# Chapter 1

## Introduction

A genome consists of all the genetic material contained in a cell of an organism and contains all of the information necessary for life. In general, this inherited information is encoded in three types of elements - genes, regulatory elements and maintenance elements. Genes contain information to code for proteins while regulatory elements control the spatial and temporal production of proteins by genes. These regulatory elements include promoters, enhancers, insulators and other elements such as non-coding regulatory RNAs. In addition, a further level of gene regulation is achieved by methylation of DNA and modification of chromatin. Maintenance elements contain information for DNA repair, replication and recombination. These elements include centromeres, telomeres, origins of replication and recombination hotspots. The sequencing of the human genome was completed in 2004 (IHGSC, 2004) and the challenge in the post-genome era lies in understanding how the information encoded in the genome sequence regulates both normal development and disease states. Identifying and annotating the human genome for all of the elements mentioned above and characterising the complex events that are associated with these elements will give us an unprecedented understanding of human biology.

### 1.1. Coding regions in the human genome

In essence, a gene is a genomic sequence directly encoding functional product molecules, either RNA or protein (Gerstein *et al*., 2007). In eukaryotes, genes are typically composed of alternating exon and intron sequences. One of the important goals in the post-genome era is to produce a definitive catalogue of genes in the human genome and many computational methods such as GeneWise (Birney and Durbin, 2000), GenomeScan (Yeh *et al*., 2001), GENSCAN (Burge and Karlin, 1997), and TWINSCAN (Flicek *et al*., 2003) have been developed to identify genes by searching for features such as splice site signals and comparing protein-coding regions in different organisms. A small percentage of the human genome sequence is currently known to code for functional products, mainly protein-coding genes (~2.2%) (Frith *et al*., 2005) and a limited number of structural and regulatory RNAs, such as snoRNAs and microRNAs (Mattick, 2007). The current estimate of protein-coding genes is ~20,000 (IHGSC, 2004; Goodstadt and Ponting 2006), but it is possible that the human genome contains a higher number of genes as microarray- based approaches have revealed the existence of many more transcribed sequences of unknown function (Cawley *et al.*, 2004; Rinn *et al.*, 2003; Kapranov *et al.*, 2007; Birney *et al.*, 2007).

## 1.2. Non-coding elements in the human genome

The current estimate of protein-coding genes in the human genome is surprisingly similar to the number of genes present in the nematode worm (~19,000) (Stein *et al.*, 2003), despite the large difference in developmental complexity and genome sizes. Furthermore, a comparison of genome organisation between the major model organisms showed that there was more than a 300 fold increase in genome size between the yeast *S. cerevisiae* and *S. pombe* and humans but only a 4-fold increase in gene number (Figure 1.1). This large increase in genome size and developmental potential is associated with a dramatic increase in non-coding and repetitive sequences (Taft *et al.*, 2007). The genomes of unicellular yeast contain little non-coding DNA compared with the genomes of multi-cellular eukaryotes. The human genome, in particular, contains large amounts of repetitive and non-coding DNA as protein-coding sequences account for only around 2% of the genome sequence. Understanding the function of the remaining 98% of the genome sequence is an important goal. Comparative analysis of the human and mouse genomes established that approximately 5% of the genomic sequences are highly conserved regions of 50-100 base pairs (bp), which is much higher than can be accounted for by protein-coding sequences alone (Waterson *et al.*, 2002). There are also shorter and weaker homologous elements found in the two genomes, some of which contain binding sites for known transcription factors and regulatory proteins, while others have as of yet no known function (Kondrashov and Shabalina, 2002). In addition, there are species-specific functional sequence elements which are not conserved. All of this, taken together, suggests that the percentage of the human genome which contains functional non-coding elements may be even higher than 5%.

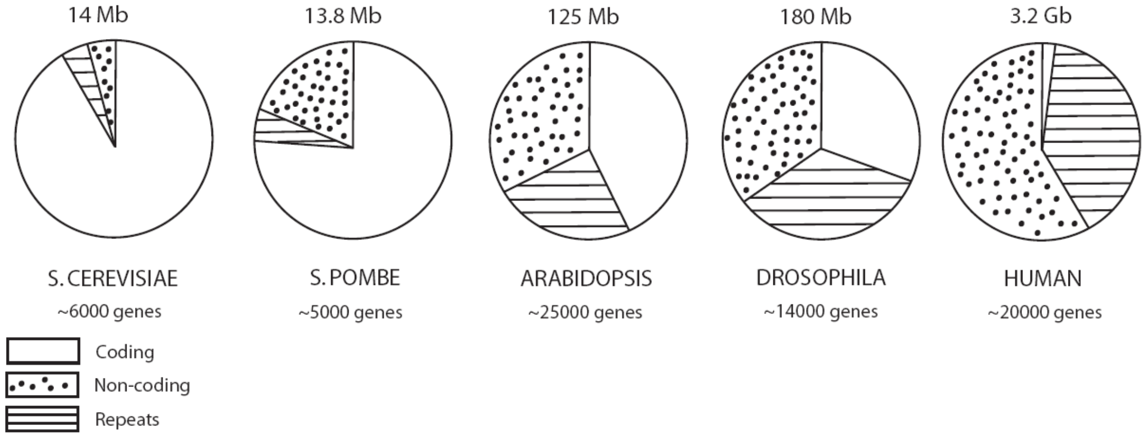

**Figure 1.1: Genome organisation of the major model organisms.** The genome sizes of the major eukaryotic model organisms are indicated above each pie-chart and the approximate number of protein-coding genes is indicated below each pie-chart. The emergence of complex multi-cellular organisms is associated with an increase in genome size. This increase in genome size is due to a large expansion in the amount of non-coding (intronic and intergenic sequences) and repeat DNA (satellite, LINE and SINE elements) sequences.

Non-protein coding sequences include regulatory and maintenance elements. A detailed discussion of regulatory elements controlling gene expression will be presented in later sections of this chapter. Maintenance elements mediate genome structure and dynamics and determine how the genome is passed on to the next generation via repair, replication and recombination. This class of "maintenance" elements includes centromeres and telomeres, origins of DNA replication, and recombination hot spots. Centromeres and telomeres are specialised structures which are involved in replication and are well characterised in terms of sequence content (Eichler and Sankoff, 2003). It is believed that the order of replication is regulated as some regions of the genome replicate earlier than others (Schubeler *et al*., 2002; Woodfine *et al*., 2004) and while the process of DNA replication is well documented, replication origins are not understood at the sequence level (Gilbert *et al*., 2004). A small number of recombination hotspots have been precisely located and as of yet there is no confirmation of sequence similarity (Jeffreys *et al*., 2001; Yauk *et al*., 2003). Thus, in order to fully understand key biological processes as well as developmental complexity, it is crucial to annotate and characterise the coding and non-coding regions of the human genome.

## 1.3. Regulation of gene expression

Multi-cellular organisms such as humans are composed of a multitude of different cell types which all develop from the same genetic template. The development of these numerous cell types is dependent upon the appropriate gene expression patterns being established within individual cells in response to external and internal cues. Some genes are expressed constantly in all cell types, while other genes are only expressed as a cell enters a particular developmental pathway. In order to understand why a particular gene is expressed or not expressed, one must be aware of what features are involved in regulating its activity. Each gene has its own associated *cis*-acting elements, which are sequences that regulate gene expression levels. *Trans*-acting factors, which include transcription factors and other proteins involved in regulating gene expression, are encoded by other genes and bind to *cis*-acting elements to control gene expression. This system allows for the cell to use multiple combinations of regulatory features in order to control gene expression. For example, a trans-acting factor can act on *cis*-elements of multiple genes or it can form a complex with another *trans*-acting factor to act on an individual gene or multiple genes.

## 1.4. Types of non-coding regulatory elements involved in controlling gene expression

Protein-coding genes which are transcribed by eukaryotic RNA polymerase II (Pol II) contain two types of *cis*-acting non-coding regulatory sequences which regulate transcription initiation: (1) a promoter composed of a core promoter and proximal regulatory elements, and (2) distal regulatory sequences, which include enhancers, insulators, silencers/repressors, locus control regions and matrix attachment regions (MARs) (Figure 1.2). These regulatory elements act in a co-ordinated manner and can be located over distances which range from several tens to several hundreds of kilobases. These elements are described below.

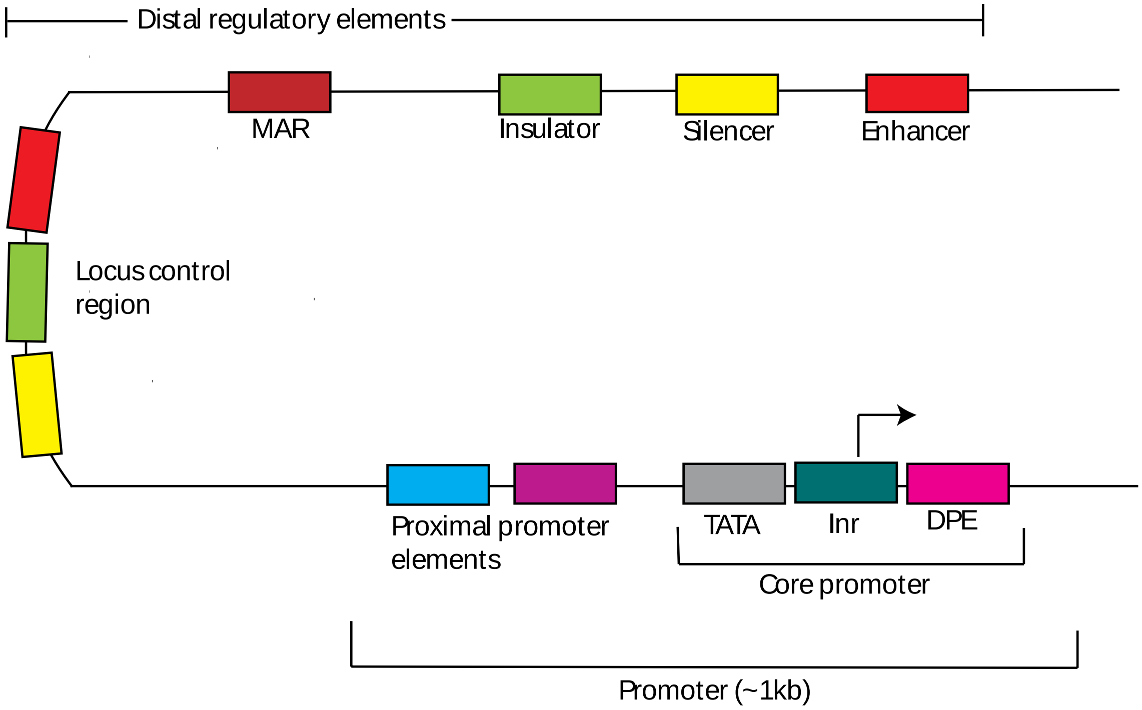

**Figure 1.2: The typical regulatory features involved in the regulation of gene expression.** The promoter is composed of a core promoter which generally contains a TATA box, initiator element (INR) and a downstream promoter element (DPE). Proximal promoter elements are usually located nearby and typically a promoter element spans less than 1 kb. A gene is also associated with several distal elements such as enhancers, silencers/repressors, insulators, and locus control regions (which are composed of several types of regulatory elements). Each element may contain several binding sites for different sequence-specific transcription factors, allowing for combinatorial control of regulation, which increases the number of possible expression patterns. Enhancers, silencers and insulators controlling the expression of a single gene can be located over several tens to several hundreds of kilobases in the human genome, making the identification of all elements controlling the expression of a particular gene a difficult task. DNA methylation at promoter regions can inhibit transcription, histone modifications at promoters, enhancers and within the gene itself can be involved in gene activation and repression. MARs anchor chromatin loops to the nuclear matrix and can shield genes from position effects. Figure adapted from Maston *et al*., 2006.

### 1.4.1. Core and Proximal Promoters

The core RNA polymerase II promoter is a region surrounding a transcription start site (TSS) and is associated with a set of DNA sequence elements (Table 1.1). Combinations of core promoter elements play a crucial role in regulating gene expression patterns (Ren and Maniatis, 1998). The TATA box is an A/T-rich sequence located approximately 25 bps upstream of the TSS.  TATA-binding protein (TBP) recognizes this sequence and begins pre-initiation complex formation. The initiator (Inr) is a pyrimidine rich sequence, which can direct transcription alone or in combination with a TATA box. The downstream promoter element (DPE) lies 28-34 bp downstream of the TSS in TATA-less promoters - it is believed to have a similar function to the TATA box in directing the pre-initiation complex to the TSS (Kadonaga, 2002). Motif ten elements (MTE) function with the Inr to enhance Pol II transcription. The downstream core element (DCE) contains three sub-elements: SI, SII, and SIII. The $TF_{II}B$ recognition element (BRE) can be found upstream (BREu) or downstream (BREd) of the TATA box and can either decrease or increase transcription (Sandelin *et al.*, 2007).

Analysis of the eukaryotic promoter database (EPD) and the database of human transcriptional start sites (DBTSS) found that 22% of human genes contain a TATA box. Of these TATA-containing promoters, 62% contained an Inr, 24% contain a DPE and 12% have a BREu (Gershenzon and Ioshikhes, 2005). Seventy-eight percent (78%) of human promoters contain no TATA box, 45% of which contain an Inr, 28% have a BREu, and 25% possess a DPE. The promoters of housekeeping genes, growth factors and transcription factors often have no TATA box (Zhou and Chiang, 2001).  The proximal promoter region is located approximately 200 bp upstream of the core promoter and contains several binding sites for activators. Approximately 60% of human promoters are found near a CpG island (Venter *et al.*, 2001), which are formally defined as genomic regions of at least 200 bp with a GC percentage greater than 50% and with an observed/expected CpG ratio greater than 60% (Gardiner-Garden and Frommer, 1987). Methylation of cytosine bases in these islands inhibits transcription (Jones *et al.*, 1998). It has been suggested that proximal promoter elements may block the local promoter region from being inappropriately methylated (Maston *et al.*, 2006).

| Core promoter element | Position | Consensus sequence (5'-3') | Interacting Protein |
|---|---|---|---|
| BREu | -38 to -32 | (G/C)(G/C)(G/A)CGCC | $TF_{II}B$ |
| TATA | -31 to -24 | TATA(A/T)A(A/T)(A/G) | TBP |
| BREd | -23 to -17 | (G/A)T(T/G/A)(T/G)(G/T)(T/G)(T/G) | $TF_{II}B$ |
| Inr | -2 to +5 | PyPyAN(T/A)PyPy | TAF1/TAF2 |
| MTE | +18 to +29 | C(G/C)A(A/G)C(G/C)(G/C)AACG(G/C) | n/a |
| DPE | +28 to +34 | (A/G)G(A/T)CGTG | TAF6/TAF9 |
| DCE | three sub-elements:<br>+6 to +11<br>+16 to +21<br>+30 to +34 | Core sequence:<br>$S_I$ CTTC<br>$S_{II}$ CTGT<br>$S_{III}$ AGC | TAF1 |

**Table 1.1: The structure of metozoan core promoters.** Core promoters are composed of a number of distinct elements, which include and upstream and downstream $TF_{II}B$-recognition element (BREu and BREd respectively), a TATA box, an Initiator element (Inr), a motif ten element (MTE), a downstream promoter element (DPE), and a downstream core element (DCE). The DCE does not occur with an MTE and DPE. With the exception of BRE motifs, all other core promoter elements are recognized by TFIID complex members. Table adapted from Thomas and Chiang, 2006.

### 1.4.2. Enhancers

An enhancer element is defined as a DNA sequence which functions in an orientation and distance independent manner to enhance expression of a gene. The first enhancer element, which could dramatically increase gene transcription from a human β-globin gene, was a genomic region of the SV40 virus (Banerji *et al*., 1981). The first human enhancer to be identified was located downstream of the immunoglobulin heavy chain locus and was found to have cell-type specific enhancer activity (Banerji *et al*., 1983). Enhancers may facilitate transcription within a specific tissue or cell type, through the recruitment of tissue-specific activators, which in turn recruit general transcription factors (Szutorisz *et al*., 2005). General transcription factor complexes aid in the binding and function of Pol II at core promoters (see section 1.5.1). Enhancers and proximal promoter elements are very similar in that they both bind activators to enhance transcription but enhancers can be located several hundred kilobases upstream, downstream or even within an intron of their target gene(s). Two models have been proposed to explain how enhancers communicate with promoters to achieve the desired level of gene expression:

i) **Looping/ direct contact model**. This model proposes that enhancers and promoters directly interact. For example, Johnson *et al.* suggest that in the case of the mouse β-globin locus, the tissue specific activators NF-E2 and GATA-1 function in the transfer of Pol II from the enhancer to the β-globin promoter during blood development (Johnson *et al*., 2001; 2002). This transfer occurs by direct physical contact between the enhancer and promoter and occurs over a 50 kb region. Therefore the intervening region of chromatin must be looped out to allow enhancers to directly interact with the promoter. This model is supported by studies carried out at the human and mouse β-globin loci in which promoters and distal enhancers were shown to co-localise within chromatin hubs which also contain RNA polymerase II (Patrinos *et al*., 2004; Osborne *et al*., 2004).

ii) **Tracking/non-contact model**. This model proposes that enhancers act as entry sites for factors that ultimately interact with the promoter (Bulger and Groudine, 1999). The transfer of general transcription factors and Pol II to the promoter occurs by a continuous linear tracking of the complex from the enhancer to the core promoter along the length of the intervening DNA sequence (Kim and Dean, 2004; Ling *et al*., 2004).

