Transcriptional Regulation of the Stem Cell Leukaemia Gene (SCL/TAL1) via Chromatin Looping

Yan Zhou

Institute of Cancer Sciences

College of Medical, Veterinary and Life Sciences

Thesis submitted for the degree of

Doctor of Philosophy

University of Glasgow

2012

Abstract

Transcriptional Regulation of the Stem Cell Leukaemia (SCL/TAL1) Gene via Chromatin Looping

The bHLH protein TAL1 (SCL) is a critical regulator of vertebrate hematopoiesis and is misregulated in T-cell acute lymphoblastic leukemia (T-ALL). This thesis studied chromatin looping interactions at the TAL1 locus – defining the first structural model which accounts for a number of phenomena associated with TAL1, its flanking genes and its relationship with its functional paralogue LYL1. The chromosome conformation capture (3C) and its high-throughput variant 4C-array technologies have been applied to characterise the chromatin interactions. Intriguing chromatin organisations have been identified at the TAL1 and LYL1 loci, which are closely associated with transcriptional regulation, chromosomal abnormality and regulatory remodelling through evolution. Firstly, in TAL1 expressing cells, the locus adopts a “cruciform” configuration – forming an active chromatin hub which brings together the TAL1 promoters, its stem cell and erythroid enhancers, and two CTCF/Rad21-bound insulators. Secondly, loss of a GATA1-containing complex bound by the TAL1 erythroid enhancer and its promoter is sufficient to disrupt the formation of the hub and the entire cruciform structure and results in decreased TAL1 expression. Thirdly, it demonstrates that genes flanking TAL1 are also dependent on this hub and that TAL1 promoters interact directly with intron 1 of the neighbouring STIL gene. This TAL1/STIL interaction also provides a structural link between the DNA sequences which mediate micro-deletions in 25% of cases of T-ALL. Finally, it demonstrates that a GATA1-dependent chromatin looping mechanism also exists at the LYL1 locus which is strikingly similar to that mediating contact between the TAL1 promoter and its erythroid enhancer. Conservation of core chromatin looping at the TAL1 and LYL1 loci may account for some aspects of their functional relationships. It also suggests that looping mechanisms at both loci could also facilitate *cis*-regulatory maintenance and/or remodelling during vertebrate evolution.

Acknowledgements

First and foremost I would like to thank my principal supervisor Dave Vetrie for all the guidance, advice and support throughout my PhD studies. His commitment to my project kept me going through all the difficult stages! I also would like to thank my co-supervisor Adam West and my adviser Jo Mountford for all their helps during the course.

I am also grateful to the funding bodies' support (the University of Glasgow PhD scholarship and the Overseas Research Students Awards Scheme) for providing me with such a great opportunity to study at the University of Glasgow.

I would also like to thank a number of individuals at the University of Glasgow who made valuable contributions to my PhD studies. In particular, I would like to thank Sreenivasulu Kurukuti who taught me how to perform 4C. In addition, I would like to thank Irene Todd who helped me with my English skills. Thanks to the former team members Kelly Chiang, Serdar Kasakyan, Peter Saffrey and Nicolas Bonhoure helped in many ways and were a pleasure to work with. I would also like to thank the former colleagues and my friends Koorosh Korfi, Susan Ridha, Ruslan Strogantsev, Gokula Mohan, Grainne Barkess, Dan Li, Meiji Ma, Carolyn Low, Elaine Gourlay for their help and support during my PhD.

Thanks to Milica Vukovic and Alison Michie at the Paul O'Gorman Leukaemia Research Centre, University of Glasgow, for providing primary murine lymphocytes.

Finally, I would like to thank my parents for their constant support and encouragement during my PhD studies.

Table of Contents

Chapter 1 Introduction

1.1 Eukaryotic transcriptional regulation

A numbers of steps including transcription initiation and elongation, mRNA processing, translation and protein stability can be modulated to regulate the expression of eukaryotic protein-coding genes. However, transcription initiation is considered as the crucial stage where most regulation occurs. The transcription machinery of eukaryotes consists of two complimentary regulatory components: the *cis*- and *trans*-acting regulatory elements. The *cis*-acting elements are DNA sequences located at the coding and non-coding regions of the genome, while the *trans*-acting elements are transcription factors or other DNA-binding proteins that recognize and bind to the consensus motifs in the *cis*-acting elements to modulate the transcription.

1.1.1 Cis-acting elements in involved in transcriptional regulation

In eukaryotic cells, RNA polymerase II (RNAP II) is responsible for the transcription of protein-coding genes. For genes transcribed by RNAP II, it typically contains two types of *cis*-acting elements (Figure 1.1) which are (i) a promoter which is composed of a core promoter and proximal elements, and (ii) a group of distal regulatory elements which may include enhancers, silencers, insulator, or locus control regions (LCR).

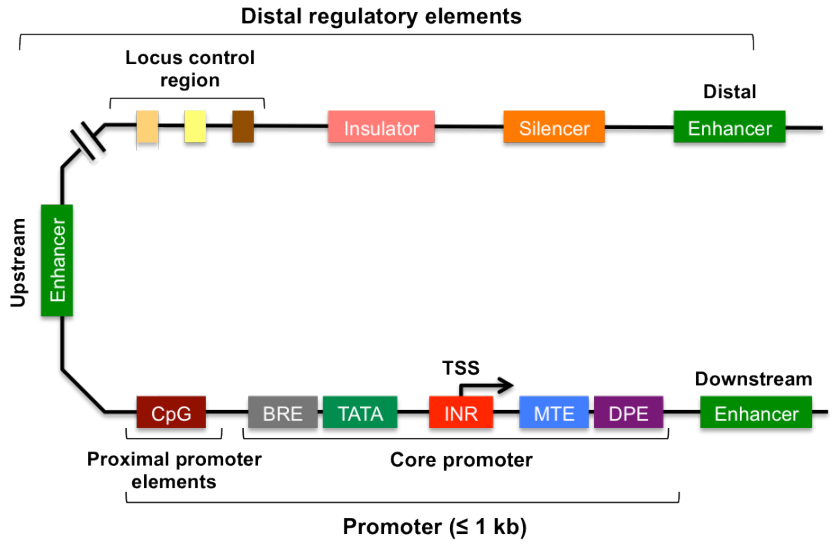


Figure 1.1: A schematic diagram of the two types of cis-regulatory elements involved in RNAP II transcription regulation. A typical promoter comprises a core promoter and proximal promoter elements such as CpG islands spanning about 1 kb around the transcription start site (TSS). The core promoter contains a TATA box (TATA), an initiator

element (INR), a downstream promoter element (DPE), a motif ten element (MTE) and a TFIIB recognition element (BRE). Distal regulatory elements including enhancer, silencer, locus control region (LCR) and insulator can be located upstream or downstream or even distant from the TSS.

1.1.1.1 Promoters

Core promoter

The core promoter locates at the start of a gene, which serves as the docking site for the basic transcriptional machinery as well as pre-initiation complex (PIC) assembly. It also defines the location of the transcription start site (TSS) and the direction of transcription (Sexton et al., 2012). The TATA box is the first identified element of the core promoter, which is bound by the TBP subunit of TFIID (Figure 1.2). Additionally, metazoan core promoters can be composed of a number of other elements (Figure 1.2), such as TFIIB-Recognition Element (BRE), Initiator (Inr), Motif Ten Element (MTE), Downstream Promoter Element (DPE), and Downstream Core Element (DCE) (Raab et al., 2012; Xu et al., 2011). Apart from the BRE is specifically recognized by TFIIB, the rest of core promoter elements are known to be TFIID-interaction sites, including TAF1/2 binding at INR, TAF6/9 binding at DPE and TAF1 binding at DCE (Figure 1.2). Further discussions about the *trans*-acting proteins mentioned above are shown in the following section at 1.1.2.

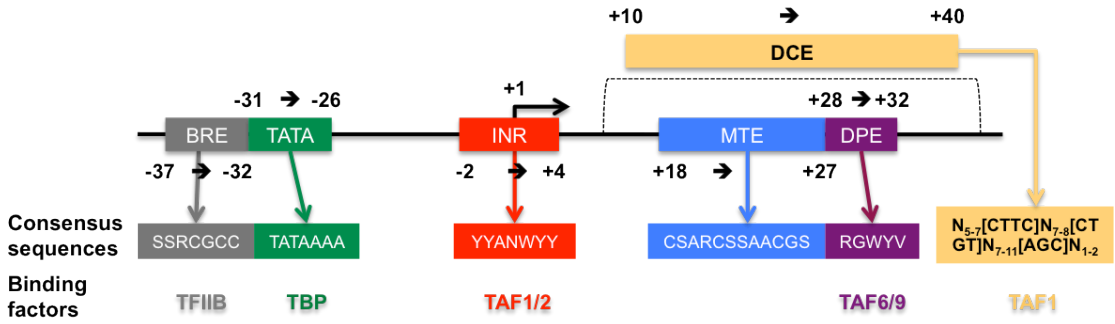


Figure 1.2: A schematic diagram of core promoter elements. A numbers of elements that compose the core promoters are shown in the diagram, including a TFIIB-Recognition Element (BRE, grey box), a TATA box (green box), an Initiator element (INR), a Motif Ten Element (MTE, blue box), a Downstream Promoter Element (DPE, violet box) and a Downstream Core Element (DCE, yellow box). The consensus sequence (human) of these elements, their relative locations to the TSS as well as the binding of transcription factors are shown. For the DC, it can be present in promoters containing a TATA box and/or INR, but presumably does not occur with a MTE or DPE.

Proximal promoter elements

The proximal promoter refers to a region with a span of several hundred base pairs upstream from the core promoter. It contains numbers of activator binding

sites and can participate in altering the rate of transcription. The CpG island is a proximal promoter elements (Figure 1.1) typically 0.5-2 kb in length with a high G+C nucleotide content (Valtieri et al., 1998), which associates with ~60% of human promoters (Vitelli et al., 2000). The presence of a CpG island is a reliable indication of the presence of a gene (Cross et al., 1994). In addition, DNA methylation at the CpG islands is associated with transcription silencing due to the methylation blocks the binding of activators to their recognition sites. Moreover, the methylation-specific binding proteins can bind to methylated CpG islands and recruit histone-modifying complex, which result in epigenetic imprinting and transcription silencing (Condorelli et al., 1997). However, the CpG islands locating at the proximal promoter stay un-methylated in the active genes.

1.1.1.2 Enhancers

Enhancers have been first characterised 30 years ago as DNA sequences of the SV40 tumour virus genome, which are capable of increasing the transcription of a heterologous human gene containing a promoter (Simonis et al., 2007). Subsequently, the first human enhancer is identified in the immunoglobin heavy-chain locus (Guillot et al., 2004; Simonis et al., 2007). Over the past three decades, a number of enhancers have been identified, which typically function in a spatial- or temporal-specific manner to regulate transcription. In addition to that, enhancers are capable of enhancing transcription independent of both the distance from and orientation relative to the promoter. Enhancers are typically composed of a tightly grouped cluster of transcription factor binding sites (TFBSs), which work cooperatively to enhance transcription. Importantly, the spatial organisation as well as orientation of TFBSs within an enhancer is crucial to its regulatory function, suggesting independences of distance and orientation can only be applied to the TFBSs cluster as a whole (Branco et al., 2008; Lecuyer et al., 2002).

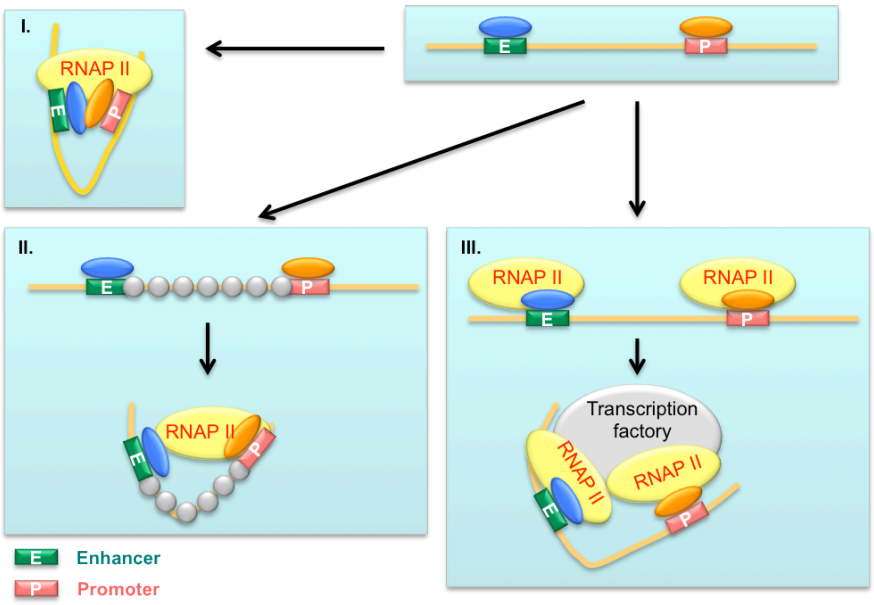


Figure 1.3: Schematic diagrams of how enhancers may interact with the promoters via looping. Chromatin fragment is shown as yellow line. Factors are bound to both the enhancer (E, green box) and promoter (P, pink box), which can potentially interact with each other and subsequently activate transcription. Panel I shows the simplest mechanism of promoter-enhancer interactions via random diffusion. Panel II shows that the promoter-enhancer interactions may be facilitated by additional factors (grey circles) that bind to the intervening sequences and bring the two elements into spatial proximity. Panel III shows that RNAP II may interact with both promoter and enhancer which bring the two elements into a common RNAP II-dependent transcription factory.

As illustrated in Figure 1.4 a, enhancers are generally long-range transcription regulatory elements, which can be situated several hundred kilobase (kb) up- or down-stream from the core promoter (Tripic et al., 2009). Since the discovery of enhancers, the dominant model for the mechanism of their action on target promoters has invoked direct contacts (Figure 1.3). Accumulated evidence favour the model termed as 'DNA-looping', which illustrates how enhancer elements function over long physical distances. In the model, the enhancers and core promoters are brought into spatial close proximity by forming DNA loops between these genomic regions (Nishimura et al., 2000; Vyas et al., 1999). A number of studies in nuclear organisation via chromosome conformation capture (3C) and its related techniques (detailed discussion in section 1.3) have provided abundant evidence for the 'DNA-looping' model. Thanks to the technical advance of 3C, it has been revealed that enhancers and promoters interact over distance within multiple loci in mammalian genomes (Anguita et al., 2004; Nishimura et al., 2000). It is now common to identify distal enhancers co-localised with their target promoters, as a result of promoter-enhancer interactions that are critical for transcription activation.

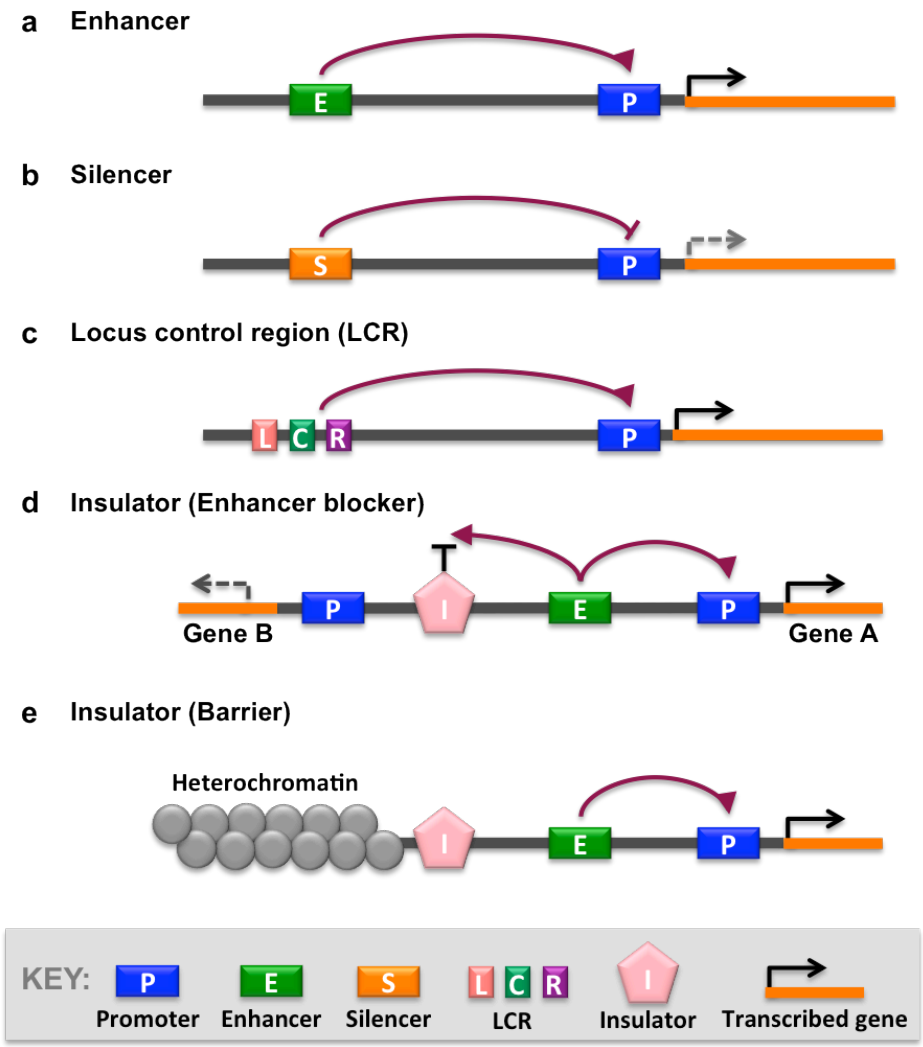


Figure 1.4: A schematic diagram of distal transcription regulatory elements. Panel a & b: enhancers and silencers function to activate and repress gene transcription, respectively. Panel c: locus control region (LCR) are composed of multiple regulatory elements that function collectively to regulate the gene transcription. Panel d & e: insulators function to either block the promoter from being contacted by the enhancer (EB) or prevent the spread of heterochromatin from the repressive region to the transcriptionally active gene (barrier).

1.1.1.3 Silencers

In contrast to enhancers, silencers are sequence-specific elements that act negatively on gene transcription (Figure 1.4 b). Alike enhancers, silencers also function independently of distance and orientation. They can either be situated as the part of a proximal promoter or a distal enhancer, or even located in distance from their target promoters as independent regulatory elements. Repressors – negative transcription factors are bound to silencers, which cooperate with other negative cofactors (co-repressors) to modulate the repression of target genes (De Gobbi et al., 2007). Numbers of models have been proposed to illustrate the mechanisms of silencers and repressors mediated transcription repression. Repressors may function either via blocking the binding of *trans*-acting elements

from their target promoters/enhancers (Aplan et al., 1992b; Xu et al., 2003) or by directly competing for the same binding sites of activators (Goldfarb et al., 1992). In addition, repressors may recruit chromatin-remodeling enzymes or chromatin modifiers to create a repressive chromatin structure, which prevents transcription factors (activators) from accessing a promoter (Grutz et al., 1998).

1.1.1.4 Locus control regions (LCRs)

Locus control regions (LCRs) are defined by their abilities to enhance transcription of an entire locus or gene cluster in a position-independent and copy-number-dependent manner (Elwood et al., 1998). LCRs are clusters of regulatory elements including enhancers, silencers, insulators, nuclear-matrix regions (MARs) or chromosome scaffold-attachment regions (SARs) (shown in Figure 1.4 c). Each of the elements is bound by different *trans*-acting proteins and regulates transcription differentially. The collective activity of all these elements defines an LCR and the most prominent feature of LCRs is the strong and specific enhancer activity. In addition, LCRs typically contain a cluster of DNase I hypersensitive sites as well as provide an open-chromatin domain for their linked genes. In mammalian genome, the first LCR has been identified in the β-globin locus (Tanigawa et al., 1993). Moreover, it has been shown that the chromatin loops form between the locus control region (LCR) and its target promoter located 40-60 kb in genomic distance (Carter et al., 2002; Tolhuis et al., 2002). In addition, all other *cis*-acting elements in the β-globin locus are also in spatial proximity, where they form an “active chromatin hub” (de Laat and Grosveld, 2003; Drissen et al., 2004). This paradigm arises from the three-dimensional co-localisation of DNase I hypersensitive sites and depends on specific DNA-protein interactions which link all the essential components for transcription initiation and activation (Tolhuis et al., 2002).

1.1.1.5 Insulators

The function of insulators is to prevent genes from being affected by the transcriptional state of their neighboring genes. Typically, insulators are approximately 0.5-3 kb in size and act in a position-dependent and orientation-independent manner. Insulators can regulate transcription in two different ways: (i) it can locate in between a promoter and a cognate enhancer to block enhancer-

promoter communication known as the enhancer blocker (Figure 1.4d); (ii) it can locate in between an active gene and the repressive chromatin region to prevent the spread of heterochromatin known as the barrier (Figure 1.4e) (Gaszner and Felsenfeld, 2006). In vertebrata genomes, a numbers of insulator elements have been characterised, such as the 5'HS4 in chicken β-globin locus, homologous elements in human and murine β-globin loci and the imprinting control region (ICR) in the Igf2/H19 locus (Phillips and Corces, 2009). In addition, the proper function of insulators requires the binding of CTCF (CCCTC-binding factor) – a *trans*-acting factor (Ono et al., 1998; Tremblay et al., 2003). Recent studies also have shown that the insulator function of CTCF is regulated by cohesin (Parelho et al., 2008; Wendt et al., 2008). The overlapped binding patterns of CTCF and cohesin appear in various cell types including ETNK2, CFTR and c-Myc genes (Gaszner and Felsenfeld, 2006; Komura et al., 2007; Valenzuela and Kamakaka, 2006; Zhou et al., 2008). Further discussions about CTCF/cohesin and insulator can be found in the Chapter 4.

1.1.2 Trans-acting elements involved in transcription regulation

Numerous trans-acting elements are required for the assembly of the entire transcriptional machinery onto a various *cis*-regulatory elements as previously discussed, in order to allow gene transcription taking place. These *trans*-acting elements are further detailed as follows.

1.1.2.1 RNA polymerase

Transcription of genes from DNA to RNA involves a three-step process including initiation, elongation and termination. Transcription initiation is associated with RNA polymerase and general transcription factors (GTFs), which are required for the formation of a pre-imitation complex (PIC) at the promoter of genes. RNA polymerases are categorised into three classes which are RNA polymerase I, II and III (RNAP I, II and III) according to the transcripts they produced. RNAP I transcribes DNA to ribosomal RNAs (rRNAs) including the 28S, 18S and 6S subunits. RNAP II transcribes DNA to messenger RNAs (mRNAs) and small nuclear RNAs (snRNAs). RNAP III transcribes DNA to transfer RNA (tRNA) and 5S rRNA. The RNAP II involves the most sophisticated transcriptional machinery among the three RNA polymerases, which is further discussed as follows.

The human RNAP II consists of 12 subunits, including Rpb1 to Rpb12 (Young, 1991). For Rbp1 to Rbp3 and Rbp11, they are sharing homologous counterparts with bacterial polymerase. Rbp5, 6, 8, 10 and 12 are common subunits which present in all three RNA polymerase classes, whereas Rpb4, 7 and 9 are unique components of the RNAP II. Each of these RNA polymerase subunits plays specific and fundamental roles in selection of transcription start site (TSS), alteration of elongation rate, interaction with activators and stability of RNA polymerase (Lee and Young, 2000). For example, Rpb1 contains a carboxyl-terminal repeat domain (CTD) consisting by repeats of the consensus sequence of Tyr-Ser-Pro-Thr-Ser-Pro-Ser (YSPTSPS) (Prelich, 2002). The CTDs can be phosphorylated at Ser2 and Ser5 and hyper-phosphorylated RNAP II is associated with transcription elongation (Sims et al., 2004).

1.1.2.2 General transcription factors (GTFs)

A number of factors including Transcription factor II A, B, D, E, F, H and RNAP II are all classified into the GTFs. GTFs assemble on the core promoter to form a transcription pre-initiation complex (PIC), which drive RNAP II to the transcription start site (TSS) and are required for transcription of almost all genes (Thomas and Chiang, 2006). Different GTFs interact with the promoter at different regions and have various functions as illustrated in Table 1.1.

Table 1.1. Functions of General transcription factors.

GTF	Functions
TFIID	• Containing two subunits: TATA-binding protein (TBP) binds to the TATA box of the core promoter in an orientation-independent manner TBP-associated factor (TAF) binds to INR and DPE and is required for promoter selection and transcriptional activation • An additional TBP-related factor (TRF) in eukaryotes recognizes DNA sequences in TATA-less promoters for transcription initiation
TFIIA	• A heterodimer interacts with TBP and stabilizes the TBP-DNA interaction • Promoting TFIID binding to DNA • Involvement in transcription activation via binding to activators
TFIIB	• Interacting with TFIID and RNAP II and being required for transcription start site selection • Binding to BRE and downstream sequences of the TATA box • A direct interacting partner of activators associates with TFIIB recruitment to the promoter
TFIIE	• Binding to DNA sequences directly upstream of the transcription start site after the formation of the PIC • Promoting the recruitment of TFIIH stimulating the CTD kinase and helicase activities of TFIIH
TFIIF	• A heterodimer contains two subunits: TFIIFβ binds to either upstream or downstream of TATA box TFIIFα binds to regions downstream of the TATA box • Binding tightly to RNAP II and being involved in avoiding non-specific DNA binding and stabilization of the PIC
TFIIH	• Containing two subunits: A core subunit with helicase activities (XPD and XPB) is required in the unwinding of DNA A kinase subunit (Cdk7) phosphorylates the CTD in RNA pol II during the transition to elongation stage

1.1.2.3 Sequence-specific transcription factors

In addition to RNA polymerase and GTFs which only account for the basal activity of transcription, sequence-specific transcription factors are required in order to fully regulate the gene transcription. These transcription factors directly bind to *cis*-regulatory elements including promoters, enhancers and silencers to execute their functions in transcriptional activation or repression, via recognising the conserved sequences within *cis*-regulatory elements known as the transcription factor binding sites (TFBSs). Some of the sequence-specific transcription factors require the recruitment of co-activators or co-repressors via protein-protein interaction to assist the performance of their functions. In addition, sequence-specific transcription factors often function in a combinatorial way. In many cases, clusters of various TFBSs are presented in the *cis*-regulatory elements forming unique motifs for binding of the transcription factor complex, in order to exert tight regulatory control on genes of interest in a temporal and/or spatial manner.

1.1.2.4 Co-activators and co-repressors

Co-activators are adaptor proteins that typically lack intrinsic sequence-specific DNA binding but provide a link between activators and the general transcriptional machinery. For instance, TBP-associated factors (TAFs) such as TAFII40 and TAFII60 can act as co-activators by linking activators such as p53 and general transcription factors to allow transmitting information between the two (Thut et al., 1995). Mediators are another type of co-activators that activate transcription by stimulating the phosphorylation of CTD of RNAP II. Moreover, they also associate with activators and transmit positive or negative signals to the promoters (Myers and Kornberg, 2000). Some co-activators or co-repressors function as docking elements on activators or repressors instead of having intrinsic enzymatic activities, which modulate the transcriptional activation or repression via mediating the recruitment of other factors to the initiation complex (Wolstein et al., 2000).

1.1.3 Epigenetic modifications and transcriptional regulation

1.1.3.1 DNA methylation

In eukaryotes, DNA methylation is one of the most widely studied epigenetic modifications, which occurs exclusively on cytosine bases at CpG dinucleotides.

The CpG dinucleotides often tend to cluster into CpG islands, which are defined as regions of more than 200 bp with over 50% of G+C content. Approximately 60% of human gene promoters are associated with CpG islands, which are usually un-methylated in normal cells; but with an exception of ~6% of them are methylated in a tissue-specific manner during early development or tissues under differentiation (Straussman et al., 2009). Generally, the methylation of CpG islands is associated with repression of gene expression. DNA methylation plays a pivotal role in gene imprinting, as hyper-methylation at one of the parental alleles results in mono-allelic expression (Kacem and Feil, 2009). Moreover, DNA methylation regulates gene expression through a number of different mechanisms. First, methylated DNA can promote the recruitment of MBD (methyl-CpG-binding domain) proteins, which can subsequently recruit complexes for histone modifications and chromatin remodelling to the methylated regions to mediate gene silencing (Esteller, 2007; Lopez-Serra and Esteller, 2008). Second, DNA methylation can repress gene expression by precluding the recruitment of DNA-binding transcription factors from their target sites (Kuroda et al., 2009). In contrast, un-methylated CpG islands providing a chromatin environment favour gene expression via the recruitment of Cfp1, which associates with histone methyl-transferase Setd1 by creating domains enriched with an active hallmark of H3K4 trimethylation (Thomson et al., 2010). DNA methylation can also occur at gene bodies, which is common in universally expressed genes and is positively associated with gene expression (Hellman and Chess, 2007). Gene body methylation has been proposed to related to RNAP II elongation efficiency and prevention of spurious transcription initiations (Lorincz et al., 2004; Zilberman et al., 2007).

DNA methylation is mediated via enzymes of the DNMT family that catalyse the transfer of a methyl group from S-adenosyl methionine to DNA. There are five members of the DNMT family proteins including DNMT1, DNMT2, DNMT3a, DNMT3b and DNMT3L which have been found in mammals. However, only DNMT1, DNMT3a and DNMT3b have methyltransferase activities, which are classified into two categories, the maintenance DNMT (DNMT1) and *de novo* DNMTs (DNMT3a and DNMT3b). DNMT1 is the most abundant DNMT proteins in the cell, which is mainly transcribed during the S phage of the cell cycle. It has high preference for hemi-methylated DNA, in order to methylate these hemi-methylated regions right after being generated from semi-conservative DNA

replication. DNMT3a and DNMT3b are proposed to be responsible for de novo methylation during embryonic development, which are highly expressed in ES cells but down-regulated in those differentiated cells (Esteller, 2007). Although the DNMT family member DNMT3L does not function as a methyltrasnferase, it has been found to be required for the establishment of genomic imprinting, which acts as a stimulator for DNMT3a and DNMT3b via interacting with them in nucleus (Bourc'his et al., 2001; Chen et al., 2005).

A number of mechanisms have been proposed for the question of how the DNA methylation machinery is targeted to specific sequences in the genome, suggesting that it's facilitated by other epigenetic factors via interactions with DNMTs (Esteve et al., 2009; Jeong et al., 2009; Ooi et al., 2007; Wang et al., 2009). In addition, siRNA-mediated/directed DNA methylation has been reported in plants, proposing the possible mechanism of siRNA-dependent DNMTs recruitment for *de novo* DNA methylation of specific regions (Matzke and Birchler, 2005; Mosher and Melnyk, 2010; Vrbsky et al., 2010). Nevertheless, it remains unclear whether similar processes in regulating DNA methylation also exist in animals.

1.1.3.2 Histone modifications

The nucleosome is the fundamental repeating unit of eukaryotic chromatin which typically consists of ~200 bp of DNA wrapped around a histone octamer. Each octamer contains two copies of the core histone proteins including H2A, H2B, H3 and H4, which group into two H2A-H2B dimers and one H3-H4 tetramer to form the nucleosome. The nucleosome is wrapped with 146 bp DNA fragment in 1.65 superhelical turns and the adjacent nucleosomes are separated by ~50 bp of linker DNA (Luger et al., 1997). Histone H1 is known as the linker histone, which binds to the linker DNA instead of forming part of the nucleosome. The N-terminal tails of histones are subject to a number of post-translational modifications, which include acetylation, methylation, phosphorylation, ubiquitination, SUMOylation and ADP-ribosylation (Kouzarides, 2007; Rando and Chang, 2009). These various histone modifications play vital roles in DNA repair and replication, transcriptional regulation, alternative splicing and chromatin condensation (Huertas et al., 2009; Kouzarides, 2007; Luco et al., 2010).

Table 1.2 Histone modifications and their possible functions (Sawan and Herceg, 2010).

Modification	Histone residue	Enzyme	Possible role
Acetylation	H2A K5	Tip60, Hat1, P300/CBP	Transcriptional activation
	H2B K5	ATF2	Transcriptional activation, DNA repair
	H2B K12	ATF2, P300/CBP	Transcriptional activation
	H2B K20	P300	Transcriptional activation
	H3 K9	Gcn5, SRC-1	Transcriptional activation, DNA repair
	H3 K14	Gcn5, PCAF, Tip60, SRC-1, TAF1, P300	Transcriptional activation, DNA repair and replication
	H3 K18	P300/CBP	Transcriptional activation, DNA repair
	H3 K23	P300/CBP	Transcriptional activation, DNA repair
	H3 K27	Gcn5	Transcriptional activation, DNA repair
	H4 K5	Hat1, Tip60, ATF2, P300	Transcriptional activation, DNA repair and replication
	H4 K8	Gcn5, PCAF, Tip60, ATF2, P300	Transcriptional activation, DNA repair and replication
	H4 K12	Hat1, Tip60	Transcriptional activation, DNA repair and replication
	H4 K16	MOF, Gcn5, Tip60, ATF2	Transcriptional activation, DNA repair
Methylation	H1 K26	EZH2	Transcriptional repression
	H3 R2	CARM1	Transcriptional activation
	H3 K4	MLL4, SET1, MLL1, SET7	Transcriptional activation
	H3 R8	PMRT5	Transcriptional repression
	H3 K9	SUV39h1/2,ESET,G9A, EZH2, Eu-HMTase1	Transcriptional activation and repression
	H3 R17	CARM1	Transcriptional activation
	H3 R26	CARM1	Transcriptional activation
	H3 K27	EZH2, G9A	Transcriptional activation and repression
	H3 K36	HYPB, NSD1	Transcriptional activation
	H3 K79	DOT1L	Transcriptional activation and repression, DNA repair
	H4 R3	PRMT1, PRMT5	Transcriptional activation
	H4 K20	PR-SET7, SUV4-20	Transcriptional activation and repression, DNA repair
Phosphorylation	H2AX S139	ATM, ATR, DNA-PK	DNA repair
	H2A T119	NHK-1	Transcriptional repression, DNA repair
	H2B S14	Mst1	DNA repair
	H3 S10	TG2, Aurora B, MSK1, MSK2	Transcriptional activation
Ubiquitination	H2A K119	Ring 1b	Transcriptional activation, DNA repair
	H2A K120	RNF 20	Transcriptional activation, DNA repair

All of these histone modifications play vital roles in transcriptional regulation, DNA repair and replications (Table 1.2). The functions of histone modifications related to transcriptional regulation are briefly summarised as follows. First, the core histones can be reversibly acetylated at lysine (K) residues such as H2A (K5, K9), H2B (K12, K15), H3 (K9, K14, K18, K56) and H4 (K5, K8, K12, K16) via histone acetyltrasferases (HATs) and histone deacetylases (HDACs). Histone acetylation is closely associated with transcriptional activation (Shahbazian and Grunstein, 2007). It has been shown that acetylation of H4K16 acts negatively on the formation of the 30 nm chromatin fibre (Shogren-Knaak et al., 2006). In addition, acetylation at histone H3K9 in promoter regions correlates with low nucleosome density in the vicinity of the transcription start sites (TSSs) (Nishida et al., 2006). Therefore, it is proposed that histone acetylation results in a 'loosening' chromatin environment via neutralising the basic charge of lysine and subsequently provides a greater access of chromatin to transcription factors. Second, methylation of histones can be found at lysine (K) or arginine (R) residues, including H3 (K4, K9, K27, K36 and K79), H4K20, H3 (R2, R8, R17 and R26) and H4R3, where lysine residues can be mono-, di- or tri-methylated but arginine residues can only be mono- or di-methylated. Comparing to histone acetylation, functions of histone methylation are much more complicated. For instance, methylation at histone H3K4 is associated with active gene expression in numerous eukaryotes (Bernstein et al., 2005; Ng et al., 2003; Pokholok et al., 2005; Santos-Rosa et al., 2002; Schneider et al., 2004; Schubeler et al., 2004), whereas H3K9 methylation is a hallmark for heterochromatin in higher eukaryotes (Richards and Elgin, 2002). Furthermore, histone methylations at arginine residuals can also contribute to both active and repressive chromatin states, such as methylations at H3R2, R17 and R26 as well as H4R4 enhance gene activation while di-methylation at H4R3 maintains silent chromatin domains (Schurter et al., 2001; Wang et al., 2001; Yu et al., 2006). Third, histone phosphorylation is found specifically targeting to serine (S) and threonine (T) residues, including H2AS1, H2BS14, H3 (T3, S10 and S28) and H4S1. Most of these studies focus on the functions of phosphorylation at histone H3S10, which demonstrate its roles in two opposing processes of gene activation during interphase and chromosome condensation during mitosis (Johansen and Johansen, 2006). Forth, histone ubiquitinations have been reported at H2A and H2B, where the carboxyl end of ubiquitin is added to H2A K199 and H2B K120 in human. The H2AK199

ubiquitination has been detected on female inactive X-chromosome, which is associated with polycomb silencing via recruiting PcG proteins PRC1-like (PRC1-L) (de Napoles et al., 2004; Wang et al., 2004). In addition, ubiquitination of a nucleosome would have a negative impact on chromatin folding and subsequently affect the regulation of gene expression, due to the ubiquitin moiety is half the size of a core histone (Shilatifard, 2006). Fifth, SUMO is an ubiquitin-related protein that induces protein sumoylation via ligating to its target protein. Sumoylation of histone H4 has been reported in mammalian cells, which correlates with transcriptional repression caused by histone deacetylation and HP1 recruitment (Shiio and Eisenman, 2003). Sixth, mono-ADP ribosylation of histone is associated with DNA repair and cell proliferation, which occurs when exposed to DNA damaging agents (Hassa et al., 2006). In addition, histone ribosylation has revealed the potential of communicating with other modification as it preferentially co-presents with acetylation at histone H4 (Golderer and Grobner, 1991).

In relation to the role of histone modification in transcriptional regulation, the human genome can be generally divided into two groups, which are transcribed euchromatin and transcriptionally inactive heterochromatin. High levels of acetylation as well as trimethylation at histone H3K4, H3K36 and H3K79 are hallmarks of the euchromatin. In contrast, heterochromatin is characterised by low levels of acetylation and high levels of methylation at histone H3K9, H3K27 and H4K20 (Li et al., 2007a). In addition to that, it has been demonstrated that levels of histone modification can be used to predict gene expression. For instance, actively transcribed genes are normally marked with high levels of H3K4me3, H3K27Ac, H2BK5Ac and H4K20me1 over the promoters and H3K79me1 and H4K20me1 along the gene body (Karlic et al., 2010). Histones can be modified at different sites simultaneously. These modifications can occur at the same site or in the same/different histone tails. Therefore, the outcome of histone modifications is dependent on the collective combination of all hallmarks, instead of a single histone mark (Duan et al., 2008; Nakanishi et al., 2009; Wang et al., 2008). The “bivalent domain” found in ES cells is a classic example of co-existing histone modifications, where the active H3K4me3 mark and the repressive H3K27me3 mark are found together at promoters of genes which are developmentally essential (Bernstein et al., 2006). The bivalent domains allow ES cells to be able to tightly regulate gene expression and rapidly response to changes of expression

pattern during different developmental processes, but no longer exist after cell commitment (Mikkelsen et al., 2007).

As a matter of fact, all the epigenetic regulators are closely associated with each other to modulate gene expression. For instance, the connection between the histone modification and DNA methylation is revealed by the relationship between DNMT3L and H3K4. The *de novo* DNA methylation is induced by the interaction between DNMT3L and histone H3 tails by recruiting DNMT3A, whereas the interaction can be repressed by methylation at H3K4 (Ooi et al., 2007). Moreover, a number of histone methyltransferases have been found to be capable of driving DNA methylation to particular genomic regions via DNMT proteins recruitment, by which to allow inactive histone marks to establish transcriptional silencing (Tachibana et al., 2008; Zhao et al., 2009). In addition, the stability of DNMTs can also be modulated by histone methyltransferases and demethylases, which indirectly affect the level of DNA methylation (Esteve et al., 2009; Wang et al., 2009). In contrast, methylated DNA can direct methylation at H3K9 via recruiting MeCP2, suggesting that DNA methylation can also modulate the histone modifications (Fuks et al., 2003).

1.1.3.3 Nucleosome positioning and remodeling

Nucleosomes are considered as a barrier to transcription that block the activator and transcription factors from accessing their binding sites on DNA. Meanwhile, they also inhibit the transcription elongation by engaging polymerase. It appears that DNA being packaged into nucleosomes can affect all stages of transcription and sequentially modulate gene expression. For instance, the precise position of nucleosomes at proximal TSSs has a great influence on transcription initiation. The displacement of nucleosome as few as 30 bp at TSS has been implicated in altering RNA polymerase II activity. In addition, the loss of a nucleosome right upstream of the TSS is closely correlated with transcription activation, whereas occupancy of nucleosome over the TSS leads to transcription repression. Furthermore, the maintenance of nucleosome-free regions at the 5’ and 3’ ends of genes is also critical for the assembly and disassembly of the transcription machinery (Cairns, 2009; Schones et al., 2008).

Nucleosome positioning not only modulates accessibility of the activators to their target binding domains but also has been demonstrated to play a key role in determining DNA methylation profiles (Chodavarapu et al., 2010). Moreover, the nucleosome remodeling machinery is also linked with DNA methylation and specific histone modifications. For example, it has been found that a key member of the SWI/SNF-related chromatin-remodeling complex associates with the methyl-CpG binding protein MeCP2 *in vivo* and is functionally linked with repression via DNA methylation (Harikrishnan et al., 2005). In addition, it has been found that the nucleosome-remodeling factor PHD is associated with H3K4me3, suggesting a new function of PHD in the highly specialized methyl-lysine-binding activity (Wysocka et al., 2006).

1.1.4 Three-dimensional (3D) chromatin organisation in eukaryotic transcriptional regulation

As described in previous sections, cross talks between *cis*- and *trans*- transcription regulatory elements play a fundamental role in regulating gene expression. The conventional idea about transcriptional regulation focuses at the lower levels of genomic organisation, which involves (i) distributions of coding and *cis*-regulatory elements of all genes in the genome and (ii) interactions between DNA elements and proteins including histones in nucleosomes as well as transcription factors (*trans*-elements) in regulating gene activation or repression. Moreover, it has been long known that DNA and histones are marked with chemical modifications which regulate transcription of genes, catalogued as ‘epigenetic modifications’ (Figure 1.5a & b). Recently, it has become clear that the 3D chromatin organisation reflects a higher-order of transcriptional regulation. Thanks to the rapid development of high-throughput technologies for genome-wide analysis, accumulated evidence now have implicated the essential relationship between higher-order nuclear organisation and genome function. Instead of picturing that transcriptional activity is entirely determined by epigenetic modification, scientists are now looking for answers in a new direction about chromatin dynamic movement and communication, described as ‘chromatin network’ or ‘chromosome interactome’ (Baker, 2011) (Figure 1.5 c & d).

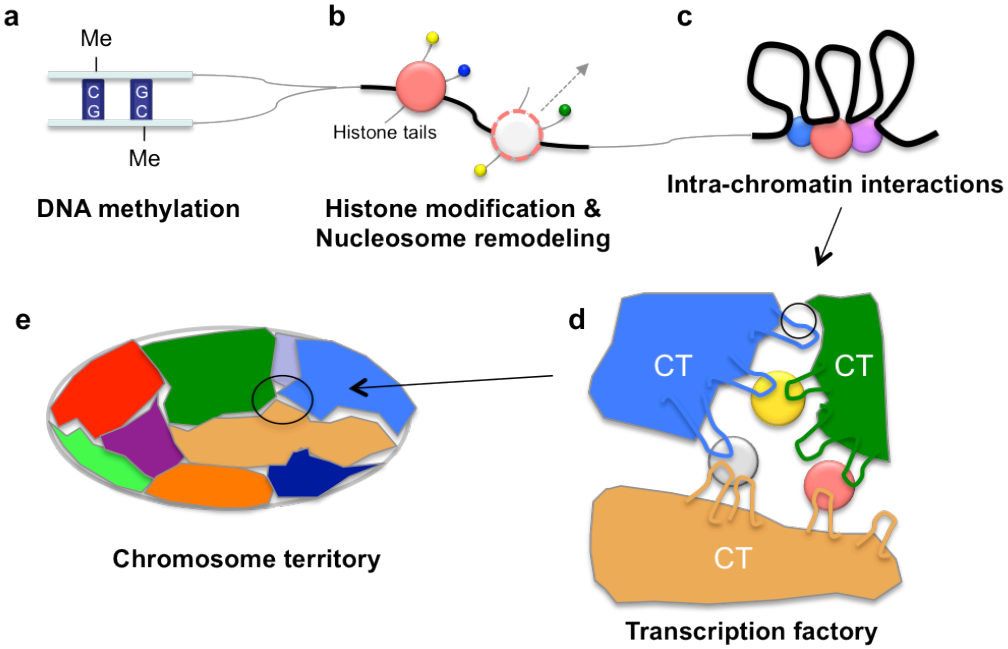


Figure 1.5: Schematic map illustrates comprehensive picture about links between epigenetic modications (DNA methylation and histone modifications) and chromatin organization in transcription regulation at different levels. Level I (panel C): chromatin interactions between cis-acting elements (e.g. promoter-enhancer). Level II (panel D): transcriptional-dependent clustering of chromatin looping interactions known as transcription factory. Level III (panel E):

The previous sections described how epigenetic modifications modulate transcriptional activation and repression. The following sections will summarise how transcription is regulated at the 3D level, from long-range chromatin interactions to transcription factory and nuclear organisation (Figure 1.5).

1.1.4.1 Chromatin interactions

As described in previous section, enhancers are *cis*-acting regulatory elements which are functionally determined by their ability for transcriptional acitivation regardless of their distances or orientations with respect to their cognate genes. Enhancers are generally separated from their cognate promoters with large genomic distance, which have been found to accurately communicate with each other via long-range chromatin interactions in order to regulate gene expression. The promoter-enhancer interaction is one of the most widely distributed types of chromatin interactions (Figure 1.5c), which have been characterised via chromosome conformation capture (3C)-based approaches (see section 1.2.4.2). Enhancer-promoter interactions and clustering of transcriptional active genes are consistent with transcription factory models, which are evident by RNAP II-associated chromatin interactome captured by ChIA-PET (Li et al., 2012).

1.1.4.2 Transcription Factories

Most protein-coding genes are transcribed by RNAP II and the distribution of discrete sites for transcription was initially identified by detection of nascent mRNA and RNAP II staining. However, it revealed a limited number of foci that were insufficient to account for the great number of active genes in the human genome (Iborra et al., 1996). Subsequent studies have shown that active genes dynamically co-localise to transcription factories for expression in a transcription-dependent way (Figure 1.5d). Transcription factories consist of numerous discrete sub-nuclear foci where nascent RNAs are produced, and contain high concentrations of RNA polymerase II and other transcriptional accessory and regulatory factors (Iborra et al., 1996; Jackson et al., 1993). Several genomic loci can share a single transcription factory non-randomly, suggesting that transcription factories may physically coordinate transcription and gene expression inside the cell nucleus. Genes sharing the same transcription factory can be either located in the same chromosome (*cis*) or on different chromosomes (*trans*) (Cherniack et al., 1991; Mitchell and Fraser, 2008; Osborne et al., 2004; Osborne et al., 2007; Papantonis et al., 2010). The discovery of long-range chromosomal interactions clustering in the nucleus of higher eukaryotes provides a functional link between nuclear architecture and gene expression regulation (Schoenfelder et al., 2010a). This idea challenges the conventional view that transcription is a one-dimensional (1D) process. Individual chromosomes occupy discrete chromosome territories in the 3D space of the cell nucleus (Bolzer et al., 2005). Growing evidence indicate that the spatial organisation of chromosome territories (CTs) plays an essential role in regulating genomic functions (Fraser and Bickmore, 2007; Lanctot et al., 2007).

1.1.4.3 Chromosome Territories

As a result of developments in fluorescence in situ hybridization (FISH) over the past two decades, specific chromosome territories (CTs) have been identified in the nuclei of mammalian cells (Cremer and Cremer, 2001). Using chromosome-specific DNA libraries, FISH experiments have demonstrated that each interphase chromosome occupies a distinct territory inside the cell nucleus (Cremer et al., 1993; Foster et al., 2005). In brief, this specific sub-nuclear domain occupied by the chromosome, is referred to as a chromosome territory (Figure 1.5e). The

structures of DNA sequences inside CTs are not random, as CTs have preferred locations relative to either the nuclear interior or periphery. Within chromosome territories, the chromosome arms are kept apart from each other as well as gene-rich regions of chromosome are separated from gene-poor regions. In addition, the interiors of CTs are permeated via highly interconnected chromatin interacting networks (Figure 1.5d & e), which allows the genomic sequences deep inside the CTs being accessible to trans-acting regulatory factors and regulating particular sets of genes (either in *cis* or in *trans*) in a coordinative manner (Meaburn and Misteli, 2007). Generally, heterochromatin and repressed regions mainly localise close to the nuclear periphery whereas the euchromatin and active regions are accommodated at the inner part of the nucleus. This pattern is conserved through the evolution from unicellular to multicellular organisms (Postberg et al., 2005).

1.2 Experimental and computational approaches to understand transcriptional regulation

1.2.1 Characterisation of regulatory elements

As described in the previous section, both *cis*- and *trans*-acting regulatory elements are the fundamental components for transcriptional regulation. Characterisation of potential *cis* regulatory elements of genes as well as identification of direct binding sites of transcription factors allow us studying to where and how the transcription factor modulate transcription of its target gene through DNA-TF binding events, which is important empirical evidence to further our understanding of transcriptional regulation networks.

1.2.1.1 DNase I hypersensitive assay

In order to allow the genomic regions of DNA being accessible to binding of proper transcription factors, nucleosomes may undergo conformational changes or be re-positioned. These nucleosome-free regions are sensitive to the digestion of nuclease enzymes, such as deoxyribonuclease (DNase), MNase and restriction enzyme. Based on that, a technique termed DNase I hypersensitive site (HSs) mapping has been developed to identify transcription factor binding sites (Hoang et al., 1996). DNase I HSs in native genomic domains have traditionally been localized by cleavage of nuclear chromatin followed by DNA purification; restriction endonuclease digestion, gel electrophoresis, southern blotting and hybridisation

with a radiolabelled DNA probe. It has been widely applied to detect genomic regions with potential regulatory activity, regardless its low-throughput and time-consuming features. More recently, however, new approaches have been developed for high-throughput identification of DNase I HSs using microarray and sequencing platforms (Robb et al., 1996; Shivdasani et al., 1995), which allow genome-wide identification of regulatory elements. It is evident that these genomic regions identified owing to their DNase I hypersensitivity are *cis*-acting elements, including promoters, enhancers, silencers, insulators and locus control regions (Tanigawa et al., 1995). Additionally, most of these DNase I HSs are cell-type specific and are mostly enhancers (Robb et al., 1995).

1.2.1.2 Reporter gene assay

The reporter gene assay is one of the most versatile approaches for detecting and analysing activity of transcription regulatory elements (Weber et al., 1984). In the assay, the genomic region of DNA to be assessed for its regulatory activity is cloned into a plasmid right upstream of a reporter gene, such as GFP (green fluorescent protein), β-galactosidase or luciferase gene. The resulting plasmid construct is subsequently transfected into cultured cell lines. The activity of the reporter gene is measured to determine whether the cloned genomic segment alters expression of reporter gene. For the genomic region being tested for core promoter activity, the segment requires to be placed immediately upstream of a reporter gene lacking an endogenous promoter. The reporter assay can also be used to detect activity for enhancers and silencers when using the appropriated strength promoter. More complex reporter systems are required for measuring activity of insulators or LCRs. For instance, insulator activity can be measured by two different ways, depending on whether enhancer-blocking or barrier function is assayed. The genomic region containing a putative enhancer-blocking insulator can be placed in between a known interacting enhancer-promoter configuration, to assess its ability of interfering the enhancer-promoter communication. By contrast, the detection of a putative barrier insulator requires a transgenic reporter assay, in order to verify its ability of shielding the transgene from the repressive effects of heterochromatin and allowing for position-independent reporter gene expression (Porcher et al., 1996). Similarly, a transgene reporter assay can also be used to determine a putative LCR based on its ability to overcome position effects and drive transgene expression (Elefanty et al., 1997). These approaches have been

extensively used to identify a number of regulatory elements in genes such as c-Myc (Begley and Green, 1999), and especially TAL1 (section 1.3).

1.2.1.3 Comparative genomic sequence analysis

Comparative sequence analysis has become widely exploited to refine searches for transcription factor binding sites. The phylogenetic footprinting is one of the comparative genomic approaches, in which genomic sequences from evolutionary-separated species are compared, and those highly conserved sequences across species are considered as candidates for being functionally important (Turpen et al., 1997). The method is built on the expectation that TFBSs are conserved through evolution and can be subsequently detected when the sequences from species with large evolutionary distances are aligned. Several programs have been developed for the analyses, such as FootPrinter and PhastCons (Aplan et al., 1997; Larson et al., 1996). However, it is evident that some of these approaches show a moderate accuracy of identifying known functional binding sites (Mead et al., 1998). These approaches are typically complicated by two challenges: (i) conserved genomic region do often contain functional regulatory elements but not always the case (Cameron et al., 2005; Dooley et al., 2005) and (ii) not all TFBSs are conserved among species (1/3 in human and rodents). In addition to that, high conversation of some important transcription regulatory elements related to human development and disease may only be found between human and other close primate relative. The example has been shown that these weakly conserved TFBSs may have important medical roles (Aplan et al., 1992c). Advances in computational methods would allow developing new analytical approaches for detection of TFBSs with a much higher sensitivity.

1.2.1.4 Identification of transcription factor binding motifs

To fully understand the regulation of a gene, it is important to not only identify its *cis*-regulatory elements but also the associated transcription factors that bind to these elements. In general, most of the transcription factor binding sites (TFBSs) are very short sequences with 6-12 bp and are degenerated in nature; but only 4-6 bp within each sites are highly conserved (Maniatis et al., 1987). The availability of consensus TFBSs allows constructing databases which can be used to identify

potential TFBSs in a given DNA sequence. There are several online transcription factor databases, such as TRANSFAC (Wingender et al., 2000), TRRD (Transcription Regulatory Region Database) and COMPEL (composite regulatory elements database) (Heinemeyer et al., 1999), which provide a comprehensive list of experimentally determined transcription factors along with their binding sites. As the result of the short length and degenerate nature of TFBSs, the output from the TFBSs database searches contains a large number of false-positive predictions. Nevertheless, the rate of false-positive predictions can be reduced by detecting clustered or composite TFBSs (Pennacchio and Rubin, 2001). The idea of detecting clustered TFBS is to identify sets of common sequence motifs in the upstream regions of a set of genes which are likely to be co-regulated. It allows researchers to identify known as well as novel motifs that might be associated with a transcription factor. A number of algorithms are available for this propose including AlignACE (Roth et al., 1998), MEME (Bailey and Elkan, 1995), MDScan (Liu et al., 2002) and NestedMICA (Down and Hubbard, 2005). Meanwhile, the false-negative prediction in TF binding sites identification is another issue when searching the databases, as the list of TFs and their binding sites is not exhaustive in these databases. Comparative sequence analysis is particularly useful in the above instances, which can identify the presence of conserved TF binding sites that have been missed from the prediction in using sequence from a single species. It is considered that highly conserved TF binding sites identified by comparative sequence analysis across species are much more reliable than those detected only in single species. These short orthologous sequences conserved over 6 bp or more are termed as ‘phylogenetic footprint’. A web-based computational tool, ConSite (http://www.phylofoot.org), allows researchers to generate their own phylogenetic footprint (Lenhard et al., 2003). In addition, more advanced tools such as JASPAR have been developed using more sophisticated statistical-based models of TFBSs (Sandelin et al., 2004).

1.2.2 Chromatin immuneprecipitation (ChIP) assay

Chromatin immunoprecipitation (ChIP) has become the most widely used approach to study protein-DNA interactions *in vivo* (O'Neill and Turner, 1996). ChIP technique has been applied for mapping the localisation of histone modifications and histone variants in the genome, for identifying DNA-binding sites

for transcription factors and other chromosome-associated proteins (Collas, 2010). The procedure of the ChIP analysis is illustrated in Figure 1.6.

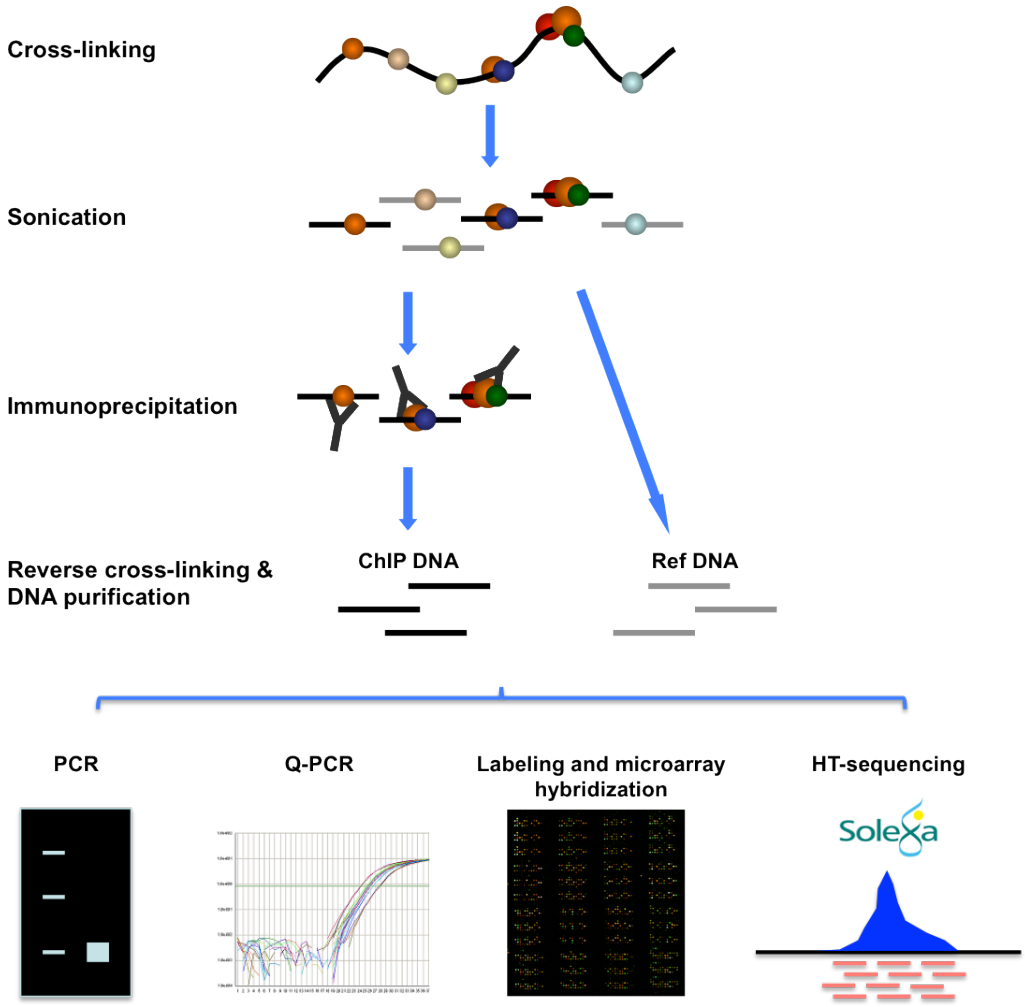


Figure 1.6: Schematic diagram of ChIP-chip assay. The description of the method is provided in the text. The basic steps involved in the method are cross-linking the DNA-protein complexes, sonication to generate sheared fragments, immunoprecipitation with a specific antibody, and DNA purification. ChIP-enriched DNA fragments can be detected by PCR, q-PCR, microarray and high-throughput sequencing.

First, protein-DNA interactions in chromatin are covalently cross-linked by formaldehyde treatment. As formaldehyde cross-linking DNA fragments and proteins within approx. 2Å of each other, it is only suitable to capture direct protein-DNA binding events. In order to overcome the limitation due to short cross-linking arm of formaldehyde, a variety of long-range bi-functional cross-linking reagents have been used in combination with formaldehyde to examine proteins both direct and indirect association with the target DNA sequences (Zeng et al., 2006). In contrast, a ChIP approach termed as native ChIP (NChIP) omits cross-linking step, which is well suited for the analysis of histones as they have high affinity for DNA. Subsequently, the chromatin fibers are fragmented either by

enzymatic digestion (e.g. MNase), or typically by physical shearing through sonication which randomly breaks chromatin into pieces of approx. 200-1000 bp, with an average of 500 bp in size. Third, the protein-bound chromatin segments are immunoprecipitated using specific antibodies to the proteins of interest. The non-specifically bound chromatin segments are removed by washing under stringent conditions, then the cross-linking is reversed and the precipitated ChIP-DNA is purified. DNA fragments specifically enriched by ChIP antibodies can be either quantified by PCR or quantitative PCR (q-PCR) at particular loci, or mapped in a genome-wide manner using DNA microarrays (ChIP-chip) (Buck and Lieb, 2004; Wu et al., 2006) and high-throughput sequencing (Barski et al., 2007).

Chromatin immunoprecipitation in combination with DNA microarray and sequencing marked the beginning of a new era for studying chromatin modifications. The technology has enabled the genome-wide mapping of transcription factor binding sites, histone modifications and nucleosome distribution. For instance, the ChIP-chip analysis has been extensively used for mapping global binding sites of c-Myc (Dang et al., 2006), elaborating transcriptional networks of Oct4, Nanog and Sox2 in ES cells (Boyer et al., 2005), identifying target genes of polycomb (Boyer et al., 2006) as well as profiling the landscape of histone modifications in T-cells (Roh et al., 2006). Thanks to the high-throughput sequencing technology, ChIP-seq has been used to generate 'chromatin-state map' for ES cell and lineage-committed cells (Mikkelsen et al., 2007). In addition, ChIP-seq has been applied to map global binding profile of transcription factor STAT1 in HeLa cells (Robertson et al., 2007). Furthermore, ChIP-seq has been used to generate a comprehensive map for histone methylations, histone variant H2A.Z, RNAP II and CTCF binding across the human genome (Barski et al., 2007). ChIP data generated using different detection platforms (e.g. qPCR, microarray and sequencing) illustrate robust overlapped profiles, reflecting the high reproducibility of ChIP-based technology.

1.2.2.1 Variations of the ChIP approach

DNA adenine methyltransferase identification (DamID)

DamID involves the labeling of DNA proximal the binding sites of proteins, such as transcription factors (Hwang et al., 1993). In DamID assay, the transcription factor under study is fused with the *Escherichia coli* DNA-adenine methyltransferase

(Dam) protein and is expressed in a cell culture system. The fusion Dam protein subsequently methylates the adenine base in GATC sites ~1.5 to 2 kb proximal the transcription factor binding sites. DNAs from this experimental sample and from a control sample, where only the Dam protein is expressed, are digested with a methylation-sensitive restriction enzyme (DpnI) and subsequently labelled and hybridised onto microarrays. DamID has been widely applied to characterize transcription factor binding sites, DNA methyltransferases as well as heterochromatin proteins in numerous species including Arabidopsis, Drosophila and mammals (Chung et al., 2002; Drake et al., 1997; Tompa et al., 2002; van Steensel, 2005; Visvader et al., 1998).

MeDIP: Methylated DNA Immunoprecipitation

MeDIP consists of the immunoprecipitation of methylated DNA segments with an antibody targeting 5-methyl cytosine (Keshet et al., 2006; Weber et al., 2005). The detection of captured methylated DNA segments can be done using PCR, microarrays or high-throughput sequencing (Taylor et al., 2007; Wilson et al., 2006). Although the assay is limited in identifying methylated regions with a CpG density below 2-3% (Keshet et al., 2006), MeDIP is being increasingly used for profiling methylation patterns of promoters in a number of organisms and cell types (Weber et al., 2005; Weber et al., 2007).

1.2.3 High-throughput approaches for genome-wide analysis

1.2.3.1 DNA microarray

DNA microarray is an important technology for investigating genome-wide gene regulation evens (Hoheisel, 2006). Generally, DNA microarrays consist of a large pool of DNA sequence which are attached to the surface of a glass silde. The DNA sequences can be large genomic clones such as BACs, PACs and cosmids, cDNA clones, PCR products or short oligonucleotides (Dhami et al., 2005; Duggan et al., 1999; Fiegler et al., 2003). High-density oligonucleotide microarrays are generated involving the direct synthesis of oligonucleotides using photolithography, optical mirrors or ink-jets on the slide surface (Hughes et al., 2001; Lipshutz et al., 1999; Singh-Gasson et al., 1999). Spotted arrays require robotic devices to spot clone sequence fragments, PCR amplicons or oligonucleotides (Schena et al., 1995). The glass surface of slides is coated with reactive molecular groups e.g. poly-L-

lysine to allow binding of DNA probes to the slide. The test sample (i.e RNA or DNA) and a reference sample are generally fluorescently labeled with nucleotide derivatives such as Cy3 and Cy5 for spotted microarrays. The repetitive DNA sequences on the microarray can be quenched using Cot1 DNA in a competitive hybridization. During hybridization, the fluorescence labeled DNA samples bind to their complimentary immobilized probes sequences and the fluorescent signal is measured in both channels to determine the enrichment in the test sample relative to the reference sample.

Microarray technology is a widely used application for studying global gene transcription regulation. For instance, microarray has been used to study gene expression in different organisms and experimental or disease systems (Shipp et al., 2002; White et al., 1999). Moreover, the microarray-based analysis has been used to profile replication timing in yeast, flies and humans based on the copy number ratio during DNA replication (Raghuraman et al., 2001; Schubeler et al., 2002; Woodfine et al., 2004). In addition, microarray technology has been extensively used for genome-wide detections in conjugating with a number of approaches, including DNase I hypersensitive site mapping, DNA methylation mapping, chromatin immunoprecipitation analysis and chromosome conformation capture-related analysis (as discussed previously and/or will be discussed in the following sections).

1.2.3.2 Next generation sequencing (NGS)

The automated Sanger method has dominated the filed of sequencing for last two decades and contributes to a number of monumental accomplishments, including the completion of the only finished-grade human genome sequence ((Jazag et al., 2005; Kim et al., 2012; Song et al., 2010). The method is known as a “first generation” sequencing technology, and the newer sequencing methods, which have been developed over last ten years, are referred as “next-generation” sequencing (NGS). The NGS technologies are commercially available from Roche/454, Illumina/Solexa, Life/APG and Helicos BioSciences, Pacific Biosciences/ SMRT® as well as Oxford Nanopore sequencing (Diakos et al., 2007; Elkon et al., 2005). These platforms coexisting in the marketplace are varied in NGS features, as the results of some technique having clear advantages for particular applications over others. These techniques offer a great capacity of

producing an enormous volume of data cheaply, in hundreds millions or even one billion short reads per instrument run, which largely expends the boundary of experimentation beyond simply determining the genome sequence of organisms. However, the voluminous data produced by these NGS platforms place substantial demands on information technology in terms of data storage, tracking and quality control (Kim and Rossi, 2007).

Microarray techniques has now been replaced by the NGS methods for studying gene expression (e.g. RNA-seq), which are capable of determining rare transcripts without prior knowledge of a particular gene or alternative splicing and sequence variations in known genes (Jiang et al., 2008; Silva et al., 2005). Additionally, its ability to sequence the whole genome cheaply has allowed conducing large-scale comparative and evolutionary studies in many related organisms. The NGS technologies can also be coupled with other methods such as ChIP and 3C-related techniques including ChIP-seq, 3C/4C/5C-seq, Hi-C and ChIA-PET (see below), which allow profiling genome-wide DNA-protein and DNA-DNA interactions in one base pair resolution, although its application is currently restrained by the limited sequencing depth.

1.2.4 Characterising of chromatin organisation in transcriptional regulation

As mentioned previously, *cis*-regulatory elements such as enhancers and promoters are capable of physically interacting with each other across large genomic distance, which is achieved by the creation of chromatin loops, to regulate transcription. The structure features and dynamics of the higher-order 3D chromatin organisation is essential for understanding cellular processes including gene transcription, DNA repair, recombination and replication. In this section, it illustrates the chromosome conformation capture (3C)-related methods that enable to characterise the chromatin organisation. Our knowledge about higher-order chromatin structure has been largely extended thank to the application of these methods during the last decade.

1.2.4.1 Overview of 3C-related techniques

3C-related approaches are now widely used in studying three-dimensional chromatin organisation, a decade after the chromosome conformation capture (3C)

technology has been initially developed (Dekker, 2002). The first 3C paper, applied this cross-linking dependent approach to study the frequency of interaction between multiple genomic loci in the budding yeast *Saccharomyces cerevisiae*, thus demonstrating the capability of studying spatial organization of entire genomes in eukaryotes (Dekker et al., 2002). Hagège and colleagues improved 3C assays to become more quantitative by introducing real-time TaqMan PCR technology, rather than using conventional measurements which rely on the intensity of stained PCR products separated by gel electrophoresis (Hagege et al., 2007). Initially, 3C protocols were only applied to yeast, drosophila and mammalian cells until recently an adapted 3C method was designed for specifically for studying plant tissues (Louwers et al., 2009a; Louwers et al., 2009b).

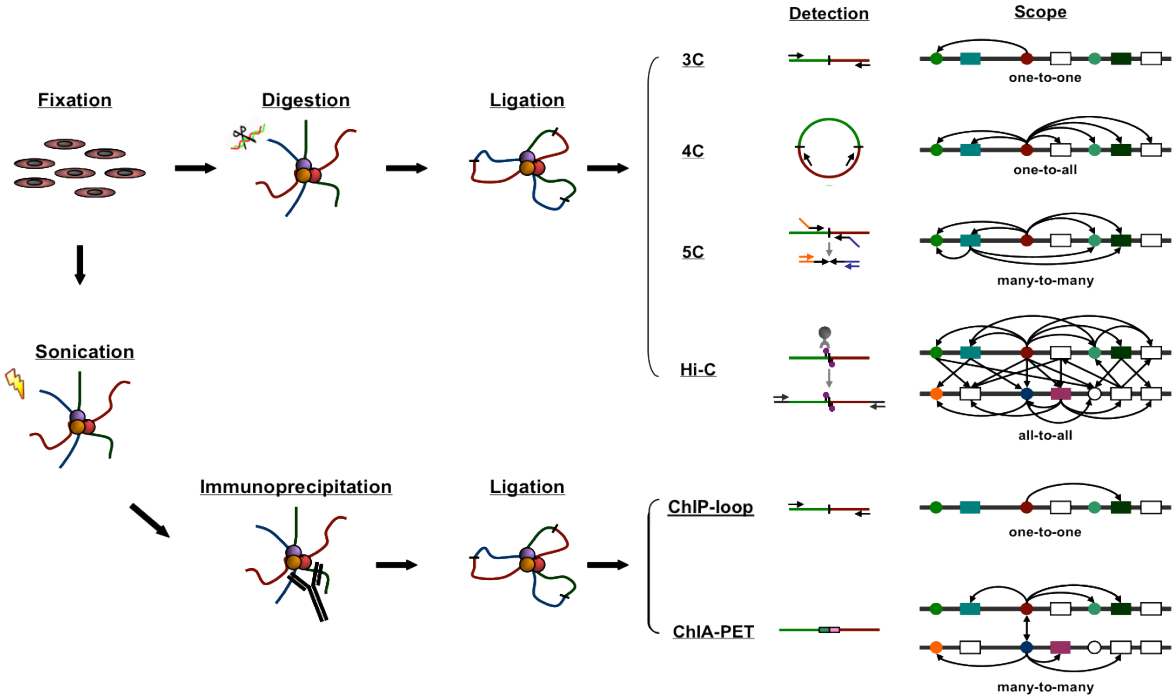


Figure 1.7: Overview of 3C/ChIP-derived methods. An overview of methods that are discussed is given. The left-panel shows the fixation, digestion, and ligation steps common to all of the 3C-dervied methods as well as the sonication, immunoprecipitation steps common to the ChIP-derived methods. The vertical panels indicate the steps that are specific to the separate methods.

The 3C-related methodologies have developed rapidly in the last five years (Figure 1.7). These approaches include 4C (circle 3C or 3C-chip, and e4C) (Gondor et al., 2008; Schoenfelder et al., 2010b; Simonis et al., 2006; Zhao et al., 2006), 5C (Carbon Copy 3C)(Dostie and Dekker, 2007; Dostie et al., 2007) and GCC (Genome Conformation Capture) (Rodley et al., 2009); the ChIP-dependent

assays such as ChIP-3C (Cai et al., 2006; Horike et al., 2005; Kumar et al., 2007) and 6C (Combined 3C-ChIP Cloning) (Tiwari et al., 2008); as well as the whole genome-wide capture approaches, like ChIA-PET (Chromatin Interaction Analysis with Paired-End Tag Sequencing) (Fullwood et al., 2009a; Fullwood and Ruan, 2009; Fullwood et al., 2009b) and Hi-C (Lieberman-Aiden et al., 2009). Each of these 3C-related techniques is described below in sections.

1.2.4.2 3C (chromosome conformation capture)

3C has now been widely used to detect and quantify physical interactions between genomic regions both in *cis* and *trans*, which consist of four major steps: cross-linking, digestion, ligation and PCR detection (Figure 1.7). First, 3C uses formaldehyde cross-linking to covalently fix interacting chromatin segments in cells. Subsequently, the fixed chromatin is digested using an appropriate restriction enzyme into smaller fragments, followed by intra-molecular ligation of cross-linked chromatin fragments. The resulting 3C library contains a large number of ligation products, representing interactions between pairs of genomic regions. These interactions can be detected by PCR with specific primer pairs, and the interaction frequency of two genomic regions is represented by the abundance of corresponding ligation product amplified by PCR. Technical details of each step are further characterised in the following sections.

Cross-linking: Formaldehyde cross-linking is used to covalently link interacting chromatin segments in living cells. Formaldehyde is a strong but reversible cross-linking agent that efficiently cross-links DNA-protein, RNA-protein and protein-protein *in vivo* (Orlando et al., 1997). Increased formaldehyde concentration and/or cross-linking time may result in over cross-linking, which may increase the chance of non-specific binding between protein and DNA as well as inhibit restriction digestion of chromatin DNA. In contrast, weak cross-linking leads to the loss of fixation stability between DNA interacting partners providing a less informative snapshot of higher-order chromatin organisation. Furthermore, evidence also suggests that cross-linking conditions can affect the sonication of the chromatin (Orlando et al., 1997) and the efficiency by which different types of proteins are cross-linked (Solomon and Varshavsky, 1985). Therefore, the quality of the 3C sample (referred to here as the 3C library) can be affected by cross-linking conditions.

Digestion: The chromatin from lysed cells is digested by a restriction endonuclease (Figure 1.7). Six-base cutters, such as BglII and EcoRI, are extensively used in preparation of the 3C library (Dekker et al., 2002; Gheldof et al., 2010; Louwers et al., 2009a; Louwers et al., 2009b; Solomon and Varshavsky, 1985; Spilianakis and Flavell, 2004; Spilianakis et al., 2005; Tolhuis et al., 2002). They normally generate digested fragments of several or even tens of kilobases (Kb) in size. In contrast, the four-base cutter restriction endonucleases, such as NlaIII and Csp6I, generate smaller digested fragments with an average size of several hundred bases. The four-base cutter is highly recommended when trying to determine interactions between closely spaced regulatory elements or study topology of small loci (Simonis et al., 2007). As if using the six-base cutter, all of those elements would lie on the same restriction fragment and would be impossible to distinguish from each other (Figure 1.8).

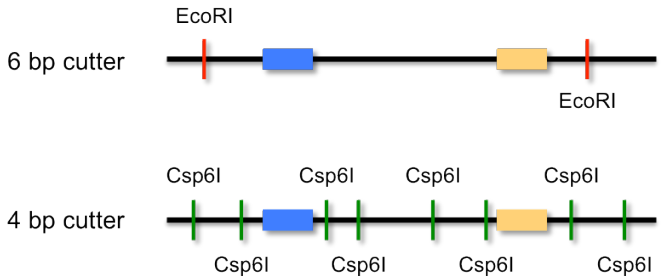


Figure 1.8: A schematics diagram of 6-bp and 4-bp cutters in assessing interactions between closely spaced regulatory elements. The black lines represent the chromatin DNA. The red and green bars represent the restriction sites of 6-bp (e.g. EcoRI) and 4-bp (e.g. Csp6I) cutters. The blue and yellow blocks represent the closely spaced regulatory elements.

Monitoring the digestion efficiency of the 3C chromatin is an important step for controlling quality of the 3C library. Insufficient digestion of 3C chromatin would subsequently result in reduction of ligation efficiency and compromise of the library quality. The restriction digestion efficiency can be affected by: i) over cross-linking (Splinter et al., 2004), ii) the presence of residual SDS (sodium dodecyl sulfate), iii) nuclei aggregation (Gondor et al., 2008; Hagege et al., 2007; Louwers et al., 2009b). Care should be taken to ensure digestion efficiency of the 3C chromatin is above 70% (Hagege et al., 2007; Simonis et al., 2007). Further discussions about control experiments for digestion efficiency are presented in section 1.2.4.3.

Ligation: The crosslinking will result in larger chromatin aggregates containing numbers of DNA fragments all together. After restriction digestion, ligation of DNA fragments is performed under the dilute condition (Figure 1.9). By doing that, the

intra-molecular ligation (refer to the ligation between DNA fragments within the same DNA-protein complex) is highly favoured over the inter-molecular ligation (refer to the ligation between DNA fragments in different DNA-protein complexes), and therefore largely reduces the chance of non-specific ligation (Splinter et al., 2004). For the intra-molecular ligation, all fragment ends compete with each other for ligation to the anchor fragment within these aggregates. Thus, even for a frequent and stable interaction between the enhancer and the promoter, it will only occasionally result in the corresponding ligation product. In addition to that, every anchor fragment is only present twice in a diploid cell, and therefore can maximally result in two ligation products of interest in a single cell. It means that the PCR amplification is required for detecting these rare “anchor-prey” ligation products from many other genome equivalents.

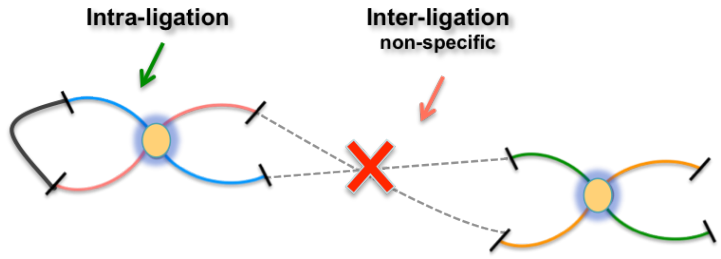


Figure 1.9: Schematics of the 3C ligation. The ligation between restriction fragments is performed in a dilute condition that favours the intra-molecular ligation over inter-molecular ligation. The yellow balls represent the protein complex and the colour lines represent the chromatin fragments. The black line indicates the intra-molecular ligation between two restriction fragments and the grey dashed lines indicate the inter-molecular ligation that is non-specific and not favoured in the dilute condition.

PCR detection: At most time, cross-linking results in large chromatin aggregates, within which all fragment ends compete with each other for ligation to the anchor. Due to that reason, even a stable and frequent interaction is only occasionally result in the corresponding ligation product that is to be amplified by specific PCR primer pairs. Additionally, a single diploid cell can only contributes maximally two ligation products of interest containing the anchor fragment. Consequently, it implies that a quantitative PCR amplification is required for detecting very rare ligation product from many genome equivalents. Therefore, it requires strict controls and careful experimental design and data interpretation (Dekker, 2006; Hagege et al., 2007; Simonis et al., 2007). After the ligation products of interest are amplified by PCR, the PCR products from each reaction are visualised by gel electrophoresis and quantified for their intensities. It is curial to appreciate that any two DNA fragments nearby on the linear chromosome are in spatial proximity, and

therefore frequently crosslink and ligate to the each other independently of the chromatin higher-order organization. Consequently, one needs to detect the anchor fragment interacting with a distal fragment more frequently than with intervening fragments in order to determine a real event of looping interaction.

1.2.4.3 Quality control and analysis of 3C experiments

Various technical issues need to be considered when interpreting 3C data. Important controls need to be performed to correctly interpret the 3C-generated data. The control experiments include: i) controls for digestion, ii) controls for ligation, iii) controls for PCR efficiency, and iv) controls for random collision (Dekker, 2006; Hagege et al., 2007; Louwers et al., 2009b; Miele et al., 2006). A set of terms used in describing control experiments which will be frequently mentioned in this thesis are listed as follows:

3C non-digested control: the genomic DNA purified from the human and murine cells.

3C digested control: a small proportion of 3C chromatin taken directly after the restriction digestion of the 3C preparation. Cross-linking is then reversed and the DNA is purified. It is also referred to as the non-ligated control

3C ligated control: a small proportion of DNA taken from the 3C library which has been digested by the restriction enzyme and ligated by DNA ligase.

3C mammalian control template: DNA is first purified from the BAC (bacterial artificial chromosome), PAC (P1-derived artificial chromosome) or YAC (yeast artificial chromosome) clone(s) encompassing the genomic regions to be assessed in the 3C assay. It is then digested with the same restriction enzyme used for the 3C assay and randomly ligated *in vitro* to generate every possible combination of ligation product. The 3C mammalian control template (also called the 3C control template) is used for calibrating the PCR efficiency of the 3C primer sets.

Controls for digestion: The quantitative-PCR (q-PCR) is used to determine the digestion efficiency. The primers are specifically designed to span a restriction site

and the digestion efficiency can be calculated by comparing the difference of PCR products generated from 3C digested control and 3C non-digested control (Hagege et al., 2007).

Controls for ligation: The ligation of chromatin to prepare the 3C library is monitored by two qualitative assays. First, the 3C non-ligated and ligated controls are visualised by gel electrophoresis and ligation is determined based on the DNA size distribution on the gel. By this analysis, the ligation reaction results in ligated DNA fragments which are relatively larger in size than the non-ligated control. Second, PCR detection of a ligation product of a chromatin loop in the *Ercc3* locus is used to check the ligation quality of the 3C library (Kurukuti, unpublished data). The *Errc3* gene is a ubiquitously expressed gene and the chromatin loops within *Ercc3* also have been extensively used as an internal control for normalising different cross-linked samples (Hagege et al., 2007; Kurukuti et al., 2006).

Controls for PCR efficiency: Ligation frequencies of interacting elements are determined by measuring the intensities of PCR products amplified with different primer pairs. PCR controls are always needed to normalise for differences in PCR efficiency so that ligation frequencies between different primer pairs can be directly compared (Dekker, 2006; Miele et al., 2006). The 3C control template is prepared for this purpose. The 3C control template is prepared either using genomic DNA (for organisms with relatively small genomes such as yeast) or alternatively using BAC/YAC clones covering the genomic segments under study (for organisms with large genomes such as human or mouse), for which the complexity of the resulting mixture of ligation products becoming too large to reliably and quantitatively detect individual ligation products (Palstra et al., 2003; Tolhuis et al., 2002). Theoretically, it contains all possible ligation products which are relevant to a genomic region of interest in equimolar amounts (Miele et al., 2006; Splinter et al., 2004). Providing the templates are in equimoalr amounts, the amount of PCR products can directly reflect the differences of PCR efficiency between different primer pairs. The primer combinations showing low PCR efficiency need to be re-designed.

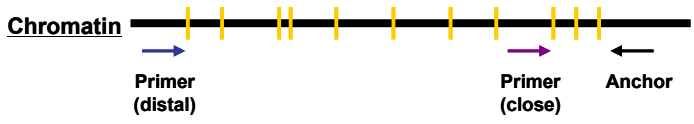


Figure 1.10: Schematic diagram of the primer pair selection for PCR titration experiment. The chromatin DNA is shown in black line and the restriction site is shown in yellow bars. A

"distal" primer (blue arrow), a "close" primer (purple arrow) and the anchor primer (black arrow) are used to determine the suitable DNA concentration for PCR amplification.

In addition, a PCR titration experiment is also critical for the 3C assay at this stage. Both the 3C control template and the 3C library need to be titrated to determine linear range of the amount of DNA template used for the PCR reactions. Titrations are performed with two different primer combinations, which should differ in genomic distance that separates the restriction fragments they recognize (as shown in Figure 1.10). Preferably, one of the primer pairs ("distal") should amplify a ligation product representing an interaction between restriction fragments that are relatively far apart (several 10 kbs) on the genomic map, whereas the other primer pair ("close") should amplify a ligation product representing an interaction between two fragments that are located close together (around 10 kb).

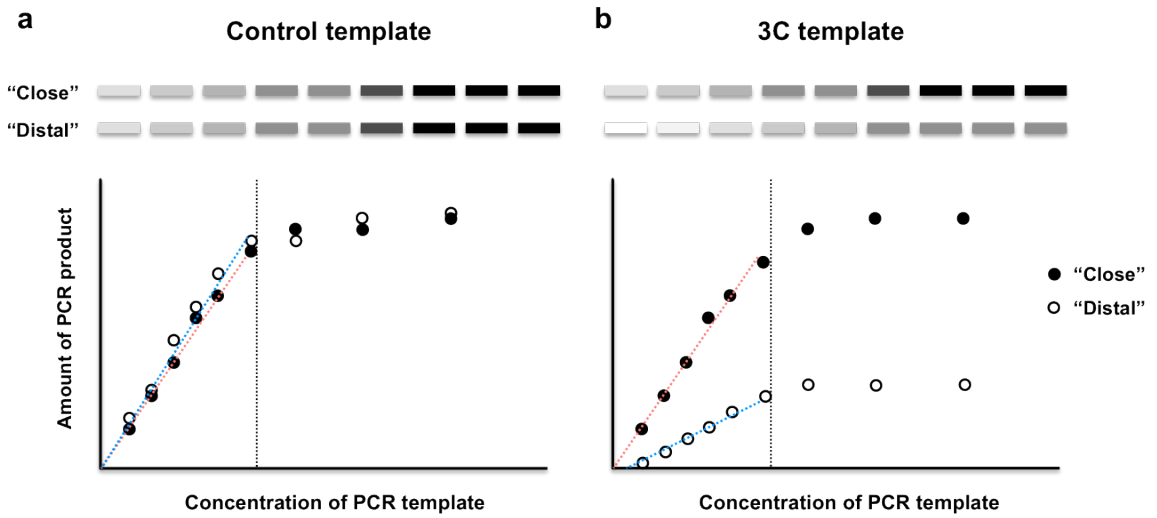


Figure 1.11: Schematics of titration of PCR templates concentration for gel quantification in the 3C. Panel a: titration plots of control template and panel b: titration plots of 3C template with two primer combinations. The PCR products visualized by gel electrophoresis are shown schematically at the top of the figure. The "close" (black ball) and "distal" (white ball) primer pairs are defined based on the relative genomic distance to the anchor. The x-axis represents the DNA concentration of templates for the 3C-PCR and the y-axis represents the amount of PCR product detected by gel quantification.

Theoretically, the amount of PCR products generated by the "close" and "distal" primers when used in combination with the "anchor" should be more or less equal when using the 3C control template (Figure 1.11a). This is because the 3C control template should provide equivalent amounts of random ligation products of all possible combinations encompassing the genomic region contained within the BAC/PAC clone (Miele et al., 2006; Splinter et al., 2004). In contrast, the "close" primer should generate more PCR product in comparison with the "distal" primer when using a 3C library as template (Figure 1.11b). As the non-specific ligation

frequency decrease as a function of distance to the anchor, the “close” fragment has a much higher likelihood of being randomly ligated to the anchor comparing to the “distal” fragment. The linear ranges of PCR amplification can be determined based on the titration plots as schematic examples illustrated in Figure 1.11. The amount of template chosen for 3C-PCR analysis and gel visualisation should be in the linear range of PCR amplification for both “distal” and “close” primer pairs (Miele and Dekker, 2009).

Controls to distinguish real looping from non-specific interactions: Detection of a ligation product between two non-contiguous DNA fragments does not necessarily mean that they are engaged in a functional looping interaction. This is because genomic regions along the chromatin fibres also randomly collide at a relatively high frequency as a result of the inherent flexibility of chromatin (Dekker, 2006; Dekker et al., 2002). Therefore, the level of random collisions needs to be assessed before determination of real interactions from those due to background levels. The frequency of non-specific interaction between two regions is inversely related to their genomic distance and is found to decrease as they moving apart from each other based on the theoretical predictions (Rippe, 2001). As shown in Figure 1.12, the interaction between a gene and its distal enhancer is determined using a fixed primer at the promoter of the gene (termed as “anchor” or “bait”) and sets of primers located with the increasing distance from the “anchor” along the chromatin. Relatively high but non-specific interactions are observed between the “anchor” and fragments separated by a very short genomic distance. However, the interaction frequency of random collisions is expected to decrease with the increasing distance along the chromatin. Therefore, a real looping interaction between the gene promoter and the enhancer is determined based on the observation of a distinctive peak above the theoretical level of background non-specific interactions. Moreover, the question about whether the looping interaction is functional can be addressed by studying whether it is transcription-dependent with additional experiments. For instance, the looping interaction between the promoter of a gene and the enhancer may be only present in certain cell types when the gene is expressed. Based on this principle, the control 3C assays are designed for detecting ligation products between the anchor and sequences which flank the target regions of interest (for example, a *cis*-regulatory element). The ligation frequencies are determined for both the target and the controls. A looping

interaction can only be determined by demonstrating the ligation frequency of the target is significantly higher in comparison with controls based on statistical analysis.

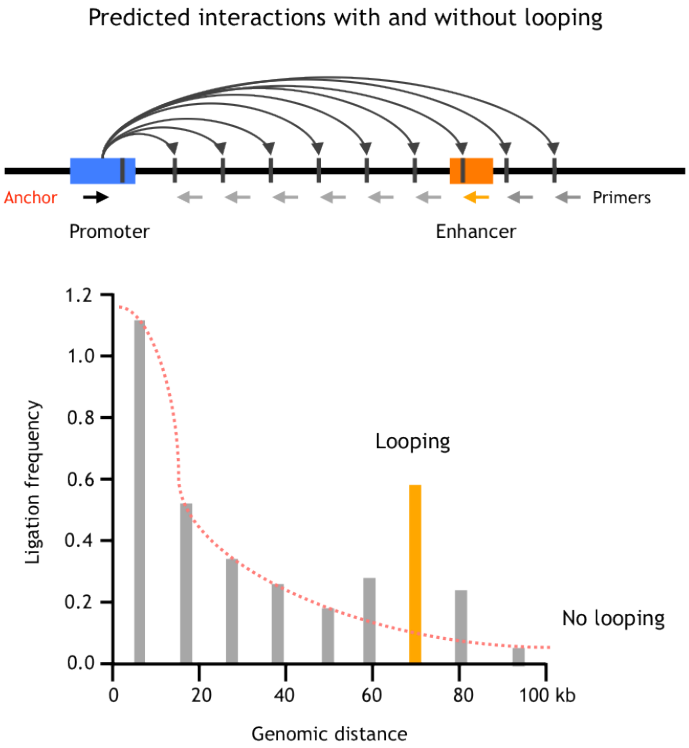


Figure 1.12: A theoretical example of a typical 3C analysis to detect in cis interactions between a gene promoter (blue) and a enhancer (orange). Red dot line illustrates the predicted pattern of random ligation events between the promoter and sites located up to 100 kb without looping. Bars indicate the expected pattern of looping interactions between the gene promoter and the distant enhancer. The presence of the true looping interaction found at higher frequency that background levels is apparent by a distinct peak in ligation frequency (the bar in orange).

Controls to allow comparison between different 3C samples: Studies of chromatin interactions are quite often conducted in either between different cell types or different experimental conditions (Chien et al., 2011; Tolhuis et al., 2002). In these instances, another internal control for normalisation, a universally transcribed locus (housekeeping gene), which is assumed to have a similar chromosome conformation and ligation frequency in different cell types, can be used as a control for comparing 3C data obtained from different samples or conditions (Dekker, 2006; Kurukuti et al., 2006). All of these control experiments discussed above serve as guidelines to avoid misinterpretation of 3C results.

1.2.4.4 4C (circular 3C/ 3C-on-chip)

Originally, 4C (chromosome conformation capture-on-chip) was established in de Laat's laboratory to analyse the spatial proximities of a selected genomic site

("viewpoint", "anchor" or "bait") with all other genomic fragments by combining the features of the 3C technology with microarrays (Simonis et al., 2006). In parallel, Ohlsson and colleagues developed a slightly different 4C variant called circular chromosome conformation capture using similar principles (Gondor et al., 2008; Zhao et al., 2006). In addition to those major 4C variants mentioned above, a modified approach termed "e4C" (enhanced chromosome conformation capture-on-chip) has recently been published (Sexton et al., 2012). This 4C variant was initially used with the help of an additional chromatin immunoprecipitation (ChIP) step in studying the Klf1-regulated DNA interactions of the active globin loci in erythroid cells (Schoenfelder et al., 2010b). Additionally, next generation sequencing (NGS) has been incorporated into various 4C approaches, which is referred as 4C-seq (Raab et al., 2012; Splinter and de Laat, 2011). In the subsequent sections, a detailed introduction to three mainstream 4C variants will be presented in a comparative way. In brief, 4C shares four common experimental steps with conventional 3C procedures, which include (in the order by which they occur): (i) formaldehyde cross-linking, (ii) restriction enzyme digestion of intact nuclei, (iii) ligation and (iv) reversal of cross-linking (details shown in Chapter 3, section 1 and summarised in Figure 1.13). However, downstream experimental steps used in 4C differ from 3C.

Chromosome conformation capture-on-chip

3C-on-chip relies on the formation of DNA circles after reversal of cross-linking (Figure 1.13a). Similar to the conventional 3C analysis, cross-linked chromatin DNA is digested with a 6-base recognizing enzyme, such as HindIII. After intra-molecular ligation, cross-links are reversed and ligated DNA fragments are trimmed with a 4-base cutter for a secondary restriction digestion. DNA circles are formed at this stage by re-ligation under conditions which favour self-ligation. Subsequently, the ligation products are inverse PCR amplified and detected using microarrays.

Circular chromosome conformation capture

For circular 3C, the formation of DNA circles is performed by the ligation of DNA fragments digested with a 4-bp restriction enzyme when chromatin DNA is still cross-linked (Gondor et al., 2008; Lomvardas et al., 2006; Zhao et al., 2006). To form such a circular structure, it requires both the anchor (also known as "bait")

sequence and captured ("prey") sequence to be ligated at both ends (Figure 1.1.3b). Circular DNA fragments are directly PCR amplified by inverse bait-specific primers after reversal of cross-linking. Detection of the PCR products is either by microarray hybridisation or DNA sequencing.

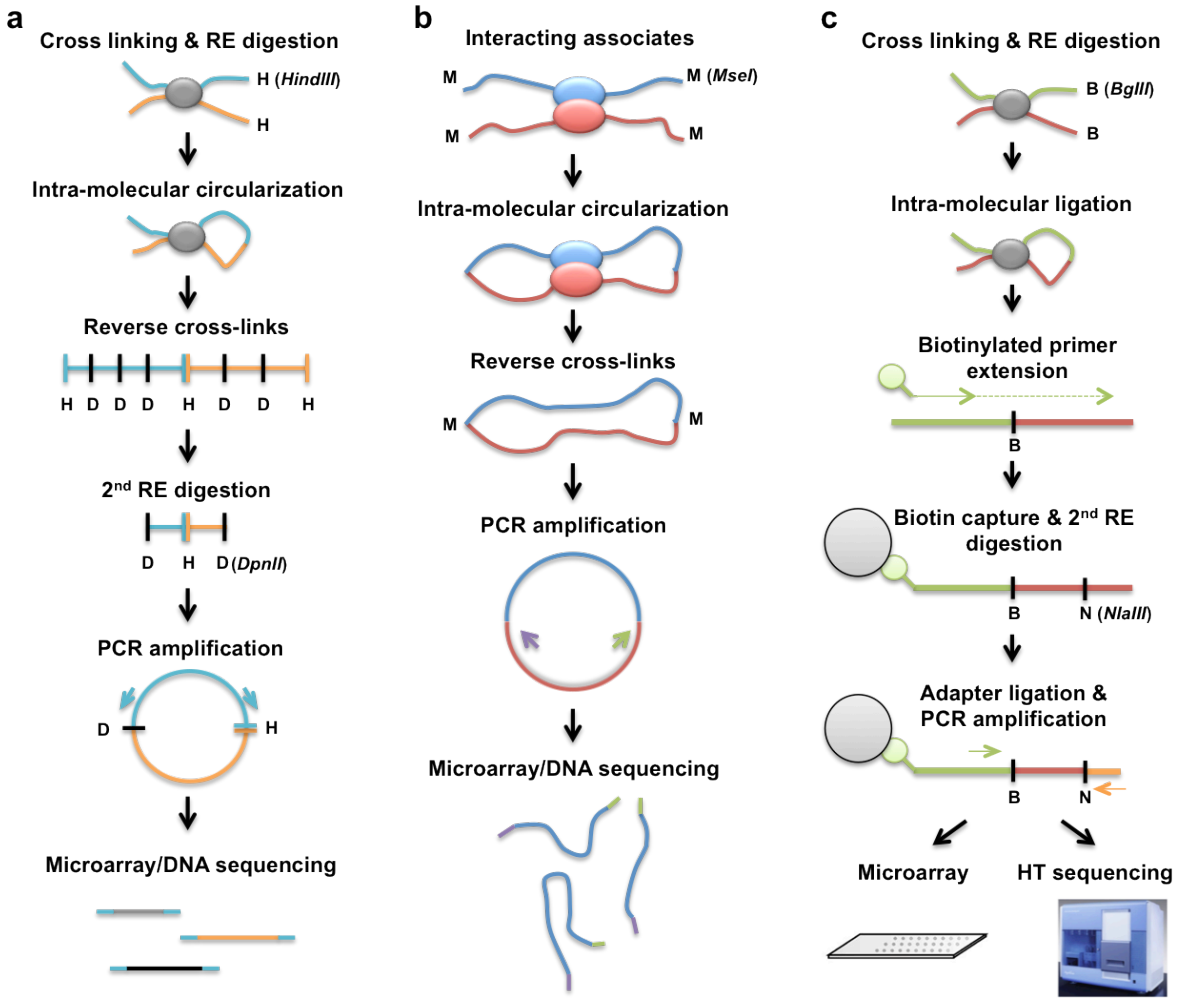


Figure 1.13: Outline of three 4C strategies. Panel a: chromosome conformation capture-on-chip (Simonis et al., 2006), panel b: circular chromosome conformation capture (Gondor et al., 2008) and panel c: enhanced 4C (Sexton et al., 2012). In panel (a) a six-base cutter enzyme, HindIII (H) is used to digest the cross-linked DNA fragments. The ligation junctions are trimmed with a second restriction digestion, DpnII (D) after de-crosslinking and then circularisation for PCR amplification. PCR products can be characterised on microarray or DNA sequencing. In panel (b), a four-base cutter enzyme, MseI (M), is used to digest the chromatin – thus providing a high resolution (256 bp in theory). This method relies on the circular ligation of cross-linked DNA fragments. PCR products are analyzed by microarray or sequencing analysis. In panel (c), e4C entails the fixation, digestion and ligation steps of conventional 3C, followed by the anchor-specific enrichment by biotinylated primer extension and pull-down, adapter ligation and PCR amplification. Detection of e4C library can be performed using either microarray or high-throughput sequencing.

Enhanced-4C

The e4C (enhanced chromosome conformation capture-on-chip) technique incorporates additional enrichment steps targeting the "anchor"-specific sequences,

which significantly improves the sensitivity of detecting distal chromatin co-localisations (Sexton et al., 2012). As shown in Figure 1.13c, e4C is composed of conventional 3C steps followed by enrichment of the “anchor” sequence using biotinylated primer extension and pull-down with streptavidin beads, adapter ligation and PCR amplification with a “anchor”-specific, nested primer. This additional biotinylated pull-down step provides over 100-fold pre-enrichment of anchor-linked sequences comparing to other 4C approaches, which facilitates the identification of both *cis* and *trans*-interactions with higher sensitivity and confidence. The “anchor”-specific chromatin interactions can be subsequently assessed via hybridizing e4C library to microarrays or high-throughput sequencing. In addition, an optional chromatin immunoprecipitation (ChIP) step can be incorporated with e4C approach, to enrich for particular subsets (e.g. a transcription factor of interest) of chromatin interactions (Schoenfelder et al., 2010b).

1.2.4.5 Comparison of the three different 4C strategies

The fight between brothers - Circular 3C versus 3C-on-chip

Circular 3C and 3C-on-chip are the two most frequently used 4C strategies for investigating DNA-DNA interactions during the last five years. Both strategies require the formation of DNA circles containing at least two sequences not normally adjacent to each other in the genome. And both strategies have their own advantages and disadvantages. One advantage of circular 3C is that it requires fewer processing steps in comparison with 3C-on-chip. Furthermore, higher resolution is one of the major advantages of circular 3C method. Restriction digestion with a 4-base cutting enzyme is preferred over a 6-base cutter for this method. This allows for the production of smaller fragments (average fragment size: 256 bp versus 4 kb) thus providing a higher resolution and accurate mapping of interaction points at a gene or loci of interest. The smaller fragment size also facilitates more uniform representation of all interacting fragments - since fragment length does not become limiting for PCR. However, the major limitation of this strategy is that the formation of DNA circles may contain multiple interacting sequences, which makes the ligation products too large to be amplified by PCR. Furthermore, the presence of protein-DNA complexes close to the restriction sites

of DNA fragments may also hamper the formation of DNA circles, thus reducing the representation of some types of DNA-DNA interactions.

The advantage of 3C-on-chip is that the formation of DNA circles is more efficient compared with circular 3C. This is because the circularisation is performed with naked DNA fragments after reversal of cross-links. Moreover, the circularisation process with this strategy relies only on ligation of a single end from both anchor and interacting fragments, while for circular 3C both ends of the two DNA fragments need to be ligated to each other (Gondor et al., 2008). In addition, the secondary restriction digestion using a 4-bp cutter remarkably reduces the size of 4C genomic library to be analysed, as 1) the size of interacting fragments is reduced and 2) the anchor-specific primer pair selectively amplifies the ligated outer ends of the interacting fragments digested by the 6-bp cutter (in 1st round of digestion, see Figure 1.1.3a) Taking advantage of this, a cost-effective tailored microarray can be designed with probes only located within 100 bp of each restriction recognition site of a given 6-bp cutter, which provides genome-wide coverage for analysing a 4C library (Simonis et al., 2006). It has been demonstrated that one can scan the entire human genome at an average resolution of 7 kb using 3C-on-chip with a specifically designed tailored array (Simonis et al., 2009; Simonis et al., 2006).

Enhanced-4C: the new-born versus old brothers

In comparison with the circular-3C and 3C-on-chip, e4C approach is fairly new in the field. The most obvious advantage of e4C is that its biotinylated enrichment for anchor-specific sequences largely has improved the sensitivity for detecting weaker and distal chromatin co-localizations. In addition to that, ligation products of e4C are linear fragments which are not relied on circularisation. It greatly improves PCR amplification efficiency at the final stage. In order to provide extra resolution of detecting, the e4C technique incorporates the step of 2nd restriction enzyme digestion from the 3C-on-chip approach. The disadvantage is that e4C technique has more sophisticated procedures comparing to other 4C approaches, which requires extra time and effort in library preparation and increases the likelihood of introducing additional bias. Despite the advantages and disadvantages of different variants of 4C technologies, it provides a powerful tool

for semi-quantitative assessments of genome-wide interacting patterns from the particular anchor points.

1.2.4.6 Potential issues of 4C techniques

The requirement for large cell numbers constrains the application of 4C for studying rare cell populations

The 4C technologies, particularly 3C-on-chip and circular-3C, are not very efficient in detecting long-range interactions due to their technical nature. The likelihood of capturing two frequently interacting fragments in the linear chromosome template is less than 1 in 500 cells, while probabilities of identifying distal *cis* or *trans* interactions are as low as 1 in 10,000 or even 1 in 100,000 cells (Simonis et al., 2007). Consequently, 4C analysis normally requires a large number of cells to provide a decent library complexity, in order to have a good coverage of genome-wide chromatin interactions. For instance, circular-3C (4C) technique may require 7 to 40 million cells as starting material for a genome-wide analysis when using tiling microarrays (Gondor et al., 2008). In contrast, much less starting material is required by 3C-on-chip (4C) strategies owing to the reduced library size as well as the use of tailored array. Generally, a single 3C-on-chip assay requires DNA equivalents from less than one million cells to sufficiently cover the possible interacting co-associates across the entire genome (Simonis et al., 2006). Nevertheless, the demand for a large number of cells is a big limitation of 4C techniques for studying some rare primary cells populations (e.g. HSC) and patient specimens. Thanks to the anchor-specific pre-enrichment of e4C technique, the specificity of PCR amplification is greatly improved and the amount of starting material required for profiling local (loci-specific) chromatin interactions are significantly reduced to the equivalent of about 100,000 cells (Kurukuti, unpublished data).

PCR over-amplification affects both sensitivity and reproducibility

Sensitivity and reproducibility are two key factors that must be monitored in order to assess the quality of 4C data. For high throughput microarray analysis, the array elements showing signals above background noise are those that have been enriched by the 4C procedure. However, given that the likelihood of interactions decreases as a function of distance from the anchor sequence, this can

compromise the detection of very long-range intra- and inter-chromosomal interactions over a background of random ligation events (since their enrichment levels will be relatively low).

In current 4C protocols, 30 to 36 cycles of PCR amplification is usually performed at the final stage of library preparation for microarray-based applications (Gondor et al., 2008; Simonis et al., 2006; Zhao et al., 2006). Moreover, an extra PCR amplification with 9-12 cycles is added onto the 4C procedure for high-throughput sequencing application (Raab et al., 2012; Sexton et al., 2012). Over-amplification of PCR is considered as a major issue that leads to compromise 4C library quality particularly for local interaction profiles, in the aspects of both sensitivity and reproducibility (unpublished data).

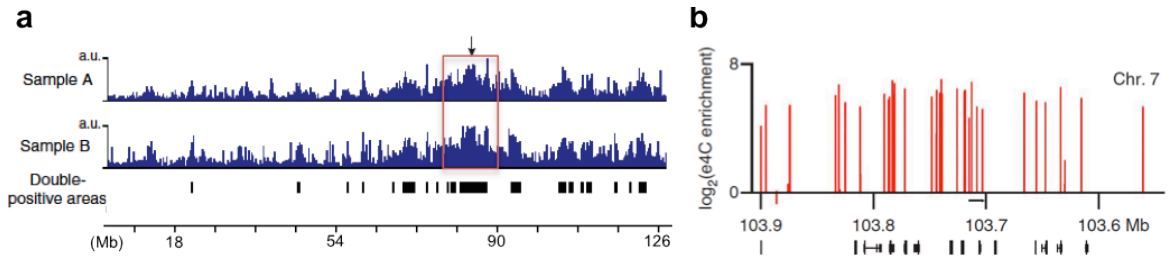


Figure 1.14: Saturation of local interaction profiles obtained by 4C technologies. Panel a: example of 4C (3C-on-chip) data analyzing mouse *Rad23a* locus (adapted from Simonis et al., 2007). Local interactions with the anchor are highlighted with the red box. Panel b: e4C microarray profile showing local interactions at the mouse *Hbb* locus (adapted from Sexton et al., 2012). The arrows denote the positions of the 4C anchors.

It has been found that genomic elements physically linked on the linear chromosome template are always heavily over-represented in 4C (3C-on-chip) analysis (Simonis et al., 2007). As a matter of fact, detection of the restriction fragments located within 5-10 Mb from the anchor is so efficient that the signal intensities of corresponding array probes are saturated, precluding a quantitative analysis of local interactions (shown in Figure 1.14a). Similarly, saturation of interactions at close-range with the anchor (Figure 1.14b) has also been found when using e4C technology with a custom microarray at the mouse *Hbb* locus. The authors suggested that the microarray application might not be suitable for assessing specific chromatin looping interactions within the gene loci (Sexton et al., 2012). However, it has been shown that the PCR over-amplification is the major reason why 4C failed to provide a quantitative assessment for the local chromatin looping interactions (further details, see Chapter 5 section 5.5 and 5.6).

High-throughput sequencing has now been applied for analysing chromatin interactions captured by 4C in a much higher (base-pair) resolution comparing to microarray application. A recent study using e4C-seq analysis illustrated a high-resolution interaction profile of the local environment of a human tRNA gene but failed to provide a genome-wide view due to lack of insufficient sequencing depth (Raab et al., 2012). However, the custom-designed microarray always provides coverage the entire genome in a cost-effective way. As the major limitation for sequencing-based analysis is the depth of sequencing, decisions between using microarrays and sequencing for 4C analysis therefore depends on the specific biological question being addressed and the budget of the experiment.

Technical limitations in verifying 4C results

3C and FISH technologies are commonly used to verify the interaction data captured by 4C assays. The technical advantage of 3C over FISH for 4C verification is its high resolution. Even with high-resolution FISH studies, such as 3D-FISH (Solovei et al., 2002) or cryo-FISH (Branco et al., 2008; Guillot et al., 2004), these methods are only able to verify the inter-chromosomal interactions captured by 4C (Simonis et al., 2006). In contrast, 3C is able to provide a similar level of resolution to 4C, as they are related technologies. As result of that, local interactions captured by 4C can only be verified using 3C, along with appropriate control assays which measure random noise (i.e., random ligation frequencies). However, 3C and 4C are not independent technologies, and may carry many of the same biases. In order to improve the reliability of 4C verification using 3C analysis, one can always i) make independent library preparation from independent biological replicates and ii) assess interactions in similar cell types (i.e., using a primary cell to compare against a similar cell line or visa versa).

1.2.4.7 Applications of 4C technology

4C approaches have been widely applied to investigate the DNA interaction profiles in a number of mammalian loci including α-globin, β-globin and Rad23a, illustrating the link between chromosome conformation and transcription regulation (Schoenfelder et al., 2010b; Simonis et al., 2006). In addition, 4C studies in *Drosophila* have found that genes repressed by Polycomb group (PcG) proteins are preferentially cluster within the same chromosome arm, which further elucidated the relationship between transcription and structure (Bantignies et al.,

2011; Tolhuis et al., 2011). 4C is the preferred approach to study the chromatin interacting profiles of individual genomic sites. However, the current protocols only allow to detecting long-range *cis*- and/or *trans*- interactions, whereas the local interactions (<50 kb) are not yet readily detected by 4C due to its limitation of resolution. This weakness has now been resolved by an improved protocol, which allows robust screening for important local interactions (see in Chapter 5).

