

Chapter 4

Identification and Characterisation of Binding Sites of the Insulator Protein CTCF in the Human Genome

4.1. Introduction

In addition to promoter and enhancer/repressor elements, insulators constitute another major class of regulatory elements found in the human genome. Insulators regulate interactions between promoters and enhancers by preventing inappropriate enhancer/promoter contact as well as acting as boundary elements to prevent the spread of silencing heterochromatin (Burgess-Beusse *et al.*, 2002). They can insulate entire genes or a cluster of genes from the influence of heterochromatin as well as facilitating the establishment of complex cell specific gene expression patterns when individual enhancers are flanked by insulator elements (Brasset and Vaury, 2005). Sequences that prevent inappropriate activation by enhancers have been termed enhancer blocking insulators and those which prevent the spread of heterochromatin have been termed barrier insulators (Sun and Elgin, 1999). Enhancer blocking insulators may function by looping of chromatin into distinct regulatory domains that prevent inappropriate enhancer-promoter communication (Cai and Shen, 2001; Gruzdeva *et al.*, 2005; Kurukuti *et al.*, 2006), while barrier elements may function by recruiting histone modifiers which deposit histone modifications associated with active chromatin, thus preventing the spread of heterochromatin (West *et al.*, 2004).

CCCTC-binding factor (CTCF) is a widely expressed 11 zinc-finger nuclear protein that was first identified by its ability to bind to the promoters of chicken, mouse, and human MYC genes (Filippova *et al.*, 1996; Klenova *et al.*, 1993, Lobanekov *et al.*, 1990). Initial characterisation of CTCF revealed that it could act as both a transcriptional repressor (Burcin *et al.*, 1997) and activator (Vostrov and Quitschke, 1997). Subsequently it was found to bind the HS4 insulator of the chicken β -globin locus (Bell *et al.*, 1999) and since then has been found to bind all known vertebrate insulator elements (Bell *et al.*, 1999; Bell *et al.*, 2001; Mukhopadhyay *et al.*, 2004) as well some insulator elements in *Drosophila* (Moon *et al.*, 2005; Holohan *et al.*, 2007). CTCF was

shown to bind to diverse and long (~50bp) DNA sequences by using different combinations of its individual zinc fingers (Burcin *et al.*, 1997; Filippova *et al.*, 1996). This multiple sequence specificity of CTCF is believed to mediate its classical transcription factor and insulator functions through the formation of distinct CTCF complexes at different CTCF sites (Gaszner and Felsenfeld, 2006; Filippova, 2008).

Binding of CTCF was found to be necessary for enhancer blocking function of the chicken beta-globin locus HS4 insulator (Bell *et al.*, 1999) and was separable from the barrier function which prevented the spread of heterochromatin (Recillas-Targa *et al.*, 2002, West *et al.*, 2002). Two models for insulator activity have been proposed - the chromatin loop domain model and the tracking model. The chromatin loop domain model is based on the ability of CTCF to form chromatin loop domains by interacting with nucleolar structural components (Dunn *et al.*, 2003; Yusufzai *et al.*, 2004) or with other CTCF sites (Kurukuti *et al.*, 2006; Ling *et al.*, 2006). Enhancer-blocking activity is conferred by positioning of promoters and enhancers into separate chromatin loop domains. The tracking model was proposed based on the ability of CTCF to interact with the transcriptional machinery and block the transfer of RNA polymerase II between an enhancer and a promoter (Zhao and Dean, 2004).

The role of CTCF in enhancer blocking is also important for the coordination of gene expression patterns at imprinted gene clusters in mammalian genomes which are regulated by imprinting control regions (ICRs) - also known as differentially methylated domains (DMDs) (Ohlsson *et al.*, 2001; Reik and Walter, 2001). CTCF was found to bind to an ICR upstream of the H19 gene and block access of Igf2 to an enhancer shared with H19 which results in no Igf2 expression from the maternal allele (Figure 4.1) (Hark *et al.*, 2000; Bell *et al.*, 2000; Kanduri *et al.*, 2000). CpG methylation of the ICR on the paternal allele prevents CTCF binding, allowing enhancer-mediated activation of the Igf2 promoter on the paternal allele (Bell *et al.*, 2000). CTCF has also been shown to prevent the spread of DNA methylation thus playing a crucial role in maintaining methylation free regions (Engel *et al.*, 2004; Pant *et al.*, 2004; Filippova *et al.*, 2005). Thus CTCF can also prevent nearby promoters from being epigenetically silenced.

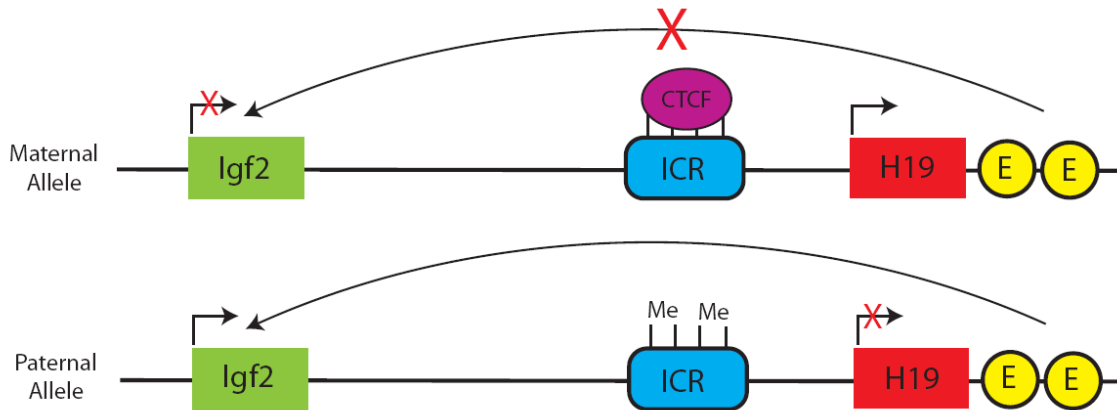


Figure 4.1: CTCF enhancer-blocking insulator at the *Igf2*/*H19* imprinted region. CTCF binds to sites in the ICR in the maternal allele and prevents downstream enhancers (E) from activating *Igf2* expression. The ICR is methylated (Me) in the paternal allele which prevents CTCF binding which means that the downstream enhancers are no longer blocked from interacting with the *Igf2* promoter.

The 5' HS4 chicken beta-globin insulator element displays both enhancer blocking and barrier insulator functions. The enhancer blocking and barrier functions were found to be separable activities and CTCF was necessary for the enhancer blocking activity of this insulator but was not required for barrier activity (Recillas-Targa *et al.*, 2002). The binding of USF1 and USF2 proteins was later found to be required for the barrier activity of the HS4 insulator and they were shown to recruit histone modifying complexes responsible for H3K4 and H4R3 methylation and histone acetylation, which prevented the spread of histone modifications associated with nearby condensed chromatin (West *et al.*, 2004, Huang *et al.*, 2007). However it is not clear if USF1 and USF2 are responsible for the barrier activity of other insulators. CTCF may be required for both enhancer blocking and barrier activity in other vertebrates. It has not been directly shown that CTCF prevents the spread of heterochromatin, but CTCF binding sites have been found close to the transition between active and silent chromatin domains (Filippova *et al.*, 2005; Barksy *et al.*, 2007). CTCF-mediated barrier activity would reconcile with the chromatin loop model of enhancer blocking as flanking a gene with a CTCF binding site could also provide barrier activity by creating an independent expression domain.

CTCF can bind to different DNA sequences using various combinations of the 11 zinc finger domains (Filippova *et al.*, 2008). CTCF complexes formed at these different DNA

sequences may be functionally different depending on CTCF interaction with other proteins. The use of different combinations of zinc fingers may mean that different combinations of zinc fingers are available for interaction. CTCF has been shown to interact with nucleophosmin (Yusufzai *et al.*, 2004), Kaiso (Defossez *et al.*, 2005), and CHD8 (Ishihara *et al.*, 2006) and these interactions have been linked to the regulation of CTCF insulator function. The transcription factor YY1 has been shown to frequently bind in close proximity to CTCF sites and is an important co-factor in CTCF mediated regulation of X inactivation (Donohoe *et al.*, 2007).

CTCF is also known to interact with the transcription factor mSin3a (Lutz *et al.*, 2000). The mammalian Sin3 proteins, mSin3A and mSin3B were discovered as a result of their interaction with the transcriptional repressors Mad1 and Mxi1 (Ayer *et al.*, 1995). They were later shown to associate with HDAC1 and HDAC2 to form a large multiprotein complex, the Sin3/histone deacetylase (HDAC) co-repressor complex (Silverstein and Ekwall, 2005). CTCF interaction with mSin3a may be responsible for CTCF mediated transcriptional repression via recruitment of HDACs.

Identifying the location of putative insulator elements in the human genome would greatly increase our understanding of how this class of *cis*-acting regulatory elements controls genome structure and function. However, computational prediction of insulator location in the human genome is a difficult task as CTCF can use various combinations of zinc fingers to bind different target sequences (Ohlsson *et al.*, 2001). Mukhopadhyay and colleagues (2004) used ChIP cloning and sequencing to identify 200 CTCF binding sequences with enhancer-blocking activity in the mouse genome but no consensus binding motif was identified by this study. Xie and colleagues used a systematic approach to discover and characterise regulatory motifs within mammalian conserved non-coding elements (CNEs) by searching for long motifs (12-22 nt) with significant enrichment in CNEs (Xie *et al.*, 2007). One of these motifs (CCACTAGATGGCA) was found to at 15,000 conserved locations in the human genome and was found to experimentally bind CTCF. Kim and colleagues performed a genome-wide ChIP-chip analysis of CTCF binding sites in the human genome and identified over 13,000 binding sites (Kim *et al.*, 2007). This large data set was used to define a 20-mer CTCF consensus binding motif, which was found at over 75% of their experimentally determined binding sites in the

human genome and was able to bind recombinant CTCF. This consensus 20 bp motif was similar to a 14bp GC rich sequence previously defined based on a limited number of characterised sites (Bell and Felsenfeld, 2000) but was refined at a number of positions (Figure 4.2). The 13 bp motif identified by Xie and colleagues (2007) was also similar to the core of this 20bp consensus motif.

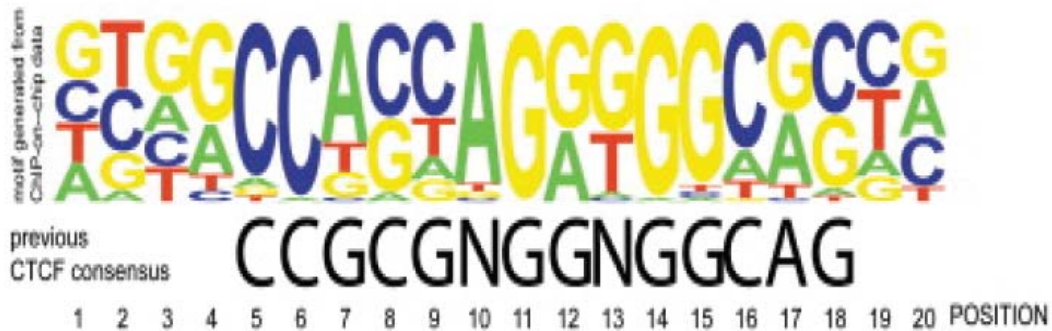


Figure 4.2: A 20-mer motif is recognised by CTCF. A DNA logo representing the CTCF-binding motif derived from a genome-wide ChIP-chip study (Kim *et al.*, 2007) and the previously reported consensus CTCF-binding sites (Bell and Felsenfeld, 2000) are shown. The relative frequency at which each nucleotide occurs in the motif is indicated by height of letter at each position. The ChIP-chip derived motif refines the previous consensus motif at six nucleotide positions (positions 7, 8, 9, 10, 13, and 17). Figure from Kim *et al.*, 2007

Kim and colleagues used their experimentally determined motif to scan the human genome sequence and a total of 31,905 sites were found to contain this motif, 12,799 of which were conserved in at least one other vertebrate genome. This suggests that the location of a large number of insulators may be conserved between vertebrates.

When work towards this PhD thesis was initiated very little was known about insulators and the binding of CTCF in the human genome. Therefore as a logical step towards understanding where insulators are found in the human genome, a ChIP-chip assay would need to be developed for CTCF. Although the work of Kim and colleagues (2007) greatly aided in the discovery of CTCF binding events in the human genome, understanding how CTCF interacts with other proteins at insulators still remains largely unexplored genome-wide. Therefore, in order to further characterise CTCF binding events, ChIP-chip assays to detect binding sites for the CTCF-associated proteins mSin3a, USF1, and USF2 were

investigated. This information in combination with data on histone modifications associated with active and inactive chromatin domains was used to further characterise CTCF binding events and determine whether there were features, apart from CTCF-binding, which distinguished insulators from the other major classes of regulatory sequences.

4.2. Aims of this chapter

In order to identify and characterise CTCF binding sites in 1% of the human genome the aims of the work presented in this chapter were as follows:

1. To develop a ChIP-chip assay for the detection of CTCF binding sites using the SCL tiling path array as a model.
2. To apply this assay for the identification of CTCF binding sites in the 1% of the human genome covered by the ENCODE regions.
3. To further characterise CTCF interactions by developing additional ChIP-chip assays to investigate the binding of the known CTCF interacting partner mSin3a and the barrier insulator proteins USF1 and USF2.
4. To investigate the conservation of CTCF binding sites in human cell lines.
5. To investigate the histone modifications associated with CTCF binding events at insulators.

4.3. Overall strategy

As discussed in Chapter 3, the SCL locus tiling path array was previously used to develop ChIP-chip assays for the detection of histone modification events associated with promoter and enhancer elements (Dhami, submitted). However, because of a lack of information in the literature on histone modifications that help define insulators, it was necessary to develop an assay which is insulator specific. Therefore a ChIP-chip assay was developed using the SCL locus as a model system for the identification of DNA sequences interacting with the insulator binding protein CTCF in K562 cells. The SCL region was used initially to develop the assay and determine whether the CTCF binding data reconciles with what is already known about regulatory features in the region. In order to gain a greater understanding of the location of putative insulator elements

relative to genes and other epigenetic and genomic features, this assay was then applied to detect CTCF interactions in K562 cells in the 1% of the human genome covered by the Sanger Institute ENCODE genomic tiling array. CTCF interactions were further characterised by performing ChIP-chip experiments to identify binding locations of the CTCF interacting partner mSin3a and the barrier proteins USF1 and USF2 (Lutz *et al.*, 2000; West *et al.*, 2004; Huang *et al.*, 2007). CTCF binding sites were also examined by in the human cell line U937 and compared with K562 binding sites to determine if the location of insulators is conserved between different cell types. All of this data was then compared to that generated from other sources, both within our laboratory and in other laboratories, which includes the location of DNase I hypersensitive sites, predicted CTCF sites, histone modifications, formaldehyde assisted isolation of regulatory elements (FAIRE), and nucleosome data.

Results

4.4. Assessing the specificity of transcription factor antibodies in western blotting assays

In order to ensure that ChIP-chip assays could detect *bona fide in vivo* CTCF, mSin3a, USF1 and USF2 binding sites, it was important to verify the specificity of the antibodies to be used in ChIP-chip assays, and to avoid cross-reactivity with proteins which share amino acid sequence similarity. To this end western blotting assays were performed. The mSin3a protein is theoretically 145 kDa in size and the antibody used in this study detected a single band at approximately the theoretical size in K562 cells (Figure 4.3). USF1 encodes a protein of molecular mass 43 kDa and this antibody detected one diffuse band in K562 nuclear extracts which was at the correct molecular weight. USF2 encodes a protein of 44 kDa molecular and a single diffuse band at the predicted molecular weight was also detected with K562 nuclear extracts– this band was of similar size as that detected for USF1. Given that USF1 and USF2 are of a similar molecular weight, it was not possible to determine for certain that the antibodies were not cross-reacting. CTCF encodes a protein of theoretical molecular mass 82 kDa and a single band at approximately 97kDa was detected in K562 cells. CTCF has been found to migrate aberrantly in SDS-PAGE as it has been observed to migrate at 130, 97, 80, 73, 70 and 55

kDa's (Klenova *et al.*, 1997). Furthermore, CTCF is known to be poly-ADP ribosylated (Yu *et al.*, 2004) and this large post-translational modification would affect the observed mass of the protein in western blotting assays. All of this data taken together suggests that the CTCF antibody used in this study was likely to be detecting the *bona fide* CTCF protein.

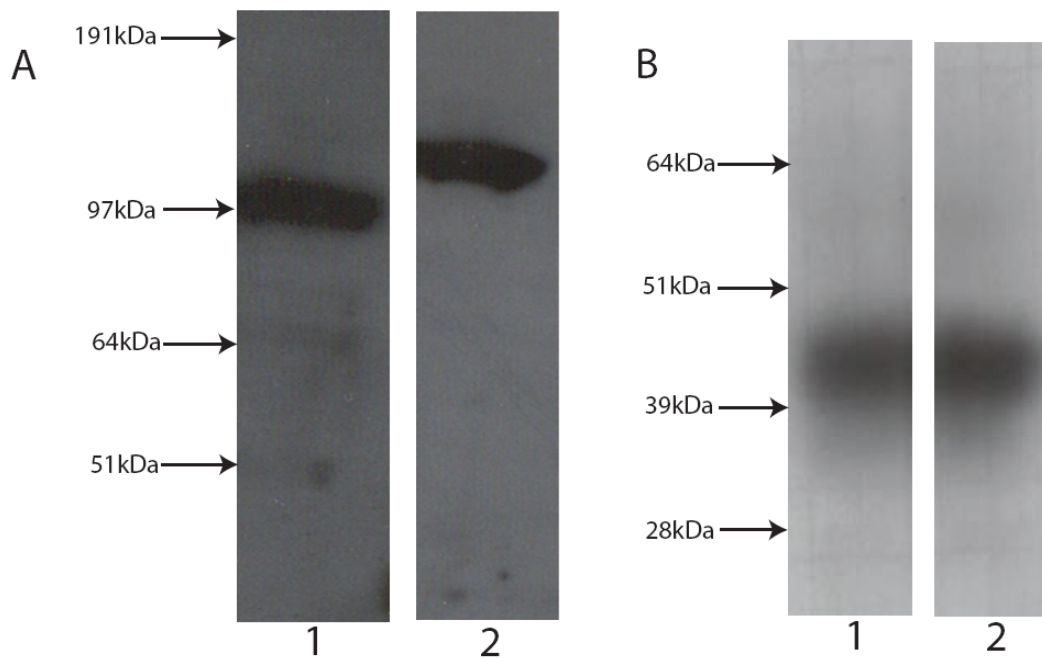


Figure 4.3: Western blot analysis of CTCF, mSin3a, USF1 and USF2 in K562 cells. The CTCF antibody used in this study detected at single band at approximately 97kDa in K562 cells (Panel A, lane 1). Panel A, lane 2 shows that the mSin3a antibody detects a single band at approximately the theoretical molecular mass of 145 kDa. Panel B lanes 1 and 2 show that the USF1 and USF2 antibodies detect single bands at the theoretical molecular masses of 43 and 44kDa's respectively.