### 1.4.3. Silencers/repressors

The opposite of an enhancer is a silencer/ repressor element, which is involved in repressing transcription. In common with enhancers, the majority of identified silencers can function independently of direction and distance. Silencer elements can be located at the proximal promoter of a target gene, within an intron, 3' untranslated region, or as part of a distal enhancer (Ogbourne and Antalis, 1998).

Silencers function as binding sites for sequence-specific transcription factors known as repressors. For example, a 21-bp DNA repressor/silencer element known as the repressor element 1 (RE-1) is found at approximately two thousand sites in the human genome (Bruce *et al.*, 2004). The repressor element-1 silencing transcription factor (REST) binds to RE-1 sites and acts as a transcriptional repressor by blocking the expression of many neuronal RE-1 containing genes in non-neuronal cells. REST recruits co-repressors via its repression domains at the amino and carboxy termini (Figure 1.3). These co-factors then facilitate the creation of a repressive chromatin state through histone deacteylation (Huang *et al.*, 1999), chromatin remodelling (Battaglioli *et al.*, 2002) and DNA methylation (Lunyak *et al.*, 2002).

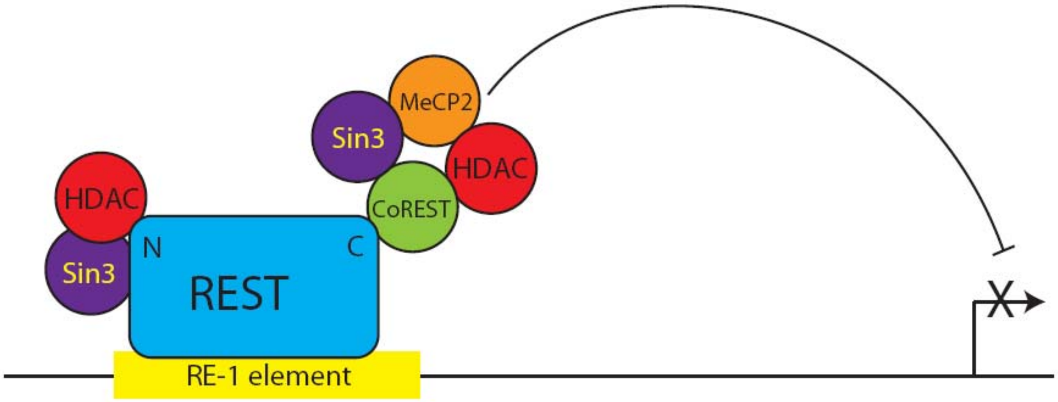

**Figure 1.3: REST-mediated repression of target genes.** The amino terminus (N) of REST recruits a Sin3-histone deacetylase (HDAC) complex to repress neuronal genes in non-neuronal cells. The carboxy terminus (C) recruits many complexes via the co-repressor CoREST, which can include Sin3-HDAC, the methyl-DNA binding protein MeCP2, SWI/SNF chromatin remodelling complex (not shown), histone H3K9 methyltransferase (not shown) and histone demethylase LSD1 (not shown) to mediate epigenetic silencing.

Other repressors work by different mechanisms, some function by blocking the binding of a nearby activator (Harris *et al.*, 2005), by directly competing for the same binding site (Li *et al.*, 2004) or by inhibiting Pol II pre-initiation complex formation (Kim *et al.*, 2003).

### 1.4.4. Insulators

There are many examples of regions in the human genome (and other eukaryotic genomes) which contain a cluster of genes that are actively expressed in a particular cell type and in close proximity to another group of genes which are not expressed in that cell type (Sproul *et al.*, 2005). Alternatively, an expressed gene may be located in a region of constitutively silent chromatin. In both situations, the maintenance of appropriate expression patterns relies upon a class of DNA sequence elements known as insulators (Gaszner and Felsenfeld, 2006). Insulators that prevent inappropriate promoter activation by enhancers have been termed enhancer blocking insulators while those which prevent the spread of silent chromatin are known as barrier insulators (Sun and Elgin, 1999) (Figure 1.4).

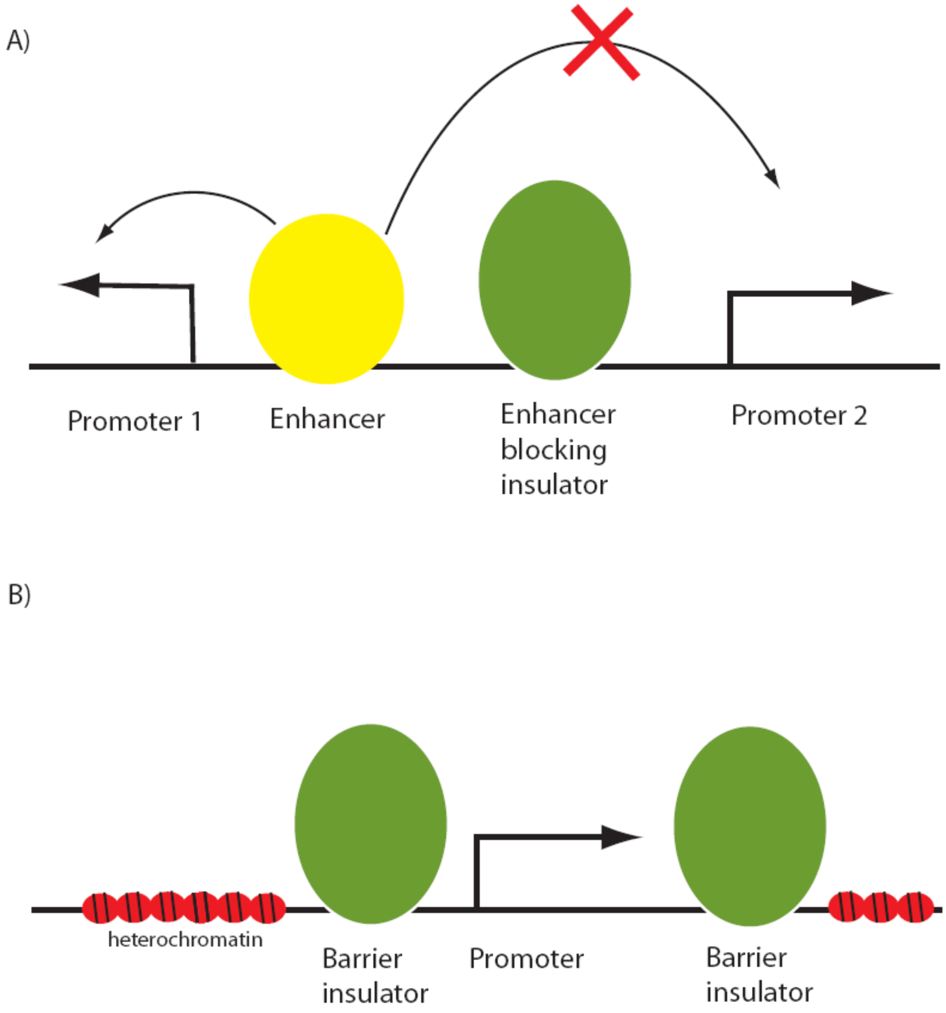

**Figure 1.4: The two types of insulator activity**. A) An enhancer-blocking insulator interferes with enhancer-promoter communication only when positioned between an enhancer and a promoter as is the case for promoter 2. In other situations (i.e. promoter 1), enhancer-promoter communication is not blocked. B) A barrier insulator prevents the spread of heterochromatin into a euchromatin region when placed at a junction between the two regions.

### 1.4.4.1 Enhancer blocking insulators

Enhancer blocking elements interfere with enhancer-promoter interactions only when located between these two elements. They function to prevent an enhancer from incorrectly activating a promoter. It has been suggested that insulators perform this function by tethering the chromatin fibre to the nuclear matrix. This may result in the formation of chromatin loops which prevent an enhancer and a promoter from communicating with each other due to their location in separate loop domains (Gaszner and Felsenfeld, 2006).

The initial work in identifying enhancer blocking elements was performed in *Drosophila* with the gypsy retro-transposon (Geyer *et al*., 1986). Insertion of this element within enhancers at the yellow locus blocked the action of those enhancers located distal to the insertion site but did not affect enhancers located proximal to the promoter (Geyer *et al*., 1986). The enhancer blocking activity of gypsy was mapped to 12 binding sites for Suppresor of Hairy Wing (Su(Hw)), which interacts with Topoisomerase-I-interacting protein (Topors). Topors can then bind the nuclear lamina (Capelson and Corces, 2005), which is consistent with the idea that chromatin loop formation is the important event in mediating enhancer-blocking activity. The first vertebrate enhancer blocking insulator to be identified, 5'HS4, was located at the 5' end of the chicken β-globin locus (Figure 1.5) (Chung *et al*., 1993; Pikaart *et al*., 1998). This complex element possessed both enhancer blocking and barrier insulator activity (discussed in section 1.4.4.2). The enhancer blocking activity was attributed to the binding of the 11 zinc finger transcription factor CTCF (Bell *et al*., 1999). CTCF can interact with itself and the nucleolar protein nucleophosmin, which may lead to the formation of discrete chromatin domains (Yusufzai *et al*., 2004). Furthermore, CTCF has been shown to mediate long range chromatin looping at the mouse β-globin locus (Splinter *et al*., 2006) and inter-chromosomal co-localisation between the Igf2/H19 and Wsb1/Nf1 loci (Ling *et al*., 2006). A more detailed discussion on CTCF is presented in section 1.4.4.3 and Chapter 4 (section 4.1).

#### 1.4.4.2 Barrier Insulators

Barrier insulators guard against position-effect variegation (PEV), which is the silencing of a euchromatic gene due to the spread of heterochromatin. Heterochromatin is more condensed than euchromatin and is associated with methylation of histone H3 lysine 9 and lysine 27 (see section 1.6.2). Heterochromatic DNA is also associated with extensive CpG methylation. Heterochromatin formation begins with the methylation of H3K9 at an initiation site, which then recruits heterochromatin protein 1 (HP1) (Grewal and Jia, 2007). HP1 can in turn recruit histone H3K9 methyltransferase activity and this cycle of events leads to the spread of heterochromatin. Euchromatin represents a less condensed form of chromatin, which is located at transcriptionally active regions of the genome and is associated with a number of histone modifications such as histone H3 acetylation and H3 K4 methylation (see section 1.6.2).

As discussed in the previous section, the chicken 5'HS4 element also displays barrier insulator activity (Pickaart *et al*., 1998), which was found to be independent of CTCF binding (Recillas-Targa *et al*., 2002). The 5' HS4 is located between a region of condensed inactive chromatin in chicken erythrocytes and the active β-globin chromatin domain (Figure 1.5). The 5' HS4 is marked by peaks of euchromatic histone modifications (Litt *et al*., 2001). These are due to the recruitment of histone acetyltransferases (HAT) and histone methyltransferases (HMT) by upstream stimulatory factor 1 (USF1) and USF2 (West *et al*., 2004; Huang *et al*., 2007). Disruption of USF1 and USF2 binding abolishes HAT and HMT recruitment along with barrier activity. CTCF binding sites have also been found close to the transition between active and silent chromatin domains in mouse and human cell types (Filippova *et al*., 2005; Barski *et al*., 2007). Although it has not been directly shown that CTCF prevents the spread of heterochromatin, it cannot be ruled out that CTCF may also have barrier activity. This hypothesis would reconcile with the chromatin loop model of enhancer blocking discussed above as flanking a gene with a CTCF binding site would provide barrier activity by creating an independent expression domain. Support for barrier insulator's functioning through the formation of chromatin loops has been provided by a study of barriers elements in yeast (Ishii *et al*., 2002). This study showed that barrier activity was linked to the anchoring of chromatin fibres to the nuclear pore.

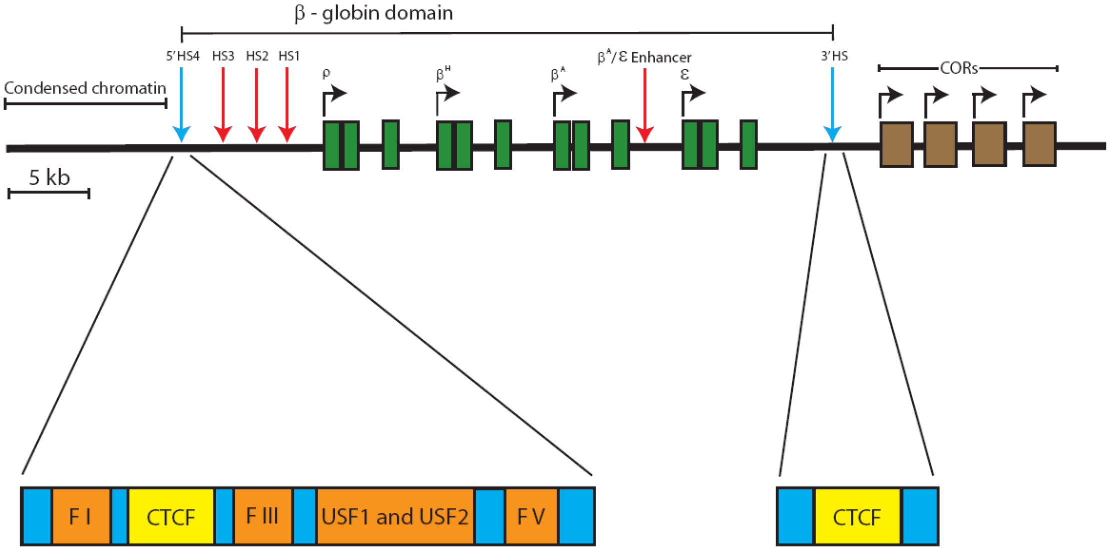

**Figure 1.5: Genomic organisation of the chicken β-globin locus**. The 5'HS4 and 3'HS insulator elements (indicated by blue arrows) define the chicken β-globin chromatin domain, which contains the developmentally regulated β-globin gene cluster and its locus control region (LCR), composed of HS1-3 and the $\beta^A$/ ε enhancers. The β-globin domain is flanked by a region of condensed chromatin at the 5' end and chicken olfactory receptor genes (CORs) at the 3' end. 5' HS4 displays both enhancer-blocking activity (mediated by CTCF) and barrier activity (mediated by USF1, USF2 and as yet uncharacterised footprint I, III and V binding proteins). 3'HS binds CTCF and displays only enhancer-blocking activity. Figure adapted from Gaszner and Felsenfeld, 2006.

### 1.4.4.3 CTCF is a multifunctional protein

Following the discovery that CTCF was responsible for the enhancer-blocking activity of the chicken β-globin locus insulators, there has been a great deal of interest in identifying other CTCF binding sites and understanding how this factor functions. CTCF enhancer blocking activity was subsequently identified in the human and mouse Igf2/H19 imprinted locus (Bell and Felsenfeld, 2000; Hark *et al*., 2000). During embryonic development Igf2 is expressed from the paternal allele only and is under the control of the imprinted control region (ICR) lying between the enhancer and Igf2 promoter. The ICR contains CTCF insulator elements, the CpG islands of which are methylated in the paternal allele. Methylation of the CpG islands abolishes CTCF binding and therefore the enhancer blocking activity of the ICR.

In addition to its role as an insulator binding protein, CTCF has been shown to function as a classical transcription factor, capable of functioning as a transcriptional repressor (Filippova *et al*., 1996; Burcin *et al*., 1997) and a transcriptional activator (Vostrov and Quitschke, 1997). The diversity of functions that CTCF performs may be attributed to its protein structure. CTCF contains 11 zinc finger domains and different combinations of zinc fingers can engage different binding sequences (Filippova *et al*., 1996). Two recent genome-wide studies have employed different approaches to identify CTCF binding sites in the human genome. Ren and colleagues used chromatin immunoprecipitation in combination with microarrays (section 1.7.3.5) to identify CTCF binding sites in human fibroblasts (Kim *et al*., 2007), while Lander and colleagues used a computational approach to identify conserved motifs and one of these motifs was able to bind CTCF (Xie *et al*., 2007). These studies identified approximately 15,000 CTCF binding sites in the human genome. The experimental identification of a large number of CTCF binding sites by Ren and colleagues allowed for a consensus CTCF binding sequence to be determined. A 20 bp consensus motif was described which refined a previously described motif (Bell and Felsenfeld, 2000) but a number of nucleotides at several sites in the consensus sequence were ambiguous suggesting that variation within the consensus sequence is responsible for the functional versatility of CTCF.

A number of factors have been shown to interact with CTCF (Wallace and Felsenfeld, 2007) and the variety of CTCF functions is likely to reflect the diversity of these interacting partners. CTCF interacts with DNA binding proteins, histone interacting proteins, histone themselves and other regulatory factors. For example the chromodomain helicase family member CHD8 interacts with CTCF and has been implicated in insulator function (Ishihara *et al*., 2006). CHD8 forms complexes with histone modifying enzymes (Dou *et al*., 2005) and may recruit these enzymes to CTCF bound insulators.

