Chapter 1

Introduction

A genome is the entire complement of the genetic material in a cell for an organism. Broadly speaking, a genome contains inherited information encoded in three types of elements – genes, regulatory elements and maintenance elements. Genes carry information to code for proteins, which are the building blocks of a cell, but also for RNA molecules of the ribosomes and other ribonucleoproteins. Regulatory elements contain information on how the genes are regulated to produce proteins with respect to time or developmental stage, location and quantity. These include, but are not limited to, promoters, enhancers and other regulatory elements and may also include regulatory RNAs. In addition to these, further regulation of genes is also achieved by modification of the DNA by methylation or by modification of the chromatin. Finally, the maintenance elements contain information for DNA repair, replication and recombination within the genome. These include centromeres, telomeres, origins of replication, replicons and recombination hotspots. The availability of the complete sequence of a genome allows the identification of all these elements - this would greatly facilitate the understanding of fundamental biological processes.

The Human Genome project was launched in 1990 with the goal of determining complete and accurate sequence of the 3 billion DNA subunits. The release of the finished version of the sequence in 2004 [IHGSC, (2004)] was an important milestone in this historic project and indeed the beginning of a very exciting era for scientists worldwide. In the post-genome era, the next challenge lies in identifying and annotating the human genome for all the above mentioned elements and characterizing the complex cellular events that are associated with them.

1.1. Coding regions of the human genome

Genes are basically sections of the DNA molecule that carry the instructions to code for a specific protein. In eukaryotes, the genes are not usually continuous and are composed of alternating sections of exons and introns. One of the key challenges in the post-genome era is to develop a definitive catalogue of protein-coding genes. Currently, the number of human genes is estimated to be between 20,000 and 25,000 [IHGSC (2004)] whereas simple organisms such as *S. cerevisiae* (yeast), *D. melanogaster* (fruitfly) or *C. elegans* (nematode worm) have only 6000, 13000 and 19000 genes respectively (Goffeau et al. 1996; Adams et al. 2000; The *C. elegans* sequence consortium 1998). These numbers seem surprising as the human genome is

around 30 times larger in size than the worm and fly genomes and 250 times larger than that of the yeast even though the number of genes in human is only two to three times more than that of the other organisms.

The large size of the human genome as compared to the total number of genes predicted so far implies that the gene density in the human genome is very low which makes finding the genes much more difficult. In addition to experimental approaches, several software programs such as GeneWise, Genomescan, and GENSCAN, among many others, provide fast and accurate computational gene predictions (Birney and Durbin 1997; Yeh et al. 2001; Burge and Karlin 1997). Many methods for predicting genes are based on compositional signals such as splice sites and coding regions associated with genes, similarity comparisons with expressed sequences (expressed sequence tags – ESTs and cDNAs) and proteins from human and other organisms.

It is possible, however, that the human genome contains a higher number of genes than what has been predicted. In recent years, there is increasing evidence that in addition to the protein-coding genes, the human genome also contains non-coding RNA transcripts (ncRNAs) including antisense RNAs (Kapranov et al. 2002; Cawley et al. 2004; Rinn et al. 2003). Hence, it seems that there is a great need for developing better gene-prediction programs that will aid in arriving at a more accurate number of genes in the human genome.

1.2. Non-coding regions of the human genome

With the genome sequence at hand, it is now known that the protein-coding sequences account for only 2% of the human genome and the remaining 98% is non-coding (Shabalina and Spiridonov 2004). Contained within the genome sequence are DNA elements that have functional importance - these include the regulatory and the maintenance elements described above. As much as a third of the human genome is believed to be involved in controlling processes such as gene expression, chromosome replication, condensation, pairing and segregation (Levine and Tjian 2003). The elucidation of the non-coding regulatory elements controlling gene expression will be discussed in detail in the later sections of this chapter. For maintenance elements, centromeres and telomeres are well characterized in terms of sequence content. Centromeres and telomeres are specialized structures that are involved in replication, with the telomeres also being responsible for the stability of linear DNA molecules. It is widely believed that the temporal order of replication is strictly regulated with some regions of the genome replicating much earlier than others (Raguraman et al. 2001; Schubeler et al. 2002; Woodfine et al. 2004). Replication has also been shown to

correlate with parameters related to gene activity, chromatin structure and nuclear position. However, not much is known about replication origins, replicons, and recombination hotspots in the human genome. Thus, in order to fully comprehend key biological processes as well as organismal complexity and diversity, it is important to fully annotate and characterize the coding as well as the non-coding regions of the human genome in greater detail.

1.3. Understanding the regulation of gene expression

In multi-cellular eukaryotic organisms, such as humans, cells with identical genetic information develop and differentiate, and the fate of each cell type is mapped out to perform a different yet critical function. The diversity of cellular function is dependent upon distinct and appropriate gene expression within the cell in response to external and internal signals. Some of the genes are expressed in all of the cells at all times, some are expressed as a cell enters a particular pathway of differentiation, and some genes are expressed only as the conditions in and around the cell change. However, to fully understand the underlying principle of genes switching on or off, it is essential to know when, where or to what extent a gene is expressed. In other words, how is the activity of the gene regulated and what sequence features are involved?

