Genomic interactions of the transcription factor VEZF1

Thesis submitted for the degree of Doctor of Philosophy

By

Carolyn M. Low

Institute of Cancer Sciences

College of Medical, Veterinary and Life Sciences

University of Glasgow

September 2013

Abstract

VEZF1 is a highly conserved vertebrate transcription factor. VEZF1 binding sites have been reported to function in gene promoter activation, insulator barrier activity and protection from *de novo* DNA methylation. This study aims to identify VEZF1 binding sites across the human genome and to develop a better understanding of the gene regulatory processes mediated by VEZF1.

ChIP-seq was used to map VEZF1 binding sites across the genome of human erythroid cells. A strong association of VEZF1 with regulatory elements was discovered, particularly at active promoters and enhancers. Investigation of the binding specificity of VEZF1 resulted in the identification of consensus DNA recognition sequences. VEZF1 has a preference for binding to homopolymeric dG motifs *in vitro* and *in vivo*. However, VEZF1 has a broader specificity for previously undiscovered degenerate motifs *in vivo*. VEZF1 interacts with degenerate motifs at erythroid-specific gene regulatory elements, and evidence is presented which indicates that VEZF1 co-operates with erythroid-specific transcription factors to enable binding to these elements *in vivo*. Investigation of erythroid-specific regulatory elements *in vivo* in embryonic chicken erythrocytes, revealed a correlation between VEZF1-enrichment of promoter elements, DNA methylation status, and expression of the β-globin genes. These analyses also indicate a role for VEZF1 in mediating DNA demethylation.

The findings presented in this thesis provide an insight into the regulatory functions of VEZF1 and indicate a role for VEZF1 in the activation of gene expression.

Table of Contents

List of figures

List of tables

Acknowledgements

I would firstly like to thank my supervisor Dr Adam West for his guidance and support throughout the course of my PhD and for performing some of the bioinformatic analyses presented in this thesis. I would also like to express my gratitude to my advisor, Dr Sarah Meek, who has been a huge support and helped me through some difficult stages. I am grateful to the Medical Research Council (MRC) who funded my studentship.

I would like to thank Dr David Vetrie and his lab members for their help. I would also like to thank Drs Andrew Crossan and Tony McBryan for their contribution to the bioinformatic analyses presented.

I am grateful to Dr Katherine West for her interest and input in this project and for all of the advice she has given me during my time here. I would also like to thank the past and present members of Drs Adam and Katherine West's lab groups for their help. I particularly wish to thank Dr Grainne Barkess for her guidance and encouragement and for organising our annual visits to Firbush where we all enjoy a little escapism and a few capsizes! I would like to thank Dr Jacqueline Dickson, who taught me when I first started in the lab, and Dr Ruslan Strogantsev for his expert advice on ChIP and for his contribution to the bioinformatic analysis of my data.

I am so lucky to have met and been surrounded by an amazing group of friends during my PhD – Elaine Gourlay, Gokula Mohan, Koorosh Korfi, Yan Zhou, Bhoomi Gor, Susan Rhida and Grainne, you kept me going, kept me laughing, and quite often kept me drinking in Otto too late! I especially want to thank Elaine who has been a fantastic friend to me through the highs and lows of the last few years, I couldn't have wished for a better bench and office-mate than you.

Thanks to my good friends Susie Walker, Natalie Ward and Lindsey Owens for always being there (usually with a bottle of Prosecco).

My final thanks go to my parents, Val and Alistair, for their tremendous support throughout my PhD and for encouraging and motivating me, especially when times were hard.

I would like to dedicate this thesis to my Grandparents, Carol, Eddie, Winnie and Charlie, who were always so supportive of me.

Author's Declaration

The research reported within this thesis is my own work, except where otherwise stated, and has not been submitted for any other degree.

Carolyn Low

Chapter 1

Introduction

1.1 Basic chromatin structure

Within the eukaryotic cell nucleus DNA exists in the form of chromatin, the basic unit of which is the nucleosome. The nucleosome consists of a stretch of ~147 base pairs (bp) of DNA wound around a histone octamer, which is comprised of 2 copies of each of the four core histone proteins – H2A, H2B, H3 and H4. Individual nucleosomes are linked to one another via internucleosomal stretches of linker DNA which range from ~10 – 80 bp in length. This basic packaging of DNA is known as the 10 nm 'beads on a string' configuration and is the primary chromatin structure. Interactions between individual nucleosomes drive further compaction of chromatin to form a secondary chromatin structure of ~30 nm diameter which is termed the 30 nm fiber. Packaging into the 30 nm fiber compacts DNA by approximately 50-fold. The stability of the 30 nm chromatin fiber is maintained by the interaction of histone H1 with linker DNA and adjacent nucleosomes (Felsenfeld and Groudine, 2003). Despite much effort, the intricacies of nucleosome positioning to form the 30 nm chromatin fiber are still unresolved with two models having been proposed (Figure 1.1) (Tremethick, 2007, Luger *et al*., 2012). In the Solenoid model neighbouring nucleosomes remain side by side as the 10 nm fiber bends to form a simple one-start helix. In the Zigzag model the 10 nm fiber adopts a zigzag structure such that two nucleosome stacks are formed. These stacks are joined by the linker DNA present between adjacent nucleosomes and as a result of the zigzag pattern this linker DNA is seen to cross back and forth between the two nucleosome stacks. In this model alternating nucleosomes interact with each other as they are ultimately positioned side by side.

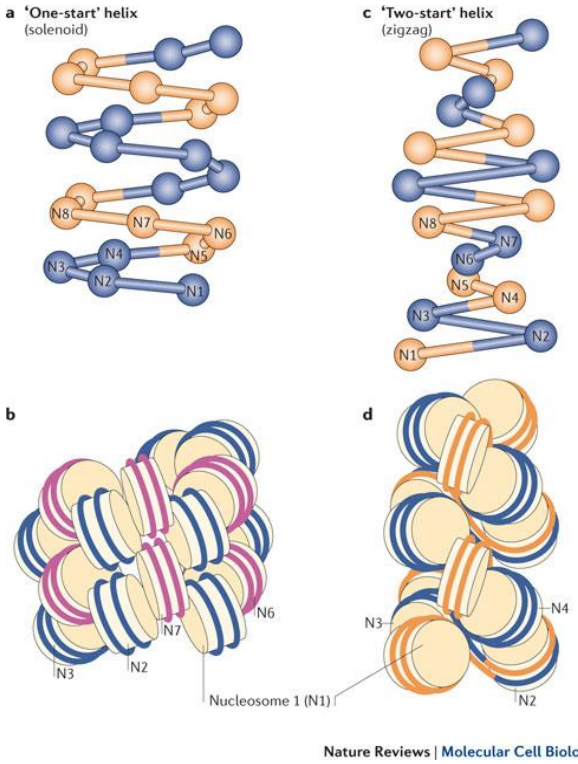


Figure 1.1 Two models for chromatin secondary structure.
The Solenoid model (A and B) consists of a one-start helix where neighbouring nucleosomes interact with each other. The Zigzag model (C and D) consists of a two-start helix where alternative nucleosomes interact as two nucleosome stacks are formed. Nucleosomes 1-8 are numbered. In (B) alternate helical gyres are coloured blue and purple. In (D) alternate nucleosome pairs are coloured blue and orange. Figure taken from (Luger *et al*., 2012).

The 30 nm chromatin fiber is further condensed to form higher order chromatin structures of 200 – 300 nm diameter. During prophase, further condensation of these 200 – 300 nm fibers is mediated by the condensin complexes I and II, Topoisomerase IIα and KIF4A proteins to form compact sister chromatids in preparation for cell division during mitosis (Vagnarelli, 2012). The sister chromatids, or daughter chromosomes that are formed during mitosis, are the highest order of condensed chromatin where the DNA is compacted 10,000 – 20,000-fold.

1.2 Structure of the eukaryotic genome

1.2.1 Coding and non-coding regions

Eukaryotic genomes contain both coding and non-coding regions. Coding regions consist of sequences which, when transcribed to form RNA or translated into protein products, serve a physical function within the cell. Protein coding sequences are found only in the exons of genes and are interspersed with non-coding intron sequences. During transcription exons and introns are linearly transcribed to form pre-messenger RNA (pre-mRNA), introns are then removed by splicing to produce mRNA, which is translated to

generate protein. Elements which regulate transcription can be found in some introns, conferring a role in the regulation of gene transcription to these non-coding sequences.

The majority of the human genome is composed of non-coding DNA sequence. In fact, it has been shown that protein-coding exons account for only ~2 % of the human genome. It has also been shown that increasing organism complexity correlates with increasing proportions of non-coding sequences in the genome (Taft *et al*., 2007). There is evidence that this increase in organism complexity is due to an increase in the number and complexity of regulatory pathways as, in addition to introns, non-coding regions are known to house numerous regulatory elements including gene promoters, enhancers and insulators.

1.2.2 Regulatory elements

Expression of each eukaryotic gene is controlled by a number of regulatory elements that are associated with that gene in the genome. Typically genes consist of a promoter from where transcription is initiated, the gene body which is comprised of protein-coding exons, non-coding introns and a transcription termination site. Gene expression can also be influenced by enhancer or silencer elements.

1.2.2.1 The core promoter

The core promoter extends ~40 bp upstream and/or downstream of the TSS and contains DNA motifs upon which direct assembly of the pre-initiation complex (PIC) occurs (Maston *et al*., 2006, Baumann *et al*., 2010) (Figure 1.2). The PIC is comprised of the general transcription factors (GTFs) TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH and RNA pol II and functions to mediate transcription initiation. The first element discovered within core promoters was the TATA box which is positioned 30 – 31 bp 5' of the TSS and provides a binding site for the TATA-binding protein (TBP), a subunit of the TFIID. TFIID is responsible for specifying the exact point of transcription initiation by promoting assembly of the PIC at a suitable region and TATA boxes can therefore be seen as directing PIC assembly to appropriate genomic sites. Initiator elements span TSSs and also function in transcription initiation by promoting assembly of the PIC. The initiator element can work cooperatively with TATA boxes or independently as it is recognised by the TAF1 and TAF2 proteins which are subunits of TFIID. A downstream promoter element (DPE) is usually found at TATA-less promoters and requires an initiator element to function. The DPE is also recognised by subunits of TFIID. TFIIB recognition elements

(BREs) have been identified upstream or downstream of TATA boxes and can influence the transcriptional activity of the core promoter. CpG islands (CGIs) have also been shown to locate to the core promoters of >70 % of human genes. CGIs commonly contain numerous SP1-like motifs, which are hypothesised to function as sites of PIC formation as the transcription factor (TF) SP1 is known to interact with TFIID via its TBP and $TAF_{II}110$ subunits (Emili *et al*., 1994, Gill *et al*., 1994).

1.2.2.2 Proximal promoter elements

Proximal promoter elements are DNA elements located immediately upstream of core promoters (Figure 1.2) and can be up to several hundred kilo bases in length. Bioinformatic analysis has shown many DNA motifs to be overrepresented at promoter proximal sites, some of which have been identified as sequence-specific TF binding sites while many have yet to be characterised (Dikstein, 2011). TF binding at promoter proximal elements can enhance transcription by promoting PIC formation or transcription initiation, elongation or re-initiation, or by recruiting proteins that contribute to the establishment of a more transcriptionally permissive chromatin structure such as histone modifying enzymes and chromatin remodelers.

1.2.2.3 Enhancer elements

Enhancer elements function to regulate transcription from their target promoters. They are typically 20 – 400 bp in length and comprise of clusters of TF binding sites which cumulatively exert affect on the activity of a target promoter (Kulaeva *et al*., 2012). Enhancers are distal regulatory elements that can be found at a distance of hundreds of kilo bases up- or downstream of their target promoters (Figure 1.2). In addition to being independent of distance from promoters, enhancer function is also independent of the orientation of enhancers relative to promoters.

3C technologies, which reveal physical interactions between elements throughout the genome, have shown that distal enhancers are in close proximity to their target promoters upon their activation (Dekker *et al*., 2002). A number of transcription factors that bind both enhancers and promoters have been shown to be required for chromosomal loop formation between these elements (Kadauke and Blobel, 2009). Some gene loci are regulated by multiple distal enhancer elements. In the case of the β-globin gene locus, multiple enhancers that constitute the locus control region (LCR) are found to

spatially cluster with one another prior to gene activation. This so-called active chromatin hub then interacts with target promoters to drive high levels of transcription (Palstra et al., 2003).

1.2.2.4 Silencer elements

Silencer elements have similar properties to enhancers but function to repress transcription from their target promoters. Silencers consist of TF binding sites to which repressor TFs bind. Similar to enhancers, silencer activity is independent of orientation and distance from target promoters and as such silencers can be located proximal to promoters, within gene introns or at distal sites. The repressor TFs which bind silencer elements can recruit co-repressors that contribute to transcriptional repression by physically blocking activator TF binding sites. Repressor TFs can also recruit histone modifying enzymes or chromatin remodelling factors that promote the formation of heterochromatin, which inhibits transcription by physically occluding the binding sites of activating TFs.

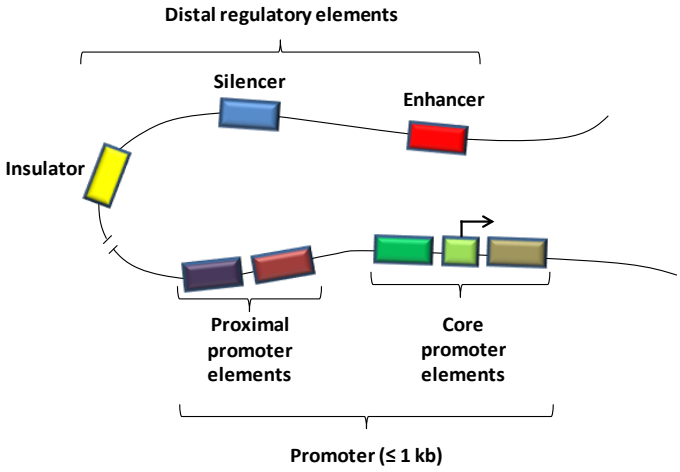


Figure 1.2 Schematic of an example gene regulatory region.
The promoter is composed of core promoter and proximal promoter elements and is typically ≤ 1 kb in length. Distal regulatory elements can include insulators, silencers and enhancers which can be found up to hundreds of kilobases away from target gene promoters. Figure adapted from (Maston *et al.*, 2006).

1.3 Chromatin and transcriptional regulation

1.3.1 Chromatin domains

In addition to compacting the genome to such an extent that ~2 metres of DNA fits into the 6 µm diameter cell nucleus (Peterson and Laniel, 2004), the formation of chromatin and regulation of its structure is essential to regulating processes such as gene expression

and DNA repair, which require DNA to be accessible for interactions with DNA binding proteins. The genome is known to be divided into transcriptionally permissive and repressive chromatin domains which differ in protein accessibility and histone modifications (section 1.3.4).

The terms heterochromatin and euchromatin are used to describe the level of chromatin compaction and transcription potential of chromatin domains. Heterochromatin is highly packaged such that DNA sequence elements are inaccessible to DNA binding proteins such as transcriptional activators. Euchromatin domains adopt a more 'open' or relaxed chromatin conformation which permits DNA binding proteins access to their cognate TF binding sites, thus euchromatin can be seen to be more transcriptionally permissive than heterochromatin.

Heterochromatin formation can be dynamic, i.e. regions of heterochromatin can be decondensed to adopt a euchromatic state and become more transcriptionally competent. However at specific genomic regions, such as gene poor or repetitive regions, heterochromatin structure is constitutive. Another well-reported feature of heterochromatin is its ability to spread throughout the genome (Grewal and Jia, 2007), this spreading can be halted by insulator elements (section 1.7).

1.3.2 Nucleosome density and positioning

Nucleosome positioning is a key regulator of gene expression, as the packaging of DNA into nucleosomes generally inhibits TF binding (Lorch *et al*., 1987, Polach and Widom, 1995, Mao *et al*., 2011). Profiling nucleosome density across the genomes of yeast and higher eukaryotes has revealed that regulatory elements such as promoters or enhancers tend to be depleted in nucleosome occupancy (Iyer, 2012, Radman-Livaja and Rando, 2010). Furthermore it is well established that in the yeast genome the -1 and +1 nucleosomes flanking promoter elements are highly positioned, and that the extent of preferred positioning of subsequent nucleosomes diminishes with increasing distance from the promoter (Struhl and Segal, 2013, Radman-Livaja and Rando, 2010). The mechanisms by which nucleosome positioning and nucleosome depleted regions (NDRs) are formed have been the subject of extensive study with DNA sequence, DNA binding proteins, chromatin remodelling enzymes and the transcription machinery being implicated, either independently or cumulatively, as affecting nucleosome positioning

(Radman-Livaja and Rando, 2010, Luger *et al*., 2012, Struhl and Segal, 2013, Segal *et al*., 2006, Bai *et al*., 2011).

In 2012, Hughes *et al* used a functional evolutionary approach to identify the determinants of nucleosome positioning using yeast as a model system (Hughes *et al*., 2012). The results of this study suggest a role for each of the afore-mentioned factors in determining nucleosome positioning. Large portions of genomic DNA from the evolutionarily distinct yeast species *K. lactis, K. waltii* and *D. hansenii* were transformed into *S. cerevisiae* and analysed for the formation and distribution of nucleosomes. Many NDRs which exist in native species were established following sequence transformation into *S. cerevisiae* host cells, indicating that intrinsic DNA sequence is involved in generating NDRs. In concordance with much of the published literature, poly(dA:dT) sequence motifs appear to act as the primary DNA sequence determinant directing the establishment of NDRs. This is believed to result from the intrinsic rigidity of poly(dA:dT) motifs which inhibits DNA bending around a histone octamer and thus inhibits nucleosome formation.

In the heterologous DNA segments from *D. hansenii*, NDRs were less conserved following transformation into *S. cerevisiae* because there are fewer poly(dA.dT) motifs in the promoters of this species. Some *D. hansenii* promoter sequences that lost NDRs in host cells contained binding sites for TFs which function as general regulatory factors in *D. hansenii* but not in *S. cerevisiae*, supporting a role for TF binding in establishing nucleosome positioning. In higher eukaryotes CTCF binding appears to function in the regulation of nucleosome positioning (Fu *et al*., 2008).

Hughes *et al* also observed that the distance between adjacent nucleosomes in heterologous DNA-associated nucleosome arrays reduced from ~178 bp in native *K. lactis* cells to ~165 bp in *S. cerevisiae* host cells (Hughes *et al*., 2012). This finding shows that DNA sequence does not dictate the majority of nucleosome positioning and indicates a role for species specific chromatin remodelers. It was also discovered that the +1 nucleosome at many TSSs differed between native and host cells, this shift was shown to correlate with shifted TSS positions indicating a mechanistic link between transcriptional initiation and positioning of the +1 nucleosome. NDRs flanked by highly positioned -1 and +1 nucleosomes were also found to form in coding regions of *D. hansenii* genes when

transferred to *S. cerevisiae*. The recruitment of transcription machinery and levels of RNA expression were found to correlate with the establishment of these novel NDRs further supporting a role for transcription machinery in directing nucleosome positioning.

1.3.3 Chromatin remodelling complexes

Chromatin remodelers are multi-protein complexes that function as DNA translocases, using the energy generated from ATP hydrolysis to reposition nucleosomes. Chromatin remodelers are divided into four families – SWI/SNF, ISW1, CHD and IN080 – which are defined by the various domains present within their ATPase subunits (Figure 1.3) (Clapier and Cairns, 2009). The catalytic ATPase domain of all remodelling enzymes is conserved and consists of one DExx and one HELICc domain. Chromatin remodelling complexes affect nucleosome repositioning by mediating sliding, destabilisation, ejection or restructuring of nucleosomes.

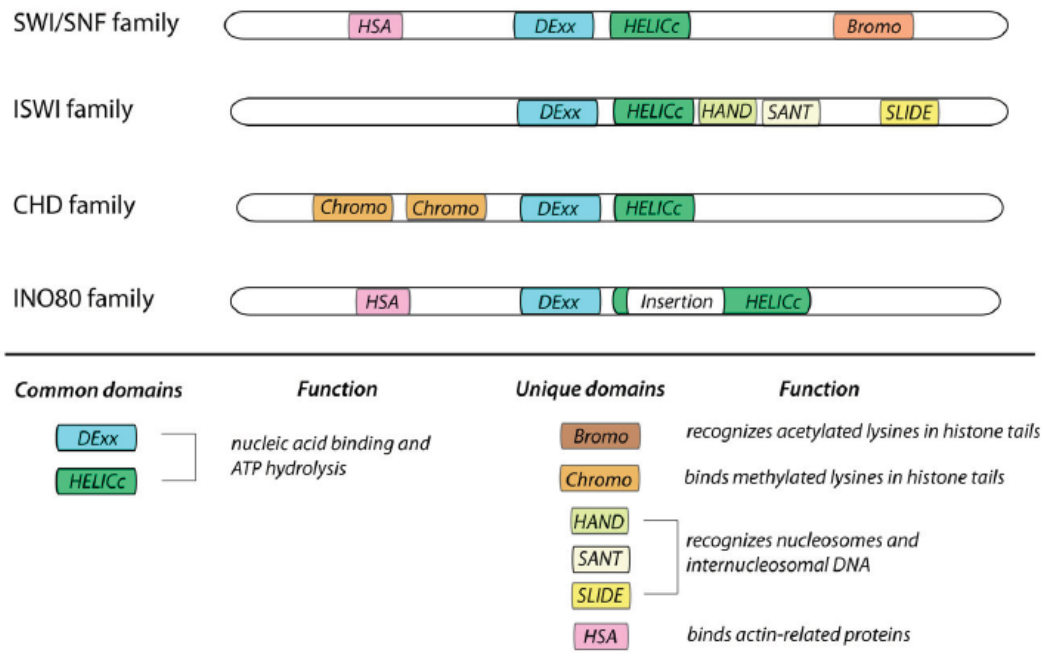


Figure 1.3 Chromatin remodeler families are defined by their catalytic domains. Figure taken from (Manelyte and Langst, 2013).

Interactions between remodelling complexes and chromatin are mediated by nucleosome recognition domains within the subunits of chromatin remodelling complexes. These include bromodomains, which recognise acetylated lysine residues on histone tails, chromodomains (CHD) and Plant Homeodomains (PHD), which bind methyl lysine, and

Hand-Sant-Slide domains that bind DNA and unmodified histone tails (Clapier and Cairns, 2009, Erdel *et al*., 2011) (Figure 1.13).

Chromatin remodelling is facilitated by the binding of remodelling complexes to nucleosomes. Proposed models suggest that the ATPase/translocase domain binds within the nucleosome and remains anchored here. Meanwhile a DNA binding domain binds linker DNA upstream of the nucleosome and draws this linker DNA into the nucleosome displacing DNA already associated with the histone octamer and creating a DNA loop (Figure 1.4 state 1-2). The ATPase/translocase domain then pumps DNA towards the nucleosome dyad, propagating the DNA loop around the nucleosome surface (Figure 1.4 state 2-4) (Clapier and Cairns, 2009). DNA-histone interactions are thus disrupted at the leading end of the DNA loop and re-established at the lagging end. This mechanism of feeding DNA around a nucleosome can be seen to facilitate nucleosome repositioning by sliding. Disruption of histone-DNA interactions may also facilitate the ejection of nucleosomes or the exchange of histone dimers, for example the exchange of histone H2A-H2B dimers for H2A.Z-H2B dimers by the human chromatin remodeler complexes SRCAP and p400 (Marques *et al*., 2010).

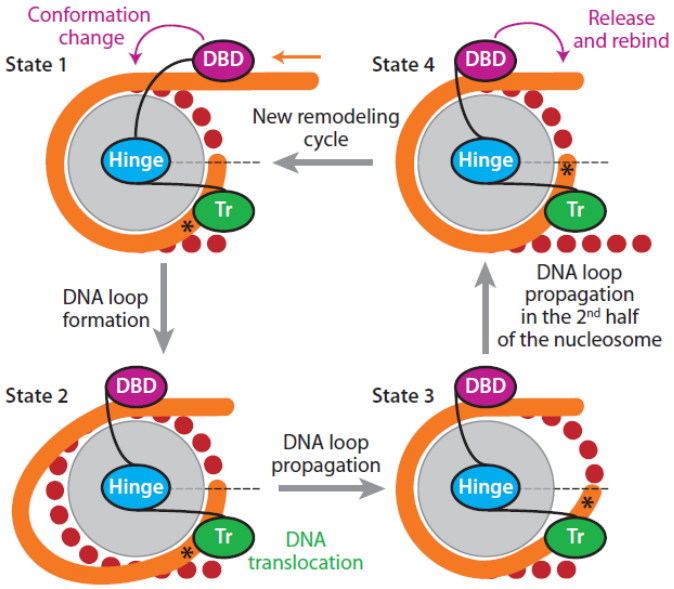


Figure 1.4 Model of a chromatin remodelling event.
The nucleosome is composed of a histone octamer, which is represented by a grey circle, and a 147 bp stretch of DNA. The first gyre of DNA is depicted by an orange line, after crossing the nucleosome dyad the second DNA gyre is represented by a red dotted line. The asterix serves as a point of reference to aid visualisation of DNA translocation. The remodelling complex interacts with the nucleosome via an ATPase/translocase (Tr) domain, which binds within the nucleosome where it remains anchored, and a DBD which binds linker DNA upstream of the nucleosome. The combined action of the DBD and the Tr creates a DNA loop which is pumped around the histone core resulting in the repositioning of nucleosomes. Figure taken from (Clapier and Cairns, 2009).

1.3.4 Histone modifications

Histone proteins have N-terminal tails which protrude from the nucleosome core. These protein tails are highly basic and can make contacts with neighbouring nucleosomes, they also serve as substrates for post-translational modification. Post-translational modification of histone tails functions to regulate chromatin structure by three mechanisms; firstly the physical presence of histone modifications can alter the contacts between neighbouring nucleosomes, secondly the post-translational modifications (PTMs) can recruit ATP-dependent chromatin remodelling enzymes and other binding proteins which can disrupt or reposition nucleosomes (section 1.3.3) and thirdly PTMs can inhibit specific TFs from interacting with DNA. At least 13 types of histone PTM have been reported to date in the published literature including acetylation, methylation, phosphorylation and ubiquitination (Tan *et al.*, 2011). Following their deposition, PTMs can induce chromatin to adopt a more condensed heterochromatin conformation or a more open euchromatin conformation depending on the modification and which residues within the histone tail are affected (Bannister and Kouzarides, 2011). In regulating

chromatin structure, PTMs have been shown to regulate the accessibility of DNA to TFs with the outcome of exerting either positive or negative effects on transcription. In the following sections I will discuss the histone PTMs that are directly relevant to this thesis.

1.3.4.1 Histone Acetylation

Lysine acetylation was the first histone PTM to be reported in 1964 and has since been found to occur in the tails of all four core histones (Tan *et al*., 2011). Lysine acetylation is a dynamic modification deposited by histone acetyl transferases (HATs). HATs transfer the acetyl group from acetyl-coenzyme A to the ε-amino group of lysine side chains. Lysine acetylation is removed by histone deacetylases (HDACs). Acetylation of lysine neutralises its positive charge, likely weakening interactions between DNA and histone proteins. The resulting reduction in nucleosome stability can lead to the adoption of a more open chromatin structure, and in accordance with this, lysine acetylation is known to positively correlate with transcriptional activity (Bannister and Kouzarides, 2011, Zentner and Henikoff, 2013).

Histone acetylation is targeted by the recruitment of HATs to specific genomic regions by transcriptional co-activators. For example the TFs BRCA1 and c-Myc interact with the TRRAP protein which is a subunit of macromolecular HAT-containing complexes. This TRRAP interaction mediates the association of BRCA1 with the GCN5 HAT, and recruitment of BRCA1/TRRAP/GCN5 complexes to BRCA1 target sites has been shown to correlate with acetylation of histone H3 (Oishi *et al*., 2006). In similar systems, association of TRRAP with C-Myc has been reported to recruit both the GCN5 and TIP60 HATs to C-Myc binding sites (McMahon *et al*., 2000, Frank *et al*., 2003).

Following their deposition, acetyl lysine marks recruit proteins containing bromodomains via an interaction between the acetyl group and the bromodomain. Bromodomains are generally found in proteins that act as regulators of chromatin structure and gene expression such as the BRG1 protein, one of the protein subunits of the ATP-dependent SWI/SNF chromatin remodelling complex. The BRG1 bromodomain is required for full binding of SWI/SNF to H3- and H4-acetylated nucleosomes (Hassan *et al*., 2006, Awad and Hassan, 2008). Subsequent chromatin remodelling by means of nucleosome sliding and eviction results in genomic DNA regions that were previously inaccessible due to nucleosome positioning becoming accessible to DNA binding factors (Wilson and Roberts, 2011).

1.3.4.2 Histone methylation

Methylation of specific lysine and arginine residues has been reported to occur in the tails of all four core histones (Tan *et al*., 2011). Lysine residues can be mono-, di-, or tri-methylated, while arginine residues can be mono-methylated and symmetrically or asymmetrically di-methylated (Bannister and Kouzarides, 2011). Histone methyl transferase enzymes catalyse the transfer of a methyl group from S-adenosylmethionine (SAM) to target residues within histones. Lysine methylation is catalysed by either SET-domain-containing proteins or DOT1-like proteins, which transfer a methyl group to the ε-amino group of lysine. Arginine methylation, meanwhile, is catalysed by members of the PRMT protein family, which transfer a methyl group to the ω-guanidino group of arginine. Histone methyl transferases tend to show high specificity towards the lysine or arginine residues that they target and the level of methylation they impart, for example the histone lysine methyl transferase (HKMT) complex SET7/9 specifically targets H3K4 for monomethylation (Xiao *et al*., 2003).

Methylation of amino acid residues within histone tails does not alter histone charge as acetylation does, however this PTM exerts effects on chromatin structure by recruiting various chromodomain-containing proteins such as chromatin remodelling complexes. Histone methylation can therefore result in either positive or negative effects on transcription depending on which lysine or arginine residues are affected and the extent of methylation (e.g. mono-, di-, or tri-methylation).

It has been proposed that DNA sequence, long non-coding RNAs (lncRNA) and DNA methylation can direct histone methylation throughout the genome (reviewed in (Greer and Shi, 2012). In Drosophila, the H3K4 methyltransferase TRX is known to be recruited to trithorax group response elements (TREs), while the Polycomb repressive complex 2 (PRC2), which catalyses trimethylation of H3K27, is recruited to Polycomb group response elements (PREs). Recruitment of methyltransferases to these specific genomic sites in Drosophila is likely mediated by sequence-specific DNA binding proteins such as the PHO protein (Brown *et al*., 1998, Fritsch *et al*., 1999). PHO is a polycomb group protein (PcG) that binds specific DNA motifs frequently found at PREs and functions in the recruitment of PcG complexes. The mechanism of PcG complex targeting in mammals is not well understood, however the protein YY1 has been identified as the mammalian functional homologue of drosophila PHO (Atchison *et al*., 2003). siRNA-mediated knock down of YY1

results in reduced recruitment of the EZH2 PcG methyltransferase and a subsequent loss of H3K27me3 at YY1 binding sites (Caretti *et al*., 2004). Some aspects of PcG complex recruitment therefore appear to be conserved between drosophila and mammals. To date only a small number of putative PREs have been identified in mammals (Sing *et al*., 2009, Woo *et al*., 2010, Bengani *et al*., 2013).

Reported interactions between lncRNAs, histone methyltransferases and demethylases have led to the hypothesis that lncRNAs may function in directing these enzymes to their genomic target sites. For example, the human lncRNA HOTAIR interacts with both the HKMT complex PRC2 and the lysine demethylase LSD1, directing these enzymes to distinct regions of the genome (Tsai *et al*., 2010). DNA methylation has also been implicated in directing histone methylation, as the methyl-CpG binding domain (MBD) of the *Arabidopsis thaliana* H3K9 histone methyltransferase SUVH4 interacts with methylated DNA, and mutation of the SUVH4 MBD results in diminished levels of H3K9me2 (Johnson *et al*., 2007).

Although for some time histone methylation was believed to be irreversible, it is now well established that this is in fact a dynamic modification but with a slower turnover than other histone PTMs (Shi *et al*., 2004, Bannister and Kouzarides, 2011). Histone demethylation is catalysed by histone demethylase enzymes and distinct demethylases act on methylated lysine or arginine substrates. Two families of lysine demethylases have been identified; amine oxidase and jumonji C (JmjC)-domain-containing, iron-dependent deoxygenases (Greer and Shi, 2012). Similarly to HKMTs, lysine demethylases show a high level of substrate specificity. The enzymes that mediate arginine demethylation have largely evaded characterisation to date, however in 2007 JMJD6 was discovered to have arginine demethylase activity (Chang *et al*., 2007).

1.3.5 Histone variants

In addition to the five core histone proteins are histone variants. Variants of the core histones H1, H2A, H2B and H3 have been discovered and characterised (reviewed in (Sarma and Reinberg, 2005, Kamakaka and Biggins, 2005, Talbert and Henikoff, 2010). The incorporation of non-canonical histones into nucleosomes has been shown to alter nucleosome properties by inducing changes in histone-DNA interactions. It is known that certain histone variants localise to specific genomic regions, indicating that they may

perform specific functions at these elements. In support of this hypothesis, histone variants have been found to function in a number of cellular processes including transcription initiation and repression, DNA repair and X-chromosome inactivation. The most studied histone variant is H2A.Z, which is deposited into nucleosomes by the exchange of H2A-H2B dimers for H2A.Z-H2B dimers. This histone dimer exchange is catalysed by both the SRCAP (Wong *et al*., 2007, Yang *et al*., 2012) and p400 (Gévry *et al*., 2007, Lee *et al*., 2012) ATP-dependent chromatin remodelling complexes in human cells. However it is unclear how H2A.Z deposition throughout the genome is targeted (Marques *et al*., 2010). H2A.Z is a key regulatory protein involved in mediating both positive and negative effects on gene expression (Marques *et al*., 2010). A positive role for H2A.Z in gene regulation is supported by the finding that this variant is frequently incorporated into the nucleosomes that flank nucleosome depleted regions at transcription start sites (TSS) where it functions in recruiting RNA pol II (Adam *et al*., 2001, Hardy *et al*., 2009, Gévry *et al*., 2009, Barski *et al*., 2007). A mutually antagonistic relationship has also been reported between H2A.Z and DNA methylation (an epigenetic modification associated with transcriptional repression, see section 1.4) in the genomes of numerous organisms including *Arabidopsis thaliana*, mouse and human (Zilberman *et al*., 2008, Zemach *et al*., 2010, Edwards *et al*., 2010, Conerly *et al*., 2010, Yang *et al*., 2012). However H2A.Z has also been reported as accumulating throughout transcriptionally silent gene bodies located within heterochromatic regions, suggesting a potential role for H2A.Z in the establishment or maintenance of the heterochromatic state (Hardy *et al*., 2009). Contradictory findings regarding the stability of H2A.Z-containing nucleosomes have been reported in the published literature. Some groups have reported H2A.Z incorporation to increase nucleosome stability (Park *et al*., 2004, Thambirajah *et al*., 2006) while others report H2A.Z-containing nucleosomes to be more unstable than those containing major H2A (Suto *et al*., 2000, Abbott *et al*., 2001, Zhang *et al*., 2005). However it appears that these disparities may in part reflect the histone partners of H2A.Z within the nucleosome. Jin and Felsenfeld have shown nucleosomes containing the histone variant H2A.Z only, to be at least as stable as nucleosomes composed solely of major histone proteins (Jin and Felsenfeld, 2007). Meanwhile nucleosomes containing both the histone variants H2A.Z and H3.3 were found to be more laible than H2A/H3 or H2A.Z/H3 nucleosomes. H2A.Z/H3.3 nucleosomes are found at the promoters, enhancers and coding regions of highly expressed genes (Jin and Felsenfeld, 2007) where the instability of these

nucleosomes presumably allows them to be easily displaced by the transcriptional machinery.

1.3.6 Histone cross-talk

Diverse combinations of histone modifications can work in concert to achieve fine-tuning of chromatin structure by both directly altering chromatin structure and regulating the recruitment of specific chromatin remodelling complexes. The deposition of some PTMs is dependent upon the presence of others, for example ubiquitination of H2B has been shown to be required for H3K4 and H3K79 methylation by specific HKMTs in yeast (Lee et al., 2007a). H2B ubiquitination is also required for the Setd1-mediated di- and tri-methylation of H3K4 in humans (Kim *et al*., 2009). Other histone modifications can be mutually exclusive such as acetylation and methylation of the same lysine residue. The presence of one PTM may inhibit protein binding to a neighbouring modification, for example the binding of Chp1 to H3K9me2/3 in yeast is inhibited by H3K4ac (Xhemalce and Kouzarides, 2010). Bivalent chromatin domains have also been reported such as those enriched by both the active H3K4me3 modification and the repressive H3K27me3 mark. Such bivalent domains have been found to be largely transcriptionally inactive but are believed to represent chromatin domains which are poised for active transcription (Greer and Shi, 2012).

1.3.7 Histone modifications at regulatory elements

NDRs found at regulatory elements such as promoters and enhancers are flanked by highly positioned nucleosomes, which contain specific histone modifications. The formation of NDRs allows TF binding to specific regulatory elements, contributing to the formation of the PIC and initiation of transcription. H3K4me3 is a characteristic mark of active promoter elements. One mechanism by which H3K4me3 can be targeted to active promoters is via initiating RNA pol II. Initiating RNA pol II is post-translationally phosphorylated at serine 5, the HKMT Set1 is recruited to sites of transcriptional initiation via interaction with this modification. Set1 tri-methylates H3K4 leading to H3K4me3 enrichment being a characteristic mark of active promoters (Ng *et al*., 2003, Hampsey and Reinberg, 2003). H3K4me2 and H3K4me1 marks are also deposited by Set1-containing complexes, H3K4me2 locates to promoter and enhancer elements while H3K4me1 preferentially associates with enhancers. The H3K27ac PTM also associates with active regulatory regions, this modification is found at promoters and has been shown to

distinguish active enhancers from inactive or poised enhancers (Creyghton *et al*., 2010). The histone H3K36me3 mark maps to transcribed regions of the genome. H3K36me3 is deposited by the Set2 HKMT which is recruited to transcribed regions by the elongating RNA pol II isoform which is phosphorylated at serine 2 (Krogan *et al*., 2003, Hampsey and Reinberg, 2003). In addition to active histone modifications, the histone variant H2A.Z is also known to be enriched at the highly positioned nucleosomes flanking NDRs at active regulatory elements (Talbert and Henikoff, 2010).

Chromatin signatures associated with repressed genomic regions include H3K9me3, which is deposited by the SUV39H1 HKMT in humans, and the H3K27me3 mark, which is deposited by the PRC2 complex. As mentioned previously, bivalent chromatin domains have been identified that are enriched by both the active PTM H3K4me3 and the repressive modification H3K27me3. Such bivalent domains are generally transcriptionally inactive but are believed to mark promoter elements that are poised for active transcription (Greer and Shi, 2012).

1.3.8 General transcription factors

In eukaryotic systems, RNA polymerase II requires the presence of transcription factor proteins to initiate transcription of mRNA from a DNA template. The six general transcription factors - TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH - are components of the basic transcription machinery and function in all RNA pol II-mediated transcription events by interacting with RNA pol II to form the PIC. TFIID binds specific DNA sequence elements (described in section 1.2.2.1) at promoters via its TBP or TAF subunits, TFIIA and TFIIB are then recruited via interactions with the TBP subunit. RNA pol II and TFIIF are then recruited via their interactions with the TFIID-TFIIB complex. TFIIE subsequently binds the complex and recruits TFIIH (Cooper, 2000). TFIIH contains ATPases and helicases that unwind double stranded DNA (Schaeffer *et al*., 1993, Schaeffer *et al*., 1994). TFIIH also contains a protein kinase that phosphorylates the C-terminal domain of the large RNA pol II subunit, this phosphorylation event releases RNA pol II from its association with the PIC allowing active transcription to proceed (Akoulitchev *et al*., 1995). This is the minimal system required for *in vitro* transcription however additional transcription factors are frequently required for the recruitment of the PIC at individual regulatory elements. There is also evidence that RNA pol II can form complexes with the

general transcription factors TFIIB, TFIIE, TFIIF and TFIIH prior to the formation of the PIC on DNA (Greenblatt, 1997, Davis *et al*., 2002). These complexes are termed RNA pol II holoenzymes and are recruited to promoter elements via direct interactions with TFIID to facilitate transcription.

1.3.9 Transcription factors and co-operativity

Transcription factor binding sites are known to cluster at regulatory genomic elements. Combinatorial binding of transcription factors to clustered binding sites enables the precise regulation of transcriptional activity (Spitz and Furlong, 2012). Co-operativity between transcription factors can increase the DNA binding affinity of these proteins (Spitz and Furlong, 2012). Direct co-operativity between transcription factors may result from direct protein-protein interactions between factors with neighbouring binding sites, these interactions can facilitate or stabilise interactions between transcription factors and their binding sites. Indirect transcription factor co-operativity may result when binding of a transcription factor to its recognition site induces local DNA bending which in turn may assist binding of another transcription factor to a neighbouring binding site. Another form of indirect transcription factor co-operativity, termed collaborative competition, has been observed when two or more proteins that compete for a single binding site are expressed simultaneously. Co-expression can increase transcription factor occupancy as multiple proteins compete more efficiently than a single protein with nucleosomes for DNA binding, thus inhibiting nucleosome repositioning over the binding site and increasing net transcription factor binding (Polach and Widom, 1996).

1.4 DNA methylation

1.4.1 Cytosine methylation

DNA methylation in vertebrates is the process by which a methyl group is covalently linked to the 5^{th} carbon of a deoxycytosine residue. DNA methylation is largely restricted to cytosine residues that exist within CpG dinucleotides and is catalysed by DNA methyl transferase enzymes (DNMTs) (Robertson and Wolffe, 2000, Robertson, 2005). This epigenetic modification is widely reported as being associated with transcriptional inhibition which can be a result of; i) the methyl group impeding transcription factor binding or ii) methyl-binding domain proteins (MBDs) associating with methylated CpG sites and recruiting transcriptional co-repressors and/or chromatin remodelers.

DNA methylation is of great importance in healthy tissues as it is involved in the regulation of a number of critical cellular processes including differentiation, X-chromosome inactivation and silencing of retrotransposons (Robertson and Wolffe, 2000, Strathdee *et al*., 2004). Aberrant cytosine methylation has also been linked to several human diseases such as Rett, fragile X and ICF syndromes (Robertson and Wolffe, 2000, Robertson, 2005). Compelling correlations have been reported between aberrant DNA methylation and cancer, where hypermethylation of tumor suppressor gene promoters is a characteristic epigenetic change observed in malignancies which can lead to the loss of tumor suppressor activity (Esteller, 2007). Loss of transcriptional activity has been widely reported as correlating with an accumulation of DNA methylation, highlighting the importance of maintaining correct DNA methylation patterns (Deaton and Bird, 2011). Aberrant DNA methylation in disease establishment and progression is of major interest as, unlike genetic alterations, DNA methylation has the potential be effectively reversed due to its dynamic nature. Indeed it has already been shown that treatment with 5-azacytidine inhibits DNA methylation and, over time, leads to a pronounced demethylation which correlates with re-expression of associated genes (Christman, 2002).

1.4.2 CpG-containing DNA elements

Approximately 70 – 80 % of CpG dinucleotides within the human genome are methylated. However regions with elevated CpG frequencies, termed CpG islands (CGIs), have been shown to remain largely unmethylated (Deaton and Bird, 2011). CGIs have been classically defined as being 0.5 – 5 kb in length, containing >50% GC content, and exhibiting an observed-to-expected CpG ratio >0.6 (Gardiner-Garden and Frommer, 1987). This ratio is elevated in comparison to non-CGI regions of the genome due to the propensity of methylated cytosine to undergo deamination, resulting in its conversion to thymine (Robertson and Wolffe, 2000). CGIs have a tendency to be located at regulatory regions of the genome, such as promoters, and it is hypermethylation of these elements that strongly correlates with promoter silencing and gene repression (Esteller, 2007). As >70% of annotated gene promoters contain a CGI (Saxonov *et al*., 2006), maintenance of appropriate CGI methylation is of great importance.

1.4.3 Directing DNA methylation

CpG methylation is established and maintained by DNMT enzymes. *De novo* DNA methylation is mediated by DNMT3A and DNMT3B, while the maintenance

methyltransferase DNMT1 acts upon hemi-methylated DNA to preserve established methylation patterns following DNA replication. DNMTs show little or no DNA sequence specificity and appear to rely on protein-protein interactions to target them to DNA elements. For example the DNMT3L protein, which lacks a catalytic domain, can interact with the amino terminus of histone H3 and recruit the *de novo* methyltransferases DNMT3A and DNMT3B thereby directing *de novo* DNA methylation (Robertson, 2002). DNMTs may also be recruited to specific genomic regions by histone methyltransferases such as EZH2. EZH2 is a polycomb group protein that functions as a histone methyltransferase mediating deposition of the repressive H3K27me modification. EZH2 can interact with DNMTs 1, 3A and 3B recruiting them to EZH2 target sites where DNA methylation marks can be deposited (Viré *et al*., 2006). DNMTs have also been also been shown to be recruited to specific genomic regions via protein-protein interactions with HDACs. In addition to these targeting mechanisms, interactions between DNMTs and TFs have been reported to direct DNA methylation. For example DNMT3A interacts with the MIZ1-MYC repressive protein complex and is recruited to MIZ1 target sites where *de novo* DNA methylation is established accompanied by transcriptional repression (Brenner *et al*., 2005).

1.4.4 Protection from DNA methylation

Several mechanisms have been described in the published literature as mediating protection from DNA methylation. The first protein-DNA interaction shown to protect DNA from *de novo* methylation was between the *E. coli lac* repressor (LacI) and *lac* operator (*lacO*) (Han *et al*., 2001). These studies showed that when *lacO* sites were introduced into mammalian cells, as episomes or via stable integration into genomic DNA, they were protected from DNA methylation via binding of LacI. IPTG treatment of the cell culture, which inhibits LacI binding to *lacO* sites, abolished this protective action demonstrating its function in prevention of *de novo* DNA methylation. A similar experimental design has also shown LacI binding to *lacO* sites to direct demethylation of pre-methylated *lacO* sites (Lin *et al*., 2000, Lin and Hsieh, 2001). Subsequent to this finding, evidence in the published literature has both suggested and demonstrated roles for numerous proteins in protecting the unmethylated state of DNA elements via interactions with associated binding sites. For example, the interaction of CTCF with specific binding sites has been shown to protect the H19 differentially methylated region (DMR) and the retinoblastoma (Rb) promoter from DNA methylation (Lewis and Murrell,

2004, Schoenherr *et al*., 2003, Pant *et al*., 2003, Dávalos-Salas *et al*., 2011, Fedoriw *et al*., 2004). Further to this, CTCF sites have been shown to mediate demethylation of a pre-methylated H19 DMR following its integration into mouse ES cells (Rand *et al*., 2004).

The unmethylated CpG binding protein Cfp1 has also been shown to function in the protection of DNA from *de novo* methylation. Cfp1 interacts with the HKMT Setd1, recruiting it to unmethylated regions of the genome whereupon Setd1 deposits the active H3K4me3 PTM (Thomson *et al*., 2010). As described previously, DNMT3L binding to the amino terminus of histone H3 can direct DNA methylation by recruiting the *de novo* DNMTs. Tri-methylation of histone H3K4 inhibits DNMT3L binding (Ooi *et al*., 2007), as a result of this DNMT3A and DNMT3B are not recruited and DNA is therefore protected from *de novo* methylation.

The histone variant H2A.Z has also been implicated as having a role in protecting DNA from methylation as H2A.Z occupancy of genomic elements has been shown to inversely correlate with DNA methylation in the *Arabidopsis thaliana*, puffer fish and human genomes (Zilberman *et al*., 2008, Zemach *et al*., 2010, Conerly *et al*., 2010). In *Arabidopsis* DNA methylation has been shown to exclude H2A.Z from associated nucleosomes, but H2A.Z nucleosomal incorporation has also been shown to prevent associated DNA from becoming methylated. Thus the histone variant H2A.Z appears to mediate another mechanism of protection from DNA methylation.

Computational studies have shown a number of DNA motifs to be over-represented within unmethylated CGIs and in their boundaries indicating a potential functional role in protection from DNA methylation. These motifs likely function as binding sites for specific TFs or other proteins whose binding may prevent the establishment of DNA methylation (Fan *et al*., 2007). To date a small number of these sites have been validated as functioning to protect DNA from the establishment of *de novo* methylation. The mouse and hamster *aprt* gene promoter-associated CGI contains three DNA motifs which match the consensus binding site of the TF SP1. Deletion and mutation of these sites have been shown to result in hypermethylation of the CGI (Macleod *et al*., 1994, Brandeis *et al*., 1994). Furthermore, the presence of these intact SP1-like binding motifs directs active demethylation of a pre-methylated *aprt* CGI following integration into the genomes of mouse ES cells, with this activity being lost in the absence or mutation of these protein

binding sites (Brandeis *et al*., 1994). These findings demonstrate that SP1-like motifs in the *aprt* promoter function in maintaining the unmethylated state of the associated CGI. Subsequently Marin *et al* (Marin *et al*., 1997) knocked out the Sp1 gene in mouse embryo's however, no effect on the methylation status of the mouse *aprt* CGI was observed, i.e. the CGI remained unmethylated. It can therefore be concluded that binding of Sp1 to the crucial Sp1-like binding sites does not maintain the unmethylated status of this CGI. In a later publication by Dickson *et al*, VEZF1 was found to bind one of the SP1-like sites in the *aprt* CGI (Dickson *et al*., 2010). All three SP1-like elements were then replaced with VEZF1-specific binding motifs and the modified *aprt* CGI was found to remain protected from DNA methylation and to direct demethylation when pre-methylated upon integration into the genome of mouse ES cells (Dickson *et al*., 2010). In other published literature, OCT and SOX protein binding sites have been shown to be required to prevent *de novo* DNA methylation (Hori *et al*., 2002, Kaufman *et al*., 2009).

1.4.5 DNA demethylation

It has been apparent for some time that DNA methylation at numerous genomic locations is dynamic. However a passive demethylation mechanism, whereby downregulation of DNMTs results in demethylation over the course of numerous cell divisions, does not explain rapid DNA demethylation events (Bhutani *et al*., 2011). An example of this is the demethylation of the paternal chromosome complement, which has been observed to occur in post-fertilisation zygotes (Oswald *et al*., 2000). The mechanism of active DNA demethylation remained unresolved until Tahiliani *et al* showed methyl cytosine (5mC) can be converted, by oxidation and hydroxylation, to hydroxymethylcytosine (5hmC) by members of the ten-eleven translocation (TET) protein family (Figure 1.5) (Tahiliani *et al*., 2009). Further studies have shown 5hmC to be deaminated by activation-induced cytidine deaminase (AID), a member of the APOBEC protein family, generating the base hydroxymethyl uracil (5hmU) (Cortellino *et al*., 2011, Guo *et al*., 2011). 5hmU is in turn recognised by members of the uracil DNA glycosylase (UDG) protein family, TDG and SMUG1, which mediate a 5hmU to cytosine conversion via the base excision repair pathway (BER), completing the conversion from methylated cytosine to unmethylated cytosine (Figure 1.5) (Cortellino *et al*., 2011, Guo *et al*., 2011).

More recently, reports have been published showing that TET1 proteins can further oxidise 5hmC, generating the newly identified 5-formylcytosine (5fC) and 5-

carboxylcytosine (5caC) (He *et al*., 2011, Ito *et al*., 2011). These bases also appear to function in DNA demethylation as 5caC can be excised and replaced by cytosine via the action of TDG in the BER pathway (Figure 1.5) (He *et al*., 2011). Ito *et al* also showed that 5caC could be directly converted to cytosine, without BER, in a reaction mediated by a putative decarboxylase (Ito *et al*., 2011).

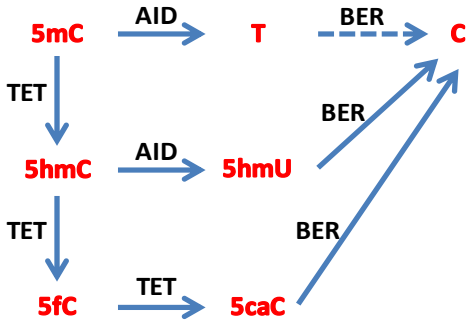


Figure 1.5 Active DNA demethylation.
5mC can be directly deaminated by AID to form thymine. Such cytosine to thymine conversions can be recognised and repaired by the BER pathway or can persist unrepaired resulting in a permanent base conversion to thymine. 5mC can also be converted to 5hmC by the TET family enzymes. 5hmC can be deaminated to form 5hmU or further oxidised by TET proteins to form 5fC and 5caC. 5hmU and 5caC are removed by the BER pathway and replaced by cytosine residues.

Genome-wide studies of 5hmC distribution have revealed a general enrichment at active enhancer elements and gene bodies (Stroud *et al*., 2011) as well as a positive correlation between 5hmC enrichment and active gene expression (Song *et al*., 2011). These findings indicate a potential role for 5hmC in the regulation of gene expression. It is currently unknown how transient 5hmC is, but it is possible that this base has a role in protecting DNA from methylation. It has been reported that DNA binding of the MBD protein Mecp2 is impaired when 5mC in the target sequence is replaced by 5hmC (Valinluck *et al*., 2004). As MBD protein binding to methylated DNA recruits histone modifying complexes which in turn mediate the establishment of a repressive chromatin state (Nan *et al*., 1998, Jones *et al*., 1998, Fuks, 2005), the inhibition of MBD-DNA interaction is likely to impair heterochromatin formation. It has also been reported that the DNA binding activity of DNMT1 is impaired by 5hmC (Valinluck and Sowers, 2007, Tahiliani *et al*., 2009), further supporting a putative role for 5hmC in protection of DNA from methylation. In this case, impaired DNA binding of DNMT1 would function to prevent maintenance DNA methylation and passive demethylation over cell divisions would be promoted.

1.4.6 Bisulphite sequencing

The bisulphite sequencing method for profiling DNA methylation was developed and first described by Clark *et al* (Clark *et al*., 1994). This technique allows the methylation status of individual cytosine residues to be elucidated. The process uses sodium bisulphite to convert unmethylated cytosine nucleosides to uracil in conditions where methylated cytosine is unchanged. The first step in this process is the sulfonation of unmethylated cystosine residues to produce cytosine sulfonate. These modified residues then undergo hydrolytic deamination which converts cytosine sulfonate to uracil sulfonate following which alkali desulphonation removes the sulfonate group yielding uracil (Figure 1.6).

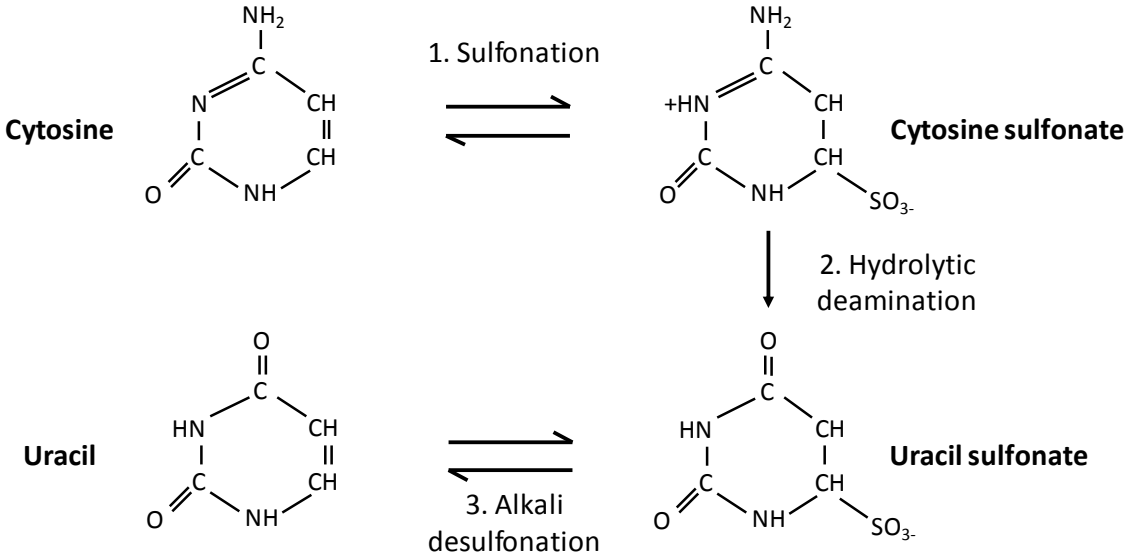


Figure 1.6 Schematic representation of the bisulphite conversion reaction.
Adapted from (Clark *et al*., 1994).

A number of kits are now commercially available with which to achieve bisulphite conversion of target DNA. For the DNA methylation studies presented in this thesis the EZ DNA Methylation Direct Kit from Zymo Research was used. The manufacturers of this kit claim that >99.5 % of non-methylated cytosine residues are converted to uracil, and that >99.5 % of methylated cytosines are protected from bisulphite conversion when using this kit.

Bisulphite conversion of DNA is followed by PCR amplification of a specific genomic region of interest. Numerous online tools are available to aid with the design of primers to amplify bisulphite modified DNA as several factors require consideration. Primers should be designed to amplify bisulphite modified DNA templates presuming that 100 %

conversion of unmethylated cytosine residues will be achieved. Thus cytosine residues that are not part of a CpG dinucleotide will be presumed to have been converted to uracil in the template DNA and must be paired with adenine residues in the primer. The presence of CpG dinucleotides within primer binding sites should be avoided as whether these cytosine residues will remain cytosine following bisulphite modification or be converted to uracil is dependent on their DNA methylation state and likely unknown at the point of primer design. The incorporation of CpG sites in primers can bias the PCR amplification of methylated or unmethylated DNA and will not allow the DNA methylation state of the associated cytosine residue to be determined by sequencing. The process of bisulphite modification is known to cause degradation of DNA (Grunau *et al*., 2001) and the presence of uracil residues slows DNA polymerase-mediated amplification of target sequences. The combination of these factors creates a correlation between shorter target amplicon size and increased PCR yield however amplicons of up to ~ 1 Kb are generally quite achievable.

1.5 Genome-wide mapping of epigenetic modifications

The identification and characterisation of epigenetic modifications across the genome allows the detection of regulatory elements and provides an insight into the regulatory mechanisms that function at these elements. Many experimental techniques have been developed to aid the genome-wide mapping of epigenetic modifications, some of which are described below.

1.5.1 Chromatin Immunoprecipitation

Chromatin immunoprecipitation (ChIP) was first described in 1988 (Solomon *et al*., 1988) and is now a standard technique employed to investigate protein-DNA interactions. Conventional ChIP uses formaldehyde to cross-link proteins to associated DNA after which chromatin is fragmented by sonication. Immunoprecipitation, using an antibody specific for a protein of interest, enriches for complexes containing that protein. Enriched DNA sequence elements are then purified away from associated proteins. Analysis of these enriched sequences provides information regarding the DNA elements with which a protein of interest interacts. This analysis can be achieved by qPCR to assay the enrichment of specific sequence elements, or by the high throughput methods of hybridising ChIP-enriched DNA onto microarrays (ChIP-chip) or using massively parallel sequencing technology to directly sequence enriched DNA elements (ChIP-seq).

1.5.1.1 ChIP-chip

High throughput quantitation of immunoprecipitated DNA can be achieved by hybridisation to tiled genomic DNA microarrays (ChIP-chip) which allows the simultaneous analysis of thousands of DNA sequence elements to determine their enrichment following ChIP. Microarrays are typically glass surfaces to which thousands of single stranded oligonucleotide probes are attached. A variety of high density DNA microarrays are commercially available that are generated by *in situ* synthesis of oligonucleotides, of typically 50-80 bases in length, directly on the chip surface. Affymetrix arrays are generated using photolithography where each DNA residue contains a photolabile protection group. A series of masks are used to direct specific patterns of UV light onto the chip in order to cleave UV-labile protection groups from targeted oligonucleotide chains allowing addition of the next nucleotide. Nimblegen arrays are synthesised using the same chemistry however a system of digital mirrors is used in place of masks to direct UV light to specific oligonucleotide chains. Agilent arrays are generated using an *in situ* synthesis printing process that utilises SurePrint Inkjet technology to deposit oligonucleotide monomers on to chips. Microarrays can also be generated by attaching double stranded PCR products of 200-400 bp in length to solid chips via 5'-(C6) amino-link modifications that are present on one strand of the PCR products (Dhami *et al*., 2005). Unmodified strands are separated from their partner strands by denaturing and are removed from the chip by washing to generate a complete microarray. The choice of DNA microarray platform is a balance between the high resolution and high genome coverage of commercial printed tiling arrays and the increased sensitivity offered by PCR product arrays (Dhami et al., 2005).

In ChIP-chip, fluorescently labelled ChIP DNA and input or control DNA are labelled with different fluorophores and co-hybridised to the genomic DNA microarray. Fluorescent spots can then be detected across the microarray corresponding to genomic sequence elements that are bound by labelled ChIP or input DNA due to their complementary sequences. The ratio of ChIP to input DNA fluorescence intensities at each element represented on the microarray is calculated to provide a measure of the extent of protein binding to the corresponding genomic locus (Iyer et al., 2001). Thus by identifying the DNA probe sequence that each fluorescent spot maps to we can identify genomic regions that are bound by the protein of interest for which ChIP was performed.

1.5.1.2 ChIP-seq

Conventional ChIP followed by massively parallel sequencing (ChIP-seq) is a technology that emerged in 2007 (Robertson *et al*., 2007) following the establishment of ChIP-chip platforms. Comparisons of these two technologies have shown the enrichment profiles of protein factors to be highly conserved between them. However these comparisons have also shown ChIP-seq to generate enrichment profiles with higher sensitivity and specificity than ChIP-chip due to the higher spatial resolution achievable with ChIP-seq profiling (Robertson *et al*., 2007, Ho *et al*., 2011). The ability of ChIP-seq to profile binding sites of a protein of interest across the whole genome simultaneously is another advantage to the ChIP-seq technology, in addition to this ChIP-seq has been reported to be the more cost-effective method of analysing mammalian genomes. This technique is now routinely used to profile the interactions of a protein of interest with DNA elements on a genome-wide scale.

There are several next generation sequencing (NGS) platforms available including Applied Biosystems' SOLiD platform, Applied Science's 454 Sequencing platform, Ion Torrent PGM and Illumina sequencing platforms. These platforms differ in many aspects including sequencing biochemistry, instrument cost, sequence yield per run, sequencing cost, run time, read length and accuracy. Illumina NGS platforms have dominated the field in recent years and as Illumina sequencing was used in this project I shall explain this method in more detail below.

In order to undergo next generation sequencing (NGS), genomic ChIP DNA must first be used to create ChIP-seq libraries. This involves the ligation of adaptor oligos to either end of ChIP-enriched genomic DNA fragments, followed by PCR amplification and size selection. ChIP-seq library samples are denatured and loaded into separate channels of a flow cell. The flow cell surface is covered by a lawn of primer pairs complementary in sequence to the adapters that are ligated to ChIP DNA during library preparation. DNA fragments from ChIP-seq libraries are bound by the flow cell primers which are extended by PCR to generate copies of ChIP-enriched DNA sequences that are covalently bound to the flow cell surface (Figure 1.7). Denaturing and washing removes the original template DNA strands leaving single stranded molecules bound to the flow cell in random patterns. The free end of each single stranded DNA molecule then binds to an adjacent primer on the flow cell to form a bridge (Figure 1.7). PCR extends the newly bound primer to create a double stranded bridge. Denaturing results in the generation of two single stranded

DNA molecules originating from one ChIP-seq library fragment. Repetition of bridge amplification generates many bridges from one original library fragment (Figure 1.7). This process of clonal bridge amplification proceeds to generate millions of unique DNA sequence clusters over the whole flow cell channel surface and is termed Cluster Generation.

Upon completion of bridge amplification, double stranded bridges are denatured. Reverse DNA strands are then cleaved and removed from the flow cell by washing, leaving whole clusters that consist only of forward DNA strands (Figure 1.7).

During the sequencing stage the millions of clusters present in each flow cell channel are sequenced simultaneously (Figure 1.7). DNA strands are bound by sequencing primers and the synthesis of complementary strands commences. The dNTPs used in these sequencing reactions are fluorescently labelled, with a different coloured fluorophore corresponding to each of the four DNA bases. Following the addition of each new nucleotide, clusters are excited by a laser which stimulates the emission of a detectable colour from the newly added nucleotide allowing its identification and the generation of DNA sequence information. The dNTPs used in sequencing also contain a terminator group which prevents more than one base being added per cycle. Once the fluorescence signal of the latest base addition has been detected, the fluorescent label and blocking group are removed from the newly added nucleotide and the next base may be added to the sequencing DNA strand.

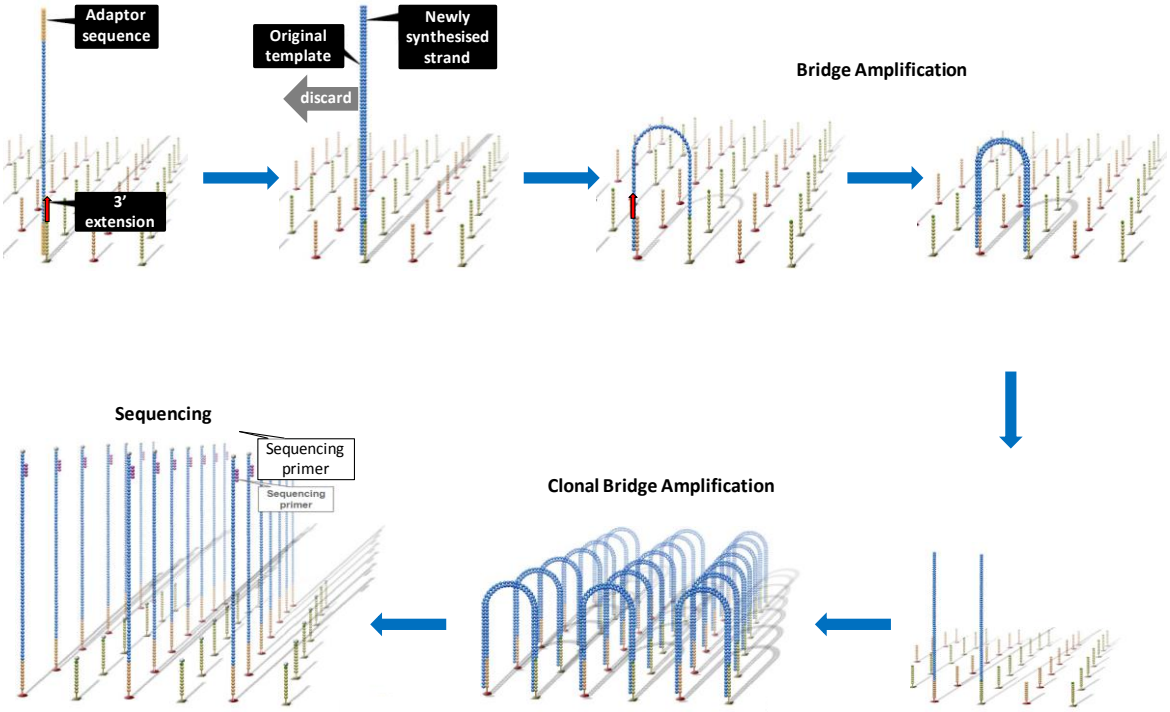


Figure 1.7 Schematic of NGS cluster generation and sequencing using an Illumina platform.
ChIP-seq library fragments bind to a lawn of primers on the flow cell surface via their complementary adapter sequences. Primer extension generates a copy of the ChIP-seq library sequence which is covalently bound to the flow cell and the original sequence is removed. Clonal bridge amplification generates millions of unique DNA sequence clusters across the flow cell channel, which are then sequenced simultaneously using fluorescently labelled dNTPs. Adapted from http://mi.caspur.it/workshop_NGS09/docs/Cappelletti_NGS09.pdf

Upon completion of NGS runs, the sequence reads generated must be aligned to a reference genome in order to map enrichment of the protein of interest across the genome. From each sequencing reaction tens of millions of reads can be expected to be generated, NGS can therefore be seen to generate a huge volume of data at relatively low cost. A major issue associated with NGS is the processing of the large amount of sequence data produced per run, to extract and interpret biologically meaningful information. This has presented bioinformatics challenges for the quality scoring of data, processing data and storing of data (Shendure and Ji, 2008).

1.5.2 MeDIP

Methylated DNA immunoprecipitation (MeDIP) is technique developed for the detection of DNA methylation (Weber *et al*., 2005). Genomic DNA is fragmented by sonication and a monoclonal antibody, which specifically recognises 5mC, is used to enrich for methylated DNA fragments. MeDIP-enriched DNA sequences can then be analysed by

QPCR, to determine the relative enrichment of individual regions, or by the high-throughput microarray or NGS methods, to profile DNA methylation on a genome-wide scale.

1.5.3 ENCODE

The Encyclopaedia of DNA Elements (ENCODE) project was launched in 2003 by the National Human Genome Research Institute (NHGRI) with the aim of generating a catalogue of functional genomic elements and mapping their precise locations. This project was a multi-national collaborative effort. During the initial pilot and technology development stage functional elements were characterised across 44 genomic regions, equating to ~30Mb or ~1% of the human genome. 14 of these regions, which covered ~15Mb, were specifically selected as well-studied loci, the remaining 30 regions were selected randomly using a stratified random sampling method (Birney *et al*., 2007, Consortium, 2004). The volume and quality of the data generated by the pilot study was unprecedented as the project mapped transcriptional activity, histone modifications, novel TSSs, chromatin structure and TF binding sites across the ~1% of the human genome corresponding to the ENCODE regions (Birney *et al*., 2007).

Upon completion of the pilot stage in 2007, the second phase of the ENCODE project was entered which aimed to catalogue functional genomic elements on a genome-wide scale. During this phase great use was made of recently developed, powerful NGS technologies and techniques including ChIP-seq, DNase-seq, FAIRE-seq etc were used to map transcribed regions, protein-coding regions, TF binding sites, histone modifications and variants, chromatin structure, DNA methylation sites and chromosome-interacting sites across the human genome. A collection of 30 papers were published in Nature, Genome Research and Genome Biology in 2012 detailing the extensive findings of this stage of the ENCODE project which ultimately assigned biochemical function to 80% of the human genome by analysis of 1,640 data sets (Dunham *et al*., 2012).

In order to allow maximal comparisons and integration of data sets, laboratories involved in the ENCODE consortium mainly focussed on a small group of selected cell types classed as ‘tier 1’ or ‘tier 2’ based on their priority. Tier 1 cell types were the erythroleukemia cell line K562, the B-lymphoblast cell line GM12878 and the H1 embryonic stem cell line H1 hESC. Tier 2 included HeLa-S3 cervical carcinoma cells, HEPG2 hepatoblastoma cells and

primary human umbilical vein endothelial cells (HUVECs). Despite this focus on tier 1 and tier 2 cell types, data generated from a total of 147 cell types was collected in this phase of the ENCODE project.

In 2010 a chromatin state annotation algorithm was generated by an ENCODE consortium lab which discovers chromatin states and assigns them to regions throughout the genome (Ernst and Kellis, 2010). ChromHMM is a multivariate Hidden Markov Model which uses combinations of chromatin marks (determined by ChIP-seq, DNase-seq etc) and the spatial context of chromatin states relative to each other to assign chromatin state annotations throughout the genome at a 200 bp resolution. ChromHMM provides a means of combining a multitude of individual data sets to identify combinations of chromatin marks that are biologically meaningful and can be used to annotate functional genomic elements and indicate what the functions of these elements are.

ChromHMM has been used to generate chromatin state maps across the genomes of nine common cell types including ENCODE tier 1 and tier 2 cell types (Ernst *et al*., 2011). Chromatin states were defined and assigned based on the presence or absence of nine chromatin marks - H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K20me1 and CTCF - as determined by ChIP-seq. Once ChromHMM had assigned chromatin state annotations to each of the nine cell types analysed a file was produced containing genome-wide tracks mapping these states, these files are publicly accessible and can be viewed in the UCSC genome browser.

1.6 The chicken β-globin locus

The chicken β-globin locus has been extensively studied as a paradigm for long-range gene regulation. This locus contains the ρ, β^H, β^A and ε genes and is located within a developmentally regulated genomic region under the control of the β-globin locus control region (LCR) (Figure 1.8). The 5' and 3' boundaries of the β-globin locus are marked by the constitutive 5' HS4 and 3' hypersensitive (HS) sites respectively.

Approximately 16 kb 5' of the chicken HS4 site is the *folate receptor 1* (*FOLR1*) gene, which is expressed in primitive erythrocytes but repressed in definitive erythrocytes (Prioleau *et al*., 1999). The 16 kb that separates the *FOLR1* gene from the β-globin locus is packaged into condensed chromatin (Prioleau *et al*., 1999). 3' of the chicken β-globin

gene locus, at approximately 3 kb downstream of the 3' HS site, is the *olfactory 51M1* (*OR51M1*) gene (Bulger *et al*., 2000), which is expressed in cells of the olfactory epithelium.

The chicken β-globin genes are expressed in a developmentally regulated manner in erythroid cells. During the early stages of chick embryogenesis primitive erythrocytes are produced from blood islands in the yolk sac from where they enter the circulation (Sheng, 2010). These primitive erythrocytes have a specific haemoglobin gene expression profile as they express the ρ and *ε* genes (Bruns and Ingram, 1973). As chick embryogenesis progresses definitive erythrocytes are produced from definitive erythropoietic clusters in the yolk sac (Sheng, 2010). These cells enter the circulation at embryonic day 5 and by embryonic day 7 most circulating erythrocytes in the chick embryo are definitive. Later in embryogenesis, at approximately embryonic day 12 (E12), definitive erythrocytes begin to be produced intra-embryonically by the bone marrow and by E13-15 most circulating erythrocytes are definitive and bone marrow-derived. As with primitive erythrocytes, definitive erythrocytes exhibit a specific haemoglobin gene expression profile. These cells strongly express the β^A gene and β^H is also expressed at very low levels in definitive chick erythrocytes (Bruns and Ingram, 1973). As a result of these changes in the proportion of primitive versus definitive erythrocytes, the ρ and ε genes appear to be highly expressed in circulating erythrocytes early in embryogenesis but this expression is seen to fall as embryogenesis progresses. In parallel with this, the β^A gene is seen to transition from non-expressed to highly expressed in the circulating erythrocyte population between E5-10.