CTCF function may be regulated through the choice of zinc fingers used in DNA binding. Furthermore, CTCF can be chemically modified which also affects its functional properties. It can be poly (ADP-ribosyl)ated, and inhibiting this modification can impair its insulator function (Yu *et al*., 2004). Poly (ADP-ribosyl)ation has been implicated in maintaining DNA hypomethylation (Zardo and Caiafa, 1998) in the genome and CTCF binding is sensitive to DNA methylation.

### 1.4.5. Locus control regions

Locus control regions (LCR) consist of a cluster of regulatory element involved in regulating an entire locus (Li *et al*., 2002). The LCR is usually composed of several *cis*-acting elements such as enhancers, insulators, silencers and nuclear matrix attachment regions (MARs) (see section 1.4.6). These elements are bound by various transcription factors and chromatin modifiers, which regulate spatial and temporal gene expression patterns. LCRs can regulate transcription from large distances in a position independent way and can be found upstream of their target locus, within an intron of a target gene (Aronow *et al*., 1992) or neighbouring gene (Adlam and Siu, 2003) or downstream of their regulated locus (Lang *et al*., 1991).

The first LCRs to be identified were the human and mouse β-globin LCR (reviewed by Chakalova *et al*., 2005). The human β-globin LCR is located ~6-25kb upstream of the locus and controls five genes that are expressed during different stages of erythrocyte development. The orientation of the LCR is critical as reversing its direction destroys most of its activity (Tanimoto *et al*., 1999). LCRs are believed to function in a similar way to the enhancer looping model discussed above as long range interactions between DNase I hypersensitive sites have been observed at the active β-globin locus (Tolhuis *et al*., 2002), which tethers these sequences into an active chromatin hub. Critically, these long range interactions are only observed when the locus is active.

### 1.4.6. Matrix Attachment regions (MARs)

The nuclear matrix refers to a structure obtained from nuclei that is resistant to extraction by high concentrations of NaCl or the non-ionic detergent-like salt lithium diiodosalicyclate (LIS) (Mirkovitch *et al*., 1984). The DNA fragments bound to this residual structure are known as MARs and serve to anchor chromatin loops to the nuclear matrix. While it is thought that there is a structural role for MARs (i.e. they mediate binding of chromatin to the nuclear matrix), biological functions have been proposed for MARs such as transcription activation as a result of chromatin remodelling (Bode *et al*., 2000) and insulation of genes from position-mediated silencing of transgenes (Allen *et al*., 2000). MARs have also been shown to antagonise DNA methylation-dependent repression of long range enhancer-promoter interactions (Forrester *et al*., 1999), and to extend an enhancer mediated region of open chromatin (Jenuwein *et al.*, 1997). CTCF has been shown to interact with the nuclear matrix (Dunn *et al.*, 2003), suggesting that CTCF may be responsible for anchoring chromatin to the nuclear matrix. Recillas-Targa and colleagues have shown that an insulator 5' upstream of the chicken α-globin gene domain co-localises with a MAR element, which binds CTCF (Valadez-Graham *et al.*, 2004). The insulator activity of this element is dependent on CTCF binding, linking MARs and CTCF once more with a role in demarcating chromosomal boundaries.

## 1.5. Proteins involved in transcriptional regulation

The proteins which bind to these regulatory elements discussed above are numerous. There may be as many as 2000 transcription factors in the human genome (Lander *et al.*, 2001) which are divided into three classes of proteins, namely, general members of the RNA polymerase complexes, sequence-specific DNA-binding proteins that mediate activation or repression of transcription and chromatin remodelling and modification complexes.

### 1.5.1. RNA polymerase complexes

At the promoter, interactions between RNA polymerase II and DNA lead to transcription initiation. Several general transcription factors (GTFs) are necessary for recognition and stable binding at the promoter (Thomas and Chiang, 2006). These GTFs were designated as $TF_{II}$ A-H (TF for transcription factor, II represents Pol II transcription and the letter refers to the nuclear extract from which the GTF was isolated). These GTFs function collectively to specify transcription start sites (TSSs). Pre-initiation complex formation begins with the binding of $TF_{II}D$ to a core promoter element such as a TATA box or Inr element followed by stepwise recruitment of the other GTFs or alternatively by recruitment of a pre-assembled Pol II holoenzyme.

$TF_{II}D$ is a complex of TBP and at least 14 TBP-associated factors (TAFs). TBP and some TAFs bind different core promoter elements, allowing $TF_{II}D$ to recognize TATA containing and TATA-less promoters. The $TF_{II}D$ complex is also associated with enzymes that post-translationally modify chromatin and other proteins involved in gene regulation. TAF1 in the $TF_{II}D$ complex is a histone acetyltransferase (HAT), which methylates histones H3 and H4 (Mizzen *et al*., 1996). $TF_{II}A$ stabilizes TBP-TATA box interactions through direct contact with TBP (Geiger *et al*., 1996). The binding of $TF_{II}B$ to TBP stabilizes $TF_{II}D$ /TBP binding to the promoter (Orphanides *et al*., 1996). $TF_{II}B$ also recruits Pol II/$TF_{II}F$ to the promoter bound $TF_{II}D$-$TF_{II}B$ complex. $TF_{II}F$ enhances the affinity of Pol II for the promoter complex and is necessary for the recruitment of $TF_{II}E$ and $TF_{II}H$ (Orphanides *et al*., 1996). $TF_{II}F$ also facilitates Pol II promoter escape (Yan *et al*., 1999). $TF_{II}E$ stimulates the ATPase, C-terminal domain (CTD) kinase, and DNA helicase activities of $TF_{II}H$ (Lee and Young, 2000), which are essential for transcription initiation and elongation.

Pol II itself is composed of 12 subunits (RPB1-12) (Young, 1991). RPB1 and RPB2 are the key catalytic units and are responsible for phosphodiester bond formation (Hampsey, 1998). The CTD of RPB1 contains tandem heptapeptide repeats of Tyr-Ser-Pro-Thr-Ser-Pro-Ser (YSPTSPS) (Prelich, 2002). The CTD can be phosphorylated at Ser2 and Ser5 and hyperphosphorylated Pol II is associated with transcription elongation (Sims *et al*., 2004). Ser5 phosphorylation levels peak early during transcript production, whereas Ser2 predominates in the middle and later stages of transcript production (Komarnitsky *et al*., 2000).

### 1.5.2. Sequence-specific transcription factors

This class includes the largest and most diverse group of factors responsible for transcription initiation, enhancement and inhibition. As discussed before, as many as 10% of the genes in the human genome - approximately 2000 genes - may encode transcription factors (TFs). Specificity of transcription is achieved by combinatorial binding of these factors to sequence-specific *cis*-regulatory elements. TFs alter transcription by interacting directly or indirectly with RNA polymerase as described in the previous section. Therefore, TFs must possess certain structural features which allow them to modulate gene expression. Currently, more than 100 different DNA-binding domains have been identified in transcription factors (Kummerfeld and Teichmann, 2006), some of which are discussed below.

The homeobox or homeodomain was originally identified in the homeotic genes of *Drosophila* (Affolter *et al*., 1994) and found to contain a DNA-binding motif called a helix-turn-helix motif. The homeobox domain is conserved in mammalian Hox gene clusters. The POU family of TFs (named after the transcription factors Pit, Oct and Unc) was subsequently found to contain both a homeobox sequence and a conserved POU-specific domain. Both domains are needed for high-specificity DNA-binding. The TFs containing these domains play important roles in development.

The leucine zipper motif was identified in several different transcription factors such as C/EBP, yeast GCN4 and the oncogenes Myc, Fos and Jun (Bosher *et al*., 1996). In this motif every seventh amino acid is a leucine, whereby leucine residues occur every two turns on the same side of a helix, allowing dimerization of TFs. Dimerization also correctly positions the nearby basic DNA-binding domains, which is necessary for DNA binding to occur.  The basic DNA-binding domain has also been identified in other transcription factors which do not contain a leucine zipper (Prendergast and Ziff, 1989). These TFs include E12 and E47 factors which are important for the development of B lymphocytes. This class of TFs also contain a different type of helix-turn-helix motif, which functions similarly to a leucine zipper. It facilitates dimerization of TFs, which in turn allows for DNA-binding by the basic motif (Jones, 1990).

In addition to these DNA-binding domains, TFs also possess distinct transcriptional activation domains (Stephanou *et al*., 2002). Analysis of many different activation domains identified three distinct categories, acidic domains, glutamine-rich domains, and proline-rich domains. Acidic activation domains contain a large number of acidic amino acids which produces a strong negative charge. The negative charge allows long range electrostatic interactions with members of the $TF_{II}D$ complex (Uesugi *et al*., 1997). Glutamine-rich activation domains have been found in TFs such as Sp1, Oct-1, and Oct-2. The  activation domain of CTF/N1, which binds to the CCAAT box motif, is proline rich and has been found in other TFs such as Jun and AP2. In summary, many different DNA binding and activation motifs exist and reflect the diversity of DNA sequences which TFs bind and the proteins that they interact with.

### 1.5.3. Proteins involved in chromatin remodelling and modification

Sequence-specific TFs need to gain access to the DNA template to initiate transcription. However, the DNA template is normally folded into a compact chromatin fibre which must be unfolded or remodelled to allow transcription to occur. (De la Serna *et al*., 2006). A description of chromatin structure is presented in section 1.6. The two classes of proteins which regulate accessibility of the DNA template – ATP-dependent chromatin remodelling enzymes and histone modifying enzymes - are described below.

#### 1.5.3.1. ATP-dependent chromatin remodelling enzymes

ATP-dependent chromatin remodelling enzymes use energy from ATP hydrolysis to remodel nucleosomes (Narlikar *et al*., 2002). Remodelling enzymes disrupt histone-DNA interactions, i.e. they promote nucleosome 'sliding', allowing transcription factors to access the DNA (Becker, 2002). They can also reposition the DNA so that it is accessible on the surface of the histone octamer (Aoyagi *et al*., 2002). In addition to these activities, remodelling complexes can transfer a histone octamer from one DNA template to another (Whitehouse *et al*., 1999) and can cause changes in super-helicity by twisting the DNA which disrupts histone-DNA interactions (Gavin *et al*., 2001).

There are three classes of ATP-dependent remodellers, the SWI/SNF (Switch/Sucrose non-fermentable), CHD (chromodomain and helicase-like domain) and ISWI (imitation SWI) families (De la Serna *et al*., 2006). The SWI/SNF family members contain a bromo domain, which binds acetylated histones (Hassan *et al*., 2002). The CHD family members have two chromodomains, which bind methylated histone tails (Bannister *et al*., 2001; Lachner *et al*., 2001; Flanagan *et al*., 2005). The ISWI family contains a SANT domain, which acts as a histone-binding domain (Boyer *et al*., 2004). Each class of enzymes forms complexes with other proteins, for example ISWI SNF2H enzymes are found in the ACF (ATP-utilizing chromatin assembly and remodelling factor) complex as well as the RSF (remodelling and spacing factor) complex. SNF2L is found as part of the NuRF (nucleosome remodelling factor) and CERF (CECR2-containing remodelling factor) complexes.

#### 1.5.3.2. Histone modifying enzymes

The other class of proteins involved in chromatin remodelling are the histone modifying enzymes. Unlike the ATP-dependent chromatin remodelling enzymes, which expose the underlying DNA by promoting nucleosome movement, histone modifying enzymes influence transcription by covalently modifying amino acid residues located in the N-terminal 'tail' and the core of histones (Kouzarides, 2007). These enzymes modify specific amino acids by adding or removing various chemical groups. The covalent modifications include acetylation, methylation, phosphorylation, ubiquitination, sumoylation and ADP-ribosylation.

Acetylation of histones was first proposed to be involved in activation of transcription over 40 years ago (Allfrey *et al*., 1964), but it wasn't until 1995 that the first histone acetyltransferase (HAT) was identified (Kleff *et al*., 1995). Since then, a large number of HATs have been characterised (Reid *et al*., 2000). HATs are divided into three main families, GNAT, MYST, and CBP/p300 (Sterner and Berger, 2000). HATs function as part of large complexes *in vivo* and different complexes are involved in distinct biological processes (Roth *et al*., 2001). These different complexes contain specific non-acetyltransferase components which interact with different sequence specific activators, targeting the complex to distinct genes. Consistently, given that histone acetylation can create a more open chromatin structure, many transcriptional co-activators, such as Gcn5/PCAF, CBP/p300 and SRC-1, have been shown to possess intrinsic HAT activity. Similar to transcriptional co-activators possessing HAT activity, many transcriptional co-repressor complexes, such as mSin3a and NURD/Mi-2, contain subunits with HDAC activity (Shahbazian and Grunstein, 2007). However, the Rpd3 small complex (Rpd3S) is an HDAC-containing complex that associates with the actively transcribing, elongating form of RNA polymerase II. Through this association, Rpd3S has been implicated in preventing inappropriate initiation within the protein-coding region of actively transcribed genes (Keogh *et al*., 2005).

Histone deacetylase (HDACs) complexes remove acetyl groups (Kurdistani and Grunstein, 2003). Deacetylation correlates with transcriptional repression and there are three catalytic groups of HDACs which are conserved from yeast to human - type I, II , and III (Narlikar *et al*., 2002). The type I family include HDACs 1, 2, 3, and 8, while type

II includes HDACs 4, 5, 6, 7, 9 and 10. These two types of enzymes have a similar mechanism of deacetylation which does not involve a co-factor. The type III family of Sir2 related enzymes require the co-factor nicotinamide adenine dinucleotide (NAD) as part of their catalytic mechanism.

Histone methyltransferases are responsible for catalysing the methylation of lysine and arginine residues in histones. Methylation modifications can result in either activation or repression of transcription (Bannister and Kouzarides, 2005). All methyltransferases which modify lysine residues, with the exception of Dot1, contain a SET domain (Marmorstein, 2003), which was named after the *Drosophila* chromatin proteins Su(var)3-9, Enhancer of zeste [E(z)], and Trithorax, in which it was first identified. Dot1 is responsible for methylating lysine 79 of histone H3, which is located in the core of the histone (Feng *et al*., 2002). These enzymes are conserved from yeast to man, for example the first H3K4 methyltransferase to be identified was yeast Set1 (Briggs *et al*., 2001; Roguev *et al*., 2001) and human MLL is highly related to yeast Set1 and is also a H3K4 specific methyltransferase (Yokoyama *et al*., 2004). Arginine methylation is catalysed by the protein arginine methyltransferases family 1 (PRMT1) of proteins (Lee *et al*., 2005).

The first histone lysine demethylase LSD1/BHC110 was identified and characterised recently (Shi *et al*., 2004; Lee et al., 2005 b). This enzyme specifically demethylates histone H3K4 and subsequently the jumonji C class of demethylases were identified which could demethylate histone H3 K4, H3 K9, H3 K27, and H3 K36 residues (Liang *et al*., 2007; Klose *et al*., 2007; Lee et al., 2007; Yamane *et al*., 2007; Secombe *et al*., 2007; Christensen *et al*., 2007; Cloos *et al*., 2006; Lan *et al*., 2007; Tsukada *et al*., 2006; Whetstine *et al*., 2006; Yamane *et al*., 2006; Hong *et al*., 2007) (section 1.6.2). While no enzymes have been identified which can reverse arginine methylation, the human enzyme peptidylarginine deiminase (PAD4/PADI4) can catalyze the conversion of an arginine residue to citrulline, antagonizing the effect of arginine methylation as citrulline prevents arginine residues from being methylated (Wang *et al*., 2004; Cuthbert *et al*., 2004).

RSK2 (Sassone-Corsi *et al*., 1999) and MSK1 (Thomson *et al*., 1999) have been identified as the mammalian kinases which perform histone H3 phosphorylation. ADP-ribosylation can be mono, catalyzed by mono-ADP-ribosyltransferases (MARTs), or poly, catalyzed by poly-ADP-ribosyltransferases (PARTs) (Hassa *et al*., 2006). Histones are mono-ubiquitynated by ubiquitin-conjugating enzymes such as the Bmi/Ring1A protein (Wang *et al*., 2004).

## 1.6. Epigenetic regulation of transcription

Conrad Waddington first coined the term 'epigenetics' in 1942 to mean "the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being" (Waddington, 1942). This definition has evolved over the years and now epigenetics is defined as changes to gene function that occur in the absence of changes to the underlying DNA sequence. Modern epigenetic research is focused on the study of covalent and non-covalent modifications of histones and DNA and how these modifications influence overall chromatin structure, gene expression and replication. The four core histone proteins which make up nucleosomes (section 1.6.1) can be modified by more than 100 different post-translational modifications (section 1.6.2). These occur mainly at specific amino acids on the N-terminal tail and recent years have seen a great increase in our understanding of these modifications. Vertebrate DNA methylation occurs almost exclusively at CpG dinucleotides (Bird, 2002) (section 1.6.3) and histone modifying proteins may be involved in directing DNA methylation to promoters (Vire *et al*., 2006).

### 1.6.1. The Nucleosome – the fundamental repeating unit of chromatin

A typical mammalian nucleus is 11-22 μm in diameter, into which two meters of DNA is packaged. The DNA is packaged in a highly ordered manner; the first level of compaction is achieved by wrapping DNA around the histone core proteins to produce a structure called the nucleosome, the basic unit of the chromatin fibre (Kornberg, 1974). Interactions between individual nucleosomes drive the folding of a nucleosomal array (11nm in diameter) into a secondary fibre, 30nm in diameter, which are then further condensed into large structures that form chromosomes.