Every gene has its own "*cis*-acting" sequence elements, which represent sequences in the vicinity of the structural portion of the gene, to regulate its expression. They vary greatly in complexity from one gene to another and range from simple structures in bacteria to more complex structures in mammalian systems. Furthermore, "*trans*-acting" factors, or transcription factors and other proteins, which are encoded by genes located at genomic locations distant from their targets, are activated in response to intra-cellular or extra-cellular signals and bind to the *cis*-acting sequences to control gene expression. There are numerous permutations and combinations as to how a gene is regulated at this level. For example, a transcription factor can act on *cis*-elements of multiple genes; or it can interact with other transcription factors to form a complex to associate with the *cis*-elements of one or multiple genes. In addition, there are often dynamic feedback loops that provide for further regulation. In summary, transcriptional activation and regulation of a eukaryotic gene involves the systematic assembly of *trans*-acting regulatory proteins on *cis*-acting regulatory sequences.

1.4. Types of non-coding regulatory elements

The highly structured, non-coding *cis*-acting regulatory elements, such as promoters, enhancers, silencers/repressors, locus control regions (LCRs), insulators, matrix attachment regions, to name a few, play pivotal roles in regulating the level, location and

chronology of expression of any given gene. These regulatory sequences act in a co-ordinated manner and are often scattered over distances of several tens to several hundreds of kilobases. Some of these non-coding regulatory elements associated with a typical human gene are illustrated in Figure 1.1 and described below.

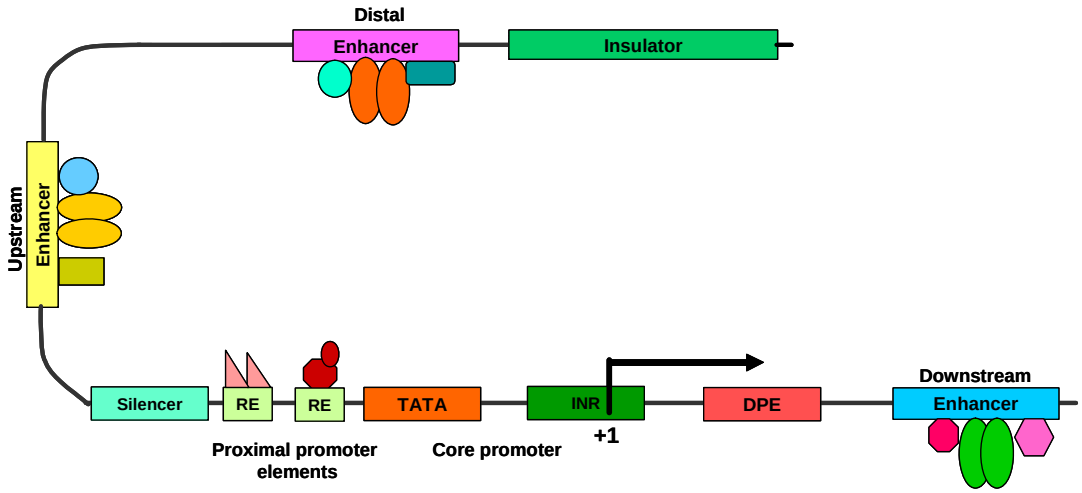

Figure 1.1: A schematic representation of the non-coding regulatory elements involved in the regulation of gene expression. The core promoter of the gene illustrated above contains a TATA box, an initiator element (INR) and a downstream promoter element (DPE), which can recruit the TATA binding protein (TBP) containing TFIID initiation complex (see text). The enhancers, insulators and silencers associated with the gene are shown by different coloured boxes. RE represents any additional regulatory elements, such as the upstream promoter elements. The +1 represents the transcriptional start site of the gene and the direction of transcription of the gene is shown by the black arrow. The coloured shapes (oblong, round, rectangle, triangle, hexagon etc.) present at some of the regulatory elements represent the proteins associated with each of these elements. The thick black line joining the boxes represents the DNA strand. This figure has been adapted from Levine and Tjian 2003.

1.4.1. Promoters

The promoter is required for accurate and efficient initiation of transcription – the process whereby a gene is encoded as a messenger RNA (mRNA). The core promoter includes DNA elements that can extend approximately 30 base pairs (bp) upstream and/or downstream of the transcription start site (labelled as +1 in Figure 1.1). A typical basal or core promoter contains a sequence of seven bases (TATAAAA) designated as the TATA box, which functions primarily to ensure that transcripts are accurately initiated. However, not all protein-coding human genes have been found to contain a TATA box (Suzuki et al. 2001). Additionally, the core promoter also contains one or more sequence elements of 8 to 12 base pairs called the upstream promoter elements (UPEs) that help to increase the rate of transcription. A number of different UPEs have been identified, for example, the CCAAT box is found in many promoters. There are at least three different sequence elements that can recruit the TATA-binding protein (TBP) containing TFIID (transcription factor-IID) initiation complex to the promoter: the

TATA box, the initiator element (INR) and the downstream promoter element (DPE) (Smale and Kadonaga 2003). TATA-less promoters use combinations of INR and DPE elements for the transcription machinery to bind (reviewed in Hahn 2004). In vertebrates, core promoters are rarely sufficient to drive faithful expression of a gene. More often than not, other *cis*-regulatory elements are required to bring about not only correct spatial and temporal expression of genes, but also increased rate of transcription.