4.5. Developing a ChIP-chip assay to detect putative insulators at the SCL locus

The SCL locus is well characterised in terms of promoter and enhancer elements based on ChIP-chip studies using a tiling array of the SCL locus (Dhami, submitted). However, no CTCF binding-elements (i.e. putative insulators) had been characterised in this region at the start of this PhD project. This same SCL tiling array was used to test for the

presence of insulators because a number of genes present on this array were expressed in a tissue specific manner, for example SCL is expressed in blood, in endothelium, and within specific regions of the central nervous system (Begley and Green, 1999) while the gene for the cytochrome P450 family member CYP4A22 (also tiled on the SCL array) is expressed in liver tissue (Savas *et al.*, 2003). Therefore it was hypothesized that the CYP4A22 and SCL genes may require insulators to regulate their expression patterns – thus, a ChIP-chip assay was developed for studying CTCF interactions in the SCL expressing cell line K562.

Eight regions were identified as significantly enriched (significant enrichments in transcription factor ChIP-chip experiments were considered to be those values that were more than three standard deviations away from the mean ratio of background levels) for CTCF binding across the SCL locus (Figure 4.4, panel A). As the CTCF antibody used in this study was raised in goat, a ChIP-chip ‘mock’ antibody control experiment was also performed with a normal goat IgG antibody to identify any non-specific enrichments associated with performing ChIP-chip experiments with a goat IgG (Figure 4.4, panel B). Data from this mock antibody control experiment was used to normalise for non-specific interactions by dividing the CTCF data set values with the corresponding goat IgG values. This approach ensured that non-specific enrichments observed in both data sets were normalised to background values. Fold enrichments for seven of the eight sites were increased following normalisation with the goat IgG data set (Figure 4.4, panel C) while one region at the CYP4A22 gene remained unchanged indicating that all 8 regions represent *bona fide* sites of CTCF interaction and non-specific interactions by a goat IgG. Significant peaks of enrichment were found at a region within the SIL gene, 31 kb upstream of SCL promoter1a (-31), and at +57 between the SCL erythroid enhancer at +51 and the CYP4A22 gene. Peaks were also found within the CYP4A22 and CYP4AZ1 genes, at the SIL and KCY promoters, at +53 region, and the MAP17 enhancer (+40).

The +53 CTCF region is associated with high levels of H3 acetylation, H3K4me2 and H3K4me3 consistent with promoter activity and novel transcripts that have been identified near this region (Dhami, submitted). In contrast the +40 CTCF region is associated with low levels of H3 acetylation and H3K4me2/H3K4me3 but displays high levels of H3K4me1, consistent with enhancer function (Chapter 3) (Follows *et al.*, 2006).

The +57 and -31 CTCF sites do not display either of these histone modification profiles suggesting that these CTCF binding regions may be functionally distinct. The genomic regions contained within +57 and -31 defines a 98kb regulatory domain containing SCL, MAP17 (thought to be co-regulated with SCL) and all known SCL regulatory elements. None of the nearby genes outside of this domain are thought to share regulatory elements, suggesting that the presence of CTCF at these sites marks the location of insulators.

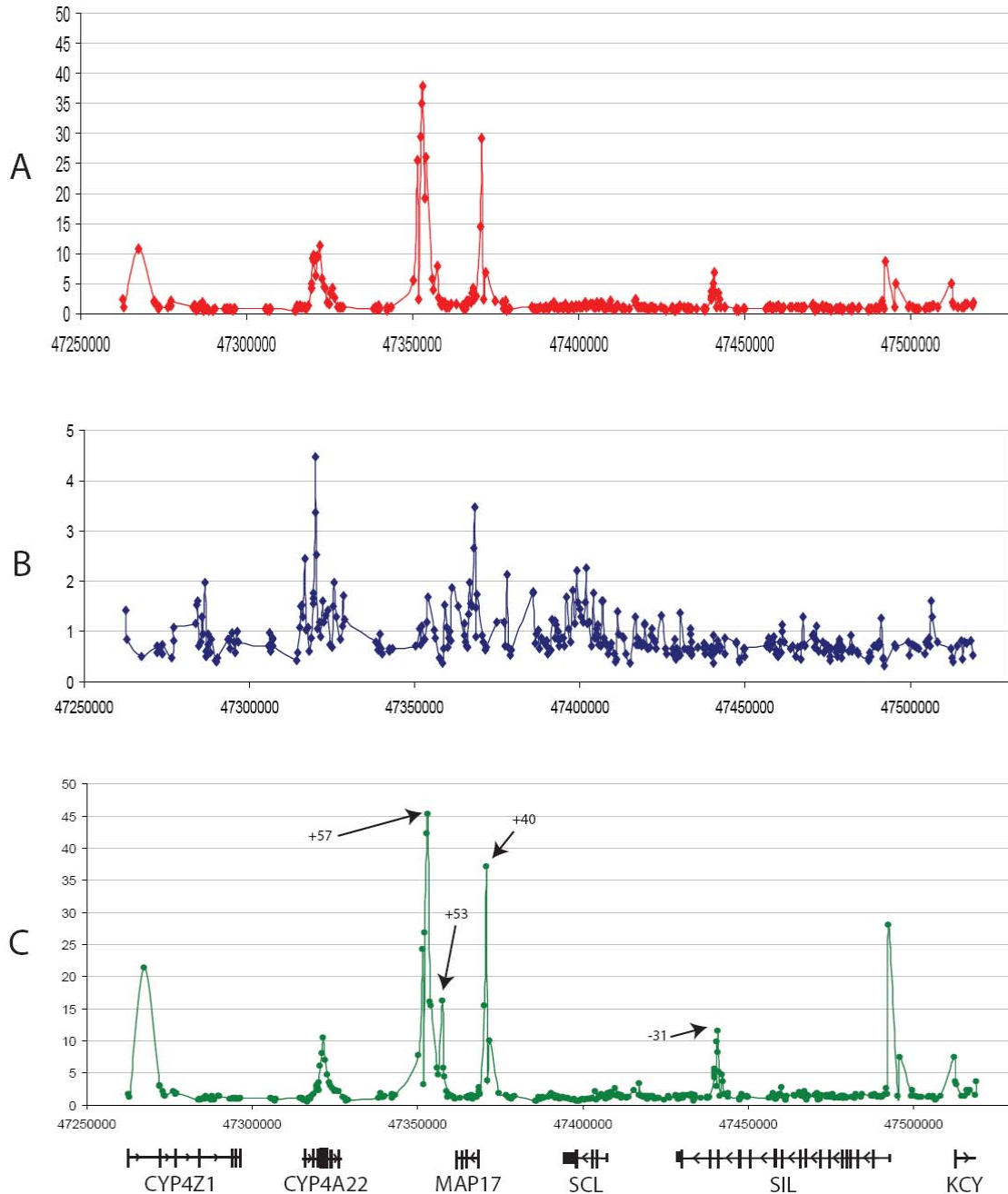


Figure 4.4: ChIP-chip profile of CTCF binding across the SCL locus in K562 before and after Goat IgG normalisation. Panel A: Fold enrichments reported for CTCF interactions before normalisation with normal goat IgG. Panel B: A normal goat IgG ChIP-chip profile showing non-specific enrichment at a number of locations. Panel C: fold enrichments increased at CTCF binding sites after normalisation with goat IgG. The location of -31, +40, +53 and +57 regions are indicated by black arrows. The human chromosome 1 genomic coordinates are indicated along the y-axes, while fold enrichments are indicated on the x-axes. Gene order and direction of transcription is shown below panel C.

4.6. Mapping and characterising CTCF binding sites in the ENCODE regions

4.6.1. Implementation and validation of the CTCF ChIP-chip assay

Following the identification of a number of CTCF sites across the SCL locus, it was important to gain a more complete understanding of the genome-wide binding patterns of CTCF. The Sanger Institute ENCODE array (chapters 1 and 3) was used to analyse the binding patterns of CTCF in 1% of the human genome. ChIP-chip assays were performed with the CTCF antibody across three biological replicates of K562 chromatin as described previously. The ChIP DNA were hybridised to the ENCODE array and the median values of the three hybridisation experiments were normalised with goat IgG values to eliminate non-specific enrichments. To identify CTCF binding sites, across 1% of the human genome, the ChIP-chip normalised data sets were analysed by CHIPOTle (Buck *et al.*, 2005). As discussed in Chapter 3, this program was developed specifically to analyse ChIP-chip data and uses a sliding windows approach to identify peaks of enrichment and then estimates the significance of enrichment for a region using a Gaussian error function. Using this analysis tool 571 CTCF sites were identified as significantly enriched across the ENCODE regions in K562 cells using a stringent p-value cut-off of p0.0001. As the ENCODE regions represent 1% of the human genome, this figure suggests that there may be more than 50,000 binding sites in the entire genome.

In order to verify the specificity of the ChIP-chip method used in this study, known CTCF binding sites in the ENCODE regions were investigated for binding in K562 cells. Eight closely associated CTCF sites have been characterised at the IGF2/H19 imprinting control region by a number of groups (Bell and Felsenfeld, 2000; Hark *et al.*, 2000, Szabo *et al.*, 2000), spanning a 4 kb region upstream of the H19 locus (coordinates

chr11:1976843-1980864). The CTCF binding profile at the IGF2/H19 locus in K562 is shown in Figure 4.5. A region upstream of the H19 gene shows a highly enriched CTCF binding peak identified by ChIPOTle, which correlates with previously characterised CTCF binding sites in the imprinting control region.

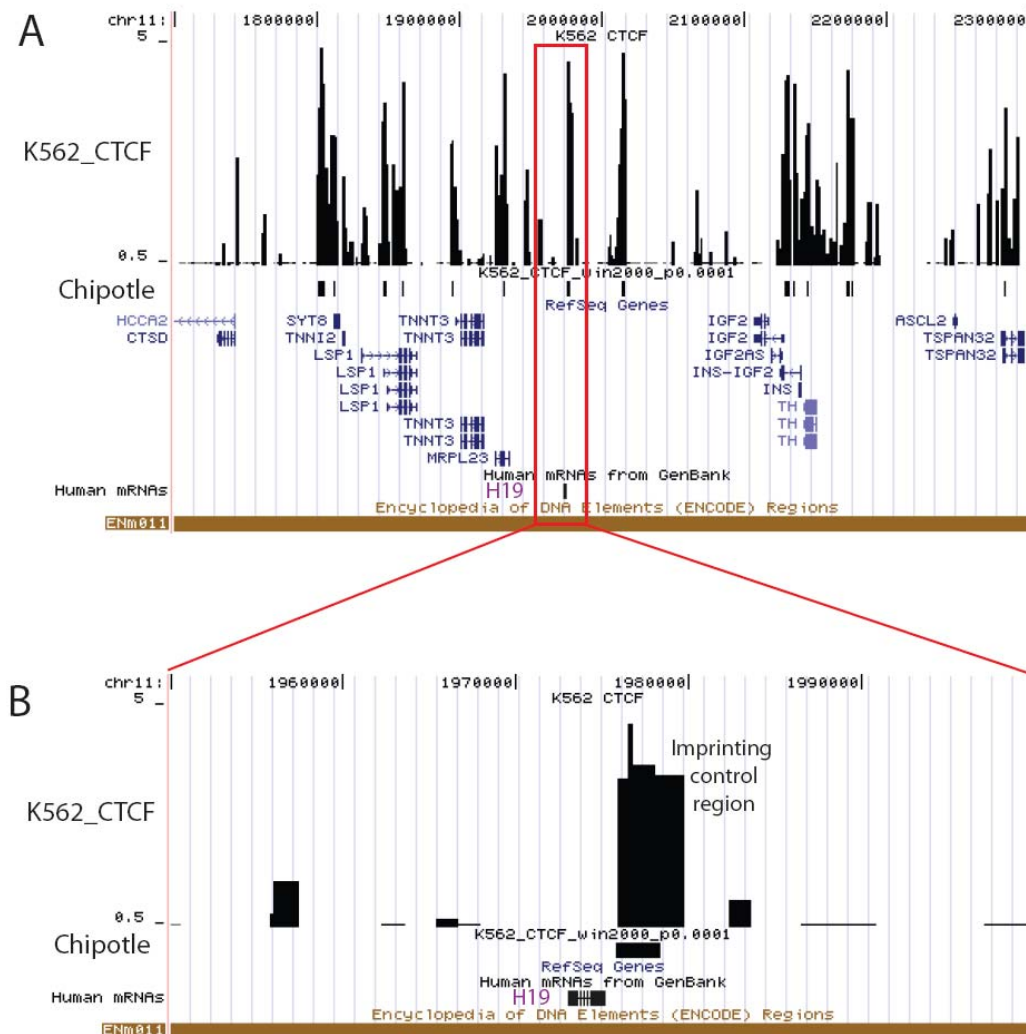


Figure 4.5: CTCF binding sites in the IGF2/H19 locus. Panel A illustrates the CTCF binding profile in ENCODE region Enm011 (IGF2/H19 locus). Log₂ fold enrichments are represented in the top half of the panel (K562_CTCF track) and sites identified by ChIPOTle (p0.0001) are indicated below the x-axis (Chipotle track). Known RefSeq genes and the location of the H19 maternally transcribed mRNA are indicated at the bottom of panel A. A magnified view of CTCF binding at the H19 locus is illustrated in panel B. A highly enriched peak of CTCF binding is observed at the imprinting control region.

In addition to the IGF2/H19 CTCF binding sites, a CTCF binding site has previously been characterised at the β -globin locus (beta-globin HS5) in K562 cells (genome coordinate's chr11:5269210-5269281) (Farrell *et al.*, 2002). Thirteen CTCF binding regions were identified at the β -globin locus in this study (Figure 4.6), one of which is located at the same coordinates as the β -globin HS5.

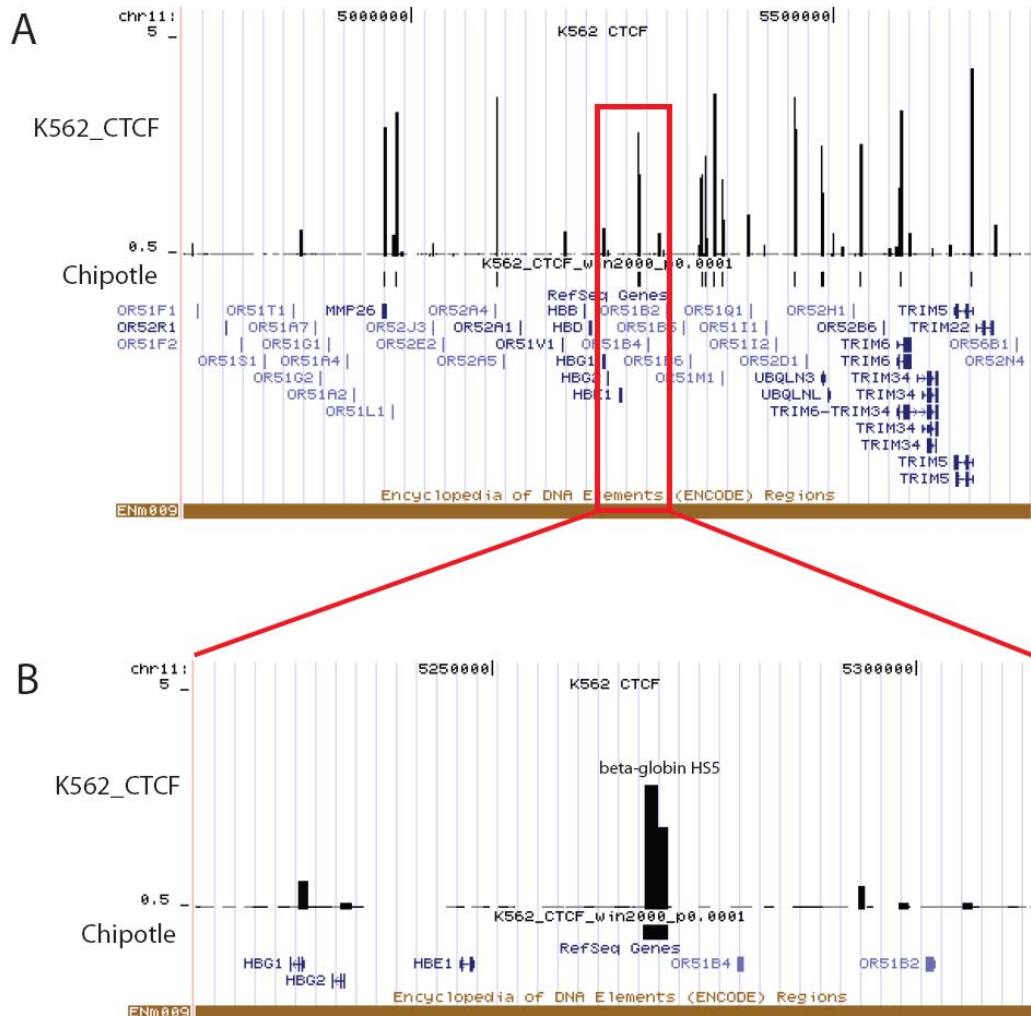


Figure 4.6: CTCF binding sites at the β -globin locus. Panel A illustrates the CTCF binding events at ENCODE region Enm009 (β -globin locus) in K562 cells. Log₂ fold enrichments are represented in the top half of the panel (K562_CTCF track) and ChIPOTle sites are indicated below the x-axis (chipotle track). Known genes are indicated at the bottom of panel A. A magnified view of CTCF binding is presented in panel B. A highly enriched peak of CTCF binding is observed at the previously characterised β -globin HS5.