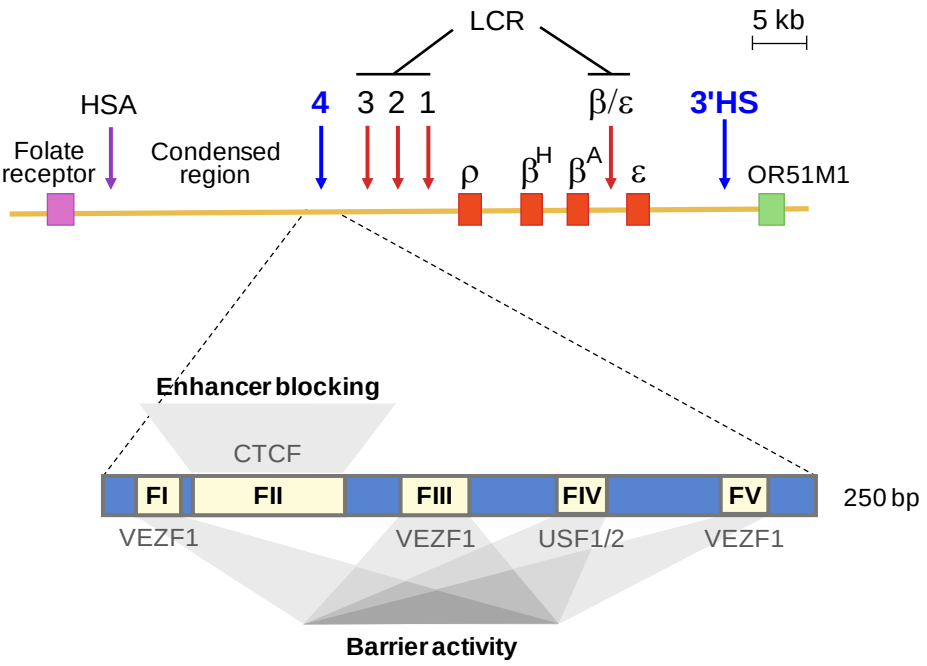


Figure 1.8 The β-globin gene locus and HS4 insulator element.
Boxes depict genes - (L-R) *Folate Receptor*, ρ, β^H, β^A, ε and *Chicken Olfactory Receptor*. Arrows show DNase I hypersensitive site, constitutive DNase I hypersensitive sites are shown in blue. Those DNase I hypersensitive sites comprising the chicken globin LCR are indicated. The 250 bp core of the HS4 element is enlarged to show the distribution of the five DNase I footprinted regions.

As the constitutive 5' HS4 site was known to mark a boundary between 3' erythroid-specific regulatory elements and a 5' region of condensed chromatin, its ability to function as an insulator was investigated and it became the first insulator element to be identified in a vertebrate species (Chung *et al.*, 1993).

1.7 Insulators

Insulators function to prevent genes from being inappropriately influenced by the transcriptional activity of their surrounding environment. These elements are generally ~0.5 – 3 kb in length and their activity is both position- and orientation-dependent. Insulators are defined as possessing one or both of two activities – enhancer blocking and barrier activity (Maston *et al.*, 2006, West *et al.*, 2002). Enhancer blocking activity functions to prevent promiscuous enhancers from interacting with promoter elements in a non-regulated manner, but only occurs when the insulator is positioned between the enhancer and promoter elements (Figure 1.9 A). Barrier activity describes the activity by which insulators can halt the spread of repressive heterochromatin, thereby preventing the inappropriate silencing of a gene or gene locus (Figure 1.9 B).

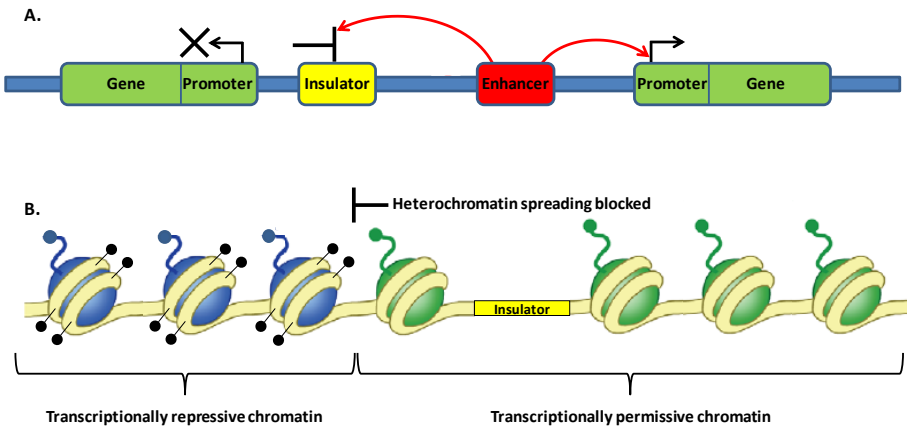


Figure 1.9 Mechanisms of insulator activity.
Insulator elements function to protect genes from inappropriate transcriptional influences in their surrounding environment. (A) Insulators can function as enhancer blockers when positioned between enhancer and promoter elements. (B) Insulators can function as barrier elements to halt the spread of repressive heterochromatin. Transcriptionally repressive nucleosomes are coloured blue, transcriptionally permissive nucleosomes are coloured green.

The chicken HS4 element functions as an insulator and is the best characterised vertebrate insulator element to date. The HS4 insulator has both enhancer blocking (Chung *et al*., 1993) and barrier activity (Pikaart *et al*., 1998), both of which are mediated by a 250 bp DNA element that can be divided into five DNase I footprinted regions (FI – V) (Figure 1.8) (Chung *et al*., 1997). Dissection of HS4 has shown its enhancer blocking and barrier activities to be regulated by separable mechanisms, which involve protein binding to specific DNA motifs within the HS4 footprinted regions. Binding of CTCF to HS4 FII mediates the enhancer blocking activity of the insulator (Bell *et al*., 1999). Meanwhile binding of USF1/2 heterodimers to FIV is necessary for barrier activity (West *et al*., 2004). USF1/2-mediated barrier activity prevents the spread of silencing histone modifications across the insulator and into downstream chromatin by co-ordinating the deposition of active histone marks on surrounding nucleosomes (West *et al*., 2004, Ma *et al*., 2011). It is believed that this collection of active histone modifications functions to prevent the deposition of repressive histone marks, thereby terminating heterochromatin assembly (West *et al*., 2004, Huang *et al*., 2007, Ma *et al*., 2011). However, binding of USF1/2 to FIV is not sufficient to mediate barrier activity of the HS4 insulator and DNA motifs within HS4 footprints I, III and V, which function as binding sites for the TF VEZF1, have been found to be essential for preventing transgene silencing in barrier assays (Dickson *et al*., 2010). Deletion of any of the three VEZF1 binding sites results in *de novo* DNA

methylation of the transgene and HS4 insulator, and silencing of the transgene (Dickson *et al*., 2010), suggesting a role for VEZF1 binding sites in protecting DNA elements from the spread of *de novo* DNA methylation.

1.8 VEZF1

In 1988 a chicken factor, named BGP1, was identified as binding to a $(dG)_{16}$ sequence in the chicken β^A promoter in erythrocytes (Lewis *et al*., 1988). Subsequently, human DB1 (Koyano-Nakagawa *et al*., 1994) and the highly homologous mouse *Vezf1* (Xiong *et al*., 1999) proteins were identified. Sequencing later showed BGP1 to be the chicken homologue of VEZF1 (Dickson *et al*., 2010). VEZF1 proteins are highly conserved across species with 89% homology between full length chicken, mouse and human VEZF1 proteins and 99% homology within their ZF domains, indicating a common and important biological function.

It was initially hypothesised that VEZF1 would function as a TF due to its glutamine stretch and proline-rich C-terminal region, which are characteristic of transcriptional activators (Gerber *et al*., 1994) (Figure 1.10). Luciferase reporter assays showed that a VEZF1 binding site activates the human endothelin 1 (EDN1) promoter in endothelial cells (Aitsebaomo *et al*., 2001) while studies by Miyashita *et al* (Miyashita *et al*., 2004, Miyashita and Sato, 2005) indicated a role for VEZF1 in regulating expression of the OP18 and Metallothionein 1 (MT1) genes, although direct interaction between VEZF1 and these gene elements has not been investigated. The $(dG)_{16}$ VEZF1 binding motif in the chicken β^A promoter does not affect transcriptional activity of this promoter (Jackson *et al*., 1989, Barton *et al*., 1993) but has been implicated in assisting nucleosome positioning at this element (Buckle *et al*., 1991).

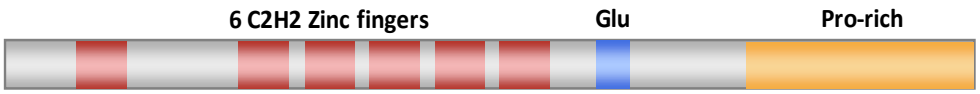


Figure 1.10 Domain organisation of VEZF1.
Red boxes represent the six C2H2 zinc fingers; blue box represents the polyglutamine chain; yellow box represents the C-terminal proline-rich domain.

The *EDN1*, *OP18* and *MT1* genes have roles in angiogenesis. In keeping with these findings, VEZF1 has been shown to be vital for the development of the vascular and lymphatic systems. The vasculature in the head, neck and back of VEZF1 null mouse

embryos is underdeveloped and poorly organised and these embryos have an embryonic lethal phenotype due to severe haemorrhaging, which is caused by a lack of vascular integrity resulting from the abnormal expression of cell junction and extra cellular matrix (ECM) proteins (Kuhnert *et al*., 2005, Zou *et al*., 2010).

There have been some contradictory findings reported regarding the expression patterns of VEZF1. Lewis *et al* (Lewis *et al*., 1988) reported that BGP1 was expressed in circulating chicken erythrocytes but not in the human HeLa cell line. This was determined by the formation of a specific complex in EMSA reactions, between the nuclear extract of embryonic or adult chicken erythrocytes and the $(dG)_{16}$ VEZF1 binding motif. No specific band appeared to form between nuclear extract from the human HeLa cell line and the $(dG)_{16}$ sequence motif, however the unpublished observations of Dr Adam West found the human VEZF1/G-string complex to have slightly higher mobility than the chicken VEZF1/G-string complex. This increased mobility leads to the specific human VEZF1/G-string complex being masked by a non-specific band of the same mobility in the EMSAs presented by Lewis *et al* (Lewis *et al*., 1988) and to the mistaken conclusion that VEZF1 was not expressed in HeLa cells. In support of this, VEZF1 mRNA and protein expression have since been profiled in the HeLa cell line by RT-QPCR and western blotting respectively (Strogantsev, 2009), these analyses show VEZF1 to be expressed in HeLa cells. VEZF1 protein expression in HeLa cells was also demonstrated by western blotting by Gowher *et al* (Gowher *et al*., 2012). Human VEZF1 has in fact been shown to be expressed across a broad range of human somatic cells by northern blotting, RT-PCR, RT-QPCR, western blotting and cDNA microarray analyses (Koyano-Nakagawa *et al*., 1994, Strogantsev, 2009) (Figure 1.11). Consistent with these findings in human, mouse VEZF1 has been shown to be broadly expressed across a range of adult mouse tissues and cell lines by northern blotting. VEZF1 expression has also been shown to be up-regulated in vascular cells during vasculogenesis in mouse, consistent with the essential function of VEZF1 in vasculogenesis (Kuhnert *et al*., 2005, Zou *et al*., 2010).

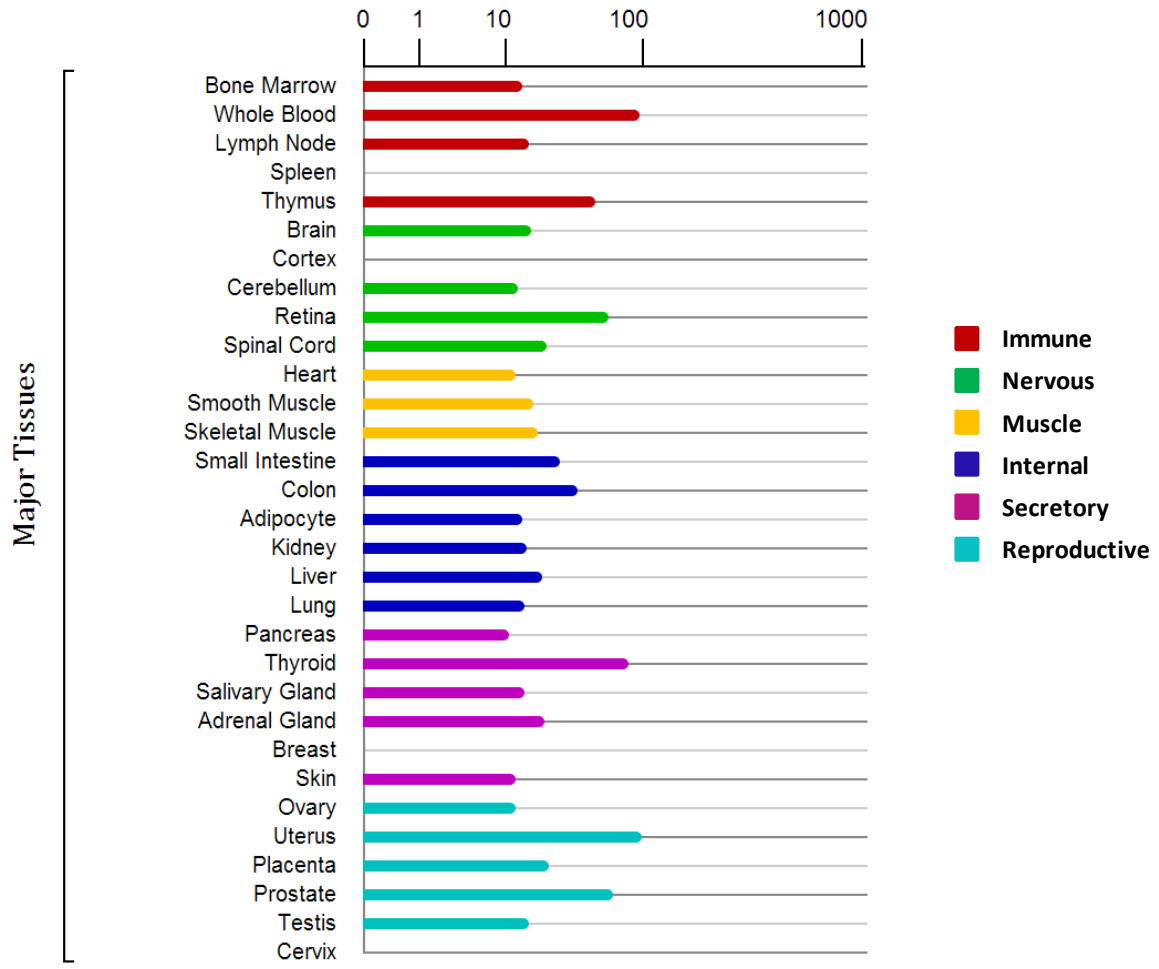


Figure 1.11 VEZF1 is broadly expressed across human somatic tissues.
Microarray analysis of VEZF1 cDNA levels in various human somatic tissues. Figure taken from GeneCards (www.genecards.org; GeneCards ID:GC17M056048).

To date a small number of VEZF1 binding sites have been identified in the human, mouse and chicken genomes. VEZF1 binding to the chicken β^A promoter and the human *IL3* and *EDN1* promoters has been shown to occur in cell types where these genes are expressed (Lewis *et al*., 1988, Koyano-Nakagawa *et al*., 1994, Aitsebaomo *et al*., 2001). These findings indicate a role for VEZF1 in positively regulating gene expression which is supported by the fact that VEZF1 binding to the *EDN1* promoter in endothelial cells activates the promoter causing expression of an associated transgene (Aitsebaomo *et al*., 2001). VEZF1 does not function as a direct transactivator of β^A globin gene expression however it has been hypothesised that its interaction with the gene promoter may aid transcriptional activation via nucleosome repositioning (Lewis *et al*., 1988, Clark *et al*., 1990).

Three VEZF1 binding sites in the chicken HS4 insulator element have been characterised. These sites have been shown to be essential for the barrier activity of HS4 and function to prevent the accumulation of *de novo* DNA methylation (Dickson *et al*., 2010) (Section 1.7). VEZF1 binding sites have also been shown to mediate protection of the hamster *Aprt* CGI from *de novo* DNA methylation and to direct demethylation of a pre-methylated CGI (Dickson et al., 2010) (section 1.4.4). Analysis of DNA methylation profiles in VEZF1 knock out or knock down cell lines is required in order to categorically define the role of this protein in protection of DNA from hypermethylation. However this is complicated by the fact that DNA methylation is deficient in *Vezf1* null mouse ES cells as Vezf1 itself is required for optimal expression of DNMT3B (Gowher *et al*., 2008).

It has been reported that VEZF1 affects alternative splicing by causing the elongating serine 2-phosphorylated RNA polymerase II isoform (Pol II-Ser2ph) to pause during transcription (Gowher *et al*., 2012). Pol II-Ser2ph pausing can alter the inclusion of alternative exons in mRNA transcripts, thus altering expression of different protein splice forms. One gene reported to be affected by this process is *Dnmt3b*. Data published by Gowher *et al* suggest that expression of the catalytically active *Dnmt3b1* isoform is mediated by VEZF1 binding to an intronic DNA sequence element located towards the 3' end of the *Dnmt3b* gene (Gowher *et al*., 2008). It was reported that VEZF1 binding to *Dnmt3b* induces pausing and accumulation of Pol II-Ser2ph at the VEZF1 binding site promoting expression of the active *Dnmt3b1* isoform. In a *Vezf1* knock out ES cell line, accumulation of Pol II-Ser2ph did not occur at the VEZF1 binding site and a decrease in levels of the *Dnmt3b1* isoform was observed which in turn lead to a global loss of DNA methylation (Gowher *et al*., 2012).

1.9 VEZF1 ChIP-chip

In a previous project in my supervisors' lab, in collaboration with the lab of Dr David Vetrie, ChIP-chip was used to map VEZF1 binding sites in the K562 cell line across the ENCODE region (see section 1.5.3), which covers ~1% of the human genome (Strogantsev, 2009, Consortium, 2004). Immunoprecipitated DNA from VEZF1 ChIP in K562 cells was fluorescently labelled with Cy3 CTP nucleotides and hybridised to microarray chips which were scanned using a confocal laser-based scanner. The intensities of fluorescent spots were quantified using ScanArray express software (Perkin Elmer) and z-scored to generate a standardised signal intensity ratio. Signal intensities across the genomic

regions represented on microarrays were visualised in the UCSC genome browser, allowing identification of genomic regions bound by VEZF1. The same experimental pipeline was followed using immunoprecipitated DNA from an anti-FLAG ChIP in the D3/6 K562 cell line (created by Dr Dan Li) which expresses a recombinant FLAG-tagged VEZF1 fusion protein at ~50 % of the level of endogenous VEZF1. An example window from the UCSC genome browser showing VEZF1 and FLAG-VEZF1 enrichment profiles across ~250 kilo bases of chromosome 1 is shown in Figure 1.12.

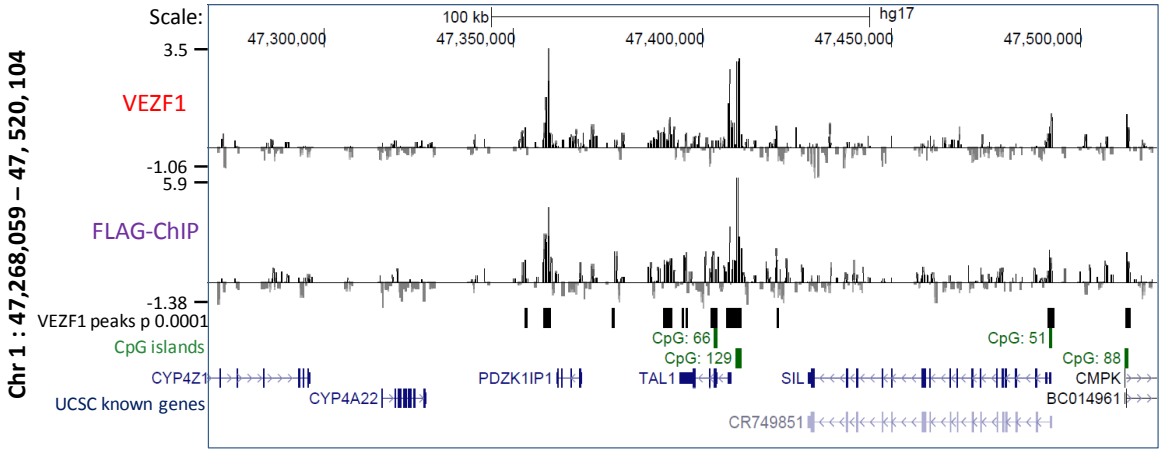


Figure 1.12 VEZF1 binding at specific elements revealed by ChIP-chip.
VEZF1 ChIP in K562 (top track) and FLAG-ChIP in D3/6 K562 (second track from top) z-scored enrichment profiles across chr 1: 47,268,059-47,520,104. Peaks identified using the ChIPOTle algorithm at p-value of significance = 0.0001 are shown as black boxes. CpG islands are shown as green boxes. UCSC annotated genes are shown in blue and direction of transcription is indicated by arrows (bottom track).

VEZF1 ChIP-chip fluorescence peaks were identified using the ChIPOTle peak-finding algorithm. This approach identified 125 VEZF1 peaks across the 1 % of the human genome profiled.

A comparison of the locations of VEFZ1 binding sites and the locations of known protein-coding genes using the Galaxy on-line interface showed that the majority of DNA elements bound by VEZF1 are gene-associated. A bias for VEZF1 binding towards the 5' end of genes was also identified, with 64% of VEZF1 binding sites mapping to promoters - defined as being within 1 kb of an annotated TSS - and a further 32 % of VEZF1 peaks mapping to distal regulatory elements such as enhancers and insulators - defined as being >1kb from a TSS and associated with DNaseI hypersensitivity and H3K4 mono-, di-, or tri-methylation (Figure 1.13). This striking association of VEZF1 with gene regulatory

elements indicates a potential role for VEZF1 in regulating gene expression, a hypothesis supported by the results of RNA-microarray analysis which showed approximately 80 % of genes that were associated with VEZF1 in K562 cells to be expressed in this cell line (Strogantsev, 2009).

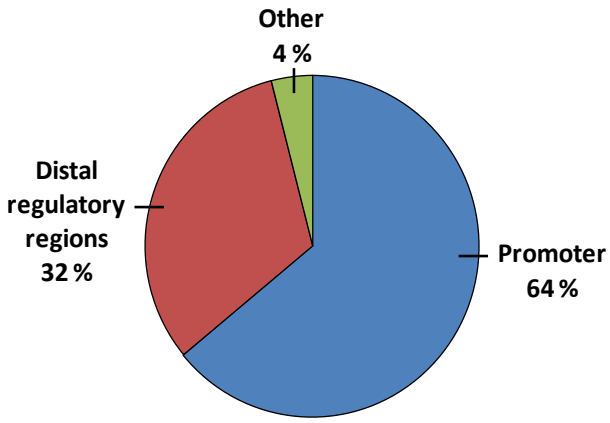


Figure 1.13 VEZF1 ChIP-chip enrichment peaks map to regulatory regions.
Location of VEZF1 ChIP-chip peaks with respect to gene regulatory elements. Promoters are defined as being within 1 Kb of a TSS; distal regulatory regions are defined as being > 1 kb from a TSS and associated with DHSs and H3K4 methylation peaks; remaining peaks as classed as 'other'.

Although very informative, the 125 VEZF1 enrichment peaks identified across the ENCODE region were not sufficient in number to allow further analyses, for example by grouping peaks based on their enrichment rank or type of associated regulatory element. The enrichment peaks generated by ChIP-chip were also very broad and often contained multiple potential VEZF1 binding sites. It was therefore difficult to predict which motifs function as VEZF1 binding sites at enriched elements. As described previously, ChIP-seq generates higher resolution enrichment profiles in which maximal points of enrichment generally correlate well with DNA motifs bound by the protein of interest. It was therefore considered that performing ChIP-seq for VEZF1 would generate higher resolution VEZF1 enrichment profiles which would enable the sequence motifs most likely to function as VEZF1 binding elements *in vivo* to be identified.

1.10 Methods for identifying TF binding sites and generating consensus binding sequences

As regulatory elements are composed of TF binding sites, identification of motifs to which TFs bind is essential to the discovery and characterisation of regulatory networks. Many

techniques have been developed to aid the identification and characterisation of TF binding sites, a selection of which are discussed below.

1.10.1 DNase I footprinting

DNase I footprinting is an *in vitro* protection assay used to identify protein-bound DNA elements. In this assay a double stranded DNA substrate is radiolabelled at its 3' phosphate and incubated in a binding reaction with nuclear extract or purified or recombinant proteins of interest. Binding reactions are then treated with a range of endonuclease DNase I concentrations. DNase I cleaves DNA substrates in a largely sequence independent manner creating a ladder of DNA fragments, however DNA elements which are bound by protein are protected from digestion as the protein prevents DNase I accessing the bound DNA. Cleavage reactions are electrophoresed through denaturing polyacrylamide gels and visualised by autoradiography or phosphorimaging. Digest reactions lacking protein appear as a ladder of DNA fragments while sequence-specific protein-DNA interactions will be revealed as 'footprints' that are protected from nuclease action. In order to identify the exact sequence of DNaseI footprints, Maxam-Gilbert sequencing marker lanes are run alongside cleavage reactions. DNaseI footprinting can be extremely helpful for the initial characterisation of novel regulatory elements and can provide a reference point from where the characterisation of specific protein binding motifs can proceed. However, the *in vitro* nature of this technique means that the binding events detected may not necessarily correspond to genuine *in vivo* binding events.

1.10.2 EMSA

Electrophoretic mobility shift analysis (EMSA) is a technique which allows the visualisation of protein-DNA interactions *in vitro*. During EMSA double stranded radiolabelled oligonucleotide (probe DNA) of a specific sequence is incubated with either a recombinant protein of interest or the nuclear lysate of a specific cell type to allow protein-DNA binding to occur. Binding reactions are then electrophoresed through a native polyacrylamide gel, following which gels are dried and the distribution of radiolabelled probe throughout the gel is visualised by phosphorimaging. During electrophoresis, free unbound probe resolves as a single band at the base of the gel, meanwhile probe that has interacted with protein to form protein-DNA complexes electrophorese at a much slower rate due to the increased mass of the complex. Protein-

probe complexes therefore appear as discrete bands which locate higher up the gel than free probe. The specificity of complexes detected by EMSA can be tested by the addition of an antibody specific for a protein of interest to the binding reaction. Association of the antibody with the protein-DNA complex will further increase the mass of this specific complex causing even slower migration and resulting in the corresponding band appearing 'super-shifted'. Interaction between protein and antibody may also interfere with protein-DNA complex formation. Unlabelled double stranded competitor oligonucleotides can be added to EMSA binding reactions in order to either demonstrate the sequence specificity of complex formation or to test for the ability of specific DNA sequences to compete with the probe sequence for binding to the protein of interest. This technique allows the validation of putative protein binding motifs in isolation, *in vitro*. Individual nucleotides or groups of nucleotide bases can be mutated from the wild type sequence to comprehensively determine which specific nucleic acid residues are important for protein binding. However, if the binding of a protein of interest to a specific binding site *in vivo* requires cooperativity with another transcription factor, it may not be detected by EMSA, for example if the co-binding protein is not present in the binding reaction or the binding site of the co-binding protein is not present on the DNA oligonucleotide. As EMSA detects protein-DNA interactions *in vitro*, however, it cannot be presumed that binding events detected by this method will correspond to those that occur *in vivo*.

1.10.3 Microwell-based protein DNA-binding specificity assay

A high throughput method for determining the DNA binding specificity of transcription factors was developed by Hallikas and Taipale (Hallikas and Taipale, 2006) (Figure 1.14). The premise of this technique is similar to that of EMSA in that it allows the validation and characterisation of putative protein binding motifs *in vitro*. A transcription factor of interest is expressed as a fusion product with the *Renilla* luciferase reporter enzyme. This fusion protein is incubated in a binding reaction with biotinylated double stranded DNA oligonucleotides, which correspond in sequence to a known consensus binding sequence for the protein. Non-biotinylated competitor oligos may also be included in the binding reaction, these oligos can correspond to putative binding sequences or contain base substitutions in the consensus binding site. The binding reactions are then transferred to individual wells of a streptavidin-coated 96-well plate, where biotinylated oligos are captured by their interactions with streptavidin. After washing to remove unbound

material, a luminometer is used to measure the amount of fusion protein captured in each well along with biotinylated sequences. This assay can be used to rapidly screen vast numbers of competitor sequence motifs, allowing the binding specificity of a specific transcription factor to be determined quickly. However, the assay can only be performed with recombinant luciferase-fusion proteins, meaning that cell nuclear extracts cannot be used to demonstrate the presence of specific proteins in specific cell types, as can be achieved by EMSA. As with EMSA, if co-binding of another protein is required for binding of the protein of interest to a putative binding site, this interaction will not be detected. Also, as this assay detects *in vitro* protein-DNA binding events, these interactions may not be conserved *in vivo*.

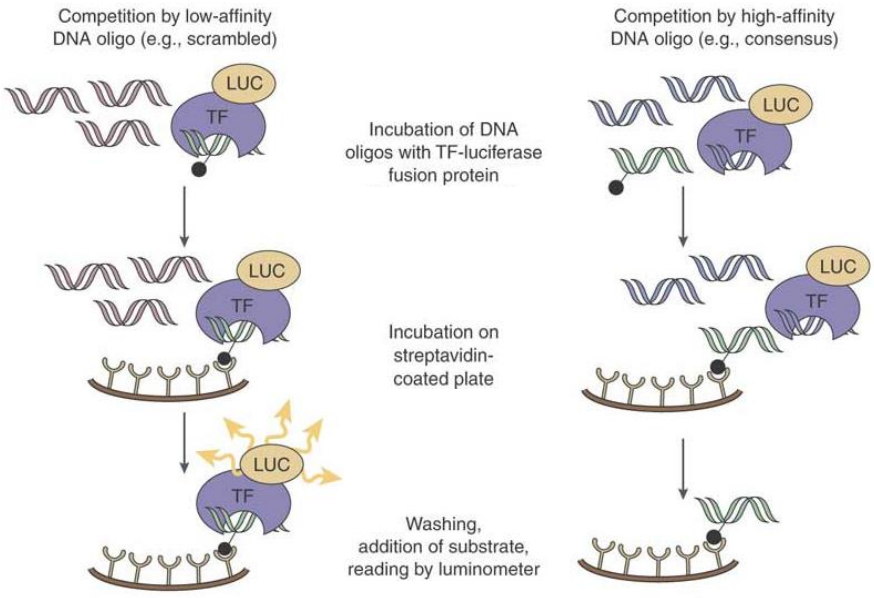


Figure 1.14 The microwell-based protein-DNA binding specificity assay.
Adapted from (Hallikas and Taipale, 2006).

1.10.4 SELEX-seq

The systematic evolution of ligands by exponential enrichment (SELEX) is a well established method for the characterisation of protein binding sites. In this method, double stranded (ds) DNA libraries are generated which contain all possible oligo sequences of a given length. A protein of interest is incubated with the dsDNA library in a binding reaction following which protein-bound DNA sequences are purified and enriched by PCR amplification. The process of protein binding, complex purification and amplification of bound sequences is repeated several times to enrich for DNA sequences that act as binding sites for the protein of interest (Wang *et al*., 2011). Enriched DNA

motifs isolated by the SELEX method can now be sequenced by NGS. Bioinformatic analyses of sequencing data can then be performed to generate consensus binding motifs or position weight matrix models. The SELEX-seq technique can produce large quantities of data regarding the *in vitro* DNA binding specificity of specific transcription factors. However, this technique does not identify specific genomic locations that may function as transcription factor binding sites *in vivo*.

1.10.5 ChIP-seq

As described previously, ChIP-seq is a powerful tool that allows genomic sites at which a protein of interest is enriched *in vivo* to be mapped on a genome-wide scale. For DNA-binding transcription factors, the peak of ChIP-seq enrichment generally correlates quite well with transcription factor binding sites (Zhang *et al*., 2008), however further experimentation (such as EMSA) is required to define and characterise specific protein binding motifs. For DNA-binding transcription factors, analysis of ChIP-seq enrichment data using motif discovery bioinformatic tools should allow the generation of a consensus binding motif for the protein of interest. As ChIP-seq profiles sites of protein enrichment *in vivo*, if binding of a protein of interest is mediated by the co-binding of another factor, these sites will be detected by ChIP-seq. However, ChIP-seq cannot reveal whether a protein interacts with DNA directly or if it is part of a multi-protein complex in which another protein factor mediates binding to DNA.

1.11 The K562 cell line

The K562 cell line was chosen as a model system for use in investigating the genomic interactions of VEZF1 for a number of reasons (section 3.1). Among these were the fact that our lab had already mapped VEZF1 binding across 1 % of the human genome by ChIP-chip in this cell line (Strogantsev, 2009). The classification of K562 as a tier 1 cell line by the ENCODE consortium has also resulted in an abundance of NGS data from this cell line being generated and made publicly available. This makes K562 an attractive system to use as ChIP-seq data generated in this project could be compared to the wealth of publicly available data. As part of the ENCODE analyses, a ChromHMM chromatin state map had also been generated for the K562 cell line which was also very useful in my project (Ernst *et al*., 2011).

K562 is an immortalised leukemic cell line which was established in 1970 from cells isolated from the pleural effusion of a female patient with chronic myeloid leukaemia (CML) during blast crisis (Lozzio and Lozzio, 1975). Karyotyping has revealed K562 cells to have a hypotriploid chromosome complement with a modal chromosome number of 67 and 21 marker chromosomes which are the product of translocations, duplications and inversions (Naumann *et al.*, 2001). Analysis of copy number variation shows the majority of the K562 genome to have a normal copy number while specific regions are amplified or heterozygously deleted and a very small proportion of chromosome 9 is homozygously deleted (Figure 1.15). From this analysis it is apparent that despite the hypotriploid karyotype of this cell line the majority of the genome is present in normal copy number complement.

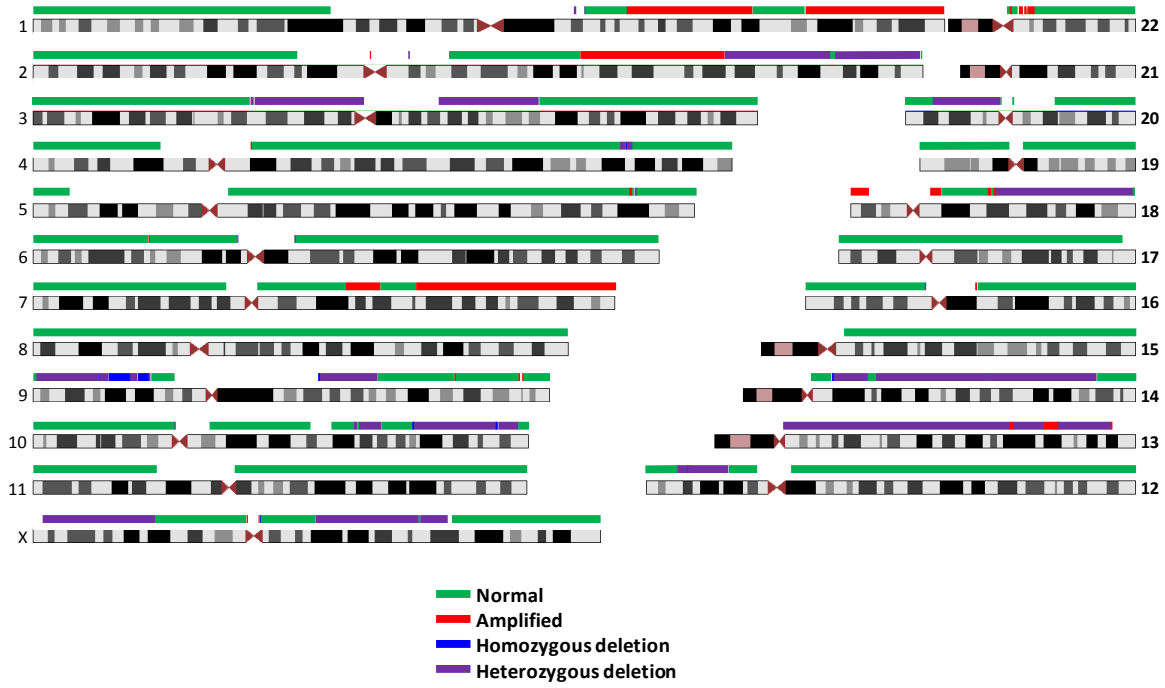


Figure 1.15 The majority of the K562 genome is present in normal copy number complement.
Chromosome numbers are shown to the right or left of each chromosome. Genomic regions present in normal copy number are indicated by a green block, amplified regions are marked by red blocks, heterozygous and homozygous deletions are marked by purple and blue blocks respectively.

K562 is a multipotent hematopoietic precursor cell line with erythroid properties that include expression of the red blood cell membrane-spanning protein glycophorin (Andersson *et al.*, 1979). K562 cells have the potential to differentiate along the megakaryocyte or erythroid lineages. Erythroid differentiation can be promoted by

treatment of K562 cells with a number of inducing agents and induces the expression of the embryonic globin genes in approximately the same proportions as are evident during early human embryogenesis. Analysis of the K562 genome has found no deletions or rearrangements of the globin genes (Rutherford *et al*., 1981), therefore their expression in this cell line is considered to represent a good system for investigating the regulation of globin gene expression.

While the K562 cell line cannot be considered as a 'normal' system, evidence in the published literature suggests that gene regulatory events are very similar between K562 cells and primary erythroid cells. ChIP-chip, ChIP-seq and DNase I HS profiles show histone modifications, TFBSs and DNase I HSs to be highly conserved between K562 and primary erythroid cells (Anguita *et al*., 2001, Follows *et al*., 2006, De Gobbi *et al*., 2007, Fujiwara *et al*., 2009). For example following ChIP-seq of the TF GATA-1 in K562 cells 36 GATA-1 binding sites were analysed for enrichment in mouse primary erythroblasts by GATA-1 ChIP-QPCR (Fujiwara *et al*., 2009). 32 of the 36 sites investigated were found to be enriched relative to a negative control element. These findings demonstrate the conservation of GATA-1 enriched sites between the K562 cell line and primary mammalian erythroblasts. GATA-1 factors are key regulators of hematopoiesis therefore the conservation of enriched sites between K562 cells and primary erythroblasts indicate the usefulness of this cell line as a model system. The treatment of K562 and primary erythroid cells with various inducing agents has also been shown to cause highly comparable effects on globin gene expression indicating that regulation of these genes is mediated by the same mechanisms in these two cell types (Lampronti *et al*., 2003, Pace *et al*., 2003, Zhou *et al*., 2000). The K562 cell line is therefore considered to be useful platform with which to investigate erythroid-specific regulatory events.

1.12 Aims and objectives of this thesis

This project was established with the aim of identifying and characterising VEZF1 binding sites in the human genome. ChIP-chip across 1 % of the human genome and EMSA analyses of specific putative VEZF1 binding sites have been performed previously (Strogantsev, 2009, Koyano-Nakagawa *et al*., 1994, Aitsebaomo *et al*., 2001). I aimed to further these findings by mapping VEZF1 binding sites on a genome-wide scale and validating a proportion of novel putative VEZF1 binding sites by EMSA analyses. We also wanted to characterise VEZF1-enriched sites – for example as promoter, enhancer, or

insulator elements, etc. – with a view to creating a better understanding of the functions that VEZF1 binding to genomic elements may mediate. Based on findings that VEZF1 binding sites within the chicken HS4 insulator element are essential for barrier activity (Dickson *et al.*, 2010), we anticipated that profiling VEZF1 binding sites across the human genome would identify putative novel insulator elements within the human genome. The general aims of this work were to:

1. Identify VEZF1 binding sites across the human genome in K562 cells using the ChIP-seq method.
2. Characterise VEZF1 binding sites by comparison to chromatin states defined by the ENCODE consortium.
3. Investigate the DNA binding specificity of VEZF1 using motif discovery tools and EMSA.
4. Utilise the *in vivo* chick embryogenesis model system to investigate the relationship between VEZF1 binding, promoter DNA methylation and transcription of the chicken β-globin genes during development.

Chapter 2

Materials and Methods

Materials

2.1 Cell lines

Cell line	Reference	Source
K562	(Lozzio and Lozzio, 1997)	Prof Tony Green, School of Clinical Medicine, University of Cambridge, via Dr David Vetrie, Institute of Cancer Sciences Epigenetics Unit, University of Glasgow

2.2 Reagents

2.2.1 Cell culture reagents

Product	**Manufacturer**	**Product Code**
RPMI Medium 1640 (1X) + GlutaMAX™ -l	Gibco	61870
Fetal Bovine Serum	Sigma	F9665
Penicillin/Streptomycin	Sigma	P0781

2.2.2 Antibodies

Anti-VEZF1 antibodies were raised against a C-terminal portion of the chicken VEZF1 protein (amino acids 377-548) (Dickson *et al*., 2010).

Antibody	**Manufacturer**	**Product Number**
FLAG (M2)	Sigma	A2220
Non-immune rabbit IgG	Sigma	I5006
Non-immune mouse IgG	Sigma	I5381

2.2.3 Enzymes

Enzyme	**Manufacturer**	**Product Number**
T4 PNK	NEB	M0201
Proteinase K	Sigma	P5568
RNase A	Fermentas	EN0531

RNase OUT	Invitrogen	160000840
RNase H	NEB	M0297
Bgl II	NEB	R0144

2.2.4 Oligonucleotides

Oligonucleotides were ordered from Biomers or MWG Eurofins. Sequences are given in appendix.

2.2.5 Plasmids

pCITE4b-VEZF1	(Dickson *et al*., 2010)	*In vitro* transcription vector containing full length cDNA encoding chicken VEZF1
pGEM®-T easy Vector System I	Promega	Product number: A1360

2.2.6 Chemicals

Reagent	**Manufacturer**	**Product Number**
Ammonium Persulfate	Fisher Scientific	A/6160/53
TEMED	Sigma	T9281
Acrylamide:Bisacrylamide (37.5:1)	Amresco	0254
Acrylamide:Bisacrylamide (19:1)	Sigma	A2917
Acrylamide:Bisacrylamide (29:1)	Sigma	A2792
Phenol-Chloroform	Sigma	P3803
Chloroform	Fisher Scientific	C/4960/17
Trizol	Life Technologies	15596026
BCP	Sigma	B9673
Ethidium Bromide	Sigma	E1510
Safeview	NBS Biologicals Ltd	NBS-SV1

2.2.7 Other molecular biology reagents

Reagent	**Manufacturer**	**Product Number**
100 bp DNA ladder	NEB	N32315
1 Kb DNA ladder	Invitrogen	10787-018
PageRuler Prestained Protein Ladder	Thermo Scientific	26616
γ^{32} P ATP	Perkin Elmer	BLU502H250UC

DTT	Melford Laboratories Ltd	MB1015
Triton X-100	Sigma	T8787
BSA	NEB	B9001S
RNasin Ribonuclease Inhibitor	Promega	N211A30683007
100 mM dNTP set	Invitrogen	10297018
Taq DNA Polymerase	NEB	M0267S
Platinum Taq DNA Polymerase	Invitrogen	10966-018
Subcloning Efficiency™ DH5α™ Competent Cells	Invitrogen	18265-017
X-gal	Fisher Scientific	7240-90-6
Glycogen	Roche	10901393007
Protein A Agarose	Millipore	16-157
Protein G Agarose	Millipore	16-266
Illumina Genomic Adaptor Oligo Mix	Illumina	1000521
Illumina Genomic PCR Primers 1.1	Illumina	1000537
Illumina Genomic PCR Primers 2.1	Illumina	1000538
FastStart Universal SYBR Green Master (Rox)	Roche	04913914001
TaqMan® Universal Master Mix II	AB	4427788

2.3 Buffer Recipes

All buffers are prepared using sterile deionised water.

2.3.1 General buffers

TBE Buffer (5X)
1.1 M Tris
900 mM Boric Acid
25 mM EDTA

TAE Buffer
2 M Tris
1 M Glacial Acetic Acid
50 mM EDTA

Glycerol Loading Buffer
30 % Glycerol
0.02 % Bromophenol Blue
0.02 % Xylene Cyanol

2.3.2 Buffers used in oligonucleotide purification

Formamide Loading Buffer
80 % formamide
0.5X TBE
10 mM NaOH

Oligo Elution Buffer
0.5 M NH_4OAc
1 mM EDTA

2.3.3 Buffers used in EMSA probe preparation

Probe Annealing and Storage Buffer (10X)
100 mM NaCl
10 mM Tris pH 7.5
1 mM EDTA

2.3.4 Buffers used in SDS-PAGE

LSB Buffer (2X) (Sample Loading Buffer)
124 mM Tris pH 6.8
4 % SDS
20 % Glycerol
20 mM DTT
0.02 % Bromophenol Blue

Resolving Buffer (4X)
1.5 M Tris pH 8.8
0.4 % SDS

Stacking Buffer (4X)
0.5 M Tris pH 6.8
0.4 % SDS

SDS-PAGE Running Buffer
25 mM Tris
250 mM Glycine
0.1 % SDS

2.3.5 Buffers used in Electrophoretic Mobility Shift Analysis (EMSA)

Binding Buffer (2X)
40 mM HEPES pH 7.9
100 mM KCl
10 % Glycerol

2.3.6 Solutions used in bacterial cell transformation and culture

SOB medium
20 g Bacto-tryptone
5 g Yeast extract
0.584 g NaCl
0.186 g KCl
Dissolve in 950 ml dH_20
Adjust to pH 7.0 with NaOH
Bring to 1000 ml with dH_20 and autoclave
For SOC medium add 1 ml of 2 M glucose to 98 ml of SOB medium

LB medium
20 g LB broth powder
Bring to 800 ml with dH_20 and autoclave

2.3.7 Solutions used in ChIP

7 % Formaldehyde Stock Solution
7 % Formaldehyde
0.1 M NaCl
0.5 mM EGTA
50 mM HEPES pH 8
1 mM EDTA

ChIP Lysis Buffer
0.25 % Triton X-100
10 mM EDTA
0.5 mM EGTA
10 mM Tris pH 8

Nuclei Rinse Buffer
0.2 M NaCl
1 mM EDTA
0.5 mM EGTA
10 mM Tris pH 8

Nuclei Lysis Buffer
50 mM Tris pH 8
10 mM EDTA
0.5 % SDS

X-ChIP Buffer
1.1 % Triton X-100
1.2 mM EDTA
16.7 mM Tris pH 8.1
167 mM NaCl

ChIP Wash Buffer 1
0.1 % SDS
1 % Triton X-100
2 mM EDTA
20 mM Tris pH 8.1
150 mM NaCl

ChIP Wash Buffer 2
0.1 % SDS
1 % Triton X-100
2 mM EDTA
20 mM Tris pH 8.1
500 mM NaCl

ChIP Wash Buffer 3
0.25 M LiCl
1 % NP-40
1 % Sodium Deoxycholate
1 mM EDTA
10 mM Tris pH 8.1

Tris-EDTA Buffer (TE)

10 mM Tris pH 8.1

1 mM EDTA

ChIP Elution Buffer

1 % SDS

0.1 M $NaHCO_3$

2.4 Cell Culture

K562 Cell Line

Wild type K562 cells and the K562 D3/6 cell line (which contains a stably integrated FLAG epitope-tagged, VEZF1 transgene that is expressed at approximately 50 % of endogenous VEZF1 levels. This cell line was made by Dr Dan Li) were cultured in RPMI 1640 growth medium supplemented with 10 % FBS, 1 % penicillin/streptomycin. Immediately following thawing, 1 ml cell stocks were diluted in 10 ml culture medium and centrifuged at 1100 x g for 5 minutes. Supernatant was discarded to remove DMSO and cell pellets were resuspended in 10 ml growth medium and cultured in a T25 flask at 37 °C with 5 % CO_2. Cultures were typically split three times a week at a 1 in 5 ratio when cells reached ~80 % density. Frozen cell line stocks were prepared by aliquoting 1 x 10^6 cells in 1 ml freezing media (90 % FBS, 10 % DMSO) into cryotubes. Tubes were placed in CoolCell® LX or CoolCell® FTS30 boxes and stored at -80 °C overnight before transfer to a liquid nitrogen storage facility.

2.5 Chromatin Immunoprecipitation (ChIP)

Formaldehyde Fixation of Cells

1. Cells were transferred from culture vessels to 50 ml falcon tubes and centrifuged at 500 x g for 5 minutes at room temperature to pellet cells.
2. Supernatant was removed and cell pellets were resuspended in fresh warmed media. An aliquot of cell suspension was used to perform a viable cell count.
3. The cell suspension was adjusted so as to contain 1 x 10^8 cells in 30 ml media.
4. 3.4 ml of a 7 % formaldehyde stock solution was added to each 30 ml cell suspension (0.8 % final concentration) and reactions were incubated on a roller at room temperature for 5 minutes.
5. The cross-linking reaction was stopped by adding 3.5 ml of 1.5 M glycine to each sample (0.175 M final concentration) and incubating on a roller for 5 minutes at room temperature. Following this incubation, samples were always kept on ice.
6. Fixed cells were centrifuged at 300 x g for 5 minutes at 4 °C.
7. Supernatant was removed and cell pellets were resuspended in 30 ml ice-cold PBS.
8. Centrifugation at 300 x g for 5 minutes at 4 °C was repeated.

Preparation of Chromatin

9. Supernatant was removed and cells were lysed by resuspending pellets in 15 ml of ice-cold ChIP Lysis Buffer.
10. Samples were incubated on ice for 10 minutes and were then centrifuged at 700 x g for 5 minutes at 4 °C.
11. Pellets were washed with 15 ml ice-cold Nuclei Rinse Buffer.
12. Samples were centrifuged at 700 x g for 5 minutes at 4 °C.
13. Supernatant was removed and nuclei were resuspended in 1.5 ml Nuclei Lysis Buffer.
14. Samples were incubated on ice for 10 minutes.

Sonication

15. Chromatin samples were transferred to 5 ml round-bottomed FACS tubes and volumes were brought to 2ml with X-ChIP buffer.
16. Sonication was performed using a Misonix automatic sonicator. The microtip was positioned to reach the bottom 3rd of the chromatin sample and the tube was secured in place on ice.
17. Sonication was performed using the following programme:

Amplitude/Output:	4.5
ON:	10 seconds
OFF:	20 seconds
Total sonication time:	5 minutes

18. Sonicated chromatin samples were transferred to 1.5 ml microcentrifuge tubes and centrifuged at 15000 x g for 10 minutes at 4 °C to pellet debris.
19. Supernatants were transferred to 50 ml falcons and brought to 20 ml with X-ChIP buffer (final concentration 5 x 10^6 cells worth of chromatin/ml).
20. Samples were snap frozen in 1 ml aliquots either on dry ice/ethanol or using a Cool Box™ 30 (Biocision) and stored at -80 °C.

Immunopurification of Cross-linked Chromatin

21. 1 ml of chromatin was thawed for each immunopurification reaction. Chromatin was pooled for preclearing.
22. 1 ml of chromatin was precleared with 5 µg non-immune IgG and 100 µl of protein A/G agarose beads. Preclearing reactions were incubated at 4 °C on a rotating wheel for 3 hours.

23. Samples were centrifuged at 800 x g for 1 minute at 4 °C to pellet agarose beads. 200 µl of supernatant was taken to be used as input template for qPCR. The remaining supernatant was aliquoted into 1.7 ml eppendorf tubes at a volume of 1 ml per IP.
24. Specific antibodies were added to each precleared chromatin sample and binding reactions were incubated on a rotating wheel at 4 °C overnight.
25. 50 µl of 50 % washed protein A/G agarose beads were added to each chromatin sample and incubated at 4 °C on a rotating wheel for 3 hours.
26. Samples were centrifuged at 800 x g for 1 minute at 4 °C to pellet agarose beads. Supernatants were discarded.
27. 700 µl of ice-cold X-ChIP buffer was used to transfer beads to a new eppendorf tube.
28. Samples were centrifuged at 800 x g for 1 minute at 4 °C to pellet agarose beads. Supernatants were discarded.
29. Beads were washed with a series of buffers as indicated below. Each wash was incubated for 3 minutes with gentle rotation at room temperature before being centrifuged at 800 x g for 1 minute at 4 °C.
    Washed once using 1 ml of ice-cold Wash Buffer 1.
    Washed once using 1 ml of ice-cold Wash Buffer 2.
    Washed once using 1 ml of ice-cold Wash Buffer 3.
    Washed twice using 1 ml of ice-cold TE buffer.

<u>Elution and Reverse Cross-linking</u>

30. Beads were resuspended in 200 µl ChIP Elution Buffer and transferred to a new eppendorf. Samples were then incubated in ChIP Elution Buffer at room temperature for 15 minutes with occasional agitation.
31. Samples were centrifuged at 800 x g for 1 minute at room temperature and eluates were transferred to new eppendorfs.
32. Steps 30 and 31 were repeated so that both eluates from each IP were combined.
33. Cross-links were reversed by adding 18 µl of 5 M NaCl to each 400 µl sample and incubating overnight at 65 °C.
34. 1 µl of RNase A was added to each sample and incubated at 37 °C for 1 hour.
35. 1 µl of Proteinase K was added to each sample and incubated at 45 °C for 2 – 4 hours.

DNA purification

36. 500 µl of phenol-chloroform was added to each IP and samples were mixed by vortexing briefly before centrifuging at max speed (16000 x g?)for 5 minutes at 4 °C.
37. The aqueous phase of each sample was transferred to a new eppendorf tube and 500 µl of chloroform was added to each. Samples were vortexed briefly and centrifuged at 16000 x g for 5 minutes at 4 °C.
38. The aqueous phase of each sample was transferred to a new eppendorf tube and DNA was precipitated by adding 1 µl of glycogen and 750 µl of 100 % ethanol.
39. Samples were vortexed to mix and incubated at room temperature for 30 minutes.
40. Precipitated DNA was pelleted by centrifuging samples at max speed for 45 minutes at room temperature. Supernatants were discarded.
41. DNA pellets were washed with750 µl of 80 % ethanol and centrifuged at max speed for 15 minutes at room temperature. Supernatant was discarded.
42. DNA pellets were air dried and resuspended in 50 µl of 10 mM tris pH8.
43. ChIP DNA samples were stored at -20 °C.

2.6 Preparation of ChIP-seq DNA libraries

In order for ChIP DNA to be analysed using the Illumina deep sequencing platform, ChIP-seq DNA libraries were constructed. Library preparation consisted of end repair, dA-tailing, adapter ligation, PCR amplification and size selection of DNA fragments as outlined below.

End Repair

The end repair reaction was assembled as follows:

ChIP DNA	30 µl
10X Phosphorylation Reaction Buffer	5 µl
dNTPs (10 mM)	2 µl
T4 DNA Polymerase	1 µl
Klenow DNA Polymerase(diluted 1:5)	1 µl
T4 PNK	1 µl
H_20	10 µl
Total volume	50 µl

Reactions were incubated at 20 °C for 30 minutes in a thermocycler.

The QIAquick PCR Purification Kit (Qiagen) was used to purify blunt-ended fragments according to the manufacturers' instructions. DNA was eluted in 34 µl of elution buffer.

dA-tailing

The dA-tailing reaction was assembled as follows:

End-repaired DNA	34 µl
dATP (1 mM)	10 µl
Klenow Fragment Buffer	5 µl
Klenow Fragment (3'→5' exo$^-$)	1 µl
Total volume	50 µl

Reactions were incubated at 37 °C for 30 minutes in a thermocycler.

The MinElute PCR Purification Kit (Qiagen) was used to purify dA-tailed fragments according to the manufacturers' instructions. DNA was eluted in 10µl of elution buffer.

Adapter Ligation

The adapter ligation reaction was assembled as follows:

Illumina Adapter Oligo Mix (diluted 1:10)	1 µl
2X Quick Ligation Reaction Buffer	15 µl
Quick T4 DNA ligase	4 µl
Total volume	20 µl

Reactions were incubated at room temperature for 15 minutes.

Unligated adapters were removed using the QIAquick PCR Purification Kit according to the manufacturer's instructions. Adapter-ligated DNA fragments were eluted in 36 µl of elution buffer.

PCR amplification of adapter-ligated ChIP DNA

The PCR reaction was assembled as follows:

5X Phusion High-Fidelity Buffer	10 µl
dNTPs (10 mM)	1.5 µl
Illumina PCR Primer 1.1	1 µl
Illumina PCR Primer 2.1	1 µ
Phusion DNA Polymerase	0.5 µl
Total volume	14 µl

PCR was performed using a thermocycler programmed to perform the following temperature cycles:

Segment	Cycles	Temperature	Time
1	1	98 °C	30 seconds
2	18	98 °C	10 seconds
		65 °C	1 minute
		72 °C	30 seconds
3	1	72 °C	5 minutes

The MinElute PCR Purification Kit (Qiagen) was used to purify PCR products according to the manufacturers' instructions. DNA was eluted in 10µl of elution buffer.

Agarose gel size selection of amplified libraries

PCR amplified ChIP-seq libraries were electrophoresed on a 2 % agarose 1X TAE gel. One well was left blank in between samples so as to avoid contamination between libraries. A 1 Kb DNA ladder (NEB) was also run on the gel to provide a marker for size selection. Gels were electrophoresed at 60 – 70 volts for 2 hours and were then stained by adding 50 µl ethidium bromide to 1X TAE running buffer and incubating on an orbital shaker for 30 minutes. DNA was visualised using a transilluminator and fragments of 200 – 300 bp were excised from the gel using a sterile scalpel. DNA was purified from agarose gel slices using the QIAquick Gel Extraction Kit (Qiagen) according to manufacturer's instructions. ChIP-seq library DNA fragments were eluted in 30 µl elution buffer.

2.7 High throughput sequencing analysis

ChIP-seq libraries were single end sequenced to 76 bp on an Illumina Genome Analyzer IIx at the Glasgow Polyomics Facility by Julie Galbraith. Sequences were converted into fastq format using CASAVA v1.8.2 software (Illumina) by Dr. Pawel Herzyk, Polyomics Facility. The performance of all sequencing runs was checked using the FastQC software (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc) as described in section 3.2.4. Sequence reads from VEZF1 and non-immune IgG ChIP-seq from K562 cells were aligned to the hg19 reference genome using Bowtie v0.12.3 software (Langmead *et al*., 2009) by Dr. Andrew Crossan, Vetrie group, ICS, University of Glasgow using the settings --solexa1.3-quals -p 6 -M 3 -3 40 –t. Alignments were formatted for the UCSC genome browser, as described in section 3.2.4, by Andy Crossan. Sequence reads from VEZF1

ChIP-seq from 5 and 10 day chicken embryonic erythrocytes were aligned to the gg4 reference genome using Bowtie v0.12.7 software by Dr. Tony McBryan, Adams group, ICS, University of Glasgow using the settings -m 3 --phred33-quals --chunkmbs 512 -p 16 --best –t. These alignments were formatted for the UCSC genome browser as described in section 3.2.4 and shown in chapter 6, by Tony McBryan.

2.8 ChIP-seq peak finding

VEZF1 ChIP-seq peaks were identified using MACS v1.4 software (Zhang *et al*., 2008) using the settings -t VEZF1.bed -c IgG.bed -n RS_macs_s25 -f BED -g hs -s 25 --bw=300 --mfold=10,30 --shiftsize=150 --keep-dup=1 by Dr Ruslan Strogantsev, University of Cambridge. Clustered ChIP-seq peaks were resolved into discrete signal peaks using PeakSplitter (http://www.ebi.ac.uk/bertone/software.html). Peak summits from MACS and PeakSplitter were merged. For summits closer than 100bp apart, the one with highest read count was retained. If there was an identical read count in these cases, the midpoint coordinate was taken to deal with slightly shifted summits.

VEZF1 ChIP-seq peaks were ranked by read enrichment levels relative to IgG control ChIP-seq. The number of VEZF1 and IgG ChIP-seq reads overlapping 50 bp either side of each summit was determined using Seqmonk software (http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/). These values were normalised for total sequence count (expressed as reads per million). The IgG rpm scores were subtracted from the VEZF1 rpm scores to allow ranking of peaks by enrichment over control.

2.9 Annotation of chromatin features at VEZF1 peaks

The gene and chromatin state annotation of VEZF1 peaks in K562 cells is described in chapter 4. The enrichment of ChIP-seq reads at VEZF1 peaks or TSS was profiled using seqMINER software (Ye *et al*., 2011). The locations of hg19 TSS were taken from the annotation included with seqMINER. The ENCODE project (Dunham *et al*., 2012) datasets analysed in this study are listed in Table 2.1.

Dataset	ENCODE reference number
K562 ChromHMM chromatin state track	wgEncodeEH000790
H3K4me1	wgEncodeEH000046
H3K4me3	wgEncodeEH000048
H3K27ac	wgEncodeEH000043
H3K27me3	wgEncodeEH000044
H3K9me3	wgEncodeEH001040
H3K36me3	wgEncodeEH000045
H2A.Z	wgEncodeEH001038
DNase-seq	wgEncodeEH000484
FAIRE-seq	wgEncodeEH000531
RNA-seq	wgEncodeEH000484
RNA Polymerase II (unphosphorylated CTD)	wgEncodeEH000727
SP1	wgEncodeEH001578
YY1	wgEncodeEH001623
ELF1	wgEncodeEH001619
EGR1	wgEncodeEH001646
HMGN3	wgEncodeEH001863
TAF1	wgEncodeEH001582
TBP	wgEncodeEH001825
CMYC	wgEncodeEH002800
TAL1	wgEncodeEH001824
GATA1	wgEncodeEH000638
BRG1	wgEncodeEH000724
P300	wgEncodeEH002834
NFE2	wgEncodeEH000624
CEBPβ	wgEncodeEH002346
BACH1	wgEncodeEH002846

Table 2.1 The ENCODE project datasets analysed in this study.

2.10 Binding site motif discovery

The preferred DNA binding specificity of VEZF1 was firstly derived from a probabilistic model (http://compbio.cs.huji.ac.il/Zinc/, (Kaplan *et al*., 2005)). DNA sequence motifs of up to 30 bp in length enriched within the 100 bp surrounding VEZF1 ChIP-seq peak summits were determined using the MEME web portal (http://meme.nbcr.net/meme/, (Machanick and Bailey, 2011) as described in section 5.3.1. DNA sequence motifs of 8 or 9 bp in length enriched within the 200 bp surrounding VEZF1 ChIP-seq peak summits were determined using the POSMO web portal (http://cb.utdallas.edu/Posmo/index.html, (Ma *et al*., 2012).

2.11 Incubation of fertilised eggs and isolation of erythrocytes from chicken embryos

Fertilised chicken eggs were a kind gift from Aviagen, Midlothian, UK. Eggs were stored at 13 °C, 80 % humidity overnight. Eggs were incubated at 37.6 °C in an Octagon 400X incubator (Brinsea UK) for 5 – 10 days. Eggs were inverted once per day.

Circulating erythrocytes were isolated from chicken embryos as follows:

1. An egg was placed apical end down in an egg tray.
2. The egg shell above the air pocket was cracked by tapping with the blunt end of a scalpel and a disc of shell was removed using fine forceps.
3. The membrane was cut with scissors and peeled back using forceps.
4. Some of the liquid was discarded by tilting the egg and allowing it to run out into a petri dish.
5. A peripheral artery was nicked using scissors and blood was collected using a P200. If the yolk sac broke the egg was discarded.
6. Blood was transferred into a 50 ml falcon tube containing 3 ml cold PBS and stored on ice.
7. Once bleeding was complete, the embryo was tipped into a petri dish and decapitated.

2.12 Bisulphite Sequencing

2.12.1 Bisulphite Conversion of DNA

Bisulphite conversion of DNA was performed using the EZ DNA Methylation-Direct Kit (Zymo Research, Cat. No. D5020). Manufacturers' instructions were followed when performing this technique.

1. 1 x 10^5 cells were isolated and resuspended in 12 μl PBS.
2. DNA was isolated by assembling the following reaction:

2X M-Digestion buffer	13 μl
Sample (1 x 10^5 cells)	12 μl
Proteinase K	1 μl

3. Reactions were incubated at 50 °C for 20 minutes.
4. Samples were mixed thoroughly and centrifuged at 10 000 x g for 5 minutes. 20 μl of supernatant was added to 130 μl of CT conversion reagent
5. Reactions were placed in a thermocycler and the following programme run.

Temperature	Time
98 °C	8 minutes
64 °C	3.5 hours
4 °C	Up to 20 hours

6. DNA was purified by spin column. DNA bound to the column filter underwent several washes and a desulphonation step before elution in 10 µl elution buffer.

2.12.2 PCR Amplification from Bisulphite Converted DNA

PCR primers for amplification of genomic regions of interest from bisulphite modified DNA were designed using the MethPrimer online tool (Li LC and Dahiya R. MethPrimer: designing primers for methylation PCRs. Bioinformatics. 2002 Nov;18(11):1427-31. PMID: 12424112). PCR reactions for amplification from bisulfite modified DNA were set up as follows:

Reaction Buffer	2.5 µl
50 mM $MgCl_2$	1.5 µl
Forward Primer (10 mM)	0.5 µl
Reverse Primer (10 mM)	0.5 µl
dNTPs (10 mM)	1 µl
Platinum Taq DNA Polymerase	0.2 µl
dH_20	15.8 µl
Bisulphite Modified DNA	3 µl
TOTAL	25 µl

PCR was performed using a thermocycler programmed to perform the following temperature cycles:

Segment	Cycles	Temperature	Time
1	1	95 °C	5 minutes
2	34	95 °C	1 minute
		55 °C	1 minute
		72 °C	2 minutes
3	1	72 °C	10 minutes

Once complete, the PCR reactions were loaded onto a 1 % agarose gel stained with Safeview (Company) and electrophoresed at 120 volts for 35 minutes.
DNA bands of the desired size were excised from the gel using a scalpel and DNA was purified from agarose fragments using the QIAquick Gel Extraction Kit (Qiagen, cat. no. 28704) according to manufacturers' instructions.

2.12.3 Ligating Bisulphite Converted DNA into pGEMT Easy Vector

Bisulphite converted PCR products were ligated into the pGEMT Easy vector by TA-cloning using the pGEMT Easy Vector System I (Promega, cat. no. A1360).

2X Rapid Ligation Buffer	5 μl
pGEMT Easy Vector (50 ng)	1 μl
PCR product	3 μl
T4 DNA Ligase	1 μl
TOTAL	10 μl

Reactions were mixed and incubated at 4 °C overnight.

2.12.4 Transformation of competent *E. Coli*

Subcloning efficiency DH5-α *E. coli* were transformed with ligations as follows:

1. Competent DH5- α cells were thawed on ice and mixed by gently flicking the tube.
2. 50 μl of competent cells were aliquoted into individual eppendorf tubes per transformation and kept on ice.
3. 2 μl of ligation reaction was added per tube and mixed by gently flicking the tube.
4. A positive control transformation using 2.5 μl (250 pg) pUC19 plasmid was performed alongside every transformation.
5. Reactions were incubated on ice for 30 minutes.
6. Cells were heat-shocked in a water bath at 42 °C for 20 seconds.
7. Reactions were incubated on ice for 2 minutes.
8. 950 μl of room temperature SOC was added to each reaction and tubes were transferred to a 37 °C bacterial shaker set to 225 rpm for 1 hour.
9. 50 μl and 200 μl of transformation reactions were spread on agar plates containing 50 μg/ml ampicillin for selection of transformed bacteria. These plates

had been coated with 40 μl of 20 mg/ml X-Gal to allow blue/white colony screening, and pre-warmed by incubating at 37 °C for 1 hour.

10. Plates were incubated at 37 °C overnight to allow colony formation and were subsequently stored at 4 °C.

2.12.5 Screening colonies for presence of insert

The multiple cloning region of the pGEMT Easy Vector lies within the α-peptide coding region of the β-galactosidase enzyme. β-galactosidase cleaves X-Gal producing a blue colour that can be easily detected by eye, thus if the pGEMT Easy Vector self-ligates and the β-galactosidase coding region is intact, active enzyme will be produced by transformed DH5-α cells and bacterial colonies will appear blue. If PCR product is successfully ligated into the pGEMT Easy Vector, the β-galactosidase coding region is interrupted such that the enzyme cannot be expressed by transformed bacterial cells and colonies will appear white. Therefore, blue/white screening of bacterial colonies was performed in order to differentiate between bacterial colonies containing DNA inserts of interest and those containing self-ligated plasmid.

Colony PCR

To ensure that the white bacterial colonies selected did contain DNA inserts of interest, colony PCR was performed.

1. A pipette tip was used to pick white bacterial colonies from X-Gal-coated agar plates.

2. The bacterial colony on the pipette tip was submerged in 50 μl dH_20 in an eppendorf tube and mixed.

3. The tip was then streaked over the surface of a pre-warmed agar plate containing 50 μg/ml ampicillin, this streak plate was incubated at 37 °C overnight to allow colony formation.

4. Colony PCR reactions were set up as follows:

10X reaction buffer	2.5 μl
Forward primer (10 mM)	0.5 μl
Reverse primer (10 mM)	0.5 μl
dNTPs (10 mM)	1 μl
Taq DNA polymerase	0.2 μl
dH_20	15.3 μl
Template (bacterial colony in H_20)	5 μl
TOTAL	25 μl

5. PCR was performed using a thermocycler programmed to perform the following temperature cycles:

Segment	Cycles	Temperature	Time
1	1	94 °C	2 minutes
2	34	94 °C	30 seconds
		55 °C	30 seconds
		72 °C	1 minute
3	1	72 °C	10 minutes

6. Once complete, the PCR products were analysed on a 1 % agarose gel stained with Safeview.

2.12.6 Purification of plasmid DNA

1. A small sample of bacterial cells from the streak plate was transferred to 2 ml of LB + 50 μg/ml ampicillin in a round bottomed falcon tube using a sterile pipette tip. This culture was incubated in a 37 °C bacterial shaker set to 225 rpm for 16 hours.
2. ~1.4 ml of overnight bacterial culture was transferred to a 1.5 ml eppendorf and centrifuged at 13000 rpm for 5 minutes to pellet bacterial cells.
3. The supernatant was removed and plasmid DNA purified from the bacterial cell pellet using the QIAprep Miniprep kit (Qiagen, cat. no. 27106) according to manufacturers' instructions.
4. The concentration of eluted DNA was measured by nanodrop.

2.12.7 Sequencing of bisulphite modified DNA

Purified plasmid DNA was sent to Source Bioscience for Sanger Sequencing of bisulphite modified DNA inserts.

- 20 μl of 100 ng/μl plasmid DNA was sent for sequencing.
- Sequencing was performed using the M13 reverse primer for all samples.
- Upon receipt of complete sequencing files, DNA sequences were reverse complemented using an online tool and aligned to the original DNA target sequence in a Microsoft Word document.
- Sequences were checked for complete conversion of non-CpG-associated cytosines to thymine.
- The methylation state of individual CpG-associated cytosine residues was then analysed.

2.13 RNA Preparation

2.13.1 RNA extraction from cell pellets

1. Cell pellets containing 5 – 10 x 10^6 cells were resuspended in 1 ml trizol (Invitrogen) by pipetting. A 'reagents only' extraction control was also set up, i.e. 1 ml trizol without cells.
2. Samples resuspended in trizol were incubated at room temperature for 5 minutes.
3. 100 μl of 1-Bromo-3-chloropropane (BCP) (Sigma) per 1 ml of trizol was added to each sample.
4. Tubes were shaken vigorously for 30 seconds, then incubated at room temperature for 15 minutes.
5. Samples were centrifuged at 12000 x g for 15 minutes at 4 °C.
6. 500 μl of isopropanol was added to a set of fresh eppendorfs.
7. The aqueous phase of each sample was transferred to a fresh eppendorf containing isopropanol, tubes were then vortexed and incubated at room temperature for 10 minutes.
8. Samples were incubated at -20 °C for a minimum 20 minutes.
9. Samples were centrifuged at 12000 x g for 15 minutes at 4 °C.
10. Supernatent was poured off as waste and tubes were touched to a clean paper towel.

11. RNA pellets were washed with 1 ml of 75 % EtOH and centrifuged at 7500 x g for 5 minutes at 4 °C.
12. Supernatent was poured off.
13. Steps 11 and 12 were repeated.
14. Samples were pulse centrifuged at 12000 x g and as much of the remaining liquid as possible was removed without disturbing the RNA pellet.
15. Tubes were left open and covered with a clean paper towel to allow RNA pellets to air dry.
16. RNA pellets were resuspended in 50 µl sterile water.
17. RNA yield was quantified using a NanoDrop.
18. Samples were stored at -80 °C.

2.13.2 RNA quality checks by agarose gel electrophoresis

1. RNA sample quality was analysed by loading 1 or 2 µg of RNA preparation per well of a 1 % agarose-TBE gel. These gels contained no nucleic acid stains.
2. Gels were run at 120 volts.
3. Ethidium bromide (Sigma) was then added to 1 x TBE running buffer at a 1 in 10000 dilution and the gel was incubated in this buffer for 30 minutes on an orbital shaker.
4. Gels were transferred to ethidium bromide-free running buffer and incubated for 10 minutes on an orbital shaker to de-stain.
5. Gels were then imaged using a Fujifilm FLA-5000 scanner.

2.14 cDNA synthesis

cDNA synthesis was performed using the Invitrogen Superscript III kit (cat #)

1. The following were combined in PCR tubes:

RNA (100 ng/µ)	4 µl
dNTPs (10 mM)	0.5 µl
Random primers (50 ng/µl)	1 µl
dH_20	1 µl
TOTAL	6.5 µl

2. Reactions were incubated at 65 °C for 5 minutes, then transferred to ice for at least 2 minutes.

3. A master mix was prepared (as shown below per reaction) of which 3.5 µl was added to each cDNA synthesis reaction.

5 x First Strand Synthesis (FS) Buffer	2 µl
RNase OUT	0.5 µl
DTT	0.5 µl
Superscript III Reverse Transcriptase	0.5 µl
TOTAL	3.5 µl
Final Volume	10 µl

4. cDNA synthesis was performed using a thermocycler programmed to perform the following temperature cycles:

Temperature	**Time**
25 °C	10 minutes
50 °C	50 minutes
85 °C	5 minutes

5. Template RNA was digested by adding 1 µl of RNase H (NEB) to each reaction and incubating at 37 °C for 20 minutes.
6. cDNA samples were then diluted with RNase- and DNase-free H_20 and either used immediately for qPCR or stored at -20 °C.

2.15 Real-time PCR

Two real-time PCR methods were used to generate the work presented and are described below.

2.15.1 SYBR® Green real-time PCR

SYBR® green is a nucleic acid stain which binds preferentially to double stranded DNA and emits a fluorescence signal. Thus as PCR products accumulate with each amplification cycle, increased SYBR Green® fluorescence signal is detected. SYBR® Green real-time PCR was used in the analysis of gene expression from diluted cDNA templates.

SYBR® Green PCR reactions were set up as follows:

2X SYBR® Green Master Mix	10 µl
Forward and Reverse Primer Mix (4 or 4.5 µM per primer)	4 µl
cDNA template	4 µl
dH_20	2 µl
TOTAL	20 µl

SYBR® Green real-time quantitative PCR reactions were run on a Stratagene Mx3000P thermocycler programmed to perform the following temperature cycles:

Segment	Cycles	Temperature	Time
1	1	95 °C	10 minutes
2	40	95 °C	15 seconds
		60 °C	30 seconds

A dissociation curve was performed upon completion of each real-time PCR in order to measure the dissociation temperature of PCR products between 60 °C – 95 °C.
Ct values were generated using the MxPro version 4.0 software.