1.4.2. Enhancers

A typical enhancer is approximately 500 base pairs in length and can interact with the *cis*-linked promoters at great distances in an orientation-independent manner. A typical vertebrate gene is likely to contain several enhancers that can be located 5' or 3' to the gene, as well as within the introns. One of the first enhancers to be characterized was in the simian virus (SV40) (Banerji et al. 1981) and the first cellular enhancer was discovered in the immunoglobulin (Ig) heavy-chain gene (Banerji et al. 1983). This was the first genetic element described to have a cell-type specific enhancer activity and was located within the gene. Inducible enhancers respond to changes in the environment, for example, the virus-inducible interferon gene regulatory element (IRE) in the human β-interferon (β-IFN) gene. Temporal and tissue-specific enhancers are active only at specific times during development or only in specific tissues, for example, the lymphoid cell-specific expression of immunoglobulin (Ig) genes (Maniatis et al. 1987).

Two models have been proposed to explain the mechanism by which distal enhancers communicate with promoters to achieve the desired level of gene expression:

i) ***Non-contact models*** propose that enhancers act at a distance to create a favourable environment for gene transcription, or act as entry sites or nucleation points for factors that ultimately communicate with the gene (Martin et al. 1996; Bulger and Groudine 1999).

ii) ***Contact models*** propose that enhancers communicate with the promoters through direct interaction by various mechanisms that 'loop-out' the intervening sequences (Choi and Engel 1988; Mueller-Storm at el. 1989). Even though a lot of attention has been focussed on the contact models, the existence and nature of such long range interactions *in vivo* is still speculative.

1.4.3. Silencers or repressors

These elements have the opposite effect of an enhancer and are involved in repressing transcription by interacting with repressor proteins. For example, REST protein (repressor element-1 silencing transcription factor) is recruited to a 21-bp DNA repressor element known as RE-1 (repressor element-1) and acts as a transcriptional repressor by blocking the expression of many neuronal genes, containing these RE-1 sites, in non-neuronal cells (Bruce et al. 2004).

1.4.4. Insulators

Some DNA sequence elements, known as insulators, possess the ability to protect genes from inappropriate signals emanating from their surrounding regulatory environment. An insulator, associated with a gene, protects it from its surroundings in one of two ways. The first way is by blocking the action of a distal enhancer on a promoter. However, this enhancer blocking only occurs if the insulator is situated between the enhancer and the promoter (Zhao and Dean 2004). The second way is by acting as a "barrier" that prevents the spreading of nearby condensed chromatin which might otherwise result in silencing of gene expression (West et al. 2002).

1.4.5. Locus control regions (LCRs)

These regions are operationally defined by their ability to enhance the expression of linked genes to physiological levels in a tissue-specific and copy number dependent manner at ectopic chromatin sites**.** The first LCR was found in the human β-globin locus (Grosveld et al. 1987).

1.4.6. Scaffold/Matrix attachment regions (S/MARs)

S/MAR regulatory elements are abundant in the eukaryote genomes and are involved in the co-ordination of gene expression within physical locations of the nuclear environment. Thus, S/MARs are genomic DNA segments which provide the anchor point for looped chromatin domains by attaching them to the nuclear matrix. The domain sizes range from a few kb to more than 100 kb; this results in the placement of a gene in close proximity to its transcription factors, providing an essential step to expression (Bode et al. 1996). The S/MARs can shield gene expression from position effects and increase transcription initiation levels (Mielke et al. 1990).

1.5. Proteins involved in transcriptional regulation

A key characteristic of the regulatory elements, mentioned in the above sections, is that they contain discrete clusters of binding sites for different sequence-specific

transcription factors which could activate or repress gene transcription. The *trans*-acting factors or transcription factors (TFs) select the genes to be activated and choreograph the assembly of a 'transcriptional machine'. Transcription factors can be grouped based on their involvement in transcription initiation:

i) the basal or general transcription factors (TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH) which assemble together with RNA polymerase to form a multi-protein complex (the pre-initiation complex or PIC) on the core promoter
ii) the sequence-specific transcription factors that have distinct domains with specific functions to regulate the formation or function of the PIC
iii) protein complexes that remodel or modify chromatin

The latter two groups act at the promoter or at other regulatory elements.

1.5.1. RNA polymerase and basal transcription factors

At the promoter (see section 1.4.1), random interactions between RNA polymerase (RNA Pol) and DNA lead to initiation of transcription; however, several general transcription factors (listed above as type (i)) are necessary for RNA Pol to recognize and bind tightly to the promoter. The transcription of a gene proceeds in three distinct phases: (i) initiation (binding of RNA polymerase to template DNA); (ii) elongation (formation of the RNA); and (iii) termination (RNA polymerase and the resulting RNA are released from the DNA template). Eukaryotes use three nuclear enzymes, Pol I, II and III, to synthesize different classes of RNA and each has their own set of associated general transcription factors. RNA Pol I transcribes ribosomal RNA (rRNAs – 28S, 18S, 5.8S), Pol II transcribes messenger RNA (mRNA) and some small nuclear RNAs (snRNAs) and Pol III transcribes transfer RNA (tRNA), 5S rRNA and U6 snRNA (Holstege et al. 1998). Of the three, the RNA Pol II transcription machinery is the most complex with a total of nearly 60 polypeptides (Hahn 2004) having thus far been characterized.