In addition, the CTCF data was compared with that of Kim and colleagues who identified nearly 14,000 CTCF binding sites in IMR90 cells and also predicted the location of over 30,000 CTCF sites based on an experimentally defined consensus sequence (Kim *et al.*, 2007). The location of predicted CTCF sites at the SCL locus was examined to determine the correlation with CTCF sites identified in K562 cells by ChIP-chip. The study by Kim and colleagues (2007) predicted five CTCF binding sites at the SCL locus (Figure 4.7). Four of these predicted sites were associated with CTCF binding in K562, namely the +57 region (associated with two copies of the consensus motif), +53 region and +40 region. These sites also bound CTCF in primary human fibroblast IMR90 cells (Kim *et al.*, 2007). Four of the five predicted CTCF binding sites were conserved at the sequence level in at least one other vertebrate genome and these four sites were bound in both K562 and IMR90 cells suggesting that they represent genuine functional elements. Sites of CTCF binding within or close to the CYP4Z1, CYP4A22, SIL, and KCY genes were not associated with predicted motifs and displayed K562-specific binding. However, Kim and colleagues also acknowledged that the consensus motif does not match all CTCF sites (Kim *et al.*, 2007). A more detailed comparison of the data generated here with that of other laboratories is presented in section 4.7.

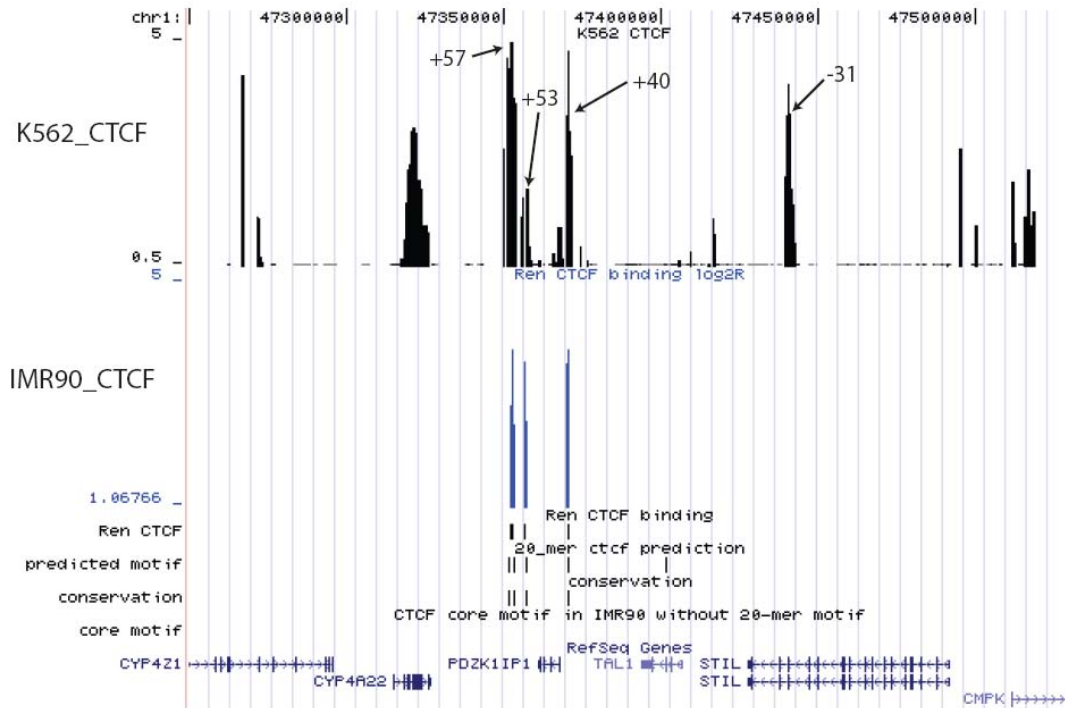


Figure 4.7: Comparison of CTCF binding sites in the K562 cell with binding sites identified in primary human fibroblast, IMR90 cells. K562 binding sites are indicated at the top of the figure by black vertical bars (K562_CTCF track). Binding sites identified in IMR90 cells by Kim and colleagues (2007) are indicated below the K562 data by blue vertical bars (IMR90_CTCF track). Reported fold enrichments are displayed as Log_2 values in both profiles. The +57, +53, and +40 binding sites are bound by CTCF in the two cell types. The +57, +53, and +40 regions are associated with the predicted 20-mer binding sequence (indicated by black vertical lines in the predicted motif track). Four of the five predicted peaks are conserved in at least one other vertebrate genome, excluding the chimpanzee genome (indicated by the conservation track). An additional five binding sites were identified in K562 cells which were not associated with the known consensus motif, which were located within the CYP4Z1, CYP4A22, and SIL (also known as STIL) genes, and at the SIL and KCY (also known as CMPK) promoters. The core motif track represents predicted CTCF binding sites based on the previously reported CTCF consensus sequence (Bell and Felsenfeld, 2000). No core motif sites are predicted in the SCL locus. Known RefSeq genes (Pruitt *et al.*, 2007) are indicated at the bottom of the figure and human chromosome 1 coordinates are displayed at the top of the figure. Note: PDZK1IP1, TAL1, STIL, CMPK are also known as MAP17, SCL, SIL, and KCY respectively.

The accurate identification of previously characterised CTCF binding sites at the IGF2/H19 and β -globin loci and overlap of CTCF sites with the predicted consensus CTCF motif at the SCL locus indicated that this ChIP-chip method in combination with

ChIPOTle analysis could be used to accurately map CTCF sites of interaction in a high-throughput manner across 1% of the human genome sequence.

4.6.2. Distribution of CTCF-binding sites in the ENCODE regions

The distribution of CTCF binding sites was investigated with respect to the location of the nearest TSSs of protein-coding genes (Figure 4.8). More than 50% of CTCF sites of interaction were found to be located 5kb or more from TSSs, consistent with CTCF functioning at distant insulator or enhancer/repressor elements.

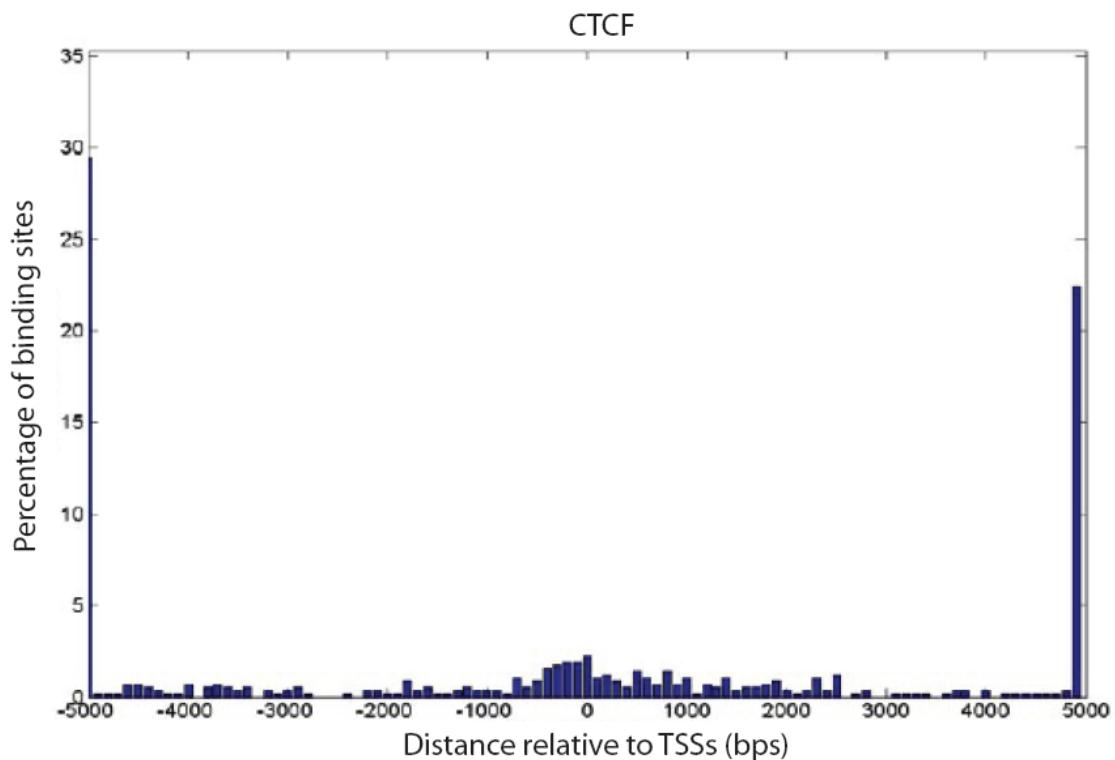


Figure 4.8: The binding intervals and distribution of CTCF sites relative to transcription start sites (TSSs) in K562. Panel B shows the relative distance of CTCF binding sites to the closest TSS of protein-coding genes. Relative distance is indicated in bp's along the x-axis and the percentage of binding sites is shown on the y-axis. Note that the total percentage of CTCF binding sites located 5 kb or more from a TSS is indicated by the blue vertical bars on the extreme left and right of the figure.

Although the CTCF sites tend to be located far from transcription start sites they are not randomly distributed across the ENCODE regions. The distribution of CTCF binding sites was examined by comparing the number of binding events with gene density in the

ENCODE regions. The distribution of CTCF sites closely follows the distribution of gene density, with a correlation coefficient of 0.76 (Figure 4.9 shows R^2 value of 0.5817). This is consistent with CTCF regulating gene expression, rather than simply performing a structural role such as the formation of chromatin loops.

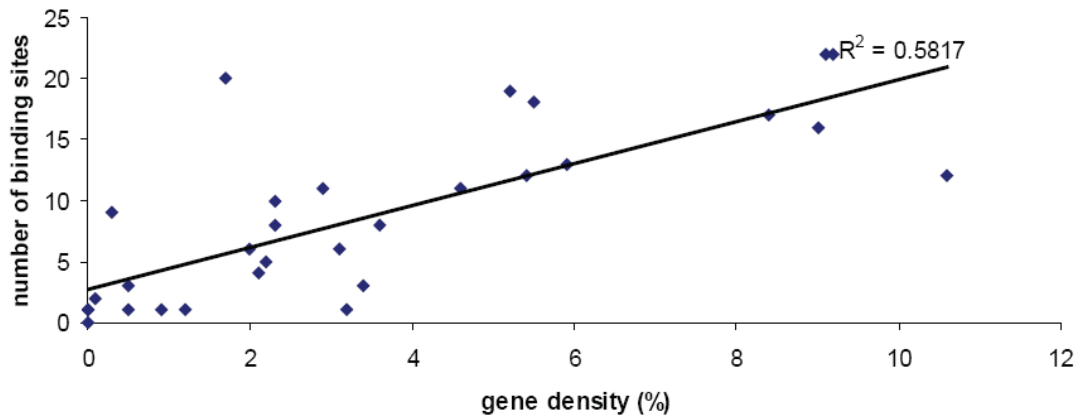


Figure 4.9: Correlation of CTCF binding events with gene density in the ENCODE regions. Gene density was expressed as a percentage of the total region size for computationally defined ENCODE regions (Chapter 3) and was plotted against the number of binding sites in each region. In general those regions associated with a high gene density also contained a high number of CTCF binding sites ($R^2=0.5817$).

The distribution of CTCF binding sites was further examined by comparing their location with respect to transcription start sites (TSSs), distal enhancer/repressor sites, and other sites (Figure 4.10). Distal enhancer/repressor sites were defined as regions associated with a peak of enrichment for H3K4me1, or H3K4me2, or H3K4me3, and not within 2.5 kb of a TSS (Chapter 3) and other sites can be located anywhere except at TSSs or distal enhancer/repressor sites. 31% of binding events were located at TSSs, consistent with a role as a transcription factor involved in gene repression/activation, while 16% of binding events were located at distal enhancer/repressor sites. Like its role at TSSs, CTCF may function as a classical transcription factor by binding to distal enhancer/repressor sites. There is also the possibility that CTCF binding at distal sites may be related to the enhancer blocking activity of this protein. Sites not at TSSs or at distal enhancers/repressors (shown as “other” in Figure 4.10) account for the largest percentage

(53%) of CTCF sites. This class of CTCF sites is not associated with any promoter or enhancer/repressor function, suggesting that they may define the location of insulators.

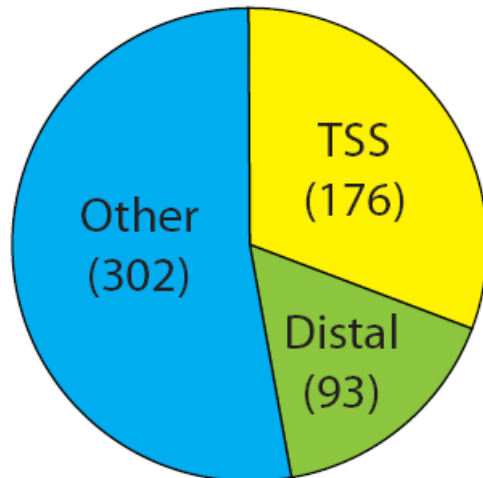


Figure 4.10: The distribution of CTCF binding sites. The pie-chart shows the distribution of CTCF binding sites mapped to transcription start sites (TSSs), distal enhancer/repressor sites, and other locations within the 44 ENCODE regions.

The possibility of multiple CTCF binding sites found in close proximity within ChIPOTle defined regions was investigated by determining the size distribution of CTCF ChIPOTle peaks as a function of 2 kb intervals (Figure 4.11). Its important to note that the size of microarray elements has an important bearing on this determination- as the average resolution of ENCODE array elements was calculated to be 1024 bp, ChIPOTle peaks within the size range of 0-2kb would suggest a single CTCF binding site. The majority (61%) of CTCF binding sites were 2kb or less in size suggesting that most CTCF binding sites were likely to be associated with one CTCF binding event. 39% were found spanning more than 2 kb suggesting that multiple copies of CTCF may be bound in close proximity at these locations, although the biological significance of these multiple close binding events is not known.

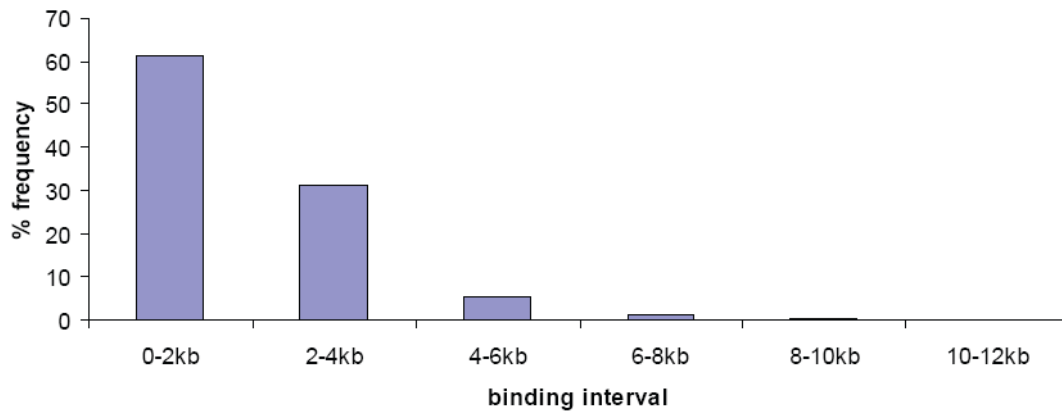


Figure 4.11: CTCF binding intervals indicate multiple CTCF sites can bind in close proximity. CTCF binding intervals defined by ChIPOTle were found to range in size from less than 2 kb up to 12 kb for one particular site. The 2kb binding intervals are indicated on the x-axis, while the percentage frequency of binding sites in each 2 kb interval is indicated on the y-axis.

4.6.3. CTCF sites at individual genes and gene clusters

While examining the distribution of CTCF sites it was noted that several individual genes were flanked in their entirety by CTCF sites and clusters of genes were also found to be flanked by two CTCF sites, as opposed to CTCF sites separating each member of the cluster. 307 genes in the ENCODE regions were flanked in their entirety by CTCF binding sites, representing 52% of the genes, while 10 clusters containing 5 or more genes were flanked by CTCF sites (Table 4.1). One cluster on the X chromosome contained 17 genes flanked by CTCF sites (Figure 4.12) and some of the other clusters were the well-characterised α -globin, β -globin and HOXA loci. The significance of CTCF binding sites at genes clusters is not known but CTCF may form chromatin domains to regulate the expression of co-transcribed genes.

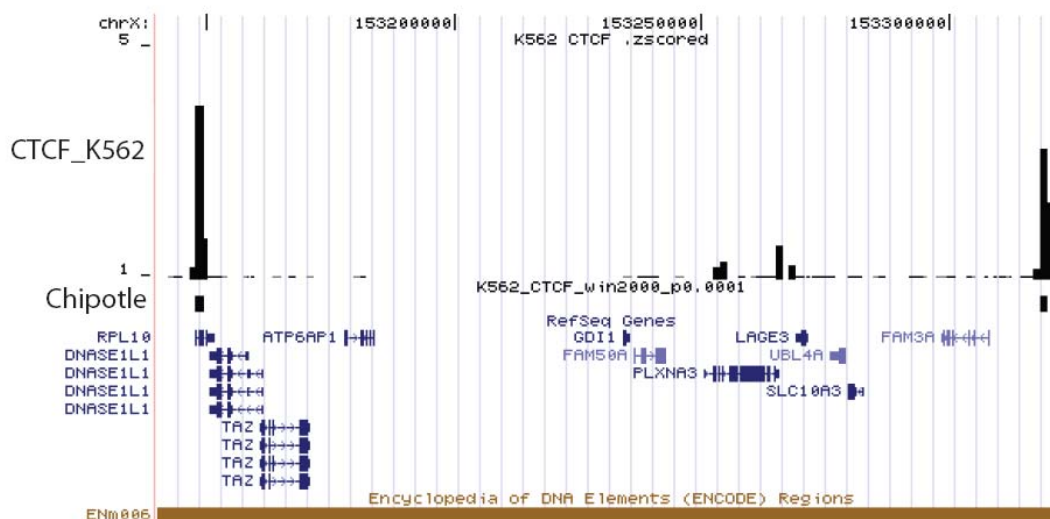


Figure 4.12: CTCF binding sites flank a cluster of genes in ENCODE region Enm006. 17 RefSeq genes (which include a number of alternative transcripts) are flanked by two CTCF sites in K562. The top track (CTCF_K562) indicates the \log_2 fold enrichments in this region and the ChIPOTle defined peaks are indicated below the x-axis. The chromosome X coordinates are indicated at the top of the figure and known RefSeq genes are indicated in blue at the bottom of the figure.

