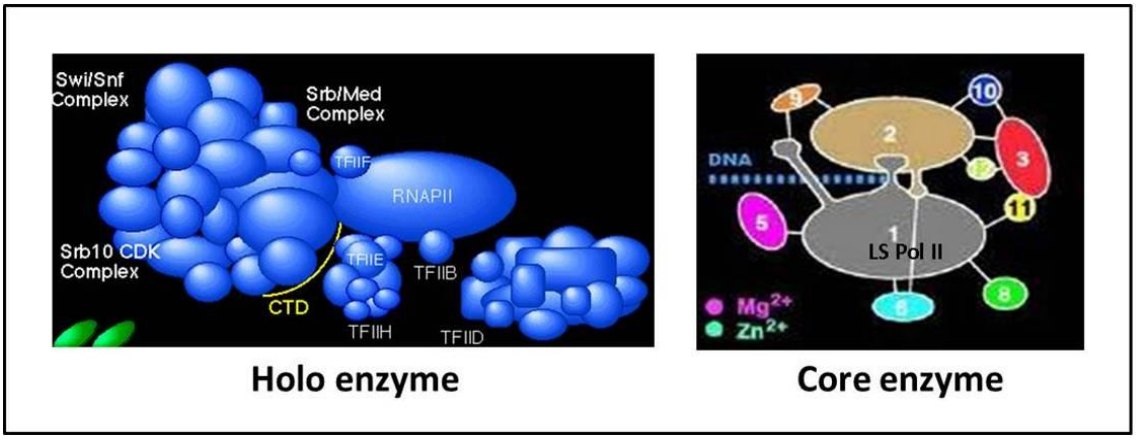

Figure 1. 4 Structure of the RNA Polymerase II (holo and core enzymes)

A schematic representation of the holo enzyme RNA Pol II (left) and the core enzyme (right). The core enzymes consists of 12 subunits (Rpb1-12), subunits 4 and 7 are not shown. The subunit 1 which represented in grey is the largest subunit of LS Pol II (Cramer *et al.*, 2000).

1.5.1.3 The largest subunit of the RNA Pol II (LS Pol II)

The largest subunit (LS Pol II) of the RNA Polymerase II, Rpb1, is the largest of all polypeptides associated with nuclear RNA polymerases. The ability of the LS Pol II to bind DNA and RNA is a unique feature of the LS Pol II, indispensable for its function (Wu and Chiang, 1998). The human LS Pol II consists of 1970 amino acids encoded by 5913 nucleotides (Mita *et al.*, 1995). It differs from other subunits in that it has a unique C-terminal domain.

1.5.1.4 The structure and functions of the C-terminal domain (CTD) of the LS Pol II

The C-terminal domain (CTD) of the Largest Subunit (LS Pol II) of the RNA Polymerase II is a unique domain of repetitive heptads (consensus sequence of $Y_1S_2P_3T_4S_5P_6S_7$) that extends from the catalytic core of the enzyme (Chapman *et al.*, 2008). This domain is heavily modified by phosphorylation, glycosylation, and proline isomerization, reviewed in (Corden, 2013). In addition to modifying enzymes, a number of RNA processing factors and chromatin modification factors also interact with this domain. This region plays a role in positioning the catalytic core and allows for interaction with other sub-units of RNA Pol II, namely Rpb2 and Rpb6 (Cramer *et al.*, 2001). The CTD consists of three regions: starting from linker (closest to the catalytic core), the heptad repeats and "tip" (Figure 1.5) (Corden, 2013). The consensus heptads have been evolutionarily conserved except for the number of repeats, which differ among the species (Dahmus, 1996).