2.15.2 Taqman® real-time PCR

Taqman® real-time PCR is designed to have a higher specificity for the target amplicon than the SYBR® Green method. In addition to forward and reverse primers, Taqman® PCR also utilises a probe oligonucleotide which is designed to bind a target DNA sequence between the forward and reverse primer binding sites. This probe is conjugated to a FAM fluorophore at its 5' end and a TAMRA quencher at its 3' end. The close proximity of these two molecules results in the quencher absorbing the emission energy of the fluorophore, thus preventing the detection of any fluorescence signal. Extension of the forward primer across the Taqman probe binding site during PCR results in degradation of the bound probe due to the 5' -> 3' exonuclease activity of the Taq polymerase enzyme used in Taqman® PCR. Degradation of the Taqman probe separates the associated fluorphore and quencher molecules. The result of this is that the emission energy of the fluorophore is no longer masked by the quencher therefore a fluorescence signal which is proportional to the amount of specific amplicon product can be detected. Taqman® real-time PCR was used in the analysis of gene expression from diluted cDNA templates and in analyses following ChIP and hMeDIP assays.

Taqman® PCR reactions were set up as follows:

2X Taqman® Master Mix	10 µl
Forward and Reverse Primer Mix (1.5 or 4.5 µM per primer)	4 µl
Taqman Probe (1.5 – 2 µM)	2 µl
DNA template	4 µl
TOTAL	20 µl

Taqman® real-time quantitative PCR reactions were run on a Roche LC480 thermocycler programmed to perform the following temperature cycles:

Segment	Cycles	Temperature	Time
1	1	95 °C	10 minutes
2	45	95 °C	30 seconds
		60 °C	1 minute
3	1	40 °C	10 seconds

Ct values were generated using the Light Cycler® 480 software.

2.15.3 Real-time PCR primer design

The following guidelines were utilised in the design of primer sets for real-time PCR:

- Primer Tm's should be 60 °C +/- 2 °C
- The G/C content of primers should be 40 – 60 %
- Primers are 18 – 28 bp in length (optimal length 20 bp)
- Target amplicon size is 50 – 150 bp
- Either the forward or reverse primer of a primer pair should span an exon-exon junction. This should prevent amplification of contaminating genomic DNA that may be present in cDNA samples.
- Taqman probes were designed such that their annealing temperatures were 10 °C higher than those of the forward and reverse primers.

The BLAST similarity search tool was used to check if designed primer sequences aligned to multiple sites in the genome.

2.15.4 Validation and optimisation of real-time PCR primers

Several of the primer sets used had previously been validated and are widely used in the lab. qPCR using newly designed primer sets was optimised using various ratios of final primer concentration per reaction (shown below):

Reverse	**Forward**		
	50 nM	**300 nM**	**900 nM**
50 nM	50/50	300/50	900/50
300 nM	50/300	300/300	900/300
900 nM	50/900	300/900	900/900

Following amplification, PCR products were analysed by agarose gel electrophoresis to ensure only one PCR product was generated.

Optimal primer concentrations were chosen based on Ct values, amplification plots, dissociation plots and agarose gel images.

Standard curves were then generated for each primer set at their chosen concentration.

2.15.5 Real-time PCR analysis

1. Gene expression in protein knock down experiments was analysed as follows:

- Expression of the gene of interest was normalised to expression of a housekeeping gene to generate a ΔCt value:

  ΔCt = Ct (gene of interest) – Ct (housekeeping gene)
- A ΔΔCt value was then calculated by subtracting the ΔCt of the gene of interest in wild type cells from the ΔCt of the gene of interest in KD cells:

  ΔΔCt = ΔCt (treated) - ΔCt (wild type)
- ΔΔCt values were then used to calculate fold change in gene expression using the following formula:

  Fold change = N^(-ΔΔCt); where 'N' = PCR amplification efficiency (typically = 2).

2. Gene expression of endogenous proteins in chicken circulating erythrocytes was analysed as follows:

- A ΔCt value was generated by normalising expression of a gene of interest to that of a housekeeping gene:

ΔCt = Ct (gene of interest) – Ct (housekeeping gene)

- Fold expression of a gene of interest relative to housekeeping gene expression was calculated using the following formula:
  Fold expression = N^(-ΔCt); where 'N' = PCR amplification efficiency.

2.16 Electrophoretic Mobility Shift Analysis (EMSA)

2.16.1 Oligonucleotide Purification

Gel Preparation

1. Glass plates were separated using 1.5 mm spacers and held in place with bulldog clips, the base of the glass plates was sealed using masking tape. 50 ml gel casting solution was prepared per gel using the SequaGel UreaGel system (National Diagnostics), as indicated below, with APS being added last. This solution was carefully pipetted into the gel casting construction and a 10 well plastic comb was inserted. The final acrylamide concentration of this preparation is 8 % and was routinely used for all oligonucleotide purifications.

Sequagel Concentrate	16 ml
Sequagel Diluent	29 ml
Sequagel Buffer	5 ml
TEMED	20 µl
10 % APS (adenosine phosphosulfate)	400 µl

2. Once gel polymerisation was complete, the comb and tape at the base of the gel were removed and the cassette was fastened into sandwich clamps and fixed into the central cooling unit of the Protean II xi XL Vertical Electrophoresis cell (Bio-Rad).
3. 5 litres of 0.5X TBE was heated to 55 – 60^{0}C and ~3 litres were poured into the electrophoresis tank, the cooling unit containing the acrylamide gel was then placed inside the tank and the upper buffer reservoir was filled with 0.5X TBE to completely submerge the electrode.
4. A syringe was used to wash out the wells with 0.5X TBE.
5. The gel was pre-run at 400 volts for 15 – 30 minutes.

Oligonucleotide Preparation

6. 100 µl of formamide loading buffer was added to each tube of lypholised oligonucleotide. Tubes were centrifuged briefly, sat at room temperature while the gel pre-ran, and mixed by pipetting prior to loading gel.

Electrophoresis of Oligonucleotides

7. The gel pre-run was stopped and 50 µl of resuspended oligonucleotide was loaded per well
8. 30 µl of 30 % glycerol loading buffer containing bromophenol blue and xylene cyanol dyes was loaded in the far left hand lane to orient gels and monitor electrophoresis.
9. Gels were run at 400 volts for ~ 50 minutes, until the bromophenol blue dye reached the base of the gel.

Purification of Oligonucleotides

10. Following electrophoresis, the gel was transferred from the glass plates to a sheet of clingfilm which was placed on top of a phosphorimaging screen.
11. In a dark room, oligonucleotide bands were visualised by UV shadowing, i.e. a UV lamp was shone over the gel and phosporimaging screen allowing the visualisation of DNA as dark purple bands.
12. Visible bands were excised from gels using a clean scalpel. Roughly the bottom quarter of each DNA band was not removed from the gel in order to ensure the exclusion of prematurely terminated oligonucleotides.
13. Excised gel bands were transferred to a 1.5 ml eppendorf (whose tip had been removed) which sat inside a second intact 1.5 ml eppendorf.
14. Double eppendorfs were placed in the inner ring of a Sigma 1-14 microfuge and pulse centrifuged for ~7 seconds. Gel fragments appeared crushed in the bottom eppendorf.
15. 400 µl of elution buffer was added to each crushed gel fragment and mixed by vortexing briefly
16. Samples were incubated at 37 °C overnight to elute oligonucleotides from the gel.
17. Filter columns (Millipore cat. no. 30HV00) were prepared by adding 400 µl of 1X TE buffer to the column, pulse centrifuging for ~10 seconds and removing the flow through from the collection tube.

18. Eluted oligonucleotide/crushed acrylamide gel mix was transferred to the filter column and centrifuged at 2000 x g for 1 minute.
19. The volume of eluted DNA was measured and transferred to a fresh eppendorf.
20. 3 x volumes of ethanol and the volume of 3 M NaAc required for a 0.3 M final concentration were calculated. 3 M NaAc was added to samples first, followed by 100 % ethanol.
21. Tube contents were mixed by inverting and sat on ice for 1 hour.
22. Samples were centrifuges at 14,000 x g, 4 °C, for 30 minutes.
23. The supernatant was poured/pipetted off and DNA pellets were washed with 100 µl of 90 % ethanol. Samples were centrifuged again as in step 22.
24. The supernatant was poured/pipetted off and DNA pellets were air dried.
25. Purified oligonucleotides were resuspended in 100 µl of 1 M tris pH8.
26. DNA concentrations were measured using a NanoDrop ND1000 spectrophotometer (Thermo).

2.16.2 EMSA Probe Generation

End-labelling top strand oligo with γ-^{32}P ATP

1. End labelling reactions were set up as indicated below.

Top strand oligo (5 pmol/µl)	2 µl
10X PNK buffer (NEB)	1.5 µl
γ-^{32}P ATP	5 µl
T4 PNK (NEB)	0.5 µl
H_20	6 µl
Total	15 µl

2. Reactions were incubated at 37 °C for 45 minutes, followed by 90 °C for 5 minutes.

Annealing of complementary oligonucleotides

3. 30 pmol (6µl) of unlabelled bottom strand oligo was added to the γ-^{32}P ATP-labelled top strand.
4. 5 µl of 10X Probe Annealing and Storage Buffer was added and reactions made to 50 µl with H_20.
5. Reactions were placed in a thermocycler and the following annealing programme was run.

Temperature	Time
90 °C	3 minutes
65 °C	10 minutes
37 °C	10 minutes
22 °C	10 minutes
4 °C	hold

Removal of Excess γ-^{32}P ATP

6. Sepharose within Roche G25 Spin Columns (cat. no. 11273949001) was resuspended by inverting and buffer drained according to the manufacturer's instructions.
7. Columns were centrifuged twice at 1100 x g for 2 minutes at 4 °C, flow through was emptied from the collection tube.
8. The radiolabelled probe sample was loaded onto the sepharose column and eluted into a screw cap tube by centrifuging at 1100 x g for 4 minutes at 4 °C.
9. Excess unincorporated γ-^{32}P ATP nucleotides are retained in the sepharose column.
10. Radiolabelled probe DNA was stored at 4 °C.

2.16.3 *In vitro* transcription of transcription factor gene coding sequences

In vitro transcription was performed using the mMessage mMachine T7 and T3 *in vitro* transcription kit (Ambion, cat. no's AM1344 and AM1348 respectively) according to the manufacturers' instructions. Briefly, reactions were assembled as below.

NTP/CAP	10 μl
10x reaction buffer	2 μl
Linear template DNA	1 μg
T7 enzyme mix	2 μl
RNasin	0.5 μl

Make to 20 μl with H_20

Reactions were mixed thoroughly by pipetting and incubated at 37 °C for 2 hours.

2.16.4 *In vitro* translation of transcription factor proteins

In vitro translation of transcription factor proteins

1. *In vitro* translation was performed using the Rabbit Reticulocyte Lysate System (Promega, product number L4960) according to manufacturers' instructions. For each protein translation two *in vitro* translation reactions were assembled, one incorporating ^{35}S methionine into the synthesised protein to allow analysis of the size of protein product generated via polyacrylamide gel electrophoresis and phosphorimaging, the second synthesising non-radiolabelled cold protein for use in EMSA. *In vitro* translation reactions were assembled as follows:

	Radiolabelled IVT	Non-radiolabelled IVT
RNA substrate	1 µl	2 µl
H_20	4.5 µl	11 µl

Samples were incubated at 65 °C for 5 minutes to denature RNA secondary structure

RRL	17.5 µl	35 µl
RNasin Ribonuclease inhibitor (Promega)	0.5 µl	1 µl
Amino Acid Mix methionine$^-$ (1 mM)	0.5 µl	1 µl
Amino Acid Mixture leucine$^-$ (1 mM)	- -	1 µl
[^{35}S] Methionine	2 µl	--

In vitro translation reactions were incubated at 30 °C for 90 minutes and subsequently stored at -80 °C.

Preparation of Protein Samples for analysis by SDS-PAGE

Radiolabelled *in vitro* translation samples were prepared for analysis by setting up reactions as described below.

Radiolabelled *in vitro* translation reaction	2 µl
2X LSB	5 µl
H_20	3 µl

Samples were mixed and incubated at 100 °C for 5 minutes, centrifuged briefly and used immediately in PAGE.

2.16.5 Denaturing Polyacrylamide Gel Electrophoresis (SDS-PAGE)

Gel Preparation

1. Resolving gel casting solutions were prepared as shown below with APS being added last. ~7.5 ml was pipetted carefully into gel casting cassettes. The final acrylamide concentration of this solution is 10 % and was routinely used for all experiments. The solution was overlaid with ~1 ml isopropanol and left to polymerise for 30 minutes at room temperature.

40 % acrylamide (37.5 acrylamide:1 bis-acrylamide)	5 ml
4X resolving buffer	5 ml
H_20	9.8 ml
TEMED	10 µl
10 % APS	100 µl

2. The isopropanol layer was poured off and residual isopropanol removed by washing with H_20. Stacking gel casting solutions were prepared as shown below with APS being added last. ~3 ml was pipetted on top of the resolving gel and a 10 well plastic comb inserted. The stacking gel was then left to polymerise for 30 minutes at room temperature.

40 % acrylamide (37.5 acrylamide:1 bis-acrylamide)	1.25 ml
4X stacking buffer	2.5 ml
H_20	6.25 ml
TEMED	10 µl
10 % APS	60 µl

Electrophoresis of Protein Samples

3. Once polymerisation of the stacking gel was complete, the plastic comb and tape at the bottom of the cassette were removed and the wells were washed with H_20 .
4. Gel cassettes were fitted into the gel tanks, locked in place and submerged in 1X SDS-PAGE running buffer.
5. Denatured protein samples were loaded into individual wells along with 5 µl of PAGEruler Prestained Protein Ladder.
6. Gels were run at 200 volts for 120 minutes.

Gel drying and imaging

7. Once electrophoresis was complete, gel cassettes were opened and the stacking gel and resolving gel foot were removed.
8. A piece of Whattman filter paper was placed on top of the gel and was peeled away from the remaining half of the gel cassette.
9. The gel was then place on top of a further two pieces of Whattman filter paper and overlaid with clingfilm before being placed in the drying unit of a gel dryer which had been pre-heated to 80 °C.
10. The gel was left in the drying unit for ~30 minutes until completely dry and was then exposed to a phosphorimaging screen overnight.
11. The phoshporimaging screen was scanned using a Fuji scanner the following day allowing visualisation of radiolabelled protein bands.

2.16.6 Electrophoretic Mobility Shift Analysis (EMSA)

Gel Preparation

1. Glass plates were separated using 1 mm spacers, positioned at both edges and the base of the plates, and held in place with bulldog clips. Gel casting solution was prepared as indicated below, with APS being added last. ~ 20 ml of this solution was carefully pipetted into each gel casting unit and a 20 well plastic comb was inserted. Gels were then left to polymerise for 1 hour at room temperature. The final acrylamide concentration of these gels was 5 % and was routinely used for all EMSAs.

40 % Acrylamide (29 Acrylamide:1 Bis-acrylamide)	6.25 ml
5X TBE	10 ml
H_20	33.2 ml
TEMED	50 µl
10 % APS	500 µl

2. Once gel polymerisation was complete, the bottom reservoir of a vertical gel tank (Whatman Biometra Model V15.17) was filled with 0.5 litres of 1X TBE buffer.
3. The spacer was removed from the gel base and the gel cassette was fastened into the gel tank.
4. The upper buffer reservoir of the gel tank was filled with 0.5 litres of 1X TBE buffer ensuring that the comb was submerged. The comb was then carefully removed and wells were rinsed with 1X TBE buffer using a syringe.

5. Gels were pre-run for a minimum of 45 minutes at 120 volts prior to sample loading.

Preparation of Probe +/- Competitor DNA

6. DNA samples were prepared for addition to protein binding reactions as described below.

Sonicated poly (dA.dT)	1 μl
Radiolabelled DNA probe (0.2 μM)	0.05 μl
Cold specific competitor DNA (if required) (2 μM)	0.25 μl
Make to 3 μl with H_20	

Preparation of Protein-DNA binding reactions

7. Binding reactions were prepared follows.

2X Binding Buffer	10 μl
DTT (100 mM)	0.2 μl
10 % Triton X-100	0.5 μl
BSA (NEB, 10 mg/ml)	0.2 μl
VEZF1 *in vitro* translation OR	2 μl
Chick red blood cell nuclear extract	0.25 μl
Make to 15 2 μl with H_20	

8. 2 μl of anti-VEZF1 antibody was added to reactions if required and all binding reactions were incubated on ice for 2 hours.
9. Reactions were then incubated at room temperature for 15 minutes.
10. 3 μl of Probe +/- competitor DNA was added to each binding reaction, mixed and incubated at room temperature for 45 minutes.

Gel Electrophoresis

11. EMSA gels were loaded while running at 120 volts.
12. Wells were rinsed again with 1X TBE using a syringe, and 10 μl of protein/DNA binding reactions were loaded per well.
13. 5 μl of bromophenol blue loading buffer was loaded into well number 20 on the left hand edge of the gel to allow gel orientation.
14. Gels were run at 120 volts for 2 hours.

Gel drying and imaging

15. Once electrophoresis was complete, the gel unit was removed from the running tank and the uppermost glass plate removed.
16. A piece of Whattman filter paper was place on top of the gel and used to separate the gel from the second glass plate.
17. The gel was then placed on top of a further two pieces of Whattman paper and overlaid with clingfilm before gel drying using gel drying equipment.
18. The gel was left in the drying unit for ~50 minutes and was then exposed to a phosphorimaging screen for ~40 hours.
19. The phoshporimaging screen was scanned using a Fuji scanner allowing visualisation of radiolabelled probe DNA.

Chapter 3

Profiling of VEZF1 DNA-binding events across the human genome in the human erythroid cell line K562

3.1 Introduction

VEZF1 is a highly conserved DNA-binding protein found in vertebrates. *Vezf1*-null mouse embryos exhibit an embryonic lethal phenotype resulting from severe haemorrhaging which is caused by a lack of vascular integrity (Kuhnert et al., 2005). Disruption of *Vezf1* also impairs vasculogenesis and haematopoeisis during embryonic stem cell differentiation (Zou *et al*., 2010). Despite its essential role in development, the gene targets of VEZF1 are relatively poorly characterised. VEZF1 binding sites have been identified within a small number of promoters (Lewis *et al*., 1988, Koyano-Nakagawa *et al*., 1994, Aitsebaomo *et al*., 2001, Dickson *et al*., 2010) and the chicken HS4 insulator element (Dickson *et al*., 2010). A preliminary VEZF1 ChIP-chip experiment in my supervisor's research group previously identified 125 VEZF1 binding sites within 1 % of the human genome in K562 cells (section 1.9). These sites included housekeeping and cell type-specific gene promoters in addition to gene distal enhancers and *potential* insulators (Strogantsev, 2009).

In this study, I wish to gain a clearer understanding of the general gene regulatory process that VEZF1 is involved in. To achieve this, I begin by using ChIP-seq to profile the binding of VEZF1 across the human genome in a highly studied cell type. This profile can then be carefully studied to determine whether VEZF1 generally associates with transcription factor functions or has a more specialised role in regulating chromosome structure like other insulator-binding proteins.

Prior to performing ChIP-seq, a cell type in which to profile VEZF1 genomic binding had to be selected. In line with the interests of my supervisor's research group and that of my sponsor (the Medical Research Council (MRC)) in human gene regulation, a human cell type was selected. We decided to focus on a cell line rather than a primary tissue at this time as cell lines offer homogeneity and the potential to scale up cell numbers when developing new methodology. We decided to perform VEZF1 ChIP-seq analysis in the

human erythroid cell line K562. There were a number of reasons for this. Firstly, my supervisor's research group has a long standing interest in erythroid gene regulation and had previously mapped VEZF1 binding to a portion of the genome in K562 cells using a ChIP-chip approach (section 1.9). This preliminary study will be used a reference when analysing VEZF1 ChIP-seq performance. The group have also generated a K562 cell line that expresses FLAG epitope-tagged recombinant VEZF1 at 50% of endogenous levels. This will allow for VEZF1 ChIP analysis with anti-FLAG antibodies, which will act as a control for potential chromatin enrichments that are not specific for VEZF1 when performing standard ChIP analysis with anti-VEZF1 antibodies. Another major factor which led to the choice of K562 cells was their status as an ENCODE tier 1 cell line. ENCODE research teams have characterised histone modifications, chromatin features and the binding profiles of many transcription factors across the whole K562 genome. This data is publically available, which therefore allows me to address my objectives of determining the chromatin state and co-binding transcription factors at VEZF1 sites in K562 cells without the need for extensive experimentation.

The principal objectives of this chapter are to:

1. Prepare crosslinked chromatin from K562 cells
2. Perform VEZF1 ChIP and validate specific enrichments using QPCR
3. Prepare and validate VEZF1 ChIP-seq libraries
4. Align ChIP-seq reads to the human genome
5. Identify peaks of VEZF1 binding in the K562 genome

3.2 Preparation of Crosslinked chromatin from K562 cells

An optimised transcription factor ChIP protocol routinely used in the lab was followed for all ChIP experiments (section 2.5). The crosslinking of K562 cell cultures in exponential growth phase in warmed serum free media with 0.8 % formaldehyde for 5 minutes was previously determined to give optimal detection of DNA-binding proteins such as VEZF1 (Strogantsev, 2009). The standard approach of 1 % formaldehyde for 10 minutes or longer may mask antibody-binding epitopes on DNA-binding proteins and may also impair chromatin sonication efficiency, resulting in reduced signal to background levels. Crosslinked chromatin was prepared and sonicated to an average fragment size of 200 – 400 bp ready for immunoprecipitation (Figure 3.1).

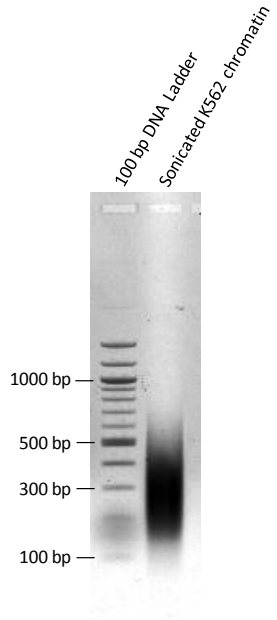


Figure 3.1 Shearing of K562 chromatin.
Agarose gel electrophoresis of K562 chromatin after sonication to an average fragment size of 200 – 400 bp for use in ChIP.

3.3 Validation of VEZF1 ChIP performance

Chromatin preparations were pre-cleared with non-immune rabbit or mouse IgG and protein A or G agarose to remove any non-specific complexes. For VEZF1 and control IgG ChIPs, chromatin from $1x10^7$ cells was used per IP. Three VEZF1 ChIPs were performed simultaneously and the final samples were combined for a final volume of 150 µl. For FLAG ChIP the amount of chromatin was dropped to $5x10^6$ cells per IP, two FLAG ChIPs were performed and the final samples were combined for a final volume of 100 µl. The anti-VEZF1 antibody 3642 was used for the ChIP of endogenous VEZF1 from K562 cells (Dickson *et al.*, 2010). The performance of VEZF1 ChIP in K562 cells was assayed by

quantitative PCR (QPCR) using primers designed against regulatory elements previously identified to bind VEZF1 (Strogantsev, 2009). Primer sets that detect the *IL3* and *EDN1* promoters serve as negative control loci as VEZF1 is known not to bind these elements in erythroid cells. The promoters of the broadly expressed genes *STAG2*, *HISPPD2A* and *POLR3K*, enhancers of the erythroid-specific gene *TAL1* and a putative insulator at the *FLNA* locus were all enriched following VEZF1 ChIP (Figure 3.2). A putative enhancer element within intron 3 of the *FOXP4* gene was mildly enriched by VEZF1 ChIP. The *IL3* and *EDN1* promoters were not enriched. These results are consistent with previous VEZF1 ChIP-QPCR and ChIP-chip analyses in K562 cells (Strogantsev, 2009), so the VEZF1 ChIP was taken forward for ChIP-seq analysis.

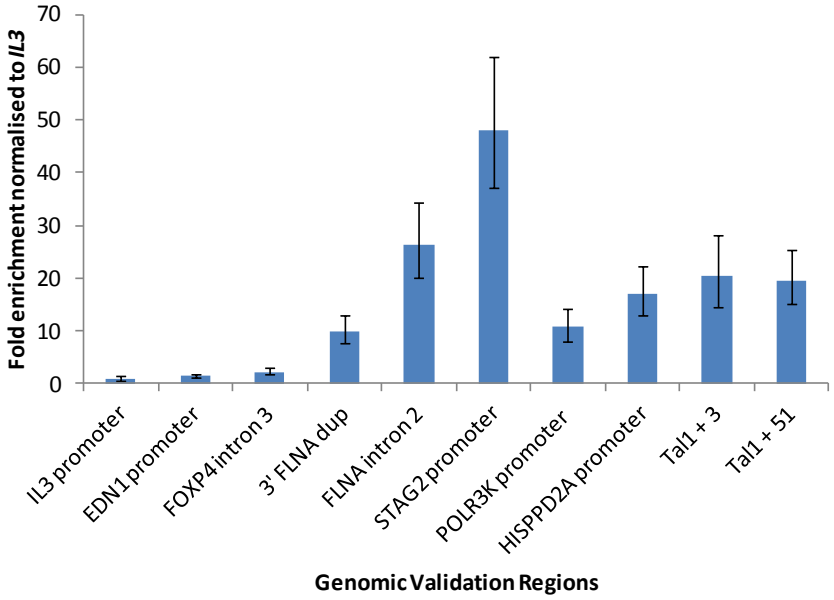


Figure 3.2 VEZF1 interacts with gene regulatory elements in K562 cells.
QPCR analysis of regulatory element enrichment after VEZF1 ChIP. The relative enrichment of each element over starting input chromatin after VEZF1 ChIP was normalised to that at the *IL3* promoter. Error bars represent standard deviation between triplicate QPCR analyses of a VEZF1 ChIP.

Agarose conjugated anti-FLAG antibodies were used to ChIP for recombinant VEZF1 from the K562 cell line D3/6, which expresses FLAG-tagged VEFZ1 protein at approximately half the level of endogenous wild type VEZF1 (Dan Li, unpublished data). A control K562 cell line S3F, which expresses the FLAG protein tag alone, was used to control for non-specific chromatin interactions with anti-FLAG agarose. QPCR analysis was used to show that the *POLR3K* and *RFX5* gene promoters and the *TAL1* +51 enhancer were all enriched following FLAG ChIP from the FLAG-VEZF1 expressing cell line D3/6 (Figure 3.3). No enrichment of the *IL3* promoter was observed. No enrichment of VEZF1-associated regulatory elements

was observed with anti-FLAG ChIP in the control cell line S3F. These results are consistent with previous ChIP analysis with anti-VEZF1 antibodies in the parental K562 cells (Figure 3.2) (Strogantsev, 2009). These results indicate that FLAG-tagged recombinant VEZF1 can interact with the same gene regulatory elements as endogenous VEZF1. The FLAG-VEZF1 ChIP was therefore taken forward for ChIP-seq analysis.

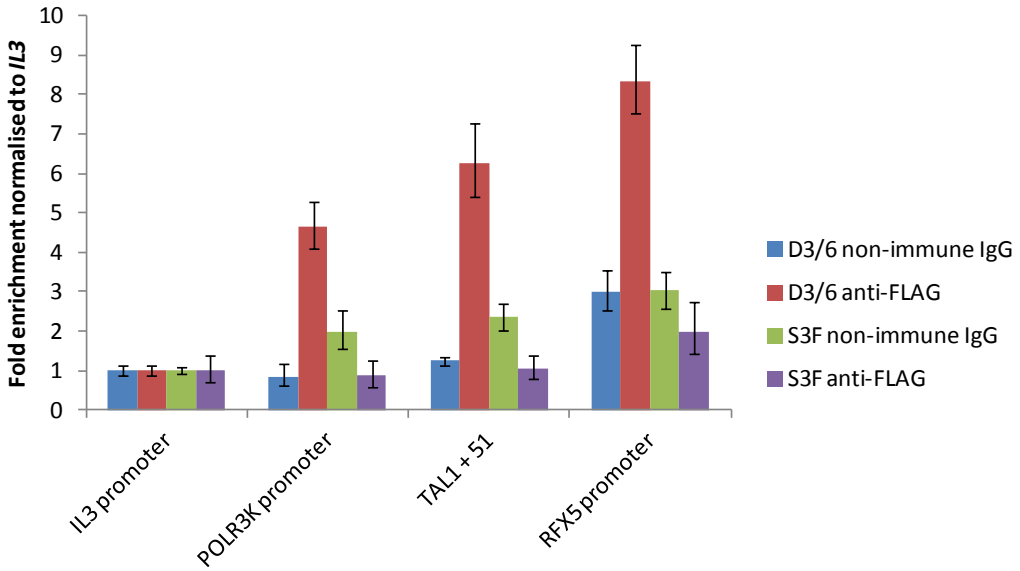


Figure 3.3 Recombinant VEZF1 can interact with the same gene regulatory elements as endogenous VEZF1.
QPCR analysis of gene regulatory element enrichment following ChIP with anti-FLAG or non-immune IgG form K562 cell lines expressing FLAG-VEZF1 (D3/6) or FLAG only (S3F). The relative enrichment of each element over starting input chromatin after VEZF1 ChIP was normalised to that at the *IL3* promoter. Error bars represent standard deviation between triplicate QPCR analyses of a VEZF1 ChIP.

3.4 ChIP-seq Library Preparation

ChIP-seq analysis was performed using the Illumina Genome Analyzer IIx platform at the University of Glasgow Polyomics Facility. The Illumina technology requires the preparation of a genomic DNA library where the enriched ChIP DNA fragments are ligated to adaptor oligonucleotides, PCR amplified and size selected prior to high throughput parallel DNA sequencing (introduction section 1.5.1.2). The standard Illumina method involves size selection prior to PCR amplification. It is known that sequencing of input ChIP-seq libraries tends to reveal increased read counts at DNaseI HSs (Vega *et al*., 2009, Liu *et al*., 2010). This bias is believed to result from regions of euchromatin being more efficiently fragmented by sonication than regions of heterochromatin. Size selection of small DNA fragments prior to PCR amplification during library preparation therefore preferentially biases towards sequencing of euchromatic genomic regions. In order to

test for bias in my ChIP-seq samples the negative control IgG ChIP-seq track will be checked for enrichment above background levels at DNaseI HSs. Size selection is also known to reduce the complexity of the library and increase the probability of the same fragments being amplified multiple times. Identical sequences are removed following sequencing as they may arise from PCR bias rather than genuine ChIP enrichment. This factor of library complexity is more of an issue when the experiment results in relatively low numbers of enrichment events, such as when a transcription factor specifically interacts with a few thousand elements, as compared with broadly distributed histones for example. In order to ensure maximal library complexity during ChIP-seq of transcription factors, we employ a rearranged library preparation procedure where PCR amplification is carried out prior to size selection on agarose gels (Dr David Vetrie personal communication). The procedure uses 30 µl of ChIP DNA – which equates to three fifths of one ChIP reaction – and reagents from the NEBNext DNA Sample Prep Reagent Set 1 kit (E6000S) (New England Biolabs) and mostly follows the manufacturer's recommendations (section 2.6).

One sixth of the volume of each ChIP-seq library (5 µl) was diluted 1 in 10 and used as template in QPCR to determine whether enrichment patterns at genomic validation regions were comparable before and after library preparation. Once again primer sets detecting the *IL3* promoter served as a negative control for enrichment. The *STAG2* promoter and *TAL1* +51 enhancer elements remained enriched following ChIP-seq library preparation of the K562:VEZF1 ChIP sample (figure 3.4). However, the relative enrichments of the *STAG2* promoter and *TAL1* +51 enhancer altered following library preparation, probably due to size selection affecting the relative abundances of genomic fragments containing the QPCR primers. The *TAL1 +51* enhancer and *RFX5* promoter remained enriched following ChIP-seq library preparation of the K562(D3/6):FLAG-VEZF1 ChIP sample (figure 3.4). The *POLR3K* promoter was slightly enriched compared to the *IL3* negative control in this library, however it appears that all three genomic validation regions tested by QPCR are less enriched relative to the *IL3* control following library preparation than in ChIP samples prior to library preparation (figure 3.3).

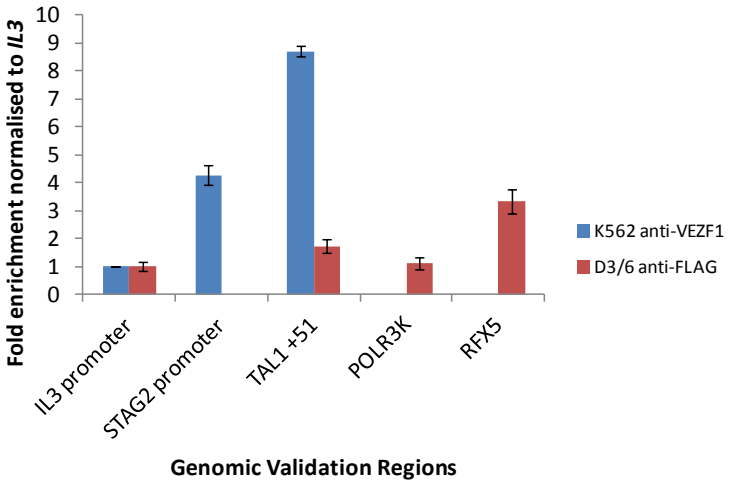


Figure 3.4 Enrichment of regulatory elements is retained following ChIP-seq library preparation.
QPCR analysis of regulatory element enrichment after library preparation using genomic DNA from VEZF1 ChIP in K562 cells and FLAG ChIP in a K562 cell line expressing FLAG-VEZF1 (D3/6). The relative enrichment of each element over starting input chromatin after VEZF1 ChIP-seq library preparation was normalised to that at the *IL3* promoter. Error bars represent standard deviation between triplicate QPCR analyses of a VEZF1 ChIP-seq library.

Many groups prefer to check the fragment size distribution of their ChIP-seq libraries prior to sequencing. This is not a standard procedure at the Glasgow Polyomics Facility, who prefer to use a QPCR based library quantification assay as the primary determinant of library quality and quantity. Nevertheless, we ran a 1 µl aliquot of the K562:VEZF1 and K562:IgG on an Agilent Bioanalyzer using the Agilent High Sensitivity DNA Kit. Despite the size selection of ChIP DNA fragments from 200 – 300 bp, the analysis unexpectedly showed that the library fragments were quite broad, ranging from ~100 – over 1000 bp with a main fragment size of ~ 300 – 400 bp (Figure 3.5). A different ChIP-seq library prepared by the Polyomics Facility showed an expected banding around ~300 bp. The differences in library appearance are probably a result of different practices in agarose gel extraction procedures during library preparation. My libraries were gel purified using the Qiagen Qiaquick gel extraction kit using the manufacturers' protocol, where agarose gel slices are heated at 50 °C to promote melting. The Polyomics Facility prefers to complete agarose melting at room temperature. It appears that the heating of short DNA fragments in 5.5 M guanidine thiocyanate (QG buffer in the kit) promotes melting of the DNA fragments. A mixture of staggered and structured annealing products may form upon cooling. Such annealing products are of no concern for the Illumina sequencing as the libraries are denatured prior to cluster formation. However, it is clear that analysis of

sequencing libraries on an Agilent Bioanalyzer can be misleading unless precautions to avoid strand melting are taken.

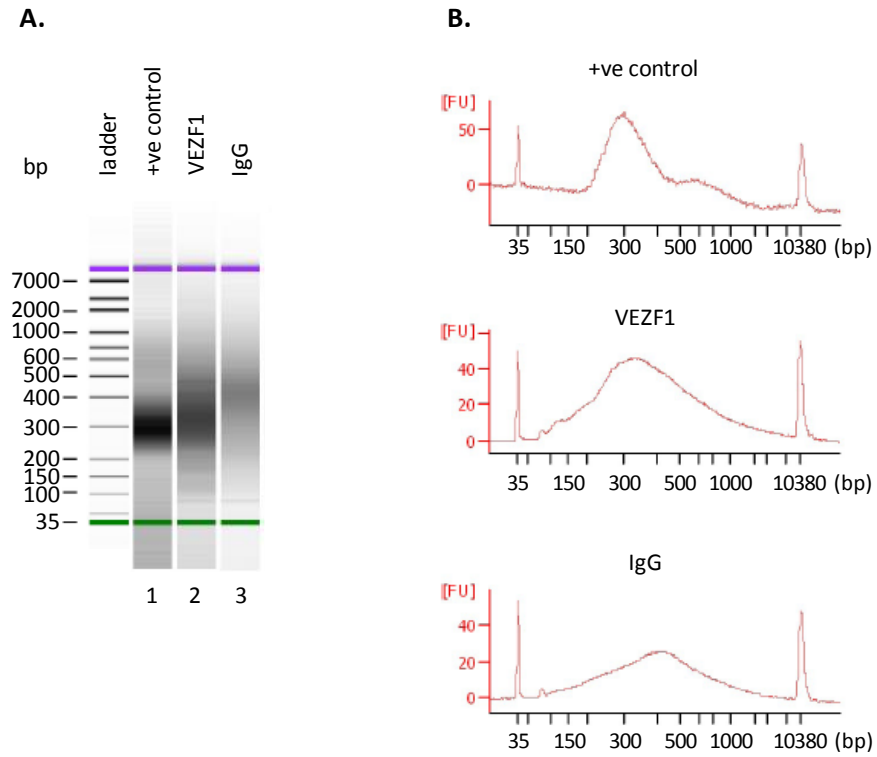


Figure 3.5 ChIP-seq library fragment size analysis.
The size distributions of K562:VEZF1 and K562:IgG ChIP-seq library fragments were determined using an Agilent Bioanalyzer. Gel (A) and electropherogram (B) results are shown for a positive control library prepared by the Polyomics Facility (lane 1) and for VEZF1 and IgG ChIP-seq libraries (lanes 2 and 3 respectively). Electropherogram peaks at 35 bp and 10,380 bp (B) correspond to size markers.

The concentrations of correctly adapted fragments in ChIP-seq libraries were quantified using the SYBR QPCR KAPA library Quantification Kit from Illumina (KK4822) by the University of Glasgow Polyomics Facility (table 3.1). An aliquot of each library was used to prepare a 20 µl solution at 1.5 nM for denaturation (table 3.1). Denatured library samples were further diluted to 12 pM concentrations for loading on the flow cell.

Sample	Concentration	Volume for 20 µl at 1.5nM
K562:VEZF1	4.88 nM	6.1 µl
K562:IgG	2.0 nM	15 µl
K562(D3/6): FLAG-VEZF1	16.57 nM	1.8 µl

Table 3.1 ChIP-seq library quantification.
The concentration of ChIP-seq libraries were quantified by QPCR and used to calculate the volume of each library required to prepare a 20 µl sample at 1.5 nM for denaturation.

3.5 VEZF1 ChIP-seq data quality

The University of Glasgow Polyomics Facility performed single end sequencing of 76 bases in length for the K562:VEZF1, K562:IgG and K562(D3/6):FLAG-VEZF1 ChIP-seq preparations. The Facility provided FASTQ formatted files generated using CASAVA software. The performance of these sequencing runs was initially checked using the FastQC software (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). Firstly, it was clear that there was a good yield, with 34-40 million reads from each run (table 3.2). Most sequencing facilities report average yields of ~35 million reads per GAIIx run. The average quality of the reads was good and there is little or no adaptor sequence contamination.

Sample	Read length	Read number	Mean Q score	Adaptor sequence
K562:VEZF1	76	34,327,963	38	0.25% of total
K562:IgG	76	34,978,434	38	No significant level
K562(D3/6): FLAG-VEZF1	76	40,327,215	38	No significant level

Table 3.2 Illumina GAIIx sequencing performance.
Data extracted from FastQC reports from 3 ChIP-seq runs.

While the overall quality of each sequence run was very good, closer inspection of the quality scores at each base is required before deciding on how much of the 76 base sequence to include in genome alignment. Low quality scores towards the end of sequences may indicate incorrect base calls, leading to an inability to align to a reference genome. The FastQC programme produces a box-and-whisker histogram of quality (Q) score distribution for each base position in a sequence run. FastQC analysis showed that the median sequence quality was in the very high range (Q scores of 28 – 40) for each position of the 76 base read length for each of the three ChIP-seq runs (Figure 3.6). However, it is clear that an increasing fraction of the reads have reduced quality after approximately 50 bases in read length, with reads in the 10^{th} and 90^{th} percentiles having poor quality scores (below 20). Sequences within the 20^{th} – 80^{th} range had poor quality scores after 70 bases in read length. This level of performance is considered to be very good for the Illumina GAIIx, but the FastQC analysis indicates that a proportion of reads might not match a reference genome if the full sequences are used for alignment. The general benchmark for high throughput sequencing quality is an average Q score of 30,

which represents a 1 in 1,000 probability of an incorrect base call. These sequencing runs fall below that threshold after ~60 bases.

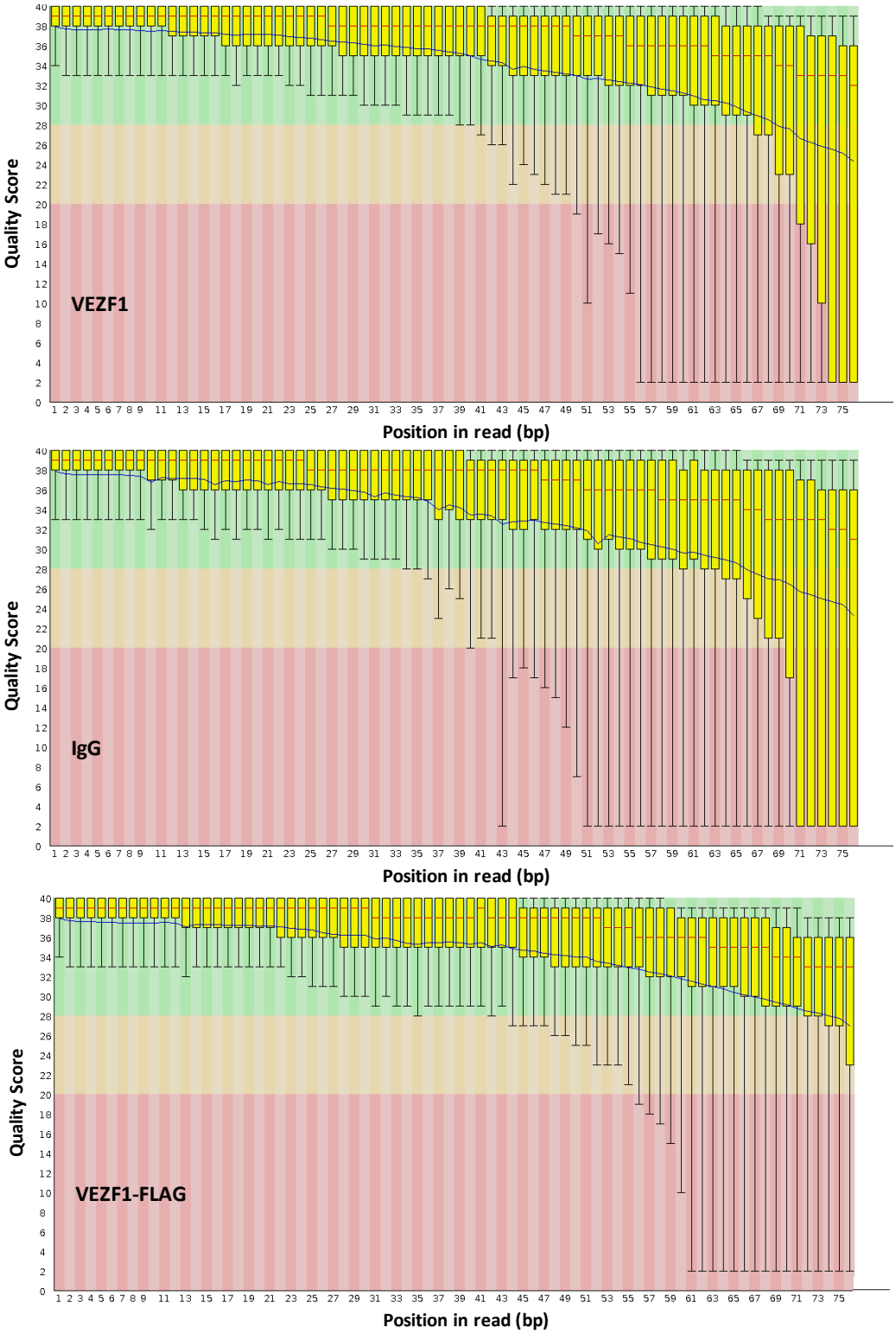


Figure 3.6 Illumina GAIIx sequencing per base quality.
FastQC report of Q score per base call for VEZF1, IgG and VEZF1-FLAG ChIP-seq runs. Horizontal red lines show the median Q score for each base position and the connected blue line indicates the average quality score. Yellow boxes represent the 20th-80th quartile range, while black whiskers represent the 10th and 90th percentiles. The background of these plots are divided into three sections where green, orange and red shading indicate Q score ranges considered to be of very good, reasonable or poor quality, respectively.

Sequences were aligned to the hg19 human reference genome using the Bowtie software package (Langmead *et al*., 2009). There are several short read alignment tools available, but we chose Bowtie due to its speed of data generation, memory efficiency and flexible options (Langmead, 2010). Bioinformatic analyses of aligned ChIP-seq reads will rely on the quantitative differences in read density at enriched sites versus non-enriched background. It is therefore important to ensure that no reads create false positive enrichments. Thus, it is standard practice in the field to consider only uniquely aligned reads by using the bowtie setting "-m" to exclude multiply aligned reads. However, this comes at the expense of unique ChIP-seq reads that align to duplicated and repetitive genomic regions. Of these, we are interested in duplicated genomic regions and can selectively include them. Example duplicated genomic regions include the α and β-globin gene clusters. We therefore use the Bowtie setting "-m 3" to report any sequence that aligns to 1, 2 or 3 genomic locations. In cases where Bowtie finds two or three genomic locations for a particular sequence, it would randomly assign the sequence to one of these locations to avoid amplifying the read density at duplicated genomic regions. Alignments to repetitive sequences are excluded.

Bowtie outputs a tab-delimited text file containing sequence alignments (SAM format), which is converted to a binary file (bam format) for subsequent bioinformatic applications. Bowtie also outputs a log file of the overall alignment performance. The alignment statistics from each alignment using the full 76 base sequences is shown in Table 3.3. 80% of the FLAG-VEZF1 ChIP-seq reads were aligned by Bowtie, so this alignment was taken forward for further analysis. However, only 31% and 44% of the VEZF1 and IgG ChIP-seq reads were aligned respectively. This is likely to be due to the reduced sequence quality at the 3'end of the VEZF1 and IgG ChIP-seq runs compared to the FLAG-VEZF1 run (Figure 3.6).

The VEZF1 and IgG ChIP-seq reads that failed to align were re-aligned using Bowtie with the setting "-3 40", which trims 3' ends of the reads to consider only first 40 bases of sequence. This ensures that all of the sequence being considered is of very high quality, but this comes at the expense of complexity. We found that trimming of sequences allowed for much greater fraction of alignment of VEZF1 and IgG ChIP-seq reads (Table 3.4). A further ~12 million VEZF1 reads were aligned following trimming, more than doubling the number of alignments. An addition 5.4 million IgG reads were also aligned

following trimming. Bowtie considers each sequence individually, so an additional step is made to filter out any identical sequences that may have arisen from PCR amplification. This removed 6-7% of the aligned reads from VEZF1 and IgG ChIP-seq runs, which our colleagues find typical fraction for good quality library preparations. However, 32% of the aligned reads from the FLAG-VEZF1 ChIP-seq run were removed due to clonal amplification. This is likely to be due to too low an amount of genomic DNA from the FLAG-VEZF1 ChIP entering into the library preparation.

Sample	Total reads	Aligned reads (Reported)	Failed to align	Reads that align >3 sites (discarded)
K562:VEZF1	34,327,963	10,503,949 (30.60%)	22,614,253 (65.88%)	1,209,761 (3.52%)
K562:IgG	34,978,434	15,299,779 (43.74%)	17,903,961 (51.19%)	1,774,694 (5.07%)
K562(D3/6): FLAG-VEZF1	40,327,215	32,035,696 (79.44%)	5,176,858 (12.84%)	3,114,661 (7.72%)

Table 3.3 Alignment of ChIP-seq reads using Bowtie.
Performance of Bowtie using the full 76 base sequences from each ChIP-seq run using the filter of multiply aligned reads –m3.

Sample	Total	Aligned (76 bases)	Aligned (40 bases)	Combined	Removed (PCR filter)	**Unique aligned**
K562:VEZF1	34,327,963	10,503,949 (30.60%)	11,996,579 (53.05%)	22,500,528	1,260,562 (5.6%)	**21,239,966**
K562:IgG	34,978,434	15,299,779 (43.74%)	5,410,839 (30.22%)	20,710,618	1,530,108 (7.4%)	**19,180,510**
K562(D3/6): FLAG-VEZF1	40,327,215	32,035,696 (79.44%)	-	32,035,696	10,363,304 (32.3%)	**21,672,392**

Table 3.4 Unique aligned ChIP-seq reads after PCR filtering.
Performance of Bowtie using the full 76 base sequence or trimmed to 40 bases (VEZF1 and IgG samples only). Sequence alignments were combined and then screened for identical sequences (PCR filter).

Following Bowtie alignment and PCR filtering, there were 19-22 million unique aligned reads from each ChIP-seq preparation. The depth of sequencing necessary to gain a full picture of the genomic interactions of a DNA binding protein largely depend upon the number of interaction sites, which is often unknown. Thorough analysis of ChIP-seq data generated by the ENCODE consortium has found that 20 million mapped reads is a good benchmark for ChIP-seq of a typical point-source DNA-binding factor. Enrichment peaks typically become saturated with additional sequencing depth, but new weak peaks continue to emerge with further sequencing (Landt *et al*., 2012). For VEZF1, a preliminary

ChIP-chip analysis of VEZF1 binding to 1% of the K562 genome identified 125 sites (section 1.9). We would therefore expected ~20,000 VEZF1 genomic binding events in K562 cells when accounting for the whole genome size and improved resolution of neighbouring binding events. If only 50% of the mapped VEZF1 ChIP-seq reads were due to specific enrichment, we would still have ~500 reads per VEZF1 binding event on average.

The .sam format alignment files produced by Bowtie were converted into binary .bam format files before conversion into plain text genome coordinate files of the .bed format to enable inspection on a genome browser. In order to visualise ChIP-seq data on a genome browser, the reads are first extended to 150 bases in length in order to more accurately represent the average size of the fragments present in the size selected ChIP library (200-300 bp – adaptors). The files were finally converted into .BigWig format files to enable rapid uploading and visualisation in the UCSC Genome Browser (http://genome.ucsc.edu/). During this conversion, the read counts that align to each genomic location were normalised for the total number of reads in the experiment to allow fairer comparison between datasets on the browser.

The ChIP-seq data were visually inspected on the UCSC Genome Browser, a representative gene rich portion of chromosome 21 is shown in Figure 3.7. The VEZF1 ChIP-seq track shows well defined peaks of enrichment. These peaks are not present in the IgG negative control track, showing that these genomic sequence elements were specifically enriched by immunoprecipitation of VEZF1 protein. The FLAG-VEZF1 ChIP-seq track displays the same peaks of enrichment as the VEZF1 ChIP-seq track. As no additional enrichment peaks are seen in the VEZF1 track compared to the FLAG track, it can be concluded that the α-VEZF1 antibody used in VEZF1 ChIP-seq interacts with the VEZF1 protein alone. The relative enrichment of the α-FLAG ChIP over background was weaker than that for the α-VEZF1 ChIP. This observation, coupled with the high levels of PCR clonality during library preparation for this ChIP (Table 3.4), suggests that further optimisation of α-FLAG ChIP would be required to increase IP yields if this approach was to be used again in future.

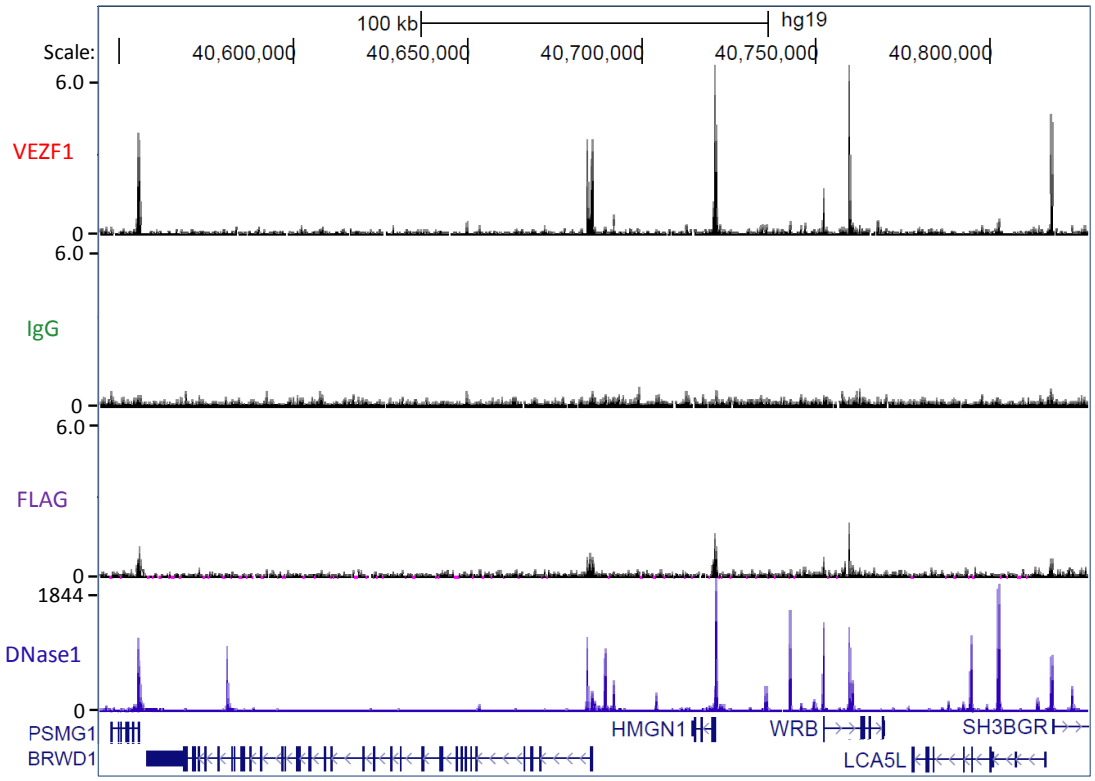


Figure 3.7 VEZF1 binding at specific elements revealed by ChIP-seq.
UCSC genome browser view of ChIP-seq data mapped to human chr21 40,547,000-40,831,000 (hg19). From top to bottom, tracks for VEZF1(K562), IgG(K562), FLAG-VEZF1 (K562:D3/6) and DNaseI (K562) are shown. The VEZF1/IgG/FLAG data are normalised to library size, where a peak height of 6 equates to 120 overlapping reads. The K562 DNaseI-seq (signal) track is from ENCODE accession wgEncodeEH000480 (Sabo *et al*., 2004). Below the tracks are UCSC annotated genes, where linked vertical blocks indicate protein-coding exons and arrows indicate the direction of transcription.

The VEZF1 peaks typically appear to be associated with the 5' ends of the annotated genes. Consistent with this, VEZF1 ChIP-seq peaks correlate with peaks of DNaseI hypersensitivity (Figure 3.7). The promoter elements present at the 5' ends of transcriptionally active genes are typically marked by DNaseI hypersensitive sites, suggesting that VEZF1 is bound at transcriptionally active promoters. It is also evident that there are a number of DNaseI hypersensitive peaks which do not align to sites of VEZF1 enrichment. The DNaseI track therefore provides further evidence that the VEZF1 ChIP-seq has specifically detected VEZF1 genomic binding events and not just all highly accessible DNA elements such as those hypersensitive to DNaseI cleavage.

I have shown that the α-VEZF1 ChIP-seq experiment resulted in high read quality and depth. Specific genomic sites with high levels of VEZF1 enrichment are evident, which are validated by similar enrichments in the α-FLAG ChIP-seq. I was therefore satisfied that the

VEZF1 ChIP-seq had identified genomic elements bound by VEZF1 and proceeded with bioinformatic analysis.

3.6 Determination of VEZF1 binding events by peak finding

A number of peak finding tools are available to identify regions of sequence enrichment from ChIP-seq data and thus identify specific sites of genomic interaction by a DNA-binding factor. We have settled on using the MACS (Model-based Analysis of ChIP-Seq) programme to identify peaks of transcription factor binding as it tends to outperform other tools (Zhang *et al*., 2008). Most peak finding tools rely solely on a control experiment, such as the IgG control in this study, to estimate background biases in sequencing technology, chromatin structure effects and genome copy number variation. However, it is difficult to provide a truly representative control that has been sequenced to the same quality and depth, so many peak finding algorithms tend to result in many false positives and negatives. MACS takes advantage of the fact that enriched ChIP-seq reads show a bimodal pattern with Watson strand tags enriched upstream of binding and Crick strand tags enriched downstream. MACS empirically models these patterns to both improve peak calling accuracy and define precise peak summits (the binding sites).

MACS peak finding on the VEZF1 ChIP-seq data was performed using settings that specified a bandwidth (chromatin fragment size) of 150 bp and an input sequence tag size of 40 bp. These settings allowed MACS to identify 10,056 VEZF1 peaks. However, inspection of the MACS peaks on the genome browser revealed three fundamental problems. First, peak resolution was poor as peak coordinates were often much wider than the original apparent regions of enrichment (Figure 3.8 A, “MACS 40nt” track). Second, independent neighbouring peaks were often called as one peak (Figure 3.8 C and D). Third, many strong single peaks were not identified at all (Figure 3.8 B, “+20” and “+51” peaks). These issues were largely overcome by instructing MACS to use an input sequence tag size of 25 bp (Figure 3.8 A – D, compare MACS 40 nt peak and MACS 25 nt peak tracks). This analysis identified 20,993 peaks. The majority of the VEZF1 enrichment peaks observed in the UCSC genome browser were detected and the size of the peaks were appropriately defined.

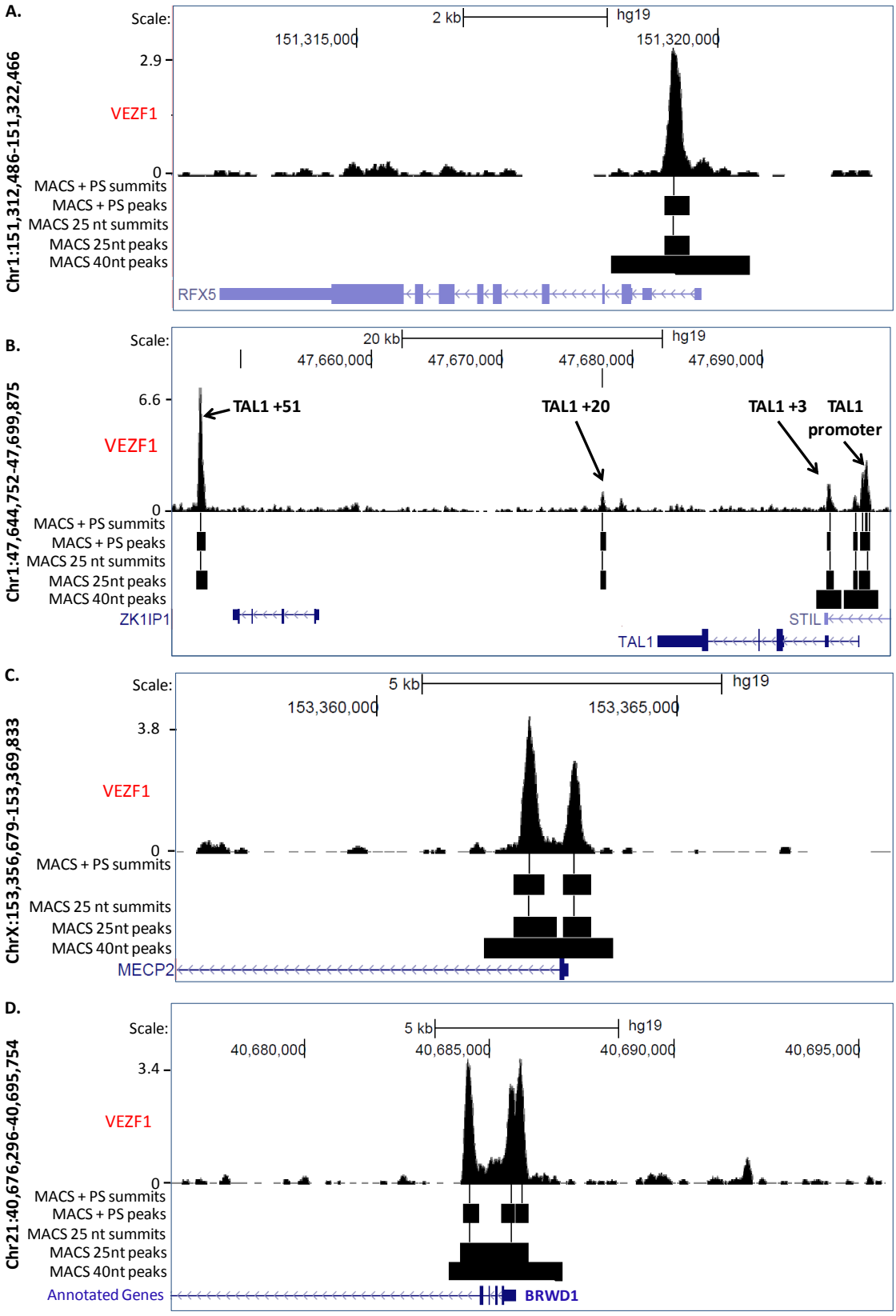


Figure 3.8 Performance of MACS peak calling of VEZF1 binding events.
VEZF1 ChIP-seq enrichment profiles across four genomic loci (A – D) in K562 cells (top tracks). The coordinates of peaks of VEZF1 enrichment identified using MACS using 25 or 40 nucleotide sequence input sizes are shown beneath. The coordinates of separated peaks from 25nt MACS resolved by PeakSplitter (PS) are also shown. The single base peak summits from 25nt MACS and PeakSplitter are shown as vertical lines. UCSC annotated genes are shown in blue and the direction of transcription is indicated by arrows (bottom track).

It was apparent however, that MACS was unable to resolve clustered VEZF1 peaks into independent peak calls (Figure 3.8 D). This lack of resolution will interfere with subsequent bioinformatic analyses that rely on precise definition of VEZF1 binding site locations (the peak summits). PeakSplitter software (Salmon-Divon *et al*., 2010) was therefore employed to resolve individual peaks in a cluster and to identify peak summits. PeakSplitter resolved an additional 5,436 VEZF1 binding locations, which mostly appear to be at the 5'ends of genes that have multiple promoters (e.g., *TAL1* and *BRWD1* promoters, Figure 3.8 B and D).

Prior to further analyses, the VEZF1 peaks were ranked by sequence tag enrichment and obvious false positive peaks were removed. There were a total of 26,429 VEZF1 peaks following peak finding and splitting. The Seqmonk tool (http://www.bioinformatics.bbsrc.ac.uk/projects/seqmonk) was used to calculate the total number of sequence reads in the 50bp around each VEZF1 peak summit, expressed as reads per million (rpm) to normalise for total ChIP-seq library size. This was also performed for corresponding genomic regions in the IgG ChIP-seq track. The IgG rpm scores for each peak were subtracted from the corresponding VEZF1 scores, essentially normalising VEZF1 specific enrichments to IgG background levels. The VEZF1 peaks were ranked by normalised rpm scores. The lowest ranked peaks were inspected on the UCSC Genome Browser. It was found that "peaks" with a VEZF1-IgG rpm ratio of 2 or less were false positive peak calls. All of the VEZF1 peaks with ratios above 2 appeared to be discernible VEZF1 peaks, albeit weak, with no equivalent peak in the IgG track (not shown). We therefore decided at this stage to retain these weak VEZF1 sites for further analysis. 88 peaks were removed as false positives, leaving a total of 26,341 identified VEZF1 peaks in the K562 genome.

3.7 Discussion

The aim of this chapter was to use ChIP-seq technology to profile VEZF1-DNA interactions in the K562 cell line on a genome-wide scale. A previously optimised VEZF1 ChIP protocol was followed and QPCR validation showed VEZF1 ChIP to be reproducible as enrichment patterns correlated highly between this and previous VEZF1 ChIP experiments in K562.

ChIP-seq libraries were generated using genomic ChIP DNA in order to prepare ChIP-enriched DNA for next generation sequencing on the Illumina GAIIX platform. During library preparation adaptor oligonucleotides were ligated to either end of ChIP-DNA fragments to allow hybridisation to complimentary primers on the surface of the sequencing flow cell. Adaptor-ligated DNA fragments were enriched by PCR amplification and fragments of 200 – 300 bp were size selected following agarose gel electrophoresis. PCR amplification was performed prior to size selection in order to maintain maximal library complexity and reduce the likelihood of clonal PCR amplification leading to discarded sequence reads. Size selection was performed for fragments of 200 – 300 bp as sequences of this length are recommended for optimal cluster formation, larger fragment sizes may result in sequence clusters overlapping which will prevent bases from being accurately called during sequencing resulting in the loss of ChIP-seq reads.

Following their preparation ChIP-seq libraries were validated by QPCR to ensure that ChIP-enriched genomic regions remained so after library preparation. While informative, this analysis can be misleading as it relies on the binding sites of primer pairs being present on the same DNA fragment. Prior to library preparation a wide variety of DNA fragment sizes are present in ChIP DNA however following size selection during library preparation the relative abundances of DNA fragments that contain both binding sites for a pair of QPCR primers is likely to be altered and may affect the enrichment value generated by QPCR.

We tried to validate the size distribution of ChIP-seq libraries as being 200 – 300 bp in length using the Agilent Bioanalyzer and were surprised to find that the fragment sizes of these libraries ranged from ~100 – over 1000 bp. We believe this unexpected fragment size distribution to be an artefact which occurred during the extraction of size-selected library DNA from agarose gel fragments. It is believed that, during the heating of agarose gel slices at 50 °C in Qiagen's QG buffer, DNA fragments within the gel slice may have

undergone some denaturation and that staggered or structured DNA products may have then formed upon cooling. It therefore appears that, if ChIP-seq library size is to be validated using an Agilent Bioanalyzer, the gel slice should be dissolved at room temperature.

During the alignment of VEZF1 and IgG ChIP-seq reads to the human reference genome, 66 % and 51 % of full length reads respectively failed to align. It was considered that these failed alignments likely resulted from mis-called bases at the 3' end of sequence reads. The reads which failed to align were therefore trimmed to 40 bp by removing the 3' 36 bases, following which 53 % and 30 % of trimmed VEZF1 and IgG sequence reads respectively were successfully aligned to the reference genome. It should be noted however that cropping sequence reads increases the likelihood of these reads aligning to multiple sites within the reference genome which, as described previously, can lead to their being discarded or randomly assigned to one of the possible alignment sites (section 3.5). It is therefore apparent that a balance between sequence length and base call accuracy must be met in order to achieve the most possible sequence read alignments.

Peak finding using the MACS programme identified 20,993 VEZF1 enrichment peaks across the reference genome. It was evident however, that MACS was unable to resolve groups of clustered peaks into independent peak calls. PeakSplitter software was therefore employed to resolve independent clustered peaks identifying a further 5,436 VEZF1 peaks. Based on these findings, the combined use of MACS and PeakSplitter is recommended for maximal resolution of ChIP-seq peaks.

On first inspection, VEZF1 enrichment peaks appear to align with DNase I hypersensitive sites, as would be expected for a sequence-specific transcription factor. As a number of DHSs do not correlate with VEZF1 peaks, we are confident that VEZF1 ChIP has specifically enriched VEZF1-associated elements and not highly accessible DNA elements in general. As VEZF1 peaks appear to be mostly situated at the 5' ends of genes, future analyses should investigate the relationship of VEZF1 peak sites to TSSs, promoter-associated epigenetic modifications and nucleosome depleted regions. VEZF1 enrichment peaks at several known enhancer elements were observed by manual inspection of the VEZF1 ChIP-seq track (data not shown), future analyses should therefore also address whether

there is a general association between VEZF1 binding and enhancer elements or whether these events are rare.

Chapter 4

VEZF1 interacts with core promoters and cell-type specific enhancers

4.1 Introduction

In this chapter, I wish to gain a clearer understanding of the general gene regulatory processes that VEZF1 is involved in. To achieve this, I carefully studied the VEZF1 ChIP-seq profile in K562 cells to determine whether VEZF1 generally functions as a transcription factor or a more specialised chromosome structure factor like other insulator-binding proteins. The principal objectives of this chapter are to:

1. Determine the genomic distribution of VEZF1 binding with respect to genes and gene regulatory elements.
2. Determine the chromatin state(s) at elements bound by VEZF1.
3. Identify transcription factors that co-bind to gene regulatory elements with VEZF1.

4.2 The genomic distribution of VEZF1 binding with respect to genes and gene regulatory elements

4.2.1 VEZF1 binding at genes

VEZF1's role in mediating barrier activity at the chicken HS4 insulator element has lead us to hypothesise that VEZF1 ChIP-seq may reveal binding at many novel insulator elements, which would be typically located in intergenic regions. However, our initial inspection of the VEZF1 ChIP-seq profile in the UCSC Genome Browser indicated that VEZF1 was frequently associated with the 5' end of genes (section 3.5). The locations of the 26,341 VEZF1 peak summits identified in chapter 3, which I refer to as VEZF1 sites, were therefore compared to the RefSeq gene annotation using the web based ChIP-seq data analysis site Nebula (http://nebula.curie.fr/) (Boeva *et al.*, 2012). This analysis reveals that only ~16 % of VEZF1 sites locate to intergenic regions (Figure 4.1). Strikingly, it was observed that 55.6% of VEZF1 sites were located within 500 bp of transcription start sites (TSS), 33% upstream and 23.6% downstream, while a further 13% of VEZF1 sites locate to the first annotated intron. Thus, more than two thirds of VEZF1 sites are gene promoter proximal and locate to potential gene regulatory elements.

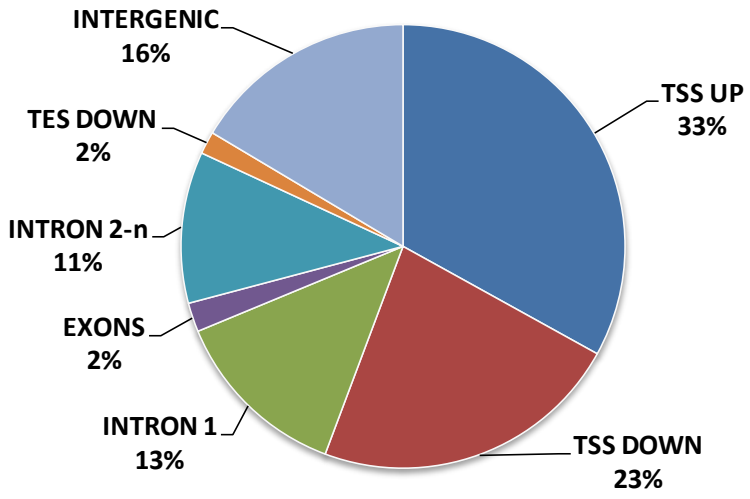


Figure 4.1 VEZF1 binding at genes.
The fraction of VEZF1 sites (ChIP-seq peak summits) mapping to annotated genes as reported by the Nebula tool. VEZF1 sites located within 500 bp upstream or downstream of transcription start sites (TSS) are indicated as TSS UP and DOWN, respectively. VEZF1 sites located within 500 bp downstream of transcription end sites (TES) are indicated as TES DOWN. The intergenic category includes VEZF1 sites that do not reside within genes or within 500 bp of their start or end.

4.2.2 VEZF1 binding at promoters

The high degree of VEZF1 association with promoters, led me to ask whether VEZF1 binding events are distributed across promoters akin to *cis*-regulatory factors or whether VEZF1 resembles a general transcription factor that is closely associated with TSS. The Nebula gene annotation of all VEZF1 peak locations was used to plot VEZF1 peaks with respect to the nearest annotated TSS. This analysis shows the highest density of VEZF1 peaks localise to the area around TSSs (Figure 4.2). This observation strongly suggests that VEZF1 is a general transcription factor with a likely role in gene promoter regulation.

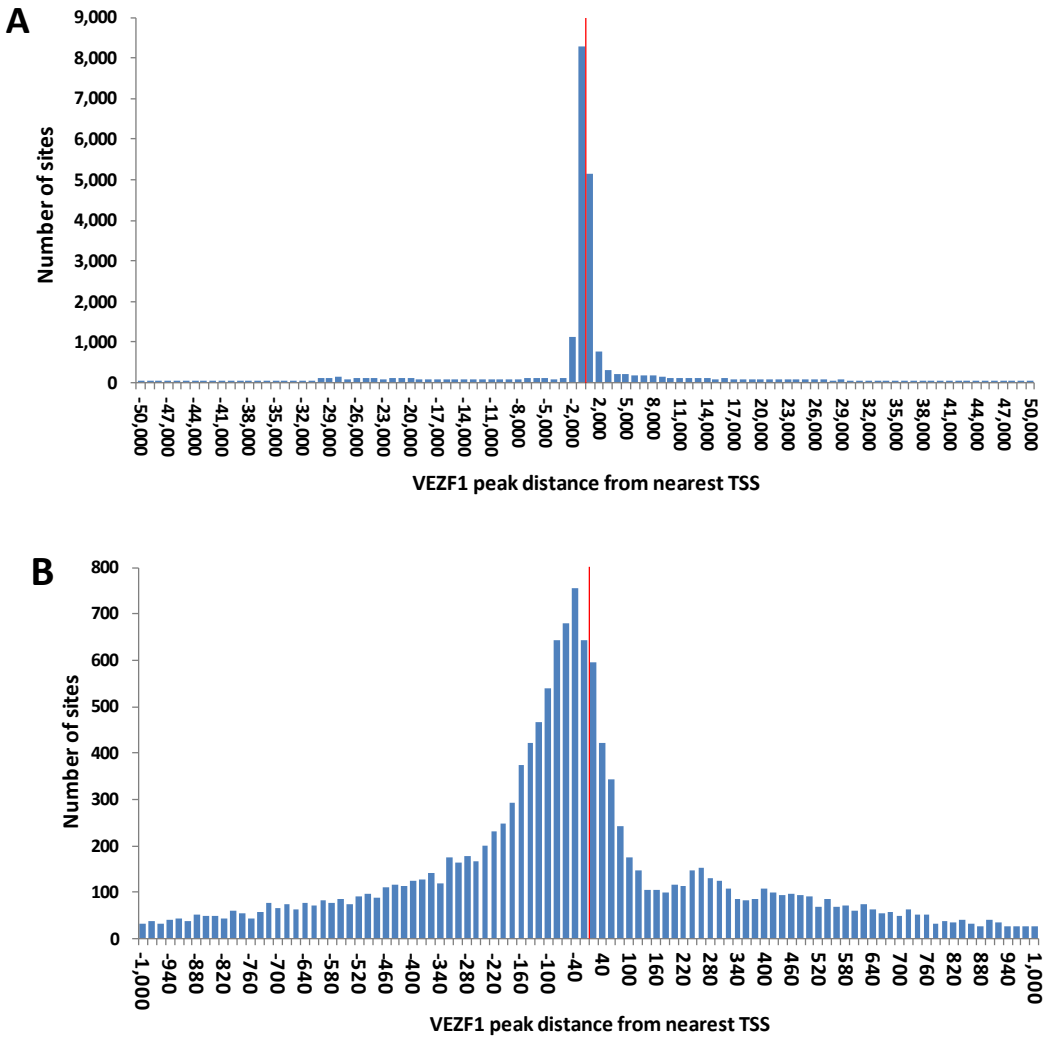


Figure 4.2 VEZF1 ChIP-seq peaks are enriched in the region of TSSs.
Distribution of VEZF1 ChIP-seq peaks with respect to nearest annotated RefSeq gene TSS. A) VEZF1 peak numbers in 1 kb windows up to 50 kb either side of a TSS. B) VEZF1 peak numbers in 20bp windows up to 1 kb either side of a TSS.