Coordinates	Description	Number of genes Flanked by CTCF
chr21:33,730,000-34,120,000	unrelated	9
chrX:153,140,000-153,323,000	unrelated	17
chrX:153,627,437-153,781,938	F8A genes	5
chr19:59,223,959-59,346,770	unrelated	8
chr19:59,701,626-59,886,317	leukocyte immunoglobulin-like receptors (LILR) cluster	8
chr16:162,168-320,833	α globin region	9
chr11:5,099,780-5,277,643	β globin region	8
chr11:5,570,932-5,669,811	tripartite motif (TRIM) cluster	7
chr7:26,852,428-26,953,093	hoxa cluster	7
chr15:41,851,756-41,958,421	unrelated	5

Table 4.1: Gene clusters flanked by CTCF binding sites. 10 gene clusters (containing five or more genes) were flanked by CTCF sites. The start and end coordinates (human genome release hg17) of the regions flanked by CTCF binding sites are indicated, along with a description and the number of genes contained within each cluster.

In contrast to CTCF sites flanking clusters of genes, a number of individual genes contained several intergenic CTCF binding sites, such as the SYN3 gene which is

associated with 16 CTCF sites (Figure 4.13). Nine ENCODE genes contained five or more intergenic CTCF binding sites (Table 4.2). Gene Ontology (GO) annotations were examined for these nine genes to determine if any GO terms were enriched in genes containing multiple CTCF sites. 7 of these contained a GO annotation and five were annotated with the integral to membrane GO term (p0.01 or less). Although the significance of multiple CTCF binding sites within genes that code for integral membrane proteins is not known, it suggests that these either these genes have an unusually complex enhancer-blocking mechanism or that CTCF may have a role as a transcription factor at enhancers/repressor elements which are often located within genes.

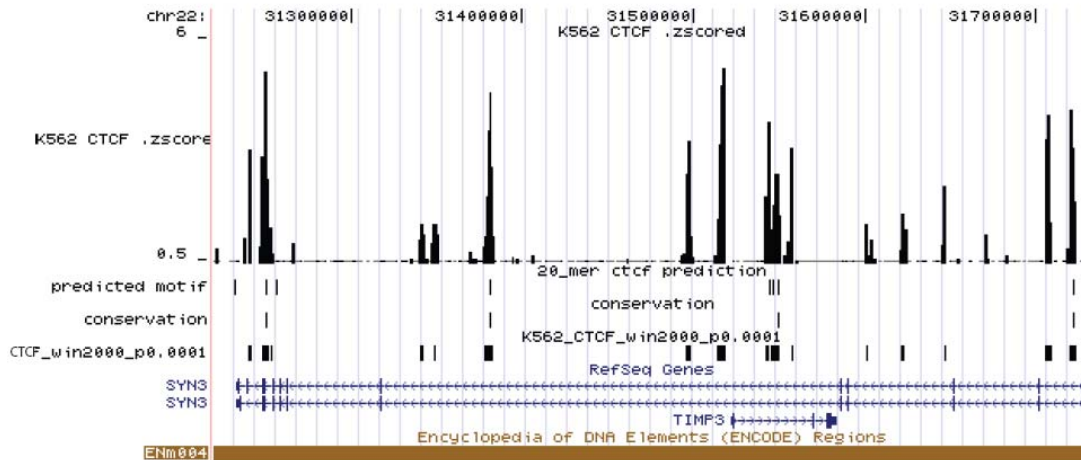


Figure 4.13: Genes can contain several CTCF binding sites. The SYN3 gene encodes a member of the synpasin family of proteins and is associated with 16 CTCF binding sites in K562 cells. The TIMP3 gene encodes a member of the tissue inhibitors of the matrix metalloproteinases family and is located within an intron of this gene. It is transcribed in the opposite direction to SYN3 and a number of the CTCF binding sites may be associated with TIMP3 regulation.

Coordinates	Gene	Description	CTCF sites
chr7:116,187,332-116,464,024	ST7	suppression of tumorigenicity 7	6
chr22:31,233,094-31,727,237	SYN3	synapsin III	16
chr5:142,130,476-142,586,243	ARHGAP26	rho GTPase-activating protein 26	8
chr7:125,672,608-126,477,261	GRM8	glutamate receptor metabotropic 8 precursor	6
chr11:130,745,779-131,710,752	AY358331	member of the IgLON (LAMP, OBCAM, Ntm) family of immunoglobulin (Ig) domain-containing glycosylphosphatidylinositol (GPI)-anchored cell adhesion molecules	8
chr13:112,392,644-112,589,467	ATP11A	integral membrane ATPase, Class VI, type 11A	7
chr2:220,204,557-220,228,997	ACCN4	amiloride-sensitive cation channel 4	5
chr5:141,953,307-142,045,812	FGF1	fibroblast growth factor 1	5
chr11:64,130,222-64,247,236	NRXN2	neurexin 2 isoform alpha-2 precursor	5

Table 4.2: ENCODE Genes containing multiple CTCF binding sites in K562 are membrane components. The genes containing multiple CTCF sites are involved in cell signaling or cell adhesion processes. GO cellular component annotation indicates that five of the nine genes are integral to membrane formation.

Thus in summary, the distribution of CTCF binding sites is complex as it can be found flanking individual genes or groups of genes or multiple CTCF sites can be found within individual genes. This complex binding pattern suggests a multi-functional role for CTCF in the regulation of gene expression.

4.7. A comparative and sequence based analysis of CTCF sites in different cell types

In order to investigate the conservation of CTCF binding in different cell types, the human cell line U937, established from a patient with generalised histiocytic lymphoma (Sundstrom and Nilsson, 1976) and displaying properties of monocytes (Anderson and Abraham, 1980) was used to perform CHIP-chip analysis of CTCF binding sites. Three

biological replicate experiments were performed as described for K562, the median data was calculated and IgG normalised as described previously. 661 U937 CTCF sites were identified in the ENCODE regions by ChIPOTle (p0.0001) compared to 571 CTCF sites in K562. The locations of CTCF sites in K562 and U937 cells were also compared with the data of Kim *et al.* (2007) and Barski *et al.* (2007). Kim *et al.* predicted the location of 412 CTCF binding sites in the ENCODE regions based on a 20-mer consensus motif. 172 of these predicted sites were conserved at the sequence level in at least one other vertebrate genome, excluding the chimpanzee genome. The number of experimentally determined K562 CTCF sites that overlapped with the location of a consensus binding motif was 187 (33% of 571 sites identified), 121 of which were conserved sites (Table 4.5). Therefore 121 of the 172 (70%) predicted and conserved CTCF sites in the ENCODE regions were bound in K562. The number of experimentally determined U937 CTCF sites that overlapped with predicted sites was 195 (29.5% of 661 sites identified), 126 of which were conserved sites (Table 4.3). Therefore 126 of the 172 (73%) predicted and conserved CTCF sites were bound in U937. This suggests that *in silico* predicted CTCF sites that are conserved in at least one other vertebrate genome is a relatively accurate predictor of CTCF binding *in vivo*. While the specificity of this approach is high (70-73% of predicted and conserved sites are bound in K562 and U937 cells respectively), the sensitivity is low as 79% and 81% of experimentally determined binding events in K562 and U937 respectively were not identified by this approach. Therefore the majority of CTCF sites identified in K562 and U937 cells were not associated with the consensus motif defined by Kim *et al.* (2007).

Cell line	Number of experimental CTCF sites	Number of predicted CTCF sites	Number of predicted and conserved CTCF sites	Number of experimental sites located at predicted sites	Number of experimental sites located at conserved sites
K562	571	412	172	187	121
U937	661	412	172	195	126

Table 4.3: Correspondence between ChIP-chip defined and predicted CTCF sites. 412 CTCF binding sites were predicted in the ENCODE regions by Kim and colleagues based on the presence of a 20-mer

consensus motif (Kim *et al.*, 2007). 172 of these predicted sites were conserved at the sequence level in at least one other vertebrate genome. 187 and 195 of the CTCF binding sites identified in K562 and U937 cells by ChIP-chip analysis overlapped with the CTCF consensus motif. 121 and 126 of the experimentally identified CTCF sites in K562 and U937 were associated with predicted CTCF sites that were also conserved in at least one other vertebrate genome.

A comparison of K562 and U937 CTCF sites identified in this study was also performed. Of the 571 and 661 binding sites identified in K562 and U937 cells respectively, 393 sites (69%) overlapped between the two cell lines (Figure 4.14). This suggests that almost a third of CTCF sites are involved in cell-type specific regulation. ENCODE region Enm004 contained the greatest number of overlapping sites, 36 of 50 sites (72%) identified in the two cell lines overlapped (Figure 4.14). As described previously, Ren and colleagues had performed a genome-wide study of CTCF binding in IMR90 cells and identified 225 CTCF sites in the ENCODE regions (Kim *et al.*, 2007). 182 (31%) of IMR90 sites overlapped with the 571 K562 binding sites identified as part of this study. Barski and colleagues used ChIP-sequencing to identify 20,262 CTCF binding sites in CD4+ cells (Barski *et al.*, 2007), of which 353 were located in the ENCODE regions. 227 and 232 of these 353 sites overlapped with CTCF sites in K562 and U937 cells respectively, representing an overlap of 40% and 35%. Therefore between 31%-40% of K562 CTCF sites identified in this study overlapped with CTCF sites reported by two other studies.

As the study of Kim *et al.* (2007) had also examined U937 CTCF binding sites in the ENCODE regions, a direct comparison of the two U937 datasets was performed. 138 of the 232 CTCF sites identified by Kim and colleagues overlapped with U937 CTCF sites identified in this study (Figures 4.14 panels B and D). The gene-rich Enm004 region on chromosome 22 contained the highest number of overlapping sites at 21 (Figure 4.14, panel D). However, 39% of U937 sites identified by Kim and colleagues were not identified in this study and almost three times more CTCF binding sites were identified in this study. The possible reasons for the differences in data sets derived from the same cell line are outlined in the discussion of this Chapter.

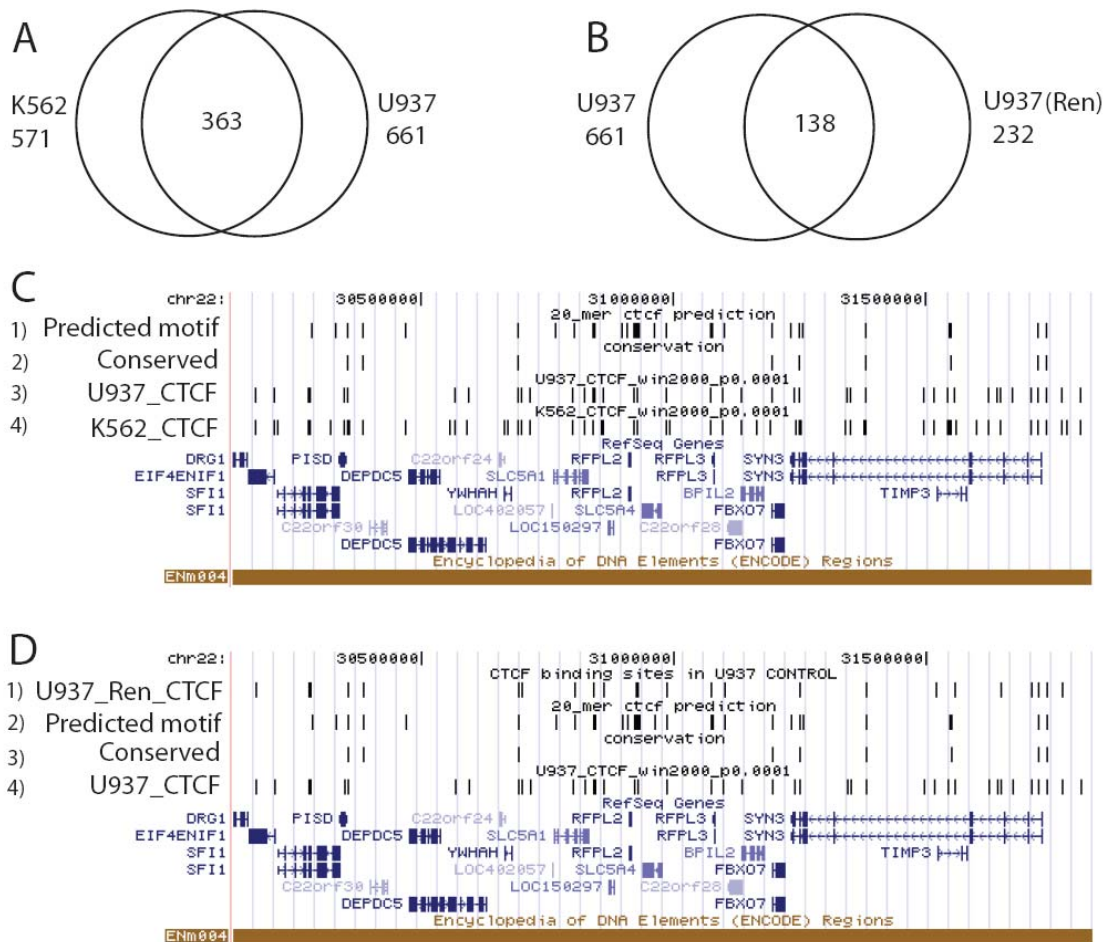


Figure 4.14: A comparison of CTCF binding sites in different cell types and between studies. Panel A: 571 and 661 binding sites were identified in K562 and U937 as part of this study. 363 of these sites overlapped in the two different cell types. Panel B: 661 CTCF sites that were identified as part of this study were compared with the 232 CTCF sites identified by Kim and colleagues in U937 cells. 138 sites overlapped between the two studies. Panel C: A UCSC screenshot of ENCODE region enm004, which contained the greatest number of overlapping CTCF sites in K562 and U937 cells (36 sites) identified in this study. Track number 1 (predicted motif) indicates the location of consensus CTCF sites, track 2 (conserved) indicates those consensus sites that are conserved in at least one other vertebrate genome, tracks 3 (U937_CTCF) and 4 (K562_CTCF) indicate the location of ChIPOTle defined CTCF binding sites in U937 and K562 cells respectively. Panel D: ENCODE region enm004 also contained the greatest number of overlapping CTCF sites in two independent studies on CTCF binding in U937 cells (21 sites). Track number 1 (U937_Ren_CTCF) indicates the location of CTCF sites identified in U937 by Kim and colleagues (2007), track 2 (predicted motif) shows the location of predicted CTCF sites, track 3 (conserved) indicates which predicted sites are conserved in at least one other vertebrate genome, and tracks 4 (U937_CTCF) indicates the location ChIPOTle defined CTCF binding sites in U937 cells as part of this

study. Enm004 chromosome coordinates are shown at the top of panels C and D, while the RefSeq genes are indicated below the data tracks.

4.8. Analysis of other transcription factors implicated in CTCF or insulator function

4.8.1. Developing assays using the SCL locus as model system

As discussed in the introduction, CTCF associates with a number of proteins and these interactions are important for modulating the function of CTCF. CTCF is known to interact with mSin3a (Lutz *et al.*, 2000) to mediate transcriptional repression. Therefore mapping sites of mSin3a interactions would allow for a more detailed picture of whether this protein is important for CTCF function genome-wide. The SCL microarray was used to develop a ChIP-chip assay to detect mSin3a interactions in K562 cells (Figure 4.15) and the data was normalised with a normal rabbit IgG control. Four peaks of significant enrichment were detected for mSin3a interaction at the SCL locus, namely the +53 region, the SCL promoter region, the SIL promoter and the KCY promoter. Three of the four regions also bound CTCF (the +53 region, the SIL promoter and the KCY promoter) suggesting that CTCF may function as a classical transcription factor at these regions. However, it was a surprising to find mSin3a at the promoter region of three actively transcribed genes (in addition the +53 region also displays bi-directional promoter activity in K562 cells; Dhimi, submitted).

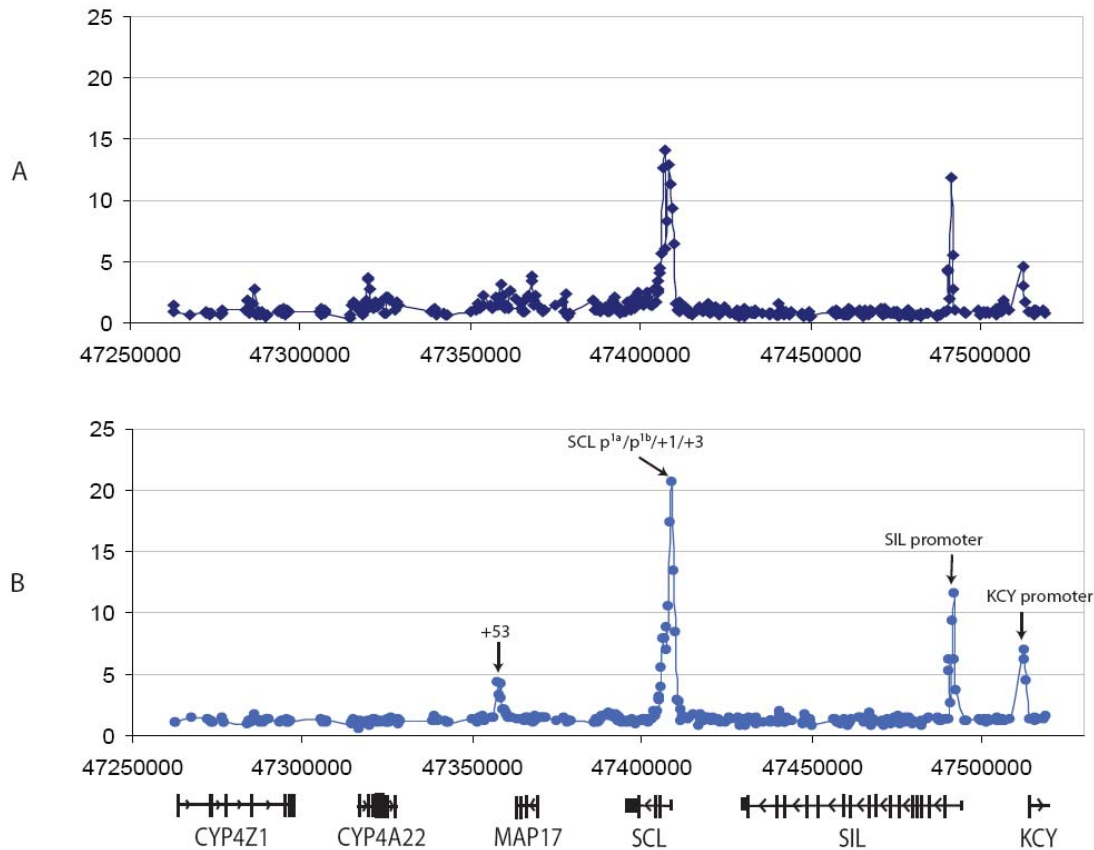


Figure 4.15: ChIP-chip profile of mSin3a interactions across the SCL locus in K562 before and after normal rabbit IgG normalisation. Panel A: Fold enrichments reported for mSin3a interactions before normalisation with normal rabbit IgG. Panel B: fold enrichments increased at mSin3a interacting sites after normalisation with normal rabbit IgG. The location of the +53 region, the SCL, SIL, and KCY promoters are indicated by black arrows. The human chromosome 1 genomic coordinates are indicated along the y-axis, while fold enrichments are indicated on the x-axis. Gene order and direction of transcription is shown below panel B.