Nucleosomes are arranged like 'beads on string' along the chromatin fibre (Kornberg and Thomas, 1974) and a typical nucleosome consists of approximately 200 bp of DNA wrapped around a histone octamer. Each histone octamer is composed of two copies of the core histone proteins H2A, H2B, H3 and H4, which wraps 146 bp of DNA in 1.7

superhelical turns (Luger *et al*., 1997), while approximately 60 bp DNA forms a linker between adjacent octamers. Nucleosomes are synthesized in an ordered manner. Firstly two heterodimers of H3 and H4 are deposited onto the DNA to form a $(H3/H4)_2$ tetramer. Then two H2A-H2B heterodimers bind on either side of the tetramer to form the octamer. Histone H1 interacts with the linker DNA and is involved in higher order folding of the chromatin fibre (Khorasanizadeh, 2004). In the nucleosome core there are 14 contact points between the histones and DNA (Luger *et al*., 1997), making nucleosomes one of the most stable protein-DNA complexes known.

Histones have N-terminal tails, which protrude from the octamer and are subject to numerous post-translational modifications, which have been implicated in a number of processes such as transcriptional activation/silencing, DNA replication and repair, and chromatin assembly (see section 1.6.2). Recent genome-wide studies in yeast (*S. cerevisiae*) have reported that nucleosomes are depleted from active regulatory elements (Lee *et al*., 2004; Pokholok *et al*., 2005; Yuan *et al*., 2005) and the chicken β-globin 5'-HS4 has been shown to be depleted of nucleosomes (Zhao *et al*., 2006). Histone replacement has been reported to mark the boundaries of *cis*-regulatory domains in *D. melanogaster* (Mito *et al*., 2007) and low nucleosome density has been observed in the vicinity of transcription start sites in human cells (Nishida *et al*., 2006; Heintzman *et al*., 2007). These numerous observations lend further support to the idea that nucleosomes are removed or moved along the chromatin fibre by chromatin remodellers to expose the underlying DNA.

The core histone proteins are expressed during the S phase and are involved in the packaging of newly synthesized DNA and were once believed to be the common components of all nucleosomes (Kornberg and Lorch, 1999). However variant forms of these histones have been identified (Kamakaka and Biggins, 2005) (Table 1.2). The chromatin fibre can be modified by the incorporation of these variants, whose expression is not restricted to the S phase. Histone variants are distinguished by amino acid sequence differences and their replication-independent deposition is important for transcriptional regulation and epigenetic maintenance. Histone H2A has the largest number of variants, which include H2A.Z and H2A.X, found in the majority of eukaryotes, and MacroH2A and H2A.Bbd, which are only found in vertebrates. In yeast, H2A.Z prevents the spread of heterochromatin (Meneghini *et al*., 2003) and can be incorporated into a nucleosome by ATP-dependent histone exchange (Mizuguchi *et al*., 2004) or a replication-independent chaperone, Nap1, can facilitate its deposition (Park *et al*., 2005). H2A.Bbd is excluded from the inactive X chromosome (Chadwick and Willard, 2001) and has been shown to confer lower stability to the nucleosome (Gautier *et al*., 2004). MacroH2A is concentrated on the inactive X chromosome (Costanzi *et al*., 2000) and interferes with transcription factor binding and SWI/SNF nucleosome remodelling (Angelov *et al*., 2003). H2A.X is phosphorylated in response to DNA double strand breaks and is recognized by the proteins involved in DNA repair (Celeste *et al*., 2003). The H3 variant H3.3 is found in transcriptionally active chromatin (Ahmad  and Henikoff, 2002; Chow *et al*., 2005). It has been suggested that replication-independent deposition and inheritance of H3.3 in regulatory regions preserves transcriptionally active chromatin (Mito *et al*., 2005). The CenH3 variant is involved in the assembly of centromeric chromatin (Ahmad and Henikoff, 2001).

| Histone | Variants | Role(s) | Localization | Structural features | Function(s) |
|---|---|---|---|---|---|
| H2A | macroH2A | Inactivation of X chromosome | Inactive X chromosome | C-terminal non-histone-like region responsible for most functions | Repressing transcription initiation, Interferes with acetylation by p300. Blocks sliding by ACF and remodelling by Swi/Snf |
| | H2A.X | Repression | General Distribution | Conserved C-terminal SQ(E/D) motif is phosphorylated upon DNA damage | |
| | H2A.Z | Transcription activation/ repression | Promoter, Hetero-chromatin boundary | C-terminal α-helix is essential for recognition | Facilitates TBP binding, is evicted upon transcription activation; Prevents elongation associated modification and remodelling at promoter |
| | H2A.Bbd | Transcription activation | Active X-chromosome and autosomes | Lack of c-terminal; Wraps 118-130 bp DNA around it. | p300 and Gal4-VP16 activated transcription is more robust on H2A.Bbd nucleosomes |
| H3 | H3.3 | Activation of transcription | Transcribed regions | Differs from H3 at only four amino acids | Transcription triggers deposition and removal |
| | CenH3 | Organization of centromeric chromatin | Centromeres | Divergent N-terminal tails | |

**Table 1.2: Histone variants and their functions**. The incorporation of histone variants into chromatin impacts on transcriptional regulation in various ways as described in the text. Adapted from Li *et al*., 2007.

### 1.6.2. Histone modifications and their functions

As discussed previously, several different types of post-translational modifications have been identified on histones. There are over 60 different amino acids on histones where modifications have been detected and lysine and arginines can be methylated in one of three different forms, resulting in more than 100 different post-translational modifications. The majority of these modifications occur on the exposed N-terminal 'tails', some of which are shown in Figure 1.6. These modifications are discussed in detail below.

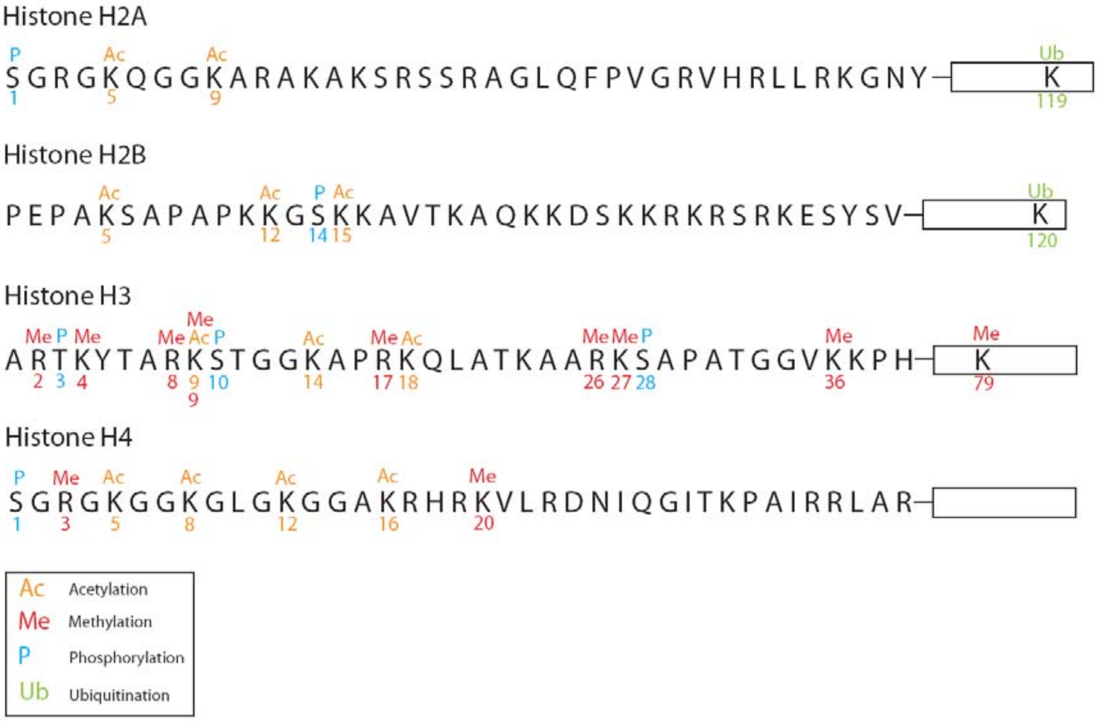

**Figure 1.6: Sites of histone modifications**. The amino terminal 'tails' of histones H2A, H2B, H3, and H4 host the vast majority of covalent histone modifications. Modifications can also occur in the globular core of histones (indicated by boxed regions). The location of acetylation (Ac), methylation (Me), phosphorylation (P) and ubiquitination (Ub) modifications are indicated above the relevant numbered amino acid residues.

The presence of these covalent modifications alters the arrangement of nucleosomes by means of *cis*- and *trans*-effects. *Cis*-effects are changes in the physical properties of nucleosomes that are brought about by the presence of a histone modification - for example, the positive charge on lysine residues is neutralised by the addition of an acetyl group which reduces the binding of basic histones to negatively charged DNA, thereby enabling transcription factors to access the DNA (Vettese-Dadey *et al*., 1996). Phosphorylation adds a net negative charge that is believed to compact nucleosome packaging (Nowak and Corces, 2004). Histone modifications can also elicit *trans*-effects by acting as a docking platform for the recruitment of enzymatic complexes that engage chromatin. For example, methylated lysine residues are recognised by chromo-like domains of the Royal family (chromo, tudor, MBT) and non-related PHD domain containing proteins. These proteins then facilitate downstream chromatin modulating events (Taverna *et al*., 2007).

#### 1.6.2.1 Acetylation

The core histones are reversibly acetylated at several lysine (K) residues. There are twelve known modification sites (Figure 1.6), two in histone H2A (K5, K9), two in histone H2B (K12, K15), four in histone H3 (K9, K14, K18, K56) and four in histone H4 (K5, K8, K12, K16). Histone acetylation is strongly associated with active transcription and histone acetylation sites are required for gene activity (Shahbazian and Grunstein, 2007). It is believed that acetylation may affect chromatin structure as it neutralizes the basic charge of lysine, which may affect the interaction of DNA with histones and nucleosome-nucleosome interactions (Tse *et al*., 1998) and recently it has been shown that acetylation of H4K16 has a negative effect on the formation of the 30 nm chromatin fibre (Shogren-Knaak *et al*., 2006). Thus, histone acetylation results in a ‘loosening’ of chromatin structure to allow greater access to transcription factors. Furthermore, histone H3 K9 acetylation in promoter regions is associated with low nucleosome density in the vicinity of transcription start sites (Nishida *et al*., 2006). Histone acetylation can also function as recognition sites for factors that promote transcription (Shahbazian and Grunstein, 2007). For example, the bromodomain in BRG1, which is the catalytic subunit of the SWI/SNF chromatin remodelling complex, binds acetylated H4 K8, while acetylation of H3 K9 and K14 is critical for the recruitment of $TF_{II}D$ (Agalioti *et al*., 2002).

Recent studies which mapped acetylated histones on a genome-wide level found that acetylation of most lysines in the histone H3 and H4 tails was observed in the 5' end of coding regions and correlated with active transcription (Roth *et al*., 2001; Kurdistani *et al*., 2004; Liang *et al*., 2004; Roh *et al*., 2004; Schubeler *et al*., 2004; Liu *et al*., 2005; Pokholok *et al*., 2005; Roh *et al*., 2006; Koch *et al*., 2007). Furthermore, many inducible genes are marked by histone acetylation even in the inactive state suggesting that the presence of histone acetylation serves to prime these genes for activation at a later stage (Roh *et al*., 2004; Vogelauer *et al*., 2000).

### 1.6.2.2 Methylation

Histones can be methylated either on lysine (K) or arginine (R) residues (Murray, 1964; Patterson and Davies, 1969). Lysine residues can be mono-, di-, or tri-methylated (Lachner *et al*., 2003) whereas arginine residues can only be mono- or di-methylated and di-methylation can occur in a symmetrical or asymmetric configuration (Zhang, 2004). Histone H3 can methylated on a number of lysine sites, which include K4, K9, K27, K36 and K79. Histone H3 can also be methylated on a number of arginine sites - R2, R8, R17 and R26. On histone H4 the main sites of methylation are K20 and R3.

Methylation at H3K4 was first observed in the trout testes (Honda *et al*., 1975) and recently studies have linked it to active gene expression in numerous eukaryotes (Santos-Rosa *et al*., 2002; Ng *et al*., 2003; Schneider *et al*., 2004; Schubeler *et al*., 2004; Bernstein *et al*., 2005; Pokholok *et al*., 2005). The consensus emerging from these large-scale studies is that high levels of H3K4 trimethylation (H3K4me3) are associated with 5' regions of actively transcribed genes. This modification is also positively correlated with histone acetylation and RNA polymerase II occupancy. However, there are differences in the patterns of H3K4 dimethylation (H3K4me2) between yeast and vertebrate chromatin. In vertebrates, most H3K4me2 co-localizes with H3K4me3 in discrete regions located nearby highly expressed genes (Schneider *et al*., 2004; Bernstein *et al*., 2005). In contrast to this H3K4me2 in *S. cerevisiae* is spread throughout genes, peaking in the middle of the coding region and can be associated with active as well as 'poised' genes (Santos-Rosa *et al*., 2002; Ng *et al*., 2003; Pokholok *et al*., 2005). H3K4 monomethylation (H3K4me1) is most abundant at the 3' end of yeast genes and outside of promoter regions and it has been correlated with functional enhancers in human cell lines (Heintzman *et al*., 2007; Roh *et al*., 2007). Clearly the number of methyl groups plays a significant role in the functional consequences of histone methylation.

Heterochromatin in higher eukaryotes is characterised by histone hypoacetylation and H3 K9 methylation (Richards and Elgin, 2002). Heterochromatin protein 1 (HP1) was shown to specifically recognize methylated H3 K9 via its chromodomain (Lachner *et al*., 2001; Nakayama *et al*., 2001). This recognition of H3 K9 by HP1 is required for the formation of heterochromatin. The production of short heterochromatic RNA (shRNA) is involved in the targeting of H3 K9 methylation to heterochromatin regions (Grewal and Moazed, 2003). However, H3 K9 methylation and HP1 binding were recently detected on active genes (Vakoc *et al*., 2005, 2006) suggesting that H3 K9 methylation is not limited to inactive regions of chromatin. Methylation of histone H3 K27 exhibits some similarities to K9 methylation. Both lysines are found within an ARKS sequence on histone H3 and K27 methylation is also associated with transcriptional silencing (Ringrose *et al*., 2004). In particular, methylation of H3 K27 is characteristic of the inactive X chromosome in female cells (Wang et al., 2001; Mak *et al*., 2002). H3 K27 methylation facilitates the binding of polycomb via its chromodomain. Polycomb is a component of the polycomb repressive complex 1 (PRC1) and is required for transcriptional silencing by the polycomb group (PcG) complex (Cao *et al*., 2002).

Methylation of H3K36 is associated with the elongating, serine 2-phosphorylated form of RNA Polymerase II (Xiao *et al*., 2003; Schaft *et al*., 2003; Krogan *et al*., 2003) and is detected across actively transcribed regions, peaking at the 3' end of genes (Bannister *et al*., 2005 b; Mikkelsen *et al*., 2007). Processivity of Pol II through coding regions requires histone acetylation. Transcriptional regulation also needs to suppress initiation from cryptic start sites that occur within coding regions. To suppress these initiation events, H3 K36 methylation creates a recognition site for the chromodomain protein Eaf3, which in turns recruits the Rpd3 HDAC complex (Keogh *et al*., 2005; Carrozza *et* *al*., 2005). The HDAC activity of this complex removes histone acetylation associated with transcriptional elongation, thereby suppressing internal initiation of transcription.

Methylation of H3 K79 is unusual because this modification lies at the core of the nucleosome rather than on the tail. Global analysis of H3 K79 methylation has shown that this modification primarily associates with the coding region of actively transcribed genes (Miao *et al*., 2005) but so far no protein has been identified that binds to this modified residue and links it to transcriptional regulation. The only evidence for how H3 K79 methylation functions in transcriptional activation comes from yeast, where it was shown that the presence of H3 K79 methylation in euchromatic regions prevents the Silent information regulator (Sir) proteins from interacting with active chromatin, thus concentrating Sir Complex binding at silent chromatin regions (Ng *et al*., 2003 b). H3 K79 methylation has also been implicated in DNA repair as the checkpoint protein P53BP1 has been shown to bind methylated H3 K79 (Martin and Zhang, 2005). H4 K20 methylation is connected to transcriptional repression and DNA repair, although very little is known about how it functions in these processes. The lysine demethylase JMJD2A binds to methylated H4 K20 via its tudor domain (Huang *et al*., 2006) and this interaction may contribute to transcriptional repression. In fission yeast, sites of DNA damage contain methylated H4K20, which is recognized by the checkpoint protein Crb2 (Sanders *et al*., 2004; Botuyan *et al*., 2006).