1.2.4.8 5C (3C carbon copy)

5C is described as a “many versus many” technology (Figure 1.7), which allows determination of interactions between multiple genomic regions (Dostie et al., 2007; O'Neil et al., 2001). In 5C assays, the 3C ligation product is firstly prepared as for conventional 3C. A matrix of oligonucleotides are designed, each of which partially overlaps a different restriction site in the genomic region of interest. Only the pairs of oligonucleotides annealed to interacting fragments, which are juxtaposed on the 3C templates can be ligated together. It allows specific detection of head-to-head 3C ligation junctions, which can largely avoid the amplification of self-ligated restriction fragments or self-ligated partial digestion products. As the 5C ligation products quantitatively present the ligation product in the 3C library, the ligation products generated yield a ‘carbon copy’ of a selected part of the 3C library, hence the name 5C. The ligation products of 5C are subsequently amplified simultaneously in a single PCR reaction because that all 5C oligos carry one of two universal tag sequences at the 5’ ends. Detection of these ligation products can be performed either on a microarray or by high-throughput sequencing.

Differ to 4C, 5C approach provides a matrix of ligation frequencies for large number of sites pair-wisely, which situates the given DNA sites in the context of those between other pairs of sites and allows reconstruct the higher-order chromatin configuration of large genomic regions. 5C analysis has been performed to study the human β-globin locus (O'Neil et al., 2001), α-globin locus (O'Neil et al., 2004) and HOXA-D gene clusters (Green and Begley, 1992; Labastie et al., 1998; Varterasian et al., 1993). The interactions between regulatory elements and genes previously detected by 3C in the globin loci, which have also been picked up by 5C technology. However, differ to 3C, 5C is not widely used for identify enhancer-promoter interactions. As 3C assay is relatively easy to design and analyze, it is

considered to be more suitable for detecting interactions between specific pair of regulatory elements such as promoters and enhancers. The capacity of screening interactions in a greater throughput and minimized bias in PCR efficiency therefore makes 5C much superior to 3C.

1.2.4.9 Hi-C (High-throughput 3C)

Hi-C technology is one of the “all versus all” approaches, which allow detection of all potential interactions between all genomic sites (Lieberman-Aiden et al., 2009). In the Hi-C preparation, a slight adjustment is made when generating the 3C template (Figure 1.7). The restriction ends are filled in with biotinylated nucleotides followed by a blunt-end ligation. DNA is then sonicated and purified with biotin beads to ensure that only ligation junctions are selected for subsequent analysis. The ligation products are detected by NGS and sequencing reads are mapped back to the genome. The interaction between two genomic fragments can then be defined when a pair of DNA is identified on two different restriction fragments.

The Hi-C technology has been applied in two human cell lines (K562 and GM06990) with about 1Mb resolution (Lieberman-Aiden et al., 2009). The data suggest that the genes with same transcription state are preferentially clustered in the nucleuses. Furthermore, it has also been revealed that nuclear organization is quite constant between different cell lines, implying the existence of a core organization in most cell lines. A variation of the Hi-C method which is partially adapted from 4C strategy with a secondary restriction digestion and ligation, has been developed for studying the all versus all genome-wide chromosome conformation with kilobase resolution (owing to smaller genome and increased sequencing depth) in both *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* (Duan et al., 2010). The Hi-C data of *S. cerevisiae* confirm the previous observation of Rabl organisation by showing clustering of the centromeres as well as the telomeres. In addition, it has also observed that the genome-wide dispersed tRNA genes are co-localised in forming two distinct nuclear clusters. Finally, the Hi-C data also suggest that the chromosomes are well situated in chromosomal territories in both *S. cerevisiae* and *S. pombe*.

1.2.4.10 ChIP-loop (chromatin immunoprecipitation loop assay)

ChIP-loop (also known as ChIP-3C) is a variant of 3C-related technology incorporating a protein precipitation step to allow identification of genome interactions that involve a specific protein (transcription factor) of interest (Horike et al., 2005; Kumar et al., 2007). In ChIP-loop (Figure 1.7), cells are fixed with formaldehyde, lysed and the crosslinked chromatin is purified by urea gradient ultracentrifugation (Horike et al., 2005). The chromatin complexes are then digested with a restriction enzyme and subjected to immuoprecipitation with specific antibodies against proteins of interest. The standard ChIP procedure is performed with pre-clearing by protein A/G beads, antibody incubation and final washing of the beads. The precipitated chromatin is ligated by T4 DNA ligase. Ligation products are then purified and detected by PCR as in usual 3C assays. The ChIP-loop technology has been mostly replaced by its high-throughput version approach known as ChIA-PET.

1.2.4.11 ChIA-PET (chromatin interaction analysis by paired-end tag sequencing)

ChIA-PET technology combines chromatin Immunoprecipitation (ChIP) with 3C-related analysis and paired-end tags (PET) sequencing, which is capable of detecting chromatin interactions exclusively to those formed between sites bound by a given transcription factor or chromatin interacting protein (Fullwood et al., 2010; Fullwood et al., 2009a; Li et al., 2010). In ChIA-PET (Figure 1.7), those chromatin interactions are captured by chemical crosslinking. Following DNA-protein complexes after sonication are pulled down with an antibody against a specific protein of interest. Two separate aliquots of tethered DNA fragments in each of the chromatin complexes are connected to two different DNA linkers independently, followed by proximity ligation with samples being pulled down by antibody. Two types of ligation products are generated including self-ligation of the same DNA fragments and inter-molecular ligation of different DNA molecules. The ligation junctions with paired-end tags (PETs) are cut by restriction digestion and extracted for high-throughput sequencing. Reads are then mapped to reference genomes to reveal relationships between distal chromosomal regions brought together into close proximity by specific proteins of interest.

ChIA-PET technology has been applied to study genomic sites bound by the estrogen receptor α (ERα), CTCF and RNA polymerase II in mammalian cells (Handoko et al., 2011; Li et al., 2012; Pulford et al., 1995). The interactome map of the ERα binding sites showed a few thousand intra-chromosomal looping interactions, which were between highly enriched binding sites located in a relatively short distance (within ~100 kb) (Fullwood et al., 2009a). For the interactome mediated by CTCF in mouse embryonic stem cells, ChIA-PET has revealed looping interactions between ~10% of the CTCF binding sites both in *cis* and *trans* (Handoko et al., 2011). It was speculated by the author that much higher number of CTCF interactions would be detected with the improvement of sequencing strategy and depth. These CTCF loops were formed between distal enhancer and genes as well as between active and inactive chromatin, which highlight the role of CTCF in spatial organization of transcription regulation. ChIA-PET has also been used to map the chromatin interactions associated with RNAP II in human cells, which uncovered promoter-centered interactomes . It has been shown that genes with promoter-promoter interactions are mostly active, and trend to be regulated and transcribed cooperatively. Cell-specific chromatin interactions have also been identified by comparative analyses of different cell lines, suggesting particular chromatin organizations for cell-specific transcription regulations. Overall, ChIA-PET studies provide insights into transcription regulation by three-dimensional chromatin interactions in mammalian cells.

1.2.4.12 Microscopy observations into mammalian genome organization

Fluorescence *in situ* hybridisation (FISH) techniques have been extensively used in detecting long-range intra-chromosomal and inter-chromosomal interactions, which have contributed an extraordinary amount of our knowledge of the nuclear architecture. Initially, FISH is developed to detect specific DNA/RNA sequences in fixed single cells. By applying the fluorescently labelled probes, FISH is allowed visualising the entire chromosomes, nuclear domains as well as individual genes within the same cell. FISH techniques have also been used in studying the relative spatial positioning of genes and their environment as illustrated in Figure 1.15. Additional levels of information have been provided through the visualisation of individual chromosomes and genes in their native spatial state in real-time via live-cell imaging, owing to the technological advances of microscopy (Cremer and Cremer, 2001).

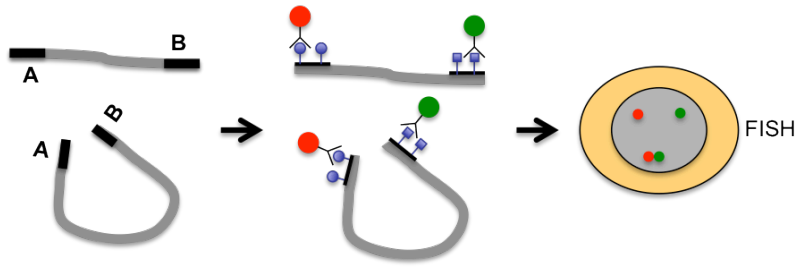


Figure 1.15: Schematics map of how FISH technique is used to assess chromosome organisation (Sofueva and Hadjur, 2012). Specific DNA/RNA sequences in the nucleus are detected by hybridisation of a sequence-specific probe labelled with digoxigenin (blue circles) or biotin (blue squares) nucleotides. The labelled probes are then specifically recognized by a fluorescently tagged anti-DIG or avidin antibody. FISH has many applications and for simplicity here we have shown how the localization between two chromatin regions can be correlated to the distance between the two probes.

FISH techniques are indispensable for identifying the spatial co-localisation of two specific regions in a single cell. However, they are limited for screening for novel chromatin interactions genome-wide for two reasons. First, they have limited spatial resolution, which would only be able to resolve interacting partners located at least 100 kb or even Mb in distance (Gilbert et al., 2004; Morey et al., 2007). Second, they are low-throughput approaches which can only assess very few chromosomes or loci at one time in few hundreds of cells.

1.3 Haematopoiesis

Haematopoiesis, the process of blood formation, has long been served as a classic model for studying sophisticated developmental processes in mammalian biology. This easy accessible system provides an ideally way to further understand the mechanisms of transcriptional regulation *in vivo*.

Haematopoiesis is a hierarchical process that haematopoietic stem cells (HSCs) give rise to multi- and bi-potent progenitor cells, which subsequently differentiate into mature and functionally distinct lineages of blood cells. In mammalian system, haematopoiesis occurs in two consecutive phases, which are primitive haematopoiesis in early embryonic development and definitive haematopoiesis in late embryonic development and in adults. Numerous tissues including yolk sac, para-aortic-spanchnopleura (PAS), aorta-gonad-mesonephros (AGM), liver, spleen and thymus have been demonstrated to serve as reservoirs of haematopoietic cells and/or sites of haematopoietic differentiation during different stages of development and differentiation (Figure 1.16).

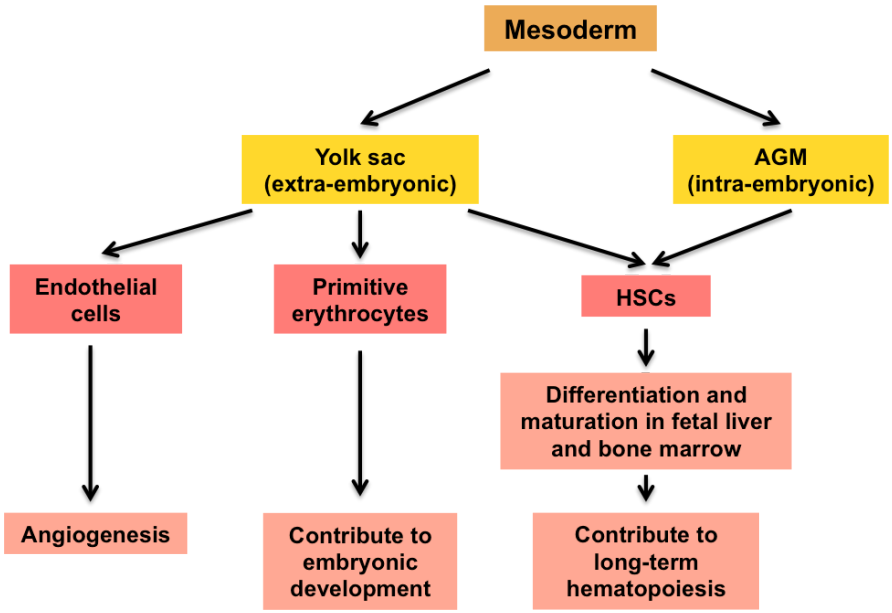


Figure 1.16: A schematic diagram of the development of the endothelial lineage and primitive and definitive haematopoiesis from their embryonic origins. The extra-embryonic yolk sac from the mesoderm gives rise to endothelial cells, primitive erythrocytes and haematopoietic stem cells (HSCs). The intra-embryonic aorta-gonad-mesonephros (AGM) also gives rise to HSCs during later embryonic development. Endothelials cells are implicated in vasculogenesis and angiogenesis. Primitive erythrocytes play a vital role for supporting embryonic development. HSCs migrate to the fetal liver or bone marrow for differentiation and maturation of various blood lineages and contribute to the maintenance of long-term haematopoiesis.

The first phase of primitive haematopoiesis takes place in the yolk sac around embryonic day 7 (E7) in mice or the 2nd to 3rd week in human gestation. In this stage, the undifferentiated mesodermal cells form extra-embryonic blood islands – where endothelial cells differentiated around the edges of the mesoderm for the formation of blood vessels, while primitive erythrocytes form in the interior regions of the blood islands. That is to say, both endothelial and haematopoietic lineages are derived from the same origin, which is evident by the existence of haemangioblast, a bi-potential common precursor for both lineages (Choi et al., 1998). Primitive haematopoiesis only occurs at very early stages of embryonic development until around E13, after then the yolk sac begins to degenerate. It is a transient but robust process that generates large amount of blood cells, including large, nucleated erythroblast as well as megakaryocytes and macrophages, for the growth and development of the early embryo (Figure 1.16). Differ to primitive haematopoiesis being mainly erythropoietic, definitive haematopoiesis gives rise to all haematopoietic lineages which occurs both in the extra-embryonic yolk sac and the intra-embryonic, PAS and later AGM (Figure 1.16). Thus, definitive progenitors are derived from a population of mesodermal cells with a fetal-adult fate instead of a purely primitive fate, suggesting that there is a temporal overlap between primitive and definitive haematopoiesis as they share a common precursor

(Kennedy et al., 1997). However, those definitive progenitors do not mature in the yolk sac, instead they migrate to other tissues for maturation. Definitive haematopoiesis is mainly derived from HSCs, a cell population which is required for haematopoietic development through the entire life span (Dzierzak et al., 1998). HSCs are characterised by their capability of self-renew and expression of cell surface markers like CD34 and c-kit. Differentiation of HSCs occurs in intra-embryonic tissues such as the fetal liver or bone marrow for maturation of haematopoietic cells (Figure 1.16). For instance, the enucleated erythrocytes produce adult globin and myeloid cells, which become mature and appear in the circulation system around E12. Meanwhile, the fetal thymus appears to be the sites for T-lymphoid development. The fetal spleen is the main site of haematopoiesis during late stage of embryogenesis before birth. Afterwards, the bone marrow acts the major location of haematopoiesis through the entire life span of the organism (Godin and Cumano, 2002; Kumaravelu et al., 2002).

1.3.1 HSC self-renewal and differentiation

All blood lineages are derived from the HSCs which are the starting point of haematopoietic differentiation. The HSCs are capable of self-renewing as well as differentiating into multiple blood cell types (shown in Figure 1.17). This differentiation process involves multiple steps which have been extensively reviewed (Ceredig et al., 2009; Seita and Weissman, 2010; Yoshida et al., 2010). According to one of the current accepted models, differentiation of short-term HSCs produces two groups of multipotent progenitors (MPPs), which are common myeloid progenitor (CMP) and lymphomyeloid bipotent progenitor (LMPP) (Figure 1.17). The CMP produces the megakaryocyte-erythrocyte progenitor, which gives rise to the erythroid and megakaryocytic lineages. The LMPP diversifies into (i) the common lymphoid progenitor, which produces B and natural killer cells; (ii) the early thymic progenitor generates T-cell lineages; (iii) the granulocyte-macrophage progenitor which gives rise to myeloid lineages (Figure 1.17). The first requisite of a haematopoietic stem cell being conducive to the process of differentiation is considered entering a state of “priming” - the concept that the chromatin allows expression of genes involved to multiple cell type specifications; followed by a progressive repression of genes linked to undesired lineages. These two independent processes reveal a major mechanism of priming gene expressions by pluripotent stem cells. The genes involved in development of stem cells are

marked with both active (e.g. H3K4me3) and repressive (e.g. H3K27me3) histone modifications, which are known as bivalent chromatin domains, to remain them poised for transcription (Azuara et al., 2006; Bernstein et al., 2006). Two levels (molecular/transcriptional and cellular) of priming can be contemplated in stem cell differentiation. Transcriptional priming is achieved by creating a transcription-conducive chromatin landscape that is ready to respond to stimuli forms, while the resultant cross-lineage transcriptome modulates the cells as a whole to react to prospective lineage commitment known as cellular priming.

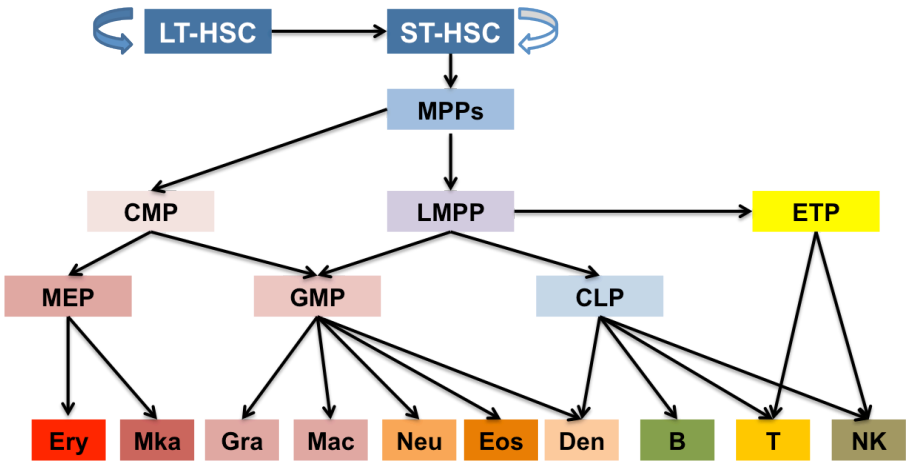


Figure 1.17: The schematic of haematopoietic stem cell (HSC) self-renewal and differentiation (Panigrahi and Pati, 2012): long-term and short-term haematopoietic stem cell (LT-HSC, ST-HSC); multipotent progenitors (MPPs); common myeloid progenitor (CMP), lymphomyeloid bipotent progenitor (LMPP); early T-lineage progenitor (ETP); megakaryocyte-erythrocyte progenitor (MEP); granulocyte-monocyte progenitor (GMP); common lymphoid progenitor (CLP); erythrocyte (Ery); megakaryocyte (Mka); granulocyte (Gra); macrophage (Mac); neutrophil (Neu); eosinophil (Eos); dendritic cell (Den); B-cells (B); T-cells (T); natural killer cells (NK).

The haematopoietic cell fates are regulated via a complex network of transcription factors, where the expression levels and activities of these factors selectively elevate or repress gene expression to determine lineage-specific commitments (Table 1.3). For instance, GATA1 (GATA-binding protein) and FOG1 (friend of GATA1) act positively for the megakaryocyte-erythrocyte-basophil and mast-cell-eosinophil development (Graf, 2002). FOG1 expression in mouse ES cells diverts haematopoietic progenitors from a programme of mast cell development to be neutrophils and redirects the development of mast cell progenitors to the erythroid, megakaryocytic and granulocytic lineages (Cantor et al., 2008; Sugiyama et al., 2008). In addition, the megakaryocytic and erythroid development is also regulated by transcription factors including MLLT3, MYB, EKLF (erythroid Kruppel-like factor) and EDAG (erythroid differentiation-associated gene) (Mukai et al., 2006; Pina et al., 2008; Yang et al., 2006). High-level of MYB and EKLF expression is

capable of driving erythropoiesis while blocking megakaryopoiesis (Frontelo et al., 2007). Moreover, EDAG is also expressed in HSCs and early progenitors, and the low-level expression of EDAG in these cell populations is required for lymphoid cell commitment (Yang et al., 2007). Increased EDAG expression directs myeloid cell development while prevents T- and B-cell development (Li et al., 2007b). The activity of myelomonocytic factor PU.1 directs HSCs or progenitors towards a myeloid commitment instead of an erythroid fate (Galloway et al., 2005). PU.1 also directs progenitors away from NK-cell and T-cell fates (Kamath et al., 2008). It acts in a concentration-dependent manner to repress erythroid, NK- and T-cell specific genes while activate myeloid associated genes. Moreover, C/EBPα diverts cells away from erythroid commitment as well as represses lymphopoiesis. Expression of C/EBPα is also required for development of myelomonocytic cells (Wang et al., 2006) as well as commitment of monocytes from neutrophil-monocyte progenitors (Mukai et al., 2006). Enforced C/EBPα expression leads to myeloid differentiation in CLPs, pro- T and B cells as well as prevents B-cell development via supressing PAX5 (Friedman, 2007). In contrast, EBF1 is critical for multi-potent progenitors to commit to a B-cell lineage and to supress progenitors from adjacent commitments such as generating T-cell or myeloid cells (Pongubala et al., 2008). In summary, the haematopoietic cell fates are tightly regulated by transcription factors, which either promoting or suppressing lineage-specific options to determining the fate boundaries and driving the final development of adjacent cell fates.

Table 1.3: Transcription factors in directing or inhabiting the development during haematopoiesis (Ceredig et al., 2009).

Lineages	Activation	Repression
Basophil/mast cell	PU.1, MYB, GATA1	FOG1
Eosinophil	C/EBPα^{p}, PU.1, MYB, GATA1	FOG1
Neutrophil	C/EBPα + AP1, PU.1, MYB, EDAG	E-proteins, EBF1, GATA1, GATA2
Monocyte	C/EBPα + AP1, PU.1, MYB, EDAG	E-proteins, PAX5, EBF1, GATA1, GATA2
Dendritic cell	C/EBPα , PU.1, MYB	E-proteins, PAX5
B cell	PAX5, EBF1, PU.1, MYB	Notch signaling, C/EBPα, EDAG
T cell	Notch signaling, E proteins, MYB	PAX5, EBF1, C/EBPα , PU.1, EDAG
Natural killer cell	MYB	Notch signaling, E protein, PAX5, C/EBPα, PU.1
Megakaryocyte	MLLT3, GATA1, GATA2, FOG1, EDAG,	MYB, EKLF, C/EBPα,
Erythrocyte	MLLT3, GATA1, GATA2, FOG1, EDAG, MYB, EKLF	C/EBPα, PU.1

1.4 The SCL/TAL1 gene

In eukaryotes, the generation of various lineages relies on establishments and maintenances of specific programmes of gene expression. Transcription factors are in the centre of this process, acting in a combinatorial manner of both activating and repressing specific sets of target genes to retain the environment for appropriate cellular functions. In particular, haematopoietic system is considered as a powerful model for characterizing the mechanisms of transcription factors in controlling differentiation and lineage commitment as well as leukemogenesis (Orkin, 2000; Tenen et al., 1997). A perfect paradigm of this is a transcription factor of the basic Helix-Loop-Helix (bHLH) family called TAL1, which plays as a fundamental regulator at several levels of the haematopoietic hierarchy. The TAL1 (SCL) gene has been first identified in a patient involved in a chromosomal translocation between chromosomes 1 and 14 t (1; 4) (p32; q11) with T-cell acute lymphoblastic leukaemia (T-ALL) (Begley et al., 1989a). It has also been independently reported by other researchers, which named as TAL1 and TCL5 (Chen et al., 1990; Finger et al., 1989). Subsequent to these discoveries, its official gene name is TAL1 – and is referred to as such for this thesis.

1.4.1 The TAL1 expression

1.4.1.1 TAL1 gene structure

The human TAL1 gene is located on chromosomal 1 (band 1p33) whereas the murine Tal1 orthologue has been mapped to the region of chromosome 4 (4qD1) (Begley et al., 1991), which is syntenic with human chromosome 1p. The murine Tal1 gene shares 88% of nucleotide homology and 94% protein homology with human TAL1 gene and the bHLH domains are identical between two species (Begley et al., 1991). The human TAL1 locus is 16 kb in size, which is composed of eight exons and the first five (Ia-III) are non-coding exons as illustrated in Figure 1.18 (Aplan et al., 1990a). The mouse Tal1 locus consists of seven exons spanning approximately 20 kb of genomic region (Begley et al., 1994). The structural organisation is highly conserved between mouse and human as shown in Figure 1.18 with the exception of an additional exon (exon IIa) at the human locus which is not present in mouse. Two alternate promoters (P^{1a} and P^{1b}) are located at the 5' end of the TAL1 gene (Figure 1.18). A third promoter (P^{Exon4})

located within exon 4 has been found being active in leukaemic T-cell lines as well as primary T-ALL cells (Bernard et al., 1992). A long 3' UTR (un-transcribed region) is a common feature for TAL1 genes in both human and mouse (Figure 1.18).

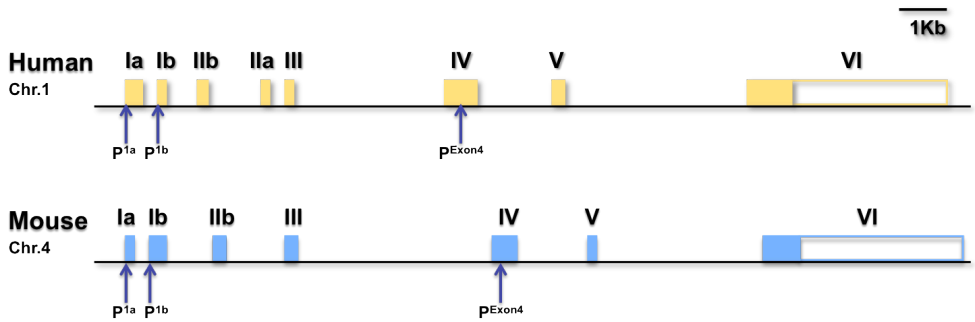


Figure 1.18: Schematic of the human and mouse TAL1 loci. The human TAL1 locus is shown at the top (light yellow) and murine Tal1 at the bottom (light blue). Each box represents an exon. The 3' UTRs are shown by the unshaded boxes. The human TAL1 gene contains an additional exon as compared to mouse (IIa). Promoters of the TAL1 gene have been mapped (shown by arrows) and are annotated as P^{1a}, P^{1b} and P^{Exon4}.

1.4.1.2 mRNA expression of TAL1

Expression of human TAL1 gene has been well characterised, and shows a complex pattern of mRNA splicing which produces a number of different RNA species varying from 3 kb, 4.8 kb to 5 kb in size (Aplan et al., 1990a). At least six alternative splicing events are generated from the 5' UTR, which reflect patterns of alternate 5' exon usage, and all but one RNA species converge on exon III (Aplan et al., 1990a; Bernard et al., 1991). Additionally, a very similar splicing pattern of Tal1 has also been observed in murine cells (Begley et al., 1994).

1.4.1.3 TAL1 expression in haematopoietic system

Originally, expression of TAL1 has been found in fetal liver, regenerative bone marrow, early myeloid cell lines and leukaemic T-cell lines using northern blot analyses (Begley et al., 1989b). Further studies have shown that TAL1 is also widely expressed in human and murine erythroid, mast and megakaryocytic cell lines (Aplan et al., 1990a; Begley et al., 1989b; Green et al., 1991). TAL1 expression in those lineages has been confirmed in human primary cells using in situ hybridization and RT-PCR (Mouthon et al., 1993). Increased expression of TAL1 has been observed during differentiation of embryonic stem cell towards to embryonic bodies and haematopoietic progenitors by in vitro differentiation analysis (Elefanty et al., 1997). Moreover, TAL1 expression has also been found in

the aorta-associated $CD34^+$ high proliferative potential haematopoietic cells, which are proposed to be HSCs present later in fetal liver and bone marrow (Labastie et al., 1998).

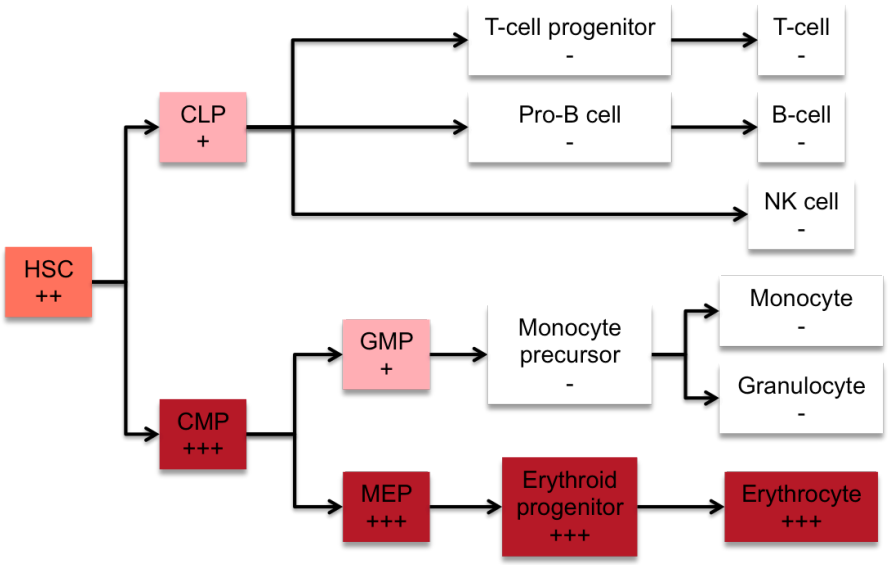


Figure 1.19: Expression of TAL1 during human haematopoietic differentiation (Zhang et al., 2005). Increased levels of TAL1 expression are shown by the colour intensities in the boxes and symbols (light pink, -, +, ++, +++, dark pink). TAL1 expression levels are highest in pluripotential (HSC, haematopoietic stem cell) and early committed erythroid precursors (CMP, common myeloid progenitor; MEP, megakaryocyte/erythrocyte progenitor) but sequentially extinguish with terminal maturation. In contrast, TAL1 expression is lower in cells committed to the lymphoid (CLP, common lymphoid progenitor) and myeloid lineages (GMP, granulocyte/monocyte progenitor) and diminishes further with maturation TAL1 protein expression.

In summary, TAL1 is generally expressed in multipotent progenitors and HSCs, in erythroid, megakaryocytic and mast cell lineages (Cross et al., 1994; Green et al., 1992; Green et al., 1991; Mouthon et al., 1993; Pulford et al., 1995; Visvader et al., 1991). As illustrated in Figure 1.19, expression of TAL1 has been observed in all haematopoietic cells with an erythroid potential and present, but is down-regulated in common lymphoid progenitors (CLPs) and granulocyte/monocyte progenitors (GMPs) (Zhang et al., 2005). Although TAL1 transcripts have been detected in some pre-B and macrophage lines (Green et al., 1992; Visvader and Begley, 1991), expression of TAL1 is absent in primary lymphocytes, macrophage lines as well as B-cell and T-cell lines, except the aberrant expression of TAL1 in T-ALL cells (Begley et al., 1989a).

1.4.1.4 TAL1 expression in extra-haematopoietic tissues

Expression of TAL1 is not only observed in haematopoietic cells but is also detected in a number of other tissues and cells. First, it has been found that TAL1

is expressed in nervous system, such as in murine brain, in post-mitotic neurons of the metencephalon and the mesencephalic roof as well as in human medulloblastoma cell lines (Begley et al., 1989b; Green and Begley, 1992; Kallianpur et al., 1994). Moreover, it has been demonstrated that TAL1 is expressed in thalamus, midbrain and hindbrain in adult and the developing embryonic central nervous system using a knock-in mouse (van Eekelen et al., 2003). Second, TAL1 is also expressed in skeletal and muscular systems, such as developing skeletal system, cells of developing cartilage and bone, melanocytes, vascular and visceral smooth muscle cells of the aorta and bladder (Kallianpur et al., 1994) as well as uterine smooth muscle (Pulford et al., 1995). Third, expression of TAL1 has been detected in endothelial cells, especially in blood vessels in the spleen (Hwang et al., 1993). In addition, expression of TAL1 has also been found in endothelial progenitors in blood islands, in endothelial cells and angioblasts within a number of organs at times coincident with their vascularization (Kallianpur et al., 1994).

1.4.1.5 TAL1 expression in other species

The expression pattern of TAL1 is also highly conserved in lower vertebrates. Its expression has been found in avian endothelial progenitors and committed cells as well as angioblasts (Drake et al., 1997). In addition, the neural pattern of TAL1 expression found at the mesencephalon/diencephalons boundary in developing chicken embryos shared high similarity with its mouse equivalent (Sinclair et al., 1999). Moreover, TAL1 expression has also been detected in *Xenopus* ventral mesoderm (sites of haematopoietic tissue development) and ventral blood island and dorso-lateral plate (sites of primitive and definitive haematopoiesis) by *in situ* studies (Mead et al., 1998; Turpen et al., 1997).

1.4.2 The TAL1 protein

The TAL1 gene encodes two isoforms of protein detected in haematopoietic lineages, which are a 42 kDa full length protein and a 22 kDa amino truncated form as results of differential mRNA splicing with translation initiation at an internal start site (Elwood et al., 1994; Goldfarb et al., 1992). Similar TAL1 protein products have also been observed in erythroleukaemia and T-ALL samples (Cheng et al., 1993b). Phosphorylation at the serine residue 122 (S122) induced by protein

kinase ERK1 has been observed in both TAL1 isoforms (Cheng et al., 1993a). The bHLH domain is retained in both full length and amino truncated proteins which is critical for DNA binding, protein-protein interaction and nuclear localisation (Goldfarb et al., 1992; Hsu et al., 1991; Hsu et al., 1994b).

1.4.2.1 TAL1 forms heterodimer with E2A proteins

The helix-loop-helix (HLH) region and the basic region are two important domains within the protein. The HLH region has the DNA-binding and dimerization motif, which is common to a large family of proteins found in species ranging from mammals to plants (Murre et al., 1989). The HLH domain is usually preceded by a highly basic motif, of 10-20 amino acids in length that determines the DNA-binding specificity. Interestingly, the HLH proteins are able to form either homo-dimeric or hetero-dimeric complexes with other family members to modulate their DNA binding activities.

As a member of the basic helix-loop-helix (bHLH) family (Begley et al., 1989b), TAL1 has been divided into two groups. Typically, Class I proteins including E2A, E2-2 and HEB, are broadly expressed and can form homo- and hetero-dimers with other HLH proteins. In contrast, Class II proteins, such as TAL1, are expressed in a tissue-specific manner and form heterodimers with Class I proteins. It has been shown that TAL1 polypeptides do not tend to form homo-dimers, but preferentially interact with the E2A proteins (E12 and E47) to form heterodimers in vitro binding assays (Hsu et al., 1991). These heterodimers bind to DNA in a sequence-specific manner, which recognize a consensus DNA sequence “AACAGATGGT” referred as the “E-box” motif (Hsu et al., 1994a). In addition, the TAL1/E2A dimers have also been found *in vivo*, which form the basis of an important transcriptional regulation mechanism as discussed in the following section.

1.4.2.2 TAL1-containing multifactorial complexes

TAL1-containing erythroid complex (TEC)

It has been shown that TAL1/E2A heterodimers do not bind to DNA alone, but also interact with other proteins including GATA1, LMO2 (LIM domain protein) and LDB1 to function as a large complex (Wadman et al., 1997), referred as “TEC” (TAL1-containing erythroid complex) in this thesis (as shown in Figure 1.18a).

Similar to TAL1, LMO2 is also critical for primitive and definitive hematopoiesis (Warren et al., 1994). LMO2 expression is limited in erythroid and myeloid lineages in committed cells and its aberrant expression in T-cells due to chromosomal translocation is also associated with T-ALL (Rabbitts, 1994; Yamada et al., 1998). LDB1 is a LIM domain binding protein and neither of them bind to DNA sequence, instead they function as bridging protein in linking DNA binding proteins between TAL1/E2A and GATA1 (Wadman et al., 1997). The TEC has been shown to bind to a bipartite DNA motif which consists of an E-box site situating ~9-12 bp upstream of a GATA site. Additionally, it has been found that the transactivity of this bipartite motif depended on the appropriate spacing as well as the orientation (Wadman et al., 1997). Furthermore, the fact that TAL1, GATA1 and LMO2 share overlapping expression domains and deficiency of each leads to failure of haematopoiesis also reinforce the idea that the regulation of genes crucial to haematopoiesis likely depends on the complex with these three key members (Barton et al., 1999). In addition, TAL1 is also associated with co-repressors including histone deacetylases, mSIN3a, BRG1, LSD1, ETO-2, GFI1b and the core-binding factor subunit CBFA2T3H during early erythropoiesis (Cai et al., 2009; Hu et al., 2009). Transcriptional co-activators (e.g. p300 and P/CAF) are also acquired by the TEC during erythroid maturation (Huang et al., 2000).

Accumulated evidence suggests that the TEC plays a critical function in transcription regulation of erythroid specific gens. For instance, GATA/E-box sites as well as the TEC have been identified in DNA elements involving in transcription regulation of p4.2 and EKLF (Anderson et al., 2000; Xu et al., 2003). Moreover, TAL1 also plays an important role in mediating long-range chromatin looping between the LCR and globin promoters, which are necessary for the developmental switch between β-globin and γ-globin as well as migration of β-globin locus to the Pol II transcription factories (Song et al., 2010). Recently, a common composite motif comprising a half E-box and GATA has been identified by ChIP-seq analyses in not only known target genes of TAL1 including TAL1 itself, EKLF, α-globin, β-globin, EPB4.2 and Glycophorin A, but also numbers of potential novel targets involved in a wide-range of biological processes (Kassouf et al., 2010). Further discussions relating to target genes of the TEC will be presented in introduction section of Chapter 6.

Auto-regulation of the TEC

Previous studies on exploring the roles of the TAL1-containing erythroid complex (TEC) in transcriptional regulation have demonstrated that the entire complex is also involved in directly regulating expression of genes of its own members (H.L.Jim's PhD thesis, University of Cambridge, 2008). For instance, it has been demonstrated that GATA1 binds to the promoter of LMO2 based on ChIP-chip and gene expression analysis in the GATA1 siRNA knockdown. In addition, it has been shown that TAL1 is directly targeted by the entire TAL1-containing erythroid complex via its binding over the GATA/E-box motifs at the TAL1 promoter and the erythroid enhancer in erythroid cell lines, suggesting that transcription of TAL1 is under auto-regulation of itself. Taken together, transcription regulation by individual members of the complex as well as by the complex as a whole provides two layers of controls to ensure that the expression level of the TEC members are tightly regulated during erythroid development. It further highlights the sophisticated regulatory network in controlling expression of the TEC.

1.4.2.3 Other TAL1 related multi-protein complexes

In addition to the TEC (Figure 1.20a), TAL1 has also been shown to involve in formation of a number of multi-protein complexes as illustrated in Figure 1.18. For instance, a complex of similar composition of TEC with the exception of GATA1 has been identified in a leukemic cell line derived from LMO2 transgenic mice, which is specifically recognize an E-box/E-box sequence motif (Grutz et al., 1998) (Figure 1.20b). The RALDH2 (Retinaldehyde dehydrogenase 2) gene can be activated by TAL1 and LMO1/2 through their recruitment to a cryptic intronic promoter via DNA-bound GATA3 transcription factor in T-ALL cells (Ono et al., 1998). Instead of being directly recruited to the E-box motif, TAL1 and LMO1/2 act as cofactors for GATA3 by forming a complex in vivo to activate the transcription of RALDH2 (Figure 1.20c). In contrast to the role of TAL1 in aberrant gene activation, a number of studies have shown that TAL1 can also exert its oncogenic functions via inhibiting the normal functions of the E-proteins E2A and HEB, which are known to be crucial regulators of lymphoid cell differentiation (Engel and Murre, 2001; Herblot et al., 2002; Herblot et al., 2000). A schematic model is showed in Figure 1.20d, and detailed descriptions about the repression mechanism via TAL1 will be further presented in section 1.4.3.3. In addition, a

multifactorial complex has been identified at the c-kit promoter containing Sp1 protein and all members of the TEC, which regulates transcriptional activation of c-kit (Lecuyer et al., 2002). Moreover, it has been demonstrated that the pRB protein can also associate with this complex in erythroid cells, which result in inhibition of c-kit promoter activity (Vitelli et al., 2000) (Figure 1.20e). Altogether, these findings highlight the complexity of the TAL1-containg multifactorial complexes and their critical roles in transcriptional regulation.

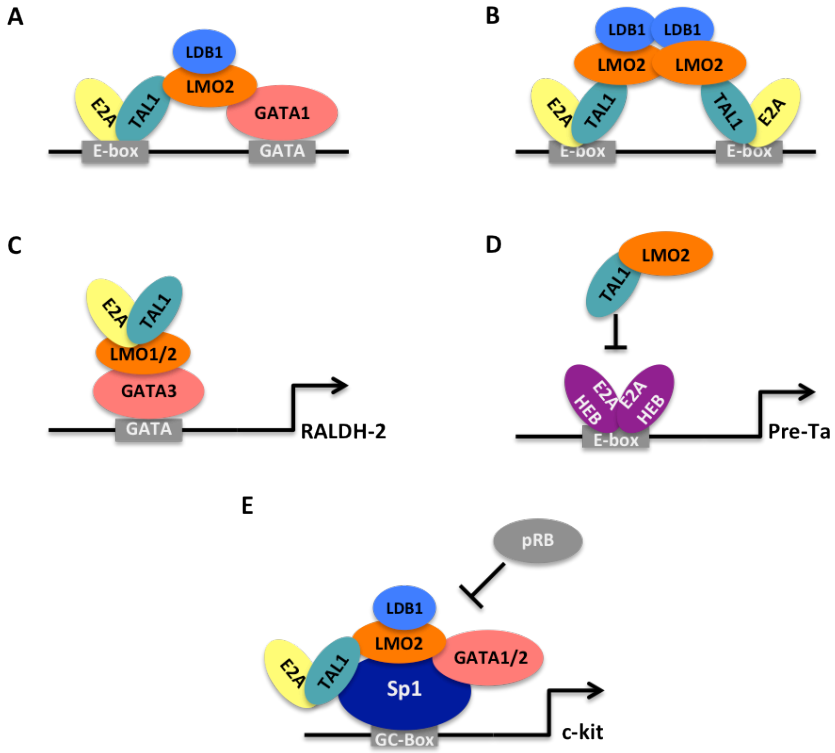


Figure 1.20: Schematic representation of the TAL1-containg multifactorial complexes. (A): the TEC associates with a composite GATA/E-box motif first identified in MEL cells (Wadman et al., 1997). (B): a multifactorial complex binds to bipartite E-box/E-box motif identified in leukemic T-cells (Grutz et al., 1998). (C): TAL1/E2A and LMO1/2 have been shown to activate the transcription of the RALDH-2 gene via recruitment by promoter-bound GATA3 protein in T-ALL cell lines (Ono et al., 1998). (D): TAL1 can also repress transcription of genes (such as Pre-Tα) regulated by E2A/HEB homo/hetero-dimers during lymphoid cell differentiation (Tremblay et al., 2003). (E): the TEC activates the c-kit gene via association with promoter bound Sp1 protein (Lecuyer et al., 2002), and the Rb protein can repress the c-kit promoter by interacting with the TAL1-containing complex during erythroid cell differentiation (Vitelli et al., 2000).

1.4.3 Functions of TAL1

1.4.3.1 Role of TAL1 in cell-proliferation and differentiation

Accumulated evidences suggest that TAL1 plays a key role in regulating cell proliferation, self-renewal and terminal differentiation (Aplan et al., 1992b; Chiba et al., 1993; Green et al., 1991). The first example of illustrating normal function of TAL1 come from a study of examining its mRNA expression in human and mouse

haematopoietic cell lines and tissues, which observe a high-level TAL1 expression in all erythroid cell lines as well as in haematopoietic cell populations enriched in erythroid precursors (Green et al., 1991). Similarly, it has been demonstrated that enforced TAL1 expression in human and mouse erythroleukaemia cell lines (K562 and MEL) promotes the erythroid differentiation without requiring additional inducer (Aplan et al., 1992b). Moreover, it has been shown that the unregulated level of TAL1 expression is associated with cytokine induced erythroid differentiation of progenitor cells, whereas TAL1 level declined during granulocytic differentiation (Cross et al., 1994). In addition, it has been found that enforced expression of TAL1 is related to erythroid and megakaryocytic differentiation in human and mouse progenitor cells (Condorelli et al., 1997; Elwood et al., 1998; Valtieri et al., 1998), whereas progenitors were able to retain undifferentiated state by expressing TAL1 constitutively (Hoang et al., 1996; Tanigawa et al., 1993). Furthermore, overexpression of TAL1 can also lead to suppression of macrophage differentiation in AML (acute myeloid leukemia) cells (Tanigawa et al., 1995) as well as granulocytic development in normal human haematopoietic progenitors (Valtieri et al., 1998).

1.4.3.2 Role of TAL1 in haematopoiesis and vasculogenesis

The TAL1 plays a critical role in haematopoiesis and vasculogenesis. It has been shown that TAL1 is not essential for maintenance and self-renewal of HSCs (Curtis et al., 2004; Mikkola et al., 2003), but is required during the differentiation from ESCs to HSCs (Endoh et al., 2002) as well as development of all haemotopoietic lineages (Robb et al., 1995; Shivdasani et al., 1995).

Haematopoietic development

The TAL1 function in normal haematopoiesis had not been fully examined until the generation of TAL1 knockout mice. It has been shown that $TAL1^{-/-}$ mice die between E9.5-10.5 due to the deficiency of blood formation (Robb et al., 1995; Shivdasani et al., 1995). In contrast, organogenesis including neural development occurred normally at this stage in $TAL1^{-/-}$ embryos, regardless of the growth of embryos was developmentally challenged, suggesting a specific function of TAL1 in haematopoiesis. In addition, the absence of haematopoietic progenitors has been confirmed by yolk sac culture, indicating that $TAL1^{-/-}$ embryonic stem cells are deficient in generating any haematopoietic lineages (Porcher et al., 1996;

Robb et al., 1996). Furthermore, it has been illustrated that TAL1$^{-/-}$ embryonic body completely abolished the expression of haematopoietic genes, demonstrating the crucial and non-redundant role of TAL1 in the formation of haematopoietic stem cells (Elefanty et al., 1997). Function of TAL1 in haematopoiesis is conserved through the vertebrate evolution such as *Xenopus* and *Danio*. A number of studies have shown that TAL1 is expressed in ventral mesoderm of early frog embryos prior to the formation of the ventral blood islands (Mead et al., 1998; Turpen et al., 1997). Additionally, TAL1 knockdown studies in Zebrafish have also demonstrated the importance of TAL1 for haematopoietic development and angiogenesis, which provides extra evidences for functional conservation of TAL1 (Dooley et al., 2005).

TAL1 is critical for the production of TIE2^{+}c-KIT^{+}CD41^{-} haemogenic intermediates that give rise CD41^{+} primitive and definitive haematopoiesis, while is dispensable for the formation of the haemangioblast (Lancrin et al., 2009). In adult HSCs, TAL1 is required for LT-HSCs in the Kit^{+}Sca1^{+}Lin^{-}CD150^{+}CD48^{-} subpopulation (Lacombe et al., 2010). Although a modest defect in the activity of ST-HSC was observed in TAL1 deletion, it did not cause stem cell failure in the adult HSCs (Curtis et al., 2004). It is speculated that the reduced expression of CDKN1a or ID1 may account for the modulation of ST-HSC function by TAL1, as both CDKN1a and ID1 are target gens of TAL1 and reduced expression impairs HSC function (Lacombe et al., 2010). In mature cell types, TAL1 expression is only found in the erythroid, megakaryocytic and mast lineages. It has been shown that TAL1 expression is not required for long-term maintenance of erythrocytes or platelets, regardless a moderate anaemia and thrombocytopenia is observed in TAL1 cKO (conditional knockout) mice (Hall et al., 2005). In addition, TAL1 is known to be very important for growth of BFU-e (blast forming units-erythroid) with sparing of CFU-e (colony forming units-erythroid). It has also been observed a similar erythroid phenotype in mice with a germline TAL1 DNA-binding mutation, implying that the important functions of TAL1 in erythropoiesis are mediated via DNA binding (Kassouf et al., 2008) (see section 1.4.2 for further details regards to TAL1 DNA-binding and its functions in transcriptional regulation of target genes). It has also been shown the importance of TAL1 in megakaryocyte progenitor growth and platelet shedding using TAL1 cKO mice (McCormack et al., 2006). TAL1 may also play a vital role in lineage commitment, as loss of TAL1 in megakaryocyte-

erythrocyte progenitors (MEP) results in aberrant mast cell differentiation probably due to the elevated level of GATA2 expression (Salmon et al., 2007).

Endothelial development

In addition to its role in regulating haematopoietic development, the importance of TAL1 has also been shown in endothelial development and angiogenesis. It has been found that TAL1 is required for the formation of yolk sac blood vessels using transgenic rescue of $TAL1^{-/-}$ embryos (Visvader et al., 1998). In addition, TAL1 has been shown to be crucial for the generation of blast colonies from blast colony forming cells (BL-CFCs), which is an in vitro equivalent of the haemangioblast (Chung et al., 2002; Robertson et al., 2000). In contrast, studies have shown that $TAL1^{-/-}$ cells initiate colony growth but fail to generate endothelial and haematopoietic lineages (D'Souza et al., 2005). These studies suggest that TAL1 is fundamental for commitment of haematopoietic and endothelial lineages from haemangioblast.

1.4.3.3 Role of TAL1 in T-cell leukaemia

Since TAL1 has been firstly discovered in T-ALL, rearrangements of this gene have been reported up to 30% of patients with T-cell leukaemia (Aplan et al., 1992c; Bernard et al., 1991; Brown et al., 1990). The translocation involving the T-cell receptor delta chain (TCR-δ) locus on chromosome 14 is one of the most frequent reported rearrangements (Begley et al., 1989a; Bernard et al., 1990; Bernard et al., 1991). These translocations mainly cause disruptions at the 5' regulatory regions of the TAL1 gene and leave the coding region unaffected, which allows generating full length TAL1 product in T-cell blasts. One additional translocation breakpoint has also been identified downstream of the TAL1 coding sequence (Begley et al., 1989a; Finger et al., 1989), which results in a truncated TAL1 product due to abnormal transcription initiation at the promoter in exon 4 (Bernard et al., 1992). Another well-characterized rearrangement at the TAL1 locus is caused by an interstitial deletion of approx. 90 kb of the entire STIL gene (located right upstream of TAL1) plus the 5' UTR (un-translated region) of TAL1 (Aplan et al., 1990b). The deletion leads to a result that the TAL1 coding sequence is situated right under the regulation of STIL promoter and drives the expression of a TAL1/STIL fusion product (Breit et al., 1993; Brown et al., 1990; Jonsson et al., 1991). As the STIL gene is transcriptionally active in T-cells, this rearrangement

results in an aberrant activation of TAL1 expression in T-cells, which consequently causes T-cell leukaemia (Kwong et al., 1995). It has been reported that ~25% of T-ALL patients carry this precise deletion designated as Tal^{d}, which was not able to be detected by standard cytogenetic analysis (Aplan et al., 1992a; Baer, 1993).

The ability of TAL1 in enhancing the tumourigenticity of a v-abl transformed T-cell line provided the first line of evidence demonstrating TAL1 as an oncogene (Elwood et al., 1993). Additionally, accumulated evidences suggest that TAL1 can contribute to carcinogenesis via multiple mechanisms, which include an enhancement in cell proliferation, cell cycling and self-renewal potential along with a reduction in cell death (Begley and Green, 1999). Generally, expression of TAL1 is limited to early T-cell development and it has no detrimental effect on T-cell development (Capron et al., 2006; Larmonie et al., 2011). However, aberrant expression of TAL1 is frequently observed in T-ALL due to chromosomal rearrangement, micro-deletions or abnormal expression of other factors within the regulatory network such as LMO2 (McCormack et al., 2010; Nagel et al., 2010). It has been revealed that interactions between TAL1 and LIM domain proteins (LMO1/2) are closely associated with T-cell malignancies in transgenic mice (Aplan et al., 1997; Larson et al., 1996). Moreover, stable heterodimers formed between TAL1 and E-proteins (E2A or HEB) have been detected in human and mouse leukaemia cells (Hsu et al., 1994b; O'Neil et al., 2001), postulating that TAL1 induces leukaemia via interfering with E2A and HEB. Further studies have demonstrated that the formation of E2A/HEB heterodimers can be sequestered by competitive interactions of TAL1, which lead to repression of E-protein target genes and subsequent acceleration of the disease (O'Neil et al., 2004). Further characterization of TAL1 transcriptional regulation would allow unravelling the mechanisms underlying normal haematopoiesis as well as leukaemogenesis.

1.4.4 Transcriptional regulation of TAL1

1.4.4.1 *Cis*-acting regulatory elements at the TAL1 locus

Due to the importance of TAL1 in haematopoiesis and T-cell leukemia, a large numbers of studies have been conducted and virtually all known regulatory sequences have been functionally tested in mouse models. As illustrated in Figure

1.21, the known TAL1 *cis*-acting regulatory elements have been previously identified using reporter assays as well as ChIP analyses, which are named based on their positions in kilobases (kb) with respect to the start of the TAL1 exon 1a.

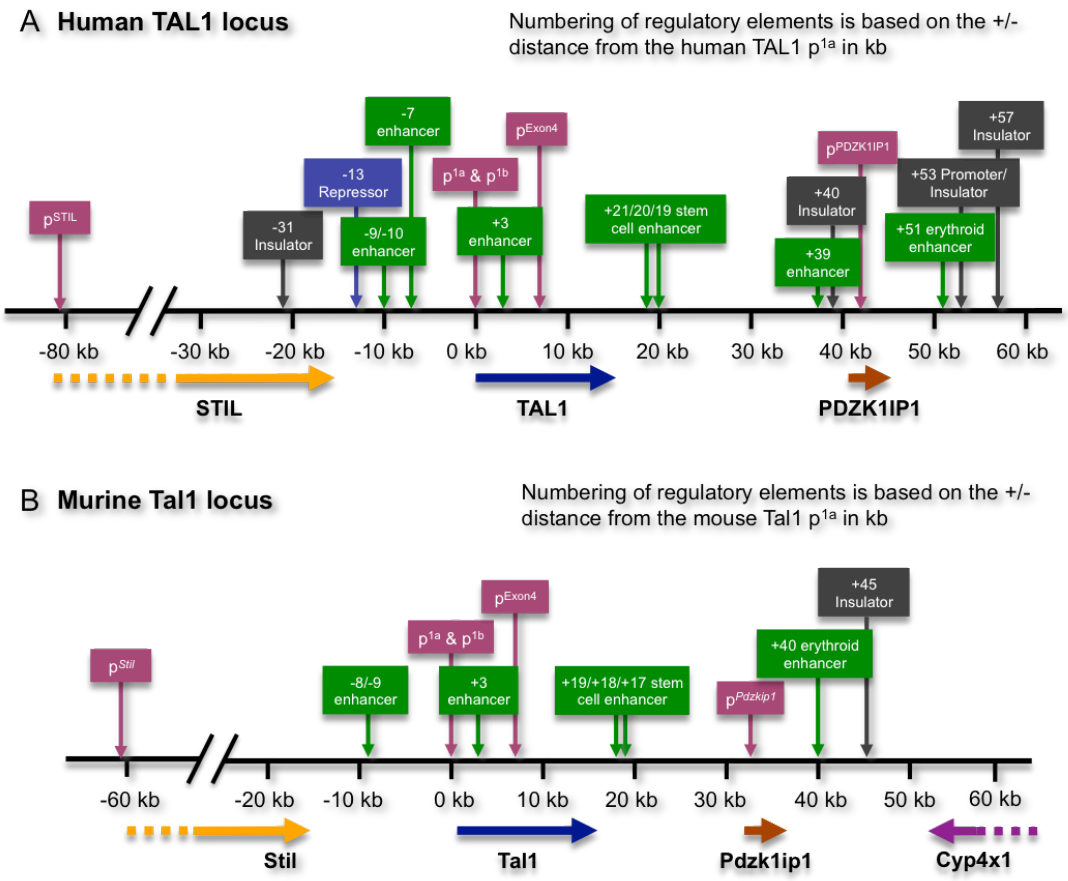


Figure 1.21: A schematic representation of the known regulatory regions at the human and mouse TAL1 loci. Panel A: human and panel B: mouse. The description of these regulatory elements has been provided in the text of this chapter. The genomic region shown in the figure encompasses the STIL (upstream), TAL1 and PDZK1IP1 (downstream) genes in human and mouse. The colour-coded arrows with the gene names show the direction of transcription of these genes. Each of the regulatory regions is colour-coded in human and mouse to show their orthologous relationship. Promoters are shown in pink, enhancers are shown in green, blue and dark grey represent the repressor and insulators, respectively.

As shown in Figure 1.21, TAL1 has two promoters located 5′ to the gene (promoter 1a and 1b), and an additional promoter within exon 4 (promoter exon4 which also known as +7 relative to the promoter 1a) (Courtes et al., 2000). The promoters of neighbouring genes including PDZK1IP1 (previously known as MAP17) and STIL (previously known as SIL) reside within highly conserved regions of DNA immediately 5' to the transcriptional start sties (TSS) (Aplan et al., 1990a). Enhancers of TAL1 are also highly conserved in vertebrates, particularly between human and mouse (Chapman et al., 2004; Gottgens et al., 2002a; Gottgens et al., 2000; Gottgens et al., 2001). In brief, two enhancers at +19/+18 and -4 were found

to direct TAL1 expression to the vast majority of HSCs along with haematopoietic progenitors and endothelium (Gottgens et al., 2004; Sanchez et al., 1999). The *cis*-acting elements at +1 and +3 have been shown to drive the TAL1 expression to the developing spinal cord in transgenic mice, whereas the +23 enhancer directs TAL1 expression to the central nervous system (CNS) (Gottgens et al., 2000). In addition, the -9 element functions as an enhancer in haematopoietic cell lines (Gottgens et al., 1997) and the enhancer at +51 in human and +40 in mouse drives erythroid-specific TAL1 expression in the primitive erythroid lineage (Delabesse et al., 2005; Dhami et al., 2010; Ogilvy et al., 2007). A repressor of TAL1 transcription was identified at the mouse +12 element, which is homologous to the human +14 element (Courtes et al., 2000). A number of CTCF insulators have been recently identified across the TAL locus by ChIP analysis and enhancer blocking assays including +57, +53, +40 and -31 elements in human (Dhami et al., 2010) and +45 elements in mouse (Follows et al., 2012). Further descriptions about these *cis*-acting elements and their roles in TAL1 transcriptional regulation are detailed in the following sections.

1.4.4.2 TAL1 promoters

Normal transcription of TAL1 can be initiated from two of its promoters, 1a and 1b, which are highly conserved in vertebrates (Bockamp et al., 1995). An additional cryptic promoter in exon4 is only found being active in T-ALL cells due to the t (1;14) translocation (Bernard et al., 1992) and it also reveals enhancer activities in a number of studies, therefore will be further discussed in the following section 1.4.4.3. Usage of the two promoters reveals a cell lineage-dependent manner. The promoter 1a is active in erythroid and megakaryocytic lineages, for which are under regulation of GATA1 (Bockamp et al., 1995; Lecointe et al., 1994). Recent ChIP-chip study in K562 cells also confirmed the binding sites of GATA1 at the promoter 1a and +3 region at the TAL1 locus (Dhami et al., 2010). In addition, the promoter 1a also functions in the central nervous system (CNS). It has also been reported that the TAL1 promoter 1a together with +1 and +3 regions direct TAL1 expression within the brain and spinal cord regions (Sinclair et al., 1999). In this study, two GATA binding motifs have been identified at the TAL1 promoter 1a, and their pivotal role in all aspects of neural expression of TAL1 has been demonstrated by transgenic analysis. Providing the fact that the expression pattern of TAL1 in the CNS overlaps with GATA2 and GATA3 but not GATA1, it

strongly suggests the involvement of GATA2 and/or GATA3 in regulating TAL1 expression in the central nervous system (CNS) (Engel et al., 1992; George et al., 1994). In contrast, no GATA binding motif has been identified at the promoter 1b but only motifs for the ETS transcription factors. Additionally, the promoter 1b has been found to be active in primitive myeloid and mast cells, which is modulated by transcription factors such as PU-1, Elf1 and Sp1/Sp3 (Bockamp et al., 1998; Bockamp et al., 1997). Moreover, the promoter 1b is normally inactive in committed erythroid cells, but active in transient assays using $CD34^+$ myeloid cells functions in a GATA-independent manner (Bockamp et al., 1997), and also exhibits a low-level activity in T-cell leukemic cells (Bernard et al., 1991). In addition, bindings of cofactors p300, CBP and HDAC2 along with TAFII-250 have also been found at both promoter 1a and 1b in K562 cells, demonstrating that both promoters are active in this erythroid progenitor cells (P. Dhami's PhD thesis, 2008).

1.4.4.3 The TAL1 enhancers

The erythroid enhancer (TAL1 +51 / Tal1 +40)

The erythroid enhancer has been firstly identified in human and mouse TAL1 locus using ChIP-chip analyses and reporter assays (Delabesse et al., 2005). The histone H3 acetylation has been identified at the +50/+40 elements in human and murine TAL1-expressing haematopoietic cell lines, whereas no peak is evident in the TAL1 non-expressing lymphoid cells, suggesting that this element is only active in TAL1-expressing cells. In addition, it has been shown that the mouse +40 element functioned as an enhancer in erythroid cell lines *in vitro* by luciferase reporter assays, which directs Tal1 expression to primitive but not definitive erythroid cells in transgenic mice.

Subsequent studies have identified a numbers of highly conserved consensus transcription factor binding sites at the erythroid enhancer, including two GATA sites, two E-boxes (CANNTG) as well as two Ets (GGAW) sites (Ogilvy et al., 2007). Both GATA and E-box sites are situated in pairs in close proximity exhibiting 9 bp and 6 bp spacing for the 5' and 3' GATA/E-box motif respectively. The spacing of 3' GATA/E-box has been previously shown to preclude recruitment of the TEC complex in erythroid cell lines (Wadman et al., 1997). Moreover, it has been demonstrated that the 3' GATA/E-box motif is particularly important for

midbrain enhancer function, whereas the 5' GATA/E-box motif is predominantly responsible for blood activity in the murine +40 erythroid enhancer (Ogilvy et al., 2007). As illustrated in Figure 1.22, the sequence alignment shows that two GATA/E-box motifs at the TAL1 erythroid enhancer are highly conserved in vertebrates from human all the way back to Chicken.

Figure 1.22: DNA sequence conservation of GATA/E-box motifs (3' and 5') at the TAL1 erythroid enhancer across six vertebrate species. Gaps in alignments are shown in dashed lines (--). The E-box/GATA composite motifs are shown in bold and highlighted in red and pink boxes respectively. The erythroid enhancer is also known as the +51 or +40 element in human and mouse, respectively.

Previous studies have shown that the histone modification of acetylation and DNase I hypersensitive site at the +50 element (Delabesse et al., 2005; Follows et al., 2007). However, recent studies have found that two highly conserved GATA/E-box motifs are located at the +51 region (also see Figure 1.22), which is 1 kb adjacent to +50 element (Dhami et al., 2010). This GATA/E-box composite site has been showed to be the canonical hallmarks of the TEC, and is consistent the murine equivalent +40 erythroid enhancer in previous studies (Ogilvy et al., 2007). Henceforward, the core erythroid enhancer of human TAL1 has been renamed as the +51 region. The enhancer activity of the +51 elements has been detected by reporter assays in human K562 but not in HPB-ALL cells, revealing its specificity to the erythroid lineage (Dhami et al., 2010). The TEC, being particularly recruited at the GATA/E-box motif, is specifically found binding at the +51 region in TAL1 expressing erythroid lineages such as K562 and HEL cells (Dhami et al., 2010). In addition, previous studies have shown the bindings of p300, CBP and HDAC2 at the +51 region. It has been found that RNA polymerase II and TAFII 250 also bind to the TAL1 +51, suggesting that it is responsible for recruiting the TEC as well as RNAP II and TAFII 250 in erythroid cells (Dhami et al., 2010).

The haematopoietic stem cell enhancer (TAL1 +21/+20 or Tal1 +19/+18)

The haematopoietic stem cell enhancer is located at +21/+20 in human and at +19/+18 in mouse. Originally, the analysis of Tal1 regulation has identified a 5.5 kb genomic element located at 3' of the Tal1 coding region, which contains two DNase I hypersensitive sites at +19 and +18 regions (Gottgens et al., 1997). It has been shown that the stem cell enhancer is able to direct reporter gene expression in mouse to hematopoietic progenitors and endothelial cells during development, long-term repopulation of HSCs and hematopoietic progenitors but not mature cells (Sanchez et al., 1999; Sanchez et al., 2001). Subsequent transgenic studies have demonstrated that a 641 bp genomic element containing the +19 hypersensitive site is sufficient to drive reporter gene expression in embryonic tissues, which is then defined as the +19 core enhancer (Gottgens et al., 2002b). In contrast, it has been observed that the +18 element has no enhancer activity in transgenic mouse embryos. Although the +18 element fails to function as a classic enhancer on it own, it role in boosting the activity of the +19 enhancer has been demonstrated by reporter assays (Smith et al., 2008). Providing the fact that the +19 and +18 elements present the identical tissue-specific pattern of DNase I hypersensitivity as well as specificity of enhancement, these two elements function as a regulatory unit to promote the TAL1 expression.

In addition, the TAL1 expression driven by the Tal1 +18/19 enhancer is able to rescue the formation of early hematopoietic progenitors and yolk sac angiogenesis but not erythropoiesis in $Tal1^{-/-}$ embryos, suggesting that this enhancer plays a vital role in progenitors but is not sufficient for supporting erythroid maturation (Sanchez et al., 2001). In addition, it has been demonstrated that the stem cell enhancer is controlled by a multi-protein complex containing GATA-2, Elf-1, Fli-1 transcription factors, which are essential for activity of this enhancer in blood and endothelial cells (Gottgens et al., 2002b). However, TAL1 expression is not dependent on the activity of this enhancer for blood cell formation *in vitro* or *vivo* (Gottgens et al., 2004; Silberstein et al., 2005), which is in agreement with the role of the erythroid enhancer in driving TAL1 expression during haematopoietic development. Most recently, a study on this enhancer has shown that the *in vivo* deletion of the +19 enhancer leads to viable mice with normal TAL1 expression in mature haematopoietic lineages, whereas the expression of TAL1 is reduced in stem and progenitor cell compartments (Spensberger et al., 2012). The results

again affirm the vital role of the stem cell enhancer in haematopoietic stem cells and progenitor cells developments but not necessary for mature haematopoietic populations.

1.4.4.4 Other enhancers

In addition to two well-characterised enhancers as present above, a number of other *cis*-acting elements across the TAL1 locus have been demonstrated their enhancer activity by the extensive studies of TAL1 transcriptional regulation in the last two decades. These elements including 3' enhancer at +23 and 5' enhancers at -4, -7 and -9 have been shown to have (i) enhancer activity by reporter assay and transgenic analysis, and (ii) active hallmarks of histone modification by ChIP analysis and DNase I hypersensitivity by microarray analysis (Dhami et al., 2010; Follows et al., 2006).

The +23 neural enhancer

Originally, a novel enhancer at the murine Tal1 +23 element has been identified using comparative sequence analysis with the syntenic TAL1 regions in human, mouse and chicken (Gottgens et al., 2000). Transgenic reporter assays conducted in *Xenopus* embryos have demonstrated that the +23 enhancer directs the reporter gene expression to the hindbrain and spinal cord, which is subsequently termed as the neural enhancer.

The +7 element (P^{Exon4})

It has been shown that the +7 element functions as a promoter only in T-ALL cells with chromosomal translocation. In addition to its role as a cryptic promoter, it has both positive and negative functions in transcription regulation. It has been evident that the +7 element functions as a powerful chromatin-dependent silencer in primitive myeloid cells, while as an enhancer in erythroid cells (Gottgens et al., 1997). In addition, a functional clustering analysis of histone modification in erythroid K562 has shown that the +7 element is clustered with enhancers such as +51 and +20/+19, instead of being categorised into the subset of promoters (Dhami et al., 2010), suggesting its possible character as an enhancer in erythroid cells.

The +3 element

First, it has been found that the sequence of the +3 element is highly conserved between human and mouse loci which indicates the presence of an important regulatory elements (Begley et al., 1994). The characteristics of the +3 element are quite different from other 3' enhancers (Gottgens et al., 1997). It has been shown that the +3 element is modestly active in erythroid cells but inactive in primitive myeloid cells in transient assays, whereas it reveals a high activity in both cell types instable transfection experiments, suggesting the full activity of this enhancer is chromatin dependent. Furthermore, it has also been evident that the +3 element functions as an enhancer to direct TAL1 expression to the developing brain in mouse (Courtes et al., 2000).

The -3/-4 endothelial enhancer

The -3/-4 element has been identified as a 5' core endothelial enhancer that is able to direct the expression to endothelial lineage (Sinclair et al., 1999). Further analyses have provided a much more comprehensive picture about its role in regulating TAL1 expression (Gottgens et al., 2004). It has been found that the -3/-4 enhancer is capable of targeting expression to hematopoietic progenitors and endothelium. In addition, deletion of the -3/-4 element leads to an abolished endothelial expression in all transgenic embryos, which further demonstrates its vital role in endothelium during embryonic development. Moreover, five conserved consensus Ets family binding motifs have been identified at this enhancer and the mutations at each of these Ets sites result in the partial disruption of its enhancer activity, which demonstrate a degree of redundancy between these Ets motifs and a critical role for Ets family factors in regulation the activity of this enhancer. Elf1 and Fli1, two Ets family protein being part of the multi-protein complex in regulating the stem cell enhancer in haematopoietic progenitor cells, have also been found to bind to the -3/-4 enhancer both *in vivo* and *vitro*, implying some degrees of functional similarities may be shared between these two enhancers.

The -7 enhancer

The -7 is a *cis*-acting regulatory element that remains to be fully studied. It has been shown that the -7 element exhibits DNAse I hypersensitivity in K562 cells (Leroy-Viard et al., 1994). Until recently, the studies have provided additional

insight about its function in TAL1 regulation using ChIP-chip analyses and reporter assays in human K562 and HPB-ALL cells (Dhami et al., 2010). It has been shown that the histone modifications including hyper mono-, di- and tri-methylation of H3K4, and hypo-acetylation of H4K16 are presented at the -7 element in K562 cells, which are all hallmarks associated to active enhancer regions, whereas in HPB-ALL cells it is marked by inactive histone modification. In addition, the -7 element clusters along with the other known TAL1 enhancers such as +51, +20/+19 and -10 in the functional clustering analysis. Moreover, it has been shown that the -7 element is capable of enhancing reporter expressing only under the control of SV40 promoter but not TAL1 promoter 1a in both K562 and HPB-ALL cells, suggesting its enhancer activity is depended on local chromatin environment or it regulates the TAL1 activity not via the promoter 1a.

The -9 enhancer

The DNaseI hypersensitivity has been detected at the -9/-10 element in human TAL1 expressing T-cell leukemia and erythroid (K562 and HEL) cell lines (Leroy-Viard et al., 1994). Subsequent studies have demonstrated that the enhancer activity of -9/-8 element is shown in murine erythroid F4N cells, but not in primitive myeloid M1 cells and lymphoid BW5147 cells (Gottgens et al., 1997). In addition, the high level of enhancer activity of the human TAL1 -9/-10 element under the control of TAL1 promoter 1a has also been demonstrated in K562 but not in HPB-ALL (lymphoid) cells by reporter assays (Dhami et al., 2010). Moreover, a number of active hallmarks of histone modifications have been shown at this region including hyper mono-, di- and tri-methylation of H3 K4 as well as hypo-acetylation of H4 K16 in K562 cells, which further support its enhancer activity (Dhami, PhD Thesis, University of Cambridge, 2005; Dhami et al., 2010). In contrast, the -9/10 enhancer shows histone hallmarks associated with inactive regions, including hyper-acetylation of H4 K16, hypo di- and tri-methylation of H3 K4 in HPB-ALL cells. It is consistent with the fact that this region is transcriptional inactive in the TAL1 non-expressing HPB-ALL cells. In addition, this enhancer has also showed GATA-1 binding activity in K562 cells, implying its possible roles in erythroid lineages (Dhami et al., 2010).

1.4.4.5 Putative CTCF insulators

The CTCF insulator plays critical roles in transcription regulation either as an enhancer-blocker or a chromatin barrier. A number of CTCF binding sites has been identified across the TAL1 locus in both human and mouse. As shown in Figure 1.23a, CTCF occupancy has been found at TAL1 +57, +53, +40 and -31 elements in human erythroid K562 cell using ChIP-chip analysis (Dhami et al., 2010). Strong insulator activity of all four CTCF-binding elements has also been confirmed by the enhancer-blocking assays. As previously mentioned, the TAL1 expression in erythroid lineages is directed by the erythroid enhancer at +51. Providing the fact that a CTCF insulator at +40 is located right in the middle of the erythroid enhancer and the TAL1 promoter, it raises a question that how the +51 enhancer manage to communicate with its target promoter regardless of the presence of a regulatory elements with enhance-blocking function in erythroid K562 cells. It is speculated that the +51 enhancer is in contact with its promoters via long-range looping interaction. Apart from the +40 element, the +53 region is another interesting element. In addition to its insulator function, it also presented various hallmarks for an active promoter, including hyper-acetylation of H3K9, hyper di- and trimethylation of H3K4, hypo-dimethylation of H3K9 and hypo-acetylation of H4K16 in K562, HL-60, HPB-ALL and Jurkat cells (Dhami, PhD Thesis, University of Cambridge, 2005). Moreover, the high ratios for tri- to mono-methylation for H3 K4 across the +53 region in K562 and HPB-ALL also indicate the transcriptional activity of a promoter region (Dhami, PhD Thesis, University of Cambridge, 2005). Indeed, this region is associated with three of novel transcripts of the unknown function that has been reported in recent studies of the TAL1 locus (Dhami et al., 2010) and annotated in the ENSEMBL (http://www.ensembl.org) database. In addition, a recent study has also shown a CTCF binding site at Tal1 +45 in mouse haematopoietic progenitors 416B cells by ChIP-chip analysis (Follows et al., 2012). The +45 element is both necessary and sufficient for insulator function in haematopoietic cells *in vitro* (Figure 1.23b). In addition, it has been found that the tissue-specific TAL1 promoter expression in brain tissue can be increased by incorporation of a fragment of the +45 element in a transgenic reporter assay. These data suggest that the +45 element functions as a boundary element to separate the active and inactive chromatin domains between Tal1/Pdzk1ip1 and Cyp4x1.

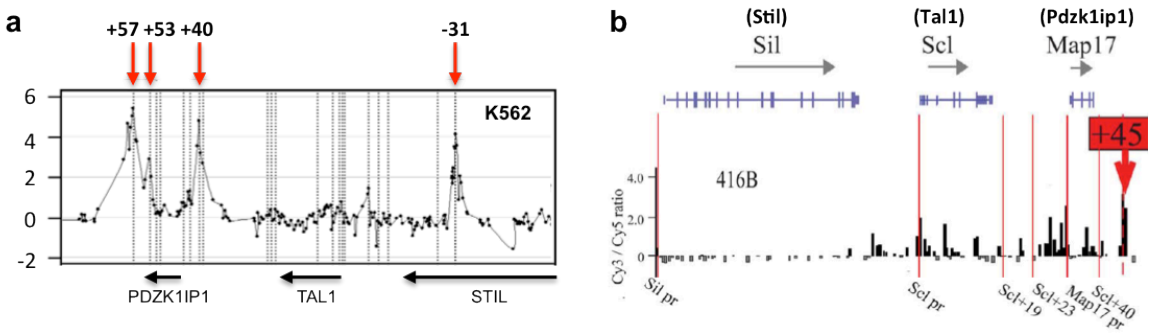


Figure 1.23: ChIP-chip profiles of CTCF occupancy at the TAL1 loci in human K562 cells (a) and mouse 416B cells (b). Figures are adapted from Dhami et al., 2010 and Follows et al., 2012. CTCF binding sites are highlighted in red arrows.

1.4.5 The TAL1 regulon

Previous studies have reported that the regulatory elements of TAL1 are confined to a 65 kb genomic region between upstream and downstream flanking genes in human (Gottgens et al., 2002a). However, this TAL1 regulon fails to include the enhancer at +51 with critical function for the TAL1 expression in erythroid lineages, which has been identified several years later (Delabesse et al., 2005). The size of the human TAL1 regulon has been redefined by a most recently study, to a region encompassing 88 kb from the -31 element upstream of TAL1 to the +57 element downstream of the TAL1 +51 enhancer (Dhami et al., 2010). As discussed in previous section, both +57 and -31 are CTCF-binding sites with insulator functions, which consistent with the role of CTCF-binding elements in defining regulatory domains. This newly proposed TAL1 regulon has encompassed yet all *cis*-regulatory elements that are known to be functionally associated with the TAL1 transcriptional regulation (Dhami et al., 2010; Gottgens et al., 2010). The study conducted in this thesis is based on the concept of the TAL1 regulon, intended to further explore the relationship between these *cis*-regulatory elements and the TAL1 expression. The size of the TAL1 regulon makes the 3C technology a good approach for assessing chromatin interactions between *cis*-acting elements within the regulon (see section 1.2.4.2 and Chapter 3 & 4).

1.4.6 The TAL1 genomic tiling path microarray

ChIP-chip technology has been used to investigate the transcriptional regulation of TAL1 during haematopoiesis (Dhami, PhD Thesis, University of Cambridge, 2005 and Dhami, et al., 2010). For that purpose, a customized tiling-path microarray has been constructed to cover the TAL1 locus in a high-resolution. The 5'-aminolink

array surface chemistry developed at the Sanger Institute has been used to construct a sensitive array platform for the TAL1 locus. The single-stranded DNA molecules which are derived from double-stranded PCR products were retained on the surface of a glass slide by using the surface chemistry technology (Dhami et al., 2005). As shown in the Figure 1.24, a 5'-(C6) amino-link modification is incorporated at the end of one strand of DNA, which allows the modified strand to be covalently attached to the surface of the slide. Meanwhile, the unmodified strand is then removed by chemical and physical denaturation. The resultant single-stranded DNA molecules are the array elements which provide an ideal target for hybridizing with the labeled DNA samples.

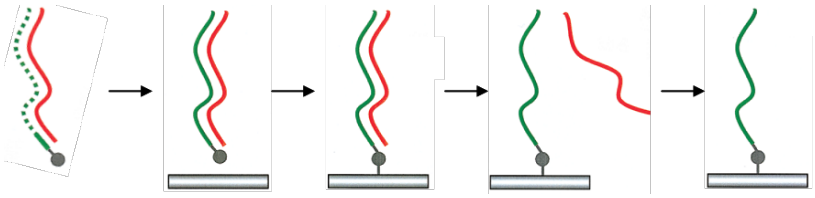


Figure 1.24: Schematic diagram shows the approach adopted to make single-stranded PCR products (Figure adapted from Dhami et al., 2005). Double-stranded PCR products (red/green denote strands) containing a 5'-(C6) amino-linked modification on one strand (grey circle on the green strand) are arrayed onto the surface of the slide (grey bar). Covalent attachment occurs via the 5' amino-link (black line) and the slide surface. Denaturation of the PCR product renders them single-stranded.

The genomic regions represented on the custom-made TAL1 tiling path array included the TAL1 gene, flanked upstream by STIL and CMPK1 and downstream by PDZK1IP1, CYP4A22 and CYP4Z1 genes (Figure 1.25). The tiling path array covered 256 kb of human chromosome 1, with 419 PCR amplicons designed at an average product size of 458 bp. The array also covered 207 Kb of mouse chromosome 4, with 530 PCR amplicons designed at an average size of 443 bp. The TAL1 tiling path microarray in combination with ChIP assays have been used to identify numbers of DNA-protein interactions as well as histone modifications at the TAL1 locus, which allow defining regulatory elements at this locus (Dhami, PhD Thesis, University of Cambridge, 2005; Dhami, et al., 2010).

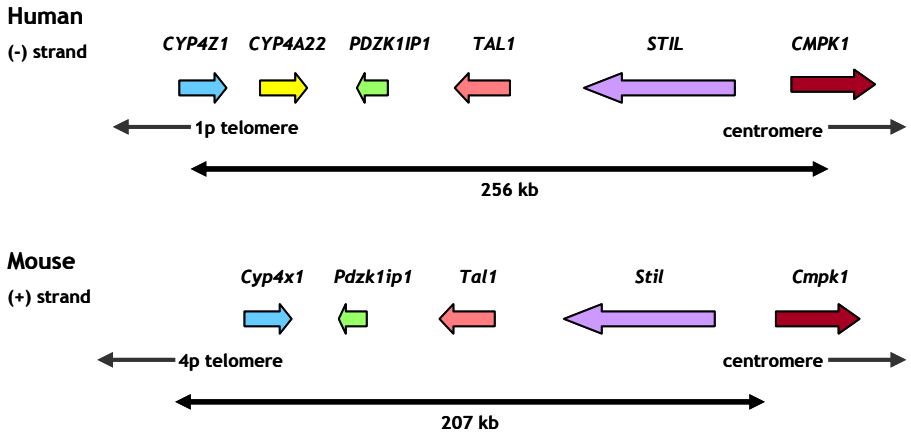


Figure 1.25: A schematic of the genomic regions across the human and mouse TAL1 loci. The size of the genomic regions in human and mouse are shown by the black double-headed arrows. The thick, coloured arrows represent the genes. The human tiling path covers the entire length of CYP4A22, PDZK1IP1, TAL1, and STIL genes, while CYP4Z1 and CMPK1 genes are partially covered. The murine tilling path covers the entire length of Cyp4x1, Pdzk1ip1, Tal1, and Stil genes. The gene order and the direction of transcription in human and mouse are shown as annotated on the chromosomes 1 and 4 in the respective genomes. The orientation of the loci with respect to the centromere and telomere is shown at the bottom with black arrows.

1.5 Aims of this thesis

At the time this PhD project was initiated, little was known mechanistically how the *cis*-acting elements separated across a ~90 kb of sequence function to regulate TAL1 expression (section 1.4). Providing that numbers of examples about gene expression regulated by promoter-enhancer loops (illustrated in section 1.1), a hypothesis was proposed that these regulatory elements may also be brought into spatial close-proximity for regulate TAL1 expression. Developments in chromosome conformation capture (3C) and its related technologies (such as 4C) provided molecular tools to examining the spatial organisation of chromatin in a high resolution (discussed in section 1.2). With this in mind, it became possible to characterise the 3-dimensional chromatin organisation of the TAL1 locus and try to understand how a number of distal *cis*-regulatory elements of TAL1 coordinate its expression via chromatin loops. The aims of this thesis were as follows:

1. To use the existing 3C-PCR approach to characterise chromatin interactions between selected *cis*-regulatory elements at the TAL1 loci in human and mouse cell lines as well as primary cells, expecting that 3D chromatin looping interactions would further our understanding about the possible mechanism of transcriptional regulation at the TAL1 loci.

2. To optimise the existing 4C approach incorporating with high-resolution TAL1 tiling path array platform in order to improve the sensitivity and reproducibility of the method.

3. Having improved 4C-array for aim 2, to then apply the method in characterising chromatin looping interactions at the human TAL1 locus. To determine the transcriptional-related chromatin organisations of the TAL1 locus by comparing between the human TAL1 active and inactive cell lines.

4. To use an existing GATA1 siRNA knockdown system to characterise the possible roles of GATA1 transcription factor and the TAL1-containing erythroid complex (TEC) in modulating the TAL1 expression in human erythroid cell line.

5. To further extend the paradigm based on the results learned from aim 1-4, in order to determine whether the LYL1 locus also adopts a similar mechanism for its transcriptional regulation as observed at the TAL1 locus.

Chapter 2 Materials and Methods

Abbreviations

ATP: adenosine triphosphate

ChIP: chromatin immunoprecipitation

Cy3: cyanine-3

Cy5: cyanine-5

dATP: 2' deoxyadenosine 5'-triphosphate

dCTP: 2' deoxycytidine 5'-triphosphate

dGTP: 2' deoxyguanosine 5'-triphosphate

dTTP: 2' deoxythymidine 5'-triphosphate

DTT: dithiothreitol

DMEM: Dulbecco's Modified Eagle Medium

mRNA: messenger ribonucleic acid

LB: lysogeny broth

PBS: phosphate buffered saline

PCR: polymerase chain reaction

RNA: ribonucleic acid

RNA Pol II: ribonucleic acid polymerase II

RT-PCR: real-time polymerase chain reaction

SCL: stem cell leukaemia

SD: standard deviation

SDS: sodium dodecyl sulfate

Materials

2.1 Composition of solutions

2.1.1 Western blotting

Cell lysis buffer (for nuclear protein extraction)

- 10 mM Tris-HCL pH 8
- 10 mM NaCl
- 0.2% Igepal (Sigma)
- 1% protease inhibitor cocktail (Sigma)

Extraction buffer (for nuclear protein extraction)

- 20 mM HEPES pH 7.9
- 0.2 mM EDTA
- 25% Glycerol
- 1.5 mM MgCl2
- 0.42 M NaCl
- 0.001 M DTT (Invitrogen)
- 1% protease inhibitor cocktail (Sigma)

1 × MOPS running buffer

1 × MOPS used in western blotting was prepared by diluting 100 ml 20 × MOPS running buffer (Invitrogen) in 1 litre of deionised water.

1 × NuPAGE transfer buffer

1 × Transfer buffer used in western blotting was prepared by diluting 50 ml 20 × NuPAGE Transfer buffer (Invitrogen) in 950 ml of deionised water and adding 100 ml of methanol.

10 × Tris-buffered saline (TBS)

10 × TBS used in western blotting was prepared by dissolving the following salts in 1 litre of deionised water and the pH was adjusted to 7.6 with HCl.

- 24.4 g Tris-HCl
- 80 g NaCl

1 × Tris-buffered saline Tween 20 (TBST)

1 × TBST used in western blotting was prepared by diluting 100 ml 10 × TBS in 1 litre of deionised water and adding 1 ml of Tween 20.

2.1.2 Microarray hybridisation

10 × dNTP mix (for DNA labelling of ChIP assays)

- 1 mM dCTP
- 2 mM each of dGTP, dTTP and dATP

Tecan-hyb buffer

- 50% formamide (Fluka)
- 5% dextran sulphate
- 0.1% Tween 20 (BDH)
- 2 X SSC
- 10 mM Tris pH 7.4

PBS/0.05% Tween 20 (Hyb wash solution 1)

PBS/0.05% Tween 20 for washing the microarrays was prepared by dissolving the following salts in one litre of HPLC water

- 7.33 g NaCl
- 2.36 g Na_2HPO_4
- 1.52 g NaH_2PO_4H2O
- 500 µl Tween 20

20 × SSC

20 × SSC used in the washing steps during TECAN hybridisation was prepared by dissolving the following salts in 1 litre of HPLC water.

- 175.3 g NaCl
- 88.2 g tri-sodium citrate

1× SSC

1 × SSC used in the washing steps during the TECAN hybridisation was prepared by by diluting 50 ml 20 × SSC in 1 litre of HPLC water.

2.1.3 Chromatin immunoprecipitation (ChIP)

Cell lysis buffer (CLB)

- 10 mM Tris-HCl pH 8.0
- 10 mM NaCl
- 0.2% Igepal (Sigma)
- 10 mM sodium butyrate (Sigma)
- 50 µg/ml PMSF (Sigma)
- 1 µg/ml leupeptin

Nuclei lysis buffer (NLB)

- 50 mM Tris-HCl pH 8.1
- 10 mM EDTA
- 1% SDS
- 10 mM sodium butyrate (Sigma)
- 50 µg/ml PMSF (Sigma)
- 1 µg/ml leupeptin

IP dilution buffer (IPDB)

- 20 mM Tris-HCl pH 8.1
- 150 mM NaCl
- 2 mM EDTA
- 1% Triton X-100 (Sigma)
- 0.01% SDS
- 10 mM sodium butyrate (Sigma)
- 50 µg/ml PMSF (Sigma)
- 1 µg/ml leupeptin

IP wash buffer 1 (IPWB1)

- 20 mM Tris-HCl pH 8.1
- 50 mM NaCl
- 2 mM EDTA

- 1% Triton X-100 (Sigma)
- 0.1% SDS

IP wash buffer 2 (IPWB2)

- 10 mM Tris-HCl pH 8.1
- 250 mM LiCl
- 1 mM EDTA
- 1% Igepal (Sigma)
- 1% deoxycholic acid

IP elution buffer (IPEB)

- 100 mM $NaHCO_3$
- 1% SDS

1 X PBS (for ChIP assays)

1 X PBS used for washing the cells in ChIP assay was prepared by dissolving the following salts in 1 litre of HPLC water and the pH was adjusted to 7.4

- 8 g NaCl
- 0.2 g KCl
- 1.44 g Na_2PO_4
- 0.24 g KH_2PO_4

*2.1.4 **Chromosome conformation capture** (3C and 4C)*

Cell lysis buffer

- 10 mM Tris-HCl (pH 8.0)
- 10 mM NaCl
- 0.2% NP-40
- Complete Protease Inhibitor Cocktail Tablet (Roche)

Washing buffer (for 4C assay)

- 10 mM Tris-HCl (pH 7.5)
- 2 M NaCl
- 1 mM EDTA

2.2 Reagents

2.2.1 Enzymes

- Csp6I (CviQI), 10 units/µl (Fermentas)
- NlaIII, 10units/µl (NEW ENGLAND BioLabs)
- T4 DNA ligase, 400000 cohesive end units/ml (NEW ENGLAND BioLabs)
- T4 DNA Polymerase, 3000 units/ml (NEW ENGLAND BioLabs)
- HotStarTaq DNA Polymerase, 5 units/µl (QIAGEN)
- Proteinase K (10mg/ml stock) (Gibco-BRL 25530-031)
- RNase A (10mg/ml stock) (ICN Biochemicals, 101076)
- Gold AmpliTaq, 5 units/µl (Perkin Elmer-Cetus)

2.2.2 Antibodies

- Anti-histone H3 tri methyl K4 antibody - ChIP Grade IgG (Abcam, ab8580)
- Anti-GATA-1 (M-20) IgG (Santa Cruz Biotechnology, sc-1234)
- Anti-LDB 1(N-18) IgG (Santa Cruz Biotechnology, sc-11198X)
- Anti-E47 (N-649) IgG (Santa Cruz Biotechnology, sc-763)
- Anti-RNA polymerase II antibody – ChIP Grade IgG (Abcam, ab5408)
- Anti-CTCF(C-20) IgG (Santa Cruz Biotechnology, sc-15914)
- Anti-Rad21 antibody - ChIP Grade IgG (Abcam, ab992)
- Donkey anti-goat IgG-HRP (Santa Cruz Biotechnology, sc-2020)
- Goat anti-rabbit IgG-HRP (Santa Cruz Biotechnology, sc-2004)

2.2.3 Other reagents

- Trizol (GibcoBRL)
- Human Cot 1 DNA (Invitrogen)
- Mouse Cot 1 DNA (Invitrogen)
- Herring sperm DNA (Sigma-Aldrich)
- Cy3-dUTP (Amersham)
- Cy5-dUTP (Amersham)
- 20% SDS (w/v)

- 20% Triton-X-100 (v/v)
- Proteinase K (50 mg/ml)
- RNase A (10 mg/ml)
- Phenol Solution (pH 8.0)(Sigma-Aldrich)
- Chloroform (Sigma-Aldrich)
- 3M sodium acetate (pH 5.2).
- RPMI 1640 (Sigma-Aldrich)
- DMEM (Invitrogen)
- Fetal bovine serum (Invitrogen)
- Horse serum (Invitrogen)
- LB (Sigma-Aldrich)
- Kanamycin (Sigma-Aldrich)
- Chloramphenicol (Sigma-Aldrich)

2.2.4 Kits

- Dynabeads M-280 (Invitrogen)
- SuperScriptTM II RNase H-reverse transcriptase (Invitrogen)
- DNeasy Blood & Tissue Kit (QIAGEN)
- RNesay Mini Kit (QIAGEN)
- QIAquick PCR purification Kit (QIAGEN)
- QIAquick Gel extraction Kit (QIAGEN)
- QIAGEN Large-Construct Kit (QIAGEN)
- EndoFree Plasmid Maxi Kit (QIAGEN)
- Fast Start Universal SYBR Green Master (ROX) (Roche)
- Quant-iT™ DNA Assay Kit (Invitrogen)
- Steady-Glo® Luciferase Assay System (Promega, E2510)
- β-Galactosidase Enzyme Assay System (Promega, E2000)
- Calf intestinal alkaline phosphatase, (CIAP) (Promega, M1821)

2.2.5 Other consumables

- Clear Seal Diamond Heat Sealing Film (Thermo Scientific)
- Thermo-Fast® 96 Semi-Skirted PCR Plate (Thermo Scientific)
- 15 ml BD Falcon™ Conical Tubes (Fisher Scientific)

2.2.6 BAC/PACs

- RP1-18D14 (Invitrogen) for the human *SCL* locus
- RP11-963I8 (Invitrogen) for the human *LYL1* locus
- RP23-453H14 (BPRC) for the mouse *Scl* locus

Methods

2.3 Tissue Culture

2.3.1 Culturing of cell lines

Table 2.1: List of all the cell lines used for the study presented in this thesis.

Cell line	Media	Serum	Supplements
K562	RPMI 1640	10 % v/v FBS	1% penicillin-streptomycin, 2 mM L-glutamine
HPB-ALL	RPMI 1640	20 % v/v FBS	1% penicillin-streptomycin, 2 mM L-glutamine
MEL	DMEM	10 % v/v FBS	1% penicillin-streptomycin, 2 mM L-glutamine
BW5147	DMEM	5% v/v FBS + 10 % v/v HS	1% penicillin-streptomycin, 2 mM L-glutamine

Note: All cell types were cultured under 5% CO_2 at 37°C.

1. Once confluent after 24 hours to 48 hours of culture, fresh media was added to each flask effecting a 1/2 dilution, and any clumps of cells were gently broken up using a syringe.

2. The number of cells needed for chromatin immunoprecipitation (ChIP) or chromosome conformation capture (3C) experiments was quite high. Therefore, culture volumes to obtain the required number of cells per flask were suitably scaled up in 225 cm^2 culture flasks with vented caps (Corning).

3. For frozen storage, cells were pelleted and resuspended at approximately 1 × 10^7 cells/ml in FBS (GibcoBRL) with 10% (v/v) DMSO. The resulting cell mixture was transferred into polypropylene cryotubes which were cooled overnight to -80°C. The cryotubes were then transferred to the gas phase of a liquid nitrogen vessel (approximately -180°C) for permanent storage. To reconstitute cultures, cells were thawed rapidly at 37°C water bath, washed once with fresh media and finally resuspended in 10 ml of fresh media.

2.3.2 Mouse primary cells

Single cell suspensions from mouse primary tissues were prepared for 3C assays. The spleen tissues were harvested from adult female pregnant mice, which were kindly provided by Dr Alison M Michie, Paul O'Gorman Leukaemia Research Centre.

1. Murine spleens were stored in DMEM medium once harvested.

2. Spleens were carefully but thoroughly crushed and re-suspended in PBS.

3. The cell suspension was passed through the 70 mm mesh into the collection tubes (15 ml Falcon) to remove cell clumps (as the single cell suspensions were required).

4. The lymphocyte cells were isolated by carefully adding 4ml of Lympholyte®-Mammal (Cedarlane Laboratory) to the bottom of the tube, and then centrifuged at 626 g (1800 rpm) for 30 minutes, room temperature.

5. The middle layer of the tube which contained the mononuclear cells (mixed population of lymphocytes) was carefully transferred into the new collection tubes without disturbing the pellet of red blood cells at the bottom.

6. The lymphocytes were washed twice with 10 ml PBS, centrifuged at 1400 rpm for 10 minutes, room temperature.

7. The concentration of the cells was counted by cytometry and prepared for RNA extraction and 3C fixation.

2.3.3 Bacterial culture

1. A single colony from a freshly streaked selective plate was used to inoculate a starter culture of 2 ml LB medium containing the appropriate selective antibiotic (25 mg/ml kanamycin for human TAL1 PAC construct or 50 mg/ml chloramphenicol for murine Tal1 BAC construct). Cultures were incubated for 16 h at 37°C with vigorous shaking (~300 rpm).

2. 0.2 ml of the starter culture was diluted into 200 ml selective LB medium (1/1000 dilution) and was incubated at 37°C for 16 hours with vigorous shaking (~300 rpm).

3. The bacterial cells were harvested by centrifugation at 6000 g for 15 minutes at 4°C. All traces of supernatant were removed by inverting the open centrifuge tube until all media had been drained. The cell pellet can be kept at -20°C.

2.4 BAC/PAC purification (QIAGEN Large-Construct Kit)

1. The bacterial pellet was re-suspend in 20 ml Buffer P1 (Resuspension buffer, 50 mM Tris·HCl, pH 8.0; 10 mM EDTA) supplied with 100 µg/ml RNase A.

2. 20 ml Buffer P2 (Lysis buffer, 200 mM NaOH, 1% SDS) was added and was mixed by inverting 4-6 times. Cells were then incubated at room temperature for 5 minutes.

3. 20 ml chilled Buffer P3 (Neutralization buffer, 3.0 M potassium acetate, pH 5.5) was added and was mixed immediately but gently by inverting 4-6 times, followed by 10 minutes incubation on ice.

4. The sample was centrifuged at 20000 g (12000 rpm in the Beckman JA-17 rotor) for 30 minutes at 4°C. The supernatant containing the BAC/PAC DNA promptly was removed.

5. The lysate was filtered through a folded filter pre-wetted with distilled water.

6. A QIAGEN-tip 500 was equilibrated by applying 10 ml Buffer QBT (Equilibration buffer, 750 mM NaCl; 50 mM MOPS, pH 7.0; 15% isopropanol; 0.15% Triton-X-100).

7. The sample from step 5 was applied to the QIAGEN-tip and was allowed to enter the resin by gravity flow.

8. The QIAGEN-tip was washed with 2 x 30 ml Buffer QC (Wash buffer, 1.0 M NaCl; 50 mM MOPS, pH 7.0; 15% isopropanol).

9. The DNA was eluted with 15 ml Buffer QF (Elution buffer, 1.25 M NaCl; 50 mM Tris·HCl, pH 8.5; 15% isopropanol), pre-warmed to 65°C.

10. 10.5 ml (0.7 volumes) of isopropanol was added to the eluted DNA. The sample was centrifuged immediately at 15000 g (9500 rpm in the Beckman JA-17 rotor) for 30 minutes at 4°C, and the supernatant was removed.

11. The DNA pellet was washed with 5 ml room temperature 70% ethanol and was centrifuged at 15000 g (9500 rpm in the Beckman JA-17 rotor) for 15 minutes. The supernatant was carefully removed without disturbing the pellet.

12. The pellet was air-dried for 5 to 10 minutes, and the DNA was dissolved in 200 µl of Buffer EB (10 mM Tris·HCl, pH 8.5).

2.5 DNA extraction (QIAGEN DNeasy Blood & Tissue Kit)

1. Cultured cells: The appropriate number of cells were centrifuged (maximum 5 × 10^6) for 5 minutes at 300 g. The pellet was resuspended in 200 µl PBS. 20 µl

proteinase K and 4 µl RNase A (100 mg/ml) were added, mixed by vortexing, and incubated for 2 minutes at room temperature.

2. 200 µl Buffer AL (without added ethanol) was added and mixed thoroughly by vortexing. The sample was then incubated at 56°C for 10 minutes.

3. 200 µl of ethanol was added to the sample and was mixed thoroughly by vortexing.

4. The mixture was transfered into the DNeasy Mini spin column placed in a 2 ml collection tube. The sample was centrifuged at 6000 g (8000 rpm) for 1 minute.

5. The DNeasy Mini spin column was placed in a new 2 ml collection tube. 500 µl Buffer AW1 was added and the sample was centrifuged for 1 minute at 6000 g (8000 rpm).

6. The DNeasy Mini spin column was placed in a new 2 ml collection tube. 500 µl Buffer AW2 was added and the sample was centrifuged for 3 minutes at 20000 g (14000 rpm) to dry the DNeasy membrane.

7. The DNeasy Mini spin column was placed in a clean 1.5 ml or 2 ml microcentrifuge tube. 200 µl Buffer AE was added directly onto the DNeasy membrane. The sample was then incubated at room temperature for 1 minute, and was centrifuged for 1 minute at 6000 g (8000 rpm) to elute.

2.6 RNA extraction and cDNA library generation

2.6.1 RNA extraction

1. 1 ml TRIZOL reagent was added per 2×10^6 cells and homogenised in a 14 ml tube. These RNA samples can be stored at -80°C.

2. The homogenised samples were incubated at room temperature for 5 minutes (or at 42°C for 15 minutes if taken from -80°C). The homogenate was split into 1ml aliquots in 2 ml round-bottom microcentrifuge tubes.

3. 0.2 ml of chloroform was added per 1 ml of TRIZOL reagent used.

4. The samples were mixed vigorously for 15 sec. and were incubated at room temperature for 2-3 minutes.

5. The samples were centrifuged at 12000 g (14000 rpm) for 15 minutes at 4°C.

6. The upper aqueous phase was transfer to a new 2 ml microcentrifuge tube and 0.5 ml isopropanol was added per 1 ml of TRIZOL reagent used. The samples were mixed by inversion.

7. The samples were incubated at room temperature for 10 minutes and centrifuged at 12000 g (14000 rpm) for 15 minutes at 4°C. The RNA was visualised as a pellet at the bottom of the tube.

8. The supernatant was removed carefully and the pellet was washed once with 75% ethanol, adding at least 1 ml per 1 ml of TRIZOL reagent used. The samples were then centrifuged at 7500 g for 5 minutes at 4°C.

9. The supernatant was removed and the RNA pellet was air-dried. The pellet was re-dissolved by adding 87.5 µl RNase-free H_2O and incubating at 55 to 60°C.

10. DNaseI treatment: 10 µl RDD and 2.5 µl DNaseI were added and the samples were incubated at room temperature for 10 minutes.

11. RNeasy mini Kit (QIAGEN) was used to clean up the samples. The RNA was eluted twice in 40 µl RNase-free H_2O (80 µl in total).

12. 7 µl of RNA was taken for quality check and quantification: 2 µl for Nanodrop and 5 µl for 1 × TAE electrophoresis at 70 volts, 1.5 hours. RNA was stored at -80°C until used for further analysis.

2.6.2 Reverse Transcription & cDNA Synthesis

First strand cDNAs were synthesized the SuperScript™ II RNase H-reverse transcriptase (Invitrogen) as follows.

1. 5 µg RNA and 5 µl 50 ng/µl random primers (Invitrogen) were mixed in a total volume of 18 µl (add RNase-free water to bring up to volume if necessary).

2. The samples were incubated at 70°C for 10 minutes and chilled on ice for 5 minutes.

3. To the RNA/oligo mix, the following was added:

- 6 µl of 5 × first strand buffer (Invitrogen)
- 3 µl of 0.1M DTT (Invitrogen)
- 1.5 µl of 10 mM dNTPs (Invitrogen)
- 1.5 µl of SuperScriptTM II RNase H-reverse transcriptase (Invitrogen)
- 30 µl TOTAL

4. The reaction mix was mixed gently by flicking and incubated at 42°C for 90 minutes on a hot block.

5. The samples were heat-inactivated at 70°C for 10 minutes and chilled on ice.

6. The resulting cDNAs were diluted to 5 ng/µl by adding 970 µl RNAse-free water to the sample. The cDNAs can be stored at -20°C.

2.7 Quantitative real-time PCR

2.7.1 Primer design

1. Primer pairs for all real-time PCR assays performed for this thesis were designed by using the IDT (integrated DNA technologies) Primer Quest software (http://eu.idtdna.com/Scitools/Applications/Primerquest/). To avoid amplifying non-specific PCR products from other parts of the genome, primer sequences were compared against the entire human/mouse genome using Primer Quest software as well as BLAT (UCSC, http://genome.ucsc.edu/cgi-bin/hgBlat?command=start).

2. The amplicons generated by these primer pairs were between 80 bp to 120 bp in length.

3. The complete lists of all the primer pair sequences, used in the real-time PCR assays, are provided in the Appendix 1 and 2.

2.7.2 Quantitative real-time PCR amplification

Quantitative real-time (qRT)-PCR was used (i) to assess the knockdown efficiency in the siRNA assays, (ii) to investigate expression of putative target genes in the siRNA assays, (iii) to investigate DNA enrichment of putative target region in the ChIP assays, (iv) to assess the 3C digestion efficiency in the 3C assays.

1. The SYBR green PCRs were set up in 96-well reaction plate (ABgene) in triplicates by mixing the following reagents on ice.

For expression analysis (cDNA quantification):

H_2O	3.5 µl
forward and reverse primer mix (final concentration 300 nM)	4 µl
2 × SYBR green PCR mastermix (Roche)	12.5 µl
cDNA (10 ng/µl)	5 µl
Total volume:	25 µl

For ChIP-qPCR analysis:

H_2O	3.5 µl
forward and reverse primer mix (final concentration 300 nM)	4 µl
2 × SYBR green PCR mastermix (Roche)	12.5 µl
ChIP (1:10/20 dilution) / input DNA (1:40 dilution)	5 µl
Total volume:	25 µl

For 3C digestion efficiency analysis:

H_2O	3.5 µl
forward and reverse primer mix (final concentration 300 nM)	4 µl
2 × SYBR green PCR mastermix (Roche)	12.5 µl
3C / non-digested DNA (10 ng/µl)	5 µl
Total volume:	25 µl

2. PCR was performed on Mx3000P QPCR Systems (Agilent Technologies) using the following conditions: 10 minutes at 95°C, 40-45 cycles of 95°C for 15 seconds and 60°C for 1 minute.

2.7.3 Data analyses

Ct values were extracted using MxPro QPCR Software (Agilent Technologies) with the same threshold and data analyses were performed as follows:

For expression assays (cDNA quantification):

$^{\Delta}Ct = Ct_{\text{house-keeping gene}} - Ct_{\text{gene of interest}}$

$^{\Delta\Delta}Ct = {}^{\Delta}Ct_{\text{luciferase control}} - {}^{\Delta}Ct_{\text{siRNA knockdown}}$

Fold repression = (1+PCR yield) $^{\Delta\Delta}CT$

% of mRNA remained after knockdown = 100/ Fold repression

For ChIP-qPCR assays:

$^{\Delta}Ct = Ct_{\text{INPUT}} - Ct_{\text{sample}}$

$^{\Delta\Delta}Ct = {}^{\Delta}Ct - {}^{\Delta}Ct_{\text{median of control regions}}$

Fold enrichment (log_2) = $^{\Delta\Delta}Ct$

For 3C digestion efficiency assays:

$^{\Delta}Ct = Ct_{site} - Ct_{control}$

$^{\Delta\Delta}Ct = {}^{\Delta}Ct_{digested} - {}^{\Delta}Ct_{non\text{-}digested}$

Digestion efficiency (%) = 100- 100/2^$^{\Delta\Delta}$Ct

2.8 Chromatin Immunoprecipitation (ChIP)

K562, HPB-ALL, MEL and BW5147 cell lines and siRNA transfected K562 cells were used for chromatin immunoprecipitation. Fresh cultures were grown (Table 2.1) for each cell line and cells were harvested for ChIP. Approximately 1×10^8 cells (1×10^7 cells used in siRNA knockdown study) were harvested for each ChIP assay which were then used to set-up five to ten immunoprecipitation (IP) conditions as described below.

2.8.1 Fixation

1. The cells were collected by centrifuging at 259 g for 8 minutes at room temperature and re-suspended in 50 ml of serum free media in a glass flask.

2. Cross-link was performed by adding formaldehyde (37%, BDH AnalaR). 1355 µl formaldehyde was added drop-wise to a final concentration of 1%.

3. The cross-linking was carried out at room temperature with constant but gentle stirring for 10 minutes.

4. 3.425 ml of 2M glycine was added to a final concentration of 0.125M constant but gentle stirring for 5 minutes at room temperature to stop the crosslinking reaction.

5. Cells were transferred to 50 ml falcon tubes and kept on ice whenever possible. The cells were pelleted by centrifuging at 259 g for 6-8 minutes at 4°C and washed with1.5 ml of ice-cold PBS.

6. After washing, the cells were pelleted at 720 g at 4°C for 5 minutes and the supernatant was removed.

2.8.2 Cell and Nuclei Lysis

7. Cells were lysed by adding 1.5 × pellet volumes of ice-cold cell lysis buffer (CLB). The cell pellets were gently resuspended and incubated on ice for 10 minutes.

8. The nuclei were recovered by centrifuging the samples at 1125 g for 5 minutes at 4°C.

9. After carefully removing the supernatant, the nuclei were lysed by resuspending the pellet in 1.2 ml of nuclei lysis buffer (NLB) and incubating on ice for 10 minutes.

2.8.3 Sonication

10. 720 µl of IP dilution buffer (IPDB) was added and the samples were transferred to 5 ml glass falcon tubes (Falcon 2058).

11. The chromatin was sonicated to reduce the DNA length to an average size of 400-600 bp using the Sanyo/MES Soniprep sonicator. The tip of the probe was dipped to reach approximately halfway down the total level of the liquid sample and the tube was kept constantly on ice (Conditions for sonication like number of bursts, length of bursts and power setting depend on the sonicator tip used). The settings used for the sonicator were:

• Amplitude: 14 microns

• Number of bursts: 8

• Length of bursts: 30 seconds

The samples were allowed to cool on ice for 1 minute between each pulse.

12. The sonicated chromatin was transferred to 2 ml microfuge tubes and spun down at 18000 g for 10 minutes at 4ºC.

2.8.4 Immunoprecipitation

13. The supernatant was transferred to a 15 ml falcon tube and 4.1 ml of IP dilution buffer was added.

14. The chromatin was pre-cleared by adding 100 µl of normal rabbit IgG (Upstate Biotechnology). The samples were incubated for 1 hour at 4°C on a rotating wheel.

15. 200 µl of homogeneous protein G-agarose suspension (Roche) was added to the pre-cleared chromatin and the samples were incubated for 3 hours to overnight at 4°C on a rotating wheel.