The description of the transcriptional process provided here is focussed on transcription by RNA pol II. The binding of RNA Pol II depends on an associated multi-subunit complex, TFIID, which is composed of TBP and TBP-associated factors (TAFs). Prior to the transcription of a gene, gene-specific regulatory factors bound near the site of transcription initiation interact either directly with the components of the transcription machinery or indirectly by recruiting factors that modify chromatin (see section 1.5.3). This, in turn, leads to the binding of the general transcription factors to the core promoter, forming a pre-initiation complex (PIC). RNA Pol II then binds tightly to the general factors and the promoter, but is not in an active conformation to start

transcription. Initiation of transcription actually begins with the unwinding and disassociation of the duplex DNA for 11-15 bases, surrounding the transcription site. The single-stranded DNA is positioned in the active site of RNA Pol II, and only then, the first phosphodiester bond of RNA is synthesized. After synthesis of approximately 30 bases of RNA, RNA Pol II is thought to release its contact with the promoter and enters the stage of elongation (Hahn 2004).

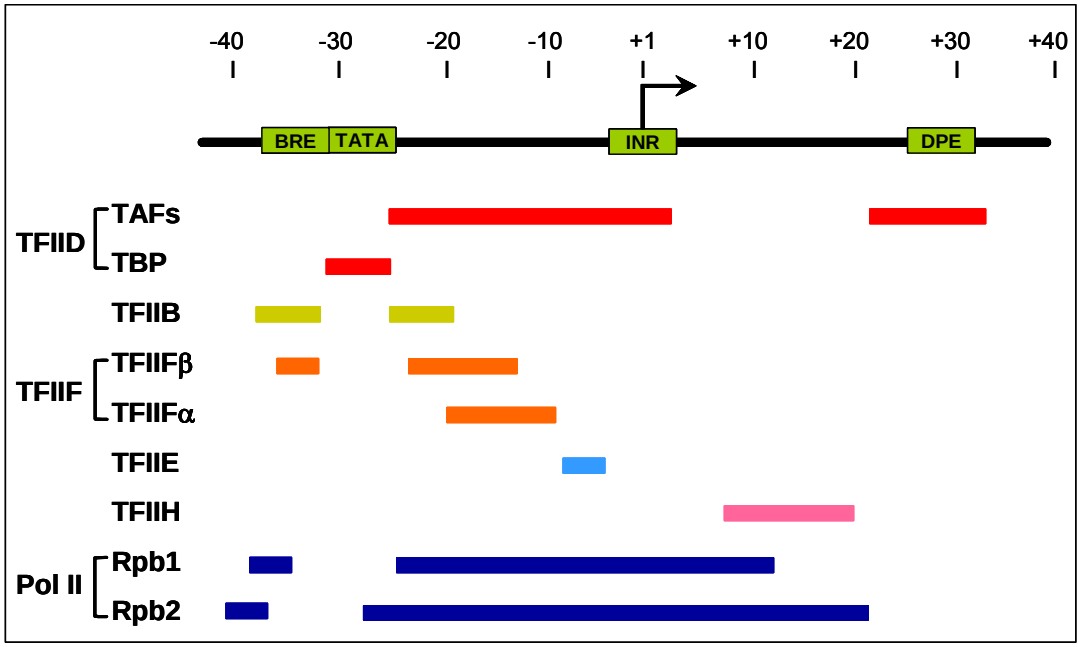

Figure 1.2: Summary of the human general transcription factor interactions at a promoter. The figure shows those proteins which interact with RNA pol II. The black, thick line at the top represents promoter DNA with the position of functional elements (scale at the top) indicated by the green boxes with the annotation. BRE: TFIIB recognition element, TATA: TATA box, INR: initiator element, DPE: downstream promoter element. The position of the transcription start site at +1 is shown by a black arrow. The proteins involved in the formation of pre-initiation complex are listed on the left. The coloured bars indicate the locations of the interactions of the respective factor with the promoter region. The description is provided in the text. This figure has been adapted from the paper: Hahn 2004.

In order to understand the orchestration of the PIC, it is important to know the functions of each of its components (listed below). The transcription machinery makes extensive interactions with the promoter DNA between positions -43 and +24 with respect to the transcription start site (see Figure 1.2).

i) First and foremost, TFIID (a complex composed of TBP and ~14 TAFs) binds the TATA element via TBP and interacts with the downstream promoter element (DPE) and initiator element (INR) through different TAFs (Smale and Kadonaga 2003). The TAFs are important for promoter recognition, in addition to both positive and negative regulation of transcription.

ii) TFIIA stabilizes TBP-DNA binding and strongly promotes binding of TFIID to DNA.

iii) TFIIB interacts with DNA on either side of the TATA element and can also make base-specific contacts with BRE (TFIIB-recognition element).

iv) TFIIE interacts with the promoter DNA just upstream of the transcription start site (+1).

v) The small subunit of TFIIF (TFIIFβ) interacts with DNA on either side of the TATA element whereas; the large subunit of TFIIF (TFIIFα) interacts with DNA downstream of the TATA element.

vi) TFIIH interacts downstream and possibly upstream of the transcription start site (reviewed in Hahn 2004).