Histone arginine methylation can contribute to active and repressive chromatin states (Strahl *et al*., 2001; Yu *et al*., 2006). Methylation of histone H3 at R2, R17, and R26 (Schurter *et al*., 2001) enhances nuclear receptor- mediated gene activation (Chen *et al*., 1999) while methylation of histone H4 R3 is involved in nuclear receptor-mediated transcription activation (Wang *et al*., 2001 b). Histone arginine methylation has recently been implicated in regulating pluripotency in the early mouse embryo (Torres-Padilla *et al*., 2007). Methylation of arginine residues is enhanced in four cell blastomeres that contribute to the inner cell mass and is minimal in the cells that contribute to the mural trophectoderm, suggesting that this modification could contribute to early cell fate determination. However, it is not understood how arginine modifications contribute to chromatin remodelling and gene activation as no proteins have been identified which bind to methylated arginine residues.

### 1.6.2.3. Phosphorylation

The core histones are phosphorylated on specific serine and threonine residues (Figure 1.6). Most studies have focused on the role of H3S10 phosphorylation (Johansen and Johansen, 2006). Phosphorylation on this residue was found to occur in tandem with the activation of immediate early genes such as c-jun and c-fos (Mahadevan *et al*., 1991) and at activated heat shock genes (Nowak and Corces, 2000). Phosphorylation of H3S10 was also observed during chromosome condensation (Wei *et al*., 1998). Therefore H3S10 phosphorylation is implicated with chromatin states, the 'open' chromatin of active genes during interphase and the 'closed' condensed chromatin of mitotic chromosomes. H3S10 seems to function by regulating a methylation/phosphorylation switch that inhibits HP1 binding to H3K9me3 (Fischle *et al*., 2003) and indeed this was shown to be true as phosphorylation of H3S10 is responsible for HP1 dissociation during mitosis (Fischle *et al*., 2005). These observations suggest a model for how H3S10 phosphorylation functions in the two opposing processes of gene activation and chromosome condensation. During interphase this modification promotes removal of HP1 from specific regions, allowing gene expression. Removal of the phosphorylation mark would therefore promote heterochromatin formation and promote chromatin condensation. Recently DNA methylation has been shown to have a role in targeting H3S10 phosphorylation to pericentromeres (Monier *et al*., 2007).

### 1.6.2.4. Ubiquitination

Histones H2A and H2B have been reported to be ubiquitinated, the carboxyl end of ubiquitin is added to K119 in H2A and K120 in H2B in humans. Histone H2A was the first ubiquitinated histone to be identified (Goldknopf *et al*., 1975) and the majority of this modification is the monoubiquitinated form. H2A ubiquitination has been linked to polycomb silencing and X-chromosome inactivation (Wang *et al*., 2004; De Napoles *et al*., 2004). Ubiquitinated H2A at K119 was found on the inactive X-chromosome in females and is correlated with the recruitment of PcG proteins PRC1-like (PRC1-L). The ubiquitin moiety is approximately half the size of a core histone, so it has been suggested that ubiquitination of a nucleosome would impact chromatin folding, thus affecting transcription (Shilatifard, 2006).

#### 1.6.2.5. Sumoylation

SUMO is a small ubiquitin-related protein of ~100 amino acids, which is capable of being ligated to its target protein. Protein sumoylation is involved in the regulation of transcription factors and components of the transcriptional machinery (Manza *et al*., 2004, Hannich *et al*., 2005) and often results in transcriptional repression (Ghioni *et al*., 2005). Histone H4 sumoylation has been reported in mammalian cells and correlates with transcriptional repressive events such as histone deacetylation and HP1 recruitment (Shiio and Eisenman, 2003). The recent description of histone sumoylation in *S. cerevisiae* and its association with transcriptional repression may address the fact that, unlike vertebrates and *S. pombe*, *S. cerevisiae* has no histone marks associated with transcriptional repression (Nathan *et al*., 2006). This evolutionarily conserved repressive mark seems to block activating acetylation and ubiquitination marks.

#### 1.6.2.6. ADP-ribosylation

Mono-ADP ribosylation of histones is linked to DNA repair and cell proliferation (Hassa *et al*., 2006). Histones are mono-ADP-ribosylated when exposed to DNA damaging agents. This modification has the potential for 'cross-talking' with other modifications as mono-ADP-ribosylation on H4 occurs preferentially when H4 is acetylated (Golderer and Grobner, 1991). Poly-ADP-ribosylation has not been confirmed on histones but it may play a role in local chromatin compaction (Hassa *et al*., 2006).

### 1.6.3. The histone code hypothesis

The initial observations that histone acetylation influenced the initation and elongation phases of transcription led to the suggestion that acetylation of histones provided an epigenetic code for the regulation of transcription (Turner, 2000). Strahl and Allis then provided a histone code model for the function of specific modifications (Strahl and Allis, 2000). The histone code hypothesis predicted that histone modifications on the same or different histones may be interdependent and that different combinations of histone modifications may act synergistically or antagonistically to affect transcription. For example, acetylation of H3K14 by GCN5 is enhanced by H3S10 phosphorylation

(Berger, 2002) while methylation of H3K9 is inhibited by phosphorylation of H3S10 and vice versa (Rea *et al*., 2000). Furthermore, the code predicted that distinct histone modifications provide a binding site for chromatin-associated ‘effector’ proteins, which mediate downstream functions. For example, the PHD (plant homeodomain) finger in BPTF, the largest subunit of the nucleosomal remodelling factor (NURF) complex, interacts specifically with peptides modified with H3K4me3 (Wysocka *et al*., 2006).

Agalioti and colleagues carried out one of the first studies to examine the histone code hypothesis (Agalioti *et al*., 2002). The human IFN-β gene is turned on by three transcription factors which form an enhanceosome at the enhancer region. The enhanceosome facilitates transcription by recruiting HATs, SWI/SNF chromatin remodelling complex and basal transcription factors in an ordered manner. GCN5 acetylated H4K8, which was required for the recruitment of the SWI/SNF complex and acetylation of H3K9 and H3K14 was required for TFIID recuitment. The authors proposed that distinct acetylation marks were required for the recruitment of transcription complexes and that this constituted a histone code. A number of large scale studies have also been carried out in budding yeast, fission yeast, *D. melanogaster*, mouse and human, which have examined a range of histone modifications (Roh *et al*., 2004, 2005, 2006; Kurdistani *et al*., 2004; Dion *et al*., 2005; Liu *et al*., 2005; Pokholok *et al*., 2005; Millar *et al*., 2006; Rao *et al*., 2005; Wiren *et al*., 2005; Sinha *et al*., 2006; Schubeler *et al*.,2004; Bernstein *et al*., 2005, 2006; Boyer *et al*., 2006; Heintzman *et al*., 2007; Barski *et al*., 2007; Koch *et al*., 2007). Some of these studies reported a clear coordination between various histone modifications and gene activity, namely histone H3 and H4 acetylation and H3K4 methylation states are found at the 5’ regions of active genes (Pokholok *et al*., 2005; Bernstein *et al*., 2005; Schubeler *et al*., 2004; Roh *et al*., 2005, 2006; Koch *et al*., 2007), whereas elevated H3K27 methylation correlates with gene repression (Boyer *et al*., 2006; Roh *et al*., 2006). In addition, H3 acetylation and H3K4me1 modifications outside of promoter regions has been correlated with enhancer elements (Heintzman *et al*., 2007; Roh *et al*., 2005). In contrast, other studies did not find a correlation between histone acetylation and gene activity (Kurdistani *et al*., 2004; Liu *et al*., 2005) and therefore dispute the existence of a histone code. Recent discoveries have complicated the simplified view that specific histone modifications underlie either an active or inactive chromatin status - for example the co-localization of the apparently contradictory modifications H3K4me3 and H3K27me3 in bivalent domains (Bernstein *et al*., 2006; Mikkelsen *et al*., 2007), while H3K9me3 has been shown to be enriched at a number of active promoters (Vakoc *et al*., 2005). In addition, the Sin3-HDAC complex has been shown to bind to H3K4me3 via the PHD (plant homeodomain) domain of the Ing2 protein to repress gene expression (Shi *et al*., 2006). This suggests that the presence of multiple modifications may be required to elicit a specific biological output and Ruthenburg and colleagues (2007) have proposed that in many cases one histone modification is not sufficient to recruit a given complex, rather multiple histone modifications all contribute to the recruitment and stabilization of chromatin complexes and dictate functional outcomes.

### 1.6.4. DNA methylation

In eukaryotes, DNA methylation is confined to cytosine bases at CpG dinucleotides and is associated with gene repression (Klose and Bird, 2006). CpGs often cluster into CpG islands and approximately 60% of human promoters are associated with CpG islands (Bird, 2002). It has been suggested that the majority of CpG islands are always unmethylated but some are methylated in a tissue-specific manner. DNA methylation functions to inhibit gene expression by two mechanisms, firstly, DNA methylation can inhibit transcription factors from binding to their DNA recognition sequence (Watt and Molloy, 1988). Secondly, proteins which recognize methyl-CpG, methyl-CpG binding proteins (MBPs) (Hendrich and Bird, 1998), can recruit transcriptional co-repressors such as histone methyltransferases and deacetylases to modify chromatin and mediate gene silencing (Jones *et al*., 1998; Nan *et al*., 1998; Sarraf and Stancheva, 2004). In these cases, DNA methylation is coupled with repressive histone modifications. Furthermore, DNA methylation within the body of a gene has been shown to alter chromatin structure and gene expression by affecting Pol II elongation efficiency (Lorincz et al., 2004).

There are two classes of mammalian DNA methyltransferases (DNMTs) - *de novo* and maintenance DNMTs. DNMT3a and DNMT3b are members of the *de novo* class, as they are responsible for methylating at previously unmethylated CpGs. DNMT1 is a maintenance enzyme as it copies methylation patterns onto newly synthesized DNA strands. *De novo* methylation is important during embryogenesis as the paternal genome is actively demethylated following fertilization and the maternal genome is demethylated passively as a result of DNA replication. Many CpG sites are re-methylated in the blastocyst resulting in the patterns observed in the adult (Reik *et al*., 2001). Alterations in DNA methylation patterns have been implicated in numerous diseases including cancer. The promoters of tumour suppressor genes are often hypermethylated, which silences their expression (Jones and Baylin, 2007) and it has recently been proposed that cancer may evolve from stem cells which carry epigenetic alterations in addition to other genomic alterations (Feinberg *et al*., 2006), but their exact origins remain controversial (Bjerkvig *et al*., 2005).

### 1.6.5. Regulation of embryonic stem cell pluripotency

One of the most important discoveries in the field of molecular biology was the development of methods during the 1980s to generate pluripotent embryonic stem (ES) cells from the inner cell mass of pre-implantation embryos (Evans and Kaufman, 1981; Martin, 1981). These cells could be grown indefinitely in culture and microinjected back into mouse blastocytes where they contributed to the formation of various cell lineages in the adult mouse. ES cells represent an undifferentiated cell type that can in theory generate daughter cells capable of differentiating into every cell type found in the adult organism, a property known as pluripotency. It is becoming clear that epigenetic mechanisms play a role in maintaining ES cells in a pluripotent state (Azuara *et al*., 2006, Mikkelsen *et al*., 2007) and that the maintenance of ES cell pluripotency depends on the transcriptional expression and silencing of a number of genes (Boyer *et al*., 2006). The key factors involved in regulating pluripotency are discussed below.

#### 1.6.5.1. Transcription factors involved in regulating pluripotency

Several pluripotency-sustaining transcription factors such as NANOG and OCT4 are expressed in ES cells and are silenced upon differentiation (Ramalho-Santos *et al*., 2002; Ivanova *et al*., 2002). Two recent genome-wide studies identified targets of NANOG, OCT4, and SOX2 in ES cells (Boyer *et al*., 2005; Loh *et al*., 2006) and demonstrated that these proteins bind to several hundred genes which can be transcriptionally active or silent. These loci are often involved in developmental processes and included ES cell specific genes and repressed tissue-specific transcription factors.

PcG proteins are required for ES cell pluripotency and are dramatically down-regulated upon differentiation (Cao and Zhang, 2004). Genome-wide studies investigating PcG protein binding have been carried out in human and mouse ES cells (Lee *et al*., 2006; Bracken *et al*., 2006; Boyer *et al*., 2006; Schwartz *et al*., 2006; Negre *et al*., 2006; Tolhuis *et al*., 2006). Genes that are required during differentiation and development, such as members of the Hox and Pax transcription families, are repressed in human and mouse ES cells by the PcG machinery and are often found in bivalent chromatin regions (Azuara *et al*., 2006; Bernstein *et al*., 2006) (see section 1.6.5.2). The majority of the genes cooperatively regulated by NANOG, OCT4, and SOX2 were repressed and overlapped with PcG protein binding sites. This suggests that these transcription factors may maintain pluripotency by coordinating the recruitment of PcG complexes to repress tissue specific genes.

### 1.6.5.2. Epigenetic regulation of pluripotency

In addition to transcription factors, chromatin structure and epigenetic modifications play a key role in the maintenance of pluripotency (Spivakov and Fisher, 2007; Reik, 2007). Evidence suggests that ES cell chromatin may be less compact and more transcriptionally ‘permissive’ than differentiated cells (Bernstein *et al*., 2007). For example, differentiated and undifferentiated human and mouse ES cells showed dramatic differences in the nuclear organization of centromeric heterochromatin and regions involved in pluripotency (Wiblin *et al*., 2005). In differentiated cells, many inactive genes are positioned close to centromeric heterochromatin (Brown *et al*., 1999), but this phenomenon has not been observed in ES cells. Furthermore, chromatin proteins are more loosely associated in ES cells when compared with differentiated cells, indicating that the chromatin of ES cells is more accessible (Meshorer *et al*., 2006).

Recent studies of histone modifications in pluripotent and differentiated cells have also advanced our understanding of the chromatin properties important for initiating and maintaining a pluripotent state (Bernstein *et al*., 2006; Azuara *et al*., 2006; Chambeyron *et al*., 2005; Szutorisz  *et al*., 2005 b; Mikkelsen *et al*., 2007). These studies have shown that inactive genes in ES cells can be associated with high levels of H3K4me2, H3K4me3 and acetylated histones H3 and H4. Some of these genes were also enriched for the repressive H3K27me3 modification. Given that histone acetylation, H3K4me2 and H3K4me3 are normally associated with expressed genes while H3K27me3 is associated with non-expressed genes, the presence of these 'contradictory' modifications on non-expressed genes in ES cells was intriguing. Consecutive or sequential ChIP reactions (re-ChIP) confirmed that these contradictory modifications were present on the same or neighbouring nucleosome (Azuara *et al*., 2006, Bernstein *et al*., 2006). These 'bivalent' domains were found to mainly overlay developmental regulator genes, the majority of which were not expressed in ES cells. In differentiated cell types, H3K27me3 alone marked several inactive developmental genes whereas H3K4me3 marked the active ones. This suggests that transcription factors involved in tissue specific development are both primed for expression and 'held back' at the same time so that their expression can be tightly regulated in ES cells.

## 1.7. Identification and characterisation of non-coding regulatory elements

It is evident that gene expression in eukaryotes is a highly controlled process requiring regulation at many different levels. The non-coding regulatory elements, in combination with the proteins that interact with them, are crucial in determining gene expression patterns. In order to understand the regulatory networks which control gene expression patterns, it is important to identify and characterise the regulatory elements associated with genes. Over the past four decades, various assays have been used to identify regulatory elements in the human genome, the majority of which have been low-throughput processes. The completion of the human genome sequence is allowing for the development of high-throughput experimental and computational methods which should ensure that these elements are identified in a more efficient manner. Discussed below are many of the low-throughput and high-throughput methods which have been used to identify non-coding regulatory elements in a variety of species.

### 1.7.1. Classical or low-throughput methods

#### 1.7.1.1. DNA footprinting

DNA footprinting is used to identify binding sites of proteins that bind to DNA. It works on the principle that when a protein binds to DNA it 'protects' the underlying DNA from cleavage by DNase I when compared with unbound DNA (Galas and Schmitz, 1978). Usually DNA fragments 200-300 bp in length are used as targets and are radiolabelled at one end. These fragments are then incubated in the presence or absence of a protein extract and then exposed very briefly to low concentrations of DNase I. Digested products are size fractionated on denaturing polyacrylamide gels and then autoradiographed. Control samples show a series of bands of different lengths due to random digestion by DNase I, whereas a test sample contains gaps where no fragments are observed (footprints). The gaps indicate the sites where protein is bound. While DNA footprinting can identify sites of DNA-protein interaction, it does not confer any functionality to the site. Therefore, it is often used in combination with other techniques such as electromobility shift assays (section 1.7.1.3) to gain a greater insight into the function of the regulatory element.

#### 1.7.1.2. DNase I hypersensitive site mapping

This technique is based on the observation that regions of chromatin which are accessible to transcription factors are more sensitive to DNase I digestion than condensed chromatin. Regions identified by their DNase I hypersensitivity are therefore likely to be involved in transcriptional regulation. This approach has the advantage of identifying most if not all *cis*-regulatory elements, but does not directly confer functionality to identified elements. DNase I hypersensitive sites (HSs) are identified by exposing chromatin to low amounts of DNase I for a short period of time, followed by DNA purification of cleaved DNA, restriction enzyme digestion, gel electrophoresis, southern blotting and hybridization with a radioactive DNA probe. This technique has been widely used to identify *cis*-regulatory elements such as promoters, enhancers, repressors, insulators and locus control regions in many cell types (Weintraub and Groudine, 1976;

Wu, 1980; Gross and Garrard, 1988). More recently, high-throughput approaches have been used to identify HSs (see section 1.7.3.3).