16. The samples were centrifuged at 1620 g for 2 minutes at 4°C to pellet the protein G-agarose beads and the supernatant was used to set up various immunoprecipitation (IP) conditions in 2 ml microfuge tubes. An aliquot of 270 µl of

chromatin was stored at -20°C to be used as input sample for array hybridisations and real-time PCR. An NLB: IPDB buffer at the ratio of 1:4 was prepared to set-up IP conditions as follows:

• No chromatin – 1350 µl NLB: IPDB buffer

• No antibody – 675 µl chromatin + 675 µl NLB: IPDB buffer

• Normal Rabbit IgG – 675 µl chromatin + 675 µl NLB: IPDB buffer + 10 µl rabbit IgG (Upstate Biotechnology)

• Test IP conditions – 675 µl chromatin + 675 µl NLB: IPDB buffer + 5-20 µg* of test antibody

(*5-20 µg antibody was used to assay for histone modifications and 10 µg antibody was used to assay for specific transcription factors). A complete list of antibodies that were tested for this study is provided in section 2.2 of this chapter.

17. The samples were incubated at 4°C overnight on a rotating wheel.

18. The samples were centrifuged at 18000 g for 5 minutes at 4ºC and the lysate/Ab samples were transferred to fresh 2 ml microfuge tubes. 100 µl of homogeneous protein G-agarose suspension was added to each sample and the samples were incubated at 4°C for at least 3 hours on a rotating wheel.

19. The samples were centrifuged at 6800 g for 30 seconds at 4°C to pellet the protein G-agarose beads.

20. The supernatant was removed and the protein G-agarose beads were carefully washed. For each wash, the wash buffer was added, the samples were vortexed briefly, were centrifuged at 6800 g for 2 minutes at 4°C and left to stand on ice for 1 minute before removing the supernatant. The washes were carried out in the following sequence:

a) The beads were washed twice with 750 µl of cold IP wash buffer 1. The beads were transferred to a 1.5 ml microfuge tube after the first wash.

b) The beads were washed once with 750 µl of cold IP wash buffer 2.

c) The beads were washed twice with 750 µl of cold TE pH 8.0.

2.8.5 Elution

21. DNA-protein-antibody complexes were eluted from the protein G-agarose beads by adding 225 µl of IP elution buffer (IPEB). The bead pellets were

resuspended in IPEB, briefly vortexed and centrifuged at 6800 g for 2 minutes at room temperature.

22. The supernatant was collected in fresh 1.5 ml microfuge tubes. The bead pellets in the original tubes were resuspended in 225 µl of IPEB again, briefly vortexed and centrifuged at 6800 g for 2 minutes. Both the elutions were combined in the same tube.

2.8.6 Reversal of cross-links

23. The reversal of cross-links was performed for the input sample which had previously been stored at -20°C. 0.1 µl of RNase A (10 mg/ml, 50 Kunitz units/mg*, ICN Biochemicals) and 16.2 µl of 5M NaCl (to the final concentration of 0.3 M) was added to the input DNA sample.

24. Similarly, 0.2 µl of RNase A (10 mg/ml, 50 Kunitz units/mg*) and 27 µl of 5M NaCl (to a final concentration of 0.3 M) was added to each of the IP test samples. All the samples including the input DNA sample were incubated at 65ºC for 6 hours to reverse the cross-links.

25. 9 µl of Proteinase K (10 mg/ml, 20 U/mg, GibcoBRL) was added to each sample and incubated at 45ºC overnight.

2.8.7 Extraction of DNA

26. 2 µl of yeast tRNA (5 mg/ml, Invitrogen) was added to each sample. Then 250 µl of phenol (Sigma) and 250 µl of chloroform were added.

27. The samples were vortexed and centrifuged at 18000 g for 5 minutes at room temperature. The aqueous layer (top layer) was collected in fresh 1.5 ml microfuge tubes and 500 µl of chloroform was added to each sample.

28. The samples were vortexed and centrifuged at 18000 g for 5 minutes at room temperature. The aqueous layer was transferred to a fresh 2.0 ml microfuge tubes.

29. 5 µg of glycogen (5 mg/ml, Roche), 1 µl of yeast tRNA (5 mg/ml, Invitrogen) and 50 µl of 3M NaAc (pH 5.2) was added to each sample and mixed well. The DNA was precipitated with 1375 µl of 100% ethanol and incubating at -70ºC for 30 minutes (or -20°C overnight).

30. The samples were centrifuged at 20800 g for 20 minutes at 4ºC. The DNA pellets were washed with 500 µl of ice-cold 70% ethanol and air-dried for 10-15 minutes.

31. The DNA pellets of the IP samples were resuspended in 50 µl of sterile filtered HPLC water and 100 µl for the Input DNA samples.

32. 5 µl of each sample was run on a 1% agarose 1 × TBE gel and visualised with ethidium bromide to check DNA size. Samples were stored at –20°C.

*The amount of enzyme causing the hydrolysis of RNA at a rate such that k (velocity constant) equals unity at 25°C and pH 5.0.

2.9 Chromosome Conformation Capture (3C)

2.9.1 Cell Fixation and Lysis

1. Single-cell suspension was prepared containing 5 × 10^7 cells in 50 ml growth media (serum-free).

2. 1.35ml of 37% formaldehyde was added to obtain a final concentration of 1% and incubated for 15 minutes at room temperature with slow agitation.

3. Fixation was stopped by adding 3.425 ml of 2 M glycine to a concentration of 0.125 M, the sample was then chilled on ice.

4. The cells were collected by centrifugation at 300 g for 5 minutes, 4°C.

5. Cells were washed with 20 ml of cold PBS and harvest again by centrifugation for 5 minutes at 300 g, 4°C.

6. The cell pellet was re-suspended in 50 ml of cold lysis buffer (10 mM Tris·HCl, pH 8.0; 10 mM NaCl; 0.2% NP-40 and 1 tablet complete protease inhibitors) and incubated for 10 minutes on ice with occasional mixing to release nuclei.

7. The nuclei were harvested by centrifugation for 5 minutes at 600 g at 4°C.

8. Aliquots of nuclei can be stored at -80°C.

2.9.2 Treatment of Nuclei with Restriction Endonuclease

9. The nuclear pellet was re-suspended in 0.5 ml of 1.2 × restriction buffer.

10. 7.5 µl of 20% SDS was added to a final concentration of 0.3% and incubated at 37°C for 1 hour with vigorous shaking (950 rpm, Thermo-mixer, Eppendorf).

11. 50 µl of 20% Triton-X-100 was added to a concentration of 1.8% and incubated at 37°C for 1 hour with vigorous shaking (950 rpm, Thermo-mixer, Eppendorf) to sequester the SDS.

12. 600 units of a highly concentrated restriction enzyme (Csp6I) was added to carry out restriction digestion overnight at 37°C with shaking (950 rpm).

13. The enzyme was inactivated by adding 40 µl of 20% SDS to a concentration of 1.3% and incubated at 65°C for 25 minutes with shaking (1050 rpm).

2.9.3 Ligation of DNA Cross-Linked to Proteins

14. 7 ml of 1.1 × ligation buffer was mixed with the ‘‘restriction’’ solution in a 15 ml tube.

15. 375 µl of 20% Triton-X-100 was added to a final concentration of 1% and incubated at 37°C for 1 hour with slow shaking to sequester the SDS.

16. 1200 units T4 DNA ligase (NEB) was added and incubated for 4-5 hours at 16°C and then for 30 minutes at room temperature with slow agitation.

2.9.4 Cross-Link Reversion and DNA Purification

17. Cross-linking was reversed by overnight incubation of the whole sample at 65°C in the presence of 40 µl of 50 mg/ml of proteinase K.

18. 40 µl of 10 mg/ml of RNase A was added to remove the RNA at 37°C for 30-45 minutes.

19. The sample was purified with 7 ml of phenol, and chloroform in the 15 ml tube. Centrifugation was performed for 10 minutes at 3000 g, room temperature.

20. The equivalent volume of sterile water was added to the sample (about 7.5 ml) in a 50 ml tube.

21. 1.5 ml of 3 M sodium acetate (pH 5.2) and 35 ml of 100% ethanol were added and incubated overnight at -20°C for DNA precipitation.

22. 3C DNA was precipitated by centrifugation for 30 minutes at 4000 g, 4°C.

23. The DNA pellet was washed with 10 ml of cold 70% ethanol and centrifuged for 30 minutes at 4000 g and 4°C.

24. The DNA pellet was air-dried and dissolved in 150 µl of buffer EB (10 mM Tris·HCl, pH 8.5).

2.9.5 BAC/PAC control templates preparation

25. 20 µg of BAC/PAC DNA was digested in 500 µl volume, by adding 50 µl 10 × restriction Buffer (Buffer B) and 200 units restriction enzyme (using the same

restriction enzyme as used for the 3C nuclei, Csp6I was used in these experiments). The sample was incubated overnight at 37°C.

26. The restriction enzyme was inactivated by incubating at 65°C for 20 minutes.

27. The sample was purified with 500 µl phenol/chloroform in a 2 ml microcentrifuge tube. Centrifugation was performed for 5 minutes at 3000 g, room temperature. The upper, aqueous layer was transferred to a fresh 2ml microcentrifuge tube, this step was repeated twice.

28. The DNA was precipitated by adding 1/10 volume of 3 M sodium acetate, pH 5.2 with 2.5 volumes of ice-cold 100% ethanol and incubated for 3 hour at -20°C. Centrifugation was performed for 30 minutes at 18000 g and 4°C.

29. The supernatant was removed and the DNA pellet was washed with 500 µl of 70% cold ethanol. Centrifugation was performed for 30 minutes at 18000 g and 4°C.

30. The DNA pellet was air-dried and re-suspended in 63 µl of H_2O (2 µl for no-ligase control) by incubating at 37°C for 5 minutes.

31. Ligation was set up with 61 µl digested BAC/PAC DNA, 7 µl 10 × T4 DNA ligase reaction buffer and 2 µl T4 DNA ligase. The sample was incubated 4 hours at 16°C and 30 minutes at room temperature.

32. The T4 DNA ligase was inactivated by incubating at 65°C for 20 minutes.

33. The DNA was purified by QIAQuick PCR purification kit and eluted with 100 µl Buffer EB (10 mM Tris·HCl, pH 8.5). The DNA sample (3C control template) obtained was stored at -20ºC.

34. 1 µl of control template and 2 µl of digestion control were analyzed on a 1% agarose 1 × TBE gel and stained with Safestain (Invitrogen) for visualization.

2.9.6 PCR Analysis of Ligation Products

35. PCR primers were designed for analysis of 3C ligation products. Annealing temperatures of primers were 56-60°C, and the size of PCR products were within the range of 100-300 bp.

36. A reaction mixture was prepared as follows:

10x buffer	2.5 µl
dNTP	0.5 µl
Qiagen HotStar Taq	0.5 µl

PCR H_2O	14.5 µl
DNA template	5.0 µl
Primers	2.0 µl
Total	25.0 µl

Three technical replicates were performed with each PCR primer pair.

37. The PCR was carried out as follows: initial denaturation for 15 minutes at 95°C; 36 cycles of 30 seconds at 94°C, 30 seconds at 56-58°C, 1 minute at 72°C; and 10 minutes at 72°C.

38. The PCR products were analysed by electrophoresis using a 2.5 % agarose gel with SYBR Safe DNA gel stain (Invitrogen).

2.9.7 Quantification of 3C PCR products

39. The gel image was captured by FLA-5000 scanner (FUJIFILM) and analysed with AIDA Image Analyzer software (ray).

40. The intensity of the PCR band was quantified and normalised against the background reading.

41. The mean and standard deviation (SD) of each PCR product were calculated in Microsoft 'Excel'.

4. The mean and SD were again normalised against the PCR efficiency of each primer pair acquired from the BAC/PAC control templates.

4. The 3C profiles were visualised by plotting the normalised mean ratios of amount of PCR product (y-axis) along with its genomic coordinates (x-axis).

All primer pairs used in the 3C-PCR analyses are provided in the Appendix 3.

2.10 4C-array

2.10.1 Biotinylated primer extension

1. 30 µg of 3C DNA (calculate based on DNA concentration acquired by PicoGreen assay) was re-suspended in 500 µl buffer EB (10 mM Tris·HCl, pH 8.5) and sonicated with Misonix Sonicator 3000.

The settings used for the sonicator were:

- Number of bursts: 8
- Length of bursts: 25 seconds

The samples were allowed to cool on ice for 10 seconds between each pulse.

2. The sample was transferred into 2ml Eppendorf tube and 50 µl (~10% volume) 3M sodium acetate (NaAc), pH 5.2 and 1.5 ml 100% Ethanol were added per sample. The sample was incubated 2 hours at -20°C for DNA precipitation.

3. The sample was centrifuged at 20000 g (14000 rpm) for 30 minutes in 4°C.

4. The DNA pellet was air-dried and re-dissolved in 100 µl H_2O.

5. The sample was purified by QIAQuick PCR purification kit and eluted in 100 µl of Buffer EB (10 mM Tris·HCl, pH 8.5).

6. The sonicated DNA was quantified by PicoGreen assay and the efficiency of sonication was verified by gel electrophoresis.

7. Primer extension was set up with 8 µg aliquots of 3C sonicated DNA.

10x PCR buffer	5.0 µl
10 mM dNTPs	1.0 µl
HotStar Taq	1.0 µl
PCR H_2O	37.5 µl
DNA template	5.0 µl
5 pmol 10µM biotin primer	0.5 µl
Total	50.0 µl

8. PCR Programme: initial denaturation for 15 minutes at 95°C; 18-36 cycles of 30 seconds at 94°C, 2 minutes at 57-60°C, 1 minute at 72°C; 10 minutes at 72°C. PCR reaction was immediately chilled on ice.

9. The sample were purified with QIAQuick PCR purification kit and eluted in 50 µl of Buffer EB (10 mM Tris·HCl, pH 8.5).

2.10.2 Capture of target fragments by binding to Dynabeads M-280 Streptavidin

10. The Dynabeads (INVITROGEN) were re-suspended by vortexing the vial to obtain a homogeneous suspension. 20 µl (200 µg) re-suspended beads per

sample were transferred to a 1.5 ml micro-centrifuge tube. The tube was placed on the magnet for 2 minutes.

11. The supernatant was carefully pipetted off while the tube remains on the magnet. 20 µl Binding Solution per sample was added along the inside wall of the tube where the beads were collected and gently re-suspended by pipetting (supplied with Dynabeads® kilobaseBINDER™ Kit).

12. The supernatant (Binding Solution) was removed while the tube was placed on the magnet. The beads were then re-suspended in 50 µl Binding Solution per sample.

13. 50 µl of a solution containing the biotinylated DNA-fragments after Biotinylated primer extension was added to the re-suspended beads and mixed carefully to avoid foaming of the solution.

14. The tube was incubated at room temperature for 3 hours on a roller to keep the beads in suspension.

15. The supernatant was removed while the tube was placed on the magnet. The Dynabeads/DNA-complex was washed twice with 500 µl Washing Solution (10 mM Tris·HCl, pH 7.5; 1 mM EDTA and 2.0 M NaCl), and once with in 100 µl 1 x NEB buffer 2 (10 mM Tris·HCl, 10 mM $MgCl_2$, 50 mM NaCl, 1 mM DTT).

2.10.3 Blunting reaction

16. Blunting reaction was performed as follows:

H_2O:	97.3 µl
10 × NEB buffer 2:	11.0 µl
10 mM dNTP:	1.0 µl
T4 DNA Polymerase:	0.2 µl
100 × BSA:	0.5 µl
Total:	110.0 µl

The tube was incubated at room temperature for 1 hour on a roller to keep the beads in suspension.

2.10.4 Ligation of adapter to the bait-prey complex

17. The supernatant was removed while the tube was placed on the magnet. The Dynabeads/DNA-complex was washed twice with 500 µl Washing Solution and once with 100 µl 1 × T4 DNA ligase reaction buffer (50 mM Tris-HCl, 10 mM $MgCl_2$, 10 mM DTT, 1 mM ATP).

18. Adapter ligation reaction was performed as follows:

H_2O:	41.0 µl
10 × T4 DNA ligase reaction buffer:	5.0 µl
100 pmol Blunt/Solexa adapter:	2.0 µl
800 units T4 DNA ligase:	2.0 µl
Total:	50.0 µl

19. The tube was incubated at room temperature overnight on a roller to keep the beads in suspension.

20. The supernatant was removed while the tube was placed on the magnet. The Dynabeads/DNA-complex was washed twice with 500 µl washing solution and once with 100 µl Buffer EB (10 mM Tris·HCl, pH 8.5). Beads can then be stored as suspension in 100 µl Buffer EB at 4°C.

2.10.5 PCR amplification

21. The supernatant was removed while the tube was placed on the magnet. Dynabeads/DNA-complex was re-suspended in 42 µl 1x PCR master mix:

10 × buffer	5.0 µl
dNTP	1.0 µl
Qiagen HotStar Taq	1.0 µl
Ampli Taq	0.5 µl
PCR H_2O	34.5 µl
Total	42.0 µl

22. Transfer beads to PCR tube with mixed primer (DNase adapter primer + nest primer) and set up reaction as follows:

Mixed Primer	8.0 µl

PCR master mix	42.0 µl
Total	50.0 µl

23. PCR Programme: initial denaturation for 15 minutes at 95°C; 15-18 cycles of 30 seconds at 94°C, 30 seconds at 57°C, 1 minute at 72°C; 10 minutes at 72°C. PCR reaction was immediately chilled on ice.

24. The PCR reaction was transferred to a new 1.5 ml tube. The PCR supernatant was collected while the tube was placed on the magnet.

25. PCR products were purified with QIAQuick PCR purification kit. DNA was eluted in 50 µl Buffer EB (10 mM Tris·HCl, pH 8.5) and stored at -20°C. The sample was ready for labelling and microarray hybridisation.

All primer pairs used in the 4C-array analyses are provided in the Appendix 4.

2.11 Microarray Hybridisation

2.11.1 Random Labeling of DNA samples

1. Two 2.0 ml microfuge tubes were used for the labelling reactions, one for ChIP or 4C DNA and one for the input DNA sample.

2. ChIP/4C DNA was random labelled by mixing 10 µl of ChIP DNA or 20 µl of 4C library with 60 µl 2.5 × Random Primers Solution and made up to 130.5 µl with water (in each tube).

3. The input DNA sample was random labelled by mixing 2 µl of input DNA with 60 µl 2.5 × Random Primers Solution and make up to 130.5 µl with water (in each tube).

4. ChIP and input DNA were denatured in a heat block for 10 minutes at 100°C, and immediately chilled on ice.

5. The following reagents were added on ice:

- 15 µl dNTP mix
- 1.5 µl Cy3 (for ChIP/4C samples) / Cy5 (for input samples) labelled dCTP (1 mM)
- 3 µl Klenow Fragment

The sample was mixed gently but thoroughly.

6. The sample was incubated at 37°C overnight and the reaction was stopped by adding 15 µl Stop buffer.

2.11.2 Purification of labeled DNA Samples

7. The unlabelled nucleotides from DNA labelling reactions were removed using G-50 columns (GE).

8. The resin in the G-50 columns was re-suspended by gentle vortexing.

9. The cap was loosened by one-quarter turn and the bottom cap was removed.

10. The columns were placed in a 1.5 ml micro-centrifuge tube for support.

11. The columns were centrifuged for 1 minute at 735 g (4000 rpm). The columns were used immediately after preparation to avoid the resin drying out.

12. 50 µl of sterile filtered HPLC water was applied to the resin bed and the columns were centrifuged at 4000 rpm for 1 minute.

13. The columns were placed in a new 1.5 ml tube and 50 µl of the labelling reactions were slowly applied to the centre of the angled surface of the compacted resin bed of each of the columns.

14. The columns were centrifuged for 2 minutes at 4000 rpm. The purified samples were collected in the bottom of the support tube.

15. The flow-through samples were retained and the columns were discarded.

16. 5 µl of each sample were run on a 1 × TBE agarose gel (1%) to check the labelling.

2.11.3 Hybridisation of the SCL genomic tiling path array

17. For each hybridisation, the precipitation reactions were set up in 2.0 ml tubes. The hybridisation DNA samples for precipitation per tube were prepared as follows:

- ~180 µl ChIP Cy3 labelled DNA
- ~180 µl input Cy5 labelled DNA
- 135 µl human/mouse Cot1 DNA
- 55 µl 3M sodium acetate (NaAc), pH 5.2

- 1200 µl 100% Ethanol (cold)

18. All the tubes were gently mixed and the rack was covered with tin foil and precipitated at -70°C for 30-40 minutes.

19. After 30-40 minutes, the precipitated DNA samples were centrifuged for 20 minutes at 13000 rpm at room temperature.

20. The supernatant was removed and the pellet was washed with 500 µl 80% Ethanol, and again centrifuged at 13000 rpm for 5 minutes at room temperature.

21. The supernatant was removed and the tube was again centrifuged at 13000 rpm for 1 minute. The liquid residue of supernatant was removed with a small tip. The pellets were air-dried for 5 minutes.

22. 100 µl of hybridization buffer was added to re-suspend the pellets of each tube. The tube was then incubated for 2-3 minutes in a 70°C heat block and the pellet was re-suspended by shaking. The pellet was completely re-suspended before continuing with the experiment.

23. The samples were pooled from two tubes (for the same hybridisation) and 3 µl yeast tRNA (100 mg/ml) was added.

24. The samples were denatured for 10 minutes at 100°C and then placed immediately on ice.

25. The samples were then placed on a 37°C heat block and incubated for 60 minutes.

26. The TECAN HS 4800 automatic hybridisation/wash station was prepared by placing the SCL genomic tiling path array into the appropriate clean chambers, priming the wash buffer pumps and loading the appropriate hybridisation/washing program protocol.

27. 175 µl of hybridisation buffer was injected into a TECAN slide chamber containing the SCL tiling array avoiding air bubbles.

28. The slides were pre-hybridised for 60 minutes at 37°C on a medium agitation setting, after which time the slides were automatically washed once in PBS/0.05% Tween 20 for 30 seconds at 37°C, and dried in preparation for injection of the labelled hybridisation mixture.

29. 175 µl labelled hybridisation sample was injected carefully into a TECAN slide chamber containing the SCL tiling array to avoid air bubble formation.

30. The slides were hybridised for 44-48 hours at 37°C on a medium agitation setting, after which time the slides were automatically washed and dried as follows:

- 10 times in PBS/0.05% Tween 20 at 37°C, each wash lastfor 1 minute with additional 30 seconds soak time
- 5 times in 0.1 × SSC at 52°C, each wash last for 1 minute with additional 2 minute soak time
- 10 times in PBS/0.05% Tween 20 at 23°C, each wash last for 1 minute with additional 30 seconds soak time
- Washing in HPLC grade water for 1 minute and 30 seconds without soaking at 23°C
- Drying with nitrogen gas

31. The slide was removed from the TECAN station and dried with compressed air.

32. The Cy3 and Cy5 images were scanned with ScanArray 4000XL Microarray Analysis System (Perkin Elmer) at 5 µm resolutions using a laser power of 100% and a photo multiplier tube (PMT) value of between 70%-85%.

33. Fluorescent intensities of each spot on the SCL tiling path array were quantitated using the ScanArray Express software (Perkin Elmer) using the adaptive circle quantitation method and the total normalisation method. The spots representing the array elements were automatically located by the software and the mean signal intensity values against background were calculated for Cy3 and Cy5 channels. The mean ratios of the Cy3/Cy5 channels were reported in the resulting Microsoft ‘Excel’ datasheet.

2.11.4 Microarray data analysis

34. Statistical analyses of the array data were conducted in Microsoft ‘Excel’. All the “unfound” spots on the array were excluded from the analyses. Quality control for the hybridisation of the arrays was performed by investigating the average intensity of the signals on the array and the signal/noise ratios. Arrays with significantly lower signal intensity and signal/ noise ratios are discarded from further analyses.

35. Mean ratios, standard deviations (SDs) and coefficients of variation (CVs) were calculated for the three replicate spots representing each array element in Microsoft 'Excel'. The mean ratios were normalised against the median values of all the mean ratios.

36. ChIP-chip or 4C-chip profiles were visualised by plotting the mean ratios (y-axis) along its genomic coordinates (x-axis) in Microsoft 'Excel'.

2.12 Quantifying transfection efficiency of siRNA by FACS analysis

2.12.1 siRNA transfection

1. Fresh medium was added to the cultured cells on the day before transfection. Cells were maintained at a density of 0.5 $\times 10^6$ /ml.

2. 5 $\times$ 10^6 of K562 cells were harvested and re-suspended in 100 µl of RPMI.1640. 2 µl of 100µM anti-GATA1 FITC-labelled siRNA were added into 0.2 ml cuvettes (Bio-Rad) along with 5 $\times$ 10^6 cells to a final concentration of 20 nM. Before transfection, the cells and siRNA were incubated on ice for 5-10 minutes. The Nucleofector^{TM} II system (Amaxa Biosystems) was used for the transfection with Programme T-16 (K562, ATCC).

3. A control was also performed with 100 µl cells incubated with FITC-labelled siRNA which was not transfected.

4. Transfected cells as well as non-transfected control were carefully transferred by 1 ml pipette and cultured in 10 ml of RPMI.1640 (10% v/v FCS and 100 µg/ml penicillin/streptomycin) in 25 cm^3 flask wit vented cap, Incubated at 37°C and 5% CO_2. The final concentration of the siRNA was 20 nM.

5. 1 ml of transfected and non-transfected cells were used for FACS assay at 0 hour as well as after 24 hours of transfection.

2.12.2 FACS analysis

6. The transfected and non-transfected cells were pelleted by centrifuged at 1300 rpm for 5 minutes at room temperature.

7. The cell pellet was washed once with 10 ml PBS, and centrifuge at 1300 rpm for 5 minutes at room temperature.

8. Cells were re-suspended in 300 µl of PBS and keep on ice for FACS.

9. The re-suspended cells were kept on ice and covered with aluminium foil.

10. Cells were flow-sorted using the BD FACS Calibur Flow Cytometer (BD Biosciences).

11. Data were analysed using WinMDI 2.9 (http://facs.scripps.edu/software.html) and percentage of transfected cells were calculated.

2.13 GATA-1 siRNA knockdown

1. 1.2 × 10^8 of K562 cells were cultured and harvested in total.

2. Half of the cells (6 × 10^7 cells) were transfected with anti-firefly luciferase siRNA (sense: 5'-CUUACGCUGAGUACUUCGAdTdT-3').

Another half (6 × 10^7 cells) was transfected with anti-GATA1 siRNA (sense: 5'-GGAUGGUAUUCAGACUCGAdTdT-3'). Procedures for transfection and cell culture were as described in section 2.13.1.

3. 3 × 10^7 cells transfected with anti-luciferase siRNA and 3 × 10^7 cells transfected with anti-GATA1 siRNA were harvested after 48-hour incubation for the further analyses:

- 2 × 10^6 cells were reserved for mRNA extraction and cDNA preparation
- 8 × 10^6 cells were reserved for protein extraction
- 1 × 10^7 cells were used for ChIP chromatin preparation
- 1 × 10^7 cells were used for 3C chromatin preparation

4. For the 96-hour time-point, 3 × 10^7 cells were re-transfected with either anti-GATA1 or anti-luciferase siRNA after the initial 48-hour incubation, and cultured for an additional 48 hours. The cells were harvested and prepared for the further analyses exactly the same as the 48-hour time-point.

5. siRNA experiments were performed in duplicate on two independent K562 cell line bio-replicates.

2.14 Western blotting

2.14.1 Protein extraction

1. Cells were harvested at by centrifugation at 1200 rpm for 5 minutes.

2. The cell pellets were washed once with 10 ml ice-cold PBS followed by centrifugation at 1200 rpm for 5 minutes.

3. Cells were re-suspended with 250 µl cell lysis buffer and incubated for 5 minutes on ice.

4. Nuclei were obtained by centrifugation at 11000 rpm for 1 minute at 4°C.

5. Nuclei were washed once with 250 µl cell lysis buffer followed by centrifugation 11000 rpm for 1 minute at 4°C.

6. Nuclei were re-suspended in 100 µl of extraction buffer and stored at -80°C.

2.14.2 Protein quantification

Nuclear proteins were quantified using the Bio-Rad Protein Assay Kit II.

1. 0, 0.625, 1.25, 2.5 or 5 µl of BSA (1.48 mg/ml) (Invitrogen) was added to the 96-well plate (Corning) in order to generate a standard curve.

2. 2 µl of nuclear protein extract was added to each well.

3. 200 µl of 1:5 diluted Dye reagent (Bio-Rad) was added to each well.

4. The reaction mixtures were mixed well by pipetting.

5. The samples were incubated for 20 minutes at room temperature.

6. The absorbance of each sample at 595 nm was measured using the MR7000 plate reader (DYNATECH).

7. The concentration of the nuclear proteins was calculated using the BSA standard curve in Microsoft 'Excel'.

2.14.3 Sample preparation for Western blotting

Nuclear protein samples were prepared under reducing conditions as follows:

30 µg of nuclear proteins	Variable
4 × LDS loading buffer (Invitrogen)	7.5 µl
sample reducing agent (Invitrogen)	3.0 µl
sterile water	Variable
Total volume:	30.0 µl

The samples were mixed by vortexing and denatured at 100°C for 2 minutes.

2.14.4 SDS-PAGE

1. NuPAGE 4-20% Novex Bis-Tris gels (Invitrogen) were washed with distilled water and combs were removed and the wells were washed with 1 × MOPS running buffer (Invitrogen) with a syringe.

2. The gels were assembled in XCell SureLock^{TM} Mini-Cell (Invitrogen).

3. The inner and outer chambers of the Mini-Cell tank were filled with 200 ml and 500 ml of 1 × MOPS Running Buffer respectively (**N.B.** 500 µl of anti-oxidant (Invitrogen) was added to the inner chambers.)

4. The denatured proteins (30 µl) and 7.5µl of See Blue Plus Standard (Invitrogen) were loaded into the wells.

5. The proteins were electrophoresed for 1 to 1.5 hours at constant 150 volts and a starting current of 125 mA/ gel at 4°C.

2.14.5 Blotting

1. 1 × NuPAGE Transfer Buffer (500ml) was prepared as follows:

20 × NuPAGE Transfer Buffer (Invitrogen)	25 ml
Methanol	50 ml
Sterile water	425 ml
Total volume:	500 ml

2. The transfer buffer and 1 litre deionised water were kept in the fridge for at least 30 minutes before use.

3. The gels were dissembled after electrophoresis.

4. Blotting pads, filter papers and gels were equilibrated in cold transfer buffer for 10 seconds.

5. PVDF membranes (Millipore) were equilibrated in 100% methanol for 10 seconds and transferred to the cold transfer buffer.

6. The blot modules were assembled as follows:

For 1 gel	**For 2 gels**
• top (+) • 3 blotting pads • 3 filter papers • transfer membrane	• top (+) • 2 blotting pads • 2 filter paper • transfer membrane

• gel • 3 filter papers • 3 blotting pads • bottom (-)	• second gel • 2 filter paper • 1 blotting pad • 2 filter paper • transfer membrane • first gel • 2 filter paper • 2 blotting pads • bottom (-)

7. Any air bubbles in blotting pads and between the gel and membrane were removed.

8. The blot module was clipped together firmly and placed into a transfer tank.

9. The blot module was filled with transfer buffer until the gel/membrane sandwich was covered in transfer buffer.

10. The outer chamber was filled with cold deionised water to the top.

11. The blotting was performed at a constant voltage of 30 volts for one gel and 35 volts for 2 gels and a starting current of 170 mA/ gel at 4°C for 2.5 hours.

2.14.6 Immunoblotting and detection

1. The membrane was blocked using blocking buffer (5% non-fat dry milk in tris-buffered-saline tween-20 (TBST)) at room temperature for 1 hour.

2. The membrane was then incubated with primary antibodies at the appropriate dilutions in 10 ml blocking buffer at 4°C overnight.

3. The membrane was then washed four times with TBST for 1 hour

4. The membrane was incubated with secondary antibodies at the appropriate dilutions in 10 ml blocking buffer at room temperature for 1 hour.

5. The membrane was then washed again four times with TBST for 1 hour.

6. The membrane was incubated with Luminata^{TM} Forte western HRP substrate (Millipore) for 5 minutes.

7. Signals were obtained using LAS-3000 imaging system (FUJIFILM) and measured with AIDA Image Analyzer software.

8. The membrane was then stained with 15 ml of water-diluted Dye Reagent Concentrate (Bio-Rad) at a dilution of 1:5, then distained with water and generally

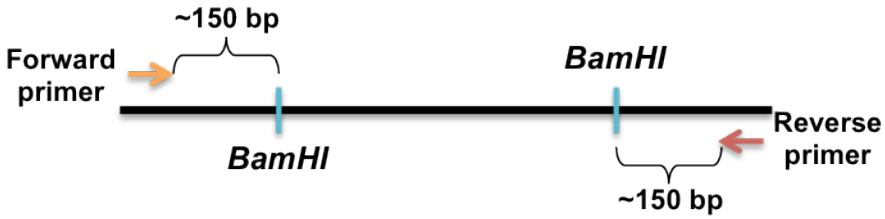
air-dried to visualise loading control. The signals were also captured using LAS-3000 imaging system and quantified with AIDA Image Analyzer software.

2.15 Enhancer trap reporter assay

2.15.1 Preparation of the insert fragment (putative enhancer)

1. The putative LYL1 enhancer fragment was amplified by PCR using primer primers designed against the genomic region in question.

2. The primers were designed flanking genomic BamHI sites (~150 bp from each sites) as illustrated beneath.



3. The reaction mixture was prepared as follows:

5 × Phusion buffer	10.0 µl
dNTP mix	1.0 µl
Phusion polymerase	0.5 µl
DNA template (BAC)	2.0 µl
primers (final con. 50 nM)	5.0 µl
H_2O	31.5 µl
Total	50.0 µl

4. The PCR was carried out as follows: initial denaturation for 2 minutes at 98°C; 30 cycles of 30 seconds at 98°C, 20 seconds at 65°C, 1 minute at 72°C; and 5 minutes at 72°C.

5. PCR products were purified by QIAquick PCR purification kit (QIAGEN), and eluted in 50 µl of nuclease-free H_2O.

6. DNA concentration was measured by NanoDrop-1000 spectrophotometer (Thermo Scientific).

7. Restriction enzyme digestion was conducted as follows:

10 × NEB buffer 4	5.0 µl
BamHI-HF	1.0 µl
PCR product (1 µg)	Variable
H_2O	Variable
Total	50.0 µl

Samples were incubated for 1 hour at 37°C.

8. The digested insert fragment was purified by gel electrophoresis on 1% TAE agarose gel. The DNA was extracted by QIAquick gel extraction kit (QIAGEN) and eluted in 50 µl of nuclease-free H_2O.

2.15.2 Preparation of reporter construct

9. The pGL3 reporter constructs containing either the SV40 promoter or SCL promoter 1a were digested by BamHI as described.

10. The linearized vectors were then de-phosphorylated using calf intestinal alkaline phosphatase (CIAP) according to manufacturer's protocol.

11. The dephosphorylated vectors were then treated with equal volume of Phenol/Chloroform to remove the residual of CIAP, and purified by QIAquick PCR purification kit (QIAGEN), eluted in 50 µl of nuclease-free H_2O.

2.15.3 Ligation and transformation

12. The insert was ligated to plasmid constructs in 2:1 and 5:1 molar ratios, respectively. The ligation reactions were conducted as follows:

10 × T4 ligase buffer	2.0 µl
T4 Ligase	1.0 µl
construct (50 ng)	Variable
insert	Variable
PCR H_2O	Variable
Total	20.0 µl

13. The samples were mixed well and incubated overnight at 16°C in a water bath.

14. Competent cells (Promega, 200 µl/vial) were thawed on ice.

15. 50 µl of compenent cells and 10 µl of ligation reactions were placed into1.5 ml tube and incubated on ice for 10 minutes.

16. Cells and ligation reactions were incubated at 42°C for 90 seconds and immediately chilled on ice for 2 minutes.

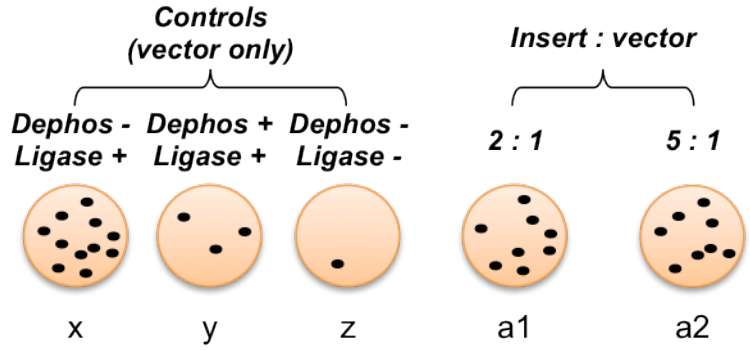
17. 940 µl of LB media was added and incubated for 1 hour at 37°C, 300 rpm.

18. A 1:5 dilution was made in culture media and 100 µl of culture (original and 1:5 dilution) was spread evenly on LB-agar plates (100 µg/ml ampicillin).

19. Non-de-phosphorylation and no-ligase controls were also transformed to the competent cells.

20. Plates were incubated at 37°C overnight.

21. Total number of colonies was counted for each plate after overnight incubation. The digestion and de-phosphorylation efficiencies of the vector were calculated as illustrated beneath.



Digestion efficiency = (1-z/x) × 100%
De-phosphorylation efficiency = [1-(y-z)/x] × 100%
x > a1 or a2 > y

2.15.4 PCR screening for detecting target insert

PCR screening was performed to select the plasmid constructs containing the target insert from the colonies after transformation.

22. A 96-well plate was set up with 200 µl of LB median (100 µg/ml ampicillin) per well.

23. The colonies were picked up using P-10 tips, and were transferred into 96-well plate accordingly.

24. The plate was incubated overnight at 37°C.

25. The PCR reaction mixture was prepared as follows:

10 × PCR buffer	5.0 µl
dNTP	1.0 µl
Ampli Taq	1.0 µl
DNA (bacteria culture)	5.0 µl
Primers	2.0 µl
PCR H_2O	36.0µl
Total	50.0 µl

26. The PCR was carried out as follows: initial denaturation for 5 minutes at 95°C; 30 cycles of 30 seconds at 95°C, 1 minute at 60°C, 1 minute at 72°C; and 5 minutes at 72°C.

27. The PCR products were visualised on 2% agarose gel, and the wells in 96-well plates which gave PCR bands of the correct size were taken further for plasmid preparation.

2.15.5 Alkaline lysis "mini-prep" Preparation

28. All the clones were grown in 10 ml cultures with 2 × LB media at 37ºC overnight supplemented with 100 µg/ml ampicillin.

29. The cultures were then transferred to new 50 ml tubes and centrifuged at 2080 g for 10 minutes at room temperature. The supernatant was poured off and 200 µl of lysis buffer was added to the pellets to lyse the cells.

30. The solution was transferred to 1.5 ml microfuges and 400 µl of freshly made 0.2M NaOH/1% SDS solution was added to each sample. The samples were incubated on ice for 10 minutes.

31. 300 µl of 3M NaAc (pH 5.2) was added to each sample and incubated on ice for 10-30 minutes.

32. The samples were centrifuged at 18000 g for 5 minutes at room temperature and the supernatant was transferred to fresh 1.5 ml microfuges and the precipitates were discarded. This step was repeated 3-4 times till the supernatant was clear.

33. 600 µl of isopropanol was added to the clear supernatant and incubated at -70ºC for 10 minutes.

34. The tubes were centrifuged at 18000 g for 5 minutes and the pellets were re-suspended in 200 µl of 0.3M NaAc, pH 7.0.

35. 200 µl of water saturated phenol/chloroform was added to the samples and vortexed briefly. Samples were centrifuged at 18000 g for 3 minutes. The aqueous (top) layer was collected in fresh 1.5 ml tubes. This step was repeated twice.

36. The DNA was precipitated from the aqueous layer by adding 200 µl isopropanol and incubating the solutions at -70ºC for 10 minutes.

37. The solutions were centrifuged at 18000 g for 5 minutes and the pellets obtained were washed with 70% ice-cold ethanol.

38. The pellets were air dried at 37ºC and finally resuspended in 50 µl of TE containing 200 µg/ml RNase A. After a quick vortex and a quick centrifuge, the tubes were incubated at 55ºC for 15 minutes in a water-bath.

39. 1 µl of each sample was analyzed on a 1% agarose TBE gel and stained with Safestain (Invitrogen) for visualization. The plasmid DNA was temporarily stored at -20ºC.

2.15.6 Plasmid Validation

40. The plasmid DNA was digested by BamHI enzyme at 37ºC for 1 hour.

41. The digestion mixture was set up as follows:

10 × NEB buffer 4	5.0 µl
BamHI-HF	1.0 µl
Plasmid DNA (100 ng)	Variable
H_2O	Variable
Total	50.0 µl

42. The digested samples were visualised on 2% agarose gel to check the release of the insert fragments.

2.15.7 Transfection and quantification of enhancer trap constructs

43. Aliquot of 5 × 10^6 K562 cells was suspended in 100 µl RMPI.1640 media.

44. 4 μg of luciferase reporter construct and 1 μg of a control plasmid expressing β-galactosidase were added for each transfection.

45. Transfection was performed in Gene Pulsar 0.2 cm cuvettes (Bio-Rad) with Nucleofection (Amaxa, protocol T-016).

46. Each transfected sample was cultured in 10 ml culture media for 48 hours.

47. Cells of each sample were harvested and cell lysates were prepared with 100 μl of 1 × lysis buffer (Promega).

48. 50 μl of cell lysates was pipetted into wells of a white-bottom 96-well plate for the Steady-Glo® Luciferase Assay System (Promega) and another 50 μl of cell lysates was pipetted into wells of an optical-bottom 96-well plate for the β-Galactosidase Enzyme Assay System (Promega).

49. 50 μl of Steady-Glo assay reagent was added to each well and Luminoskan Ascent micro-plate luminometer (Thermo Scientific) was used to measure luciferase activity.

50. 50 μl of Assay 2 × Buffer was added to each well.

51. Samples were mixed by pipetting and incubated at 37ºC for 30 minutes.

52. The reactions were stopped by adding 150 μl of 1M sodium carbonate.

53. The β-galactosidase activity was measured by reading the absorbance at 420 nm.

54. Relative enhancer activities were determined as the ratio of luciferase activity to β-galactosidase activity.

2.16 Sequence Analysis

2.16.1 Sequence alignments and TF binding sites

The multi-species sequence alignments of TAL1 and LYL1 regulatory regions found in human and other vertebral species were obtained from the UCSC Genome Browser (http://www.genome.ucsc.edu/).

TF binding sites including GATA, E-box and Ets at the TAL1 and LYL1 loci were identified using TFSEARCH (http://www.cbrc.jp/research/db/TFSEARCH.html) and TESS (http://www.cbil.upenn.edu/cgi- bin/tess/tess).

2.16.2 Public datasets

ChIP-seq datasets (ENCODE project) in human and mouse were acquired and visualised in UCSC Genome Browser (http://www.genome.ucsc.edu/).

2.17 Statistical nalysis

2.17.1 Student T-test

The student T-test was used for comparing the difference between the means of enrichments of the control region and the target region in 3C analyses. The two-sample heteroscedastic test was performed using two-tailed distribution in Microsoft 'Excel'. It was calculated using the TTEST function, and a p-value cut-off of 0.01 was used to determine a significant interaction in the 3C assay.

2.17.2 Other statistical tests

Other statistical analyses of data including the calculation of standard deviations and Pearson correlation coefficients were conducted using Microsoft 'Excel' in this thesis. Standard deviations were calculated using the STDEV function, Pearson correlation coefficients were calculated using the CORREL function.

Chapter 3: Identification of looping interactions at the TAL1 loci in human and murine cells by 3C

Summary

First, RT-qPCR was performed to assess the expression of TAL1 and its neighbouring genes in human and murine erythroid and lymphoid cells. Second, ChIP-chip of H3K4me3 modification was performed to determine the chromatin landscape of the TAL1 loci and transcriptional activity of its regulatory elements in these cell lines. It was observed that the epigenetic profiles at the promoters agreed with the gene expression patterns in these cell lines. In addition, it was identified that the enhancers of TAL1 were only active in the TAL1 expressing erythroid lineages. Third, 3C was used to study chromatin looping interactions between *cis*-acting regulatory elements across the TAL1 loci in human and murine cells. A number of control experiments were conducted to avoid misinterpretation of the 3C data. Looping interactions between the TAL1 promoter and the erythroid enhancers (+51/+40) were observed only in the erythroid cells, while interactions between the TAL1 promoter and the stem cell enhancers (+19/+18) were identified in both the erythroid and lymphoid cells. Based on the interaction profiles captured by the 3C, structural models were proposed for putative chromatin configurations of the TAL1 locus in the erythroid and lymphoid cell lines.

3.1 Introduction

3.1.1 Enhancer-promoter interactions and its roles in gene function

3.1.1.1 Theoretical models of enhancer-promoter communication

The physical contacts between the promoter and its enhancers are considered as a crucial step for transcriptional initiation (Maston et al., 2006). However, how an enhancer can activate its cognate genes across a distance of up to several hundred kilo-bases has long been disputed. To date, four different hypotheses

including “linking model”, “tracking model”, “looping model” and “facilitated tracking model” have been proposed to describe how the distal enhancers may be in contact with its promoter (shown in Figure 3.1).

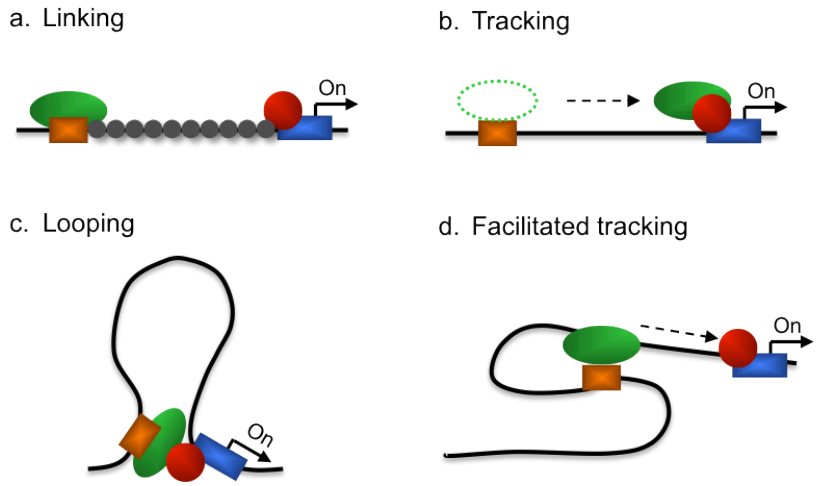


Figure 3.1: The four different models show a distal enhancer activating its cognate gene over a long distance. The enhancer is represented in the orange rectangle and the activation complex recruited by the enhancer is shown in the green oval. The genes are represented in the blue rectangle and the complex formed on gene promoter is shown in the red circle. The linking proteins are shown in the small grey circles.

The linking model (Figure 3.1 a) proposes that the binding of facilitator proteins links an enhancer and its cognate gene to mediate transcription activation (Bulger and Groudine, 1999). The tracking model (Figure 3.1 b) hypothesizes that the transcription-activating complex recruited at an enhancer is able to track along chromatin fibre until encountering the promoter of cognate gene, but without altering the spatial proximity between the enhancer and the promoter (Tuan et al., 1992). The looping model (Figure 3.1 c) suggests that a chromatin loop is formed between the enhancer and promoters, which brings together all the elements required for transcription activation (Ptashne and Gann, 1997). The facilitated tracking model (Figure 3.1 d) incorporates aspects of both the tracking and looping models, proposing that the enhancer together with its activation complex move along the chromatin until they reaches the cognate promoter (Blackwood and Kadonaga, 1998). The enhancer and the promoter are brought into a spatial proximity by the chromatin fibre being pulled through. Despite the different models of the enhancer-promoter interaction which are proposed, a number of recent studies by 3C have shown that *cis*-acting regulatory elements can be brought closer together via chromatin loops (Drissen et al., 2004; Murrell et al., 2004; Splinter et al., 2006; Vakoc et al., 2005).

3.1.1.2 Enhancer-promoter cross-talk in the TAL1 locus

The TAL1 locus is a well-defined genomic region with numbers of *cis*-regulatory elements which have been identified using different approaches, including enhancer trap assays, transgenic mice analyses, sequence conservation analyses, array-based DNaseI hypersensitive site mapping (ADHM) as well as ChIP-chip for histone modifications and transcription factor bindings (Chapman et al., 2004; Dhami et al., 2010; Follows et al., 2006; Gottgens et al., 2002a; Gottgens et al., 2001; Ogilvy et al., 2007). It is known that these *cis*-regulatory elements participate in regulating transcription of the TAL1 gene at different developmental stages or in different cell lineages (Delabesse et al., 2005; Gottgens et al., 2010; Gottgens et al., 2002b; Sanchez et al., 1999; Silberstein et al., 2005; Smith et al., 2008). However, there is very limited knowledge about how these regulatory elements communicate with each other to modulate TAL1 expression in blood lineages.

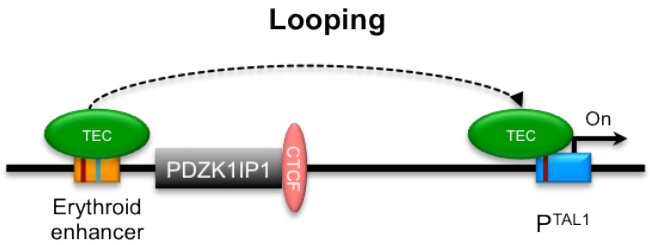


Figure 3.2: A schematic map of the erythroid enhancer activating the TAL1 gene via looping in erythroid cells. The *PDZK1IP1* gene (grey rectangle) and a CTCF insulator (pink oval) located in the middle of the erythroid enhancer (orange rectangle) and the TAL1 promoter (blue rectangle). The TAL1 erythroid complex (TEC) is represented in green ovals. The GATA and E-box binding sites are represented in dark red and light blue lines, respectively. The dashed arrow illustrates the movement of the transcription-activating complex from the enhancer to the promoter via looping.

Several lines of evidence imply that chromatin looping may be required for TAL1 expression in human and murine erythroid cells (shown in Figure 3.2). First, a CTCF binding site (+40) previously identified was located at the middle of the erythroid enhancer (+51) and the TAL1 promoters (1a/1b). This CTCF site may act as an enhancer-blocking insulator which prevents the erythroid enhancer from activating the TAL1 promoter (Bell et al., 1999). In addition, the PDZK1IP1 gene is also located right in between the TAL1 enhancer and promoter. Both of the CTCF insulator and PDZK1IP1 may act as physical barriers which prevent the enhancer-promoter communication via tracking or linking models. In order to deliver the transcriptional activating complex recruited at the erythroid enhancer to the TAL1 promoter, chromatin looping is the most likely way for them to overcome these

obstacles. Second, the GATA/E-box binding motifs were found at the erythroid enhancer as well as the GATA binding sites at the TAL1 promoter (Bockamp et al., 1995; Gottgens et al., 2010). Third, the GATA1 transcription factors and other members of the TAL1-containing erythroid complex (TEC) including TAL1, E47, LDB1 and LMO2 have been shown to bind at both the TAL1 promoter and the +51/+40 erythroid enhancer *in vivo* in both human and murine erythroid cells (Dhami et al., 2010; Ogilvy et al., 2007). Fourth, RNA polymerase II has been shown to bind to both TAL1 promoters and its enhancers (the +51, +20 and -10 enhancers) in human erythroid cells, which may also indicate that there is physical contact between these elements to regulate TAL1 expression (Dhami et al., 2010). Based on this circumstantial evidence, it is likely that looping interactions exist at the TAL1 locus in erythroid lineages. In addition, it is known that the TAL1 *cis*-acting elements are distrusted with a span of ~100 kb in the human genome. The size of the TAL1 locus makes it a good experimental paradigm for studying chromatin interactions between regulatory elements using 3C technology.

3.2 Aims

Although a number of TAL1 *cis*-acting elements have long been identified and characterised for their regulatory functions, little is known about how these regulatory elements communicating with each other and subsequently controlling the transcription of TAL1. As described in Chapter 1, 3C technology is a robust and powerful tool for mapping specific chromatin interactions between *cis*-acting regulatory elements with prior knowledge of their genomic positions. Thus, it was considered as a suitable method for assessing chromatin interactions at the TAL1 locus. The aims of this chapter were:

1. To characterise the transcriptional and chromatin status of the TAL1 locus in selected human and mouse cell lines

2. To establish 3C-PCR assays for characterising putative looping interactions at the human and murine TAL1 loci.

3. To characterise promoter-enhancer looping interactions of the TAL1 locus in human and murine TAL1-expressing and non-expressing cells.

3.3 Overall strategy

First, the transcriptional state and chromatin landscape were determined by RT-qPCR and ChIP-chip analysis in human and mouse erythroid and lymphoid cell lines (Figure 3.3, left panel). Second, the 3C method was established with numbers of control experiments (Figure 3.3, middle panel), including i) the restriction digestion efficiency of 3C chromatin DNA was assessed by q-PCR; ii) the ligation of 3C library was assessed by gel electrophoresis and the ERCC3 control PCR; iii) the DNA concentration of PCR template was titrated and optimised for a relatively quantitative detection by PCR and gel electrophoresis; iv) the 3C control templates (BAC/PAC) which covered the TAL1 loci were prepared to normalise the differences in primer efficiency between individual primer combinations. Third, looping interactions between the TAL1 promoter and its three enhancers were assessed by the 3C (Figure 3.3, right panel). The TAL1 promoter was selected as the anchor of the 3C. The erythroid enhancer (TAL1 +51/ Tal1 +40), the stem cell enhancer (TAL1 +19/ Tal1 +18) and the TAL1 -10/ Tal1 -9 enhancer were chosen as the target sites of the 3C analysis. In addition, another five flanking regions of these enhancers were located at +64, +46, +5, -8, -41 in human and +55, +30, +15, -5, -28 in mouse (numbers represent the distance in kb from TAL1 promoter 1a) were selected as the control sites. The locations of PCR primers designed for the anchor, three enhancers, and control regions were illustrated in Figure 3.4.

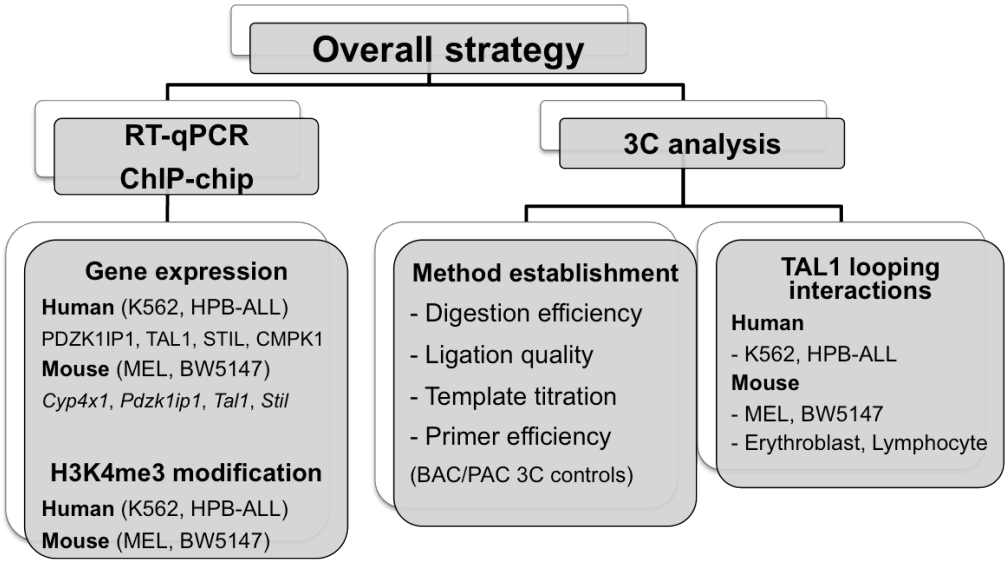


Figure 3.3: Schematics map of overall strategy of Chapter 3.

The 3C analysis was performed on four human and murine cell lines as well as murine primary erythroid (TAL1 expressing) and lymphoid (TAL1 non-expressing) cells. In this chapter, all experiments were performed with two biological replicates for each cell type. In addition, the q-PCR and 3C analysis was conducted with three technical replicates.

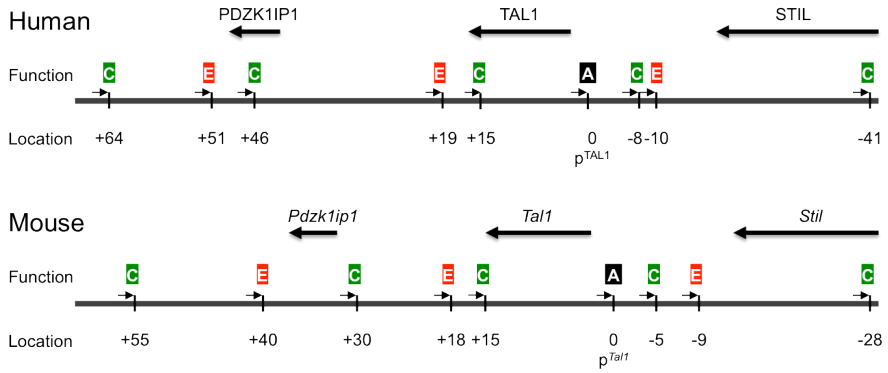


Figure 3.4 Design of 3C primers across the TAL1 loci. Human and mouse TAL1 loci are shown to illustrate the location of 3C primers used in primer combinations with the anchor (P^{TAL1}). Genes across the TAL1 locus are shown with coloured arrows (denoting direction of transcription). Colour-coded boxes with arrows indicating the position and function of the 3C primers: C = control, E = enhancer, A = anchor. 3C primers are named based on their distance upstream (-) or downstream (+) in kilobases from the TAL1 promoter 1a.

Results

3.4 Characterisation of human and murine cell lines

3.4.1 Characterising gene expression across the TAL1 loci

The expression level of TAL1 and its neighbouring genes were characterised in the erythroid and lymphoid cell lines. The mRNA extracted from two bio-replicates of each cell line was reverse-transcribed to generate the cDNA libraries. Subsequently, the expression of TAL1 and its neighbouring genes was assessed using quantitative-PCR (q-PCR). The relative expression level was calculated by normalising the Ct value of target genes against the housekeeping genes (ACTB for human and Gapdh for mouse).

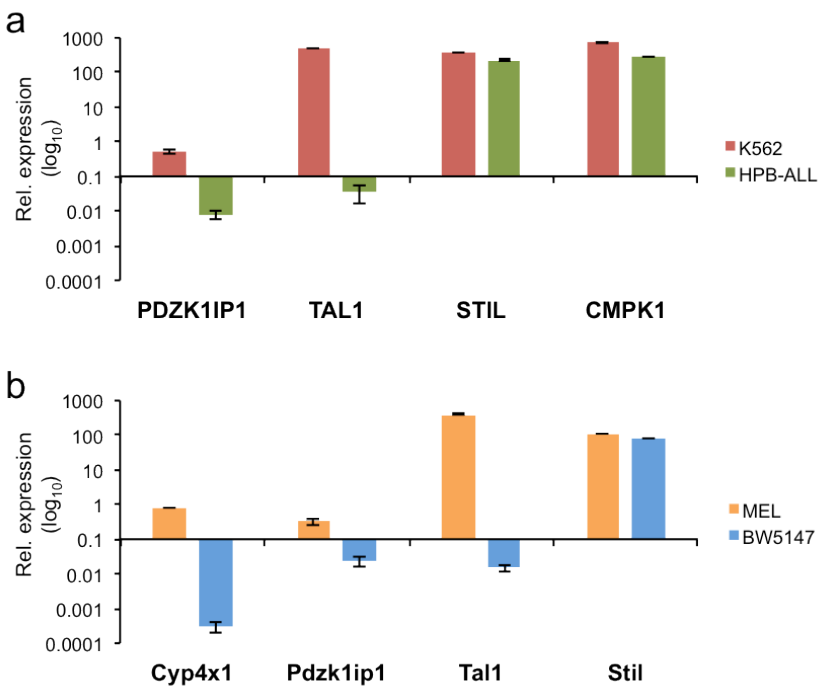


Figure 3.5: Gene expression profiles across the TAL1 loci. A. Human K562 and HPB-ALL cell lines and B. Murine MEL and BW5147 cell lines. The x-axis represents the genes within the TAL1 loci and the y-axis represents the relative expression level (log_{10} scaled) after being normalised against the housekeeping genes. Error bars represent the stand error calculated for two bio-replicates.

Expression level of PDZK1IP1, TAL1, STIL and CMPK1 was determined in human K562 and HPB-ALL cells. As illustrated in Figure 3.5 a, PDZK1IP1 expressed at a lower level (less than 1) in K562 cells and was inactive in HPB-ALL cells (less than 0.01). High-level of TAL1 expression was observed in K562 but not HPB-ALL cells. As it was reported, the TAL1 gene was expressed in erythroid but not lymphoid lineages (Delabesse et al., 2005). Adequate expressions of STIL and CMPK1 were identified in both K562 and HPB-ALL cells.

Similarly, expression of *Cyp4x1*, *Pdzk1ip1*, *Tal1* and *Stil* was assessed in murine erythroid and lymphoid cell lines (Figure 3.5 b). Low-level of *Cyp4x1* and *Pdzk1ip1* expressions were observed in the MEL cells. Similar to the human K562 cells, significant *Tal1* expression was only detected in the erythroid lineage (MEL). No expression of *Cyp4x1*, *Pdzk1ip1*, *Tal1* was observed in the lymphoid BW5147 cells. *Stil* was the only gene with high-level of expression in both MEL and BW5147 cells.

3.4.2 Determination of the active chromatin landscape of human and murine cell lines by ChIP-chip assays

ChIP analysis of histone modifications across genomic regions is a robust method of identifying *cis*-regulatory elements and assessing their activity *in vivo* (described in Chapter 1, section 1.2.2). Histone modifications have previously been used extensively to characterise the TAL1 locus in human and mouse cells, providing evidence of the location and activity of regulatory elements known to drive TAL1 expression (Dhami et al., 2010; Spensberger et al., 2012). The histone H3 Lys4 (K4) methylation is a conserved marker for transcriptionally active regions which is primarily found to be associated with the 5' region of genes (Ng et al, 2003), is directly related to mRNA expression levels and is also associated with active enhancer elements (Pekowska et al., 2011). ChIP-chip assays for H3K4me3 modification were performed on selected human and murine cell lines in this thesis. In combination with the mRNA expression analysis, it was to provide confirmation of transcriptional states of genes at the TAL1 loci. Most importantly, it was to investigate chromatin landscape of the TAL1 locus in selected human cell lines and their murine equivalent. The ChIP-chip profiles would not only reveal the distribution of *cis*-acting elements across the TAL loci, but also provide the additional information about the transcriptional activity of both known regulatory elements and newly identified novel regions. Consequently, it was important to determine the transcriptional activities of cis-acting elements within the TAL1 loci before studying the chromatin interactions and configurations between these elements using the 3C analysis.

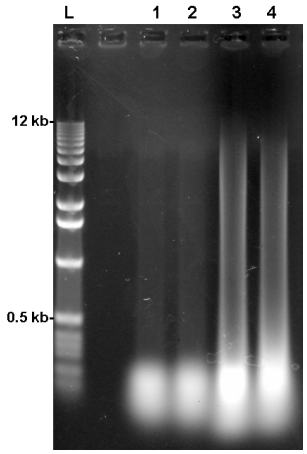


Figure 3.6 Gel electrophoretic analysis of ChIP DNAs. Lane L = DNA ladder, lane 1 & 2 = ChIP DNAs isolated using H3K4me3 antibody for two biological replicates of K562 cells, lane 3 & 4 = input DNAs (sheared purified total genomic DNA). Bright bands at the bottom of

gels are the yeast tRNA which was used as carrier to aid in precipitation of the ChIP DNAs during purification. Faint smears for ChIP DNA can be observed for H3K4me3. The samples were visualised by electrophoresis on 1% agarose gels with SYBR Safe stain.

For the ChIP-chip analysis, the chromatin was first fixed with formaldehyde, followed by fragmented by sonication. The fixed chromatin fragments were immunopreciptated by using antibody against H3K4me3 and were visualised by agarose gel electrophoresis along with the input DNAs. For each cell lines, ChIP-chip analysis was performed with two bio-replicates. A gel image of two bio-replicates of K562 ChIP and input DNAs is displayed as an example (shown in Figure 3.6). Subsequently, the ChIP and input DNAs were labelled by Cy3 and Cy5 dyes and prepared for microarray hybridisation. The TAL1 genomic tiling path arrays (see Chapter 1, section 1.4.6) were used to assess the H3K4me3 modification of the TAL1 loci in four of selected human and murine cell lines. A scanned composite image of H3K4me3 ChIP profile in one of the K562 bio-replicate is presented as an example (Figure 3.7). The green spots on the image show enrichments in the H3K4me3 ChIP sample as compared to the input DNA. The yellow spots represent equal hybridisation of the ChIP and input DNA to the spot. Orange/red spots show regions which are under-represented in the ChIP sample. The white spots reflect saturated spots in the ChIP sample.

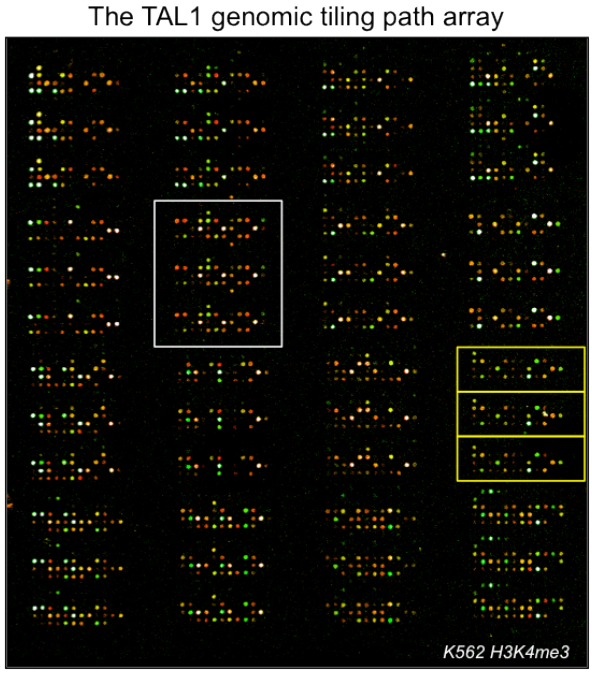


Figure 3.7 A composite image of the TAL1 genomic tiling array. The array was hybridised with H3K4me3 ChIP DNA in cell line K562 along with K562 input DNA. Each spot on the array represents an array element. Each array element was spotted in triplicate (a triplicate set of array elements is shown within the three yellow boxes) in a 16 sub-grid format (a single sub-grid is shown within the white box).

The K562 cells used in this thesis were at relatively later passage stocks made by Dr. P.Dhami. In order to validate this batch of cells acting as the early passage, ChIP-chip assays were performed with H3K4me3 antibody on two K562 bio-replicates as illustrated beneath (Figure 3.8). Significant enrichments of H3K4me3 were observed over numbers of *cis*-acting elements across the TAL1 locus (Figure 3.8). The highest peaks were found at the promoter regions of the TAL1 gene, including the promoter 1a, 1b and TAL1 +3/+1. The high H3K4me3 enrichments were also observed at the STIL and CMPK1 promoters and the TAL1 +53 promoter. In addition, significant enrichment was observed at the TAL1 +51 enhancer. Moreover, three minor peaks were identified at the TAL1 +21/20/19 and TAL1 -10 enhancers as well as the PDZK1IP1 promoter. The H3K4me3 profiles are not only highly similar between two bio-replicates, but also perfectly reassembled to those ChIP-chip profiles obtained previously (Dhami et al., 2010). In addition, the previous ChIP-chip studies in K562 cells also observed high levels of H3K4me1/2 and low levels of H3K27me3 modifications at these regulatory elements enriched for H3K4me3 modifications, reflecting the active chromatin states at these regions.

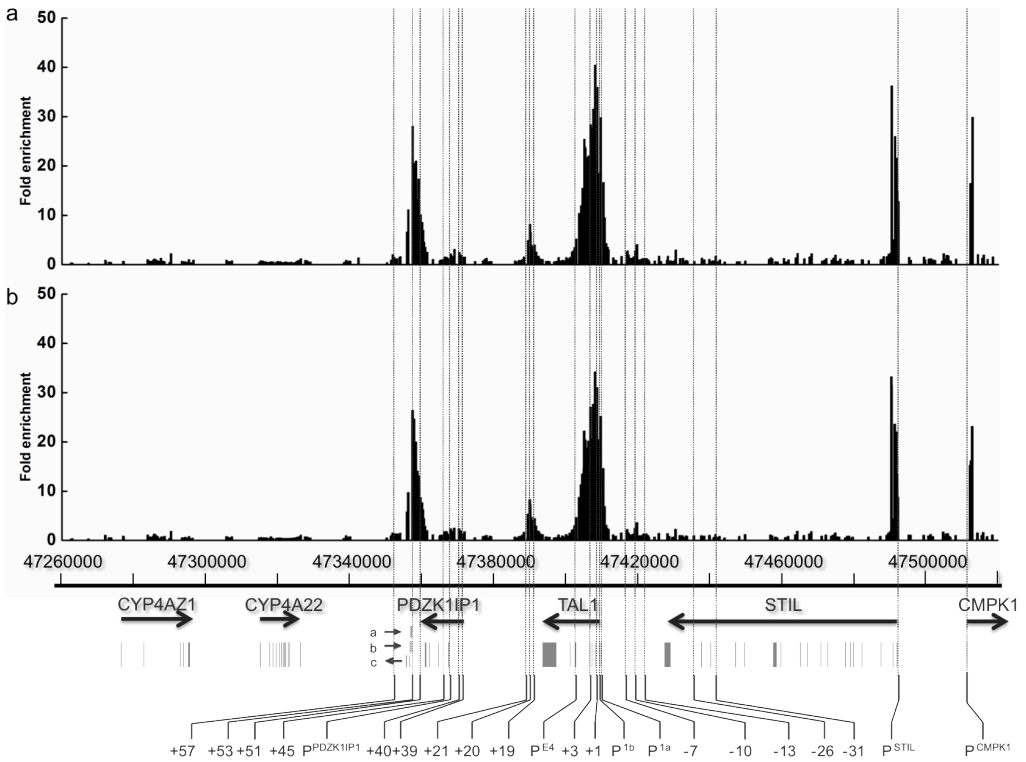


Figure 3.8 Histone H3K4me3 modification profiles across the human TAL1 locus in the K562 cells. Panel a & b correspond to two bio-replicates. Histograms represent the data obtained for each genomic tiling array element. In each panel, the x-axis is the genomic sequence co-ordinate (NCBI build 35) and the y-axis is the enrichment obtained in ChIP-chip assays. Schematic diagram at the bottom of the figure shows the genomic organisation of TAL1 and its neighbouring genes. Exons are shown as vertical blocks with gene names and direction

of transcription shown above. Transcripts denoted by a,b and c refer to transcripts of unknown function (Dhami et al., 2010). Vertical lines at the bottom (with dotted lines through all the panels) show the location of known and novel regulatory regions at the TAL1 locus. Promoters are denoted by P. Other nomenclature refers to the distance in kb from TAL1 promoter 1a (P^{1a}).

The ChIP-chip assays of H3K4me3 were also performed on the HPB-ALL cells. The combined profiles of two bio-replicates in the HPB-ALL and K562 cells are illustrated beneath (Figure 3.9). In HPB-ALL cells, high-level enrichments were identified at the STIL and CMPK1 promoters (Figure 3.9 a). Comparing with K562 cells, the H3K4me3 peaks were disappeared from the entire promoter region (TAL1 +3/1 regions, promoter 1a and 1b) and enhancers (TAL1 +51, +21/20/19, -10) of the TAL1 gene in HPB-ALL cells. The ChIP-chip profile of HBP-ALL agrees with its mRNA expression pattern, as high-level expressions of STIL and CMPK1 but not TAL1 and PDZK1IP1 were observed in HPB-ALL cells (shown in section 3.4.1). Additionally, significant H3K4me3 enrichments were observed at the TAL1 +53 region in both cell lines. The TAL1 +53 was identified as an active region in all cell types (K562, HPB-ALL, HL-60 and Jurkat) in previous studies (Dhami's PhD thesis, University of Cambridge 2005) and was proven to have promoter activity in both K562 and HPB-ALL cells (Dhami et al., 2010).

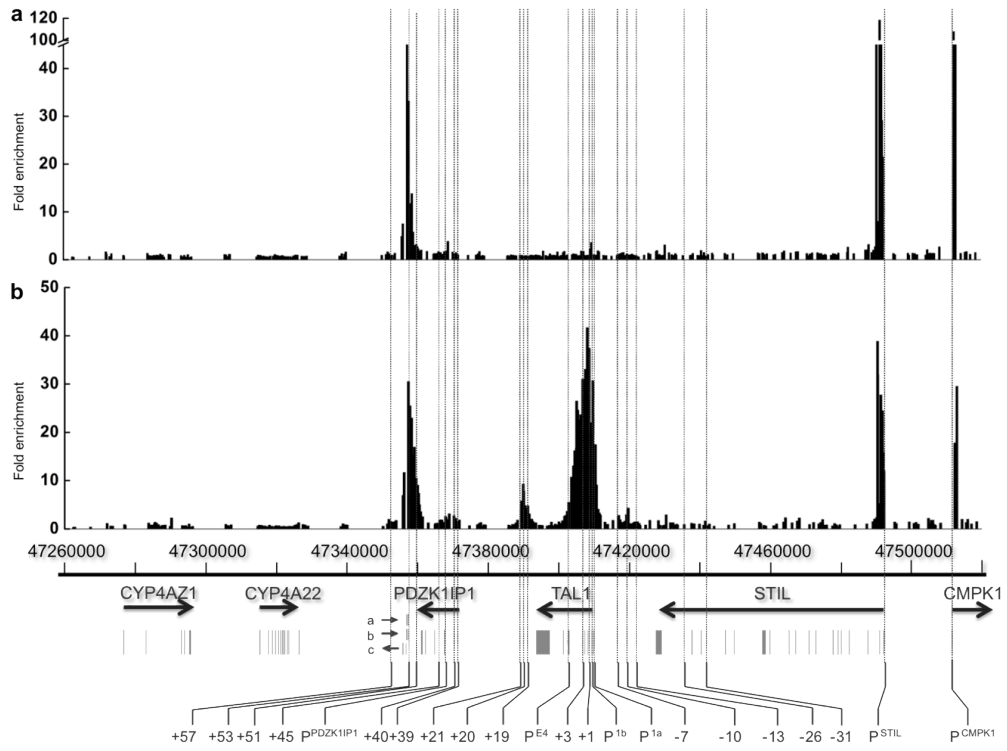


Figure 3.9 Histone H3K4me3 modification profiles across the human TAL1 locus. Panel A: HPB-ALL (TAL1 non-expressing) and panel B: K562 (TAL1 expressing). Histograms represent the data obtained for each genomic tiling array element. In each panel, the x-axis is the genomic sequence co-ordinate (NCBI build 35) and the y-axis is the enrichment obtained in ChIP-chip assays. All other aspects of the figure are as in Figure 3.8.

The ChIP-chip assays were also conducted in murine MEL (*Tal1* expressing) and BW5147 (*Tal1* non-expressing) cells to histone modification of H3K4me3. The analysis was performed in two bio-replicates for each cell type and combined profiles were illustrated in Figure 3.10. In MEL cells, a significant peak was identified at the *Stil* promoter as well as a moderate peak at the *Cyp4x1* promoter (Figure 3.10b). In addition, high-level enrichments were also observed at the TAL1 promoter region (*Tal1* +3/1, promoter 1a and 1b), the *Tal1* +43 region (the murine equivalent of the TAL1 +53 promoter) as well as at the *Tal1* +40, +19/18 and -8/9 enhancers. In contrast, H3K4me3 peaks were only identified at the *Stil* promoter and *Tal1* +43 promoter (a minor peak) in the BW5147 cells (Figure 3.10a). Similar to the human equivalent, the peaks were absent at the entire promoter region and enhancers of *Tal1* in the BW5147 cells comparing to the MEL cells.

These observations of H3K4me3 modification perfectly line up with the gene expression patterns in two mouse cell lines, as the *Cyp4x1* (lower level), *Tal1* and *Stil* genes were expressed in the MEL cells, while only expression of the *Stil* gene was observed in the BW5147 cells.

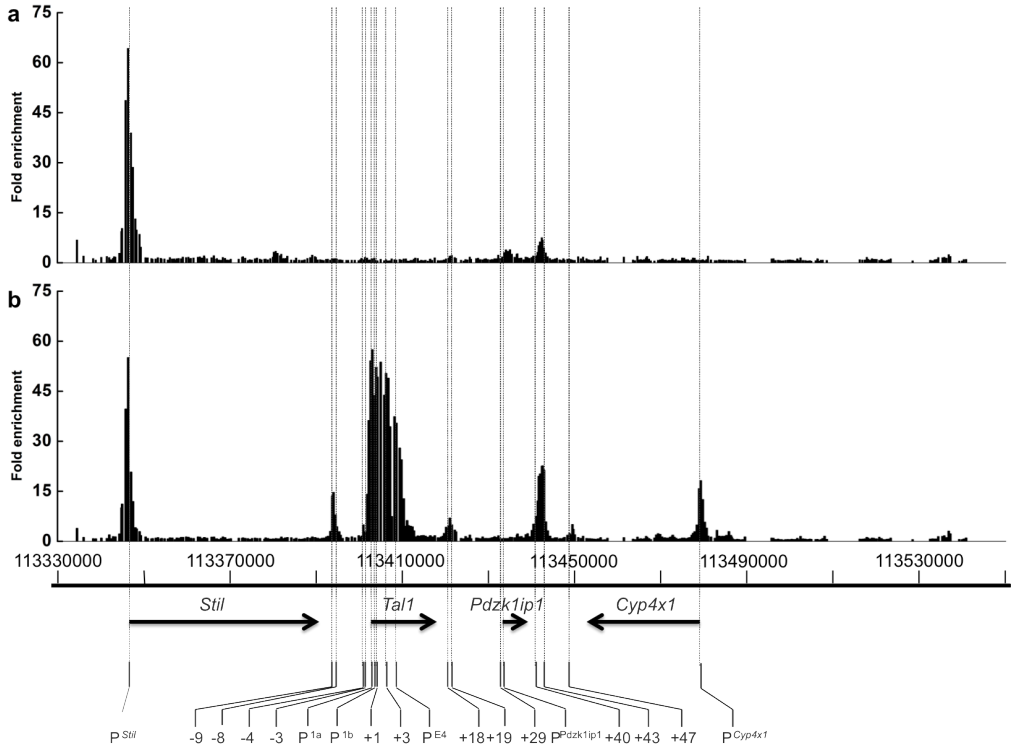


Figure 3.10 Histone H3K4me3 modification profiles across the murine *Tal1* locus. Panel A: BW5147 (*Tal1* non-expressing) and panel B: MEL (*Tal1* expressing). Histograms represent the data obtained for each genomic tiling array element. In each panel, the x-axis is the genomic sequence co-ordinate (NCBI build 35) and the y-axis is the enrichment obtained in ChIP-chip assays. All other aspects of the figure are as in Figure 3.8.

Firstly, the ChIP enrichments of H3K4me3 at the promoters across the TAL1 loci corresponding to the mRNA expression patterns illustrated at section 3.4.1. Secondly, with respects to the transcriptional state of the TAL1 loci, the profiles of H3K4me3 modification displayed high similarities between human and murine cell lines (Figure 3.9 and 3.10). To sum up, the H3K4me3 modification marked numbers of promoters and enhancers across the TAL1 loci in human K562 and murine MEL cells, indicating that these *cis*-acting elements are active in the TAL1 expressing cell lines (Table 3.1). Although additional ChIP analyses on histone H3K27me3, H3K4me1/2 and H3K27Ac modifications in MEL cells could provide further evidence for active chromatin states of these *cis*-acting elements. In contrast, the promoter and enhancers of TAL1 were marked as inactive in human and murine TAL1 non-expressing lymphoid cell lines (Table 3.1).

Table 3.1: Transcriptional states of *cis*-acting elements across the TAL1 loci in human and murine cell lines.

Human	K562	HPB-ALL	Mouse	MEL	BW5147
TAL1 +53	active	active	*Tal1* +43	active	active
TAL1 +51	active	inactive	*Tal1* +40	active	inactive
TAL1 +21/20/19	active	inactive	*Tal1* +19/18	active	inactive
P^{TAL1}	active	inactive	P^{Tal1}	active	inactive
TAL1 -9/-10	active	inactive	*Tal1* -8/-9	active	inactive
P^{SIL}	active	active	P^{Stil}	active	active
P^{CMPK1}	active	active	P^{Cyp4x1}	active	inactive

3.5 Establishment of the 3C-PCR assay

3.5.1 Quality control of the 3C library

3.5.1.1 Assessment of digestion efficiency by q-PCR

In 3C, the DNA contained within the cross-linked chromatin is digested by a suitable restriction enzyme during library preparation. Unlike purified DNA, the DNA in cross-linked chromatin is complex with histones and other proteins. Thus, it is thought that digestion of this DNA substrate by restriction enzymes is less efficient and must be monitored carefully to ensure a high degree of digestion (Gondor et al., 2008). Insufficient digestion of DNA in chromatin has downstream consequence on the remaining steps of the 3C procedure and results in

insufficient ligation. This ultimately results in the inability to detect appreciable levels of ligation products by PCR for some or all primer pair combinations. As a consequence, it may lead to misinterpretation of DNA-DNA interactions that occur *in vivo* (Dekker, 2006). Thus, it is crucial to quantitatively assess the restriction digestion efficiency of chromatin DNA during preparation of the 3C.

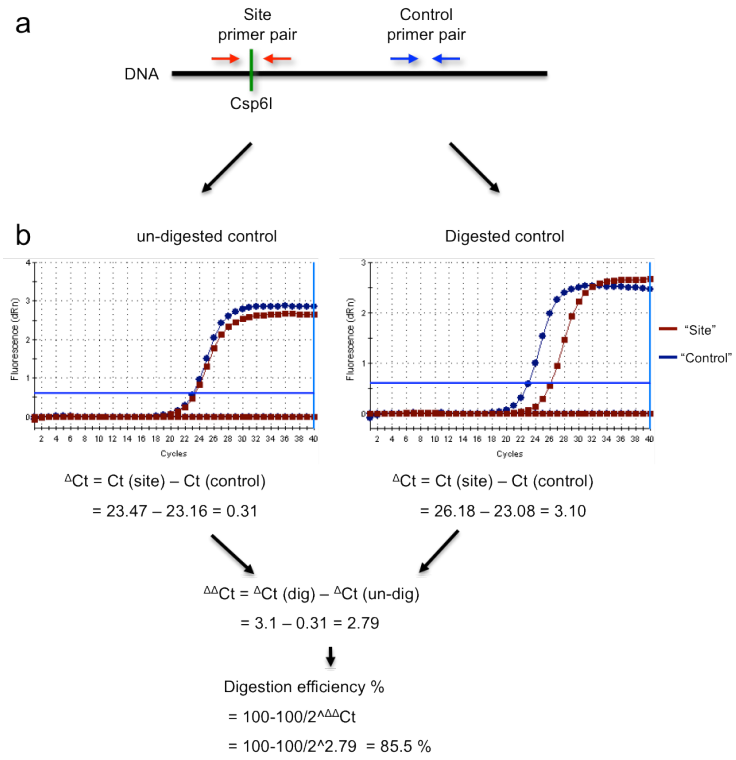


Figure 3.11: A schematic diagram of assessing digestion efficiency. The figure illustrates the assessment of digestion efficiency of the K562 bio-replicate 1. The "site" primer pair (red arrows) and "control" primer pair (blue arrows) are the two sets of primers used for this assay. The DNA is represented by a black line and the restriction site is represented by a green line. The amplification plots of q-PCR analysis are shown corresponding to the non-digested and digested controls using "site" and "control" primer pairs. The details of calculation are shown step by step at the bottom of the figure.

For this study, the efficiency of restriction enzyme digestion of the 3C library was assessed by q-PCR. Two sets of primers were designed for the q-PCR assays as shown in Figure 3.11a. The "site" primer pair was designed flanking a Csp6I restriction enzyme site. Theoretically, the digested chromatin DNA would yield less PCR product in comparison to the un-digested DNA when using the "site" primer pair. The "control" primer pair was designed to amplify a region within a Csp6I fragment. Thus, it would make no difference of the yield of PCR product between digested and un-digested chromatin DNA when using the "control" primer pair. Both of the primer pairs were designed for amplifying the similar sized (~100 bp) PCR product in order to minimise any biases generated during the PCR

amplification. During the 3C preparation, two small aliquots of chromatin DNA were retained for the control experiments. One aliquot was not subjected to digestion known as un-digested control. Another aliquot was retained after restriction digestion by Csp6I but was not subjected to DNA ligation, which was known as digested/no-ligase control. Both controls were used as the PCR templates for measuring digestion efficiency of the 3C library. The q-PCR analysis was performed using the "site" and "control" primer pairs to assess digestion efficiency of all 3C libraries presented in this thesis. The data of K562 "bio-replicate1" is used as an example (Figure 3.11b) to illustrate the procedure of determining digestion efficiency step by step from the q-PCR readings (Ct values). The $^{\Delta}$Ct values are calculated using the Ct value of "site" primer pair minuses the Ct value of "control" primer pair. The $^{\Delta\Delta}$Ct value between un-digested and digested controls is subsequently calculated. The formula used in this thesis is as follows: Digestion efficiency% = 100–100/2^((Ct_S–Ct_C)$_D$–(Ct_S–Ct_C)$_{UD}$) (Hagege et al., 2007).

Table 3.2: Digestion efficiency of the 3C libraries used for the work presented in this thesis. Two 3C biological replicates were prepared for all human and mouse cell types.

3C library	Digestion efficiency	3C library	Digestion efficiency
K562 Biorep1	85.5%	K562 Biorep2	88.4%
HPB-ALL Biorep1	72.9%	HPB-ALL Biorep2	75.3%
MEL Biorep1	88.2%	MEL Biorep2	80.8%
BW5147 Biorep1	81.2%	BW5147 Biorep2	85.5%
Erythrocyte Biorep1	90.5%	Erythrocyte Biorep2	91.2%
Lymphocyte Biorep1	92.1%	Lymphocyte Biorep2	83.4%

All of the 3C libraries generated for the current study were tested for digestion efficiency, and those libraries with digestion efficiency lower than 70% were discarded. The digestion efficiencies of the 3C libraries used in this thesis are shown in Table 3.2.

3.5.1.2 Assessment of ligation in the 3C libraries

Visualisation of restriction enzyme digestion and ligation

A direct comparison of the size of DNA fragments between digested/no-ligase controls (DNA purified from the libraries after digestion but prior to ligation) and 3C libraries (i.e., ligated controls - 3C libraries that had been both digested and ligated)

was visualised by agarose gel electrophoresis (Figure 3.12). For ligated controls, an intensely staining high molecular weight (>12 kb) DNA smear (ligated DNA stained by SYBR Safe) was observed on top of each lane. This smear was absent from the digested controls. The decreased electrophoretic mobility of DNA species in the ligated controls as compared to the digested controls was a reflection of ligation events in the 3C libraries (shown in Figure 3.12). All 3C libraries prepared for the present study were monitored qualitatively in this manner.

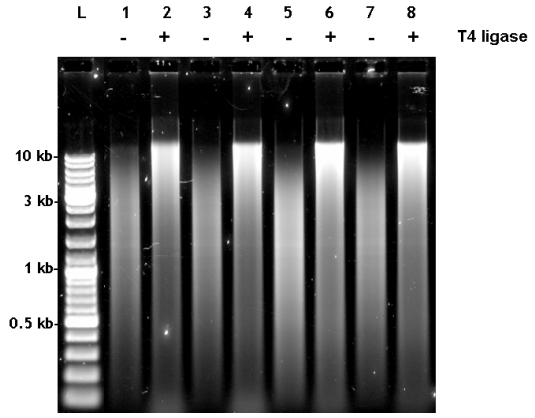


Figure 3.12 Visualisation of the digestion and ligation profiles by gel electrophoresis. Lane L = DNA ladder for size reference; Lane 1, 3, 5 and 7 = 3C digested controls and lane 2, 4, 6 and 8 = 3C libraries (ligated controls); "+" and "-" = with or without T4 ligase. The size of DNA markers are labelled on the left side of the image.

Determination of 3C ligation efficiency using the ERCC3 gene

The quality of the 3C ligation was also assessed by PCR in detecting a ligation event between two restriction fragments of the ERCC3 gene. It is known that the 3C library is a mixed population of ligated restriction fragments – with the frequency of ligation products reflecting either random or functional interactions based on the *in vivo* spatial proximity of the ligation partners. A specific 3C primer pair was designed to detect a known ligation event between two fragments of the ERCC3 gene which are separated in several kilobases from each other in their genomic locations (shown in Figure 3.13A). The PCR amplification was performed on digestion/no-ligase controls and 3C libraries. As the schematic map shown in Figure 3.13A, the specific ERCC3 fragment can be PCR amplified only in the 3C library but not in the digestion (non-ligated) control. The sizes of PCR products from the various templates were visualised by agarose gel electrophoresis. The bands representing the ligation product of the ERCC3 gene (~150 bp) were only observed in 3C libraries but not in the digestion (non-ligated) controls (shown in Figure 3.13B), which are highlighted by yellow boxes. The detection of a known

ligation event of the ERCC3 gene also provides a second more quantitative measurement of the efficiency of the ligation reaction in generating 3C libraries. This control experiment was used to assess the ligation efficiency of all 3C libraries prepared for this study.

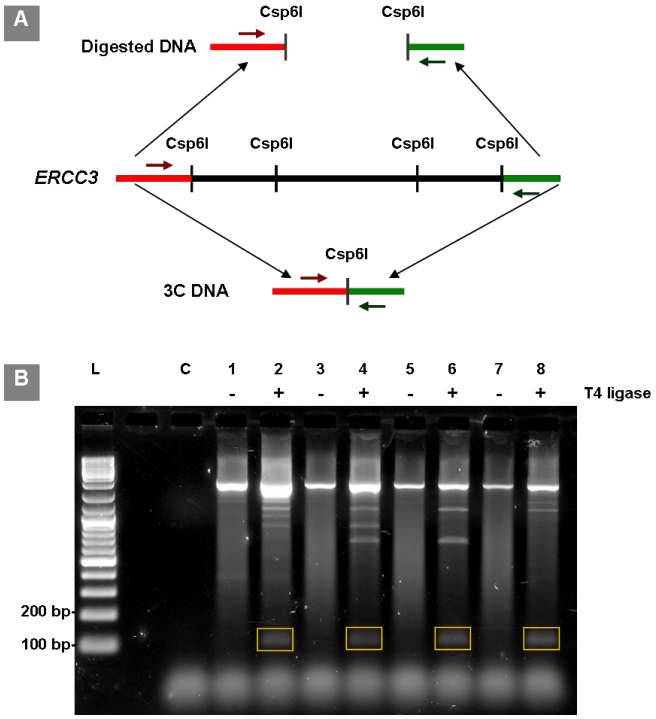


Figure 3.13 Detecting ligation products of the ERCC3 gene. A: schematic map of ERCC3 control PCR. Target fragments of PCR and primers are colour-coded in red and green. The Csp6I sites are shown in black bars across the ERCC3 gene. B: visualisation of the ERCC3 amplicon by gel electrophoresis. The 3C control PCR amplification was performed using 200 ng DNA of the 3C digested and ligated controls. Lane L = DNA ladder for size reference; lane C = water control; lane 1, 3, 5 & 7 = the 3C digested controls and lane 2, 4, 6 & 8 = the 3C ligated controls; "+" and "-" = with or without T4 ligase. The yellow boxes indicate the PCR amplicons of ERCC3. The size of DNA markers is labelled on the left side of the image.

3.5.2 Preparation of 3C control templates from BAC/PAC DNA

3.5.2.1 Identification of PAC and BAC clones for preparing 3C control templates

As discussed previously, for 3C analysis in organisms with larger genomes, such as mouse or human, the PCR efficiency of 3C primer pair combinations are calculated by utilizing 3C control templates made from BAC (YAC or PAC) clones covering the genomic segments under study (Dekker, 2006). Using online tools from the UCSC Genome Browser (http://genome.ucsc.edu/) and Ensembl (http://www.ensembl.org/index.html) websites, a 193 kb PAC clone (RP1-18D14) and a 196 kb BAC clone (RP23-453H14) were identified and subsequently

obtained which covered the human and mouse TAL1 loci respectively (shown in Figure 3.14).

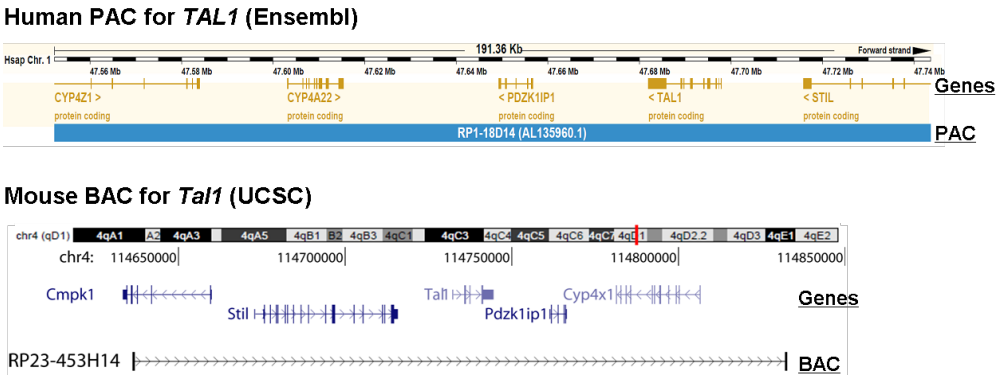


Figure 3.14 Schematic of the human PAC and mouse BAC used for the 3C control template. The genomic coordinates and genes covered by the human PAC and mouse BAC clones are shown on top of each panel. The PAC and BAC clones are shown in blue bar and arrow line, respectively.

Digested PAC and BAC DNA was ligated in a relatively high concentration of approximately 200 to 300 ng/µl (20 to 30 fold higher than the concentration for intra-molecular ligation in a mammalian 3C library) to obtain all possible ligation products between restriction fragments covering the TAL1 locus. A small proportion of digested and ligated 3C BAC and PAC control templates were visualised by gel electrophoresis (shown in Figure 3.15).

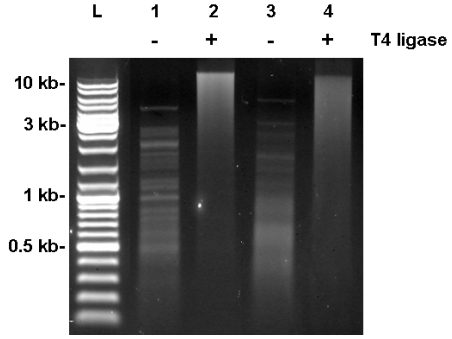


Figure 3.15 Visualisation of the digestion and ligation profiles of the 3C control templates by gel electrophoresis. Lane L = DNA ladder for size reference; Lanes 1 and 3 = 3C digested controls and lanes 2 and 4 = 3C ligated controls (3C control libraries); "+" and "-" = with or without T4 ligase. The size of DNA markers was labelled on the left side of the image.

Similar to the quality assessment of 3C libraries, the size distributions of non-ligated controls after restriction digestion were visualised on the gel and the size-shift between no-ligase controls and 3C control templates was observed indicating digestion and ligation events of 3C control templates.

3.5.2.2 Optimisation of the usage of 3C and BAC/PAC control templates in PCR amplification

The 3C interaction is detected by quantifying PCR products from gel images. As the ligation frequency can be extremely low between two distal elements, it is crucial to ensure that all PCR products, from both target (e.g. *cis*-acting elements) and control regions, can be visualized by agarose gel electrophoresis (Dekker, 2006). In addition, it is also important to optimize the amount of template for PCR reaction, in order to achieve relatively quantitative results. Thus, the DNA templates being used for PCR amplification, which were from either the 3C library or the BAC/PAC control template, were collaborated for their usage with a serial of titration experiments (discussed in Chapter 1, section 1.2.4).

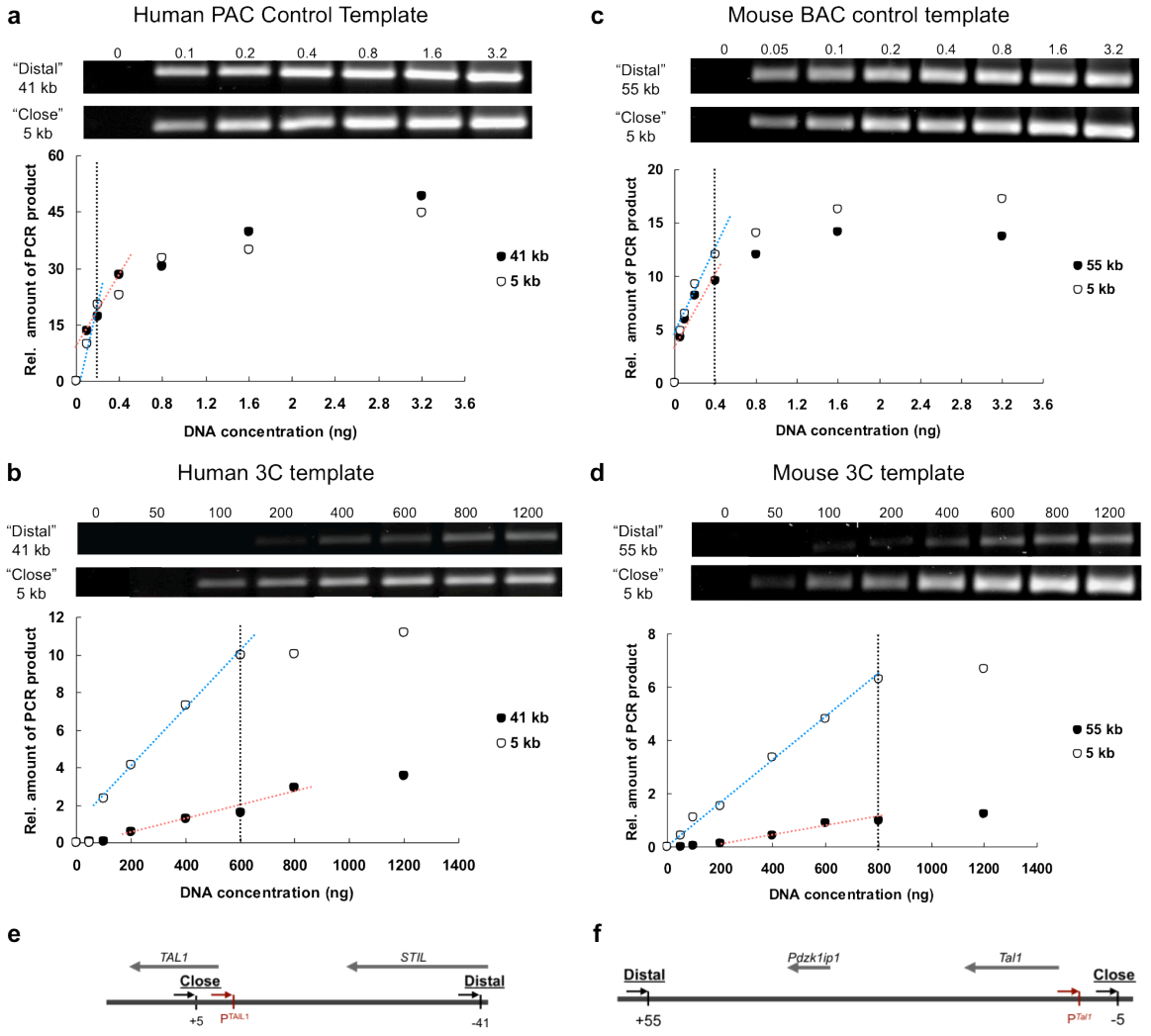


Figure 3.16 PCR titrations of 3C libraries and 3C control templates. Panel a and b: human PAC control template and 3C library. Panel c and d: mouse BAC control template and 3C library. PCR bands from gel images represent PCR products of 3C libraries or control templates using "distal" and "close" primer sets. The numbers on top of the gel represent the amount of PCR template used in PCR amplification. Titration results are shown by dot plots. The x-axis represents the DNA concentration used for PCR amplification and the y-

axis represents the relative amount of PCR product. The exponential range of PCR amplification for both "distal" and "close" primer sets are approximated by red dash line. Panel e and f: schematic diagrams of locations of "distal" and "close" primers in mouse and human TAL1 loci.

The "distal" and "close" primers were designed accordingly for both human and mouse (shown in Figure 3.16 e and f). In human TAL1 locus, the "distal" and the "close" primers are located 41 kb upstream and 5 kb downstream relative to the anchor (P^{TAL1}), respectively. In mouse *Tal1* locus, the "distal" and the "close" primers are located 55 kb downstream and 5 kb upstream relative to the anchor (P^{Tal1}), respectively. The experiments were performed on the 3C control templates as well as human and mouse 3C libraries. For human and mouse 3C libraries, the amount of DNA template used for PCR reaction was titrated from 50 ng to 1.2 ug (Figure 3.16 b and d). For 3C control templates, the amount of DNA template used for PCR reaction was titrated from 0.1 ng (human PAC) / 0.05 ng (mouse BAC) to 3.2 ng (Figure 3.16 a and c). Gel images of PCR products as well as composite plots of quantification data were shown in Figure 3.16. Trend lines were drawn to facilitate the determination of the linear range of DNA concentration (for further details, refer to Chapter 1, section 1.2.4). For PAC/BAC control templates, the similar titration profiles were observed between samples using "distal" (41/ 55 kb) and "close" (5 kb) primer pairs (Figure 3.16 a and c). Providing the representation of all possible ligation products is theoretically equal in the 3C control template, the minor differences were likely due to differences of primer efficiencies (this will be discussed in the following section 3.5.3). The linear ranges of PCR amplification was determined based on the trend line (Figure 3.16a & c), corresponding to 0.1-0.2 ng human PAC template and 0.05-0.2 ng of mouse BAC template, respectively. In contrast, significant difference was observed for the yields of PCR products between "distal" (41 kb or 55 kb) and "close" (5 kb) primer pairs when using 3C templates (Figure 3.16 b & d), as the ligation frequency of two distal fragments is always lower than two proximal fragments. For human and mouse 3C DNA templates, the linear ranges of PCR amplification for both primer pairs were between 200-600 ng and 200-800 ng accordingly (shown in Figure 3.16 b and d). Therefore, the optimized DNA concentration for subsequent 3C analysis should be in the ranges of 200-600 ng and 200-800 ng for human and mouse 3C libraries.

3.5.3 Determination of PCR efficiency of 3C primers and 3C data normalization

For the 3C assay, the final ligation frequency between the anchor and each region of interest is determined by quantifying PCR products obtained from images of agarose gels and normalising these quantified values against their respective primer efficiencies. For each 3C primer sets, PCR reactions were performed in triplicate. This section illustrates how PCR efficiencies of all primer sets using 3C control templates were determined, and how these were used to normalise the 3C data. A schematic map is shown to illustrate the whole process of 3C-PCR detection (Figure 3.17). It has the following steps: i) PCR amplification; ii) gel quantification; iii) calculation of relative amount of PCR products; iv) calculation of relative ligation frequency and v) data interpretation (shown in Figure 3.18 left panel).

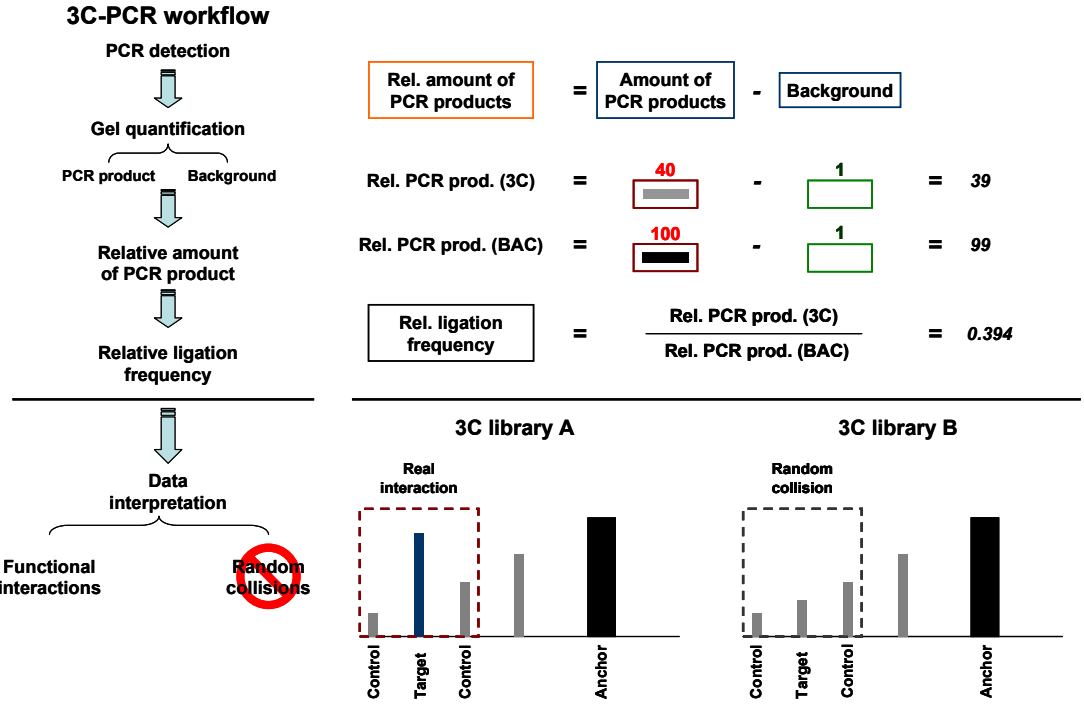


Figure 3.17 Procedures of PCR efficiency determination and 3C normalisation. The flowchart on the left panel shows the workflow of 3C-PCR detection and the schematic map on the right panel displays an example of 3C normalisation and how to distinguish functional interactions from random collisions. The red boxes with black or grey bands represent the PCR product of 3C-PCR and green boxes represent the background from the gel and the colour-coded numbers on top of those boxes represent the intensity values given by gel quantification. Numbers labelled in orange and black represent the relative amount of PCR and relative ligation frequency respectively after calculation and normalisation.

An example is given with metadata to demonstrate the procedures of data analysis (shown in Figure 3.17 right panel). The PCR amplification was performed using two biological replicates 3C libraries and their corresponding 3C control templates, and PCR products were subjected to agarose gel electrophoresis. The intensity values for PCR products and for background from the gel were quantified and the relative amount of PCR product was calculated by the intensity value of PCR product minus the background from the gel. The relative ligation frequency for each primer sets was determined by dividing the value of the relative amount of PCR product of the 3C library by the 3C control template, in order to normalise the differences of PCR efficiency between primer sets. Eventually, any *bona fide* functional interactions were determined by comparing the relative ligation frequency between target regions and their control regions which were juxtaposed between the target and the anchor region. Random collisions would decrease as a function of distance from the anchor while the relative ligation frequencies of true functional interactions would increase regardless of their genomic distance to the anchor when compared to the juxtaposed control regions (Figure 3.17 bottom panel). Statistical tests (unpaired two-tailed t-tests) were used to determine the significance of differences in ligation frequencies between target amplicons and control amplicons.

The PCR efficiency of 3C primer pairs was determined using the 3C control templates which were generated from human PAC and mouse BAC DNA. The steps of PCR detection, gel quantification and calculation of relative amount of PCR products were performed as described above. Figure 3.18 shows a typical gel image of PCR products obtained from amplification with the control templates with 3C primer combinations. The quantified data is also shown plotted in the accompanying histograms thus revealing differences in PCR efficiency of 3C primer sets across the TAL1 locus for both human and mouse. As shown in Figure 3.18, the bright bands within the red boxes indicate PCR products of the predicted amplicon size. Other bands in each lane correspond to non-specific amplification or partial digestion products detected by 3C-PCR.