In order to take a closer look at VEZF1 enrichment at gene TSS and to compare different ChIP-seq datasets, further analyses were performed using the seqMINER tool (Ye *et al.*,

2011). SeqMINER allows ChIP-seq read density to be plotted relative to a set of chosen 'peaks'. Firstly, the density of all VEZF1, FLAG-VEZF1 and IgG ChIP-seq reads were plotted relative to the 26,341 VEZF1 peak summits identified in chapter 3 (Figure 4.3). As expected, VEZF1 ChIP-seq reads accumulated in a sharp symmetrical bell-shaped distribution over the defined VEZF1 peak summits (Figure 4.3). The FLAG-VEZF1 reads showed the same enrichment at the VEZF1 ChIP-seq peak summits, albeit with a lower signal to noise ratio due to lower ChIP yield, as discussed previously. The IgG negative control reads showed the background level of reads generated from non-specific ChIP enrichment (Figure 4.3). This analysis further demonstrates that peak finding accurately identified VEZF1 binding sites, that the α-VEZF1 and α-FLAG-VEZF1 ChIPs identify the same sites and that there are few/no false positive peaks contaminating the identified collection of peaks.

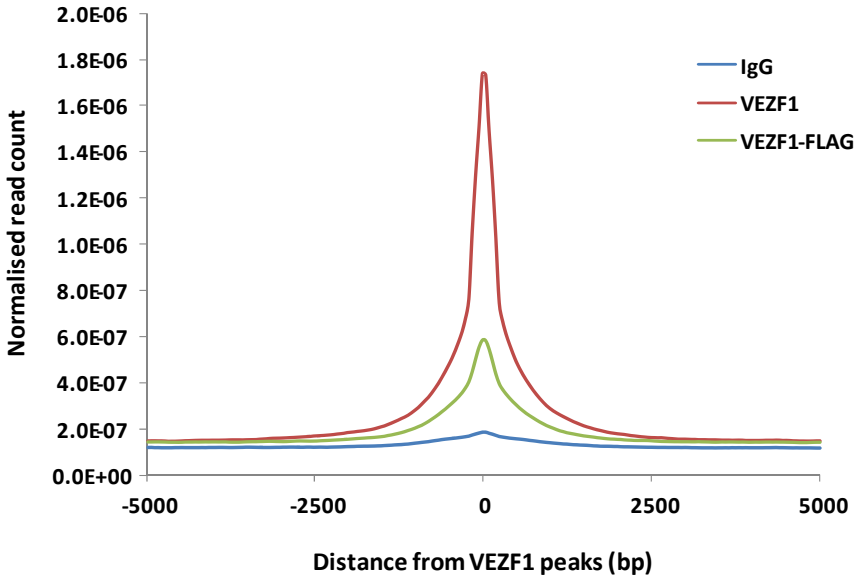


Figure 4.3 VEZF1 and FLAG-VEZF1 ChIP-seq read densities are maximal at the sites of VEZF1 ChIP-seq peaks.
SeqMINER analysis of VEZF1 (red), FLAG-VEZF1 (green) and IgG (blue) ChIP-seq read densities in regions 5 kb either side of all the 26,341 VEZF1 peak summits identified in chapter 3. Mean read densities (reads/500 bp windows) were normalised to total aligned read count for each ChIP-seq library.

The distribution of VEZF1, FLAG-VEZF1 and IgG ChIP-seq reads was then compared to the location of all RefSeq annotated gene TSSs. VEZF1 ChIP-seq reads accumulate directly over TSSs (Figure 4.4) correlating with the similar analysis performed using Nebula (Figure 4.2). FLAG ChIP-seq read distribution was very similar but with a lower signal to noise ratio. As before the IgG plot serves as a negative control showing the distribution of non-

specifically enriched reads (Figure 4.4). It is apparent from this data that VEZF1 is highly associated with gene transcription start sites.

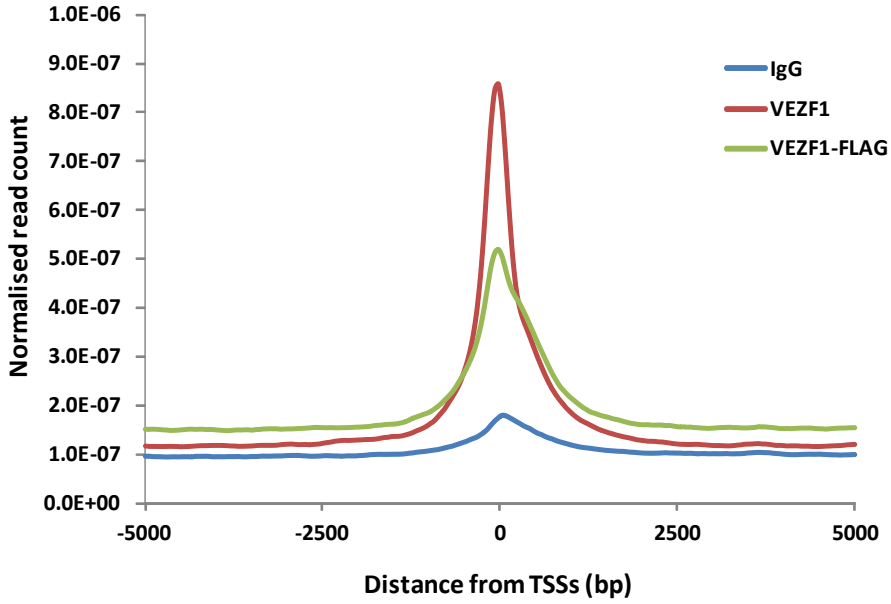


Figure 4.4 VEZF1 and FLAG ChIP-seq read densities are maximal at TSSs.
seqMINER analysis mapped ChIP-seq read densities relative to distance from TSSs. VEZF1 ChIP-seq reads from K562 cells and FLAG ChIP-seq reads from D3/6 K562 cells were maximal at TSSs. Sequence reads from the negative control IgG ChIP in K562 cells were not substantially enriched at VEZF1 peaks. Mean read densities (reads/50 bp windows) were normalised to total aligned read count for each ChIP-seq library.

4.3 The chromatin states at elements bound by VEZF1

4.3.1 The ENCODE ChromHMM chromatin state map

While the comparison of VEZF1 peaks to gene annotation can give a quick overview of VEZF1 gene association, it lacks key information about gene regulation. The gene annotation does not take account of alternative promoter usage in different cell types, nor does it include information relating to enhancer, silencer or insulator element locations. Indeed, the majority of distal gene regulatory elements remain to be identified and annotated (Dunham *et al*., 2012). We were therefore interested in comparing the locations of VEZF1 peak summits with the ChromHMM chromatin state map for K562 cells defined by the ENCODE project (Ernst and Kellis, 2010, Ernst *et al*., 2011). This map defines predicted promoter, transcribed, enhancer, insulator and heterochromatin states from high quality ChIP-seq datasets for H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K20me1 and CTCF (section 1.5.3). Prior to comparison with all the VEZF1 peaks, we compared the K562 ChromHMM track on the UCSC browser

to the *TAL1* gene locus, which has multiple well characterised promoters and enhancers. (Figure 4.5). The ChromHMM model accurately calls the alternative *TAL1* promoters p1a and p1b, the strong erythroid-specific enhancer +51. The haematopoietic stem cell enhancers +18.5 and +20 are correctly called, despite being weak in K562 cells. The +3 element is called as a strong promoter. It appears that the ChromHMM model will be useful in predicting gene regulatory functions that associate with VEZF1 sites at genomic locations that otherwise lack functional characterisation.

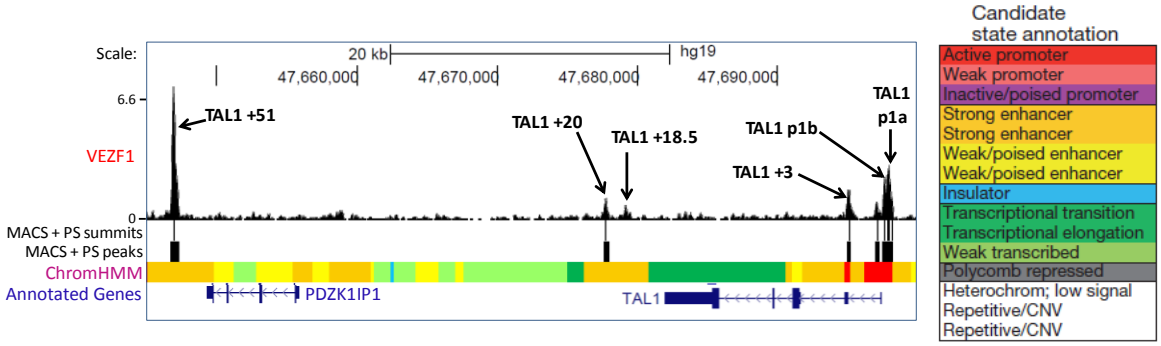


Figure 4.5 The ChromHMM chromatin state map accurately calls known regulatory regions of the *TAL1* locus in K562 cells.
UCSC genome browser view of VEZF1 ChIP-seq data mapped to human chr1 47,645,000-47,700,000 (hg19). VEZF1 ChIP-seq peaks and peak summits are represented by black boxes and vertical black lines respectively. The K562 ChromHMM chromatin state track is from ENCODE accession wgEncodeEH000790 (Ernst and Kellis, 2010, Ernst *et al*., 2011). Below the tracks is the UCSC gene annotation, where linked vertical blocks indicate protein-coding exons and arrows indicate the direction of transcription.

4.3.2 The chromatin state distribution of VEZF1 sites

The 26,341 VEZF1 peak summits defined in chapter 3 were overlapped with the ENCODE HMM chromatin state map for K562 cells using the genomic intersection tool in Galaxy (http://galaxyproject.org/) (Taylor *et al*., 2007). Single base peak summits were used to prevent the counting of more than one chromatin state per VEZF1 peak and should most accurately reflect the chromatin events at VEZF1 binding sites. This analysis revealed a striking association between VEZF1 binding and genomic elements that carry the chromatin hallmarks of promoters that are strong (11,284 sites), weak (3,042) or poised (293 sites) (Figure 4.6 A). Collected together, 55.4% of all the identified VEZF1 sites locate to promoter-associated elements (Figure 4.6 B). This is consistent with 55.6% of VEZF1 sites being located within 500 bp of transcription start sites (section 4.2.1).

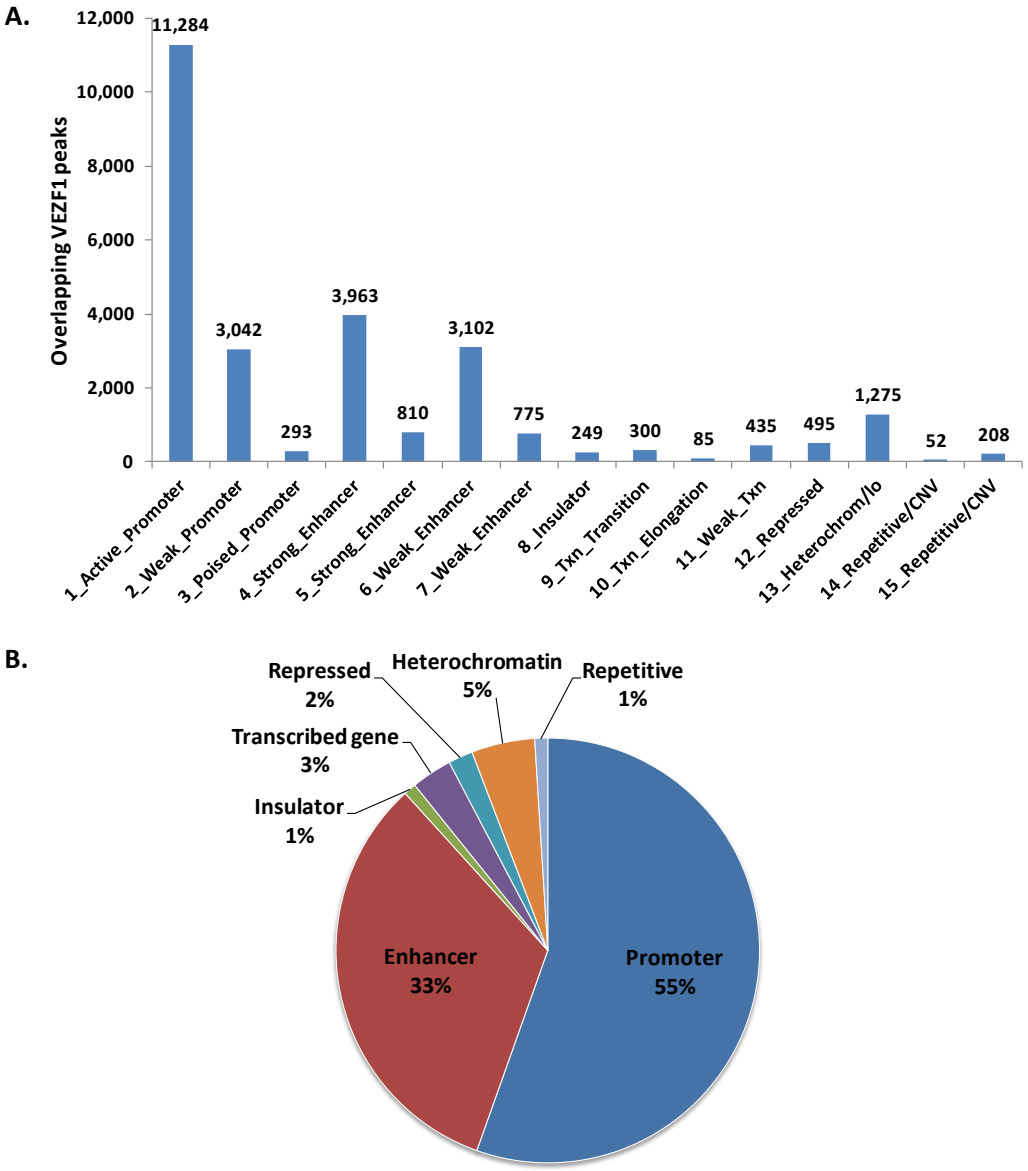


Figure 4.6 VEZF1 ChIP-seq peaks are highly associated with gene regulatory elements. A) The overlap between 26,341 VEZF1 peaks and 15 ChromHMM chromatin states for K562 cells (ENCODE UCSC accession wgEncodeEH000790). B) Summary of the chromatin states at VEZF1 sites, where all ChromHMM-defined classes of promoter were grouped together as were enhancers, transcribed genes and repetitive elements.

Furthermore, there is also a striking association between VEZF1 and genomic elements that carry the chromatin hallmarks of enhancers that are strong (3,963 + 810 sites) or weak (3,102 + 775 sites). Collected together, 32.8% of all the identified VEZF1 sites locate to 8,650 enhancer-associated elements (Figure 4.6 B). There were only minor associations with other chromatin state types. Only 85 sites locate to regions of transcriptional elongation, suggesting that VEZF1 does not play a major role in transcriptional pausing (section 1.8). There is a weak association with other genomic elements including insulator-like elements, to which only 249 out of the total 26,341 VEZF1 peaks mapped. We have noticed from genome browsing that many CTCF sites overlap other chromatin

states, especially enhancer states (not shown). The very low degree of overlap between VEZF1 peaks and the “insulator” chromatin state therefore under-reports the true levels of association between VEZF1 and CTCF.

4.3.3 Chromatin features at promoter-associated VEZF1 elements

The levels of various chromatin features at 14,619 VEZF1 peaks that overlap promoter-associated chromatin states (section 4.3.2) were analysed using seqMINER. The aim of this analysis was to both validate the promoter-associated chromatin state and to look at the relationship between VEZF1 ChIP-seq peaks and adjacent chromatin. In order to achieve this, ChIP-seq and DNase-seq read alignment files were downloaded from the UCSC genome browser. These data sets were loaded into seqMINER and the read density from 5 kb upstream to 5 kb downstream of the 14,619 VEZF1 peaks was determined. The average read profiles were standardised to allow comparison between different ChIP-seq experiments.

This analysis showed that there are high levels of the histone H3 modifications H3K4me3 and H3K27ac and the variant histone H2A.Z at promoter-associated VEZF1 elements (Figure 4.7). Conversely, H3K27me3 and H3K9me3, both associated with repressive chromatin, are not enriched. The H3K4me1 histone modification is present at low levels at promoter-associated VEZF1 peaks. These findings are entirely consistent with the chromatin signatures observed at transcriptionally active gene promoters (section 1.3.7). Interestingly, the levels of H2A.Z, H3K4me3 and H3K27ac are all depleted at the centre of the VEZF1 peaks (Figure 4.7) suggesting that VEZF1 binding sites are situated within nucleosome depleted regions (NDRs). These observations closely correlate with VEZF1 binding in close proximity to transcription start sites (section 4.2.2), which are characteristic NDRs (Lee *et al*., 2004, Iyer, 2012, Struhl and Segal, 2013).

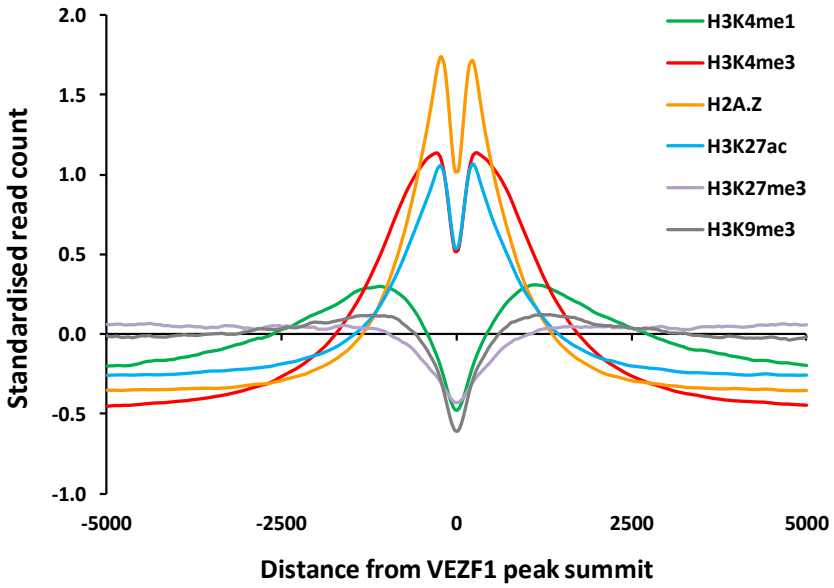


Figure 4.7 Chromatin signatures of promoter-associated VEZF1 peaks.
VEZF1 sites locate within nucleosome depleted regions of active promoters in K562 cells. Mean read density profiles of H3K4me1 (wgEncodeEH000046), H3K4me3 (wgEncodeEH000048), H2A.Z (wgEncodeEH001038), H3K27ac (wgEncodeEH000043), H3K27me3 (wgEncodeEH000044) and H3K9me3 (wgEncodeEH001040) ChIP-seq from 5 kb upstream to 5 kb downstream of the 14,619 VEZF1 peak summits that overlap ChromHMM promoter states. Standardised read counts represent Z-scored read densities in 50 bp windows.

Next, we studied the relationship between VEZF1 binding and the average transcription state at gene TSS. We made use of Affymetrix cDNA microarray data for K562 cells from our colleague Dr. David Vetrie, which identified a total of 11,714 genes as being expressed . We use the RefSeq coordinates for these genes in SeqMINER, such that the analysis focussed on their annotated TSS with the bodies of all the genes oriented to the right hand side of the TSS. Consistent with previous reports (section 1.3.7), this analysis found that the TSS of active genes are flanked by nucleosomes enriched in H2A.Z, H3K4me3 and H3K27ac (Figure 4.8 B). The consistent depletion in histone protein enrichment directly over the TSS denotes the nucleosome depleted region (NDR) which is found at the TSSs of active genes (Segal et al., 2006, Lee et al., 2007b). RNA polymerase II was observed to bind at the TSS of these genes, providing further indication of active gene transcription at these gene promoters (Figure 4.8 A). Furthermore, RNA-seq levels were elevated into the gene bodies, with high enrichments immediately downstream of the TSS. The H3K36me3 histone modification, which marks regions of active gene transcription, was also enriched within the bodies of expressed genes (figure 4.8 B). These observations are entirely consistent with the transcriptional activity of the 11,714 selected genes. VEZF1 ChIP-seq reads were seen to form a peak directly over these active

TSSs (Figure 4.8 A). It is apparent that VEZF1 binds at the nucleosome depleted TSSs of actively expressed genes in K562 cells.

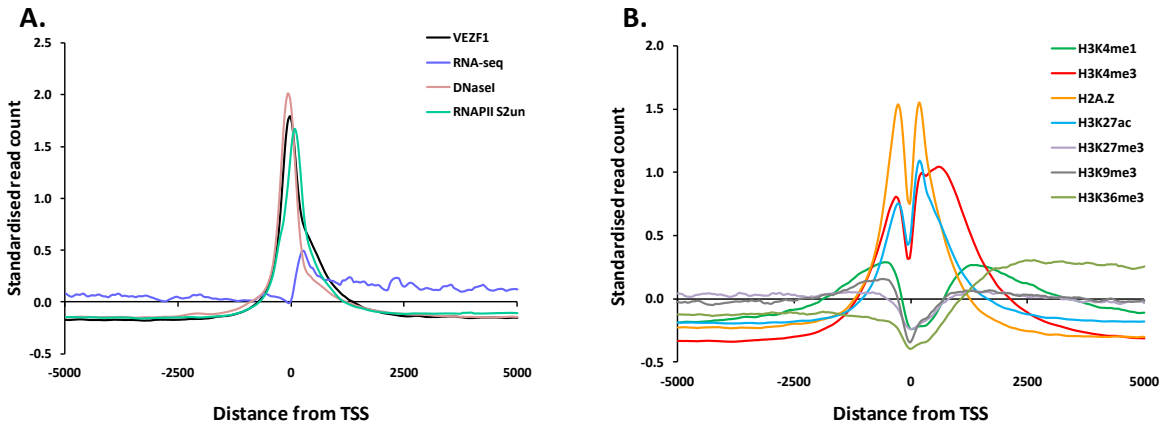


Figure 4.8 Chromatin signature over the TSS of actively expressed genes.
VEZF1 binds at the nucleosome depleted TSSs of actively expressed genes in K562 cells. Mean read density profiles from 5 kb upstream to 5 kb downstream of the TSS of 11,714 transcriptionally active genes in K562 cells. (A) VEZF1 and RNA Polymerase II (unphosphorylated CTD) ChIP-seq (wgEncodeEH000727), DNase-seq (wgEncodeEH000484) and RNA-seq (wgEncodeEH000484). (B) H3K4me1 (wgEncodeEH000046), H3K4me3 (wgEncodeEH000048), H2A.Z (wgEncodeEH001038), H3K27ac (wgEncodeEH000043), H3K27me3 (wgEncodeEH000044), H3K9me3 (wgEncodeEH001040) and H3K36me3 (wgEncodeEH000045)ChIP-seq are shown. Standardised read counts represent Z-scored read densities in 50 bp windows.

4.3.4 Relationship between VEZF1 binding to promoters and gene expression

In order to determine whether VEZF1 binding at gene promoters is associated with their transcriptional levels, VEZF1 ChIP-seq read density at TSS was compared to gene expression levels. Rather than plot average enrichment levels of ChIP-seq reads, as shown above, SeqMINER was used to generate heatmaps where each TSS is presented individually. 11,714 TSS were ordered by associated gene expression level, with the most highly expressed genes at the top of the heatmap plots and lowest expressers at the bottom. K562 RNA-seq data from the ENCODE project was plotted to validate the ordering by gene expression levels determined by Affymetrix microarrays. It can be seen from the resulting heatmap that RNA-seq reads are enriched in the proximity of the TSS for most of the elements analysed, with especially high levels for the top one third of expressed genes (Figure 4.9). RNA-seq reads also align to the gene bodies of the most expressed genes, with more RNA-seq reads at the most highly expressed genes. Furthermore, the levels of RNA polymerase II, open chromatin (FAIRE) and the promoter-

associated histone marks H3K4me3, H3K27ac and H2A.Z are all highest at the most expressed genes, with gradual decreases down to the lowest expressed genes (Figure 4.9). The levels of VEZF1 enrichment correlate directly with gene expression as the most highly expressed genes are those that are most highly enriched by VEZF1 (Figure 4.9). These data also show VEZF1 enrichment to directly correlate with RNA pol II enrichment and with levels of active chromatin marks. VEZF1 enrichment levels were most directly comparable to the degree of open chromatin determined by the FAIRE-seq assay. Together, these data demonstrate that VEZF1 binding at gene TSSs is closely associated with the chromatin opening and transcription activity of gene promoters.

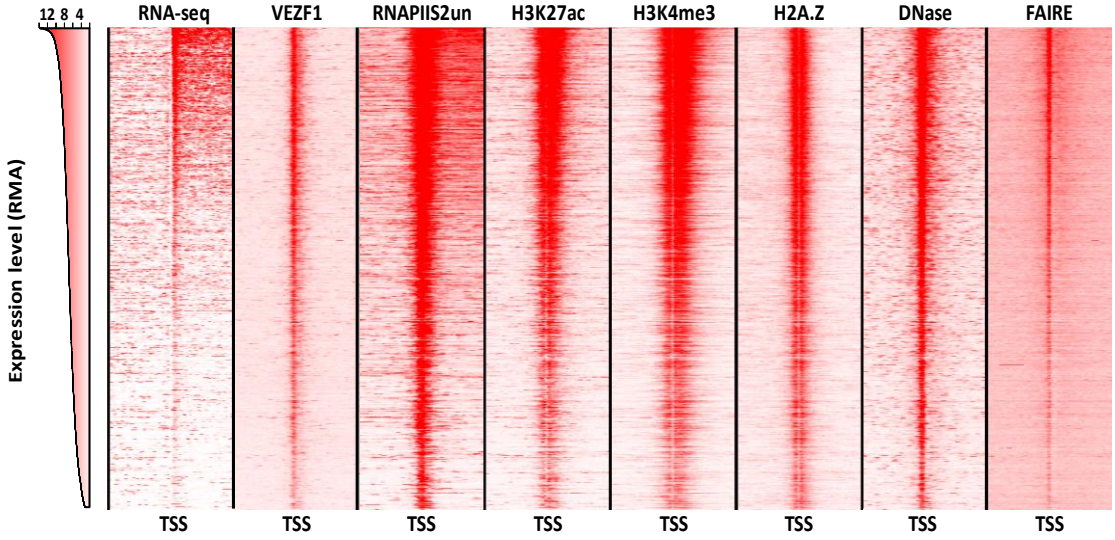


Figure 4.9 VEZF1 binding at TSS is associated with chromatin opening and transcription. Read density profiles of RNA-seq (wgEncodeEH000484), ChIP-seq of VEZF1, RNA Polymerase II (unphosphorylated CTD) (wgEncodeEH000727), H3K27ac (wgEncodeEH000043), H3K4me3 (wgEncodeEH000048) and H2A.Z (wgEncodeEH001038) with DNase-seq (wgEncodeEH000484) and FAIRE-seq (wgEncodeEH000531) from 5 kb upstream to 5 kb downstream of the TSS of 11,714 transcriptionally active genes in K562 cells. Read densities are shown as a heat map where red is highest and white is lowest. Reade densities for the 11,714 TSS are arranged as a stack and ordered by the expression level of the associated gene (RMA value from Affymetrix cDNA microarray analysis).

4.3.5 Chromatin features at enhancer-associated VEZF1 elements

The comparison of VEZF1 ChIP-seq peaks with the ENCODE ChromHMM chromatin state maps for K562 cells found that VEZF1 bound at 8,650 elements that carry the chromatin hallmarks of enhancer elements (section 4.3.2). I wished to study the chromatin features at VEZF1 sites associated with gene distal enhancers. We found that 2,205 of the ChromHMM enhancer-like elements were located within 1kb of an annotated TSS. We therefore analysed the levels of various chromatin features at the 6,355 enhancer-associated VEZF1 sites >1kb from a TSS using seqMINER.

The average read profile for H3K4me1 ChIP-seq showed this modification to be enriched in the nucleosomes at the enhancer-associated VEZF1 peaks (Figure 4.10). This is in contrast to the lack of H3K4me1 enrichment at promoter-associated VEZF1 sites and TSS (Figures 4.7 and 4.8). In addition, the H3K27ac modification was also found to be highly enriched at the enhancer sites (Figure 4.10). This chromatin profile of high [H3K27ac/H3K4me1] is consistent with previous studies of enhancer-associated histone modifications (section 1.3.7). H3K27ac enrichment at enhancer elements has been shown to differentiate active from poised enhancers (Creyghton *et al*., 2010). The high level of H3K27ac present at VEZF1-enriched enhancers therefore indicates the active state of these enhancer elements. H3K4me3 and the histone variant H2A.Z was also enriched at the VEZF1 enhancer sites, but there was no apparent enrichment of the repressive histone modifications H3K9me3 and H3K27me3 (Figure 4.10).

It is interesting to note that the levels of H3K4me1, H4K4me3, H3K27ac and H2A.Z are enriched in the nucleosomes either side of the VEZF1 peak, but are depleted over the site of VEZF1 binding (Figure 4.10). This profile strongly indicates that enhancers bound by VEZF1 also contain nucleosome depleted regions much like promoter elements.

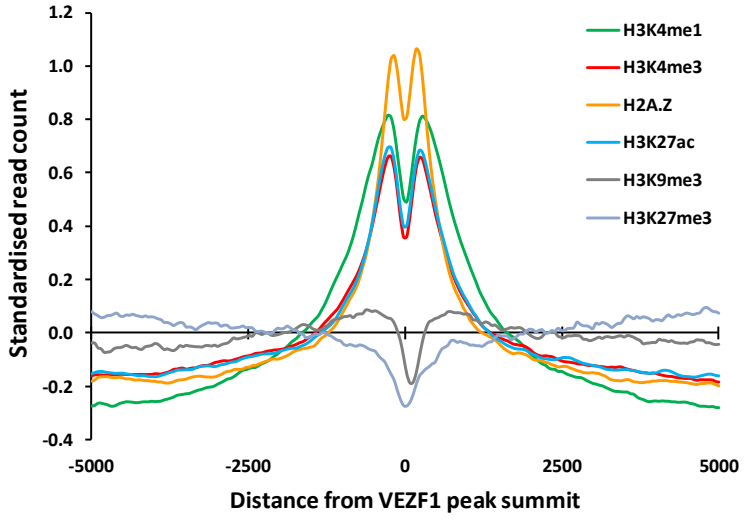


Figure 4.10 Chromatin signatures of enhancer-associated VEZF1 peaks.
VEZF1 sites locate within nucleosome depleted regions of active enhancers in K562 cells. Mean read density profiles of H3K4me1 (wgEncodeEH000046), H3K4me3 (wgEncodeEH000048), H2A.Z (wgEncodeEH001038), H3K27ac (wgEncodeEH000043), H3K9me3 (wgEncodeEH001040) and H3K27me3 (wgEncodeEH000044) and ChIP-seq from 5 kb upstream to 5 kb downstream of the 6,355 VEZF1 peak summits that overlap ChromHMM enhancer states located >1kb from an annotated TSS. Standardised read counts represent Z-scored read densities in 50 bp windows.

4.4 Co-binding transcription factors at VEZF1 sites

4.4.1. Co-binding of VEZF1 with general transcription factors at promoters

Gene regulatory elements tend to include clusters of transcription factor binding sites that act in concert to regulate chromatin remodelling, transcription complex loading and activity. I wished to gain some insight into which transcription factors share potential physical and functional relationships with VEZF1. The locations of all 26,321 VEZF1 ChIP-seq peaks or the 14,619 promoter-associated peaks were therefore compared to transcription factor binding site (TFBS) peaks collated from the ENCODE project. It was most practical to start by comparing VEZF1 peaks to the full ENCODE version 2 release of transcription factor ChIP-seq peaks. It should be cautioned that while this list contains ~2.7 million identified TFBS, it is not specific to K562 cells. A window of 200 bp around each VEZF1 peak summit was generated and analysed for overlap with any ENCODE TFBS using the genomic intersection tool in Galaxy (http://galaxyproject.org/) (Taylor *et al.*, 2007). The incidences of overlapping peaks were added and expressed as a percentage of VEZF1 binding events (Table 4.1).

This analysis found that VEZF1 binding frequently overlapped the binding of many well characterised general transcription factors. RNA polymerase II is the factor that most frequently interacts with VEZF1 elements, with an overlap with 93% of promoter-associated and 73% of all VEZF1 binding events (Table 4.1). TFIID components TAF1 and TBP and the transcription elongation factor pTEFb complex component cyclin T2 (CCNT2) all frequently interact with elements bound by VEZF1. These frequencies of overlap with general transcription factors are consistent with VEZF1 binding at sites of transcription initiation, such as gene TSS. The remainder of the factors that most frequently interact at VEZF1 sites are typically transcription factors with broad functions like, CMYC, EGR1, FOS/JUN, SP1, USF and YY1. GATA1, GATA2 and TAL1 are tissue-specific transcription factors that regulate erythroid-specific genes and are expressed in K562 cells. These factors were found to associate with between 3,800 and 4,500 VEZF1 elements (Table 4.1). Interestingly, these elements do not appear to be associated with promoter elements.

All 26428 sites			
Rank	Factor	Overlaps	Fraction
1	POL2	19164	72.5%
2	TAF1	16309	61.7%
3	HEY1	15049	56.9%
4	EGR-1	14107	53.4%
5	E2F6	13543	51.2%
6	ELF1	12873	48.7%
7	ZBTB7A	12507	47.3%
8	TBP	11892	45.0%
9	SIN3A	11140	42.2%
10	C-MYC	10659	40.3%
11	CCNT2	10095	38.2%
12	E2F1	9673	36.6%
13	YY1	8899	33.7%
14	CTCF	8593	32.5%
15	GABP	8368	31.7%
16	SP1	7926	30.0%
17	HMGN3	7695	29.1%
18	IRF1	7455	28.2%
19	MAX	7240	27.4%
20	NFKB	7182	27.2%
21	USF-1	6850	25.9%
22	PAX5	6569	24.9%
23	ETS1	6445	24.4%
24	NRSF	6182	23.4%
25	POL2(Ser2p)	6118	23.1%
26	ZNF263	5512	20.9%
27	PU.1	5308	20.1%
28	MXI1	5262	19.9%
29	USF1	5145	19.5%
30	CEBPB	5073	19.2%
31	CHD2	4690	17.7%
32	P300	4676	17.7%
33	HDAC2	4655	17.6%
34	AP-2γ	4507	17.1%
35	**GATA1**	4481	17.0%
36	RAD21	4470	16.9%
37	POU2F2	4382	16.6%
38	JUND	4377	16.6%
39	TCF12	4192	15.9%
40	EBF1	4154	15.7%
41	TAF7	4076	15.4%
42	ZNF143	4056	15.3%
43	NRF1	4034	15.3%
44	E2F4	4006	15.2%
45	TCF4	3901	14.8%
46	C-FOS	3857	14.6%
47	**TAL1**	3769	14.3%
48	**GATA2**	3747	14.2%
49	CTCFL	3639	13.8%
50	SMC3	3579	13.5%

All 14619 promoters			
Rank	Factor	Overlaps	Fraction
1	POL2	13521	92.5%
2	TAF1	13027	89.1%
3	HEY1	12375	84.7%
4	E2F6	10399	71.1%
5	EGR-1	10093	69.0%
6	ELF1	9563	65.4%
7	SIN3A	9408	64.4%
8	TBP	9339	63.9%
9	ZBTB7A	9170	62.7%
10	C-MYC	8313	56.9%
11	E2F1	8189	56.0%
12	CCNT2	7468	51.1%
13	YY1	7102	48.6%
14	GABP	6174	42.2%
15	CTCF	6097	41.7%
16	HMGN3	5939	40.6%
17	NFKB	5795	39.6%
18	IRF1	5772	39.5%
19	SP1	5761	39.4%
20	MAX	5405	37.0%
21	POL2(Ser2p)	5389	36.9%
22	PAX5	5374	36.8%
23	ETS1	5259	36.0%
24	USF-1	4619	31.6%
25	MXI1	4156	28.4%
26	ZNF263	4088	28.0%
27	NRSF	3987	27.3%
28	CHD2	3844	26.3%
29	TAF7	3559	24.3%
30	POU2F2	3541	24.2%
31	NRF1	3489	23.9%
32	AP-2γ	3467	23.7%
33	E2F4	3465	23.7%
34	ZNF143	3309	22.6%
35	PU.1	3225	22.1%
36	CEBPB	3118	21.3%
37	TCF12	3014	20.6%
38	JUND	2952	20.2%
39	E2F1	2895	19.8%
40	RAD21	2876	19.7%
41	CTCFL	2873	19.7%
42	TCF4	2873	19.7%
43	EBF1	2872	19.6%
44	P300	2843	19.4%
45	RFX5	2648	18.1%
46	HDAC2	2616	17.9%
47	GTF2F1	2598	17.8%
48	SRF	2535	17.3%
49	ZEB1	2512	17.2%
50	C-FOS	2447	16.7%

Table 4.1 Overlap between VEZF1 and other transcription factor binding events. Incidences of ChIP-seq peaks for VEZF1 (peak summits expanded to 200bp) that overlap with other factors, as defined by the ENCODE project (EncodeRegTfbsClusteredV2). The top 50 factors that most frequently overlap all 26,341 (left) or the 14,619 promoter-associated VEZF1 binding events (right) are shown. Known erythroid-specific transcription factors are highlighted in red.

An initial goal of this study was to identify novel insulator elements. The initial analysis of the overlap between VEZF1 peaks in K562 cells with CTCF peaks identified in all studies found 8,593 overlapping events (Table 4.1). In order to look more accurately at the potential co-binding of VEZF1 and CTCF in K562 cells, 26,341 VEZF1 peaks were compared with 31,849 CTCF peaks from K562 cells (wgEncodeEH002797, FDR>2). It was found that 2,759 VEZF1/CTCF sites overlapped when 200 bp around each peak summit was used. Most of these are promoter-associated, but 578 are gene distal. Some of these may be insulators akin to the chicken HS4 element.

In order to determine the relationship between VEZF1 and general transcription factor binding at TSS, the ChIP-seq read densities of these factors in K562 cells were profiled using seqMINER. The TSSs of 11,714 genes that are transcriptionally active in K562 cells were arranged by associated gene expression level as before. We selected SP1, YY1, CMYC, ELF1, EGR1, HMGN3, TBP and TAF1 for this analysis along with the unphosphorylated transcription initiation form of RNA polymerase II. These factors were chosen as they are well characterised and each bind to different classes of DNA sequence element at core promoters. It is evident from this analysis that VEZF1 enrichment of active gene promoters directly correlates with the levels of other general transcription factors (Figure 4.11). All of these factors are found to be most greatly enriched around the TSS of highly expressed genes with enrichment levels falling in conjunction with decreasing gene expression levels (Figure 4.11).

In order to gain an understanding of the average spatial position of each factor relative to TSS, the mean read density profiles of each factor in K562 cells were calculated. This analysis showed that while the average ChIP-seq profile of VEZF1 overlaps that of general transcription factors at TSS, the position of maximal enrichment differs for each factor (Figure 4.12). The maximal point of enrichment of each factor relative to the TSS was calculated and these values are presented in table 4.2.

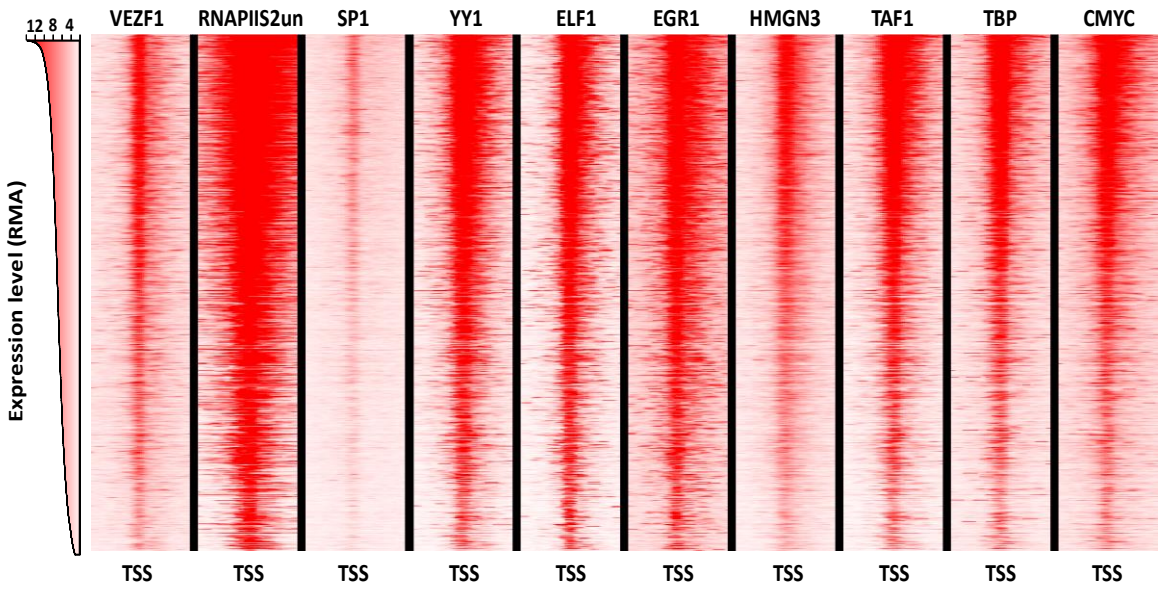


Figure 4.11 VEZF1 binds alongside other general transcription factors at transcriptionally active gene promoters.
Read density profiles of VEZF1, RNA Polymerase II (unphosphorylated CTD) (wgEncodeEH000727), SP1 (wgEncodeEH001578), YY1 (wgEncodeEH001623), ELF1 (wgEncodeEH001619), EGR1 (wgEncodeEH001646), HMGN3 (wgEncodeEH001863), TAF1 (wgEncodeEH001582), TBP (wgEncodeEH001825) and CMYC (wgEncodeEH002800) ChIP-seq from 5 kb upstream to 5 kb downstream of the TSS of 11,714 transcriptionally active genes in K562 cells. Read densities are shown as a heat map where red is highest and white is lowest. Read densities for the 11,714 TSS are arranged as a stack and ordered by the expression level of the associated gene (RMA value from Affymetrix cDNA microarray analysis).

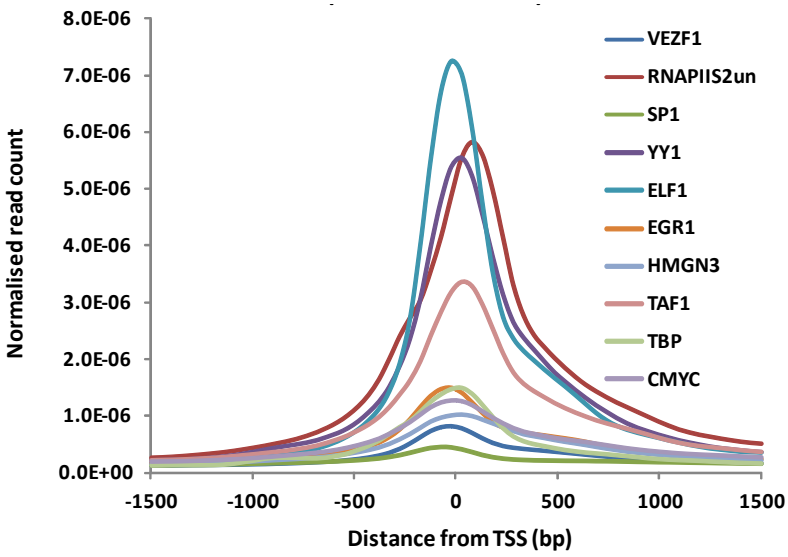


Figure 4.12 Transcription factor binding profiles at TSS.
Mean read density profiles of VEZF1, RNA Polymerase II (unphosphorylated CTD) (wgEncodeEH000727), SP1 (wgEncodeEH001578), YY1 (wgEncodeEH001623), ELF1 (wgEncodeEH001619), EGR1 (wgEncodeEH001646), HMGN3 (wgEncodeEH001863), TAF1 (wgEncodeEH001582), TBP (wgEncodeEH001825) and CMYC (wgEncodeEH002800) ChIP-seq from 1.5 kb upstream to 1.5 kb downstream of RefSeq TSS. Standardised read counts represent Z-scored read densities in 15 bp windows.

TF	Base position at which read density is at maximum
SP1	-65 bp
EGR1	-40 bp
VEZF1	-35 bp
ELF1	-15 bp
CMYC	-15 bp
TBP	+15 bp
YY1	+20 bp
HMGN3	+20 bp
TAF1	+40 bp
RNAPII	+85 bp

Table 4.2 Average binding positions of transcription factors at gene promoters.
The positions of maximal mean ChIP-seq read density for VEZF1 and general transcription factors relative to transcription start sites (rounded to 5 bp).

4.4.2 Co-binding of VEZF1 with transcription factors at enhancers

The sets of transcription factors which operate at gene distal enhancers tend to differ from those which function at core promoters. In order to gain insight into which transcription factors might operate with VEZF1 at enhancers, the locations of 8,650 enhancer-associated VEZF1 ChIP-seq peaks (section 4.3.2) were compared to transcription factor binding site peaks collated from the ENCODE project. The analysis was performed as described in the previous section, where the incidences of overlapping peaks were added and expressed as a percentage of VEZF1 binding events (Table 4.3).

This analysis found that VEZF1 binding at enhancer-associated elements frequently overlapped the binding of many well characterised general transcription factors. Similar to promoters, RNA polymerase II is the factor that most frequently interacts with VEZF1 enhancer sites, with an overlap 76% at enhancer-associated VEZF1 elements (Table 4.3). The binding of general transcription factors such as EGR1, TAF1, TBP, cyclin T2 (pTEFb), CMYC, SP1, USF1 and YY1 is also a regular feature (27-55% frequency) at enhancer-associated VEZF1 elements. It should be noted that 2,205 of the 8,560 elements with enhancer-like chromatin states studied here locate within 1kb of an annotated TSS (section 4.3.5). Much of the overlap with general transcription factors may therefore be a result of TSS activity. In contrast to promoter-associated elements (Table 4.1), the tissue-

specific transcription factors GATA1, GATA2 and TAL1 were found to bind at a subset of 25-30% of VEZF1 enhancers (Table 4.3).

8650 enhancers			
Rank	**Factor**	**Overlaps**	**Fraction**
1	POL2	6598	76.3%
2	EGR1	4723	54.6%
3	TAF1	4335	50.1%
4	ZBTB7A	4133	47.8%
5	E2F6	3899	45.1%
6	ELF1	3887	44.9%
7	HEY1	3697	42.7%
8	CCNT2	3328	38.5%
9	TBP	3245	37.5%
10	C-MYC	2800	32.4%
11	**TAL1**	2608	30.2%
12	GABP	2585	29.9%
13	SP1	2529	29.2%
14	USF-1	2417	27.9%
15	CTCF	2409	27.8%
16	SIN3A	2360	27.3%
17	JUND	2359	27.3%
18	YY1	2307	26.7%
19	NRSF	2296	26.5%
20	HMGN3	2293	26.5%
21	**GATA2**	2260	26.1%
22	E2F1	2216	25.6%
23	**GATA1**	2196	25.4%
24	IRF1	2184	25.2%
25	PU.1	2120	24.5%
26	MAX	2093	24.2%
27	HDAC2	2079	24.0%
28	CEBPB	2001	23.1%
29	E2F6	1903	22.0%
30	P300	1855	21.4%
31	NFKB	1696	19.6%
32	CJUN	1597	18.5%
33	ETS1	1583	18.3%
34	CFOS	1567	18.1%
35	ZNF263	1566	18.1%
36	PAX5	1501	17.4%
37	RAD21	1421	16.4%
38	EBF1	1357	15.7%
39	MXI1	1322	15.3%
40	USF1	1317	15.2%
41	TCF12	1293	14.9%
42	AP-2γ	1246	14.4%
43	TCF4	1150	13.3%
44	JUNB	1147	13.3%
45	CHD2	1146	13.2%
46	SMC3	1141	13.2%
47	FOXA1	1137	13.1%
48	POL2(Ser2p)	1102	12.7%
49	FOSL1	1092	12.6%
50	USF2	1079	12.5%

Table 4.3 Overlap between VEZF1 and other transcription factor binding events at enhancers.

Incidences of ChIP-seq peaks for VEZF1 (peak summits expanded to 200bp) that overlap with other factors, as defined by the ENCODE project (EncodeRegTfbsClusteredV2). The

top 50 factors that most frequently overlap 8,650 enhancer-associated VEZF1 binding events are shown. Known erythroid-specific transcription factors are highlighted in red.

We wished to look more closely at the relationship between VEZF1 and transcription factors associated with enhancer activity (P300, BRG1, CEBPβ) and erythroid-specific gene regulatory elements (GATA1, TAL1, NFE2 and BACH1) (Fujiwara *et al*., 2009, Hu *et al*., 2011). SeqMINER clustering of ChIP-seq read densities for these factors were profiled at 6,355 enhancer-like VEZF1 elements located >1kb from an annotated TSS. The analysis was set to identify 10 clusters, each of which reveals groups of elements with different transcription factor associations. VEZF1 was found to bind with comparable intensity to all of the 6,355 elements studied (Figure 4.13). P300, a histone acetyltransferase responsible for depositing H3K27ac marks (Jin *et al*., 2011), is bound at all of the VEZF1 enhancers (Figure 4.13) which is consistent with the high level of H3K27ac at these elements (Figure 4.10). The VEZF1 enhancers are also bound by RNA polymerase II.

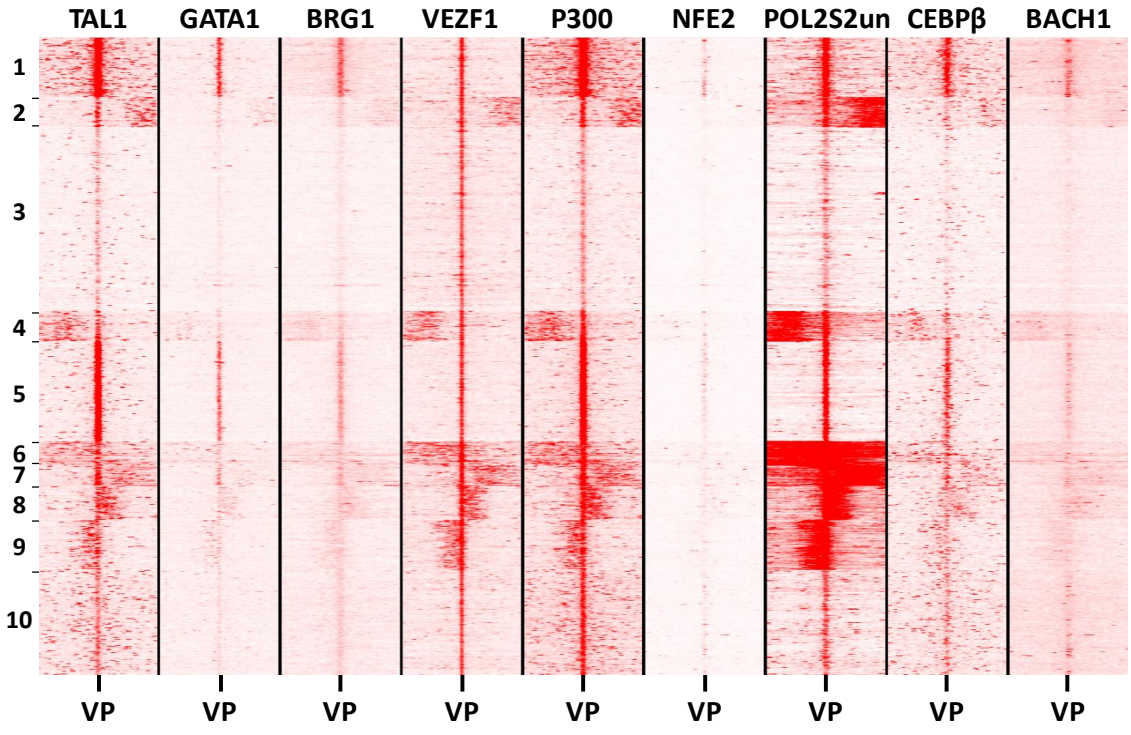
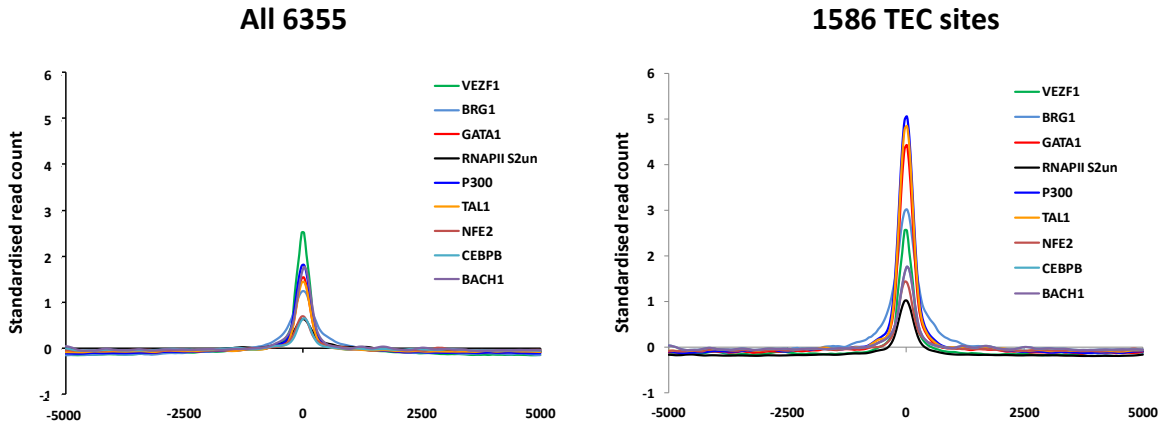


Figure 4.13 VEZF1 and erythroid transcription factor co-binding at enhancer elements. Read density profiles of TAL1 (wgEncodeEH001824), GATA1 (wgEncodeEH000638), BRG1 (wgEncodeEH000724), VEZF1, P300 (wgEncodeEH002834), NFE2 (wgEncodeEH000624),RNA Polymerase II (unphosphorylated CTD) (wgEncodeEH000727), CEBPβ (wgEncodeEH002346) and BACH1 (wgEncodeEH002846) ChIP-seq from 5 kb upstream to 5 kb downstream of 6,355 VEZF1 peak summits that both overlap ChromHMM enhancer states and are located >1kb from annotated TSS. Read densities are shown as a heat map where red is highest and white is lowest.

The binding of the other factors studied varies, where four patterns of binding are observed. The first group totalling 1,280 elements (clusters 6-9, Figure 4.13) appear to be TSS as judged by the high levels of RNA pol II extending from the VEZF1 peak summits. However, an interesting group of 1,586 elements (clusters 1 and 5, Figure 4.13) are bound by the erythroid-specific transcription factors GATA1, TAL1 and NFE2 in addition to the enhancer associated transcription factors P300, BRG1 and CEBPβ. GATA1 and TAL1 are known to frequently co-operate at erythroid gene regulatory elements in a TAL1-Erythroid-Complex (TEC) (Wadman *et al*., 1997). A third group of 1,356 elements is also bound by TAL1 at low levels (clusters 2 and 10, Figure 4.13), but generally lacks binding of GATA1, NFE2, CEBPβ or BACH1. Finally, a group of 2,133 elements (clusters 3 and 4) lacks binding of any of the cell-type specific transcription factors studied, indicating that VEZF1's association with enhancer elements is probably not exclusive to erythroid-specific elements.

The mean ChIP-seq read density profiles for the transcription factors studied were compared between all 6,355 TSS-distal VEZF1 enhancers and the 1,586 "TEC enhancers" identified from clustering. This analysis showed that the GATA1, TAL1, P300 and BRG1 transcription factors were all substantially enriched at the TEC enhancers compared to all the TSS-distal VEZF1 enhancers studied (Figure 4.14). The levels of NFE2, CEBPβ and RNA polymerase II were also substantially increased. Taken together, these findings indicate that the 1,586 elements identified in this analysis are bound by erythroid-specific transcription factors, probably collectively as a TEC complex. These elements have enhancer-associated chromatin signatures and are bound by known enhancer factors such as P300, BRG1 and CEBPβ. Consistent with this, the 1,586 element group includes the well characterised erythroid-specific enhancers of the TAL1 (+51 element), HBA (-8, -33 and -40 elements) and HBB (HS1, HS2 and HS3 elements) genes.



ChIP target	Peak Maxima (Z score) All 6355	1586 TEC
VEZF1	2.51	2.56
RNAPII S2un	0.63	1.01
GATA1	1.53	4.43
TAL1	1.43	4.85
P300	1.81	5.06
BRG1	1.23	3.02
CEBPB	0.65	1.76
NFE2	0.68	1.44
BACH1	1.76	1.76

Figure 4.14 Transcription factor binding at 1,586 erythroid enhancer elements.
Mean read density profiles of TAL1 (wgEncodeEH001824), GATA1 (wgEncodeEH000638), BRG1 (wgEncodeEH000724), VEZF1, P300 (wgEncodeEH002834), NFE2 (wgEncodeEH000624),RNA Polymerase II (unphosphorylated CTD) (wgEncodeEH000727), CEBPβ (wgEncodeEH002346) and BACH1 (wgEncodeEH002846) ChIP-seq 5 kb either side of either 6,355 TSS VEZF1 enhancers or 1,586 TEC-VEZF1 enhancers. Standardised read counts represent Z-scored read densities in 50 bp windows. The peak maxima for each read profile is tabulated. Increases in transcription factor enrichment in the TEC enhancer group are highlighted red.

4.5 Discussion

The aim of this chapter was to gain further insight into the types of regulatory processes that VEZF1 is involved in. This was to be accomplished by determining: i) the locations of VEZF1 binding sites relative to genes and gene regulatory elements, ii) the chromatin states at VEZF1-bound elements and iii) the TFs which co-bind with VEZF1.

As discussed in chapter 3, peak finding following VEZF1 ChIP-seq identified a total of 26,341 VEZF1 enrichment peaks across the K562 genome. Analysis of the locations of these peaks relative to the RefSeq human genome annotation and the ChromHMM chromatin state map for the K562 cell line revealed a striking correlation between VEZF1 binding and promoter and enhancer regulatory elements, with 55 % of VEZF1 peaks (14,619 peaks) locating to promoters and 33% (8,650 peaks) locating to enhancers. Furthermore, the distribution of VEZF1 ChIP-seq reads at promoter elements is highly concentrated around TSSs and was found to peak at an average distance of -35 bp from the TSS of VEZF1-associated promoters. These findings indicate a role for VEZF1 in the regulation of gene expression.

VEZF1 peaks map to NDRs at both promoter and enhancer elements. The promoter-associated NDRs to which VEZF1 maps are highly enriched on either side by the active chromatin marks H2A.Z, H3K4me3 and H3K27ac and are depleted in the repressive H3K27me3 and H3K9me3 marks. The enhancer-associated NDRs to which VEZF1 maps are also surrounded by nucleosomes enriched in the active chromatin marks H2A.Z and H3K4me3 as well as H3K27ac and H3K4me1 which are reported as being markers of active enhancer elements. VEZF1-associated enhancers are also depleted in repressive histone modifications. The association of VEZF1 with the TSSs of genes that are actively expressed in K562 cells combined with the finding that levels of VEZF1 enrichment at active TSSs positively correlate with gene expression levels, RNA pol II enrichment and levels of active chromatin marks, further supports a link between VEZF1 binding and active gene expression.

Analysis of VEZF1 co-binding factors identified RNA pol II as the TF which most frequently interacts with VEZF1-associated promoter elements. This analysis also revealed VEZF1-bound promoters to be co-occupied with a number of general TF proteins whose level of enrichment positively correlate with those of VEZF1. However a major limitation of this

analysis lies in the source of the TFBS locations. The full ENCODE version 2 release of TF ChIP-seq peaks was used to identify sites of TF binding proximal to VEZF1 ChIP-seq peaks, however this release contains the locations of ~2.7 million TFBSs identified by ChIP-seq of 119 TFs across 77 cell types and is therefore not specific to binding events in K562 cells. A proportion of the 'co-binding' events between VEZF1 and general TFs is therefore likely to be misleading as these proteins may bind the same genomic element but in different cell types. This co-binding screen is also limited by the accuracy of peak calling within the ChIP-seq data sets used to generate the ENCODE version 2 release of TFBSs. Examination reveals peak calling in many tracks to be accurate, but there is evidence of false positive and false negative peak calling and poor peak definition in some datasets. Such poor peak calling performances may affect the identification of sites of co-occupancy with VEZF1 and other TFs. This co-occupancy analysis was useful as an initial screen to identify proteins which may co-bind with VEZF1 at different types of genomic element. Unfortunately, it is not possible to extract K562-specific TFBS data from the ENCODE version 2 release of TFBSs. Further investigation will be required to study the co-occupancy of VEZF1 and specific proteins of interest across the K562 cell genome. In order to achieve this, publicly available ChIP-seq reads for factors of interest should be aligned to the hg19 human reference genome and peak calling performed using MACS and PeakSplitter softwares. Enrichment overlaps between VEZF1 and factors of interest can then be studied and the juxtaposition of binding motifs investigated. This type of analysis can be used to reveal co-operativity between TFs at specific types of genomic elements.

A similar analysis of VEZF1 co-binding factors at enhancer elements identified RNA pol II as being the TF which most frequently interacts with VEZF1-associated enhancers. A number of general TFs were also revealed to bind genomic DNA near VEZF1 peaks as were the tissue-specific TFs GATA1, GATA2 and TAL1. A group of 1,586 VEZF1 enrichment peaks were discovered to associate with enhancer elements which appear to be bound by the TEC complex. The TEC is an erythroid cell-specific TF complex which functions to regulate expression of a group of erythroid-specific genes (Wadman *et al*., 1997, Lahlil *et al*., 2004, Xu *et al*., 2007, Dhami *et al*., 2010). This group of VEZF1-enriched TEC-associated enhancers were co-occupied by the erythroid-specific gene regulatory elements GATA1, TAL1, NFE2 and BACH1 in addition to the general active enhancer associated TFs BRG1, P300 and CEBPβ.

It can be concluded from the findings presented in this chapter that VEZF1 primarily binds NDRs within active promoter and enhancer elements. Levels of VEZF1 enrichment are seen to positively correlate with those of co-binding TFs and a specific subgroup of VEZF1-associated enhancers appears to be comprised of TEC-associated erythroid specific enhancer elements.

Chapter 5

Definition of a VEZF1 DNA-binding consensus motif

5.1 Introduction

In chapter 3, ChIP-seq analysis identified 26,341 peaks of VEZF1 enrichment across the genome of human K562 cells. While the average width of the VEZF1 ChIP-seq peaks identified by MACS analysis was 433 bp, the analysis did generate peak summits where VEZF1 is most likely to bind. However, the specific DNA motifs with which VEZF1 interacts remain undefined. The bioinformatic analysis of chromatin and gene regulatory events at VEZF1 peaks presented in Chapter 4 found that ~55% of VEZF1 sites locate to TSSs at promoters, while ~33% of sites locate to enhancers. Around 1,600 of these enhancers are bound by erythroid-specific factors. A previous ChIP analysis of VEZF1 binding in different cell types found VEZF1 enrichment at housekeeping promoters in all cell types studied, whereas VEZF1 binding at erythroid-specific promoters and enhancers was restricted to erythroid cells (section 1.9) (Strogantsev, 2009). The fact that VEZF1 interacts with different regulatory element types and that binding can be either constitutive or cell-type specific suggests that VEZF1 may interact with different DNA motifs.

Chicken VEZF1 has been found to interact with multiple sites within the β-globin gene locus *in vitro*. Originally known as Beta Globin Protein 1 (BGP1), chicken VEZF1 interacts with a minimum of seven dG.dC bases within a long homopolymeric string upstream of the β^A globin gene promoter *in vitro* (Lewis *et al*., 1988, Clark *et al*., 1990) and has also been shown to interact with three sites within the HS4 β-globin insulator. One of these sites, FI, contains a homopolymeric 9(dG.dC) string, but the FIII and FV sites each have two shorter dG.dC runs (Dickson *et al*., 2010). The binding of human VEZF1 to DNA elements in the core promoters of the *IL3* and *EDN1* genes have also been characterised. These sites are G-rich, but also lack a minimum of seven contiguous dG.dC bases (Koyano-Nakagawa *et al*., 1994, Aitsebaomo *et al*., 2001). The VEZF1 binding motifs identified in these studies are shown in Table 5.1. It is clear that while there is a relationship between these motifs, no clear binding site consensus can be identified from this small number of sites.

Genomic Element	VEZF1-binding motif
β-Adult promoter	CGGGGGGGGGGGGGGGT
HS4 FI	CTTTGGGGGGGGGCTGTC
HS4 FIII	TCGGGGATGCGGGGGGAG
HS4 FV	TGGGGGATACGGGGAAAA
IL3 promoter	GGCGGGGGGAGGTGGTGG
EDN1 promoter	GACAAATGGGGGTGAGAT

Table 5.1 VEZF1 binding motifs identified in previous studies.
dG nucleotides within VEZF1 binding motifs are shown in red. Refer to text for references.

The analysis of enriched DNA motifs within VEZF1 ChIP-seq peaks should allow the definition of a robust consensus sequence. Studies in this chapter therefore aim to:

- Determine a consensus DNA binding site motif, or motifs, for VEZF1 from ChIP-seq data using bioinformatic tools
- Identify putative VEZF1 recognition motifs within ChIP peaks
- Analyse the relative DNA-binding affinity of VEZF1 *in vitro*
- Develop a model of VEZF1 DNA-binding affinity and its relationship to gene regulatory mechanisms

5.2 *ab initio* prediction of VEZF1 DNA-binding specificity

VEZF1 contains six Cys_2-His_2 (C2H2) zinc finger (ZF) domains, but no other recognisable DNA-binding motifs. Protein structure studies and mutagenesis have resulted in the development of a generic C2H2 DNA-binding model, which maps the exact interactions between nucleotides and specific ZF domain residues (Pavletich and Pabo, 1991, Elrod-Erickson *et al*., 1998). This model, in combination with the experimentally defined sequence specificity of C2H2 factors allows for *ab initio* prediction of the DNA-binding specificity of uncharacterised C2H2 proteins.

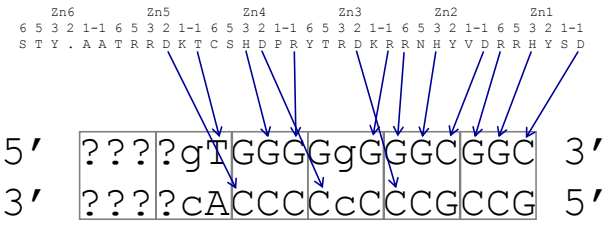
We firstly compared the VEZF1 protein sequence to a straightforward C2H2 code based on mutagenic studies of EGR1 (zif268) (Wolfe *et al*., 2000). This analysis predicted that the optimal recognition sequence for VEZF1 would be gTGGGGgGGGCGGC (Figure 5.1B). Most C2H2 domains studied have a canonical arrangement of tandem C2H2 domains. Given that ZF1 of VEZF1 has a non-canonical spacing with respect to the other ZF domains (figure 5.1A), the model might not accurately predict the potential specificity of this domain. The specificity of ZF2-ZF6 would therefore be gTGGGGgGGG. Next, we used a more sophisticated *ab initio* prediction tool that used a probabilistic model built from 455 C2H2 protein–DNA pairs from the TRANSFAC 7.3 database of experimentally defined sequence specificities (Kaplan *et al*., 2005). This analysis found that the optimal recognition sequence for VEZF1 would be GGGGgGGGcg (Figure 5.1C). Both of these predictions are highly convergent and indicate that an optimal VEZF1 recognition motif consists of homopolymeric strings of eight dG.dC bases. It is apparent from both models that the probability of a central guanosine base is weaker than the surrounding bases. This suggests that the central base position is not as critical for VEZF1 binding as surrounding positions and that other bases might be tolerated at this position.

A.

>Hs_VEZF1
MEANWTAFLFQAHEASHHQQQAAQNSLLPLLSSAVEPPDQKPLLPIPITQKPQGAPETLKDAIGIKKEKP
KTSFV**CTYCSKAFRDSYHLRRHESCH**TGIKLVSRPKKTPTTVVPLISTIAGDSSRTSLVSTIAGILSTVT
TSSSGTNPSSSASTTAMPVTQSVKKPSKPVKKNHA**CEMCGKAFRDVYHLNRHKLSH**SDEKPFE**CPICNQR**
FKRKDRMTYHVRSHEGGITKPYT**CSVCGKGFSRPDHLSCHVKHVH**STERPFK**CQTCTAAFATKDRLRTHM**
VRHEGKVS**CNICGKLLSAAYITSHLKTH**GQSQSINCNTCKQGISKTCMSEETSNQKQQQQQQQQQQQQQH
VTSWPGKQVETLRLWEEAVKARKKEAANLCQTSTAATTPVTLTTPFSITSSVSSETMSNPVTVAAAMSMR
SPVNVSSAVNITSPMNIGHPVTITSPLSMTSPLTLTTPVNLPTPVTAPVNIAHPVTITSPMNLPTPMTLA
APLNIAMRPVESMPFLPQALPTSPPW

B.

Manual alignment
using C2H2 zinc finger recognition code



C.

ab initio prediction
(C2H2 zinc finger database)

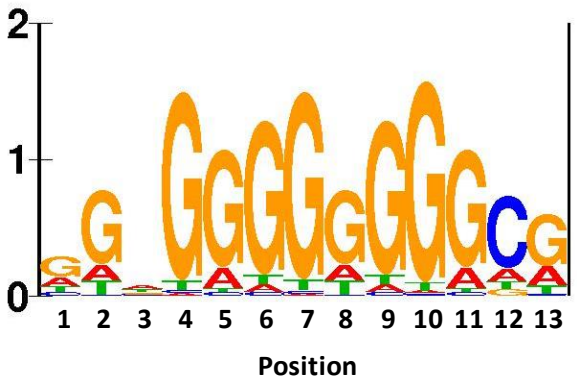


Figure 5.1 VEZF1 is predicted to prefer a homopolymeric deoxyguanosine motif.
A) The amino acid sequence of human VEZF1, C2H2 domains are highlighted in red. B) Prediction of the preferred DNA binding specificity of VEZF1 derived from comparison to a C2H2 DNA recognition code (Wolfe *et al*., 2000). C) Prediction of the preferred DNA binding specificity of VEZF1 derived from a probabilistic model (http://compbio.cs.huji.ac.il/Zinc, (Kaplan *et al*., 2005)).

5.3 Identification of enriched DNA sequence motifs at VEZF1 ChIP-seq peak summits

The discovery of regulatory DNA motifs, such as TF binding motifs, has been a major focus of biologists and bioinformaticians over recent years. The identification of TF binding motifs is important to achieving a better understanding of the regulatory mechanisms in which TFs function. The great number of computational motif discovery tools that have been developed and are publicly available for use reflects the complexity of the challenge that creating these tools poses. With the advent of ChIP-seq and other high throughput sequencing techniques, motif discovery tools are required which are capable of efficiently processing the huge amounts of data generated by these techniques.

5.3.1 Motif discovery using MEME

MEME (Multiple EM for Motif Elicitation) is the most commonly used motif discovery tool due to its ease of use via the MEME web portal (http://meme.nbcr.net/meme/, (Machanick and Bailey, 2011) . MEME was used to search for enriched DNA sequence motifs of up to 30 bp in length within the 100 bp surrounding VEZF1 ChIP-seq peak summits. Prior to motif discovery, the 26,341 VEZF1 ChIP-seq peak summits were ranked by read enrichment levels normalised against non-immune IgG ChIP-seq enrichments at the same genomic locations.

Motif discovery was performed for groups of 1000 VEZF1 ChIP-seq peaks based on their peak strength. MEME analysis was performed on (1) the 1000 most highly enriched VEZF1 peaks followed by peaks ranked 5000-6000 (2), 10,000-11,000 (3), 15,000-16,000 (4), 20,000-21,000 (5) and 25,000-26,000 (6). The aim was identify the most optimal VEZF1 motif from the top 1000 VEZF1 peaks followed by potentially divergent motifs as VEZF1 enrichments decrease. Eight sequence motifs were generated for each group. The highest ranking motif was found in 70 to 80 % of the 1000 elements tested in most of the groups of peaks (Figure 5.2). The additional motifs discovered were unlikely to represent VEZF1 motifs as they were found in a minor fraction of elements tested (4.8 % of input on average, not shown).