### 1.7.1.3. Electromobility shift assays

An electromobility shift assay (EMSA), also known as a gel-shift or gel retardation assay, is another technique used for studying DNA-protein interactions. It relies on the principle that a DNA sequence bound by a protein(s) migrates slower than unbound DNA during electrophoresis (Garner and Revzin, 1981). In a typical experiment, radiolabelled DNA fragments suspected of containing a regulatory sequence are incubated with or without protein extracts and then size-fractionated by polyacrylamide gel electrophoresis. DNA fragments which bind protein are identified by their lower mobility band on a gel. This type of assay can be used to quickly identify the presence of a specific DNA-protein interaction but with the caveat that *in vitro* identified binding sites do not always reflect *in vivo* binding sites (Lieb *et al*., 2001). Both EMSA and footprinting assays can often detect unintended DNA-protein interactions as result of non-specific proteins, such as DNA repair proteins, binding to the end of DNA probes (Klug, 1997).

### 1.7.1.4. PCR-based methods for detecting DNA-protein interactions

PCR can be used to amplify DNA fragments that bind proteins. Several PCR-based techniques have been developed for this purpose, namely **s**ystematic **e**volution of **l**igands by **ex**ponential enrichment (SELEX) (Tuerk and Gold, 1990), **s**election **a**nd **a**mplification **b**inding site (SAAB), **c**yclic **a**mplification **s**election **t**argets (CASTing) (Wright *et al*., 1991), and **mu**ltiplex **s**election **t**arget (MUST) (Nallur *et al*., 1996). In these methods DNA-protein complexes are isolated by techniques such as immunoprecipitation and binding to affinity columns. The DNA is then recovered, PCR amplified, then mixed with fresh proteins and after many rounds of this DNA fragments interacting with a specific protein are enriched. Quantitative PCR in combination with chromatin immunoprecipitation can also be used to identify DNA-protein interactions (section 1.7.3.5).

### 1.7.1.5. Reporter-gene assays which confer function

There are many variations to the reporter-gene assay, which allow it to be used for identifying the majority of regulatory elements (Matson *et al*., 2006) (See Figure 1.7). The 'test' genomic region is cloned into a plasmid upstream of a reporter gene such as the luciferase, chloramphenicol acetyltransferase (CAT), β-galactosidase, green fluorescent protein (GFP), or G418 resistance gene. The construct is then transfected transiently or stably into cultured cells and the output of the reporter gene is assayed to determine if the test sequence has functional activity.

The arrangement of the construct depends on the regulatory activity being tested for (Fig 1.7), for example if a DNA segment is being tested for core promoter activity it is cloned immediately upstream of a reporter gene which lacks a promoter. Proximal promoters are assayed by cloning them upstream of a reporter gene whose expression is driven by a weak heterologous core promoter. This system can also be used to test for gene enhancer and silencer activity, by cloning these putative elements in a reporter construct whose expression is driven by a weak or strong promoter. An increase/decrease in reporter gene expression is used to determine enhancer/silencer activity. These approaches have been used to characterise these elements in a number of genes including c-Myc and SCL (Mautner *et al*., 1995; Gottgens *et al*., 1997).

Insulators can be measured for enhancer-blocking (Bell *et al*., 1999) or heterochromatin barrier activity (Recillas-Targa *et al*., 2002). In enhancer-blocking assays, the putative insulator sequence is cloned between an enhancer and promoter that are known to interact. If the element has enhancer-blocking activity, then it will interfere with enhancer-promoter communication and reporter gene expression will be reduced. Assaying for heterochromatin barrier activity requires an assay in which the reporter construct is stably integrated into the genome. Barrier elements flanking the reporter gene would shield it from position effects which cause transgene silencing. Thus barrier elements allow for position-independent expression of the reporter gene. Conferring LCR activity requires the identification of a genomic segment that can overcome position effects to confer temporal and tissue specific expression of a reporter gene (Grosveld *et al*., 1987).

There are several disadvantages associated with using reporter assays to identify regulatory elements. Firstly, the location of these elements is often not known and they can be found close to and far from a gene. Furthermore, regulatory elements can de different sizes and knowing what size segment to test can be a problem. Secondly, chromatin context plays a key role in regulating gene expression patterns and these reporter constructs do not reflect the correct context. Thirdly, if the cell culture system used does not match developmental conditions under which the regulatory element is normally active, then the activity may not be detected. Despite these disadvantages, reporter genes assays are still the most accurate way of conferring functionality upon a putative regulatory element.

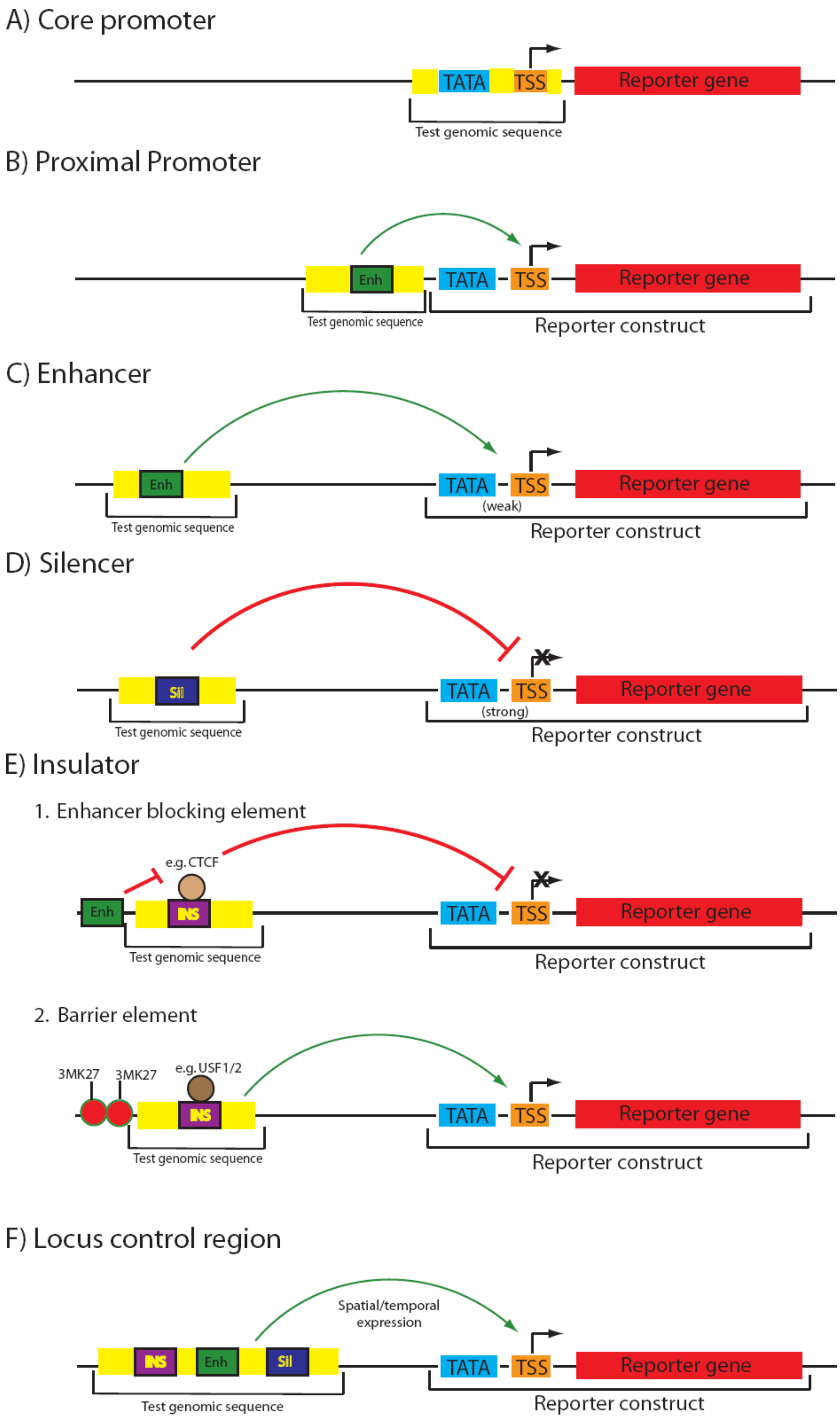
A) Core promoter
TATA
TSS
Reporter gene
Test genomic sequence
B) Proximal Promoter
Enh
TATA
TSS
Reporter gene
Test genomic sequence
Reporter construct
C) Enhancer
Enh
TATA
TSS
Reporter gene
(weak)
Test genomic sequence
Reporter construct
D) Silencer
Sil
TATA
TSS
Reporter gene
(strong)
Test genomic sequence
Reporter construct
E) Insulator
1. Enhancer blocking element
e.g. CTCF
Enh
INS
TATA
TSS
Reporter gene
Test genomic sequence
Reporter construct
2. Barrier element
3MK27
3MK27
e.g. USF1/2
INS
TATA
TSS
Reporter gene
Test genomic sequence
Reporter construct
F) Locus control region
Spatial/temporal
expression
INS
Enh
Sil
TATA
TSS
Reporter gene
Test genomic sequence
Reporter construct

**Figure 1.7: Functional plasmid-based reporter gene assays for the identification of regulatory elements**. A) A genomic sequence (yellow box) is tested for core promoter activity by cloning it immediately upstream of a reporter gene which lacks an endogenous promoter. B-D) Proximal promoters, enhancers and silencers can be assayed for by cloning a genomic segment upstream of a reporter gene driven by the appropriate strength promoter. E) Enhancer-blocking insulator elements interfere with enhancer-promoter communication, thereby down-regulating reporter gene expression. Barrier insulators shield a reporter gene from heterochromatin-mediated gene inactivation. F) Locus control regions can overcome position effects and confer correct expression patterns in reporter assays. Enh: enhancer; INS: insulator; Sil: silencer.

### 1.7.2. Computational detection of regulatory elements

A complete computational approach to studying transcriptional regulatory elements (Elnitski *et al*., 2005) often requires diverse data sets in order to determine (1) promoter location (2) predicted and verified transcription factor binding sites (3) gene expression profiles and (3) sequence conservation. The availability of such data sets has allowed rigorous computational prediction of regulatory elements in vertebrate genomes as discussed below (Prakash and Tompa, 2005; Cora *et al*., 2005; Hallikas *et al*., 2006; Prabhakar *et al*., 2006).

#### 1.7.2.1. Promoter location prediction

Identifying the promoter of a particular gene can be a difficult task as core promoter sequences can be located a large distance from the first coding exon due to 5'-untranslated regions and introns (Maston *et al*., 2006). As discussed in section 1.4.1, promoters contain various combinations of core promoter motifs (Gershenzon *et al*., 2005) and searching for the co-occurance of these motifs has had limited success in predicting promoter locations (Fickett and Hatzigeorgiou, 1997). Promoter prediction programs based on the analysis of known core promoters have been most successful and include PromoterInspector (Scherf *et al*., 2000), First EF (Davuluri *et al*., 2001) and Eponine (Down and Hubbard, 2002). However, the sensitivity and specificity of these programs is limited by the number of known core promoters and are limited to finding new promoters that are similar to ones in the training data. Approximately 60% of human genes lie near CpG islands and a comparison of promoter-prediction programs found that promoters associated with CpG islands are predicted well. However, prediction of the other 40% of promoters is much less reliable (Bajic *et al*., 2004). Therefore, the availability of more experimental data in the form of novel transcripts (Carninci *et al*., 2005), more precise mapping of 5' ends of transcripts and ChIP-chip data for factors which bind to promoters (Kim *et al*., 2005) will aid the training of these programs.

### 1.7.2.2. Prediction of transcription factor binding sites

Genome sequences can be scanned for sequence motifs which match experimentally verified transcription factor binding sites (TFBSs). Most TFBSs are very short sequences and are often degenerate, usually only 4-6 bp within each TFBS are fully conserved and these sites often occur in clusters (Maniatis *et al*., 1987). Experimental data on the location of transcription factor binding sites has been compiled in databases such as TRANSFAC (Wingender *et al*., 2000) and information on various TFBSs can be used to build a position-specific scoring matrix (PSSM) for a particular TF (Vavouri and Elgar, 2005). Programs such as MATCH (Kel *et al*., 2003) can scan an input sequence for matches to all PSSMs in TRANSFAC. However programs based on PSSMs often identify a high number of false positives due to the quality of data used to build a matrix (Fogel *et al*., 2005). To overcome this problem, more sophisticated statistical models have been used to predict TFBSs such as the program JASPAR (Sandelin *et al*., 2004). Hallikas and colleagues have recently used the well defined binding specificities of several transcription factors involved in the Hedgehog, Wnt and Ras/MAPK signalling pathways to develop a computational tool which could identify mammalian enhancer elements regulated by these pathways (Hallikas *et al*., 2006). The authors made use of the observation that TF binding sites tend to be found in clusters when forming a tissue-specific enhancer element, to produce a computational tool capable of accurately identifying enhancers. However the problem still exists that the majority of TF binding specificities have not been determined.

### 1.7.2.3. Comparative sequence analysis

With the completion of genome sequences for a number of organisms, it has become easier to compare genomic sequences and identify regions of high sequence conservation across species. The use of comparative sequence analysis or comparative genomics to identify non-coding regulatory elements has recently become very popular. The rationale behind this approach is that, just like coding sequences, regulatory sequences are under evolutionary selective pressure and so should have evolved at a slower rate than other non-coding sequences. A number of programs have been developed to identify sequences which have been conserved through evolution such as PhastCons (Siepel et al., 2005), Footprinter (Blanchette and Tompa, 2003), SynPlot (Gottgens *et al*., 2001) and VISTA (Visel *et al*., 2007). Comparative sequence analysis has been used in many cases to identify *bona fide* regulatory elements (Gottgens *et al*., 2000; Martin *et al*., 2004; Nobrega *et al*., 2003; Woolfe *et al*., 2005; Pennacchio *et al*., 2006, 2007). For example, whole genome comparisons between humans and the pufferfish, *Fugu rubripes*, identified 1400 highly conserved non-coding sequences (Woolfe *et al*., 2005). Many of the sequences displayed tissue-specific enhancer activity when tested in functional assays. Pennachio and colleagues (2006) used a similar approach to expand the number of characterized human enhancers and the data derived from such studies will be extremely useful in the training of enhancer prediction programs.

However, conserved sequence elements do not always correspond to functional regulatory regions (Balhoff and Wray, 2005). It has been suggested that there is a high rate of evolutionary turnover of mammalian TFBSs (Dermitzakis and Clark, 2002). This may be due to degeneracy of TFBSs or a specific regulatory element may not be conserved (Follows *et al*., 2003). A recent study by Odom and colleagues, utilized ChIP-chip to map the binding sites of conserved tissue-specific transcription factors (FOXA2, HNF1A, HNF4A, and HNF6A) in human and mouse hepatocytes (Odom *et al*., 2007). It was observed that many of the binding events for the transcription factors were species-specific. This analysis also showed that for many TF binding events in human, the orthologous gene in mouse is bound but not at the conserved sequence element. Therefore it seems that many of the transcriptional regulatory elements relevant to human development may not be highly functionally conserved between evolutionarily distant species. Furthermore, sequence comparisons showed that 21,855 of the sequences associated with histone acetylation islands are not conserved between the human and mouse genomes and random sampling showed that 50% of these non-conserved sequences function as enhancers (Roh *et al*., 2007). Therefore, histone modification data can be used to identify species-specific regulatory elements that would otherwise be missed by comparative sequence analysis alone. Phylogenetic shadowing (Boffelli *et al*., 2003), which analyses sequences from closely related species such as primates, may be required to identify human specific regulatory elements (Prabhakar *et al*., 2006).

### 1.7.3. Applications of genomic DNA microarrays to identify regulatory elements

With the completion of many genome sequences, DNA microarray technology has emerged as an important technology for investigating global gene regulation events as described below (Hoheisel, 2006). Typically, DNA microarrays consist of a large collection of DNA sequences that are fixed to the surface of a glass slide. The DNA sequences can be comprised of large genomic clones (BACs, PACs and cosmids), cDNA clones, primer specific PCR products, or short oligonucleotides (Fiegler *et al*., 2003; Duggan *et al*., 1999; Dhami *et al*., 2005; Lipshutz *et al*., 1999). High-density oligonucleotide microarrays are manufactured by the direct synthesis of oligonucleotides on the slide surface. Photolithography, optical mirrors or ink-jets can be used to synthesize oligonucleotides (Lipshutz *et al*., 1999; Singh-Gasson *et al*., 1999; Hughes *et al*., 2001). Spotted arrays rely on robotic devices to spot clone fragments, PCR products, or oligonucleotides (Schena *et al*., 1995). The glass slides are coated with reactive molecular groups such as poly-L-lysine, which allows the DNA probes to bind to the slide.