It is evident that each of the general transcription factors plays a simple but essential role in the initiation mechanism. It is also noteworthy that the promoter elements, such as the TATA-element, DPE and INR not only serve as binding sites for the subunits of the transcription machinery, but also aid in orienting the transcription machinery to direct unidirectional transcription. The carboxyl-terminal domain (CTD) of RNA Pol II is critical for elongation and undergoes phosphorylation during this process. CTD also acts as a platform for the assembly of factors that regulate elongation and termination. The mechanism of termination, however, is not very clear.

In mammalian systems, there are at least six TAFs with alternative subunits that can change the composition of the TFIID complex. These unique TFIID complexes function either in specific cell-types or at specific developmental stages (Levine and Tjian 2003). Furthermore, molecular studies have shown that promoters vary widely in the requirement for TAFs to promote normal gene regulation. Therefore, diverse TFIID complexes can function at distinct promoters through the use of tissue-specific TAFs (Freiman et al. 2001).

1.5.2. Sequence-specific transcription factors

In the elaborate mechanism of gene transcription and regulation, described so far, the sequence-specific transcription factors (referred to as “transcription factors” or TFs hereafter) function to initiate, enhance, or inhibit the transcription of a gene. It is obvious that specificity of gene expression is primarily achieved by combinatorial control which means that the transcription factors interact, directly or indirectly, with other factors that are simultaneously or sequentially bound to the *cis*-regulatory elements of the same gene. There may be as many as 3,000 transcription factors in humans (Lander et al. 2001). Given the fact that there are only 20,000-25,000 genes in the human genome (IHGC, 2004), means that there appears to be one factor for every

ten genes. Put together, this concept of one factor per ten genes and the combinatorial nature of transcriptional regulation could mean a dramatic expansion in regulatory complexity.

In order to gain an insight into how the transcription factors work, it is important to first understand their structure and organization. Most of the transcription factors are modular. A typical TF has a DNA binding module linked to one or more activation or repression modules as well as perhaps a multimerization module and a regulatory module. This remarkable modular nature was first revealed in the yeast GAL4 transcription factor (Brent and Ptashne 1985). There is a huge structural diversity in these various TF modules. The helix-turn-helix motif was the first well defined DNA binding module which was originally discovered in prokaryotic DNA-binding proteins. However, many more DNA binding modules have been discovered which include the homeodomain (a variant of the helix-turn-helix, for example, OCT-2), zinc finger (for example, GATA family), leucine zipper (for example, Fos and Jun families) and helix-loop-helix (for example, MyoD, E2A) (Murre et al. 1989). Furthermore, other studies revealed activation modules with acidic motifs (Ma and Ptashne 1987; Hope et al. 1988), non-acidic motifs (example, specificity protein-1, Sp1) (Courey and Tjian 1988) and other motifs including proline-rich regions (Mermod et al. 1989) and hydrophobic β-sheets (Leuther et al. 1993).

In order to make sure that the regulation of gene expression is precise and tightly controlled, the transcription factors mostly function in clusters with other factors. The reason behind this lies in the fact that these factors generally bind to the DNA with relatively low specificity; therefore, binding of a single factor to the DNA does not confer the specificity required for the desired regulation. Hence, transcription factors functioning in a cluster would ensure that they bind to the DNA with high specificity. When these factors are part of such clusters, they also usually function synergistically and therefore activate transcription more strongly than a single factor functioning on its own (references in Kadonaga 2004). One of the ways by which transcription factors mediate in the transcriptional regulation of a gene is by recruiting transcriptional co-activators and co-repressors to the DNA template via protein-protein interactions. The co-factors subsequently act both directly and indirectly, to regulate the activity of the RNA Pol II transcriptional machinery at the core promoter (Levine and Tjian 2003).

Another mode of regulation by the transcription factors is to undergo posttranslational modifications. A very good example of this phenomenon is provided by the phosphorylation of the cyclic AMP response element binding protein (CREB). Cyclic AMP activates protein kinase A, which in turn phosphorylates CREB at serine residue

133. As a result of the phosphorylation, CREB becomes activated and stimulates transcription from cyclic AMP-responsive promoters (Gonzalez and Montminy 1989). Similarly, some transcription factors can be acetylated. For example, the acetylation of p53 increases the affinity of its binding to the DNA (Gu and Roeder 1997).

In addition to the structural and functional diversity, a further level of complexity is obtained by virtue of many transcription factors being members of multi-protein families. A few examples are: the GATA family (with GATA-1,-2,-3,-4,-5 and -6) of transcription factors, the Sp1 family of proteins, and the AP-1 family (containing Fos, Jun and other related proteins). Most often, the members of the same family display closely related or essentially identical DNA binding properties but have distinct activation functions. The molecular bases for these unique functional specificities need to be elucidated in detail but these distinct activation modules may be helpful in conferring cell- or tissue-specific regulation or even at different stages of development of an organism. Thus, it is quite clear that in order to orchestrate precise and appropriate expression of a gene, the sequence-specific factors or TFs must work in a synchronized manner with the transcription machinery as well as the co-factors.

1.5.3. Proteins involved in chromatin remodelling

For the transcription factors to begin the assembly of transcriptional machinery, they first need to gain access to the DNA template which is normally folded into compact chromatin fibres. This process referred to as chromatin remodelling or unfolding of the chromatin, is not an inherent feature of transcriptional activators - rather, they recruit specialized protein complexes to carry out this process. Two major classes of protein complexes regulate the accessibility of the DNA template to the DNA-binding TFs - chromatin remodelling complexes and the histone modifying enzymes.