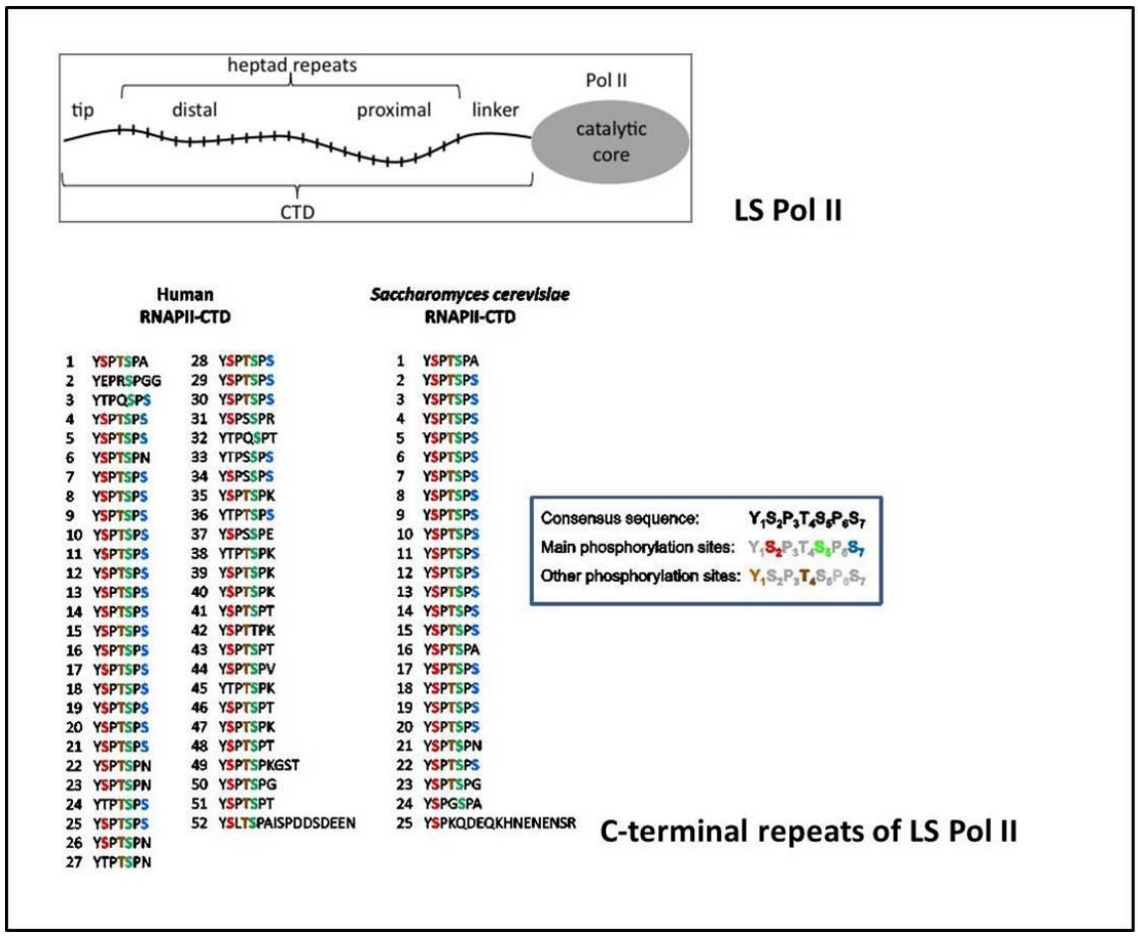

Figure 1. 5 Structure of the LS Pol II and the C-terminal domain (CTD)

The above figure illustrates a schematic representation of the largest subunit (LS Pol II) of the RNA Pol II and the C-terminal domain with repetitive heptads (Dahmus, 1996).

The CTD of the LS Pol II in mammalian cells has 52 repeats whereas yeast has 26-27 copies and other eukaryotes contain intermediate number of repeats (Dahmus, 1996). There is a clear relationship between the length of the CTD repeats and genomic complexity of an organism. There was a study suggested that CTD is largely unstructured however, phosphorylation can influence the propensity of the CTD peptides to form β-turns and polyproline helices (Buratowski, 2003).

The previous study reported that, multiple forms of the LS Pol II were identified by ion exchange chromatography, resulting in different forms termed as Pol IIo, IIa, and IIb which migrated differently in the SDS-PAGE. The most rapidly migrating IIb form was a proteolytic breakdown product lacking the CTD, whereas both forms IIo and IIa had the CTD but differed in their phosphorylation, in which IIo being highly phosphorylated. It was reported that, IIo form could be converted to the IIa form by a phosphatase treatment. The LS Pol II undergoes a reversible phosphorylation with the hypophosphorylated IIa form functioning in initiation and the hyperphosphorylated IIo form in elongation processes (Dahmus, 1995).