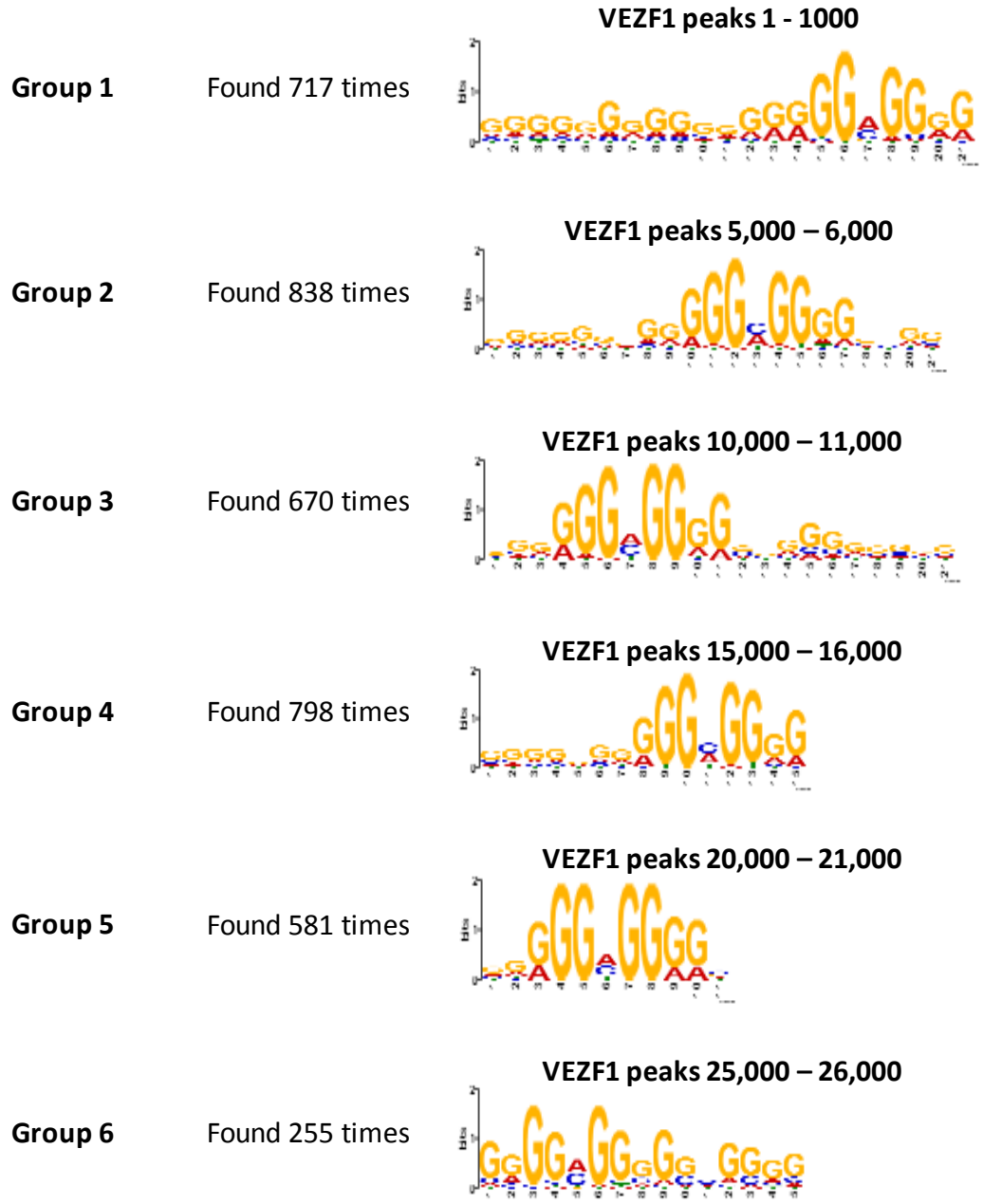


Figure 5.2 Motif discovery using MEME identifies G-rich motifs.
Motif discovery was performed using the 100 bp of DNA sequence surrounding VEZF1 ChIP-seq peaks as input for MEME. Sequences were analysed in groups of 1000 based on relative ChIP-seq enrichment levels. Weblogo presentations of the discovered position weight matrices of base enrichment are shown, where the height of each letter indicates the relative occurrence of nucleotides in enriched motifs.

The most highly represented DNA motifs identified by MEME for each group of 1000 VEZF1 ChIP-seq peaks analysed are G-rich. A common pattern of gGGnGGgg was apparent within groups 2, 3, 4 and 5, however this pattern is less apparent within the motifs discovered from groups 1 or 6 (Figure 5.2). Most of the DNA motifs discovered by MEME are very long in length when compared to previously characterised VEZF1 binding sites

(Table 5.1) or the motif predicted from the protein amino acid sequence (Figure 5.1). The DNA motif discovered from the 1000 most enriched VEZF1 sites is largely comprised of a long string of poorly enriched G bases that are consistently underlain by A bases (Figure 5.2, group 1). I am therefore concerned that MEME has had difficulty in aligning discovered G-rich motifs when forming a consensus. Another unexpected finding from these MEME analyses was that no obvious motif change was apparent between the highest and lowest VEFZ1-enriched sites.

Further constraints of the MEME programme are that it is very slow in performing motif finding analyses and that the MEME-ChIP tool actually only uses 600 sequences to generate a motif, i.e. in analyses where 1000 DNA sequences are used as input MEME-ChIP randomly selects 600 sequences and uses them to perform its analysis, this may lead to specific motifs being missed or under-/over-reported.

5.3.2 Motif discovery using POSMO

5.3.2.1 Motif discovery in all VEZF1 ChIP-seq peaks

We decided to make use of the recently developed motif discovery programme POSMO as it has been reported to be more efficient than iterative discovery tools like MEME (Ma *et al*., 2012). POSMO incorporates two advances over existing motif discovery methods. First is that POSMO uses a motif occurrence model that reflects the fact that transcription factor motifs typically cluster around ChIP-seq peak summits. This approach is computationally efficient, permitting analysis of full ChIP-seq datasets in minutes. Second, POSMO incorporates a word clustering that improves the alignment of discovered motifs when forming position weight matrices (PWMs) that describe a discovered motif.

POSMO can be used to search for motifs of 7 to 9 bp in length with the width of the DNA sequence around ChIP-seq peak summits set by the user. Following comparisons of different settings combinations it was concluded that using a 200 bp window around ChIP-seq peak centres for motif finding generated the most consistent results between 8-mer and 9-mer motif searches and subsequent analyses were performed searching for 9-mer motifs in a 200 bp window around VEZF1 peaks. Figure 5.3 shows the most represented 8 and 9-mer motifs discovered by POSMO analyses of a 200 bp window surrounding all of the total 26,341 VEZF1 ChIP-seq peak summits. Similar to outputs of

the prediction models and MEME analysis, homopolymeric strings of dG residues were apparent with a central residue (position 6 in the POSMO web logos) being less prevalent (Figure 5.3). Strikingly, POSMO discovered a far greater number of motif occurrences than MEME, with an average of approximately two occurrences for each VEZF1 peak.

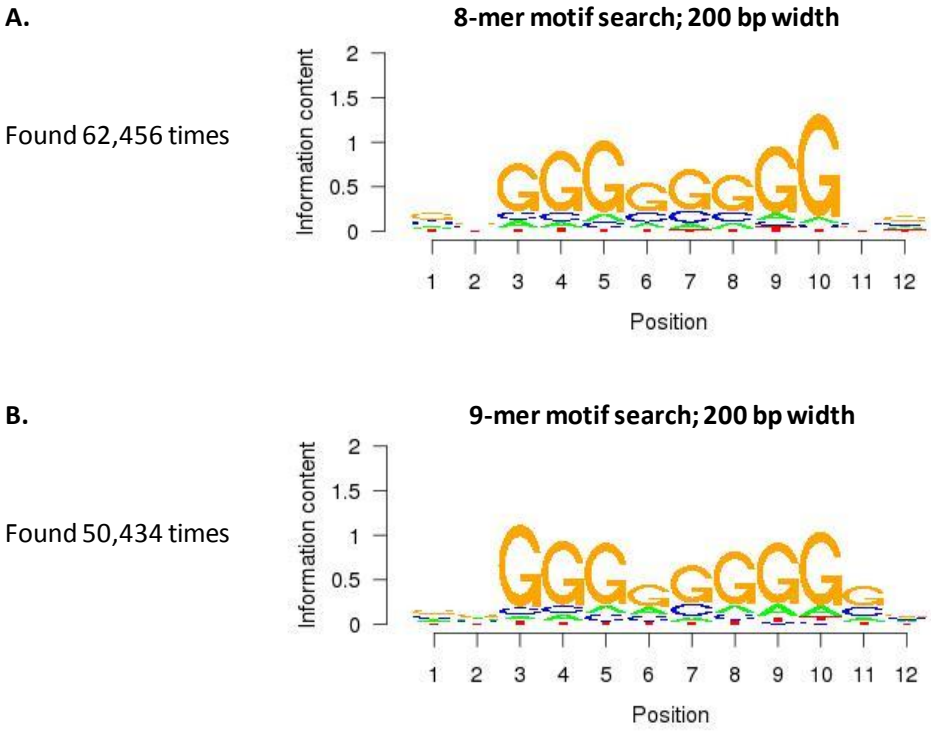


Figure 5.3 Motif discovery using POSMO identifies homopolymeric dG.dC motifs. Overrepresented motifs of 8 or 9 nucleotides in length were identified within a 200 bp window surrounding 26,341 VEZF1 ChIP-seq peak summits. The most frequently occurring motifs are shown along with the number of times each motif was found. Weblogo presentations of the discovered position weight matrices of base enrichment are shown, where the height of each letter indicates the relative occurrence of nucleotides in enriched motifs.

5.3.2.2 Motif discovery within groups of VEZF1 ChIP-seq peaks ranked by read enrichment

POSMO was used to discover motifs within groups of VEZF1 ChIP-seq peaks based on their peak strength. POSMO analysis was performed on (1) the 250 most highly enriched VEZF1 peaks, (2) the top 1000 peaks, (3) peaks ranked 8001-9000 and (4) peaks ranked 20,000-21,000. The aim was to identify the most optimal VEZF1 motif from the most enriched VEZF1 peaks followed by potentially divergent motifs as VEZF1 enrichments decrease. These analyses discovered highly enriched GGGGNGGGG motifs in all of the groups of VEZF1 sites, but the relative weight of each G base altered with decreasing VEZF1 enrichment (Figure 5.4). The primary motif discovered within the 250 most

enriched VEZF1 sites appears as a homopolymeric run of 9(dG.dC) (Figure 5.4A). This motif matches the high affinity VEZF1 sites found at the chicken β^A-globin promoter and the HS4 insulator FI site (Table 5.1) (Dickson *et al*., 2010).

The motifs within the top 1000 peaks can be homopolymeric but may consist of two shorter stretches of dG residues flanking an A or C residue shown at position 5 of the web logo (Figure 5.4B). The VEZF1 motifs within peaks of intermediate enrichment (ranked 8,000-9,000) no longer appear to be homopolymeric but consist of two highly conserved stretches of four G residues which were separated by either a C or an A nucleotide (Figure 5.4C). Finally, the most frequently occurring motif identified within VEZF1 peaks of lower enrichment (ranked 20,000-21,000) consist of two highly conserved stretches of four G residues which were separated by an A base (Figure 5.4D).

The abundance of the discovered motifs at VEZF1 ChIP-seq peak summits indicates that multiple target sites can be found at the most highly enriched elements. The most frequently occurring motif from the 250 highest enriched VEZF1 sites was found 571 times in these 200 bp elements. Likewise, the most frequently occurring motif generated using the top 1000 sites was found 1941 times in those elements. However, the best motif found from 1000 intermediate sites was found approximately once per element on average. Interestingly, the motif that occurred most often in 1000 sites of weaker VEZF1 enrichment was only found 392 times in these elements. These observations indicate that decreasing enrichment levels correlate with VEZF1 binding to increasingly degenerate DNA motifs.

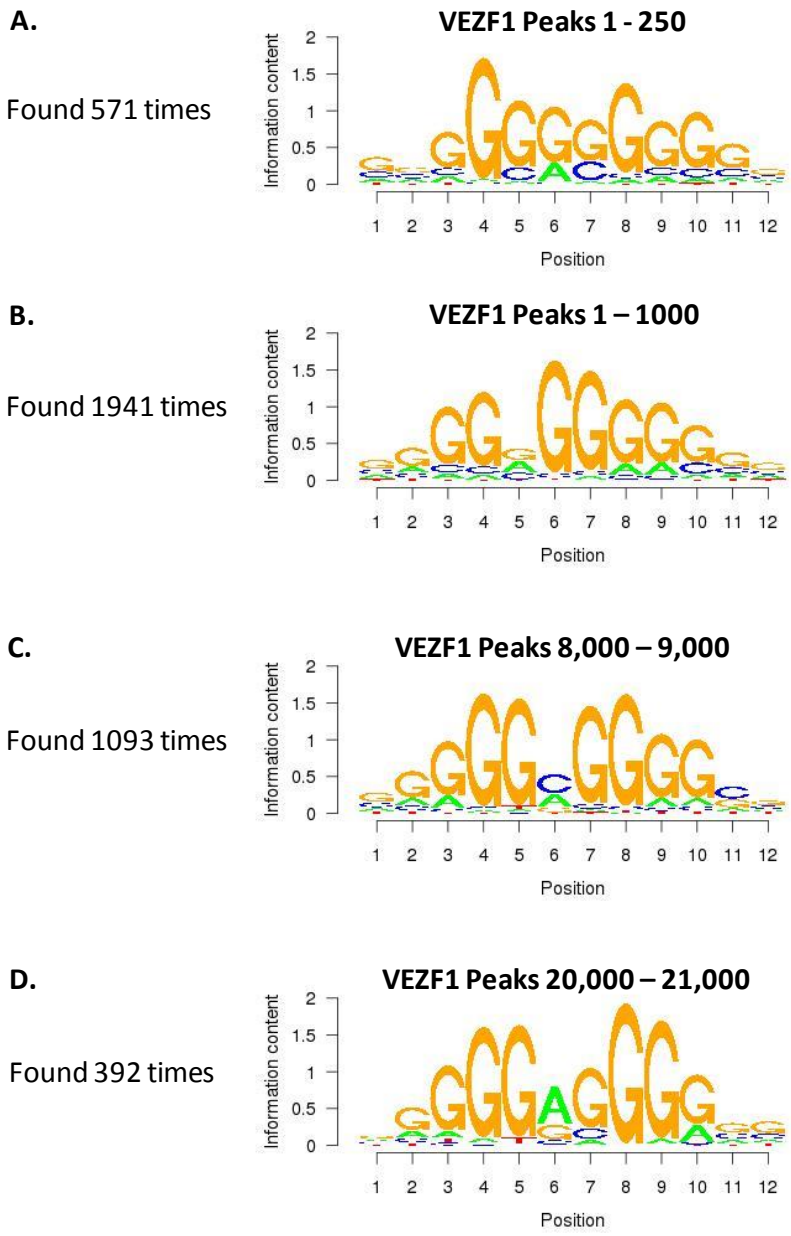


Figure 5.4 VEZF1 binding motifs differ in correlation with levels of VEZF1 enrichment. DNA sequence motifs identified by POSMO as being enriched within 200 bp sequences surrounding VEZF1 ChIP-seq peak summits. VEZF1 sites were ranked based on ChIP-seq read enrichments and groups of 250 or 1000 elements studied, as shown. The most frequently occurring motif from each group is presented along with the number of times each motif was found. Weblogo presentations of the discovered position weight matrices of base enrichment are shown, where the height of each letter indicates the relative occurrence of nucleotides in enriched motifs.

5.3.2.3 Motif discovery in VEZF1 ChIP-seq peaks that overlap promoter or enhancer HMM chromatin states

Comparison of VEZF1 ChIP-seq peak summit locations with the ChromHMM chromatin state map for K562 cells (Ernst *et al.*, 2011) identified 14,619 VEZF1 sites that overlap with

promoter states and 8,560 sites that overlap with enhancer states (section 4.3.2). POSMO motif discovery was performed on these different groups of VEZF1 sites to determine whether VEZF1 interacts with different motifs when performing different gene regulatory tasks. Similar to the motifs found at the highest affinity VEZF1 sites described above, the most frequent motif at all promoter-associated VEZF1 sites appears as a homopolymeric run of 9(dG.dC) (Figure 5.5A). In order to identify VEZF1 binding site motifs associated with enhancers, we excluded 2,205 elements that were located within 1kb of an annotated TSS from the analysis. We also excluded a further 1,280 elements that strongly resembled TSSs in SeqMINER cluster analysis due to the high levels of RNA polymerase II (clusters 6-9, Figure 4.13). The most frequent motif discovered at 3,489 TSS distal enhancer-associated elements consists of two highly conserved stretches of four G residues which were separated by an A or T base (Figure 5.5B). This enhancer motif is reminiscent of that discovered for 1000 weak VEZF1 sites (figure 5.4D), however these enhancer-associated sites have the same distribution of read enrichment as for all VEZF1 sites (not shown). It therefore appears that VEZF1 interacts with a divergent motif gGGGwGGGg at enhancers.

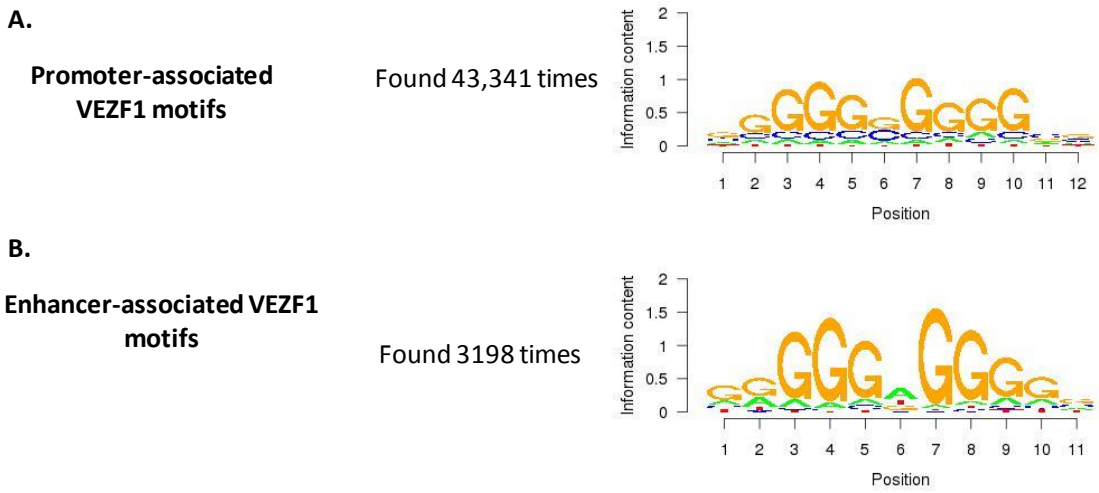


Figure 5.5 VEZF1 interacts with divergent motifs at enhancer-associated elements. DNA sequence motifs identified by POSMO as being enriched within 200 bp sequences surrounding VEZF1 ChIP-seq peak summits. 14,619 promoter-associated (A) and 3,489 TSS-distal enhancer-associated (B) VEZF1 sites were analysed. The most frequently occurring motif from each group is presented along with the number of times each motif was found. Weblogo presentations of the discovered position weight matrices of base enrichment are shown, where the height of each letter indicates the relative occurrence of nucleotides in enriched motifs.

5.4 Identification of putative VEZF1 recognition motifs within ChIP peaks

POSMO DNA sequence motif discovery analyses indicate that VEZF1 interacts with GGGGNGGGG motifs. However, the relative enrichment of guanosine bases at the edge of the motif and the base at the centre of the motif alter depending on the strength or regulatory function of VEZF1 sites. The most highly enriched sites tend to contain homopolymeric 9(dG.dC) motifs, while weak sites tend to contain divergent gGGGaGGGg motifs (section 5.3.2.2). Enhancer sites, which can have high, intermediate or low levels of VEZF1 ChIP-seq read enrichment, harbour a gGGGwGGGg motif that is similar to that enriched at weak VEZF1 sites (section 5.3.2.3). We need to determine whether the divergent motifs identified are true VEZF1 binding motifs and whether there are differences in VEZF1's DNA-binding affinity for homopolymeric 9(dG.dC) versus gGGGwGGGg sequence motifs.

As this phase of work started before VEZF1 ChIP-seq analysis was complete, putative VEZF1 binding elements were identified within 125 VEZF1 ChIP-chip peaks (section 1.9). A GGGGNGGGG motif discovered from these 125 VEZF1 binding sites was identified as a putative binding motif (Figure 5.6). The ChIP-chip peaks had an average size of 400 bp, so

multiple putative VEZF1 binding elements were identified within many of the peaks. The aim of the experiment was to increase our understanding of VEZF1 specificity, so a variety of putative VEZF1 target elements were selected that contained either homopolymeric dG strings or two shorter dG strings with a central dC, dT or dA base (Table 5.2). The target elements were identified from ChIP-chip peaks that had been validated by independent ChIP assays in 9 different cell types (Strogantsev, 2009).

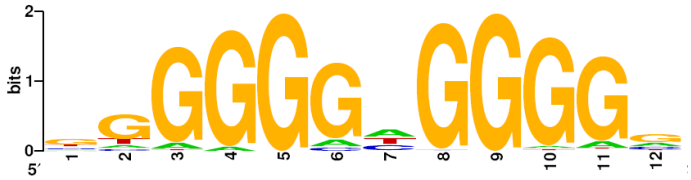


Figure 5.6 Enriched sequence motif identified in 125 VEZF1 ChIP-chip peaks.
The most enriched DNA sequence motif identified by MEME within 125 VEZF1 ChIP-chip peaks (Strogantsev, 2009). Weblogo presentations of the discovered position weight matrix of base enrichment are shown, where the height of each letter indicates the relative occurrence of nucleotides in enriched motifs.

VEZF1 site	Sequence	Cell type
Controls		
No competitor		
HS4 FI	TGGGGGCTTTGGGGGGGGGCTGTCCCCGTG	Broad
HS4 F1aaa1	TGGGGGCTTTGGGGGTTTGCTGTCCCCGTG	Broad
HS4 FIII	CTCGGGGATGCGGGGGGGAGCGCCGGACCGG	Broad
G string	nnnnnnnnnnnGGGGGGGGGnnnnnnnn	
EDN1 proA	CCCCTATTAGAGTGGGGGTAAACAGCTC	Vascular*
TAL1 pro1bA	GGGGGGGGCGGTGGGGGGGCATTTTCCG	Myeloid*
HMGN1 proA	GGCGCCCGGGGGGGGGGGGGGGGCCCCG	Broad
STAG2 proA	GCCACCCATGGGGGGGGGGGGTCTCCGG	Broad*
GC	nnnnnnnnnnGGGGCGGGGnnnnnnnn	
EHD1 16A	GGGGTAATGGGGGGCGGGGCGGGGGGC	Broad
DNMT3B proB	GGGGAACGGGGGGGCGGGGACGAGGGA	Broad
HBA p380	GTGCCAGGCCGGGGCGGGGGTGCGGGC	Erythroid*
FLNA int	GGGGTGGGATGGGGCGGGGCCATCCAG	Broad*
TAL1 pro1bB	GGCGGCAGCCGGGGCGGGGGCGTCCGT	Myeloid*
FLNA IRA	TGCCGAGGCGGGGGCGGGGCGTGGAGG	Broad*
MECP2 proB	CCCTTGCCGGGGGGCGGGGGTCAGGGG	Broad*
BRWD1 pro	CGGCGCGGGGGGGGCGGGGGGCGGGGG	Broad*
HBA p40	TTATGCTTGGGGCGCGGGGGCACGCCG	Erythroid*
HBA p177	GGGTGCACGCGGGGCGGGGGCCAGGAC	Erythroid*
GT	nnnnnnnnnnGGGGTGGGGnnnnnnnn	
GMCSF +30B	GTGGGTGGGGGGGGTGGGGAAAGGGGT	Lymphoid*
RFX pro	AAGTGGAGCGGGGGTGGGGCGGGGTAG	Broad*
GMCSF +30A	CCTGCCTCTAGGGGTGGGGTAGGTGAG	Lymphoid*
FLNA IRC	GGCTCCCGGGGGGGGTGGGTGGTGGCGC	Broad*
POLR3K pro	CGCGCCGGGGGGGGTGGGGCGGCTGCG	Broad
HBB HS3GT	CAGGGAGGGTGGGGTGGGGTCAGGGCT	Erythroid*
CKMT C	AAGGCCGGGGGCGGTGGGGGAGTGGCT	Broad*
HBB HS2	CCAGAAGGCGGGGGTGGGGCACTGACC	Erythroid
CKMT A	TGCGCCAGGAGGGGGTGGGGCTGGAGTC	Broad*
GA	nnnnnnnnnnGGGGAGGGGnnnnn	
TAL1 +3	AGGCGGGTGGGGGGAGGAGGGGGTA	Myeloid*
IL3 proP	GGCAAGGCGGGGGGAGGTGGTGGTG	Lymphoid*
HBA HS14	GGGGCCGGGGGCGGAGGGGGCCAGA	Broad*
IL3 proD	CTGGGAGCTGGGGGAGGGGCTGGCC	Lymphoid*
MECP2 proA	GGGGAGGACGGGGGAGGGGGGAGGT	Broad*
GMPPA	GGGGCCGGGGGCGGAGGGGGCCAGA	Broad*
TAL1 +51	CCCAGGGCCTGGGGAGGGGGAGCCT	Erythroid*
DNMT3B proA	GGGAGTGGGTGGGGAGGGGGCGGTG	Broad
TAL1 +20	CTGGTCCAAAGGGGAGGGGAGGAGT	Myeloid*
IL3 -37	TGATTTTGTGGGGGAGGTTGTTTGA	Lymphoid*
HBZ pro	GGTCAGGTGAGGGGAGGGGGCTGCA	Erythroid*
HBA HS40	TCCTGTGGGGGTGGAGGTGGGACAA	Erythroid
HBB pro	TCCCAGGAGCAGGGAGGGCAGGAGC	Erythroid
GnnG	nnnnnnnnnGGGGnnGGGGnnnnnnnnnn	
CKMT B	GGAGTGGCTGGGGCTGGGGGCGGTATCGG	Broad*
HBG pro	AGATAGTGTGGGGAAGGGGCCCCCAAGAG	Erythroid
FLNA IRB	CGCGTCTGGGGGGTCGTGGGGAAGCAGGG	Broad*

Table 5.2. Putative VEZF1 binding motifs identified at validated ChIP elements.
Predicted VEZF1 binding motifs are underlined and associated dG nucleotides are shown in red. The cell type specificity of VEZF1 binding for many sites (asterisks) was previously determined (Strogantsev, 2009). The cell type specificity of other elements is inferred from ENCODE chromatin state maps.

5.5 Analysis of the *in vitro* relative DNA-binding affinity of VEZF1

Electrophoretic mobility shift analyses (EMSA) were employed to validate whether the putative VEZF1 recognition sites are indeed bound by VEZF1 *in vitro*. EMSA involves the formation of protein:DNA complexes in solution followed by electrophoresis on native polyacrylamide gels. The short double-stranded DNA oligonucleotides are radiolabelled to allow the visualisation of both bound and unbound DNA. Unbound DNA probe migrates quickly, whereas the binding of protein would shift, or retard, the mobility of the labelled DNA. The mobility of the protein:DNA complexes are a function of size, shape and charge. EMSA was chosen as this assay allows the direct visualisation of distinct protein:DNA complexes, this avoiding the need for purified VEZF1. However, the assay is limited by its relatively low throughput capacity and the fact that the *in vitro* nature of the system makes it difficult to replicate *in vivo* binding events, where post-translational modifications and cooperativity with co-binding factors are difficult to reproduce.

A two stage approach to studying the relative affinity of VEZF1 towards the panel of putative recognition sites identified in section 5.4 was taken. In the first stage, the putative target sites were used as non-labelled competitors of VEZF1 binding to a labelled positive control site. Footprint I from the chicken HS4 insulator element was used as the positive control site. FI contains a homopolymeric 9(dG.dC) motif that is specifically recognised by VEZF1 (Dickson *et al*., 2010). Substitution of guanosine bases in the FIaaa1 mutant (Table 5.2) disrupts VEZF1 binding, thus forming a negative control. In the second stage, any motifs meriting further investigation were directly labelled for use as probes in EMSA assays.

5.5.1 Production of recombinant VEZF1 protein

Recombinant VEZF1 protein was produced by *in vitro* translation using rabbit reticulocyte lysate. This *in vitro* translation system can rapidly produce sufficient quantities of full length mammalian proteins for EMSA that have correct protein folding. Full length VEZF1 cannot be produced in bacteria due to insolubility arising from improper folding (Adam West, unpublished observation). A cDNA encoding chicken VEZF1 was previously cloned into the pCITE4b plasmid (Dickson *et al*., 2010). Chicken VEZF1 is 98.5% identical with human VEZF1, with identical DNA binding residues in the C2H2 zinc finger domains. The

pCITE4b-GgVEZF1 plasmid contains chicken VEZF1 cDNA cloned downstream of a T7 expression promoter and a CITE element which functions to enhance the efficiency of *in vitro* translation (Figure 5.7). Recombinant full length VEZF1 was produced following *in vitro* transcription and translation. SDS-PAGE analysis of ^{35}S-methionine labelled translations showed the production of a polypeptide of the expected size of 64kDa for full length VEZF1, with some truncation products (Figure 5.8).

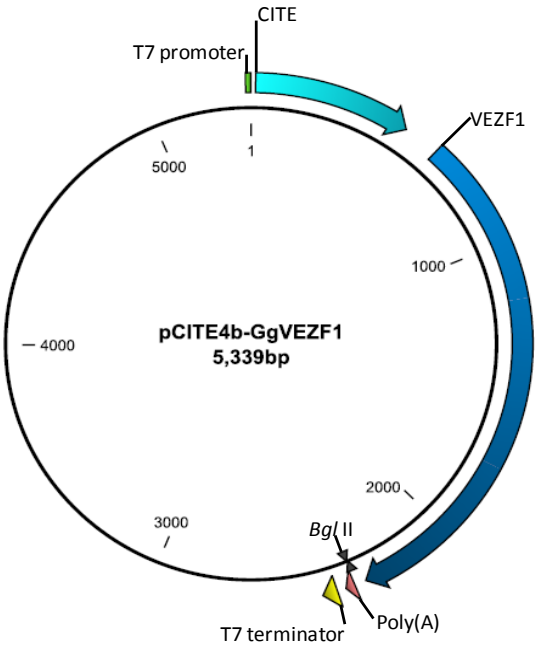


Figure 5.7 The pCITE4b-*Gg*VEZF1 plasmid.
Scaled map of pCITE4b-GgVEZF1 showing the T7 promoter from where transcription initiates, the CITE element which functions to enhance *in vitro* translation efficiency and the chicken VEZF1 coding sequence. The plasmid was linearised with *Bgl* II prior to *in vitro* transcription to prevent unproductive read-through transcription.

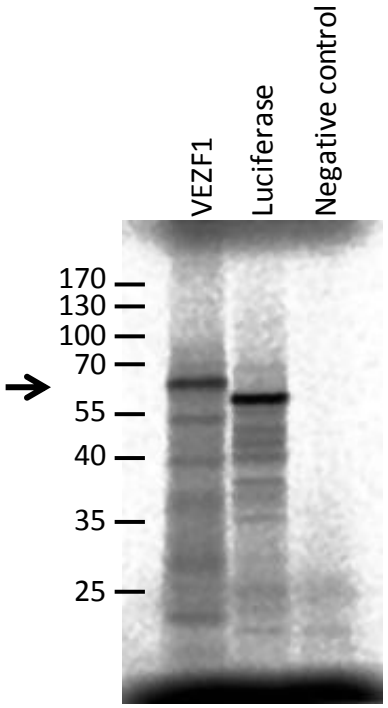


Figure 5.8 *In vitro* translation of full length VEZF1.
Phosphorimaged radiogram showing ^{35}S methionine-labelled *in vitro* translated proteins from VEZF1, luciferase or no cRNAs electrophoresed on an SDS-PAGE gel. The expected size of full length VEZF1 (64 kDa) is indicated by an arrow.

5.5.2 Establishment of EMSA assays

The EMSA assay was established as previously described (Dickson *et al*., 2010), using ^{32}P end-labelled HS4 FI double stranded 40-mer oligonucleotide as the probe DNA. The probe was incubated with either chicken red blood cell nuclear extract or *in vitro* translated VEZF1 and the resulting protein:DNA complexes electrophoresed using 5% (29:1 Ac:Bis) native PAGE in 1X TBE. The initial performance of the EMSA assay was disappointing as the unbound DNA probe failed to resolve as a discrete band near the dye front. Instead, the unbound probe often appeared as a retarded smear (Figure 5.9). The formation of complexes with *in vitro* translated recombinant VEZF1 was also much weaker than anticipated. The smearing of the DNA probe was variable in nature and some days extended up to the well. This problem had not been previously observed with this protocol.

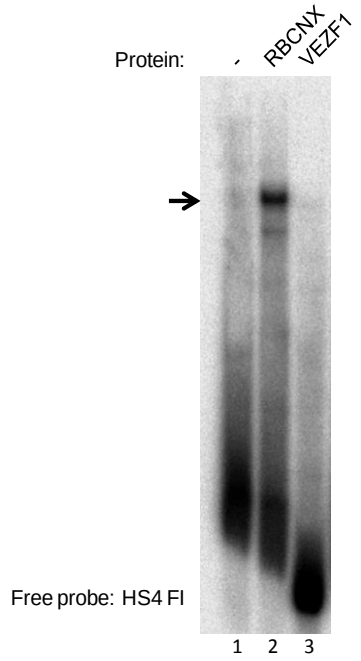


Figure 5.9 Improper resolution of DNA probes during initial EMSA assays.
Phosphorimaged radiogram of native PAGE analysis of ^{32}P-labelled HS4-FI probe incubated with no protein (lane1), red blood cell nuclear extract (lane 2) or *in vitro* translated VEZF1 (lane 3). The well is at the top and the dye front is at the bottom of this and subsequent EMSA images. Retarded protein:DNA complexes of the mobility expected for full length VEZF1 are indicated by an arrow.

The replacement of PAGE buffers, EMSA reagents and DNA probe purification reagents failed to identify a cause for poor probe resolution during PAGE. It was considered that the high GC content of the FI oligonucleotides may support mis-annealing, resulting in staggered or structured multimers of varying lengths, which may account for the retarded mobility of the free probe. The conditions for the annealing of the top and bottom strands were adjusted from a gradual cooling from 90 degrees Celsius at one degree per minute to a stepped programme of 90, 65, 37 and 22 degrees on a thermocycler. There was also concern that repeated freeze-thawing of prepared probe DNA was promoting strand separation as the smearing consistently deteriorated with time after probe preparation. DNA probe preparations were therefore stored at 4 rather than 20 degrees.

The formation of complexes between *in vitro* translated recombinant VEZF1 and the FI probe was initially weaker than anticipated. Reduced yield from the *in vitro* translation of complementary RNAs can occur if the RNA contains secondary structures. It was notable that a number of truncated VEZF1 polypeptides were produced in the original *in vitro* translation. cRNAs were therefore incubated at 65 °C for 5 minutes immediately prior to their addition into *in vitro* translation reactions.

EMSA analysis of VEZF1 interaction with the HS4 FI probe was repeated following the above protocol adjustments. The combination of stepwise probe annealing and freeze-thaw avoidance resulted in probe DNA resolution as a discrete band towards the dye front (Figure 5.10). Efficient complex formation between FI and full length VEZF1 from the optimised *in vitro* translation can now also be observed (Figure 5.10, lane 3). The addition of polyclonal antibodies raised against the C-terminus of VEZF1 resulted in a "supershift" of full length recombinant VEZF1:DNA complexes to slower migrating complexes (Figure 5.10, compare lanes 3 and 4). Faster migrating complexes that formed with *in vitro* translated VEZF1 are not supershifted, presumably because they contain C-terminal truncation products of VEZF1. A portion of the complexes between nuclear proteins extracted from chicken red blood cells and the FI probe are also supershifted by anti-VEZF1 antibodies (Figure 5.10, compare lanes 1 and 2). It has previously been shown that a mixture of VEZF1, SP1, SP3 and other related factors in nuclear extracts can form complexes with the FI site *in vitro* (Dickson *et al*., 2010). These results are entirely consistent with previous published observations, so the optimised EMSA assay is suitable for studying VEZF1 interactions with putative binding motifs identified from ChIP peaks.

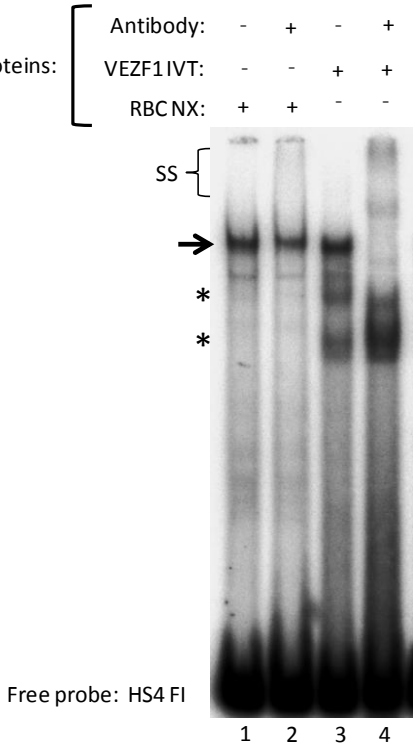


Figure 5.10 Optimised EMSA analysis of VEZF1 DNA binding.
Radiogram of native PAGE analysis of ^{32}P-labelled HS4-FI probe incubated with red blood cell nuclear extract (lanes 1 and 2) or *in vitro* translated VEZF1 (lanes 3 and 4). Retarded protein:DNA complexes of the mobility expected for full length VEZF1 are indicated by an arrow. Faster migrating DNA complexes with truncated *in vitro* translations of VEZF1 are indicated by asterisks. Supershifted VEZF1:DNA complexes with anti-VEZF1 antibodies (supplemented in lanes 2 and 4) are indicated (SS).

5.5.3 Competition EMSA analysis of VEZF1 binding to putative target sites

40 bp double stranded oligonucleotides containing 49 putative VEZF1 binding sites were prepared without radiolabelling. These sites include the 39 sites described in Table 5.2 plus a further 10 sites that subsequently failed independent ChIP validation, as discussed later. A 50 fold excess of each "cold" competitor was individually added to radiolabelled HS4-FI probe prior to incubation with recombinant VEZF1. The complexes from these 49 EMSA reactions were analysed over four PAGE gels. We only have capacity to run two large format PAGE gels at any one time, so a panel of control binding reactions were included to monitor consistency and ensure comparability between gels.

Addition of a 50-fold excess of cold HS4-FI efficiently competes for VEZF1 interaction with radiolabelled FI (Figure 5.11A, compare lanes 3 and 5). FI contains a homopolymeric 9(dG.dC) motif that is specifically recognised by VEZF1 (Dickson *et al*., 2010). Substitution of guanosine bases in the FIaaa1 mutant (Table 5.2) disrupts VEZF1 binding. Addition of a 50-fold excess of cold F1aaa1 has a minimal competition effect on VEZF1 interaction with radiolabelled FI (Figure 5.11A, compare lanes 3, 5 and 6). These controls show that competition EMSAs can be used to profile the relative DNA-binding affinity of VEZF1 for putative binding sites relative to its affinity for HS4-FI.

It is evident from the four EMSA gels that a number of the putative VEZF1 binding sequences compete for VEZF1 binding to the HS4 FI sequence (Figures 5.11 and 5.12). In order to generate a quantitative competition score for each VEZF1 binding sequence, the intensity of the VEZF1-FI specific complex bands were quantified from the phosphorimage. The VEZF1-FI specific band intensities for each competition were normalised to that of the specific band without competition to generate a score of the percentage of VEZF1-HS4 FI band remaining after competition with each DNA sequence tested. This analysis was performed for all competitor sequences used in the EMSA screen. The IGF2, NRXN, SPA1, GMCSF pro, LYL1 and EHD17 sequences were removed from the analysis at this stage as it became apparent that these ChIP-chip peaks failed to show reproducible VEZF1 enrichment in the independent ChIP-seq or ChIP-qPCR assays.

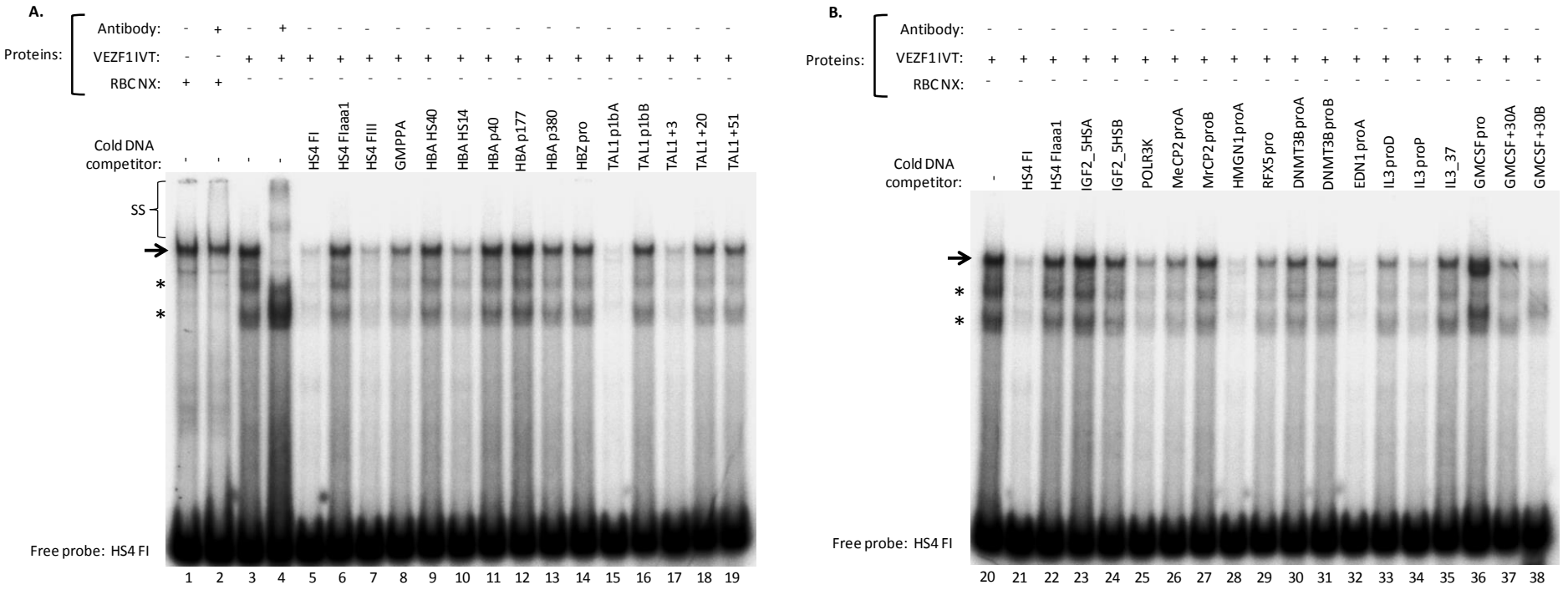


Figure 5.11 Competition EMSA analysis of VEZF1 interaction with putative binding motifs.
Radiogram of native PAGE analysis of ^{32}P-labelled HS4-FI probe incubated with red blood cell nuclear extract (lanes 1 and 2) or *in vitro* translated VEZF1 (lanes 3-38). Retarded protein:DNA complexes of the mobility expected for full length VEZF1 are indicated by arrows. Faster migrating DNA complexes with truncated *in vitro* translations of VEZF1 are indicated by asterisks. Supershifted VEZF1:DNA complexes with anti-VEZF1 antibodies (supplemented in lanes 2 and 4) are indicated (SS). 50 fold excess of the non-labelled competitor DNAs indicated above each lane were added in lanes 5-19 and 21-38.

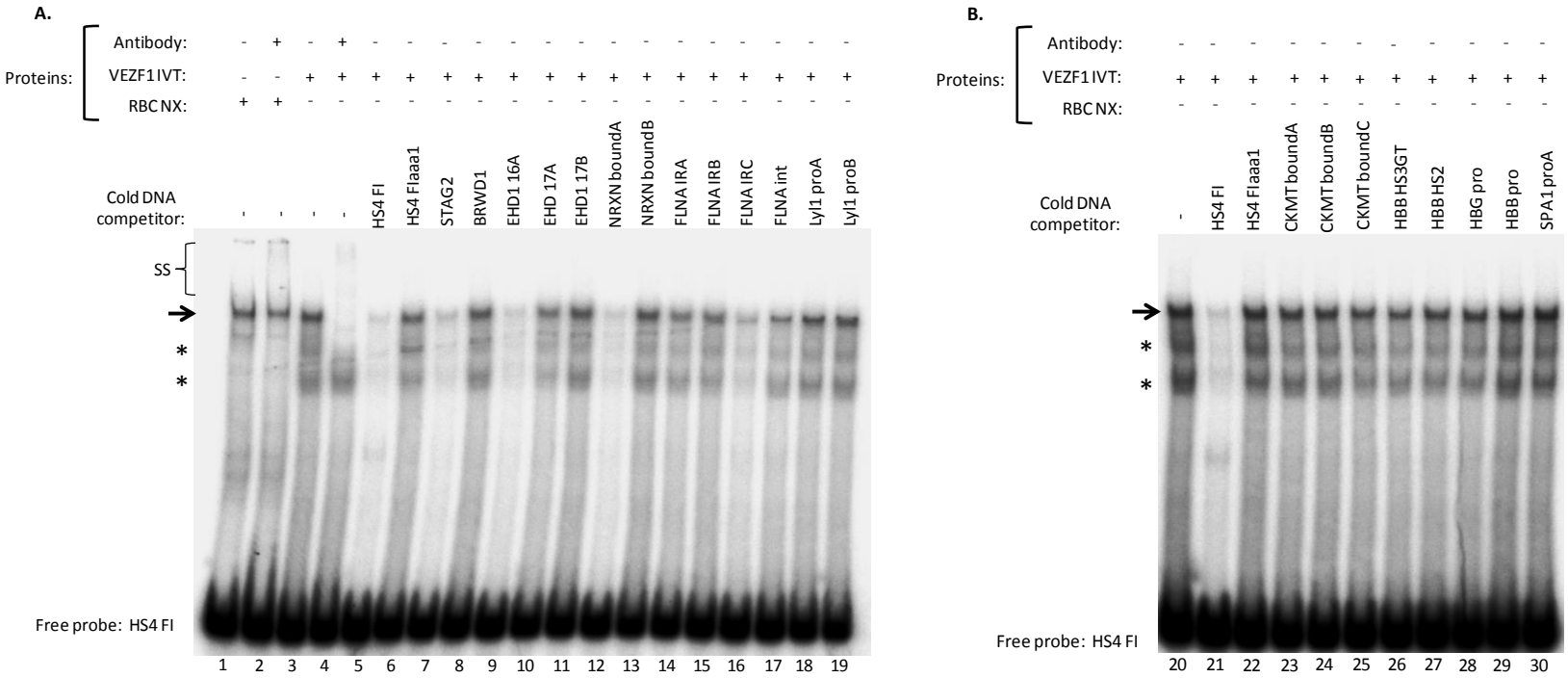


Figure 5.12 Competition EMSA analysis of VEZF1 interaction with putative binding motifs.
Radiogram of native PAGE analysis of ^{32}P-labelled HS4-FI probe incubated with red blood cell nuclear extract (lanes 1 and 2) or *in vitro* translated VEZF1 (lanes 3-39). Retarded protein:DNA complexes of the mobility expected for full length VEZF1 are indicated by arrows. Faster migrating DNA complexes with truncated *in vitro* translations of VEZF1 are indicated by asterisks. Supershifted VEZF1:DNA complexes with anti-VEZF1 antibodies (supplemented in lanes 2 and 4) are indicated (SS). 50 fold excess of the non-labelled competitor DNAs indicated above each lane were added in lanes 5-19 and 21-30.

The competition scores for the HS4 FI and HS4 FIaaa1 mutant control competitors on all four EMSA gels was found to be consistent. The 50 fold excess cold HS4 FI sequence reproducibly competed VEZF1 binding to radiolabelled FI by 93 to 94 % across all four EMSA gels (Figure 5.13). The FIaaa1 mutant competed by 22 to 41% across all four EMSA gels. This level of reproducibility indicates that competition EMSAs are a reliable way of screening for the relative affinity of putative VEZF1 binding sequences.

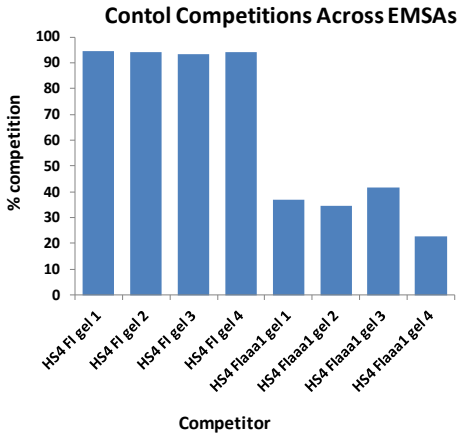


Figure 5.13 Competition scores for the HS4 FI and HS4 FI mutant sequences were consistent between the four EMSA competition gels.

The competition scores for each of the 39 putative VEZF1 binding sequences were compared to the types of sequence motif represented (Figure 5.14). The scores were categorised based on a comparison to the high affinity VEZF1 site HS4-FI ("strong") and the FIaaa1 mutant which disrupts the majority of VEZF1 binding ("very weak"). Each of the four elements containing homopolymeric G strings are efficient competitors for VEZF1 binding (Figure 5.14, G-string group). Conversely, the sequences containing SP1 consensus GC boxes are typically weak competitors for VEZF1 binding (Figure 5.14, GC group). Sequences containing GA or GT motifs compete for VEFZ1 binding with variable affinity (Figure 5.14, GA and GT groups). It is not immediately apparent from looking at the sequence motifs why some GA or GT motifs are apparently higher affinity VEZF1 binding sequences than others (Table 5.3). However, it is clear that most of the weakest competitor sequences (classed as "very weak" or "negative") diverge from the VEZF1 consensus sequence (Table 5.3).

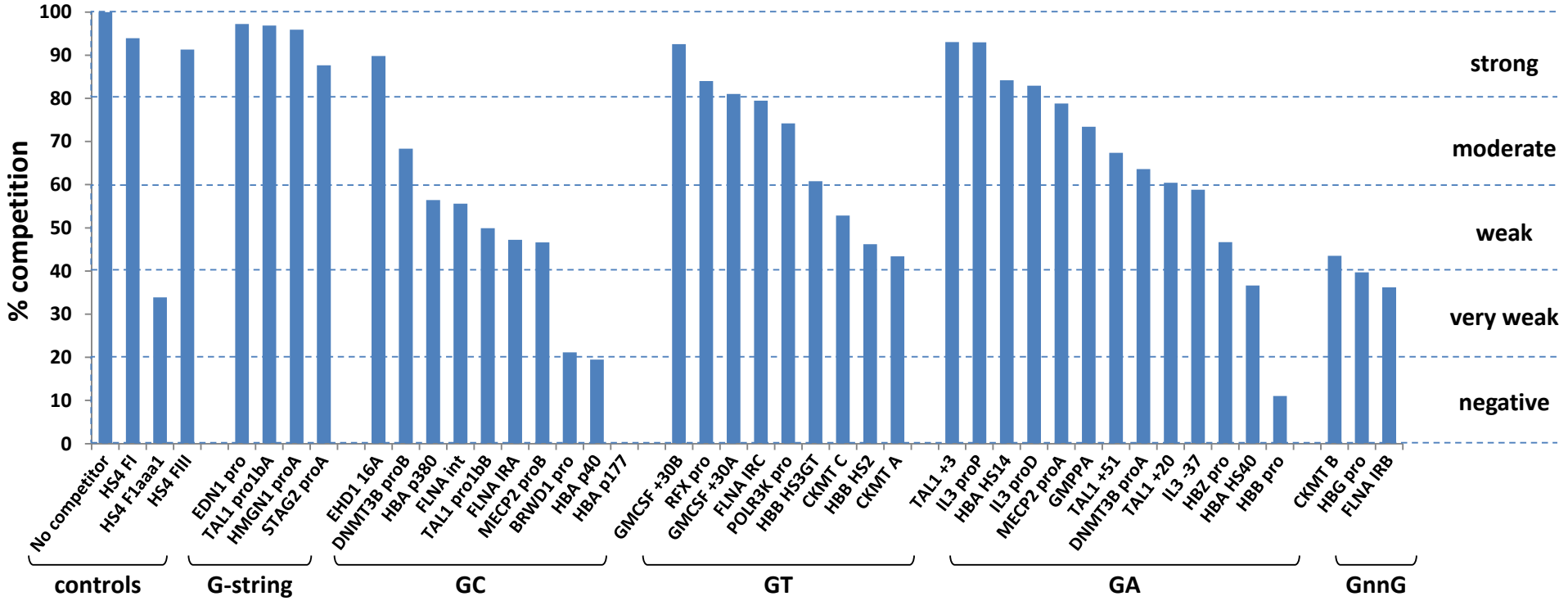


Figure 5.14 Putative VEZF1 binding sequences compete for VEZF1 binding with different efficiencies.
Quantification of VEZF1's relative affinity for putative binding sequences. The intensities of the complexes formed between full length recombinant VEZF1 and the HS4 FI sequence (from the gels shown in Figures 5.11 and 5.12) for each competition experiment were normalised to no competitor to form a percentage competition value. The competition scores have been defined as 'strong' (80 – 100 % competition), 'moderate' (60 – 80 % competition), 'weak' (40- 60 % competition), 'very weak' (20 – 40 % competition) or 'negative' (0 – 20 % competition). Competition data are grouped based on the class of sequence motif found in each competitor.

VEZF1 site	Sequence	% comp	Score	Selected
Controls				
No competitor		0	negative	
HS4 FI	TGGGGGCTTTGGGGGGGGGCTGTCCCCGTG	93.9	strong	
HS4 F1aaa1	TGGGGGCTTTGGGGGTTTGCTGTCCCCGTG	33.0	very weak	
HS4 FIII	CTCGGGGATGCGGGGGGAGCGCCGGACCGG	91.3	strong	
G string	nnnnnnnnnnnGGGGGGGGGnnnnnnnn			
EDN1 proA	CCCCTATTAGAGTGGGGGTAAACAGCTC	97.2	strong	YES
TAL1 pro1bA	GGGGGGGGCGGTGGGGGGGCATTTTCCG	96.8	strong	YES
HMGN1 proA	GGCGCCCGGGGGGGGGGGGGGGCCCCG	95.9	strong	YES
STAG2 proA	GCCACCCATGGGGGGGGGGGGTCTCCGG	87.6	strong	YES
GC	nnnnnnnnnnGGGGCGGGGnnnnnnnn			
EHD1 16A	GGGGTAATGGGGGGCGGGGCGGGGGGC	89.8	strong	YES
DNMT3B proB	GGGGAACGGGGGGGCGGGGACGAGGGA	68.3	moderate	YES
HBA p380	GTGCCAGGCCGGGGCGGGGGTGCGGGC	56.4	weak	
FLNA int	GGGGTGGGATGGGGCGGGGCCATCCAG	55.6	weak	YES
TAL1 pro1bB	GGCGGCAGCCGGGGCGGGGGCGTCCGT	49.9	weak	
FLNA IRA	TGCCGAGGCGGGGGCGGGGCGTGGAGG	47.2	weak	
MECP2 proB	CCCTTGCCGGGGGGCGGGGGTCAGGGG	46.6	weak	
BRWD1 pro	CGGCGCGGGGGGGGCGGGGGGCGGGGG	21.2	very weak	
HBA p40	TTATGCTTGGGGCGCGGGGGCACGCCG	19.5	negative	
HBA p177	GGGTGCACGCGGGGCGGGGGCCAGGAC	0	negative	
GT	nnnnnnnnnnGGGGTGGGGnnnnnnnn			
GMCSF +30B	GTGGGTGGGGGGGGTGGGGAAAGGGGT	92.5	strong	
RFX pro	AAGTGGAGCGGGGGTGGGGCGGGGTAG	84.0	strong	
GMCSF +30A	CCTGCCTCTAGGGGTGGGGTAGGTGAG	81.0	strong	
FLNA IRC	GGCTCCCGGGGGGGGTGGGTGGTGGCGC	79.5	moderate	YES
POLR3K pro	CGCGCCGGGGGGGGTGGGGCGGCTGCG	74.2	moderate	YES
HBB HS3GT	CAGGGAGGGTGGGGTGGGGTCAGGGCT	60.8	moderate	YES
CKMT C	AAGGCCGGGGGCGGTGGGGGAGTGGCT	52.8	weak	
HBB HS2	CCAGAAGGCGGGGGTGGGGCACTGACC	46.2	weak	YES
CKMT A	TGCGCCAGGAGGGGGTGGGGCTGGAGTC	43.4	weak	
GA	nnnnnnnnnnGGGGAGGGGnnnnn			
TAL1 +3	AGGCGGGTGGGGGGAGGAGGGGGTA	93.0	strong	YES
IL3 proP	GGCAAGGCGGGGGGAGGTGGTGGTG	93.0	strong	YES
HBA HS14	GGGGCCGGGGGCGGAGGGGGCCAGA	84.2	strong	
IL3 proD	CTGGGAGCTGGGGGAGGGGCTGGCC	82.9	strong	
MECP2 proA	GGGGAGGACGGGGGAGGGGGGAGGT	78.8	moderate	YES
GMPPA	GGGGCCGGGGGCGGAGGGGGCCAGA	73.3	moderate	
TAL1 +51	CCCAGGGCCTGGGGAGGGGGAGCCT	67.3	moderate	YES
DNMT3B proA	GGGAGTGGGTGGGGAGGGGGCGGTG	63.5	moderate	
TAL1 +20	CTGGTCCAAAGGGGAGGGGAGGAGT	60.4	moderate	
IL3 -37	TGATTTTGTGGGGGAGGTTGTTTGA	58.8	weak	
HBZ pro	GGTCAGGTGAGGGGAGGGGGCTGCA	46.7	weak	
HBA HS40	TCCTGTGGGGGTGGAGGTGGGACAA	36.6	very weak	
HBB pro	TCCCAGGAGCAGGGAGGGCAGGAGC	11.0	negative	
GnnG	nnnnnnnnnGGGGnnGGGGnnnnnnnnn			
CKMT B	GGAGTGGCTGGGGCTGGGGGCGGTATCGG	43.5	weak	
HBG pro	AGATAGTGTGGGGAAGGGGCCCCCAAGAG	39.6	very weak	
FLNA IRB	CGCGTCTGGGGGGTCGTGGGGAAGCAGGG	36.2	very weak	

Table 5.3. The competition scores of putative VEZF1 binding sequences.
Predicted VEZF1 binding motifs are underlined and associated stretches of dG nucleotides are shown in red. Scores from EMSA competition assays are shown and those sequences selected for use as probe in subsequent EMSAs are indicated.

5.5.4 Direct EMSA analysis of VEZF1 binding to putative target sites

It is inferred from competition EMSAs that a reduction in the formation of complexes between VEZF1 and the radiolabelled HS4 FI probe is due to interaction of VEZF1 with non-labelled competitor DNA sequences. In order to directly demonstrate that VEZF1 interacts with the putative binding sequences, 15 sequences were selected for use as radiolabelled probes in a second series of EMSAs (Table 5.3). A series of six binding reactions were set up using each probe sequence. The first four reactions test whether endogenous proteins in chicken red blood cell nuclear extract interact with the test sequence, whether any resulting complexes contain VEZF1 (supershift with anti-VEZF1 antibodies) and whether they have the same DNA sequence specificity as reported for VEZF1 (competition with F1 and F1aaa1). The latter two reactions test whether recombinant VEZF1 interacts with the test sequence. The same concentration of DNA probes and proteins were used in all the direct EMSA experiments below.

5.5.4.1 Direct EMSA analysis of VEZF1 binding to G string sequences

Four sequences that contain G strings were analysed for VEZF1 binding by direct EMSA (Table 5.3). Competition analysis demonstrated that the EDN1 proA, TAL1 pro1bA, HMGN1 proA and STAG2 proA sequences are all efficient competitors for VEZF1 binding (section 5.5.3). All four elements contain homopolymeric G strings. Direct EMSA analysis shows that VEZF1 does indeed interact with each of these elements, but that different relative affinities for each site can be observed (Figure 5.15). It is apparent that the TAL1 pro1bA element is the highest affinity site for recombinant VEZF1, followed by HMGN1 proA and EDN1 proA, then STAG2 proA (Figure 5.15, compare lane 5 in each panel). Incubation with anti-VEZF1 antibodies results in inhibition of binding of full length VEZF1 or supershifted VEZF1:DNA complexes for each of these elements (Figure 5.15, compare lanes 5 and 6 in each panel). The relative binding of recombinant VEZF1 to each of the G string motifs correlates with the competition scores for each sequence. The three highest affinity sequences EDN1, TAL1p1bA and HMGN1 all achieved competition scores of 96 – 97 %, whereas the weaker STAG2 sequence scored a lower competition value of 87 % (Table 5.3).

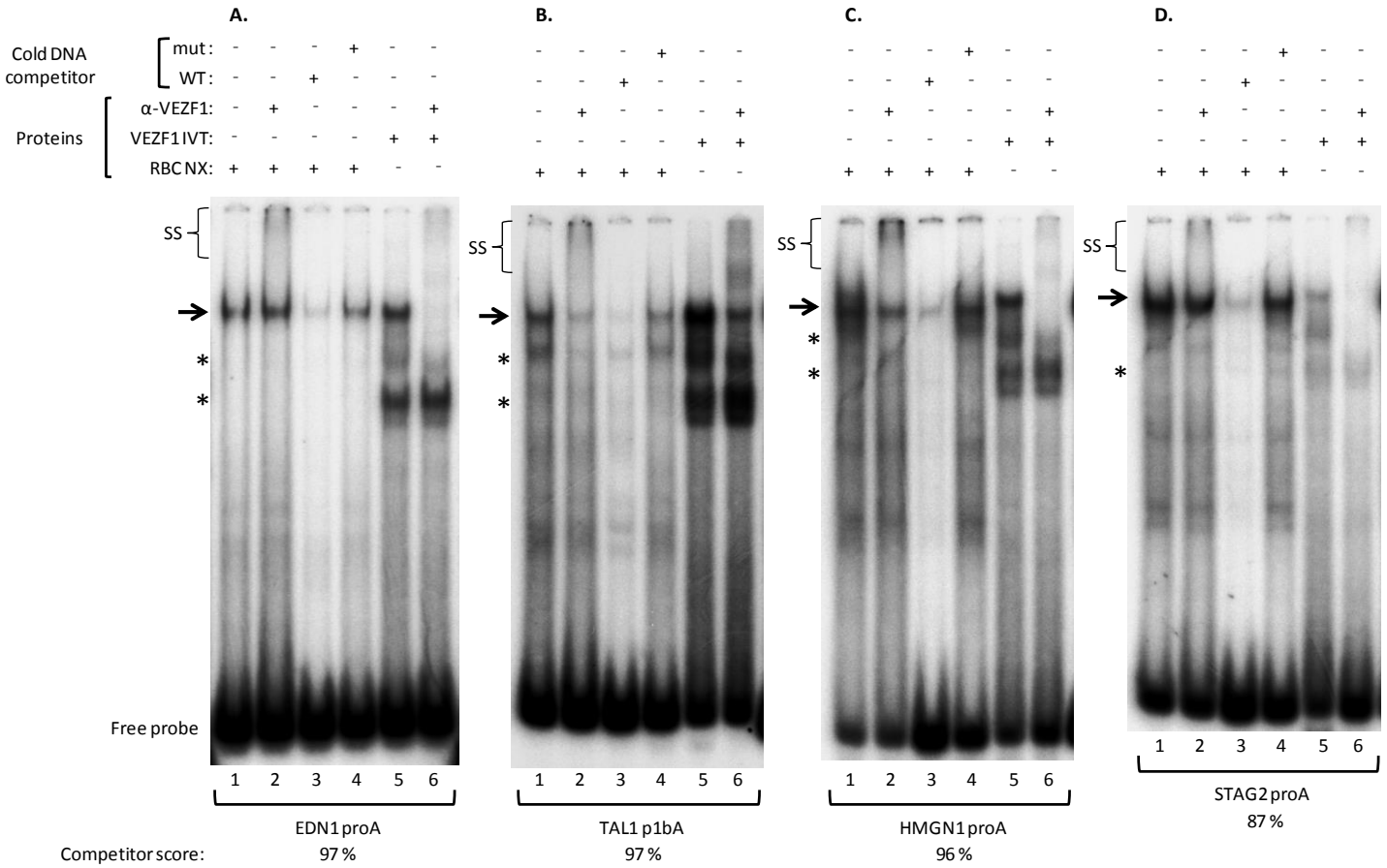


Figure 5.15 VEZF1 interacts with novel G string sequences found at VEZF1 ChIP peaks.
Radiograms of native PAGE analysis of ^{32}P-labelled EDN1 pro (panel A), TAL1 pro1bA (panel B), HMGN1 proA (panel C) and STAG2 (panel D) probes incubated with red blood cell nuclear extract (lanes 1 to 2) or *in vitro* translated VEZF1 (lanes 5 and 6). Retarded protein:DNA complexes of the mobility expected for full length VEZF1 are indicated by arrows. Faster migrating DNA complexes with truncated *in vitro* translations of VEZF1 are indicated by asterisks. Supershifted VEZF1:DNA complexes with anti-VEZF1 antibodies (supplemented in lanes 2 and 6) are indicated (SS). 50 fold excess of the non-labelled competitor DNAs indicated above each lane were added in lanes 3 and 4.

The results of direct EMSA analysis of nuclear extract protein interactions with each of the G string sequences are more complex to interpret. All of the G string sequences form complexes of a mobility expected for full length VEZF1, but different degrees of complex formation are observed. The STAG2 proA element is the highest affinity site for nuclear proteins, followed by HMGN1 proA, TAL1 p1bA and EDN1 pro (Figure 5.15, compare lane 1 in each panel). The complexes that form with each of the four sequences are specific for the G string motifs as they are efficiently competed by the G string containing FI site, but much less so by the FIaaa1 mutant (Figure 5.15, compare lanes 1, 3 and 4 in each panel). Supershift analysis indicates that VEZF1 is present in each of the complexes between the G string sequences and nuclear extract proteins, but that other proteins also bind to these sequences (Figure 5.15, compare lanes 1 and 2 in each panel). The majority of the complexes formed between nuclear proteins and the TAL1 pro1bA or HMGN1 proA sequences are supershifted by anti-VEZF1 antibodies, indicating that VEZF1 is the main constituent of these complexes. This observation is consistent with the high affinity of recombinant VEZF1 for these elements. VEZF1 appears to be a minor constituent of the complexes formed between nuclear proteins and the EDN1 proA and STAG2 proA sequences, this is consistent with the lower affinity of recombinant VEZF1 for the STAG2 proA element compared to the TAL1p1bA and HMGN1 proA sequences.

Taken together, it can be concluded that VEZF1 interacts with the G string sequences EDN1 proA, TAL1 pro1bA, HMGN1 proA and STAG2 proA. The relative DNA-binding affinity of VEZF1 for the EDN1 proA, TAL1 pro1bA and HMGN1 proA sites is comparable to that of the HS4 FI element. Binding to the STAG2 proA element is lower. Other nuclear proteins with similar specificity and EMSA mobility also interact with each of these G string sequences *in vitro*. This was previously observed for the HS4 FI element, which can also be bound by SP1 and SP3 *in vitro* (Dickson *et al*., 2010).

5.5.4.2 Direct EMSA analysis of VEZF1 binding to GC motif sequences

Three sequences that contain GC motifs akin to the SP1 consensus binding motif were analysed for VEZF1 binding by direct EMSA. Competition analysis demonstrated that sequences like FLNA int are typically weak competitors for VEZF1 binding, although two sequences with longer G strings, EHD1 16A and DNMT3B proB, were stronger competitors (Table 5.3). Direct EMSA analysis shows that recombinant VEZF1 does interact with the DNMT3B proB sequence with high affinity, but little or no binding of VEZF1 is observed at the EHD1 16A or FLNA int sequences (Figure 5.16, compare lane 5 in each panel). All of the GC sequences form strong complexes with nuclear proteins. These are of a mobility expected for full length VEZF1, but supershift analyses indicate that very little VEZF1 is present in these complexes (Figure 5.16, compare lanes 1 and 2 in each panel). It is likely that other transcription factors such as the SP1 family proteins are present in these complexes.

Taken together, the competition and direct EMSA analyses demonstrate that unlike homopolymeric G string sequences, VEZF1 does not efficiently recognise isolated GGGGCGGGG motif sequences *in vitro*. Direct VEZF1 binding to the multiple GC sequence EHD1 16A was not observed, in contradiction to the efficient competition previously observed with this sequence (Table 5.3). VEZF1 does interact with the GC sequence DNMT3B proB reasonably well, but it should be noted that there is a homopolymeric 7(dG.dC) string in this sequences with a further short G string a three bases away (Table 5.3). Mutagenesis experiments would be required to determine whether VEZF1 is interacting with the GC motif in DNMT3B proB or the separated G strings.

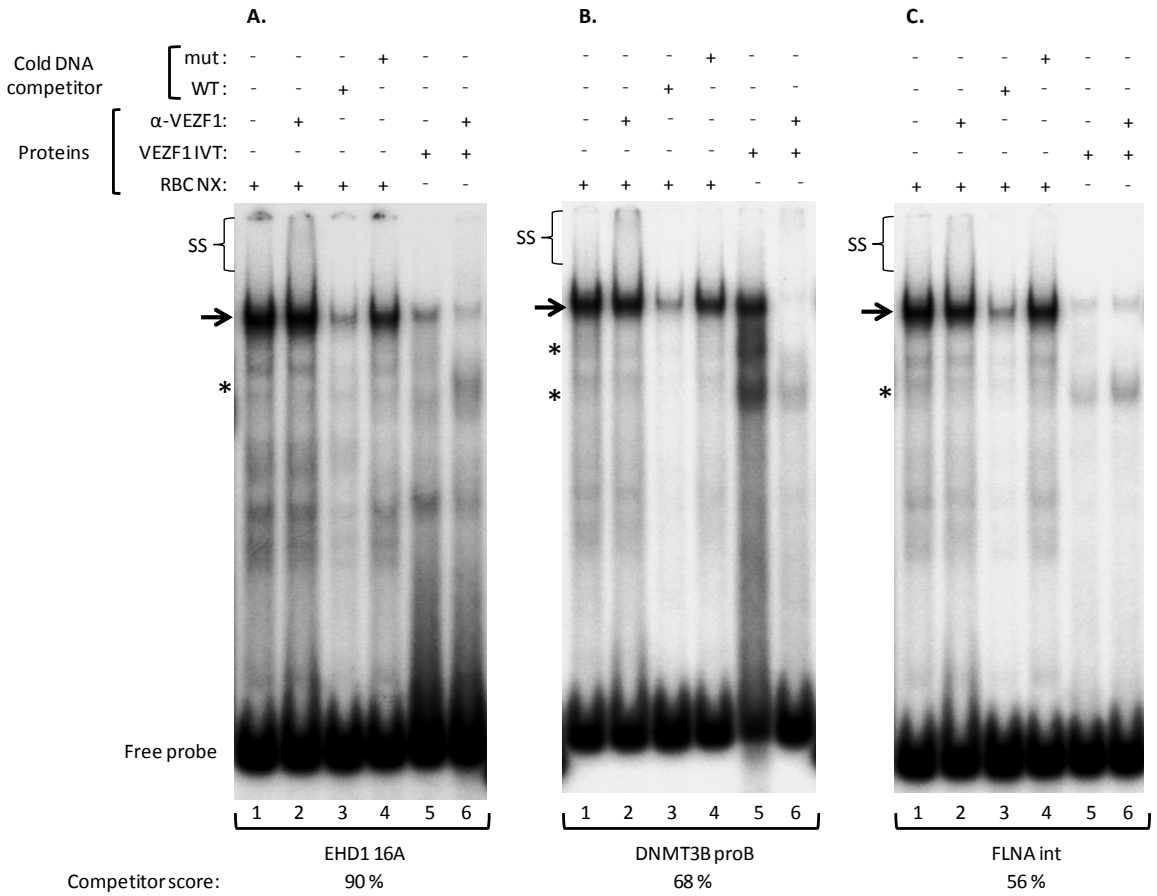


Figure 5.16 VEZF1 interaction with GC sequences found at VEZF1 ChIP peaks is variable.
Radiograms of native PAGE analysis of ^{32}P-labelled EHD 16A (panel A), DNMT3B B (panel B) and FLNA int (panel C) probes incubated with red blood cell nuclear extract (lanes 1 to 2) or *in vitro* translated VEZF1 (lanes 5 and 6). Retarded protein:DNA complexes of the mobility expected for full length VEZF1 are indicated by arrows. Faster migrating DNA complexes with truncated *in vitro* translations of VEZF1 are indicated by asterisks. Supershifted VEZF1:DNA complexes with anti-VEZF1 antibodies (supplemented in lanes 2 and 6) are indicated (SS). 50 fold excess of the non-labelled competitor DNAs indicated above each lane were added in lanes 3 and 4.

5.5.4.3 Direct EMSA analysis of VEZF1 binding to GT motif sequences

Four GT motif sequences that had moderate or weak competition activity were analysed for VEZF1 binding by direct EMSA. The binding of recombinant VEZF1 to the FLNA IRC and POLR3K pro sequences was weak. Furthermore, there was no apparent binding of recombinant VEZF1 to the HBB HS3GT or HBB HS2 sequences (Figure 5.17, compare lane 5 in each panel). All of the GT sequences form strong complexes with nuclear proteins. These are of a mobility expected for full length VEZF1, but supershift analyses show that while VEZF1 is present in these complexes, other factors probably make up the majority of the complexes (Figure 5.17, compare lanes 1 and 2 in each panel).

Taken together, the competition and direct EMSA analyses demonstrate that unlike homopolymeric G string sequences, VEZF1 does not efficiently recognise isolated GGGGTGGGG motifs *in vitro*. A previous direct EMSA study in our research group found that VEZF1 strongly interacts with the GT motif sequence GMCSF +30B under the same experimental conditions for direct EMSA shown here (Adam West, unpublished observations). The GMCSF +30B sequence does include a homopolymeric 8(dG.dC) string with a further short G string one base away (Table 5.3). Mutagenesis experiments would be required to determine whether VEZF1 is interacting with the GT motif in GMCSF +30B or the separated G strings.