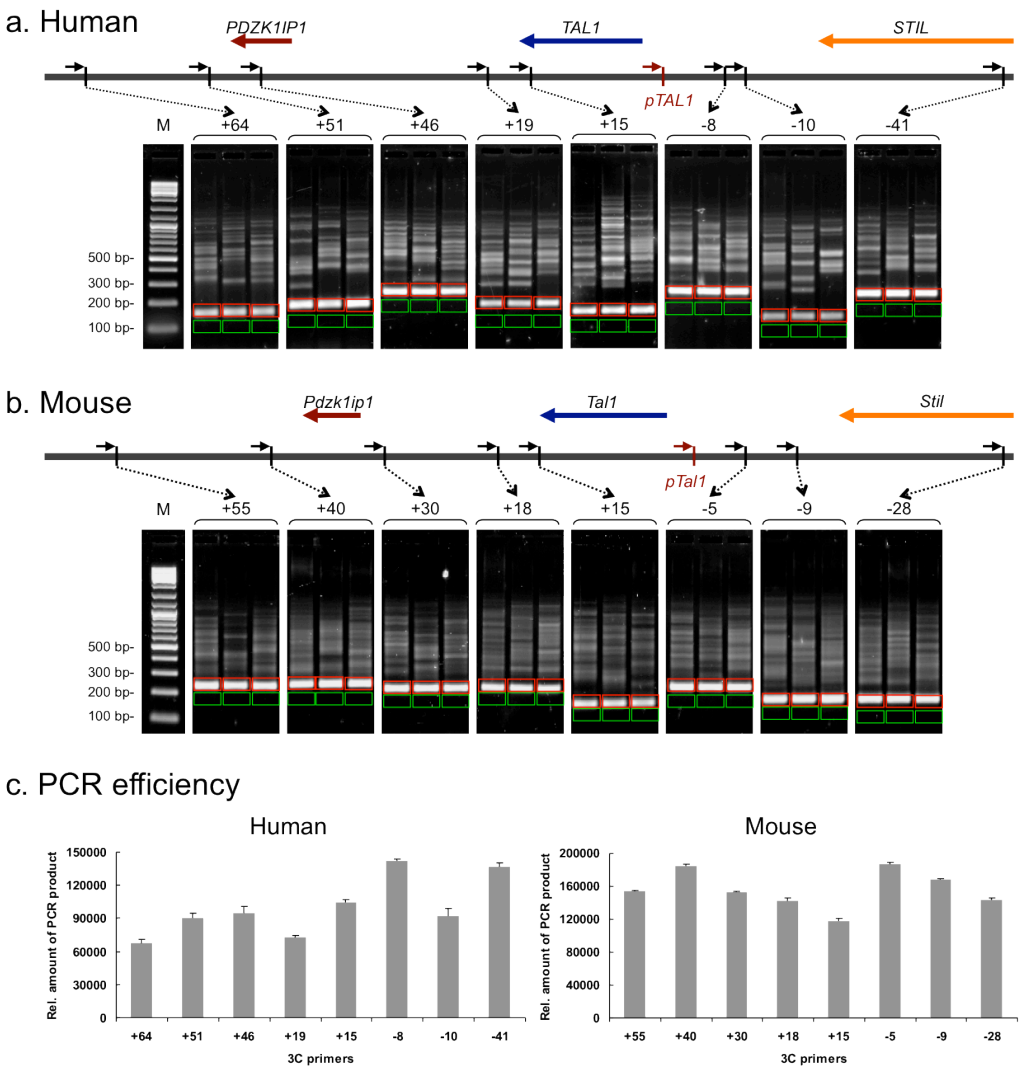


Figure 3.18 Determining PCR efficiencies of 3C primers using 3C control templates. Panel A and B: 3C-PCR analysis of 3C control templates (human PAC and mouse BAC respectively) imaging and quantifying by gel electrophoresis. The Schematic map on top of each panel shows the location of 3C primers which are labelled by their genomic positions to the TAL1 promoter (the anchor) in +/- kb distance. The label on each section corresponds to the 3C primers and the "M" equal to DNA marker. The size of DNA markers is labelled on the left-side of the gel image. The red and green boxes represent the areas of the PCR products and the backgrounds respectively. PCR efficiencies of the human and mouse 3C primers are shown on the right side of the figure. The x-axis represents the relative amount of 3C PCR product quantified using gel electrophoresis (shown in panel A and B, mean ± S.E.M, n = 3) and the y-axis represents the 3C primers labelled by their genomic location to the TAL1 promoter (the anchor) in +/- kb distance.

3.6 Determination of looping interactions in human and murine TAL1 loci by 3C

3.6.1 Determination of looping interactions in human K562 and HPB-ALL cell lines

The 3C was performed on human erythroid (K562) and lymphoid (HPB-ALL) cell lines to determine the looping interactions between the TAL1 promoter and three enhancers regions - the +51 erythroid enhancer, the +20 stem cell enhancer and the -10 enhancer. The relative amount of PCR product of all primers sets (three enhancers and five control regions) were acquired as previously described. Two biological replicates of each cell line are analysed in parallel and a high level of reproducibility in the 3C profiles were observed between replicates (Figure 3.19). Statistical analysis was performed to validate the reproducibility of the 3C analysis. Pearson correlation coefficients for bio-replicates comparisons in K562 and HPB-ALL were 0.94 ($p = 2.5 \times 10^{-4}$) and 0.72 ($p = 2.2 \times 10^{-2}$), respectively. Raw data of the 3C were then normalised against the PCR efficiency differences based on the PCR results obtained using the 3C control templates.

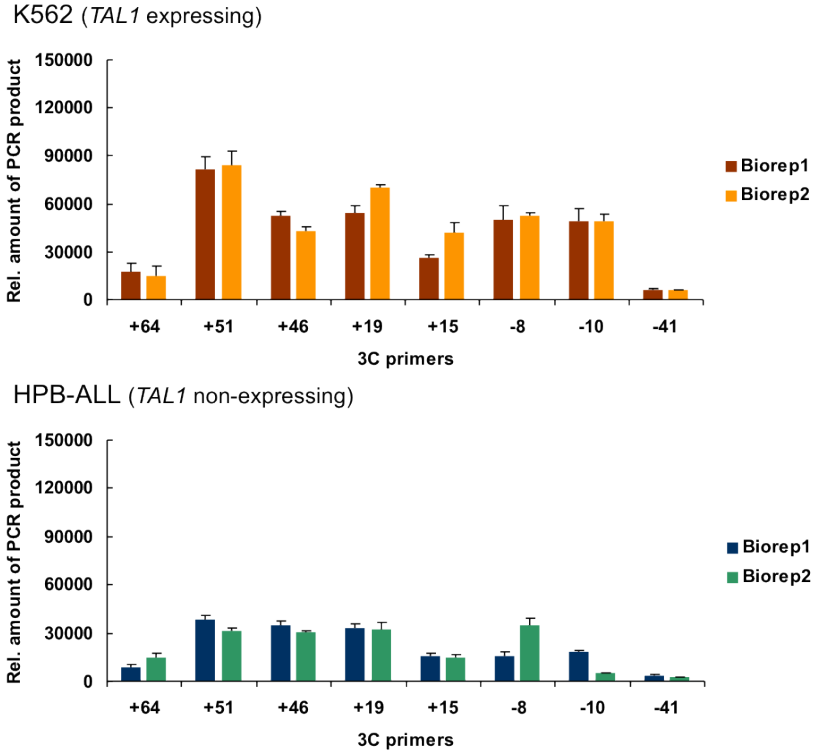


Figure 3.19 Raw band intensities (i.e., not normalised for PCR efficiencies) of the 3C-PCR assays in human K562 and HPB-ALL cell lines. Two biological replicates of each cell line are shown side-by-side in histograms with coloured labels. The y-axis represents the relative amount of 3C PCR product (counts) quantified using gel electrophoresis (mean + S.E.M, n =

3) and the x-axis represents the labelled 3C primers based on their distance in kilobases upstream (-) or downstream (+) from the TAL1 promoter 1a.

As shown in Figure 3.20, the final 3C-PCR profiles of K562 and HPB-ALL (the averages of two bio-replicates) are presented in the histograms (Figure 3.20a) along with the ERCC3 internal quality control (Figure 3.20c). In K562, the highest interacting partner for the TAL1 promoter 1b was observed at the +51 erythroid enhancer (p = 3.1×10^{-4}). In addition, high-level of interactions were also detected at the +19 and -10 enhancers (p = 9.3×10^{-5} and 9.9×10^{-3}, respectively). All of these interactions were statistically significant above background (control regions). These data suggest that *bona fide* looping interactions occur *ex vivo* in a TAL1 expressing cell line between the TAL1 promoter 1b region and regions involving the +51, +19 and -10 enhancers. In contrast, no significant differences in ligation frequencies were observed between control regions and neither the +51 enhancer (p = 0.25) nor the -10 enhancer (p = 0.29) in TAL1 non-expressing HPB-ALL. However, the significant interaction frequency was observed at the +19 enhancer (p = 1.7 x 10^{-4}), suggesting that the +19 enhancer interacts via looping with the TAL1 promoter 1b region in HPB-ALL cells *ex vivo*. In addition, this enhancer-promoter interaction is not context-dependent, as the TAL1 expression is below the level of detection in HPB-ALL cells. Using the same logic, the interaction profiles of K562 and HPB-ALL also suggest that interactions between the TAL1 promoter and either of +51 and -10 are context-dependent – since they are found only in TAL1-expressing K562 cells.

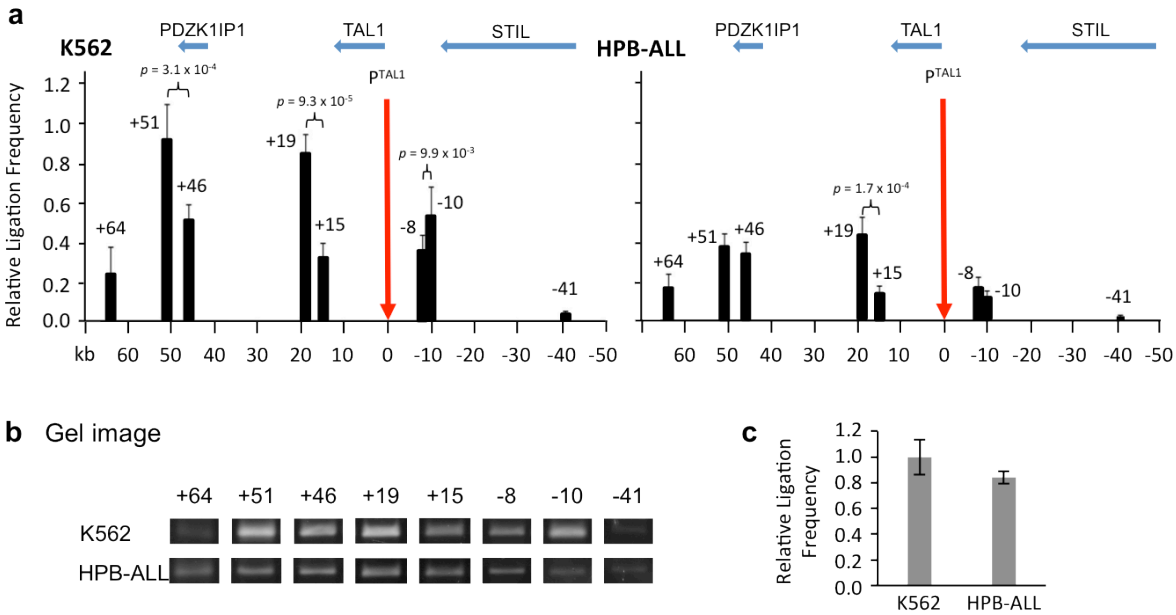


Figure 3.20: 3C-PCR assays of the TAL1 promoter across the TAL1 locus in human K562 and HPB-ALL cells. (A) 3C profiles: schematic diagram at the top of the figure shows the

genomic organisation of the TAL1 and its neighbouring genes. 3C primer pairs are denoted with numbers based on their distance in kilobases upstream (-) or downstream (+) from the TAL1 promoter. The position of the 3C anchor (TAL1 promoter 1b) is denoted in red bars. The y-axis represents the relative ligation frequency after PAC normalisation (mean + S.E.M, n = 2 bio-replicates) and the x-axis represents the kb distance from the 3C anchor. The p-values indicate the statistical significance determined by student T-test. (B) A Gel image of 3C PCR products shows the raw data obtained from agarose gel electrophoresis. (C) Bar diagrams show the interaction frequencies between two non-adjacent Csp6I fragments at the Ercc3 locus determined by 3C. Interaction frequencies (grey bars) are shown with standard errors and represent the mean from two biological replicate samples.

3.6.2 Determination of looping interactions in murine MEL and BW5147 cell lines

In order to determine whether the looping interactions which were identified in human cell lines were conserved at the murine *Tal1* locus, 3C analysis was also performed on murine MEL (Tal1 expressing) and BW5174 (Tal1 non-expressing) cells. For each cell line, data sets from two biological replicates were combined and statistical analysis (t-test) was performed as described. 3C-PCR profiles of MEL and BW5147 are shown in the histograms (Figure 3.21). For MEL cells, the three murine Tal1 enhancers at +40, +18 and -9 (equivalent to human TAL1 +51, +19 and -10) showed the highest ligation frequencies to the TAL1 promoter 1b region (shown in Figure 3.21). Two of these interactions were statistically significant - the +40 erythroid enhancer when compared with its control region +30 (p = 1.2 x 10^{-3}), and the +18 stem cell enhancer and its control region +15 (p = 2.7 x 10^{-3}), suggesting that the TAL1 promoter shows bona fide looping interactions with both the +40 and +18 enhancers in MEL cells. This was in complete agreement with the data obtained from the human K562 cell line. However, no statistical difference in ligation frequencies was observed between the -9 enhancer and it control region -5 (p = 0.18), which might suggest a biological difference between K562 and MEL cells with respect to *ex vivo* interactions between the TAL1 promoter and this upstream enhancer region. Similar to HPB-ALL, no significant interactions were observed at the +40 and -9 enhancer in the *Tal1* non-expressing BW5147 cells. In addition, the ligation frequency of the +18 enhancer was significantly higher than its control (p = 1.6 x 10^{-3}) indicating a true looping interaction between the +18 enhancer and the *Tal1* promoter in BW5147 cells. This was also in agreement with the data obtained in human HPB-ALL cells, further reinforcing the idea that the TAL1 stem cell enhancer (+19 in human, +18

in mouse) interacts with the TAL1 promoter 1b in a transcriptionally independent manner.

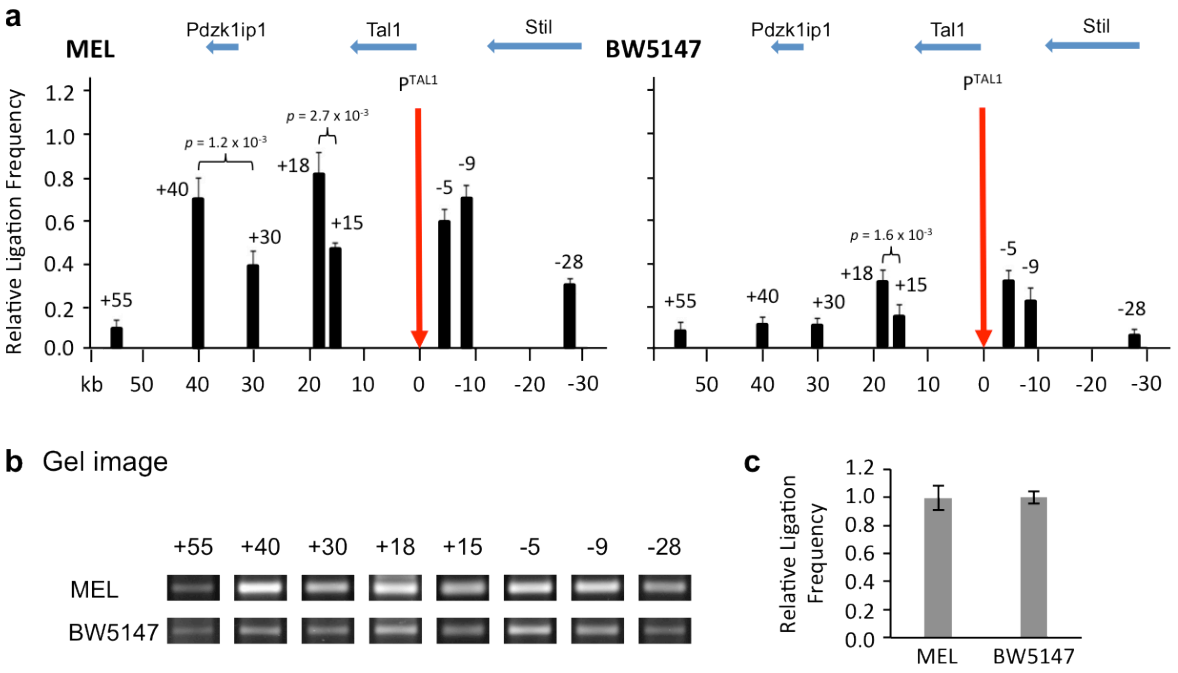


Figure 3.21: 3C-PCR assays of the Tal1 promoter across the Tal1 locus in murine MEL and BW5147 cells. (A) 3C profiles: schematic diagram at the top of the figure shows the genomic organisation of the Tal1 and its neighbouring genes. 3C primer pairs are denoted with numbers based on their distance in kilobases upstream (-) or downstream (+) from the Tal1 promoter. The position of the 3C anchor (Tal1 promoter 1b) is denoted in red bars. The y-axis represents the relative ligation frequency after BAC normalisation (mean + S.E.M, n = 2 bio-replicates) and the x-axis represents the kb distance from the 3C anchor. The p-values indicate the statistical significance determined by student T-test. (B) A Gel image of 3C PCR products shows the raw data obtained from agarose gel electrophoresis. (C) Bar diagrams show the interaction frequencies between two non-adjacent *Csp*6I fragments at the Ercc3 locus determined by 3C. Interaction frequencies (grey bars) are shown with standard errors and represent the mean from two biological replicate samples.

3.6.3 Determination of looping interactions in murine erythroid and lymphoid cells

In order to determine whether the looping interactions that had been identified were restricted to cell lines (*ex vivo*) or whether they were present *in vivo* in primary haematopoietic cells, primary bulk erythroblasts (a gift from Peter Fraser's laboratory – Babraham, U.K.) and lymphocytes were analyzed by 3C across the murine *Tal1* locus. Strikingly, the interacting patterns found in mouse primary cells were highly similar to their corresponding murine cell lines (Figure 3.21). In primary erythroblasts, the highest ligation frequency was observed at the +40 enhancer which was statistically different from its control region at +30 (p = 1.2 x 10^{-4}). In contrast, no significant difference was shown between the +40 and +30

regions in primary lymphoid cells. These results confirmed that +40 erythroid enhancer interacts with the *Tal1* promoter 1b in a context-dependent manner *in vivo*. Ligation frequencies of the +18 stem cell enhancer were shown to be significantly higher that its control region in both erythroid and lymphoid cells (p = 9.4 x 10^{-3} and 3.1 x10^{-4}, respectively). This again confirmed that the +18 enhancer interacts with *Tal1* promoter 1b irrespective of the transcription of TAL1. No *bona fide* looping interaction between *Tal1* promoter 1b and the -9 enhancer was identified in either primary erythroid or lymphoid cells, as the ligation frequencies for this region were significantly lower than its control at the -5 region (p = 2.5 x 10^{-2} and 4.6 x 10^{-3}, respectively). Thus, while human K562 cells showed a significant looping interaction between -10 and TAL1 promoter 1b, this was not confirmed in either murine erythroid cells or a murine erythroid cell line – indicative of either a species-specific difference, or a difference due to the behavior of cells grown in culture.

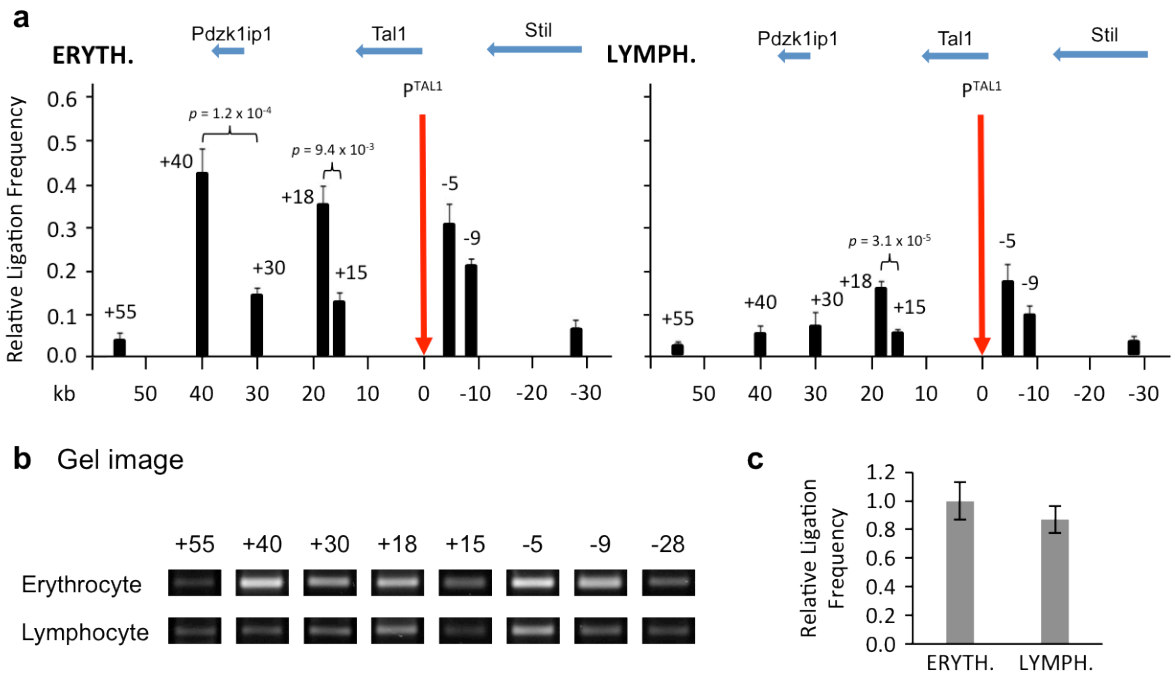


Figure 3.22: 3C-PCR assays of the TAL1 promoter across the Tal1 locus in murine primary erythroblasts and lymphoid cells. (A) 3C profiles: schematic diagram at the top of the figure shows the genomic organisation of the Tal1 and its neighbouring genes. 3C primer pairs are denoted with numbers based on their distance in kilobases upstream (-) or downstream (+) from the Tal1 promoter. The position of the 3C anchor (Tal1 promoter 1b) is denoted in red bars. The y-axis represents the relative ligation frequency after BAC normalisation (mean + S.E.M, n = 2 bio-replicates) and the x-axis represents the kb distance from the 3C anchor. The p-values indicate the statistical significance determined by student T-test. (B) A Gel image of 3C PCR products shows the raw data obtained from agarose gel electrophoresis. (C) Bar diagrams show the interaction frequencies between two non-adjacent *Csp*6I fragments at the Ercc3 locus determined by 3C. Interaction frequencies (grey bars) are shown with standard errors and represent the mean from two biological replicate samples.

The Pearson correlation analysis was performed to statistically assess the similarity of 3C interaction patterns between murine cell lines and primary cells. The 3C interaction profile of MEL cells was compared to the erythroblasts, while the profile of BW5147 cells was compared with the lymphocytes. The pairwise comparisons showed high-level correlations between cell lines and primary cells. The correlation co-efficiencies were 0.878 ($p < 0.004$) and 0.963 ($p < 0.0001$) for the murine erythroid lineages and lymphoid lineages, respectively. For the TAL1 locus, the cell lines used here do appear to reflect the true nature of in vivo chromatin looping events.

Discussions

3.7 Differences and similarities of the TAL1 looping interactions in human and mouse cells

Based on 3C profiles shown in section 3.5, Table 3.3 summarises the TAL1 looping interactions observed in human and murine cell lines as well as murine primary cells. In general, looping interactions of the erythroid enhancer are only presented in human and murine erythroid lineages, but completely absent in the lymphoid lineages. Interactions between the TAL1 promoter and the stem cell enhancer are presented in all cell types, regardless of the TAL1 expression. In addition, the interaction of the -10 enhancer is only present in human erythroid K562 cells, but absent in lymphoid and murine erythroid lineages.

	K562	HPB-ALL	MEL	BW5147	Erythroid	Lymphoid
TAL1 / *Tal1* expression	+	-	+	-	+	-
TAL1 +51 / *Tal1* +40	Present	Absent	Present	Absent	Present	Absent
TAL1 +20 / *Tal1* +18	Present	Present	Present	Present	Present	Present
TAL1 -10 / *Tal1* -9	Present	Absent	Absent	Absent	Absent	Absent

Table 3.3 Comparison of looping interactions at the TAL1 locus between human and murine cells. The cell lines and primary cells are shown on the top of the table. The expressions of TAL1 in each cell types are listed in the first row. Three colour-coded (orange, green and blue) enhancers of the TAL1 locus are shown on the left side. The presence (green) or absence (red) of bona fide looping interactions between the TAL1 promoter and enhancers in each cell type are shown.

3.7.1 Looping interactions between the TAL1 promoter and the erythroid enhancer (+51/ +40)

Looping interactions between the TAL1 promoter and its erythroid enhancer (+51/+40) were observed in both the human K562 and mouse MEL cell lines as well as mouse primary erythroid cells, all of which express the TAL1. In contrast, no interaction between the TAL1 promoter and +51/+40 enhancers was observed in the TAL1 non-expressing cells.

3.7.1.1 The TEC complex in mediating the looping interaction via GATA/E-box motif

It has been shown that the erythroid enhancer has two GATA/E-box motifs (the 5' GATA/E-box and the 3' GATA/E-box), which are conserved in human, dog, mouse and rat (Ogilvy et al., 2007). Moreover, GATA/E-box motifs have been identified in numbers of hematopoietically expressed genes, such as EKLF (Anderson et al., 1998, 2000), GATA-1 (Nishimura et al., 2000; Vyas et al., 1999), α-globin (Anguita et al., 2004), glycophorin A (Lahlil et al., 2004), ECPR (Mollica et al., 2006), Cdc6 (Vilaboa et al., 2004) and protein 4.2 (Xu et al., 2003). GATA/E-box motif is also known to mediate the recruitment of the TAL1 erythroid complex (TEC), containing TAL1, E2A, LDB1, LMO2 and GATA1 (Rodriguez et al., 2005; Wadman et al., 1997). In particular, it has been shown that the TAL1 erythroid complex trans-activates the murine protein 4.2 gene via two GATA/E-box motifs in its proximal promoter. The formation of a DNA loop between two GATA/E-box elements is mediated by ldb1 bridging protein, in linking two TEC complexes (Xu et al., 2003). This provides a model which shows how the erythroid enhancer may possibly activate the TAL1 transcription.

However, the 5' GATA/E-box motif exhibits 9-bp spacing while the 3' GATA/E-box motif shows only 6-bp spacing (Ogilvy et al., 2007). Such spacing was previously shown to preclude recruitment of the TEC complex in erythroid cells (Wadman et al., 1997), and the 3' GATA/E-box was found to be particularly critical for the enhancer activity in the midbrain but not in erythroid cells (Ogilvy et al., 2007). It also has been shown that both TAL1 and GATA1 proteins bind to the mouse +40 enhancer in definitive erythroid cells, and mutation of the 5' GATA/E-box motif

results in loss of enhancer function (Ogilvy et al., 2007). It suggests that the 5' GATA/E-box motif may be the only element functioning in the erythroid cells. Moreover, no GATA/E-box motif is present in the TAL1 promoter or other TAL1 enhancers characterized so far (Bockamp et al., 1995; Gottgens et al., 2000; Gottgens et al., 2004; Gottgens et al., 2002b). Thus, the model of the murine protein 4.2 gene may not be applicable for the TAL1 locus.

Additionally, it has been found that both the TAL1 promoter and its erythroid enhancer (+51) are directly bound with the TEC complex and RNA polymerase II in human K562 cells (Dhami et al., 2010). Providing the fact that the erythroid enhancer and the TAL1 promoter are in close-proximity in erythroid cells, it implies that the looping interaction between the TAL1 promoter and the +51 enhancer may still be mediated by GATA1 and other members of the TEC complex. Although without the GATA/E-box element over the promoter, the TEC complex may be recruited at the GATA/E-box motif of the erythroid enhancer; and in some way, be brought to the promoter and enhancing TAL1 transcription in the erythroid lineage.

3.7.1.2 Role of CTCF in TAL1 regulation

CTCF acts either as the enhancer-blocking insulator in preventing the distal enhancer from interacting with the gene promoter or as boundary element in blocking the spread of heterochromatin (Wallace and Felsenfeld, 2007). As mentioned in the section 3.1 of this chapter, the CTCF binding site at the TAL1 +40 may act as an enhancer-blocking insulator in preventing the +51 enhancer from activating the TAL1 promoter. Moreover, it has been shown recently that CTCF sites form chromatin loops by interacting with each other in the globin and H19/Igf2 loci in mammals (Splinter et al., 2006; Zhao et al., 2006). In addition, it also has been illustrated that cohesion and CTCF are able to facilitate formation of *cis*-interactions in regulating the IFNG locus (Hadjur et al., 2009). The latest study on CTCF was showed that CTCF was able to function as a bridge in connecting enhancers to their target gene for transcription activation (Handoko et al., 2011). , In contrast to the well-accepted enhancer-blocking model, this observation suggests that CTCF may play a positive regulatory role in at least a subset of genes.

A number of CTCF binding sites (+57, +53, +40 and -31) have been identified at the TAL1 locus in human K562 cells (Dhami et al., 2010), all of which are also experimentally confirmed or predicted based on the consensus binding motif of CTCF (Dhami et al., 2010; Kim et al., 2007). It is speculated that these CTCF sites may also be involved in the looping configuration of the TAL1 locus, by either stabilizing the active loops to facilitate the TAL1 transcription or isolating the *cis*-acting elements (e.g. enhancers) from the promoters by forming inactive loops.

Providing the looping interaction exists between the enhancer and promoter, the presence of the +40 insulator must not be able to interfere with this interaction. It is speculated that the CTCF binding at the +40 may not be in the same cells of the enhancer-promoter looping structure. Alternatively, the CTCF site at the +40 region may not participate in any looping structure, neither interacting between CTCF sites to form the inactive loops nor being in contact with the +51 enhancer to block the enhancer-promoter interaction. Furthermore, the +40 insulator may even play a positive regulatory role in mediating the enhancer-promoter interactions as shown previously. Thus, it is important to further determine the possible interactions between these prominent CTCF sites using the 3C analysis. The role of these CTCF sites in local looping structures at the TAL1 locus will be examined in Chapter 5.

3.7.2 Looping interactions between the TAL1 promoter and the stem cell enhancer (+19/ +18)

Interaction between the TAL1 promoter regions and the stem cell enhancer (+19/+18) was observed in all human and murine cells studied irrespective of the transcription of TAL1. Furthermore, the +19/+18 enhancer was the only interaction detected at statistically significant levels in the TAL1 non-expressing cells. The murine stem cell enhancer has been shown to drive expression of TAL1 in HSCs and hematopoietic progenitors but not in mature cells, indicating that this enhancer has an important role in progenitors but is not sufficient to drive erythroid maturation (Sanchez et al., 1999; Sanchez et al., 2001). Thus, the +19/+18 enhancer cannot drive the TAL1 expression in erythroid cells such as K562 and MEL. In addition, it has previously been reported that the TAL1 +19 enhancer is not required for TAL1 transcriptional initiation or hematopoiesis (Gottgens et al.,

2004). All of these observations agree with the 3C interaction patterns in erythroid and lymphoid cell lines regardless of the TAL1 expression.

Nevertheless, how this looping interaction is related to TAL1 function remains uncertain at this stage. It is speculated that the P^{TAL1}- +19/+18 interaction may act as a repression loop to suppress the TAL1 transcription in the lymphoid lineages (HPB-ALL, BW5147 and murine lymphocytes) as well as the fraction of erythroid cells in which TAL1 is not expressed. Recently, Brown and colleagues have identified a looping interaction between the promoter and terminator region of BRCA1 (breast cancer-associated gene) using the 3C analysis, which was associated with silencing of the BRCA1 gene (Tan-Wong et al., 2008). In addition, the stem cell enhancer has also been shown to be able to drive the expression of the TAL1 flanking gene, PDZK1IP1 *(also known as Pdzk1ip1 or MAP17)* in mouse *Pdzk1ip1* promoter-TAL1 +19 fusion transgenic embryos (Tijssen et al., 2011). However, the mechanism by which a looping interaction between +19/+18 and the TAL1 promoters could facilitate *PDZK1IP1* expression is not clear. Consequently, the formation of chromatin loops can either facilitate or repress the gene transcription. Further work is required to elucidate the function of chromatin loops involving the stem cell enhancer.

3.7.3 Looping interactions between the TAL1 promoter and the -10/-9 enhancer

Significant interactions were identified between the -10 enhancer and the TAL1 promoter in K562 but not in HPB-ALL cells. In addition, no interaction was observed at the paralogous -9 enhancer in all four murine cell types. The -10 enhancer was firstly identified using ChIP-chip and enhancer trapping assays in K562 cells (Dhami et al., 2010). The -10 region showed hallmarks of the enhancer including the significant binding of RNA polymerase II as well as high-level of H3K4 methylation and low-level of H3K27me3 modifications.

In both of these murine cell types, the -9 region (the equivalent of human -10) did not show statistically significant increases in ligation frequencies above its control. This suggested that the results we obtained for -10 in human were species-specific or reflected differences in interactions found in cell lines and primary cells.

However, it cannot exclude a possibility that the small statistically significant increases in ligation frequencies between the TAL1 -10 region and its control may be due to a mechanistic difference as it was only observed in human but not in murine erythroid lineages.

3.7.4 Reduced ligation frequency in TAL1 non-expressing cells

Overall, lower levels of 3C interactions were observed in the TAL1 non-expressing (lymphoid) cells relative to TAL1 expressing (erythroid) cells. This observation extends to both test and control elements across the TAL1 locus. This effect is not due to issues with 3C library quality as a control gene, ERCC3, which is outside the TAL1 locus and did not show significantly reduced relative ligation frequencies in lymphoid cells when compared to erythroid (data shown in Appendix Figure 1). It appears that a TAL1 non-expressing cell type has lower levels of both functional interactions and random collisions, therefore, resulting in reduced levels of ligation events which may be due to the reduced spatial complexity and proximity between chromatin fibres in the inactive loci.

This observation also lines up with previous studies; the interaction frequencies between promoters and their flanking regions up to 400 kb in both directions were significantly lower for inactive genes in comparison to active genes using the Hi-C approach (Yaffe and Tanay, 2011). Thus, inactive loci are likely to have a reduced 3-dimensional looping structure overall – and this is reflected in lower 3C ligation events than observed for active genes. Given that only the TAL1 stem cell enhancer (+19 or +18/human or mouse) appears to be in contact with the TAL1 promoter in lymphoid cells further supports this idea. However, as the number of 3C primer combinations used here was relatively low, it is unclear whether other regions across the locus that were not tested here, may show much higher levels of interactions. Thus lymphoid cells may have complex looping structures – but they are very different from those observed in erythroid cells and may involve other genomic sequences that were not tested here.

3.7.5 Primary models of the TAL1 looping configurations

Assuming all the looping interactions captured by the 3C are presented in the single cells, primary models of the TAL1 looping configuration are illustrated in Figure 3.23. For TAL1 non-expressing lymphoid cells, the stem cell enhancer is in close proximity with the TAL1 promoter, mediated by some unknown transcription factors. For TAL1 expressing erythroid cells, two chromatin loops formed in between the erythroid and stem cell enhancers and the TAL1 promoter, mediated by undefined transcription factors.

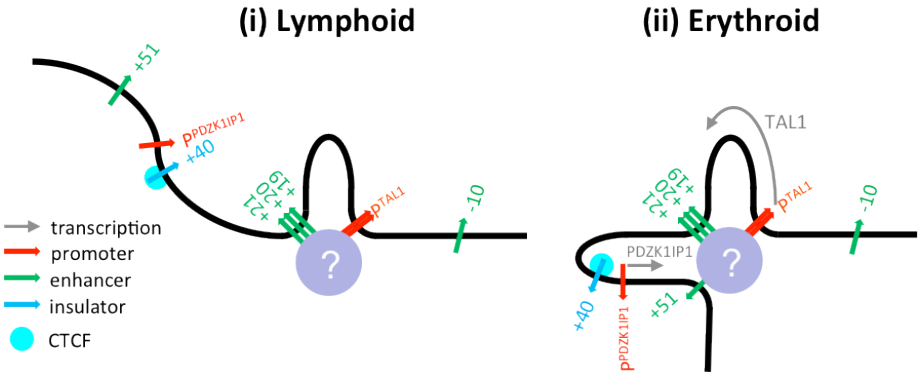


Figure 3.23: Putative structural organisation of the TAL1 chromatin hub. Organization of the TAL 1 chromatin hub in TAL1^{-} lymphoid cells (i) and TAL1^{+} erythroid cells (ii). Promoters, enhancers and CTCF binding sites are indicated by the red, green and blue arrows respectively. The erythroid (+51) and the stem cell (+19 → +21) enhancers are highlighted. Direction of transcription of relevant genes are indicated by the grey arrows. Unknown factors mediating the interaction of the TAL1 promoters and other regulatory elements in lymphoid (i) and erythroid cells (ii) are represented by the blue-grey ball.

Further analysis may be required to characterize the looping configuration of the TAL1 locus in detail. Interactions between other *cis*-regulatory elements, including enhancers, CTCF insulators and promoters of neighboring genes will be examined by both 3C and 4C analysis in the following chapters to build up a more sophisticated model of the TAL1 locus. In addition, it is proposed that the GATA1 protein and the TEC complex are involved in mediating the +51-P^{TAL1} interactions in erythroid cells. The GATA1 siRNA knockdown experiments will be performed in the K562 cells to validate whether the formation of the +51-P^{TAL1} interaction were dependent on the GATA1/TEC complex.

3.8 Weakness of the 3C-PCR method

The results of 3C-PCR between independent biological replicates demonstrated high reproducibility of the method. However, several disadvantages of method should be discussed based on the results described in this chapter.

3.8.1 The 3C is a low-throughput assay

The conventional 3C is a low throughput method that is only able to assess a limited number of elements within a limited span (commonly several hundred Kb to 1 Mb) of genomic region in a single run of analysis (de Wit and de Laat, 2012). As shown in overall strategy (section 3.3), the 3C was performed with eight primer pairs for each species in this chapter. Given the fact that a 4-bp cutting restriction enzyme (i.e., Csp6I) was used in 3C preparation, it provided the capability for capturing ligation products with a relatively high-resolution. In reality, the resolution of 3C was also compromised because of its limited capacity for detection.

3.8.2 The 3C is a population-based assay

As the 3C is a population-based assay starting with 10 million cells, the looping interactions detected by the 3C may also be presented in different proportions of the cells. For instance, the +19/+18 interaction may be a repressive loop which is found in lymphoid cells as well as a proportion of erythroid cells that do not express TAL1. At this stage, it is impossible to conclude that the looping interactions detected by the 3C all occur in the same proportion of cells at the same time. Thus, it requires further 3C analysis in combination with other approaches, such as siRNA knockdown, to disrupt the chromatin interactions by depleting the key transcription factors that are mediating the loops. The looping structures affected by knocking-down of the key factor protein are the chromatin interactions exiting in the same cells.

The total numbers of known regulatory elements across the TAL1 loci are 17 and 14 in human and mouse, respectively. So far, only three enhancers of each species were tested by the 3C in this chapter. Thus, the 3C profiles generated in this chapter only provide a very limited snapshot of the looping interactions in the

TAL1 loci. The 4C is a high-throughput method that is able to capture all chromatin interactions from one particular viewpoint of the entire genome. The 4C analysis (Schoenfelder et al., 2010) will be applied in combination with the high-resolution TAL1 tiling path array in order to provide a broader view of chromatin interactions at the TAL1 locus. Furthermore, 4C analysis is also amenable for the use of restriction enzymes such as Csp6I. Thus, 4C is fully able to make use of the higher resolution provided by 4 bp cutting restriction enzymes – since the final number of data points assayed would be equivalent to the number of Csp6I sites across the TAL1. There are approximately 450 restriction sites within the human TAL1 locus with an average resolution of approx. 600 bp.

Conclusions

The work presented in this Chapter demonstrated that the 3C is a relatively efficient technology that can be used to characterise chromatin interactions locally, providing that the one is prepared to perform the appropriate controls and has prior knowledge of the locus under study in order to target the regulatory elements of interest. The 3C analysis described here has provided primary looping models of the TAL1 locus in both human and mouse TAL1-expressing and non-expressing cells. The subsequent chapters would further elaborate these models and assess whether the formation of the chromatin structure is dependent on the particular transcription factors.

Chapter 4 The role of CTCF and cohesin (Rad21) in transcription regulation of the TAL1 locus

Summary

A number of CTCF binding sites (CTSs) have been previously identified across the TAL1 locus, of which the role in transcription regulation is still undetermined. Human TAL1 expressing erythroid cell line (K562) and non-expressing lymphoid cell line (HPB-ALL) have been used for the study. The ChIP-qPCR has been performed to assess the binding of CTCF and Rad21 (cohesin) at four CTSs (TAL1 +57, +53, +40 and -31) at the TAL1 locus. In K562 cells, significant ChIP enrichments of CTCF and Rad21 have been observed at all four CTSs. In contrast, high-levels of ChIP enrichments have been shown at +57, +40 and -31 for CTCF and at +57, +40 for Rad21. In addition, 3C has also been performed to determine the chromatin interactions between the four CTSs in the TAL1 locus. Long-range looping interactions have been identified only between +57/53 and -31 in K562 cells. In contrast, no interaction has been observed between all four CTSs in HPB-ALL cells. These observations suggest that CTCF and/or cohesin play a key role in mediating looping interactions between CTSs at the TAL1 locus in K562 cells. Additionally, these looping interactions may also participate in transcription regulation of the TAL1 locus in K562 cells.

4.1 Introduction

In previous chapter, it has been shown that the TAL1 promoter physically interacts with three of its enhancers (+51, +21/20/19 and -10/9) in the erythroid K562 cells. In addition, recent studies have demonstrated the presence of CTCF-bound elements at the TAL1 locus which display insulator enhancer-blocking or barrier activity *in vitro* or *in vivo* (Dhami et al., 2010; Follows et al., 2012). However, the presence of CTCF at +40 in K562 cells (Dhami et al., 2010) makes it difficult to envisage how the +51 erythroid enhancer communicates with its promoters to regulate TAL1 expression in this lineage. Therefore, it is necessary to determine the involvement of CTSs in the looping structure of the TAL1 locus, in order to better understand the relationship between insulator behaviour and transcriptional regulation of TAL1 through looping with its known *cis*-regulatory elements.

4.1.1 Role of CTCF in transcriptional regulation

CTCF has a number of regulatory roles including transcriptional regulation of gene expression at the β-globin and IFN-γ loci, V(D)J recombination at the immunoglobulin-encoding Igh and Igk loci, mono-allelic expression of imprinted genes and X-chromosome inactivation (Phillips and Corces, 2009). Most importantly, the role of CTCF insulators in mediating intra- and inter-chromosomal interactions has been well established in vertebrate genomes using 3C, 4C as well as FISH technologies.

At the β-globin locus, it has been shown that a number of CTCF insulators including 5'HS5 and 3'HS1 sites as well as other distal CTSs flanking the locus interact with each other (Tolhuis et al., 2002). Conditional deletion of CTCF and targeted disruption of a CTCF-binding site destabilised the long-range interactions between CTCF binding sites at the mouse β-globin locus (Splinter et al., 2006), which demonstrated the direct involvement of CTCF in chromatin organisation. In addition, the interactions between 5'HS5 or 3'HS1 and other CTSs are cell-type specific which only exist in erythroid cells, not in non-expressing brain cells (Hou et al., 2010; Splinter et al., 2006). Knockdown of CTCF leads to a global reduction of CTCF-mediated interactions, which negatively affects β-globin transcription (Hou et al., 2010). A most recent study demonstrated that CTCF/cohesin mediated looping structures, facilitating the interactions between regulatory elements and promoter and regulating gene expression at mammalian β-globin locus (Chien et al., 2011).

It has also been shown that CTCF-mediated chromatin interactions are required for modulating gene expression in several other loci. For example, it has been found that CTCF facilitates the contacts between CTSs and is responsible for the establishment and maintenance of a specific three-dimensional organisation of silenced cluster of Hox genes in human cells (Ferraiuolo et al., 2010). Moreover, CTCF-mediated chromatin interactions also have a crucial function in imprinted gene expression at the H19/Igf2 locus. It has been demonstrated that these CTSs can mediate allele-specific interactions that may restrict the accessibility of the Igf2 promoter to the shared enhancers by keeping Igf2 in an enclosed domain (Kurukuti et al., 2006; Murrell et al., 2004). Similarly, CTCF binding and its mediated chromatin loops is necessary for activation of IFNG during T-cell

differentiation (Hadjur et al., 2009). Although a number of experiments have proposed a model in which CTCF-binding sites mark the boundaries of separate domains, it is less clear what interactions stabilize CTCF/insulator interactions. As will be discussed in next the section, CTCF can recruit many cofactors to its binding sites, and the mechanisms stabilizing long-range interactions could be correspondingly complex.

4.1.2 Role of cohesin in CTCF-mediated looping interactions and transcription regulation

A variety of proteins such as CH8, YY1, Oct4 and RNA PolII interact with CTCF, which appears to be important or essential in assays for insulator activity or stabilization of long-range contacts (Wallace and Felsenfeld, 2007). Although the detailed molecular mechanisms that stabilize CTCF looping interaction are still under investigation, it has been discovered recently that the cohesin complex also interacts with CTCF. The CTCF-dependent cohesin binding has been observed at most of the CTSs, which is essential for looping formation as well as for insulating activity (Parelho et al., 2008; Stedman et al., 2008; Wendt et al., 2008). As illustrated in Figure 4.1a, the cohesin complex is composed of four protein components (SMC1, SMC3, SCC1/Rad21 and SA2) and only SA2 makes direct contact with a site on the C terminal tail of CTCF (Xiao et al., 2011). In addition, it has also been shown that the p68/SRA complex interacts with both CTCF and cohesin complex to stabilise the binding of cohesin (Figure 4.1b), and depletion of p68 or SRA results in loss of cohesin binding and imprinted expression at the IGF2/H19 locus (Yao et al., 2010).

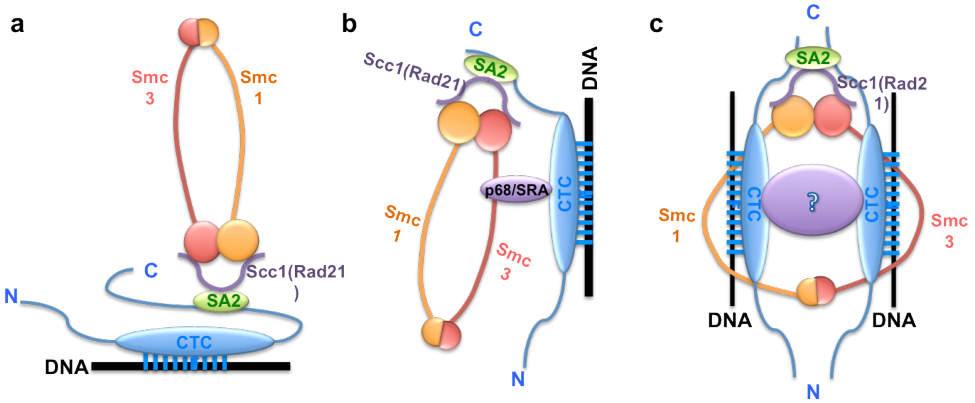


Figure 4.1: Schematics of the interaction between CTCF and cohesin. Panel a: via SA2 at C terminal tail of CTCF and panel b: p68/SRA mediated CTCF-cohesin interaction. Panel c: speculative model demonstrating how cohesin might help form and stabilise CTCF-dependent long-range chromosome loops.

It is not well understood how an encounter between chromatin fibres is initiated and subsequently stabilised, although it has been proposed to involve stochastic or directed chromatin movements. Once the spatial distance between two chromatin fibres is sufficiently close, the interaction can then be stabilized by protein-DNA complexes. The cohesin is one of the protein complexes which plays such a role in stabilising interaction between chromatin fibres. Various views of how the cohesin ring interacts with the chromosomes have been proposed (Onn et al., 2008). The handcuff model dictates that the cohesin ring may encircle a single sister chromatid (30 nm) and interact with a second ring containing the other sister when mediating cohesion between sister chromatids (Zhang et al., 2008). Another model, known as the embrace model, predicts that a single ring structure formed by the cohesin complex is capable of accommodating two 10-nm chromatin fibres, in which DNA is wrapped into nucleosomes (Gruber et al., 2003). The cohesin complex has demonstrated its ability to mediate interactions between chromosomes. Additionally, it has been proposed that the cohesin complex is able to stabilize the loop interaction between two CTCF insulators with the presence of SRA RNA and p68, by forming the base of the loop (Yang and Corces, 2011). Therefore, it is speculated that cohesin plays a key role in facilitating the formation of DNA loops via physically connecting different CTCF-binding sites (Figure 4.1 c).

A number of cases have been reported about the role of cohesin in CTCF mediated looping interactions. CTCF recruits the cohesin complex *in vivo* at most CTSs and the depletion of CTCF results in loss of cohesin binding (Kang and Lieberman, 2009; Rubio et al., 2008; Stedman et al., 2008; Wendt and Peters, 2009). Most importantly, cohesin is necessary for almost all the intra-chromosomal interactions mediated by CTCF, including H19/Igf2, β-globin, IFNG and APO loci (Hadjur et al., 2009; Hou et al., 2010; Mishiro et al., 2009; Nativio et al., 2009). For example, it has been shown that the depletion of cohesin components causes loss of imprinted gene expression at the Igf2-H19 locus, as well as loss of CTCF mediated looping interactions within the nucleus (Hadjur et al., 2009; Mishiro et al., 2009; Nativio et al., 2009). It suggests that the cohesin complex is essential for CTCF function and its ring structure may contribute to the ability of bringing distal CTSs to interact. However, the binding of CTCF is not affected by depletion of cohesin in most cases. Additionally, it is known that some cohesin can remain bound to DNA during interphase, suggesting that the binding of cohesin may be indirect and largely related to CTCF. Most recently, it has been shown that the

cohesin complex is localized at enhancers and promoters by interaction with mediator and subsequently stabilizes loop structure formed between enhancers and promoters, which demonstrate the role of cohesin in facilitating long-range interactions (Kagey et al., 2010).

4.1.3 Four CTCF bound elements in the TAL1 locus

Previous ChIP-chip studies on the TAL1 locus have identified a number of *cis*-regulatory elements bound by CTCF *in vivo* in erythroid K562 cells (Dhami et al., 2010). As shown in Figure 4.2 a, highly enriched CTCF bindings are observed at +57, +53, +40 and -31. For the CTCF binding site at +57, the reporter assay has shown no enhancer activity but insulator enhancer-blocking activity at this region, indicating its role as an enhancer-blocking insulator. In addition, high levels of repressive H3K27me3 marks and active H3K4me1, 2 marks have been observed upstream and downstream of the location of +57 respectively in K562 cells, which suggests the role of +57 as a boundary element. In the TAL1 expressing K562 cells, it is important to have a barrier insulator at this location in order to isolate the regulatory elements of TAL1 (such as the +51 and +20 enhancers) from those of CYP4A22 upstream as well as prevent the spread of repressive histone modifications into the TAL1 regulatory domain. For the binding of CTCF at +53 and +40, insulator enhancer-blocking activity has been shown for both regions. The +53 is also an active bio-directional promoter located at 2 kb upstream from the +51 enhancer. The PDZK1IP1 gene (+43) and its enhancer (+40) are also locate about 10 kb in distance (Follows et al., 2006). Given the complexity of the regulatory elements at this genomic region, it would be necessary for CTCF to mediate insulator functions at +53 and +40 to compartmentalise regulation of the TAL1 and PDZK1IP1 genes. For the CTCF binding at -31, multiple peaks of enrichment have been identified over at least a 2 kb interval centred at -31 by ChIP-chip assay, suggesting the role of CTCF may be complex at this element. Insulator enhancer-blocking activity of -31 has been confirmed using enhancer-trapping assays. Moreover, no other TAL1 regulatory elements have been identified so far, from -31 towards the STIL gene, suggesting the binding of CTCF at -31 and/or its nearby sequences functions as a boundary insulator between TAL1 and STIL regulatory domains. To sum up, an 88 kb TAL1 regulatory domain has been determined between two boundary elements at +57 and -31 based on the previous study.

4.1.4 The cohesin complex (Rad21) also binds to the CTSs in the TAL1 locus

Recently, the ChIP-seq data from the ENCODE project has revealed that Rad21 (a key component of cohesin) also binds to those four CTSs at the TAL1 locus in K562 cells. The ChIP-seq profiles of CTCF and Rad21 have been visualised using the UCSC Genome Browser (http://genome.ucsc.edu/) as illustrated in Figure 4.2 b. CTCF and Rad21 peaks identified by ChIP-seq at +57, +53, +40 and -31 align with the previous CTCF ChIP-chip data shown in Figure 4.2 a (Dhami et al., 2010). Motif sequences of these CTSs have also been denoted based on the consensus CTCF motif sequence from the Ren laboratory website (http://bioinformatics-renlab.ucsd.edu/rentrac/wiki/CTCF_Project). However, no corresponding motif is presented at -31 (Shane's thesis). No public ChIP-chip/seq data is available for binding of CTCF and Rad21 in HPB-ALL cells.

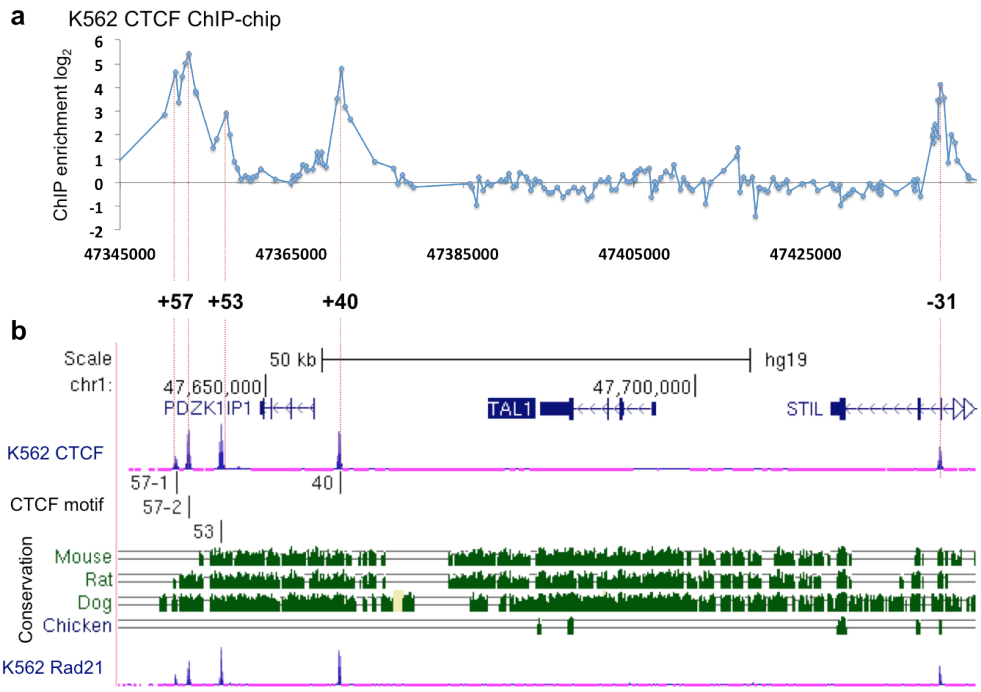


Figure 4.2: Diagrams of CTCF and Rad21 binding at the TAL1 locus. Panel a: ChIP-chip profile of CTCF binding in K562 cells. The x-axis represents the human chromosome 1 coordinates and the y-axis represents the ChIP fold enrichment (log_2). ChIP peaks are denoted based on the kb distance (+) upstream and (-) downstream to the TAL1 p1a. Panel b: A UCSC screenshot of ENCODE ChIP-seq profiles of CTCF (track 1) and Rad21 (track 4) at the TAL1 locus in K562 cells. Genes within the TAL1 locus are shown on top of the panel. CTCF motif (track 2) indicates the location of consensus CTCF sites based on data from the Ren laboratory website. Conservation (track 3) indicates those consensus sites that are conserved in other vertebrate genomes.

To summarise, a number of known facts about CTCF and cohesin in transcription regulation and looping formation as well as their relationship to the TAL1 locus are listed as follows:

1. A number of examples have shown that both CTCF and cohesin proteins are involved in mediating looping interactions as well as regulating gene expression (shown in section 4.1.1 and 4.1.2).

2. Previous studies using both ChIP-chip and ChIP-seq have shown that CTCF protein binds at four regions including +57, +53, +40 and -31 at the TAL1 locus in K562 cells. All these CTSs demonstrate either enhancer blocking or barrier insulator activities and define a TAL1 regulon containing all known regulatory elements related to TAL1 transcription.

3. The ChIP-seq data (ENCODE) has shown that the cohesin also binds to all four CTSs in K562 cells.

Thus, based on all of this, the binding of CTCF and cohesin may play a critical role in looping and transcriptional regulation at the TAL1 locus.

4.2 Aims of the chapter

A number of examples have illustrated that CTCF and/or cohesin play a key role in transcription regulation, either through mediating looping interactions between CTSs to facilitate or repress transcription activation or by stabilizing the chromatin loops formed between regulatory elements. It is known that the four CTSs at the TAL1 locus are all bound both by CTCF and cohesin in TAL1 expressing K562 cells. The barrier activity of insulators at +57 and -31 allows defining an 88-kb transcription domain at the TAL1 locus, which includes all known regulatory elements for the TAL1 transcription. In addition, CTCF and Rad21 binding at +40 is a blocker for crosstalk between the +51 enhancer and the TAL1 promoter. Consequently, it raises the question of how the CTCF and Rad21 may facilitate or prevent the transcriptional regulation at the TAL1 locus. Therefore, it is useful to determine whether the CTCF/cohesin mediated loops were formed between the four CTSs, and most importantly, their role in TAL1 transcription regulation.

The aims of the work presented in this chapter were as follows:

1. To apply ChIP-qPCR assay for the assessment of CTCF and cohesin binding sites at the TAL1 locus in both TAL1 expressing K562 and non-expressing HPB-ALL cells.

2. To investigate the possible looping interactions between CTSs at the TAL1 locus in K562 and HPB-ALL cells.

3. To investigate how CTCF and/or cohesin facilitate the transcriptional regulation of TAL1

4.3 Overall strategy

As discussed in chapter 3, the TAL1 promoter was able to interact with the +51 enhancer via looping in the erythroid K562 cells, regardless of the presence of a putative CTCF insulator (+40) in between. In K562, four key CTSs were previously identified by ChIP-chip assays, two of which (+53 and +40) were found to have enhancer-blocking activities while the other two CTSs (+57 and -31) were found to define insulator boundaries for an 88 kb TAL1 regulatory domain (Dhami et al., 2010). However, little was known about the role of CTCF and cohesin in regulating TAL1 expression. Therefore, it was necessary to determine whether the binding of CTCF and cohesin was dependent on TAL1 expression as well as to understand how CTCF and cohesin participated in transcription regulation of TAL1.

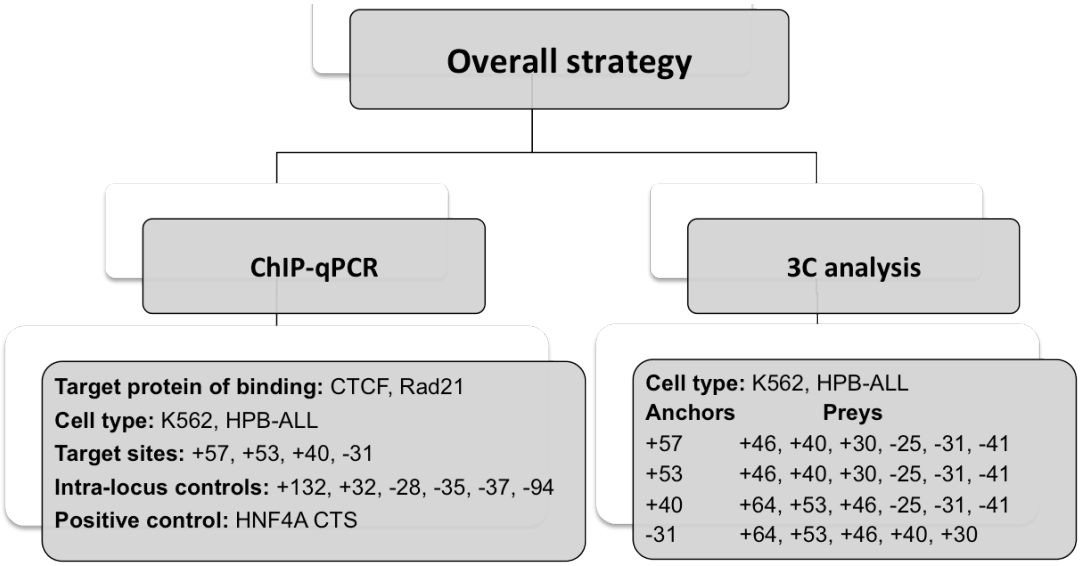


Figure 4.3: Schematics map of overall strategy of Chapter 4.

First, the ChIP-qPCR analysis was performed on TAL1 expressing and non-expressing cell lines with antibodies against CTCF and Rad21 proteins (Figure 4.3, left panel). The CTCF binding has been previously determined using the ChIP-chip assay in K562 cells (Dhami et al., 2010). For K562 cells, the ChIP-qPCR on CTCF was performed to analyse and validate the CTCF binding profile at the TAL1 locus. In addition, CTCF ChIP was also performed in HPB-ALL cells to determine the

CTCF binding profile across the TAL1 locus in a non-expressing cell line, and whether there were differences between K562 and HPB-ALL cells, which could relate to TAL1 expression. Moreover, Rad21 ChIP was performed in K562 and HPB-ALL cells to determine whether the CTCF and Rad21 bound in a cell-type specific means at the TAL1 locus. Second, the 3C analysis was performed in K562 and HPB-ALL cell lines to determine the looping interactions between four CTCF binding sites at the TAL1 locus and how these interactions could relate to transcription regulation of TAL1 (Figure 4.3 right panel). Each of the CTSs was treated as the 3C anchors accordingly. The 3C primers were designed for individual anchors to detect the looping interactions between the CTSs.

For each cell types, two bio-replicates were used to perform the ChIP and 3C analysis. In addition, three technical replicates were performed in the q-PCR and PCR detections. As illustrated in Figure 4.4, eleven genomic sites including six intra-locus control sites, four TAL1 CTSs and one positive control region (HNF4A CTS) were monitored for the CTCF and Rad21 binding using ChIP-qPCR. Furthermore, nine genomic regions including four TAL1 CTSs and five 3C control regions were assessed for the chromatin interactions using 3C.

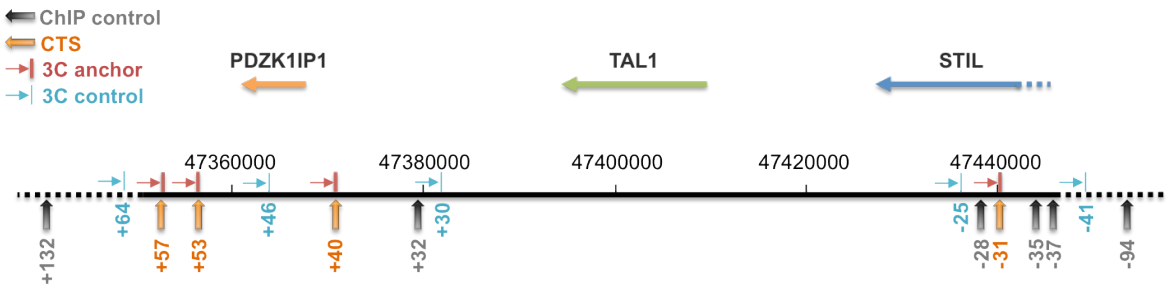


Figure 4.4: Schematics diagram of the CTSs and control regions assessed by the ChIP-qPCR and 3C analysis. The colour-coded arrows represent the genes across the TAL1 locus on top of the diagram. The black line represents the DNA fragment and the black bars and numbers correspond to the genomic coordinates of the TAL1 locus (NCBI Build 35). The grey and orange arrows represent the control regions and the CTSs of the ChIP-qPCR analysis respectively. The blue and red arrows represent the 3C control regions and the 3C anchors accordingly. The genomic regions are denoted based on the kb distance (+) upstream and (-) downstream to the TAL1 p1a.

Results

4.4 Characterisation of the CTCF and cohesin bindings at the *TAL1* locus by ChIP-qPCR assay

4.4.1 Assessing the specificity of Rad21 antibody in western blotting assays

In order to ensure that assays could detect *bona fide* CTCF and Rad21 (cohesin) binding sites, it was important to verify the specificity of the antibodies to be used in the ChIP-qPCR analysis and to avoid cross-reactivity with proteins which share amino acid sequence similarity. The CTCF antibody used in this thesis was previously validated (Shane D. PhD thesis, 2008), thus to this end western blotting assay was only performed for the Rad21 antibody (details shown in Appendix 2a). The predicted size of Rad21 is ~72 kDa and the antibody used in this study detected a major band at approximately the theoretical size in K562 cells (Figure 4.5). An additional band at ~50 kDa was also observed which is possibly a cleavage fragment of Rad21 protein (Pati et al., 2002). Taken together, it suggested that the antibodies used in this study were able to detect the *bona fide* CTCF and Rad21 proteins.

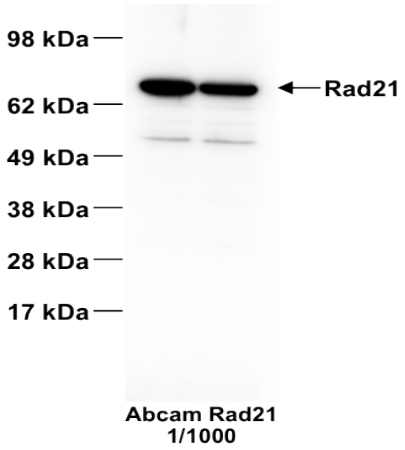


Figure 4.5: Western blot analysis for the characterisation of antibodies against Rad21.The nuclear protein extracts of two K562 bio-replicates were used for the western blot analysis. The predicted size of Rad21 protein is ~72 kDa. The arrow on the right of the blot indicates the predicted protein size of Rad21. Size markers are shown on the left of each panel.

4.4.2 Motif analysis of CTCF binding sites

In the previous study, the motif analysis had identified four consensus CTCF motifs corresponding to +57, +53 and +40, but no consensus CTCF binding motif was identified at -31. With the further understanding of CTCF binding and the high-throughput ChIP-seq technology, numbers of CTCF binding sites detected by ChIP analysis do not match the known consensus motif.

Given that abundant ChIP-seq data has largely extended our understanding of CTCF binding motif, the motif analysis was performed based on the latest CTCF consensus motif (Essien et al., 2009) to reanalyse the motifs of CTSs at the TAL1 locus.

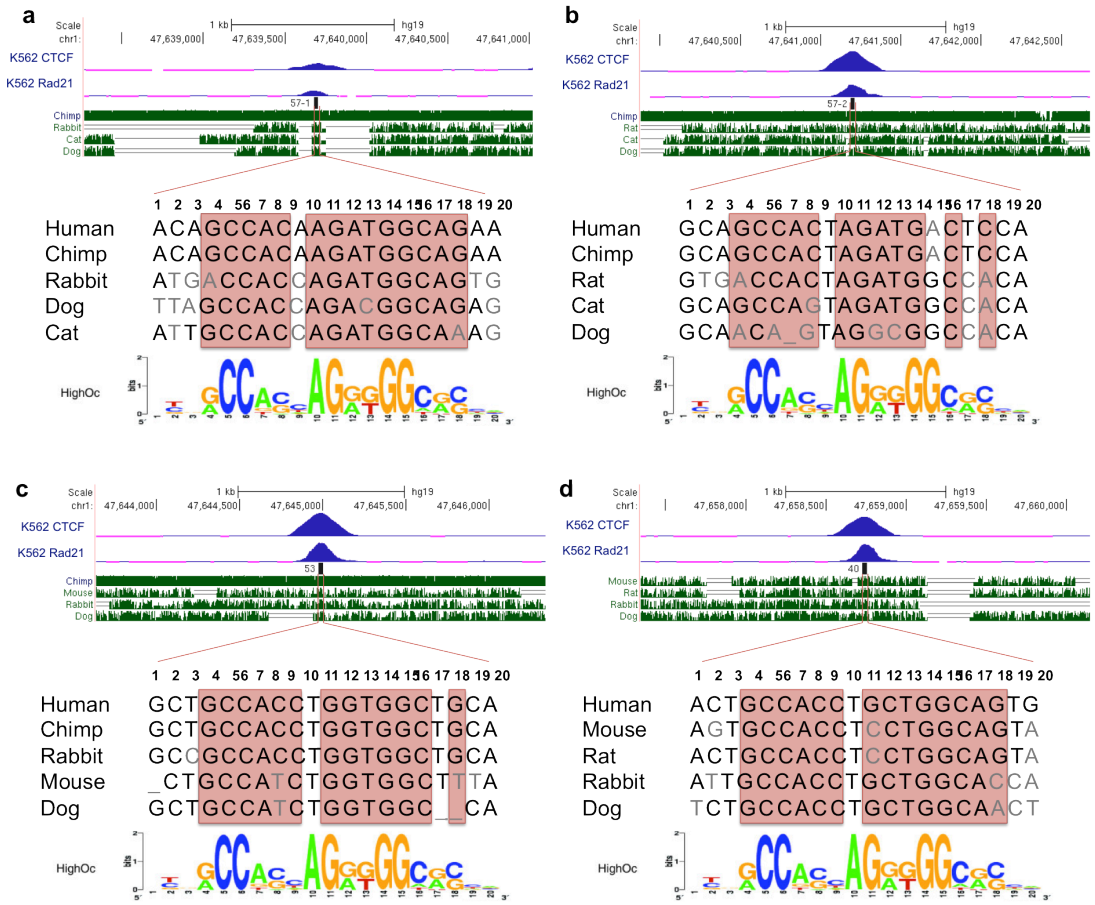


Figure 4.6: Motif analysis for CTCF binding sites at the TAL1 +57, +53 and +40. Panel a: CTCF motif 1 at +57, panel b: CTCF motif 2 at +57, panel c: CTCF motif at +53 and panel d: +40 CTCF motif at +40. On top of each panel, is shown a UCSC screenshot of ENCODE ChIP-seq profiles of CTCF (track 1) and Rad21 (track 2). CTCF motif (track 3) indicates the location of consensus CTCF sites based on data from Ren laboratory website. Conservation (track 3) indicates those consensus sites that are conserved in other vertebrate genome. On bottom of each panel, is shown a 20 mer human CTCF binding motif sequence of corresponding CTCF site as well as the results of multiple sequence alignments across vertebrate genome. The motif logo of consensus CTCF binding site is also represented at the bottom (Essien et al., 2009).

Based on the latest CTCF consensus motif sequence, four CTCF binding motifs were detected at +57 (two motifs corresponding to two ChIP-seq peaks), +53 and +40 (shown in Figure 4.6). The motif sequences identified at these four sites perfectly matched up with the results of previous analyses (S.Dillon's PhD thesis, 2008) which were performed based on the ChIP-seq data from Ren's laboratory (Kim et al., 2007). Additionally, multi-sequence alignments also illustrated that these CTCF motif sequences were highly conserved across the species (Figure 4.6).

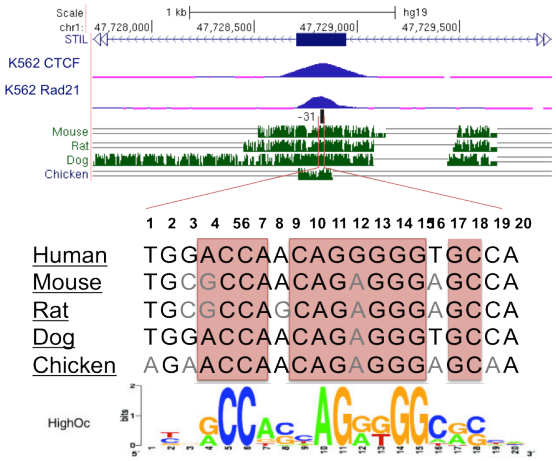


Figure 4.7: Motif analysis for CTCF binding site at the TAL1 -31 region. Rest of the details are as described previously.

Furthermore, the accumulated ChIP-seq data of CTCF provided a better insight of identifying the CTCF binding motifs which were previously omitted. In this assay, a highly conserved CTCF binding motif was identified at -31 as illustrated in Figure 4.7. The CTCF binding motif at -31 aligns up with the CTCF and Rad21 peaks as shown in Figure 4.7. High levels of sequence conservation have been shown across the vertebrate genomes. The sequence of CTCF motif at -31 matches up with the sequence feature of high-occupancy CTCF sites identified by ChIP-seq in human CD4+ T-cells (Essien et al., 2009). It also agrees with the high-level CTCF occupancy at -31 (as shown in section 4.3.2) detected by ChIP-qPCR analysis.

4.4.3 Determining CTCF and Rad21 binding patterns at the TAL1 locus

Four CTSs (+57, +53, +40 and -31) along with six negative control regions within the TAL1 locus and one positive control (CTS at the HNF4A locus) were assessed by the ChIP-qPCR. The q-PCR primer pairs for these regions were designed (described in Chapter 2) and ChIP assays using CTCF and Rad21 antibodies were performed as shown in the previous chapter (Chapter 3, section 3.4.1). ChIP enrichments ($\log_2$ scaled) of CTCF and Rad21 at all sites were presented in the histogram as shown in Figure 4.8.

High-level ChIP enrichments of CTCF were observed at +57, +40, -31 and a known HNF4A CTS in both K562 and HPB-ALL cells. The enrichment of CTCF at +53 was as strong as expected in K562 cells (Dhami et al., 2010). In TAL1 non-expressing HPB-ALL, CTCF enrichment was about 2-fold lower in comparison to TAL1 expressing K562 cells at +53 (Figure 4.8), which might suggest that the binding of CTCF at this site is transcription-specific.

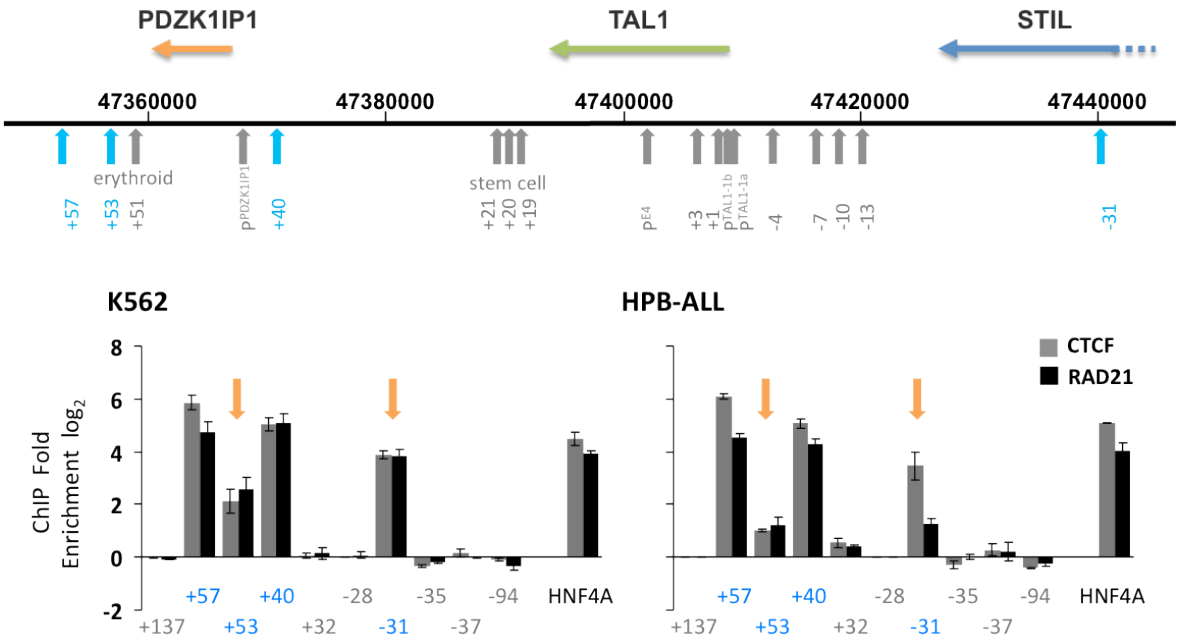


Figure 4.8: ChIP enrichments (log_2) of CTCF and Rad21 at sites of insulator activity (+57, +53, +40, -31) in human erythroid (K562) and lymphoid (HPB-ALL) cell lines. On top of the figure, a genomic track illustrates the known regulatory elements across the TAL1 locus. The CTSs are highlighted in blue. Orange arrows show Rad21 recruitment at -31 is lower in HPB-ALL. ChIP enrichments (log_2) are shown with standard errors. Annotation of test and negative control regions is denoted in black and grey text respectively.

Similarly, significant ChIP enrichments of Rad21 were identified at +57, +40, -31 and the HNF4A CTS in K562 (Figure 4.8). ChIP enrichment of Rad21 at +53 was to the similar level of CTCF in K562. In HPB-ALL, high-level of Rad21 bindings were only observed at the +57, +40 and HNF4A control. In additional, it showed approximately 4-folds lower of Rad21 occupancy at +53 and -31 in HPB-ALL than K562 cells (Figure 4.8). It also implied that Rad21 occupancy at +53 and -31 might depended on expression of TAL1, as higher enrichments of Rad21 at these two sites only appeared in TAL1 expressing cells. Additionally, it also suggested that the significant binding of CTCF does not necessarily contribute to the high-level of Rad21 binding at the CTSs.

4.5 Determination of looping interactions between CTSs at the *TAL1* locus

In order to further characterise the chromatin configuration of the TAL1 locus and determine the role of these looping interactions in regulating the TAL1 expression, the 3C analysis was performed using four CTSs as “anchors” to determine whether these genomic regions also participated in chromatin looping interactions in TAL1 expressing and non-expressing cell lines. It would help to understand how these looping interactions between CTSs might contribute to the TAL1 expression.

The 3C-PCR assays were performed as previously described in Chapter 3 and ligation events between the anchors and CTSs or the control regions were detected and presented as histograms in Figure 4.9. Using a Csp6I restriction fragment containing the TAL1 +57 as the anchor, 3C-PCR analyses were performed to detect the interactions between +57 and other CTCF sites across the TAL1 locus. In both human K562 and HPB-ALL cell lines, the ligation frequency decreased with respect to distance from the anchor for all three control regions located at +46, +30 and -25 (Figure 4.9 a). However, it was significantly (p = 1.6 x 10^{-5}) increased for the CTSs at the -31 region in K562 cells while no significant interaction was observed in HPB-ALL cells (Figure 4.9 a). It indicated that a looping interaction between CTSs at +57 and -31 occurred specifically in the TAL1 expressing cell line (K562).

Similarly, 3C was performed using +53 (a adjacent CTS of +57) as the anchor. Looping interactions were assessed between the anchor (+53) and +40 and -31 along with the appropriate controls juxtaposed between them. As shown in Figure 4.9 b, the 3C ligation frequency gradually ceased as a function of distance from +46 to +30 in both K562 and HPB-ALL cells, indicating that no looping interactions occurred between +53 and +40 in both cell lines. In addition, significant interaction between the anchor and -31 (p = 2.4 x 10^{-5}) was only observed in K562 cells but not in HPB-ALL cells (Figure 4.9 b). These observations were in agreement with interaction profiles using +57 as the anchor (Figure 4.9 a). Thus, both +57 and +53, two closely spaced CTSs, showed identical 3C patterns of interactions in both the TAL1 expressing and TAL1 non-expressing cell lines – at least for the regions tested here.

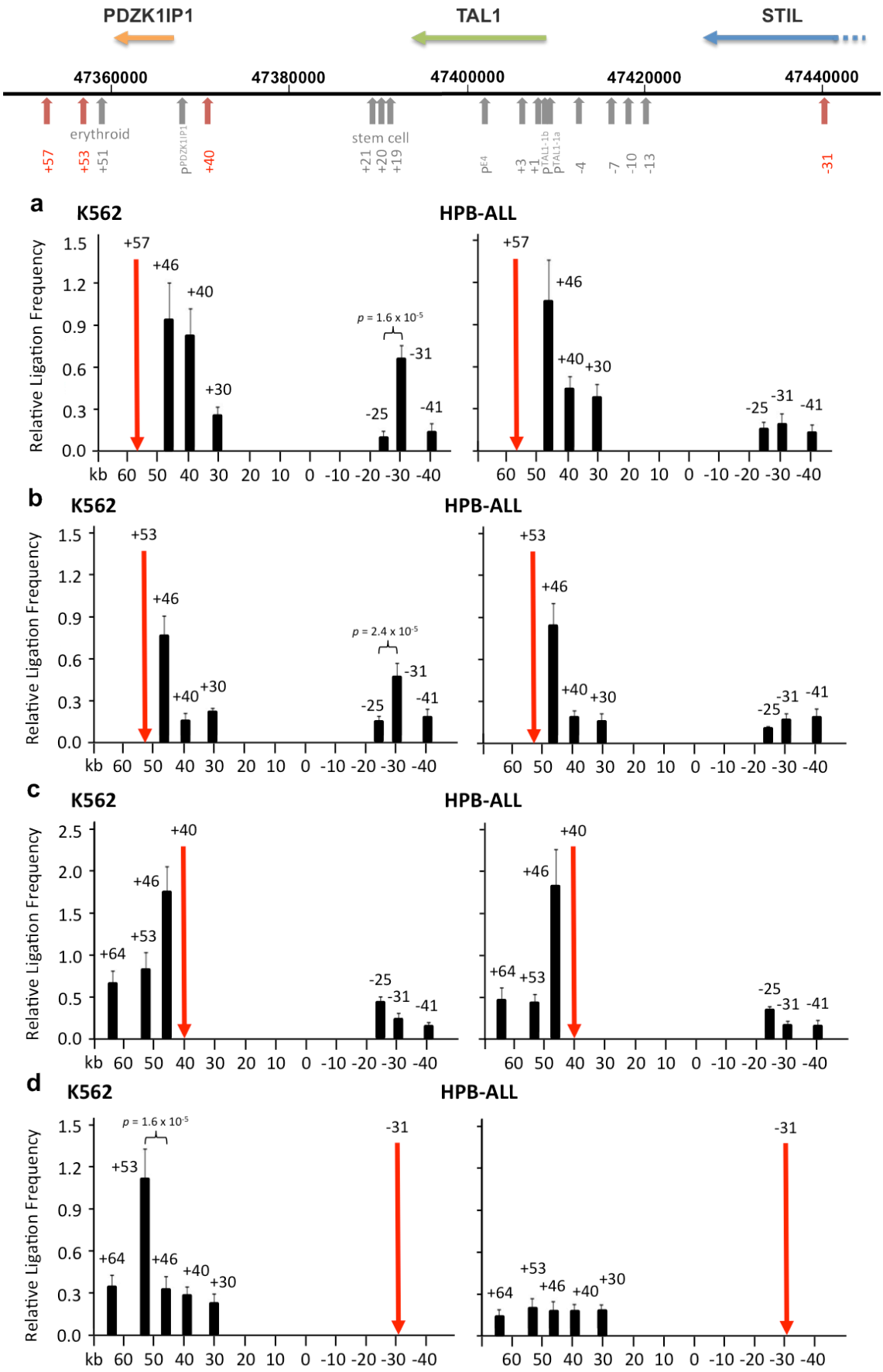


Figure 4.9: Bar diagrams of looping interactions between insulators determined by 3C in human erythroid (K562) and lymphoid (HPB-ALL) cell lines. On top of the figure, a genomic track illustrates the known regulatory elements across the TAL1 locus. The 3C anchors (CTSs) are highlighted in red with arrows. Interaction frequencies (black bars) are shown with standard errors and normalized relative to BAC controls. Locations of 3C anchors are denoted with vertical red arrows. Locations of genes and their directions of transcription

are shown at the top of figures. p values are indicated for interaction frequencies which are significantly higher for test regions when compared to those of control regions. Scales (in kb) are shown at the bottom.

The lack of looping involving the CTS at +40 was further validated using the +40 region as the anchor for 3C analysis. As shown in Figure 4.9c, the ligation frequencies between the anchor (TAL1 +40) and all regions tested decreased as the function of distance from the anchor in both K562 and HPB-ALL cells – with no significant increases in ligation frequencies at the CTSs. This further confirmed that the +40 region did not participate in any looping interactions with other CTSs at the TAL1 locus in K562 and HPB-ALL cells.

Similarly, further validation of looping interactions involving the CTS at -31 was performed by taking the -31 region as the 3C anchor. As shown in Figure 4.9d, ligation frequency was relatively low at the +30, +40 and +46 regions, but significantly increased at the +53 region in K562 cells (p = 1.6 x 10^{-5}). In contrast, ligation frequencies were consistently at low levels across the genomic region from +30 to +64 in HPB-ALL cells, indicating detection of random ligation events with no *bona fide* looping interaction observed in HPB-ALL cells (Figure 4.9d). These results further supported the existence of looping interactions involving CTSs at -31 and +53 only in TAL1 expressing cell line (K562).

Discussions

4.6 Transcriptional dependent binding of CTCF and Rad21 at the TAL1 locus

CTCF and Rad21 binding profiles in K562 cells captured by ChIP-qPCR are in agreement with the previously reported profiles using both ChIP-chip (only CTCF) (Dhami et al., 2010) and ChIP-seq (ENCODE) methods. As shown in result sections, CTCF and Rad21 bind to all four CTSs in both K562 and HPB-ALL. However, there are similarities as well as differences in the level of enrichments at these sites between the cell lines. The similarities between K562 and HPB-ALL cells including both CTCF and Rad21 are highly enriched at +57 and +40 as well as a high-level of CTCF binding at -31. The differences between K562 and HPB-ALL cells including ChIP enrichments of both CTCF and Rad21 at +53 as well as ChIP enrichment Rad21 at -31. The reduced level of CTCF and Rad21 occupancies at +53 in HPB-ALL may suggest that the binding of CTCF at this site

is cell-type specific and recruitment of Rad21 is CTCF-dependent. In addition, it has been noticed that the level of CTCF occupancy at -31 is very similar between K562 and HPB-ALL, which implies that the reduced level of Rad21 occupancy at -31 in HPB-ALL are unlikely to be due to the level of CTCF binding. It is speculated that -31 may be a tissue-specific binding site to the cohesin complex, and Rad21 can be loaded to different levels at -31, probably depending on the transcriptional state of TAL1.

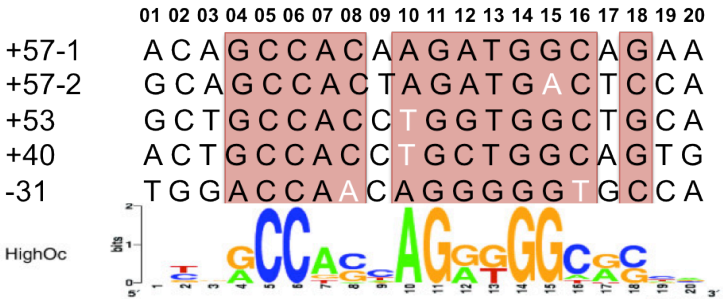


Figure 4.10: Comparison of CTCF binding motifs at the CTSs. Sequence alignment of 20-mer CTCF binding motifs is shown between CTSs at +57, +53 +40 and -31. DNA bases aligned up between sequences and consensus CTCF motif (Essien et al., 2009) are highlighted in pink.

The level of CTCF and Rad21 occupancy also varies between CTSs at the TAL1 locus. The lowest bindings of CTCF and Rad21 have been observed at +53. Initially, it is speculated that the reduced binding efficiency of CTCF at this site may due to the variety of CTCF motif sequence between CTSs. However, the comparison between CTCF binding motifs at different CTSs has observed that all five motif-sequences share high similarity and align up with the consensus CTCF motif sequence (Figure 4.10). It demonstrates that the cell-type specific CTCF binding at +53 cannot be due to the sequence difference of CTCF motif at the CTSs. Thus, it is proposed that the binding of CTCF at +53 may be facilitated by co-factors of CTCF or even other transcription factors such as GATA1, as much lower CTCF occupancy has been shown at +53 in the lymphoid HPB-ALL cells where GATA1 is not expressed. Overall, the differences in binding between K562 and HPB-ALL for CTCF and/or cohesin suggest that these may be sites (+53 and -31), which are context-dependent.

4.7 Regulating TAL1 expression via looping interactions between CTSs

The role of CTCF in mediating formation of chromatin loops have been previously demonstrated at the H19/IGF2, β-globin and major histocompatibility complex

(MHC) class II loci (Kurukuti et al., 2006; Majumder et al., 2008; Splinter et al., 2006). Furthermore, the first direct evidence of loop formation *in vivo* has been provided by Dean and colleague, showing that two CTCF-bound insulators of a human ectopical insert forms a loop in transgenic mice (Hou et al., 2008). The chromatin loops formed between CTSs at the TAL1 locus can either facilitate or prevent the transcription, depending on the combinations of their interacting partners. There are three different speculated models that elucidate how the looping configurations formed between CTSs can possibly alter the expression of TAL1 in the cells where it is appropriate.

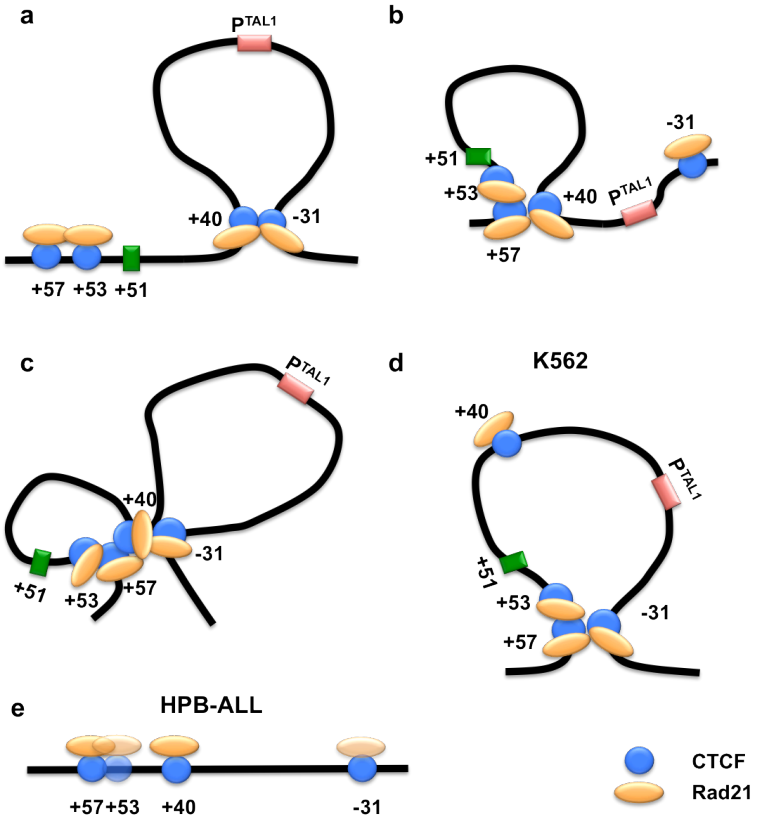


Figure 4.11: Speculated models of looping interactions between CTSs at the TAL1 locus. Panel a: loop between CTSs at +40 and -31. Panel b: loop between CTSs at +57/53 and +40. Panel c: loops between all four of CTSs. Panel d: loops detected by 3C in K562 cells. Panel e: no loops detected by 3C in HPB-ALL cells. The chromatin is represented by black line. The enhancer (+51) and the TAL1 promoter are shown in green and red boxes respectively.

For the first looping model, a repressive loop can be formed between CTSs at +40 and -31 as illustrated in Figure 4.11a. This looping configuration greatly reduced the chance of contact between the TAL1 promoter and its +51 enhancer and possibly results in down-regulation of the TAL1 expression. Alternatively, the second looping model illustrates how the +51 enhancer is isolated from its promoter via a loop formed between CTSs at+57/53 and +40 (Figure 4.11 b). Thus, the involvement of the CTS at +40 in any looping interactions with the rest of

the CTSs would be a major obstacle for TAL1 expression in the erythroid lineage, where the TAL1 transcription is under regulation of the +51 erythroid enhancer. Similarly, 3C experiments in the H19/Igf2 locus have shown that the CTCF and/or cohesin mediated interactions plays an important role in imprinted gene expression by relocating relative regulatory elements into different loops. It can either keep the Igf2 promoter in an enclosed loop from its shared enhancers in the allele-specific manner (Kurukuti et al., 2006) which is similar to the first model (Figure 4.11a); or create a loop between CTSs at the ICR (imprinting control region) and downstream shared enhancer enclosing all the enhancers to prevent contacts from the Igf2 promoter (Nativio et al., 2009), which is similar to the second model as proposed (Figure 4.11b). In addition, enhancer-blocking function via CTCF loops has also been demonstrated in vivo at a non-imprinted β-globin (Hou et al., 2008).

Nevertheless, it has been shown that the TAL1 promoter is capable of interacting with the +51 enhancer regardless of the +40 CTCF insulator in K562 cells. This illustrated the fact that in order to crosstalk to its promoter, the enhancer was able to circumvent the effect of a CTCF insulator at +40 by utilising a looping mechanism. However, contact between the +51 enhancer and its promoter may become more difficult, if CTS at +40 participates in looping interactions with other CTSs as illustrated in first two models. Similar to the above two models, a more sophisticated looping model is proposed as shown in Figure 4.11c. It brings all CTSs together and situates the +51 enhancer and the TAL1 promoter into two independent loops, which also has negative impact on the TAL1 transcription.

For the fourth looping model, it demonstrates that looping interactions between CTSs excluding the one at +40 is also the observed configuration in K562 cells. In K562 cells, looping interactions have been identified between three CTSs at +57, +53 and -31 and the speculated looping configuration is shown in Figure 4.11d. For this configuration, the CTS at +40 does not form any loop with other three sites, as it is highly favoured for the TAL1 transcription in the erythroid cells, by disposing the +51 erythroid enhancer and the TAL1 promoters into the same loop along with other known regulatory elements of TAL1. In K562 cells, the chromatin loop formed between CTSs at +57/53 and -31 is the only configuration that can possibly accommodate both the +51 enhancer and the TAL1 promoter within a single chromatin loop. In addition, this looping configuration also defines a TAL1

regulon that accommodates all known *cis*-acting elements related to transcription regulation of TAL1 (Figure 4.11 d). The similar chromatin architecture facilitated by CTCF and/or cohesin has also been reported at the β-globin locus as previously discussed in section 4.1.1.

Although high level of CTCF and Rad21 binding were observed at most of these sites, no looping interaction formed between CTSs in HPB-ALL cells where the TAL1 expression is repressed (Figure 4.11e). As in the β-globin locus, it has shown a linear chromatin structure in non-expressing brain cells, irrespective of CTCF binding at HS85 (Phillips and Corces, 2009). Looping interactions between CTSs are only observed in the TAL1 expressing K562 cell but not in non-expressing HPB-ALL cells, indicating that these interactions are context-dependent. In particular, it is speculated that the lower level of Rad21 binding at -31 can be either the reason or the consequence related to the lack of CTCF loops in HPB-ALL cells.

The context-dependent looping interaction may suggest that the loops formed between other cis-acting elements at the TAL1 locus may be prior to the formation of CTCF loops. However, the molecular machinery of how the looping interactions between CTSs regulate the TAL1 transcription in the erythroid K562 cells is still unclear. It is postulated that the CTCF loop can facilitate the TAL1 transcription by accommodating the +51 enhancer and its promoter into the same chromatin loop. Additionally, the cohesin ring structure formed at the bottom of this CTCF loop can help stabilising the GATA1/TEC-mediated chromatin loops formed between other regulatory elements within the TAL1 regulon (Dhami et al., 2010).

4.8 A putative 3D organisation of the TAL1 locus

So far, looping interactions determined by 3C in human erythroid cells (K562) are listed as follows:

- +57 & -31 (mediated by CTCF and cohesin)
- +53 & -31 (mediated by CTCF and cohesin)
- P^{TAL1} & +51 (mediated by the TEC complex)

- P^{TAL1} & +21/20/19
- P^{TAL1} & -10/9

By integrating new loops identified in this chapter, a modified model of TAL1 looping configuration has been proposed as illustrated in Figure 4.12.

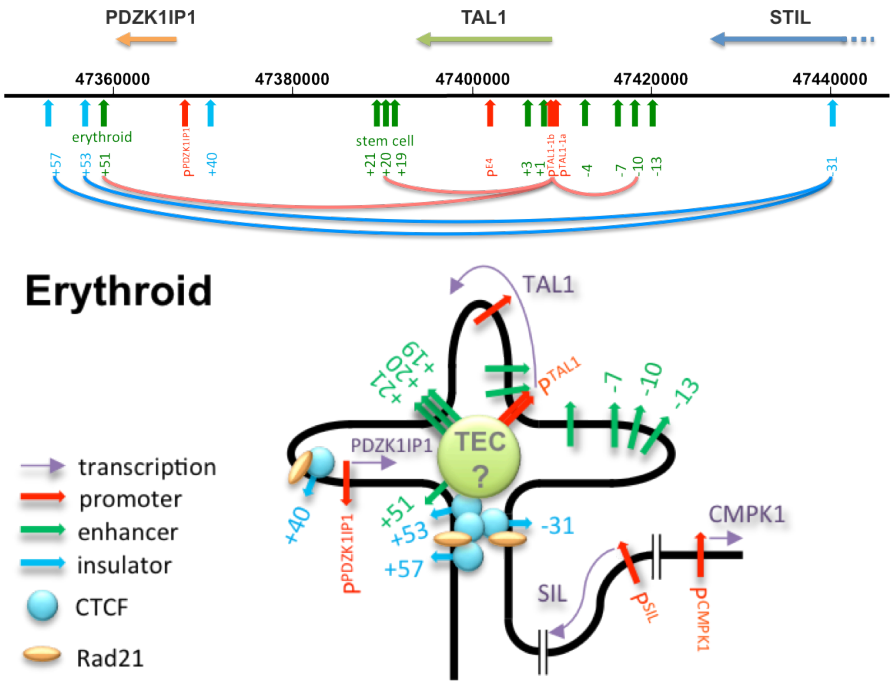


Figure 4.12: Putative structural organisation of the TAL1 chromatin hub in $TAL1^{+}$ erythroid cells. Promoters, enhancers and CTCF binding sites are the vertical red, green and blue arrows respectively. The erythroid (+51) and the stem cell (+19 to +21) enhancers are highlighted. Interactions between the TAL1 promoter and its regulatory elements and between CTCF binding sites are shown with red and blue loops respectively. Direction of transcription of relevant genes (purple arrows) and CTCF and Rad21 binding at insulators are also shown and detailed in the key. The TEC (putative) mediating the interaction of the TAL1 promoters and other regulatory elements in erythroid cells are represented by the light-green ball.

In TAL1 expressing K562 cells, three enhancers (+51, +21/20/19 and -10) and the promoters (1a and 1b) of the TAL1 are brought into the spatial close-proximity by the TEC complex to the centre of the active chromatin hub (as described in Chapter 3, section 3.7.5). Additionally, the looping interactions between the +57/53 and the -31 insulator elements form the stem of the active chromatin hub (Figure 4.12). The CTCF and cohesin mediated interactions can facilitate and stabilize the formation of a “cruciform” structure between *cis*-acting elements at the TAL1 locus.

As shown in Figure 4.12, the +57/-31 interaction accompanied with the binding of CTCF and cohesin provides a stem structure to help in stabilising the “cruciform”

configuration that forms between TAL1 *cis*-regulatory elements. However, the -31 region is located at exon16 of the STIL gene. Given that the STIL gene is also transcribed in K562, the interaction between CTSs at +57 and -31 may become an obstacle for full-length transcription of STIL. Consequently, it is speculated that the STIL gene is only transcribed in a proportion of cells that are not expressing TAL1. The dynamic of chromatin configuration is able to accommodate possible looping formations at the TAL1 locus as well as coordinate the transcription of TAL1 and its neighbouring genes.

Additionally, it has been shown that the +51 enhancer is brought together with -31, as +57/53 located only several kb upstream of the +51 enhancer (Figure 4.12). This may also imply that the TAL1 promoter and -31 should also be in relatively close-proximity as a result of +51-P^{TAL1} interaction. Furthermore, it is also important to demonstrate the GATA1/TEC-mediated three-way interactions between P^{TAL1} and two enhancers (+51 and +21/20/19). Therefore, in order to demonstrate the TAL1 “cruciform” configuration in K562 cells, further analysis is required to prove the additional co-localizations between P^{TAL1} and -31, between the erythroid (+51) and stem cell (+21/20/19) enhancers as well as between the P^{TAL1} and +57/53.

Most importantly, the entire “cruciform” model of the TAL1 locus is built on the assumption that all the looping interactions detected by 3C are present in the same cells at the same time. As discussed in section 4.5.2, it is speculated that the context-dependent binding of Rad21 at -31 and CTCF/Rad21 at +53 may be facilitated by the GATA1 and TEC complex. A recent study has shown that the looping interaction between HS5 and 3’HS1 in the b-globin as well as the binding of CTCF at these sites are significantly reduced by GATA1 knockdown in K562 cells (Woon Kim et al., 2011). It supports this speculation by illustrating the way a transcription factor can facilitate CTCF recruitment and looping interaction. In order to prove that in the TAL1 locus, one needs to demonstrate that the entire cruciform configuration can be disrupted by knocking-down the key transcription factors mediating the looping interactions. Given that GATA-1 is speculated to mediate the primary looping structure between the TAL1 promoter and the +51 enhancer, 3C can be performed to determine to what extend the TAL1 “cruciform” configuration is being disrupted after knocking-down GATA-1 by siRNA, in order to prove whether all interactions were in the same cells at the same time.

Conclusions

The work described in this chapter has demonstrated that CTCF and cohesin bind to the CTSs in TAL1 expressing and non-expressing cell lines in a context-dependent manner. Consequently, a context-dependent looping configuration formed between four CTSs has been identified in TAL1 expressing K562 cells, which is speculated to being mediated by CTCF and cohesin. Knocking-down of CTCF and/or cohesin will enable us to determine the involvement of these two proteins in formation of loops between CTSs. Based on the interaction data detected by 3C, a modified TAL1 looping configuration has been proposed that accommodates both the TEC complex (shown in Chapter 3) as well as the CTCF and cohesin complex at the TAL1 locus in K562 cells. So far, the chromatin interactions between known *cis*-acting elements at the TAL1 regulon have been determined. However, it is important to understand whether or not other regulatory elements outside of the regulon were also involved in TAL1 transcription. In the following chapter, a high-throughput 4C technique will be applied, which will further extend our understanding of TAL1 regulation, to look at the distal regulatory elements located outside of the TAL1 regulon. In addition, it has not yet been proven that all interactions exist in the same cells as those expressing TAL1 at this stage. The GATA1 siRNA knocking-down experiments conducted in chapter 6 would allow addressing these particular issues.

Chapter 5 Establishment and optimisation of the 4C-array method to study the *TAL1* Locus

Summary

A 4C variant was adapted from the enhanced-4C (e4C) technique in order to study the chromatin *cis*-interactions at the TAL1 locus. With the adjustments of starting 3C material and PCR cycling numbers, the optimised 4C method provided high sensitivity and reproducibility in detecting interactions using the TAL1 microarray platform. Subsequently, the optimised 4C-array method was applied to study chromatin interactions between the TAL1 promoter (anchor) and other genomic elements across the TAL1 locus in human TAL1 expressing (K562) and non-expressing (HPB-ALL) cell lines. A number of significant interactions were captured by 4C-array. They were classified into three groups including i) the known interacting partners previously detected via 3C (TAL1 enhancers at +51, +21/+20/+19, +10); ii) the known regulatory elements predicted to be interacted with P^{TAL1} (the CTSs at +57/+53 and -31 and the repressor at -13) and iii) the novel interacting co-associates ($P^{PDZK1IP1}$ and STIL exons and introns, especially the STIL+1). The number of significant interactions and the level of ligation frequencies were much higher in the TAL1-expressing K562 cells comparing to the non-expressing HPB-ALL cells. The interaction data detected by 4C-array were not only in agreement with the previous 3C profiles but also identified numbers of new interacting partners, which fully supported the “cruciform” model at the TAL1 locus. Additionally, it was found that the P^{TAL1} and STIL+1 elements aligning up with the breakpoints of STIL-TAL1 fusion which were situated in spatial close-proximity, suggesting a possible mechanism of STIL-TAL1 rearrangement in the T-ALL patients.

5.1 Introduction

The 3C-PCR results provide us with a snapshot of some of the looping interactions in the TAL1 locus. However, the method is limited because it requires prior knowledge of all the locations of regulatory elements being examined in the specific locus of interest. Different from 3C (“one versus one”), 4C is known as a “one versus all” strategy which selects a defined single “anchor” to screen for its

contacting partners genome-wide. The 4C approach used in this Chapter, has obvious advantages which are as follows:

(i) No prior knowledge of the interacting looping partners is required apart from the selected anchor sequence.

(ii) The method is higher-throughput comparing to 3C and is used in combination with the genomic tiling microarray or next-generation sequencing.

(iii) The method is capable of detecting interactions between the anchor sequence and any other genomic regions specified on the microarray. When used with whole genome microarrays or next-generation sequencing, it is capable of detecting interactions between the bait and any other region genome-wide.

In this Chapter, the 4C-array method (4C in combination with microarrays) was used to monitor the looping interactions across the TAL1 locus. The compositions of the 4C libraries were assayed by using a TAL1 genomic tiling path array covering the human and murine TAL1 loci (spanning 256 kb and 207 kb of the human and murine TAL1 loci, respectively). As part of this analysis, improvements to the method were developed to make it more robust and reproducible in assaying for chromatin looping interactions.

5.2 **Aim of the chapter**

Looping interactions between the TAL1 promoter and its three distal enhancers (+51, +19, -10) have been identified in human and murine cells using 3C analysis as illustrated in Chapter 3. Additionally, it has also been detected that CTSs at +57/+53 interact with the CTS at -31 in human K562 cells in the previous chapter. By viewing the data as a whole, it suggests that the CTSs at -31 may also be in close proximity with the +51 element and consequently be co-localised with the TAL1 promoter in K562 cells. Although primary 3C analyses provide abundant information about chromatin interactions between several regulatory elements at the TAL1 locus, the spectrum of interaction detection is restrained by limited throughput of 3C technology and relationships between the TAL1 promoter and a number of other known regulatory elements are still unclear. The technical advance offers a high-throughput 4C analysis, which is capable of profiling

genome-wide chromatin co-localisation from one viewpoint (anchor) to other genomic elements without requiring prior knowledge of the location of putative interacting co-associates. Providing that the TAL1 genomic tiling path microarray provides a high-resolution and robust analysis platform (P. Dhami's PhD thesis, University of Cambridge, 2008), 4C in conjunction with array platform can be used to identify loci-wide interactions from any chosen viewpoint (anchor) to the rest of genomic elements within the TAL1 locus. To this end, the aims of the work described in this chapter were:

1. To assess performance of the 4C-array method in terms of its sensitivity and reproducibility

2. To optimise the 4C-array method using the sensitivity and the reproducibility as benchmarks

3. To further characterise chromatin interactions between the P^{TAL1} and other genomic region across the TAL1 locus using 4C-array

5.3 **Overall strategy**

This Chapter describes the use of 4C in combination with microarrays to further elaborate *cis*-regulatory looping interactions at the TAL1 locus. For this work, the promoter of TAL1 was chosen as the anchor to explore all possible interactions between the anchor and other genomic DNA sequence elements within a 256 kb segment spanning containing the TAL1 locus. A genomic tiling path array spanning this region (resolution approx. 400 bp) was used to detect these interactions captured from the 4C samples.

The 4C-array method used in this chapter was adapted from the enhanced-4C (e4C) procedure (Fraser's laboratory). To establish the method, K562 cell line was used in the subsequent assessments. The method was systematically assessed at four different levels (Figure 5.1, step 1) using two criteria as follows:

(i) Sensitivity (Comparison with 3C data in turns)

(ii) Reproducibility (Comparison between replicates)

Subsequently, the 4C-array method was optimised for its library complexity and PCR cycling numbers (Figure 5.1, step 2) and was then re-assessed with the criteria listed above. After optimisation, 4C-array was applied to study chromatin interactions at the TAL1 locus in both K562 and HPB-ALL cells (Figure 5.1, step 3).

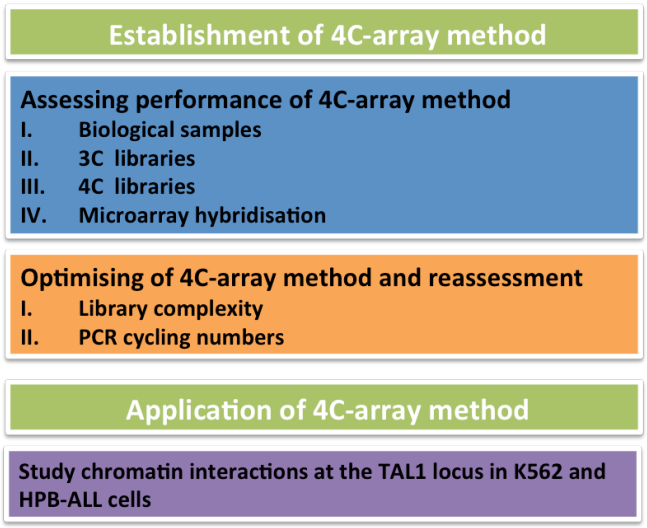


Figure 5.1: Overall strategy employed to establish and optimise the 4C-array method. This flow diagram shows an overview of the strategy employed to establish and optimise the 4C-array method in K562 cells. Subsequently, the optimised 4C-array method was used to study chromatin interactions at the TAL1 locus in K562 and HPB-ALL cells.

Results

5.4 Establishment of 4C-array

5.4.1 *Overall procedure of 4C-array*

The concept of 4C is to capture the interactions between one particular region, normally the promoter of a gene and its all-possible interacting partners genome-wide. The 3C library contains ligation products of all interactions in a particular cell sample while the 4C library contains a subset of the interactions contained within a 3C library. As shown in Figure 5.2 (steps 1 to 3), generating an initial 3C library was part of the 4C procedure. For all available 4C variants (circular-3C, 3C-on-chip and enhanced-4C), the chromatin is digested with a 6-base cutter, such as HindIII (Gondor et al., 2008; Simonis et al., 2006; Zhao et al., 2006); or subsequently followed by a second digestion with a frequent 4-base cutter, such as DpnII or NlaIII (Simonis et al., 2006). The latest variant of 4C technologies (e4C) has been used for profiling genome-wide chromatin co-associations (Sexton et al., 2012). In this Chapter, the e4C has been subsequently optimised for studying *cis*-interactions at the TAL1 locus in combination with the TAL1 specific tiling path

array, known as 4C-array. In the 4C-array method, two major alternations have been made comparing to its original version. Firstly, a 4-base cutter (Csp6I) was introduced to the step of chromatin digestion in order to provide a higher resolution comparing to the 6-base cutter used in circular-3C, as the average size of the Csp6I-digested fragments was over 10-fold smaller than HindIII-digested ones, with only approx. 700 bp at the TAL1 locus. Secondly, the resultant 3C DNA was than sonicated (as shown in Figure 5.2, step 4) instead of secondary restriction digestion (3C-on-chip and e4C) during the 4C-array preparation (note: the starting amount of 3C material for making a 4C library is referred to as “library complexity” in this Chapter). Sonication reduced the size of ligation products to fragments to less than 500 bp (to ensure adequate primer extension of the complete fragment at later steps during the 4C procedures). In comparison with the secondary restriction digestion, sonication has two advantages in fragmenting the DNA molecules. First, fragmentation by sonication is independent of the DNA sequence itself while cutting by endonucleases are restricted to the location of their DNA recognition sites. It means that sonication cuts more randomly and uniformly in comparison with endonucleases. Secondly, a sonication step only takes less than 10 minutes in comparison with a restriction digestion which takes over an hour. It has been shown that sonication is an un-biased way in comparison to the second restriction digestion for fragmenting 3C chromatin (Fullwood and Ruan, 2009). Based on the reasons stated above, sonication is considered as a relatively un-biased and more efficient way in fragmenting 3C DNA molecules.