Previous studies with deletions of the CTD repeats showed that deletion of the 31 repeats interferes with viability in the cell (Meininghaus *et al.*, 2000). The nonconsensus 52^{nd} CTD repeat has an unusual 10 amino acids extension and shows a CKII phosphorylation (Chapman *et al.*, 2004), and is also a binding site for Abl1 tyrosine kinase (Baskaran *et al.*, 1999).

Deletion of the repeats 1 to 3 (Chapman *et al.*, 2005) and the repeat 52 resulted in the LS Pol II degradation yielding the LS Pol IIb form. The LS Pol II without 52^{nd} repeat has the reduced ability to transcribe (Chapman *et al.*, 2004) splice and 3'-end process pre-mRNA (Fong *et al.*, 2003).

1.5.2 Y-Box Protein Binding 1 (YB-1)

The study on YB-1 and CTCF interaction was initiated after the discovery that CTCF and YB-1 exist in a complex *in vivo* (Chernukhin *et al.*, 2000). YB-1 is a member of the family of Y-Box (an inverted CCAAT-box) binding factors (Matsumoto and Wolffe, 1998). It is a multifunctional factor which is expressed in a wide range of cells. It participates in the regulation of diverse cell functions such as transcription, replication cell proliferation DNA repair, DNA unwinding (Ladomery and Sommerville, 1995) . YB-1 is also shown to be linked to cancer in which it controls the *MDR1* gene which is involved in multi-drug resistance development in cancer cells (Matsumoto and Bay, 2005).

YB-1 was found to interact with other transcription factors, thus the direct interaction of YB-1 with the p65 (*relA*) protein activates the transcription of the human polyomavirus JC promoter (Raj *et al.*, 1996). The synergistic interaction between YB-1 and Ap2 transcription factor mediates high level of expression of the matrix-metalloproteinase-gelatinase A gene (Mertens *et al.*, 1998).

YB-1 was also found to form a functional heteromer with Purα and the HIV-1 transcription factor Tat (Ansari *et al.*, 1999). Interaction of YB-1 with YY-1 is required for regulation of the *grp78* gene controlling pathways involved in the stress protection (Brehm *et al.*, 1997).

1.5.2.1 Y-Box protein family

The Y-box protein is a family of DNA binding proteins that contain a highly conserved 70 amino acid DNA domain, the so-called "cold shock domain". The "cold shock domain" (CSD) region was identified as one of the most evolutionary conserved nucleic acid binding domains between bacteria, plants and animals. The name "Y box proteins" comes from the ability of the CSD to bind to the Y-box sequence [5'– CTGATTGG – 3'] of DNA, which is an inverted CCAAT box, in the promoter region of many genes. Among vertebrates, several Y Box protein genes have been cloned and characterized which includes include rat (Hasegawa *et al.*, 1991), chicken (Grant and Deeley, 1993). The human YB-1 was originally cloned from human B cell expression library for a protein that binds to the Y-Box in the promoter region of MHC class II gene. Then, two homologues of this gene *dbpA* and *dbpB* was further isolated from human placental cDNA expression libraries using the *c-erb2* gene promoter as the probe. Sequence analysis revealed that *dbpB* was identical to the *YB-1*.

Previously reported study has isolated two clones identical to *dbpA* and *dbpB* (*YB-1*) cDNA after screening a λ-phage expression library with a DNA probe containing the human leukosialin CD43 promoter sequence which were found to have expression pattern, which are not restricted to only hematopoietic cells (Kudo *et al.*, 1995). The *dbpA* was located on the chromosome 12p13 and *dbpB* on chromosome 1p34.

The dbpA was expressed in skeletal and heart muscles and the dbpB (YB-1) is found in skeletal muscle, kidney, lung and liver cells. The YB-1 protein is mainly located in the cytoplasm of the cell (Koike *et al.*, 1997). Interestingly it was detected in both the cytoplasm and nucleus in various types of human cancers (Ohga *et al.*, 1996).

1.5.2.2 *dbpA* : organization, structure and function

All Y-Box proteins contain three domains: the N terminal domain, cold shock domain (CSD) and C terminal domain (Figure 1.6). The *dbpA* gene consists of 10 exons spanning 24 kb of the genomic DNA. Its N terminal domain is encoded by the first exon, the cold shock domain (CSD) is encoded by the 2^{nd} to the 5^{th} exon and the C terminal domain by the 5^{th} to 8^{th} exons. However, the exon sequences of the N terminal domain of *dbpA* are found to have only 15 % homology with the *dbpB*'s (*YB-1*) N-terminal domain, whereas its C terminal has 30 % homology compared to CSD, which shows 96% sequence homology with its *YB-1* counterpart.