For spotted microarrays, the test sample (RNA or DNA) and a reference sample are normally fluorescently labelled with nucleotide derivatives, usually containing Cy3 and Cy5 (Figure 1.8). DNA sequences on the microarray may contain repetitive sequences so the binding of repetitive DNA elements is suppressed by using Cot1 DNA in a competitive hybridization. During hybridization, the labelled samples bind to their complimentary immobilized probes sequences and the fluorescent signal is calculated in the two channels to determine which sequences are enriched in the test sample relative to the reference sample.

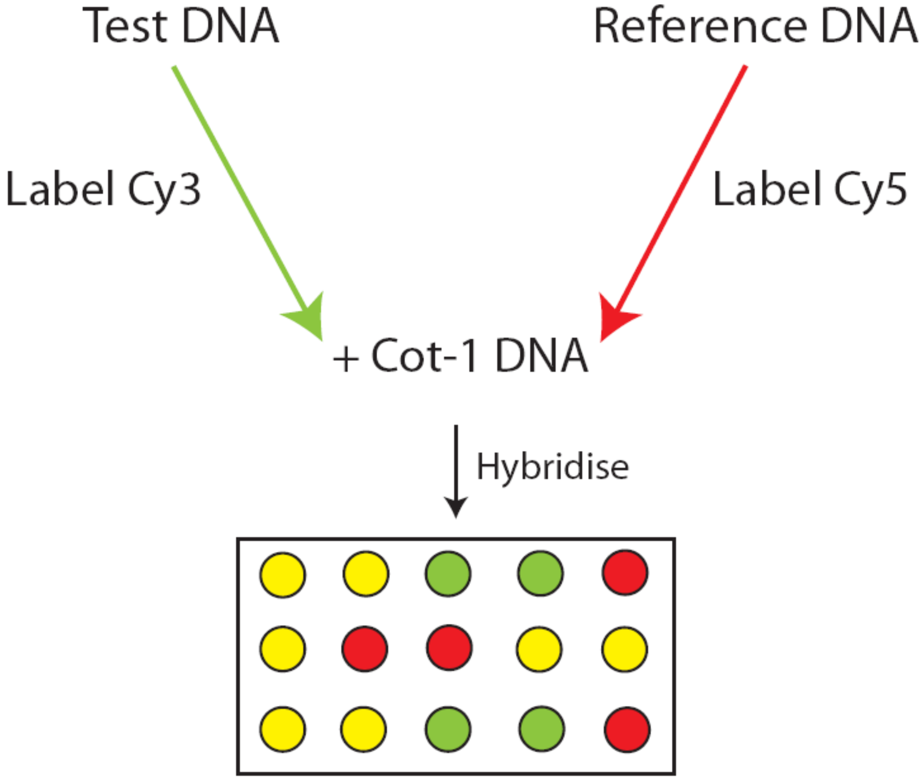

**Figure 1.8: The basic principles of two-colour competitive hybridisation**. The test DNA sample is labelled with Cy3 nucleotide derivatives while the reference DNA sample is labelled with Cy5 nucleotide derivatives. The two DNA samples are mixed together with Cot-1 DNA and are competitively hybridised to a microarray. Green spots represent genomic regions that are enriched in the test sample relative to the reference sample, red spots represent DNA sequences depleted in the test sample relative to the reference and yellow spots represented DNA sequences that are present in equal amounts in the test and reference samples.

### 1.7.3.1. Transcript profiling

Microarray-based monitoring of gene expression was first reported for *Arabidopsis thaliana* in 1995 (Schena *et al*., 1995). Since then transcriptional profiling has become the most widely used application of microarray technology, and has been used to study gene expression in a number of different organisms in numerous experimental systems. For example, it has been used to study gene expression patterns during normal development (White *et al*., 1999) and disease (Shipp *et al*., 2002). Transcriptional profiling can also be used to identify genes which display significant changes in gene expression upon inactivation of a particular transcription factor (Young, 2000). Once a list of perturbed genes is compiled, it allows you to search for sequence motifs present in their upstream/downstream regions and may result in the discovery of novel TFBSs. Several programs are available for motif discovery such as MEME (Bailey and Elkan, 1995), AlignACE (Hughes *et al*., 2000), and NestedMICA (Down and Hubbard, 2005). However, the number of direct targets identified by this method can vary depending on TF redundancy, and developmental stage being studied. Furthermore, the expression of many genes may be perturbed due to secondary effects observed upon inactivating a particular factor.

### 1.7.3.2. Replication timing

It has been known for some time that different parts of the genome replicate at different times during S phase. Actively transcribed regions replicate early during S phase while inactive regions replicate late in S phase. Microarray-based assays have been used to analyse replication timing in yeast (Raghuraman *et al*., 2001), *Drosophila* (Schubeler *et al*., 2002) and humans (Woodfine *et al*., 2004; White *et al*., 2004). Woodfine *et al*. adapted the technique of comparative genomic hybridization to assess replication timing in human cells. S-phase cells were isolated from asynchronously growing cells; the DNA was extracted and labelled. This was hybridized with DNA isolated from G1-phase cells and relative fluorescence at each array spot can be used to infer replication timing. The earlier a locus replicates the more DNA content it will have relative to the same locus in a G1 cell. In theory, an early replicating locus will have a copy number ratio of 2:1, whereas a late replicating locus will have a 1:1 ratio and intermediate replicating loci will be in between. A correlation between early replication and high gene density, high GC content, low SINE repeat content and high transcriptional activity was observed.

White and colleagues used a similar method to analyse replication timing in two different cell lines, but instead of comparing S1:G1 ratios, they isolated DNA from early and late S phase and compared these ratios. They concluded that early replication and transcriptional activity are often correlated but found that genes differentially transcribed between the two cell lines were replicated at the same time. This implies that cell-specific transcription does not alter replication timing but global changes in chromatin architecture may be needed.  A study by Gilbert and colleagues (2004) used random fragmentation of chromatin followed by sucrose centrifugation to separate 'open' and 'closed' chromatin fragments based on mass and density. Replication timing was also assessed and the authors suggested that there was a link between open chromatin and early replication.

### 1.7.3.3. DNase I hypersensitive site microarrays

Traditional methods used for mapping of DNase I hypersensitive sites (HSs) rely on laborious techniques (see section 1.7.1.2) and can only be applied to study small genomic regions in a single experiment (Cockerill, 2000). To circumvent these problems, several protocols have recently been developed which allow DNase I HSs to be mapped in a high-throughput manner (Crawford *et al*., 2004, 2006, 2006 b; Sabo *et al*., 2004, 2006; Follows *et al*., 2006). A number of studies have used cloning of DNase I HSs coupled with large-scale sequencing (Crawford *et al*., 2004, 2006 b; Sabo *et al*., 2004). Crawford and colleagues developed a method in which nuclei that had been digested with DNase I were blunted ended by T4 DNA polymerase. Blunted ended DNA was digested with restriction enzymes and cloned into a vector for sequencing. Sabo and colleagues (2004) attached a biotinylated linker to DNA that had been exposed to DNase I and then cut with a restriction enzyme. The DNase I cut fragments were then captured using streptavidin beads, a second linker was attached and the DNA fragments were amplified and cloned for sequencing. This method was then used in combination with tiled microarrays to identify DNase I HSs (Crawford *et al*., 2006, Follows *et al*., 2006). Sabo and colleagues developed a novel method to isolate DNase I HSs which were then mapped using microarrays (Sabo *et al*., 2006).  This method isolated DNA fragments associated with two DNase I cuts that occurred in close proximity (<1200 bp). These short fragments were isolated by size fractionation on a sucrose gradient. Chromatin and equally sized non-chromatin fragments were then labelled and hybridized to a microarray to identify DNase I HSs.

#### 1.7.3.4. Matrix attachment regions microarrays

Eukaryotic chromatin is organized into loops by attachment to a chromosome scaffold or matrix (Mirkovitch et al., 1984). The DNA and proteins associated with this nuclear scaffold/matrix can be isolated by extraction of histones with high salt or mild detergent followed by restriction enzyme treatment to remove all DNA except matrix attached DNA. The AT-rich DNA segments that mediate attachment of chromatin to the nuclear matrix are known as matrix attachment regions (MARs) and occur on average every 50-200kb in the human genome (Bode, 2000). PML and SATB1 have recently been identified as MAR-binding proteins that regulate transcription by orchestrating chromatin loop formation (Kumar *et al*., 2007). Sumer and colleagues isolated MAR DNA and hybridized it to a BAC/PAC array to define a 2.5 Mb region of MAR enriched chromatin at a human neocentromere (Sumer *et al*., 2003). Ioudinkova and colleagues have also used arrays to map MARs at the chicken α-globin domain (Ioudinkova *et al*., 2005) suggesting that high-resolution microarrays could be used to map MAR sites throughout the human genome.

#### 1.7.3.5. Chromatin immunopreciptation microarrays (ChIP-chip)

Another recent application of microarray technology has been to study chromatin structure and function. DNA microarrays in combination with chromatin immunoprecipitation (ChIP) have been used to investigate the *in vivo* interactions of transcription factors or other regulatory complexes with genomic DNA. The use of ChIP in combination with microarrays has been termed ChIP-on-chip or ChIP-chip.

ChIP is one of the most powerful and widely used techniques for investigating *in vivo* DNA-protein interactions as these events are cross-linked in the native chromatin environment. Solomon and Varshavsky (1985) pioneered the development of a ChIP procedure and since then ChIP has been used in organisms ranging from yeast to human cells. The ChIP procedure is typically performed by cross-linking DNA-protein interactions using formaldehyde. The chromatin is then extracted by lysing the cells and nuclei. The chromatin is then sonicated into sheared fragments of approximately 300 bp to 1000 bp in size. The cross-linked protein-DNA complexes of interest are then immunoprecipitated with a specific antibody, the cross-links are reversed and the enriched ChIP DNA is recovered. DNA that has not been immunoprecipitated or immunoprecipitated with a mock-antibody is used as a reference. Both the ChIP DNA and reference DNA can be fluorescently labelled and hybridized to a DNA microarray to identify the *in vivo* interactions of regulatory proteins with DNA.

ChIP-chip was first pioneered for the study of yeast transcription factors (Ren *et al*., 2000; Iyer *et al*., 2001; Wyrick *et. al*., 2001; Damelin *et al*., 2002). The ChIP-chip method was subsequently used to study chromatin structure and function in yeast (Robyr *et al*., 2002, Bernstein *et al*., 2002; Nagy *et al*., 2003; Robert *et al*., 2004; Kurdistani *et al*., 2004; Pokholok *et al*., 2005; Liu *et al*., 2005; Lee *et al*., 2004; Bernstein *et al*., 2004; Yuan *et al*., 2005) and has also been applied to study DNA-protein interactions in other genomes, including the human genome. When analysing larger genomes, two main approaches have been taken:

**Biased approach**: This approach uses arrays containing sub-sets of regulatory elements from across the genome such as promoter regions or CpG islands. Promoter arrays, have been used to identify E2F (Ren *et al*., 2002), c-Myc (Li *et al*., 2003), and HNF transcription factors (Odom *et al*., 2004) binding sites in human cells. CpG island microarrays have also been used to identify c-Myc and E2F target genes (Mao *et al*., 2003; Weinmann *et al*., 2002; Wells *et al*., 2002). However the disadvantage of these types of microarrays is that they are inherently biased for the regions of the genome selected to study. Promoter or CpG islands arrays represent a particular set of regulatory elements, so their use in ChIP-chip is restricted to associating function with these elements.

**Unbiased approach**: The unbiased approach uses arrays containing entire genomic regions in the form of tiling paths of DNA sequences. Tiling arrays were first used in a mammalian system to map GATA-1 binding sites across the human β-globin locus (Horak *et al*., 2002). Entire chromosomal tiling paths of chromosome 21 and 22 (Martone *et al*., 2003; Euskirchen *et al*., 2004; Cawley *et al*., 2004) have also been constructed, which allowed for the unbiased identification of NFκB (Martone *et al*., 2003), CREB (Euskirchen *et al*., 2004), Sp1, c-myc, and p53 binding sites (Cawley *et al*., 2004), and estrogen receptor targets (Carroll *et al*., 2005). Many of the binding sites mapped to 3’ ends of genes and within introns, which would have been missed by promoter or CpG arrays. Similarly, tiling microarrays covering the entire human genome have been used to identify active core promoters across the entire human genome in human fibroblast cells (Kim *et al*., 2005). Genome-wide tiling arrays have also been used to identify target sites for the transcription factor p63 (Yang *et al*., 2006), estrogen receptor binding sites (Carroll *et al*., 2006) and the insulator binding protein CTCF (Kim *et al*., 2007), amongst many other examples. Thus, tiling path microarrays can be used to comprehensively map DNA-protein interactions across genomes in an unbiased way.

Despite the rapid advances in generating large datasets, there are several disadvantages associated with current ChIP-chip methods, which limit its application. Firstly, efficiency of the ChIP reaction depends on antibody quality and epitope accessibility and formaldehyde fixation may introduce biases by 'masking' epitopes of chromatin proteins. Alternate techniques such as N-ChIP, biotin-tag affinity purification, or DamID can overcome these problems (Mito *et al*., 2005; O'Neill and Turner, 2003; Van Steensel *et al*., 2000). The N-ChIP method uses native or uncross-linked chromatin and offers a major advantage in terms of antibody specificity as epitopes that are recognised by antibodies can be disrupted by formaldehyde cross-linking (O'Neill and Turner, 2003). However, N-ChIP can only be used to investigate histone proteins as the majority of non-histone proteins are not retained on the DNA during nuclease digestion. Biotin-tag affinity purification has been used to map histone variants by fusing a biotin ligase recognition peptide to the histone H3.3 protein and streptavidin pull-down achieves high-specificity (Mito *et al*., 2005). DamID maps DNA binding proteins by fusing a protein of interest to DNA adenine methylase, which then methylates adenine bases at binding sites. These sites are then identified by digestion with adenine methylation sensitive restriction enzymes (Van Steensel *et al*., 2000).

Secondly, because of the small DNA yields obtained after a ChIP reaction, immunoprecipitated DNAs are usually PCR amplified (Horak *et al*., 2002). This may result in amplification bias and an increase in false positives and false negatives. Alternatively, many sample DNAs are pooled (Weinmann *et al*., 2002), before being labelled with fluorescent cyanine-conjugated dyes and hybridized to DNA microarrays. Thirdly, the number of cells needed for a ChIP-chip assay is somewhere between $10^7$ and

$10^8$ for a single assay. This constraint prevents the analysis of cell populations where cell numbers are limited or rare - for example, cells found in the early stages of embryonic development. A modified ChIP method has recently been developed, which allows histone modifications to be studied from as few as 100 mouse embryonic stem cells (O'Neill *et al.*, 2006). This carrier ChIP (CChIP) procedure involves mixing a large number (5 x $10^7$) of *Drosophila* cells with a small number ($10^2$ - $10^3$) of mammalian cells before preparing nuclei and chromatin (Figure 1.9). Native chromatin fragments were prepared by nuclease digestion and immunoprecipitated with an antibody to a histone modification. Mammalian DNA fragments were quantified by radioactive PCR, electrophoresis and phosphorimaging. The presence of a large excess of *Drosophila* DNA in the ChIP DNA samples may prevent this method from being used in combination with microarray hybridization to identify interactions in a high-throughput manner as fluorescently labelled nucleotides would be preferentially incorporated into *Drosophila* DNA samples during the labelling process.

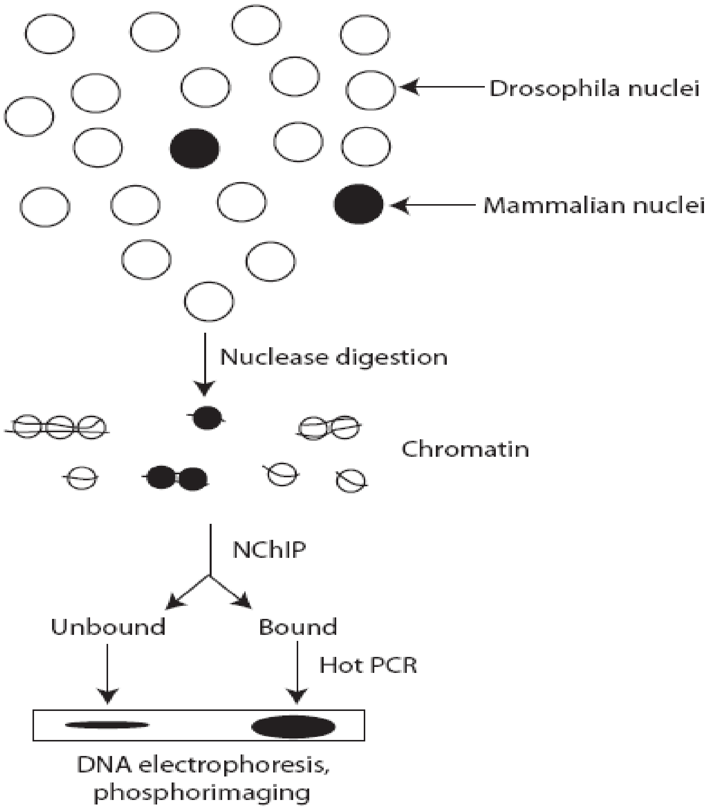

**Figure 1.9: Outline of the carrier ChIP (CChIP) method**. In the CChIP method, 5 x $10^7$ *Drosophila* SL2 cells are mixed with a small number of mammalian cells ($10^2$-$10^3$) and nuclei are prepared. Nuclease digested chromatin is then prepared and immunoprecipitated with an antibody to a specific histone modification. Mammalian antibody bound and unbound fractions are then quantified by radioactive PCR, electrophoresis and phosphorimaging. Bound/unbound ratios are used to represent histone modification levels.