In contrast to mSin3a, USF1 and USF2 are known to bind in close proximity to CTCF at the chicken β -globin HS4 insulator and are responsible for the chromatin barrier activity of this insulator (West *et al.*, 2004; Huang *et al.*, 2007). Thus, mapping USF1 and USF2 interactions in combination with that of CTCF and histone modification data may allow for chromatin barrier insulators to be identified in the human genome. USF1 and USF2 ChIP-chip assays were tested using the SCL microarray but no significant enrichments were detected (data not shown). However, the ENCODE array had previously been used in ChIP-chip assays to detect a number of USF1 binding sites in HepG2 cells (Rada-

Iglesias *et al.*, 2005). USF1 and USF2 ChIP-chip assays were therefore tested using the ENCODE array and several hundred USF1 and USF2 binding sites were identified in K562 cells using this array as described in the following section.

4.8.2. Mapping the distribution of mSin3a, USF1, and USF2 binding sites in the ENCODE regions

In order to further characterise the 571 K562 ENCODE CTCF sites, ChIP-chip experiments were performed with K562 cells to detect mSin3a, USF1 and USF2 interactions in the ENCODE regions. Microarray experiments were performed and the data normalized with the relevant mock IgG data as described earlier in this Chapter. ChIPOTle was then used to identify peaks of enrichment at a high confidence ($p < 0.0001$) for each transcription factor. 483 binding sites were identified for USF1, 219 sites were identified for USF2, and 310 mSin3a interactions were mapped. This represented a substantial number of peaks in 1% of the genome, suggesting that like CTCF these transcription factors may regulate the expression of a large number of genes across the genome. The distributions of mSin3a, USF1, and USF2 binding sites were determined with respect to the location of promoters, distal enhancer/ repressors, and other sites (Figure 4.16)

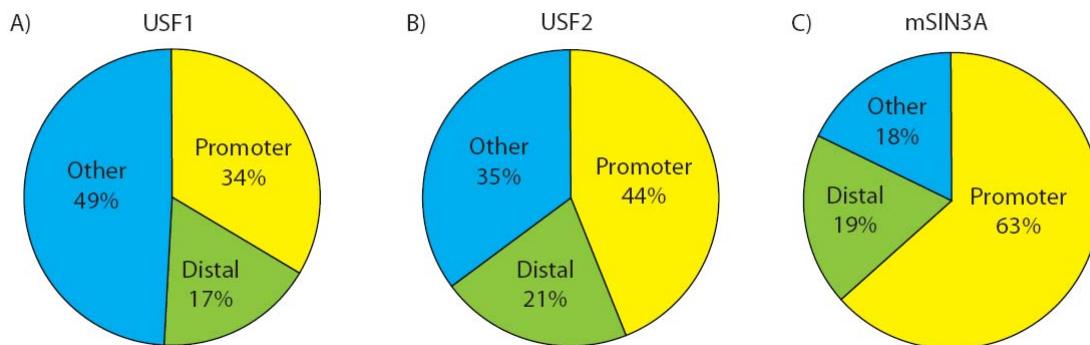


Figure 4.16: The distribution of USF1, USF2, and mSin3a sites of interaction in the ENCODE regions in K562 cells. The ChIPOTle sites for USF1 (A), USF2 (B) and mSin3a (C) were mapped to promoters (within 2.5kb of known transcription start sites), distal enhancer/repressors (associated with H3K4me1, H3K4me2, or H3K4me3 and not within 2.5kb of a TSS) and other sites which were not associated with H3K4 methylation or within 2.5kb of promoters. The percentages of sites that map to each genomic feature are indicated.

The distribution of USF1 binding sites mirrored that of CTCF as approximately equal percentages of binding events were identified (31% of CTCF sites were located at promoters, 16% were at distal sites and 53% were located at other sites). USF2 binding was marginally more biased towards promoter elements than USF1 but broadly similar to the pattern of binding observed for USF1 and CTCF. However, in contrast, the distribution of mSin3a interactions was found to be heavily biased towards promoters. This finding is consistent with the known role of mSin3a in binding close to or at promoters to repress transcription by recruiting other proteins such as histone deacetylases (Dannenber *et al.*, 2005).

To more accurately map the locations of mSin3a, USF1, and USF2 binding sites at promoters, their binding patterns were mapped with respect to the location of the nearest TSSs (Figure 4.17). The majority of mSin3a binding sites were located within 2 kb of TSSs and further analysis determined that 82% of the TSSs associated with an mSin3a binding event (less than 1 kb from a TSS) in K562 cells were also associated with CpG islands. In contrast only about 50% of the ENCODE gene promoters are associated with a CpG island. This suggests that mSin3a associates more readily with promoters with CpG islands and may interact with some proximal promoter sequence binding proteins that are specific to CpG-containing promoters. In contrast, as discussed above, the binding pattern of USF1 and USF2 was somewhat similar to CTCF as binding sites were not restricted to promoter regions - approximately half of the sites for these TFs were located 5 kb or more from TSSs.

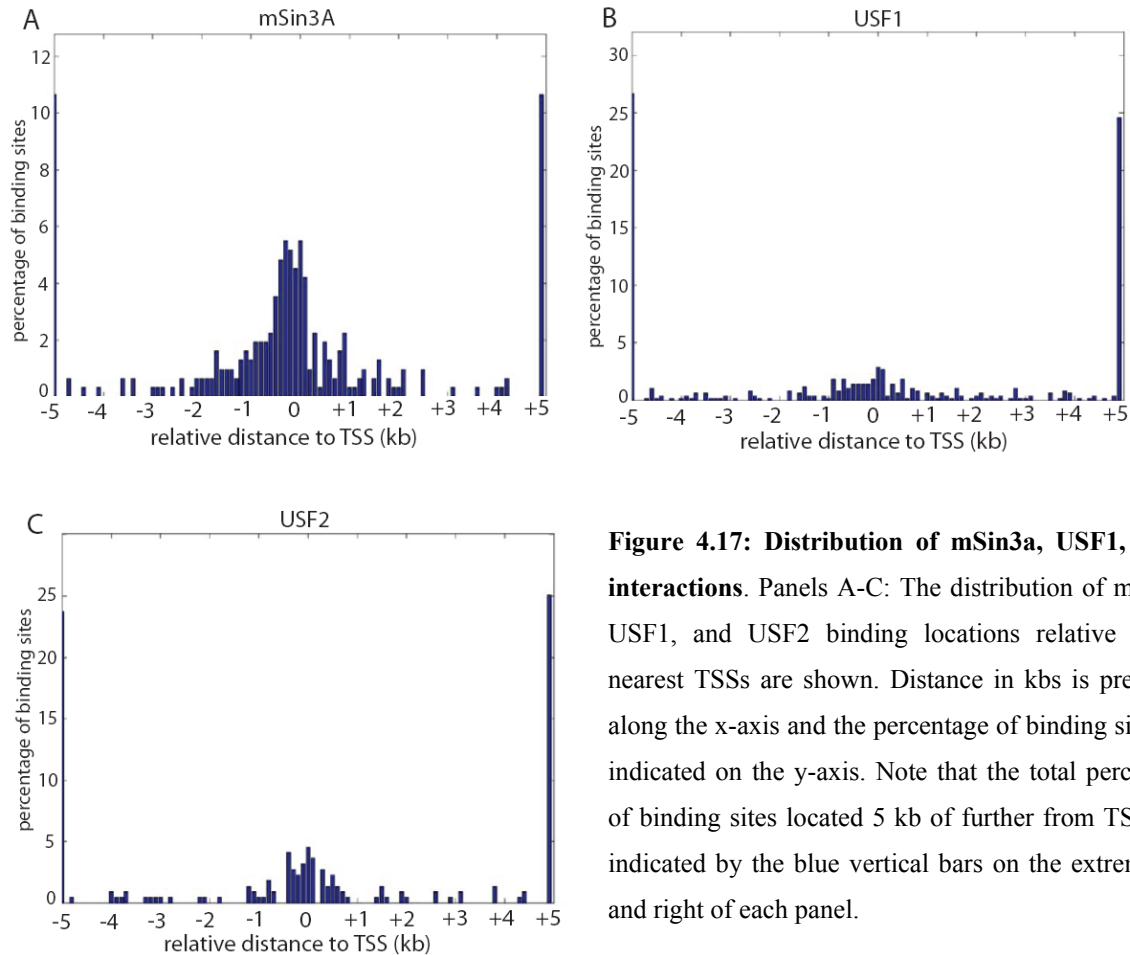


Figure 4.17: Distribution of mSin3a, USF1, USF2 interactions. Panels A-C: The distribution of mSin3a, USF1, and USF2 binding locations relative to the nearest TSSs are shown. Distance in kbs is presented along the x-axis and the percentage of binding sites are indicated on the y-axis. Note that the total percentage of binding sites located 5 kb of further from TSSs are indicated by the blue vertical bars on the extreme left and right of each panel.

4.8.3. Analysing interactions between CTCF and mSin3a, USF1, and USF2

CTCF is known to interact with mSin3a (Lutz *et al.*, 2000) while USF1 and USF2 bind in close proximity to CTCF at the HS4 chicken β -globin insulator (West *et al.*, 2004). Therefore, potential interaction or co-localisation of CTCF with mSin3a, USF1, and USF2 was examined by analysing the extent of overlapping ChIPOTle sites. The number of CTCF sites which overlapped with interactions for one or more of the other factors were determined (Table 4.4). The binding events were then categorised into those found at transcription start sites (TSSs), at distal enhancer/repressor sites, and “other” sites. Distal enhancer/repressor sites were defined as regions associated with a peak of enrichment for H3K4me1, or H3K4me2, or H3K4me3, and not within 2.5 kb of a TSS (Chapter 3). Sites defined as “other” are not TSSs or distal enhancer/repressors (according to the definitions used in this study). Overlapping interactions at TSSs were

further sub-categorised into those located at active and inactive TSSs (as determined by available gene expression data for K562).

	CTCF only	CTCF +mSin3a only	CTCF +USF1 only	CTCF +USF2 only	CTCF +mSin3a +USF1	CTCF +mSin3a +USF2	CTCF +USF1 +USF2	CTCF +mSin3a +USF1 +USF2
All sites	384	47	37	0	15	31	31	26
TSSs	73	33	13	0	11	14	14	18
Active TSSs	29	17	3	0	8	7	7	12
Inactive TSSs	11	1	0	0	0	1	1	3
Distal enhancer/repressors	52	9	11	0	4	5	5	7
Other sites	259	5	13	0	0	12	12	1

Table 4.4: Overlapping combinations of CTCF, mSin3a, USF1, and USF2 binding sites in K562.

ChIPOTle hits for each of the four transcription factors were analysed and the number of overlapping binding events are shown for the different factor combinations. The data is mutually exclusive for each of the seven combinations, for example binding sites shared by CTCF and mSin3a indicates that USF1 and USF2 binding sites do not overlap. Sites at TSSs are defined as binding events within 2.5kb of transcription start sites (TSSs), Distal enhancer/repressor sites are defined as sites containing either a H3K4me1, or H3K4me2, or H3K4me3 ChIPOTle peak not within 2.5 kb of a TSS. The remaining sites were classified as other sites. Where possible, TSSs associated with binding events were classified as active or inactive using Affymetrix expression microarray data (Chapter 3). Note: binding sites that spanned a TSS and a distal site were not counted twice but were preferentially assigned as a TSS binding event.

A number of observations arose from this analysis, demonstrating that the distribution of CTCF, mSin3a, USF1, and USF2 overlapping sites is complex. Furthermore, given that the number of sites analysed in the groups shown in Table 4.4 were relatively small, it is difficult to determine an accurate picture of the relationship of these four regulators and how they act in combination to regulate gene expression.

The majority (67.3%) of CTCF binding sites (384/571) do not overlap with either mSin3a, USF1 or USF2, whilst the remaining 32.7% of CTCF sites (187/571) do overlap

with sites of interaction for one or more of the three factors. This latter figure was broken down into the following proportions:

- 8.2% of CTCF binding sites overlap with mSin3a binding sites only (CTCF+mSin3a)
- 6.5% of CTCF binding sites overlap USF1 binding sites only (CTCF+USF1)
- 2.6% of CTCF binding sites overlap with both mSin3a and USF1 binding sites (CTCF+mSin3a+USF1)
- 5.4% of CTCF binding sites overlap with both mSin3a and USF2 binding sites (CTCF+mSin3a+USF2)
- 5.4% of CTCF binding sites overlap with both USF1 and USF2 binding sites (CTCF+USF1+USF2)
- 4.5% of CTCF binding sites overlap with all three of mSin3a, USF1 and USF2 binding sites (CTCF+mSin3a+USF1+USF2)

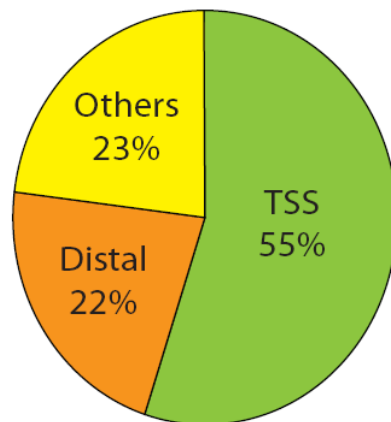


Figure 4.18: CTCF binding sites that overlap with binding sites of mSin3a, USF1, and USF2 are found in diverse locations. 55% of CTCF sites that overlap with one or more of the other factors are located at TSSs, 22% are at distal sites and 23% are found at other locations.

These co-localising interactions were located at TSSs, distal sites and “other” sites (Figure 4.18) and implicated CTCF and these other factors in gene activation, gene repression, and/or insulator functions. The majority (55%) of CTCF binding sites that overlapped with mSin3a or USF1 or USF2 were located at TSSs and may be involved in regulating gene expression. 22% of overlapping interactions were located at distal enhancer/repressor sites and may be also involved in regulating gene expression or may act as enhancer-blocking insulators by binding at or in close proximity to enhancers. The remaining 23% of overlapping interactions were located at “other” sites, which represent

good candidates for insulator function as they are not associated with promoter or enhancer/repressor activity.

Upon examination of the distribution of co-localisation sites at TSSs, over two-thirds of CTCF+mSin3a overlapping interactions (33/47) were located at TSS consistent with the finding that CTCF interacts with mSin3a to repress transcription (Lutz *et al.*, 2000). However, data was available for the expression status of 18 of the genes associated with these TSS and 17 were associated with active and only 1 was inactive. This suggests that CTCF together with mSin3a may be involved in activating gene expression rather than repression. This was a surprising finding as mSin3a is best known as a co-repressor protein that recruits histone deacetylases to silence gene expression (Heinzel *et al.*, 1997) although more recent evidence suggests that its yeast homolog can also function in gene activation (De Nadal *et al.*, 2004).

Overall, of those CTCF sites which overlapped with at least one other of the three TFs (mSin3a, USF1, or USF2) at TSSs (103/187 as described above) - 52.4% (54/103) of these sites were associated with active gene expression. Specifically, approximately one-third of CTCF+USF1 overlapping sites (13/37) were at TSSs but no conclusive evidence regarding their association with active gene expression or repression could be determined as data was only available for three of these genes. Nearly three-quarters of CTCF+mSin3a+USF1 overlapping sites were located at TSSs (11/15) and 8/8 (for which expression data was available) were associated with active gene expression. Approximately half of CTCF+mSin3a+USF2 overlapping sites were also located at TSSs (14/31) and 7/8 were associated with active gene expression (as was the case for genes where CTCF+USF1+USF2 co-localised). Finally, approximately 70% of the CTCF+mSin3a+USF1+USF2 overlapping sites (18/26) were also located at TSSs, the majority of which were associated with actively expressed genes (12/15).

Overlapping sites for the four regulators were then examined at distal enhancer/repressor elements. Fewer overlapping sites were observed at these locations. 16-30% of overlapping interactions were located at distal sites but the functional relevance of these interactions is not known as the sample sizes were small. CTCF and these factors may be involved in regulating gene expression binding acting as classical transcription factors

that bind enhancer/repressor elements or alternatively may be acting as enhancer-blocking insulators.