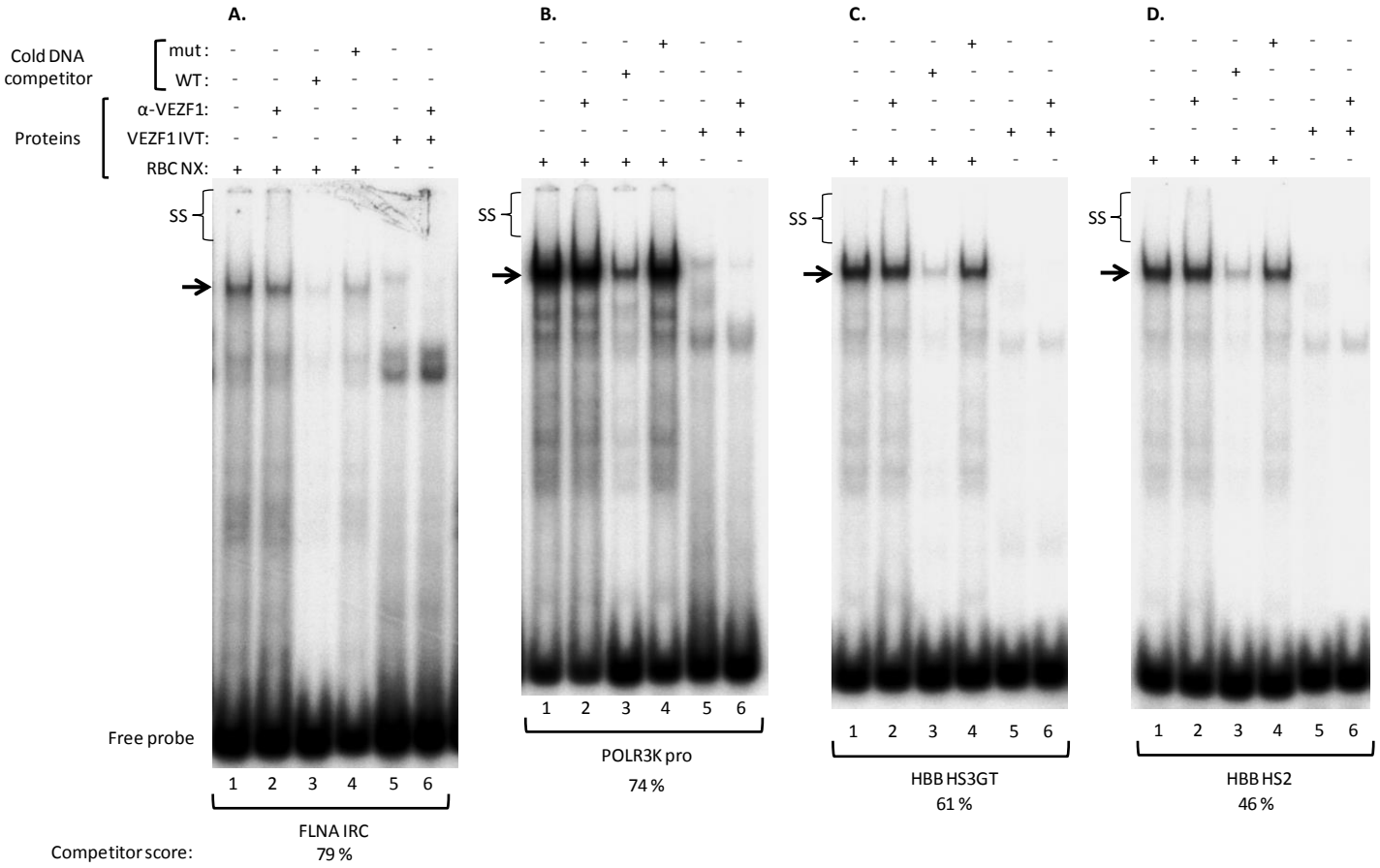


Figure 5.17 VEZF1 interactions with GT sequences found at VEZF1 ChIP peaks are variable.
Radiograms of native PAGE analysis of ^{32}P-labelled FLNA IRC (panel A), POLR3K pro (panel B), HBB HS3GT (panel C) and HBB HS2 (panel D) probes incubated with red blood cell nuclear extract (lanes 1 to 2) or *in vitro* translated VEZF1 (lanes 5 and 6). Retarded protein:DNA complexes of the mobility expected for full length VEZF1 are indicated by arrows. Faster migrating DNA complexes with truncated *in vitro* translations of VEZF1 are indicated by asterisks. Supershifted VEZF1:DNA complexes with anti-VEZF1 antibodies (supplemented in lanes 2 and 6) are indicated (SS). 50 fold excess of the non-labelled competitor DNAs indicated above each lane were added in lanes 3 and 4.

5.5.4.4 Direct EMSA analysis of VEZF1 binding to GA motif sequences

Four GA motif sequences that had strong or moderate competition activity were analysed for VEZF1 binding by direct EMSA. Strong binding of recombinant VEZF1 to the TAL1 +3 and IL3 proP was observed, however little or no VEZF1 binding was observed at the MeCP2 proA and TAL1 +51 sequences (Figure 5.18, compare lane 5 in each panel). All of the GA sequences tested form complexes with nuclear proteins. Supershift analyses show that VEZF1 is a major constituent of the complexes formed with the TAL1 +3 and IL3 proP sequences. VEZF1 accounts for approximately half of the complexes formed with MeCP2 proA, but is a minor component of complexes formed with TAL1 +51 (Figure 5.18, compare lanes 1 and 2 in each panel).

Taken together, the competition and direct EMSA analyses demonstrate that VEZF1 can efficiently recognise isolated GGGGAGGGG motifs *in vitro*. However, not all sequences that contain isolated GGGGAGGGG motifs are bound efficiently by VEZF1 *in vitro*. It is difficult to determine why VEZF1 has preference for one GGGGAGGGG sequence over another *in vitro*. It is possible that bases which flank the GGGGnGGGG motif discovered from ChIP-seq studies play a role. The high affinity TAL1 +3 and IL3 proP sequences do contain additional flanking G bases, unlike TAL1 +51 (Table 5.3).

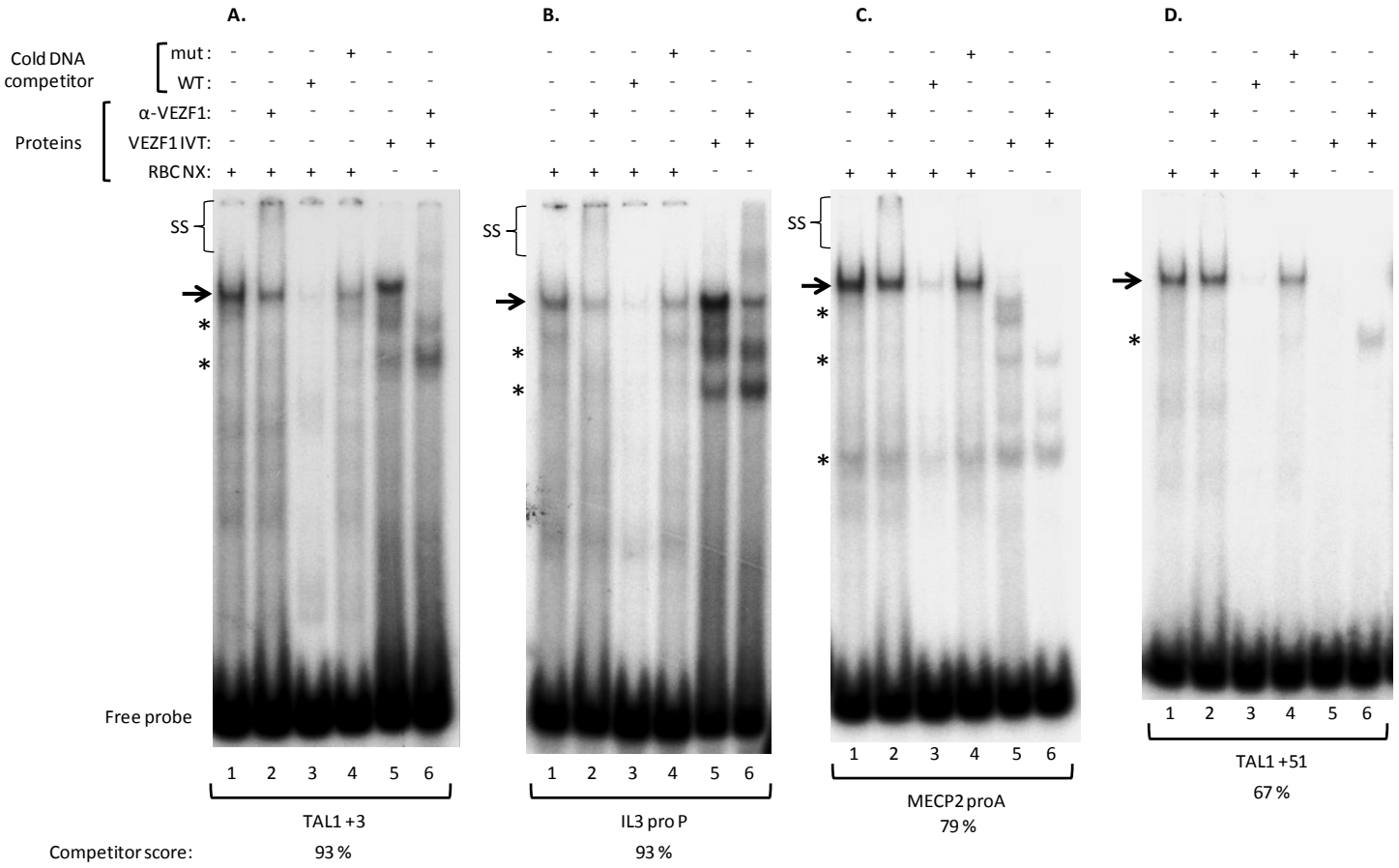


Figure 5.18 VEZF1 interaction with GA sequences found at VEZF1 ChIP peaks is variable.
Radiograms of native PAGE analysis of ^{32}P-labelled TAL1 +3 (panel A), IL3 proB (panel B), MECP2 proA (panel C) and TAL1 +51 (panel D) probes incubated with red blood cell nuclear extract (lanes 1 to 2) or *in vitro* translated VEZF1 (lanes 5 and 6). Retarded protein:DNA complexes of the mobility expected for full length VEZF1 are indicated by arrows. Faster migrating DNA complexes with truncated *in vitro* translations of VEZF1 are indicated by asterisks. Supershifted VEZF1:DNA complexes with anti-VEZF1 antibodies (supplemented in lanes 2 and 6) are indicated (SS). 50 fold excess of the non-labelled competitor DNAs indicated above each lane were added in lanes 3 and 4.

5.6 Integration of EMSA and ChIP-seq data

The EMSA experiments described in this chapter were designed using VEZF1 ChIP-chip data from a previous study. The ~400 bp size of ChIP-chip peaks may result in the selection of sequence motifs for EMSA analysis that are not actually bound by VEZF1 *in vivo*. ChIP-seq analysis now provides data that can map *in vivo* sites of VEZF1 binding at higher resolution than ChIP-chip. It is expected that for most VEZF1 binding events, the distribution of ChIP-seq reads will follow a Gaussian distribution around a binding event, where the summit of a given peak of enriched reads would closely indicate the site of VEZF1 binding. Working on this assumption, the distance between each putative VEZF1 binding motif studied relative to the nearest VEZF1 ChIP-seq peak summit was calculated (Table 5.4).

This review shows that the majority of the sequences selected for EMSA analyses map directly to VEZF1 ChIP-seq peaks, apparently confirming that the putative VEZF1 binding DNA motifs identified from analysis of ChIP-chip data were indeed bound by VEZF1. The *EDN1* proA sequence does not overlap a ChIP-seq peak in K562 cells as VEZF1 binding to this site is specific to vascular cell types (Strogantsev, 2009). Likewise, VEZF1 binding at the three *IL3* sequences studied is specific to lymphoid cell types. 11 of the 39 elements selected for EMSA analysis are located more than 100 bp from the nearest VEZF1 ChIP-seq peak summit and are unlikely to represent the true *in vivo* targets of VEZF1 at these elements. These include the G string element *STAG2* proA and the multiple GC elements *EHD1* 16A and *BRWD1* pro (Figure 5.19). Each of these elements had lower affinity for recombinant VEZF1 than expected from their sequences.

Some of the putative VEZF1 binding sequences studied in EMSA analyses derive from well characterised gene enhancer elements. Included among these are the *TAL1* erythroid enhancer element *TAL1* +51, and the β-globin LCR enhancer elements *HBB* HS2 and *HBB* HS3. Close inspection of the genomic location of these sequence elements show that they are located at the centre of the VEZF1 ChIP-seq peaks identified at these loci (Figure 5.20). The inability of VEZF1 to interact with these elements when isolated *in vitro* suggest that VEZF1 may require cooperative interactions with co-binding factors when interacting with these enhancer elements *in vivo*. This is consistent with a previous ChIP analysis

which showed that VEZF1 binding at these enhancers was restricted to erythroid cell types (Strogantsev, 2009).

VEZF1 site	Sequence	Distance to peak	Competition EMSA	Direct EMSA
G string				
EDN1 proA	CCCCTATTAGAGTGGGGGTAAACAGCTC	No peak	strong	strong
TAL1pro1bA	GGGGGGGGCGGTGGGGGGGCATTTTCCG	0 bp	strong	strong
HMGN1 proA	GGCGCCCGGGGGGGGGGGGGGGGCCCCG	57 bp	strong	strong
STAG2 proA	GCCACCCATGGGGGGGGGGGGTCTCCGG	308 bp	strong	moderate
GC				
EHD1 16A	GGGGTAATGGGGGGCGGGGCGGGGGGC	578 bp	strong	weak
DNMT3B proB	GGGGAACGGGGGGGCGGGGACGAGGGA	80 bp	moderate	strong
HBA p380	GTGCCAGGCCGGGGCGGGGGTGCGGGC	181 bp	weak	-
FLNA int	GGGGTGGGATGGGGCGGGGCCATCCAG	160 bp	weak	weak
TAL1pro1bB	GGCGGCAGCCGGGGCGGGGGCGTCCGT	0 bp	weak	-
FLNA IRA	TGCCGAGGCGGGGGCGGGGCGTGGAGG	0 bp	weak	-
MECP2 proB	CCCTTGCCGGGGGGCGGGGGTCAGGGG	270 bp	weak	-
BRWD1 pro	CGGCGCGGGGGGGGCGGGGGGCGGGGG	402 bp	very weak	-
HBA p40	TTATGCTTGGGGCGCGGGGGCACGCCG	121 bp	negative	-
HBA p177	GGGTGCACGCGGGGCGGGGGCCAGGAC	0 bp	negative	-
GT				
GMCSF +30B	GTGGGTGGGGGGGGTGGGGAAAGGGGT	48 bp	strong	-
RFX pro	AAGTGGAGCGGGGGTGGGGCGGGGTAG	70 bp	strong	-
GMCSF +30A	CCTGCCTCTAGGGGTGGGGTAGGTGAG	10 bp	strong	-
FLNA IRC	GGCTCCCGGGGGGGGTGGGTGGTGGCGC	41 bp	moderate	weak
POLR3K pro	CGCGCCGGGGGGGGGTGGGGCGGCTGCG	24 bp	moderate	weak
HBB HS3GT	CAGGGAGGGTGGGGTGGGGTCAGGGCT	30 bp	moderate	negative
CKMT C	AAGGCCGGGGGCGGTGGGGGAGTGGCT	63 bp	weak	-
HBB HS2	CCAGAAGGCGGGGGTGGGGCACTGACC	0 bp	weak	negative
CKMT A	TGCGCCAGGAGGGGTGGGGCTGGAGTC	435 bp	weak	-
GA				
TAL1 +3	AGGCGGGTGGGGGGAGGAGGGGGTA	80 bp	strong	moderate
IL3 proP	GGCAAGGCGGGGGGAGGTGGTGGTG	No peak	strong	strong
HBA HS14	GGGGCCGGGGCGGAGGGGGCCAGA	364 bp	strong	-
IL3 proD	CTGGGAGCTGGGGGAGGGGCTGGCC	No peak	strong	-
MECP2 proA	GGGGAGGACGGGGGAGGGGGGAGGT	15 bp	moderate	weak
GMPPA	GGGGCCGGGGGCGGAGGGGGCCAGA	298 bp	moderate	-
TAL1 +51	CCCAGGGCCTGGGGAGGGGGAGCCT	0 bp	moderate	negative
DNMT3B proA	GGGAGTGGGTGGGGAGGGGGCGGTG	0 bp	moderate	-
TAL1 +20	CTGGTCCAAAGGGGAGGGGAGGAGT	30 bp	moderate	-
IL3 -37	TGATTTTGTGGGGGAGGTTGTTTGA	No peak	weak	-
HBZ pro	GGTCAGGTGAGGGGAGGGGGCTGCA	34 bp	weak	-
HBA HS40	TCCTGTGGGGGTGGAGGTGGGACAA	10 bp	very weak	-
HBB pro	TCCCAGGAGCAGGGAGGGCAGGAGC	No peak	negative	-
GnnG				
CKMT B	GGAGTGGCTGGGGCTGGGGGCGGTATCGG	85 bp	weak	-
HBG pro	AGATAGTGTGGGGAAGGGGCCCCCAAGAG	No peak	very weak	-
FLNA IRB	CGCGTCTGGGGGGTCGTGGGGAAGCAGGG	255 bp	very weak	-

Table 5.4 The distance of putative VEZF1 binding motif to the nearest ChIP-seq peak. The VEZF1 binding motifs studied by EMSA analyses are shown. The distance of each sequence motif from the nearest VEZF1 ChIP-seq peak summit in K562 cells is shown alongside the scoring of the relative binding to recombinant VEZF1 in EMSA assays. Distances of >100 bp to the nearest VEZF1 ChIP-seq peak summit and negative direct EMSA scores are shown in blue.

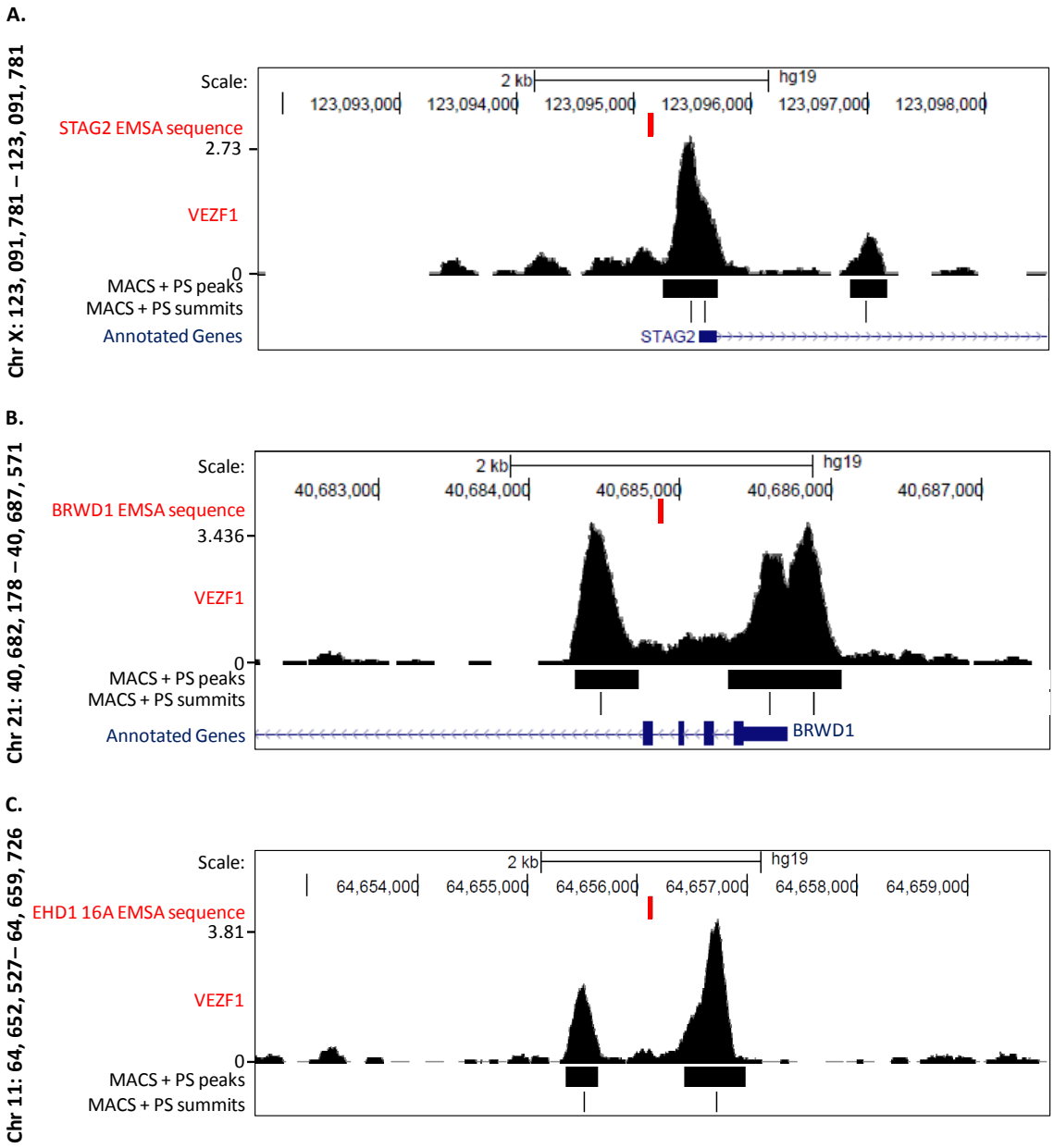


Figure 5.19 Some EMSA sequences do not locate to VEZF1 ChIP-seq peak summits. UCSC Genome browser views of VEZF1 ChIP-seq profiles from K562 cells at the (A) STAG2 (B) EHD1 and (C) BRWD1 loci. ChIP-seq peaks and summits identified by MACS and PeakSplitter analysis (section 3.6) are shown in black below the ChIP-seq track. UCSC annotated genes are shown in blue and the direction of transcription is indicated by arrows (bottom track). The location of 40 bp sequences selected for EMSA analyses are indicated above the ChIP-seq track in red. Genomic locations (hg19) are indicated.

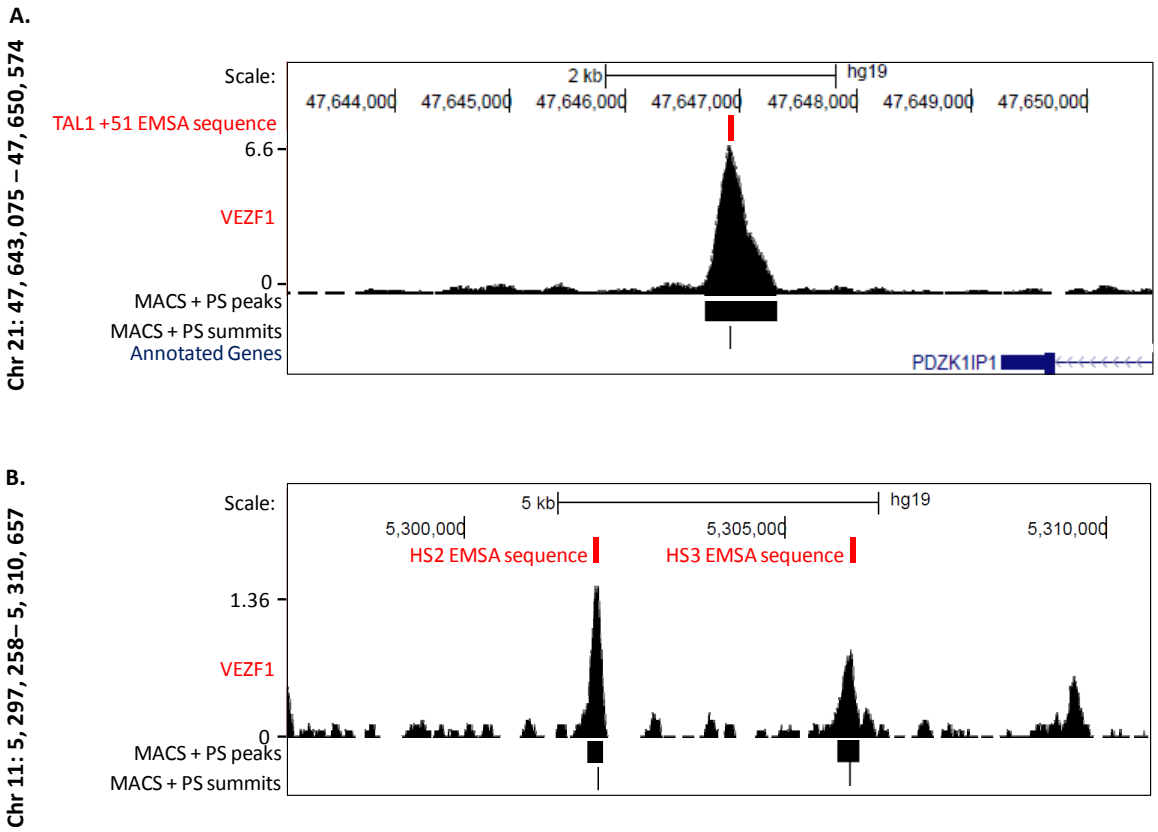


Figure 5.20 Enhancer-associated VEZF1 binding motifs locate to VEZF1 ChIP-seq peaks. UCSC Genome browser views of VEZF1 ChIP-seq profiles from K562 cells at the TAL1 +51 enhancer (A) and HBB HS2 and HS3 enhancers (B). ChIP-seq peaks and summits identified by MACS and PeakSplitter analysis (section 3.6) are shown in black below the ChIP-seq track. UCSC annotated genes are shown in blue and the direction of transcription is indicated by arrows (bottom track). The location of 40 bp sequences selected for EMSA analyses are indicated above the ChIP-seq track in red. Genomic locations (hg19) are indicated.

5.7 Discussion

The aims of this chapter were to use VEZF1 ChIP-seq data to define a VEZF1 consensus binding motif or motifs, to analyse the relative DNA binding affinity of VEZF1 for putative VEZF1-binding motifs *in vitro*, and to develop a model of DNA binding affinity and its relationship to gene regulatory mechanisms. A small number of VEZF1 binding sites have been identified and characterised in the published literature, these DNA motifs are consistently G-rich. A minimum stretch of seven contiguous dG.dC bases is required for VEZF1 binding to the chicken β^A promoter (Clark *et al*., 1990). Consistent with this, putative VEZF1 consensus binding sites generated by prediction tools consist of homopolymeric runs of eight dG.dC bases (Figure 5.1). Most validated VEZF1 binding sites however lack this homopolymeric run of dG bases and consist of G-rich motifs interspersed with A, C or T bases (table 5.1). It is apparent therefore that while VEZF1 clearly interacts with G-rich DNA elements, no clear consensus binding motif can be identified from this small collection of DNA binding sites.

It was discovered that the performance of different motif discovery tools can be greatly variable. Initial motif discovery performed by MEME using VEZF1 ChIP-seq data discovered G-rich motifs as being most highly represented as expected, however these motifs were very long in length and the majority of G residues within them were poorly conserved. This unexpected finding raised great concern regarding the ability of MEME to accurately align G-rich sequences during motif discovery. A second motif finding tool, POSMO, discovered far more believable enriched DNA motifs. POSMO consistently discovered DNA motifs of ~9 highly conserved bases as being enriched within VEZF1 ChIP-seq peaks. Outside these 9 bases no sequence conservation is apparent indicating that enriched putative protein-binding motifs were accurately aligned and representative PWMs generated.

VEZF1 motif discovery by POSMO using the most highly enriched ChIP-seq sites found homopolymeric G string sequences to be the most frequently occurring motifs. POSMO analysis also showed that decreasing VEZF1 enrichment correlates with binding to increasingly degenerate motifs which consist of two runs of four G bases separated by an A or C residue. Homopolymeric 9(dG) motifs were found to be highly enriched at VEZF1-associated promoters while a gGGGA/TGGGg motif was discovered to occur with greatest

frequency at VEZF1-associated enhancers. These findings indicate that VEZF1 generally binds high affinity homopolymeric DNA motifs at promoters and weaker more divergent motifs at enhancers.

Consistent with the results of POSMO motif discovery, VEZF1 was found to consistently form strong interactions with homopolymeric G string motifs *in vitro* by a series of competition and direct EMSAs. VEZF1 interacted weakly with GC motifs and formed interactions of variable strength with GA and GT motifs. It is also noteworthy that most of the weakest VEZF1 binding sites diverge from the GGGGNGGGG consensus motif.

Some putative VEZF1-binding motifs, which mapped directly to VEZF1 ChIP-seq peaks, were unable to interact with VEZF1 when isolated *in vitro*, these included the erythroid-specific enhancer elements *TAL1*+51, *HBB* HS2 and *HBB* HS3. The TAL1+51-associated VEZF1 ChIP-seq peak ranked within the top 13 % of all 26,429 VEZF1 ChIP-seq peaks, while *HBB* HS2- and HS3-associated peaks ranked within the top 32 % and 81 % of peaks respectively. The sequence motifs of these putative VEZF1-binding elements correlate with the enhancer-associated divergent VEZF1 motif identified by POSMO motif discovery (figure 5.5). Failure of VEZF1 to interact with these motifs *in vitro* may indicate that additional co-binding factors are required for VEZF1 binding to these sites *in vivo*. Such co-factors may include components of the TEC complex, which were found to interact with VEZF1-enriched erythroid-specific enhancer elements in chapter 4 (section 4.4.2).

The *HBB* HS2 and HS3 elements are components of the human β-globin locus control region (LCR), which regulates the erythroid and developmental-specific expression of the human β-globin genes. The functional core of HS2 has been mapped to a 375 bp fragment which contains a number of protein binding sites (Reddy and Shen, 1991). The putative VEZF1 binding GT motif that we have identified within the HS2 element is footprinted *in vivo* in K562 cells (Ikuta and Kan, 1991), however deletion of this element has been reported to have little effect on HS2 enhancer function (Sorrentino *et al*., 1990). The functional core of HS3 maps to a 225 bp region which contains six footprinted elements in erythroid cells (Philipsen *et al*., 1990, Strauss and Orkin, 1992). The putative VEZF1-binding GT motif that we have identified within the HS3 element is located within footprint II. Mutational analyses of HS3 in the murine erythroleukemia cell line Hu11, which contains a portion of the human chromosome 11 that includes the human β-globin

locus, have shown the putative VEZF1-binding motif to be essential for HS3 activity (Philipsen *et al*., 1993). Two consensus GATA motifs, which are located within footprint I and footprint III and flank the putative VEZF1 site, are also essential for HS3 activity (Philipsen *et al*., 1993). These GATA sites are located at a distance of 19 bp 5' and 17 bp 3' of the VEZF1 site. These findings support the hypothesis that co-binding of erythroid-specific factors may mediate VEZF1 binding to erythroid-specific sites.

Chapter 6

The relationship between VEZF1, promoter DNA methylation and transcription of the chicken β-globin genes

6.1 Introduction

VEZF1 has previously been reported to interact with gene regulatory elements at the chicken β-globin locus. VEZF1 is the factor originally identified as Beta Globin Protein 1 (BGP1), which binds to a long homopolymeric G-string upstream of the β^A gene promoter *in vitro* (Lewis *et al*., 1988, Dickson *et al*., 2010). VEZF1 has also been shown to bind to three G-string sequences within the HS4 insulator element *in vitro* and *in vivo* (Dickson *et al*., 2010). The binding of VEZF1 at gene regulatory elements across the chicken β-globin locus in primary chick erythrocytes was previously profiled by ChIP-qPCR (Figure 6.1, Ruslan Strogantsev, unpublished data). This analysis was performed in circulating nucleated erythrocytes from 5 day old embryos, which are reported to express the embryonic β-globin genes *ρ* and *ε*, and 10 day embryos, which are reported to express the β^A gene. It was found that VEZF1 interacts with the HS4 insulator in both 5 and 10 day erythrocytes. Interestingly, VEZF1 was found to interact with the β-globin gene promoters in a stage-specific manner. VEZF1 binding is observed at the *ρ* and *ε* promoters only in 5 day erythrocytes, the stage at which they are expressed. Conversely, VEZF1 interacts with the β^A promoter only when it is expressed in 10 day erythrocytes. VEZF1 is also found to interact with the HS2 and $\beta^{A/\varepsilon}$ enhancer elements at both stages. These results show that VEZF1 binding to gene regulatory elements can be regulated during development and that binding to gene promoters correlates with their expression.

VEZF1 binding to each of the three sites in the HS4 element is essential for the barrier activity of this insulator. Deletion of any one VEZF1 binding site from HS4 results in an accumulation of DNA methylation and silencing of a transgene flanked by mutant insulators (Dickson *et al*., 2010). Further support of a role for VEZF1 as a regulator of DNA methylation is provided by evidence that VEZF1 binding sites in the hamster *APRT* gene promoter are sufficient to protect this CpG island from *de novo* DNA methylation or to direct demethylation of a pre-methylated *APRT* CGI upon integration into the genome of mouse ES cells (Dickson *et al*., 2010). Given the apparent relationship between VEZF1

binding and protection against DNA methylation, it is of interest to investigate whether the DNA methylation status of the *ρ* and β^A promoters changes during erythroid development and how this correlates with VEZF1 binding and gene expression.

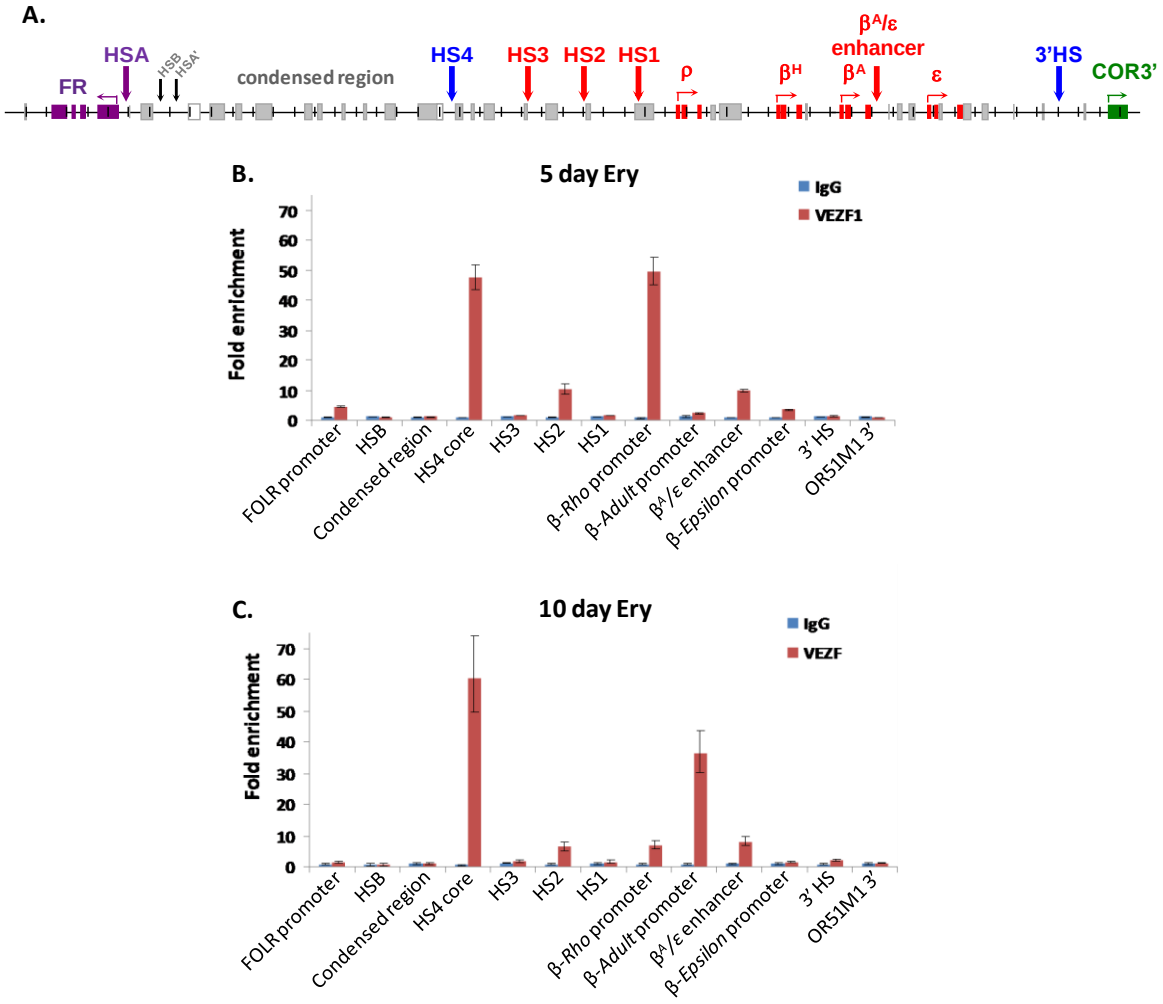


Figure 6.1 VEZF1 interactions at the β-globin locus during chick erythroid development. (A) Schematic of the chicken globin gene locus. ChIP-qPCR analysis of VEZF1 interactions (red) with DNA elements across the chicken β-globin gene cluster and surrounding loci in (B) 5 day chicken embryonic erythrocytes or (C) 10 day chicken embryonic erythrocytes. IgG ChIP-qPCR is included as a negative control for non-specific immunoprecipitation of DNA sequences (blue). Error bars reflect standard deviation between triplicate technical replicates (Dr Ruslan Strogantsev, unpublished data).

The studies in this chapter aim to address the following:

1) Determine the expression of *VEZF1* and the β-globin genes in circulating erythrocytes during chicken embryonic development

2) Profile the genomic binding of VEZF1 in circulating chicken embryonic erythrocytes during chicken embryonic development

3) Identify the specific DNA motifs bound by VEZF1 at β-globin gene regulatory elements

4) Examine the affinity of VEZF1 for putative binding sequences found at β-globin gene promoters.

5) Study the relationship between VEZF1 binding, promoter DNA methylation and transcription of the ρ and β^A genes.

6.2 The expression of VEZF1 and the β-globin genes in circulating erythrocytes during chicken embryonic development

In order to understand the regulation of VEZF1 binding and DNA methylation of the β-globin genes, the expression of VEZF1 and each of the β-globin genes needed to profiled in our erythrocyte preparations. Circulating erythrocytes were isolated from chick embryos that had been incubated for between 5 and 10 days post fertilisation. The mRNA levels of ρ, ε, β^H and β^A globin were determined at different stages of erythrocyte development (Figure 6.2). The "embryonic" β-globin genes *ρ* and *ε* are reported to be highly expressed at day 5 and progressively repressed throughout the process of embryonic development (Bruns and Ingram, 1973, Sheng, 2010). The expression of these genes in our erythrocyte preparations correlate with this pattern. ρ globin expression at day 5 is 204-fold greater than β-actin (ACTB) and increases to a peak of 897-fold above ACTB at day 6. ρ globin expression progressively silences during later stages of erythroid development, falling to 34-fold expression above ACTB at day 10. ε globin expression followed a very similar pattern. ε globin mRNA levels were 532-fold above ACTB at day 5 with expression peaking at 1136-fold at day 6. ε globin also progressively silences during later stages of erythroid development, falling to 43-fold expression above ACTB at day 10.

Conversely, the expression of the "adult" β-globin genes β^H and β^A are reported to be unexpressed in chick embryos at day 5 post-fertilisation. Expression of the β^A gene increases progressively throughout the course of embryonic development until it is strongly expressed at day 10. Expression of the β^H gene also occurs throughout chick embryonic development but reaches only very low expression levels in 10 day erythrocytes (Bruns and Ingram, 1973, Sheng, 2010). The expression of these genes in our erythrocyte preparations correlate with this pattern (Figure 6.2). The expression of β^A globin in erythrocytes increased from only 0.9-fold relative to ACTB in 5 day erythrocytes to a peak of 1043-fold above ACTB in 10 day erythrocytes. β^H globin expression levels followed a similar pattern, but at a much lower level than β^A, rising from 0.06-fold relative to ACTB at day 5 to 17.5-fold at day 10.

These gene expression analyses confirm that the developmentally regulated switching of β-globin gene expression in our embryonic erythrocyte preparations matches those previously reported (Bruns and Ingram, 1973, Sheng, 2010).

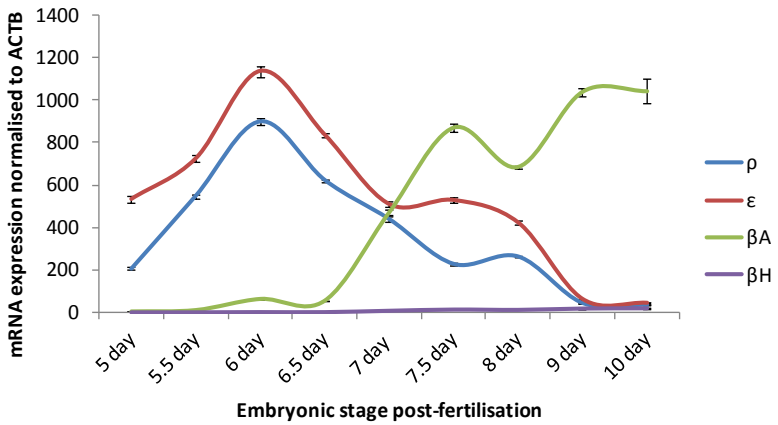


Figure 6.2 β-globin gene expression in embryonic chicken erythrocytes.
Quantitative RT-PCR analysis of β-globin mRNA levels in erythrocytes isolated from chicken embryos between day 5 and 10 post fertilisation. β-globin mRNA levels were normalised to those of ACTB in each preparation. Error bars represent standard deviation between three technical PCR replicates. The expression profile is representative of erythrocyte preparations from three independent clutches.

Embryonic stage-specific binding of VEZF1 was previously observed at the β-globin promoters but not at the HS4 insulator, so it is important to understand whether the expression of VEZF1 changes during the period of embryogenesis studied. RT-PCR analysis shows that VEZF1 mRNA levels are relatively unchanged in circulating erythrocytes between 5 and 10 days post fertilisation (Figure 6.3).

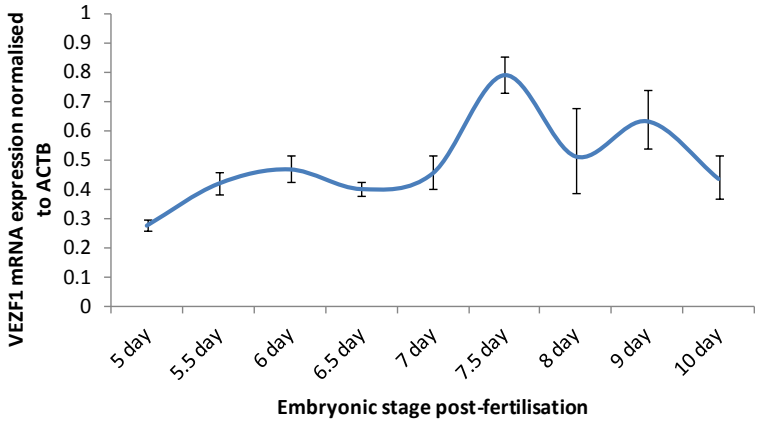


Figure 6.3 VEZF1 gene expression in embryonic chicken erythrocytes.
Quantitative RT-PCR analysis of VEZF1 mRNA levels in erythrocytes isolated from chicken embryos between day 5 and 10 post fertilisation. VEZF1 levels were normalised to those of ACTB in each preparation. Error bars represent standard deviation between three technical PCR replicates.

6.3 The genomic binding of VEZF1 in circulating chicken embryonic erythrocytes during chicken embryonic development

It is important to determine the exact sequences bound by VEZF1 *in vivo* during development in order to gain a clear understanding of how VEZF1 binding may be regulated. VEZF1 may interact with divergent low affinity sequences at stage-specific sites and therefore rely on cooperative interactions with co-binding transcription factors. Alternatively, VEZF1 binding sequences may be selectively masked by stage-specific nucleosome positioning, for example. ChIP-seq analysis should define VEZF1 binding events more accurately than the ChIP-QPCR method we have previously employed. This analysis would also allow the identification of stage-specific binding events at other gene loci, such as the α-globin gene cluster.

6.3.1 Preparation of crosslinked chromatin from embryonic erythrocytes

$1x10^8$ circulating embryonic erythrocytes were collected from 5 and 10 day embryos and formaldehyde crosslinked chromatin was prepared using previously optimised conditions (2.5). Sonication of chromatin purified from 5 day erythrocytes yielded fragments with an average size of ~400 – 700 bp, while sonication of chromatin from 10 day erythrocytes yielded fragments with an average size of ~300 – 600 bp (Figure 6.4). Fragmented chromatin from $5x10^6$ 5 and 10 day erythrocytes was used per ChIP. VEZF1 ChIPs from both preparations were performed in duplicate and pooled.

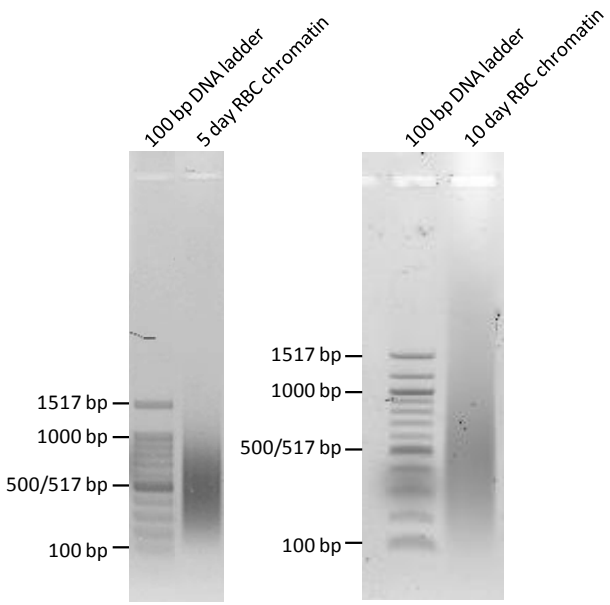


Figure 6.4 Sonication of chick erythrocyte chromatin.
Agarose gel electrophoresis of 5 and 10 day embryonic chick erythrocyte chromatin after sonication for use in ChIP.

6.3.2 Validation of VEZF1 ChIP performance and library preparations

The enrichment of previously identified VEZF1 binding targets following VEZF1 ChIP was validated using QPCR analysis. The HS4 insulator element was found to be bound by VEZF1 in both 5 and 10 day erythrocytes (Figure 6.5). VEZF1 was also bound at the ρ-globin promoter in 5 day erythrocytes, but not at the β^A-globin promoter. Conversely, VEZF1 was bound at the β^A-globin promoter in 10 day erythrocytes, but not at the ρ-globin promoter. These results are consistent with previous findings in the West laboratory (Figure 6.1). No VEZF1 binding was observed within the condensed chromatin region upstream of the β-globin gene locus. These analyses show that genomic elements bound by VEZF1 in erythrocytes were isolated efficiently by VEZF1 ChIP and that stage-specific binding patterns were still being identified. One third of each ChIP DNA pool was used in ChIP-seq library preparation.

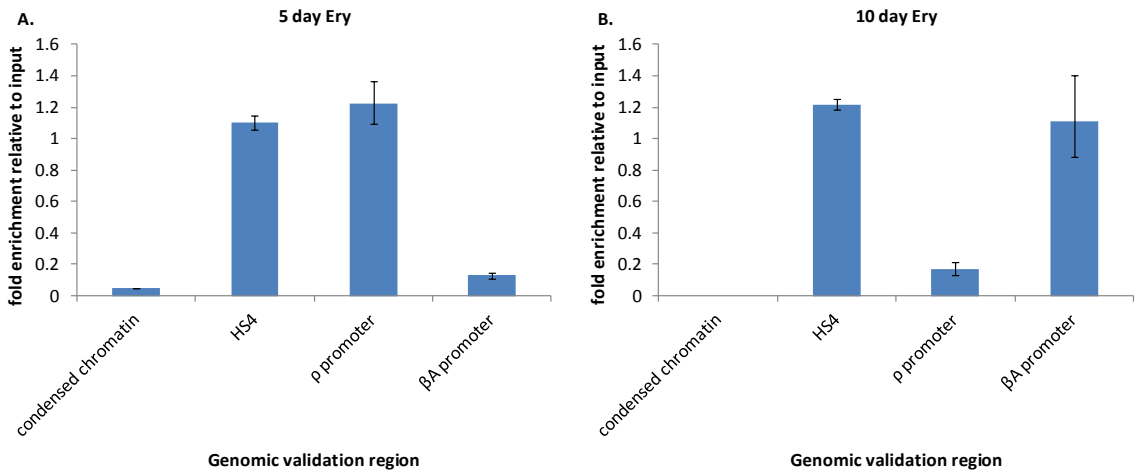


Figure 6.5 Validation of VEZF1 ChIP from embryonic erythrocytes.
QPCR analysis of β-globin gene locus sequences following VEZF1 ChIP in (A) 5 day and (B) 10 day embryonic erythrocytes. Primer sets targeting the condensed chromatin 5' of HS4, the HS4 insulator, the ρ promoter and the β^A promoter were used for the analysis. The relative enrichment of each element after VEZF1 ChIP was normalised to that of starting input chromatin. Error bars represent standard deviation between triplicate QPCR analyses of a VEZF1 ChIP.

VEZF1 ChIP-seq libraries were validated for the enrichment of previously identified VEZF1 binding targets using QPCR analysis. This analysis confirmed that HS4 insulator sequences remain highly enriched in the VEZF1 ChIP-seq libraries from 5 and 10 day erythrocytes (Figure 6.6). The ρ-globin promoter also remains highly enriched in the 5 day erythrocyte library, but the enrichment of β^A promoter sequences appear to be lost following library preparation from 10 day erythrocytes. The apparent alteration in relative enrichments may be due to issues relating to sequence representation or PCR bias during library preparation. However, it may simply be due to the location of the QPCR primer set relative to the true VEZF1 binding event(s) at the β^A promoter. The initial QPCR on the VEZF1 ChIP has a full range of sonicated chromatin fragment sizes to detect from, whereas the ChIP-seq libraries are size selected, so QPCR primer sets need to overlap the VEZF1 binding peak quite accurately. Given that we are unsure of the true VEZF1 binding event(s) at the β^A promoter and that the enrichments of the HS4 insulator remain consistently high, we decided to proceed with sequencing of both the VEZF1 ChIP-seq libraries from 5 and 10 day erythrocytes.

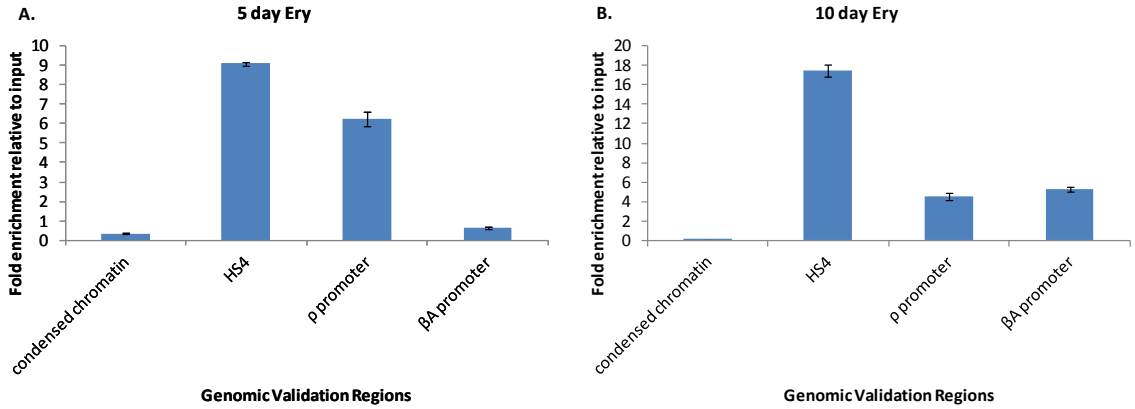


Figure 6.6 Validation of VEZF1 ChIP-seq library preparations.
QPCR analysis of β-globin gene locus sequences following VEZF1 ChIP in (A) 5 day and (B) 10 day embryonic erythrocytes. Primer sets targeting the condensed chromatin 5' of HS4, the HS4 insulator, the ρ promoter and the β^A promoter were used for the analysis. The relative enrichment of each element after VEZF1 ChIP-seq library preparation was normalised to that of starting input chromatin. Error bars represent standard deviation between triplicate QPCR analyses of a VEZF1 ChIP.

The VEZF1 ChIP-seq libraries from 5 and 10 day embryonic erythrocytes were given to the University of Glasgow Polyomics Facility for sequencing on an Illumina GAIIx sequencer. The concentration of correctly adapted fragments in each ChIP-seq library was quantified using the SYBR QPCR KAPA library Quantification Kit from Illumina (KK4822) by the University of Glasgow Polyomics Facility (Table 6.1). An aliquot of each library was used to prepare a 20 µl solution at 1.5 nM for denaturation (Table 6.1). Denatured library samples were further diluted to 12 pM concentrations for loading on the flow cell.

Sample	Concentration	Volume for 20 µl at 1.5nM
5 day:VEZF1	6.93 nM	4.3 µl
10 day:VEZF1	6.78 nM	4.4 µl

Table 6.1 ChIP-seq library quantifications.
The concentration of ChIP-seq libraries were quantified by QPCR and used to calculate the volume of each library required to prepare a 20 µl sample at 1.5 nM for denaturation.

6.3.3 Chicken VEZF1 ChIP-seq data quality

The performance of the Illumina sequencing was assessed using FastQC. Both of the VEZF1 ChIP-seq runs yielded around 47 million reads, with high average qualities and no adaptor sequence contamination (Table 6.2)

Sample	Read length	Read number	Mean Q score	Adaptor sequence
5 day:VEZF1	73	46,891,117	39	No significant level
10 day:VEZF1	76	47,534,861	39	No significant level

Table 6.2 Illumina GAIIx sequencing performance.
Data extracted from FastQC reports from 2 ChIP-seq runs.

The box-and-whisker plots of phred quality scores throughout each sequencing run show that the median sequence quality remains within the very good range throughout the length of both sequence runs (Q scores of 28 – 40) (Figure 6.7). However, the base calls of sequences in the 10^{th} and 90^{th} percentiles have poor Q scores after ~64 bases. These results indicate that very good performance levels have been achieved in these sequencing runs, although a proportion of full length reads may not align to the chicken reference genome due to lower Q scores at the 3' ends of some reads.

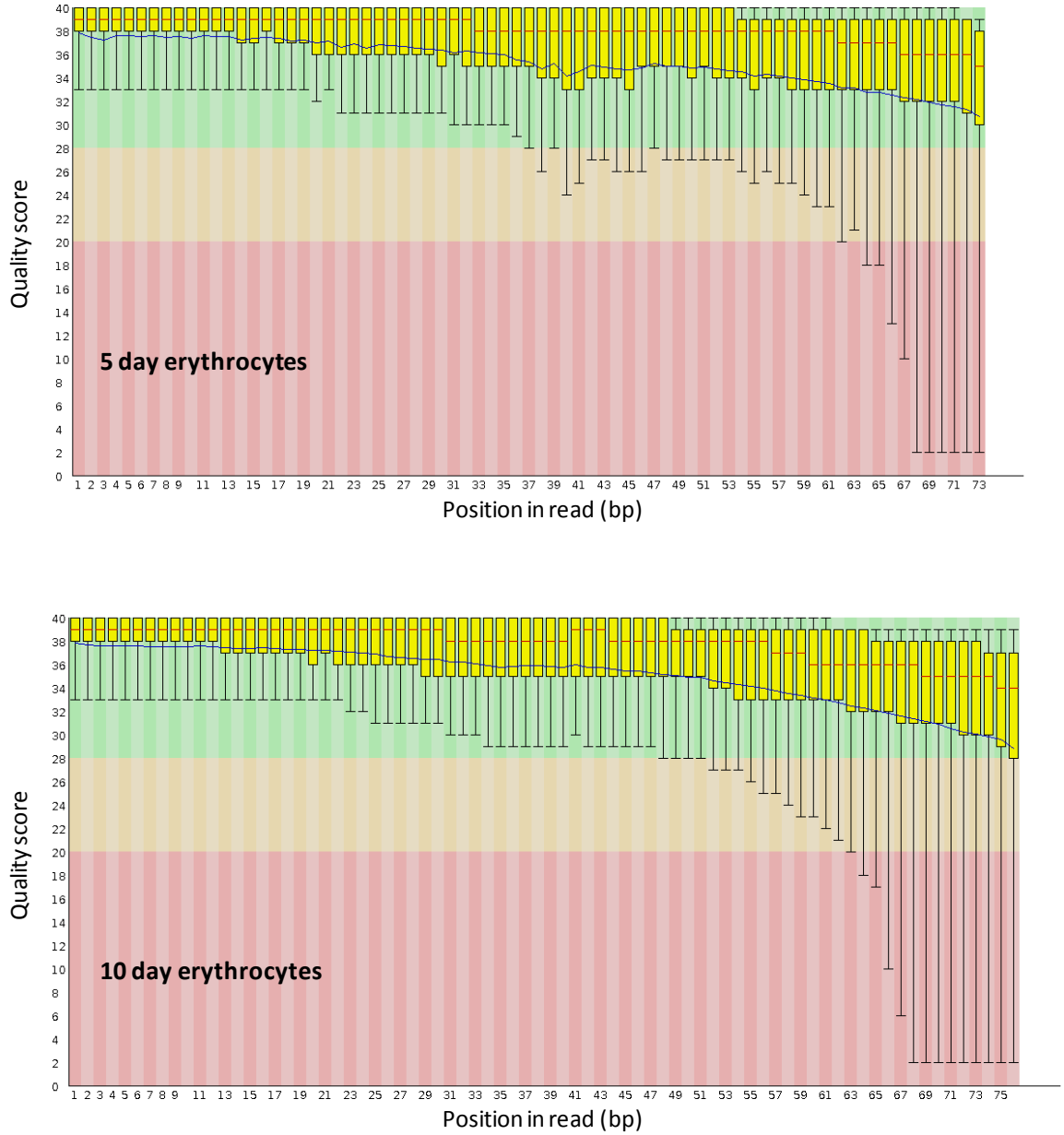


Figure 6.7 Illumina GAIIx sequencing per base quality.
FastQC report of Q score per base call for 5 day and 10 day VEZF1 ChIP-seq runs. Horizontal red lines show the median Q score for each base position and the connected blue line indicates the average quality score. Yellow boxes represent the 20th-80th quartile range, while black whiskers represent the 10th and 90th percentiles. The background of these plots are divided into three sections where green, orange and red shading indicate Q score ranges considered to be of very good, reasonable or poor quality, respectively.

Full length ChIP seq reads were aligned to the galGal4 chicken reference genome by the Bowtie software package using the "-m3" command to allow for alignment to the duplicated globin genes. 71% and 74% of VEZF1 ChIP-seq reads from the 5 and 10 day preparations were successfully aligned to the chicken reference genome (Table 6.3). Unfortunately, 80 % of VEZF1:5 day aligned reads and 84% of 10 day aligned reads were removed as they were not unique, leaving 6,438,091 unique aligned reads from 5 day

VEZF1 ChIP-seq and 5,725,147 unique aligned reads from 10 day VEZF1 ChIP-seq. The apparent high clonality in these libraries is likely to be due to too low an amount of VEZF1 ChIP DNA entering the library preparation. We will seek to sonicate chromatin to a lower average fragment size in future to increase the concentration of fragments that fall within the Illumina library size selection.

Sample	Total reads	Aligned reads (Reported)	Failed to align	Removed (PCR filter)	Unique aligned
5 day:VEZF1	46,891,117	33,355,819 (71.13%)	13,535,298 (28.87%)	26,917,728 (80.1%)	6,438,091
10 day:VEZF1	47,534,861	35,144,419 (73.93%)	12,390,442 (26.07%)	29,419,272 (83.7%)	5,725,147

Table 6.3 Performance of Bowtie ChIP-seq read alignment

Peaks of VEZF1 ChIP-seq read enrichment were identified using MACS (see section 3.6 for a description). We did not have an input or non-immune IgG ChIP-seq reference dataset for 5 or 10 day erythrocytes at this stage, so we asked MACS to calculate peaks from the background signal in the VEZF1 ChIP-seq datasets. This approach can result in false positive and false negative peak calling, but it allows us to make a start in understanding VEZF1 binding events. Initial MACS peak finding was performed on reads that were aligned to defined chromosomes only (~4.5% of reads removed) using default settings. Using these parameters 4,878 and 5,879 VEZF1 peaks were identified in the 5 and 10 day ChIP-seq samples respectively. PeakSplitter was used to resolve clustered peaks to yield a total of 5,398 and 6,131 peaks for VEZF1 in 5 and 10 day erythrocytes, respectively. Further peak finding is planned once non-immune IgG ChIP-seq data are available for use in creating a more representative MACS background model.

6.4 Identification of specific DNA elements bound by VEZF1 at β-globin gene regulatory elements

The enrichment of VEZF1 ChIP-seq reads across the β-globin gene cluster was compared between 5 and 10 day erythrocytes. Consistent with previous ChIP-QPCR analyses, VEZF1 enrichments are observed at the HS4 insulator at both embryonic stages (Figure 6.8). Striking enrichments of VEZF1 are also observed at the both the ρ and ε embryonic globin gene promoters in 5 day embryonic erythrocytes. Little or no VEZF1 biding is observed at these elements in 10 day embryonic erythrocytes. This profile is consistent with VEZF1 binding to the embryonic globin gene promoters specifically when they are transcriptionally active (Figure 6.2). The specific binding of VEZF1 to the ρ-globin promoter is consistent with previous ChIP-QPCR analyses (Figures 6.1 and 6.5), but the discovery of strong VEZF1 binding at the ε-globin promoter is novel. A minor enrichment of VEZF1 at the ε-globin promoter has been observed by ChIP-QPCR previously (Figure 6.1), but close inspection using the UCSC genome browser shows that the QPCR primer set is located 270 bp away from the VEZF1 peak centre.

VEZF1 enrichment is also observed at the β^A globin promoter in 10 day embryonic erythrocytes, but not in 5 day embryonic erythrocytes (Figure 6.8). The specific binding of VEZF1 to the β^A globin promoter is consistent with previous ChIP-QPCR analyses (Figures 6.1 and 6.5). Again, this profile is consistent with VEZF1 binding to the β^A gene promoter specifically when it is transcriptionally active (Figure 6.2). VEZF1 binding is also observed at β^A/ε and HS2, the predominant enhancers of the β-globin locus (Figure 6.8).

Comparison of the VEZF1 binding peaks, identified by MACS, with the ChIP-seq tracks on the UCSC genome browser shows that MACS has identified the majority of the observed peaks of VEZF1 ChIP-seq read enrichment (Figure 6.8). However, the width of the peaks has not been defined accurately and the binding of VEZF1 to HS2 in 5 day erythrocytes and to the β^A/ε enhancer in 10 day erythrocytes has not been identified. These problems are likely to be due to the background model being created from the VEZF1 ChIP-seq tracks rather than a negative control. Once non-immune IgG ChIP-seq samples have been sequenced, the MACS analysis will be re-run. Further analyses of peak numbers, locations

and underlying sequence motifs will then be feasible.

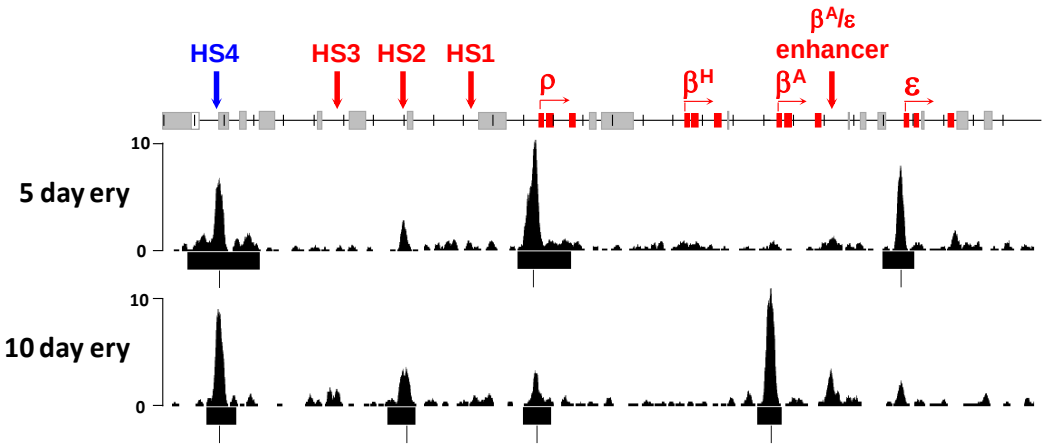


Figure 6.8 The interaction of VEZF1 with β-globin gene regulatory elements is developmentally regulated.
UCSC genome browser view of VEZF1 ChIP-seq data from 5 and day 10 erythrocytes mapped to the β-globin gene cluster (gg4, chr1: 193,703,000-193,733,000). Peaks of ChIP-seq read enrichment identified by MACS are shown below each track. ChIP-seq read numbers are normalised to total library size. A scale schematic of the chicken β-globin locus annotated to indicate the position of known gene regulatory elements. A gg4 assembly error that creates a 690 bp duplication at chr1:193711749-193712439 has been removed from this view.

6.4.1 VEZF1 binding elements at the HS4 insulator element

VEZF1 binding at the HS4 insulator has previously been mapped to three DNaseI footprinted sequences FI, FIII and FV by EMSA analyses (Dickson *et al*., 2010). We therefore compared the VEZF1 ChIP-seq profile in 5 and 10 day erythrocytes to these sequences to test the validity of using ChIP-seq to pinpoint VEZF1 binding elements *in vivo*. The maximal VEZF1 ChIP-seq read enrichment at HS4 aligns to footprints FI and FIII, with a lesser enrichment at FV (Figure 6.9). This profile is entirely consistent with the known affinity of VEZF1 to each of these sequence elements. I was therefore confident to use this approach to pinpoint the *in vivo* targets of VEZF1 binding to other gene regulatory elements at the β-globin gene cluster.

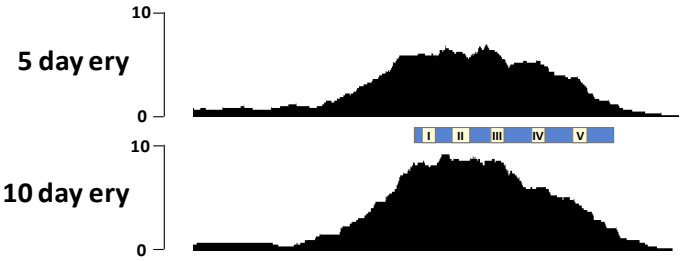


Figure 6.9 The interaction of VEZF1 with the HS4 insulator element.
UCSC genome browser view of VEZF1 ChIP-seq data from 5 and day 10 erythrocytes mapped to the HS4 insulator element (gg4, chr1:193,704,573-193,705,164). ChIP-seq read numbers are normalised to total library size. A scale schematic of the 276 bp core HS4 insulator element with previously characterised footprint elements is shown.

6.4.2 Putative VEZF1 binding elements at the embryonic ρ and ε-globin promoters

ChIP-QPCR and ChIP-seq experiments have demonstrated that VEZF1 interacts with the promoters of the ρ and ε-globin genes specifically in 5 day embryonic erythrocytes (Figures 6.5 and 6.8). Previous characterisation of the ρ-globin promoter failed to identify discrete protein binding sites by DNaseI footprinting as the majority of the promoter was footprinted by erythrocyte nuclear extracts (Minie *et al.*, 1992). Four putative VEZF1 binding elements can be found within 300 bp of the ρ-globin transcription start site. Of these, it is apparent that a GGGGTGGGG motif centred 55 bp upstream of the TSS is the site of maximal VEZF1 ChIP-seq read enrichment in 5 day erythrocytes (Figure 6.10). It is notable that this VEZF1 site is 23 bp from a GATA motif. The asymmetry of the VEZF1 peak suggests that VEZF1 may also occupy additional motifs further upstream of the TSS.

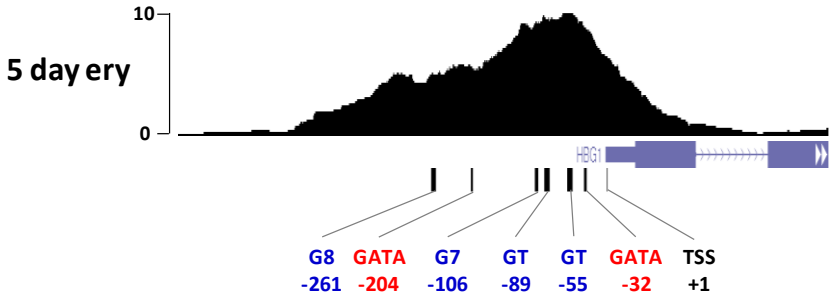


Figure 6.10 The interaction of VEZF1 with the ρ-globin gene promoter.
UCSC genome browser view of VEZF1 ChIP-seq data from 5 day erythrocytes mapped to the ρ-globin (annotated at *HBG1*) gene promoter (gg4, chr1:193,715,634-193,716,633). ChIP-seq read numbers are normalised to total library size. The coordinates of putative VEZF1 binding sites (blue) and GATA motifs (red) are shown beneath.

The arrangement of VEZF1 and GATA motifs at the core ε-globin promoter is very similar to that of the ρ-globin promoter. Previous characterisation of the ε-globin promoter

found that a GGGGTGGGG motif centred 55 bp upstream of the TSS is footprinted by embryonic erythrocyte extracted proteins (Mason *et al*., 1996). It is apparent that this GT motif precisely located to the peak of maximal VEZF1 ChIP-seq read enrichment in 5 day erythrocytes (Figure 6.11). Again this GT motif is located 23 bp from a GATA motif, which is also footprinted by erythrocyte extracts (Mason *et al*., 1996). In contrast to the ρ-globin promoter, there are no other apparent VEZF1 motifs further upstream of the core GT motif.

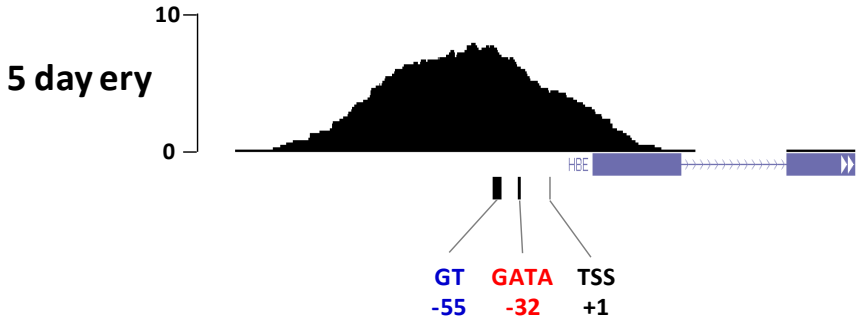


Figure 6.11 The interaction of VEZF1 with the ε-globin gene promoter.
UCSC genome browser view of VEZF1 ChIP-seq data from 5 day erythrocytes mapped to the ε-globin (annotated at *HBE*) gene promoter (gg4, chr1:193,728,157-193,728,906). ChIP-seq read numbers are normalised to total library size. The coordinates of putative VEZF1 binding sites (blue) and GATA motifs (red) are shown beneath.

6.4.3 Putative VEZF1 binding elements at the β^A globin promoter

A number of previous studies have carefully characterised the regulatory elements at the β^A-globin promoter. One of the earliest studies identified a DNA-binding activity, termed Beta Globin Protein 1 (BGP1), that interacts with a long homopolymeric G string located 186 bp upstream of the TSS (Lewis *et al*., 1988). Recently, BGP1 was identified to be VEZF1 (Dickson *et al*., 2010). Subsequent studies identified a number of additional elements that are footprinted by erythrocyte factor binding *in vitro* and make varying contributions to promoter activity in reporter assays (Emerson *et al*., 1985, Gallarda *et al*., 1989). There are five putative VEZF1 binding elements within the sequences 200 bp upstream of the β^A globin gene TSS. All of these elements fall within the peak of VEZF1 ChIP-seq read enrichment in 10 day erythrocytes (Figure 6.12). The VEZF1 ChIP-seq peak at the β^A-globin promoter is broad and likely incorporates multiple VEZF1 binding events. Close inspection indicates that the 17 base G-string at -186, the GGGGGAGGG motif at -115 and the GAGGAGGGG motif at -46 lie at the three apparent sub-peaks. The GA site at

-46 is 18 bp from a GATA motif. Both lie within the Stage Selector Element (SSE), which is likely to determine stage-specific binding of VEZF1 (Choi and Engel, 1988).

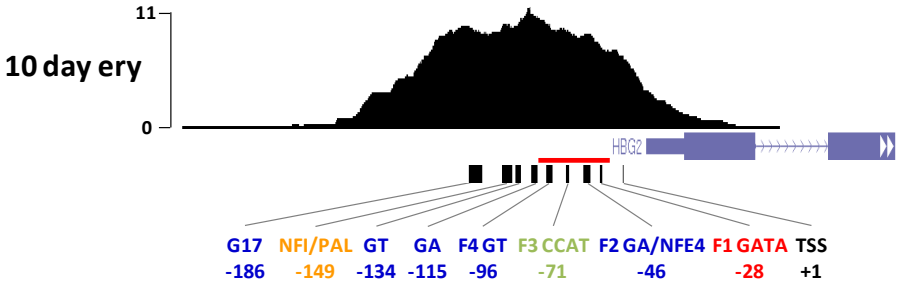


Figure 6.12 The interaction of VEZF1 with the β^A globin gene promoter.
UCSC genome browser view of VEZF1 ChIP-seq data from 10 day erythrocytes mapped to the adult β-globin (annotated at *HBG2*) gene promoter (gg4, chr1:193,723,622-193,724,621). ChIP-seq read numbers are normalised to total library size. The coordinates of putative VEZF1 binding sites (blue), NFI (orange), CTF (green) and GATA motifs (red) are shown beneath. Promoter numbering is taken from (Emerson *et al*., 1985). Footprints F1 to F4 are taken from (Gallarda *et al*., 1989). Sequences incorporating the stage selector element (SSE) defined by (Choi and Engel, 1988) are indicated with a horizontal red line.

6.4.4 Putative VEZF1 binding elements at the β-globin HS2 and $\beta^{A/\varepsilon}$ enhancers

Previous studies have mapped the binding of at least five factors to the $\beta^{A/\varepsilon}$ enhancer in erythrocyte extracts (Emerson *et al*., 1987, Gallarda *et al*., 1989). The peak of maximal VEZF1 ChIP-seq read enrichment in 10 day erythrocytes maps to a GGGTGGGGG motif known as F3 in the $\beta^{A/\varepsilon}$ enhancer (Figure 6.13). The lack of additional sub-peaks or additional putative VEZF1 motifs indicates that F3 is the only VEZF1 site at this element. F3 is located 16 bp away from a palindromic GATA-TATC motif, which may be responsible for the erythroid-specific binding of VEZF1 to this element.

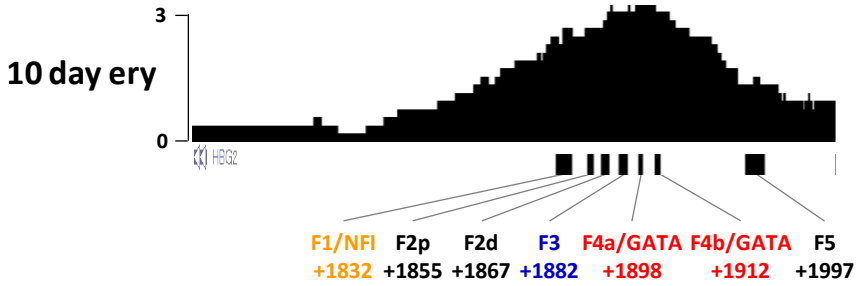


Figure 6.13 The interaction of VEZF1 with the $\beta^{A/\varepsilon}$ enhancer.
UCSC genome browser view of VEZF1 ChIP-seq data from 10 day erythrocytes mapped to the $\beta^{A/\varepsilon}$ enhancer located downstream of the adult β-globin (annotated at *HBG2*) gene (gg4, chr1:193,725,790-193,726,340). ChIP-seq read numbers are normalised to total library size. The coordinates of putative VEZF1 binding sites (blue), NFI (orange) and GATA motifs (red) are shown beneath. The location of DNaseI footprints F1 to F5 are taken from (Emerson *et al*., 1987, Gallarda *et al*., 1989). Numbering is relative to adult β-globin TSS.