1.5.3.1 Chromatin remodelling complexes

The chromatin remodelling complexes (also called the ATP-dependent remodelling complexes) utilize energy from ATP-hydrolysis to modify chromatin structure by remodelling nucleosomes. The chromatin remodelling complexes can be broadly categorized based on the identity of their catalytic ATPase subunit: SWI2/SNF2, ISWI family, Mi-2 family (Narlikar et al. 2002). The SWI/SNF and the ISWI-based family of complexes are the best studied families of remodelling complexes.

ATP-dependent complexes can move nucleosomes, thereby exposing DNA sequences to facilitate the binding of transcription factors. They can also create conformations where DNA is accessible on the surface of the histone octamer (Narlikar et al. 2002). A variety of assays have shown that in addition to mobilizing and repositioning the

nucleosomes (Whitehouse et al. 1999), the chromatin remodelling complexes can also transfer a histone octamer from a nucleosome to a separate DNA template and can facilitate access of nucleases to nucleosomal DNA. For instance, SWI/SNF complexes have the ability to cause conformational changes thereby increasing the DNase and restriction enzyme sensitivity of DNA sites within a mononucleosome (Kingston and Narlikar 1999). In addition, the remodelling complexes have the ability to create dinucleosome-like structures from mononucleosomes (Lorch et al. 1998; Schnitzler et al. 1998), and have the ability to cause changes in superhelicity by twisting the DNA which in turn, disrupts the histone-DNA contacts (Havas et al. 2000; Gavin et al. 2001).

The most obvious mechanism for increasing DNA exposure, by nucleosome repositioning, entails “sliding” of the DNA with respect to the histone octamer (Meersseman et al. 1992). The “sliding” mechanism involves identical amounts of movement of the entry and exit points of the DNA in the same direction. This results in an octamer that is transcriptionally repositioned and the DNA that was originally interacting with the histones becomes non-nucleosomal. All three families of ATP-dependent remodelling complexes can change the translational position of nucleosomes on DNA (Whitehouse et al. 1999; Langst and Becker 2001; Schnitzler et al. 2001). In addition to the sliding mechanism, other mechanisms can also result in changes in translational position: a conformationally altered nucleosome could collapse to a canonical nucleosome structure that has an altered position or there could be partial or complete release of the histone octamer followed by rebinding at a new location (Studitsky et al. 1994; Lorch et al. 1999).

1.5.3.2 Histone modifying enzymes

The other class of proteins, which are involved in chromatin remodelling, include the histone modifying enzymes. Unlike the chromatin remodelling complexes, which expose the underlying chromatin by directly mobilizing the nucleosomes; the histone modifying enzymes influence transcription indirectly by covalent modifications of the histones themselves. These proteins add or remove many chemical moieties at specific residues of the N-terminal tails of histones. The covalent modifications include acetylation, methylation, phosphorylation, ubiquitination, and ADP-ribosylation. Some enzymes in this class also alter the chromatin structure by directly affecting the DNA template (for example, methylation of CpG islands).

Hyperacetylation of the lysine residues in the N-terminal tails of the core histones was proposed to be involved in activation of transcription over 40 years ago and has subsequently been strongly correlated with active genes (Allfrey et al. 1964). The steady

state level of acetylation in transcriptionally active genes is maintained by the opposing actions of histone acetyl transferases (HATs) and histone deacetylase complexes (HDACs) (Reid et al. 2000; Vogelauer et al. 2000).

GNAT, MYST and p300/CBP are a few of the main families of HATs (Narlikar et al. 2002) and the members of these families share a highly conserved motif containing an acetyl-CoA binding site. They also always act as part of large complexes *in vivo*. It has been suggested that the different complexes have different subunit compositions and different histone specificities. Correspondingly, these complexes appear to be involved in distinct biological functions (Roth et al. 2001). For instance, studies have shown that the SAGA complex might function as a co-activator at the site of initiation, in addition to its acetylation activity. Similarly, mice that are homozygous for deletion of the HAT proteins p300, CBP, PCAF or GCN5 exhibit distinct developmental defects suggesting differences in function of these otherwise highly related HATs (Narlikar et al. 2002). The different functions of HAT complexes are likely to be caused by their non-HAT subunits. Each complex contains a specific set of non-HAT subunits, which might interact with different sequence-specific activators that target the complexes to distinct genes or they may also differentially modulate HAT activity. For example, GCN5-containing complexes have different substrate specifities than GCN5 on its own (Grant et al. 1999).

Mammalian HDAC proteins identified to date can be divided into three main classes – class I, class II and class III HDACs (Narlikar et al. 2002). Class I includes HDACs 1, 2, 3 and 8 which share homology in their catalytic sites; class II includes HDACs 4, 5, 6, 7, 9 and 10 (De Rujiter et al. 2003). The class III of HDACs is the conserved nicotimamide adenine dinucleotide-dependent Sir2 family of deacetylases (De Rujiter et al. 2003). Sir2 is involved in heterochromatin silencing at silent mating loci, telomeres and ribosomal DNA in yeast (Moazed 2001). The histone specificities of the class I HDACs are just beginning to be characterized. For example, the yeast homolog of HDAC1, Rpd3, deacetylates all sites except lysine 16 on histone H4 which is linked to heterochromatic silencing and the data are consistent with Rpd3 having distinct function from Sir2 deacetylases (Narlikar et al. 2002).