Finally, CTCF overlapping sites were then examined at the “other” sites, which may represent true insulator elements that could be distinguished from promoters or distal enhancers/repressors. Again, very few overlapping CTCF+mSin3a sites were observed at “other” sites (only 5 of 47), consistent with CTCF interacting primarily with mSin3a at promoters to regulate gene expression. However, one-third of CTCF+USF1 overlapping sites (13/37) were located at “other” sites and may be involved in insulator functions. No CTCF+mSin3a+USF1 sites or CTCF+USF2 sites were located at “other” sites. Of CTCF+mSin3a+USF2 sites and CTCF+USF1+USF2 sites, one-third were at “other” locations (12/31 in both cases). Overall, only 14.6% of “other” sites (38/259) showed co-localisation of CTCF with either USF1 or USF2, suggesting that their co-localisation may not be a general paradigm for insulator barrier function in the human genome (see section 4.9).

4.8.4. Transcription factor binding and gene expression status

In the previous section it was noted that CTCF sites at TSSs that overlapped with mSin3a or USF1 or USF2 sites were predominantly associated with active gene expression. This observation was explored further by investigating whether CTCF, mSin3a, USF1, or USF2 binding at TSSs could be used to predict the expression status of a gene by plotting receiver operating characteristic (ROC) curves (Figure 4.19) (Analysis performed by Dr. Ulas Karaöz, Boston University). The ROC of a classifier shows its performance as a compromise between selectivity and sensitivity. In this case CTCF, mSin3a, USF1, and USF2 binding at TSSs are used as a classifier of gene expression status at 238 K562 Affymetrix probe sets (on/off state based on present or absent MAS5 calls as described in Chapter 2). The plots illustrated in Figure 4.19 show sensitivity at all possible specificities and indicate that mSin3a binding at TSSs is highly predictive of active gene expression. The maximum accuracy value is reported for each TF and a value of 0.7824 was obtained for mSin3a, which is identical to the value obtained for H3K4me3 (Chapter 3). Therefore the binding of mSin3a at TSSs is highly predictive of active gene expression. USF2 shows a relatively high value also of 0.7098 while USF1 is less

predictive (0.6788) and CTCF is the least predictive (0.658). CTCF displays a similar value to H3K4me1 (0.6477) and the CTCF ROC curve is found to be negatively correlated with gene expression at low thresholds as observed by the curve located below the diagonal line. Therefore CTCF binding at TSSs is the least accurate predictor of active transcriptional state. This finding implicates mSin3a in active gene expression which is surprising given its accepted role in gene repression (as mentioned above). However, recent work has implicated the yeast homolog Sin3 in gene activation (De Nadal *et al.*, 2004) and this is discussed in greater detail in section 4.10.

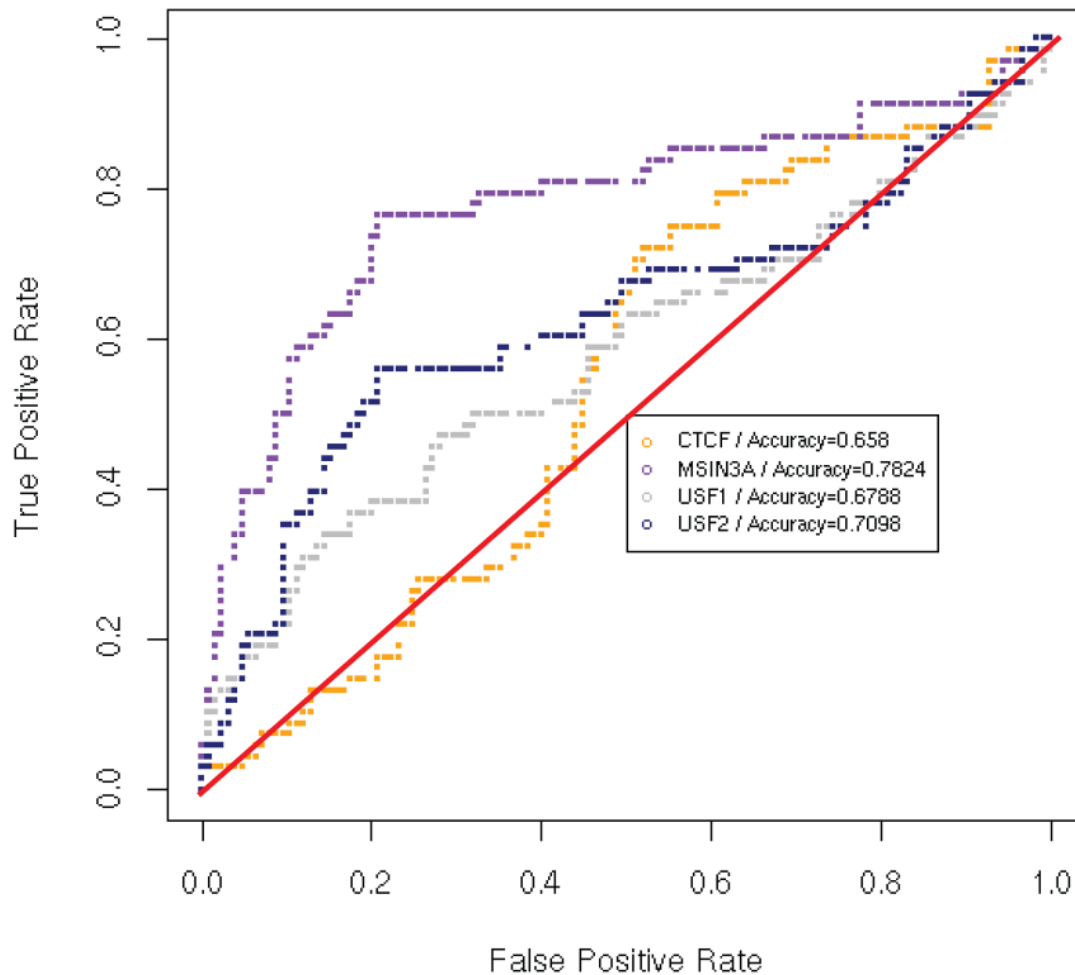


Figure 4.19: Predictive power of transcription factor binding for gene expression in K562. Receiver operating characteristic (ROC) curves were plotted for 1kb regions around TSSs to determine the association between transcription factor (TF) binding and gene expression. ROC curves illustrate the predictive accuracy of TF binding on classifying the expression states of genes (on/off). The red diagonal line represents the ROC curve of a TF that is randomly associated with active or inactive gene expression.

TF binding that is positively associated with the gene on state will have a ROC curve above the diagonal line, while a TF associated with the gene off state will have a ROC curve below the diagonal. A TF associated randomly with expression state of genes achieves a ROC score of 0.5 (red line) while a TF that is a perfect predictor of gene expression state receives a ROC score of 1.0. The maximum ROC score is indicated for each factor and the highest score was obtained for mSin3a (0.7824).

4.8.5. Motifs analysis of CTCF, mSin3a, USF1 and USF2 binding sites

As 384 (67%) and 466 (70%) CTCF sites identified in K562 and U937 cells did not contain the predicted consensus CTCF motif, a computational analysis was performed to determine if any other known motifs were enriched in the CTCF binding sites. Motif matrices from TRANSFAC (Matys *et al.*, 2006) and JASPER (Bryne *et al.*, 2007) databases were used to scan the CTCF sites. 1 kb centred sequences from the ChIPOTle hits were extracted as the foreground sequence, while 1 kb sequences flanking the ChIPOTle sites were defined as background sequences. The best known motif which distinguishes the foreground and background sequences with a relative error rate of 0.381 was motif PF0045 from the JASPAR database (Table 4.5). This motif was previously identified in a study of regulatory motifs identified in human promoters and 3' UTRs by comparative analysis of several mammalian genome sequences (Xie *et al.*, 2005) and more recently was shown to bind CTCF (Xie *et al.*, 2007). This motif also matches the core of the longer 20-mer CTCF motif reported by Ren and colleagues. One other motif from JASPAR (PF0156) and three other motifs from TRANSFAC were also enriched, however, the motif that matched the known CTCF consensus sequence distinguished the foreground and background sequences with the highest sensitivity and specificity coupled with the lowest relative error rate.

Name	Logo	Sn	Sp	Error	pvalue
1. PF0045		0.651	0.586	0.381	0
2. PF0156		0.578	0.554	0.434	0
3. M00325		0.653	0.473	0.437	0
4. M00411		0.511	0.603	0.443	0
5. M01057		0.392	0.714	0.447	0

Table 4.5: Known motifs associated with CTCF binding sites. Motif matrices from JASPAR and TRANSFAC databases were used to search for enrichment in CTCF binding sites. The PF0045 motif in the JASPAR database, which matches the known CTCF motif, was found to be the best motif for distinguishing the foreground and background sequences with a relative error rate of 0.381, sensitivity (Sn) of 0.651, and specificity (Sp) of 0.586. The sensitivity associated with a motif and p-value cut-off is the proportion of true foreground sequences that were classified as foreground; the specificity is the proportion of true background sequences classified as background sequences. The relative error rate (Error) for a given motif and associated with a particular p-value cut-off is $1-(Sn+Sp)/2$. A good motif has a low error rate and balanced Sn and Sp values (Smith *et al.*, 2005). DNA logos are presented for each motif and the height of each letter indicates relative occurrence of nucleotides in the binding sites.

Because TRANSFAC and JASPAR can only be used to scan sequences for the presence of known motifs, the DME (discriminating matrix enumerator) program (Smith *et al.*, 2005) was used to search for the presence of novel motifs in the CTCF binding sites. The DME motif discovery algorithm calculates motif relative over-representation between two sets of sequences- foreground (identified binding sites) and background (sequences in the vicinity of identified binding sites) - using a strategy of enumerating position-weight matrices. Motifs with length from 8-12 nucleotides were searched for by extracting 1 kb centred sequences from the ChIPOTle sites and defining these as foreground sequences while 1 kb sequences flanking the ChIPOTle sites were defined as background sequences. The top DME motif which distinguished foreground and background sequences with a relative error rate of 0.381 was consistent with the core of the known CTCF consensus motif (Table 4.6) (Note DME retrieves the reverse

compliment of a binding motif). Therefore DME identified no novel motifs in the CTCF binding sites.






Name	Logo	Sn	Sp	Error	pvalue
1. DME0001		0.555	0.683	0.381	0
2. DME0002		0.566	0.574	0.43	0.0002
3. DME0003		0.357	0.78	0.431	0.0003
4. DME0004		0.737	0.399	0.432	0.0003
5. DME0005		0.405	0.728	0.434	0.0008

Table 4.6: De novo CTCF motif discovery. The DME program (Smith A PNAS) was used to identify novel motifs in CTCF binding sites. DME0001 motif was found to be the best motif for distinguishing the foreground and background sequences with a relative error rate of 0.381 at a p-value cutoff of 0. This motif was consistent with the core of the known CTCF motif (positions 5-16) as it is identified in reverse compliment by the DME program. The sensitivity (Sn) associated with a motif and p-value cut-off is the proportion of true foreground sequences that were classified as foreground; the specificity is the proportion of true background sequences classified as background sequences. The relative error rate (Error) for a given motif and p-value cut-off is $1-(Sn+Sp)/2$. A good motif has a low error rate and balanced Sn and Sp values (Smith *et al.*, 2005). DNA logos are presented for each motif and the height of each letter indicates relative occurrence of nucleotides in the binding sites.

USF1 and USF2 are basic helix-loop-helix TFs and this family of TFs generally binds to a consensus sequence of 5'-CANNTG-3' (where N is any nucleotide) called enhancer box (E-box) motif which is the second most conserved motif in higher eukaryotes (Xie *et al.*, 2005). A large number of USF1 and USF2 binding sites were identified in the course of this study and this data could be used to confirm the presence of E-box motifs in USF sites or identify a novel USF binding motif. Similarly a large number of mSin3a interactions were identified in the ENCODE regions but as mSin3a does not directly interact with DNA itself but instead interacts with sequence specific DNA binding proteins, an analysis of mSin3a interacting sequences may identify a motif associated with its DNA-binding partner. To these ends, the mSin3a, USF1 and USF2 data sets from the K562 cell line were analysed for enrichment of known motifs contained in JASPAR

(Bryne *et al.*, 2007) and TRANSFAC (Matys *et al.*, 2006) databases and for novel motifs using the DME program (Smith *et al.*, 2005) as described previously. No known motif in JASPAR or TRANSFAC displayed high sensitivity and specificity in distinguishing foreground sequences from background sequences in the mSin3a dataset (data not shown). Therefore the identification of novel motifs by the DME program was investigated. Several novel motifs were discovered to be associated with mSin3a interactions and may represent potential binding motifs for an mSin3a interacting factor (Table 4.7). However, it must be noted that many of these motifs have a high GC content and many of the mSin3a binding sites overlap with promoter regions that have CpG islands.






Name	Logo	Sn	Sp	Error	pvalue
1. DME0001		0.626	0.565	0.404	0
2. DME0002		0.432	0.752	0.408	0
3. DME0003		0.371	0.806	0.411	0.0002
4. DME0004		0.619	0.556	0.412	0.0003
5. DME0005		0.565	0.594	0.421	0.0019

Table 4.7: Discovery of novel motifs associated with mSin3a interactions. The DME program was used to identify novel motifs in genomic regions associated with mSin3a binding. The GC-rich DME0001 motif was found to be the best motif for distinguishing the foreground and background sequences with a relative error rate of 0.404 at a p-value cutoff of 0. The sensitivity (Sn) associated with a motif and p-value cut-off is the proportion of true foreground sequences that were classified as foreground; the specificity is the proportion of true background sequences classified as background sequences. The relative error rate (Error) for a given motif and p-value cut-off is $1-(Sn+Sp)/2$. A good motif has a low error rate and balanced Sn and Sp values (Smith *et al.*, 2005). DNA logos are presented for each motif and the height of each letter indicates relative occurrence of nucleotides in the binding sites.

The E-box motif (MA0093) in the JASPAR database best distinguished foreground sequences from background sequences for the USF1 binding sites (Table 4.8), followed closely by the USF1 motif from TRANSFAC (M00121), which contains an E-box motif at its core. No known USF2 binding motif is present in these databases but this study shows that the USF1 motif (M00121) also best identifies USF2 binding sites. Only 19

USF2 only sites were identified when studying co-localisation with CTCF, mSin3 and USF1 and 172 USF2 sites overlapped with USF1 sites. This suggests that USF1-USF2 heterodimers are more common than USF2 binding alone and is consistent with the same binding motif being identified for the two factors. Alternatively, this could also suggest that the two antibodies used in ChIP-chip for USF1 and USF2 cross-reacted to some degree. Novel motif analysis using the DME program also identified the same E-box motifs for USF1 and USF2 (data not shown). Therefore CTCF and USF1 sites were associated with known binding motifs, while USF2 binding was also associated with an E-box motif and novel GC rich motifs were associated with mSin3a interactions.

A

Name	Logo	Sn	Sp	Error	pvalue
1. MA0093		0.679	0.602	0.359	0
2. M00121		0.565	0.707	0.364	0
3. M01029		0.571	0.693	0.368	0
4. M00187		0.584	0.673	0.372	0
5. MA0058		0.54	0.702	0.379	0

B






Name	Logo	Sn	Sp	Error	pvalue
1. M00121		0.484	0.845	0.336	0
2. MA0104		0.484	0.813	0.352	0
3. M00055		0.457	0.836	0.354	0
3. MA0093		0.406	0.886	0.354	0
5. M00217		0.484	0.806	0.355	0

Table 4.8: Searching for known motifs in USF1 and USF2 binding sites. Known motif matrices from JASPAR and TRANSFAC databases were enriched in USF1 (panel A) and USF2 (panel B) binding sites. The E-box motif in the JASPAR database (MA0093) and the USF1 motif in TRANSFAC (M00121), which contains the E-box sequence at its core, were found to be the best performing known motifs for distinguish foreground sequences from background sequences in USF1 and USF2 ChIP-chip data sets respectively.

4.9. Chromatin structure at insulators

CTCF has been implicated in establishing local chromatin structure at a number of loci (Cho *et al.*, 2005; Filippova *et al.*, 2001) and has also been found at the transition regions between chromosomal domains of X inactivation and escape (Filippova *et al.*, 2005) suggesting that CTCF may influence local chromatin conformation to facilitate barrier insulator function. In addition, chromatin structure at the nucleosome level has been shown to be important for the binding of CTCF (Kanduri *et al.*, 2002). Therefore, the chromatin properties of CTCF binding sites were examined to gain further insights into these processes.

4.9.1. CTCF binding sites are located in accessible chromatin domains

While regulatory elements such as promoters and enhancers are known to be associated with DNase I hypersensitive regions in the human genome (Follows *et al.*, 2006) it is not clear if chromatin accessibility is a general feature of insulators elements in the human genome. Chromatin accessibility at CTCF sites was examined using three different but complementary ENCODE datasets generated using K562 cells – DNase I hypersensitive data (Xi *et al.*, 2007), histone H2B and H3 data (Chapter 3), and formaldehyde assisted isolation of regulatory elements (FAIRE) data (obtained courtesy of Dr Pawan Dhami and Dr. Alex Bruce, Wellcome Trust Sanger Institute). Xi and colleagues recently performed a study of DNase I hypersensitive sites in the ENCODE regions using the DNase-chip method and identified over 1200 hypersensitive sites in K562 cells (Xi *et al.*, 2007). This data was publicly available and the location of DNase I hypersensitive sites were compared with the location of CTCF sites. More than half (296 of 571) of the CTCF sites were found to overlap with one or more hypersensitive sites. Histone density and FAIRE are inversely correlated at regions of open chromatin (Dhami, submitted; Giresi *et al.*, 2007) so the results of both assays were compared to CTCF binding. The averaged H2B/H3 z-scored \log_2 values were calculated across a 20 kb window across all 571 CTCF sites. A depletion of these core nucleosome proteins was observed approximately 2 kb upstream and downstream of CTCF binding sites, with the greatest depletion observed at the centre of the CTCF binding sites (Figure 4.20). In the FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) procedure, formaldehyde

cross-linked and sonicated chromatin was phenol-chloroform extracted to identify nucleosome-depleted DNA (Giresi *et al.*, 2007). Genomic regions depleted of nucleosome proteins were enriched in the aqueous phase following phenol-chloroform extraction and the purified DNA was hybridised to a microarray in a similar fashion to a ChIP experiment. CTCF sites were associated with a peak of enrichment in this assay, further supporting the hypothesis that CTCF binds at regions of open chromatin in the human genome.