A previous study mapped at least seven transcription factor binding events at the HS2 enhancer in erythrocyte extracts (Abruzzo and Reitman, 1994). Two of these motifs, a GGTGCGGTGGG motif called F2 and a GGGCGTGGGG motif called F5, appear to overlap the peaks of maximal VEZF1 ChIP-seq read enrichment in 10 day erythrocytes (Figure 6.14). However, VEZF1 appears to only interact with the F2 site in 5 day erythrocytes. Both the F2 and F5 GT motifs are located 24 bp from GATA binding sites.

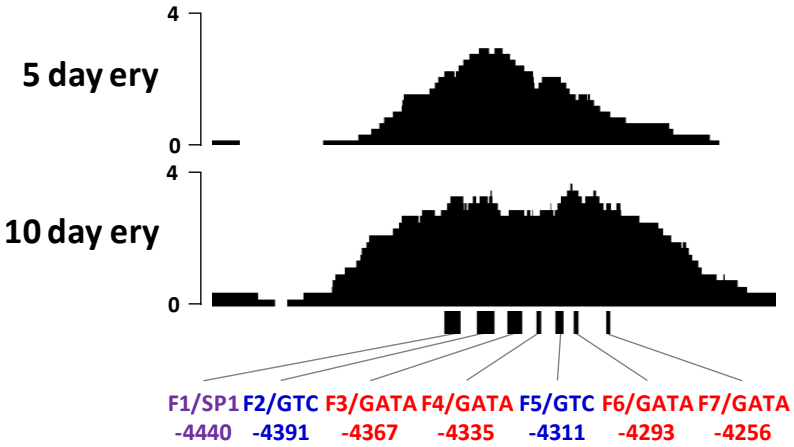


Figure 6.14 The interaction of VEZF1 with the HS2 β-globin enhancer.
UCSC genome browser view of VEZF1 ChIP-seq data from 5 and 10 day erythrocytes mapped to the HS2 β-globin enhancer located upstream of the ρ-globin gene (gg4, chr1:193,710,790-193,711,550). ChIP-seq read numbers are normalised to total library size. The coordinates of putative VEZF1 binding sites (blue), SP1 (purple) and GATA motifs (red) are shown beneath. The location of DNaseI footprints F1 to F7 are taken from(Abruzzo and Reitman, 1994). Numbering is relative to ρ-globin TSS.

6.5 The affinity of VEZF1 for putative binding sequences found at β-globin gene promoters

The binding of VEZF1 to the embryonic ρ- and ε-globin promoters specifically in primitive erythrocytes from 5 day stage embryos is novel and unexpected. In order to define the specificity and affinity of VEZF1 for putative sites in these promoters, sequences were selected for EMSA analyses of VEZF1 binding *in vitro*. The selection of putative VEZF1 binding sites was performed prior to the completion of the VEZF1 ChIP-seq analysis described above. We had ChIP-QPCR data that revealed VEZF1 binding at the ρ-globin promoter in 5 day erythrocytes and we had the VEZF1 consensus motif GGGGnGGGG from an earlier VEZF1 ChIP-chip study (Strogantsev, 2009). Three 40 bp putative VEZF1 binding sequences were selected for EMSA analysis. These incorporated the G8, G7 and GT motifs located 261, 106 and 55 bp upstream of the ρ-globin TSS (Figure 6.10) (Table 6.4).

Probe	Probe sequence
Rho_G8-261	GAATGCATCA**CG**CAGAGGGGGGGGTTTGGTGCCTTCTGCA
Rho_G7-106	**G**GGTGGGGGTC**CG**TGC**C**GGGGGGGTC**CG**TCCATGGGGTGG
Rho_GT-55	TGACCCCACAGCATGGGGTGGGGAGGAGCTGTCAG**CG**GTG
VEZF1 consensus	nnnnnnnnnnnnnnnnGGGGnGGGGnnnnnnnnnnnnnnnn

Table 6.4 Putative VEZF1 binding motifs within the ρ-globin gene promoter.
Putative VEZF1 binding motifs are underlined, G residues associated with these motifs are shown in red.

Each of the three sites were prepared as double stranded oligonucleotides that were radiolabelled for use in direct EMSA analyses as described previously (section 5.5). A series of six binding reactions were set up using each probe sequence, as before. The first four reactions test whether endogenous proteins in chicken erythrocyte nuclear extract interact with the test sequence, whether any resulting complexes contain VEZF1 (supershift with anti-VEZF1 antibodies) and whether they have the same DNA sequence specificity as reported for VEZF1 (competition with F1 and F1aaa1). The latter two reactions test whether recombinant VEZF1 interacts with the test sequence. The same concentration of DNA probes and proteins were used in each direct EMSA experiment below. In addition, the Rho_G7-106 site was methylated at the CpG dinucleotide associated with the putative VEZF1 binding motif, to examine whether methylation

overlapping this VEZF1 site might be responsible for regulating VEZF1 binding at the ρ-globin promoter during development.

Direct EMSA analysis shows that VEZF1 does interact with each of the ρ-globin elements *in vitro*, but that different relative affinities for each site can be observed (Figure 6.15). The Rho_G8-261 sequence is the highest affinity site for recombinant VEZF1, followed by Rho_G7-106 (Figure 6.15, compare lane 5 in panels B and D). Incubation with anti-VEZF1 antibodies results in the inhibition of binding of full length VEZF1 or supershifted VEZF1:DNA complexes for each of these elements (Figure 6.15, compare lanes 5 and 6 in panels B and D). This is consistent with high affinity of VEZF1 for isolated G string sequences described earlier (section 5.5.4.1). Direct EMSA analysis of erythrocyte nuclear protein interactions with each of the G string sequences confirms the formation of complexes of the expected mobility for full length VEZF1, which are supershifted by anti-VEZF1 antibodies (Figure 6.15, compare lanes 1 and 2 in panels B and D). However, the degree of VEZF1 binding to these elements is much lower than observed for longer G-string elements studied in chapter 5, as shown by significant competition with the F1aaa1 mutant VEZF1 site. CpG methylation of the Rho_G7-106 sequence has no effect of VEZF1 binding (Figure 6.15, compare panels B and C).

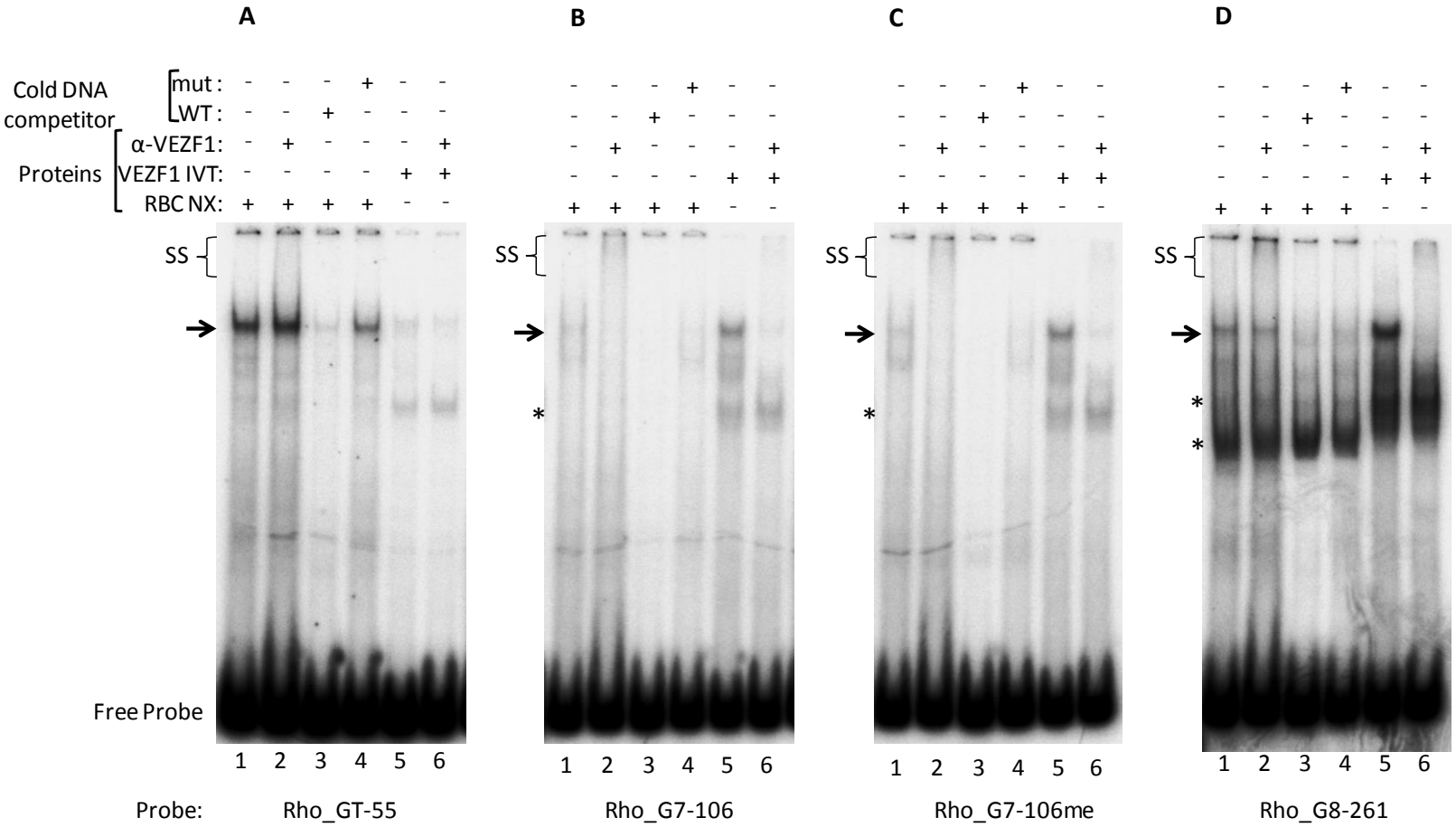


Figure 6.15 VEZF1 interactions with putative VEZF1 binding sites in the ρ-globin gene promoter. Oligonucleotides corresponding in sequence to putative VEZF1 binding sites in the β-*Rho* promoter were radiolabelled and used as probe DNA in a series of EMSAs to determine whether a direct interaction between each DNA motif and VEZF1 protein can occur *in vitro*. VEZF1-probe complexes are indicated by an arrow and confirmed by the supershift (SS) of specific bands by the addition of a VEZF1-specific antibody to binding reactions (A-D lanes 2 and 6). Asterisks indicate non-specific bands.

Direct EMSA analysis shows that VEZF1 binding to the Rho_GT-55 element in isolation is very weak *in vitro*. Complexes with recombinant VEZF1 are barely detectable (Figure 6.15, panel A, lanes 5 and 6). Some of the complexes formed between this site and erythrocyte nuclear proteins are supershifted by VEZF1 antibodies, but the contribution of VEZF1 to these complexes is minor (Figure 6.15, panel A, compare lanes 1 and 2). These findings are in agreement with my earlier results which show that VEZF1 does not efficiently recognise isolated GGGGTGGGG motifs *in vitro* (Section 5.5.4.3).

The 40 bp ρ-globin core promoter sequence Rho_GT-55 studied by direct EMSA is identical to the GT-55 element found at the ε-globin promoter. Both elements are found at the peak of the VEZF1 ChIP-seq enrichments observed at these promoters in 5 day erythrocytes (Figures 6.10 and 6.11). The inability of VEZF1 to interact with these elements when isolated *in vitro* suggests that VEZF1 may require cooperative interactions with co-binding factors when interacting with these promoter elements *in vivo*. Such cooperativity is the likely mechanism that controls the stage-specific binding of VEZF1 to these elements. Both elements share identical spacing to a proximal GATA binding site. The two additional G7 and G8 motifs further upstream in the ρ-globin promoter are weak VEZF1 sites in isolation, but appear to be bound by VEZF1 in 5 day erythrocytes (Figure 6.10). These additional weak VEZF1 sites may contribute to cooperative interactions that stabilise VEZF1 binding to the GT -55 element at the ρ-globin promoter.

6.6 The relationship between VEZF1 binding, promoter DNA methylation and transcription of the *ρ* and β^A genes

6.6.1 Establishment of bisulphite DNA sequencing to study β-globin promoter methylation

VEFZ1 binding sequences have been shown to function in the protection of CG-rich elements at promoters and insulators from *de novo* DNA methylation (Dickson *et al*., 2010). Early studies discovered that DNA methylation patterns at the β-globin genes change during chicken embryonic development (McGhee and Ginder, 1979). More recent studies show that the silencing of ρ-globin gene expression in definitive erythrocytes is dependent on methylation and conversely that activation of the β^A globin gene at this same stage involves some form of DNA demethylation (Singal *et al*., 2002, Ramachandran *et al*., 2007). Given that we have shown that VEZF1 binding to the β-globin gene promoters is tightly regulated during erythrocyte development, it is of interest to study the relationship with DNA methylation.

Given that circulating erythrocytes may contain a mixture of primitive and definitive erythrocytes expressing embryonic or adult β-globin genes, it is important to use an assay that allows the study of individual DNA molecules. We also wanted to use a technique that could study the methylation status of the bulk of CG dinucleotides in β-globin promoters to both reveal methylation patterns with respect to individual VEZF1 sites and to avoid the reliance on individual CGs as reporters for the whole promoter. We therefore opted to employ bisulphite DNA sequencing. Genomic DNA was prepared from $1x10^5$ circulating erythrocytes collected from embryos aged between 5 and 10 days post fertilisation. DNA was bisulphite modified and subjected to Taq polymerase PCR with primers that amplify a 510 bp CpG-rich portion of the ρ-globin promoter and a 257 bp CpG-rich portion of the adult β-globin promoter (Figure 6.16).

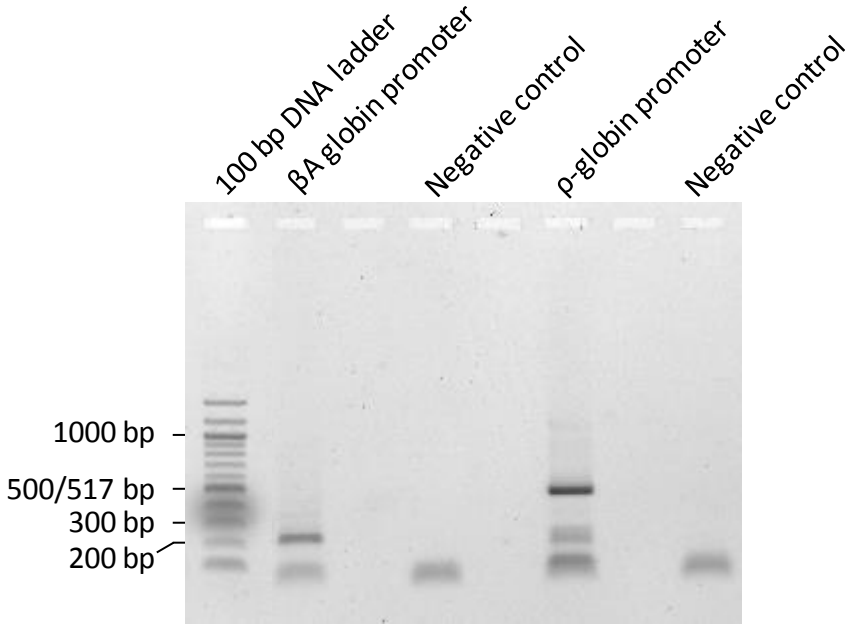


Figure 6.16 PCR amplification of β-globin promoter sequences from bisulphite modified erythrocyte DNA.
Agarose gel electrophoresis of PCR products amplified from bisulphite modified 5 day embryonic erythrocyte DNA. 257 and 510 bp products were expected for the β^A globin and ρ-globin promoter fragments, respectively. No DNA template negative control reactions are also shown for each set of PCR primers.

Bisulphite PCR products were gel purified and TA cloned into the pGEM-T easy vector. Transformed bacterial colonies were screen with blue/white selection and then subjected to colony PCR to check for the insertion of a PCR product of expected size (Figure 6.17). Plasmid DNA was prepared from colonies that passed this screening, with an aliquot checked for correct fragment insertion by restriction analysis (Figure 6.18). Plasmids containing bisulphite modified β-globin promoter fragments were subject to Sanger sequencing. The sequencing chromatograms were inspected to check the performance of sequencing over the cloned PCR products (Figure 6.19). At least 10 plasmid clones were sequenced for each bisulphite PCR in order to gain an average view of the DNA methylation status of 10 individual β-globin promoter molecules at each stage of development studied.

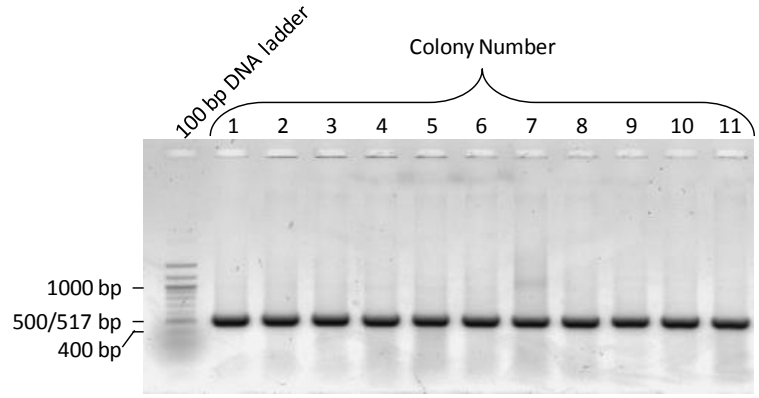


Figure 6.17 Colony PCR analysis of TA cloned bisulphite PCR products.
Agarose gel electrophoresis of colony PCR products from white bacterial colonies transformed with TA cloning reactions. The ligation of the β^A globin promoter bisulphite PCR product into pGEM-T easy would yield a 510 bp colony PCR product. Self ligation of pGEM-T easy would yield a 253 bp colony PCR product.

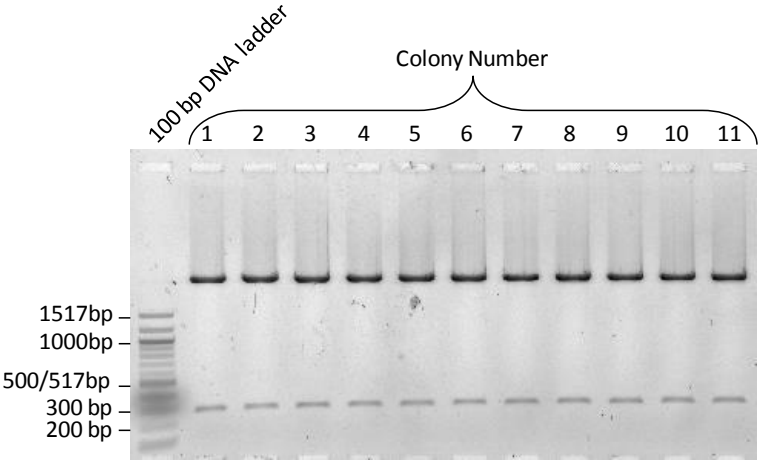


Figure 6.18 Restriction analysis of TA cloned bisulphite PCR products.
Agarose gel electrophoresis of *Eco* RI digested plasmid DNA prepared from white bacterial colonies transformed with TA cloning reactions. The ligation of the β^A globin promoter bisulphite PCR product into pGEM-T easy would yield a 274 bp DNA fragment and a 3006 bp linear vector backbone after *Eco* RI digestion.

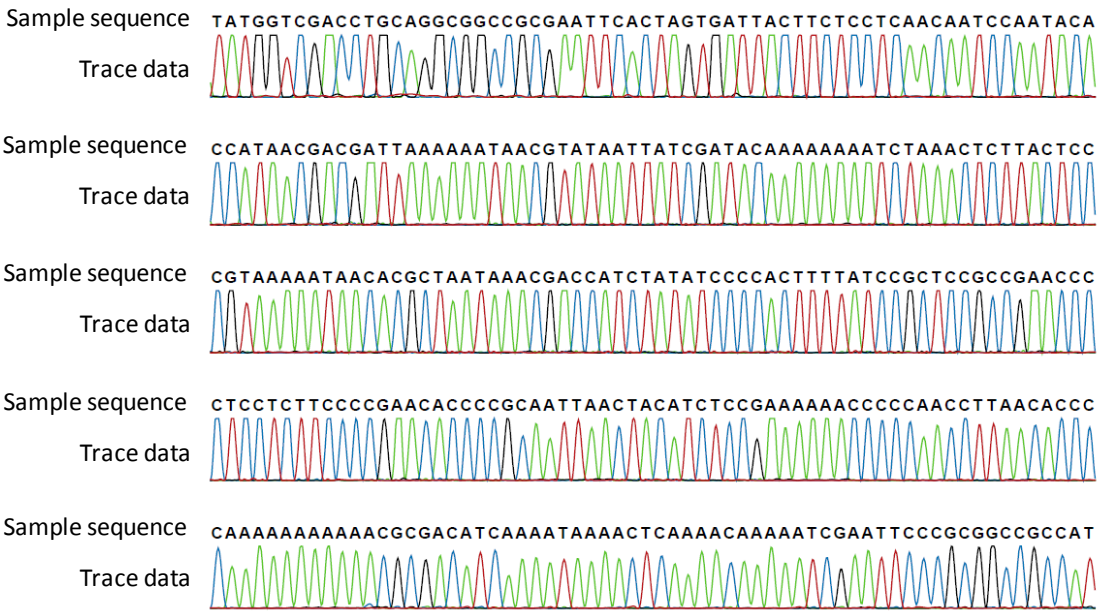


Figure 6.19 Sequence chromatogram of a 5 day β^A promoter sequencing clone.

Bisulphite modified DNA sequences were aligned with their corresponding genomic DNA sequence. Each sequence was firstly checked for the conversion of all non-CpG dinucleotide cytosine residues to thymine indicating complete bisulphite conversion. Fully converted molecules were then scored for the methylation status of CpG-associated cytosines, where TG reports non-methylation and CG reports methylation (Figure 6.20 A). The DNA methylation status of each CpG dinucleotide can then be summarised as a shaded box plot (Figure 6.20 B). Box plots are ordered by the most heavily methylated molecules at the top to the least methylated molecules at the bottom. Finally, an average percentage methylation score was calculated for each CpG based on the methylation status throughout all the molecules sequenced. These scores can be presented as pie charts to allow quick comparisons between analyses (Figure 6.20 C).

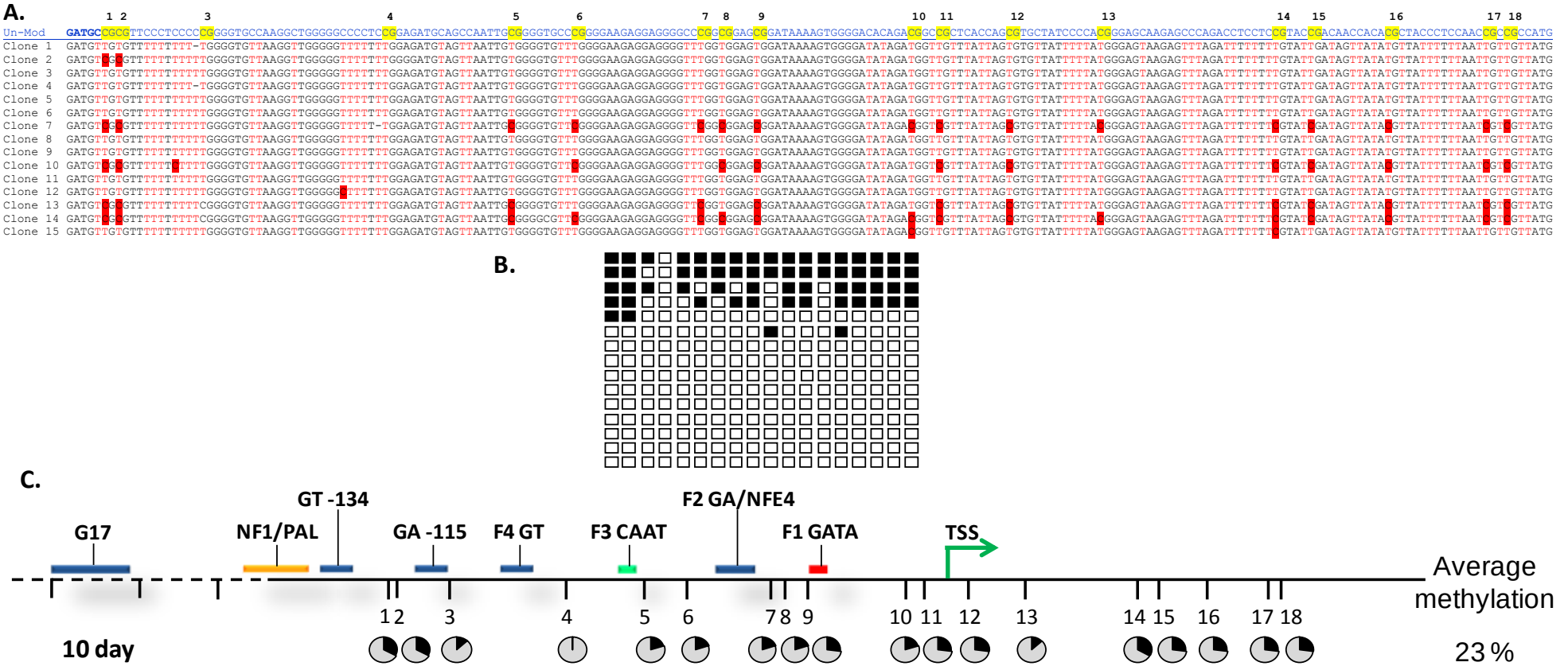


Figure 6.20 Pipeline of analysis of bisulphite converted DNA sequencing samples.
(A) Alignment of bisulphite modified β^A promoter DNA sequences to their corresponding genomic sequence. Bisulphite conversion of C to T is shown in red text. Non-conversion of CpG-associated cytosine residues (methylated) are highlighted. (B) Box plots summarising the non-methylation (open) or methylation (filled) of CpG-associated cytosines. Each row of boxes represents one sequenced clone. (C) Pie chart presentation of percentage methylation scores for each CpG dinucleotide; black represents methylated cytosine, grey represents unmethylated cytosine. The solid black horizontal line represents the region of the βA promoter amplified by PCR, the dashed line represents sequence outwith the PCR amplified region. PCR amplified CpG dinucleotides are numbered 1 – 18. The coordinates of putative VEZF1 binding sites (blue boxes), NF1 (orange box), CTCF (green box) and GATA motifs (red box) are shown above. Promoter numbering is taken from (Emerson *et al.*, 1985). Footprints F1 – F4 are taken from (Gallarda *et al.*, 1989).

6.6.2 DNA methylation status of the ρ-globin promoter in erythrocytes during embryonic development

The DNA methylation status of the ρ-globin promoter element was assessed in circulating erythrocytes collected from embryos 5, 5.5, 7, 8 and 10 days post fertilisation. 10 promoter molecules were subjected to bisulphite sequencing for the 5 and 5.5 day samples and 15 molecules were sequenced from the 7, 8 and 10 day samples (Figure 6.21). This analysis found that the ρ-globin promoter was entirely free from CpG methylation in 5 day erythrocytes. This is consistent with previous analyses that studied the methylation status of individual CpG dinucleotides across the ρ promoter at this time point (Singal *et al*., 1997). At 5.5 days post-fertilisation three of the ten clones sequenced showed very low levels of DNA methylation however by the 7 day time point seven of the fifteen clones sequenced were heavily methylated. In 8 day erythrocytes seven of the fifteen clones were still heavily methylated but showed higher methylation levels than those at embryonic day 7, i.e. there are fewer unmethylated cytosine residues within these clones. By embryonic day 10, when the ρ globin gene is essentially silenced, 13 of the 15 sequenced clones are extensively methylated and the remaining 2 clones show light methylation of the ρ globin promoter, this finding is in line with previous analyses which studied the methylation status of CpG dinucleotides across the ρ promoter at this time point however these analyses found this regulatory element to be completely methylated 10 days after fertilisation (Singal *et al*., 1997) whereas our analysis found it to be 68 % methylated.

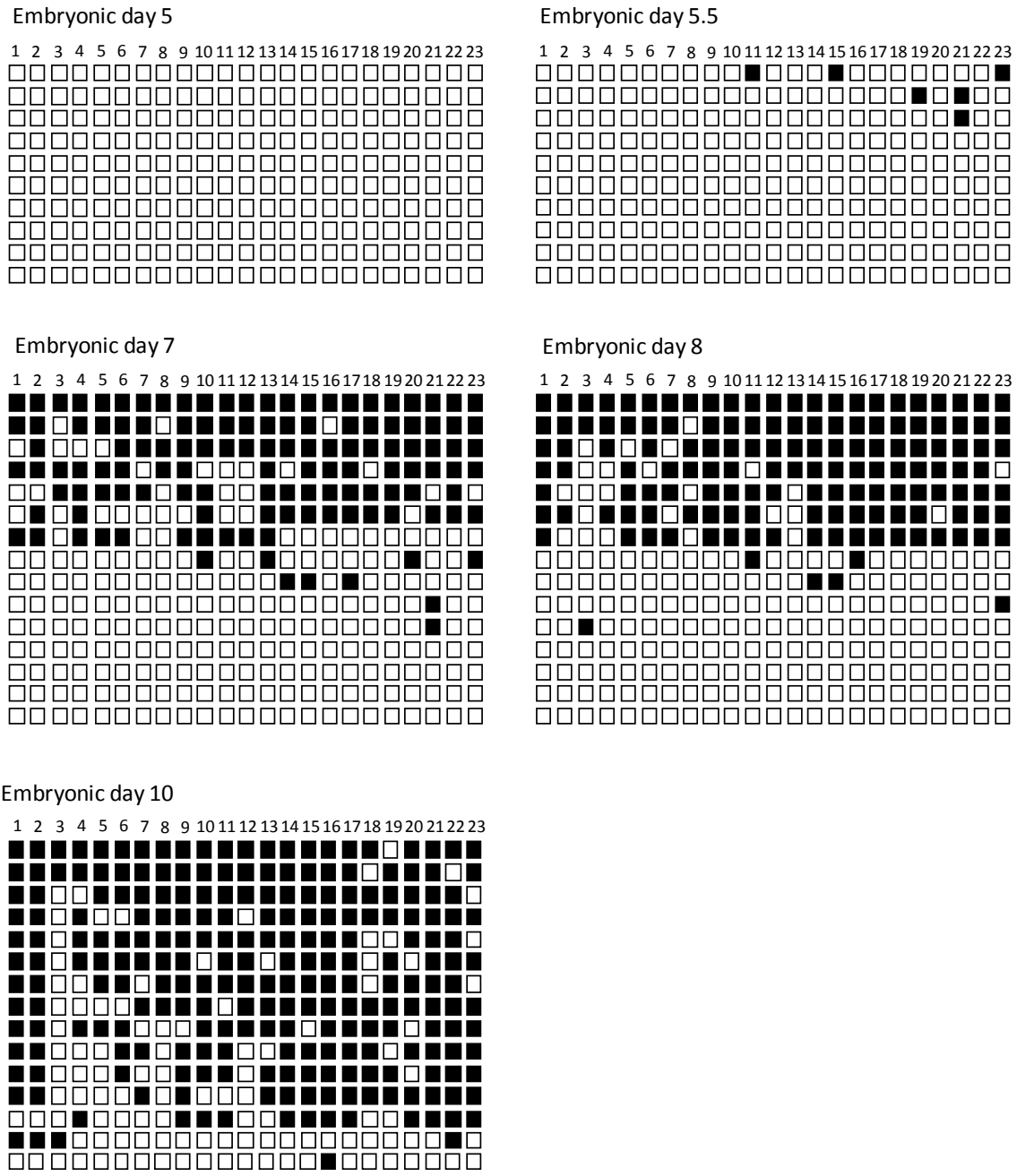


Figure 6.21 The DNA methylation status of the ρ-globin promoter is dynamic during chick embryogenesis.

10 – 15 clones were sequenced for each time point. Each box represents the cytosine residue of one CpG dinucleotide and each row represents one sequenced clone. Black boxes denote methylated cytosine while white boxes denote unmethylated cytosine. CpG dinucleotides are numbered 1 – 23.

The average percentage methylation scores for each CpG site were calculated at each time point and are presented as individual pie charts in figure 6.22. The average over all methylation score for each time point is also shown in this figure. From the data presented in figure 6.22 it is evident that the ρ promoter is entirely unmethylated in circulating red blood cells at day 5 of chick embryogenesis and accumulates DNA methylation throughout the developmental time course studied, peaking at a level of 68 % over all DNA methylation at the latest time point of embryonic day 10.

When these DNA methylation status results are considered alongside the gene expression and VEZF1 ChIP-seq results (figures 6.2 and 6.8 respectively) it is apparent that VEZF1 binding at the ρ promoter correlates with active expression of this gene and a wholly unmethylated promoter element at embryonic day 5. As embryogenesis progresses and the ρ gene is silenced, a steady increase in DNA methylation within the ρ promoter in erythrocytes is apparent. This methylation reaches its peak at embryonic day 10 at which point VEZF1 protein is no longer associated with the ρ promoter element. Thus the association of VEZF1 with the ρ promoter can be seen to correlate with active gene expression and an unmethylated or lightly methylated DNA state.

It is indeterminable from this data whether the increase in DNA methylation levels seen across the ρ promoter results from an accumulation of CpG methylation in erythrocytes throughout the time course studied or whether the changes seen reflect the presence of two distinct cell populations, i.e. primitive and definitive. In the case of the latter scenario, one population of cells which is unmethylated across the ρ promoter and is predominant early in embryogenesis may be outgrown over this developmental time course by a second population of cells which is heavily methylated across the ρ promoter.

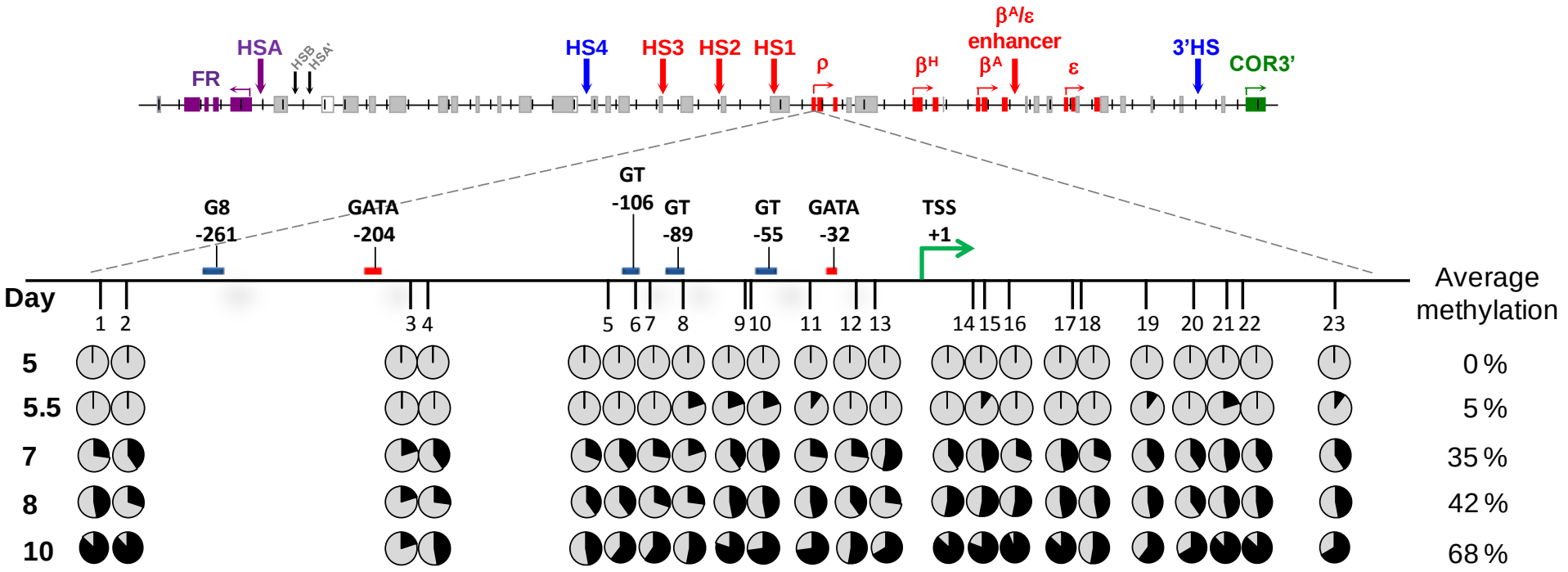


Figure 6.22 The DNA methylation status of the ρ-globin gene promoter is dynamic throughout chick embryogenesis.
Solid horizontal line denotes the genomic region studied by bisulphite sequencing. Vertical lines mark sites of CpG dinucleotides. Pie charts represent individual CpG dinucleotides at indicated time points during chick embryogenesis. Pie chart shading indicates the average DNA methylation status of each CpG dinucleotide; black represents methylated cytosine, grey represents unmethylated cytosine. Average overall DNA methylation of all sequenced clones at each time point is shown on the right of the figure. 10 or 15 clones were sequenced per time point. The coordinates of putative VEZF1 binding sites (blue boxes), GATA motifs (red boxes) and the ρ globin TSS are indicated.

6.6.3 DNA methylation status of the β^A promoter throughout chick embryogenesis

The DNA methylation status of the β^A-globin promoter element was assessed in circulating erythrocytes collected from embryos 5, 5.5, 7, 8, 9 and 10 days post fertilisation. 10 promoter molecules were subjected to bisulphite sequencing for the 5, 5.5 and 9 day samples and 15 molecules were sequenced from the 7, 8 and 10 day samples (Figure 6.23). This analysis found that the β^A promoter was 83 % methylated in 5 day erythrocytes. At 5.5 day post-fertilisation all sequenced clones remained heavily methylated with 81 % overall methylation. At embryonic day 7 the majority of clones remain heavily methylated however three of the fifteen clones sequenced showed very little DNA methylation. By embryonic day 8, seven of the fifteen sequenced clones are lightly methylated or unmethylated. At embryonic day 9 only two of the ten clones sequenced showed any DNA methylation and at embryonic day 10 only four out of fifteen sequenced clones were heavily methylated. Thus the frequency of methylated clones is seen to decrease over this developmental time course, these findings are consistent with previous analyses which studied the methylation status of CpG dinucleotides across the β^A promoter at 5, 8 and 10 days post-fertilisation by bisulphite sequencing (Ramachandran *et al*., 2007).

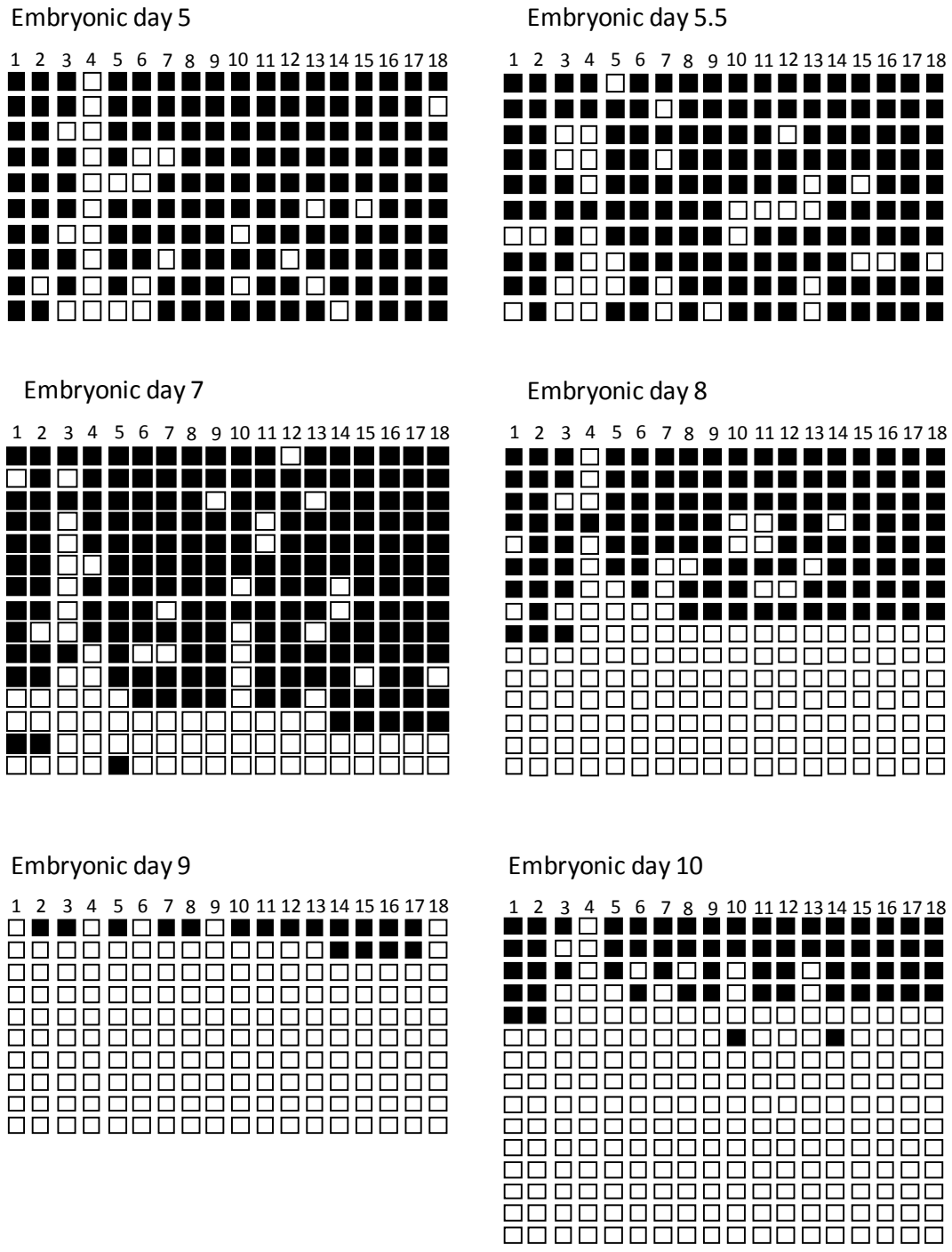


Figure 6.23 The DNA methylation status of the β^A promoter is dynamic during chick embryogenesis.

10 – 15 clones were sequenced for each time point. Each box represents the cytosine residue of one CpG dinucleotide and each row represents one sequenced clone. Black boxes denote methylated cytosine while white boxes denote unmethylated cytosine. CpG dinucleotides are numbered 1 – 18.

The average percentage methylation scores for each CpG site were calculated at each time point and are presented as individual pie charts in figure 6.24. The average over all methylation score for each time point is also shown in this figure. It is clear from the data collected that the β^A promoter is heavily methylated in circulating erythrocytes at embryonic day 5 and that this genomic element becomes progressively less methylated as embryogenesis proceeds reaching just 9 % over all methylation at embryonic day 9. At embryonic day 10, which was the latest time point analysed, the average level of overall DNA methylation across the β^A promoter was 23 %.

When these results are considered collectively with the β^A gene expression data and VEZF1 ChIP-seq enrichment profiles at the β^A promoter (figures 6.2 and 6.8 respectively) it is apparent that in 5 day red blood cells where the β^A gene is not expressed its promoter is not bound by VEZF1 and is heavily methylated. During the window of embryonic development investigated here the β^A gene is seen to become actively expressed and reaches a plateau in mRNA expression at days 9 and 10 (figure 6.2). This activation correlates with reduced DNA methylation across the β^A promoter in circulating erythrocytes isolated from chick embryos (figures 6.23 and 6.24). The observed activation of β^A gene expression and reduction in DNA methylation levels of this gene promoter element can also be seen to correlate with the absence of association with VEZF1 protein in 5 day red blood cells and the enrichment of the β^A promoter with VEZF1 in 10 day red blood cells (figure 6.8).

We cannot tell from this data whether the progressive reduction in DNA methylation of the β^A gene promoter is the result of active DNA demethylation or is due to the expansion of a population of cells where the β^A promoter is unmethylated.

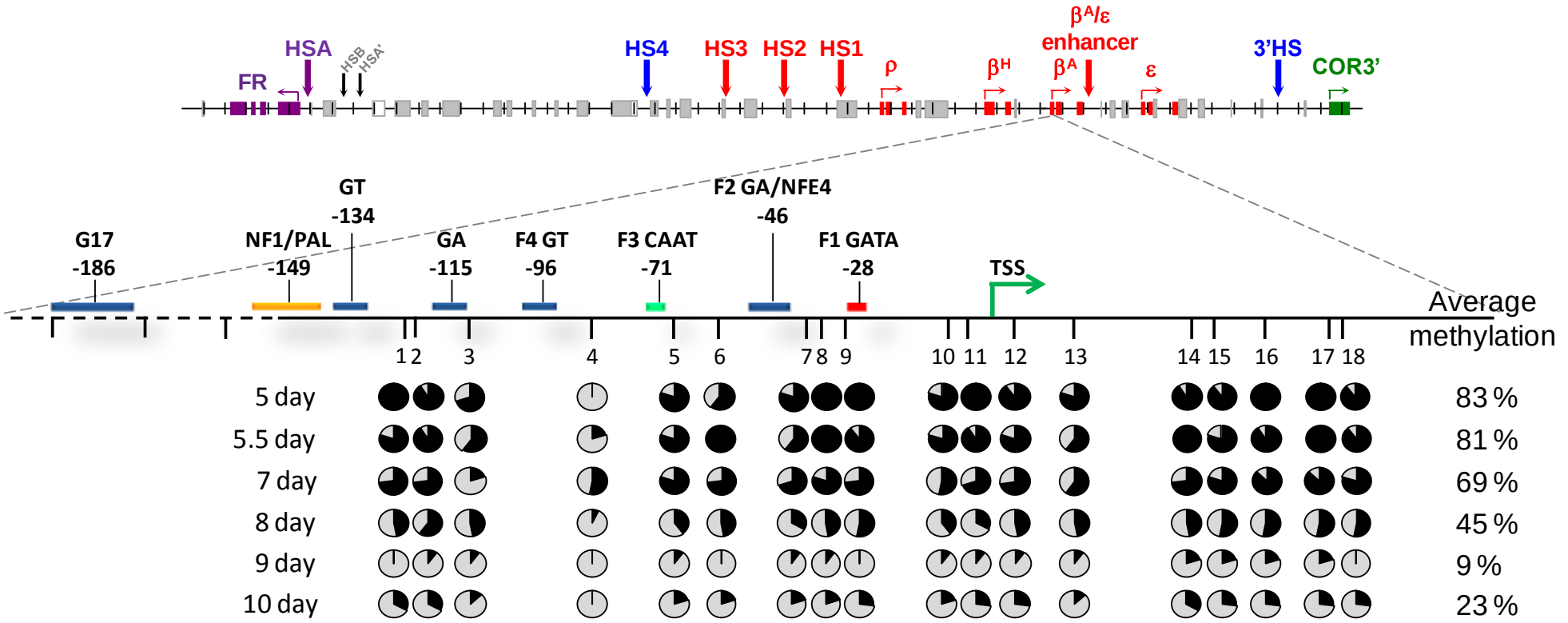


Figure 6.24 The DNA methylation status of the β^A gene promoter is dynamic throughout chick embryogenesis. Solid horizontal line denotes genomic region studied by bisulphite sequencing. Dashed horizontal line represents sequence outwith the amplicon studied by bisulphite sequencing. Vertical lines mark sites of CpG dinucleotides. Pie charts represent individual CpG dinucleotides at indicated time points during chick embryogenesis. Pie chart shading indicates the average DNA methylation status of each CpG dinucleotide; black represents methylated cytosine, grey represents unmethylated cytosine. Average over all DNA methylation of all sequenced clones at each time point is shown on the right of the figure. 10 or 15 clones were sequenced per time point. The coordinates of putative VEZF1 binding sites (blue boxes), NF1 (orange box), CTCF (green box) and GATA motifs (red box) are shown above. Promoter numbering is taken from (Emerson *et al*., 1985). Footprints F1 – F4 are taken from (Gallarda *et al*., 1989).

6.7 Discussion

The aims of this chapter were to profile VEZF1 expression levels in circulating erythrocytes during chicken embryonic development, to identify VEZF1 binding sites at gene regulatory elements across the chicken β-globin locus, and to study the relationship between VEZF1 binding, promoter DNA methylation and transcription of the ρ and β^A globin genes.

VEZF1 mRNA levels were found to be relatively unchanged in circulating chick erythrocytes between 5 and 10 days post-fertilisation by RT-PCR (Figure 6.3). It can therefore be concluded that any changes in VEZF1 enrichment at specific genomic elements across this timecourse do not arise due to changes in VEZF1 expression.

ChIP-seq libraries were generated using VEZF1 ChIP DNA from chick erythrocytes 5 and 10 days post-fertilisation, and used for NGS on the Illumina GAIIX platform. During Bowtie alignment of ChIP-seq reads, it became apparent that the 5 and 10 day VEZF1 ChIP-seq libraries had high levels of sequence clonality, which resulted in 80 % and 84 % of reads respectively being discarded from each sequencing run. This clonality is believed to have resulted from too little ChIP DNA being used in the library preparation. Size selection during library preparation enriches for small ChIP DNA fragments in ChIP-seq libraries and I believe that the cross-linked chromatin isolated from erythrocytes was not sonicated to small enough average fragment sizes to retain a sufficient concentration of DNA fragments following size selection (Figure 6.4). In light of this, chromatin samples will be sonicated to a lower average fragment size in the future, in order to reduce clonality and increase unique sequence reads and sequencing depth. After the removal of clonal reads 6,438,091 unique sequence reads from the 5 day VEZF1 ChIP-seq sample aligned to the chicken reference genome and 5,725,147 aligned from the 10 day ChIP-seq sample. These sequencing depths are below the guidelines of the ENCODE consortium, which aim for ≥10 million unique aligned reads per ChIP-seq dataset for a point-source DNA-binding transcription factor (Landt *et al*., 2012). Therefore, weak VEZF1 enrichment signals may be lost amongst background signal, however enrichment peaks should still be apparent at more highly enriched genomic elements.

Peak finding, using the MACS programme followed by PeakSplitter, identified a total of 5,398 and 6,131 peaks from 5 and 10 day VEZF1 ChIP-seq samples respectively. The efficiency of VEZF1 peak finding was limited by the lack of input or IgG ChIP-seq reference samples, meaning that enrichment peaks were calculated using the background signal in the VEZF1 ChIP-seq datasets, which is less reliable. It is intended that this peak calling will be repeated once IgG ChIP-seq reference samples are available.

Plotting normalised VEZF1 ChIP-seq reads across the chicken reference genome revealed well defined VEZF1 enrichment peaks (Figure 6.8). Enrichment across the chicken β-globin locus in both 5 and 10 day erythrocytes revealed VEZF1 binding at the HS4 insulator and the HS2 enhancer element at both stages. The ρ and ε gene promoters were enriched by VEZF1 in 5 day erythrocytes, where these genes are expressed. The β^A promoter and $\beta^{A/\varepsilon}$ enhancer elements were enriched in 10 day erythrocytes, where the β^A gene is expressed. We identified putative VEZF1 binding motifs within each of the VEZF1-enriched elements across the β-globin locus.

At the ρ promoter, a GT motif (GT-55) was found to correlate with the site of maximal VEZF1 ChIP-seq read enrichment in 5 day erythrocytes. This GT motif is located 23 bp 5' of a GATA motif (GATA-32). Minie *et al* found mutation of either the GT-55 or GATA-32 motifs reduce ρ promoter activity by ≥10-fold, showing that these sequence motifs are essential for the regulation of ρ-globin expression (Minie *et al*., 1992). At the ε promoter an identical GT motif (GT-55) was found to correlate with the peak of VEZF1 enrichment in 5 day erythrocytes. This GT motif was also 23 bp 5' of a GATA motif (GATA-32). Mutation of the GT-55 and GATA-32 motifs at the ε promoter results in reduced promoter activity (Mason *et al*., 1996), showing that these motifs are essential for regulation of ε-globin expression.

VEZF1 is known to bind a homopolymeric G17 motif in the β^A promoter (Lewis et al., 1988). However, ChIP-seq revealed a broad VEZF1 enrichment peak across this promoter, within which three putative VEZF1 binding motifs (G17-186, GA-115 and GA-46) appeared to locate to three VEZF1 subpeaks (Figure 6.12). The G17-186 motif is the validated VEZF1 binding site identified by Lewis *et al* (Lewis *et al*., 1988). The homopolymeric G-string sequence of this motif correlates with those motifs characterised in chapter 5, by POSMO motif discovery and EMSA, with which VEZF1 consistently forms strong interactions. The

other two putative VEZF1 binding sites are GA motifs, which locate to a Stage Selector Element (SSE) described by Choi and Engel. The SSE mediates interaction between the β^A promoter and the $\beta^{A/\varepsilon}$ enhancer, regulating β^A gene expression in a developmental stage-specific manner (Choi and Engel, 1988). A GATA motif also lies within the SSE at 18 bp 3' of the GA-46 putative VEZF1 binding motif. The GA-46 motif is required for transcriptional activity of the β^A promoter in definitive embryonic chick erythrocytes (Jackson *et al*., 1989). Mutation of both the GA-46 and GATA motifs results in a ~10-fold reduction in β^A promoter-mediated transcription of a transgene in HD3 cells. Deletion of the upstream promoter beyond -112 bp (i.e. upstream of the SSE) reduces β^A promoter-mediated expression by only ~2-fold demonstrating that the SSE is the main element responsible for controlling β^A promoter activity and that the G17-186 VEZF1-binding motif is not critical for β^A promoter activity. Barton *et al* demonstrated that both the GATA and GA-46 motifs are essential for transcriptional expression of the β^A gene (Barton *et al*., 1993). Work published by Galladra *et al* (Gallarda *et al*., 1989) implicates a 65 kDa protein present in the cell extract of definitive erythrocytes as binding this GA-46 motif and referred to this protein as NF-E4. The GA-46 motif is of the sequence GAGGAGGGG, which is very similar to the VEZF1 consensus binding motif (GGGGNGGGG), we therefore propose that this previously defined NF-E4 binding motif is actually a VEZF1 binding site within the β^A promoter.

A single putative VEZF1 binding motif was identified at the $\beta^{A/\varepsilon}$ enhancer. This motif lies 16 bp 5' of a GATA motif and maps to the point of maximal VEZF1 enrichment in 10 day erythrocytes (Figure 6.13). The binding of VEZF1 to the $\beta^{A/\varepsilon}$ enhancer appeared to be very low in the ChIP-seq analysis from 5 day erythrocytes (Figure 6.8). However, a strong peak of VEZF1 enrichment is seen at this element in the 5 day ChIP-seq data prior to the removal of identical reads (PCR clonality, not shown). Furthermore, VEZF1 ChIP-qPCR experiments found comparable levels of VEZF1 binding at the $\beta^{A/\varepsilon}$ enhancer in both 5 and 10 day erythrocytes (Figure 6.1). Repetition of VEZF1 ChIP-seq in 5 day and 10 day erythrocytes, to achieve lower read clonality and higher sequencing depth would reveal whether the $\beta^{A/\varepsilon}$ enhancer is enriched by VEZF1 at both timepoints. As this enhancer is active at both 5 and 10 days to mediate expression of the ε and β^A genes respectively, I expect this element to be enriched by VEZF1 at both timepoints.

Two putative VEZF1 motifs were identified within the HS2 enhancer (F2/GTC-4391 and F5/GTC-4311), which map to two VEZF1 enrichment summits in 10 day erythrocytes. In the 5 day ChIP-seq dataset, VEZF1 appears to interact with the F2/GTC-4391 motif only. Both putative VEZF1 sites lie within 18 – 24 bp of a GATA motif.

Most of the putative VEFZ1 binding sites which map to ChIP-seq enrichment peaks across the β-globin locus are GT or GA box motifs. POSMO motif discovery and EMSA experiments have already shown VEZF1 to form relatively weak interactions with these types of motifs when isolated *in vitro* (chapter 5). This was confirmed by a set of EMSA experiments, which used three putative VEZF1 binding motifs in the ρ-globin promoter as probe DNA and tested for complex formation with VEZF1 protein (6.15). VEZF1 showed the highest relative DNA-binding affinity for the homopolymeric G8-261 motif (GGGGGGGG) followed by the homopolymeric G7-106 motif (GGGGGGG). VEZF1 interacted weakly with the GT-55 motif (GGGGTGGGG) *in vitro*. As the GT-55 sequence is found at the VEZF1 ChIP-seq enrichment peaks at both the ρ and ε gene promoters, the inability of VEZF1 to interact with this sequence when isolated *in vitro* suggests that VEZF1 requires cooperativity with co-binding factors when interacting with this element *in vivo*. It has been shown in the published literature that the putative VEZF1 binding site GA-46 and the GATA-28 sequence motifs are both required for β^A promoter activity (Jackson *et al*., 1989). As all of the VEZF1-enriched elements across the β-*globin* locus contain GATA motifs, most of which are within 16 – 24 bp from a putative VEZF1 binding motif, and as the GATA-1, -2 and -3 proteins are known to function in the regulation of hematopoiesis, we suggest a role for the binding of GATA factors to these motifs in facilitating VEZF1 binding to these erythroid-specific elements. This hypothesis is supported by unpublished GATA-1 ChIP-qPCR data produced by Dr Ruslan Strogantsev, which shows the ρ, ε and β^A gene promoters to be enriched by GATA-1 in embryonic chick erythrocytes at time points during when these genes are actively expressed, and the $\beta^{A/\varepsilon}$ enhancer to be constitutively enriched throughout this timecourse i.e. GATA-1 enrichment of these genomic elements correlates with VEZF1 enrichment (data not shown).

As VEZF1 binding sites have been shown to protect a transgene and a CGI from DNA methylation (Dickson *et al*., 2010) analysis of the methylation status of individual cytosine residues across the β^A and ρ promoters by bisulphite sequencing was performed. These

analyses showed that the DNA methylation status of the β^A and ρ promoters change over the course of embryonic development (figures 6.22 and 6.24). A correlation between DNA methylation status, gene expression and VEZF1-enrichment of promoter elements became apparent at this point, i.e. when genes are transcriptionally inactive their promoters are not enriched by VEZF1 and are heavily methylated. Conversely, when genes are actively expressed their promoters are enriched with VEZF1 and are unmethylated or lightly methylated. When these findings are considered alongside those of Dickson *et al* (Dickson *et al*., 2010), a model whereby VEZF1 association with DNA elements functions to protect those elements from *de novo* DNA methylation may be proposed.

Chapter 7

Summary and Conclusions

7.1 Summary of work presented in this thesis

Identifying transcription factor binding sites is essential to the identification of regulatory elements and the understanding of diverse and complex regulatory networks which operate in eukaryotic organisms. Work detailed in this thesis sought to reveal the regulatory elements with which the highly conserved transcription factor VEZF1 associates. A small number of VEZF1 binding sites have been reported in the published literature where they have been shown to function in the regulation of gene expression (Aitsebaomo *et al*., 2001) and in the protection of DNA from *de novo* methylation (Dickson *et al*., 2010). VEZF1 is also known to have essential functions in vasculogenesis and lymphogenesis, with VEZF1 null mice exhibiting an embryonic lethal phenotype (Kuhnert *et al*., 2005, Zou *et al*., 2010). A pilot project in my supervisor's lab previously profiled VEZF1 binding sites across the ENCODE regions, which cover 1% of the human genome, in the K562 cell line by ChIP-chip. This study found the majority of VEZF1 binding sites to map to promoters and other putative gene regulatory elements (Strogantsev, 2009).

The principal aims of this thesis were to identify and validate putative VEZF1 binding sites across the human genome and to determine the relative affinities with which VEZF1 binds to a range of G-rich motifs. Sites of VEZF1 binding across the K562 genome were identified by ChIP-seq and used to generate VEZF1 consensus binding motifs. Motif discovery was also performed within groups of VEZF1 peaks ranked by ChIP-seq read enrichment in order to determine whether VEZF1 binding motifs differ in correlation with levels of VEZF1 enrichment. EMSA was employed to validate the relative affinity of VEZF1 interactions with putative VEZF1 binding motifs *in vitro*. Finally, ChIP-seq was used to identify VEZF1 binding sites across the chicken genome in an *in vivo* developmental system where the relationship between VEZF1 binding, gene expression and DNA methylation was investigated.

7.2 Identification and characterisation of VEZF1 binding sites (Chapters 3 and 4)

In chapter 3, ChIP-seq was performed using an anti-VEZF1 antibody in order to map VEZF1 binding sites across the human genome in K562 cells. In total, 26,341 VEZF1 binding sites were identified. The VEZF1 ChIP-seq dataset contained 21,239,966 uniquely aligned reads, this is more than double the minimum read depth set by the ENCODE consortium for a point-source transcription factor, we are therefore confident that regions of both high and low levels of VEZF1 enrichment should be evident in this dataset.

Analyses presented in chapter 4 show the majority of VEZF1 binding sites to be associated with regulatory elements, as 55 % of VEZF1 binding sites associate with gene promoter elements and 33 % associate with enhancers. Within promoters and enhancers, VEZF1 binding sites map to nucleosome depleted regions, which are essential for the correct regulation of gene expression (Bai *et al*., 2010). The chromatin features surrounding these nucleosome depleted regions indicate that VEZF1-enriched promoter and enhancer elements are active in K562 cells. In support of this, VEZF1 enrichment levels were found to positively correlate with transcript levels from associated genes and with levels of RNA pol II and active chromatin marks. These findings are consistent with a role for VEZF1 binding in facilitating transcription activity. Previous investigations into the effect of VEZF1 binding sites on transcription activity have generated varying results. The G17 VEZF1 binding motif within the β^A promoter was found to be dispensable for promoter activity, while the stage selector element is essential for promoter activity (Jackson *et al*., 1989, Barton *et al*., 1993). However, we have identified a putative VEZF1 binding site within the stage selector element, which correlates with a VEZF1 ChIP-seq enrichment peak. Therefore, this putative VEZF1 binding site may function in the regulation of β^A promoter activity. Mutation of a CT/GC-rich region of the human IL3 promoter, to which VEZF1 binds, reduces the basal transcription activity of the IL3 promoter, while deletion of a 53 bp region of the EDN1 promoter, to which VEZF1 binds, abolishes EDN1 transcriptional activity (Aitsebaomo *et al*., 2001, Koyano-Nakagawa *et al*., 1994). In order to determine the effects of VEZF1 binding on target gene expression, VEZF1 protein levels should be knocked down in the K562 cell line and the effect on target gene transcription measured by RT-QPCR. This VEZF1 knock down may be achieved using lentiviral systems

or transcription activator-like effector nucleases or repressors (TALENs and TALE repressors respectively).

VEZF1-enriched promoter and enhancer elements were found to be bound by numerous additional protein factors. Most were general transcription factors that are known to associate with active promoter or enhancer elements. The erythroid-specific transcription factors GATA1, TAL1, E2A, LMO2 and LDB1 are known to associate to form a TAL1-erythroid complex which functions to regulate expression of a subset of erythroid-specific genes (Wadman *et al*., 1997, Lahlil *et al*., 2004, Xu *et al*., 2007, Dhami *et al*., 2010). A group of 1,586 VEZF1-associated enhancer elements were found to be bound by erythroid-specific transcription factors, possibly as a TAL1-erythroid complex. Consistent with this finding, the erythroid-specific *TAL1* +51 and *HBB* gene enhancers are found in this group.

Only 249 VEZF1 ChIP-seq peaks mapped to insulator elements by comparison to the K562 ChromHMM chromatin state map. This was much lower than expected as it was predicted that profiling VEZF1 enrichment across the genome would identify novel insulator elements, due to VEZF1 binding sites being essential for barrier activity of the chicken HS4 insulator (Dickson *et al*., 2010). When analysed by ChromHMM, the chicken HS4 insulator is assigned a promoter state. As we know this element to have no transcriptional activity, I believe that the 14,619 VEZF1 ChIP-seq peaks reported to locate to promoter elements by comparison to the ChromHMM map contain a proportion of mis-identified insulator elements. In order to investigate this further, VEZF1-enriched ChromHMM-defined promoters that are distal from annotated TSSs should be identified. The locations of these elements could be compared to the GENCODE human reference genome annotation (Harrow *et al*., 2012) to filter out un-annotated gene promoters and promoters of non-coding genes. Following this filtering step, we would be left with a list of VEZF1-enriched potential insulator elements. A proportion of the VEZF1-associated enhancer elements, which were identified by comparison to the ChromHMM map, appeared to be TSSs as judged by the high levels of RNA pol II extending from the VEZF1 peak summits (Section 4.4.2, Figure 4.13). It is therefore apparent that, while the ChromHMM chromatin state maps are an excellent tool for analyzing the distribution of ChIP-seq enrichment peaks amongst regulatory elements on a genome-wide scale, there

appears to be some accuracy issues in the definition of some insulator and enhancer elements.

During the course of this project Gowher *et al* published results from VEZF1 ChIP-seq in the human cervical cancer HeLa cell line (Gowher *et al* 2012). In this paper VEZF1 ChIP-seq data was used to support a role for VEZF1 in regulating splicing events by mediating pausing of the elongating RNA pol II, however the function of VEZF1 binding DNA was not investigated beyond this. When the publicly available files containing sequence reads from ChIP-seq in HeLa cells are viewed in the UCSC genome browser, the negative control input track shows substantial read enrichment at the same sites that appear enriched in the VEZF1 ChIP-seq track. This indicates that VEZF1 ChIP did not specifically enrich for DNA elements that are bound by VEZF1 in HeLa cells. If facilitating RNA pol II pausing was a major function of VEZF1 binding, we would expect a large proportion of the VEZF1 ChIP-seq enrichment peaks in K562 to locate to regions of transcription elongation by comparison to the ChromHMM chromatin state map, however only 85 out of 26,341 VEZF1 peaks do (section 4.3.2, figure 4.6). When sites reported as VEZF1-mediated RNA pol II pause sites in HeLa by Gowher *et al* were investigated using publicly available ChIP-seq tracks in the UCSC genome browser, I observed them to map to the 5' ends of genes and to have the characteristics of promoters as they were enriched by RNA pol II, H3K4me3, H2A.Z and H3K27ac (data not shown). No other RNA pol II peaks were present in the vicinity of these genes, indicating to us that regions observed to be simultaneously enriched by VEZF1 and RNA pol II by Gowher *et al* actually mark active promoter elements rather than RNA pol II pause sites.

7.3 Definition of VEZF1 binding motifs and VEZF1 binding site validation (Chapter 5)

In chapter 5, a VEZF1 consensus binding motif of GGGGNGGGG was identified from VEZF1 ChIP-seq binding sites. This motif is consistent with validated chicken VEZF1 binding sites in the β^A promoter (Lewis *et al*., 1988) and the HS4 insulator FI (Dickson *et al*., 2010), and with a motif generated using VEZF1 ChIP-chip binding sites from a previous study in our lab (Strogantsev, 2009). Further analyses showed the most highly enriched VEZF1 sites to be homopolymeric 9(dG) motifs while decreasing VEZF1 enrichment levels were found to correlate with VEZF1 binding to degenerate motifs where G-strings were replaced by

GGGGNGGGG sequences. Likewise, motif searching using promoter-associated VEZF1 peaks identified a high specificity 9(dG) binding site, while at enhancers a divergent gGGGWGGGg site was discovered.

EMSA assays confirmed these findings, as VEZF1 was found to have the highest relative DNA-binding affinity for homopolymeric G-strings *in vitro*. Meanwhile, VEZF1 binding to GA and GT motifs was more variable, and VEZF1 was found to interact weakly with GC motifs *in vitro*. VEZF1 was unable to interact with a number of putative binding sites when isolated *in vitro*. These sites included GA and GT motifs found at the centre of VEZF1 ChIP-seq peaks at erythroid-specific gene associated elements including the *TAL1* +51 and *HBB* enhancer elements. These results indicate a potential role for co-binding factors in mediating VEZF1 binding to low-affinity, non-homopolymeric dG motifs *in vivo*. As these sites mostly locate to erythroid-specific regulatory elements, co-binding proteins such as the TAL1-erythroid complex factors identified in chapter 4 may be involved in the regulation of VEZF1 binding to divergent sites at erythroid-specific regulatory elements. In support of this hypothesis, a GATA factor consensus binding motif was identified 38 bp 3' of the putative VEZF1 binding motif at the *TAL1* +51 enhancer. GATA binding motifs were also discovered 44 bp 3' of the putative VEZF1 site at the *HBB* HS2 enhancer, and 17 bp 3' and 19 bp 5' of the putative VEZF1 site at the *HBB* HS3 enhancer element. As each of these GATA motifs are situated in close proximity to putative VEZF1 binding sites, cooperative interactions between GATA factors and VEZF1 at erythroid-specific regulatory elements are a feasible possibility. In order to investigate this further, GATA1 protein expression could be knocked down in the K562 cell line and VEZF1 ChIP-seq performed to determine whether VEZF1 enrichment of erythroid-specific regulatory elements is lost in the absence of GATA1.

7.4 The relationship between VEZF1, promoter DNA methylation and transcription of the chicken β-globin genes (Chapter 6)

In chapter 6, ChIP-seq was performed using an anti-VEZF1 antibody to profile VEZF1 binding sites across the chicken genome in circulating erythrocytes isolated from chick embryos 5 and 10 days post-fertilisation. In total 5,398 and 6,131 VEZF1 peaks were identified in the 5 and 10 day datasets respectively. The 5 day sample contained 6,438,091 unique aligned reads and the 10 day sample contained 5,725,147. As these are

below the minimum read depth of 10,000,000 unique aligned reads set by the ENCODE consortium for a point-source transcription factor (Landt *et al*., 2012), weak VEZF1 enrichment signals may be lost amongst background signal however enrichment peaks should still be apparent at more highly enriched elements.

The developmentally regulated chicken β-globin locus shows stage-specific patterns of VEZF1 enrichment at 5 and 10 day timepoints. The HS4 insulator and HS2 enhancer elements are enriched at both 5 and 10 days. In the 5 day sample the promoters of the ρ- and ε-globin genes are enriched, correlating with active expression of these genes. In the 10 day sample the β^A promoter is highly enriched correlating with the active expression of this gene, and the $\beta^{A/\varepsilon}$ enhancer is enriched to a lower extent. Putative VEZF1 binding motifs were identified at VEZF1 enrichment peaks across the β-globin locus, the majority were GA or GT motif sites. Despite these motifs being enriched by VEZF1 *in vivo*, as demonstrated by ChIP-seq, VEZF1 formed weak interactions with these elements in *in vitro* EMSA binding reactions. As β-globin gene expression is erythroid cell-specific, these results provide support for VEZF1 having low affinity for cell type-specific VEZF1 sites *in vitro* and requiring co-binding factors to facilitate interaction with these sites *in vivo*. Consensus GATA motifs were identified within all VEZF1-enriched elements across the β-globin locus, mostly at a distance of 16 – 24 bp from putative VEZF1 binding sites. This spacing of GATA and VEZF1 binding motifs should allow cooperative interactions to occur. We know the ρ, ε and β^A promoters are enriched by GATA1 in circulating erythrocytes at the developmental timepoints when these genes are actively expressed and associated with VEZF1 (Strogantsev *et al* unpublished). These findings support a role for erythroid-specific TFs in mediating VEZF1 binding to erythroid-specific genomic elements.

Consistent with the proposed role for VEZF1 in mediating protection from *de novo* DNA methylation, VEZF1 binding at the ρ and β^A gene promoters was found to correlate with an unmethylated promoter state and active gene expression. At the ρ-globin promoter, loss of VEZF1 enrichment correlated with an accumulation of DNA methylation. Meanwhile, at the βA promoter, VEZF1 binding correlated with demethylation of the promoter element. Given the findings of Dickson *et al*, which showed VEZF1 binding sites to protect a transgene and a CGI from *de novo* DNA methylation and to direct active demethylation (Dickson *et al*., 2010), a model can be conceived whereby VEZF1

association with its cognate binding sites mediates both protection from DNA methylation and active demethylation.

7.5 Conclusions

Prior to this study, VEZF1 had been found to bind at a small number of cell type specific promoters and an insulator element but its role in gene regulation was unclear. In this thesis I present evidence that VEZF1 is broadly associated with the activation of gene transcription. Of the more than 26,000 identified in this thesis 88 % map to promoter and enhancer elements. VEZF1 binding positively correlates with RNA pol II enrichment, gene transcription levels and with a non-methylated DNA state. The discovery of VEZF1 binding at enhancer elements is novel and has not been reported to date in the published literature.

VEZF1 binds a range of G-rich motifs with varying relative affinities. VEZF1 has high affinity for binding to homopolymeric stretches of ≥9 dG residues *in vitro* but shows variable or low affinity for binding to GA, GT or GC motifs. This implicates co-binding factors in facilitating VEZF1 binding to these lower affinity sites *in vivo*. Evidence to support this hypothesis comes from my finding that a number of VEZF1-enriched erythroid-specific enhancer elements are also bound by erythroid-specific transcription factors in the human genome, and from a previous finding in our lab that GATA1 is bound to the chicken ρ, ε and β^A promoters at the same developmental timepoints as VEZF1. I therefore propose a model whereby VEZF1 is recruited to low affinity binding sites *in vivo* via protein-protein interactions with co-binding cell type-specific transcription factors. Candidate co-binding transcription factors at erythroid enhancers bound by VEZF1 are GATA1, GATA2 and TAL1. Given that VEZF1 is expressed at similar levels in most somatic cell types, we can hypothesise that other cell-type-specific transcription factors may mediate VEZF1 binding to low affinity sites at enhancer elements in other cell types.

I present evidence that VEZF1 binding at gene regulatory elements occurs within nucleosome depleted regions flanked by positioned nucleosomes enriched in the variant histone H2A.Z. This nucleosome profile is observed at both promoter and enhancer elements. We hypothesise that VEZF1 has general role in establishing the specialised chromatin structure at regulatory elements. The underlying DNA sequence determines whether VEZF1 acts in a constitutive or cell type-specific manner at any given element. The exact nature of VEZF1 action remains to be determined but may relate to events such

as H2A.Z recruitment, nucleosome positioning, establishment of nucleosome depleted regions or recruitment of RNA pol II.

APPENDIX I: Primer Sequences

A. ChIP-qPCR human primer sequences

Abbreviations: P – promoter; dup – duplicated region; F – forward primer; R – reverse primer; T – TaqMan® probe.