Subsequently, a quality control (QC) step was conducted to assess the quality of sonication (described in 5.4.3). As illustrated in Figure 5.2 (step 5), a biotinylated primer complementary to the “anchor” sequence was then used for primer extension – thus priming only those ligation products containing the region of interest. In step 6, the ligation products of interest marked at their 5’ end with biotin were purified from the remainder of the 3C DNA by streptavidin beads. The ends of interacting partners were polished by blunt-ending and ligated to a PCR adapter. In step 7, the “anchor”–specific 4C fragments were PCR amplified with an anchor-specific nested primer in combination with a primer directed to the adapter sequence. This was performed while the template DNA was still attached to the beads. The PCR products, so called the 4C library was taken for the further downstream application. The 4C library was then labelled and applied for hybridisation to the TAL1 genomic tiling microarray.

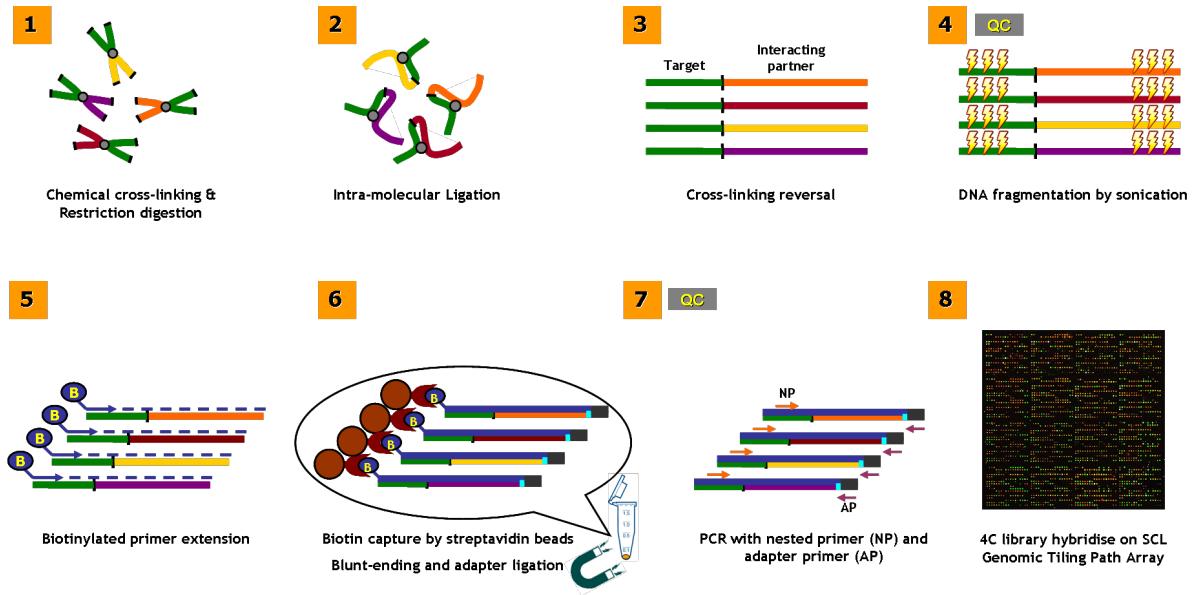


Figure 5.2: Overall procedure of the 4C library preparation. Steps 1 to 3 demonstrate the 3C procedures of chemical cross-linking, restriction endonuclease digestion, intra-molecular ligation and denaturation of crosslinked protein. Step 4 shows the sonication step which reducing the average size of ligated fragments to avoid the incomplete primer extension in the following step. Step 5 shows the primer extension step with a biotinylated primer directed against the anchor (or "bait") sequence (i.e., the region of interest). Step 6 shows the isolation of hybrid fragments, blunt-ending and adapter ligation. The fragments with anchor sequences after primer extension (biotinylated 5' ends modified) are extracted from the pool of 3C DNA by streptavidin beads, followed by blunt-ending and blunt adapter ligation. Step 7 shows the PCR amplification of hybrid fragments containing the anchor sequence to generate the pool called the "4C library". Nested primers against the anchor sequence and the adapter are used to amplify the specific hybrid fragments. Step 8 shows the detection of DNA species in the 4C library by hybridising onto the TAL1 genomic tiling path microarray.

5.4.2 *Designing primers for the 4C-array assay*

The TAL1 promoter was chosen as the anchor sequence to explore looping interactions at the TAL1 locus. To begin with, three types of primers were designed for the 4C-array assay (Figure 5.3), which were

(i) A 5'-biotinylated primer for primer extension of the anchor sequence,

(ii) An anchor-specific nested primer for PCR amplification with

(iii) An adapter-specific primer for PCR amplification

The biotinylated primer was designed approximately 150 bp toward the nearest restriction endonuclease (Csp6I) recognition sites of the anchor sequence (the TAL1 promoters). The anchor-specific nested primer was designed about 100 bp toward the same restriction recognition site. In combination with the adapter-

specific primer, a complex mixture of anchor-prey hybrid fragments could be amplified be PCR amplified. A high complexity mixture of PCR species was favoured because the size of the primer extended template was less than, on average, approx. 500 bp. The 4C primer was designed amplifying the promoter 1b of the TAL1 gene based on the criteria of GC content and annealing template for the PCR reaction. The anchor (fragment of the promoter 1b) was termed as P^{TAL1} in the later 4C-array analysis.

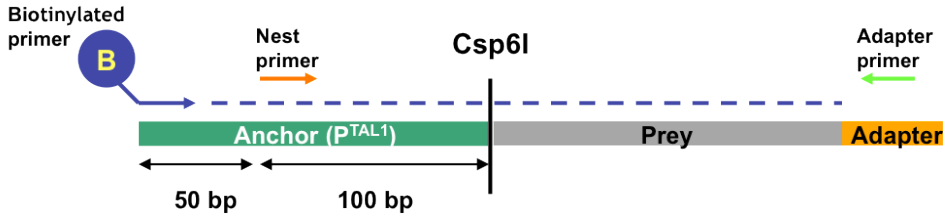


Figure 5.3: Schematic diagram of 4C primer design. Two types of 4C primers were designed for a designated region of interest, the "anchor (green bar)". i) 5'-biotinylated primer (blue arrow), located ~150 bp in distance to the nearest restriction site (Csp6I) of the anchor fragment; ii) the nested anchor primer (orange arrow), was located ~100 bp from the Csp6I site. The adapter primer (green arrow) was used in combination with nested anchor primer to generate the 4C library by PCR.

5.4.3 *Quality control of the sonication of 3C DNA*

For the 4C-array analysis, sonication was used to fragment the 3C DNA (step 4 in Figure 5.2) instead of a secondary restriction digestion. The sonicated DNA fragments were visualised by gel electrophoresis to determine the efficiency of sonication, based on the distribution of sonicated fragments by size (QC of step 4 in Figure 5.2). As shown in Figure 5.4, the size distribution of four sonicated 3C DNA samples were predominantly within the region of 500 bp to 150 bp, suggesting that sonication provided a randomly fragmented mixture where the size of vast majority of the DNA fragments was similar.

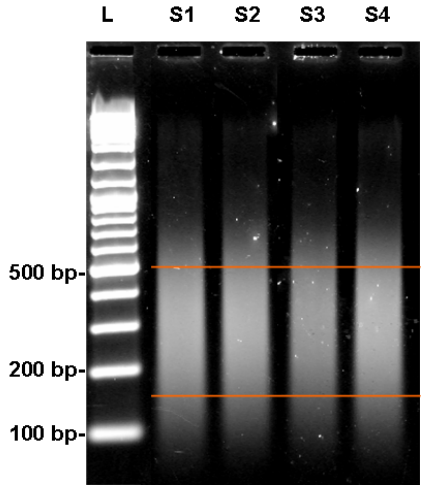


Figure 5.4: Agarose gel electrophoresis of sonicated 3C DNA from K562 and HPB-ALL cells. Lane L = DNA ladder for size reference; lane S1 & S2 (K562) and S3 & S4 (HBP-ALL) are the

sonicated 3C DNA samples. The areas marked by orange lines indicate the main size distribution of sonicated 3C DNA fragments which are regions between 500 bp to 150 bp. The size of DNA markers is shown on the left of the image. The samples were electrophoresed on 1.5% agarose (1 x TBE) gels and visualised with DNA Safe-stain.

5.5 Systematically assessing the performance of 4C-array

A number of criteria were applied in order to systematically evaluate the performance of 4C-array. In the schematic shown in Figure 5.5, there are four major levels that could introduce experimental variation during the 4C-array analysis, which could consequently affect sensitivity and reproducibility of the 4C data obtained from the array analysis. These included:

(i) The biological samples (referred to as level I),

(ii) Generation of the 3C library (referred to as level II),

(iii) Generation of the 4C library (referred to as level III), and

(iv) Microarray hybridisation (referred to as level IV).

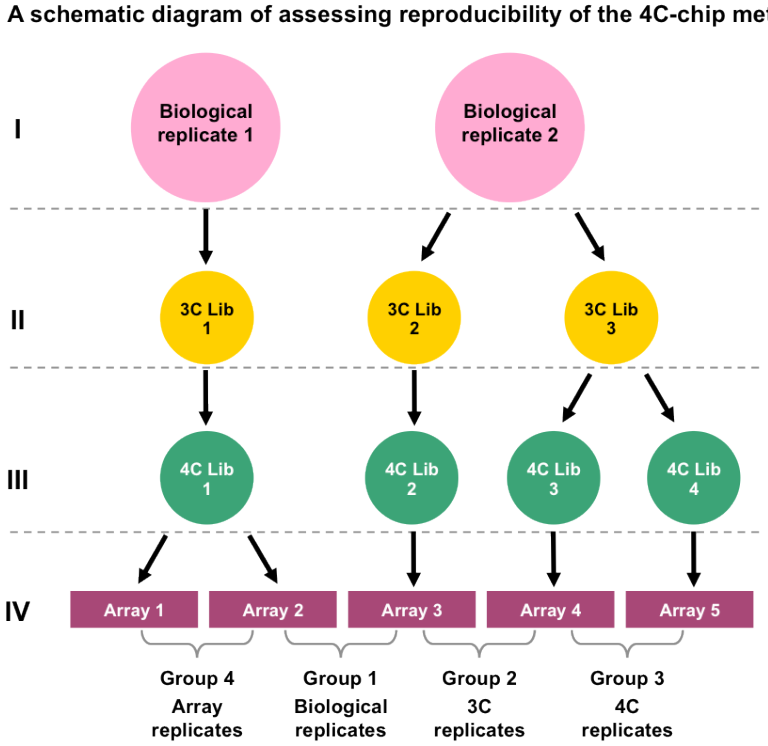


Figure 5.5: Flow chart used to systematically assess the reproducibility of the 4C-array method. The circles in pink represent independent biological replicates of K562 cells (level 1). The circles in gold represent independent 3C libraries and two 3C replicate libraries (right), which were made from same batch of K562 nucleus (level 2). The circles in green

represent independent 4C libraries, two of which are 4C replicates generated from the same 3C DNA (level 3). The squares in purple represent the independent hybridisations using TAL1 tiling path arrays which were used to detect interaction profiles from 4C libraries (level 4).

Performing biological and technical replicates for the 4C-array experiments allowed for biological and experimental variations at these different levels to be identified. Each of these four levels of variation was assessed as follows.

Level I: biological replicates included multiple batches of 3C nuclei which were generated from independent cross-linking experiments from different passages of a cell line grown at different times. Two biological replicates of K562 cells were prepared as shown in Figure 5.5.

Level II: technical 3C replicates included: two independent 3C libraries which were derived from using different aliquots of the same batch of nuclei ("3C Lib 2 & 3" in Figure 5.5).

Level III: two independent 4C libraries which were derived from using different aliquots of the same 3C library ("4C Lib 3 & 4" in Figure 5.5).

Level IV: two independent microarray hybridisations derived from using different aliquots of the same 4C library derived from a single K562 bio-replicate ("TAL1 array 1 & 2" in Figure 5.5).

In this section, 4C libraries were prepared using K562 3C material under the following conditions: i) 1 µg of sonicated 3C DNA was used for the 5'-biotinylated primer extension and subsequent manipulations, ii) 4C library was amplified using anchor-specific nested primer and adapter-specific primer with 36 cycles of PCR. The 3C libraries used for 4C-array preparation were checked for digestion and ligation quality as described in Chapter3. The TAL1 genomic tiling path array was used for the detection of 4C libraries described in this Chapter. The array contained both human and mouse genomic array elements spanning the TAL loci in both species (described in detail in section 1.4.6 of Chapter 1). For the work presented here, the TAL1 tiling arrays were used to study 4C interaction patterns for human cell lines only.

Labelling: 4C libraries generated from human K562 cells and genomic DNA extracted from corresponding cells (input) were differentially labelled with Cy3 (4C libraries) and Cy5 (input) by a random priming method before hybridised onto the TAL1 genomic tiling path array. In this method a DNA template is denatured allowing random primers to hybridize to complimentary sequences. The random primers are then extended by the 5'-3' polymerase activity of Klenow resulting in a strand displacement activity with the incorporation of fluorescently labelled nucleotides. The efficiency of labelling was monitored qualitatively by gel electrophoresis before hybridising onto arrays (Figure 5.6). Visual inspection of samples determined whether DNA fragments had been generated by labelling and the size distribution. A majority of labelled 4C DNA located below 150 bp with a large smear extended over 12-kb was evident from this electrophoretic analysis.

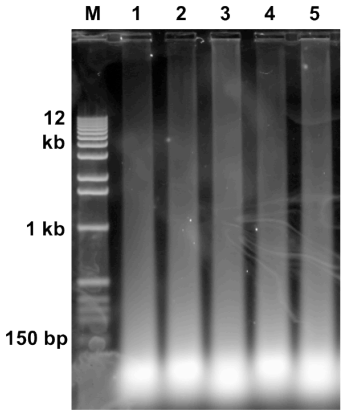


Figure 5.6: Electrophoresis of fluorescently labelled DNA of 4C-array samples. Lane M = DNA ladder, Lane 1 to 5: Cy3-labelled 4C-array samples 1 to 5. The samples were electrophoresed on 1% agarose 1 x TBE gels and visualised with DNA Safe-stain.

Hybridisation: A scanned composite image of the human TAL1 genomic tiling array is shown in Figure 5.7 as the example of 4C-array hybridisation. Each spot on the array represents an array element. Each array element was spotted in triplicate (yellow boxes) in a 16 sub-grid format (white box). The green spots on the image show enrichments in the 4C library as compared to the input DNA. The yellow spots represent equal hybridisation of the 4C and input DNA to the spot. Orange/red spots show regions which are under-represented in the 4C library. The white spots reflect saturated spots in the 4C sample. As the 4C-array library is prepared using human K562 cells, the absence of array spots over the image is corresponding to the array elements detecting mouse genome.

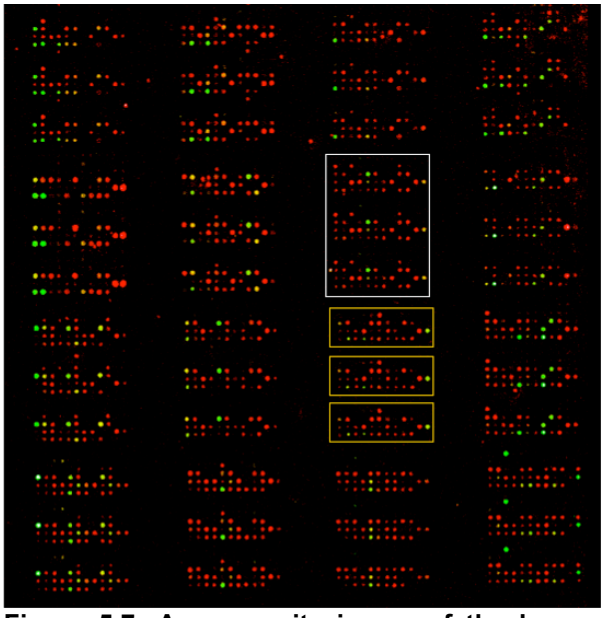


Figure 5.7: A composite image of the human TAL1 genomic tiling array. The array was hybridised with a 4C sample from the K562 cell line using the TAL1 promoter as the anchor. A single sub-grid is shown within the white box which can break down into a triplicate set of array elements is shown within the three yellow boxes.

5.5.1 *Assessing the 4C-array method at the level of reproducibility*

The overall interacting profiles of five 4C libraries were presented in histograms as shown in Figure 5.8. Comparisons are made between different biological samples, different 3C library prepared from the same batch of cells, different 4C libraries prepared from the same 3C library as well as the same 4C library were hybridised on different TAL1 arrays.

To evaluate the 4C-array in terms of reproducibility, the dependence between these 4C-array samples were calculated based on the pair-wise comparisons of array profiles. The most common way of measuring dependence is the Pearson's correlation, which can be used to estimate the linear relationship between two sets data. However, the Pearson' correlation was not suitable in this case, as the level of interaction frequency between genomic regions may vary between samples and not follow the linear pattern. This is because that chromatin interaction in the inter-phase nuclei is a dynamic process and the signal detected by 4C represents an average level of interaction frequency based on a population-based assay. Therefore, a non-linear dependence test such as the Spearman's rank correlation can provide a more accurate measurement of the relationship between two 4C libraries (Simonis et al., 2006).

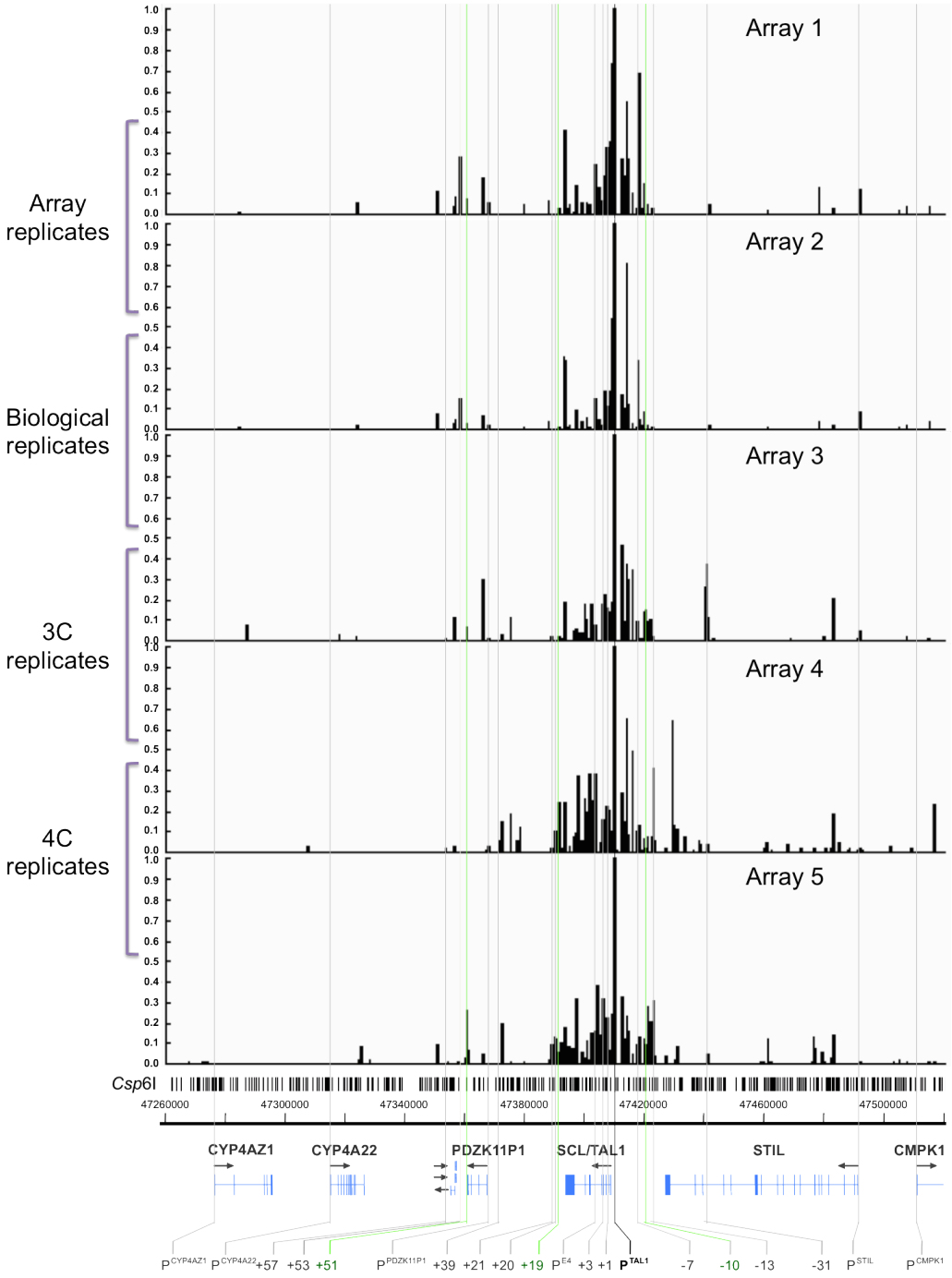


Figure 5.8: 4C-array looping interaction profiles across the human TAL1 locus in the K562 cell line generated corresponding to the experimental conditions listed in Figure 5.9. Relationships between Array 1 to 5 are illustrated briefly on the left of the figure. The histograms in each panel represent the data obtained for each genomic tiling array element. In each panel, the x-axis is the genomic sequence co-ordinate (NCBI build 35) and the y-axis is the enrichment obtained in 4C-array assays expressed as the relative ratio to the anchor (p1b). Schematic diagram at the bottom of the figure shows the genomic organisation of TAL1 and its neighbouring genes. Exons are shown as vertical blocks (blue) with gene names and direction of transcription shown above. Vertical lines at the bottom (with dotted lines through all the panels) show the location of known and novel regulatory regions at the TAL1 locus. Promoters are denoted by P. Other nomenclature refers to the distance in kb from TAL1 promoter 1a (P^{1a}). The anchor is highlighted in bold. Known interacting regions (based on 3C results) in K562 cells are highlighted in green at the bottom of the figure.

The spearman's rank correlation was calculated between four levels of the 4C-array profiles, which were biological replicates (level I), 3C replicates (level II), 4C

replicates (level III) and array replicates (level IV). As shown in Table 5.1, low level of correlation co-efficiencies were observed between biological replicates (r_s = 0.360, p < 0.001), 3C replicates (r_s = 0.460, p < 0.001) and 4C replicates (r_s = 0.506, p < 0.001). Much higher correlation co-efficiency was observed between array replicates (r_s = 0.823, p < 0.001) suggesting variation at the level of hybridisation was very low as previously reported (Dhami et al., 2010). Most importantly, the reproducibility between two 4C replicates is as low as ~50%, even though the 4C replicates used material from a single 3C library, indicating the 4C methodology introduces the major source of variation. The variation between the 3C replicates may be due to the inherited major variation from the 4C manipulation in combination with variation at the level of 3C methodology. Although the biological variation had the lowest correlation, its variation could mainly due to the inherited variation from 3C and 4C methodology in combination with differences between biological samples.

Table 5.1 Spearman' rank correlation between the 4C-array samples

Spearman's rank	[1] r_s	[2] p
Biological replicates	0.360	< 0.001
3C replicates	0.460	< 0.001
4C replicates	0.506	< 0.001
Array replicates	0.823	< 0.001

[1] **Spearman's rank co-efficiency**
[2] **Statistical significances of Spearman's rank correlation**

5.5.2 *Assessing the 4C-array method at the level of sensitivity*

Although the reproducibility is critical for a robust method being able to produce data at similar level of quality, the sensitivity is much more fundamental for the method being able to produce data in a decent quality. For instance, a method with high reproducibility but low sensitivity would only provide bad quality data repetitively. Therefore, sensitivity of 4C-array method was assessed using known looping interactions detected by 3C as a benchmark.

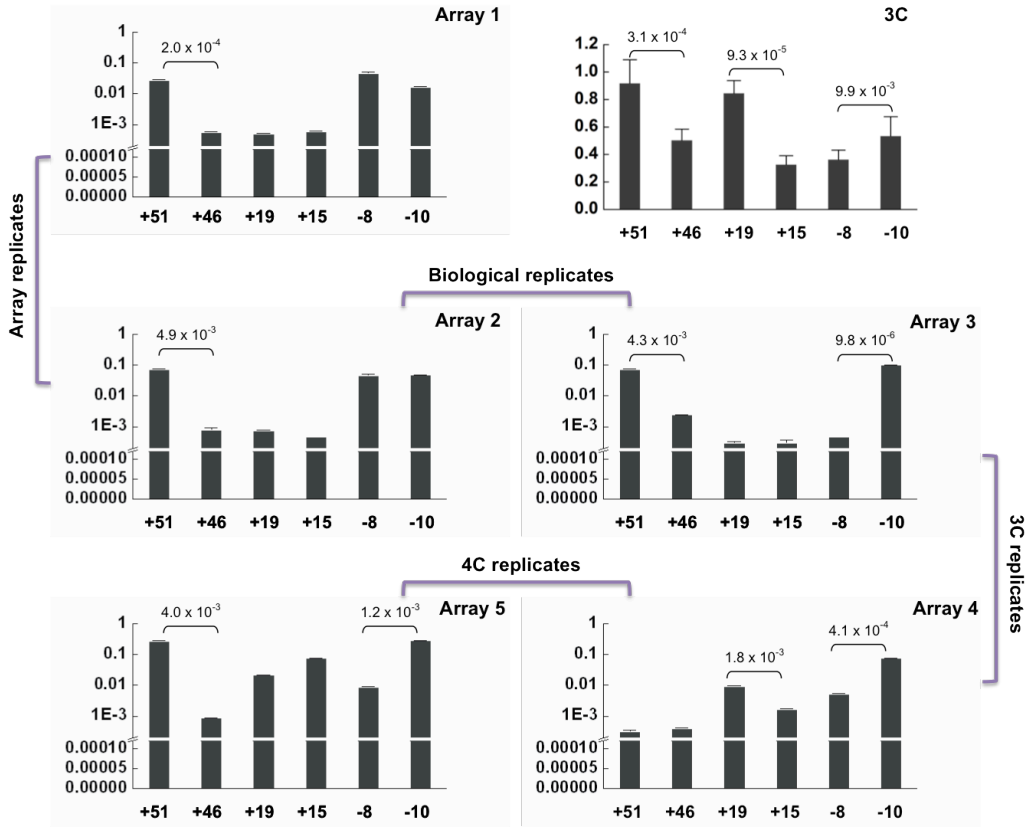


Figure 5.9: Matrix of 4C-array assessment profiles by histograms. Relationships between Array 1 to 5 and their corresponding experimental conditions are illustrated in Figure 5.9. The x-axis represents the regions examined in 3C-PCR assays. The y-axis represents the ligation frequency relative to the anchor (log_{10} scaled after breakpoint). The p values of student's T-test between enhancers (+51, +19 and -10) and their control regions were labelled on top of the histograms.

Six 3C data points were used which included three known promoter-enhancer interactions (+51, +19 and -10) as well as the control regions (+46, +15 and -8) for each enhancer regions. The array tiles representing the enhancers and control regions are pinpointed based on the location of primers used in the 3C analysis. These array tiles were also checked for their relative location to the nearest Csp6I restriction site to ensure the representation of a single Csp6I fragment. Interaction frequencies of corresponding array tiles were subsequently plotted in histograms as shown in Figure 5.9. The student t-test was performed as described previously to measure statistical significance of enrichment at the target regions comparing against its control regions.

Overall, all five 4C-array libraries failed to detect all three known interactions previously captured by 3C (Figure 5.9), indicating a low sensitivity of 4C-array at this stage. For array replicates, two 4C-array profiles (Array 1 and 2) are very similar in terms of the ranking of enrichment and both of them detected only one significant interaction at the +51 enhancer (p = 2.0 × 10^{-4} and 4.9 × 10^{-4}, t-test) as

shown in Figure 5.9. It suggested that sensitivity of Array 1 and 2 was very low in terms of detecting known looping interactions captured by 3C, and additionally, robust microarray hybridisation provided highly reproducible 4C profiles but with low sensitivity. For 4C replicates (Array 4 and 5), although starting material used for 4C libraries preparation was the same sonicated 3C DNA, two significant interactions detected in Array 4 were the +51 and -10 enhancers (p = 4.0 × 10^{-3} and 1.2 × 10^{-3}, t-test), whereas the +19 and -10 enhancers (p = 1.8 × 10^{-3} and 4.1 × 10^{-4}, t-test) were captured significantly in Array 5 (Figure 5.9). Similarly, insensitivity and irreproducibility were observed between the 3C replicates (Array 3 and 4) and biological replicates (Array 2 & 3) as shown in Figure 5.9. In summary, these data suggested that 4C-array was failed to provide the high-level sensitivity and reproducibility at this stage as a robust method. Thus, it is necessary to optimise the method in order to improve its sensitivity and reproducibility.

5.6 Optimisation of 4C library complexity and PCR amplification conditions

5.6.1 *Rationality and experiment design of 4C optimisation*

Possible causes

Initially, 1 μg of sonicated 3C DNA was used for a 4C-array assay (described above). A series of steps were then conducted including biotinylated primer extension, streptavidin beads extraction and PCR adapter ligation. Subsequently, 4C material was PCR amplified with 36 cycles before hybridising on the TAL1 array. Library complexity and PCR cycling condition were considered as two conditions, which could affect the sensitivity and reproducibility of the 4C-array assay. With high cycle numbers and relatively low complexity of PCR template, the ligation products corresponding to frequent interactions are likely to be saturated at early stage of PCR amplification which will lead to less quantitative presentation of the real interaction frequency. The complexity of PCR template can be improved by increasing the starting 3C material for the 4C preparation. This is because 3C-based analysis has a very low-efficiency of generating proper ligation products between two interacting partner fixed in the genome. Thus, additional ligation products can be provided by increasing the amount of starting 3C material, which can subsequently improve the complexity of 4C library for the PCR amplification. Therefore, the adjustments of the 4C-array protocol were made accordingly via

increasing the library complexity and reducing the cycle number of PCR amplification. These were then tested systematically with respect to reproducibility and sensitivity as described for the series of experiments presented above.

Experiment design

As previously described above, three criteria were used to assess the 4C procedure through a series of optimisation experiments where PCR cycling conditions and library complexity were altered.

(i) The ability to detect known interactions identified by 3C;

(ii) The ability to reproducibly detect these known interactions;

(iii) The overall reproducibility of 4C-array results using Spearman's correlation analysis.

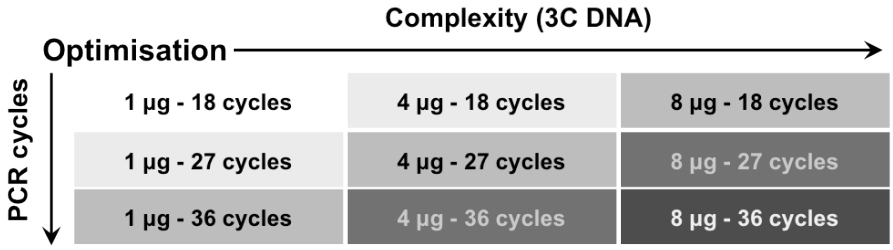


Figure 5.10: Matrix of experiments to assess the PCR amplification conditions of the 4C method. Nine independent 4C libraries were derived from a single sonicated 3C library. The amount of starting 3C DNA and the number of PCR cycles were tested at three levels.

The experimental matrix is show in Figure 5.10. The starting amount of sonicated 3C DNA was tested at three levels: 1 µg (original amount according to the protocol in Fraser's laboratory), 4 µg, and finally 8 µg. Although it is always better to start with more 3C DNA, 8 µg is the maximum amount of sonicated 3C material selected for this serial of 4C-array analysis due to the total amount of DNA in a single 3C library is limited (~50-100 µg). In addition, the step of sonication also results in a loss of ~20-30% of total 3C DNA. Thus, the amount of DNA material from a single 3C library would only be able to accommodate a range of 4C-array analysis at the scale described above. The number of PCR cycles was tested at 36 cycles (original amount according to the protocol in Fraser's laboratory), at 27 cycles, and eventually at 18 cycles. Further deductions of the PCR cycle numbers might result in producing 4C library with inadequate amount of DNA, which would subsequently affect downstream detection using the microarray platform.

5.6.2 *Optimising experimental conditions for the 4C-array assay in K562 cells*

The 3C library was generated using K562 cells and the library quality was checked for digestion and ligation efficiency as described in Chapter 3. Subsequently, 3C DNA was sonicated to within a range of 150-500 bp as previously described. The sonicated DNA was then used to generate nine 4C libraries following the conditions as previously shown in Figure 5.10.

Table 5.2 Assessment of DNA concentrations of the 4C-array libraries

No.	4C library	DNA con.[1] (ng/µl)	Total amount[2] (ng)	Amount for hyb[3] (ng)
1	1 µg - 18 cycles	N/A	N/A	N/A
2	1 µg - 27 cycles	0.4	20	8
3	1 µg - 36 cycles	5.5	275	110
4	4 µg - 18 cycles	N/A	N/A	N/A
5	4 µg - 27 cycles	3.8	190	76
6	4 µg - 36 cycles	11.9	595	238
7	8 µg - 18 cycles	N/A	N/A	N/A
8	8 µg - 27 cycles	7.1	355	142
9	8 µg - 36 cycles	10.0	500	200

[1] DNA concentration in each 4C library measured by PicoGreen® DNA quantification kit
[2] Total amount of DNA in each 4C library (50 µl) estimated based on its DNA concentration.
[3] Amount of DNA used for labelling and microarray hybridisation (20 µl of 4C DNA)

The 4C libraries prepared under different complexity and cycling conditions were quantified and the DNA concentration and total amount of DNA generated in each libraries were shown in Table 5.2. The DNA concentration was below the threshold of detection in all three 4C libraries prepared under 18 cycles of PCR amplification (Table 5.2 and Figure 5.11a). The 4C libraries generated with 27 and 36 cycles of PCR amplification produced the detectable amount of DNA as shown in Table 5.2 and Figure 5.11a.

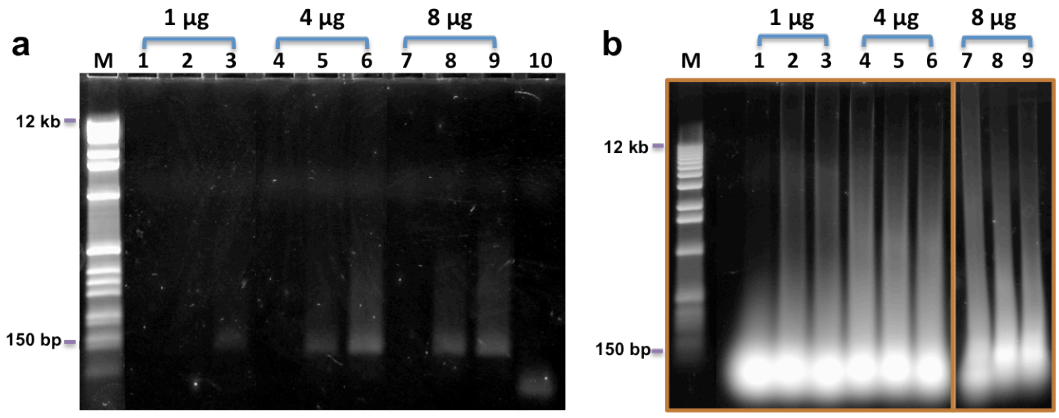


Figure 5.11: Electrophoresis of 4C-array DNA samples before and after labelling. Panel a: 2 µl of 4C libraries DNA before labelling and Panel b: 5 µl of Cy3 labelled 4C DNA. Lane M =

DNA ladder, lane 1, 4 and 7: 18 cycles of PCR, lane 2, 5 and 8: 27 cycles of PCR, lane 3, 6 and 9: 36 cycles of PCR and lane 10: PCR H_2O control. The samples were electrophoresed on 1% agarose 1 x TBE gels and visualised with DNA Safe-stain.

For the 4C libraries with 18 cycles PCR amplification, although the amount of 4C DNA used for microarray hybridisation was too low to be detected, the labelling was almost unaffected (Figure 5.11b), as the random priming method used for labelling is able to provide a DNA amplification of at least four-fold over starting amounts (Lieu et al., 2005). The only exception was the labelled 4C DNA (1 µg - 18 cycles) as shown in lane 1, which seemed to have a relatively narrow size distribution (< 1 kb) of Klenow-amplified Cy3-labelled fragments in comparison with other 4C-array libraries. However, this would not affect the detection of 4C library using the TAL1 array as the majority of the labelled fragments were distributed in the size range of 80-150 bp (S. Dillon's PhD thesis, University of Cambridge, 2008). In addition, similar examples can be found in ChIP-chip analysis when trying to determine the genome-wide binding pattern of transcription factors. Although ChIP DNA immunoprecipitated using the antibody against the particular transcription factor was not able to be visualised using gel electrophoresis, the efficiency of DNA labelling as well as the quality of array hybridisation were not compromised, comparing to the ChIP analysis using the antibody against histone modifications (ChIP DNA is visible on gel), owing to the robust microarray platform for detection (P. Dhami's PhD thesis, University of Cambridge, 2005). Taken together, the 4C libraries generated under all levels of complexity as well as PCR cycling conditions provided adequate quantities of DNA for downstream detection using the microarray platform.

5.6.2.1 Assessing the ability to detect known interactions identified by 3C (Criteria I)

The 4C-array profiles prepared under nine different conditions were assessed as previously described in section 5.5, in order to determine which conditions allow significantly detection of three known interactions captured by 3C. Four of the 4C libraries were able to recapitulate the data obtained by 3C with respect to detection of all three TAL1 promoter-enhancer interactions (Figure 5.12), which were prepared under conditions of 4 µg - 36 cycles, 8 µg - 18 cycles, 8 µg - 27 cycles and 8 µg - 36 cycles, respectively.

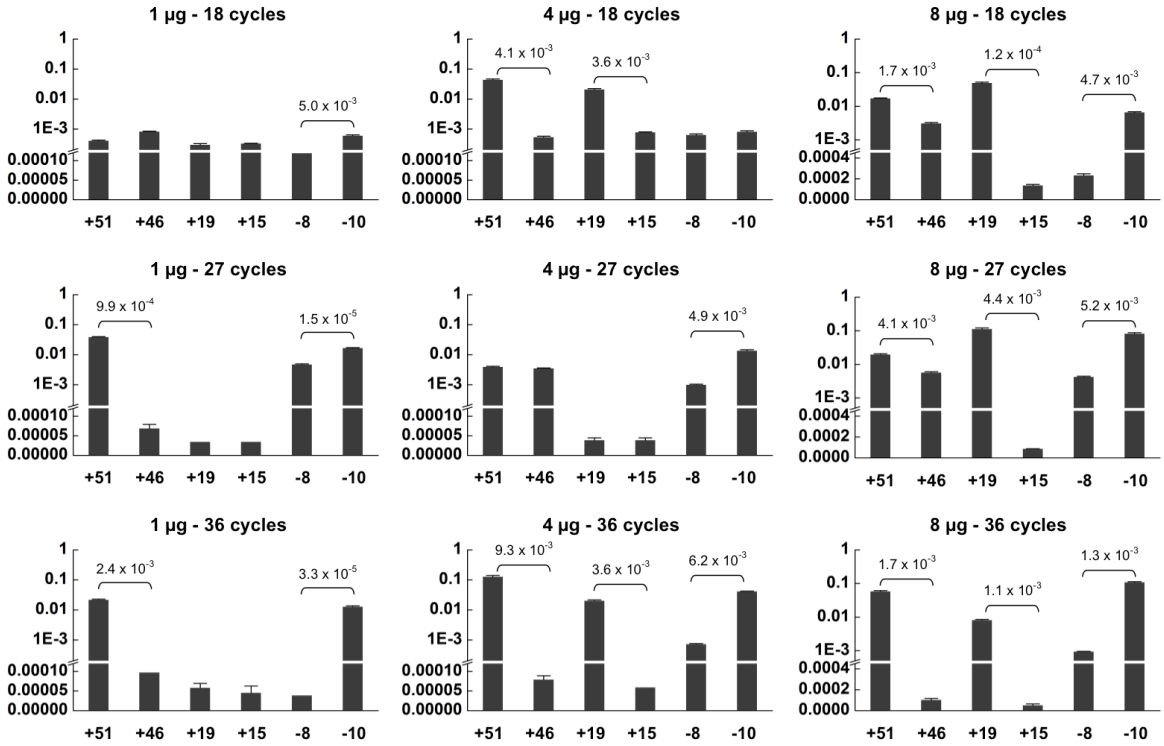


Figure 5.12: Matrix of 4C-array adjustment profiles by histograms. Panel A: 4C-array adjustment of the amount of starting 3C material and the PCR cycling conditions (1 µg to 8 µg and 18 cycles to 36 cycles, respectively). The x-axis represents the regions examined by 3C. The y-axis represents the ligation frequency relative to the anchor (log10 scaled after breakpoint). The p values of student's T-test between enhancers (+51, +19 and -10) and their control regions were labelled on top of the histograms. The error bars are calculated based on the triplicates of the array elements.

Subsequently, the Spearman's rank correlation was calculated between the 3C and these four conditions to measure the similarity between 3C and 4C profiles (shown in Table 5.3). 4C-array profile prepared under the condition of 8 µg - 18 cycles has the highest correlation co-efficiency (r_s = 0.914) with the 3C profile, followed by 8 µg - 27 cycles and 4 µg - 36 cycles (r_s = 0.829 and 0.8), accordingly. The lowest correlation co-efficiency was observed between 3C profile and 8 µg - 36 cycles (r_s = 0.686). Therefore, two conditions (8 µg - 18 cycles and 8 µg - 27 cycles) were selected because of their highest correlation co-efficiency with the 3C profile and would be preceded to the next level of assessment.

Table 5.3 Spearman's rank correlation between 3C-PCR and 4C-array profiles.

Spearman's rank	r_s	p
3C vs. 8 µg - 18 cycles	0.914	< 0.001
3C vs. 8 µg - 27 cycles	0.829	< 0.001
3C vs. 4 µg - 36 cycles	0.800	< 0.001
3C vs. 8 µg - 36 cycles	0.686	< 0.001

[1] **Spearman's rank co-efficiency**
[2] **Statistical significances of Spearman's rank correlation**

5.6.2.2 Assessing the ability to reproducibly detect known interactions (Criteria II)

In order to work out the best condition for a robust 4C-array analysis, two additional 4C libraries were generated under the selected conditions in the previous section. The 3C DNA used for this 4C analysis was generated from a different K562 biological replicate. The 3C library preparation and quality control assays were performed as described in Chapter 3. Using the second K562 bio-replicate, 4C-array assays showed its ability to detect all three known interactions repetitively under both conditions (8 µg - 18 and 27 cycles) as shown in Figure 5.13. In addition, the Spearman's rank correlation co-efficiencies were calculated for 4C-array profiles prepared using the second bio-replicate, with respect to the 3C profile. It showed that both 8 µg - 18 cycles and 8 µg - 27 cycles' 4C profiles were significantly correlated with the 3C profile (r_s = 0.714 and 0.686, $p < 0.001$).

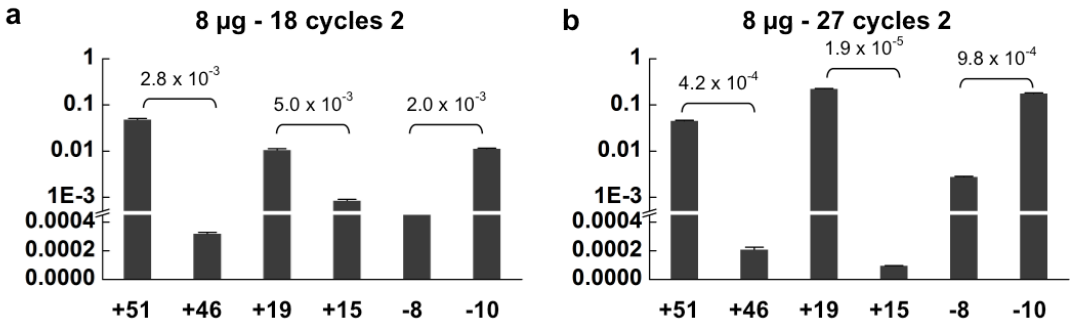


Figure 5.13: 4C-array profiles of the second biological replicates by histograms. Panel a: 4C-array prepared condition of 8 µg - 18 cycles and Panel b: 8 µg - 27 cycles. The x-axis represents the regions examined by 3C. The y-axis represents the ligation frequency relative to the anchor (log_{10} scaled after breakpoint). The p values of student's T-test between enhancers (+51, +19 and -10) and their control regions were labelled on top of the histograms. The error bars are calculated based on the data sets of array element triplicates.

For both biological replicates, 4C-array analyses conducted under the conditions of 8 µg - 18 cycles and 27 cycles had the highest sensitivity and reproducibility comparing to other conditions (8 µg - 36 cycles 2, r_s = 0.629, data not shown), in terms of detecting known 3C interactions. Taken together, it suggested that these two conditions were the best candidates so far based on the criteria I. and II.

5.6.2.3 Assessing overall reproducibility of 4C-array profiles (Criteria III)

The reproducibility of 4C-array conditions was only validated in terms of repetitively detecting known 3C interactions. However, the 3C-PCR results only provided a snapshot of overall interacting pattern between TAL1 promoter and its

co-associates. Therefore, the method was further assessed for its reproducibility at the level of overall interacting patterns under the above two conditions.

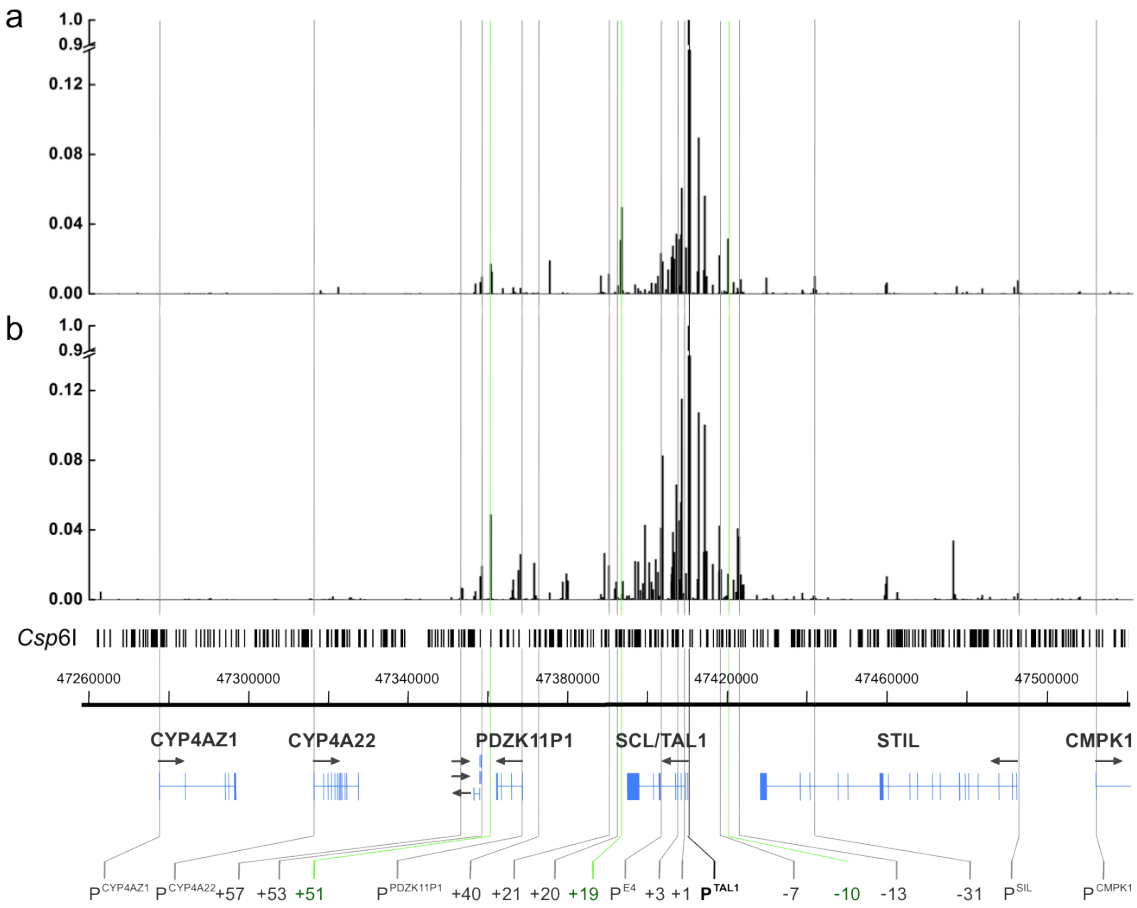


Figure 5.14: 4C-array looping interaction profiles across the human TAL1 locus in the K562 cell line generated with 8 µg and 18 cycles of PCR. Two biological replicates were shown in panel A & B. The histograms in each panel represent the data obtained for each genomic tiling array element. In each panel, the x-axis is the genomic sequence co-ordinate (NCBI build 35) and the y-axis is the enrichment obtained in 4C-array assays expressed as the relative ratio to the anchor (p1b). Schematic diagram at the bottom of the figure shows the genomic organisation of TAL1 and its neighbouring genes. Exons are shown as vertical blocks (blue) with gene names and direction of transcription shown above. Vertical lines at the bottom (with dotted lines through all the panels) show the location of known and novel regulatory regions at the TAL1 locus. Promoters are denoted by P. Other nomenclature refers to the distance in kb from TAL1 promoter 1a (P^{1a}). The anchor is highlighted in bold. Known interacting regions are highlighted in green at the bottom of the figure.

The comparison was made between 4C-array profiles of two bio-replicates under the conditions of 8 µg - 18 cycles and 27 cycles as illustrated in Figure 5.14 and 5.15 respectively. The level of reproducibility was analysed by calculating spearman's correlation coefficients between the two biological replicates under each set of conditions. The profiles of 4C-array performed under 18 cycles of PCR amplification showed the highest correlation between two replicates (r_s = 0.73, p < 0.001). The reproducibility of the 4C-array with 27 cycles of PCR amplification was relative lower (r_s = 0.53, p < 0.001).

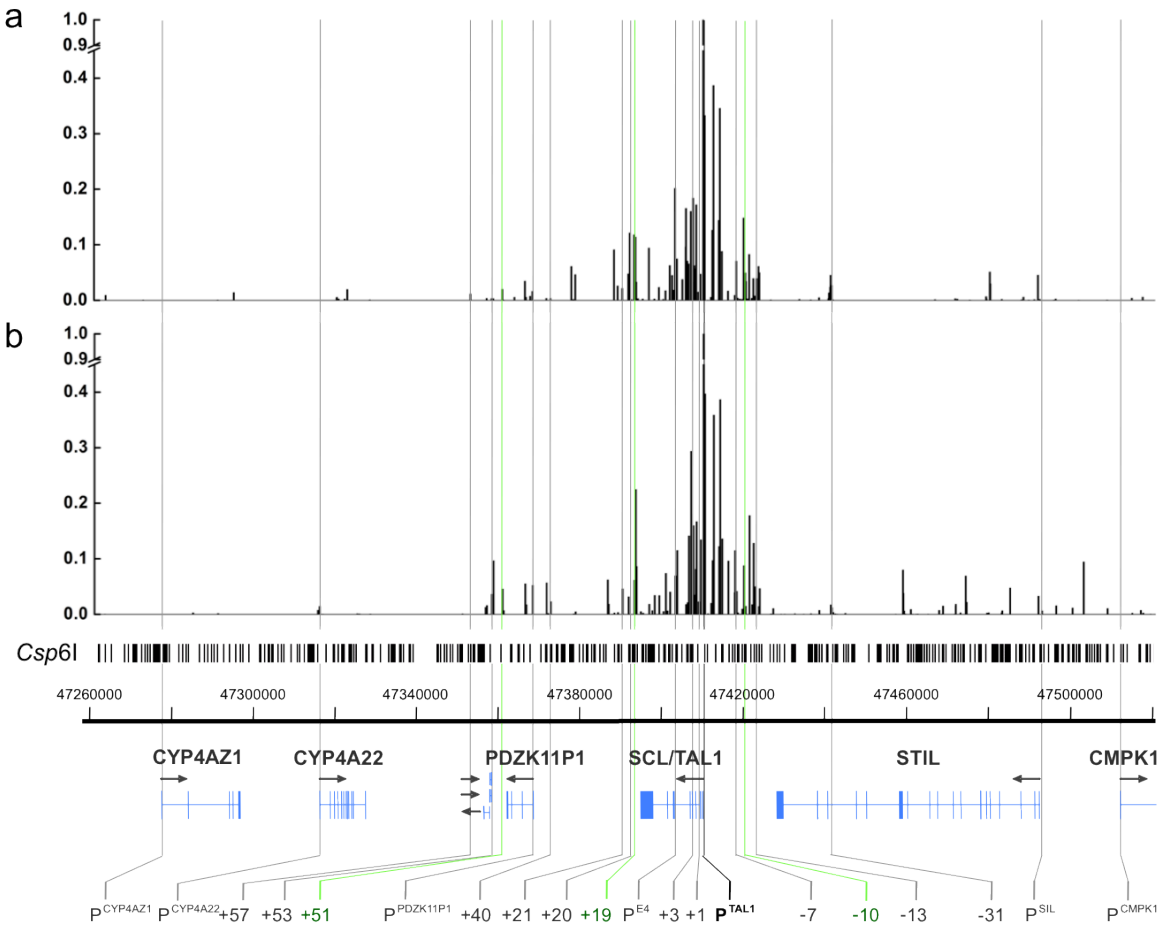


Figure 5.15: 4C-array looping interaction profiles across the human TAL1 locus in the K562 cell line generated with 8 µg and 27 cycles of PCR. Two biological replicates were shown in panel A & B. Other details of description are as shown in Figure 5.14.

In summary, the sensitivity and reproducibility of the 4C-array method was assessed under nine different complexity and PCR cycling conditions based on three criteria listed at the beginning of this section. The 4C library prepared under the condition of 8 ug of 3C material and 18 cycles of PCR demonstrated the highest sensitivity and reproducibility in capturing looping interactions across the TAL1 locus using microarray analysis. This condition was subsequently used to study the interacting partners of the TAL1 promoter in K562 and HPB-ALL cells.

5.7 Studying looping interactions at the TAL1 locus using 4C-array in TAL1 expressing and non-expressing cell lines.

5.7.1 *Overall interaction patterns*

The optimised conditions (8 µg of 3C DNA and 18 cycles of PCR amplification) described in the previous section were used for preparing 4C-array library to determine the looping interaction profiles for the TAL1 promoter in TAL1

expressing (K562) and non-expressing (HPB-ALL) cell lines. For both cell lines, 4C-array analysis was performed with two biological replicates. The 3C DNA used for 4C-array library preparation was from the same 3C libraries prepared and described in Chapter 3. After microarray hybridisation and data analysis, interaction profiles from two biological replicates were averaged and plotted as shown in Figure 5.16.

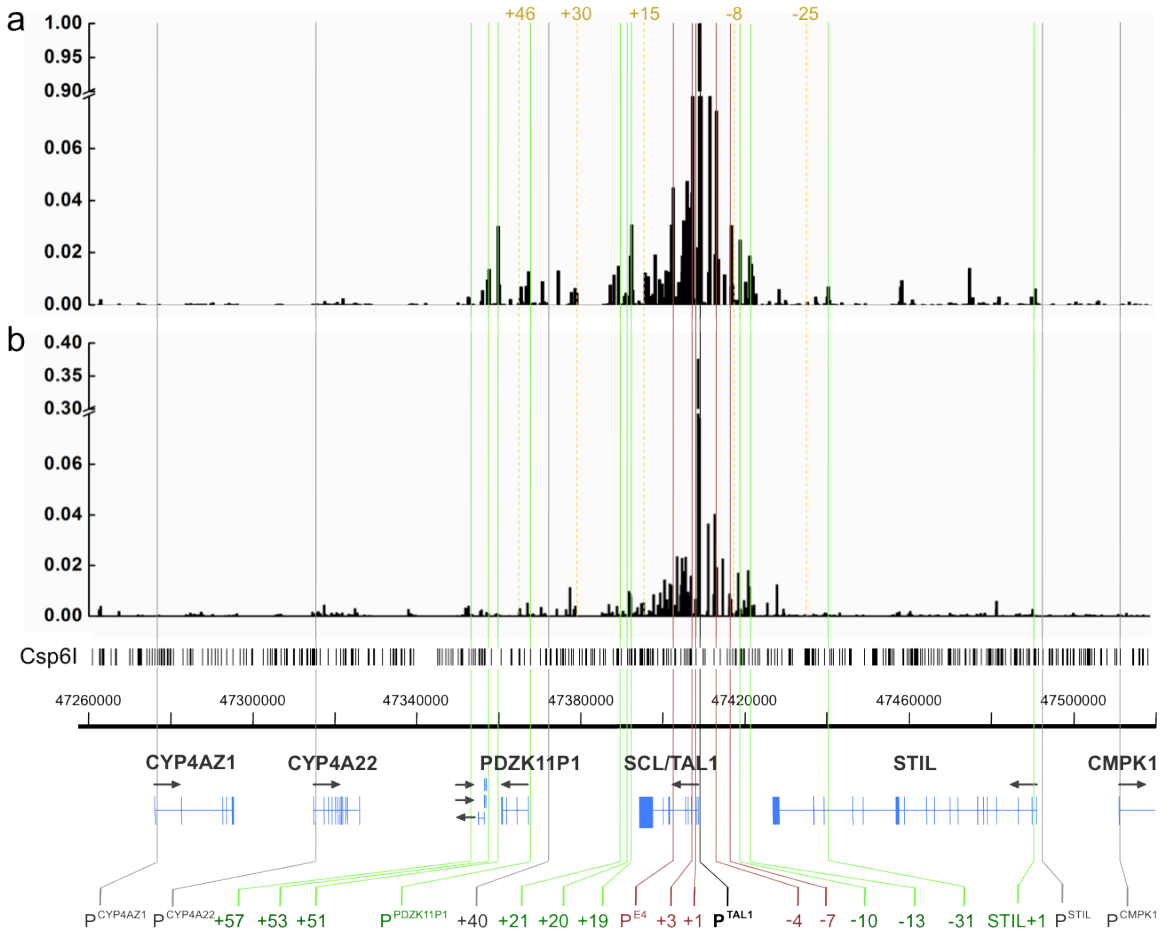


Figure 5.16: 4C-array looping interaction profiles across the human TAL1 locus. Panel A: K562 cells; panel B: HPB-ALL cells. Other details of description are as shown in Figure 5.14. Known regulatory elements significantly interacting (T-test, shown in Table 5.4) with the TAL1 promoter in K562 and/or HPB-ALL are highlighted in green lines and denoted at the bottom of the figure. The location of control regions (used in 3C analysis) used for determine the real looping interactions are highlighted in yellow dashed lines and denote at the top of the figure. Known regulatory elements with high enrichment located around the TAL1 promoter or not interacting with the TAL1 promoter in K562 and/or HPB-ALL are highlighted in red and dark grey lines respectively and denoted at the bottom of the figure.

In TAL1-expressing K562 cells, the overall interaction profile of the TAL1 promoter to the rest of genomic regions within the locus showed three major interactions clusters corresponding to the +57/+53/+51 regions, the +21/+20/19 regions and the regions around the TAL1 promoter (shown in Figure 5.16a). In addition, a number of other enrichments were also present at the PDZK1IP1 promoter as well

as genomic regions across the STIL gene body. In TAL1 non-expressing HPB-ALL cells, most of the enrichments are observed at the genomic regions clustered around the anchor (P^{TAL1}) and interaction frequencies of these regions are decreased as a function of distance when moving away from the TAL1 promoter. Additionally, the overall interaction frequency is relatively lower in comparison with K562 cells (shown in Figure 5.16b), which agrees with the similar patterns observed in 3C profiles between the two cell lines. As shown in Chapter 3 and appendix, it has been proven that the reduced interaction frequency in HPB-ALL cells is not due to the quality differences (digestion and/or ligation) of 3C libraries, but related to a context-dependent manner. For the same reason, the reduced overall interaction frequency detected by 4C-array in HPB-ALL suggests that the spatial contacts between the promoter and other regulatory elements are much less frequent when transcription of the gene is inactive.

Overall, the interactions between the TAL1 promoter and other genomic regions at the TAL1 locus detected by 4C-array were classified in three groups (shown in Table 5.4):

1. Interactions with known regulatory elements which have been previously determined by the 3C analysis (coded orange)

2. Interactions of known regulatory elements which may or may not interact with the TAL1 promoter predicted based the 3C profiles and the cruciform model (coded in blue)

3. Novel interactions with known and novel genomic elements across the TAL1 locus (coded in violet)

The statistical analysis (T-test) was performed as described in Chapter 3 and 4 using control regions which had been validated in 3C analysis, including +46, +30, +15, -8 and -25 (as denoted in Figure 5.16). Array tiles used to represent the control regions were determined by aligning with the PCR primer of corresponding control regions used in 3C analysis. In addition to that, these array tiles were also checked for their relative positions to the nearest Csp6I restriction sites, in order to ensure that all tiles can be assigned to a Csp6I fragment. It subsequently means that the signal of each region is derived from a single fragment.

Table 5.4 Interactions between the TAL1 promoter and genomic elements across the TAL1 locus by 4C-array

Regulatory elements[1]	Control regions[2]	Function[3]	4C-array[4]		3C-PCR[5]	
			K562 (TAL1+)	HPB-ALL (TAL1-)	K562 (TAL1+)	HPB-ALL (TAL1-)
TAL1 +57	vs. TAL1 +46	CTCF binding site	2.3×10^{-3}	6.6×10^{-4}		
TAL1 +53	vs. TAL1 +46	Promoter/CTCF binding site	1.3×10^{-3}	no		
TAL1 +51	vs. TAL1 +46	Erythroid enhancer	2.8×10^{-3}	no	yes	no
$P^{PDZK1IP1}$	vs. TAL1 +30	Promoter	2.9×10^{-5}	1.1×10^{-3}		
TAL1 +40	vs. TAL1 +30	CTCF binding site	no	no		
TAL1 +21/+20/+19	vs. TAL1 +15	Stem cell enhancer	6.3×10^{-4}	4.5×10^{-4}	yes	yes
P^{TAL1}		Promoter	Anchor	Anchor	Anchor	Anchor
TAL1 -10	vs. TAL1 -8	Active enhancer	4.0×10^{-5}	1.2×10^{-3}	yes	no
TAL1 -13	vs. TAL1 -8	Putative repressor	2.3×10^{-3}	1.3×10^{-4}		
STIL exon18	vs. TAL1 -8	STIL exon	1.4×10^{-3}	8.1×10^{-4}		
TAL1 -27	vs. TAL1 -25	unknown	1.8×10^{-3}	no		
TAL1 -31/ STIL exon16	vs. TAL1 -25	CTCT binding site	4.6×10^{-3}	1.1×10^{-3}		
STIL exon13	vs. TAL1 -25	STIL exon	3.4×10^{-3}	2.3×10^{-4}		
TAL1 -56	vs. TAL1 -25	unknown	2.9×10^{-4}	no		
STIL exon4	vs. TAL1 -25	STIL exon	4.7×10^{-3}	3.2×10^{-3}		
STIL exon2	vs. TAL1 -25	STIL exon	3.5×10^{-3}	no		
STIL +1	vs. TAL1 -25	STIL intron 1	1.1×10^{-3}	1.3×10^{-6}		

[1] Genomic element across the TAL1 locus
[2] Control regions used to determine looping interaction via T-test
[3] Function of genomic elements at the TAL1 locus
[4] p-values (T-test) for interactions detected by 4C-array
[5] Interactions previously detected by 3C

The 4C profiles showed evidence for a number of looping interactions between the TAL1 promoter and a number of known regulatory elements and novel regions previously unstudied by 3C, which will be discussed accordingly in the following sections.

5.7.2 *Interactions with known regulatory elements supported by 3C*

In Chapter 3, it was illustrated by 3C that three enhancers (+51, +19 and -10) interacted with the TAL1 promoter in K562, while only the +19 enhancer was in contact with the promoter in HPB-ALL. As shown in Figure 5.16 and Table 5.4, significant enrichments are shown at all three enhancers (+51, +21/+20/+19 and -10) in K562 and two of the enhancers (+21/+20/+19 and -10) in HPB-ALL, which mostly agree with the 3C results in both cell lines. However, in addition to supporting what has been known based on 3C, the enhancer at -10 is determined as a significant looping interaction in HPB-ALL by 4C-array analysis.

5.7.3 *Interactions with known regulatory elements not analysed by 3C*

Based on the 3C data shown in Chapter 3 and 4, a primary interaction model has been proposed and it predicts that the CTSs at +57/+53 and -31 but not +40 are likely to be in close proximity with the TAL1 promoter in K562 cells if all the interactions detected by 3C were presented in the same cells. As a high-throughput technology, 4C-array profiles provide the additional scope in detecting looping interactions, which allow pinpointing the contact points of the TAL1 promoter across the locus (as illustrated in Figure 5.16).

In K562 cells, significant enrichments were observed at +57/+53 as well as -31 regions, indicating that the TAL1 promoter is in contact with these CTSs frequently (Figure 5.16a and Table 5.4). As expected, no contact was observed between the TAL1 promoter and the CTS at +40 (Figure 5.16a and Table 5.4). Although no interaction was detected between all CTSs at the TAL1 locus as well as between the +51 enhancer and the TAL1 promoter, it was observed that low level but significant interactions between the TAL1 promoter and the CTSs at +57 and -31 in HPB-ALL cells (Table 5.4). However, the level of interaction is significantly much lower (7-folds, p = 0.0030) at -31 in comparison with K562. The underlying mechanism of these contacts is unclear. Nevertheless, the P^{TAL1}/+57 and P^{TAL1}/-31 interactions are separated events which would not exist in the same cells, as no contact was detected between the CTSs at +57 and -31. It is assumed that interaction between the TAL1 promoter and CTSs at +57 and -31 in a TAL1 non-expressing cell line may due to the insulator function for transcription repression.

In addition to the interactions with CTCF binding sites, the -13 region known as a repressor was also found to have significant interactions with the TAL1 promoter in both K562 and HPB-ALL cells (Figure 5.16 and Table 5.4). It is unclear whether the interaction between the P^{TAL1} and the repressor at -13 has a functional relationship or is just a bystander event, as the -13 element is located in a close genomic distance to the -10 enhancer which was in contact with the P^{TAL1} in both cell lines.

For those genomic elements located around the anchor (P^{TAL1}), high levels of enrichments were observed at +7 (P^{Exon4}), +3, +1, -4 and -7 (enhancer) in both cell lines (Figure 5.16). However, it was difficult to determine whether these

interactions were statistically significant and functionally linked to the P^{TAL1} due to the following reasons. First, it was extremely difficult to define control regions for statistical analysis of these *cis*-acting elements, as there was limited genomic distance to the P^{TAL1}. Second, functional interactions can only be determined between distal elements, as the adjacent genomic regions tend to be co-localised with each other by their very nature. Therefore, the 4C-array data presented here was only able to demonstrate that all these adjacent *cis*-acting elements were in spatial close-proximity with the P^{TAL1}.

5.7.4 ***Novel interactions with known and novel regions***

The TAL1 promoter was also found to interact with the adjacent genes across the TAL1 locus. First, significant enrichment was observed at the promoter of PDZK1IP1 in both K562 and HPB-ALL cells (Figure 5.16 and Table 5.4), suggesting the promoters of TAL1 and PDZK1IP1 genes are in spatial proximity regardless of transcriptional states. Second, a number of contact points were observed across the STIL gene body in both cell types. In addition, some of these interacting partners corresponded to a number of exons across the STIL gene body.

As shown in Figure 5.17a, eight significant enrichments were observed in K562 cells, five of which occurred at the STIL exons 2, 4, 13, 16 and 18. In comparison with K562, five out eight significant interactions were also observed in HPB-ALL, occurring at the STIL exons 4, 13, 16 and 18 (Figure 5.17b). Statistical analysis was performed in order to rule out the possibility that the GC content of the exon may bias the PCR amplification as well as microarray hybridization. The GC content of array elements corresponding to all 18 exons of STIL was calculated. Subsequently, the STIL exons were divided into two groups based on whether they interacted with the TAL1 promoter in K562 and HPB-ALL cells. The unpaired t-test was then performed to determine whether GC contents were significantly different between exons with and without interactions. No statistical significance of GC contents was observed between two groups of exons in both K562 and HPB-ALL ($p = 0.901$ and 0.669, respectively), suggesting that these interacting sites over exons were not artefacts due to the GC content differences.

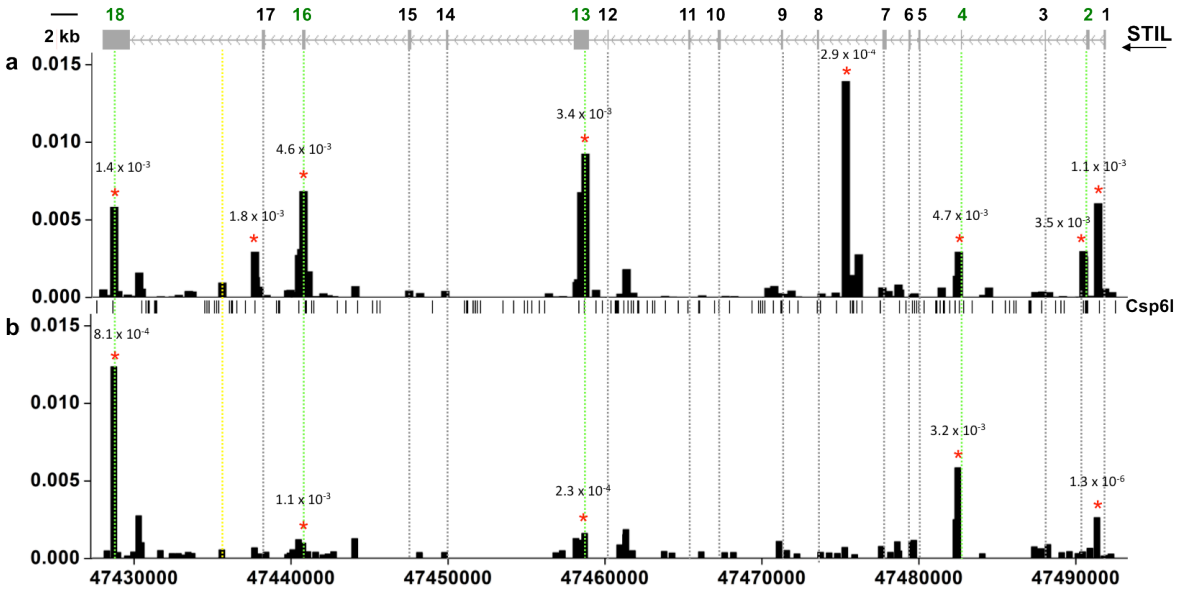


Figure 5.17: 4C-array interactions across the STIL gene. Panel A: K562 cells; panel B: HPB-ALL cells. The histograms in each panel represent the data obtained for each genomic tiling array element. In each panel, the x-axis is the genomic sequence co-ordinate (NCBI build 35) and the y-axis is the enrichment obtained in 4C-array assays expressed as the relative ratio to the anchor (TAL1 P^{1b}). Exons are shown as vertical blocks with gene name and direction of transcription on top of the diagram. Recognising sites of restriction enzyme used in the 4C-array (Csp6I) are shown as vertical lines in the middle of the panels. Dotted lines (green and grey) through all the panels show the location of STIL exons. The lines in green indicate that the exons exactly align up with the significant novel interaction regions (Csp6I fragments) in K562 and/or HPB-ALL cells. The yellow dotted line indicates the location of control region at TAL1 -25 which was used for determining the statistical significance of all upstream contact points across the STIL gene body. The control region for the interaction at STIL exon18 was at TAL1 -8 as shown in Figure 5.16. The red asterisks indicate the enrichments are statistically significant (t-test, $p < 0.01$)

The STIL +1 region, one of the interacting partners of the TAL1 promoter, was found significantly enriched in both K562 and HPB-ALL cells (Table 5.4). This particular region located at the intron 1 of STIL was related to the STIL-TAL1 rearrangement. As shown in Figure 5.18a, the 5' part of TAL1 is involved in a very precise site-specific deletion of all coding STIL exons, which results in fusion of the coding exons of TAL1 to the first non-coding exon of STIL (Breit et al., 1993; Brown et al., 1990; Jonsson et al., 1991). It was reported that 25% of T-ALL patients had these precise deletions designated as Tal^{d} (Brown et al., 1990) and these rearrangement were not detected by standard cytogenetic analysis (Aplan et al., 1992; Baer, 1993). As a result of that, the TAL1 gene becomes activated and is causative in leukaemogenesis (Kwong et al., 1995).

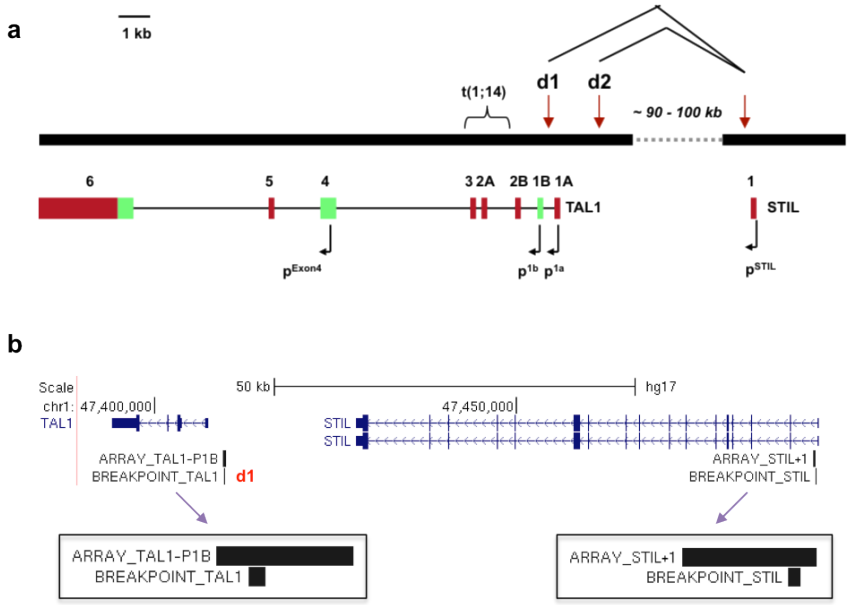


Figure 5.18: Detection of P^{TAL1}-STIL+1 interaction aligns up with the breakpoints of STIL-TAL1 rearrangements. Panel a: a schematic map of STIL-TAL1 rearrangements. The exons and promoters of STIL and TAL1 are displayed below the map; red boxes denote non-coding sequence and green boxes denote coding sequence. The chromosome 1 breakpoints (1;14)(p34;q11) translocations fall within the region indicated by a bracket. The endpoints of the tal^{d1} and tal^{d2}rearrangements are marked with red arrows. These interstitial deletions of 90-100 kb place the STIL promoter and the non-coding STIL exon 1 within or immediately upstream of the TAL1 transcription unit. Panel b: a snapshot of USCS genomic browser shows that the array elements of the anchor and STIL+1 region exactly align up with the deletion points at d1 and STIL intron 1, respectively.

As shown in Figure 5.18b, the STIL +1 region and the TAL1 promoter exactly correspond to the STIL-TAL1 rearrangement sites detected in T-ALL (Kwong et al., 1995). The co-localisation of the STIL+1 region and the TAL1 promoter were identified by the 4C-array assays in K562 (p = 1.1 × 10^{-3}) and HPB-ALL (p = 1.3 × 10^{-6}) as shown in Figure 5.16 and Table 5.4. This suggests that high physical proximity of two endpoints of the *tal* deletion might increase the likelihood of STIL-TAL1 fusion and might also explain why this deletion has a specific breakpoint at the STIL intron1.

5.8 Validation of novel looping interactions detected by 4C-array

A number of significant interactions were detected between the TAL1 promoter and known regulatory elements by 4C-array analysis, which had not been previously assessed by 3C. Two novel interacting regions, the CTS at -31 and the STIL intron 1 (STIL+1), were selected for the further validations due to their potential roles of transcriptional regulation at the TAL1 locus. As previously shown, in this thesis and elsewhere, the -31 region is a known CTCF binding site in K562

cells and also has enhancer activity *in vitro* (Dhami et al., 2010; Solomon and Varshavsky, 1985). Its interaction with the TAL1 promoter may suggest its involvement in regulation of TAL1 transcription. As explained above, the STIL+1 region located at 1 kb downstream of the STIL promoter also aligns up with the deletion point of STIL-TAL1 fusion (Brown et al., 1990; Kwong et al., 1995). Its interaction with the TAL1 promoter not only provides a possible mechanism for the STIL-TAL1 rearrangement, but also implies that TAL1 and STIL might be co-regulated.

3C primers for these two regions and flanking control regions were designed as described in Chapter 3. 3C assays were performed, analysed and quantified by agarose gel electrophoresis, and the relative ligation frequency of each region was determined by normalisation against BAC/PAC control templates as previously described. Combined 3C profiles of two biological replicates of both K562 and HPB-ALL cells are shown as histograms (Figure 5.19). Statistical significance tests of target signals were calculated in comparison with the control region close to the anchor by student's T-test as previously described.

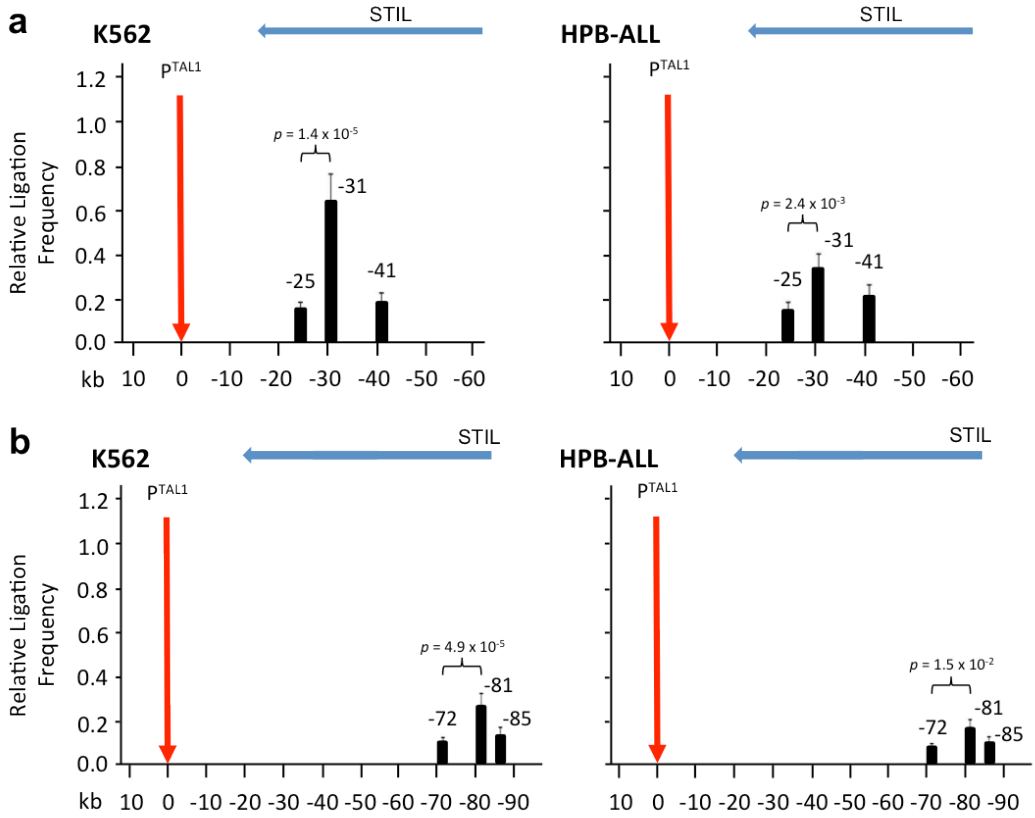


Figure 5.19: 3C validation of the looping interactions captured by 4C-array analysis. Panel A: interaction between PTAL and -31 and Panel B: interaction between PTAL1 and STIL+1. Interaction frequencies (black bars) are shown with standard errors and normalized relative to BAC controls. The location of 3C "anchor" is denoted by vertical red arrow. Locations of genes and their directions of transcription are shown by blue arrows at the top of each panel. The p-values are indicated for interaction frequencies which are significantly higher

for test regions when compared to those of control regions. Scales (in kb) are shown at the bottom of each panel.

As shown in Figure 5.19a, the frequencies of interaction between the TAL1 promoter and CTS at -31 are significantly enriched in comparison with the control region (TAL1 -25) in both K562 and HPB-ALL cells (p = 1.4 x 10^{-5} and 2.4 x 10^{-3}, respectively). Additionally, the relative ligation frequency was much higher in K562 than in HPB-ALL cells, which agrees with observations of the 4C-array results. Similarly, significant interactions between the TAL1 promoter and the genomic region at the STIL intron 1 were also validated by 3C. As shown in Figure 5.19, the interaction frequencies of STIL+1 are significantly higher than the control region in both K562 and HPB-ALL cells (p = 4.9 x 10^{-4} and 1.5 x 10^{-2}, respectively). In summary, these two looping interactions detected by 4C-array method were verified using 3C analysis. Together with the high reproducibility and sensitivity, it has been demonstrated that 4C-array technology is capable of determining *de novo* chromatin interactions between *cis*-acting elements.

Discussions

5.9 Optimisation of the 4C-array method

In comparison with the conventional conditions (1 µg - 36 cycles) of the 4C-array method, the sensitivity and reproducibility has been extensively improved after optimisation in studying *cis*-interactions by increasing library complexity (8 µg) and reducing PCR cycling numbers (18 cycles).

5.9.1 *Sensitivity*

Theoretically speaking, the library complexity is not enlarged with an increase of starting 3C material, based on the assumption of equal representations for all possible ligation products. However, the frequency of two interacting partner being ligated in the genome as well as its efficiency of being subsequently captured by 3C and following 4C manipulations are largely under-represented due to the technical limitation and complex organisation of the genome. Thus, the library complexity of 4C is directly depended on the amount of starting 3C material. For instance, 4C libraries prepared with inadequate starting material (1 µg, as shown in section 5.5.2) failed to capture all interactions detected by 3C. As presented in section 5.6.2, the best sensitivity in terms of detecting known looping interactions

previously captured by 3C was achieved under the conditions of 4 µg - 36 cycles as well as the 8 µg with all cycling conditions. Comparing to 1 µg of starting material, 4 µg and 8 µg of 3C DNA potentially contains additional less-frequent ligation products, which provided a much greater complexity for PCR amplification as well as consequently improved the sensitivity of 4C-array method.

5.9.2 *Reproducibility*

The reproducibility of 4C-array method under the condition of 4 µg - 36 cycles (r_S = 0.800) is much better than the 8 µg - 36 cycles (r_S = 0.686) in terms of its ability to reassemble the interaction profiles captured by 3C. Increasing the amount of 3C material means simultaneously increasing the representation and subsequent concentration of the abundant templates in the PCR amplification. Consequently, with higher numbers of the PCR cycles (36 cycles), over-saturation of these abundant interactions was in a much higher degree for the 4C library prepared with 8 µg 3C DNA than 4 µg 3C DNA. Within the group of using 8 µg 3C DNA, 4C libraries amplified with 18 cycles of PCR illustrated the highest reproducibility of both local (comparison with 3C profile) and overall (comparison with biological replicates) profiles, in comparison with the 4C-array profiles generated with 27 and 36 cycles of PCR amplification. It suggests that over-cycling with 27 or 36 rounds of PCR has a high-probability of saturating the abundant templates in the 4C library at very early stage of amplification, which consequently compresses the dynamic range and makes the detection of interactions less quantitative. In addition, unpublished data (S.Kurukuti) of 4C-seq libraries prepared using 36 cycles of PCR amplification illustrates over-saturated of *cis*-interaction profiles, so it is impossible to determine any real looping interaction locally. It also suggests that the quality of 4C data could be largely compromised due to over amplification.

In section 5.6, it has demonstrated that the sensitivity and reproducibility are greatly improved by a combination of increased library complexity as well as reduced PCR cycling conditions. As a result of that, the abundant templates are less likely to be over-amplified on one hand, while on the other hand the less-frequent ligation templates are also included for a better representation of the overall interaction profile.

5.10 **Possible ways for further optimisation of the 4C-array method**

Although a significant improvement has been achieved for the sensitivity and reproducibility of the 4C-array method by adjusting the amount of starting material as well as PCR cycling conditions, a number of other steps and experimental conditions can potentially introduce bias to the method that may require further optimisation. For instance, the efficiency and fidelity of the biotinylated primer extension step may require to be validated by sequencing, in order to ensure carbon-copying all co-associates with the anchor. In addition, the sonication step can also be optimised for generating DNA fragments within a suitable size-range. One needs to consider both the restriction enzyme being used and the size distribution of restriction fragments within the particular genomic region of interest. Based on that, the sonication conditions can be optimised to achieve suitable sized fragments for downstream primer extension but without losing interacting co-associates due to over-sonication.

Last but not the least, generating a custom designed array with probes only recognising the regions around the restriction sites of Csp6I (or other restriction enzyme being used) can be a cost-effective way for profiling genome-wide interacting co-associates (Simonis et al., 2007). Nevertheless, compared to the designed tailored array (resolution ~7 kb, see Chapter 1, section 1.2.4), the TAL1 tiling path array provides a much better resolution (~400 bp) for studying local interactions.

5.11 **A full profile of the TAL1 promoter interactions and the model of TAL1 “cruciform” configuration in erythroid K562 cells**

In K562, distribution of interacting partners of the P^{TAL1} across the TAL1 locus fully supports the previously proposed “cruciform” configuration as shown in Figure 5.20. Three major clusters of interactions detected by 4C-array are centred at +51, +19 and P^{TAL1}, which are the three key contact points of the “cruciform” configuration speculated to be linked by the TEC (Figure 5.20). 4C-array profile also confirms the speculation that CTSs at +57/+53 are brought into close proximity with the TAL1 promoter by the +51 enhancer. In addition, the TAL1 promoter is found to be in contact with the TAL1 -31, which is in agreement with

the observation of chromatin loops between CTSs +57/+53 and -31 (see Chapter 4), as all these *cis*-acting regulatory elements are situated at the centre of this configuration (Figure 5.20). The *cis*-acting regulatory elements including P^{Exon4}, +3, +1, -4, -7, -10 and -13 are also found to interact with the TAL1 promoter by 4C-array and/or 3C, which also fit the TAL1 "cruciform" model. However, it's virtually impossible to determine whether these interactions are functionally related or only because of being in close-proximity to the TAL1 promoter (anchor) by their very nature. 3C analysis assessed interactions between TAL1 promoter and its enhancers related to the TAL1 transcriptional regulation (see Chapter 3). In addition to that, 4C-array assays illustrate an extra layer of insight, by integrating the TAL1 neighbouring genes into the center of the cruciform organization (Figure 5.20). No data so far has suggested that TAL1 and its neighboring genes are co-transcribed, although one previous study does show evidence that the expression of STIL and PDZK1IP1 is always found in cells that express TAL1 (Delabesse et al., 2005). Chromatin interactions between the promoters of TAL1 and neighboring genes may provide a possible mechanistic solution of transcriptional co-regulation at the TAL1 locus.

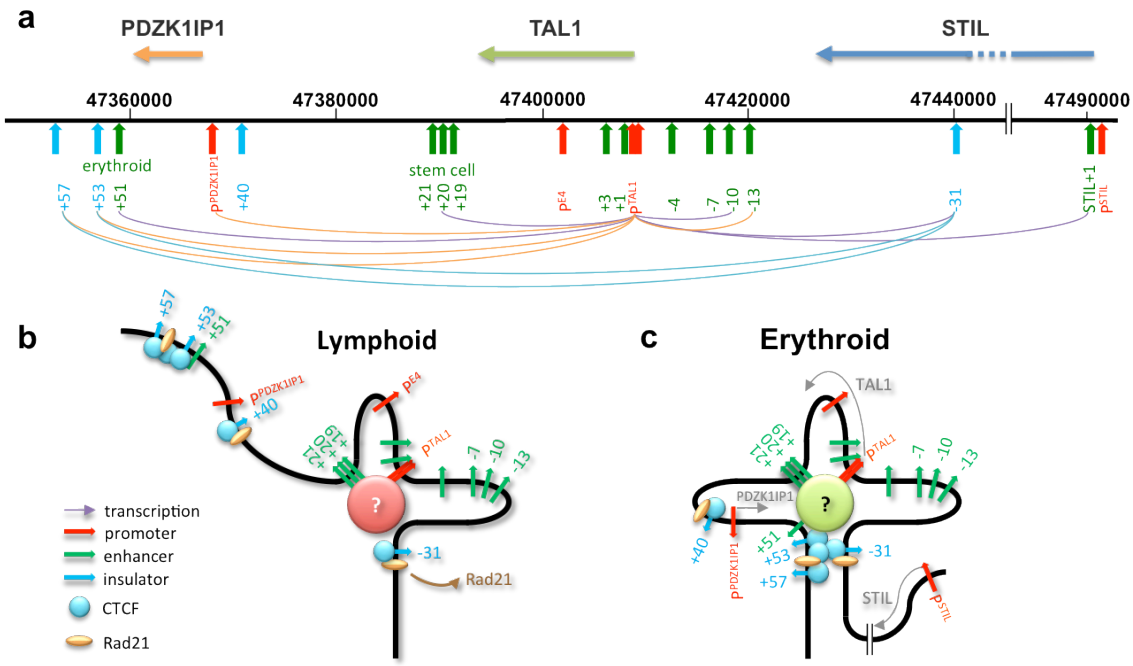


Figure 5.20: Predicted structural model of TAL1 chromatin organization. (a) Schematic diagram depicting the organization of the human TAL1 locus and looping interactions between its regulatory elements in K562 cells based on accumulated 3C and 4C data (Chapter 3,4 and 5). The scale is genome co-ordinates (bp) for human chromosome 1 (hg.17). Promoters, enhancers and CTCF binding sites are the vertical red, green and blue arrows respectively. The cyan and orange lines represent the interactions detected only by 3C-PCR or 4C-array, whereas the violet lines represent the interactions capture in both analysis. (b) Organization of the chromatin hub in $TAL1^-$ lymphoid cells. Unknown factors mediating the interaction of the stem cell enhancers, the TAL1 promoters and the -31 region in lymphoid cells are represented by the red ball. (c) Organization of the chromatin hub in $TAL1^+$ erythroid cells. Locations of promoters, enhancers, and putative insulators are

depicted as above. Direction of transcription of relevant genes (grey arrows), unknown transcriptional complex are situated at the centre of the hub, and CTCF and Rad21 binding at insulators are also shown and detailed in the key.

In contrast, the chromatin configuration of TAL1 locus in the TAL1 non-expressing lymphoid cells (i.e. HPB-ALL) is much less sophisticated in terms of the numbers of interacting partners (*cis*-acting elements) as well as the formation of chromatin loops (Figure 5.20). The interaction between the TAL1 promoter and the stem cell enhancer (+20/+19) is maintained, although TAL1 is not expressed in the lymphoid lineage. Similarly, the stem cell enhancer also does not require for TAL1 expression in the erythroid lineages (Ogilvy et al., 2007; Spensberger et al., 2012) However, this specific interaction has been observed in both human and mouse erythroid and lymphoid lineages, suggesting that it may not be a loop with context-dependent roles but a universal structure that exists in different lineages. It is speculated that this chromatin loop may also exist in the hematopoietic stem cell (HSC) and progenitor populations, where this particular enhancer is important for driving TAL1 transcription (Gottgens et al., 2002). In addition, this particular looping structure may be inherited from the HSC populations and conserved through differentiation. Nevertheless, little is known about how this particular looping structure is formed and maintained in cell types where TAL1 is not even transcribed.

In addition, the erythroid-specific looping interaction between the +51 erythroid enhancer and the TAL1 promoter does not exist in lymphoid cells, where TAL1 is transcriptionally inactive. As a result, the genomic regions downstream of the TAL1 gene (+20 enhancer onwards) tend to move away from the TAL1 looping structure, which is reflected by much lower level of interaction frequencies between TAL1 promoter and these regions comparing to erythroid cells (K562). On the other side, both 3C and 4C results indicate that the TAL1 promoter also weakly interacts with the -31 and STIL +1 in lymphoid cells. Functional related studies are required to explore the underlying mechanisms of these looping interactions in transcription regulation of the TAL1 locus in lymphoid cells.

5.12 **Cross-talk between TAL1 and STIL genes and the TAL1-STIL rearrangement**

A number of contact points have been identified between the STIL gene body and the TAL1 promoter in both K562 and HPB-ALL cells, most of which align with

exons of the STIL gene. These contact points with TAL1 promoter imply that the STIL gene may get fully transcribed when moving through the active TAL1 chromatin hub. It is speculated that these contain points may be related to the nonsense-mediated decay of STIL during co-transcription process with TAL1.

Most interestingly, the contact point at the STIL +1 region perfectly aligns up with one of the breakpoints of STIL-TAL1 micro-deletion in T-ALL (see section 5.7.4). Although a perfect match has been found between the breakpoints of STIL-TAL1 rearrangement and the particular looping interactions (P^{TAL1}/STIL +1) in K562 and HPB-ALL cells, the STIL-TAL1 micro-deletion deletion is only detected in some of the T-ALL patients but not in patients which do not have the disease. In addition, K562 is derived from a CML patient (Lozzio and Lozzio, 1975), instead of T-ALL cell lines such as HPB-ALL, which suggests that the looping interaction between deletion points may pre-exist in patients with haematological malignancies and even T-ALL patients without deletion. This particular interaction brings two breakpoints into a spatial close-proximity that increase the likelihood of chromosome rearrangement between these two sites, although the underlying mechanism driving formation of this particular loop remains to be fully studied. Moreover, the number of contact points across the STIL gene in HPB-ALL is much less than in K562, suggesting the likelihood of having the particular STIL-TAL1 rearrangement in T-ALL cells is increased comparing to the CML progenitor cells. Taken together, these novel interactions between TAL1 promoter and the STIL gene not only provide a deep understanding about co-transcription but also propose a potential mechanism of chromatin rearrangement via looping.

5.13 **Study intra- and inter-chromosomal interaction in using 4C technology**

4C is a powerful technology for identifying local chromatin looping interactions in a high-resolution using high-throughput sequencing or genomic tiling microarray (e.g. the TAL1 tiling array). For example, the local chromatin organisations of the Hox gene clusters have been characterised by 4C-seq, revealing a dynamic chromatin configuration that is transcriptionally related, as gene activation is paralleled by a transition from one repressed domain (marked with H3K27me3) to another active domain (marked with H3K4me3) (Noordermeer et al., 2011). It suggests that genes with the same transcriptional state tend to co-localise in three-dimensional

space, and both transcriptionally “positive” and “negative” gene compartments have their own chromatin organisations in order to maintain their own transcriptional states. This observation is also in agreement with the local interaction profiles detected at the TAL1 locus. Although TAL1 is transcriptionally inactive in the lymphoid lineage, the chromatin still preserves a secondary structure in the TAL1 locus. In a broad picture, it is speculated that the TAL1 locus in erythroid and lymphoid lineages may be situated in different chromatin domains, which is similar to the Hox gene clusters.

Although the scope of 4C-array is currently limited within the TAL1 locus in using the TAL1 array, it can be adapted for high-throughput sequencing application based on the e4C protocol (Sexton et al., 2012), although inadequate sequencing depth is still the barrier for the 4C technologies in profiling inter-chromosomal interactions at base-pair resolution (Raab et al., 2012; Sexton et al., 2012). Nevertheless, further studies about profiling long-range intra-chromosomal interactions of the TAL1 gene by 4C-seq in combination with gene expression analysis will provide a much broader picture about local transcriptional environment and its relationship with chromatin organisation.

In addition, it will be interesting to explore other interacting partners of the TAL1 gene genome-wide, especially for those erythroid-specific genes being regulated by same transcription factors (Kerenyi and Orkin, 2010). A good example of studying the transcriptional interactome was provided by e4C incorporated with RNAP II ChIP using the mouse globin gene in the erythroid cells (Schoenfelder et al., 2010). It was found that the active globin genes associate with hundreds of other transcribed genes, which illustrated context-dependent intra- and inter-chromosomal interactomes. In addition, genes regulated by a particular transcription factor Klf1 are preferentially co-localised at specialized transcription factories, suggesting active co-regulated genes cooperate with their transcription factors to establish specialized compartments for efficient and cooperative transcription regulation. Thus, 4C-array analysis can be incorporated with a restriction enzyme-specific tailored microarray platform for profiling genome-wide interacting co-associates of the TAL1 gene. This will enhance our understandings about how a group of functionally related genes can be co-regulated for their transcription in a three-dimensional space.

Last but not the least, it will be also interesting to look at the TAL1 chromatin configuration in the HSC populations. As previously discussed, the looping interaction between the TAL1 promoter and the stem cell enhancer (+20/+19) exists in both erythroid and lymphoid lineages. To verify whether this particular interaction is derived from the stem cell population and preserved all the way through the differentiation, 4C-array can be applied to determine the TAL1 chromatin configuration in the HSC populations. In addition to that, 4C-array can be also be used to characterise the transitions of chromatin configuration at the TAL1 locus during the haematopoietic development.

Conclusions

The work described in this chapter present an improved 4C method with high resolution, sensitivity and reproducibility on the TAL1 tiling array platform. Furthermore, 4C-array captured not only loop interactions previously identified by 3C-PCR, but also novel interactions with known *cis*-regulatory elements as well as novel regions across the entire TAL1 locus. It suggests that 4C-array technology is capable of providing new insights to the local chromatin organisation in a high-resolution. In addition, 4C-array identified a chromatin interaction between breakpoints at the TAL1 promoter and the STIL +1 element in the haematopoietic lineages, proposing a possible mechanistic solution for the STIL-TAL1 rearrangement in T-ALL. Last but not the least, the interaction profiles captured by 4C-array further support the predicted TAL1 cruciform model in erythroid cells and provide an additional layer of evidence in linking transcription regulation of TAL1 and its neighbouring genes via chromatin loops. In the subsequent chapter, the study will focus on revealing the relationship between transcription regulation and chromatin organisation at the TAL1 locus.

Chapter 6 Looping interactions at the TAL1 locus are GATA1 dependent

Summary

There are a number of evidence which imply that the promoter-enhancer looping interactions at the TAL1 locus in the erythroid cells may be mediated via the TAL1-containing erythroid complex (TEC). To further elucidate the dependence of TAL1 transcriptional regulation and chromatin looping configuration on the TEC, siRNA knockdown system has been applied to deplete a key TEC member - GATA1 in K562 cells. FACS analysis was used to monitor the siRNA transfection efficiency of electroporation. RT-qPCR and western blotting results illustrated that the highly efficient knockdown of GATA1 (over 95% depletion at mRNA and protein levels) was achieved after two rounds of siRNA electroporation at the 96-hour time-point. Results of RT-qPCR, 3C-PCR and ChIP-qPCR analysis illustrated that the complete depletion of GATA1 resulted in (i) reduced expression of TAL1 and its adjacent genes, (ii) reduction of RNA polymerase II recruitment at promoters and enhancers, (iii) loss of TEC (GATA1, LDB1 and TCF3) occupancy at its target sites at the P^{TAL1} and the TAL1 +51 enhancer, iv) reduction of CTCF/Rad21 occupancy at the CTS at -31 and reduction of interacting frequency between CTSs at +57 and -31, (v) loss of looping interactions between the TAL1 promoter and its enhancers and (vi) reduction of interacting frequency between the TAL1 promoter and the STIL +1. The analysis in this chapter demonstrated that the “cruciform” configuration and the transcription of TAL1 and its adjacent genes are GATA1-dependent in erythroid K562 cells.

6.1 Introduction

The previous chapters have proposed a sophisticated chromatin interacting model for the TAL1 locus in the erythroid K562 cells, which is known as the “cruciform” configuration. This predicted model relies on the assumption that all chromatin interactions detected by the 3C and 4C analyses are existed in the same cells at the same time. In addition, accumulated evidences suggest that the major chromatin interaction (between the TAL1 promoter and the +51 enhancer) within the “cruciform” structure might be mediated via a transcription factor complex known as the TAL1-containing erythroid complex (TEC).

The siRNA-induced knockdown is a straightforward tool used to rapidly assess gene function via inhibition of the target gene expression at the post-transcriptional level. Therefore, this approach is considered to be a useful approach to elucidate the possible roles of TEC in facilitating chromatin interactions and consequently regulating transcription of the TAL1 locus by knocking-down key members of the complex.

6.1.1 *The TAL1-containing erythroid complex (TEC)*

TAL1, a haematopoietic specific bHLH transcription factor forms a trans-activating protein complex in erythroid cells along with other transcription factors including E2A/TCF3 (a member of the E-protein family), LMO2 (LIM domain only 2), LDB1 (LIM domain-binding protein 1) and GATA1, which specifically recognizes a characteristic DNA motif consisting of a GATA and an E-box (CAGGTC) sites (Wadman et al., 1997). This GATA/E-box motif is normally separated by 9 to 12 bp, which can be found at a number of erythroid-specific genes (Tripic et al., 2009).

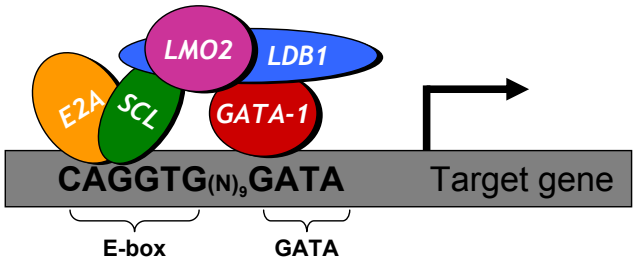


Figure 6.1 Model of the TAL1 erythroid complex. The complex binds to the GATA/E-box motif spaced at 9-12 bp. The DNA binding modules are provided by the E2A/TCF3-TAL1 heterodimer (E-box binding) and GATA1 (GATA binding). The LIM-only protein (LMO2) together with LDB1 links the two DNA-binding modules. Target genes with GATA/E-box motif(s) provide recognition sites for the TAL1 erythroid complex. Transcriptions of target genes are activated by the TEC.

As shown in Figure 6.1, TAL1 binds to the E-box elements by forming the heterodimers with E-protein family members E2A/TCF3 (transcript variants E12 and E47). GATA1 and TAL1 associate at regulatory regions containing juxtaposed GATA and E-box elements, and subsequently assemble higher-order protein structures anchored to DNA via GATA1 and TAL1/E2A heterodimer. LMO2 and LDB1 do not directly bind to DNA - instead, they act as bridging protein in linking between TAL1-E2A heterodimer and GATA1. In this chapter, this multifactorial complex is thereafter termed as TAL1-containing erythroid complex (TEC).

6.1.1.1 Roles of the TEC in transcription regulation

The TEC plays a key role in activating erythroid-specific genes, in most of the cases via occupying at the consensus GATA/E-box motifs at the gene promoters or other cis-acting elements.

Kit cytokine receptor gene (c-kit): The ChIP analysis has revealed that the TEC together with an additional member SP1, assemble at the c-kit promoter and the combinatorial occupancy of entire complex is critical for the synergistic transactivation of c-kit (Lecuyer et al., 2002). However, the c-kit gene is regulated in a manner that requires the Sp1-binding site instead of a GATA or E-box motifs. In addition, it has been demonstrated that the retinoblastoma protein (pRb) can associate with TAL1-cotaiining complex in erythroid cells, which leads to repression of c-kit promoter activity (Vitelli et al., 2000).

α- and β-globin genes: It has been shown that TEC binds to both the locus control region (LCR) and promoters of the human β-globin locus during erythroid differentiation and formation of long-range chromatin looping interaction between the LCR and the β-globin promoter is meditated by LDB1 (Song et al., 2007). In addition, occupancy of the TEC has also been reported at both mouse and human α-globin loci by DNase I hypersensitivity assays and ChIP-chip analysis (Anguita et al., 2004; De Gobbi et al., 2007). However, no functional analysis of the α-globin clusters has been performed to investigate roles of the TEC in transcriptional regulation.

Protein 4.2 (P4.2): The Protein 4.2 (P4.2) gene encodes an important component of the erythrocyte cell membrane skeleton, which is also regulated by the TEC. It has been demonstrated that the TEC activates transcription of P4.2 via two GATA/E-box motifs at its promoter in the erythroid cells (Xu et al., 2003). Both GATA/E-box motifs and all members of TEC are required for the maximal transcription of P4.2. TEC has been found to associate with the SWI/SNF protein Brg1 and down-regulate P4.2 expression by recruiting chromatin-remodelling complexes and histone modification enzymes (Xu et al., 2006).

GATA1: The *cis*-acting regulatory element (HS I) upstream of the GATA1 in mouse contains a consensus GATA/E-box motif and is shown to be bound by the SCL/TAL1 eyrthroid complex (Vyas et al., 1999). In addition, point mutations of the

GATA1 but not the E-box motif result in a complete abolishment of the function of the element. A numbers of examples listed above reveal a genome-wide mechanism of the TEC in transcription regulation of the erythroid-related genes.

6.1.1.2 The TEC regulates the TAL1 transcription in erythroid lineages

Two highly conserved consensus GATA/E-box motifs have been identified at the TAL1 erythroid enhancers (+51 in human and +40 in mouse) by a four-way (human/dog/mouse/rat) sequence alignment (Ogilvy et al., 2007). In all four species, the 5' GATA site and E-box motif is separated by 9 bp while the 3' GATA/E-box motif had a spacing of 6 bp. In addition, the ChIP analysis has demonstrated that the Tal1 erythroid enhancer (+40) is physically bound by GATA1 and TAL1 in the murine erythroid F4N cells. Moreover, it has been shown by *in vivo* transgenic analysis that the mutation of the 5' GATA/E-box motif results in almost a complete abolishment of the erythroid enhancer activity whereas the effects of mutating 3' GATA/E-box motif is minimal, suggesting that the 5' GATA/E-box is essential for maintenance of enhancer function. It has been speculated that the structure of 5' GATA/E-box is in accordance with the consensus sequence (9 bp reported by Wadman) whereas the 3' GATA/E-box motif exhibits 6-bp spacing which is previously shown to preclude recruitment of the TEC in erythroid lineages (Wadman et al., 1997).

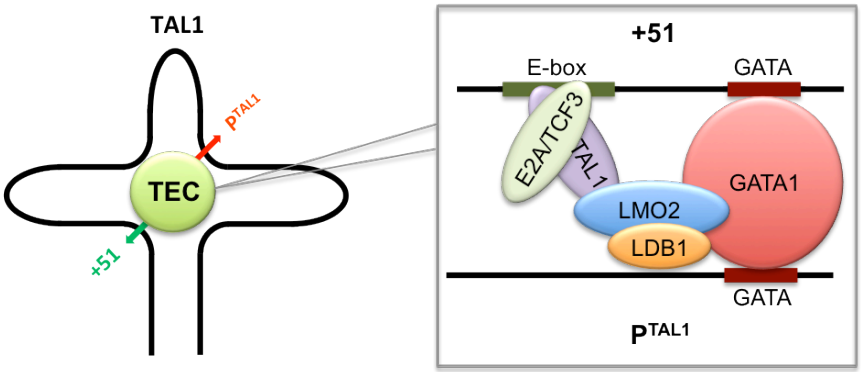


Figure 6.2: A model of the TAL1-containing erythroid complex at the TAL1 locus. Predicted occupancy patterns of TEC members at the TAL1 promoter (1a) and the +51 enhancer are illustrated on the right panel.

Recently, it has been shown that the TEC is bound to the GATA/E-box motif at the TAL1 +51 enhancer in human erythroid cell lines (Dhami et al., 2010). In addition, occupancy of the TEC has also been observed at the TAL1 promoter 1a in human erythroid cells (H.L.Jim's PhD thesis, University of Cambridge, 2008). However, it is intriguing that the binding of the entitle TEC can be detected at the promoter 1a, providing the fact that only GATA sites but no consensus juxtaposed E-box motif

has been identified at the TAL1 promoter so far. Additionally, it has been found that the TAL1 promoter 1b locates ~1 kb upstream of the promoter 1a, is physically in contact with the +51 erythroid enhancer captured by the 3C-PCR and 4C-array analyses in K562 cells (see Chapter 3 and 4).

Based on the above evidence, a predicted model of the TEC-mediated contact between the TAL1 promoter and the +51 enhancer in K562 cells is illustrated in Figure 6.2. A GATA1 protein has multiple zinc fingers, which are capable of binding to different DNA domains. Accumulated evidences promote the speculation that the GATA1 protein may act as a bridge in linking the two elements via chromatin loop in the first place (see Chapter 3). Subsequently, recruitment of other TEC members, such as E2A and TAL1 at the +51 GATA/E-box motif leads to formation of the entire TEC, which co-occupies at both the TAL1 promoter and the +51 enhancer to regulate the TAL1 transcription in erythroid K562 cells (Figure 6.2). Additionally, it is also speculated that transcriptional regulation of theTAL1 neighbouring genes may be related to the TEC, as previous results illustrated that promoters of these genes are in contact with the TAL1 promoters in the “cruciform” model (see Chapter 5).

6.1.1.3 Why is GATA1 is the best candidate for studying regulation of TAL1 expression via TEC?

The previous study about the TEC in regulating transcription of its target genes (H.L. Jim’s PhD thesis, 2008) provides an outline on selecting the best candidate for the knockdown analysis of the TEC in this thesis:

1. Availability of specific antibodies for ChIP and western blotting analysis: antibodies for GATA1, TAL1, LDB1 and E2A detect the protein unambiguously in western blotting. It reported the lack of working antibody for LMO2. ChIP antibodies for GATA1, LDB1 and E2A gave high enrichments across the TAL1 locus. In contrast, the antibody for TAL1 showed much weaker enrichments, and again lacked for the working antibody for LMO2 ChIP analysis.