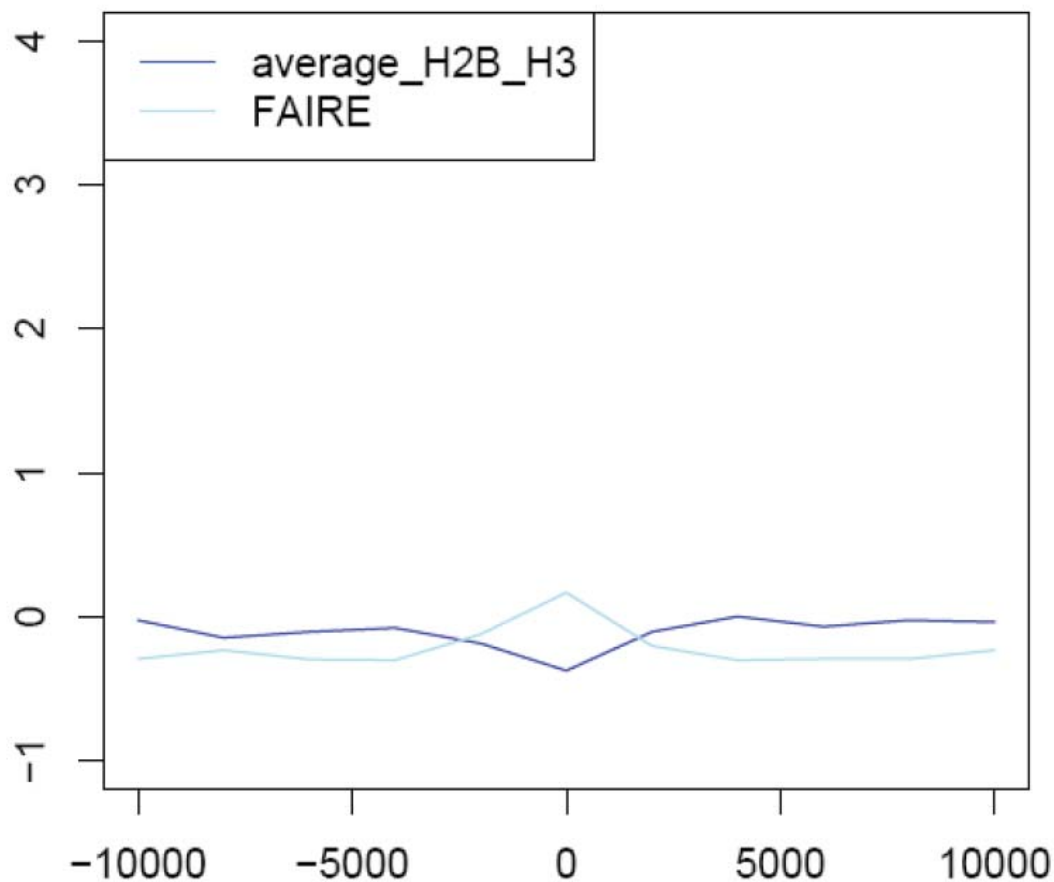


Figure 4.20: Nucleosome density and FAIRE profile at CTCF binding sites in the ENCODE regions. CTCF binding sites in K562 are associated with a large region of histone H2B and H3 depletion (approximately 4 kb) as can be seen by the depletion in the average histone H2B and H3 signal (dark blue profile). This depletion in nucleosomes correlates with a peak of FAIRE enrichment (light blue profile). The values along the x-axis represent distance in base pairs up and downstream of CTCF binding sites. The scale on the y-axis represents z-scored log₂ ratios.

Several ENCODE regions have been well-characterised in terms of the location and function of DNase I hypersensitive sites. For example, DNase I hypersensitive sites have been systematically mapped at the human α -globin locus and functionally tested in a number of studies (Higgs *et al.*, 1990; Jarman *et al.*, 1991; Sharpe *et al.*, 1993; Vyas *et al.*, 1995). CTCF binding sites at the α -globin locus in K562 cells were compared with the location of DNase I hypersensitive sites and all CTCF sites overlapped with hypersensitive sites in this region (Figure 4.21). Four of the CTCF binding sites were associated with previously identified hypersensitive sites at HBAs/2 located near the HBA2 gene, the -33 region, the -48/-46 region, and the -55 region, all of which are located within the c16orf35 gene. Three other hypersensitive sites located further upstream also contained CTCF binding sites, namely the -87 region, -130 region and the -138/-140 region. The -87 region was located within the RHBDF1 gene, while the -130 and -138/-140 regions were located approximately 30 and 40 kb from the POLR3K promoter. CTCF binding sites have been identified upstream of the chicken alpha-globin locus (Valadez-Graham *et al.*, 2004; Klochkov *et al.*, 2006) and, whilst this region is conserved and located upstream of the human HBZ gene, no binding site was identified in this region in K562 cells.

In erythroid cells, a region upstream of the β -globin genes contains a series of DNase I hypersensitive sites that comprise the locus control region (LCR), which regulate β -globin gene expression in erythroid cells. The chicken β -globin LCR contains hypersensitive site 5' HS4 which binds CTCF while another site outside of the LCR, 3' HS1, also binds CTCF and together these two sites form the boundaries of the β -globin locus in chicken erythrocytes (Bulger *et al.*, 1999). Homologous hypersensitive sites are also observed in the human β -globin locus, HS5 and 3'HS, which also bound CTCF in K562 cells (Farrell *et al.*, 2002). Unlike the chicken β -globin locus, neither element possesses barrier activity and both have been proposed to function as enhancer blocking elements. In this study, HS5 was also found to bind CTCF, but no CTCF binding was identified at the 3'HS (Figure 4.12). However, a novel CTCF binding site was identified upstream of the locus between the olfactory receptor genes OR52A4 and OR52A5, which also correlated with a hyper-sensitive site in K562 cells. The work presented in this thesis

genomic region containing actively expressed genes in K562 that are involved in erythroid development (Dhami, submitted). Therefore it was hypothesized that the +57 region may act as barrier insulator to prevent the spread of silencing histone modifications associated with the inactive CYP4Z1 and CYP4A22 genes. The formation of transcriptionally inactive chromatin domains is an important mechanism for silencing of specific genes during developmental programming (Craig, 2005) and as the H3K27me3 modification has been implicated in the formation and maintenance of silent chromatin regions via the recruitment of Polycomb group proteins (Cao *et al.*, 2002; Lee *et al.*, 2006; Boyer *et al.*, 2006), ChIP-chip experiments were performed with K562 cells to identify regions of the SCL locus associated with elevated levels of H3K27me3 (Dhami, submitted) (Figure 4.22). As can be seen in the Figure, the liver-specific CYP genes are associated with high levels of H3K27me3 in K562 cells, while SCL and other active genes contain low levels of this modification. This is consistent with the formation of an inactive chromatin domain at the CYP region in K562 cells. This modification is enriched over the entire body of both genes, while little enrichment for this modification is observed over the expressed SCL, SIL and KCY genes. There seemed to be a clear transition point at +57 from high levels of H3K27me3 associated with the CYP genes to low levels associated with actively transcribed genes. This data suggest that this CTCF binding site acts as a barrier element to prevent the spread of the silencing H3K27me3 modification associated with the inactive CYP region into the nearby actively expressed MAP17 and SCL genes.

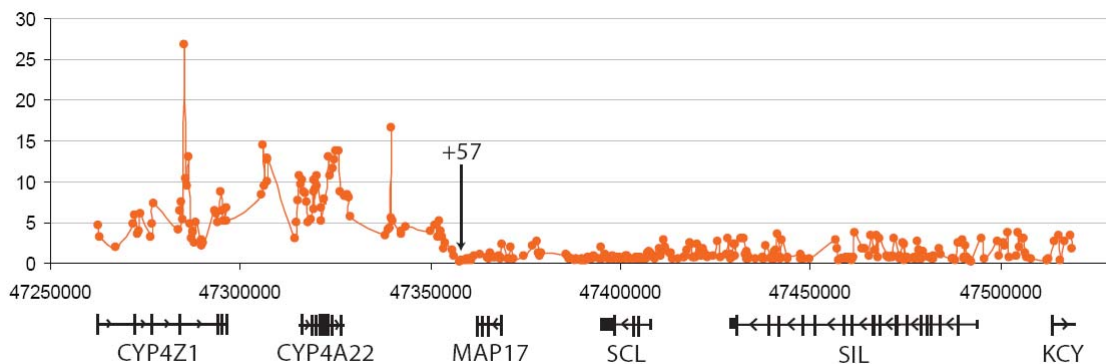


Figure 4.22: ChIP-chip profile of Histone H3K27me3 across the SCL locus in K562. The CYP region is associated with high levels of H3K27me3, while the MAP17, SCL, SIL and KCY genes are associated with low levels of H3K27me3. The transition from a high to low H3K27me3 enrichment coincides with the

binding of CTCF at the +57 region. The human chromosome 1 genomic coordinates, gene order and direction of transcription are indicated along the y-axis, while fold enrichments are indicated on the x-axis.

The distribution of CTCF binding sites in the ENCODE regions was compared with K562 histone H3K27 methylation profiles generated by others at the Sanger Institute (courtesy of Drs. Pawan Dhama and Alex Bruce) to determine the location of other putative barrier insulators. H3K27me1 and H3K27me3 have been found to be associated with neighbouring active and inactive chromatin domains respectively in K562 cells (Dhama and Bruce, unpublished data). These domains were therefore examined for binding of CTCF. 38 CTCF binding sites were identified which demarcated active chromatin regions from inactive chromatin regions (Table 4.9), two of which are presented in Figure 4.23. Panel A illustrates the dynamic change in H3K27me3 and H3K27me1 states at ENCODE region Enm003 and a CTCF binding sites were located at chromatin state transition points. This region contains the APO cluster of genes (APOA1, APOA5, APOA4, and APOC3) involved in apolipoprotein metabolism. These genes are mainly expressed in the liver and small intestine and are associated with elevated levels of H3K27me3 in K562. However the neighbouring BUD13, ZNF259 and KIAA0999 genes are associated with elevated levels of H3K27me1. CTCF binding sites are located at boundaries between active and inactive chromatin domains suggesting that CTCF in combination with other factors may prevent the spread of H3K27me3 into active chromatin regions or vice versa. Figure 4.23 panel B illustrates the same phenomenon in another ENCODE region (Enm334). The fork-head box P4 transcription factor (FOXP4) is involved in cancer progression and is expressed in K562. This gene is associated with elevated levels of H3K27me1 while the nearby MDFI gene, a MyoD inhibitor involved in cartilage formation is not expressed in K562 and is associated with high levels of H3K27me3. CTCF once again forms a barrier between an inactive chromatin region associated with high H3K27me3 levels and a neighbouring active chromatin region associated with high H3K27me1 levels.

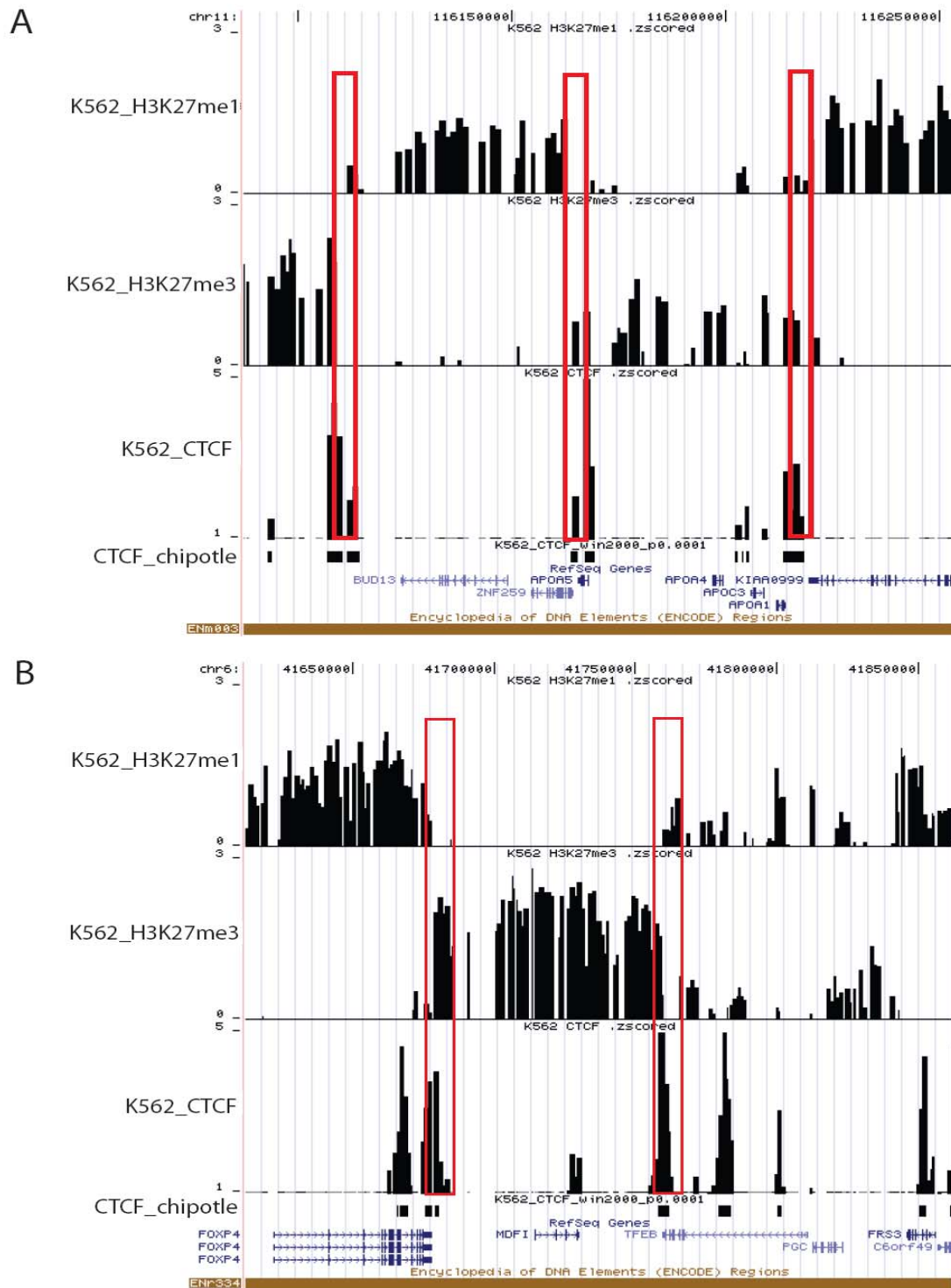


Figure 4.23: CTCF binding sites demarcate the boundary between chromatin regions associated with active and inactive histone modifications. Panel A: the top track (K562_H3K27me1) represents histone H3K27me1 enrichments in ENCODE region enm003, the middle track (K562_H3K27me3) represents

H3K27me3 enrichments and the bottom track represents CTCF enrichments and ChIPOTle defined CTCF sites are indicated below this track as black bars. CTCF sites located at the boundary between chromatin regions associated with H3K27me1 or H3K27me3 are indicated by red boxes. Panel B: ENCODE region enr334 contains CTCF binding sites that demarcate active and inactive chromatin regions. RefSeq genes in the two regions are presented below the CTCF ChIPOTle tracks, while chromosome coordinates are presented at the top of each panel. Fold enrichments for histone modifications and CTCF are presented as \log_2 values (scale indicated on the y-axis of each panel).

Twenty one of the 44 ENCODE regions contained at least one putative CTCF chromatin barrier elements while 9 ENCODE regions contained multiple barrier elements (Table 4.10). USF1 and USF2 recruit histone modifying enzymes to the chicken beta-globin barrier insulator (West *et al.*, 2004; Huang *et al.*, 2007) so the pattern of USF1 and USF2 binding was examined at the 38 putative barrier elements (Table 4.10). 6 of the 38 sites were found to co-localise with a USF1 binding site, while 7 other putative barrier insulators overlapped with both USF1 and USF2 binding sites. This suggests that USF factors may play a role in recruiting histone modifying enzymes at 13 of the 38 (34%) putative barrier elements. In total 140 of the 571 CTCF sites co-localised with USF1 and/or USF2 sites, representing 25% of CTCF interactions. As 34% of CTCF sites located at putative barriers co-localised with USF1 or USF2, this suggested that there may be an over-representation of USF binding at putative barrier elements relative to the total percentage of co-localised CTCF and USF sites. However, the vast majority (91%) of CTCF sites which co-localised with USF1 and/or USF2 do not seem to be involved in barrier function.

Region	CTCF boundary coordinates	Overlapping USF1 coordinates	Overlapping USF2 coordinates
Enm001	chr7:115760053-115761552	N/A	N/A
Enm003	chr11:116111357-116114356	N/A	N/A
	chr11:116166857-116169356	N/A	N/A
	chr11:116213357-116218356	chr11:116216857-116218356	N/A
Enm004	chr22:30690594-30692093	N/A	N/A
	chr22:30693094-30694593	chr22:30694094-30694593	N/A
	chr22:30698594-30700093	N/A	N/A
	chr22:31252844-31253343	Chr22:31253094-31258093	Chr22:31253594-31254093
Enm005	chr21:32687032-32688531	N/A	N/A
	chr21:32906532-32907531	chr21:32905998-32907497	chr21:32905998-32907497
	chr21:33492032-33494531	N/A	N/A
Enm006	ChrX:152696278-152697777	ChrX:1562694278-152696277	N/A
	ChrX:153093278-153094777	N/A	N/A
Enm007	Chr19:59406293-59407792	Chr19:59403793-59407792	Chr19:59403793-59406292
Enm009	Chr11:5101319-5102818	N/A	N/A
Enm011	Chr11:1720702-1723701	N/A	N/A
	Chr11:1750702-1751201	N/A	N/A
	Chr11:1914702-1917201	N/A	N/A
	Chr11:1978202-1979701	N/A	N/A
Enm014	Chr7:126629702-126631201	N/A	N/A
Enr121	Chr2:118309511-118312010	N/A	N/A
	Chr2:118286511-118290010	Chr2:118288511-118290010	Chr2:118288511-118290010
Enr131	Chr2:234522984-234524483	Chr2:234521984-234523483	N/A
Enr132	Chr13:112390066-112391565	N/A	N/A
Enr133	Chr21:39444495-39450994	N/A	N/A
Enr223	Chr6:74075593-74077592	N/A	N/A
Enr232	Chr9:129014732-129016231	Chr9:129014732-129015231	N/A
	Chr9:129243732-129244731	N/A	N/A
Enr233	Chr15:41668924-41671423	Chr15:41668527-41676026	Chr15:41668527-41670526
	Chr15:41768424-41770923	Chr15:41768527-41776026	Chr15:41768527-41771026
	Chr15:41981424-41984923	Chr15:41982527-41986026	N/A
Enr322	Chr14:98921983-98931982	N/A	N/A
Enr331	Chr2:220309066-220310565	N/A	N/A
Enr332	Chr11:64267062-64268561	Chr11:64267062-64268561	Chr11:64266812-64267311
Enr333	Chr20:33352934-33358433	N/A	N/A
	Chr20:33499434-33500933	N/A	N/A
	Chr20:33668434-33669933	N/A	N/A
Enr334	Chr6:41678954-41680453	N/A	N/A

Table 4.9: CTCF binding sites in K562 located at the boundary between active and inactive genomic regions defined by enriched levels of H3K27me1 and H3K27me3 respectively. The 21 ENCODE regions which contained CTCF sites at chromatin boundaries are indicated in the first column, while the second column contains the genomic coordinates of the 38 proposed CTCF boundary elements. The third

and fourth columns contain USF1 and USF2 binding coordinates which co-localise with these CTCF boundary sites. N/A: not applicable, i.e. no co-localisation of USF1/2.