P-IL3_F:	GGTTGTGGGCACCTTGCT
P-IL3_R:	TCTGTCTTGTTCTGGTCCTTCGT
P-IL3_T:	FAM-ACATATAAGGCGGGAGGTTGTTGCCAA-TAMRA
P-EDN1_F:	TGCCCCCGAATTGTCAGA
P-EDN1_R:	CAGGCCCGAAAGGAAATCA
P-EDN1_T:	FAM-CGGGCGTCTGCCTCTGAAGTTAGCA-TAMRA
FOXP4 intron 3_F:	GAGATGTTTCCGGCGTGTGT
FOXP4 intron 3_R:	TCTGAGTGGTGCGTCTGCTAA
FOXP4 intron 3_T:	FAM-ACGGGTATTGAGATTTCCACGGG-TAMRA
FLNA 3′ dup_F:	CCTTGTGTTTTGGGTGTGG
FLNA 3′ dup_R:	CACGACCTCTGGACGTTTCT
FLNA 3′ dup_T:	FAM-TGCGCTTCTCTAAGCGTTCCATTCC-TAMRA
FLNA intron 2_F:	CCCAAAAGAGAGAGGGAAGG
FLNA intron 2_R:	ATGTTGGAAAGCAGCAGTGA
FLNA intron 2_T:	FAM-CCCAGGGCTGGGAGACTGTCTG-TAMRA
P-STAG2_F:	CGATCTCTCCATCCCTTCC
P-STAG2_R:	CCAGACCCCAGCCAATTTAG
P-STAG2_T:	FAM-TCAGTTTTCTTCGGGCAACAATTT-TAMRA
P-POLR3K_F:	GGGTGATGTTGTGCACGTAG
P-POLR3K_R:	ATCGTGGAGGAGGGACAAC
P-POLR3K_T:	FAM-TGTTGCAGGAGAAGCGGTGGC-TAMRA
P-HISPPD2A_F:	CCCTCCTCGGTACTCTCCTC
P-HISPPD2A_R:	GGGCCTTAGTGTCAGCGTAG
P-HISPPD2A_T:	FAM-CTCGCATCCCGACTCCACTAGCCTT-TAMRA
TAL1+3_F:	AGGAAAGGCTCCAAACACCT
TAL1+3_R:	ATGGCTGGGAATTACCTCCT
TAL1+3_T:	FAM-CGATTCCCTGGACTGGTTGGTCG-TAMRA
TAL1+51_F:	TTACAGCCCTTCACCCTCAC
TAL1+51_R:	TGGGAATGAGCGATAAGGAT
TAL1+51_T:	FAM-ATGTTCCTGCCCTGATCCAGAGGG-TAMRA
P-RFX5_F:	CTGGGCCCAAGTTCTCATTA
P-RFX5_R:	CCCTACGTCATCTCCCACAA
P-RFX5_T:	FAM-TTAATCTCGCCACGACTTCCCCC-TAMRA

B. ChIP-qPCR chicken β-globin locus primer sequences

15.9 (16 kb cond.)_F:	CAGCAGACGCTGTGGTGAA
15.9 (16 kb cond.)_R:	CTTGCAGGATGCAGACTGGA
15.9 (16 kb cond.)_T:	FAM-ATCCCATCGGTGCCACCCTGAG-TAMRA
21.7 (HS4)_F:	CGGGATCGCTTTCCTCTGA
21.7 (HS4)_R:	CCGTATCCCCCCAGGTGTCT
21.7 (HS4)_T:	FAM-CGCTTCTCGCTGCTCTTTGAGCCTG-TAMRA

31.9 (P-β-Rho globin)_F: CAGAGGAGCCAACATTTGGG
31.9 (P-β-Rho globin)_R: CCCCTCTGGGTGATGCATT
31.9 (P-β-Rho globin)_T: FAM–CGCTGCAGGCGTGAAGCCATT–TAMRA

39.8 (P-β^A globin)_F: CTGTGGTCTCCTGCCTCACA
39.8 (P-β^A globin)_R: AGGCTGGGTGCCCCTC
39.8 (P-β^A globin)_T: FAM–CAATGCAGAGTGCTGTGGTTTGGAACTG–TAMRA

C. RT-qPCR chicken gene expression primer sequences

β-Actin_F: TGCTGCGCTCGTTGTTGA
β-Actin_R: TCGCCGGGGACGATG
β-Actin_T: TGGCTCCGGTATGTGCAAGGCC

β-Rho_F: CAGCGTGGTGGCCCAT
β-Rho_R: TGACTTTCACACTGTGTCCTGCT
β-Rho_T: FAM – CCTGGCCTACAAGTACCACTGAGCTC – TAMRA

β-Epsilon_F: GGCAGAAGCTGGTCAACGTT
β-Epsilon_R: CCACGGCTGTGCTGCAG
β-Epsilon_T: FAM – TGCTCTGGCCCGCAAGTACCACTG – TAMRA

β-Adult_F: CGCGTGGTGGCCCAT
β-Adult_R: AGGTGCTCCGTGATCTTTGGT
β-Adult_T: FAM – CCCTGGCTCGCAAGTACCACTAAGCA – TAMRA

β-Hatching_F: AGAAGATGGTGCGTGTGGTG
β-Hatching_R: CAGGAGCATCTCCAAGTGGCT
β-Hatching_T: FAM – CCCACGAGTACCACTGAGCCCCA – TAMRA

VEZF1_F: AAAGGATCGCATGACCTACC
VEZF1_R: AATGATCAGGCCTCGAGAAG

D. Bisulfite Sequencing primer sequences

P-β-Rho_F: GAATGTTGTAGAGGAGTTAATATTTGG
P-β-Rho_R: TCAATACCCAAAATACAACTAAACC

P-β-Adult_F: TTTGTTTTGAGTTTTATTTTGATGT
P-β-Adult_R: ACTTCTCCTCAACAATCCAATACAC

M13_F: GTTTTCCCAGTCACGAC
M13_R: CAGGAAACAGCTATGAC

APPENDIX II: EMSA oligonucleotide sequences

A. Chicken sequences

Locus

Beta Globin

HS4_FI_T	GGAGCTCACGGGGACAGCCCCCCCCCAAAGCCCCCAGGGA
HS4_FI_B	TCCCTGGGGGCTTTGGGGGGGGGCTGTCCCCGTGAGCTCC
HS4_FIaaa1_T	GGAGCTCACGGGGACAGCAAACCCCCAAAGCCCCCAGGGA
HS4_FIaaa1_B	TCCCTGGGGGCTTTGGGGGTTTGCTGTCCCCGTGAGCTCC
HS4_FIII_T	AGGCGCGCCCCGGTCCGGCGCTCCCCCCGCATCCCCGAGC CGGGGCGCGCCT
HS4_FIII_B	AGGCGCGCCCCGGCTCGGGGATGCGGGGGGAGCGCCGGAC CGGGGCGCGCCT

B. Human sequences

Locus

Alpha Globin

HBA_HS40_T	ATCCTGTGGGGGTGGAGGTGGGACAAGGGA
HBA_HS40_B	TCCCTTGTCCCACCTCCACCCCCACAGGAT
HBA_HS14_T	GCCGTGGGGCCGGGGCCGGGGGCGGAGGGGGCCAGA
HBA_HS14_B	TCTGGCCCCCTCCGCCCCCGGCCCCGGCCCCACGGC
HBA_P-40_T	CCGGGCGTGCCCCCGCGCCCCAAGCATAAA
HBA_P-40_B	TTTATGCTTGGGGCGCGGGGGCACGCCCGG
HBA_P-177_T	GACGTCCTGGCCCCCGCCCCGCGTGCACCC
HBA_P-177_B	GGGTGCACGCGGGGCGGGGGCCAGGACGTC
HBA_P-380_T	GTGCCAGGCCGGGGCGGGGGTGCGGGCTGA
HBA_P-380_B	TCAGCCCGCACCCCCGCCCCGGCCTGGCAC
HBZ_Pro_T	ACCCCTGCAGCCCCCTCCCCTCACCTGACC
HBZ_Pro_B	GGTCAGGTGAGGGGAGGGGGCTGCAGGGGT

Beta Globin

HBB_HS3GT_T	CAGGGAGGGTGGGGTGGGGTCAGGGCTGGC
HBB_HS3GT_B	GCCAGCCCTGACCCCACCCCACCCTCCCTG
HBB_HS2_T	GGGGTCAGTGCCCCACCCCCGCCTTCTGGT
HBB_HS2_B	ACCAGAAGGCGGGGGTGGGGCACTGACCCC
HBG_Pro_T	CCTCTTGGGGGCCCCTTCCCCACACTATCT
HBG_Pro_B	AGATAGTGTGGGGAAGGGGCCCCCAAGAGG
HBB_Pro_T	TCCCAGGAGCAGGGAGGGCAGGAGCCAGGG
HBB_Pro_B	CCCTGGCTCCTGCCCTCCCTGCTCCTGGGA

TAL1

TAL1_P1bA_T	TTCGATGGCCGGGGGGGGCGGTGGGGGGGCATTTTCCACG
TAL1_P1bA_B	CGTGGAAAATGCCCCCCCACCGCCCCCCCCGGCCATCGAA
TAL1_P1bB_T	TCCACGGACGCCCCCGCCCCGGCTGCCGCC
TAL1_P1bB_B	GGCGGCAGCCGGGGCGGGGGCGTCCGTGGA

TAL1_+3_T	AAGGCGGGTGGGGGGAGGAGGGGGTAGAGG
TAL1_+3_B	CCTCTACCCCCTCCTCCCCCCACCCGCCTT
TAL1_+20_T	TCTCCACTCCTCCCCTCCCCTTTGGACCAG
TAL1_+20_B	CTGGTCCAAAGGGGAGGGGAGGAGTGGAGA
TAL1_+51_T	TCCCAGGGCCTGGGGAGGGGGAGCCTCTGG
TAL1_+51_B	CCAGAGGCTCCCCCTCCCCAGGCCCTGGGA

FLNA

FLNA_IRA_T	TCCCTCCACGCCCCGCCCCCGCCTCGGCAC
FLNA_IRA_B	GTGCCGAGGCGGGGGCGGGGCGTGGAGGGA
FLNA_IRB_T	GCGCGTCTGGGGGGTCGTGGGGAAGCAGGG
FLNA_IRB_B	CCCTGCTTCCCCACGACCCCCCAGACGCGC
FLNA_IRC_T	CAGCGCCACCACCCACCCCCCCGGGAGCCG
FLNA_IRC_B	CGGCTCCCGGGGGGGTGGGTGGTGGCGCTG
FLNA_int_T	GCTGGATGGCCCCGCCCCATCCCACCCCCC
FLNA_int_T	GGGGGGTGGGATGGGGCGGGGCCATCCAGC

EHD1

EHD1_bound16A_T	TGGGGTAATGGGGGGCGGGGCGGGGGGCCAGGCAGG
EHD1_bound16A_B	CCTGCCTGGCCCCCCGCCCCGCCCCCCATTACCCCA
EHD1_bound17A_T	GGATATCATCCCCTCCTCCCCCCAGACACA
EHD1_bound17A_B	TGTGTCTGGGGGGAGGAGGGGATGATATCC
EHD1_bound17B_T	GGCCCAGGTCGGGGGAGGGGCAGTGTCCTT
EHD1_bound17B_B	AAGGACACTGCCCCTCCCCCGACCTGGGCC

CKMT

CKMT_boundA_T	CTGGACTCCAGCCCCACCCCTCCTGGCGCA
CKMT_boundA_B	TGCGCCAGGAGGGGTGGGGCTGGAGTCCAG
CKMT_boundB_T	CCGATACCGCCCCCAGCCCCAGCCACTCCC
CKMT_boundB_B	GGGAGTGGCTGGGGCTGGGGGCGGTATCGG
CKMT_boundC_T	AGCCACTCCCCCACCGCCCCCGGCCTTCAC
CKMT_boundC_B	GTGAAGGCCGGGGGCGGTGGGGGAGTGGCT

IL3-GMCSF

IL3_proP_T	GCCTGCCCCACACCACCACCTCCCCCCGCCTTGCCCGGGG
IL3_proP_B	CCCCGGGCAAGGCGGGGGGAGGTGGTGGTGTGGGGCAGGC
IL3_proD_T	AACCTCCCAGGCCAGCCCCTCCCCCAGCTCCCAGTGACAG
IL3_proD_B	CTGTCACTGGGAGCTGGGGGAGGGGCTGGCCTGGGAGGTT
IL3_-37_T	AGGTCTTTCTCAAACAACCTCCCCCACAAAATCATATTGA
IL3_-37_B	TCAATATGATTTTGTGGGGGAGGTTGTTTGAGAAAGACCT
GMCSF+30A_T	CCCTGCCTGCCTCTAGGGGTGGGGTAGGTGAGGTAGGAGT
GMCSF+30A_B	ACTCCTACCTCACCTACCCCACCCCTAGAGGCAGGCAGGG
GMCSF+30B_T	AGGAACCCCTTTCCCCACCCCCCCCACCCACTGCAGAAAC
GMCSF+30B_B	CTTTCTGACAGTGGGTGGGGGGGGTGGGGAAAGGGGTCCT
GMCSF_pro_T	TAAGTGTCTCCCACGCCCCACCCCAGCCATTCCAGGCCAG
GMCSF_pro_B	CTGGCCTGGAATGGCTGGGGTGGGGCGTGGGAGACACTTA

DNMT3B

DNMT3B_proA_T	TCGGCACCGCCCCCTCCCCACCCACTCCCG
DNMT3B_proA_B	CGGGAGTGGGTGGGGAGGGGGCGGTGCCGA
DNMT3B_proB_T	TCCCTCGTCCCCGCCCCCCCGTTCCCCCCGCGCCCGG
DNMT3B_proB_B	CCGGGCGCGGGGGGAACGGGGGGGCGGGGACGAGGGA

MECP2

MECP2_proA_T	CTGAGACCTCCCCCCTCCCCCGTCCTCCCC
MECP2_proA_B	GGGGAGGACGGGGGAGGGGGGAGGTCTCAG
MECP2_proB_T	ACCCTTGCCGGGGGGCGGGGGTCAGGGGCG
MECP2_proB_B	CGCCCCTGACCCCCGCCCCCCGGCAAGGGT

NRXN2

NRXN_boundA_T	GCCTTGAGTGTGGGGAGGGGTGCAGGCAGG
NRXN_boundA_B	CCTGCCTGCACCCCTCCCCACACTCAAGGC
NRXN_boundB_T	CTGCCCGCCCCATGGCCCCCCCAGACTGCC
NRXN_boundB_B	GGCAGTCTGGGGGGGCCATGGGGCGGGCAG

IGF2

IGF_5HSA_T	TGGAGCCGGGTGGGGCTGGGGAAGCGAGTT
IGF_5HSA_B	AACTCGCTTCCCCAGCCCCACCCGGCTCCA
IGF_5HSB_T	GGGTTAGCCCCCAGGCCCCTCCTGCCCATC
IGF_5HSB_B	GATGGGCAGGAGGGGCCTGGGGGCTAACCC

LYL1

LYL1_proA_T	CTCACTTTCCCTGGGGGGGACCTACGCGGA
LYL1_proA_B	TCCGCGTAGGTCCCCCCCAGGGAAAGTGAG
LYL1_proB_T	CCGGGCCGCCGGGGGCGGGGAGAAGGCGGG
LYL1_proB_B	CCCGCCTTCTCCCCGCCCCCGGCGGCCCGG

GMPPA-ACCN4

GMPPA_boundA_T	CATGGGGTGTGGGGGGCGGGGGCTCTGTCA
GMPPA_boundA_B	TGACAGAGCCCCCGCCCCCCACACCCCATG

POLR3K

POLR3K_pro_T	CTGCAGCCGCCCCACCCCCCCCGGCGCGCC
POLR3K_pro_B	GGCGCGCCGGGGGGGGTGGGGCGGCTGCAG

SPA1

SPA1_boundA_T	GACCAGGTTGGGGGAGGGGGTGAATGTCAC
SPA1_boundA_B	GTGACATTCACCCCCTCCCCCAACCTGGTC

HMGN1

HMGN_proA_T	GGCGCCCGGGGGGGGGGGGGGGCCCCCGG
HMGN_proA_B	CCGGGGGCCCCCCCCCCCCCCCCGGGCGCC

RFX5

RFX5_proA_T	CCGCTACCCCGCCCCACCCCCGCTCCACTT
RFX5_proA_B	AAGTGGAGCGGGGGTGGGGCGGGGTAGCGG

STAG2

STAG2_proA_T	CTACGGAGACCCCCCCCCCCCATGGGTGGC
STAG2_proA_B	GCCACCCATGGGGGGGGGGGGTCTCCGTAG

BRWD1

BRWD1_proA_T	CGCCCGCCCCCGGCCCCCCGCCCCCCGCCCCCCCCCGCGCCGGAG
BRWD1_proA_B	CTCCGGCGCGGGGGGGGCGGGGGGCGGGGGGCCGGGGGCGGGCG

EDN1

EDN1_proA_T	CGGAGCTGTTTACCCCCACTCTAATAGGGG
EDN1_proA_B	CCCCTATTAGAGTGGGGGTAAACAGCTCCG

References

ABBOTT, D. W., IVANOVA, V. S., WANG, X., BONNER, W. M. & AUSIÓ, J. 2001. Characterization of the stability and folding of H2A.Z chromatin particles: implications for transcriptional activation. *J Biol Chem,* 276**,** 41945-9.

ABRUZZO, L. V. & REITMAN, M. 1994. Enhancer activity of upstream hypersensitive site 2 of the chicken beta-globin cluster is mediated by GATA sites. *J Biol Chem,* 269**,** 32565-71.

ADAM, M., ROBERT, F., LAROCHELLE, M. & GAUDREAU, L. 2001. H2A.Z is required for global chromatin integrity and for recruitment of RNA polymerase II under specific conditions. *Mol Cell Biol,* 21**,** 6270-9.

AITSEBAOMO, J., KINGSLEY-KALLESEN, M. L., WU, Y., QUERTERMOUS, T. & PATTERSON, C. 2001. Vezf1/DB1 is an endothelial cell-specific transcription factor that regulates expression of the endothelin-1 promoter. *J Biol Chem,* 276**,** 39197-205.

AKOULITCHEV, S., MÄKELÄ, T. P., WEINBERG, R. A. & REINBERG, D. 1995. Requirement for TFIIH kinase activity in transcription by RNA polymerase II. *Nature,* 377**,** 557-60.

ANDERSSON, L. C., NILSSON, K. & GAHMBERG, C. G. 1979. K562--a human erythroleukemic cell line. *Int J Cancer,* 23**,** 143-7.

ANGUITA, E., JOHNSON, C. A., WOOD, W. G., TURNER, B. M. & HIGGS, D. R. 2001. Identification of a conserved erythroid specific domain of histone acetylation across the alpha-globin gene cluster. *Proc Natl Acad Sci U S A,* 98**,** 12114-9.

ATCHISON, L., GHIAS, A., WILKINSON, F., BONINI, N. & ATCHISON, M. L. 2003. Transcription factor YY1 functions as a PcG protein in vivo. *EMBO J,* 22**,** 1347-58.

AWAD, S. & HASSAN, A. H. 2008. The Swi2/Snf2 bromodomain is important for the full binding and remodeling activity of the SWI/SNF complex on H3- and H4-acetylated nucleosomes. *Ann N Y Acad Sci,* 1138**,** 366-75.

BAI, L., CHARVIN, G., SIGGIA, E. D. & CROSS, F. R. 2010. Nucleosome-depleted regions in cell-cycle-regulated promoters ensure reliable gene expression in every cell cycle. *Dev Cell,* 18**,** 544-55.

BAI, L., ONDRACKA, A. & CROSS, F. R. 2011. Multiple sequence-specific factors generate the nucleosome-depleted region on CLN2 promoter. *Mol Cell,* 42**,** 465-76.

BANNISTER, A. J. & KOUZARIDES, T. 2011. Regulation of chromatin by histone modifications. *Cell Res,* 21**,** 381-95.

BARSKI, A., CUDDAPAH, S., CUI, K., ROH, T. Y., SCHONES, D. E., WANG, Z., WEI, G., CHEPELEV, I. & ZHAO, K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell,* 129**,** 823-37.

BARTON, M. C., MADANI, N. & EMERSON, B. M. 1993. The erythroid protein cGATA-1 functions with a stage-specific factor to activate transcription of chromatin-assembled beta-globin genes. *Genes Dev,* 7**,** 1796-809.

BAUMANN, M., PONTILLER, J. & ERNST, W. 2010. Structure and basal transcription complex of RNA polymerase II core promoters in the mammalian genome: an overview. *Mol Biotechnol,* 45**,** 241-7.

BELL, A. C., WEST, A. G. & FELSENFELD, G. 1999. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell,* 98**,** 387-96.

BENGANI, H., MENDIRATTA, S., MAINI, J., VASANTHI, D., SULTANA, H., GHASEMI, M., AHLUWALIA, J., RAMACHANDRAN, S., MISHRA, R. K. & BRAHMACHARI, V. 2013. Identification and Validation of a Putative Polycomb Responsive Element in the Human Genome. *PLoS One,* 8**,** e67217.

BHUTANI, N., BURNS, D. M. & BLAU, H. M. 2011. DNA demethylation dynamics. *Cell,* 146**,** 866-72.

BIRNEY, E., STAMATOYANNOPOULOS, J. A., DUTTA, A., GUIGÓ, R., GINGERAS, T. R., MARGULIES, E. H., WENG, Z., SNYDER, M., DERMITZAKIS, E. T., THURMAN, R. E., KUEHN, M. S., TAYLOR, C. M., NEPH, S., KOCH, C. M., ASTHANA, S., MALHOTRA, A., ADZHUBEI, I., GREENBAUM, J. A., ANDREWS, R. M., FLICEK, P., BOYLE, P. J., CAO, H., CARTER, N. P., CLELLAND, G. K., DAVIS, S., DAY, N., DHAMI, P., DILLON, S. C., DORSCHNER, M. O., FIEGLER, H., GIRESI, P. G., GOLDY, J., HAWRYLYCZ, M., HAYDOCK, A., HUMBERT, R., JAMES, K. D., JOHNSON, B. E., JOHNSON, E. M., FRUM, T. T., ROSENZWEIG, E. R., KARNANI, N., LEE, K., LEFEBVRE, G. C., NAVAS, P. A., NERI, F., PARKER, S. C., SABO, P. J., SANDSTROM, R., SHAFER, A., VETRIE, D., WEAVER, M., WILCOX, S., YU, M., COLLINS, F. S., DEKKER, J., LIEB, J. D., TULLIUS, T. D., CRAWFORD, G. E., SUNYAEV, S., NOBLE, W. S., DUNHAM, I., DENOEUD, F., REYMOND, A., KAPRANOV, P., ROZOWSKY, J., ZHENG, D., CASTELO, R., FRANKISH, A., HARROW, J., GHOSH, S., SANDELIN, A., HOFACKER, I. L., BAERTSCH, R., KEEFE, D., DIKE, S., CHENG, J., HIRSCH, H. A., SEKINGER, E. A., LAGARDE, J., ABRIL, J. F., SHAHAB, A., FLAMM, C., FRIED, C., HACKERMÜLLER, J., HERTEL, J., LINDEMEYER, M., MISSAL, K., TANZER, A., WASHIETL, S., KORBEL, J., EMANUELSSON, O., PEDERSEN, J. S., HOLROYD, N., TAYLOR, R., SWARBRECK, D., MATTHEWS, N., DICKSON, M. C., THOMAS, D. J., WEIRAUCH, M. T., GILBERT, J., et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature,* 447**,** 799-816.

BOEVA, V., LERMINE, A., BARETTE, C., GUILLOUF, C. & BARILLOT, E. 2012. Nebula--a web-server for advanced ChIP-seq data analysis. *Bioinformatics,* 28**,** 2517-9.

BRANDEIS, M., FRANK, D., KESHET, I., SIEGFRIED, Z., MENDELSOHN, M., NEMES, A., TEMPER, V., RAZIN, A. & CEDAR, H. 1994. Sp1 elements protect a CpG island from de novo methylation. *Nature,* 371**,** 435-8.

BRENNER, C., DEPLUS, R., DIDELOT, C., LORIOT, A., VIRÉ, E., DE SMET, C., GUTIERREZ, A., DANOVI, D., BERNARD, D., BOON, T., PELICCI, P. G., AMATI, B., KOUZARIDES, T., DE LAUNOIT, Y., DI CROCE, L. & FUKS, F. 2005. Myc represses transcription through recruitment of DNA methyltransferase corepressor. *EMBO J,* 24**,** 336-46.

BROWN, J. L., MUCCI, D., WHITELEY, M., DIRKSEN, M. L. & KASSIS, J. A. 1998. The Drosophila Polycomb group gene pleiohomeotic encodes a DNA binding protein with homology to the transcription factor YY1. *Mol Cell,* 1**,** 1057-64.

BRUNS, G. A. & INGRAM, V. M. 1973. Erythropoiesis in the developing chick embryo. *Dev Biol,* 30**,** 455-9.

BUCKLE, R., BALMER, M., YENIDUNYA, A. & ALLAN, J. 1991. The promoter and enhancer of the inactive chicken beta-globin gene contains precisely positioned nucleosomes. *Nucleic Acids Res,* 19**,** 1219-26.

BULGER, M., BENDER, M. A., VAN DOORNINCK, J. H., WERTMAN, B., FARRELL, C. M., FELSENFELD, G., GROUDINE, M. & HARDISON, R. 2000. Comparative structural and functional analysis of the olfactory receptor genes flanking the human and mouse beta-globin gene clusters. *Proc Natl Acad Sci U S A,* 97**,** 14560-5.

CARETTI, G., DI PADOVA, M., MICALES, B., LYONS, G. E. & SARTORELLI, V. 2004. The Polycomb Ezh2 methyltransferase regulates muscle gene expression and skeletal muscle differentiation. *Genes Dev,* 18**,** 2627-38.

CHANG, B., CHEN, Y., ZHAO, Y. & BRUICK, R. K. 2007. JMJD6 is a histone arginine demethylase. *Science,* 318**,** 444-7.

CHOI, O. R. & ENGEL, J. D. 1988. Developmental regulation of beta-globin gene switching. *Cell,* 55**,** 17-26.

CHRISTMAN, J. K. 2002. 5-Azacytidine and 5-aza-2'-deoxycytidine as inhibitors of DNA methylation: mechanistic studies and their implications for cancer therapy. *Oncogene,* 21**,** 5483-95.

CHUNG, J. H., BELL, A. C. & FELSENFELD, G. 1997. Characterization of the chicken beta-globin insulator. *Proc Natl Acad Sci U S A,* 94**,** 575-80.

CHUNG, J. H., WHITELEY, M. & FELSENFELD, G. 1993. A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila. *Cell,* 74**,** 505-14.

CLAPIER, C. R. & CAIRNS, B. R. 2009. The biology of chromatin remodeling complexes. *Annu Rev Biochem,* 78**,** 273-304.

CLARK, S. J., HARRISON, J., PAUL, C. L. & FROMMER, M. 1994. High sensitivity mapping of methylated cytosines. *Nucleic Acids Res,* 22**,** 2990-7.

CLARK, S. P., LEWIS, C. D. & FELSENFELD, G. 1990. Properties of BGP1, a poly(dG)-binding protein from chicken erythrocytes. *Nucleic Acids Res,* 18**,** 5119-26.

CONERLY, M. L., TEVES, S. S., DIOLAITI, D., ULRICH, M., EISENMAN, R. N. & HENIKOFF, S. 2010. Changes in H2A.Z occupancy and DNA methylation during B-cell lymphomagenesis. *Genome Res,* 20**,** 1383-90.

CONSORTIUM, E. P. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science,* 306**,** 636-40.

COOPER, GM. 2000. The Cell: A Molecular Approach. 2nd edition. Sunderland (MA): Sinauer Associates. Available from: http://www.ncbi.nlm.nih.gov/books/NBK9839/

CORTELLINO, S., XU, J., SANNAI, M., MOORE, R., CARETTI, E., CIGLIANO, A., LE COZ, M., DEVARAJAN, K., WESSELS, A., SOPRANO, D., ABRAMOWITZ, L. K., BARTOLOMEI, M. S., RAMBOW, F., BASSI, M. R., BRUNO, T., FANCIULLI, M., RENNER, C., KLEIN-SZANTO, A. J., MATSUMOTO, Y., KOBI, D., DAVIDSON, I., ALBERTI, C., LARUE, L. & BELLACOSA, A. 2011. Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. *Cell,* 146**,** 67-79.

CREYGHTON, M. P., CHENG, A. W., WELSTEAD, G. G., KOOISTRA, T., CAREY, B. W., STEINE, E. J., HANNA, J., LODATO, M. A., FRAMPTON, G. M., SHARP, P. A., BOYER, L. A., YOUNG, R. A. & JAENISCH, R. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A,* 107**,** 21931-6.

DÁVALOS-SALAS, M., FURLAN-MAGARIL, M., GONZÁLEZ-BUENDÍA, E., VALDES-QUEZADA, C., AYALA-ORTEGA, E. & RECILLAS-TARGA, F. 2011. Gain of DNA methylation is enhanced in the absence of CTCF at the human retinoblastoma gene promoter. *BMC Cancer,* 11**,** 232.

DAVIS, J. A., TAKAGI, Y., KORNBERG, R. D. & ASTURIAS, F. A. 2002. Structure of the yeast RNA polymerase II holoenzyme: Mediator conformation and polymerase interaction. *Mol Cell,* 10**,** 409-15.

DE GOBBI, M., ANGUITA, E., HUGHES, J., SLOANE-STANLEY, J. A., SHARPE, J. A., KOCH, C. M., DUNHAM, I., GIBBONS, R. J., WOOD, W. G. & HIGGS, D. R. 2007. Tissue-specific histone modification and transcription factor binding in alpha globin gene expression. *Blood,* 110**,** 4503-10.

DEATON, A. M. & BIRD, A. 2011. CpG islands and the regulation of transcription. *Genes Dev,* 25**,** 1010-22.

DEKKER, J., RIPPE, K., DEKKER, M. & KLECKNER, N. 2002. Capturing chromosome conformation. *Science,* 295**,** 1306-11.

DHAMI, P., BRUCE, A. W., JIM, J. H., DILLON, S. C., HALL, A., COOPER, J. L., BONHOURE, N., CHIANG, K., ELLIS, P. D., LANGFORD, C., ANDREWS, R. M. & VETRIE, D. 2010.

Genomic approaches uncover increasing complexities in the regulatory landscape at the human SCL (TAL1) locus. *PLoS One,* 5**,** e9059.

DHAMI, P., COFFEY, A. J., ABBS, S., VERMEESCH, J. R., DUMANSKI, J. P., WOODWARD, K. J., ANDREWS, R. M., LANGFORD, C. & VETRIE, D. 2005. Exon array CGH: detection of copy-number changes at the resolution of individual exons in the human genome. *Am J Hum Genet,* 76**,** 750-62.

DICKSON, J., GOWHER, H., STROGANTSEV, R., GASZNER, M., HAIR, A., FELSENFELD, G. & WEST, A. G. 2010. VEZF1 elements mediate protection from DNA methylation. *PLoS Genet,* 6**,** e1000804.

DIKSTEIN, R. 2011. The unexpected traits associated with core promoter elements. *Transcription,* 2**,** 201-6.

DUNHAM, I., KUNDAJE, A., ALDRED, S. F., COLLINS, P. J., DAVIS, C. A., DOYLE, F., EPSTEIN, C. B., FRIETZE, S., HARROW, J., KAUL, R., KHATUN, J., LAJOIE, B. R., LANDT, S. G., LEE, B. K., PAULI, F., ROSENBLOOM, K. R., SABO, P., SAFI, A., SANYAL, A., SHORESH, N., SIMON, J. M., SONG, L., TRINKLEIN, N. D., ALTSHULER, R. C., BIRNEY, E., BROWN, J. B., CHENG, C., DJEBALI, S., DONG, X., ERNST, J., FUREY, T. S., GERSTEIN, M., GIARDINE, B., GREVEN, M., HARDISON, R. C., HARRIS, R. S., HERRERO, J., HOFFMAN, M. M., IYER, S., KELLLIS, M., KHERADPOUR, P., LASSMANN, T., LI, Q., LIN, X., MARINOV, G. K., MERKEL, A., MORTAZAVI, A., PARKER, S. C., REDDY, T. E., ROZOWSKY, J., SCHLESINGER, F., THURMAN, R. E., WANG, J., WARD, L. D., WHITFIELD, T. W., WILDER, S. P., WU, W., XI, H. S., YIP, K. Y., ZHUANG, J., BERNSTEIN, B. E., GREEN, E. D., GUNTER, C., SNYDER, M., PAZIN, M. J., LOWDON, R. F., DILLON, L. A., ADAMS, L. B., KELLY, C. J., ZHANG, J., WEXLER, J. R., GOOD, P. J., FEINGOLD, E. A., CRAWFORD, G. E., DEKKER, J., ELINITSKI, L., FARNHAM, P. J., GIDDINGS, M. C., GINGERAS, T. R., GUIGÓ, R., HUBBARD, T. J., KELLIS, M., KENT, W. J., LIEB, J. D., MARGULIES, E. H., MYERS, R. M., STARNATOYANNOPOULOS, J. A., TENNEBAUM, S. A., WENG, Z., WHITE, K. P., WOLD, B., YU, Y., WROBEL, J., RISK, B. A., GUNAWARDENA, H. P., KUIPER, H. C., MAIER, C. W., XIE, L., CHEN, X., MIKKELSEN, T. S., et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489**,** 57-74.

EDWARDS, J. R., O'DONNELL, A. H., ROLLINS, R. A., PECKHAM, H. E., LEE, C., MILEKIC, M. H., CHANRION, B., FU, Y., SU, T., HIBSHOOSH, H., GINGRICH, J. A., HAGHIGHI, F., NUTTER, R. & BESTOR, T. H. 2010. Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. *Genome Res,* 20**,** 972-80.

ELROD-ERICKSON, M., BENSON, T. E. & PABO, C. O. 1998. High-resolution structures of variant Zif268-DNA complexes: implications for understanding zinc finger-DNA recognition. *Structure,* 6**,** 451-64.

EMERSON, B. M., LEWIS, C. D. & FELSENFELD, G. 1985. Interaction of specific nuclear factors with the nuclease-hypersensitive region of the chicken adult beta-globin gene: nature of the binding domain. *Cell,* 41**,** 21-30.

EMERSON, B. M., NICKOL, J. M., JACKSON, P. D. & FELSENFELD, G. 1987. Analysis of the tissue-specific enhancer at the 3' end of the chicken adult beta-globin gene. *Proc Natl Acad Sci U S A,* 84**,** 4786-90.

EMILI, A., GREENBLATT, J. & INGLES, C. J. 1994. Species-specific interaction of the glutamine-rich activation domains of Sp1 with the TATA box-binding protein. *Mol Cell Biol,* 14**,** 1582-93.

ERDEL, F., KRUG, J., LÄNGST, G. & RIPPE, K. 2011. Targeting chromatin remodelers: signals and search mechanisms. *Biochim Biophys Acta,* 1809**,** 497-508.

ERNST, J. & KELLIS, M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol,* 28**,** 817-25.

ERNST, J., KHERADPOUR, P., MIKKELSEN, T. S., SHORESH, N., WARD, L. D., EPSTEIN, C. B., ZHANG, X., WANG, L., ISSNER, R., COYNE, M., KU, M., DURHAM, T., KELLIS, M. & BERNSTEIN, B. E. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature,* 473**,** 43-9.

ESTELLER, M. 2007. Epigenetic gene silencing in cancer: the DNA hypermethylome. *Hum Mol Genet,* 16 Spec No 1**,** R50-9.

FAN, S., FANG, F., ZHANG, X. & ZHANG, M. Q. 2007. Putative zinc finger protein binding sites are over-represented in the boundaries of methylation-resistant CpG islands in the human genome. *PLoS One,* 2**,** e1184.

FEDORIW, A. M., STEIN, P., SVOBODA, P., SCHULTZ, R. M. & BARTOLOMEI, M. S. 2004. Transgenic RNAi reveals essential function for CTCF in H19 gene imprinting. *Science,* 303**,** 238-40.

FELSENFELD, G. & GROUDINE, M. 2003. Controlling the double helix. *Nature,* 421**,** 448-53.

FOLLOWS, G. A., DHAMI, P., GÖTTGENS, B., BRUCE, A. W., CAMPBELL, P. J., DILLON, S. C., SMITH, A. M., KOCH, C., DONALDSON, I. J., SCOTT, M. A., DUNHAM, I., JANES, M. E., VETRIE, D. & GREEN, A. R. 2006. Identifying gene regulatory elements by genomic microarray mapping of DNaseI hypersensitive sites. *Genome Res,* 16**,** 1310-9.

FRANK, S. R., PARISI, T., TAUBERT, S., FERNANDEZ, P., FUCHS, M., CHAN, H. M., LIVINGSTON, D. M. & AMATI, B. 2003. MYC recruits the TIP60 histone acetyltransferase complex to chromatin. *EMBO Rep,* 4**,** 575-80.

FRITSCH, C., BROWN, J. L., KASSIS, J. A. & MÜLLER, J. 1999. The DNA-binding polycomb group protein pleiohomeotic mediates silencing of a Drosophila homeotic gene. *Development,* 126**,** 3905-13.

FU, Y., SINHA, M., PETERSON, C. L. & WENG, Z. 2008. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet,* 4**,** e1000138.

FUJIWARA, T., O'GEEN, H., KELES, S., BLAHNIK, K., LINNEMANN, A. K., KANG, Y. A., CHOI, K., FARNHAM, P. J. & BRESNICK, E. H. 2009. Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. *Mol Cell,* 36**,** 667-81.

FUKS, F. 2005. DNA methylation and histone modifications: teaming up to silence genes. *Curr Opin Genet Dev,* 15**,** 490-5.

GALLARDA, J. L., FOLEY, K. P., YANG, Z. Y. & ENGEL, J. D. 1989. The beta-globin stage selector element factor is erythroid-specific promoter/enhancer binding protein NF-E4. *Genes Dev,* 3**,** 1845-59.

GARDINER-GARDEN, M. & FROMMER, M. 1987. CpG islands in vertebrate genomes. *J Mol Biol,* 196**,** 261-82.

GERBER, H. P., SEIPEL, K., GEORGIEV, O., HÖFFERER, M., HUG, M., RUSCONI, S. & SCHAFFNER, W. 1994. Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science,* 263**,** 808-11.

GÉVRY, N., CHAN, H. M., LAFLAMME, L., LIVINGSTON, D. M. & GAUDREAU, L. 2007. p21 transcription is regulated by differential localization of histone H2A.Z. *Genes Dev,* 21**,** 1869-81.

GÉVRY, N., HARDY, S., JACQUES, P. E., LAFLAMME, L., SVOTELIS, A., ROBERT, F. & GAUDREAU, L. 2009. Histone H2A.Z is essential for estrogen receptor signaling. *Genes Dev,* 23**,** 1522-33.

GILL, G., PASCAL, E., TSENG, Z. H. & TJIAN, R. 1994. A glutamine-rich hydrophobic patch in transcription factor Sp1 contacts the dTAFII110 component of the Drosophila TFIID complex and mediates transcriptional activation. *Proc Natl Acad Sci U S A,* 91**,** 192-6.

GOWHER, H., BRICK, K., CAMERINI-OTERO, R. D. & FELSENFELD, G. 2012. Vezf1 protein binding sites genome-wide are associated with pausing of elongating RNA polymerase II. *Proc Natl Acad Sci U S A,* 109**,** 2370-5.

GOWHER, H., STUHLMANN, H. & FELSENFELD, G. 2008. Vezf1 regulates genomic DNA methylation through its effects on expression of DNA methyltransferase Dnmt3b. *Genes Dev,* 22**,** 2075-84.

GREENBLATT, J. 1997. RNA polymerase II holoenzyme and transcriptional regulation. *Curr Opin Cell Biol,* 9**,** 310-9.

GREER, E. L. & SHI, Y. 2012. Histone methylation: a dynamic mark in health, disease and inheritance. *Nat Rev Genet,* 13**,** 343-57.

GREWAL, S. I. & JIA, S. 2007. Heterochromatin revisited. *Nat Rev Genet,* 8**,** 35-46.

GRUNAU, C., CLARK, S. J. & ROSENTHAL, A. 2001. Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acids Res,* 29**,** E65-5.

GUO, J. U., SU, Y., ZHONG, C., MING, G. L. & SONG, H. 2011. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell,* 145**,** 423-34.

HALLIKAS, O. & TAIPALE, J. 2006. High-throughput assay for determining specificity and affinity of protein-DNA binding interactions. *Nat Protoc,* 1**,** 215-22.

HAMPSEY, M. & REINBERG, D. 2003. Tails of intrigue: phosphorylation of RNA polymerase II mediates histone methylation. *Cell,* 113**,** 429-32.

HAN, L., LIN, I. G. & HSIEH, C. L. 2001. Protein binding protects sites on stable episomes and in the chromosome from de novo methylation. *Mol Cell Biol,* 21**,** 3416-24.

HARDY, S., JACQUES, P. E., GÉVRY, N., FOREST, A., FORTIN, M. E., LAFLAMME, L., GAUDREAU, L. & ROBERT, F. 2009. The euchromatic and heterochromatic landscapes are shaped by antagonizing effects of transcription on H2A.Z deposition. *PLoS Genet,* 5**,** e1000687.

HARROW, J., FRANKISH, A., GONZALEZ, J. M., TAPANARI, E., DIEKHANS, M., KOKOCINSKI, F., AKEN, B. L., BARRELL, D., ZADISSA, A., SEARLE, S., BARNES, I., BIGNELL, A., BOYCHENKO, V., HUNT, T., KAY, M., MUKHERJEE, G., RAJAN, J., DESPACIO-REYES, G., SAUNDERS, G., STEWARD, C., HARTE, R., LIN, M., HOWALD, C., TANZER, A., DERRIEN, T., CHRAST, J., WALTERS, N., BALASUBRAMANIAN, S., PEI, B., TRESS, M., RODRIGUEZ, J. M., EZKURDIA, I., VAN BAREN, J., BRENT, M., HAUSSLER, D., KELLIS, M., VALENCIA, A., REYMOND, A., GERSTEIN, M., GUIGÓ, R. & HUBBARD, T. J. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res,* 22**,** 1760-74.

HASSAN, A. H., AWAD, S. & PROCHASSON, P. 2006. The Swi2/Snf2 bromodomain is required for the displacement of SAGA and the octamer transfer of SAGA-acetylated nucleosomes. *J Biol Chem,* 281**,** 18126-34.

HE, Y. F., LI, B. Z., LI, Z., LIU, P., WANG, Y., TANG, Q., DING, J., JIA, Y., CHEN, Z., LI, L., SUN, Y., LI, X., DAI, Q., SONG, C. X., ZHANG, K., HE, C. & XU, G. L. 2011. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science,* 333**,** 1303-7.

HO, J. W., BISHOP, E., KARCHENKO, P. V., NÈGRE, N., WHITE, K. P. & PARK, P. J. 2011. ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC Genomics,* 12**,** 134.

HORI, N., NAKANO, H., TAKEUCHI, T., KATO, H., HAMAGUCHI, S., OSHIMURA, M. & SATO, K. 2002. A dyad oct-binding sequence functions as a maintenance sequence for the unmethylated state within the H19/Igf2-imprinted control region. *J Biol Chem,* 277**,** 27960-7.

HU, G., SCHONES, D. E., CUI, K., YBARRA, R., NORTHRUP, D., TANG, Q., GATTINONI, L., RESTIFO, N. P., HUANG, S. & ZHAO, K. 2011. Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. *Genome Res,* 21**,** 1650-8.

HUANG, S., LI, X., YUSUFZAI, T. M., QIU, Y. & FELSENFELD, G. 2007. USF1 recruits histone modification complexes and is critical for maintenance of a chromatin barrier. *Mol Cell Biol,* 27**,** 7991-8002.

HUGHES, A. L., JIN, Y., RANDO, O. J. & STRUHL, K. 2012. A functional evolutionary approach to identify determinants of nucleosome positioning: a unifying model for establishing the genome-wide pattern. *Mol Cell,* 48**,** 5-15.

IKUTA, T. & KAN, Y. W. 1991. In vivo protein-DNA interactions at the beta-globin gene locus. *Proc Natl Acad Sci U S A,* 88**,** 10188-92.

ITO, S., SHEN, L., DAI, Q., WU, S. C., COLLINS, L. B., SWENBERG, J. A., HE, C. & ZHANG, Y. 2011. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science,* 333**,** 1300-3.

IYER, V. R. 2012. Nucleosome positioning: bringing order to the eukaryotic genome. *Trends Cell Biol,* 22**,** 250-6.

IYER, V. R., HORAK, C. E., SCAFE, C. S., BOTSTEIN, D., SNYDER, M. & BROWN, P. O. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature,* 409**,** 533-8.

JACKSON, P. D., EVANS, T., NICKOL, J. M. & FELSENFELD, G. 1989. Developmental modulation of protein binding to beta-globin gene regulatory sites within chicken erythrocyte nuclei. *Genes Dev,* 3**,** 1860-73.

JIN, C. & FELSENFELD, G. 2007. Nucleosome stability mediated by histone variants H3.3 and H2A.Z. *Genes Dev,* 21**,** 1519-29.

JIN, Q., YU, L. R., WANG, L., ZHANG, Z., KASPER, L. H., LEE, J. E., WANG, C., BRINDLE, P. K., DENT, S. Y. & GE, K. 2011. Distinct roles of GCN5/PCAF-mediated H3K9ac and CBP/p300-mediated H3K18/27ac in nuclear receptor transactivation. *EMBO J,* 30**,** 249-62.

JOHNSON, L. M., BOSTICK, M., ZHANG, X., KRAFT, E., HENDERSON, I., CALLIS, J. & JACOBSEN, S. E. 2007. The SRA methyl-cytosine-binding domain links DNA and histone methylation. *Curr Biol,* 17**,** 379-84.

JONES, P. L., VEENSTRA, G. J., WADE, P. A., VERMAAK, D., KASS, S. U., LANDSBERGER, N., STROUBOULIS, J. & WOLFFE, A. P. 1998. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat Genet,* 19**,** 187-91.

KADAUKE, S. & BLOBEL, G. A. 2009. Chromatin loops in gene regulation. *Biochim Biophys Acta,* 1789**,** 17-25.

KAMAKAKA, R. T. & BIGGINS, S. 2005. Histone variants: deviants? *Genes Dev,* 19**,** 295-310.

KAPLAN, T., FRIEDMAN, N. & MARGALIT, H. 2005. Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput Biol,* 1**,** e1.

KAUFMAN, Y., HELED, M., PERK, J., RAZIN, A. & SHEMER, R. 2009. Protein-binding elements establish in the oocyte the primary imprint of the Prader-Willi/Angelman syndromes domain. *Proc Natl Acad Sci U S A,* 106**,** 10242-7.

KIM, J., GUERMAH, M., MCGINTY, R. K., LEE, J. S., TANG, Z., MILNE, T. A., SHILATIFARD, A., MUIR, T. W. & ROEDER, R. G. 2009. RAD6-Mediated transcription-coupled H2B

ubiquitylation directly stimulates H3K4 methylation in human cells. *Cell,* 137**,** 459-71.

KOYANO-NAKAGAWA, N., NISHIDA, J., BALDWIN, D., ARAI, K. & YOKOTA, T. 1994. Molecular cloning of a novel human cDNA encoding a zinc finger protein that binds to the interleukin-3 promoter. *Mol Cell Biol,* 14**,** 5099-107.

KROGAN, N. J., KIM, M., TONG, A., GOLSHANI, A., CAGNEY, G., CANADIEN, V., RICHARDS, D. P., BEATTIE, B. K., EMILI, A., BOONE, C., SHILATIFARD, A., BURATOWSKI, S. & GREENBLATT, J. 2003. Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II. *Mol Cell Biol,* 23**,** 4207-18.

KUHNERT, F., CAMPAGNOLO, L., XIONG, J. W., LEMONS, D., FITCH, M. J., ZOU, Z., KIOSSES, W. B., GARDNER, H. & STUHLMANN, H. 2005. Dosage-dependent requirement for mouse Vezf1 in vascular system development. *Dev Biol,* 283**,** 140-56.

KULAEVA, O. I., NIZOVTSEVA, E. V., POLIKANOV, Y. S., ULIANOV, S. V. & STUDITSKY, V. M. 2012. Distant activation of transcription: mechanisms of enhancer action. *Mol Cell Biol,* 32**,** 4892-7.

LAHLIL, R., LÉCUYER, E., HERBLOT, S. & HOANG, T. 2004. SCL assembles a multifactorial complex that determines glycophorin A expression. *Mol Cell Biol,* 24**,** 1439-52.

LAMPRONTI, I., BIANCHI, N., BORGATTI, M., FIBACH, E., PRUS, E. & GAMBARI, R. 2003. Accumulation of gamma-globin mRNA in human erythroid cells treated with angelicin. *Eur J Haematol,* 71**,** 189-95.

LANDT, S. G., MARINOV, G. K., KUNDAJE, A., KHERADPOUR, P., PAULI, F., BATZOGLOU, S., BERNSTEIN, B. E., BICKEL, P., BROWN, J. B., CAYTING, P., CHEN, Y., DESALVO, G., EPSTEIN, C., FISHER-AYLOR, K. I., EUSKIRCHEN, G., GERSTEIN, M., GERTZ, J., HARTEMINK, A. J., HOFFMAN, M. M., IYER, V. R., JUNG, Y. L., KARMAKAR, S., KELLIS, M., KHARCHENKO, P. V., LI, Q., LIU, T., LIU, X. S., MA, L., MILOSAVLJEVIC, A., MYERS, R. M., PARK, P. J., PAZIN, M. J., PERRY, M. D., RAHA, D., REDDY, T. E., ROZOWSKY, J., SHORESH, N., SIDOW, A., SLATTERY, M., STAMATOYANNOPOULOS, J. A., TOLSTORUKOV, M. Y., WHITE, K. P., XI, S., FARNHAM, P. J., LIEB, J. D., WOLD, B. J. & SNYDER, M. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res,* 22**,** 1813-31.

LANGMEAD, B. 2010. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics,* Chapter 11**,** Unit 11.7.

LANGMEAD, B., TRAPNELL, C., POP, M. & SALZBERG, S. L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol,* 10**,** R25.

LEE, C. K., SHIBATA, Y., RAO, B., STRAHL, B. D. & LIEB, J. D. 2004. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat Genet,* 36**,** 900-5.

LEE, J. S., SHUKLA, A., SCHNEIDER, J., SWANSON, S. K., WASHBURN, M. P., FLORENS, L., BHAUMIK, S. R. & SHILATIFARD, A. 2007a. Histone crosstalk between H2B monoubiquitination and H3 methylation mediated by COMPASS. *Cell,* 131**,** 1084-96.

LEE, K., LAU, Z. Z., MEREDITH, C. & PARK, J. H. 2012. Decrease of p400 ATPase complex and loss of H2A.Z within the p21 promoter occur in senescent IMR-90 human fibroblasts. *Mech Ageing Dev,* 133**,** 686-94.

LEE, W., TILLO, D., BRAY, N., MORSE, R. H., DAVIS, R. W., HUGHES, T. R. & NISLOW, C. 2007b. A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet,* 39**,** 1235-44.

LEWIS, A. & MURRELL, A. 2004. Genomic imprinting: CTCF protects the boundaries. *Curr Biol,* 14**,** R284-6.

LEWIS, C. D., CLARK, S. P., FELSENFELD, G. & GOULD, H. 1988. An erythrocyte-specific protein that binds to the poly(dG) region of the chicken beta-globin gene promoter. *Genes Dev,* 2**,** 863-73.

LIN, I. G. & HSIEH, C. L. 2001. Chromosomal DNA demethylation specified by protein binding. *EMBO Rep,* 2**,** 108-12.

LIN, I. G., TOMZYNSKI, T. J., OU, Q. & HSIEH, C. L. 2000. Modulation of DNA binding protein affinity directly affects target site demethylation. *Mol Cell Biol,* 20**,** 2343-9.

LIU, E. T., POTT, S. & HUSS, M. 2010. Q&A: ChIP-seq technologies and the study of gene regulation. *BMC Biol,* 8**,** 56.

LORCH, Y., LAPOINTE, J. W. & KORNBERG, R. D. 1987. Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. *Cell,* 49**,** 203-10.

LOZZIO, C. B. & LOZZIO, B. B. 1975. Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. *Blood,* 45**,** 321-34.

LUGER, K., DECHASSA, M. L. & TREMETHICK, D. J. 2012. New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nat Rev Mol Cell Biol,* 13**,** 436-47.

MA, M. K., HEATH, C., HAIR, A. & WEST, A. G. 2011. Histone crosstalk directed by H2B ubiquitination is required for chromatin boundary integrity. *PLoS Genet,* 7**,** e1002175.

MA, X., KULKARNI, A., ZHANG, Z., XUAN, Z., SERFLING, R. & ZHANG, M. Q. 2012. A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information. *Nucleic Acids Res,* 40**,** e50.

MACHANICK, P. & BAILEY, T. L. 2011. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics,* 27**,** 1696-7.

MACLEOD, D., CHARLTON, J., MULLINS, J. & BIRD, A. P. 1994. Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. *Genes Dev,* 8**,** 2282-92.

MANELYTE, L. & LANGST, G. 2013. Chromatin Remodelers and Their Way of Action.

MAO, C., BROWN, C. R., GRIESENBECK, J. & BOEGER, H. 2011. Occlusion of regulatory sequences by promoter nucleosomes in vivo. *PLoS One,* 6**,** e17521.

MARIN, M., KARIS, A., VISSER, P., GROSVELD, F. & PHILIPSEN, S. 1997. Transcription factor Sp1 is essential for early embryonic development but dispensable for cell growth and differentiation. *Cell,* 89**,** 619-28.

MARQUES, M., LAFLAMME, L., GERVAIS, A. L. & GAUDREAU, L. 2010. Reconciling the positive and negative roles of histone H2A.Z in gene transcription. *Epigenetics,* 5**,** 267-72.

MASON, M. M., GRASSO, J. A., GAVRILOVA, O. & REITMAN, M. 1996. Identification of functional elements of the chicken epsilon-globin promoter involved in stage-specific interaction with the beta/epsilon enhancer. *J Biol Chem,* 271**,** 25459-67.

MASTON, G. A., EVANS, S. K. & GREEN, M. R. 2006. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet,* 7**,** 29-59.

MCGHEE, J. D. & GINDER, G. D. 1979. Specific DNA methylation sites in the vicinity of the chicken beta-globin genes. *Nature,* 280**,** 419-20.

MCMAHON, S. B., WOOD, M. A. & COLE, M. D. 2000. The essential cofactor TRRAP recruits the histone acetyltransferase hGCN5 to c-Myc. *Mol Cell Biol,* 20**,** 556-62.

MINIE, M. E., KIMURA, T. & FELSENFELD, G. 1992. The developmental switch in embryonic rho-globin expression is correlated with erythroid lineage-specific differences in transcription factor levels. *Development,* 115**,** 1149-64.

MIYASHITA, H., KANEMURA, M., YAMAZAKI, T., ABE, M. & SATO, Y. 2004. Vascular endothelial zinc finger 1 is involved in the regulation of angiogenesis: possible contribution of stathmin/OP18 as a downstream target gene. *Arterioscler Thromb Vasc Biol,* 24**,** 878-84.

MIYASHITA, H. & SATO, Y. 2005. Metallothionein 1 is a downstream target of vascular endothelial zinc finger 1 (VEZF1) in endothelial cells and participates in the regulation of angiogenesis. *Endothelium,* 12**,** 163-70.

NAN, X., NG, H. H., JOHNSON, C. A., LAHERTY, C. D., TURNER, B. M., EISENMAN, R. N. & BIRD, A. 1998. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature,* 393**,** 386-9.

NAUMANN, S., REUTZEL, D., SPEICHER, M. & DECKER, H. J. 2001. Complete karyotype characterization of the K562 cell line by combined application of G-banding, multiplex-fluorescence in situ hybridization, fluorescence in situ hybridization, and comparative genomic hybridization. *Leuk Res,* 25**,** 313-22.

NG, H. H., ROBERT, F., YOUNG, R. A. & STRUHL, K. 2003. Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Mol Cell,* 11**,** 709-19.

OISHI, H., KITAGAWA, H., WADA, O., TAKEZAWA, S., TORA, L., KOUZU-FUJITA, M., TAKADA, I., YANO, T., YANAGISAWA, J. & KATO, S. 2006. An hGCN5/TRRAP histone acetyltransferase complex co-activates BRCA1 transactivation function through histone modification. *J Biol Chem,* 281**,** 20-6.

OOI, S. K., QIU, C., BERNSTEIN, E., LI, K., JIA, D., YANG, Z., ERDJUMENT-BROMAGE, H., TEMPST, P., LIN, S. P., ALLIS, C. D., CHENG, X. & BESTOR, T. H. 2007. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature,* 448**,** 714-7.

OSWALD, J., ENGEMANN, S., LANE, N., MAYER, W., OLEK, A., FUNDELE, R., DEAN, W., REIK, W. & WALTER, J. 2000. Active demethylation of the paternal genome in the mouse zygote. *Curr Biol,* 10**,** 475-8.

PACE, B. S., QIAN, X. H., SANGERMAN, J., OFORI-ACQUAH, S. F., BALIGA, B. S., HAN, J. & CRITZ, S. D. 2003. p38 MAP kinase activation mediates gamma-globin gene induction in erythroid progenitors. *Exp Hematol,* 31**,** 1089-96.

PALSTRA, R. J., TOLHUIS, B., SPLINTER, E., NIJMEIJER, R., GROSVELD, F. & DE LAAT, W. 2003. The beta-globin nuclear compartment in development and erythroid differentiation. *Nat Genet,* 35**,** 190-4.

PANT, V., MARIANO, P., KANDURI, C., MATTSSON, A., LOBANENKOV, V., HEUCHEL, R. & OHLSSON, R. 2003. The nucleotides responsible for the direct physical contact between the chromatin insulator protein CTCF and the H19 imprinting control region manifest parent of origin-specific long-distance insulation and methylation-free domains. *Genes Dev,* 17**,** 586-90.

PARK, Y. J., DYER, P. N., TREMETHICK, D. J. & LUGER, K. 2004. A new fluorescence resonance energy transfer approach demonstrates that the histone variant H2AZ stabilizes the histone octamer within the nucleosome. *J Biol Chem,* 279**,** 24274-82.

PAVLETICH, N. P. & PABO, C. O. 1991. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science,* 252**,** 809-17.

PETERSON, C. L. & LANIEL, M. A. 2004. Histones and histone modifications. *Curr Biol,* 14**,** R546-51.

PHILIPSEN, S., PRUZINA, S. & GROSVELD, F. 1993. The minimal requirements for activity in transgenic mice of hypersensitive site 3 of the beta globin locus control region. *EMBO J,* 12**,** 1077-85.

PHILIPSEN, S., TALBOT, D., FRASER, P. & GROSVELD, F. 1990. The beta-globin dominant control region: hypersensitive site 2. *EMBO J,* 9**,** 2159-67.

PIKAART, M. J., RECILLAS-TARGA, F. & FELSENFELD, G. 1998. Loss of transcriptional activity of a transgene is accompanied by DNA methylation and histone deacetylation and is prevented by insulators. *Genes Dev,* 12**,** 2852-62.

POLACH, K. J. & WIDOM, J. 1996. A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. *J Mol Biol,* 258**,** 800-12.

POLACH, K. J. & WIDOM, J. 1995. Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation. *J Mol Biol,* 254**,** 130-49.

PRIOLEAU, M. N., NONY, P., SIMPSON, M. & FELSENFELD, G. 1999. An insulator element and condensed chromatin region separate the chicken beta-globin locus from an independently regulated erythroid-specific folate receptor gene. *EMBO J,* 18**,** 4035-48.

RADMAN-LIVAJA, M. & RANDO, O. J. 2010. Nucleosome positioning: how is it established, and why does it matter? *Dev Biol,* 339**,** 258-66.

RAMACHANDRAN, K., VAN WERT, J., GOPISETTY, G. & SINGAL, R. 2007. Developmentally regulated demethylase activity targeting the betaA-globin gene in primary avian erythroid cells. *Biochemistry,* 46**,** 3416-22.

RAND, E., BEN-PORATH, I., KESHET, I. & CEDAR, H. 2004. CTCF elements direct allele-specific undermethylation at the imprinted H19 locus. *Curr Biol,* 14**,** 1007-12.

REDDY, P. M. & SHEN, C. K. 1991. Protein-DNA interactions in vivo of an erythroid-specific, human beta-globin locus enhancer. *Proc Natl Acad Sci U S A,* 88**,** 8676-80.

ROBERTSON, G., HIRST, M., BAINBRIDGE, M., BILENKY, M., ZHAO, Y., ZENG, T., EUSKIRCHEN, G., BERNIER, B., VARHOL, R., DELANEY, A., THIESSEN, N., GRIFFITH, O. L., HE, A., MARRA, M., SNYDER, M. & JONES, S. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods,* 4**,** 651-7.

ROBERTSON, K. D. 2002. DNA methylation and chromatin - unraveling the tangled web. *Oncogene,* 21**,** 5361-79.

ROBERTSON, K. D. 2005. DNA methylation and human disease. *Nat Rev Genet,* 6**,** 597-610.

ROBERTSON, K. D. & WOLFFE, A. P. 2000. DNA methylation in health and disease. *Nat Rev Genet,* 1**,** 11-9.

RUTHERFORD, T., CLEGG, J. B., HIGGS, D. R., JONES, R. W., THOMPSON, J. & WEATHERALL, D. J. 1981. Embryonic erythroid differentiation in the human leukemic cell line K562. *Proc Natl Acad Sci U S A,* 78**,** 348-52.

SABO, P. J., HAWRYLYCZ, M., WALLACE, J. C., HUMBERT, R., YU, M., SHAFER, A., KAWAMOTO, J., HALL, R., MACK, J., DORSCHNER, M. O., MCARTHUR, M. & STAMATOYANNOPOULOS, J. A. 2004. Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc Natl Acad Sci U S A,* 101**,** 16837-42.

SALMON-DIVON, M., DVINGE, H., TAMMOJA, K. & BERTONE, P. 2010. PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. *BMC Bioinformatics,* 11**,** 415.

SARMA, K. & REINBERG, D. 2005. Histone variants meet their match. *Nat Rev Mol Cell Biol,* 6**,** 139-49.

SAXONOV, S., BERG, P. & BRUTLAG, D. L. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A,* 103**,** 1412-7.

SCHAEFFER, L., MONCOLLIN, V., ROY, R., STAUB, A., MEZZINA, M., SARASIN, A., WEEDA, G., HOEIJMAKERS, J. H. & EGLY, J. M. 1994. The ERCC2/DNA repair protein is associated with the class II BTF2/TFIIH transcription factor. *EMBO J,* 13**,** 2388-92.

SCHAEFFER, L., ROY, R., HUMBERT, S., MONCOLLIN, V., VERMEULEN, W., HOEIJMAKERS, J. H., CHAMBON, P. & EGLY, J. M. 1993. DNA repair helicase: a component of BTF2 (TFIIH) basic transcription factor. *Science,* 260**,** 58-63.

SCHOENHERR, C. J., LEVORSE, J. M. & TILGHMAN, S. M. 2003. CTCF maintains differential methylation at the Igf2/H19 locus. *Nat Genet,* 33**,** 66-9.

SEGAL, E., FONDUFE-MITTENDORF, Y., CHEN, L., THÅSTRÖM, A., FIELD, Y., MOORE, I. K., WANG, J. P. & WIDOM, J. 2006. A genomic code for nucleosome positioning. *Nature,* 442**,** 772-8.

SHENDURE, J. & JI, H. 2008. Next-generation DNA sequencing. *Nat Biotechnol,* 26**,** 1135-45.

SHENG, G. 2010. Primitive and definitive erythropoiesis in the yolk sac: a bird's eye view. *Int J Dev Biol,* 54**,** 1033-43.

SHI, Y., LAN, F., MATSON, C., MULLIGAN, P., WHETSTINE, J. R., COLE, P. A. & CASERO, R. A. 2004. Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell,* 119**,** 941-53.

SING, A., PANNELL, D., KARAISKAKIS, A., STURGEON, K., DJABALI, M., ELLIS, J., LIPSHITZ, H. D. & CORDES, S. P. 2009. A vertebrate Polycomb response element governs segmentation of the posterior hindbrain. *Cell,* 138**,** 885-97.

SINGAL, R., FERRIS, R., LITTLE, J. A., WANG, S. Z. & GINDER, G. D. 1997. Methylation of the minimal promoter of an embryonic globin gene silences transcription in primary erythroid cells. *Proc Natl Acad Sci U S A,* 94**,** 13724-9.

SINGAL, R., WANG, S. Z., SARGENT, T., ZHU, S. Z. & GINDER, G. D. 2002. Methylation of promoter proximal-transcribed sequences of an embryonic globin gene inhibits transcription in primary erythroid cells and promotes formation of a cell type-specific methyl cytosine binding complex. *J Biol Chem,* 277**,** 1897-905.

SOLOMON, M. J., LARSEN, P. L. & VARSHAVSKY, A. 1988. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell,* 53**,** 937-47.

SONG, C. X., SZULWACH, K. E., FU, Y., DAI, Q., YI, C., LI, X., LI, Y., CHEN, C. H., ZHANG, W., JIAN, X., WANG, J., ZHANG, L., LOONEY, T. J., ZHANG, B., GODLEY, L. A., HICKS, L. M., LAHN, B. T., JIN, P. & HE, C. 2011. Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat Biotechnol,* 29**,** 68-72.

SORRENTINO, B. P., NEY, P. A. & NIENHUIS, A. W. 1990. Localization and characterization of the DNase I-hypersensitive site II (HS II) enhancer. A critical regulatory element within the beta-globin locus-activating region. *Ann N Y Acad Sci,* 612**,** 141-51.

SPITZ, F. & FURLONG, E. E. 2012. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet,* 13**,** 613-26.

STRATHDEE, G., SIM, A. & BROWN, R. 2004. Control of gene expression by CpG island methylation in normal cells. *Biochem Soc Trans,* 32**,** 913-5.

STRAUSS, E. C. & ORKIN, S. H. 1992. In vivo protein-DNA interactions at hypersensitive site 3 of the human beta-globin locus control region. *Proc Natl Acad Sci U S A,* 89**,** 5809-13.

STROGANTSEV, R. S. 2009. *Mapping and characterisation of genomic binding sites of the chromatin barrier protein VEZF1.* Thesis (Ph D), University of Glasgow.

STROUD, H., FENG, S., MOREY KINNEY, S., PRADHAN, S. & JACOBSEN, S. E. 2011. 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biol,* 12**,** R54.

STRUHL, K. & SEGAL, E. 2013. Determinants of nucleosome positioning. *Nat Struct Mol Biol,* 20**,** 267-73.

SUTO, R. K., CLARKSON, M. J., TREMETHICK, D. J. & LUGER, K. 2000. Crystal structure of a nucleosome core particle containing the variant histone H2A.Z. *Nat Struct Biol,* 7**,** 1121-4.

TAFT, R. J., PHEASANT, M. & MATTICK, J. S. 2007. The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays,* 29**,** 288-99.

TAHILIANI, M., KOH, K. P., SHEN, Y., PASTOR, W. A., BANDUKWALA, H., BRUDNO, Y., AGARWAL, S., IYER, L. M., LIU, D. R., ARAVIND, L. & RAO, A. 2009. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science,* 324**,** 930-5.

TALBERT, P. B. & HENIKOFF, S. 2010. Histone variants--ancient wrap artists of the epigenome. *Nat Rev Mol Cell Biol,* 11**,** 264-75.

TAN, M., LUO, H., LEE, S., JIN, F., YANG, J. S., MONTELLIER, E., BUCHOU, T., CHENG, Z., ROUSSEAUX, S., RAJAGOPAL, N., LU, Z., YE, Z., ZHU, Q., WYSOCKA, J., YE, Y., KHOCHBIN, S., REN, B. & ZHAO, Y. 2011. Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell,* 146**,** 1016-28.

TAYLOR, J., SCHENCK, I., BLANKENBERG, D. & NEKRUTENKO, A. 2007. Using galaxy to perform large-scale interactive data analyses. *Curr Protoc Bioinformatics,* Chapter 10**,** Unit 10.5.

THAMBIRAJAH, A. A., DRYHURST, D., ISHIBASHI, T., LI, A., MAFFEY, A. H. & AUSIÓ, J. 2006. H2A.Z stabilizes chromatin in a way that is dependent on core histone acetylation. *J Biol Chem,* 281**,** 20036-44.

THOMSON, J. P., SKENE, P. J., SELFRIDGE, J., CLOUAIRE, T., GUY, J., WEBB, S., KERR, A. R., DEATON, A., ANDREWS, R., JAMES, K. D., TURNER, D. J., ILLINGWORTH, R. & BIRD, A. 2010. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature,* 464**,** 1082-6.

TREMETHICK, D. J. 2007. Higher-order structures of chromatin: the elusive 30 nm fiber. *Cell,* 128**,** 651-4.

TSAI, M. C., MANOR, O., WAN, Y., MOSAMMAPARAST, N., WANG, J. K., LAN, F., SHI, Y., SEGAL, E. & CHANG, H. Y. 2010. Long noncoding RNA as modular scaffold of histone modification complexes. *Science,* 329**,** 689-93.

VAGNARELLI, P. 2012. Mitotic chromosome condensation in vertebrates. *Exp Cell Res,* 318**,** 1435-41.

VALINLUCK, V. & SOWERS, L. C. 2007. Endogenous cytosine damage products alter the site selectivity of human DNA maintenance methyltransferase DNMT1. *Cancer Res,* 67**,** 946-50.

VALINLUCK, V., TSAI, H. H., ROGSTAD, D. K., BURDZY, A., BIRD, A. & SOWERS, L. C. 2004. Oxidative damage to methyl-CpG sequences inhibits the binding of the methyl-CpG binding domain (MBD) of methyl-CpG binding protein 2 (MeCP2). *Nucleic Acids Res,* 32**,** 4100-8.

VEGA, V. B., CHEUNG, E., PALANISAMY, N. & SUNG, W. K. 2009. Inherent signals in sequencing-based Chromatin-ImmunoPrecipitation control libraries. *PLoS One,* 4**,** e5241.

VIRÉ, E., BRENNER, C., DEPLUS, R., BLANCHON, L., FRAGA, M., DIDELOT, C., MOREY, L., VAN EYNDE, A., BERNARD, D., VANDERWINDEN, J. M., BOLLEN, M., ESTELLER, M., DI CROCE, L., DE LAUNOIT, Y. & FUKS, F. 2006. The Polycomb group protein EZH2 directly controls DNA methylation. *Nature,* 439**,** 871-4.

WADMAN, I. A., OSADA, H., GRÜTZ, G. G., AGULNICK, A. D., WESTPHAL, H., FORSTER, A. & RABBITTS, T. H. 1997. The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. *EMBO J,* 16**,** 3145-57.

WANG, J., LU, J., GU, G. & LIU, Y. 2011. In vitro DNA-binding profile of transcription factors: methods and new insights. *J Endocrinol,* 210**,** 15-27.

WEBER, M., DAVIES, J. J., WITTIG, D., OAKELEY, E. J., HAASE, M., LAM, W. L. & SCHÜBELER, D. 2005. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet,* 37**,** 853-62.

WEST, A. G., GASZNER, M. & FELSENFELD, G. 2002. Insulators: many functions, many mechanisms. *Genes Dev,* 16**,** 271-88.

WEST, A. G., HUANG, S., GASZNER, M., LITT, M. D. & FELSENFELD, G. 2004. Recruitment of histone modifications by USF proteins at a vertebrate barrier element. *Mol Cell,* 16**,** 453-63.

WILSON, B. G. & ROBERTS, C. W. 2011. SWI/SNF nucleosome remodellers and cancer. *Nat Rev Cancer,* 11**,** 481-92.

WOLFE, S. A., NEKLUDOVA, L. & PABO, C. O. 2000. DNA recognition by Cys2His2 zinc finger proteins. *Annu Rev Biophys Biomol Struct,* 29**,** 183-212.

WONG, M. M., COX, L. K. & CHRIVIA, J. C. 2007. The chromatin remodeling protein, SRCAP, is critical for deposition of the histone variant H2A.Z at promoters. *J Biol Chem,* 282**,** 26132-9.

WOO, C. J., KHARCHENKO, P. V., DAHERON, L., PARK, P. J. & KINGSTON, R. E. 2010. A region of the human HOXD cluster that confers polycomb-group responsiveness. *Cell,* 140**,** 99-110.

XHEMALCE, B. & KOUZARIDES, T. 2010. A chromodomain switch mediated by histone H3 Lys 4 acetylation regulates heterochromatin assembly. *Genes Dev,* 24**,** 647-52.

XIAO, B., JING, C., WILSON, J. R., WALKER, P. A., VASISHT, N., KELLY, G., HOWELL, S., TAYLOR, I. A., BLACKBURN, G. M. & GAMBLIN, S. J. 2003. Structure and catalytic mechanism of the human histone methyltransferase SET7/9. *Nature,* 421**,** 652-6.

XIONG, J. W., LEAHY, A., LEE, H. H. & STUHLMANN, H. 1999. Vezf1: A Zn finger transcription factor restricted to endothelial cells and their precursors. *Dev Biol,* 206**,** 123-41.

XU, Z., MENG, X., CAI, Y., LIANG, H., NAGARAJAN, L. & BRANDT, S. J. 2007. Single-stranded DNA-binding proteins regulate the abundance of LIM domain and LIM domain-binding proteins. *Genes Dev,* 21**,** 942-55.

YANG, X., NOUSHMEHR, H., HAN, H., ANDREU-VIEYRA, C., LIANG, G. & JONES, P. A. 2012. Gene reactivation by 5-aza-2'-deoxycytidine-induced demethylation requires SRCAP-mediated H2A.Z insertion to establish nucleosome depleted regions. *PLoS Genet,* 8**,** e1002604.

YE, T., KREBS, A. R., CHOUKRALLAH, M. A., KEIME, C., PLEWNIAK, F., DAVIDSON, I. & TORA, L. 2011. seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic Acids Res,* 39**,** e35.

ZEMACH, A., MCDANIEL, I. E., SILVA, P. & ZILBERMAN, D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science,* 328**,** 916-9.

ZENTNER, G. E. & HENIKOFF, S. 2013. Regulation of nucleosome dynamics by histone modifications. *Nat Struct Mol Biol,* 20**,** 259-66.

ZHANG, H., ROBERTS, D. N. & CAIRNS, B. R. 2005. Genome-wide dynamics of Htz1, a histone H2A variant that poises repressed/basal promoters for activation through histone loss. *Cell,* 123**,** 219-31.

ZHANG, Y., LIU, T., MEYER, C. A., EECKHOUTE, J., JOHNSON, D. S., BERNSTEIN, B. E., NUSBAUM, C., MYERS, R. M., BROWN, M., LI, W. & LIU, X. S. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol,* 9**,** R137.

ZHOU, W., CLOUSTON, D. R., WANG, X., CERRUTI, L., CUNNINGHAM, J. M. & JANE, S. M. 2000. Induction of human fetal globin gene expression by a novel erythroid factor, NF-E4. *Mol Cell Biol,* 20**,** 7662-72.

ZILBERMAN, D., COLEMAN-DERR, D., BALLINGER, T. & HENIKOFF, S. 2008. Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. *Nature,* 456**,** 125-9.

ZOU, Z., OCAYA, P. A., SUN, H., KUHNERT, F. & STUHLMANN, H. 2010. Targeted Vezf1-null mutation impairs vascular structure formation during embryonic stem cell differentiation. *Arterioscler Thromb Vasc Biol,* 30**,** 1378-88.