2. The efficiency of siRNA knockdown in K562 cells: it has been reported that the siRNA for GATA1 has the highest knockdown efficiency among all five members of the TEC (as illustrated in Table 6.1).

Table 6.1: The percentage of mRNA remained after siRNA knockdown (KD) at 12, 24, 36 and 48 h time-points are shown for each of the five transcription factors (H.L. Jim's PhD thesis, 2008).

	12 h	24 h	36 h	48 h
GATA1 KD	13%	11%	18%	19%
E2A KD	38%	21%	26%	51%
TAL1 KD	28%	40%	36%	50%
LDB1 KD	16%	13%	18%	28%
LMO2 KD	31%	23%	21%	32%

3. Disrupting a protein that binds to DNA directly such as GATA1, might have a higher likelihood of disrupting the whole complex comparing to co-factor protein such as LDB1.

Providing the facts that listed above, GATA1 is considered as one of the best candidate for siRNA knockdown study in human erythroid K562 cells for this thesis.

6.1.2 *Studying transcriptional regulation complexes in using RNA interference technology*

The β-globin locus is a well-established model for studying transcription regulation. During the last decade, it has been assessed for its chromatin loop formation between regulatory elements, especially the interactions between the LCRs and promoters that form a "chromatin hub" in actively transcribed globin genes in erythroid cells (Palstra et al., 2003; Tolhuis et al., 2002). A number of erythroid activators including EKLF, GATA1 and FOG1 have been shown to be required for the lineage-specific close-proximity of regulatory elements in the β-globin locus (Drissen et al., 2004; Vakoc et al., 2005).

Most interestingly, a recent study in Dean's laboratory has shown that occupancy of the entire TEC at the β-globin locus LCR and promoter is critical for transcription activation during erythroid differentiation (Song et al., 2007). This study also reveals that TEC is responsible for RNAP II recruitment at β-globin promoter, transcription of β-globin genes and formation of chromatin loops between promoters and its LCR during the erythroid differentiation by knocking-down LDB1 using RNAi. A subsequent study from the same laboratory has further illustrated that the LDB1 is capable of modulating the gene expression at multiple levels (Song et al., 2010). As a key member of TEC, the LDB1 stabilizes the entire complex on the β-globin locus. Moreover, it is also required for the recruitment of

P-TEFb, which phosphorylates the C-terminal domain of RNAP II at Ser2 for efficient elongation. Most importantly, the study also revealed the role of LDB1 in controlling migration of the β-globin locus from the nuclear periphery to transcription factories for robust transcription.

Latest studies conducted in Kim's laboratory have revealed that an additional erythroid activator, NF-E2, associated with GATA1 is involved in transcription of the human globin genes (Kim et al., 2012; Woon Kim et al., 2011). RNAi assay was applied to study the role of two erythroid-specific transcriptional activators, GATA1 and NF-E2 in the human K562 erythroid cells. It demonstrated that the GATA1 and NF-E2 knockdown inhibited the transcription of globin genes as well as the formation of chromatin loops at the β-globin locus. Models of chromatin looping configuration of the human β-globin locus were proposed based on these results. It illustrated that the entire chromatin structure was abolished by GATA1 knockdown whereas the chromatin loop mediated by CTCF between insulator elements at HS5 and 3'HS1 was not affected by NF-E2 knockdown.

These studies provide good examples for using RNAi technologies to assess the role of transcription factor complex in regulating gene expression as well as modulating chromatin looping interactions. Approaches including RT-qPCR, western blotting, ChIP and 3C analyses have been widely used in these studies to monitor the alternations of target gene transcription, TF occupancy over target sites and looping interactions.

6.1.3 *RNA interference (RNAi)*

The RNA interference (RNAi) technology has been developed based on the discovery of silencing of specific genes via double-stranded RNA (dsRNA) in *Caenorhabditis elegans* (Fire et al., 1998). Fire and colleagues have shown that only a few dsRNA molecules are sufficient to almost completely suppress the expression of a specific gene that is homologous to the dsRNA. Subsequently, this technology has been adapted to be used in the mammalian system, for which allows not only elucidating gene functions but also developing antiviral therapeutics (McManus and Sharp, 2002). In the section, the main focus is on discussing the use of siRNA for RNA interference.

6.1.3.1 Mechanism of siRNA for RNA interference and its applications

The mechanisms of RNAi functioning are illustrated in Figure 6.3. RNA interference (RNAi) is a cellular process whereby introduced double-stranded RNA can induce degradation of target RNA with sequence similarity (Fire et al., 1998). Firstly, the dsRNA is broken down into 21-23 nt RNAs by the RNase III-like protein Dicer which is known as the short interfering RNAs (siRNAs). Subsequently, the siRNAs associate with the RNAi silencing complex (RISC) in recognition of the complementary target mRNA. Eventually, the cleavage of the target mRNA occurs in the center of the region complementary to the siRNA and leads to degradation of the target mRNA and recycling of the RISC complex (Hamilton and Baulcombe, 1999; Hammond et al., 2000; McManus and Sharp, 2002).

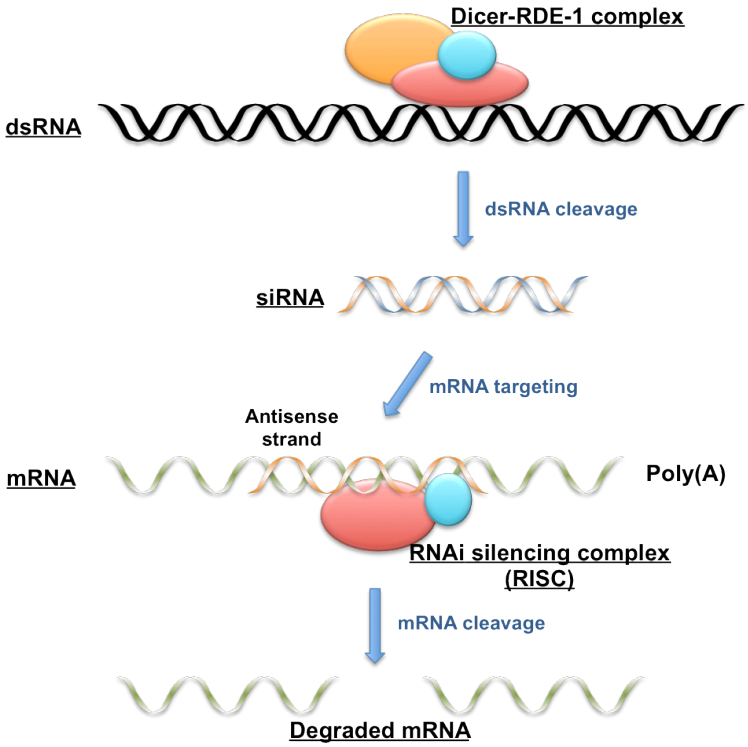


Figure 6.3: Schematic diagram of short interfering RNAs (siRNAs) for RNA interference: The long double-stranded RNA (dsRNA) is cleaved by the Dicer–RDE-1 (RNAi deficient-1) complex to form siRNAs. The antisense strand of the siRNA is used by the RNAI silencing complex for mRNA cleavage, resulting in mRNA degradation.

Since this post-transcriptional gene silencing strategy initially derived from *C. elegans*, it has been rapidly developed as a tool used for identifying and assigning gene functions in mammalian cell lines (Elbashir et al., 2001; Holen et al., 2002). It has also been used to knockdown the expression of particular genes, which are involved in certain cell signalling pathways (Jazag et al., 2005) or used in relation to the understanding of mechanisms of disease (Diakos et al., 2007). Additionally,

in combination with expression analysis by microarray and computational or experimental promoter studies, RNAi has been used to dissect transcriptional regulation network of key factors, which are involved in apoptosis as well as self-renewal of ES cells (Elkon et al., 2005; Jiang et al., 2008). Moreover, RNAi technologies have been widely used in genetic screens (Silva et al., 2005) and therapeutically treatment for various human diseases (Kim and Rossi, 2007).

6.1.3.2 Establishing siRNA knockdown

It is important to perform a time-course experiment for the siRNA knockdown to determine the time-point yield of the highest level of knockdown at levels of both the mRNA and protein. Normally, the siRNA knockdown results in rapid reduction at the mRNA level, which is readily observable within 18 hours or less (Follows et al., 2006). Nevertheless, the turnover time for different proteins can be different, as stable proteins may require a longer period of time for siRNA suppressing before being knocked down in comparison to less stable ones. Generally, the siRNA-induced knockdown lasts approximately 3-5 days for most cell lines (McManus and Sharp, 2002). However, the longer the time the target gene is silenced via siRNA knockdown, the more downstream genes will get affected. Thus, it is a general view that using an earlier, rather than a later, time point in perturbation studies reduces the likelihood of identifying non-specific effects at the level of gene expression. Transient knockdown by siRNA occurs between 24 to 120 hours after transfection of the siRNA to the cells (Holen et al., 2002). This time interval is sufficient for the experimental observation of changes of downstream effects. Indeed, the time required for the maximum RNAi efficiency was shown to be proportional to the half-life of the target protein (Choi et al., 2005). Thus, all of these factors need to be considered when deciding on the appropriate experimental conditions to analyse the biological effects of siRNA knockdowns.

6.2 Aims of the chapter

This chapter used siRNA knockdown of GATA1 to determine whether the TEC is critical for mediating chromatin looping at the TAL1 locus. The aims to be addressed by the study were as follow:

1. To determine the time point of maximum siRNA knockdown of GATA1 in terms of mRNA and protein expression, occupancy over target sites and disruption of the known looping interaction.

2. To further characterize the possible effects of GATA1 knockdown including

   i) Disruption of transcription at the TAL1 locus,

   ii) Disruption of recruitments of related transcription factors,

   iii) Disruption of the known looping interactions and the “cruciform” configuration

6.3 Overall strategy

Efficient knockdown for GATA1 in the TAL1 erythroid complex (TEC) was required for determining the GATA1/TEC-dependence of transcription regulation and chromatin configuration at the TAL1 locus. GATA1 siRNA knockdown system was previously established in human K562 cells (H. L. Jim’s PhD thesis, University of Cambridge). To this end, the overall strategy of this chapter is summarised in Figure 6.4. Five experimental approaches were used for this study, which included siRNA knockdown, RT-qPCR for detecting mRNA expression, western blotting for detecting protein expression, ChIP-qPCR for detecting occupancy of target transcription factors and 3C-PCR for deterring chromatin looping interactions (Figure 6.4 panel a). The siRNAs against either GATA1 or luciferase (control) were transfected into K562 cells by electroporation (Figure 6.4).

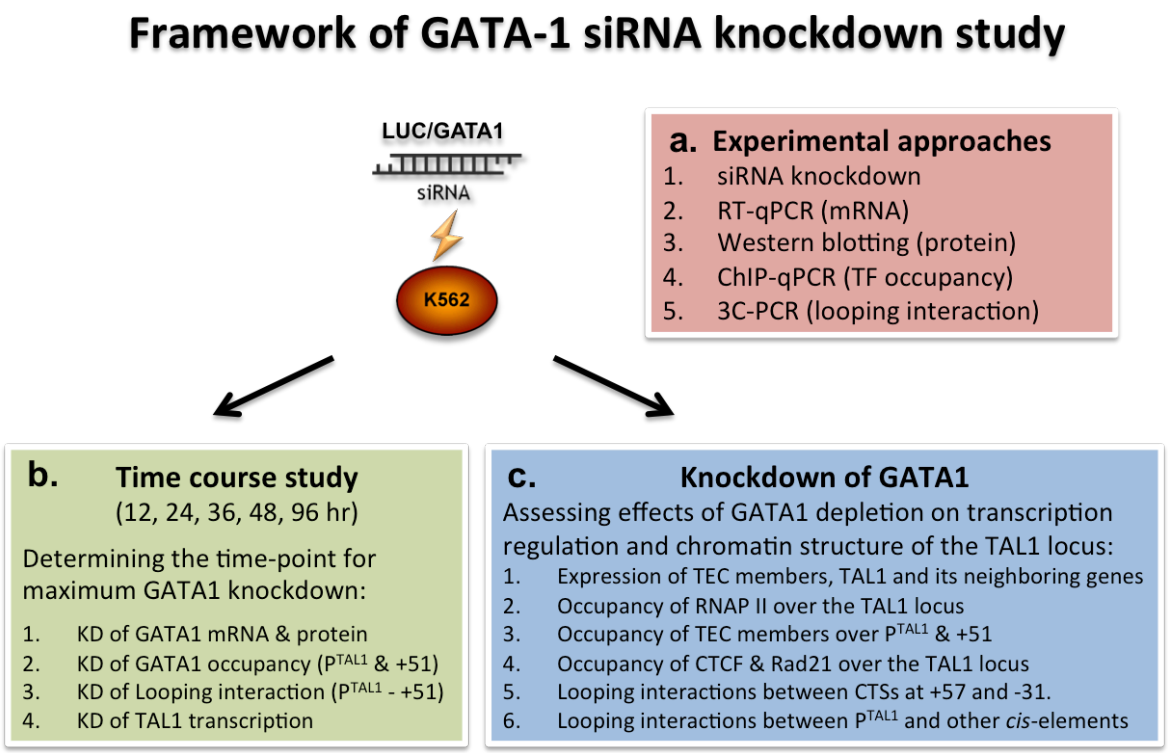


Figure 6.4: Overall strategy of siRNA knockdown analysis of GATA1. Panel A: Five experimental approaches used in the study. The human erythroid K562 cells transfected with GATA1 or luciferase (control) siRNA are used for the analysis listed in panel B and C. Panel B: Time-course study to determine the time-point for maximum GATA1 knockdown. Panel C: Determination of the effects of GATA1 depletion on transcription regulation of TAL1 for the chosen time-point based time-course study.

Firstly, time course study was performed to determine the time-point for maximum GATA1 knockdown (Figure 6.4, panel b). The knockdown of GATA1 was monitored at four levels, including i) mRNA and protein expression, ii) ChIP-occupancy at target sites, iii) disruption of a particular known looping interaction and iv) disruption of TAL1 expression (see results section 6.4). Secondly, the effects of GATA1 depletion on transcription regulation of the TAL1 locus were assessed at three different levels (Figure 6.4, panel c). The 1st level is disruption of expression of target genes (bullet point 1); the 2nd level is disruption of bindings of target transcription factors over their binding sites (bullet points 2-4) and the 3rd level is disruption of looping interactions mediated by those transcription factors (bullet points 5-6).

Results

It has been proposed that the transcription of TAL1 is regulated via chromatin loops between the P^{TAL1} and the +51 erythroid enhancer mediated by the TAL1 erythroid complex (TEC) in erythroid K562 cells. Whether the TAL1 active loops are dependent on the TEC can be verified by knockdown study. One of the TEC key components, GATA1, was selected as the best candidate for siRNA knockdown study based on the reasons illustrated in section 6.1. Two siRNAs were used in this thesis; one against GATA1, and one against firefly luciferase used as a negative control for all the siRNA experiments performed in this thesis (Table 6.2). These siRNAs had been used for previous studies, and ~90% knockdown of mRNA had been achieved in K562 cells using GATA1 siRNA (H.L. Jim's PhD thesis, 2008, University of Cambridge).

Table 6.2 Characterisation of siRNAs for knockdown of the GATA1. The siRNA sequences, target gene are shown in the table

Target gene	Sense sequence (5' to 3')	Antisense sequence (5' to 3')
Firefly Luciferase	CUUACGCUGAGUACUUCGAtt	UCGAAGUACUCAGCGUAAGtt
GATA1	GGAUGGUAUUCAGACUCGAtt	UCGAGUCUGAAUACCAUCCtt

The siRNA knockdown analysis performed in this chapter followed the procedures as previously described (H.L. Jim's PhD thesis). The transfection efficiency was re-assessed by FACS (Fluorescence Activated Cell Sorting) analysis using a 3' fluorescein (FITC)-labelled GATA1 siRNA in K562 cells (details shown in Chapter 2, section 2.12). Over 90% of transfection efficiency (91%) was achieved which was in agreement with the previous result (93%).

6.4 Time-course analysis of knockdown with GATA1

Time-course analysis of GATA1 siRNA knockdown was conducted to determine the optimum time-point for maximum GATA1 depletion in K562 cells. Knockdown of GATA1 was monitored at the following time-points: 12 hr, 24 hr, 36 hr, 48 hr and 96 hour. In particular, a second transfection was performed after 48 hour for the 96 hour time-point, as it was previously demonstrated that re-transfection for a second time offered an additional degree of knockdown at both mRNA and protein levels (Bruce et al., 2009). The following criteria were used to determine the optimal time-point:

1. Maximum knockdown of GATA1 for ~90% at the mRNA and protein levels (this level was achievable with this GATA1 siRNA based on previous results, H.L. Jim's PhD thesis). The earliest time point(s) fulfilled this criteria should be chosen, where the additional three criteria were applied as follows.

2. The time-point at which the TAL1 expression being affected.

3. The time-point at which GATA1 occupancy at either the +51 enhancer or the P^{TAL1} being reduced.

4. The time-point at which the looping interaction between the P^{TAL1} and the +51 enhancer being affected.

6.4.1 *Monitoring the mRNA and protein level of GATA1 during the time-course study*

After being transfected with GATA1 and luciferase siRNAs, cells were subsequently harvested for mRNA and nuclear protein at 12, 24, 36, 48 and 96-hour time-points. The quantitative reverse-transcribe PCR (qRT-PCR) and Western blotting assays were performed to assess the efficiency of GATA1 siRNA knockdown relative to cells transfected with the luciferase siRNA.

To determine mRNA expression of GATA1, three housekeeping genes (β-actin, β-tubulin and GAPDH), which are highly and constitutively expressed in most tissues or cell lines, were used as internal controls to accurately normalise the sample-to-sample variations in mRNA levels. Comparing to luciferase siRNA control, decreased expression of GATA1 at the mRNA level were observed across all five time-points after transfection with GATA1 siRNA (Figure 6.5a). The highest knockdown of GATA1 mRNA was achieved at the time-point of 96 hour (96.7%), followed by 48 hour (92.9%), 24 hour (92.8%), 12 hour (91.1%), and 36 hour (89.3%) time-points. In addition, knockdown efficiency of GATA1 siRNA was assessed at protein level by western blotting analysis (Figure 6.5b & c). The remaining GATA1 protein was quantified and normalised relative to the loading control (Bradford stained) and knockdown efficiency of GATA1 was calculated based on the GATA1 level in the luciferase control. The 96-hour time-point again showed the highest knockdown of GATA1 at protein level (96.3%), followed by 48

hour (88.1%), 24 hour (82.0%), 36 hour (80.6%) and 12 hour (56.0%) time points. Taken together, the earliest time-points that fulfilled the first criterion were 48 hour and 96 hour. These two time-points were then taken into the subsequent analyses to further assess the optimal time-point for GATA1 knockdown.

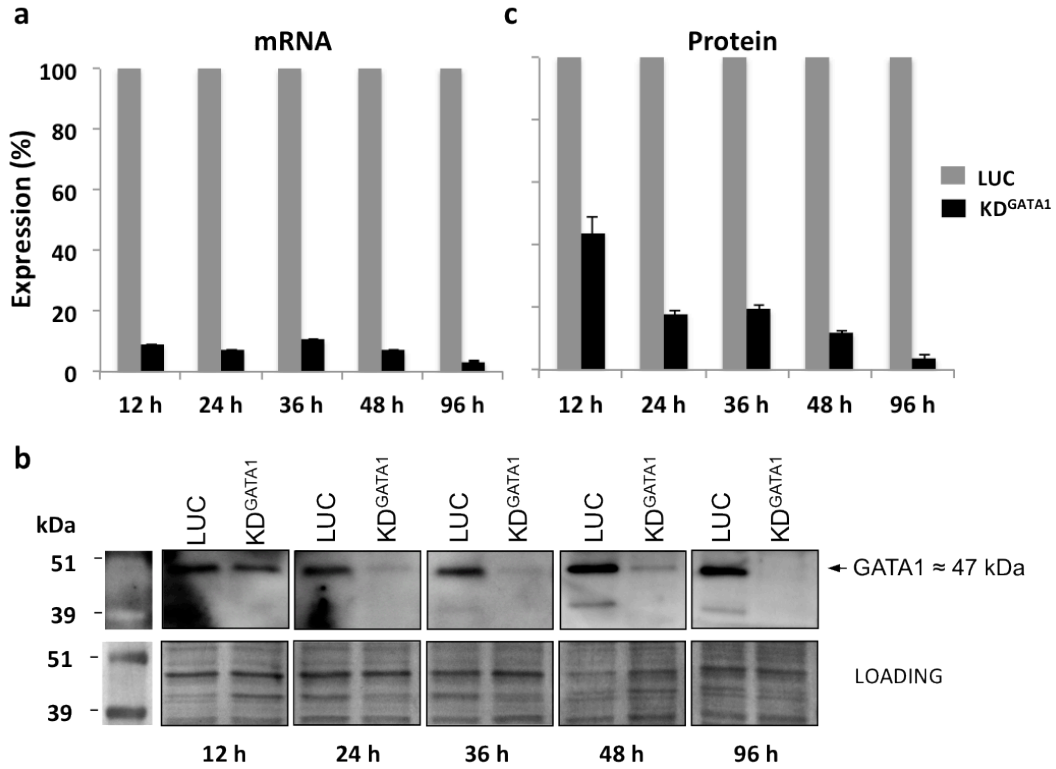


Figure 6.5: siRNA knockdown time-course study of GATA1. A: Knockdown of GATA1 at the mRNA level was quantified by quantitative PCR as described in the text. Histogram shows the mRNA level of GATA1 remaining (y-axis) after siRNA transfection relative to luciferase siRNA transfection across the time points (x-axis). B: Western blotting analyses of GATA1 protein knockdowns. Upper panel shows the bands detected by the GATA1 antibody. The lower panel shows the blot stained with Bradford reagent as loading control. Protein size-makers (kDa) are shown at the left side of the panels. The arrow shows the predicted size of the GATA1 protein. C: Knockdown of GATA1 at the protein level was quantified by western blotting bands and normalised against the loading control. Histogram shows the protein level of GATA1 remaining (y-axis) after siRNA transfection relative to luciferase siRNA transfection across the time points (x-axis). The error bars show the standard error of the mean between the two independent biological replicates.

6.4.2 *Monitoring the expression of TAL1 at the 48 and 96 hour time-points*

As the TEC was previously found to be associated with TAL1 expression in the erythroid lineages (H.L. Jim's PhD thesis), the expression of TAL1 was monitored for cells knocked-down by the GATA1 siRNA after 48 and 96 hour. The histogram in Figure 6.6 illustrates the TAL1 expression at mRNA level in GATA1 knockdown K562 relative to the luciferase controls. The percentage of remaining TAL1 mRNA was down to ~60% at 48 hour time-point and further decreased to ~50% at 96 hour time-point (Figure 6.6).

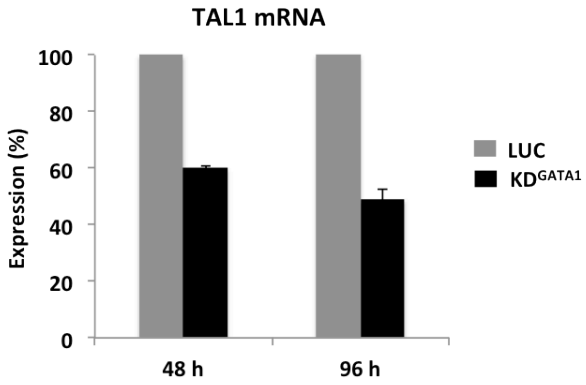


Figure 6.6: RT-qPCR analysis of mRNA expression of TAL1 in K562 cells 48 and 96 hours after siRNA transfection with GATA1 (KD) or with luciferase (LUC). Expression of TAL1 in KD^{GATA1} is shown relative to LUC control. Histogram shows the percentage of mRNA remaining (y-axis) after GATA1 siRNA knockdown the time points (x-axis). The error bars show the standard error of the mean between the two biological replicates.

It had been reported that depletion of GATA1 after 48 hour of siRNA knockdown resulted in the reduction of TAL1 expression to 63% (H.L. Jim's PhD thesis). The observation was in agreement with the previous studies, suggesting that the TAL1 expression was significantly affected by GATA1 knockdown at both 48 and 96-hour time-points. Therefore, the second criterion was fulfilled by GATA1 knockdown at 48 and 96 hour.

6.4.3 *Monitoring the GATA1 ChIP occupancy at 48 and 96 hour*

ChIP occupancy of GATA1 was previously detected at the P^{TAL1} and the +51 enhancer in erythroid K562 cells (H.L. Jim's PhD thesis and Dhami et al., 2010). In order to further characterize knockdown of GATA1 at the level of occupancy over its binding sites, ChIP-qPCR analysis was performed using K562 cells harvested 48 and 96 hour after siRNA transfection.

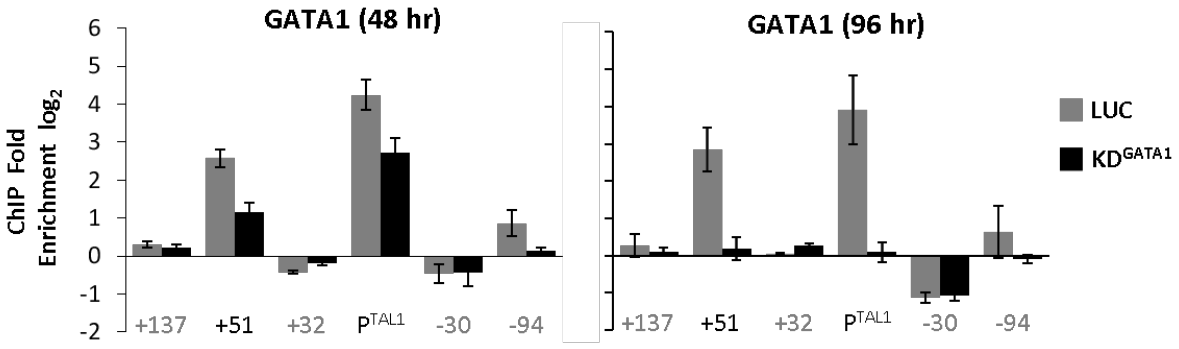


Figure 6.7: ChIP analysis of GATA1 occupancy at the +51 erythroid enhancer and TAL1 promoter 1a (P^{TAL1}) in K562 cells 48 and 96 hours after siRNA transfection with GATA1 (KD) or with luciferase (LUC). ChIP enrichments (log_2) are shown with standard errors. Annotation of test and negative control regions is denoted in black and grey text respectively. Luciferase treated samples are shown as grey bars, KD^{GATA1} treated samples are shown as black bars.

Knockdown efficiency (KD%) was calculated by comparing ChIP enrichments of GATA1 KD against luciferase control and statistical significance (p-value) was determined by t-test as illustrated in Table 6.2. As shown in Figure 6.7, significant ChIP enrichments of GATA1 were detected at the +51 enhancer and the P^{TAL1} in K562 cells transfected with luciferase siRNA at 48 and 96-hour time-points. For GATA1 knockdown after 48 hours, ~65 % of GATA1 occupancy was depleted from the TAL1 promoter and the +51 enhancer (Table 6.3) in comparison with the luciferase control (Figure 6.7, left panel). Based on the reasons stated previously, the time-course studies were extended to 96 hour with a second electroporation conducted at the 48 hr time point. As shown in Figure 6.7 (right panel) and Table 6.2, over 90 % of the GATA1 protein was depleted from the TAL1 promoter and 85% of GATA1 was removed from the +51 enhancer in comparison with the luciferase control after knockdown with GATA1 siRNA at 96-hour time-point. Regarding to the third criterion, GATA1 occupancy at the TAL1 promoter and the +51 enhancer was affected at both 48 and 96-hour time-points, and the highest knockdown efficiency of GATA1 occupancy was observed at the 96-hour time-point.

Table 6.3 Percentages of GATA1 protein lost from the binding target sites at 48 and 96 hour after GATA1 knockdown.

	KD^{GATA1} 48 hr		KD^{GATA1} 96 hr	
	KD[1] %	*p-value*[2]	KD[1] %	*p-value*[2]
P^{TAL1}	65.4 %	1.4×10^{-3}	94.0 %	3.3×10^{-5}
+51	62.6 %	1.6×10^{-2}	85.0 %	7.4×10^{-5}

[1]Percentage of depletion of GATA1 occupancy is calculated based on GATA1 ChIP enrichment of GATA1 siRNA knockdown samples versus Luciferase siRNA controls.
[2]P-value is calculated based on student T-test (2-tails)

6.4.4 *Monitoring the looping interaction between the P^{TAL1} and the +51 enhancer at the 48 and 96 hour time-points*

Providing the facts that i) the TEC (including GATA1) binds at the P^{TAL1} and the +51 enhancer in the erythroid K562 cells and ii) the P^{TAL1}/+51 interaction is a known major interaction of "cruciform" structure in K562 cells (see Chapter 3 & 5), this particular looping interaction was monitored, in order to further assess knockdown efficiency of GATA1 at the 48 and 96 hour time-points. 3C-PCR was performed to determine the P^{TAL1}/+51 interaction in the GATA1 knockdown and the luciferase control K562 cells using the TAL1 promoter 1b as the 3C "anchor".

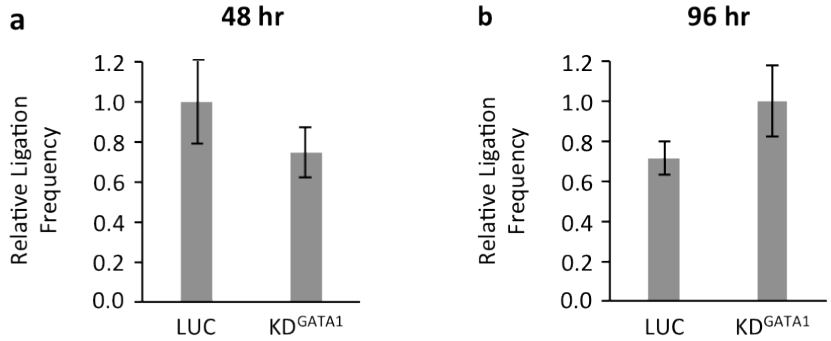


Figure 6.8: Assessing the quality of 3C libraries. Bar diagrams show the interaction frequencies between two non-adjacent Csp6I fragments at the ERCC3 locus determined by 3C. (A) K562 cells 48 hrs after transfection with siRNA for luciferase (LUC) or GATA1 (B) K562 cells 96 hour after transfection with siRNA for luciferase (LUC) or GATA1 (KD^{GATA1}). Interaction frequencies (grey bars) are shown with standard errors and represent the mean from two bio-replicate samples.

The quality of the 3C libraries prepared from the GATA1 knockdown and luciferase control K562 cells was assessed as previously described in Chapter 3. As shown in Figure 6.8 a, the interaction frequency between the ERCC3 fragments in GATA1 knockdown (KD^{GATA1}) sample was lower than in luciferase control (LUC) at the 48 hour time-point, suggesting that the reduced level of interaction frequency in KD^{GATA1} corresponding to LUC could be due to the differences of library quality between two group of samples. On the contrary, the 3C library quality of KD^{GATA1} was better than LUC for the 96 hour time-point, which was determined based on ERCC3 interaction frequencies as shown in Figure 6.8 b, implying that the interaction frequency in LUC would be relatively higher if the library quality was taken into account.

At the 48 hour time-point, significant interactions were observed between the +51 enhancer and the TAL1 promoter in both luciferase and GATA1 siRNA transfected K562 cells (shown in Figure 6.9 a), in agreement with the observations captured by 3C-PCR and 4C-array in the wild type K562 cell. At the 96-hour time-point (Figure 6.9 b), a significant interaction at the +51 enhancer was observed only in the luciferase control K562 cells (student's t-test, $p = 2.1 \times 10^{-9}$). In contrast, a complete loss of looping interaction was found at the TAL1 locus in K562 cells after 96 hour of GATA1 knockdown (Figure 6.9 b), suggesting that the looping interaction between the P^{TAL1} and the +51 enhancer in K562 was not abolished until GATA1 knockdown at 96-hour time-point. Thus, the forth criterion was only fulfilled by the 96 hour time-point.

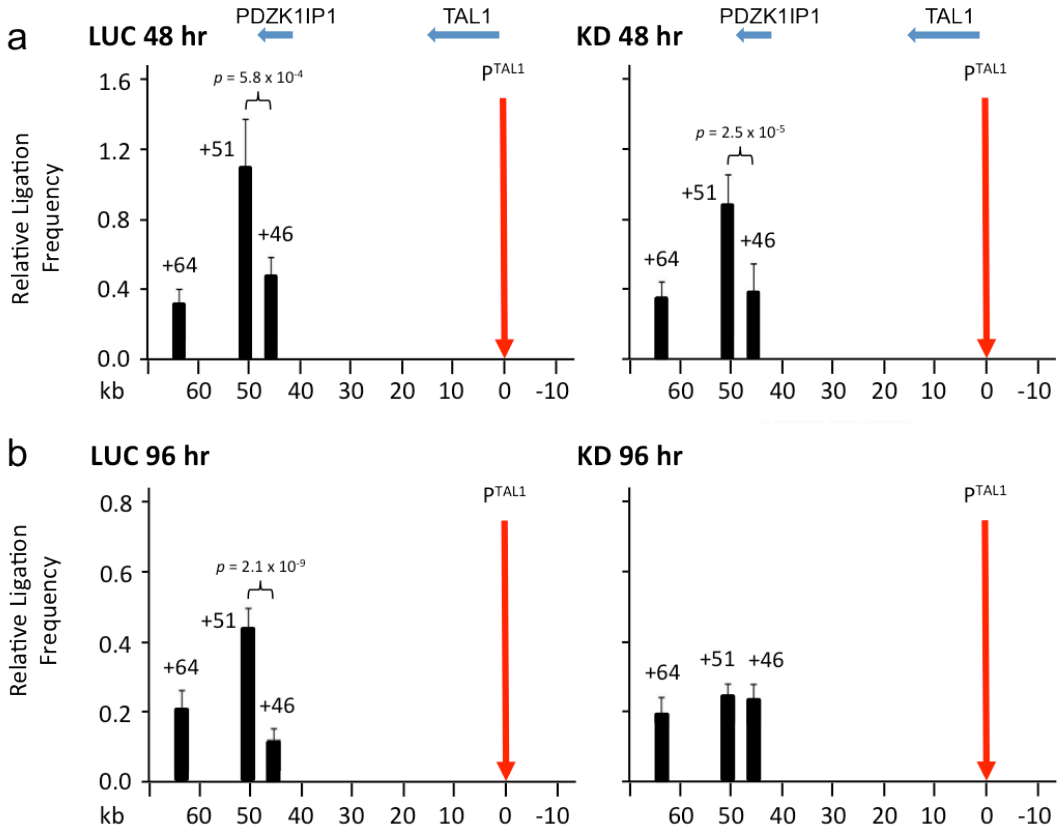


Figure 6.9: Looping interaction between the TAL1 erythroid enhancer (+51) and TAL1 promoter (P^{TAL1}) 48 and 96 hours after transfection with siRNA for GATA1 (KD^{GATA1}) and luciferase (LUC). Histograms show looping interaction between the P^{TAL1} and +51 are determined by 3C. (A) Interaction between the TAL1 promoter 1b (P^{TAL}) and the erythroid enhancer 48 hours after siRNA transfection. (B) Interaction between the TAL1 promoter 1b (P^{TAL}) and the erythroid enhancer 96 hours after siRNA transfection. Interaction frequencies (black bars) are shown with standard errors and normalized relative to BAC controls. Locations of 3C "anchor" regions are denoted with vertical red arrows. Locations of genes and their directions of transcription are shown at the top of each panel. *p* values are indicated for interaction frequencies which are significantly higher for test regions when compared to those of control regions. Scales (in kb) are shown at the bottom of each panel.

Taken together, the 96-hour time-point for GATA1 knockdown was the optimized time-point that fulfilled all four criteria. GATA1 knockdown at the 96 hour resulted in not only the reduced level of TAL1 expression but also the significant depletion of GATA1 occupancy at the P^{TAL1} and the +51 enhancer as well as the loss of looping interaction between these two particular binding sites. Taken together, it implied that GATA1 involved in regulating expression of TAL1, most likely via GATA1-mediated promoter-enhancer interaction. Further analyses including RT-qPCR, ChIP and 3C were conducted in order to comprehensively characterize the downstream affects of GATA1 knockdown at 96 hour on transcriptional regulation and chromatin structure at the TAL1 locus.

6.5 GATA1 knockdown at the 96 hour time-point

As in the 3C and 4C data illustrated in previous chapters, it has been proposed that a cruciform configuration may be adapted at the TAL1 locus in erythroid K562 cells. In addition, it was also speculated that TEC might be involve in mediating looping interactions between TAL1 promoter and its enhancers and subsequently facilitating expression at the TAL1 locus (see Chapter 5). Providing the fact that all four criteria were achieved in at the 96-hour time-point of GATA1 knockdown, this time-point was selected to assess how GATA1 and the TEC could be involved in transcriptional regulation of the TAL locus. It was assessed at five different levels as listed in section 6.3, including i) transcription of TAL1 and its neighbouring genes, ii) RNAP II occupancy at the promoters and enhancers at the TAL1 locus, iii) occupancy of other TEC members at the P^{TAL1} and the +51 enhancer, iv) occupancy of CTCF/Rad21 at the TAL1 locus and chromatin interactions between the CTSs, v) the foundation of "cruciform" structure - looping interactions between the P^{TAL1} and its enhancers.

6.5.1 *Depletion of GATA1 affects the transcription of TAL1 and its neighbouring genes*

Expression of two groups of genes after 96-hour knockdown of GATA1 was assessed. In the first group there were two key components of the TEC, the bridging protein LDB1 and the DNA binding protein TCF3, as these were known target genes of GATA1 previously reported (H.L. Jim's PhD thesis, 2008, University of Cambridge). In the second group there were TAL1 and its neighbouring genes including PDZK1IP1 (downstream of TAL1), STIL and CMPK1 (upstream of TAL1). In particular, it had previously been shown that the PDZK1IP1 and STIL genes co-expressed with TAL1 in K562 cells (Delabesse et al., 2005). Additionally, the 4C-array profiles of K562 (see Chapter 5) also suggested that the promoters of PDZK1IP1 and STIL might be in the spatial close-proximity, implying that their co-expression might be regulated through the same mechanism.

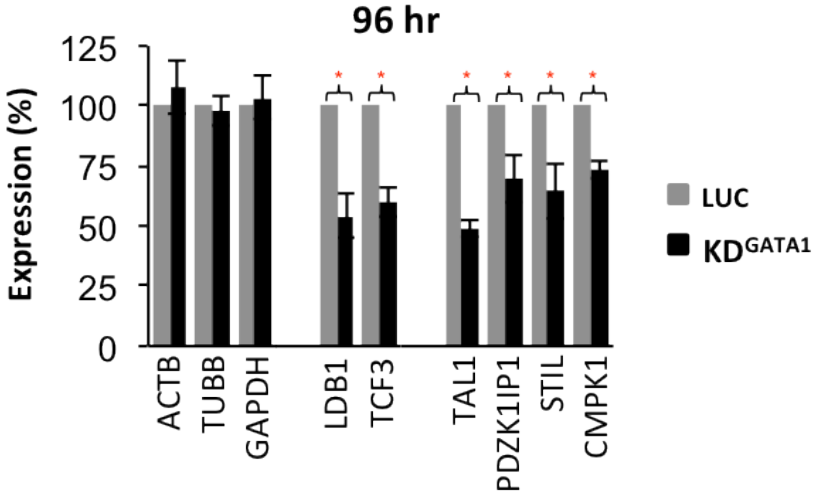


Figure 6.10: RT-qPCR analysis of mRNA expression of TEC members, TAL1 and its neighbouring genes in K562 cells 96 hours after siRNA transfection with GATA1 (KD) or with luciferase (LUC). ACTB, TUBB and GAPDH were used as gene expression controls. Expression of LDB1, E2A/TCF3, TAL1, PDZK1IPI, STIL and CMPK1 (with standard errors) in KD^{GATA1} are shown relative to levels in LUC control. Red asterisks indicate statistically significant ($p < 0.01$) differences of expression between KD^{GATA1} and LUC. LDB1 and TCF3 levels were also measured as which are known genes targets of GATA1 (Hau Ling Jim, 2008, PhD thesis).

As mentioned in section 6.4.1, the three housekeeping genes (β-actin, β-tubulin and GAPDH) were used as the internal controls to normalise the expression data. The t-test was performed to determine the statistical significance (red asterisk) of the expression levels between GATA1 knockdown and luciferase control. Firstly, no significant difference was observed between GATA1 knockdown and luciferase control samples for three housekeeping genes (Figure 6.10). Secondly, expression of LDB1 and E2A/TCF3 was significantly decreased after GATA1 knockdown for 96 hour, which were down to 53.9% and 60.0% respectively (Figure 6.10). A similar level of LDB1 reduction (down to 52%) was previously reported in the GATA1 knockdown K562 cells (H.L. Jim's PhD thesis, 2008). It suggested that the depletion of GATA1 resulted in down-regulation of other members of the TEC, which might even consequently lead to the loss of the TEC at its the binding sites. Thus, subsequent ChIP-qPCR analysis was performed to assess the loss of the TEC (LDB1 and TCF3) at the P^{TAL1} and the +51 enhancer, which will be discussed in the following section. Thirdly, significant reduction of expression was also observed for TAL1 along with its neighboring genes. For TAL1, its expression was declined to less than 50% (49.1%) after GATA1 knockdown in comparison with the luciferase control. Similarly, the transcript levels of the PDZK1IP1, STIL and CMPK1 genes were also down to 69.7%, 64.6% and 73.4% in comparison with their normal levels respectively (Figure 6.10). Taken together, it suggested that the transcriptional regulation of TAL1 and its neighbouring genes was GATA1 and/or

the TEC dependent, which might be via the “cruciform” configuration formed in K562 cells. However, no interaction was observed between the promoters of TAL1 and CMPK1 genes. Additionally, it revealed relatively less dependence of GATA1/TEC for the CMPK1 expression compared to the TAL1 expression (73.4% vs. 49.1%). Nevertheless, the mechanism of the involvement of GATA1 and/or the TEC in regulating expression of CMPK1 remained unclear.

6.5.2 *Depletion of GATA1 affecting recruitment of RNA polymerase II over promoters and enhancers*

As TAL1 and STIL are co-expressed in almost all hematopoietic lineages including K562 (Delabesse et al., 2005), recruitments of RNA polymerase II (RNAP II) at the promoters via enhancers are required for these genes to be transcribed. It has been demonstrated that the RNAP II bind at the promoters of TAL1, STIL and CMPK1 as well as the +51 and +20/+19 enhancers (Dhami et al., 2010). Additionally, it has been illustrated that the TAL1 promoter 1b, the +51 and +20/+19 enhancers as well as the STIL intron1 (1 kb downstream of promoter) are co-localised based on the 4C-array analysis. Thus, it is speculated that RNAP II can be loaded to these genes at the same time when they are in spatial proximity by forming the “cruciform” configuration. Providing the facts that expression of TAL1 and its neighbouring genes was affected by GATA1 knockdown, it was interesting to determine how occupancy of RNAP II was affected by depletion of GATA1.

ChIP-PCR was performed to determine the RNAP II occupancy at the +51 and +19/20 enhancers, TAL1 promoter 1a, along with promoters of STIL and CMPK1 (as shown in Figure 6.11). For the luciferase control, high-level ChIP enrichments (approx. 3.5-folds, $\log_2$ scaled) of RNAP II were observed at the +51 enhancer, promoters of TAL1, STIL and CMPK1, while relatively lower RNAP II occupancy (~1.5-folds, $\log_2$ scaled) was also observed at the +20 enhancer (Figure 6.11). The level of RNAP II enrichments was in agreement with the RNAP II ChIP-chip profile in wild-type K562 cells (Dhami et al., 2010). In contrast, significant reductions of RNAP II occupancy were observed over the enhancers (+51 and +20) as well as promoters of TAL1 (promoter 1a) and STIL in GATA1 knockdown K562 cells (Figure 6.11).

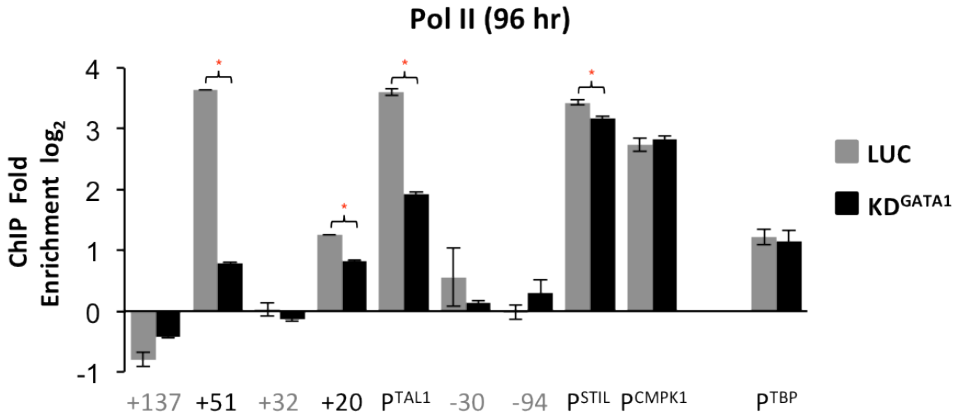


Figure 6.11: ChIP analysis of RNA polymerase II (RNAP II) occupancy of the +51 and +20 enhancers, the TAL1 promoter 1a (P^{TAL1}) and promoters of STIL and CMPK1 in K562 cells 96 hours after siRNA transfection with GATA1 (KD) or with luciferase (LUC). Positive control for RNAP II was the promoter of the TBP gene (P^{TBP}). ChIP enrichments (log_2) are shown with standard errors. Annotation of test and negative control regions is denoted in black and grey text respectively. Luciferase treated samples are shown as grey bars, KD^{GATA1} treated samples are shown as black bars. T-test was performed to determine the statistical significance (indicated by red asterisks, p values shown in Table 6.3) of reduction of RNAP II after GATA1 knockdown.

RNAP II occupancy at the +51 enhancer and the promoter 1a were down-regulated 86.2% (p = 4.9 x 10^{-6}) and 68.8% (p = 7.2 x 10^{-4}) accordingly after GATA1 knockdown for 96 hour (Table 6.4), suggesting that GATA1 and TEC might be important for RNAP II recruitment at the TAL1 promoter and its erythroid enhancer. Additionally, as shown in Table 6.4, minor depletions of RNAP II were observed at the +20 enhancer (25.8%, p = 9.0 x 10^{-4}) and STIL promoter (17.3%, p = 2.7 x 10^{-4}), which implied a much less dependence on GATA1 and TEC for the RNAP II recruitment at these two sites. This is because that the +20/+19 stem cell enhancer is not required for the expression of TAL1 in erythroid lineage (Sanchez et al., 1999; Sanchez et al., 2001). The observation of low-level of RNAP II occupancy at the +20 enhancer could probably be due to its spatial close-proximity to the TAL1 promoter within the “cruciform” configuration. In addition, although the expression of STIL in GATA1 knockdown K562 was down regulated over 35% relative to luciferase control, less than 20% of the RNAP II was depleted from the STIL promoter, implying that RNAP II might be also recruited at the STIL promoter in a GATA1/TEC-independent manner in a certain proportion of cells. Interestingly, no significant reduction of RNAP II occupancy was observed at the CMPK1 promoter between the GATA1 knockdown and the luciferase control (Figure 6.10), regardless of the fact that over 25% reduction of CMPK1 expression was observed after GATA1 knockdown. However, it is unclear how the expression of CMPK1 was decreased without affecting RNAP II occupancy at its promoter.

Table 6.4 Percentages of RNA polymerase II protein lost from its binding sites at 96 hour after GATA1 knockdown.

	KD[1] %	*p-value*[2]
+51	86.2 %	4.9×10^{-6}
+20	25.8 %	9.0×10^{-4}
P^{TAL1}	68.8 %	7.2×10^{-4}
P^{STIL}	17.3 %	2.7×10^{-4}

[1]Percentage of depletion of RNA polymerase II occupancy is calculated based on RNA pol II ChIP enrichment of GATA1 siRNA knockdown samples versus luciferase siRNA controls.
[2]P-value is calculated based on student T-test (2-tails)

Providing two facts that the PDZK1IP1 promoter was also situated in the "cruciform" configuration and the expression of PDZK1IP1 was affected by GATA1 knockdown, it would also be interesting to assess the RNAP II occupancy over the PDZK1IP1 promoter in GATA1 knockdown K562 cells. However, previous ChIP-chip studies had been shown that no detectable enrichment of RNAP II was observed at the PDZK1IP1 in K562 cells (Dhami et al., 2010). Therefore, it is not possible to assess whether RNAP II occupancy at the PDZK1IP1 promoter was also affected along with its gene expression by GATA1 knockdown.

6.5.3 *Depletion of GATA1 by siRNA knockdown results in loss of occupancy of other members of the TAL1 erythroid complex (TEC)*

The occupancy of other members of the TAL1 erythroid complex (TEC) was also examined as a result of loss of GATA1 during siRNA knockdown in K562 cells. ChIP-qPCR analysis was performed for both LDB1 and E2A/TCF3 (E47 isoform) at both the TAL1 promoter 1a and the +51 enhancer. ChIP enrichments (log_2) of LDB1 and E2A/TCF3 are shown in Figure 6.12.

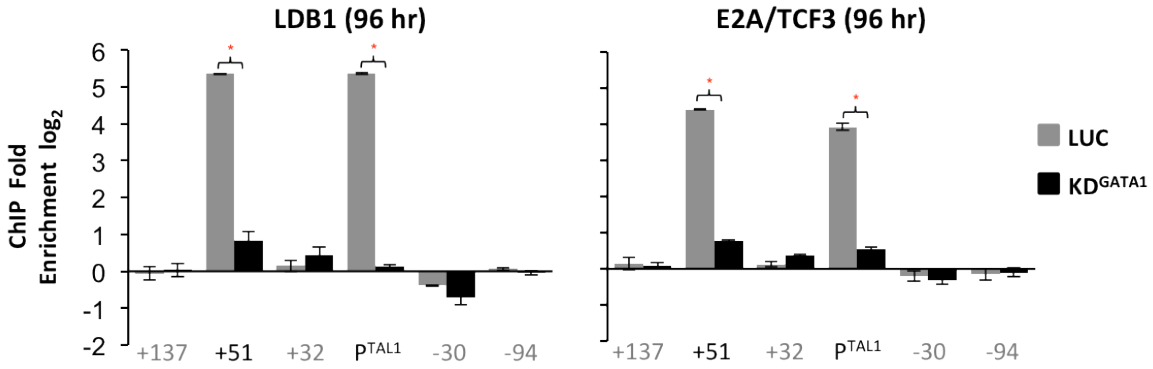


Figure 6.12: ChIP analysis of LDB1 and E2A/TCF3 occupancy of the +51 erythroid enhancer and the TAL1 promoter 1a (P^{TAL1}) in K562 cells 96 hours after siRNA transfection with GATA1 (KD) or with luciferase (LUC). ChIP enrichments (log_2) are shown with standard errors. Annotation of test and negative control regions is denoted in black and grey text respectively. Luciferase treated samples are shown as grey bars, KD^{GATA1} treated samples are shown as black bars.

As expected, significant ChIP enrichments of LDB1 and E2A/TCF3 were observed at the +51 region and at promoter 1a in luciferase control K562 cells (Figure 6.12). In contrast, significant reductions of LDB1 and E2A/TCF3 occupancy were observed at the TAL1 promoter 1a and its +51 enhancer after GATA1 knockdown (Figure 6.12). Over 95% of LDB1 and over 85% of E2A/TCF3 were depleted from their binding sites at the promoter 1a and the +51 enhancer due to GATA1 knockdown (Table 6.5).

Table 6.5 Percentages of LDB1 and TCF3/E47 proteins lost from the binding target sites at 96 hour after GATA1 knockdown.

	LDB1 KD^{GATA1} 96 hr		E2A/TCF3 KD^{GATA1} 96 hr	
	KD[1] %	*p-value*[2]	KD[1] %	*p-value*[2]
P^{TAL1}	97.6 %	4.2×10^{-6}	89.0 %	1.8×10^{-4}
+51	96.0 %	1.8×10^{-6}	86.6 %	5.3×10^{-6}

[1]**Percentage of depletion of LDB1 or TCF3/47 occupancy is calculated based on ChIP enrichments of GATA1 siRNA knockdown samples versus luciferase siRNA controls.**
[2]***p*-value is calculated based of student T-test (2-tails)**

As reported in the previous section, GATA1 knockdown resulted in 46% and 40% reductions of LDB1 and TCF3 expression. However, ChIP occupancy of LDB1 and E2A/TCF3 occupancy were depleted over 96% and 86% respectively from their binding sites. Therefore, changes in LDB1 and E2A/TCF3 expression alone do not adequately account for the loss of TF binding at the +51 enhancer and the TAL1 promoter 1a (although analysis at the protein level would have helped resolve this issue). Nonetheless, loss of three members of the TEC, either directly through the absence of GATA1, or indirectly through down-regulation of other members of the complex due to GATA1, resulted in substantial loss of the TEC at the TAL1 locus.

6.5.4 *Depletion of GATA1 affects the long-range looping interactions of the TAL1 locus*

It had previously been shown that the GATA1 and TEC occupancy was greatly depleted from both the +51 erythroid enhancer and from promoter 1a as a result of GATA1 siRNA knockdown, and that this loss was also associated with loss of chromatin looping between the TAL1 promoters and its erythroid enhancer (see section 6.4.4). Based on the cruciform model that had been proposed to account for all the looping interactions observed at the TAL1 locus, it would also be expected that loss of the TEC during GATA1 knockdown would affect the entire cruciform looping structure. Therefore, additional 3C analysis was performed to determine to what extend the looping structure between other *cis*-regulatory

elements was affected by GATA1 knockdown. In addition to the +51 enhancer, the +20/+19 stem cell enhancer was another major interacting partner of the TAL1 promoter (P^{TAL1}) in wild-type K562 cells determined by both 3C and 4C (see Chapters 3 and 5). The three-way interaction between these regulatory elements (+51, +20/+19 and P^{TAL1}) formed the fundamental structure of the "cruciform" configuration in K562 cells.

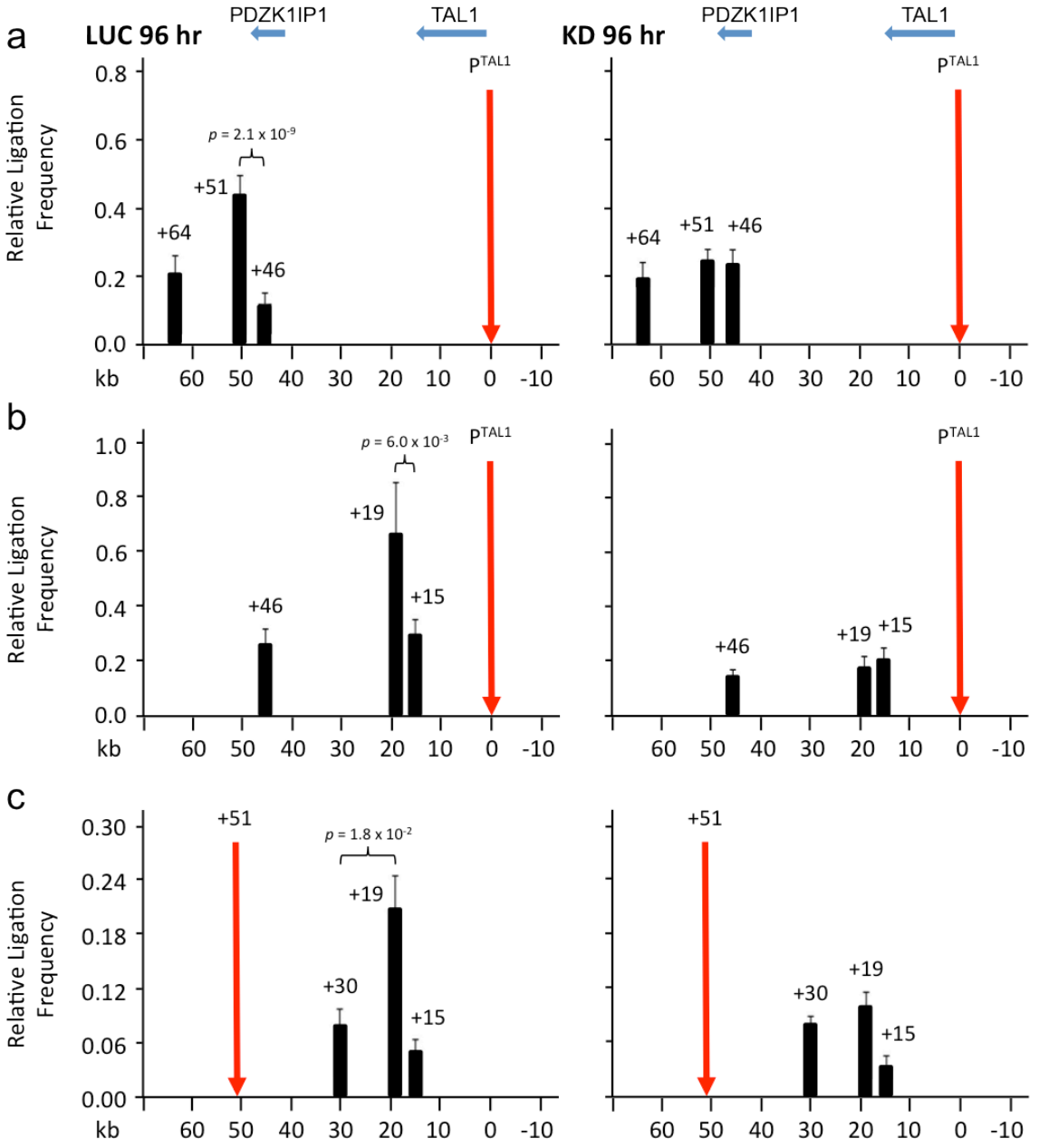


Figure 6.13: GATA1 knockdown results in the loss of looping interactions between the P^{TAL1} and its enhancers at +51 and +19. Histograms of interaction patterns across the human TAL1 locus after siRNA transfection with GATA1 (KD) or with luciferase (LUC) were determined by 3C. Panel A and B: Interactions between the TAL1 promoter 1b (P^{TAL}) and the +51 erythroid enhancer and the +19/+20 stem cell enhancer 96 hours after siRNA transfection. Panel C: Interaction between the erythroid enhancer and the stem cell enhancer 96 hours after siRNA transfection. Interaction frequencies (black bars) are shown with standard errors and normalized relative to BAC controls. Locations of the 3C "anchor" regions are denoted with vertical red arrows. Locations of genes and their directions of transcription are shown at the top of each panel. *p* values are indicated for interaction

frequencies which are significantly higher for test regions when compared to those of control regions. Scales (in kb) are shown at the bottom of each panel.

3C-PCR analyses were performed to assess the three-way looping interaction combinations in both GATA1 knockdown and luciferase control K562 cells. As expected, significant interaction frequencies were observed between the TAL1 promoter 1b and the +51 and +20/+19 enhancers in luciferase control (Figure 6.13a & b, left panels), in agreement with the 3C and 4C profiles of wild-type K562 presented in previous chapters. In contrast, the interaction frequencies at the +51 and +20/+19 enhancers were dramatically reduced in GATA1 knockdown K562 cells, and no statistically significant difference was observed with respect to the interaction frequency at the control region (Figure 6.13a & b, right panels). In addition, significant interaction frequency was also observed between the +51 and +20/+19 enhancers in luciferase control as illustrated in Figure 6.13c (left panel). However, this looping interaction was also disrupted by GATA1 knockdown in K562 cells (Figure 6.13c, right panel).

As being summarised in Figure 6.14, disruptions of the three-way interactions were observed between the TAL1 promoter 1b, the +51 and +20/+19 enhancers, implying that disassociation of cruciform configuration was in progress due to GATA1 knockdown in K562 cells. Taken together, one can conclude based on these observations that (i) the TAL1 promoter-enhancer interactions do existed within the same proportion of cells at the same time, and (ii) at least a major part of the "cruciform" configuration is GATA1/TEC-dependent.

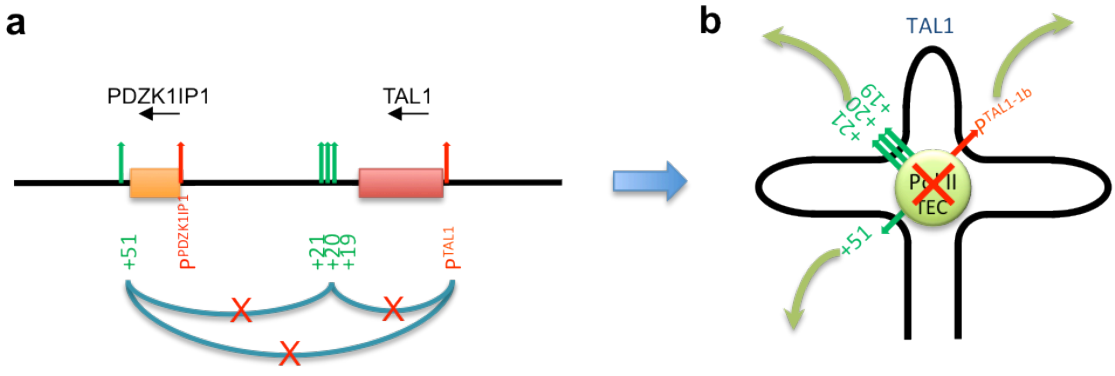


Figure 6.14: Schematics diagrams of the TAL1 promoter-enhancer interactions disassociation after GATA1 knockdown. Panel A: a genomic map illustrates locations of three elements across the TAL1 locus and the loss of three-way interactions. Panel B: a predicted model shows the disengagement of three-way looping interactions results in disassociation of the "cruciform" configuration when the TEC and RNAP II were depleted due to GATA1 knockdown.

6.5.5 *Depletion of GATA1 affecting CTCF and Rad21 occupancy over CTS at -31*

The data presented in the previous sections of this Chapter illustrated that depletion of GATA1 resulted in loss of TAL1 expression, the loss of the TEC occupancy as well as the loss of looping interactions between the TAL1 promoters and its enhancers. It provided further evidence in supporting the "cruciform" model where all these looping interactions detected by 3C and 4C-array (data shown in Chapter 3 & 4) was occurring in the same K562 cells at the same time. However, whilst looping interactions also seemed to occur at the +57/+53 and -31 insulators - sites of CTCF and Rad21 binding (see Chapter 5) - it was not known whether these looping interactions also occurred in the same cells as those mediated by the TAL1 promoter and its enhancers, and whether the recruitment of CTCF and Rad21 was also dependent on GATA1 and/or the TEC. Thus, it was important to determine whether disruption of GATA1 could affect recruitments of CTCF and Rad21 (cohesin subunit) proteins at the TAL1 locus, and whether this resulted in the loss of looping interactions between the insulator elements at +57/+53 and -31. Therefore, further analyses were conducted to assess whether GATA1 knockdown affected the bindings of CTCF and Rad21 and the looping interactions between the CTSs in K562 cells.

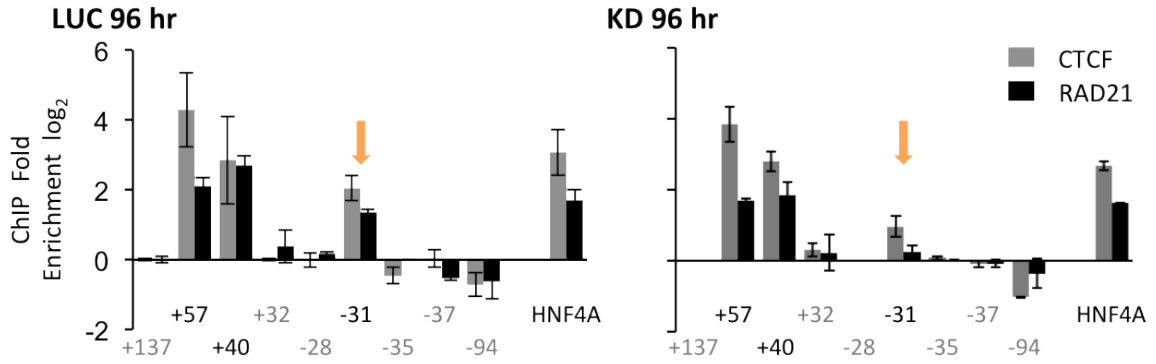


Figure 6.15: ChIP analysis of CTCF and Rad21 occupancy of CTSs at +57, +40 and -31 in K562 cells 96 hours after siRNA transfection with GATA1 (KD) or with luciferase (LUC). ChIP enrichments (log_2) are shown with standard errors. Orange arrows indicate reduction in CTCF and Rad21 occupancy at -31 in GATA1 knockdown comparing to luciferase control. Annotation of test and negative control regions is denoted in black and grey text respectively. A known site of CTCF and Rad21 binding at the HNF4A locus was used as a positive control.

The ChIP analysis of CTCF and Rad21 occupancy was presented for only three CTSs at +57, +40 and -31, as CTS at +53 showed no significant enrichment above the control regions (background). This was because that ChIP-enrichments of CTCF and Rad21 were much lower at +53 in the first place, comparing to other

three CTSs in wild-type K562 (see Chapter 5). In addition, the overall ChIP efficiency was reduced due to the limited number of cells available after siRNA transfection for GATA1 knockdown and luciferase control samples. Thus, the assessments in this section were focused on the CTSs at +57, +40 and -31.

CTCF and Rad21 occupancy at the insulators/CTCs (+57, +40 and -31) at the TAL1 locus were tested by ChIP-qPCR in the GATA1 knockdown K562 cells and their levels compared to that of the luciferase control. Occupancy patterns in the luciferase control mirrored those seen in wild-type K562 cells (see Chapter 4). No significant differences of CTCF and Rad21 ChIP enrichments were observed at +57, +40 and insulator elements of HNF4A (positive control) when comparing GATA1 knockdown cells with respect to the control (Figure 6.15). However, substantial reductions (>50% loss) of both CTCF and Rad21 occupancy were observed at -31 in GATA1 knockdown K562 cells, suggesting that the depletion of GATA1/TEC substantially affected the occupancy of CTCF and Rad21 at the TAL1 locus.

6.5.6 *Depletion of GATA1 affecting looping interaction between CTSs*

As the CTCF and Rad21 occupancy at -31 were affected as the result of GATA1 knockdown, it was important to assess whether the looping interaction between the +57 and -31 elements were also affected as a result of loss of CTCF and Rad21. The 3C analysis was performed on both GATA1 knockdown and luciferase control K562 cells, using +57 element as the anchor.

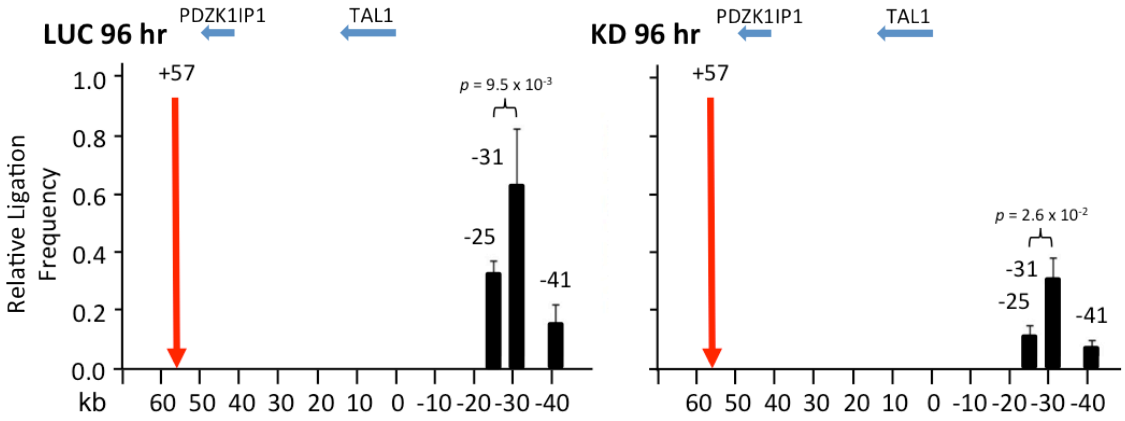


Figure 6.16: GATA1 knockdown results in the loss of looping interactions between CTCF/Rad21 bound insulators. Histograms of interaction between +57 and -31 insulators at the human TAL1 locus after siRNA transfection with GATA1 (KD) or with luciferase (LUC) were determined by 3C. Interaction frequencies (black bars) are shown with standard errors and normalized relative to BAC controls. Locations of 3C "anchor" regions are denoted with vertical red arrows. Locations of genes and their directions of transcription are shown at the

top of each panel. *p* values are indicated for interaction frequencies which are significantly higher for test regions when compared to those of control regions. Scales (in kb) are shown at the bottom of each panel.

Significant interaction frequencies were observed at -31 with respect to its control region at -25 in both luciferase control and GATA1 knockdown K562 cells (Figure 6.16). However, over 50% reduction of interaction frequency at -31 was observed at GATA1 knockdown cells comparing to luciferase control. Additionally, overall interaction frequencies of not only -31 but also its control regions at -25 and -41 were largely decreased. This suggested that the looping interaction between CTSs at +57 and -31 was disrupted as a result of GATA1 knockdown and subsequent affects on stability of cruciform configuration. However, the reduced frequency of interaction also indicated that the particular "+57/-31" interaction was not completely abolished, as otherwise no significant interaction would be detected at -31 with respect to the control region.

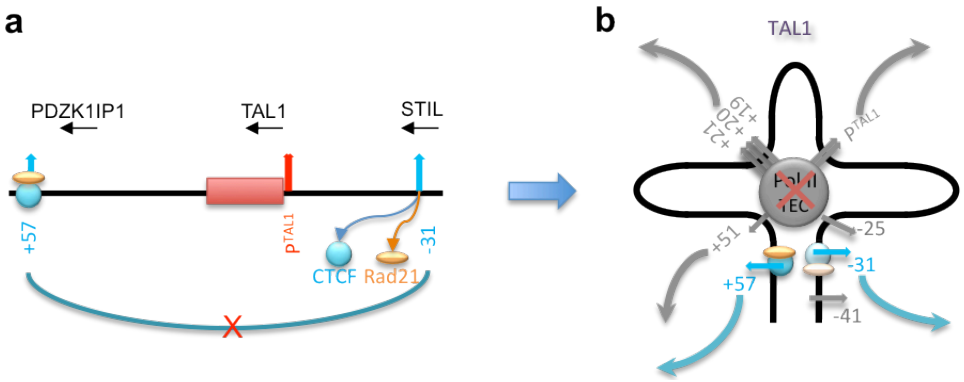


Figure 6.17: Schematics diagrams of GATA1 knockdown disrupting CTCF/Rad21 occupancy and CTSs interaction. Panel A: a genomic map illustrates the loss of CTCF/Rad21 occupancy at -31 and looping interactions between CTSs at +57 and -31 due to GATA1 knockdown. Panel B: a predicted model shows the disengagement of CTSs interaction (the "stem" structure), which results in further disassociation of the "cruciform" configuration.

Observations in this section are summarised in Figure 6.17a, which illustrates the decreased occupancy of CTCF and Rad21 at -31 as well as the reduced interaction frequency between +57 and -31. In wild-type K562 cells, the looping interaction between +57/53 and -31 was likely mediated by CTCF and cohesin, which acted as the "stem" in stabilising the entire cruciform configuration (see Chapter 4 and 5). In contrast, GATA1 knockdown K562 cells showed not only the disengagement of three-way promoter-enhancer interactions (see previous section) but also the trend of disassociation of the cruciform structure (Figure 6.17b). This was because the interaction frequencies of the CTS at -31 and its adjacent control regions were all reduced, implying that the entire -31 part of "stem" was about to move away from the CTSs at +57.

6.5.7 *Depletion of GATA1 affecting the TAL1-STIL interactions*

The speculation was that the interaction was GATA1/TEC-mediated at least in a proposition of cells by which RNAP II was recruited at the STIL promoter (located within 1 kb upstream of the STIL intron 1). Providing the fact that the model of cruciform configuration proposed the co-localisation of the TAL1 promoter 1b and the STIL intron 1, the disassembling of cruciform structure in GATA1 knockdown K562 might subsequently affect this particular interaction. Based on the assumption that all interactions captured by the 3C-PCR and 4C-array in wild-type K562 cells were existed in the same cells at the same time, it was expected to observe the reduction of the STIL expression accompanied by the depletion of RNAP II occupancy over the STIL promoter would be observed, as in the results of GATA1 knockdown (see section 6.5.1 and 2). Therefore, 3C analysis was performed to assess the interaction between the STIL intron 1 (-81) and the TAL1 promoter 1b (3C anchor).

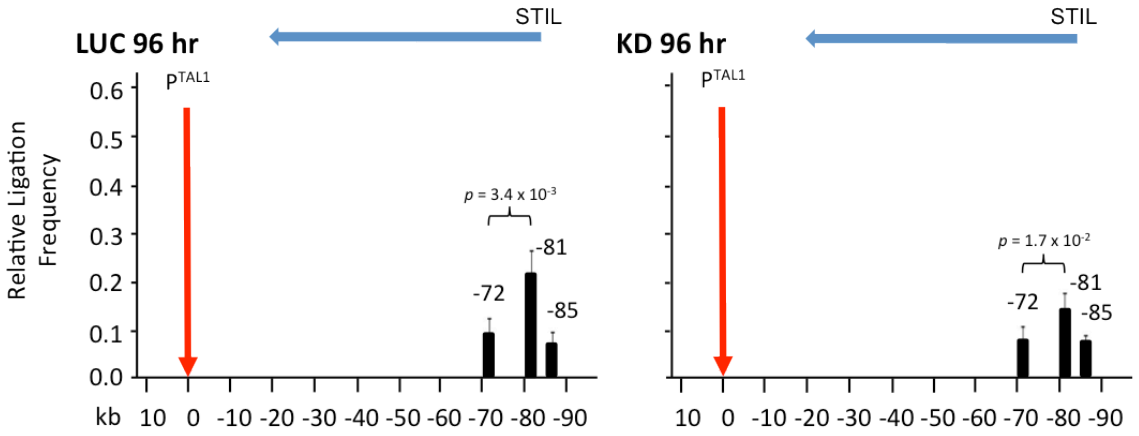


Figure 6.18: GATA1 knockdown results in lost of interaction between TAL1 promoter 1b and STIL intron 1. Histograms of the looping interaction after siRNA transfection with GATA1 (KD) or with luciferase (LUC) were determined by 3C. Interaction frequencies (black bars) are shown with standard errors and normalized relative to BAC controls. The location of 3C "anchor" is denoted with the vertical red arrow. Locations of genes and their directions of transcription are shown at the top of each panel. *p* values are indicated for interaction frequencies which are significantly higher for test regions when compared to those of control regions. Scales (in kb) are shown at the bottom of each panel.

Significant interactions were observed at the STIL intron 1 (-81) with respect to its control region at -72 in both luciferase control and GATA1 knockdown K562 cells (Figure 6.18). Similar to the interacting pattern between CTSs at +57 and -31, it was observed a decreased level of interaction frequency at the STIL intron 1 in GATA1 knockdown cells, suggesting that the STIL-TAL1 interaction existed in the same cells with other GATA/TEC-dependent interactions for at least a fraction of K562 cells.

6.5.8 *Summary of GATA1 knockdown in human K562 erythroid cells*

The predicted model in Figure 6.19 illustrated the effect of GATA1 knockdown on the cruciform structure of the TAL1 locus in human erythroid cells (based on the model proposed in Chapters 3, 4 and 5). The loss of GATA1 results in

(i) The loss of TAL1 and its neighbouring genes expression

(ii) The loss of the TAL1 erythroid complex bound at both the +51 enhancer and the TAL1 promoters,

(iii) The loss of RNAP II occupancy at the TAL1 regulatory elements

(iv) The loss of looping interactions between TAL1 promoters and its regulatory elements,

(v) The loss of CTCF/Rad21 occupancy at the -31 element,

(vi) The loss of looping interactions between the +57 and -31 insulator elements.

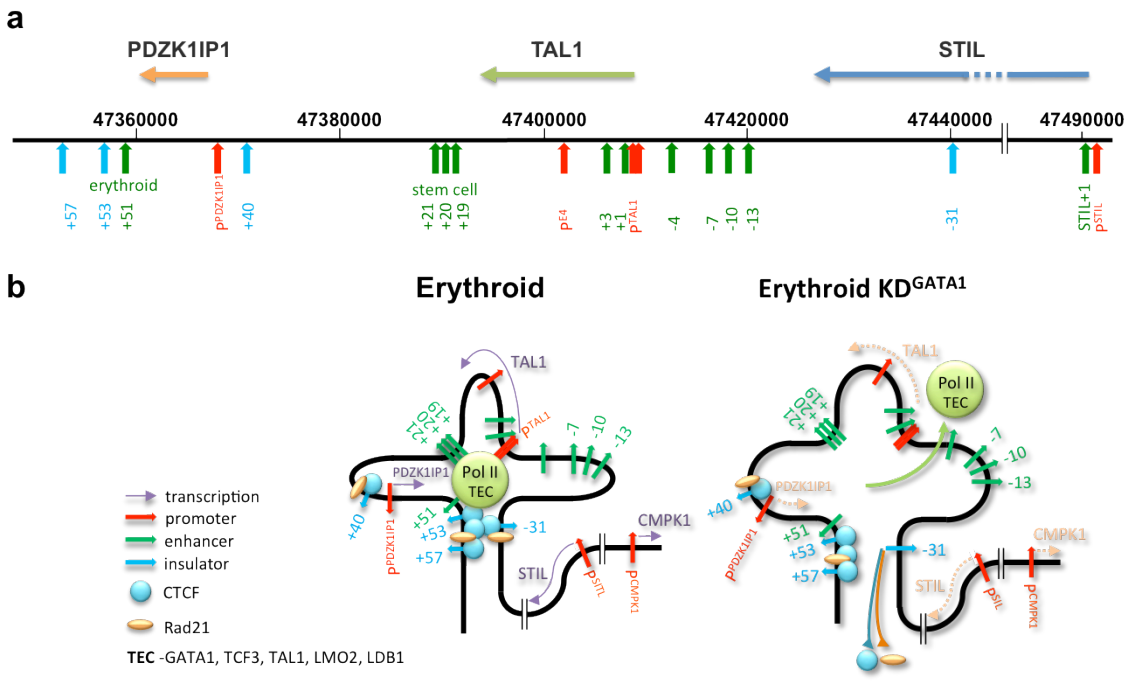


Figure 6.19: Structural organization of the TAL1 locus. (a) Schematic diagram depicting the organization of the human TAL1 locus. The scale is genome co-ordinates (bp) for human chromosome 1 (hg.17). Promoters, enhancers and CTCF binding sites are the vertical red, green and blue arrows respectively. (b) Chromatin organization of the TAL 1 locus in TAL1^{+} erythroid cells and in TAL1^{+} erythroid cells with a GATA1 knockdown (KD^{GATA1}). Locations of promoters, enhancers, and putative insulators are depicted as above. Direction of transcription of relevant genes (purple arrows), TEC and Pol II recruitment/loss at the hub,

and CTCF and Rad21 binding at insulators are also shown and detailed in the key. Perturbation of transcription by GATA1 knockdown is shown by the dotted orange arrows.

Overall, all of these events resulted in the dissolution of the cruciform configuration as illustrated in the Figure 6.19b. Furthermore, they provide strong evidence that the chromatin structure described as the TAL1 "cruciform hub" is present in a proportion of K562 cells expressing TAL1 and that this structure is important in the transcriptional regulation of the TAL1 locus.

Discussions

6.6 Looping interaction between TAL1 promoter and +51 enhancer was not affected until GATA1 knockdown at 96 hour

The chromatin loop formed between the TAL1 promoter 1b and the +51 enhancer is the major interaction detected in erythroid K562 cells, which plays a key role in erythroid-specific transcription of TAL1. The results presented in this chapter have demonstrated that this interaction is mediated by GATA1. However, the significant knockdown of GATA1 at mRNA and protein levels as well as the drastic depletion of GATA1 occupancy at the TAL1 promoter 1a (1 kb from 1b) and the +51 enhancer had no effect on this particular looping interaction at 48 hr time-point. This apparent "delay" in the loss of chromatin loops can be interpreted in a number of ways. Firstly, it may suggest that the chromatin loop is stabilised by additional transcription factors other than GATA1 itself, such as other members of TEC. It is why the disengagement of this particular loop was observed only at the 96-hour time-point when the entire TEC was depleted from the TAL1 locus.

Moreover, this chromatin loop is not stand-alone. It is also involved in a more sophisticated cruciform chromatin configuration, which includes other looping interactions between not only promoters and enhancers but also CTCF-binding sites (CTSs). Thus, spatial co-localisation of the TAL promoter 1b and +51 is also dependent on the interacting partners as a whole. The loss of GATA1 may abolish the direct physical link between the TAL promoter 1b and +51 through GATA1 protein at 48 hr time-point. However, the temporal effects of GATA1 knockdown may not subsequently abolish the entire cruciform structure, which is still maintained by other proteins, most likely via CTCF and Rad21 at the "stem" of the

structure. As a result, the TAL promoter 1b and +51 can remain in spatial close-proximity when still being situated in the cruciform structure.

6.7 Looping interactions of the TAL1 "chromatin hub" are dependent on GATA1 and the TEC

The depletion of GATA1 and the TEC at the 96-hour time-point has dramatically disrupted the three-way promoter-enhancer looping interactions between the TAL1 promoter 1a, +51 and +20/+19, which resulted in disassociation of the cruciform configuration. It is in agreement with the loss of chromatin looping configurations between LCRs and its promoters in human and mouse β-globin loci, as the results of GATA1 or LDB1 knockdown (Song et al., 2007; Woon Kim et al., 2011).

6.7.1 *Is the loop between TAL1 promoter 1b and the +20/+19 enhancer directly dependent on GATA1/TEC?*

It is known that the P^{TAL}-+51 is a GATA1-dependent loop as stated in the previous section. However, no direct evidence has shown that looping interaction between TAL1 promoter 1b and the +20/+19 enhancer is mediated directly via GATA1 protein in this study, although the loss of GATA1/TEC indeed accounts for the loss of P^{TAL1}-+20/+19 interaction. Previous studies in Green's laboratory have demonstrated that the TAL1 stem cell enhancer (+20/+19) contains multiple GATA and Ets motifs, but is regulated by GATA2 and Ets factors Fli-1 and Elf1 for TAL1 expression during the HSC formation (Gottgens et al., 2002b). This enhancer is mainly active in the HSC and progenitor compartments but not in terminal differentiated lineages. However, this looping configuration between the TAL1 promoter and +20/+19 is still existed, regardless of its enhancer activity in erythroid K562 and lymphoid HPB-ALL cells. Particularly, this looping interaction was observed in HPB-ALL cells (see Chapter 3 and 5) where the TAL1 and GATA1 were not expressed. Providing that this interaction is able to exist in the absence of the GATA1 and the TEC, it suggests that the loop formation between the TAL1 promoter 1b and +20/+19 enhancer can exist in a GATA1/TEC-independent manner. Further ChIP analysis can be performed to assess the potential candidate of the TFs (preferable GATA and Ets factors) that may mediate the looping interactions in K562 and HPB-ALL, although the TFs being used in different lineages may vary. Taken together, the loss of "P^{TAL1}-+20/+19" interaction

during GATA1 knockdown is considered as the consequence of disengagement of the entire cruciform configuration, which is GATA1/TEC-dependent.

6.7.2 *GATA1 works along with other TEC member in maintaining the cruciform configuration*

As discussed in the previous section, the loss of GATA1 does not have an immediate affect on the loops formed between the TAL1 promoter 1b and +51 at 48 hr time-point. Given the fact that drastic removal (over 60%) of GATA1 occupancy was observed at the TAL1 promoter and the +51 enhancer, it suggests that the loss of GATA1 itself may not be enough to abolish the loop in presence of other TEC members, although GATA1 is critical for the initial recruitment of the TEC in mediating formation of the erythroid-specific loop.

6.7.3 *Additional experiments for determining temporal relationship between the loss of TEC occupancy and the loss of chromatin loops*

ChIP-occupancy of the TEC members including LDB1 and E2A/TCF3 were not assessed at the 48-hour time-point. Thus, it is unable to determine whether the TEC has lost at the earlier time-points other than 96 hour. In addition, all looping interactions apart from the P^{TAL1}-+51 were also assessed at 96 hour but not at 48 hour, which may fail in providing further evidence to determine the temporal relationship of the loss between looping interactions and occupancy of the TECs. Moreover, it would be interesting to study the temporal effect of GATA1 knockdown in a reversed way. As it has been determined the complete disassociation of TAL1 chromatin configuration at 96 hour, the process of restoring this cruciform structure at the TAL1 locus can be studied simply by stopping the GATA1 knockdown. In addition, as result of the cruciform configuration, the TAL1 promoter and other genomic elements within loops are also satiated in close-proximity. It is expected that reductions of interaction frequencies between these elements accompanied with the disengagement of the cruciform structure. Therefore, the high-throughput 4C-array analysis can also be used to study the overall loss of the entire “cruciform” structure on GATA1 knockdown K562 cells, which would provide much more comprehensive and informative interaction profiles.

6.8 GATA1 and TEC are required for TAL1 expression in human erythroid lineage.

It has been shown that the reduction of TAL1 expression is accompanied with the depletion of GATA1 and the TEC as well as the disengagement of looping interaction between TAL1 promoter and the +51 erythroid enhancer. Providing the fact that the critical roles of GATA1 and the TEC in erythroid-specific gene expression (Lahlil et al., 2004), the observations in this chapter delivered two lines of conclusions. Firstly, the knockdown of GATA1 and a subsequent loss of the TEC occupancy at the TAL1 promoter and the +51 enhancer lead to the completely abolishment of looping interaction between them, which demonstrated that this particular interaction can be GATA1/TEC-dependent. Secondly, the expression of TAL1 was affected at 48 hr, which was before the interaction being affected. In addition, there was approx. 50% of the TAL1 expression remained at the 96-hour time-point when the interaction was completely abolished, indicating there might be an allele-specific expression for the TAL1 gene in the erythroid K562 cells. It was speculated that only half of the TAL1 expression was dependent on the “cruciform” configuration, while another allele of TAL1 was transcribed under totally different mechanism that remained to be uncovered.

Alternatively, as two GATA sites were identified at the TAL1 promoter 1a, it may suggested that the TAL1 promoter 1a is capable of recruiting GATA1 by its own, in order to initiate the transcription independent of the TEC in some of the K562 cells. Consequently, knockdown of GATA1 may also lead to a direct affect on the GATA1 recruitment at the TAL1 promoter and the subsequent transcription initiation, which would explain why the TAL1 expression was decreased before disassociation of the TEC-dependent cruciform configuration at the 96-hour time-point. For the proposition of the cell that the TAL1 expression in looping-dependent, the substantial loss of interactions between the TAL1 promoter and its enhancers resulted in further down-regulation of the TAL1 expression as being observed at the 96-hour time-point.

Nevertheless, the loss of this particular erythroid-specific interaction due to GATA1 knockdown resulted in the reduction of the TAL1 expression, which suggests that the GATA1/TEC is required for the appropriate transcription of TAL1.

6.9 Expression and RNAP II recruitment at the TAL1 locus are partially dependent on GATA1/TEC

Reduction of RNAP II occupancy along with reduced level of gene expression at the TAL1 locus during GATA1 knockdown suggest that GATA1 is essential for recruitment of RNAP II as well as maintaining active transcription of TAL1 and its neighboring genes. Similarly, knockdown of GATA1 and NF-E2 proteins at the human β–globin loci led to the transcriptional reduction of the γ-globin genes paralled by the reduction of RNAP II occupancy across the gene. In addition, the adjacent ε-globin gene was also showed a decreased transcription in knockdown cells (Woon Kim et al., 2011).

TAL1 expression was dramatically decreased to half of its normal level after knocking-down GATA1 at 96 hour, implying its transcription is closely related to GATA1 in K562 cells, in agreement with the role of GATA1 in activating erythroid-specific genes (Courtes et al., 2000). Relatively, the occupancy of RNAP II at the TAL1 enhancers and promoter 1a was also subject to depletion of GATA1. Although the percentages of RNAP II reduction seem to be more drastic at the +51 enhancer comparing to the TAL1 promoter 1a (86.2% vs. 68.8%), there is merely enough evidence to distinguish the initial RNAP II recruitments between these two regions as the ChIP efficiency can vary between different binding sites. Thus, it is unclear whether the RNAP II recruitment were initiated either at +51 and being delivered to the promoter 1a, or recruited at promoter 1a independent of the +51 enhancer. However, it is clear that the recruitment of RNAP II at these two sites is also GATA1-dependent based on the observation in this chapter. In addition, the degree of reduction of RNAP II occupancy at promoter 1a agrees with the level of down-regulation of TAL1 expression after GATA1 knockdown.

The partial loss of RNAP II at the +20/19 enhancer is possibly due to the dissociation of the entire cruciform structure after GATA1 knockdown in K562 cells. Similarly, although it was unable to assess the degree of loss of RNAP II occupancy at the PDZK1IP1 promoter, it is speculated that the reduction of PDZK1IP1 expression may also be related to the loss of the entity of the cruciform structure at 96 hour. As the PDZK1IP1 is situated at this TAL1-centered transcription hub, a pool of highly enriched RNAP II and other transcriptional activators actually provides a positive environment for its low-level transcription

(over 1000-folds less than the level of TAL1 transcription). In addition, GATA1 depletion also resulted in a minor reduction of the RNAP II occupancy at the promoter of STIL, whereas expression of STIL was affected in a much drastic way. It suggests that although the STIL transcription and RNAP II recruitment are partially dependent on the GATA1/TEC transcription hub, other different machineries may also involved in its regulation. Interestingly, the expression of CMPK1 is partly GATA1-dependent, however, GATA1 depletion had no affect on RNAP II recruitment at its promoter, suggesting that GATA1 may not regulate CMPK1 directly. Instead, it may involve in modulation CMPK1 expression through the downstream target genes of GATA1, without affecting RNAP II occupancy.

Taken together, it can be concluded that loss of RNAP II and expression of TAL1 and its neighbouring genes are not only depended on GATA1/TEC but also very much relied on the entity of the TAL1 “cruciform” configuration. The relationship between TAL1 chromatin organisation and co-regulation of TAL1 and neighbouring genes is further discussed in the following section.

6.10 CTCF and Rad21 occupancy and looping interactions between CTSs are GATA1/TEC-dependent in TAL1 expressing K562 cells

Both CTCF and Rad21 (cohesin) are required for chromatin organization and loops formed between CTSs involved in transcription regulation via either situated the promoter and enhancers within a same active loop or separate the promoters and *cis*-element into different loops (Phillips and Corces, 2009). Regardless the fact that significant CTCF occupancy was observed at +57 and -31 in both K562 and HPB-ALL cells, the looping interaction between this two regions was observed only in erythroid K562 cells where TAL1 is transcriptionally active (see Chapter 4). In addition, Rad21 occupancy at -31 was much lower in HPB-ALL than in K562 cells (see Chapter 4), suggesting the recruitment of Rad21 at -31 may be TAL1 transcription-dependent. Moreover, It was also found that GATA1 knockdown for 96 hour resulted in reduction of CTCF and Rad21 occupancy at -31 as well as the interaction between +57 and -31. All these evidence suggest that the lower level of Rad21 occupancy at -31 is related to its reduced interaction frequency with the CTS at +57.

Reduction of CTCF and Rad21 at -31 is accompanied with the depletion of GATA1/TEC, further elucidated that the recruitments of CTCF and Rad21 at the TAL1 locus are very likely dependent on GATA1/TEC-related machinery. Similarly, reduction of looping interaction between CTSs at +57 and -31 along with disassociation of the cruciform structure further supports the theory of this CTS-specific loop playing a critical role in maintenance of chromatin structure at the TAL1 locus. It agrees with the previous observation that CTCF and Rad21 played a key role in maintaining the "stem" structure of the multi-looping configuration between the LCRs and globin genes at the human β-globin locus (Kim et al., 2012). The depletion of GATA1 led to a decreased affinity of physical links between the TAL1 promoter and +51 by the TEC. However, the occupancy of CTCF and Rad21 at the "stem" of the cruciform structure may not be affected at the 48-hour knockdown of GATA1, suggesting that CTCF/Rad21 play a key role in stabilizing this configuration by maintaining all the regulatory elements situated still within the spatial close-proximity, even without the direct link by GATA/TEC.

Taken together, it provides a structure link between CTCF/Rad21 recruitments, interactions between CTSs and transcriptional regulation of TAL1. Looping interaction between +57 and -31 is served as a stabilizer in maintaining the cruciform configuration that favored for TAL1 transcription. This argument is fully supported by a number of recent studies on CTCF/cohesin that have demonstrated their critical roles in facilitating and maintaining the chromatin-looping configuration in mammalian gene loci (Chien et al., 2011; Degner et al., 2011; Guo et al., 2011; Hadjur et al., 2009; Handoko et al., 2011).

6.11 Models of co-transcriptional regulation of the TAL1 locus in a cruciform structure dependent manner

Two potential models were proposed as shown in Figure 6.15, in order to illustrate how genes flanking TAL1 may use the TAL1 chromatin hub to facilitate their transcription.

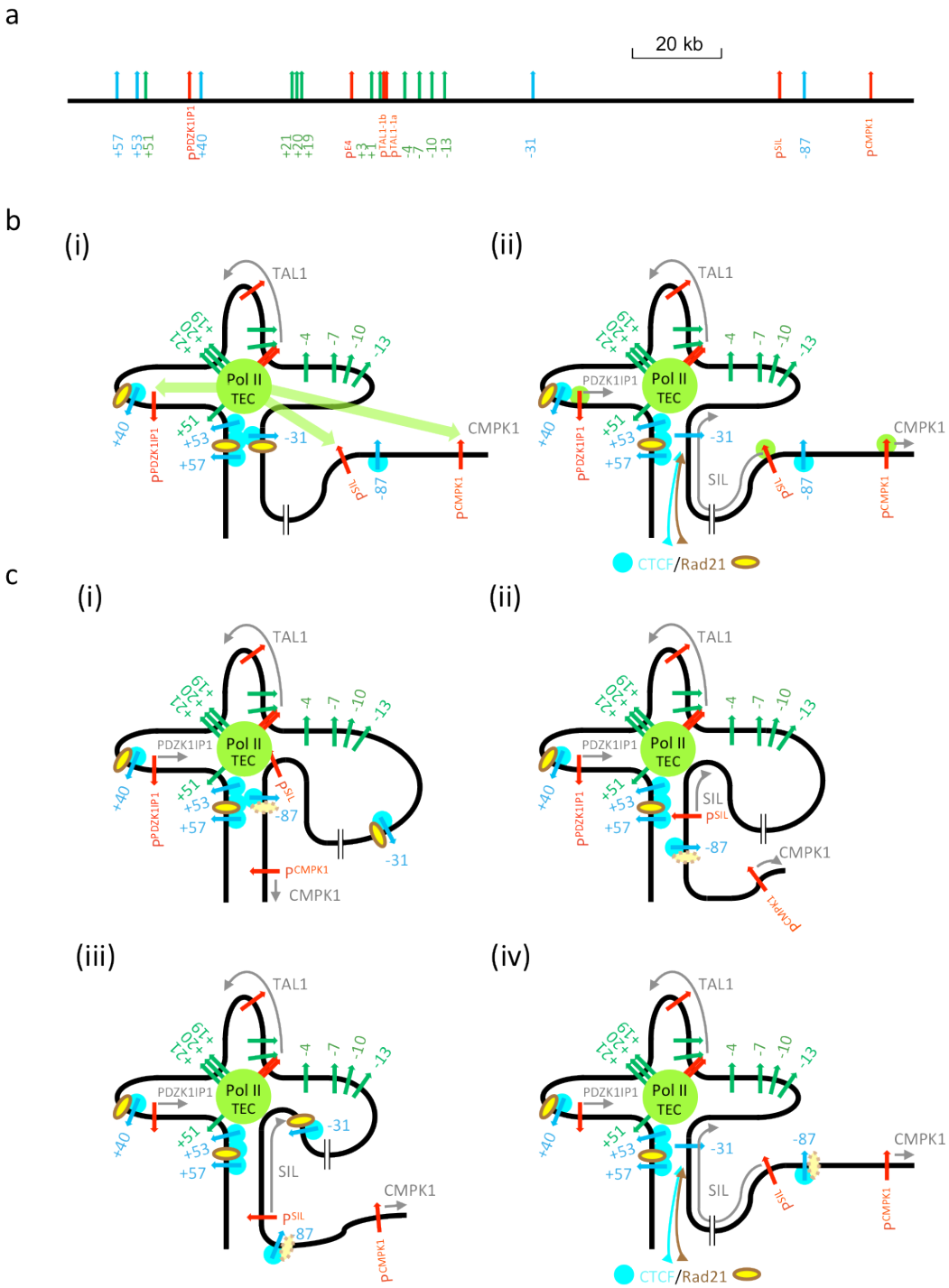


Figure 6.20: Models of STIL, CMPK1 and PDZK1IP1 transcription which is dependent on the TAL1 chromatin hub. (A) Linear schematic diagram showing the organization of the human TAL1 locus. Details as described previously (B) The recruitment model. The proximity of the STIL, CMPK1 and PDZK1IP1 promoters to the TAL1 chromatin hub favours the recruitment of Pol II and other factors to their respective promoters (shown by light green arrows connecting the hub to the promoters) in a hub-dependent step (i). Transcription can then occur from these promoters in a hub-independent manner (ii). (C) Direct interaction model. The promoters of STIL, CMPK1 and PDZK1IP1 engage directly with the Pol II machinery within TAL1 chromatin hub which is entirely hub-dependent at all stages of transcription. Transcription is facilitated by the movement of chromatin through the hub with loops becoming large or smaller accordingly (shown in this figure with respect to SIL transcription). Locations of promoters, enhancers, and insulators are depicted as in Figure 6.14. Direction of transcription of relevant genes (grey arrows), TEC and Pol II recruitment at the hub, and CTCF/Rad21 binding at insulators are also shown as in Figure 6.14. Pol II recruitment to gene promoters via the recruitment model in A is shown by green balls. Binding of CTCF at -87 between the STIL and CMPK1 promoters (Dhami et al. 2010) is also shown – Rad21 binding to this region is not known (and shown partially shaded).

6.11.1 *The recruitment model*

The first model is termed the “recruitment model”, which illustrated a hub-dependent recruitment of Pol II and other factors over promoters of TAL1 neighbouring genes including STIL, CMPK1 and PDZK1IP1 (Figure 6.20b, panel i). Once loaded with Pol II and other transcription factor complexes, the following transcription processes occurred independently from these promoters (Figure 6.20b, panel i). The RNAP II was recruited and enriched at the centre of the TAL1 chromatin hub. It is speculated that when the promoters of STIL, PDZK1IP1 even CMPK1 were brought into the spatial close-proximity to the TAL1 promoter via looping, they were situated in the centre of the hub, from where the RNAP II being loaded. As soon as the RNAP II was recruited at the promoters of neighbouring genes, the looping structure may no longer be required for the process of transcriptional elongation of these neighboring genes (Figure 6.20b, panel ii).

In supporting this model, evidences collected in this study are listed as followed. First, looping interaction was detected between the TAL1 promoter 1b and the PDZK1IP1 promoter by 4C-array in K562 cells (see Chapter 5). Second, the interaction was also detected between the TAL1 promoter and the STIL intron 1 region (1 kb in distance to the STIL promoter) by 3C-PCR and 4C-array in K562 cells. In addition, a significant reduction of this looping interaction was observed along with disassociate of the cruciform structure during GATA1 knockdown, implying this interaction is very likely dependent on existence of the cruciform hub. Third, the RNAP II recruitment at the STIL promoter was also affected by the loss of cruciform structure. In summary, these evidence suggest that the promoters of PDZK1IP1 and STIL are indeed brought into the spatial close-proximity with the TAL1 promoter, from where the RNAP II is being recruited.

However, there are also several lines of observations that cannot be explained based on this model. First, no evidence suggests that the CTCF and Rad21 can be removed from the CTS at -31 during STIL transcription, providing the fact that interaction between CTSs at +57 and -31 stabilizes the cruciform structure and high-level of CTCF and Rad21 occupancy are observed at these two sites. Thus, this particular interaction between CTSs at +57 and -31 becomes an obstacle for the RNAP II elongation across the gene body to get full transcripts of STIL. Second, no evidence has been shown how the RNAP II gets recruited at the

promoter of CMPK1 in a cruciform-dependent way, specifically no chromatin interaction was observed between promoters of TAL1 and CMPK1. In addition, disruption of cruciform configuration led to no effect on RNAP II recruitment at CMPK1 promoter, also suggesting that the CMPK1 promoter recruits RNAP II through a different mechanism. Third, the complete disassociation of TAL1 cruciform configuration only accounts for the partial reduction of RNAP II over the STIL promoter indicating its recruitment is related but not completely dependent on the cruciform hub.

6.11.2 *The direct interaction model*

The second model is called the “direct interaction model”, which proposed that the promoters of STIL and PDZK1IP1 might be in directly contact with the TAL1 chromatin hub via the Pol II machinery or other factors which bring them into contact (Figure 6.20c). Transcription of these genes acted in an entirely hub-dependent manner. This model mainly accounts for the possible transcription mechanism of upstream STIL gene. Initially, the promoter of STIL was brought into the centre of the TAL1 cruciform hub via chromatin looping (Figure 6.20c, panel i). It is speculated that the structure may be stabilized via interactions between CTSs at +57/+53 and, the -87 - a CTCF-binding site previously identified which is situated in between the STIL and CMPK1 promoters (Dhami et al., 2010). In this model, STIL, TAL1 and PDZK1IP1 share the same transcriptional hub as they are all brought into a spatial close-proximity. Subsequently, the STIL gene is getting transcribed when moves outward the transcriptional hub (Figure 6.20c, panel ii & iii). Similar to the recruitment model, it also requires the removal of CTCF/Rad21 occupancy at -31 in order to get full-length transcripts of STIL (Figure 6.20c, panel iv).

A number of evidence derived from 3C and 4C studies in this thesis are listed as followed to support this model. First, the 4C-array of P^{TAL1} detected not only the interactions with the STIL intron1 but also multiple contact points across the STIL gene body, of which aligned up with the STIL exons (data showed in chapter 4). It is speculated that these regions are all in spatial close-proximity with the TAL1 promoter at certain time-points during STIL transcription, as the entire gene body has gone through the transcription hub where the TAL1 promoter located. It is also speculated that these contact points may correspond to locations of where RNAP

II poses during transcription. Second, an additional CTCF binding site at -87 was previously identified by ChIP-chip assay in K562, which might serve as a stabilizer for the initial contact between the P^{TAL1} and STIL intron1 via interacting with CTSs at +57/+53. Third, it has been reported that the PDZK1IP1 and TAL1 are frequently co-expressed in a number of hematopoietic lineages, but PDZK1IP1 is transcribed in a much lower level (Delabesse et al., 2005). This model explains the low-level transcription of PDZK1IP1 as a bystander effect of being situated in the active TAL1 transcriptional hub.

On the other hand, this model suggests that the STIL transcription is fully depended on the TAL1 transcriptional hub. However, disruption of the entire TAL1 transcriptional hub do not result in a dramatic reduction of STIL expression, indicating this model is insufficient in covering the whole transcriptional machinery of STIL. Furthermore, as previously discussed, the production of a full-length STIL mRNA will require the transient removal of CTCF and Rad21 from the TAL1 -31 in both models. Although it has been demonstrated that CTCF and RAD21 binding at -31 is dynamic; transient loss of both proteins from -31 would allow the entire STIL gene passing through the hub to produce full-length transcripts.

In summary, each proposed model has its own imperfection in fitting all the results. Two models in combination would cover the most likely situation in K562 cells at different time points and/or different proportions of cells. However, it cannot rule out the possibility that these, or other models, may account for transcription of TAL1flanking genes in a co-ordinated way.

6.12 The *cis*-acting regulatory elements remodelling at the TAL1 locus during vertebrate evolution

The locations of transcription factor binding motifs at the TAL1 locus, including GATA, E-box, and Ets sites, had been allocated based on their conservation at the level of DNA sequence (Gottgens et al., 2010). Based on the literatures, a schematic map (Figure 6.20) illustrates the distribution of these TF binding sites which involves in transcriptional regulation of the TAL1 locus through the vertebrate evolution. In Frog and Chicken, the GATA and Ets motifs were heavily situated at the TAL1 promoters, and GATA/E-box motifs are located at two of its stem cell (+20/+19) and erythroid (+51) enhancers. In Platypus and Opossum, the

some of the Ets sites were shifted from the TAL1 promoters to the +20/+19 enhancer, whereas the E-box was no longer existed at the +20/+19 enhancer. In Opossum, the loss of one additional GATA motif was observed at the TAL1 promoters comparing to its ancestors. In human and mouse, the motifs at two enhancers were as same as in Opossum and Platypus, whereas only two GATA motifs were still conserved at the TAL1 promoter. Loss and relocation of these TF binding motifs could due to both point mutations and DNA rearrangement through evolution (Esteller, 2007; Park et al., 2008). However, the function of TAL1 and its pattern of expression are highly conserved throughout vertebrates from mammals to teleost fish (Chapman et al., 2004; Gottgens et al., 2002a), regardless of rearrangement between the TF binding motifs. As demonstrated in previous section, a “cruciform” configuration exists in TAL1 expressing erythroid cells which links all *cis*-acting elements. Although it is unclear whether all these TF motifs are required for TAL1 transcription at the same developmental or temporal stages, it may provide a possible mechanism of how the important *cis*-elements/TF motifs required for transcription can be brought together via the chromatin looping mechanism.

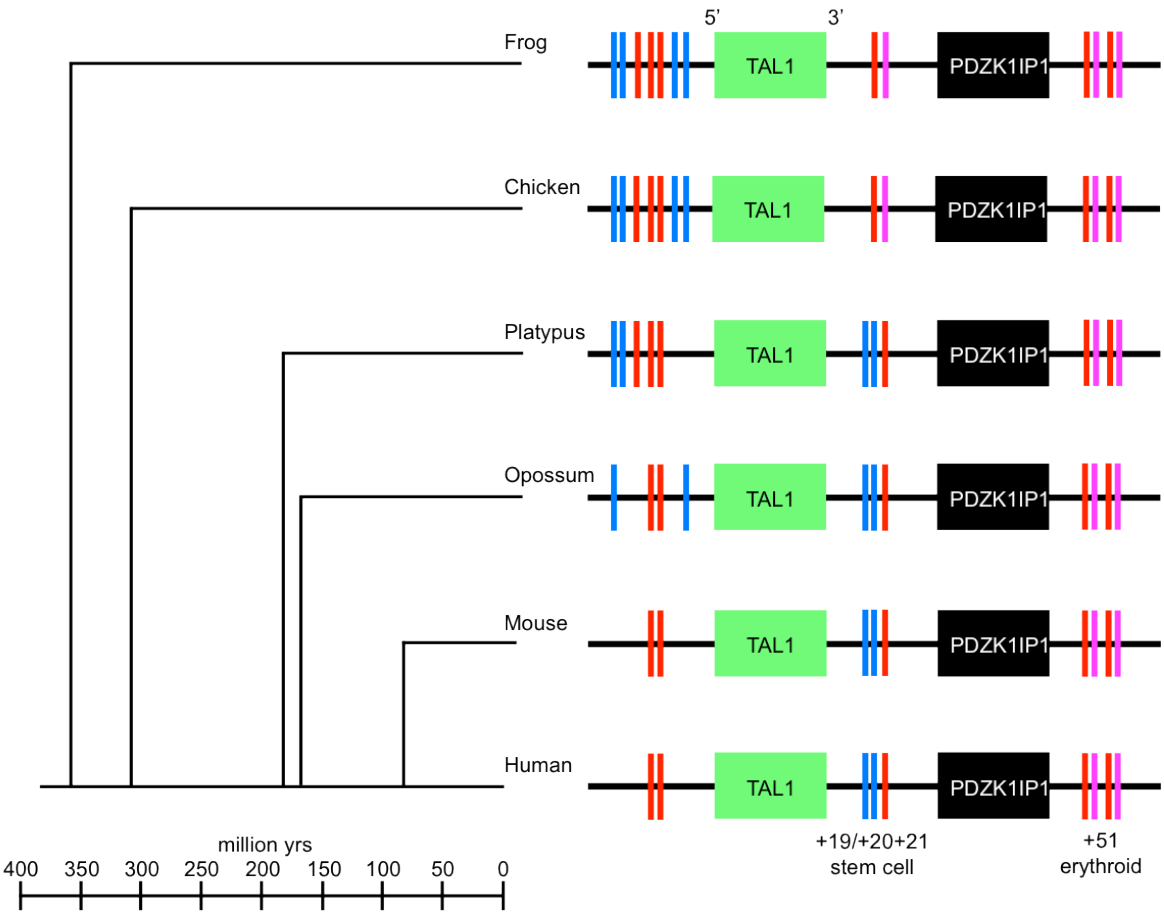


Figure 6.20: The TAL1 locus during vertebrate evolution. Left of the schematic shows the evolutionary tree of TAL1 across 400 million years of vertebrate evolution. Right of the

schematic shows the organization of the TAL1 locus. The TAL1 gene is shown in the dark green box. The PDZK1IP1 gene (black boxes) and it locations with respect to the TAL1 +51 erythroid enhancer is also shown. Ets, GATA and E-box motifs conserved at the level of DNA sequence are also shown.