4.10. Discussion

Insulator elements, which can have either enhancer-blocking activity or act as barriers between genomic regions of active and inactive chromatin, represent an important class of regulatory element for which little information is available in the literature. Therefore, in this study K562 and U937 cells were used to map the location of the insulator binding protein, CTCF, using a ChIP-chip strategy. The Sanger Institute ENCODE microarray was used to map the location of 571 and 661 CTCF sites in K562 and U937 cells respectively. The distribution of CTCF was examined and motifs associated with CTCF binding sites were analysed. In addition, CTCF binding sites were further characterised by investigating the binding distribution of mSin3a, USF1, and USF2, factors which have been implicated in CTCF transcriptional repression and insulator functions (Lutz *et al.*, 2000, West *et al.*, 2004). As a means of identifying putative barrier insulators, the chromatin structure at CTCF binding sites was also investigated

4.10.1. Widespread distribution of CTCF binding sites in the human genome

CTCF is known to be a highly versatile transcription factor that can regulate gene expression by different modes of action, such as promoter activation and repression, and constitutive- and methylation-dependent chromatin insulation (Filippova, 2008). In order to gain a greater understanding of insulators, the study presented in this thesis used a ChIP-chip method to map CTCF binding sites in 1% of the human genome and approximately 600 CTCF binding sites were identified in two different cell lines. The distribution of CTCF sites in K562 was examined and while most CTCF binding sites were located far from TSSs, CTCF binding was not randomly distributed across the genome but followed the distribution of genes. A similar correlation was also observed by Xie and colleagues (Xie *et al.*, 2007). Further analysis showed that CTCF sites classified into three categories: promoters bound by CTCF, distal enhancers/repressors bound by CTCF and “other” locations. The majority of CTCF sites were located at “other” locations which is consistent with an insulator function as these sites are not

associated with promoter or enhancer activity. The varied localisation of CTCF binding sites is consistent with previous studies which reported that CTCF can bind at intron, exon, promoter, and intergenic locations (Mukhopadhyay *et al.*, 2004, Kim *et al.*, 2007, Barks *et al.*, 2007). This study also confirmed the findings of Kim and colleagues (2007) who noted that CTCF binding sites often occurred in what they described as CTCF-paired domains (CPDs). They found that 74% of genes in the human genome are surrounded completely by CTCF binding sites. In this study approximately half of the ENCODE genes were flanked in their entirety by CTCF binding sites, while a number of clusters of related (and unrelated genes) were also observed to be flanked by CTCF in this study. This suggests that CTCF can regulate individual genes or groups of genes. CTCF sites flanking clusters of related or apparently unrelated genes may function to ensure that enhancers associated with a gene cluster cannot inadvertently activate other genes outside the cluster and similarly enhancers associated with genes outside the cluster cannot influence the regulation of the genes within a cluster. Multiple CTCF binding sites can also be located across individual genes and may be used to control the interaction of multiple lineage- or temporal-specific enhancers, or some may be located at enhancers where their function is not insulator activity but transcription factor activity.

4.10.2. Cross-study comparison of CTCF binding sites

This study has demonstrated the accuracy of the ChIP-chip method in detecting previously characterised CTCF binding sites located in the ENCODE regions. In addition to identifying known CTCF binding sites, approximately 600 novel CTCF interactions were identified in K562 and U937 cells in this study. As the ENCODE regions were chosen to contain features representative of the entire human genome sequence, this suggests that there may be as many as 50,000-60,000 CTCF binding sites in the human genome. Mukhopadhyay and colleagues had previously mapped the location of 200 CTCF binding sites in the mouse genome using a ChIP-chip approach, the majority of which display insulator functions in assays, and estimated that the total number of CTCF binding sites in the mouse genome would be around 4,000 (Mukhopadhyay *et al.*, 2004). More recently Xie and colleagues used sequence conservation analysis to predict the location of 15,000 CTCF binding sites in the human genome (Xie *et al.*, 2007). However,

this number does not take into account CTCF binding sites which are not conserved in other mammalian species. Recent genome-wide ChIP studies derived data sets have identified more CTCF binding sites in the human genome. Barski and colleagues identified approximately 20,000 CTCF binding sites in human CD4⁺ T cells using a high-throughput ChIP-sequencing method (Barski *et al.*, 2007) while Kim and colleagues used a whole genome ChIP-chip approach to map CTCF binding sites in human cells and used their experimentally derived consensus motif to predicted the location of approximately 30,000 CTCF binding sites in the human genome (Kim *et al.*, 2007). However the authors also note that 25% of their experimentally defined CTCF sites do not contain the consensus motif indicating that more than 30,000 CTCF binding sites are located in the human genome.

Analysis of CTCF binding sites in the ENCODE regions revealed that Kim *et al.* identified 225 and 232 sites in IMR90 and U937 cells respectively, while Barski and colleagues identified 353 CTCF binding sites in CD4⁺ T cells. In contrast this study identified approximately 600 CTCF binding sites in K562 and U937 cells, 500 of which were not detected by Kim and colleagues when a direct comparison of CTCF binding sites in U937 cells was performed. This study and that of Kim and colleagues used similar ChIP protocols suggesting that antibody quality or sensitivity of array platform may be responsible for the large difference in the number of CTCF sites identified between the two ChIP-chip studies. In addition, the possibility that ChIPOTle is over-calling CTCF sites cannot be excluded, although a stringent p value threshold was used to minimize this possibility. The CTCF antibody used by Kim and colleagues was a mixture of monoclonal antibodies while the antibody used in this study was a polyclonal one. Small changes in the structure or accessibility of an epitope upon cross-linking can dramatically affect the function of a monoclonal antibody. In contrast, because polyclonal antibodies recognise a number of antigenic epitopes, the effects of cross-linking are less of a problem. Therefore, the reduced number of CTCF sites identified by Kim and colleagues relative to this study may be the due to the type of antibody used in the ChIP procedure.

Different array platforms may also have affected the number of CTCF sites identified by the two studies. Kim and colleagues used Nimblegen oligonucleotide arrays and

oligonucleotide probes are not as good a hybridisation element as larger PCR product probes. Nimblegen arrays also require ChIP DNAs to be amplified prior to labeling with fluorescent dyes. Any bias during the amplification process could change the representation of the sample and some CTCF interacting sequences may not be detected following microarray hybridisation. Sanger Institute in-house constructed PCR product microarrays do not require amplification of ChIP DNAs prior to hybridisation illustrating that this platform may be more sensitive than commercially-available microarrays.

In addition, different analysis methods were used in the two studies which may have affected how many CTCF sites were identified. Kim *et al.* normalised microarray data by a Lowess (locally weighted regression) normalisation method (Berger *et al.*, 2004.), which removes intensity-dependent effects in \log_2 ratios. Probability statistics were then calculated for each probe based on a single array error model (SAEM) (Li *et al.*, 2003). Kim *et al.* (2007) also used a very stringent statistical threshold (p 0.000001) to obtain a similar number of CTCF binding sites in U937 cells as detected in IMR90 cells. In contrast microarray data obtained in this study was total intensity normalised (Quackenbush, 2002), which assumes that total hybridisation intensities summed over all array elements should be the same for each sample. A normalisation factor is calculated by summing measured intensities in both the Cy3 and Cy5 channels, which adjusts each ratio such that the mean ratio is equal to 1. In addition each data point was then normalised by dividing with the corresponding value obtained from a goat IgG ChIP-chip experiment which can help to identify authentic sites that are subtly enriched prior to normalisation. Goat IgG normalisation can also help to remove non-specific enrichments from a ChIP-chip data set. Binding sites were then detected in normalised data using the ChIPOTle program with a high stringency threshold (p 0.0001). ChIPOTle has been shown to be more accurate than SAEM in accurately detecting TF binding events using ChIP-chip data (Buck *et al.*, 2005). Therefore a combination of the very high stringency threshold chosen by Kim and colleagues coupled with the use of SAEM and no species-specific IgG normalisation may have resulted in the identification of fewer CTCF sites compared to this study.

4.10.3. CTCF co-localisation with mSin3a, USF1 and USF2 binding sites

As described in the introduction to this Chapter, CTCF interacts with a diverse number of factors which modulate the activity of this multifunctional protein. In this study, CTCF co-localisation/co-operation with other factors was examined by performing ChIP-chip experiments to identify mSin3a, USF1 and USF2 interactions that overlapped with CTCF sites to gain a greater understanding of CTCF function. CTCF binding sites which overlapped with mSin3a binding sites may be involved in transcriptional regulation while those CTCF sites which overlapped with USF1/2 binding may be barrier elements similar to the chicken β -globin locus. One third of CTCF binding sites co-localised with binding sites for one or more of these transcription factors and these overlapping sites were located at active and inactive TSSs, distal enhancer/repressor sites and “other” locations further demonstrating that that CTCF is a highly versatile factor in terms of binding location, function and potential interacting partners. No clear correlation between CTCF sites overlapping with mSin3a and gene repression was observed – however, surprisingly, CTCF and mSin3a co-localisation at promoters was associated with active gene expression (see section 4.10.4). However, the picture is indeed complex as CTCF and mSin3a binding sites also overlapped at diverse genomic locations including active and inactive TSSs and distal sites. Similarly, USF1 and USF2 binding sites overlapped at a number of CTCF sites, a small percentage of which may function as barrier insulators based on the presence of histone H3K27 methylation modifications. In addition CTCF, mSin3a, USF1 and USF2 overlapping interactions were observed in various combinations suggesting that the modulation of CTCF function is highly complex and that no clear functional categorisation of CTCF binding sites can be performed based solely on co-localisation with mSin3a or USF1 or USF2.

4.10.4. mSin3a interaction at promoters is associated with active gene expression

While mSin3a co-localisation with CTCF was limited, due to the number of sites analysed in this study, and did not allow for a clear functional categorisation of CTCF sites, an interesting observation arose from this study which suggested that the accepted view of mSin3a acting as co-repressor of gene expression may not be always correct. In

this study the majority of mSin3a binding events at transcription start sites were associated with actively expressed genes and a ROC analysis identified that the presence of mSin3a at TSSs was as accurate an indicator of active gene expression as the presence of H3K4me3 or H3 acetylation (see Chapter 3). Further evidence supporting a role for mSin3a involvement in gene activation comes from studies of the yeast homologue Sin3. De Nadal and colleagues proposed that yeast mitogen-activated protein kinases Hog1 induces gene expression by recruiting the Rpd3-Sin3 histone deacetylase complex complex to the promoters of genes regulated by osmostress (De Nadal *et al.*, 2004). More recently Sharma and colleagues described how the Rpd3-Sin3 complex is required for the activation of DNA damage inducible genes (Sharma *et al.*, 2007) and a similar phenomenon was observed by Sertil and colleagues who described that Rpd3-Sin3 was required for the transcriptional induction of anaerobic genes (Sertil *et al.*, 2007). An emerging theme in yeast studies is that histone deacetylase (HDAC) complexes containing Sin3 are required for resetting promoters in the wake of elongating RNA Polymerase II to prevent spurious transcription initiation (Lee and Shilatifard, 2007). Thus far there are no examples in the literature of mSin3a associating with active promoters in mammalian genomes but the data presented in this thesis suggests that mSin3a is predominantly associated with actively expressed genes and may recruit HDACs to ‘reset’ these promoters following the passage of RNA polymerase II.

4.10.5. Identification of transcription factor consensus binding motifs

Approximately one third of the CTCF binding sites identified in this study overlapped with the CTCF consensus binding motif identified by Kim and colleagues (2007). This suggested that other CTCF recognition motifs exist, a fact pointed out by Kim and colleagues who noted that approximately 20% of CTCF binding sites did not contain this motif but could bind CTCF when further characterised *in vitro*. In addition a number of previously characterised CTCF binding sites do not contain this motif suggesting that CTCF recognises a number of DNA sequences (Filippova, 2008). *De novo* motif analysis in this study failed to identify a novel motif suggesting that many CTCF sites may not be associated with a consensus motif.

USF1 binding sites identified by ChIP-chip experiments were enriched for an E-box motif present in the TRANSFAC database (Matys *et al.*, 2006). In addition the ChIP-chip approach was used to define an identical consensus sequence for USF2 for which no *in vitro* motif had been previously established. While USF1 and USF2 were not extensively linked to insulators, they are important regulators of genes expression (Di Duca *et al.*, 2006; Pezzolesi *et al.*, 2007) and identifying the genome-wide binding sites of these TFs is important for understanding their function. A number of novel GC rich motifs associated with mSin3a interaction were identified as part of this study. As mSin3a is not known to interact directly with DNA itself, but is recruited by other sequence-specific TFs, this suggests that the mSin3a binding partner(s) may have a preference for GC rich binding sequences.

4.10.6. CTCF and chromatin structure

CTCF has been implicated in regulating local chromatin domains (reviewed in Filippova 2008) and in this study the local chromatin structure at CTCF binding sites was assessed using nucleosome data, DNase I hypersensitive site data and FAIRE data. CTCF binding sites were depleted of histones H2B and H3 and were enriched in FAIRE assays consistent with being located in ‘open’ chromatin regions like other regulatory elements (Dhami, submitted; Giresi *et al.*, 2007). This is consistent with the observation that binding of CTCF to its target sites may be controlled by nucleosome occupancy as CTCF is unable to interact with a target site if it is occupied with a nucleosome (Kanduri *et al.*, 2002). In addition more than 50% of K562 CTCF binding sites were associated with DNase I hypersensitive sites. A recent study also showed that approximately 70% of the 225 CTCF binding sites identified in IMR90 cells in the ENCODE regions overlapped with DNase I hypersensitive sites suggesting that CTCF preferentially binds in accessible regions of the genome (Xi *et al.*, 2007). A number of well-characterised DNase I hypersensitive sites at the α - and β -globin loci were associated with CTCF binding in K562 cells. CTCF sites were located at α -globin sites which included HBAs/2 at the HBA2 promoter, -33, -48/-46 and -55 regions (all within C16orf35). Three uncharacterised hypersensitive sites located further upstream, at -87, -130 and -138/-140 and were also associated with CTCF binding. The HS5 of the β -globin LCR was

associated with CTCF in K562, but unlike a previous report (Farrell CM 2002) the 3'HS was not associated with CTCF in this study. The nearest site to the 5'HS that was associated with CTCF binding in this study was located further upstream between two olfactory receptor genes. However, the presence of CTCF sites at these two loci in a relevant cell type provides evidence that CTCF may be involved in the regulation of haemoglobin synthesis.

CTCF has also been implicated in the formation of chromatin domains that escape X-inactivation during early development (Filippova *et al.*, 2005) and a number of CTCF binding sites have been detected between active and silent chromatin domains in a recent study (Barski *et al.*, 2007). Barski and colleagues described how several large regions of chromatin containing inactive genes were associated with high levels of H3K7me3 and neighbouring chromatin regions containing actively transcribed genes were associated with H3K27me1 (Barski *et al.*, 2007). These active and inactive chromatin domains were separated by CTCF binding sites. A similar phenomenon was observed in K562 cells in this study. Nearly 40 CTCF sites were located at the boundary between regions of active and inactive chromatin associated with H3K27me1 and H3K27me3 respectively. While USF1 and USF2 function as barrier proteins by recruiting a histone methyltransferases and histone acetyltransferases to prevent the spread of heterochromatin at the chicken β -globin insulator (West *et al.*, 2004; Huang *et al.*, 2007), only one third of CTCF boundary sites co-localised with either USF1 or USF2 in a human cell type. This suggests that the chicken β -globin USF-mediated barrier insulator model may not be applicable to the majority of barrier elements in the human genome. Perhaps other proteins are responsible for recruiting chromatin modifying enzymes such as histone H3K27 demethylase enzymes to these barriers to maintain K27 methylation states. There is precedence for this in other eukaryotic genomes as lysine-specific histone demethylase 1 (LSD1), which removes methyl groups from H3K4me1, H3K4me2, H3K9me1, and H3K9me2 (Shi *et al.*, 2004; Metzger *et al.*, 2005), has been implicated in the formation of boundaries between euchromatin and heterochromatin in *S. pombe* and *Drosophila* (Lan *et al.*, 2007 b; Rudolph *et al.*, 2007). In *S. pombe* the Lsd1/2 complex is recruited to boundary elements and limits the formation of heterochromatin perhaps by demethylating H3K9. In contrast *Drosophila* LDS1 homologue SU(VAR)3-3 is required for heterochromatin

formation by demethylating H3K4me1 and H3K4me2 and does not demethylate H3K9me1 or H3K9me2. This prevents the spread of H3K4 methylation into heterochromatin regions. In this study CTCF is associated with boundaries between active and inactive chromatin regions defined by the presence of H3K27me1 and H3K27me3 respectively. The recently identified H3K27 demethylases UTX and JMJD3 (Agger *et al.*, 2007), which are capable of demethylating H3K27me3, may be recruited to these chromatin boundaries, thus preventing the spread of H3K27me3 into active chromatin domains.