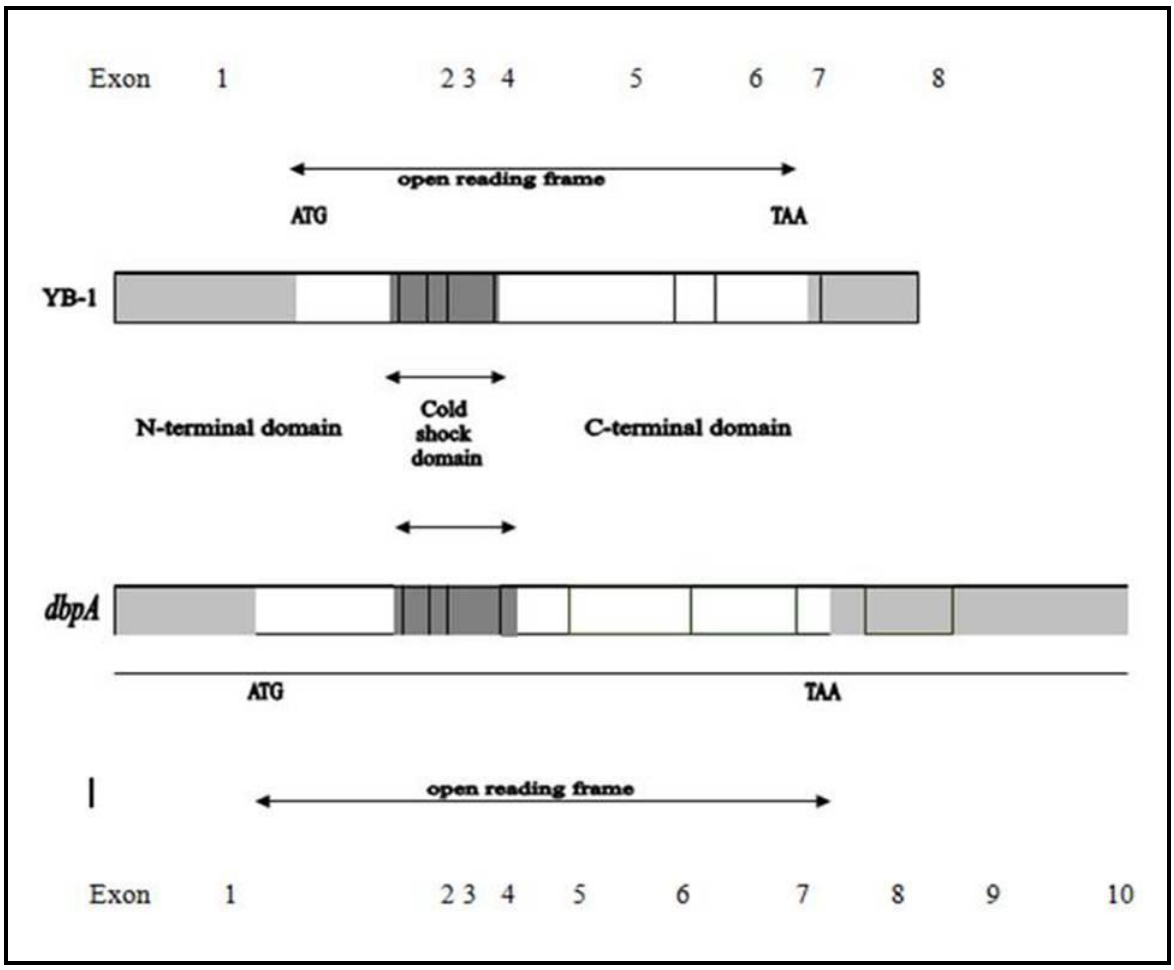

Figure 1. 6 Domain structure of human *dbpB* (YB-1) and *dbpA*

These genes consist of 3 main domains which are N, CSD and C-terminal domain. The light and dark shading represent non-coding sequences and the cold shock domains (CSD), respectively (Toh *et al.*, 1998).