Despite the highly conserved expression pattern of TAL1 across species, each independent *cis*-acting element reveals a considerable functional differences and an unexpected regulatory plasticity as a whole. These functional alternations are due to the switch of TF binding sites between *cis*-acting elements (Figure 6.20). For instance, the loss of Ets sites at the mammalian TAL1 promoter, which may be acquired by its stem cell +20/+19 enhancer. In turn, the loss of an E-box at the mammalian +20/+19 enhancer results in functional switch from an erythroid element to a hematopoietic stem cell/endothelial enhancer. In the ancestral TAL1 locus (Frog/Chicken, Figure 6.20), it contains two enhancers (orthologous to the human +20/+19 and +51 elements) with erythroid-specific GATA/E-box motifs. The functional redundancy may subsequently lead to the accelerated evolutionary alternation of the +20/+19 equivalent in losing its E-box without deleterious consequences. The exchanges of TF-binding motifs between *cis*-acting elements at the TAL1 locus, and how their functions are altered as the consequence of that during the evolution are illustrated in Table 6.5. It has been proposed that the *cis*-remodelling can be facility by the exchange of regulatory elements as the result of their physical interactions (Cameron and Davidson, 2009). In order to allow these site movements being accumulated through the evolution, the exchanges much occur at the very early state of the embryogenesis or in the germlines (LaVoie, 2003). The ancestral TAL1 promoter is active in midbrain, haematopoietic stem cell and endothelial cells, whereas in mammals the TAL1 promoters function only in midbrain as a result of losing the Ets motifs. There are two erythroid enhancers in ancestral TAL1 locus, whereas one of them is converted to stem cell enhancer as the erythroid-specific GATA/E-box motifs are replaced by the Ets and E-box motifs (Table 6.5).

Table 6.5: TF-binding motif exchanges and functional switches between cis-regulatory elements at the TAL1 locus during vertebrate evolution.

	TAL1 promoter	Stem cell enhancer	Erythroid enhancer
Ancestor	GATA/Ets	GATA/E-box	GATA/E-box
	Midbrain/HSC/Endothelial	Erythroid	Erythroid/Midbrain
Mammals	GATA	Ets/E-box	GATA/E-box
	Midbrain	HSC/Endothelial	Erythroid

Providing the fact that redundancy of regulatory elements has been found in many other vertebrate gene loci (Kuroda et al., 2009; Lorincz et al., 2004; Mayer et al., 2010), *cis*-regulatory remodelling is likely to be a widespread phenomenon. The TAL1 "cruciform" configuration presented in this thesis has provided an excellent paradigm for explaining the possible cause of *cis*-regulatory remodelling as well as how genes manage to cope with its consequences. The evidence are listed as follows:

1. In the cruciform hub, the TAL1 promoter and two of its enhancers are situated in a spatial close-proximity, which are highly increase the odds of DNA rearrangement. In addition, it is speculated that the exchange of TF-binding motifs between these regulatory elements through the evolution can also be a consequence of the structural links between these *cis*-acting elements.

2. In human and mouse, the TAL1 promoter itself has lost the ability of driving TAL1 expression in HSC and endothelial lineages as the results of losing Ets motif. Co-localisation with the stem cell enhancer may allow the TAL1 promoter to compensate its functional loss via looping, and using its Ets motif to recruit the corresponding transcription factors for the proper transcription of TAL1.

3. The interactions between the TAL1 promoter and enhancers containing the GATA/E-box motif are critical for enhancing the TAL1 transcription in the erythroid lineages. In addition, it is speculated that the ability of the TAL1 promoter interacts with the stem cell enhancer may have long been established, as it contains GATA/E-box in its ancestral version in chicken TAL1 locus may be acting as an erythroid-specific enhancer. And one can speculate that TAL1 promoter and this enhancer may interact through looping in chicken TAL1 locus.

4. In erythroid cells, the TAL1 "cruciform" hub accommodates these three key *cis*-regulatory elements, containing all the TF-binding motifs that are required for TAL1 transcription. In contrast, in lymphoid cells, the erythroid-specific loop (the +51 erythroid enhancer) is not activated. However, the "stem cell" loop remains even when TAL1 is not expressed, suggesting this

loop is transcriptional independent. It is speculated that this looping structure may be predisposed in the haematopoietic stem cells (HSCs), which is required to drive the TAL1 transcription, and subsequently being passed to the terminally differentiated cells such as erythroid and lymphoid lineages. The key point to be highlighted here is that the cruciform hub model provides a possible mechanism of how the TAL1 gene overcomes evolutionary constraints in order to gather all its functional elements into the same transcriptional machinery for the proper transcription of TAL1.

Conclusions

The work presented in this chapter have illustrated that TEC regulated the expression of TAL1 and its neighbouring genes in the erythroid K562 cells. It has also demonstrated that the formation of the TAL1 cruciform configuration is also GATA/TEC-dependent. In addition, the synchronous loss or at least decreased interacting frequency of all loop interactions between the TAL1 regulatory elements previously defined by 3C and 4C due to the GATA1 knockdown provides direct lines of evidence to suggest that all these interactions exist in the same cells at the same time. Taken together, it concludes that transcription regulation of the TAL1 locus is depended on the recruitment of the TEC accompanied with formation of loops between regulatory elements within the locus.

Providing the fact that the GATA/E-box motif shows a genome-wide distribution, it suggested that transcription regulation in the TAL1 locus could be one of these examples which under regulation of the TEC. As it was previously discovered, the entire TEC was also bound to the LYL1 promoter 1 and its putative +33 enhancer, depended on the same GATA and GATA/E-box motifs found in the TAL1 locus. It would be interesting to determine whether the TEC acting the similar role in a different locus. Thus, LYL1, also a paralogous gene of TAL1 would be studied using GATA1 knockdown system as presented in the subsequent chapter.

Chapter 7 Chromatin looping at the LYL1 locus: characterization of a putative LYL1 enhancer element with functional similarities to the TAL1 erythroid enhancer

Summary

Based on results showed in Chapter 6 and previous studies about erythroid-specific transcriptional regulation of the TEC, a genomic element located 33 kb downstream of the LYL1 promoter was speculated to have a similar function as the TAL1 +51 enhancer and involve chromatin looping interactions with the LYL1 promoter to modulate the LYL1 expression. Enhancer activity of the LYL1 +33 element was determined using transient reporter assays in human K562 cells. In addition, chromatin loop formed between the LYL1 +33 and its cognate promoter was detected by 3C in K562 cells. Further siRNA knockdown analyses demonstrated that this looping structure was GATA1 and/or TEC-dependent - as was the case for the erythroid enhancer-promoter interactions at the TAL1 locus (described in Chapter 6). The work in this chapter demonstrates a putative transcriptional regulation of the LYL1 gene via chromatin looping interactions.

7.1 Introduction

Similar to TAL1, the LYL1 (Lymphoblastic leukemia derived sequence 1) gene encodes an haematopoietic-restricted bHLH transcription factor, which was first identified at chromosomal translocation breakpoints t(7;19)(q35;p13) occurred in T-ALL patients (Mellentin et al., 1989).

7.1.1 Gene and protein structure of LYL1

The human LYL1 gene maps to the short arm of chromosome 19 (19p13.2), which consists of four exons and produces a ~1.5 kb RNA transcript (Cleary et al., 1988; Mellentin et al., 1989). The LYL1 gene encodes a bHLH protein with 267 amino acids and molecular weight of ~28 kDa (Visvader and Begley, 1991). The mouse LYL1 gene is located on chromosome 8 and its protein shares 78% amino acid sequence identity with human LYL1 (Kuo et al., 1991). The LYL1 protein is preferentially located in the nucleus during normal blood development, however

ectopic expression of the protein can be detected in the cytoplasm of the myeloid leukemia cells.

7.1.2 Expression of the LYL1 gene

7.1.2.1 Normal expression of LYL1

LYL1 expression is restricted to haematopoietic lineages such as myelocytes and erythrocytes in adults (Visvader et al., 1991). In particular, LYL1 has been found to be expressed in most B-cell lineages, which is down-regulated during terminal differentiation, but not expressed in most T-cell lineages (Kuo et al., 1991). Although LYL1 is broadly expressed in the haematopoietic system, the highest expression is detected in bone marrow progenitors and pro-B cells (Capron et al., 2006; Chambers et al., 2007). LYL1 expression has also been reported in angiogenic and mature adult endothelium (Pirot et al., 2010). In addition, it has also been shown that LYL1 is highly expressed in stem cells and progenitor cells, consistent with its critical role in controlling the size and function of the stem/progenitor compartment (see section 7.1.3).

7.1.2.2 Aberrant expression of LYL1

Ectopic expression of LYL1 has been implicated in haematopoietic malignancy. In T-ALL, translocation situates the LYL1 gene under regulatory control of β-TCR, which leads to its ectopic expression (Mellentin et al., 1989). Moreover, over-expression of LYL1 has also been found in T-ALL patients in the absence of chromosomal abnormalities (Ferrando et al., 2002). In addition, it has been shown that the LYL1 expression is observed in most patients with acute myeloblastic leukemia or high-risk myelodysplastic syndrome (Meng et al., 2005). It has also reported that LYL1 may contribute to the growth and differentiation patterns and drug resistance of AML cells, proposing a potential role of LYL1 as an oncogenic factor in AML. Recently, the important role of LYL1 in AML cell proliferation has been demonstrated with further knockdown study using a lentiviral shRNA recombinant vector in $CD34^+$ cells from AML patients (Meng et al., 2009).

7.1.3 LYL1 functions

Numerous studies have illustrated the critical roles of LYL1 associated with blood-related development and maintenance. First, it has been shown that the LYL1-null mice display normal level of blood cell counts apart from the reduced B-lymphoid population (Capron et al., 2006). In addition, the loss of LYL1 also leads to reduction of the frequency of immature progenitors, CFU-S12 (S12 colony-forming unit) and LTC-IC (long-term culture-initiating cell) content in the E14 fetal liver and bone marrow. As a result, the competitive reconstituting abilities of E14 fetal liver and bone marrow were severely impaired, particularly to B- and T-lymphoid lineages. Taken together, it suggests that the LYL1 is functionally important for HSC properties and lymphoid differentiation, which is largely distinct from TAL1 functions. Second, LYL1 is essential for the maintenance of normal HSC functions in the absence of TAL1 (Souroullas et al., 2009). It has been found that the adult Tal1-null HSCs can survive with a single allele expression of Lyl1 whereas the $Lyl1^{-/-}$ Tal1 null HSCs undergo rapid apoptosis, implying its role in adult HSC maintenance. In contrast, the Lyl1 null HSCs engraft very poorly when the single allele Tal1 is expressed. Although it has been reported that the LYL1 is unable to rescue TAL1 null haematopoiesis (Chan et al., 2007), LYL1 has a more fundamental role in maintaining adult HSCs, implying that TAL1 and LYL1 differ in their relative roles during the development and adult haematopoiesis. Most recently, a novel role of LYL1 has been reported that it functions as a major regulator in the postnatal maturation of newly formed blood vessels (Pirot et al., 2010). The depletion of LYL1 in human endothelial cells has demonstrated that it regulates the expression of molecules that are associated with the stabilisation of vascular structure. A null allele of LYL1 has been generated to study the residual function of the N-terminus in the absence of the bHLH region (Souroullas and Goodell, 2011). The function of LYL1 is detectable but relatively weak without the bHLH, resulting in a reduced function in lymphoid development as well as in haematopoietic repopulation. It demonstrates that the bHLH region is critical for LYL1 function in blood development.

7.1.4 LYL1, a Class II bHLH transcriptional factor

Class II bHLH transcription factors like LYL1 and its paralogue TAL1, regulate transcription by binding to target gene sequences as heterodimers with E-proteins

including E2A, E12 and E47 (Blackwell and Weintraub, 1990; Lassar et al., 1991). They recognize an E-box domain (CANNTG) in order to activate or repress transcription (Goldfarb and Lewandowska, 1995; Hofmann and Cole, 1996; Hsu et al., 1994; Wadman et al., 1997).

7.1.4.1 LYL1 modulates transcription via interacting with E2A proteins

It has been shown that LYL1 can inhibit the regulatory activity of the bHLH factor E2A/HEB (Zhong et al., 2007). Some of the E2A/HEB target genes such as CD5, pTa and RAG1/2 were subsequently found to be down-regulated in the thymus of lymphomagenic LYL1 mice. In addition, transcriptional activity of CD4 can be repressed by LYL1 regardless of the absence of exogenous E2A/HEB, implying that LYL1 may use a different mechanism other than competitive binding to E2A/HEB heterodimers.

7.1.4.2 Co-regulation of LYL1 and other haematopoietic transcription factors

Genome-wide binding profiles of LYL1 along with other nine key haematopoietic transcriptional regulators (i.e. TAL1, LMO2, GATA2, RUNX1, MEIS1, PU.1, ERG, FLI-1 and GFI1B) have been reported in haematopoietic progenitor cells using ChIP-seq (Wilson et al., 2010). It has been shown that numerous significant overlaps of the binding patterns are observed for pairs of TFs, not only involving known partners such as TAL1/LMO2/LYL1/GATA2, but also involving an additional novel partner RUNX1 with either TAL1, LMO2, LYL1 or GATA2. This comprehensive analysis reveals that tight combinatorial interactions among TAL1, LYL1, LMO2, GATA2, RUNX1, FLI-1 and ERG play a crucial role in transcriptional regulation of haematopoietic stem/progenitor cells.

7.1.4.3 LYL1-CREB1 protein complex in transcriptional regulation

In addition to the associations with E-proteins and/or haematopoietic factors, LYL1 also interacts with other proteins such as CREB1. It has been shown that LYL1 and CREB1 form protein complexes on the ID1 promoter via interaction between N-terminal domain of LYL1 and the Q2/KID domain of CREB1 (San-Marina et al., 2008). Moreover, the LYL1-CREB1 complexes have also been recently identified at the promoter of STMN1 using ChIP-chip analysis (San-Marina et al., 2012). Further shRNA knockdown analyses have demonstrated that LYL1 and CREB1

play critical roles in regulating STMN1 expression. As the STMN1 gene product is a vital regulator of cell cycle and cellular motility, the results suggest that modulating STMN1 expression via disruption of LYL1-CREB1 complex may be beneficial for regulating the proliferation of leukemic cells.

7.1.5 LYL1, a paralogue of TAL1

The LYL1 protein shares 82% amino acid identity with TAL1 in the critical bHLH domains which includes the conserved domains important for interaction with LMO2 (Wadman et al., 1994), suggesting that these two protein share at least part of their target genes as well as some biological functions (Porcher et al., 1999; Schlaeger et al., 2004).

In fact, like TAL1, LYL1 is required for HSC maintenance (Souroullas et al., 2009). LYL1 and TAL1 also display overlapping expression patterns across several haematopoietic lineages as well as developing endothelial cells (Chapman et al., 2003; Visvader et al., 1991). In addition, both LYL1 and TAL1 are found to interact with the lim-only-domain leukemia oncogenes LMO1 and LMO2 (Wadman et al., 1994). Similar to TAL1, LYL1 is also regulated by Ets and GATA family transcription factors in endothelial, haematopoietic progenitor and megakaryocytic cells (Chan et al., 2007). However, functions of LYL1 and TAL1 do also differ in some aspects. It has been found that TAL1 expression is initiated prior to LYL1 during early haemangioblast specification. In addition, the forced expression of LYL1 in ES cells is unable to rescue the haematopoietic defect of $Tal1^{-/-}$ ES cells, suggesting a unique role of TAL1 in development of HSCs (Chan et al., 2007). TAL1 and LYL1 also differ in that the latter is not expressed in the central nervous system and is more ubiquitously expressed across the adult haematopoietic sub-compartments (Giroux et al., 2007) – thus resulting in adult haematopoietic defects in LYL1 null alleles (Capron et al., 2011; Capron et al., 2006). Taken together, these studies reveal that LYL1 and TAL1 share overlapping functions (i.e., partial functional redundancy) to some degree during haematopoiesis.

Given these striking parallels between TAL1 and LYL1, it can be hypothesized that these two gene loci may be under similar regulatory or structural control. To support this, the structural and regulatory similarities between two genes are evident in the following sections.

7.1.5.1 Sequence conservation and motif arrangement of the LYL1 promoter is similar to the TAL1 promoter 1a

Both LYL1 and TAL1 contain two core-promoters, termed as LYL1 P1/P2 and TAL1 P1a/P1b respectively. For the LYL1 promoters, P1 contains two conserved ETS family binding sites while P2 contains three conserved ETS sites and two GATA binding sties (Chan et al., 2007). In addition, both genes have shown to be co-regulated by GATA and ETS factors (Chan et al., 2007; Gottgens et al., 2004; Gottgens et al., 2002). The LYL1 promoter (a 464 bp fragment containing both P1 and P2) has been shown to be highly-conserved across five species (human, dog, cow, mouse and rat) using competitive sequence analysis (Chapman et al., 2003). Intriguingly, it has been shown that the arrangement of the two GATA binding motifs located at LYL1 P2 highly resembles the pattern at TAL1 promoter 1a (Figure 7.1a). In addition, this arrangement has also been found to be conserved in both TAL1 and LYL1 promoters across vertebrates (Figure 7.1b), implying that it may have been existed in the promoter of the common ancestral gene which is thought to have given rise to both TAL1 and LYL1 via duplication (Chapman et al., 2003).

Figure 7.1: Sequence alignment of TAL1 and LYL1 promoters. (A) Alignment of a portion of the TAL1 promoter 1a with a fragment of the LYL1 promoter proximal region. (B) Competitive sequence alignment between vertebrate organisms. Blocks containing the two conserved GATA motifs on the antisense strands are highlighted in red.

This LYL1 promoter fragment is sufficient to drive LYL1 expression in transgenic mouse embryos in developing endothelial and haematopoietic cells (Chan et al., 2007). In addition, erythroid-specific LYL1 expression is modulated by GATA

factors via conserved GATA motifs located at the LYL1 promoter. During erythroid maturation, the binding of GATA2 activates LYL1 expression whereas GATA1 binding results in LYL1 repression.

7.1.5.2 Similarity of TF motifs and binding patterns of the LYL1 +33 putative enhancer with the TAL1 +51 enhancer

Previous ChIP-chip studies have identified the LYL1 gene as a target of four (GATA1, TAL1, E2A and LDB1) out of five TEC members (H.L. Jim's PhD thesis, University of Cambridge, 2008). In addition, it has been demonstrated that the siRNA knockdown of each TEC member can affect the expression of LYL1, which further suggests a role for the TEC in its regulation. However, the promoter of LYL1 has no GATA/E-box composite motif, which is unable to support its role in recruiting the TEC. However, it has been speculated that a *cis*-acting element may have canonical GATA/E-box motifs which mediating the binding of the TEC, similar to what has been found at the TAL1 +51 enhancer (H.L. Jim's PhD thesis, as shown in Figure 7.2).

Figure 7.2: Multiple sequence alignments of the (putative) enhancers of TAL1 and LYL1. E-box and GATA motifs were identified by TESS and TFSearch and by viewing the conserved TFBS track on the UCSC genome browser. Multiple species sequence alignments were taken from the UCSC genome browser. Red boxes indicate the conserved nucleotides across species in the E-box or GATA motifs.

Subsequently, an element located at approximately 33 kb downstream of the LYL1 promoter had been identified, containing a highly conserved GATA/E-box motif with a consensus 8-bp spacing (Figure 7.2). Previous ChIP-qPCR studies have also confirmed the occupancy of the TEC members including GATA1, E2A/TCF3 (isoforms E12 and E47), TAL1 and LDB1 at the LYL1 +33 enhancer in human erythroid K562 (Figure 7.3) and HEL cell lines. These previous studies failed to

demonstrate the occupancy of LMO2 at this element due to a lack of a good quality ChIP-grade antibody for ChIP analysis (H.L. Jim's PhD thesis, 2008).

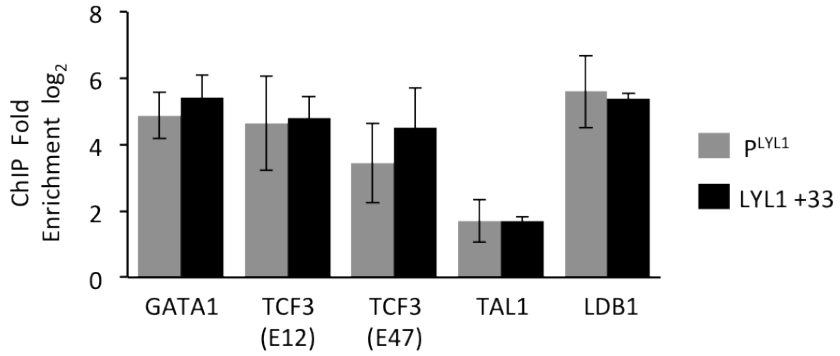


Figure 7.3: ChIP enrichments (log_2) with standard errors for members of the TAL1 erythroid complex (TEC) at the LYL1 +33 elements (black bars) as well as LYL1 promoter (P^{LYL1}) (grey bars) in K562 cells.

The LYL1 +33 element is located within an intron of a neighbouring gene (NFIX), which may create difficulties with it communicating with the LYL1 promoter (Figure 7.2). In addition, there are also several CTCF binding sites situated between the +33 element and the LYL1 promoter (Figure 7.4). This arrangement is similar to the TAL1 +51 enhancer, which is separated from its cognate promoters by the PDZK1IP1 (MAP17) gene and a CTCF binding element. Taking together, it implies that the LYL1 locus may adopt a similar looping mechanism observed at the TAL1 locus, which allows the LYL1 +33 putative enhancer to communicate with its promoter.

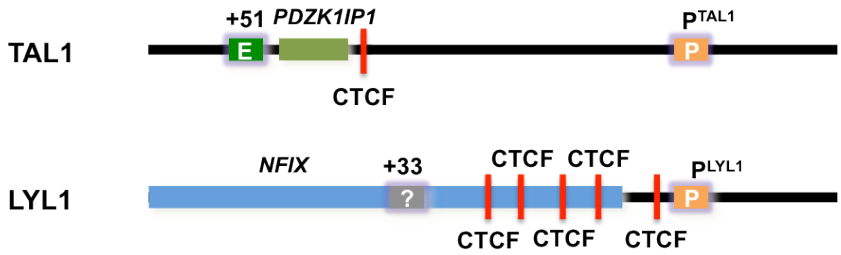


Figure 7.4: Schematics map shows structures of the TAL1 and LYL1 genes. "E" = enhancer, "P" = promoter and "?" = putative enhancer. The green and blue boxes represent the PDZK1IP1 and NFIX genes located in between the promoters and (putative) enhancers of TAL1 and LYL1. The red bars represent CTCF insulators located between the (putative) enhancers and cognate promoters.

A number of intriguing similarities between LYL1 and TAL1 have been illustrated, from sequence conservation of TF motifs in regulatory elements to the binding patterns of the TEC complex to their regulatory elements (Table 7.1). Given that the transcriptional regulation of the TAL1 locus is GATA1/TEC-dependent (see

Chapter 6), it would be interesting to explore whether the LYL1 locus is also regulated through a similar mechanism – and via looping.

Table 7.1: Structure similarity between the TAL1 and LYL1 loci

Similarity	TAL1	LYL1
A gene located between (putative) enhancer and its cognate promoter	PDZK1IP1	NFIX
CTCF binding site(s) between (putative) enhancer and its cognate promoter	+40 insulator	Five CTCF binding sites
Conserved TF binding motifs at the promoter	GATA	GATA
Conserved TF binding motifs at the (putative) enhancer (TAL1 +51/LYL1 +33)	Two GATA/E-box	One GATA/E-box
TF occupancy over the (putative) enhancer and its cognate promoter	TEC	TEC

7.2 Aims of the chapter

The overall aim of the work presented in this Chapter was to determine whether there were functional similarities between the TAL1 and LYL1 loci in how they are regulated using chromatin looping. To these ends, the aims of this chapter are as follows:

1. To determine the enhancer activity of the LYL1 +33 element by reporter assays

2. To determine chromatin looping interactions between the LYL1 promoter and the +33 element by 3C-PCR

3. To determine whether GATA1 was required for the transcription regulation of the LYL1 gene by GATA1 siRNA knockdown

4. To determine whether this regulation was mediated through the loss of the TEC and the loss of chromatin looping.

5. To assess whether a novel region designated LYL1 +24 was also involved in LYL1 regulation via chromatin looping.

7.3 Overall strategy

As illustrated in Figure 7.5, three types of experiments were conducted to study the involvement of chromatin looping at the LYL1 locus. First, enhancer activity of the LYL1 +33 was studied *in vitro* using luciferase reporter assays. Second, computational analysis was performed to further characterise the similarity of sequence conservation and TF binding between the TAL1 +51 and the LYL1 +33 elements. Third, looping interaction involving the *LYL1* promoters and the +33 enhancer was identified by 3C-PCR assays. Forth, a GATA1 knockdown was used to determine whether it resulted in alterations of LYL1 expression, the binding of the TEC and looping interactions between the LYL1 promoters and the +33 element. In addition, a similar analysis was also performed for a novel region, called "LYL1 +24" as described in section 7.4.5.

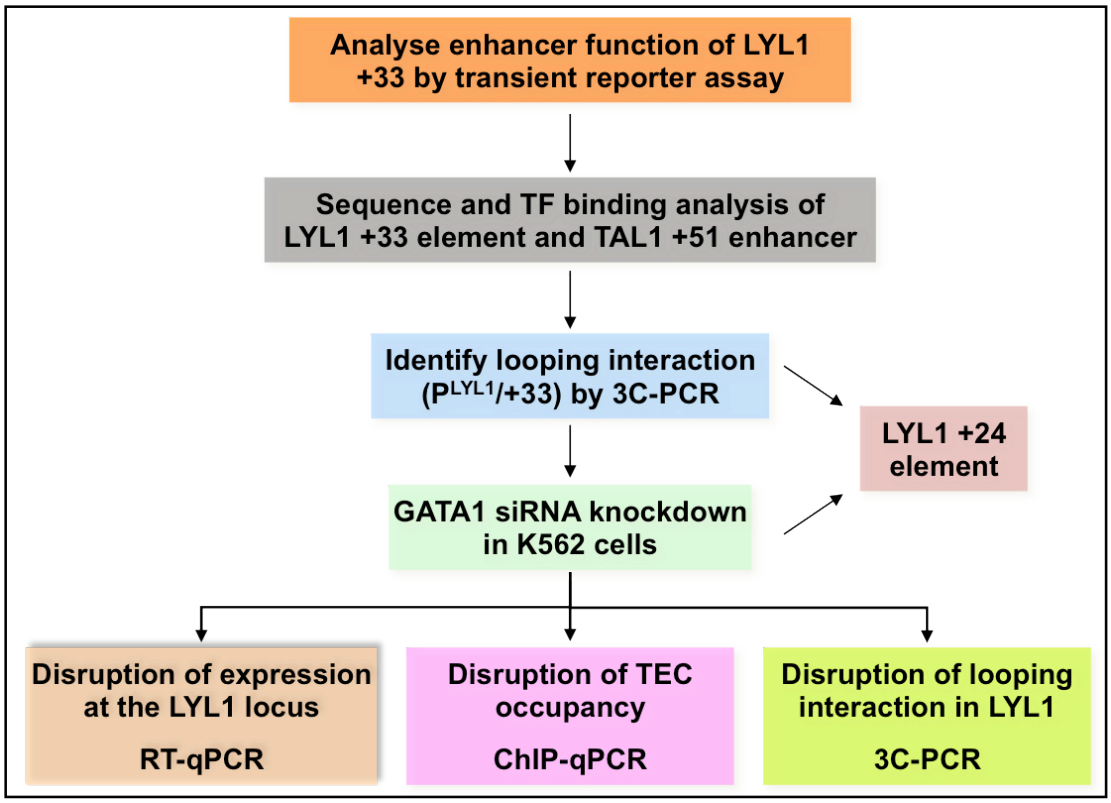


Figure 7.5: Overall strategy for analysing the transcription regulation of the LYL1 locus. Step 1: determination of the enhancer activity of the LYL1 +33 element using enhancer trap reporter assay. Step 2: further analyze the similarity of sequence conservation and TF binding between TAL1 +51 and LYL1 +33. Step 3: determination of the looping interaction between the LYL1 promoters and the +33 element by 3C-PCR assays. Step 4: determination of the effects of GATA1 depletion on the transcription regulation of the LYL1 locus. A similar strategy was employed to also study a novel region – LYL1 +24 (see section 7.4.5).

Results

7.4 Determining enhancer activity of the LYL1 +33 element by transient reporter assays

A 2596 bp fragment spanning the human LYL1 +33 putative enhancer element was cloned into a luciferase reporter gene vector under the control of the SV40 promoter. The vector containing LYL1 +33 element was tested along with another two control vectors, which contained i) a human DNA sequence with no regulatory function (negative control) and ii) the human TAL1 +51 enhancer (positive control) in human erythroid K562 cells. Relative enhancer activities of three constructs were determined as the ratio of luciferase activity to β-galactosidase activity as illustrated in Figure 7.6. In comparison with the negative control region, both LYL1 +33 and TAL1 +51 elements were able to increase luciferase expression to a similar levels under the control of the SV40 promoters in K562 cells, suggesting the LYL1 +33 had enhancer activity *in vitro*.

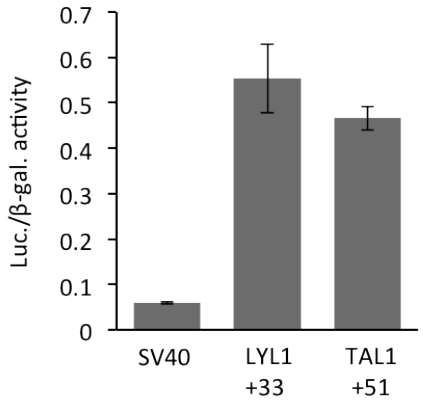


Figure 7.6: Transient reporter assays used to determine enhancer activity of the LYL1 +33 element. The y-axis is the enhancer activity measured as the luciferase reporter activity relative to β-galactosidase activity (the latter used as a transfection control). The three constructs tested were (i) luciferase under the control of the SV40 promoter with a human DNA sequence with no regulatory function cloned into the enhancer cloning site (SV40), (ii) luciferase under the control of the SV40 promoter with the human +33 element cloned into the enhancer cloning site (LYL1 +33), (iii) luciferase under the control of the SV40 promoter with the human TAL1 erythroid enhancer cloned into the enhancer cloning site (TAL1 +51).

7.5 Computational analysis of the LYL1 +33 element using public data

7.5.1 Comparative sequence analysis of the TAL1 +51 and LYL1+33 elements

Further comparative sequence analysis revealed that the GATA/E-box binding motifs are highly conserved across the vertebrates, which can be traced all the

way back to chicken for the TAL1 +51 and to lizard for the LYL1 +33 enhancers (Figure 7.7). These two vertebrate species diverged from human over 300 million years ago (further details see section 7.6.2). In TAL1 +51 enhancer, the spacing between the GATA and E-box at the 3' and 5' composite motifs is 6 bp and 9 bp (consensus) respectively, whereas the LYL1 +33 enhancer has the consensus 9 bp spacing (Figure 7.7). This high sequence conservation of the GATA/E-box motif provides further indirect evident that the LYL1 +33 element was very likely to function as an enhancer.

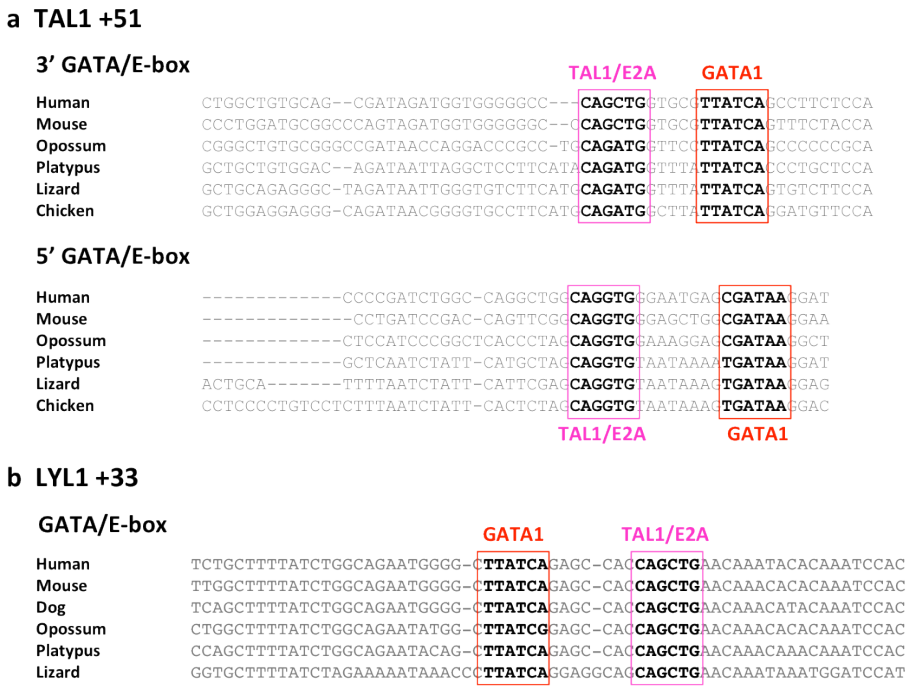


Figure 7.7: DNA sequence conservation of GATA/E-box motifs at the TAL1 +51 enhancer and LYL1 +33 enhancer across six vertebrate species. Gaps in alignments are shown by dashed lines (--). E-box/GATA composite motifs are shown in bold and highlighted by red and pink boxes respectively.

7.5.2 Occupancy of GATA1 and TAL1 transcription factors at the TAL1 and LYL1 loci

TAL1 and GATA1 ChIP-seq binding profiles for TAL1 and LYL1 were analysed based on public available datasets (ENCODE project) (Figure 7.8). The binding profiles of these two TFs were compared across the promoters of TAL1 and LYL1, the TAL1 +51 enhancer and the LYL1 +33 element. It was found that the TAL1 and GATA1 binding profiles were different at those four target sites. Although the GATA/E-box at 3' end of the +51 enhancer lacks the normal spacing in comparison with the consensus sequence of the TEC binding motif as previously discussed in Chapter 6, levels of GATA1 and TAL1 binding appeared to favour this motif as the peak of binding resides closer to it than to the 5' GATA/E-box motif

(Figure 7.8a) – although this is difficult to discriminate given that the two GATA/E-box motifs are very close to one another and the read densities observed are cumulative for both motifs. The levels of GATA and TAL1 occupancies at the LYL1 +33 element were much lower than at the TAL1 +51 enhancer. This could probably due to the fact that the LYL1 +33 only had one GATA/E-box motif (Figure 7.8 b, left panel) while the TAL1 +51 has two GATA/E-box motifs (Figure 7.8 a, left panel). In contrast, it was observed that the levels of GATA1 and TAL1 occupancies at the TAL1 and LYL1 promoters were very similar, as both promoters had two GATA binding motifs (Figure 7.8 a & b, right panels). Furthermore, the binding patterns of GATA1 and TAL1 at the promoters of TAL1 and LYL1 were independent from the levels of GATA1 and TAL1 at their enhancers. These data suggested that although similar motifs were present in the regulatory elements of both the TAL1 and LYL1 loci, the levels of GATA1 and TAL1 occupancies as well as the machineries involved in TAL1 and LYL1 transcriptional regulation might be slightly different.

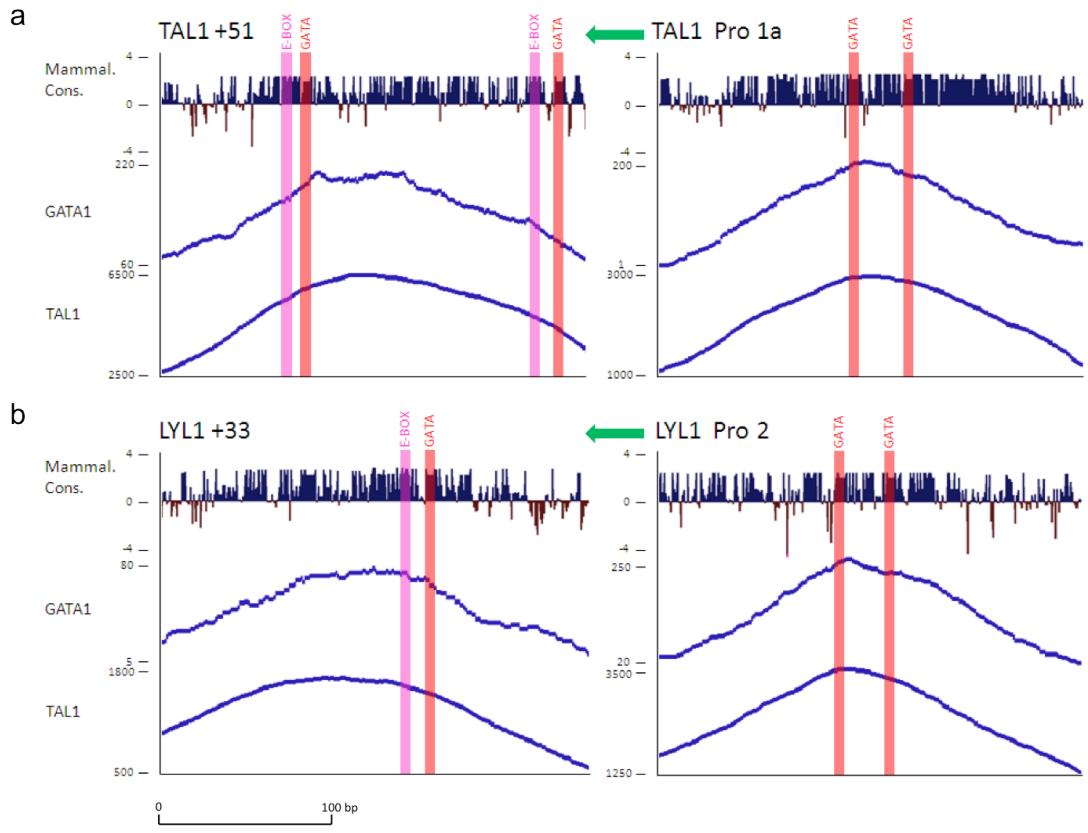


Figure 7.8: Occupancy of TAL1 and GATA1 at the TAL1 and LYL1 loci. (A) Occupancy at the +51 erythroid enhancer and at TAL1 promoter 1a. (B) Occupancy at the +33 enhancer and the LYL1 promoter. At the top of each panel is the level of placental mammalian DNA sequence conservation (scores based on phyloP data taken from UCSC genome browser). The location of GATA and E-box motifs are highlighted in red and pink respectively. The directions of transcription for TAL1 and LYL1 with respect to the orientation of the +51 and +33 elements respectively are shown by the green arrows. Blue line graphs show the frequency of individual bases (y axes) in ChIP-seq reads for GATA1 and TAL1 obtained from publicly available ENCODE datasets. Scale (in bp) is shown at the bottom.

7.6 Determine looping interaction between the +33 enhancer and promoter of the LYL1 gene

In order to provide direct evidence that the LYL1+33 element was involved in LYL1 regulation, 3C-PCR assays were performed using the LYL1 promoter fragment as the anchor to determine whether it was in close physical contact with the LYL1 +33 region via chromatin looping. Control 3C assays were designed for two regions flanking LYL1 +33: LYL1 +45 and LYL1 +10, respectively. The 3C-PCR assays were conducted as described in Chapter 3 of this thesis. As shown in Figure 7.9, the relative ligation frequency was significantly increased (p = 4.6 x 10^{-3}, student's T-test) at the LYL1 +33 element in comparison with the +10 region (which is only 10 kb away from the LYL1 promoter), indicating that the LYL1 +33 enhancer interacted with the LYL1 promoter at a level significantly above levels expected by random ligation events in K562 cells. This data, together with the occupancy patterns of the TEC (at both the LYL1 promoters and the +33 enhancer - discussed above), and the results of enhancer trap reporter assays, provided three lines of evidence that there indeed were structural and functional similarities between the TAL1 +51 erythroid enhancer and the LYL1 +33 element. These data also suggest that LYL1 +33 enhancer may have a direct role in LYL1 transcriptional regulation.

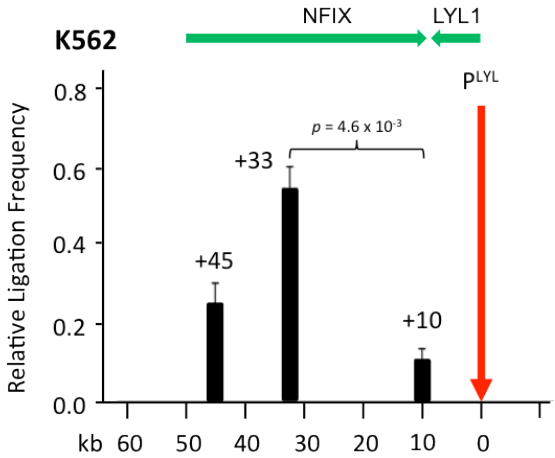


Figure 7.9: Looping interactions between the LYL1 promoter (P^{LYL1}) and the LYL1 +33 element in K562 cells. Bar diagram of interaction pattern in K562 cell lines determined by 3C. Interaction frequencies (black bars) at three locations across the locus are shown with standard errors and normalized relative to BAC controls. The location of the 3C "anchor" (P^{LYL1}) is denoted with the vertical red arrow. The locations of the LYL1 and NFIX genes and their directions of transcription are shown at the top of each panel. The *p* values are indicated for interaction frequencies which are significantly higher for test regions when compared to those of control regions (controls defined as regions located between the "anchor" and test regions). Scales (in kb) are shown at the bottom of each panel.

7.7 Loss of GATA1 occupancy at the LYL1 locus: Similar consequences to that observed at the TAL1 regulon

As showed in Chapter 6, the transcription and chromatin configuration of the TAL1 locus were significantly disrupted due to the loss of the TEC after siRNA knockdown of GATA1 for 96 hours in K562 cells. Given that LYL1 was also known to be a target gene of GATA1 and the TEC (H.L.Jim's PhD thesis, University of Cambridge, 2008), it was speculated that LYL1 expression and looping interactions between the LYL1 promoter and the +33 enhancer might also be disrupted by GATA1 knockdown. All the experiments were performed in parallel with the studies for the TAL1 locus, using same controls and knockdown samples as previously described in Chapter 6.

7.7.1 Depletion of GATA1 affects expression at the LYL1 locus

As shown in Chapter 6, it was demonstrated that the expression level of TAL1 and it neighbouring genes were down-regulated by GATA1 knockdown in K562 cells. In parallel, the expression of LYL1 along with its neighbouring genes, NFIX (downstream of LYL1) and TRMT1 (upstream of LYL1) were also assessed in K562 cells 96 hour post-transfection with luciferase and GATA1 siRNAs.

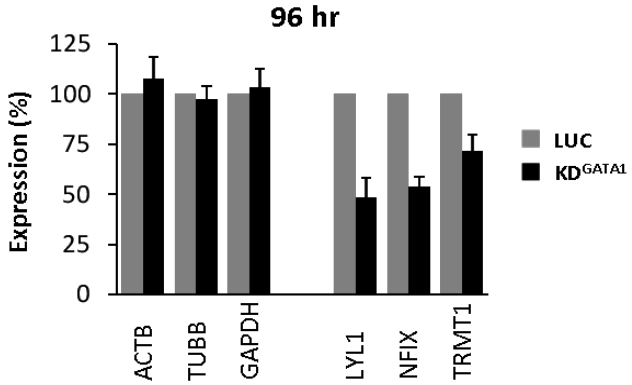


Figure 7.10: Effect of GATA1 knockdown on gene expression at the LYL1 locus. LYL1, NIFX and TRMT1 transcript levels (with standard errors) in GATA1 knockdown samples are shown relative to levels in the luciferase control samples 96 hours after transfection with siRNA for GATA1 (KD^{GATA1}) or luciferase (LUC). ACTB, TUBB and GAPDH were used as gene expression controls (also shown in Figure 6.10).

A significant reduction (over 50%) of LYL1 expression was observed in GATA1 knockdown K562 cells in comparison to luciferase control (Figure 7.10). Similarly, expression level of the NFIX and TRMT1 genes were also significantly down regulated to 53.8 % and 71.8% in GATA1 knockdown K562 cells, comparing to the level of expressions in the luciferase control (Figure 7.10).

The loss of GATA1 had a direct effect on LYL1 expression in agreement with previously results, implying that LYL1 may be under the regulation of GATA1 (H.J. Jim's PhD thesis, University of Cambridge, 2008). More importantly, GATA1 knockdown also had significant effects on expression of neighbouring genes of both TAL1 and LYL1, suggesting that not only TAL1 and LYL1 were dependent on GATA1 for expression, but that their neighbouring genes were also dependent on GATA1. As was shown in Chapter 6 for TAL1, the "neighbouring gene effect" at the LYL1 locus suggested that the flanking genes were dependent on LYL1 expression, to some degree, for their own expression.

7.7.2 Depletion of GATA1 affects the RNAP II occupancy

As shown in Chapter 6, TAL1 expression and RNA pol II (Pol II) recruitment to the TAL1 chromatin hub were decreased due to the loss of GATA1 during knockdown. Similarly, as LYL1 transcription was also down regulated during GATA1 knockdown (as shown above), it was necessary to determine whether the recruitment of RNA pol II was also affected by GATA1 depletion. ChIP analyses were performed with RNA pol II antibody as described previously in chapter 6. The qPCR assays were conducted with the primers amplifying the LYL1 promoter and the +33 enhancer as well as negative and positive controls.

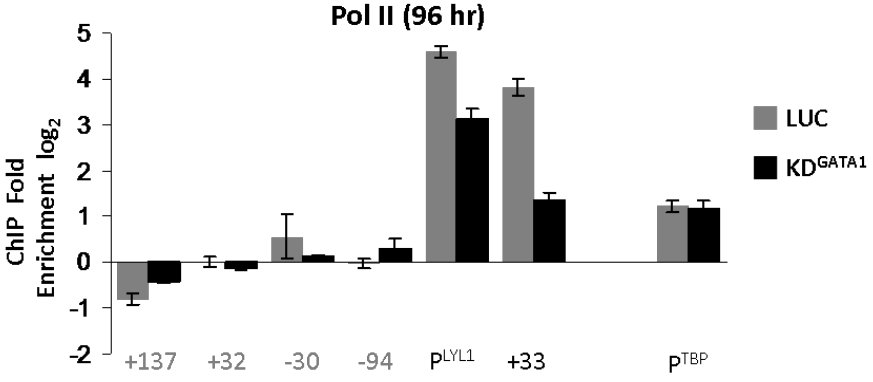


Figure 7.11: Effect of GATA1 knockdown on recruitment Pol II at the LYL1 locus. Pol II recruitment at the LYL1 promoter (P^{LYL1}) and the +33 enhancer 96 hours after transfection with siRNA for GATA1 (KD^{GATA1}) or luciferase (LUC) was determined by ChIP-qPCR. Positive control for Pol II occupancy was the promoter of the TBP gene (P^{TBP}). ChIP enrichments (log_2) are shown with standard errors. Annotation of test and negative control regions is denoted in black and grey text respectively. Luciferase treated samples are shown as grey bars, KD^{GATA1} treated samples are shown as black bars. All the negative control regions was previously used in studying the TAL1 locus (as described in Figure 6.11)

Significant Pol II occupancy was detected at the LYL1 promoter and the +33 enhancer in luciferase control K562 cells. However, in GATA1 knockdown cells, ChIP enrichments of Pol II had significantly declined more than 60% ($p = 2.2 \times 10^{-4}$)

and 80% (3.1 x 10^{-3}) over the LYL1 promoter and +33 enhancer respectively (Figure 7.11 and Table 7.2). Thus, loss of GATA1 during knockdown resulted in loss of Pol II from the Lyl1 locus, providing yet another line of evidence of the similarities between TAL1 regulation and LYL1 regulation.

Table 7.2 Percentages of RNA polymerase II (Pol II) protein lost from the binding target sites at 96 hour after GATA1 knockdown.

	Pol II KD^{GATA1} 96 hr	
	KD[1] %	***p-value*[2]**
P^{LYL1}	63.5 %	2.2 x 10^{-4}
+33	82.1 %	3.1 x 10^{-3}

[1]**Percentage of depletion of PolII occupancy is calculated based on PolII ChIP enrichment of GATA1 siRNA knockdown samples versus luciferase siRNA controls.**
[2]**P-value is calculated based on student T-test (2-tails)**

7.7.3 GATA1 occupancies at the LYL1 locus after GATA1 siRNA knockdown

The levels of GATA1 occupancy at the LYL1 locus were analysed by ChIP-qPCR assays from ChIP material obtained from K562 cells, 48 hour and 96 hour after transfection with GATA1 and luciferase siRNAs. The q-PCR assays were conducted with primers amplifying the LYL1 promoter and the +33 enhancer along with four negative control regions from the TAL1 locus which are not bound by GATA1 (controls taken from Dhami et al., 2010).

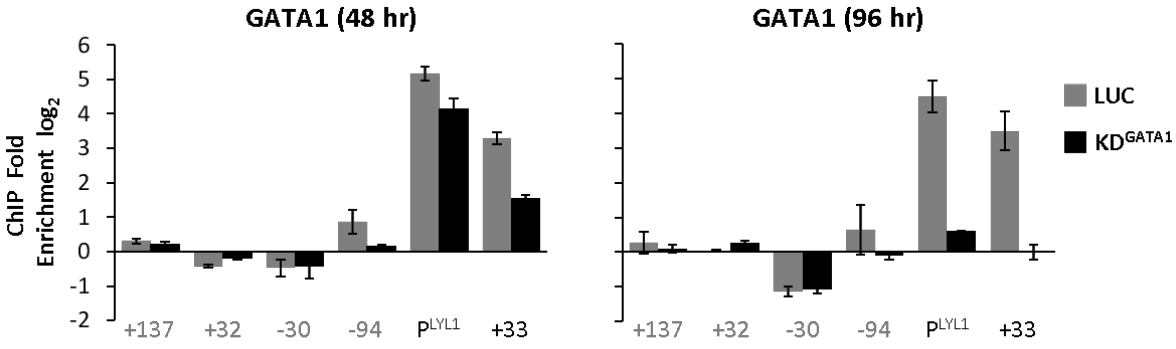


Figure 7.12: GATA1 occupancy at the LYL1 promoter (P^{LYL1}) and LYL1 +33 element, 48 and 96 hours after transfection with siRNAs for GATA1 (KD^{GATA1}) or luciferase (LUC). ChIP enrichments (log_2) are shown with standard errors. Annotation of test and negative control regions is denoted in black and grey text respectively. Luciferase treated samples are shown as grey bars, KD^{GATA1} treated samples are shown as black bars. All the negative control regions was previously used in studying the TAL1 locus (as described in Figure 6.7).

High levels of binding of GATA1 was detected at the LYL1 promoter and at the +33 enhancer in the luciferase control K562 cells at both 48 hour and 96 hour time-points (Figure 7.12). However, GATA1 occupancy over the LYL1 promoter and its +33 enhancer were significantly reduced at 48 and 96 hours after

transfection with GATA1 siRNA. At the 48-hour time-point, GATA1 occupancies were significantly down to 50.2% (p = 4.7 x 10^{-3}) and 29.6% (p = 4.1 x 10^{-4}) at the LYL1 promoter and at the +33 enhancer, respectively (Figure 7.12 GATA1 48hr and Table 7.3). The most dramatic reductions of GATA1 occupancy at these regions were achieved at the 96-hour time-point. Compared to luciferase controls, over 90% of GATA1 was significantly removed at both the LYL1 promoter (6.5% remained, p = 5.8 x 10^{-4}) and the +33 enhancer (8.4% remained, p = 5.4 x 10^{-4}) (Figure 7.12 GATA1 96hr and Table 7.3). These data mirrored the GATA1 clearance kinetics observed at the TAL1 promoter 1a and the +51 erythroid enhancer (see Chapter 6, section 6.5).

Table 7.3 Percentages of GATA1 protein lost from the binding target sites at 96 hour after GATA1 knockdown.

	KD^{GATA1} 48 hr		KD^{GATA1} 96 hr	
	KD[1] %	*p-value*[2]	KD[1] %	*p-value*[2]
P^{LYL1}	49.8 %	4.7 x 10^{-3}	93.5 %	5.8 x10^{-4}
+33	70.4 %	4.1 x 10^{-4}	91.6 %	5.4 x10^{-4}

[1]**Percentage of depletion of GATA1 occupancy is calculated based on GATA1 ChIP enrichment of GATA1 siRNA knockdown samples versus luciferase siRNA controls.**
[2]**P-value is calculated based on student T-test (2-tails)**

7.7.4 Depletion of GATA1 results in the loss of members of the TAL1 erythroid complex at the LYL1 locus.

As previously described, ChIP occupancy of two TEC members, LDB1 and E2A/TCF3 was also assessed for the LYL1 promoter and the +33 enhancer in K562 cell after 96 hour of GATA1 siRNA knockdown and luciferase control (Figure 7.13). ChIP enrichments of LDB1 and E2A/TCF3 (E47) were observed at the LYL1 promoter and at the +33 enhancer in luciferase control K562 cells, which were in agreement with that reported by others in our laboratory using K562 cells (H.L.Jim, PhD Thesis, 2010).

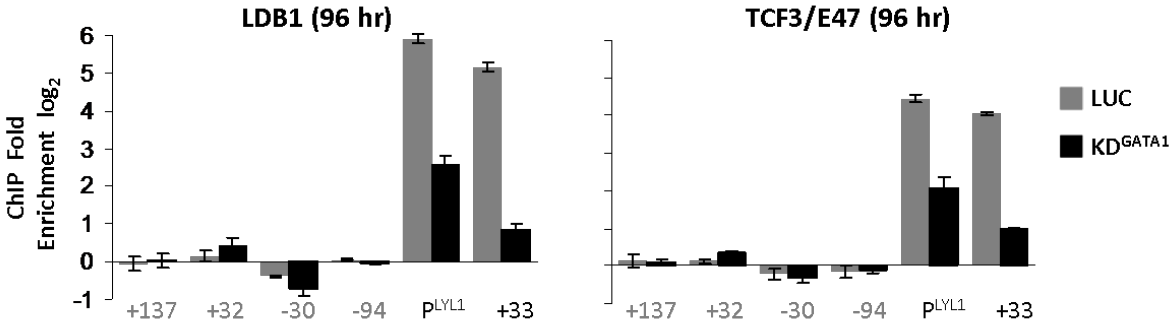


Figure 7.13: LDB1 and E2A/TCF3 (E47) recruitments at the LYL1 enhancer (+33) and LYL1 promoter (P^{LYL1}) 96 hours after transfection with siRNA for GATA1 (KDGATA1) or luciferase

(LUC). ChIP enrichments (log2) are shown with standard errors. Annotation of test and negative control regions is denoted in black and grey text respectively. Luciferase treated samples are shown as grey bars, KD^{GATA1} treated samples are shown as black bars. Same negative control regions were used as previously described in Chapter 6.

Both of these members of the TEC were significantly affected by GATA1 knockdown at the LYL1 locus. In comparison with luciferase control, ChIP enrichments for LDB1 were less than 10% ($p = 7.6 \times 10^{-7}$) and 4.7% ($p = 2.2 \times 10^{-7}$) at the LYL1 promoter and the +33 enhancer respectively (Figure 7.13, LDB1 96hr; Table 7.4). Similarly, occupancy of E2A/TCF3 (E47) was only 15% (LYL1 promoter; $p = 3.0 \times 10^{-4}$) and 9.4% (+33; $p = 9.2 \times 10^{-8}$) of the levels detected in the control (Figure 7.13, E2A/TCF3 (E47) 96hr and Table 7.4). Although it had previously been shown that both of these genes are down-regulated at the mRNA level as a result of GATA1 knockdown (Chapter 6), the level of loss of these members of the TEC at the LYL1 locus was far greater than their level of down-regulation. Thus, it was likely that the loss of GATA1 had a direct effect on the recruitment of other members of the TEC at the LYL1 locus. However, assessment of the levels of the proteins for these two transcription factors during GATA1 knockdown would help determine whether their loss of occupancy at LYL1 was due to (i) their inability to be recruited in the absence of GATA1, or (ii) due to their down-regulation by GATA1. As stated in Chapter 6, these confirmatory experiments were not performed here. Irrespective of this last point, these data are consistent with that observed at the TAL1 locus – again providing further evidence of the functional similarities between LYL1 and TAL1 in their regulatory mechanisms.

Table 7.4 Percentages of LDB1 and E2A/TCF3 (E47) proteins lost from the binding target sites at 96 hour after GATA1 knockdown.

	LDB1 KD^{GATA1} 96 hr		E2A/TCF3 (E47) KD^{GATA1} 96 hr	
	KD[1] %	*p-value*[2]	KD[1] %	*p-value*[2]
P^{LYL1}	90.9 %	7.6×10^{-7}	85.0 %	3.0×10^{-4}
+33	95.3 %	2.2×10^{-7}	90.6 %	9.2×10^{-6}

[1]**Percentage of depletion of LDB1 or E2A/TCF3 (E47) occupancy is calculated based on LDB1 or E2A/TCF3 (E47) ChIP enrichment of GATA1 siRNA knockdown samples versus Luciferase siRNA controls.**
[2]**P-value is calculated based on student T-test (2-tails)**

7.7.5 Depletion of GATA1 results in loss of looping interactions at the LYL1 locus

Based on the results obtained at the TAL1 locus, it was hypothesized that loss of GATA1 during knockdown would also effect the chromatin looping between the

LYL1 promoter and the +33 enhancer. 3C-PCR assays were performed using the luciferase and GATA1 knockdown K562 cells at 48 hour and 96 hour after siRNA transfection. As described in section 7.4.3, the LYL1 promoter was used as the anchor for these 3C assays along with the appropriate test and control regions.

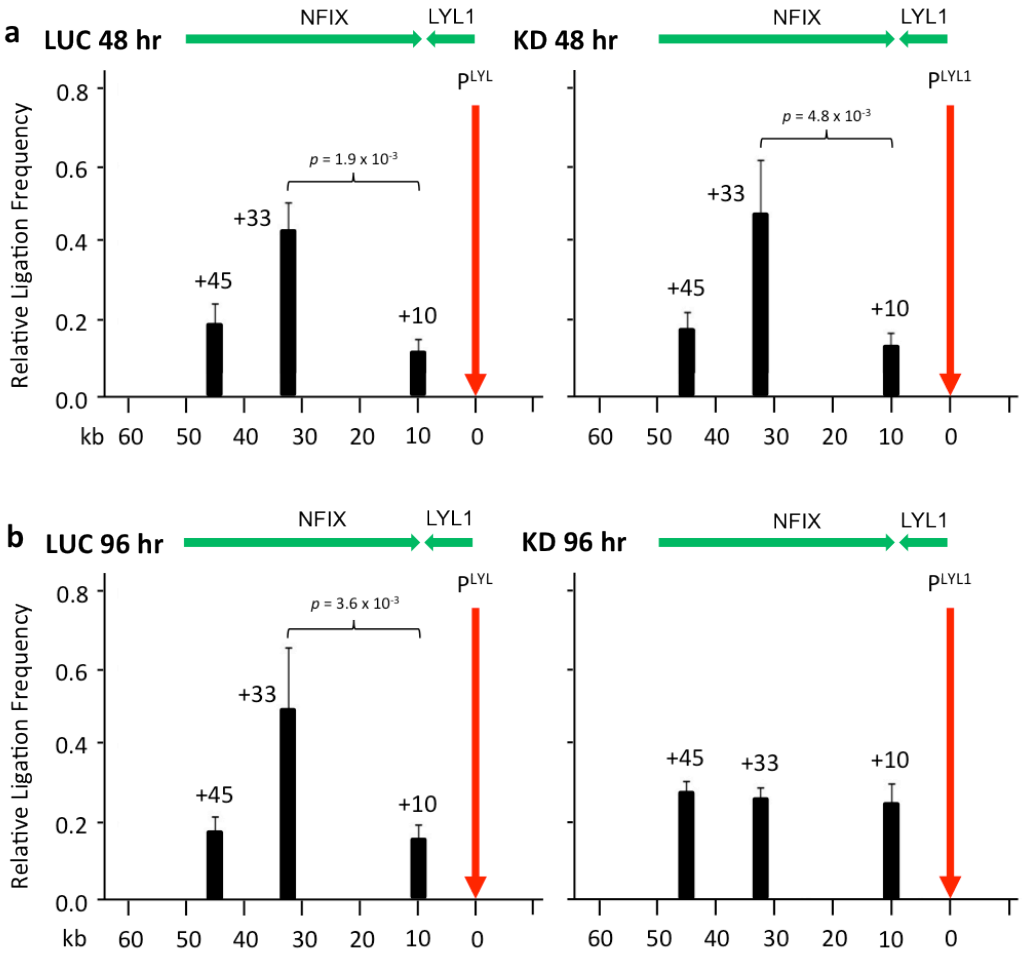


Figure 7.14: GATA1 knockdown results in loss of looping interactions between the +33 enhancer and the LYL1 promoter (P^{LYL1}). Bar diagrams of interaction patterns across the human LYL1 locus after siRNA transfection with GATA1 (KD) or with luciferase (LUC) was determined by 3C. (A) Interactions 48 hours after siRNA transfection. (B) Interactions 96 hours after siRNA transfection. Interaction frequencies (black bars) are shown with standard errors and normalized relative to BAC controls. The location of 3C "anchor" is denoted by vertical red arrow. Locations of genes and their directions of transcription are shown at the top of each panel. *p* values are indicated for interaction frequencies which are significantly higher for test regions when compared to those of control regions. Scales (in kb) are shown at the bottom of each panel.

Figure 7.14 shows the interaction results obtained in the GATA1 knockdown study using 3C-PCR. The luciferase controls consistently showed significant interaction frequencies above background between the LYL1 promoter and the +33 enhancer at both the 48 and 96-hour time-points. These data were in agreement with the interaction profiles obtained by 3C-PCR in the wild-type K562 cell (see section 7.4). However, in the GATA1 knockdown cells, the 3C interaction profiles between the LYL1 promoter and the +33 enhancer showed remarkable similarities to the

interactions between the TAL1 promoters and the +51 erythroid enhancer. While LYL1 looping interactions appeared to be unaffected at the 48 hour time-point in GATA1 knockdown cells (Figure 7.14, panel A), the level of the P^{LYL1}/+33 interaction after 96 hours of GATA1 knockdown was no longer significant above background (Figure 7.14 panel B). Although GATA1 occupancy at the LYL1 promoter and the +33 enhancer was significantly decreased at the 48 hour time point, the complete depletion of the TEC from its target sites was not achieved until 96 hour time point. Similar to as previously shown in the TAL1 locus during GATA1 knockdown (see Chapter 6, section 6.4), it suggested that looping interaction between P^{LYL1} and +33 was not fully abolished until the complete knockdown of GATA1 after 96 hours. These data confirmed that GATA1 is also required for mediating looping interactions between the LYL1 +33 enhancer and the promoter of LYL1, and that the LYL1 locus shows similar kinetics to analogous looping interactions found at the TAL1 locus. This data, taken together with the clearance of the TEC from the LYL1 locus during GATA1 knockdown, would further suggest that the TEC is mediating the interaction between the LYL1 promoter and the +33 enhancer.

7.8 Assessment of the LYL1 +24 element

As described in previous section, a TEC-dependent chromatin looping interaction between the LYL1 promoter and the +33 enhancer was identified, implying similarity in chromatin structure between the TAL1 and LYL1 loci. As a result of that, it was speculated that the LYL1 locus might contain an additional element functioning as the equivalent to the TAL1 +20/+19 stem cell enhancer. A genomic region located 24 kb downstream of the LYL1 promoter (i.e., the +24 element) was identified based on its transcription factor binding patterns and histone modification profiles. As shown in Figure 7.15, the +24 element is the only genomic region between the +33 enhancer and the LYL1 promoter that contains majority of the enhancer hallmarks observed at the +33 enhancer. For example, the RNAP II and GATA1 occupancies were observed at the +24 element. In addition, histone modifications of H3K4me1 and H3K27ac, hallmarks of active enhancers, were also observed at this element.

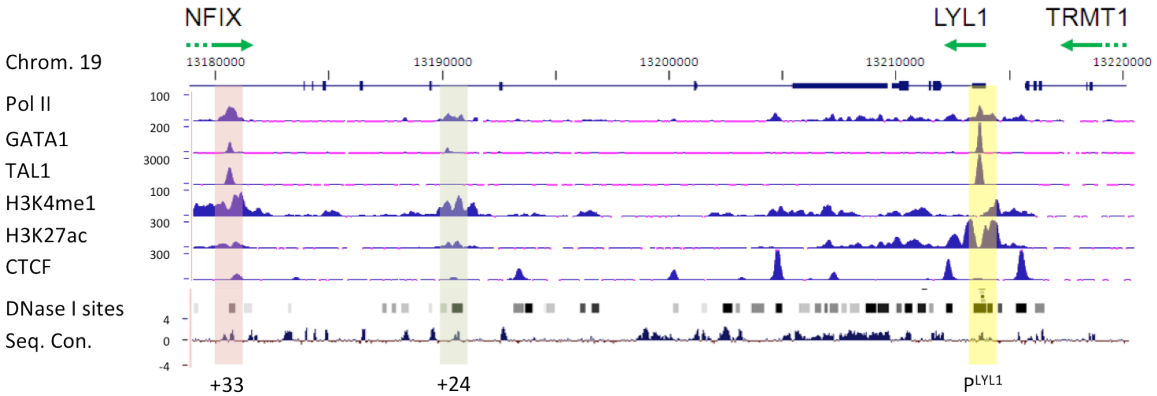


Figure 7.15: Snapshot taken from the UCSC genome browser showing the organization of the human LYL1 locus. The LYL1, NFIX and TRMT1 genes are shown at the top and their directions of transcription are denoted by the green arrows. Scale is genome co-ordinates (bp) from human chromosome 19 (hg.19). Image also shows ENCODE tracks of ChIP-seq data for a number of proteins and histone modifications, DNase I hypersensitive sites and level of evolutionary sequence conservation all detailed at the left. The locations of the +33 enhancer, the +24 element and the LYL1 promoter (P^{LYL1}) are shown by the blue, green and yellow bars respectively.

To further determine whether the +24 element was equivalent to the TAL1 +20/+19 enhancer, 3C analysis was performed to assess the potential looping interaction between the LYL1 promoter and the +24 element. As previously described, two adjacent elements (+10 and +28) of the +24 element were selected, serving as the controls to help distinguish functional looping interaction from non-specific background interactions. As shown in Figure 7.17, the interaction frequencies at the +10, +24 and +28 elements decrease as function of distance to the LYL1 promoter, suggesting no significant looping interaction was observed between the +24 element and the LYL1 promoter.

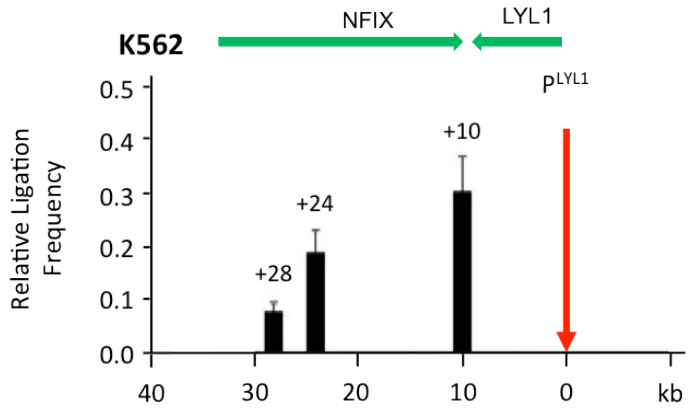


Figure 7.16: 3C analysis of interaction between the LYL1 promoter (P^{LYL1}) and the LYL1 +24 element in K562 cells. Bar diagram of interaction pattern in K562 cell lines determined by 3C. Interaction frequencies (black bars) at three locations across the locus are shown with standard errors and normalized relative to BAC controls. Location of the 3C "anchor" region (P^{LYL1}) is denoted by the vertical red arrow. The locations of the LYL1 and NFIX genes and their directions of transcription are shown at the top of each panel. p values are indicated for interaction frequencies which are significantly higher for test regions when compared to those of control regions (controls defined as regions located between the "anchor" and test regions). Scales (in kb) are shown at the bottom of each panel.

In addition, ChIP-qPCR analysis was also conducted to further assess GATA1 and RNAP II occupancies at the LYL1+24 element in GATA1 knockdown and luciferase control K562 cells. As shown in Figure 7.17, ChIP-occupancy of GATA1 was significantly depleted from the LYL1 +24 element as excepted due to the general loss of GATA1 protein. In contrast, RNAP II occupancy at the +24 element remains unaffected after 96 hour knockdown of GATA1, suggesting that the recruitment of RNAP II is independent of GATA1 binding at the +24 element.

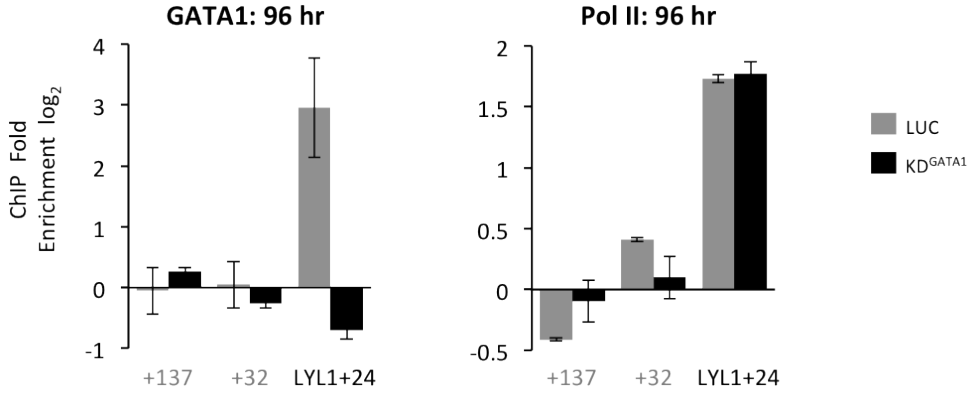


Figure 7.17: GATA1 and RNAP II occupancies at the LYL1 +24 element 96 hours after transfection with siRNA for GATA1 (KD^{GATA1}) or luciferase (LUC). ChIP enrichments (log2) are shown with standard errors. Annotation of test and negative control regions is denoted in black and grey text respectively. Luciferase treated samples are shown as grey bars, KD^{GATA1} treated samples are shown as black bars. Same negative control regions were used as previously described in Chapter 6.

Taking together, this indicates that the LYL1 +24 element cannot be the equivalent element of the TAL1 +20/+19 enhancer – at least not in terms of its looping dynamics with the promoter. However, it cannot be ruled out that this region is an enhancer which regulates LYL1 via a looping-independent mechanism.

Overall, the results presented in this Chapter suggest that the LYL1 locus adopts a single loop chromatin structure, which is much less complicated when compared to its paralogous TAL1 locus. Thus, although there are similarities between the two loci as far as their dependency on GATA/E-box motifs involved in regulating looping between a single enhancer and its cognate promoter, there may be other differences in terms of regulation between two genes which have yet to be identified.

Discussion

7.9 LYL1 and TAL1 share similar regulatory machineries

The data presented in Chapters 6 and this Chapter show that the TAL1 and LYL1 loci share a series of similarities associated with the TEC-dependent regulatory machinery in human erythroid cells. These similarities are listed in Table 7.5 and summarised in the following sections.

Table 7.5: TAL1 and LYL1 share similar regulatory machineries

TAL1	LYL1
Looping interaction between the +51 enhancer and TAL1 promoter 1b is GATA1/TEC dependent	Looping interaction between the +33 enhancer and LYL1 promoter 2 is GATA1/TEC dependent
Expression of TAL1 and its neighboring genes (PDZK1IP1, STIL and CMPK1) is GATA1/TEC dependent	Expression of LYL1 and its neighboring genes (NFIX and TRMT1) is GATA1/TEC dependent
Depletion of GATA1 results in disassociation of TEC at the +51 enhancer and TAL1 promoter 1a	Depletion of GATA1 results in disassociation of TEC at the +33 enhancer and LYL1 promoter 2
Multiple chromatin loops (e.g. P^{TAL1}/+51 & +20) forming into a cruciform configuration in erythroid K562 cells	A single chromatin loop (P^{LYL1}/+33)

7.9.1 GATA1/TEC-dependent chromatin loops at both LYL1 and TAL1 locus

As presented in this Chapter, looping interaction between the P^{LYL1} and the +33 enhancer was decreased along with significant reductions of GATA1, LDB1 and TCF3/E47 occupancies after 96 hour of GATA1 knockdown. These data suggest that GATA1/TEC plays a critical role in mediating chromatin loop between the promoter and the +33 enhancer at the LYL1 locus in erythroid K562 cells. At the paralogous TAL1 locus, a complex cruciform chromatin looping configuration is also modulated via GATA1/TEC in K562 cells. In the case of TAL1, this structure involves the +51 enhancer which shares striking similarities with the LYL +33 element in terms of its TEC binding patterns and clearance of the regulatory machinery during GATA1 knockdown. Therefore, the data presented in this Chapter suggests that there are some parallels in the regulation of both loci – at least with respect to looping governed through the recruitment of the TEC. In both instances, looping mechanisms would allow communication between promoters and their distal *cis*-regulatory elements – thus overcoming obstacles imposed by the presence of other genes and CTCF binding sites juxtaposed between these elements at their respective loci.

7.9.2 Expression of both the LYL1 and TAL locus is GATA1/TEC-dependent

LYL1 expression was significantly down-regulated to about 50% of its wild-type level as a result of GATA1 knockdown. Its expression level was indirectly affected via dissociation of the TEC complex, loss of Pol II and loss of looping interaction at the P^{LYL1} and the +33 enhancer. In addition, expression levels of its neighbouring genes NFIX and TRMT1 were also down regulated in agreement with observation in the TAL1 locus. Theoretical models of how neighbouring genes of TAL1 were getting involved via the cruciform configuration have been proposed and discussed in Chapter 6. However, it is unclear whether the neighbouring genes of LYL1 are also co-regulated via the similar chromatin looping machinery as illustrated at the TAL1 locus. Thus, ChIP-qPCR and 4C analysis could be used to i) characterise the RNAP II occupancy at the neighbouring genes after GATA1 knockdown and ii) identify the chromatin looping interactions between LYL1 and its neighbouring genes (Chapter 8, section 8.2). This would facilitate the determination of whether genes flanking LYL1 are in close contact with the regulatory apparatus of LYL1 – and whether loss of these contacts results in the disassembly of transcriptional complexes at their respective promoters.

7.9.3 The LYL1 +24 element is not a functional equivalent of the TAL1 +20/+19 enhancer

Although the LYL1 +24 element has a number of enhancer hallmarks and was initially considered to be a functional equivalent of the TAL1 +20/+19 enhancer, no detectable chromatin loop was formed between LYL1 promoter and the +24 element (see section 7.4.5). In addition, GATA1 and RNAP II occupancies were observed at the +24 element, suggesting its putative role as an enhancer which may be regulated by GATA1. However, the function of the +24 element remains to be further characterised. Report assay could be used to determine the function of this element *in vitro*. Moreover, the interacting partners of the +24 element can be further determined by 4C-array analysis using the +24 element as the anchor. Ultimately *in vivo* studies (transgenic reporter assays, etc.) would be required to determine the role of both the +24 element as well as the +33 enhancer element.

7.9.4 Models of the TEC-dependent chromatin hubs at the TAL1 and LYL1 loci

Based on the results as shown in Chapter 6 and this Chapter, the chromatin hubs of the TAL1 and LYL1 loci are illustrated in Figure 7.18. In contrast to a highly sophisticated cruciform configuration at the TAL1 locus, the chromatin hub in the LYL1 locus is much simpler which consists of a single loop between the LYL1 promoter and the +33 enhancer. Although the neighbouring genes of LYL1 are also affected when the hub was disrupted by GATA1 knockdown, the mechanism of how these genes get involved in this chromatin hub remains unclear. It is proposed that further characterisation of the LYL1 locus using 4C analysis will provide a much broader view of its chromatin organisation.

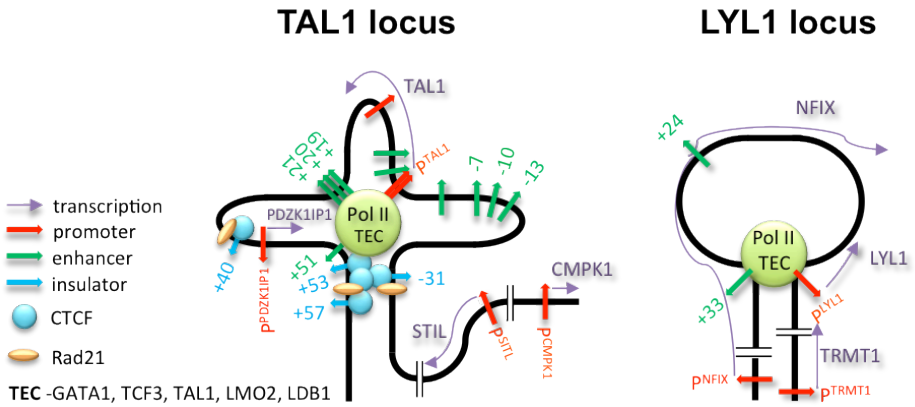


Figure 7.18: Chromatin organisation of the TAL1 and LYL1 loci in erythroid cells. Locations of promoters, enhancers, and putative insulators are depicted as above. Direction of transcription of relevant genes (purple arrows), TEC and Pol II recruitment/loss at the hub, and CTCF and Rad21 binding at insulators are also shown and detailed in the key.

Given the fact that LYL1 and TAL1 are regulated via the same transcription factor complex in erythroid cells, could the two genes be situated in a single TEC-associated transcription factory via inter-chromosomal interactions? Co-regulation of genes that share transcription factor complexes has been previously shown for the β-globin loci and its co-regulated genes situated at other chromosomal regions. Several approaches could be applied to further support the idea that a similar situation occurs for TAL1 and LYL1. These include (i) DNA-FISH to determine the co-localisation of two loci in interphase nuclei *in vivo,* along with (ii) ChIA-PET or ChIP-e4C analysis (Schoenfelder et al., 2010) to determine whether there are chromatin interactions between these two loci, and whether this is mediated via the TEC (see Chapter 8).

7.10 Ectopic expression of LYL1 in T-ALL driven by TRMT1 may share a similar mechanism with the TAL1-STIL micro-deletion

As demonstrated in Chapter 6 and this Chapter, the TAL1 and LYL1 loci have a similar TEC-dependent looping mechanism, which not only involve in transcriptional regulation of TAL1 and LYL1 but also closely relate to the expression of their neighbouring genes. Given the fact that TAL1 is physically in contact with the PDZK1IP1 and STIL genes, it is speculated that the relationships between LYL1 and its neighbouring genes might also via looping interactions. In particular, co-localisation of two genomic elements (STIL intron 1 and TAL1 promoter 1b) at breakpoint most common in TAL1/STIL deletions of T-ALL has been detected by 4C. TRMT1 gene is located upstream of LYL1 in the same transcriptional orientation - this resembles the configuration of STIL and TAL1. Similar to STIL, widespread expression of TRMT1 was observed in all 24 T-ALL cell lines, whereas LYL1 expression was only found in 11/24 of cell lines (Nagel et al., 2010). Strikingly, deletion of the TRMT1 gene was also identified in 5/24 of T-ALL cell lines, correlating with ectopic expression of LYL1. However, in contrast to STIL deletions involving expression of TAL1-STIL fusion transcripts, no TRMT1-LYL1 fusion mRNA have thus far been detected in T-ALL cell lines that have been studied. Nevertheless, it was speculated that the promoter activity of TRMT1 could drive ectopic expression of LYL1 in these cell lines (Nagel et al., 2010).

These observations of TAL1 and LYL1 in T-ALL, suggest that there may be similar mechanisms in how alterations of these genes could drive T-ALL. As illustrated in Chapter 5, the looping interaction captured between the TAL1 promoter and the STIL +1 aligns very close to the deletion breakpoints of STIL in T-ALL cases. From the 3D point of view, these deletion events could be explained as a consequence of the increased chance of co-localisation between two points in a three-dimensional space via looping. Thus, the deletion of TRMT1 may also reveal a co-localisation event between the deletion points, similar to the TAL1/STIL deletion. Altogether, it suggests that ectopic expression of TAL1 and LYL1, via deletion events involving their neighbouring genes may be promoted by same machinery via looping in T-ALL cells. Further 3C and/or 4C analysis should be performed to determine the co-localisation of the LYL1 and TRMT1 promoters, which would

elucidated the relationship between micro-deletion, chromatin spatial organisation and ectopic expression in T-ALL cells (see Chapter 8).

7.11 Evolutionary conservation of gene structures at the TAL1 and LYL1 loci

A large body of functional and structural evidence points to TAL1 and LYL1 arising from a common ancestral bHLH protein during vertebrate evolution. For example, the near-perfect similarity and functional inter-changeability of the bHLH domains has been demonstrated (Capron et al., 2006; Porcher et al., 1999; Schlaeger et al., 2004). Furthermore, at the level of regulation, the arrangement of two highly conserved GATA binding sites in the TAL1 and LYL1 promoters has been identified and shown to bind GATA to a similar level (Chapman et al., 2003) (this Chapter).

The evolutionary evidence also suggests that the TAL1 locus preceded the emergence of LYL1 – and that LYL1 emerged as a duplication of TAL1. LYL1 emerged at some point between chicken and marsupial divergence, as sequence evidence of the LYL1 gene and protein can only be found as far back as marsupials (platypus and opossum). TAL1, however, has been conserved as far back as both amphibian and chicken (which diverged from mammals over 310 million years ago). The data presented in this Chapter for the LYL1 +33 element shows high sequence identity across a number of species including lizard, points to LYL1 being conserved as far back as reptiles which diverged from mammals around ~276 million years ago (Kumar and Hedges, 1998). This would imply that LYL1 arose somewhere between chicken and reptile speciation. However, there is limited evidence to suggest that lizard has a functional LYL1 gene. There is some evidence of conserved sequences between human, mouse and lizard, when translated, which shows identity to peptide sequences from mammalian LYL1. These sequence alignments map upstream of the lizard NFIX orthologue – in the location where a LYL1 lizard orthologue would be expected to have resided (data not shown).

However, the inability to identify a lizard LYL1 gene could be due to the current depth and assembly of genome sequence information in lower vertebrates. This was not such a serious problem for the analysis of TAL1, since this locus was

sequenced and assembled locally across the TAL1 region in a number of vertebrates very early in the human genome project through collaborative work between the B. Gottgens laboratory in Cambridge and the Sanger Institute. Furthermore, examination of TAL1 across species at the level of comparative sequence analysis has been performed using local alignment tools – which perform more robustly in identifying local sequence conservation than global sequence alignment algorithms (the results of the latter are normally displayed on genome browsers). Thus, whether a lizard LYL1 gene exists, awaits further sequence analysis and comparative studies.

The question also arises that, if LYL1 did not emerge as a functional protein in lizard, why has the +33 enhancer region been conserved as far back as lizard? Furthermore, given that the +33 element is embedded in an intron of the NFIX gene (which bears no relationship to genes flanking TAL1), the idea that LYL1 arose from TAL1 with an “intact” erythroid enhancer is hard to comprehend. An alternative explanation is that the duplication of the ancestral TAL1 at the time of the emergence of mammals, serendipitously placed LYL1 under the control of a highly-conserved erythroid GATA/E-box-driven module. This resulted in this newly duplicated bHLH gene acquiring a “new” regulatory function in the erythroid lineage which only exerted selection pressure and functionality during mammal speciation. This may explain why evidence for extensive LYL1 gene sequence conservation is only found as far back at marsupials.

By comparing the organization of TAL1 and LYL1 in species which have thus far been studied, there is further evidence which also points to the two genes sharing other similar regulatory modules (Figure 7.19) as described below:

a. The two genes share similarity in GATA1 sites within their promoters. These similarities date all the way back to the ancestral TAL1 gene found in frogs and chicken.

b. The two genes also show similarities in the +51 and +33 GATA/E-box driven elements. Again this motif is shared with the ancestral TAL1 gene of frog and chicken. Notably chicken and frog possess a second enhancer with a single GATA/E-box (discussed in Chapter 6) which is evidence for the substantial regulatory remodelling that has accompanied TAL1

evolution in mammals (which do not have this enhancer). However, given that mammalian TAL1 has an erythroid enhancer with two GATA/E-boxes while LYL1 appears to have an element with only one GATA-E-box – it is not known which of these two ancestral GATA/E-box modules has been retained in LYL1 (or whether this module is a gain of function as described above).

c. Both TAL1 and LYL1 possess GATA/Ets modules. Although the TAL1 promoter has diverged substantially through evolution from the ancestral promoter structure – it has retained the GATA/Ets module within the stem cell enhancer which is critical for blood, brain and endothelial cells function(Bockamp et al., 1998; Bockamp et al., 1997; Bockamp et al., 1995; Gottgens et al., 2002; Sinclair et al., 1999). For LYL1, this GATA/Ets module is retained in the promoter – and this shows striking similarity to the TAL1 ancestral promoter structure of both frogs and chicken. This evidence would suggest that while the TAL1 promoter has diverged substantially through cis-regulatory remodelling, the LYL1 promoter has retained the structure of the ancestral bHLH gene. Based on current sequence assemblies, identification of this LYL1 promoter configuration could only be confirmed as far back at opossum speciation.

Intriguingly, it has been shown that the chicken TAL1 locus is more similar to the TAL 1 frog locus at the level of sequence conservation than it is to human TAL 1 (Gottgens et al., 2010). This observation is surprising since chicken and mammal evolved from each other 310 million years ago, while chicken and frog have evolved from each other more than 350 million years ago. One explanation for this could have been the early emergence of LYL1, which provided a level of functional redundancy in haematopoietic development to the TAL1 protein, which explains why the rate of TAL1 divergence in mammals has been allowed to accelerate. Thus, the emergence of LYL1 could partially compensate for the function of TAL1, which reduced the selective pressures of TAL1 to maintain its original linear configuration of TF binding motifs. However, this does not explain why LYL1 has retained a “primitive” promoter configuration – as overlap in function between LYL1 and TAL1 may also mean that LYL1 structure could also have diverged more rapidly.

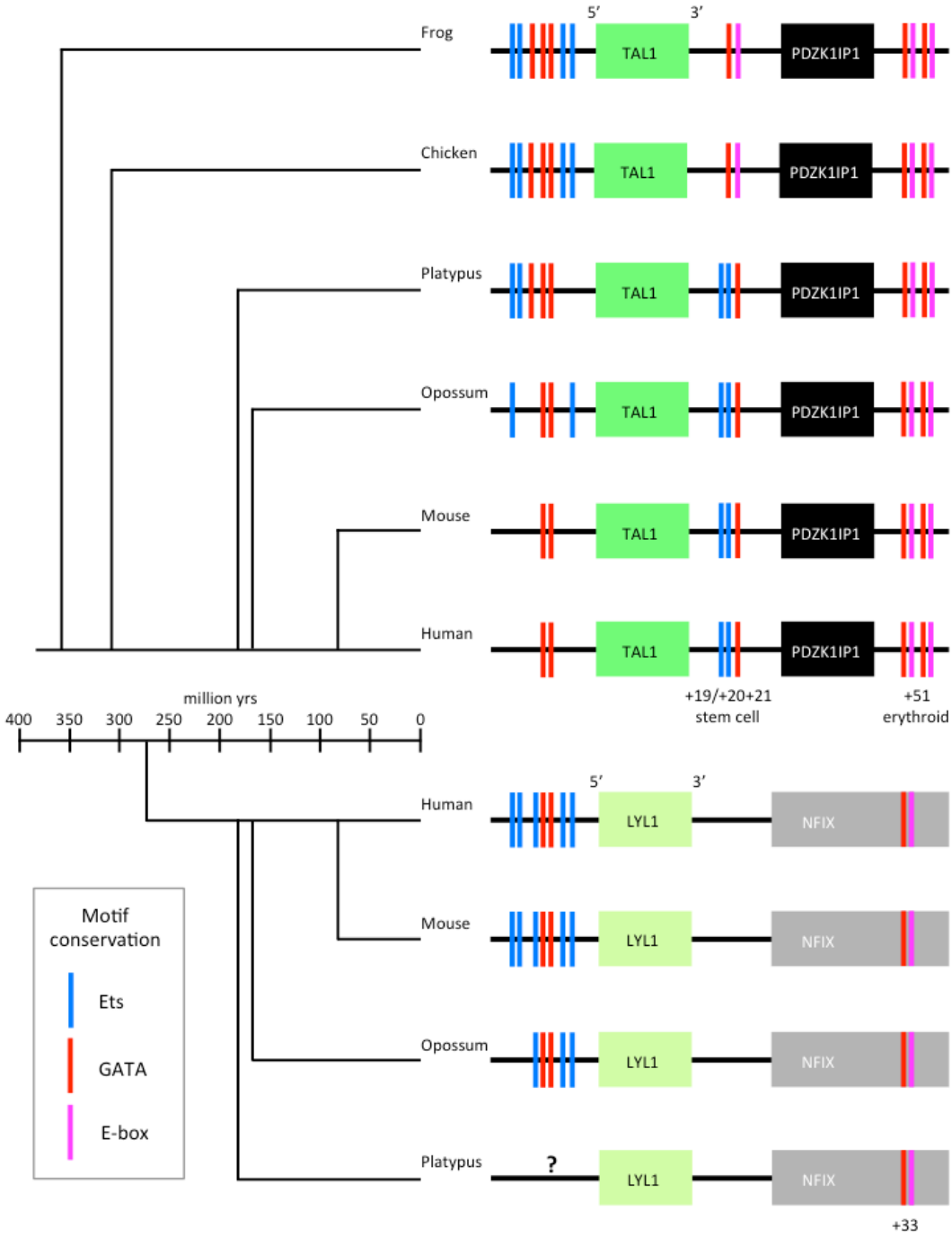


Figure 7.19: The TAL1 and LYL1 loci during vertebrate evolution. Left of the schematic shows the evolutionary tree and divergence of TAL1 and LYL1 across 400 million years of vertebrate evolution. Right of the schematic shows the organization of the TAL1 and LYL1 loci. TAL1 and LYL1 genes are shown by the dark and light green boxes respectively. The PDZK1IP1 and NFIX genes (black and grey boxes respectively) and their locations with respect to the TAL1 +51 erythroid enhancer and the LYL1 +33 element are also shown. Ets, GATA and E-box sequence motifs which are conserved at the level of DNA sequence are also shown. The LYL1 gene emerged at some point between chicken and marsupial divergence. However, conservation of genomic DNA sequences in lizard with high homology to the LYL1 protein (not shown) and the LYL1 +33 element suggests that LYL1 may have emerged at some point between chicken and reptile divergence – more than 270 million yrs ago (Kumar and Hedges, 1998).

If, however, chromatin looping could somehow facilitate cis-regulatory remodelling, the complex cruciform structure at the TAL1 locus – which brings all of its regulatory circuitry into close proximity - may have further facilitated the

divergence of TAL1 regulation (discussed in Chapter 6). Since human LYL1 does not have such a complex structure – rapid evolution of LYL1 regulation may not have occurred. Clearly there are still many unanswered questions about the relationship between TAL1 and LYL1, their arrival during vertebrate evolution, and whether their chromatin configurations have played any roles in the circuitry they use for their regulation.

Conclusion

Similar to TAL1, there is now evidence for a GATA/TEC-dependent regulatory looping mechanism at the LYL1 locus in erythroid cells. However, the LYL1 locus adopts a TEC-mediated single chromatin loop between the +33 enhancer and its cognate promoter, which is much less complex in comparison to the TAL1 cruciform chromatin configuration. Taken together, a common transcriptional cis-regulatory mechanism has been identified for the first time at these two loci.

Chapter 8 Final discussion and future works

8.1 General discussion

Although the regulation of TAL1 has been extensively studied over the last two decades, this thesis presents here the first looping models of its transcriptional regulation (Chapter 6). The looping models detected at the TAL1 locus are context-dependent with respect to its transcription which is consistent with the observation from other gene loci (e.g the globin loci). The experimental data established that GATA1 was a critical requirement for both the transcription of the TAL1 and neighbouring genes and the maintenance of the transcriptional looping configuration, although the data do not exclude the possibility that the integrity of the looping configuration is governed by other factors. As the matter of fact, the TEC has been shown to involve in mediating interactions via LDB1 in regulating gene targets (Song et al., 2007; Xu et al., 2003). Alternatively, it would also be predicted based on the transcription factory model that local concentrations of RNAP II, GATA1 as well as other factors are sufficient for providing a loose scaffold which stabilise the spatial proximity between these regulatory elements (Osborne et al., 2004; Schoenfelder et al., 2010).

This thesis has demonstrated that the TAL1 active hub (see the models proposed in Chapter 6) in erythroid cells also co-regulates, to some extend, the transcription of neighbouring genes of TAL1 including PDZK1IP1, STIL and CMPK1. In particular, the active hub model adequately describes the event of TAL1 and PDZK1IP1 co-regulation without the requirement to propose that TAL1 regulatory elements directly control PDZK1IP1 expression (Tijssen et al., 2011), although it cannot rule out the possibility that their role may be indirect as a result of hub formation.

One of the most important implications of this study is closely related to the molecular pathology of T-ALL. The TAL1/STIL micro-deletions (Tal^{d} deletions), in cases of T-ALL (25%) are well documented (Chapter 5). It has been reported that site-specific Ig/TCR recombinase activity results in the juxtaposition of the 5' end of the STIL gene with the TAL1 promoter and production of a STIL/TAL1 fusion transcript under the control of the STIL promoter (Aplan et al., 1992; Borkhardt et al., 1992; Breit et al., 1993; Brown et al., 1990; Janssen et al., 1993; Kikuchi et al.,

1993). These deletions occur in a remarkably high frequency which suggests that sequences at the breakpoints are specifically prone to recombination by aberrant Ig/TCR recombinase activity. Nevertheless, it is countered by the observation that sequences at the breakpoints are relatively poor substrates for the Ig/TCR recombinase (Brown et al., 1990). The 3C and 4C data presented in this thesis provides an different angle of view. It has been shown that chromatin configurations which modulate transcription of TAL1 situated the TAL1 and STIL genes into spatial close-proximity in both CML- (K562) and T-ALL (HPB-ALL)-derived cell lines. The juxtapositions of the TAL1 promoters and STIL are detected at several sites across the STIL gene, and one of these co-localising to the location of the STIL breakpoint in TAL^{d} deletions. Therefore, spatial proximity between these sites may predispose to TAL^{d} deletions via increasing the likelihood of their recombination. Latest evidence indicates that chromosomal proximity may be a critical determinant in double-stranded breakage and re-joining which result in translocations in cancer genomes (Zhang et al., 2012). The view of the events leading to TAL^{d} deletions in this thesis agrees to these reports and demonstrated that local context-dependent looping topology may be deterministic in molecular pathology.

8.2 Future works

8.2.1 Identification of the long-range intra- and inter-chromosomal interactions

Recently, the combination of high-throughput sequencing and 4C technology was provided a comprehensive map of genome-wide high-resolution interaction networks (Hakim et al., 2011; Noordermeer et al., 2011; Schoenfelder et al., 2010; Xu et al., 2011). The 4C-array method used in this thesis has also been adapted for use with high-throughput sequencing purpose. The 4C-seq assays could provide much more information regarding the three-dimensional chromatin organisation not only of particular locus (TAL1) but also of long-range intra- and inter-chromosomal interactions. The 4C-seq will be able to tell us, other than its neighbouring genes, what other regions or genes are also in contact or even co-transcribed with the TAL1 gene. In addition, it has been shown that both the TAL1 and LYL1 loci are regulated via the TEC (Chapter 6 and 7). As a result of this, it is very likely that 4C-seq assays may also capture the co-localisation of the TAL1 and LYL1 genes, which could be greatly in favour of the “transcription factory”

model. In addition to 3C-based technology, the 3D super-resolution imaging technology (Huang et al., 2008) can be applied in studying 3D chromatin folding dynamics of these loci at the single cell level. A most recent example of using this technology has presented the first non-biochemical evidence that the formation of active chromatin hub at the β-globin locus (van de Corput et al., 2012). In addition, it could be intriguing to further characterise how the chromatin organisation at the TAL1 locus would change at different stages of HSC differentiation using 4C-seq. Induced pluripotent stem cells (iPSCs) technology could provide a great advantage in generating adequate amount of HSCs from human or mouse blood cells (Takahashi et al., 2007; Takahashi and Yamanaka, 2006).

8.2.2 Further characterisation of the roles of CTCF and cohesin in facilitating looping structures

To date, the roles that CTCF and cohesin play in modulating chromosomal organisation are still unclear. It is believed that CTCF and/or cohesin can facilitate the formation of looping structures by mediating interactions between CTSs across the genome. However, there is no evidence as to whether CTCF and/or cohesin are initiating looping structure by bringing other CTSs together or if they only function to stabilise the structure once it is formed. In our study also, the data could not give an exclusive answer to this question. Therefore, it would be desirable to perform knockout experiments to further illustrate possible roles of CTCF and cohesin in the formation of looping structures. It has been found that two CTSs at +57/+53 and -31 are in physical contact, which may stabilise the "cruciform" configuration at the TAL1 locus.

In order to assess the role of CTCF/cohesin at the TAL1 locus, the genomic sequence of the CTS at -31 in K562 cells could be either deleted or mutated using Zn finger nucleases (ZFNs) (Carroll, 2011). The ChIP-qPCR and 3C-PCR assays would be performed in the -31 knockout K562 cells to characterise how the loss of -31 will affect the formation of the TAL1 cruciform configuration. If the cruciform structure is disrupted as a result of the loss of -31, this immediately demonstrates the predominate role of CTCF and/or cohesin. It suggests that the formation of a looping structure in the TAL1 locus is promoted by the interactions between CTSs. Otherwise, it supports the idea that the interactions between CTSs stabilise the TAL1 looping structure.

8.2.3 Identification of regulator elements in the LYL1 locus

Although LYL1 also has fundamental roles in haematopoiesis, little is known about the *cis*-acting regulatory elements across the LYL1 locus. Base on the ENCODE data from the public database (UCSC, http://genome.ucsc.edu/), multiple CTCF binding sites and H3K4me1 peaks are shown across the LYL1 locus, suggesting the presence of putative enhancers and other regulatory elements.

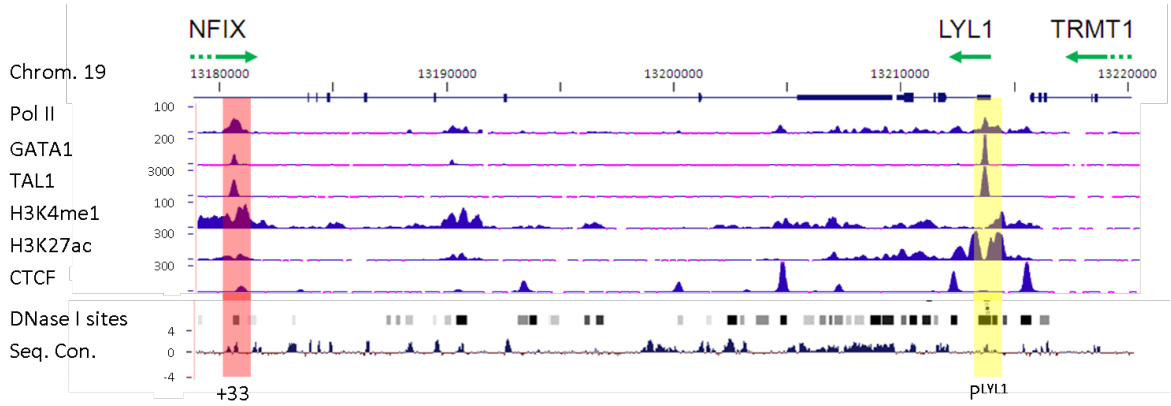


Figure 8.1: Snapshot taken from the UCSC genome browser showing the organization of the human LYL1 locus. The LYL1, NFIX and TRMT1 genes are shown at the top and their directions of transcription are denoted by the green arrows. Scale is genome co-ordinates (bp) from human chromosome 19 (hg. 19). Image also shows ENCODE tracks of ChIP-seq data for a number of proteins and histone modifications, DNase I hypersensitive sites and level of evolutionary sequence conservation all detailed at the left. The locations of the putative LYL1 enhancer (+33 element) and LYL1 promoter 2 (PLYL1) are shown by the red and yellow bars respectively.

The putative regulatory elements can be identified based on the binding of transcription factors (e.g. CTCF, TAL1 and GATA) or the active marks of histone modifications, such as H3K27ac and H3K4me1/2/3. The activity of these putative regulatory elements (e.g. LYL1 +24) could be assayed by reporter assays in *vitro* and *vivo*. In addition, the biological functions of LYL1 regulatory elements could be assessed using transgenic mice. For instance, the role of LYL1 +24 element could be determined by assessing the LYL1 expression through the HSC differentiation in $LYL1^{\Delta 24/\Delta 24}$ mice. Further ChIP-qPCR and 3C analyses could be conducted to identify the involvement of these putative regulatory elements in the chromatin looping structure of the LYL1 locus. These studies could be conducted not only in a tissue-specific way by comparing between different lineages, but also through an evolutionary perspective by comparing across species.

8.2.4 Investigation of looping interactions at the LYL1 locus

The primary data in this thesis implies that the machinery of transcriptional regulation at the LYL1 locus might be similar to that at the TAL1 locus. Identifying a common mechanism sharing between different genes would allow extending our understanding about the role of chromatin organisation in transcriptional regulation. The 3D chromatin configuration can be assessed by both 3C-qPCR and 4C-seq approaches. With very limited prior knowledge of the *cis*-acting regulatory elements within the LYL1 locus, 4C-seq can be a suitable technique for the initial screen of the genome-wide interaction pattern of the LYL1 locus taking the LYL1 promoter and/or the +33 enhancer as viewpoints. Although the majority of interactions captured by 4C-seq are intra-chromosomal interactions, it still provides informative data of not only local contact points within a few hundred kb in distance but also of interactions for all genes and genomic elements associated with the viewpoint across the entire genome (in here, P^{LYL1} and/or +33).

8.2.5 Investigation of evolutionary conservation between the TAL1 and LYL1 genes

There are many similarities between TAL1 and LYL1 in both their gene structures and their transcription factor binding patterns. It would be interesting to further explore the functional conservation between these two genes and to further characterise the possible mechanisms of how the divergence of two genes promotes the evolutionary process. Frog and chicken could be used to validate the hypothesis proposed in chapter 7. Chromatin configuration of the TAL1 locus could be studied in erythroblasts of these two species. The looping interactions between the TAL1 promoter and its enhancers could be determined using 3C assays. The binding pattern of transcription factors involved in TAL1 regulation such as GATA1, TAL1 and LDB1 could be assessed using ChIP analysis.

8.2.6 Investigation of the co-regulation mechanism of TAL1 and STIL

3C and 4C data in this thesis show the promoters of two adjacent genes, TAL1 and STIL, are in spatial proximity, providing a possible explanation of the TAL1-STIL deletions that are frequently found in T-ALL patients (Breit et al., 1993; Brown et al., 1990; Janssen et al., 1993). However, the mechanism driving the

interaction between the TAL1 and STIL promoters is still unclear. An ETS-binding protein, ELF-1 was previously found to bind over the STIL promoter region (STIL+1) in both erythroid (K562 & HL-60) and T-ALL (HPB-ALL and Jurkat) cell lines using ChIP-chip analyses (P. Dhami's PhD thesis, 2005). The recent ChIP-seq data illustrate the occupancy of ELF-1 at both the STIL+1 region and the TAL1 +53 region in K562 cells (ENCODE project, UCSC, Figure 8.2). It is speculated that ELF-1 may mediate looping interaction between STIL +1 and +53, which provides a possible mechanism to allow the TAL1 promoter and STIL +1 being situated in the spatial proximity, as the P^{TAL1}-+51 interaction in K562 cells.

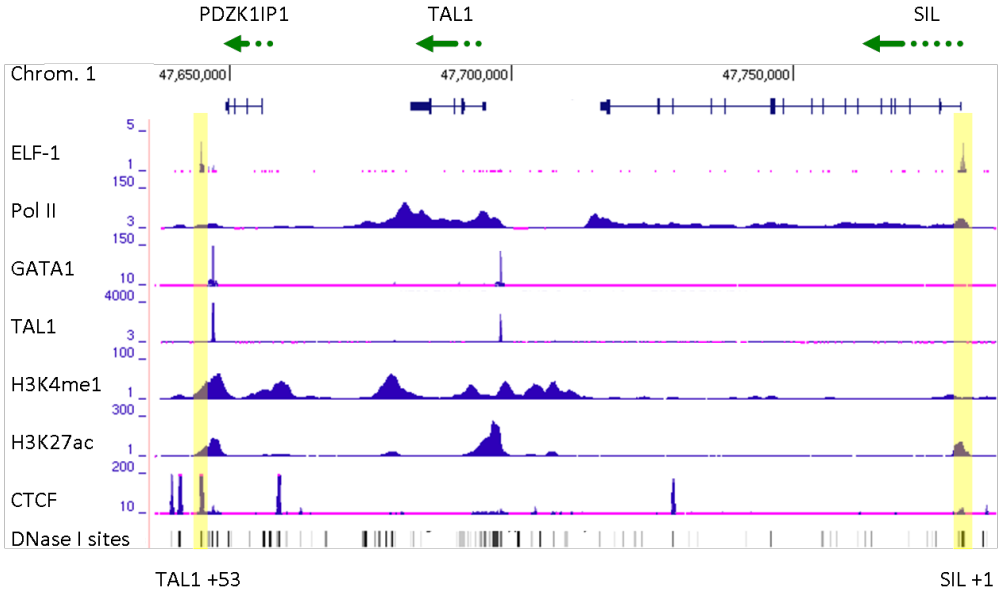


Figure 8.2: Snapshot taken from the UCSC genome browser showing the organization of the human TAL1 locus. The PDZK1IP1, TAL1 and STIL genes are shown at the top and their directions of transcription are denoted in green arrows. Scale is genome co-ordinate (bp) from human chromosome 1 (hg. 19). Image also shows ENCODE tracks of ChIP-seq data for a number of proteins and histone modifications, DNase I hypersensitive sites in K562 cells all detailed at the left. Locations of the TAL1 +53 element and STIL +1 element bound by ELF-1 protein are highlighted in yellow bars.

Further experiments will be required to answer this hypothesis. First, ELF-1 ChIP would be performed to confirm ELF-1 binding at the STIL +1 and the TAL1 +53 in K562 cells. Second, 3C analyses would be conducted to determine the possible interactions between STIL +1 and TAL1 +53. Third, the ELF-1 protein would be knocked-down by siRNA in K562 cells. The knockdown cells would be examined for the loss of ELF-1 binding over the target sites, the loss of interaction between the STIL +1 and TAL1 +53, and a subsequent loss of interactions between the TAL1 and STIL promoters. Further questions would be addressed such as whether knockdown of ELF-1 affects the formation of the TEC and the TAL1

cruciform configuration. The ChIP and 3C experiments on ELF-1 knockdown K562 cells would be able to address these questions.

The interaction between promoters of TAL1 and STIL was not only detected in wild-type K562 cells but was also found in HPB-ALL cells. However, the above explanation could only fit into the TAL1-STIL model for K562 cells, as no interaction was detected between the TAL1 promoter and the +51 enhancer in HPB-ALL cells. Nevertheless, the promoters of TAL1 and STIL are still in contact in HPB-ALL cells, though with less frequently than in K562 cells. Moreover, the TAL1 promoter is in contact with the +20/+19 stem cell enhancer in HPB-ALL cells. The murine Tal1 +19/+18 stem cell enhancer is controlled by a multi-protein including GATA2, Fli-1 and Elf-1 (Gottgens et al., 2004). It has been shown by recent ChIP-seq studies that the Ets factors Pu.1 and Erg bind to the Tal1 +19 enhancer (Tijssen et al., 2011; Wilson et al., 2010). The assumption is that the ELF-1 protein mediates the interaction between the promoters of TAL1 and STIL through binding to the stem cell enhancer instead of TAL1 +53 in HPB-ALL cells. ELF-1 ChIP could be performed to detect binding over the TAL1 +20/19 stem cell enhancer in HPB-ALL and the 3C assays could be conducted to determine the interactions between STIL +1 and TAL1 +20/19.

8.2.7 Investigation of the structural mechanism of TAL1-STIL deletion in T-ALL

The TAL1-STIL deletion was found in over 25% of T-ALL patients (Breit et al., 1993; Brown et al., 1990; Janssen et al., 1993). The underlying mechanism of this deletion is still unclear. 3C and 4C results in this thesis have provided a possible mechanism of the TAL1-STIL deletion from a chromatin structure point of view. Further experiments need to be conducted using T-ALL cells.

Table 8.1: Three types of T-ALL cell lines categorised based on TAL1-STIL deletion and expression level of TAL1 and STIL.

Type	TAL1-SIL deletion	TAL1 expression	SIL expression	Cell lines
A	yes	yes	yes	CCRF-CEM, RPMI 8402
B	no	yes	yes	Jurkat, REX
C	no	no	yes	HPB-ALL, MOLT

There are three types of the T-ALL cell lines that could be studied as well as primary $CD4^+$ T-cells. The three types of T-ALL cells are categorised as shown in Table 8.1. The 3C assays would be conducted to detect chromatin interactions between the P^{TAL1} and the P^{STIL} in different types of T-ALL and in the primary T-cells. The relationship between the P^{TAL1}/P^{STIL} interaction and the TAL1-STIL micro-deletion events would be examined. Further ELF-1 ChIP assays would also be conducted to determine ELF-1 occupancy at its target sites, which would provide a possible mechanism of ELF-1 mediated looping interactions.

8.3 Final thoughts

The results presented in this thesis illustrated the relationships between long-range chromatin interactions, higher-order chromosomal configuration and transcriptional regulation at the TAL1 locus in erythroid and lymphoid lineages. This is the initial effort in determining the structural links between *cis*-acting regulatory elements in the three-dimensional perspective and underlying transcription machinery. Given that there are more than 2000 transcription factors and over 30000 human genes, it requires a huge effort to understand how gene expression is regulated via long-range intra- and inter-chromosomal interactions mediated by transcription factors between consensus and tissue-specific regulatory elements. Advances in technology (Hi-C, ChIA-PET and ChIP-seq) and computational tools are required to integrate all the data in a biologically meaningful way. The outputs in this thesis provide a foundation and will make a valuable contribution towards this goal.

Reference

Anderson, K.P., Crable, S.C., and Lingrel, J.B. (1998). Multiple proteins binding to a GATA-E box-GATA motif regulate the erythroid Kruppel-like factor (EKLF) gene. J Biol Chem *273*, 14347-14354.