Specific residues on the N-terminal tails of histones can also be methylated (see section 1.7.2). The members of the family of histone methyltransferases (HMTs) (Lachner and Jenuwein 2002) which carry out methylation of the lysine residues of histones contain a conserved SET domain that is flanked by cysteine-rich regions (Rea et al. 2002). The methylases required to catalyze arginine methylation include the CARM1/PRMT1 family

of HMTs. Of all of the HMTs, Dot1 lacks a SET domain and targets the lysine 79 residue of histone H3 that resides within the core domain (Ng et al. 2002).

It was believed until recently that methylation marks are irreversible as enzymes responsible for removing methyl groups had not been identified. But this concept of irreversibility does not account for situations where rapid reversal of gene expression takes place. To solve this paradox, mechanisms including enzyme-catalysed demethylation, replacement of methylated histones by unmodified histones, and clipping of methylated histone tails have been proposed (Banister et al. 2002; Henikoff 2004). However, recently few studies have reported the identification of demethylases which implies that methylation might not be an irreversible process as believed thus far. The protein LSD1, a nuclear homolog of amine oxidases, was identified which specifically demethylates histone H3 K4 and thus functions as a histone demethylase (Shi et al. 2004). Additionally, it has been found that the human enzyme peptidylarginine deiminase 4 (PAD4/PAD14) can catalyse the conversion of methylated arginines to citrulline, providing yet another mechanism by which histone methylation can be controlled (Wang et al. 2004; Cuthbert et al. 2004).

Rsk-2 (Sassone-Corsi et al. 1999) and Msk1 (Thomson et al. 1999) have been identified as the kinases that carry out histone H3 phosphorylation. Overall, it is unknown whether there are a large number of kinases that target histones as substrates. In addition, the co-activator and acetylatransferase $Taf_{II}250$ has been identified as a histone H1 ubiquitin-conjugating enzyme (Pham and Sauer 2000).

1.6. Identification of genomic non-coding regulatory elements

It is evident that gene expression in eukaryotes is a highly controlled and co-ordinated 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 the level, location and chronology of gene expression. Therefore, in order to understand the regulatory networks which facilitate gene expression, it is important to identify and characterize the DNA regulatory elements associated with the genes. Over the years, various experimental systems and assays have been used to identify non-coding regulatory elements – the majority of which have been quite low-throughput and laborious processes. However, with the completed human genome sequence at hand, new analysis tools and the development of high-throughput methods, both experimental and computational, mean that it should be more efficient to identify and characterize these elements. All of these methods are discussed below.

1.6.1. Classical methods

1.6.1.1 DNase I hypersensitive assay

As discussed previously, nucleosomes undergo conformational changes or are repositioned making the underlying region of DNA free of nucleosomes and accessible to TFs (see also section 1.5.3.1). These nucleosome-free regions are sensitive to nuclease enzymes and are therefore, readily degraded by the enzyme DNase I. It is evident that regions identified owing to their DNase I hypersensitivity are involved in transcriptional regulation and most types of regulatory elements (such as promoters, enhancers, suppressors, insulators, and locus control regions) have been shown to be associated with DNase I hypersensitive sites (HSs) (Gross and Garrard 1988).

DNase I HSs in native genomic domains have traditionally been localized by cleavage of nuclear chromatin followed by DNA purification, restriction endonuclease digestion, gel electrophoresis, southern blotting and hybridisation with a radiolabelled DNA probe (Wu et al. 1979). The five 5' HSs of the human β-globin LCR were identified by DNase I hypersensitive analysis and were later characterized for their function by deletion analysis (Tuan et al. 1985; Grosveld et al. 1987).

Although mapping DNase I HSs to identify regions with potential regulatory activity has been widely used, this technique is low-throughput and time-consuming. More recently, however, several studies have described approaches to map DNase I HSs in a high throughput, sequence-specific manner which are easily scalable to genome-wide studies. These involve generating genome-wide libraries of regulatory sequences by cloning DNase I HSs (Crawford et al. 2004) and computational analysis of 'chromatin-profiles' which are generated across a region by real-time PCR (Dorschner et al. 2004).

1.6.1.2 DNA footprinting

The DNA footprinting assay is used to identify the binding sites of proteins or transcription factors (TFs) that bind to the DNA. It works on a simple principle: when a TF binds to the DNA, the contact point is usually over a few nucleotides. The bound protein, however, renders the underlying DNA segment relatively resistant to cleavage by DNase I when compared with naked DNA (Galas and Schmitz 1978).

Usually, short cloned DNA fragments (usually a few hundred bases long) are used as targets and are end-labelled at one end only. The cloned fragments are individually incubated in the presence or absence of a protein extract and subsequently exposed briefly to low concentrations of DNase I. Such partial digestion conditions ensure that, for any one DNA fragment, each DNA molecule is cut only rarely and at a random

position if no protein is bound. The digestion products are then size-fractionated on long denaturing polyacrylamide gels, prior to autoradiography. Control samples show a series of bands corresponding to DNA fragments of every possible length whereas the corresponding test lanes reveal gaps where no fragments are seen (footprints), as the DNase I is not able to cut at these positions because of steric inhibition by the bound protein.