### 1.7.4. Other ChIP-based methods

ChIP has also been combined with sequencing to determine the location of protein-DNA interactions. Sequencing can be performed from individually cloned ChIP fragments (Weinmann *et al*., 2001), from cloned concatenations of single tags, where each tag is a signature from a ChIP DNA (ChIP-STAGE). ChIP-STAGE has been used to map several histone modifications (Roh *et al*., 2005, 2006) and transcription factor binding sites (Impey *et al*., 2004). ChIP in combination with sequencing of concatenated paired-end ditags (ChIP-PET) has also been used to map transcription factor target sites in the human genome (Wei *et al*., 2006; Loh *et al*., 2006). In both ChIP-PET and ChIP-STAGE methods, relative tag representation is used to calculate binding enrichment.

Advances in sequencing technology have seen ChIP combined with massively parallel sequencing (ChIP-Seq) to identify various protein-DNA interactions across the entire human genome. Unlike ChIP-STAGE or ChIP-PET, the ChIP-Seq method does not involve plasmid library construction as the very large number of short sequence reads produced by sequencing allows for the direct quantification of all DNA sequences present in a ChIP sample. Barski and colleagues recently generated high-resolution maps for the genome-wide distribution of 20 histone lysine and arginine methylation states as well as histone variant H2A.Z, RNA polymerase II, and CTCF across the human genome using massively parallel Solexa sequencing technology (Barski *et al*., 2007). This technology attaches randomly fragmented ChIP DNAs to an optically transparent surface, followed by solid-phase amplification of DNAs to create more than 10 million clusters which are then sequenced using four colour sequencing by synthesis technology. Short sequence reads are then aligned against the reference genome sequence to calculate relative enrichment levels in a ChIP sample. This method has also been used to map REST

(Johnson *et al*., 2007) and STAT1 (Robertson *et al*., 2007) transcription factor binding sites in the human genome.

## 1.8. Genomic microarray platforms used in this study

### 1.8.1. The SCL genomic tiling path microarray

The development of ChIP-chip technology was used at the Sanger Institute to investigate regulatory elements at the SCL locus (Dhami, PhD Thesis, University of Cambridge, 2005; Dhami, submitted). The Stem Cell Leukemia (SCL) gene (also known as TAL1) is a basic helix-loop-helix transcription factor (TF) that is considered to be a master regulator of haematopoiesis (Robb and Begley, 1997). Over-expression of the SCL gene is the most common molecular abnormality found in human acute T-cell leukaemia and this TF is required for the normal development of all adult haematopoietic lineages. A tiling-path microarray was constructed to understand the regulation of SCL during haematopoiesis. The construction of a sensitive array platform for the SCL locus was made possible by using the 5'-aminolink array surface chemistry developed at the Sanger Institute. This surface chemistry allowed for single-stranded DNA molecules (derived from double-stranded PCR products) to be retained on the surface of a glass slide (Dhami *et al*., 2005). A 5'-(C6) amino-link modification is incorporated at the end of one strand of DNA, which allows the modified strand to be covalently attached to the surface of the slide (Figure 1.10). During slide processing, chemical and physical denaturation removes the unmodified strand, while the strand attached to the slide is preserved. The single-stranded DNA molecules provide an ideal hybridisation target for a labelled DNA sample.

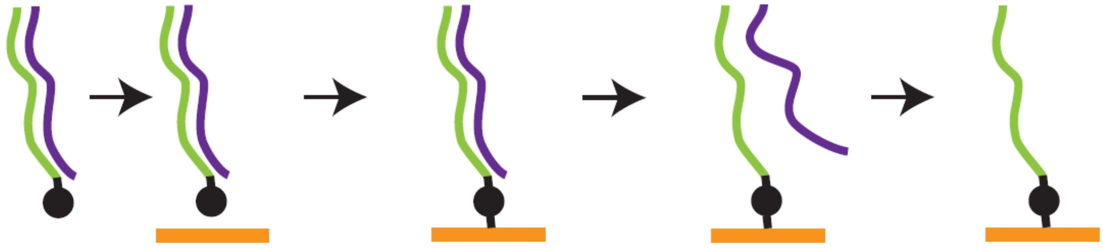

**Figure 1.10: Microarray surface chemistry.** Double-stranded PCR products (denoted by green and purple strands) containing a 5'-(C6) amino linker on one strand (black circle) are arrayed onto the surface

of the slide (orange rectangle). Covalent attachment of the PCR products occurs via the 5'-amino-link and the surface of the amine coated slide. Denaturation removes the strand that is not covalently attached to the slide surface (purple strand), resulting in a single-stranded DNA probe.

The genomic regions represented on the SCL tiling path array included the SCL gene, flanked upstream by SIL and KCY and downstream by MAP17, CYP4A22 and CYP4Z1 genes (Figure 1.11). The tiling path array covered 256 kb of human chromosome 1, with 419 PCR amplicons designed at an average product size of 458 bp.

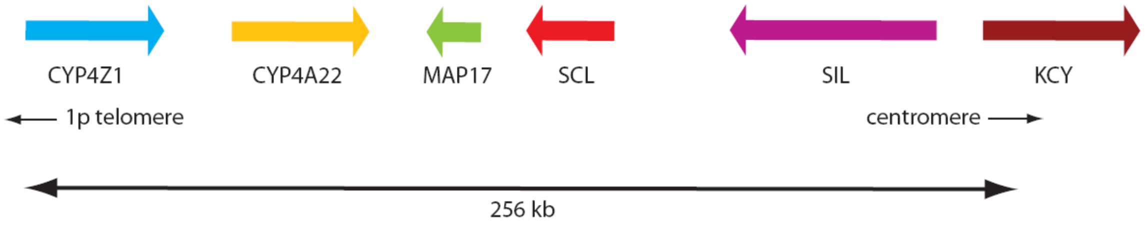

**Figure 1.11: The genomic region included on the human SCL tiling path microarray.** The genomic region contained on the SCL tiling path microarray is indicated by double-headed arrows. This 256 kb region contains 6 genes, represented by coloured arrows. The tiling path covers the entire length of CYP4A22, MAP17, SCL, and SIL genes, while CYP4Z1 and KCY genes are partially covered. The gene order and direction of transcription is indicated by coloured arrows. The region is annotated on the negative strand of chromosome 1, and the orientation, with respect to the 1p telomere and centromere, is indicated by black arrows.

The SCL tiling path array and ChIP assays were used to identify a number of protein-DNA interactions which defined regulatory elements in the locus (Dhami, PhD thesis, University of Cambridge, 2005; Dhami, submitted). ChIP-chip assays were used to detect regions enriched for H3 K9/K14 diacetylation (H3 acetylation) and H4 K5/8/12/16 tetra-acetylation (H4 acetylation) at the SCL locus (Figure 1.12). It was observed that the most prominent enrichments for H3 acetylation and H4 acetylation were located at known and novel promoters. Lower enrichments were also detected at known and novel enhancer elements. The methylation status of histone H3K4 was also investigated. H3K4me1 was found to be enriched at known and novel enhancers (Figure 1.12). It was also shown that H3K4me2 and H3K4me3 occurred at the promoters of transcriptionally active genes across the SCL locus.

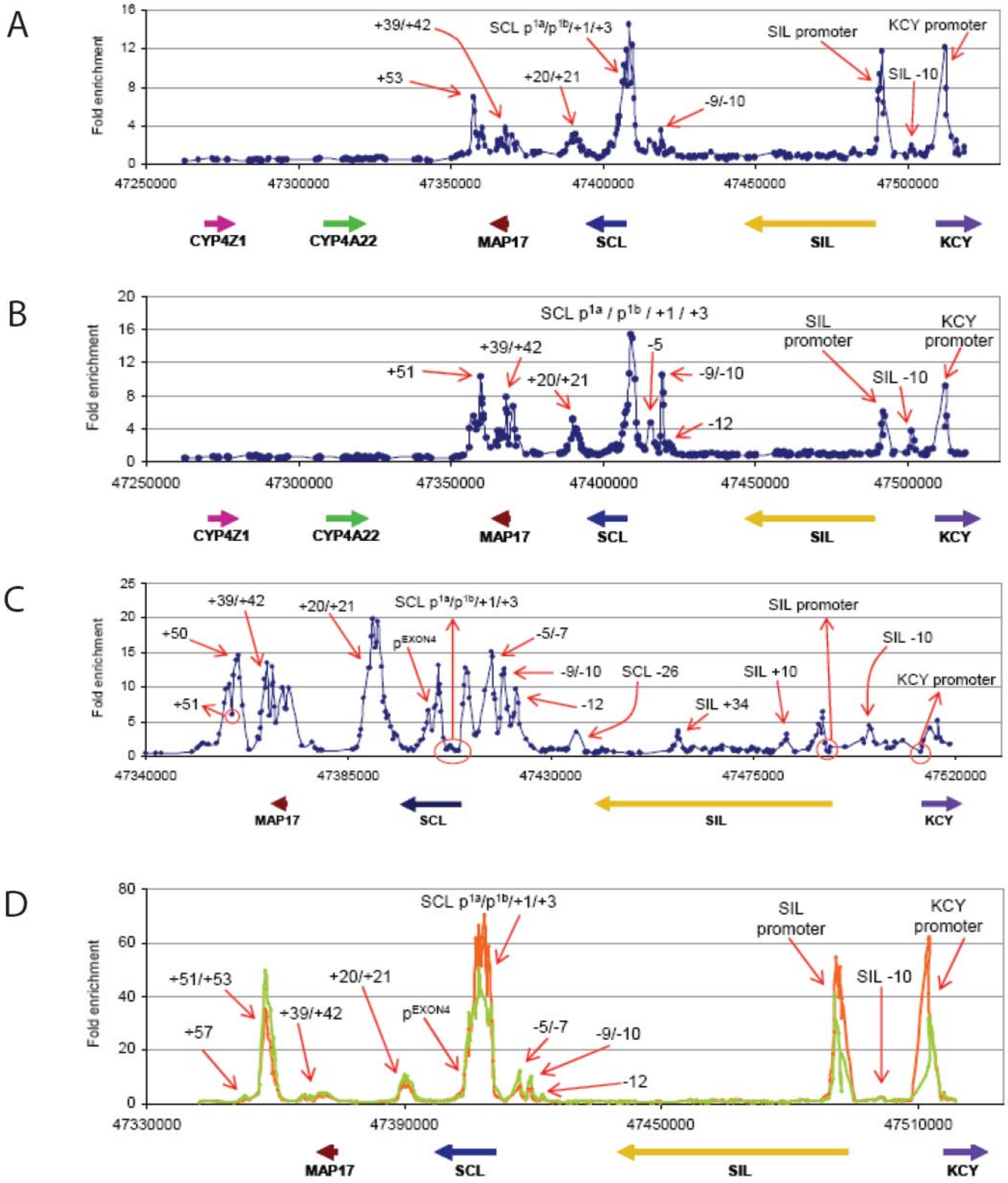

**Figure 1.12. Histone acetylation and methylation states define regulatory elements at the SCL locus**. Panels A and B show the histone H3 acetylation and H4 acetylation profiles across the SCL locus in K562 cells. Panel C shows the H3K4me1 profile across the MAP17, SCL, SIL and KCY genes in K562 cells. Panel D shows the H3K4me2 (green) and H3K4me3 (orange) profiles across the MAP17, SCL, SIL and KCY genes in K562 cells. The location of promoter and other regulatory elements (many of which were already known) are indicated by red arrows. The x-axes represent genomic coordinates along human

chromosome 1 and fold enrichments are displayed on the y-axes. The regulatory elements are denoted based on their distance upstream (-) or downstream (+) in kilobases from the SCL promoter 1a. The coloured arrows below each panel represent the gene order and direction of transcription. Figure from Dhami 2005 thesis.

Those previously known regulatory elements detected by these ChIP-chip assays included (names based on distances upstream (-) or downstream (+) in kilobases from SCL promoter 1a):

(i) **Promoters**: SCL has three promoters p1a, p1b (Aplan *et al*., 1990) and pEXON4 (Bernard *et al*., 1992). p1a and p1b are active in erythroid and megakaryocytic lineages while pEXON4 is active in leukaemic T-cells.

(ii) **Stem cell enhancer**: The stem cell enhancer is located at +20/+21 and directs SCL expression to most haematopoietic progenitors and endothelium (Sanchez *et al*., 1999; Sanchez *et al*., 2001; Gottgens *et al*., 2001; Pimanda *et al*., 2006).

(iii) **Neural regulators**: Regulatory elements located at SCL p1a, +1, and +3 direct SCL expression to regions within the brain and spinal cord (Sinclair *et al*., 1999).

(iv) **Erythroid enhancer**: The +51 erythroid enhancer targets SCL expression to primitive erythroblasts (Delabesse *et al*., 2005).

(v) **-9/-10 enhancer**: The -9/-10 region shows enhancer activity in reporter assays (Gottgens *et al*., 1997).

### 1.8.2. The ENCODE project

#### 1.8.2.1. A summary of the pilot phase findings

In 2003, an international consortium of research groups established a pilot study to evaluate a number of experimental approaches to catalogue all functional elements in 30 Mb (1%) of the human genome, comprised of 44 distinct genomic regions (The ENCODE Project Consortium, 2004) (see Chapter 3). The goal of this project was to develop efficient approaches for the large-scale identification and characterisation of regulatory elements, with the expectation of adopting these methods to analyse the whole genome. Thirty of the 44 regions were randomly picked by the ENCODE consortium to satisfy various non-exonic conservation and gene-density rates across the genome. The remaining 14 regions were manually chosen because of their important biological or pathological role - and extensive regulatory information already exists for some of these regions. These regions include the CFTR locus, the interleukin cluster, the α- and β-globin loci, the HOXA cluster and the IGF2/H19 imprinted region. The inclusion of these regions allowed for data obtained from the ENCODE project to be validated with respect to previously characterised regulatory elements. Methodologies used by groups in the consortium included using tiling microarrays to identify transcribed regions (Emanuelsson *et al*., 2007), high-throughput mapping of DNase I hypersensitive sites (Sabo *et al*., 2006; Crawford *et al*., 2006), comparative sequence analysis (Margulies *et al*., 2007; King *et al*., 2007), computational analysis (Greenbaum *et al*., 2007; Bajic *et al*., 2006; Zheng *et al*., 2007) replication timing assays (Karnani *et al*., 2007), and ChIP-chip assays to detect histone modifications (Koch et al., 2007; Rada-Iglesias *et al*., 2007) and sequence-specific transcription factors (Bieda *et al*., 2006). Over 200 data sets were generated by the consortium members and analysed (Birney *et al*., 2007). The principle findings of this analysis are summarised as follows:

(i) The majority of the human genome sequence is transcribed.

(ii) Many non-coding transcripts were identified, many of which overlapped with coding regions.

(iii) Many novel transcription start sites were identified, many of which were associated with a chromatin structure similar to well characterised promoters.

(iv) Chromatin accessibility and histone modification patterns can be used to accurately predict the location and activity of transcription start sites.

(v) Distal sites are associated with a characteristic histone modification pattern

(vi) Replication timing correlates with chromatin structure.

(vii) The majority of evolutionarily constrained sequences are associated with an experimentally determined function while many other functional elements are not under evolutionary constraint.

#### 1.8.2.2. The Sanger Institute ENCODE Microarray

Identification of regulatory elements in the ENCODE regions at the Sanger Institute focused on using ChIP-chip assays to detect a wide range of DNA-protein interactions. An array containing the 44 regions was constructed at the Sanger Institute (Koch *et al*., 2007). Double-stranded PCR products were spotted on microarrays using the same 5'-aminolink array surface chemistry used to construct the SCL tiling path array, which was then processed to generate single-stranded DNA probes. The Sanger Institute ENCODE microarray consisted of 24,005 PCR fragments with an average size of 1024 bp (average non-overlapping tile length = 992 bp). The array covered approximately 80% of the targeted regions and over 90% of non-repetitive regions. The Sanger Institute ENCODE array provides a new resource for investigators interested in identifying functional elements, and formed the basis for much of the work presented in this thesis.

## 1.9. Aims of this thesis

At the time this PhD project was initiated, there was relatively little information known about non-coding regulatory elements across the human genome. Furthermore, given some of the limitations of high-throughput approaches such as ChIP-chip (discussed in section 1.7.3.5), it was necessary to improve existing methods in order to identify and characterise non-coding elements in a systematic way. Therefore, with these views in mind, the aims of this thesis were as follows:

1. To use existing ChIP-chip approaches to characterise a variety of types of regulatory elements (promoters, enhancers and insulators) across selected regions of the human genome.
2. To develop further existing ChIP-chip approaches in order to improve sensitivity of the method when using cell types which are limiting in number.
3. Having improved ChIP-chip for aim 2, to then apply these methods to study cell types which are limiting in number.
4. To analyse the ChIP-chip data obtained for non-coding regulatory elements in the human genome and thereby understand fundamental principles of gene regulation.