Anderson, K.P., Crable, S.C., and Lingrel, J.B. (2000). The GATA-E box-GATA motif in the EKLF promoter is required for in vivo expression. Blood *95*, 1652-1655.

Anguita, E., Hughes, J., Heyworth, C., Blobel, G.A., Wood, W.G., and Higgs, D.R. (2004). Globin gene activation during haemopoiesis is driven by protein complexes nucleated by GATA-1 and GATA-2. EMBO J *23*, 2841-2852.

Aplan, P.D., Begley, C.G., Bertness, V., Nussmeier, M., Ezquerra, A., Coligan, J., and Kirsch, I.R. (1990a). The SCL gene is formed from a transcriptionally complex locus. Mol Cell Biol *10*, 6426-6435.

Aplan, P.D., Jones, C.A., Chervinsky, D.S., Zhao, X., Ellsworth, M., Wu, C., McGuire, E.A., and Gross, K.W. (1997). An scl gene product lacking the transactivation domain induces bony abnormalities and cooperates with LMO1 to generate T-cell malignancies in transgenic mice. EMBO J *16*, 2408-2419.

Aplan, P.D., Lombardi, D.P., Ginsberg, A.M., Cossman, J., Bertness, V.L., and Kirsch, I.R. (1990b). Disruption of the human SCL locus by "illegitimate" V-(D)-J recombinase activity. Science *250*, 1426-1429.

Aplan, P.D., Lombardi, D.P., Reaman, G.H., Sather, H.N., Hammond, G.D., and Kirsch, I.R. (1992a). Involvement of the putative hematopoietic transcription factor SCL in T-cell acute lymphoblastic leukemia. Blood *79*, 1327-1333.

Aplan, P.D., Nakahara, K., Orkin, S.H., and Kirsch, I.R. (1992b). The SCL gene product: a positive regulator of erythroid differentiation. EMBO J *11*, 4073-4081.

Aplan, P.D., Raimondi, S.C., and Kirsch, I.R. (1992c). Disruption of the SCL gene by a t(1;3) translocation in a patient with T cell acute lymphoblastic leukemia. J Exp Med *176*, 1303-1310.

Azuara, V., Perry, P., Sauer, S., Spivakov, M., Jorgensen, H.F., John, R.M., Gouti, M., Casanova, M., Warnes, G., Merkenschlager, M.*, et al.* (2006). Chromatin signatures of pluripotent cell lines. Nat Cell Biol *8*, 532-538.

Baer, R. (1993). TAL1, TAL2 and LYL1: a family of basic helix-loop-helix proteins implicated in T cell acute leukaemia. Semin Cancer Biol *4*, 341-347.

Bailey, T.L., and Elkan, C. (1995). The value of prior knowledge in discovering motifs with MEME. Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology *3*, 21-29.

Baker, M. (2011). Genomics: Genomes in three dimensions. Nature *470*, 289-294.

Bantignies, F., Roure, V., Comet, I., Leblanc, B., Schuettengruber, B., Bonnet, J., Tixier, V., Mas, A., and Cavalli, G. (2011). Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. Cell *144*, 214-226.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Barton, L.M., Gottgens, B., and Green, A.R. (1999). The stem cell leukaemia (SCL) gene: a critical regulator of haemopoietic and vascular development. The international journal of biochemistry & cell biology *31*, 1193-1207.

Begley, C.G., Aplan, P.D., Davey, M.P., Nakahara, K., Tchorz, K., Kurtzberg, J., Hershfield, M.S., Haynes, B.F., Cohen, D.I., Waldmann, T.A*., et al.* (1989a). Chromosomal translocation in a human leukemic stem-cell line disrupts the T-cell antigen receptor delta-chain diversity region and results in a previously unreported fusion transcript. Proc Natl Acad Sci U S A *86*, 2031-2035.

Begley, C.G., Aplan, P.D., Denning, S.M., Haynes, B.F., Waldmann, T.A., and Kirsch, I.R. (1989b). The gene SCL is expressed during early hematopoiesis and encodes a differentiation-related DNA-binding motif. Proc Natl Acad Sci U S A *86*, 10128-10132.

Begley, C.G., and Green, A.R. (1999). The SCL gene: from case report to critical hematopoietic regulator. Blood *93*, 2760-2770.

Begley, C.G., Robb, L., Rockman, S., Visvader, J., Bockamp, E.O., Chan, Y.S., and Green, A.R. (1994). Structure of the gene encoding the murine SCL protein. Gene *138*, 93-99.

Begley, C.G., Visvader, J., Green, A.R., Aplan, P.D., Metcalf, D., Kirsch, I.R., and Gough, N.M. (1991). Molecular cloning and chromosomal localization of the murine homolog of the human helix-loop-helix gene SCL. Proc Natl Acad Sci U S A *88*, 869-873.

Bell, A.C., West, A.G., and Felsenfeld, G. (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell *98*, 387-396.

Bernard, O., Azogui, O., Lecointe, N., Mugneret, F., Berger, R., Larsen, C.J., and Mathieu-Mahul, D. (1992). A third tal-1 promoter is specifically used in human T cell leukemias. J Exp Med *176*, 919-925.

Bernard, O., Guglielmi, P., Jonveaux, P., Cherif, D., Gisselbrecht, S., Mauchauffe, M., Berger, R., Larsen, C.J., and Mathieu-Mahul, D. (1990). Two distinct mechanisms for the SCL gene activation in the t(1;14) translocation of T-cell leukemias. Genes Chromosomes Cancer *1*, 194-208.

Bernard, O., Lecointe, N., Jonveaux, P., Souyri, M., Mauchauffe, M., Berger, R., Larsen, C.J., and Mathieu-Mahul, D. (1991). Two site-specific deletions and t(1;14) translocation restricted to human T-cell acute leukemias disrupt the 5' part of the tal-1 gene. Oncogene *6*, 1477-1488.

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas, E.J., 3rd, Gingeras, T.R*., et al.* (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. Cell *120*, 169-181.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K*., et al.* (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315-326.

Blackwell, T.K., and Weintraub, H. (1990). Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. Science *250*, 1104-1110.

Blackwood, E.M., and Kadonaga, J.T. (1998). Going the distance: a current view of enhancer action. Science *281*, 60-63.

Bockamp, E.O., Fordham, J.L., Gottgens, B., Murrell, A.M., Sanchez, M.J., and Green, A.R. (1998). Transcriptional regulation of the stem cell leukemia gene by PU.1 and Elf-1. J Biol Chem *273*, 29032-29042.

Bockamp, E.O., McLaughlin, F., Gottgens, B., Murrell, A.M., Elefanty, A.G., and Green, A.R. (1997). Distinct mechanisms direct SCL/tal-1 expression in erythroid cells and CD34 positive primitive myeloid cells. J Biol Chem *272*, 8781-8790.

Bockamp, E.O., McLaughlin, F., Murrell, A.M., Gottgens, B., Robb, L., Begley, C.G., and Green, A.R. (1995). Lineage-restricted regulation of the murine SCL/TAL-1 promoter. Blood *86*, 1502-1514.

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Muller, S., Eils, R., Cremer, C., Speicher, M.R.*, et al.* (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. PLoS biology *3*, e157.

Borkhardt, A., Repp, R., Harbott, J., Keller, C., Berner, F., Ritterbach, J., and Lampert, F. (1992). Frequency and DNA sequence of tal-1 rearrangement in children with T-cell acute lymphoblastic leukemia. Ann Hematol *64*, 305-308.

Bourc'his, D., Xu, G.L., Lin, C.S., Bollman, B., and Bestor, T.H. (2001). Dnmt3L and the establishment of maternal genomic imprints. Science *294*, 2536-2539.

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G.*, et al.* (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. Cell *122*, 947-956.

Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K.*, et al.* (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature *441*, 349-353.

Branco, M.R., Branco, T., Ramirez, F., and Pombo, A. (2008). Changes in chromosome organization during PHA-activation of resting human lymphocytes measured by cryo-FISH. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology *16*, 413-426.

Breit, T.M., Mol, E.J., Wolvers-Tettero, I.L., Ludwig, W.D., van Wering, E.R., and van Dongen, J.J. (1993). Site-specific deletions involving the tal-1 and sil genes are restricted to cells of the T cell receptor alpha/beta lineage: T cell receptor delta gene deletion mechanism affects multiple genes. J Exp Med *177*, 965-977.

Brown, L., Cheng, J.T., Chen, Q., Siciliano, M.J., Crist, W., Buchanan, G., and Baer, R. (1990). Site-specific recombination of the tal-1 gene is a common occurrence in human T cell leukemia. EMBO J *9*, 3343-3351.

Bruce, A.W., Lopez-Contreras, A.J., Flicek, P., Down, T.A., Dhami, P., Dillon, S.C., Koch, C.M., Langford, C.F., Dunham, I., Andrews, R.M.*, et al.* (2009). Functional diversity for REST (NRSF) is defined by in vivo binding affinity hierarchies at the DNA sequence level. Genome Res *19*, 994-1005.

Buck, M.J., and Lieb, J.D. (2004). ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics *83*, 349-360.

Bulger, M., and Groudine, M. (1999). Looping versus linking: toward a model for long-distance gene activation. Genes Dev *13*, 2465-2477.

Cai, S., Lee, C.C., and Kohwi-Shigematsu, T. (2006). SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes. Nat Genet *38*, 1278-1288.

Cai, Y., Xu, Z., Xie, J., Ham, A.J., Koury, M.J., Hiebert, S.W., and Brandt, S.J. (2009). Eto2/MTG16 and MTGR1 are heteromeric corepressors of the TAL1/SCL transcription factor in murine erythroid progenitors. Biochem Biophys Res Commun *390*, 295-301.

Cairns, B.R. (2009). The logic of chromatin architecture and remodelling at promoters. Nature *461*, 193-198.

Cameron, R.A., Chow, S.H., Berney, K., Chiu, T.Y., Yuan, Q.A., Kramer, A., Helguero, A., Ransick, A., Yun, M., and Davidson, E.H. (2005). An evolutionary constraint: strongly disfavored class of change in DNA sequence during divergence of cis-regulatory modules. Proc Natl Acad Sci U S A *102*, 11769-11774.

Cameron, R.A., and Davidson, E.H. (2009). Flexibility of transcription factor target site position in conserved cis-regulatory modules. Developmental biology *336*, 122-135.

Cantor, A.B., Iwasaki, H., Arinobu, Y., Moran, T.B., Shigematsu, H., Sullivan, M.R., Akashi, K., and Orkin, S.H. (2008). Antagonism of FOG-1 and GATA factors in fate choice for the mast cell lineage. J Exp Med *205*, 611-624.

Capron, C., Lacout, C., Lecluse, Y., Wagner-Ballon, O., Kaushik, A.L., Cramer-Borde, E., Sablitzky, F., Dumenil, D., and Vainchenker, W. (2011). LYL-1 deficiency induces a stress erythropoiesis. Exp Hematol *39*, 629-642.

Capron, C., Lecluse, Y., Kaushik, A.L., Foudi, A., Lacout, C., Sekkai, D., Godin, I., Albagli, O., Poullion, I., Svinartchouk, F.*, et al.* (2006). The SCL relative LYL-1 is required for fetal and adult hematopoietic stem cell function and B-cell differentiation. Blood *107*, 4678-4686.

Carroll, D. (2011). Genome engineering with zinc-finger nucleases. Genetics *188*, 773-782.

Carter, D., Chakalova, L., Osborne, C.S., Dai, Y.F., and Fraser, P. (2002). Long-range chromatin regulatory interactions in vivo. Nat Genet *32*, 623-626.

Ceredig, R., Rolink, A.G., and Brown, G. (2009). Models of haematopoiesis: seeing the wood for the trees. Nature reviews Immunology *9*, 293-300.

Chambers, S.M., Boles, N.C., Lin, K.Y., Tierney, M.P., Bowman, T.V., Bradfute, S.B., Chen, A.J., Merchant, A.A., Sirin, O., Weksberg, D.C.*, et al.* (2007). Hematopoietic fingerprints: an expression database of stem cells and their progeny. Cell stem cell *1*, 578-591.

Chan, W.Y., Follows, G.A., Lacaud, G., Pimanda, J.E., Landry, J.R., Kinston, S., Knezevic, K., Piltz, S., Donaldson, I.J., Gambardella, L.*, et al.* (2007). The paralogous hematopoietic regulators Lyl1 and Scl are coregulated by Ets and GATA factors, but Lyl1 cannot rescue the early Scl-/- phenotype. Blood *109*, 1908-1916.

Chapman, M.A., Charchar, F.J., Kinston, S., Bird, C.P., Grafham, D., Rogers, J., Grutzner, F., Graves, J.A., Green, A.R., and Gottgens, B. (2003). Comparative and functional analyses of LYL1 loci establish marsupial sequences as a model for phylogenetic footprinting. Genomics *81*, 249-259.

Chapman, M.A., Donaldson, I.J., Gilbert, J., Grafham, D., Rogers, J., Green, A.R., and Gottgens, B. (2004). Analysis of multiple genomic sequence alignments: a web resource,

online tools, and lessons learned from analysis of mammalian SCL loci. Genome Res *14*, 313-318.

Chen, Q., Cheng, J.T., Tasi, L.H., Schneider, N., Buchanan, G., Carroll, A., Crist, W., Ozanne, B., Siciliano, M.J., and Baer, R. (1990). The tal gene undergoes chromosome translocation in T cell leukemia and potentially encodes a helix-loop-helix protein. EMBO J *9*, 415-424.

Chen, Z.X., Mann, J.R., Hsieh, C.L., Riggs, A.D., and Chedin, F. (2005). Physical and functional interactions between the human DNMT3L protein and members of the de novo methyltransferase family. J Cell Biochem *95*, 902-917.

Cheng, J.T., Cobb, M.H., and Baer, R. (1993a). Phosphorylation of the TAL1 oncoprotein by the extracellular-signal-regulated protein kinase ERK1. Mol Cell Biol *13*, 801-808.

Cheng, J.T., Hsu, H.L., Hwang, L.Y., and Baer, R. (1993b). Products of the TAL1 oncogene: basic helix-loop-helix proteins phosphorylated at serine residues. Oncogene *8*, 677-683.

Cherniack, R.M., Crystal, R.G., and Kalica, A.R. (1991). NHLBI Workshop summary. Current concepts in idiopathic pulmonary fibrosis: a road map for the future. The American review of respiratory disease *143*, 680-683.

Chiba, T., Nagata, Y., Kishi, A., Sakamaki, K., Miyajima, A., Yamamoto, M., Engel, J.D., and Todokoro, K. (1993). Induction of erythroid-specific gene expression in lymphoid cells. Proc Natl Acad Sci U S A *90*, 11593-11597.

Chien, R., Zeng, W., Kawauchi, S., Bender, M.A., Santos, R., Gregson, H.C., Schmiesing, J.A., Newkirk, D.A., Kong, X., Ball, A.R., Jr*., et al.* (2011). Cohesin mediates chromatin interactions that regulate mammalian beta-globin expression. J Biol Chem *286*, 17870-17878.

Chodavarapu, R.K., Feng, S., Bernatavichute, Y.V., Chen, P.Y., Stroud, H., Yu, Y., Hetzel, J.A., Kuo, F., Kim, J., Cokus, S.J*., et al.* (2010). Relationship between nucleosome positioning and DNA methylation. Nature *466*, 388-392.

Choi, I., Cho, B.R., Kim, D., Miyagawa, S., Kubo, T., Kim, J.Y., Park, C.G., Hwang, W.S., Lee, J.S., and Ahn, C. (2005). Choice of the adequate detection time for the accurate evaluation of the efficiency of siRNA-induced gene silencing. Journal of biotechnology *120*, 251-261.

Choi, K., Kennedy, M., Kazarov, A., Papadimitriou, J.C., and Keller, G. (1998). A common precursor for hematopoietic and endothelial cells. Development *125*, 725-732.

Chung, Y.S., Zhang, W.J., Arentson, E., Kingsley, P.D., Palis, J., and Choi, K. (2002). Lineage analysis of the hemangioblast as defined by FLK1 and SCL expression. Development *129*, 5511-5520.

Cleary, M.L., Mellentin, J.D., Spies, J., and Smith, S.D. (1988). Chromosomal translocation involving the beta T cell receptor gene in acute leukemia. J Exp Med *167*, 682-687.

Collas, P. (2010). The current state of chromatin immunoprecipitation. Molecular biotechnology *45*, 87-100.

Condorelli, G.L., Tocci, A., Botta, R., Facchiano, F., Testa, U., Vitelli, L., Valtieri, M., Croce, C.M., and Peschle, C. (1997). Ectopic TAL-1/SCL expression in phenotypically normal or leukemic myeloid precursors: proliferative and antiapoptotic effects coupled with a differentiation blockade. Mol Cell Biol *17*, 2954-2969.

Courtes, C., Lecointe, N., Le Cam, L., Baudoin, F., Sardet, C., and Mathieu-Mahul, D. (2000). Erythroid-specific inhibition of the tal-1 intragenic promoter is due to binding of a repressor to a novel silencer. J Biol Chem *275*, 949-958.

Cremer, T., and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nature reviews Genetics *2*, 292-301.

Cremer, T., Kurz, A., Zirbel, R., Dietzel, S., Rinke, B., Schrock, E., Speicher, M.R., Mathieu, U., Jauch, A., Emmerich, P.*, et al.* (1993). Role of chromosome territories in the functional compartmentalization of the cell nucleus. Cold Spring Harbor symposia on quantitative biology *58*, 777-792.

Cross, M.A., Heyworth, C.M., Murrell, A.M., Bockamp, E.O., Dexter, T.M., and Green, A.R. (1994). Expression of lineage restricted transcription factors precedes lineage specific differentiation in a multipotent haemopoietic progenitor cell line. Oncogene *9*, 3013-3016.

Curtis, D.J., Hall, M.A., Van Stekelenburg, L.J., Robb, L., Jane, S.M., and Begley, C.G. (2004). SCL is required for normal function of short-term repopulating hematopoietic stem cells. Blood *103*, 3342-3348.

D'Souza, S.L., Elefanty, A.G., and Keller, G. (2005). SCL/Tal-1 is essential for hematopoietic commitment of the hemangioblast but not for its development. Blood *105*, 3862-3870.

Dang, C.V., O'Donnell, K.A., Zeller, K.I., Nguyen, T., Osthus, R.C., and Li, F. (2006). The c-Myc target gene network. Semin Cancer Biol *16*, 253-264.

De Gobbi, M., Anguita, E., Hughes, J., Sloane-Stanley, J.A., Sharpe, J.A., Koch, C.M., Dunham, I., Gibbons, R.J., Wood, W.G., and Higgs, D.R. (2007). Tissue-specific histone modification and transcription factor binding in alpha globin gene expression. Blood *110*, 4503-4510.

de Laat, W., and Grosveld, F. (2003). Spatial organization of gene expression: the active chromatin hub. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology *11*, 447-459.

de Napoles, M., Mermoud, J.E., Wakao, R., Tang, Y.A., Endoh, M., Appanah, R., Nesterova, T.B., Silva, J., Otte, A.P., Vidal, M.*, et al.* (2004). Polycomb group proteins Ring1A/B link ubiquitylation of histone H2A to heritable gene silencing and X inactivation. Dev Cell *7*, 663-676.

de Wit, E., and de Laat, W. (2012). A decade of 3C technologies: insights into nuclear organization. Genes Dev *26*, 11-24.

Degner, S.C., Verma-Gaur, J., Wong, T.P., Bossen, C., Iverson, G.M., Torkamani, A., Vettermann, C., Lin, Y.C., Ju, Z., Schulz, D.*, et al.* (2011). CCCTC-binding factor (CTCF) and cohesin influence the genomic architecture of the Igh locus and antisense transcription in pro-B cells. Proc Natl Acad Sci U S A *108*, 9566-9571.

Dekker, J. (2006). The three 'C' s of chromosome conformation capture: controls, controls, controls. Nat Methods *3*, 17-21.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306-1311.

Delabesse, E., Ogilvy, S., Chapman, M.A., Piltz, S.G., Gottgens, B., and Green, A.R. (2005). Transcriptional regulation of the SCL locus: identification of an enhancer that targets the primitive erythroid lineage in vivo. Mol Cell Biol *25*, 5215-5225.

Dhami, P., Bruce, A.W., Jim, J.H., Dillon, S.C., Hall, A., Cooper, J.L., Bonhoure, N., Chiang, K., Ellis, P.D., Langford, C.*, et al.* (2010). Genomic approaches uncover increasing complexities in the regulatory landscape at the human SCL (TAL1) locus. PLoS One *5*, e9059.

Dhami, P., Coffey, A.J., Abbs, S., Vermeesch, J.R., Dumanski, J.P., Woodward, K.J., Andrews, R.M., Langford, C., and Vetrie, D. (2005). Exon array CGH: detection of copy-number changes at the resolution of individual exons in the human genome. Am J Hum Genet *76*, 750-762.

Diakos, C., Krapf, G., Gerner, C., Inthal, A., Lemberger, C., Ban, J., Dohnal, A.M., and Panzer-Grumayer, E.R. (2007). RNAi-mediated silencing of TEL/AML1 reveals a heat-shock protein- and survivin-dependent mechanism for survival. Blood *109*, 2607-2610.

Dooley, K.A., Davidson, A.J., and Zon, L.I. (2005). Zebrafish scl functions independently in hematopoietic and endothelial development. Developmental biology *277*, 522-536.

Dostie, J., and Dekker, J. (2007). Mapping networks of physical interactions between genomic elements using 5C technology. Nat Protoc *2*, 988-1002.

Dostie, J., Zhan, Y., and Dekker, J. (2007). Chromosome conformation capture carbon copy technology. Curr Protoc Mol Biol *Chapter 21*, Unit 21 14.

Down, T.A., and Hubbard, T.J. (2005). NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. Nucleic Acids Res *33*, 1445-1453.

Drake, C.J., Brandt, S.J., Trusk, T.C., and Little, C.D. (1997). TAL1/SCL is expressed in endothelial progenitor cells/angioblasts and defines a dorsal-to-ventral gradient of vasculogenesis. Developmental biology *192*, 17-30.

Dreyer, A.K., and Cathomen, T. (2012). Zinc-finger nucleases-based genome engineering to generate isogenic human cell lines. Methods Mol Biol *813*, 145-156.

Drissen, R., Palstra, R.J., Gillemans, N., Splinter, E., Grosveld, F., Philipsen, S., and de Laat, W. (2004). The active spatial organization of the beta-globin locus requires the transcription factor EKLF. Genes Dev *18*, 2485-2490.

Duan, Q., Chen, H., Costa, M., and Dai, W. (2008). Phosphorylation of H3S10 blocks the access of H3K9 by specific antibodies and histone methyltransferase. Implication in regulating chromatin dynamics and epigenetic inheritance during mitosis. J Biol Chem *283*, 33585-33590.

Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y.J., Lee, C., Shendure, J., Fields, S., Blau, C.A., and Noble, W.S. (2010). A three-dimensional model of the yeast genome. Nature *465*, 363-367.

Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J.M. (1999). Expression profiling using cDNA microarrays. Nat Genet *21*, 10-14.

Dzierzak, E., Medvinsky, A., and de Bruijn, M. (1998). Qualitative and quantitative aspects of haematopoietic cell development in the mammalian embryo. Immunology today *19*, 228-236.

Elbashir, S.M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., and Tuschl, T. (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature *411*, 494-498.

Elefanty, A.G., Robb, L., Birner, R., and Begley, C.G. (1997). Hematopoietic-specific genes are not induced during in vitro differentiation of scl-null embryonic stem cells. Blood *90*, 1435-1447.

Elkon, R., Rashi-Elkeles, S., Lerenthal, Y., Linhart, C., Tenne, T., Amariglio, N., Rechavi, G., Shamir, R., and Shiloh, Y. (2005). Dissection of a DNA-damage-induced transcriptional network using a combination of microarrays, RNA interference and computational promoter analysis. Genome Biol *6*, R43.

Elwood, N.J., Cook, W.D., Metcalf, D., and Begley, C.G. (1993). SCL, the gene implicated in human T-cell leukaemia, is oncogenic in a murine T-lymphocyte cell line. Oncogene *8*, 3093-3101.

Elwood, N.J., Green, A.R., Melder, A., Begley, C.G., and Nicola, N. (1994). The SCL protein displays cell-specific heterogeneity in size. Leukemia *8*, 106-114.

Elwood, N.J., Zogos, H., Pereira, D.S., Dick, J.E., and Begley, C.G. (1998). Enhanced megakaryocyte and erythroid development from normal human CD34(+) cells: consequence of enforced expression of SCL. Blood *91*, 3756-3765.

Endoh, M., Ogawa, M., Orkin, S., and Nishikawa, S. (2002). SCL/tal-1-dependent process determines a competence to select the definitive hematopoietic lineage prior to endothelial differentiation. EMBO J *21*, 6700-6708.

Engel, I., and Murre, C. (2001). The function of E- and Id proteins in lymphocyte development. Nature reviews Immunology *1*, 193-199.

Engel, J.D., Beug, H., LaVail, J.H., Zenke, M.W., Mayo, K., Leonard, M.W., Foley, K.P., Yang, Z., Kornhauser, J.M., Ko, L.J.*, et al.* (1992). cis and trans regulation of tissue-specific transcription. Journal of cell science Supplement *16*, 21-31.

Essien, K., Vigneau, S., Apreleva, S., Singh, L.N., Bartolomei, M.S., and Hannenhalli, S. (2009). CTCF binding site classes exhibit distinct evolutionary, genomic, epigenomic and transcriptomic features. Genome Biol *10*, R131.

Esteller, M. (2007). Epigenetic gene silencing in cancer: the DNA hypermethylome. Human molecular genetics *16 Spec No 1*, R50-59.

Esteve, P.O., Chin, H.G., Benner, J., Feehery, G.R., Samaranayake, M., Horwitz, G.A., Jacobsen, S.E., and Pradhan, S. (2009). Regulation of DNMT1 stability through SET7-mediated lysine methylation in mammalian cells. Proc Natl Acad Sci U S A *106*, 5076-5081.

Ferraiuolo, M.A., Rousseau, M., Miyamoto, C., Shenker, S., Wang, X.Q., Nadler, M., Blanchette, M., and Dostie, J. (2010). The three-dimensional architecture of Hox cluster silencing. Nucleic Acids Res *38*, 7472-7484.

Ferrando, A.A., Neuberg, D.S., Staunton, J., Loh, M.L., Huard, C., Raimondi, S.C., Behm, F.G., Pui, C.H., Downing, J.R., Gilliland, D.G.*, et al.* (2002). Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. Cancer Cell *1*, 75-87.

Fiegler, H., Carr, P., Douglas, E.J., Burford, D.C., Hunt, S., Scott, C.E., Smith, J., Vetrie, D., Gorman, P., Tomlinson, I.P.*, et al.* (2003). DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. Genes Chromosomes Cancer *36*, 361-374.

Finger, L.R., Kagan, J., Christopher, G., Kurtzberg, J., Hershfield, M.S., Nowell, P.C., and Croce, C.M. (1989). Involvement of the TCL5 gene on human chromosome 1 in T-cell leukemia and melanoma. Proc Natl Acad Sci U S A *86*, 5039-5043.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391*, 806-811.

Follows, G.A., Dhami, P., Gottgens, B., Bruce, A.W., Campbell, P.J., Dillon, S.C., Smith, A.M., Koch, C., Donaldson, I.J., Scott, M.A*., et al.* (2006). Identifying gene regulatory elements by genomic microarray mapping of DNaseI hypersensitive sites. Genome Res *16*, 1310-1319.

Follows, G.A., Ferreira, R., Janes, M.E., Spensberger, D., Cambuli, F., Chaney, A.F., Kinston, S.J., Landry, J.R., Green, A.R., and Gottgens, B. (2012). Mapping and functional characterisation of a CTCF-dependent insulator element at the 3' border of the murine Scl transcriptional domain. PLoS One *7*, e31484.

Follows, G.A., Janes, M.E., Vallier, L., Green, A.R., and Gottgens, B. (2007). Real-time PCR mapping of DNaseI-hypersensitive sites using a novel ligation-mediated amplification technique. Nucleic Acids Res *35*, e56.

Foster, H.A., Abeydeera, L.R., Griffin, D.K., and Bridger, J.M. (2005). Non-random chromosome positioning in mammalian sperm nuclei, with migration of the sex chromosomes during late spermatogenesis. J Cell Sci *118*, 1811-1820.

Fraser, P., and Bickmore, W. (2007). Nuclear organization of the genome and the potential for gene regulation. Nature *447*, 413-417.

Friedman, A.D. (2007). Transcriptional control of granulocyte and monocyte development. Oncogene *26*, 6816-6828.

Frontelo, P., Manwani, D., Galdass, M., Karsunky, H., Lohmann, F., Gallagher, P.G., and Bieker, J.J. (2007). Novel role for EKLF in megakaryocyte lineage commitment. Blood *110*, 3871-3880.

Fuks, F., Hurd, P.J., Wolf, D., Nan, X., Bird, A.P., and Kouzarides, T. (2003). The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. J Biol Chem *278*, 4035-4040.

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

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H*., et al.* (2009a). An oestrogen-receptor-alpha-bound human chromatin interactome. Nature *462*, 58-64.

Fullwood, M.J., and Ruan, Y. (2009). ChIP-based methods for the identification of long-range chromatin interactions. J Cell Biochem *107*, 30-39.

Fullwood, M.J., Wei, C.L., Liu, E.T., and Ruan, Y. (2009b). Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome Res *19*, 521-532.

Galloway, J.L., Wingert, R.A., Thisse, C., Thisse, B., and Zon, L.I. (2005). Loss of gata1 but not gata2 converts erythropoiesis to myelopoiesis in zebrafish embryos. Dev Cell *8*, 109-116.

Gaszner, M., and Felsenfeld, G. (2006). Insulators: exploiting transcriptional and epigenetic mechanisms. Nature reviews Genetics *7*, 703-713.

George, K.M., Leonard, M.W., Roth, M.E., Lieuw, K.H., Kioussis, D., Grosveld, F., and Engel, J.D. (1994). Embryonic expression and cloning of the murine GATA-3 gene. Development *120*, 2673-2686.

Gheldof, N., Smith, E.M., Tabuchi, T.M., Koch, C.M., Dunham, I., Stamatoyannopoulos, J.A., and Dekker, J. (2010). Cell-type-specific long-range looping interactions identify distant regulatory elements of the CFTR gene. Nucleic Acids Res *38*, 4325-4336.

Gilbert, N., Boyle, S., Fiegler, H., Woodfine, K., Carter, N.P., and Bickmore, W.A. (2004). Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. Cell *118*, 555-566.

Giroux, S., Kaushik, A.L., Capron, C., Jalil, A., Kelaidi, C., Sablitzky, F., Dumenil, D., Albagli, O., and Godin, I. (2007). lyl-1 and tal-1/scl, two genes encoding closely related bHLH transcription factors, display highly overlapping expression patterns during cardiovascular and hematopoietic ontogeny. Gene Expr Patterns *7*, 215-226.

Godin, I., and Cumano, A. (2002). The hare and the tortoise: an embryonic haematopoietic race. Nature reviews Immunology *2*, 593-604.

Golderer, G., and Grobner, P. (1991). ADP-ribosylation of core histones and their acetylated subspecies. The Biochemical journal *277 (Pt 3)*, 607-610.

Goldfarb, A.N., Goueli, S., Mickelson, D., and Greenberg, J.M. (1992). T-cell acute lymphoblastic leukemia--the associated gene SCL/tal codes for a 42-Kd nuclear phosphoprotein. Blood *80*, 2858-2866.

Goldfarb, A.N., and Lewandowska, K. (1995). Inhibition of cellular differentiation by the SCL/tal oncoprotein: transcriptional repression by an Id-like mechanism. Blood *85*, 465-471.

Gondor, A., Rougier, C., and Ohlsson, R. (2008). High-resolution circular chromosome conformation capture assay. Nat Protoc *3*, 303-313.

Gottgens, B., Barton, L.M., Chapman, M.A., Sinclair, A.M., Knudsen, B., Grafham, D., Gilbert, J.G., Rogers, J., Bentley, D.R., and Green, A.R. (2002a). Transcriptional regulation of the stem cell leukemia gene (SCL)--comparative analysis of five vertebrate SCL loci. Genome Res *12*, 749-759.

Gottgens, B., Barton, L.M., Gilbert, J.G., Bench, A.J., Sanchez, M.J., Bahn, S., Mistry, S., Grafham, D., McMurray, A., Vaudin, M.*, et al.* (2000). Analysis of vertebrate SCL loci identifies conserved enhancers. Nature biotechnology *18*, 181-186.

Gottgens, B., Broccardo, C., Sanchez, M.J., Deveaux, S., Murphy, G., Gothert, J.R., Kotsopoulou, E., Kinston, S., Delaney, L., Piltz, S.*, et al.* (2004). The scl +18/19 stem cell enhancer is not required for hematopoiesis: identification of a 5' bifunctional hematopoietic-endothelial enhancer bound by Fli-1 and Elf-1. Mol Cell Biol *24*, 1870-1883.

Gottgens, B., Ferreira, R., Sanchez, M.J., Ishibashi, S., Li, J., Spensberger, D., Lefevre, P., Ottersbach, K., Chapman, M., Kinston, S.*, et al.* (2010). cis-Regulatory remodeling of the SCL locus during vertebrate evolution. Mol Cell Biol *30*, 5741-5751.

Gottgens, B., Gilbert, J.G., Barton, L.M., Grafham, D., Rogers, J., Bentley, D.R., and Green, A.R. (2001). Long-range comparison of human and mouse SCL loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences. Genome Res *11*, 87-97.

Gottgens, B., McLaughlin, F., Bockamp, E.O., Fordham, J.L., Begley, C.G., Kosmopoulos, K., Elefanty, A.G., and Green, A.R. (1997). Transcription of the SCL gene in erythroid and CD34 positive primitive myeloid cells is controlled by a complex network of lineage-restricted chromatin-dependent and chromatin-independent regulatory elements. Oncogene *15*, 2419-2428.

Gottgens, B., Nastos, A., Kinston, S., Piltz, S., Delabesse, E.C., Stanley, M., Sanchez, M.J., Ciau-Uitz, A., Patient, R., and Green, A.R. (2002b). Establishing the transcriptional programme for blood: the SCL stem cell enhancer is regulated by a multiprotein complex containing Ets and GATA factors. EMBO J *21*, 3039-3050.

Graf, T. (2002). Differentiation plasticity of hematopoietic cells. Blood *99*, 3089-3101.

Green, A.R., and Begley, C.G. (1992). SCL and related hemopoietic helix-loop-helix transcription factors. International journal of cell cloning *10*, 269-276.

Green, A.R., Lints, T., Visvader, J., Harvey, R., and Begley, C.G. (1992). SCL is coexpressed with GATA-1 in hemopoietic cells but is also expressed in developing brain. Oncogene *7*, 653-660.

Green, A.R., Salvaris, E., and Begley, C.G. (1991). Erythroid expression of the 'helix-loop-helix' gene, SCL. Oncogene *6*, 475-479.

Gruber, S., Haering, C.H., and Nasmyth, K. (2003). Chromosomal cohesin forms a ring. Cell *112*, 765-777.

Grutz, G.G., Bucher, K., Lavenir, I., Larson, T., Larson, R., and Rabbitts, T.H. (1998). The oncogenic T cell LIM-protein Lmo2 forms part of a DNA-binding complex specifically in immature T cells. EMBO J *17*, 4594-4605.

Guillot, P.V., Xie, S.Q., Hollinshead, M., and Pombo, A. (2004). Fixation-induced redistribution of hyperphosphorylated RNA polymerase II in the nucleus of human cells. Experimental cell research *295*, 460-468.

Guo, C., Yoon, H.S., Franklin, A., Jain, S., Ebert, A., Cheng, H.L., Hansen, E., Despo, O., Bossen, C., Vettermann, C*., et al.* (2011). CTCF-binding elements mediate control of V(D)J recombination. Nature *477*, 424-430.

Hadjur, S., Williams, L.M., Ryan, N.K., Cobb, B.S., Sexton, T., Fraser, P., Fisher, A.G., and Merkenschlager, M. (2009). Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. Nature *460*, 410-413.

Hagege, H., Klous, P., Braem, C., Splinter, E., Dekker, J., Cathala, G., de Laat, W., and Forne, T. (2007). Quantitative analysis of chromosome conformation capture assays (3C-qPCR). Nat Protoc *2*, 1722-1733.

Hakim, O., Sung, M.H., Voss, T.C., Splinter, E., John, S., Sabo, P.J., Thurman, R.E., Stamatoyannopoulos, J.A., de Laat, W., and Hager, G.L. (2011). Diverse gene reprogramming events occur in the same spatial clusters of distal regulatory elements. Genome Res *21*, 697-706.

Hall, M.A., Slater, N.J., Begley, C.G., Salmon, J.M., Van Stekelenburg, L.J., McCormack, M.P., Jane, S.M., and Curtis, D.J. (2005). Functional but abnormal adult erythropoiesis in the absence of the stem cell leukemia gene. Mol Cell Biol *25*, 6355-6362.

Hamilton, A.J., and Baulcombe, D.C. (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. Science *286*, 950-952.

Hammond, S.M., Bernstein, E., Beach, D., and Hannon, G.J. (2000). An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. Nature *404*, 293-296.

Handoko, L., Xu, H., Li, G., Ngan, C.Y., Chew, E., Schnapp, M., Lee, C.W., Ye, C., Ping, J.L., Mulawadi, F.*, et al.* (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. Nat Genet *43*, 630-638.

Harikrishnan, K.N., Chow, M.Z., Baker, E.K., Pal, S., Bassal, S., Brasacchio, D., Wang, L., Craig, J.M., Jones, P.L., Sif, S.*, et al.* (2005). Brahma links the SWI/SNF chromatin-remodeling complex with MeCP2-dependent transcriptional silencing. Nat Genet *37*, 254-264.

Hassa, P.O., Haenni, S.S., Elser, M., and Hottiger, M.O. (2006). Nuclear ADP-ribosylation reactions in mammalian cells: where are we today and where are we going? Microbiology and molecular biology reviews : MMBR *70*, 789-829.

Heinemeyer, T., Chen, X., Karas, H., Kel, A.E., Kel, O.V., Liebich, I., Meinhardt, T., Reuter, I., Schacherer, F., and Wingender, E. (1999). Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. Nucleic Acids Res *27*, 318-322.

Hellman, A., and Chess, A. (2007). Gene body-specific methylation on the active X chromosome. Science *315*, 1141-1143.

Herblot, S., Aplan, P.D., and Hoang, T. (2002). Gradient of E2A activity in B-cell development. Mol Cell Biol *22*, 886-900.

Herblot, S., Steff, A.M., Hugo, P., Aplan, P.D., and Hoang, T. (2000). SCL and LMO1 alter thymocyte differentiation: inhibition of E2A-HEB function and pre-T alpha chain expression. Nat Immunol *1*, 138-144.

Hoang, T., Paradis, E., Brady, G., Billia, F., Nakahara, K., Iscove, N.N., and Kirsch, I.R. (1996). Opposing effects of the basic helix-loop-helix transcription factor SCL on erythroid and monocytic differentiation. Blood *87*, 102-111.

Hofmann, T.J., and Cole, M.D. (1996). The TAL1/Scl basic helix-loop-helix protein blocks myogenic differentiation and E-box dependent transactivation. Oncogene *13*, 617-624.

Hoheisel, J.D. (2006). Microarray technology: beyond transcript profiling and genotype analysis. Nature reviews Genetics *7*, 200-210.

Holen, T., Amarzguioui, M., Wiiger, M.T., Babaie, E., and Prydz, H. (2002). Positional effects of short interfering RNAs targeting the human coagulation trigger Tissue Factor. Nucleic Acids Res *30*, 1757-1766.

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

Hou, C., Dale, R., and Dean, A. (2010). Cell type specificity of chromatin organization mediated by CTCF and cohesin. Proc Natl Acad Sci U S A *107*, 3651-3656.

Hou, C., Zhao, H., Tanimoto, K., and Dean, A. (2008). CTCF-dependent enhancer-blocking by alternative chromatin loop formation. Proc Natl Acad Sci U S A *105*, 20398-20403.

Hsu, H.L., Cheng, J.T., Chen, Q., and Baer, R. (1991). Enhancer-binding activity of the tal-1 oncoprotein in association with the E47/E12 helix-loop-helix proteins. Mol Cell Biol *11*, 3037-3042.

Hsu, H.L., Huang, L., Tsan, J.T., Funk, W., Wright, W.E., Hu, J.S., Kingston, R.E., and Baer, R. (1994a). Preferred sequences for DNA recognition by the TAL1 helix-loop-helix proteins. Mol Cell Biol *14*, 1256-1265.

Hsu, H.L., Wadman, I., Tsan, J.T., and Baer, R. (1994b). Positive and negative transcriptional control by the TAL1 helix-loop-helix protein. Proc Natl Acad Sci U S A *91*, 5947-5951.

Hu, X., Li, X., Valverde, K., Fu, X., Noguchi, C., Qiu, Y., and Huang, S. (2009). LSD1-mediated epigenetic modification is required for TAL1 function and hematopoiesis. Proc Natl Acad Sci U S A *106*, 10141-10146.

Huang, B., Wang, W., Bates, M., and Zhuang, X. (2008). Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy. Science *319*, 810-813.

Huang, S., Qiu, Y., Shi, Y., Xu, Z., and Brandt, S.J. (2000). P/CAF-mediated acetylation regulates the function of the basic helix-loop-helix transcription factor TAL1/SCL. EMBO J *19*, 6792-6803.

Huertas, D., Sendra, R., and Munoz, P. (2009). Chromatin dynamics coupled to DNA repair. Epigenetics : official journal of the DNA Methylation Society *4*, 31-42.

Hughes, T.R., Mao, M., Jones, A.R., Burchard, J., Marton, M.J., Shannon, K.W., Lefkowitz, S.M., Ziman, M., Schelter, J.M., Meyer, M.R.*, et al.* (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nature biotechnology *19*, 342-347.

Hwang, L.Y., Siegelman, M., Davis, L., Oppenheimer-Marks, N., and Baer, R. (1993). Expression of the TAL1 proto-oncogene in cultured endothelial cells and blood vessels of the spleen. Oncogene *8*, 3043-3046.

Iborra, F.J., Pombo, A., Jackson, D.A., and Cook, P.R. (1996). Active RNA polymerases are localized within discrete transcription "factories' in human nuclei. J Cell Sci *109 (Pt 6)*, 1427-1436.

Jackson, D.A., Hassan, A.B., Errington, R.J., and Cook, P.R. (1993). Visualization of focal sites of transcription within human nuclei. EMBO J *12*, 1059-1065.

Janssen, J.W., Ludwig, W.D., Sterry, W., and Bartram, C.R. (1993). SIL-TAL1 deletion in T-cell acute lymphoblastic leukemia. Leukemia *7*, 1204-1210.

Jazag, A., Ijichi, H., Kanai, F., Imamura, T., Guleng, B., Ohta, M., Imamura, J., Tanaka, Y., Tateishi, K., Ikenoue, T.*, et al.* (2005). Smad4 silencing in pancreatic cancer cell lines using stable RNA interference and gene expression profiles induced by transforming growth factor-beta. Oncogene *24*, 662-671.

Jeong, S., Liang, G., Sharma, S., Lin, J.C., Choi, S.H., Han, H., Yoo, C.B., Egger, G., Yang, A.S., and Jones, P.A. (2009). Selective anchoring of DNA methyltransferases 3A and 3B to nucleosomes containing methylated DNA. Mol Cell Biol *29*, 5366-5376.

Jiang, J., Chan, Y.S., Loh, Y.H., Cai, J., Tong, G.Q., Lim, C.A., Robson, P., Zhong, S., and Ng, H.H. (2008). A core Klf circuitry regulates self-renewal of embryonic stem cells. Nat Cell Biol *10*, 353-360.

Johansen, K.M., and Johansen, J. (2006). Regulation of chromatin structure by histone H3S10 phosphorylation. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology *14*, 393-404.

Jonsson, O.G., Kitchens, R.L., Baer, R.J., Buchanan, G.R., and Smith, R.G. (1991). Rearrangements of the tal-1 locus as clonal markers for T cell acute lymphoblastic leukemia. J Clin Invest *87*, 2029-2035.

Kacem, S., and Feil, R. (2009). Chromatin mechanisms in genomic imprinting. Mamm Genome *20*, 544-556.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S.*, et al.* (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430-435.

Kallianpur, A.R., Jordan, J.E., and Brandt, S.J. (1994). The SCL/TAL-1 gene is expressed in progenitors of both the hematopoietic and vascular systems during embryogenesis. Blood *83*, 1200-1208.

Kamath, M.B., Houston, I.B., Janovski, A.J., Zhu, X., Gowrisankar, S., Jegga, A.G., and DeKoter, R.P. (2008). Dose-dependent repression of T-cell and natural killer cell genes by PU.1 enforces myeloid and B-cell identity. Leukemia *22*, 1214-1225.

Kang, H., and Lieberman, P.M. (2009). Cell cycle control of Kaposi's sarcoma-associated herpesvirus latency transcription by CTCF-cohesin interactions. Journal of virology *83*, 6199-6210.

Karlic, R., Chung, H.R., Lasserre, J., Vlahovicek, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. Proc Natl Acad Sci U S A *107*, 2926-2931.

Kassouf, M.T., Chagraoui, H., Vyas, P., and Porcher, C. (2008). Differential use of SCL/TAL-1 DNA-binding domain in developmental hematopoiesis. Blood *112*, 1056-1067.

Kassouf, M.T., Hughes, J.R., Taylor, S., McGowan, S.J., Soneji, S., Green, A.L., Vyas, P., and Porcher, C. (2010). Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. Genome Res *20*, 1064-1083.

Kennedy, M., Firpo, M., Choi, K., Wall, C., Robertson, S., Kabrun, N., and Keller, G. (1997). A common precursor for primitive erythropoiesis and definitive haematopoiesis. Nature *386*, 488-493.

Kerenyi, M.A., and Orkin, S.H. (2010). Networking erythropoiesis. J Exp Med *207*, 2537-2541.

Keshet, I., Schlesinger, Y., Farkash, S., Rand, E., Hecht, M., Segal, E., Pikarski, E., Young, R.A., Niveleau, A., Cedar, H.*, et al.* (2006). Evidence for an instructive mechanism of de novo methylation in cancer cells. Nat Genet *38*, 149-153.

Kikuchi, A., Hayashi, Y., Kobayashi, S., Hanada, R., Moriwaki, K., Yamamoto, K., Fujimoto, J., Kaneko, Y., and Yamamori, S. (1993). Clinical significance of TAL1 gene alteration in childhood T-cell acute lymphoblastic leukemia and lymphoma. Leukemia *7*, 933-938.

Kim, D.H., and Rossi, J.J. (2007). Strategies for silencing human disease using RNA interference. Nature reviews Genetics *8*, 173-184.

Kim, S., Kim, Y.W., Shim, S.H., Kim, C.G., and Kim, A. (2012). Chromatin structure of the LCR in the human beta-globin locus transcribing the adult delta- and beta-globin genes. The international journal of biochemistry & cell biology *44*, 505-513.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V., and Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell *128*, 1231-1245.

Komura, J., Ikehata, H., and Ono, T. (2007). Chromatin fine structure of the c-MYC insulator element/DNase I-hypersensitive site I is not preserved during mitosis. Proc Natl Acad Sci U S A *104*, 15741-15746.

Kouzarides, T. (2007). Chromatin modifications and their function. Cell *128*, 693-705.

Kumar, P.P., Bischof, O., Purbey, P.K., Notani, D., Urlaub, H., Dejean, A., and Galande, S. (2007). Functional interaction between PML and SATB1 regulates chromatin-loop architecture and transcription of the MHC class I locus. Nat Cell Biol *9*, 45-56.

Kumar, S., and Hedges, S.B. (1998). A molecular timescale for vertebrate evolution. Nature *392*, 917-920.

Kumaravelu, P., Hook, L., Morrison, A.M., Ure, J., Zhao, S., Zuyev, S., Ansell, J., and Medvinsky, A. (2002). Quantitative developmental anatomy of definitive haematopoietic stem cells/long-term repopulating units (HSC/RUs): role of the aorta-gonad-mesonephros (AGM) region and the yolk sac in colonisation of the mouse embryonic liver. Development *129*, 4891-4899.

Kuo, S.S., Mellentin, J.D., Copeland, N.G., Gilbert, D.J., Jenkins, N.A., and Cleary, M.L. (1991). Structure, chromosome mapping, and expression of the mouse Lyl-1 gene. Oncogene *6*, 961-968.

Kuroda, A., Rauch, T.A., Todorov, I., Ku, H.T., Al-Abdullah, I.H., Kandeel, F., Mullen, Y., Pfeifer, G.P., and Ferreri, K. (2009). Insulin gene expression is regulated by DNA methylation. PLoS One *4*, e6953.

Kurukuti, S., Tiwari, V.K., Tavoosidana, G., Pugacheva, E., Murrell, A., Zhao, Z., Lobanenkov, V., Reik, W., and Ohlsson, R. (2006). CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. Proc Natl Acad Sci U S A *103*, 10684-10689.

Kwong, Y.L., Chan, D., and Liang, R. (1995). SIL/TAL1 recombination in adult T-acute lymphoblastic leukemia and T-lymphoblastic lymphoma. Cancer Genet Cytogenet *85*, 159-160.

Labastie, M.C., Cortes, F., Romeo, P.H., Dulac, C., and Peault, B. (1998). Molecular identity of hematopoietic precursor cells emerging in the human embryo. Blood *92*, 3624-3635.

Lacombe, J., Herblot, S., Rojas-Sutterlin, S., Haman, A., Barakat, S., Iscove, N.N., Sauvageau, G., and Hoang, T. (2010). Scl regulates the quiescence and the long-term competence of hematopoietic stem cells. Blood *115*, 792-803.

Lahlil, R., Lecuyer, E., Herblot, S., and Hoang, T. (2004). SCL assembles a multifactorial complex that determines glycophorin A expression. Mol Cell Biol *24*, 1439-1452.

Lancrin, C., Sroczynska, P., Stephenson, C., Allen, T., Kouskoff, V., and Lacaud, G. (2009). The haemangioblast generates haematopoietic cells through a haemogenic endothelium stage. Nature *457*, 892-895.

Lanctot, C., Cheutin, T., Cremer, M., Cavalli, G., and Cremer, T. (2007). Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. Nature reviews Genetics *8*, 104-115.

Larmonie, N.S., Dik, W.A., van der Velden, V.H., Hoogeveen, P.G., Beverloo, H.B., Meijerink, J.P., van Dongen, J.J., and Langerak, A.W. (2011). Correct interpretation of T-ALL oncogene expression relies on normal human thymocyte subsets as reference material. Br J Haematol.

Larson, R.C., Lavenir, I., Larson, T.A., Baer, R., Warren, A.J., Wadman, I., Nottage, K., and Rabbitts, T.H. (1996). Protein dimerization between Lmo2 (Rbtn2) and Tal1 alters thymocyte development and potentiates T cell tumorigenesis in transgenic mice. EMBO J *15*, 1021-1027.

Lassar, A.B., Davis, R.L., Wright, W.E., Kadesch, T., Murre, C., Voronova, A., Baltimore, D., and Weintraub, H. (1991). Functional activity of myogenic HLH proteins requires hetero-oligomerization with E12/E47-like proteins in vivo. Cell *66*, 305-315.

LaVoie, H.A. (2003). The role of GATA in mammalian reproduction. Exp Biol Med (Maywood) *228*, 1282-1290.

Lecointe, N., Bernard, O., Naert, K., Joulin, V., Larsen, C.J., Romeo, P.H., and Mathieu-Mahul, D. (1994). GATA-and SP1-binding sites are required for the full activity of the tissue-specific promoter of the tal-1 gene. Oncogene *9*, 2623-2632.

Lecuyer, E., Herblot, S., Saint-Denis, M., Martin, R., Begley, C.G., Porcher, C., Orkin, S.H., and Hoang, T. (2002). The SCL complex regulates c-kit expression in hematopoietic cells through functional interaction with Sp1. Blood *100*, 2430-2440.

Lee, T.I., and Young, R.A. (2000). Transcription of eukaryotic protein-coding genes. Annual review of genetics *34*, 77-137.

Lenhard, B., Sandelin, A., Mendoza, L., Engstrom, P., Jareborg, N., and Wasserman, W.W. (2003). Identification of conserved regulatory elements by comparative genome analysis. Journal of biology *2*, 13.

Leroy-Viard, K., Vinit, M.A., Lecointe, N., Mathieu-Mahul, D., and Romeo, P.H. (1994). Distinct DNase-I hypersensitive sites are associated with TAL-1 transcription in erythroid and T-cell lines. Blood *84*, 3819-3827.

Li, B., Carey, M., and Workman, J.L. (2007a). The role of chromatin during transcription. Cell *128*, 707-719.

Li, C.Y., Zhan, Y.Q., Li, W., Xu, C.W., Xu, W.X., Yu, D.H., Peng, R.Y., Cui, Y.F., Yang, X., Hou, N.*, et al.* (2007b). Overexpression of a hematopoietic transcriptional regulator EDAG induces myelopoiesis and suppresses lymphopoiesis in transgenic mice. Leukemia *21*, 2277-2286.

Li, G., Fullwood, M.J., Xu, H., Mulawadi, F.H., Velkov, S., Vega, V., Ariyaratne, P.N., Mohamed, Y.B., Ooi, H.S., Tennakoon, C.*, et al.* (2010). ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biol *11*, R22.

Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J.*, et al.* (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell *148*, 84-98.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O.*, et al.* (2009). Comprehensive

mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289-293.

Lieu, P.T., Jozsi, P., Gilles, P., and Peterson, T. (2005). Development of a DNA-labeling system for array-based comparative genomic hybridization. Journal of biomolecular techniques : JBT *16*, 104-111.

Lipshutz, R.J., Fodor, S.P., Gingeras, T.R., and Lockhart, D.J. (1999). High density synthetic oligonucleotide arrays. Nat Genet *21*, 20-24.

Liu, X.S., Brutlag, D.L., and Liu, J.S. (2002). An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nature biotechnology *20*, 835-839.

Lomvardas, S., Barnea, G., Pisapia, D.J., Mendelsohn, M., Kirkland, J., and Axel, R. (2006). Interchromosomal interactions and olfactory receptor choice. Cell *126*, 403-413.

Lopez-Serra, L., and Esteller, M. (2008). Proteins that bind methylated DNA and human cancer: reading the wrong words. British journal of cancer *98*, 1881-1885.

Lorincz, M.C., Dickerson, D.R., Schmitt, M., and Groudine, M. (2004). Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. Nature structural & molecular biology *11*, 1068-1075.

Louwers, M., Bader, R., Haring, M., van Driel, R., de Laat, W., and Stam, M. (2009a). Tissue- and expression level-specific chromatin looping at maize b1 epialleles. Plant Cell *21*, 832-842.

Louwers, M., Splinter, E., van Driel, R., de Laat, W., and Stam, M. (2009b). Studying physical chromatin interactions in plants using Chromosome Conformation Capture (3C). Nat Protoc *4*, 1216-1229.

Lozzio, C.B., and Lozzio, B.B. (1975). Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. Blood *45*, 321-334.

Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J., Pereira-Smith, O.M., and Misteli, T. (2010). Regulation of alternative splicing by histone modifications. Science *327*, 996-1000.

Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature *389*, 251-260.

Majumder, P., Gomez, J.A., Chadwick, B.P., and Boss, J.M. (2008). The insulator factor CTCF controls MHC class II gene expression and is required for the formation of long-distance chromatin interactions. J Exp Med *205*, 785-798.

Maniatis, T., Goodbourn, S., and Fischer, J.A. (1987). Regulation of inducible and tissue-specific gene expression. Science *236*, 1237-1245.

Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional regulatory elements in the human genome. Annual review of genomics and human genetics *7*, 29-59.

Matzke, M.A., and Birchler, J.A. (2005). RNAi-mediated pathways in the nucleus. Nature reviews Genetics *6*, 24-35.

Mayer, L., Kazantzidis, S., Escala, A., and Callegari, S. (2010). Direct formation of supermassive black holes via multi-scale gas inflows in galaxy mergers. Nature *466*, 1082-1084.

McCormack, M.P., Hall, M.A., Schoenwaelder, S.M., Zhao, Q., Ellis, S., Prentice, J.A., Clarke, A.J., Slater, N.J., Salmon, J.M., Jackson, S.P.*, et al.* (2006). A critical role for the transcription factor Scl in platelet production during stress thrombopoiesis. Blood *108*, 2248-2256.

McCormack, M.P., Young, L.F., Vasudevan, S., de Graaf, C.A., Codrington, R., Rabbitts, T.H., Jane, S.M., and Curtis, D.J. (2010). The Lmo2 oncogene initiates leukemia in mice by inducing thymocyte self-renewal. Science *327*, 879-883.

McManus, M.T., and Sharp, P.A. (2002). Gene silencing in mammals by small interfering RNAs. Nature reviews Genetics *3*, 737-747.

Meaburn, K.J., and Misteli, T. (2007). Cell biology: chromosome territories. Nature *445*, 379-781.

Mead, P.E., Kelley, C.M., Hahn, P.S., Piedad, O., and Zon, L.I. (1998). SCL specifies hematopoietic mesoderm in Xenopus embryos. Development *125*, 2611-2620.

Mellentin, J.D., Smith, S.D., and Cleary, M.L. (1989). lyl-1, a novel gene altered by chromosomal translocation in T cell leukemia, codes for a protein with a helix-loop-helix DNA binding motif. Cell *58*, 77-83.

Meng, Y.S., Hu, X.J., and Liu, W. (2009). Knockdown of LYL1 impaired proliferation of CD34(+) myeloid leukemia cells. Leukemia & lymphoma *50*, 1896-1899.

Meng, Y.S., Khoury, H., Dick, J.E., and Minden, M.D. (2005). Oncogenic potential of the transcription factor LYL1 in acute myeloblastic leukemia. Leukemia *19*, 1941-1947.

Miele, A., and Dekker, J. (2009). Mapping cis- and trans- chromatin interaction networks using chromosome conformation capture (3C). Methods Mol Biol *464*, 105-121.

Miele, A., Gheldof, N., Tabuchi, T.M., Dostie, J., and Dekker, J. (2006). Mapping chromatin interactions by chromosome conformation capture. Curr Protoc Mol Biol *Chapter 21*, Unit 21 11.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P.*, et al.* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553-560.

Mikkola, H.K., Klintman, J., Yang, H., Hock, H., Schlaeger, T.M., Fujiwara, Y., and Orkin, S.H. (2003). Haematopoietic stem cells retain long-term repopulating activity and multipotency in the absence of stem-cell leukaemia SCL/tal-1 gene. Nature *421*, 547-551.

Mishiro, T., Ishihara, K., Hino, S., Tsutsumi, S., Aburatani, H., Shirahige, K., Kinoshita, Y., and Nakao, M. (2009). Architectural roles of multiple chromatin insulators at the human apolipoprotein gene cluster. EMBO J *28*, 1234-1245.

Mitchell, J.A., and Fraser, P. (2008). Transcription factories are nuclear subcompartments that remain in the absence of transcription. Genes Dev *22*, 20-25.

Mollica, L.R., Crawley, J.T., Liu, K., Rance, J.B., Cockerill, P.N., Follows, G.A., Landry, J.R., Wells, D.J., and Lane, D.A. (2006). Role of a 5'-enhancer in the transcriptional regulation of the human endothelial cell protein C receptor gene. Blood *108*, 1251-1259.

Morey, C., Da Silva, N.R., Perry, P., and Bickmore, W.A. (2007). Nuclear reorganisation and chromatin decondensation are conserved, but distinct, mechanisms linked to Hox gene activation. Development *134*, 909-919.

Mosher, R.A., and Melnyk, C.W. (2010). siRNAs and DNA methylation: seedy epigenetics. Trends in plant science *15*, 204-210.

Mouthon, M.A., Bernard, O., Mitjavila, M.T., Romeo, P.H., Vainchenker, W., and Mathieu-Mahul, D. (1993). Expression of tal-1 and GATA-binding proteins during human hematopoiesis. Blood *81*, 647-655.

Mukai, H.Y., Motohashi, H., Ohneda, O., Suzuki, N., Nagano, M., and Yamamoto, M. (2006). Transgene insertion in proximity to the c-myb gene disrupts erythroid-megakaryocytic lineage bifurcation. Mol Cell Biol *26*, 7953-7965.

Murre, C., McCaw, P.S., and Baltimore, D. (1989). A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins. Cell *56*, 777-783.

Murrell, A., Heeson, S., and Reik, W. (2004). Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. Nat Genet *36*, 889-893.

Myers, L.C., and Kornberg, R.D. (2000). Mediator of transcriptional regulation. Annual review of biochemistry *69*, 729-749.

Nagel, S., Venturini, L., Meyer, C., Kaufmann, M., Scherr, M., Drexler, H.G., and MacLeod, R.A. (2010). Multiple mechanisms induce ectopic expression of LYL1 in subsets of T-ALL cell lines. Leuk Res *34*, 521-528.

Nakanishi, S., Lee, J.S., Gardner, K.E., Gardner, J.M., Takahashi, Y.H., Chandrasekharan, M.B., Sun, Z.W., Osley, M.A., Strahl, B.D., Jaspersen, S.L.*, et al.* (2009). Histone H2BK123 monoubiquitination is the critical determinant for H3K4 and H3K79 trimethylation by COMPASS and Dot1. J Cell Biol *186*, 371-377.

Nativio, R., Wendt, K.S., Ito, Y., Huddleston, J.E., Uribe-Lewis, S., Woodfine, K., Krueger, C., Reik, W., Peters, J.M., and Murrell, A. (2009). Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus. PLoS Genet *5*, e1000739.

Ng, H.H., Robert, F., Young, R.A., and Struhl, K. (2003). Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. Mol Cell *11*, 709-719.

Nishida, H., Suzuki, T., Kondo, S., Miura, H., Fujimura, Y., and Hayashizaki, Y. (2006). Histone H3 acetylated at lysine 9 in promoter is associated with low nucleosome density in the vicinity of transcription start site in human cell. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology *14*, 203-211.

Nishimura, S., Takahashi, S., Kuroha, T., Suwabe, N., Nagasawa, T., Trainor, C., and Yamamoto, M. (2000). A GATA box in the GATA-1 gene hematopoietic enhancer is a critical element in the network of GATA factors and sites that regulate this gene. Mol Cell Biol *20*, 713-723.

Noordermeer, D., Leleu, M., Splinter, E., Rougemont, J., De Laat, W., and Duboule, D. (2011). The dynamic architecture of Hox gene clusters. Science *334*, 222-225.

O'Neil, J., Billa, M., Oikemus, S., and Kelliher, M. (2001). The DNA binding activity of TAL-1 is not required to induce leukemia/lymphoma in mice. Oncogene *20*, 3897-3905.

O'Neil, J., Shank, J., Cusson, N., Murre, C., and Kelliher, M. (2004). TAL1/SCL induces leukemia by inhibiting the transcriptional activity of E47/HEB. Cancer Cell *5*, 587-596.

O'Neill, L.P., and Turner, B.M. (1996). Immunoprecipitation of chromatin. Methods in enzymology *274*, 189-197.

Ogilvy, S., Ferreira, R., Piltz, S.G., Bowen, J.M., Gottgens, B., and Green, A.R. (2007). The SCL +40 enhancer targets the midbrain together with primitive and definitive hematopoiesis and is regulated by SCL and GATA proteins. Mol Cell Biol *27*, 7206-7219.

Onn, I., Heidinger-Pauli, J.M., Guacci, V., Unal, E., and Koshland, D.E. (2008). Sister chromatid cohesion: a simple concept with a complex reality. Annual review of cell and developmental biology *24*, 105-129.

Ono, Y., Fukuhara, N., and Yoshie, O. (1998). TAL1 and LIM-only proteins synergistically induce retinaldehyde dehydrogenase 2 expression in T-cell acute lymphoblastic leukemia by acting as cofactors for GATA3. Mol Cell Biol *18*, 6939-6950.

Ooi, S.K., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S.P., Allis, C.D.*, et al.* (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature *448*, 714-717.

Orkin, S.H. (2000). Diversification of haematopoietic stem cells to specific lineages. Nature reviews Genetics *1*, 57-64.

Orlando, V., Strutt, H., and Paro, R. (1997). Analysis of chromatin structure by in vivo formaldehyde cross-linking. Methods *11*, 205-214.

Osborne, C.S., Chakalova, L., Brown, K.E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J.A., Lopes, S., Reik, W.*, et al.* (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. Nat Genet *36*, 1065-1071.

Osborne, C.S., Chakalova, L., Mitchell, J.A., Horton, A., Wood, A.L., Bolland, D.J., Corcoran, A.E., and Fraser, P. (2007). Myc dynamically and preferentially relocates to a transcription factory occupied by Igh. PLoS biology *5*, e192.

Palstra, R.J., Tolhuis, B., Splinter, E., Nijmeijer, R., Grosveld, F., and de Laat, W. (2003). The beta-globin nuclear compartment in development and erythroid differentiation. Nat Genet *35*, 190-194.

Panigrahi, A.K., and Pati, D. (2012). Higher-order orchestration of hematopoiesis: is cohesin a new player? Exp Hematol *40*, 967-973.

Papantonis, A., Larkin, J.D., Wada, Y., Ohta, Y., Ihara, S., Kodama, T., and Cook, P.R. (2010). Active RNA polymerases: mobile or immobile molecular machines? PLoS biology *8*, e1000419.

Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H.C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T.*, et al.* (2008). Cohesins functionally associate with CTCF on mammalian chromosome arms. Cell *132*, 422-433.

Park, J.C., Chae, Y.K., Son, C.H., Kim, M.S., Lee, J., Ostrow, K., Sidransky, D., Hoque, M.O., and Moon, C. (2008). Epigenetic silencing of human T (brachyury homologue) gene in non-small-cell lung cancer. Biochem Biophys Res Commun *365*, 221-226.

Pati, D., Zhang, N., and Plon, S.E. (2002). Linking sister chromatid cohesion and apoptosis: role of Rad21. Mol Cell Biol *22*, 8267-8277.

Pekowska, A., Benoukraf, T., Zacarias-Cabeza, J., Belhocine, M., Koch, F., Holota, H., Imbert, J., Andrau, J.C., Ferrier, P., and Spicuglia, S. (2011). H3K4 tri-methylation provides an epigenetic signature of active enhancers. EMBO J *30*, 4198-4210.

Pennacchio, L.A., and Rubin, E.M. (2001). Genomic strategies to identify mammalian regulatory sequences. Nature reviews Genetics *2*, 100-109.

Phillips, J.E., and Corces, V.G. (2009). CTCF: master weaver of the genome. Cell *137*, 1194-1211.

Pina, C., May, G., Soneji, S., Hong, D., and Enver, T. (2008). MLLT3 regulates early human erythroid and megakaryocytic cell fate. Cell stem cell *2*, 264-273.

Pirot, N., Deleuze, V., El-Hajj, R., Dohet, C., Sablitzky, F., Couttet, P., Mathieu, D., and Pinet, V. (2010). LYL1 activity is required for the maturation of newly formed blood vessels in adulthood. Blood *115*, 5270-5279.

Pokholok, D.K., Harbison, C.T., Levine, S., Cole, M., Hannett, N.M., Lee, T.I., Bell, G.W., Walker, K., Rolfe, P.A., Herbolsheimer, E.*, et al.* (2005). Genome-wide map of nucleosome acetylation and methylation in yeast. Cell *122*, 517-527.

Pongubala, J.M., Northrup, D.L., Lancki, D.W., Medina, K.L., Treiber, T., Bertolino, E., Thomas, M., Grosschedl, R., Allman, D., and Singh, H. (2008). Transcription factor EBF restricts alternative lineage options and promotes B cell fate commitment independently of Pax5. Nat Immunol *9*, 203-215.

Porcher, C., Liao, E.C., Fujiwara, Y., Zon, L.I., and Orkin, S.H. (1999). Specification of hematopoietic and vascular development by the bHLH transcription factor SCL without direct DNA binding. Development *126*, 4603-4615.

Porcher, C., Swat, W., Rockwell, K., Fujiwara, Y., Alt, F.W., and Orkin, S.H. (1996). The T cell leukemia oncoprotein SCL/tal-1 is essential for development of all hematopoietic lineages. Cell *86*, 47-57.

Postberg, J., Alexandrova, O., Cremer, T., and Lipps, H.J. (2005). Exploiting nuclear duality of ciliates to analyse topological requirements for DNA replication and transcription. J Cell Sci *118*, 3973-3983.

Prelich, G. (2002). RNA polymerase II carboxy-terminal domain kinases: emerging clues to their function. Eukaryotic cell *1*, 153-162.

Ptashne, M., and Gann, A. (1997). Transcriptional activation by recruitment. Nature *386*, 569-577.

Pulford, K., Lecointe, N., Leroy-Viard, K., Jones, M., Mathieu-Mahul, D., and Mason, D.Y. (1995). Expression of TAL-1 proteins in human tissues. Blood *85*, 675-684.

Raab, J.R., Chiu, J., Zhu, J., Katzman, S., Kurukuti, S., Wade, P.A., Haussler, D., and Kamakaka, R.T. (2012). Human tRNA genes function as chromatin insulators. EMBO J *31*, 330-350.

Rabbitts, T.H. (1994). Chromosomal translocations in human cancer. Nature *372*, 143-149.

Raghuraman, M.K., Winzeler, E.A., Collingwood, D., Hunt, S., Wodicka, L., Conway, A., Lockhart, D.J., Davis, R.W., Brewer, B.J., and Fangman, W.L. (2001). Replication dynamics of the yeast genome. Science *294*, 115-121.

Rando, O.J., and Chang, H.Y. (2009). Genome-wide views of chromatin structure. Annual review of biochemistry *78*, 245-271.

Richards, E.J., and Elgin, S.C. (2002). Epigenetic codes for heterochromatin formation and silencing: rounding up the usual suspects. Cell *108*, 489-500.

Rippe, K. (2001). Making contacts on a nucleic acid polymer. Trends Biochem Sci *26*, 733-740.

Robb, L., Elwood, N.J., Elefanty, A.G., Kontgen, F., Li, R., Barnett, L.D., and Begley, C.G. (1996). The scl gene product is required for the generation of all hematopoietic lineages in the adult mouse. EMBO J *15*, 4123-4129.

Robb, L., Lyons, I., Li, R., Hartley, L., Kontgen, F., Harvey, R.P., Metcalf, D., and Begley, C.G. (1995). Absence of yolk sac hematopoiesis from mice with a targeted disruption of the scl gene. Proc Natl Acad Sci U S A *92*, 7075-7079.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A.*, et al.* (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods *4*, 651-657.

Robertson, S.M., Kennedy, M., Shannon, J.M., and Keller, G. (2000). A transitional stage in the commitment of mesoderm to hematopoiesis requiring the transcription factor SCL/tal-1. Development *127*, 2447-2459.

Rodley, C.D., Bertels, F., Jones, B., and O'Sullivan, J.M. (2009). Global identification of yeast chromosome interactions using Genome conformation capture. Fungal Genet Biol *46*, 879-886.

Rodriguez, P., Bonte, E., Krijgsveld, J., Kolodziej, K.E., Guyot, B., Heck, A.J., Vyas, P., de Boer, E., Grosveld, F., and Strouboulis, J. (2005). GATA-1 forms distinct activating and repressive complexes in erythroid cells. EMBO J *24*, 2354-2366.

Roh, T.Y., Cuddapah, S., Cui, K., and Zhao, K. (2006). The genomic landscape of histone modifications in human T cells. Proc Natl Acad Sci U S A *103*, 15782-15787.

Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nature biotechnology *16*, 939-945.

Rubio, E.D., Reiss, D.J., Welcsh, P.L., Disteche, C.M., Filippova, G.N., Baliga, N.S., Aebersold, R., Ranish, J.A., and Krumm, A. (2008). CTCF physically links cohesin to chromatin. Proc Natl Acad Sci U S A *105*, 8309-8314.

Salmon, J.M., Slater, N.J., Hall, M.A., McCormack, M.P., Nutt, S.L., Jane, S.M., and Curtis, D.J. (2007). Aberrant mast-cell differentiation in mice lacking the stem-cell leukemia gene. Blood *110*, 3573-3581.

San-Marina, S., Han, Y., Liu, J., and Minden, M.D. (2012). Suspected leukemia oncoproteins CREB1 and LYL1 regulate Op18/STMN1 expression. Biochim Biophys Acta *1819*, 1164-1172.

San-Marina, S., Han, Y., Suarez Saiz, F., Trus, M.R., and Minden, M.D. (2008). Lyl1 interacts with CREB1 and alters expression of CREB1 target genes. Biochim Biophys Acta *1783*, 503-517.

Sanchez, M., Gottgens, B., Sinclair, A.M., Stanley, M., Begley, C.G., Hunter, S., and Green, A.R. (1999). An SCL 3' enhancer targets developing endothelium together with embryonic and adult haematopoietic progenitors. Development *126*, 3891-3904.

Sanchez, M.J., Bockamp, E.O., Miller, J., Gambardella, L., and Green, A.R. (2001). Selective rescue of early haematopoietic progenitors in Scl(-/-) mice by expressing Scl under the control of a stem cell enhancer. Development *128*, 4815-4827.

Santos-Rosa, H., Schneider, R., Bannister, A.J., Sherriff, J., Bernstein, B.E., Emre, N.C., Schreiber, S.L., Mellor, J., and Kouzarides, T. (2002). Active genes are tri-methylated at K4 of histone H3. Nature *419*, 407-411.

Sawan, C., and Herceg, Z. (2010). Histone modifications and cancer. Advances in genetics *70*, 57-85.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*, 467-470.

Schlaeger, T.M., Schuh, A., Flitter, S., Fisher, A., Mikkola, H., Orkin, S.H., Vyas, P., and Porcher, C. (2004). Decoding hematopoietic specificity in the helix-loop-helix domain of the transcription factor SCL/Tal-1. Mol Cell Biol *24*, 7491-7502.

Schneider, R., Bannister, A.J., Myers, F.A., Thorne, A.W., Crane-Robinson, C., and Kouzarides, T. (2004). Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. Nat Cell Biol *6*, 73-77.

Schoenfelder, S., Clay, I., and Fraser, P. (2010a). The transcriptional interactome: gene expression in 3D. Current opinion in genetics & development *20*, 127-133.

Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N.F., Horton, A., Andrews, S., Kurukuti, S., Mitchell, J.A., Umlauf, D., Dimitrova, D.S.*, et al.* (2010b). Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. Nat Genet *42*, 53-61.

Schones, D.E., Cui, K., Cuddapah, S., Roh, T.Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. Cell *132*, 887-898.

Schubeler, D., MacAlpine, D.M., Scalzo, D., Wirbelauer, C., Kooperberg, C., van Leeuwen, F., Gottschling, D.E., O'Neill, L.P., Turner, B.M., Delrow, J.*, et al.* (2004). The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. Genes Dev *18*, 1263-1271.

Schubeler, D., Scalzo, D., Kooperberg, C., van Steensel, B., Delrow, J., and Groudine, M. (2002). Genome-wide DNA replication profile for Drosophila melanogaster: a link between transcription and replication timing. Nat Genet *32*, 438-442.

Schurter, B.T., Koh, S.S., Chen, D., Bunick, G.J., Harp, J.M., Hanson, B.L., Henschen-Edman, A., Mackay, D.R., Stallcup, M.R., and Aswad, D.W. (2001). Methylation of histone H3 by coactivator-associated arginine methyltransferase 1. Biochemistry *40*, 5747-5756.

Seita, J., and Weissman, I.L. (2010). Hematopoietic stem cell: self-renewal versus differentiation. Wiley interdisciplinary reviews Systems biology and medicine *2*, 640-653.

Sexton, T., Kurukuti, S., Mitchell, J.A., Umlauf, D., Nagano, T., and Fraser, P. (2012). Sensitive detection of chromatin coassociations using enhanced chromosome conformation capture on chip. Nat Protoc *7*, 1335-1350.

Shahbazian, M.D., and Grunstein, M. (2007). Functions of site-specific histone acetylation and deacetylation. Annual review of biochemistry *76*, 75-100.

Shiio, Y., and Eisenman, R.N. (2003). Histone sumoylation is associated with transcriptional repression. Proc Natl Acad Sci U S A *100*, 13225-13230.

Shilatifard, A. (2006). Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression. Annual review of biochemistry *75*, 243-269.

Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S., *et al.* (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature medicine *8*, 68-74.

Shivdasani, R.A., Mayer, E.L., and Orkin, S.H. (1995). Absence of blood formation in mice lacking the T-cell leukaemia oncoprotein tal-1/SCL. Nature *373*, 432-434.

Shogren-Knaak, M., Ishii, H., Sun, J.M., Pazin, M.J., Davie, J.R., and Peterson, C.L. (2006). Histone H4-K16 acetylation controls chromatin structure and protein interactions. Science *311*, 844-847.

Silberstein, L., Sanchez, M.J., Socolovsky, M., Liu, Y., Hoffman, G., Kinston, S., Piltz, S., Bowen, M., Gambardella, L., Green, A.R., *et al.* (2005). Transgenic analysis of the stem cell leukemia +19 stem cell enhancer in adult and embryonic hematopoietic and endothelial cells. Stem Cells *23*, 1378-1388.

Silva, J.M., Li, M.Z., Chang, K., Ge, W., Golding, M.C., Rickles, R.J., Siolas, D., Hu, G., Paddison, P.J., Schlabach, M.R., *et al.* (2005). Second-generation shRNA libraries covering the mouse and human genomes. Nat Genet *37*, 1281-1288.

Simonis, M., Klous, P., Homminga, I., Galjaard, R.J., Rijkers, E.J., Grosveld, F., Meijerink, J.P., and de Laat, W. (2009). High-resolution identification of balanced and complex chromosomal rearrangements by 4C technology. Nat Methods *6*, 837-842.

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

Simonis, M., Kooren, J., and de Laat, W. (2007). An evaluation of 3C-based methods to capture DNA interactions. Nat Methods *4*, 895-901.

Sims, R.J., 3rd, Belotserkovskaya, R., and Reinberg, D. (2004). Elongation by RNA polymerase II: the short and long of it. Genes Dev *18*, 2437-2468.

Sinclair, A.M., Gottgens, B., Barton, L.M., Stanley, M.L., Pardanaud, L., Klaine, M., Gering, M., Bahn, S., Sanchez, M., Bench, A.J., *et al.* (1999). Distinct 5' SCL enhancers direct transcription to developing brain, spinal cord, and endothelium: neural expression is mediated by GATA factor binding sites. Developmental biology *209*, 128-142.

Singh-Gasson, S., Green, R.D., Yue, Y., Nelson, C., Blattner, F., Sussman, M.R., and Cerrina, F. (1999). Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. Nature biotechnology *17*, 974-978.

Smith, A.M., Sanchez, M.J., Follows, G.A., Kinston, S., Donaldson, I.J., Green, A.R., and Gottgens, B. (2008). A novel mode of enhancer evolution: the Tal1 stem cell enhancer recruited a MIR element to specifically boost its activity. Genome Res *18*, 1422-1432.

Sofueva, S., and Hadjur, S. (2012). Cohesin-mediated chromatin interactions--into the third dimension of gene regulation. Briefings in functional genomics *11*, 205-216.

Solomon, M.J., and Varshavsky, A. (1985). Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. Proc Natl Acad Sci U S A *82*, 6470-6474.

Solovei, I., Cavallo, A., Schermelleh, L., Jaunin, F., Scasselati, C., Cmarko, D., Cremer, C., Fakan, S., and Cremer, T. (2002). Spatial preservation of nuclear chromatin architecture during three-dimensional fluorescence in situ hybridization (3D-FISH). Experimental cell research *276*, 10-23.

Song, S.H., Hou, C., and Dean, A. (2007). A positive role for NLI/Ldb1 in long-range beta-globin locus control region function. Mol Cell *28*, 810-822.

Song, S.H., Kim, A., Ragoczy, T., Bender, M.A., Groudine, M., and Dean, A. (2010). Multiple functions of Ldb1 required for beta-globin activation during erythroid differentiation. Blood *116*, 2356-2364.

Souroullas, G.P., and Goodell, M.A. (2011). A new allele of Lyl1 confirms its important role in hematopoietic stem cell function. Genesis *49*, 441-448.

Souroullas, G.P., Salmon, J.M., Sablitzky, F., Curtis, D.J., and Goodell, M.A. (2009). Adult hematopoietic stem and progenitor cells require either Lyl1 or Scl for survival. Cell stem cell *4*, 180-186.

Spensberger, D., Kotsopoulou, E., Ferreira, R., Broccardo, C., Scott, L.M., Fourouclas, N., Ottersbach, K., Green, A.R., and Gottgens, B. (2012). Deletion of the Scl +19 enhancer increases the blood stem cell compartment without affecting the formation of mature blood lineages. Exp Hematol *40*, 588-598 e581.

Spilianakis, C.G., and Flavell, R.A. (2004). Long-range intrachromosomal interactions in the T helper type 2 cytokine locus. Nat Immunol *5*, 1017-1027.

Spilianakis, C.G., Lalioti, M.D., Town, T., Lee, G.R., and Flavell, R.A. (2005). Interchromosomal associations between alternatively expressed loci. Nature *435*, 637-645.

Splinter, E., and de Laat, W. (2011). The complex transcription regulatory landscape of our genome: control in three dimensions. EMBO J *30*, 4345-4355.

Splinter, E., Grosveld, F., and de Laat, W. (2004). 3C technology: analyzing the spatial organization of genomic loci in vivo. Methods in enzymology *375*, 493-507.

Splinter, E., Heath, H., Kooren, J., Palstra, R.J., Klous, P., Grosveld, F., Galjart, N., and de Laat, W. (2006). CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes Dev *20*, 2349-2354.

Stedman, W., Kang, H., Lin, S., Kissil, J.L., Bartolomei, M.S., and Lieberman, P.M. (2008). Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators. EMBO J *27*, 654-666.

Straussman, R., Nejman, D., Roberts, D., Steinfeld, I., Blum, B., Benvenisty, N., Simon, I., Yakhini, Z., and Cedar, H. (2009). Developmental programming of CpG island methylation profiles in the human genome. Nature structural & molecular biology *16*, 564-571.

Sugiyama, D., Tanaka, M., Kitajima, K., Zheng, J., Yen, H., Murotani, T., Yamatodani, A., and Nakano, T. (2008). Differential context-dependent effects of friend of GATA-1 (FOG-1) on mast-cell development and differentiation. Blood *111*, 1924-1932.

Tachibana, M., Matsumura, Y., Fukuda, M., Kimura, H., and Shinkai, Y. (2008). G9a/GLP complexes independently mediate H3K9 and DNA methylation to silence transcription. EMBO J *27*, 2681-2690.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell *131*, 861-872.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663-676.

Tan-Wong, S.M., French, J.D., Proudfoot, N.J., and Brown, M.A. (2008). Dynamic interactions between the promoter and terminator regions of the mammalian BRCA1 gene. Proc Natl Acad Sci U S A *105*, 5160-5165.

Tanigawa, T., Elwood, N., Metcalf, D., Cary, D., DeLuca, E., Nicola, N.A., and Begley, C.G. (1993). The SCL gene product is regulated by and differentially regulates cytokine responses during myeloid leukemic cell differentiation. Proc Natl Acad Sci U S A *90*, 7864-7868.

Tanigawa, T., Nicola, N., McArthur, G.A., Strasser, A., and Begley, C.G. (1995). Differential regulation of macrophage differentiation in response to leukemia inhibitory factor/oncostatin-M/interleukin-6: the effect of enforced expression of the SCL transcription factor. Blood *85*, 379-390.

Taylor, K.H., Kramer, R.S., Davis, J.W., Guo, J., Duff, D.J., Xu, D., Caldwell, C.W., and Shi, H. (2007). Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. Cancer Res *67*, 8511-8518.

Tenen, D.G., Hromas, R., Licht, J.D., and Zhang, D.E. (1997). Transcription factors, normal myeloid development, and leukemia. Blood *90*, 489-519.

Thomas, M.C., and Chiang, C.M. (2006). The general transcription machinery and general cofactors. Critical reviews in biochemistry and molecular biology *41*, 105-178.

Thomson, J.P., Skene, P.J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A.R., Deaton, A., Andrews, R., James, K.D.*, et al.* (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature *464*, 1082-1086.

Thut, C.J., Chen, J.L., Klemm, R., and Tjian, R. (1995). p53 transcriptional activation mediated by coactivators TAFII40 and TAFII60. Science *267*, 100-104.

Tijssen, M.R., Cvejic, A., Joshi, A., Hannah, R.L., Ferreira, R., Forrai, A., Bellissimo, D.C., Oram, S.H., Smethurst, P.A., Wilson, N.K.*, et al.* (2011). Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. Dev Cell *20*, 597-609.

Tiwari, V.K., Cope, L., McGarvey, K.M., Ohm, J.E., and Baylin, S.B. (2008). A novel 6C assay uncovers Polycomb-mediated higher order chromatin conformations. Genome Res *18*, 1171-1179.

Tolhuis, B., Blom, M., Kerkhoven, R.M., Pagie, L., Teunissen, H., Nieuwland, M., Simonis, M., de Laat, W., van Lohuizen, M., and van Steensel, B. (2011). Interactions among Polycomb domains are guided by chromosome architecture. PLoS Genet *7*, e1001343.

Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol Cell *10*, 1453-1465.

Tompa, R., McCallum, C.M., Delrow, J., Henikoff, J.G., van Steensel, B., and Henikoff, S. (2002). Genome-wide profiling of DNA methylation reveals transposon targets of CHROMOMETHYLASE3. Curr Biol *12*, 65-68.

Tremblay, M., Herblot, S., Lecuyer, E., and Hoang, T. (2003). Regulation of pT alpha gene expression by a dosage of E2A, HEB, and SCL. J Biol Chem *278*, 12680-12687.

Tripic, T., Deng, W., Cheng, Y., Zhang, Y., Vakoc, C.R., Gregory, G.D., Hardison, R.C., and Blobel, G.A. (2009). SCL and associated proteins distinguish active from repressive GATA transcription factor complexes. Blood *113*, 2191-2201.

Tuan, D., Kong, S., and Hu, K. (1992). Transcription of the hypersensitive site HS2 enhancer in erythroid cells. Proc Natl Acad Sci U S A *89*, 11219-11223.

Turpen, J.B., Kelley, C.M., Mead, P.E., and Zon, L.I. (1997). Bipotential primitive-definitive hematopoietic progenitors in the vertebrate embryo. Immunity *7*, 325-334.

Vakoc, C.R., Letting, D.L., Gheldof, N., Sawado, T., Bender, M.A., Groudine, M., Weiss, M.J., Dekker, J., and Blobel, G.A. (2005). Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. Mol Cell *17*, 453-462.

Valenzuela, L., and Kamakaka, R.T. (2006). Chromatin insulators. Annual review of genetics *40*, 107-138.

Valtieri, M., Tocci, A., Gabbianelli, M., Luchetti, L., Masella, B., Vitelli, L., Botta, R., Testa, U., Condorelli, G.L., and Peschle, C. (1998). Enforced TAL-1 expression stimulates primitive, erythroid and megakaryocytic progenitors but blocks the granulopoietic differentiation program. Cancer Res *58*, 562-569.

van de Corput, M.P., de Boer, E., Knoch, T.A., van Cappellen, W.A., Quintanilla, A., Ferrand, L., and Grosveld, F.G. (2012). Super-resolution imaging reveals three-dimensional folding dynamics of the beta-globin locus upon gene activation. J Cell Sci *125*, 4630-4639.

van Eekelen, J.A., Bradley, C.K., Gothert, J.R., Robb, L., Elefanty, A.G., Begley, C.G., and Harvey, A.R. (2003). Expression pattern of the stem cell leukaemia gene in the CNS of the embryonic and adult mouse. Neuroscience *122*, 421-436.

van Steensel, B. (2005). Mapping of genetic and epigenetic regulatory networks using microarrays. Nat Genet *37 Suppl*, S18-24.

Varterasian, M., Lipkowitz, S., Karsch-Mizrachi, I., Paterson, B., and Kirsch, I. (1993). Two new Drosophila genes related to human hematopoietic and neurogenic transcription factors. Cell growth & differentiation : the molecular biology journal of the American Association for Cancer Research *4*, 885-889.

Vilaboa, N., Bermejo, R., Martinez, P., Bornstein, R., and Cales, C. (2004). A novel E2 box-GATA element modulates Cdc6 transcription during human cells polyploidization. Nucleic Acids Res *32*, 6454-6467.

Visvader, J., and Begley, C.G. (1991). Helix-loop-helix genes translocated in lymphoid leukemia. Trends Biochem Sci *16*, 330-333.

Visvader, J., Begley, C.G., and Adams, J.M. (1991). Differential expression of the LYL, SCL and E2A helix-loop-helix genes within the hemopoietic system. Oncogene *6*, 187-194.

Visvader, J.E., Fujiwara, Y., and Orkin, S.H. (1998). Unsuspected role for the T-cell leukemia protein SCL/tal-1 in vascular development. Genes Dev *12*, 473-479.

Vitelli, L., Condorelli, G., Lulli, V., Hoang, T., Luchetti, L., Croce, C.M., and Peschle, C. (2000). A pentamer transcriptional complex including tal-1 and retinoblastoma protein downmodulates c-kit expression in normal erythroblasts. Mol Cell Biol *20*, 5330-5342.

Vrbsky, J., Akimcheva, S., Watson, J.M., Turner, T.L., Daxinger, L., Vyskot, B., Aufsatz, W., and Riha, K. (2010). siRNA-mediated methylation of Arabidopsis telomeres. PLoS Genet *6*, e1000986.

Vyas, P., McDevitt, M.A., Cantor, A.B., Katz, S.G., Fujiwara, Y., and Orkin, S.H. (1999). Different sequence requirements for expression in erythroid and megakaryocytic cells within a regulatory element upstream of the GATA-1 gene. Development *126*, 2799-2811.

Wadman, I., Li, J., Bash, R.O., Forster, A., Osada, H., Rabbitts, T.H., and Baer, R. (1994). Specific in vivo association between the bHLH and LIM proteins implicated in human T cell leukemia. EMBO J *13*, 4831-4839.

Wadman, I.A., Osada, H., Grutz, G.G., Agulnick, A.D., Westphal, H., Forster, A., and Rabbitts, T.H. (1997). The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. EMBO J *16*, 3145-3157.

Wallace, J.A., and Felsenfeld, G. (2007). We gather together: insulators and genome organization. Current opinion in genetics & development *17*, 400-407.

Wang, D., D'Costa, J., Civin, C.I., and Friedman, A.D. (2006). C/EBPalpha directs monocytic commitment of primary myeloid progenitors. Blood *108*, 1223-1229.

Wang, H., Huang, Z.Q., Xia, L., Feng, Q., Erdjument-Bromage, H., Strahl, B.D., Briggs, S.D., Allis, C.D., Wong, J., Tempst, P.*, et al.* (2001). Methylation of histone H4 at arginine 3 facilitating transcriptional activation by nuclear hormone receptor. Science *293*, 853-857.

Wang, H., Wang, L., Erdjument-Bromage, H., Vidal, M., Tempst, P., Jones, R.S., and Zhang, Y. (2004). Role of histone H2A ubiquitination in Polycomb silencing. Nature *431*, 873-878.

Wang, J., Hevi, S., Kurash, J.K., Lei, H., Gay, F., Bajko, J., Su, H., Sun, W., Chang, H., Xu, G.*, et al.* (2009). The lysine demethylase LSD1 (KDM1) is required for maintenance of global DNA methylation. Nat Genet *41*, 125-129.

Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Peng, W., Zhang, M.Q.*, et al.* (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet *40*, 897-903.

Warren, A.J., Colledge, W.H., Carlton, M.B., Evans, M.J., Smith, A.J., and Rabbitts, T.H. (1994). The oncogenic cysteine-rich LIM domain protein rbtn2 is essential for erythroid development. Cell *78*, 45-57.

Weber, F., de Villiers, J., and Schaffner, W. (1984). An SV40 "enhancer trap" incorporates exogenous enhancers or generates enhancers from its own sequences. Cell *36*, 983-992.

Weber, M., Davies, J.J., Wittig, D., Oakeley, E.J., Haase, M., Lam, W.L., and Schubeler, D. (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat Genet *37*, 853-862.

Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Paabo, S., Rebhan, M., and Schubeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet *39*, 457-466.

Wendt, K.S., and Peters, J.M. (2009). How cohesin and CTCF cooperate in regulating gene expression. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology *17*, 201-214.

Wendt, K.S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T.*, et al.* (2008). Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature *451*, 796-801.

White, K.P., Rifkin, S.A., Hurban, P., and Hogness, D.S. (1999). Microarray analysis of Drosophila development during metamorphosis. Science *286*, 2179-2184.

Wilson, I.M., Davies, J.J., Weber, M., Brown, C.J., Alvarez, C.E., MacAulay, C., Schubeler, D., and Lam, W.L. (2006). Epigenomics: mapping the methylome. Cell Cycle *5*, 155-158.

Wilson, N.K., Foster, S.D., Wang, X., Knezevic, K., Schutte, J., Kaimakis, P., Chilarska, P.M., Kinston, S., Ouwehand, W.H., Dzierzak, E.*, et al.* (2010). Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. Cell stem cell *7*, 532-544.

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Pruss, M., Reuter, I., and Schacherer, F. (2000). TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res *28*, 316-319.

Wolstein, O., Silkov, A., Revach, M., and Dikstein, R. (2000). Specific interaction of TAFII105 with OCA-B is involved in activation of octamer-dependent transcription. J Biol Chem *275*, 16459-16465.

Woodfine, K., Fiegler, H., Beare, D.M., Collins, J.E., McCann, O.T., Young, B.D., Debernardi, S., Mott, R., Dunham, I., and Carter, N.P. (2004). Replication timing of the human genome. Human molecular genetics *13*, 191-202.

Woon Kim, Y., Kim, S., Geun Kim, C., and Kim, A. (2011). The distinctive roles of erythroid specific activator GATA-1 and NF-E2 in transcription of the human fetal gamma-globin genes. Nucleic Acids Res *39*, 6944-6955.

Wu, J., Smith, L.T., Plass, C., and Huang, T.H. (2006). ChIP-chip comes of age for genome-wide functional analysis. Cancer Res *66*, 6899-6902.

Wysocka, J., Swigut, T., Xiao, H., Milne, T.A., Kwon, S.Y., Landry, J., Kauer, M., Tackett, A.J., Chait, B.T., Badenhorst, P.*, et al.* (2006). A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. Nature *442*, 86-90.

Xiao, T., Wallace, J., and Felsenfeld, G. (2011). Specific sites in the C terminus of CTCF interact with the SA2 subunit of the cohesin complex and are required for cohesin-dependent insulation activity. Mol Cell Biol *31*, 2174-2183.

Xu, Z., Huang, S., Chang, L.S., Agulnick, A.D., and Brandt, S.J. (2003). Identification of a TAL1 target gene reveals a positive role for the LIM domain-binding protein Ldb1 in erythroid gene expression and differentiation. Mol Cell Biol *23*, 7585-7599.

Xu, Z., Meng, X., Cai, Y., Koury, M.J., and Brandt, S.J. (2006). Recruitment of the SWI/SNF protein Brg1 by a multiprotein complex effects transcriptional repression in murine erythroid progenitors. The Biochemical journal *399*, 297-304.

Xu, Z., Wei, G., Chepelev, I., Zhao, K., and Felsenfeld, G. (2011). Mapping of INS promoter interactions reveals its role in long-range regulation of SYT8 transcription. Nature structural & molecular biology *18*, 372-378.

Yaffe, E., and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet *43*, 1059-1065.

Yamada, Y., Warren, A.J., Dobson, C., Forster, A., Pannell, R., and Rabbitts, T.H. (1998). The T cell leukemia LIM protein Lmo2 is necessary for adult mouse hematopoiesis. Proc Natl Acad Sci U S A *95*, 3890-3895.

Yang, J., and Corces, V.G. (2011). Chromatin insulators: a role in nuclear organization and gene expression. Advances in cancer research *110*, 43-76.

Yang, L., Wang, L., Kalfa, T.A., Cancelas, J.A., Shang, X., Pushkaran, S., Mo, J., Williams, D.A., and Zheng, Y. (2007). Cdc42 critically regulates the balance between myelopoiesis and erythropoiesis. Blood *110*, 3853-3861.

Yang, L.V., Wan, J., Ge, Y., Fu, Z., Kim, S.Y., Fujiwara, Y., Taub, J.W., Matherly, L.H., Eliason, J., and Li, L. (2006). The GATA site-dependent hemogen promoter is transcriptionally regulated by GATA1 in hematopoietic and leukemia cells. Leukemia *20*, 417-425.

Yao, H., Brick, K., Evrard, Y., Xiao, T., Camerini-Otero, R.D., and Felsenfeld, G. (2010). Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA. Genes Dev *24*, 2543-2555.

Yoshida, T., Ng, S.Y., and Georgopoulos, K. (2010). Awakening lineage potential by Ikaros-mediated transcriptional priming. Current opinion in immunology *22*, 154-160.

Young, R.A. (1991). RNA polymerase II. Annual review of biochemistry *60*, 689-715.

Yu, M.C., Lamming, D.W., Eskin, J.A., Sinclair, D.A., and Silver, P.A. (2006). The role of protein arginine methylation in the formation of silent chromatin. Genes Dev *20*, 3249-3254.

Zeng, P.Y., Vakoc, C.R., Chen, Z.C., Blobel, G.A., and Berger, S.L. (2006). In vivo dual cross-linking for identification of indirect DNA-associated proteins by chromatin immunoprecipitation. Biotechniques *41*, 694, 696, 698.

Zhang, N., Kuznetsov, S.G., Sharan, S.K., Li, K., Rao, P.H., and Pati, D. (2008). A handcuff model for the cohesin complex. J Cell Biol *183*, 1019-1031.

Zhang, Y., McCord, R.P., Ho, Y.J., Lajoie, B.R., Hildebrand, D.G., Simon, A.C., Becker, M.S., Alt, F.W., and Dekker, J. (2012). Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Cell *148*, 908-921.

Zhang, Y., Payne, K.J., Zhu, Y., Price, M.A., Parrish, Y.K., Zielinska, E., Barsky, L.W., and Crooks, G.M. (2005). SCL expression at critical points in human hematopoietic lineage commitment. Stem Cells *23*, 852-860.

Zhao, Q., Rank, G., Tan, Y.T., Li, H., Moritz, R.L., Simpson, R.J., Cerruti, L., Curtis, D.J., Patel, D.J., Allis, C.D*., et al.* (2009). PRMT5-mediated methylation of histone H4R3 recruits DNMT3A, coupling histone and DNA methylation in gene silencing. Nature structural & molecular biology *16*, 304-311.

Zhao, Z., Tavoosidana, G., Sjolinder, M., Gondor, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K.S., Singh, U*., et al.* (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet *38*, 1341-1347.

Zhong, Y., Jiang, L., Hiai, H., Toyokuni, S., and Yamada, Y. (2007). Overexpression of a transcription factor LYL1 induces T- and B-cell lymphoma in mice. Oncogene *26*, 6937-6947.

Zhou, X., Weatherford, E.T., Liu, X., Born, E., Keen, H.L., and Sigmund, C.D. (2008). Dysregulated human renin expression in transgenic mice carrying truncated genomic constructs: evidence supporting the presence of insulators at the renin locus. American journal of physiology Renal physiology *295*, F642-653.

Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T., and Henikoff, S. (2007). Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nat Genet *39*, 61-69.

Appendix 1 Oligonucleotide primer pairs used to determine the expression levels of gene transcripts using SyBr green-based quantitative PCR.

GENE	FORWARD 5' → 3'	REVERSE 5' → 3'
ACTB	AGAAGGAGATCACTGCCCTGG	CACATCTGCTGGAAGGTGGAC
TUBB	GCAGATGCTTAACGTGCAGA	CAATGAAGGTGACTGCCATC
GAPDH	AGGTCCACCACTGACACGTTG	AGCTGAACGGGAAGCTCACT
TAL1	TTTTGTGAAGACGGCACGG	TGAGAGCTGACAACCCCAGG
PDZK1IP1	TTGCAATCGCCTTTGCAGTC	TCCATCTGCCTTGTTTCCGA
STIL	ATGCACATAACGTGGATCACG	TCCATGCTCAAATCCACACC
CMPK1	TCTCATGAAGCCGCTGGT	TCCTGCAGAAAGGTGTGTGT
LYL1	TTGAAGCGGAGACCAAGCCA	CGAAGGCGCCGTTAACGTTCT
NFIX	ACAAGGTGTGGCGGCTGAAT	CGTTGGGCAGTGGTTTGATGTC
TRMT1	TACAGCCGAGCCAAGCAGAT	GTCCAGATCGATGACGTCAAACCT
GATA1	CAAGCTACACCAGGTGAACCG	AGCTGGTCCTTCGGCTGC
LDB1	CCAGCTAGCACCTTCGCC	GTCGTCAATGCCGTTGGC
TCF3	AGGTGCTGTCCCTGGAGGAG	CCGACTTGAGGTGCATCTGG

First column shows the gene name. Second and third columns show the DNA sequences for the forward and reverse primers respectively.

Appendix 2 Oligonucleotide primer pairs used to determine chromatin immunoprecipitation (ChIP) enrichment levels across the TAL1 and LYL1 loci.

GENE	ASSAY	FEATURE	FORWARD 5' → 3'	REVERSE 5' → 3'
LYL1	P^{LYL1}	promoter	CCCGGTTTCCTCCCTCTCAC	TGGTTTCCTCCGGGGTCAG
	LYL1+24	putative enhancer (GATA1)	TCAGTGTCCCAGTTGAGAAGGT	AAAGTGAAGGAGTGACAGGGCT
	LYL1+24	putative enhancer (Pol II)	CCAGAGAGAAGCCAAGTTTCCTCA	CCTCCTCCCGTCTGTTTCCAAT
	LYL1 + 33	putative enhancer	GGGCCTGCGAACAGGAGATA	CCTCGTGGCTGCTCTGCTTT
TAL1	TAL1 +137	neg. control	TTTGCAGTGCCCTGTTCTTAG	TGTTGGCTACCTTGATCATGTG
	TAL1 +57	insulator	CTGCAATATCTCGAGCAGCCAC	GAACAACACGGGCATGGAGATG
	TAL1 +51	erythroid enhancer	TGACCTTACAGCCCTTCACCC	AGCTCCCTGCTCCCAGCAC
	TAL1 +40	insulator	GTCAATGTCCACCGTCCCTTTC	GGAGCCAGTTTGCTGCTGAAG
	TAL1 +32	neg. control	GGATTGAGGAGAGGGCATGTG	GCACGGCTGTGGAGCTATG
	TAL1 +20	stem cell enhancer	TTCGAACGGATCACATCCTG	TTGGTCCGAGCTCTGCCTC
	P^{TAL1}	promoter	CGCCGCAGAGATAAGGCACT	CCCACTCCCTCCGGTGAAAT
	TAL1 -28	neg. control	TGTCACGCAGGATATAGTGGCA	TTAGGAGGCTGAAGTAGGAGGAC
	TAL1 -30	neg. control	GTGCCCTTGAGAGCCTAGGG	CCTCAACAGCCTGTCTTATAATTG
	TAL1 -31	insulator	CAACCAGGTGCTGCTTGAGTC	GAGAAGAGCTGCTGGGAAGG
	TAL1 -35	neg. control	TGGTAACCTGGGAACAAGGTGT	ACTGGCTCCTTCTCATCATTCAGG
	TAL1 -37	neg. control	CCACTGTGCCCAGCCTATTT	GTGAGCCAAGACAGTGCCATT
	TAL1 -94	neg. control	CAGGGTATATCTATGTTCCTAGCAC	GATTGATGAATGGTGACAAAGC
CMPK1	P^{CMPK1}	promoter	GCGCAGAGGTTAGCGTGTC	GCCTCTAACCCAAATCCGC
STIL	P^{STIL}	promoter	GCTCCTACCCTGCAAACAGAC	GGAAACCAGGAGCACAAAGC
TBP	P^{TBP}	promoter (pos. control)	GACCTATGCTCACACTTCTCATGG	CGTTGATAATGTCACTTCCGCCAG
HNF4A	HNF4A	CTCF/Rad21 (pos. control)	GATTATCACACCTTGAGGGTAGGG	ACTGTCCTGTACATTGTCCCTG

First column shows the gene name. Second column shows the region assayed relative to the gene locus (numerical designations refer to distance in kb from the relevant gene promoter; - = upstream from promoter, + = downstream from promoter). Third column describes the function of the element assayed. Fourth and fifth columns show the DNA sequences for the forward and reverse primers respectively.

Appendix 3 Oligonucleotide primer pairs used for 3C analysis in human and mouse cells

BAIT	PREY	SPECIES	BAIT PRIMER 5' → 3'	PREY PRIMER 5' → 3'
P^{TAL1}		human	CTCTGTGTCCGAGTGTGGTG	
	TAL1 +64			TCTTCCTAGCCTCGATGGTC
	TAL 1 +51			CGCAGAAAAGCAAGGATAGG
	TAL 1 +46			GTGAGAACCAGGACCCAGAA
	TAL 1 +19			CCCACAATGGAGAGGATGAC
	TAL 1 +15			AGCCTGAGTGCTACAAAGGT
	TAL1 -8			GCGTGAAAGTCAACCATGTG
	TAL1 -10			CCTGAACCAGGAGTTTGTCAC
	TAL1 -25			TGGCAAGTAGGCTGGAACTT
	TAL1 -31			GTTACTGGCACCCCCTGTT
	TAL1 -41			AGTGGAAGAGCCTCCCTTTG
	TAL1 -72			GGTGATCCACCTGCCTCAT
	TAL1 -81			ATGCTCGCTCTTGCATTCCT
	TAL1 -85			TGCAAAGGCCCTGAGTTACA
TAL1 +57		human	GGCAACCATGGGTCTAAAGCAT	
	TAL1 +46			GTGAGAACCAGGACCCAGAA
	TAL1 +40			GAAACCTGGGAGTCACCTGAA
	TAL1 +30			TTACAGACGCATGCCACCTC
	TAL1 -25			TGGCAAGTAGGCTGGAACTT
	TAL1 -31			GTTACTGGCACCCCCTGTT
	TAL1 -41			AGTGGAAGAGCCTCCCTTTG
TAL1 +53		human	TGGGAAGAAATGGCATCTACGC	
	TAL1 +46			GTGAGAACCAGGACCCAGAA
	TAL1 +40			GAAACCTGGGAGTCACCTGAA

	TAL1 +30			TTACAGACGCATGCCACCTC
	TAL1 -25			TGGCAAGTAGGCTGGAACTT
	TAL1 -31			GTTACTGGCACCCCCTGTT
	TAL1 -41			AGTGGAAGAGCCTCCCTTTG
TAL1 +40		human	GAAACCTGGGAGTCACCTGAA	
	TAL1 +64			TCTTCCTAGCCTCGATGGTC
	TAL1 +53			TGGGAAGAAATGGCATCTACGC
	TAL1 +46			GTGAGAACCAGGACCCAGAA
	TAL1 -25			TGGCAAGTAGGCTGGAACTT
	TAL1 -31			GTTACTGGCACCCCCTGTT
	TAL1 -41			AGTGGAAGAGCCTCCCTTTG
TAL1 -31		human	GTTACTGGCACCCCCTGTT	
	TAL1 +64			TCTTCCTAGCCTCGATGGTC
	TAL1 +53			TGGGAAGAAATGGCATCTACGC
	TAL1 +46			GTGAGAACCAGGACCCAGAA
	TAL1 +40			GAAACCTGGGAGTCACCTGAA
	TAL1 +30			TTACAGACGCATGCCACCTC
TAL1 +51		human	CGCAGAAAAGCAAGGATAGG	
	TAL1 +30			TTACAGACGCATGCCACCTC
	TAL1 +19			CCCACAATGGAGAGGATGAC
	TAL1 +15			AGCCTGAGTGCTACAAAGGT
p^{LYL1}		human	ATCGCTACAAGGAGGGTCCTAA	
	LYL1 +45			TGTCTCAGAGTCCTGTGGGT
	LYL1 +33			TTGCTCCCACTTGCTCCTTT
	LYL1+28			TTCCCTAGTGGGATGGAATCCT
	LYL1+24			AGGGCGATCTAGGTGTTCTCAT

	LYL1 +10			TCAATTGACCGGTTGGACTTGG
ERCC3		human	CCCTGGACATGTCGGAAA	AGGGGTTTGCTCTTTGAGGT
P^{TAL1}		mouse	TGCCCCTTAAGCTTGGTTTC	
	TAL1 +55			TGGGAACAGATTGTGGGACT
	TAL1 +40			TGCTGGCTTCCTCTCTTTTC
	TAL1 +30			AAAAGCCTCTCCCTCTCCAG
	TAL1 +18			CCTAGATGAGGGGTGAGAGC
	TAL1 +15			AGCCTTTCCCCTTGATGTTC
	TAL1 -5			CGACCTTCCCTACGTCTTTG
	TAL1 -9			GAGAACAGATGGGCTTGGTC
ERCC3		mouse	AACGGACAGCTTTAGGCAGA	TGGCTGTAGTTGTGCCTTCTC

First column shows the 3C “bait” region and gene locus from which it is derived. Naming system is as per Supplemental Table 3. Second column shows the “prey” region used in 3C primer combinations with the “bait”. Third column is the species in which the assays were performed. The fourth and fifth columns are the “bait” and “prey” primer sequences respectively.

Appendix 4 Oligonucleotide primers used for 4C sample preparation.

Primer Name	Sequence 5' → 3'
$P^{TAL1\text{-}1b}$ (primer extension)	biot-GGCGGCGTTGGCTGCTTCTAAGTG
$P^{TAL1\text{-}1b}$ (nested PCR primer)	GACAGGCTCTGTGTCCGAGT
Blunt-ended adapter (forward)	ACAGGTTCAGAGTTCTACAGTCCGAC
Blunt-ended adapter (reverse)	p-GTCGGACTGTAGAACTCTGAAC
Adapter PCR primer	GGTTCAGAGTTCTACAGTCCGAC

First column shows the primer name and its use in constructing the 4C library. Second column shows the primer sequence. Biotinylation = biot, p = 5' phosphate.