DNA footprinting is useful only for the identification of the protein-DNA binding sites and does not provide any insight into the functionality of that site. It is apparent that such sites could be putative regulatory regions involved in regulation of gene expression and therefore, usually other methods are used in combination with DNA footprinting to further analyse the identified sites. For instance, 19 protein binding sites were identified in the 5' end of the human factor VIII gene by DNA footprinting. However, further analysis using gel-shift assays (see section 1.6.1.3) and deletion mutants revealed the functions of the identified sites which were involved in gene regulation in a tissue-specific manner (Figueiredo and Brownlee 1995).

1.6.1.3 Gel-shift assays

A gel-shift assay (also known as a gel retardation or electromobility shift assay) is a simple and highly sensitive technique for studying DNA-protein complexes utilizing the difference in migration speed of DNA-protein complexes and free DNA during electrophoresis. The DNA molecules that are bound by a protein complex migrate more slowly than the unbound DNA, in a non-denaturing gel matrix (Garner and Revzin 1981). In a typical gel shift assay, protein extracts are incubated with labelled DNA fragments from a genomic clone or labelled oligonucleotides, which are suspected of containing regulatory sequences. The resulting preparation is size-fractionated by PAGE (polyacrylamide gel electrophoresis) in parallel with a control sample in which the DNA was not mixed with the protein. DNA fragments which have bound protein are identifiable as low mobility bands (Garner and Revzin 1981).

The advantages of the gel-shift assay are:

i) it can be used to identify the presence of a specific interaction even in crude extracts

ii) it is relatively quick and simple

iii) only active protein is present in the complex bands as the free protein is separated (Carey 1991)

However, a possible caveat of this *in vitro* assay is that the identified binding site of a TF does not always reflect its actual binding to its target *in vivo* (Lieb et al. 2001).

1.6.1.4 Enhancer/promoter trap reporter assays

The enhancer trap assay was established to define the transcriptional activity of cloned genomic sequences (an enhancer or enhancer-like element) from a mixture of DNA fragments (Weber et al. 1984). This assay is based on an expression system in which the vector is designed so that the reporter gene is controlled almost entirely by the sequences integrated upstream of the gene. The reporter gene is under the control of an inactive, minimal or incomplete promoter and is not activated by this truncated promoter *per se*. However, the reporter gene is activated by the insertion of a putative regulatory element (enhancers, promoters, repressors) upstream of the promoter. The vector-reporter gene construct is then transfected into cultured target cells by methods such as electroporation or using liposomes. The reporter gene is expressed, and its level quantified, if all the necessary regulatory elements are present. Commonly used reporter genes in human cells include, the bacterial CAT (chloramphenicol acetyl transferase) gene, the β-galactosidase gene and the firefly luciferase gene. Luciferase has the added advantage of providing a very sensitive assay - it catalyzes the oxidation of luciferin with the emission of yellow-green light which can be detected easily and at low levels. (Alam and Cook 1990; Pardy 1994). These approaches have been used to characterize regulatory elements in a number of genes including SCL and c-Myc (Mautner et al. 1995; Sinclair et al. 1999; Gottgens et al. 1997).

After the identification of a putative regulatory element, it may be possible to further define the minimal region required by the regulatory element to drive transcription by generating a series of progressive deletion constructs. For example, a number of 5' and 3' putative regulatory elements have been identified for the stem cell leukaemia (SCL) gene which is expressed in haematopoietic development (discussed in section 1.9.5). A series of deletion constructs of a putative enhancer revealed that only a fragment of 640 base pairs was sufficient to enhance transcription in the experimental system in which it was tested. Additionally, this fragment also contained a cluster of binding sites for various haematopoiesis-related sequence-specific TFs (Gottgens et al. 2002).

1.6.1.5 PCR-based methods

PCR can be applied to amplify genomic DNA fragments that bind with proteins. A series of PCR-based methods have been developed, such as **s**ystematic **e**volution of **l**igands by **ex**ponential enrichment (SELEX) (Tuerk and Gold 1990), **t**arget **d**etection **a**ssay (TDA) (Thiesen and Bach 1990), **s**election **a**nd **a**mplification **b**inding site (SAAB), **c**yclic **a**mplification **s**election **t**argets (CASTing) (Wright et al. 1991), **mu**ltiplex **s**election **t**arget (MuST) (Nallur et al. 1996), and other similar methods (Xu et al. 2002).

All of these methods share some common steps. Random oligonucleotides or genomic DNA are mixed with purified proteins or nuclear extracts. Then, DNA-protein complexes are isolated by one of the various isolation methods such as immunoprecipitation, EMSA, filter binding assay, binding to affinity columns etc. The DNA recovered from the first cycle of selection is PCR-amplified, mixed with fresh proteins and the process is repeated. DNA fragments binding with specific proteins can be enriched after several rounds.

Another PCR-based method to detect genomic sequences showing *in vivo* DNA-protein interactions is to analyse chromatin immunoprecipitated (ChIP) samples (see section 1.6.4.3) by using sequence-specific primers. By comparing the enriched sample with the control sample (not enriched), the regions specifically bound by the proteins of interest can be identified (Delabesse et al. 2005).

1.6.2. Computational methods

Given that functional regulatory elements are embedded in non-coding sequence (which makes up 98% of the human genome), it would be no exaggeration to say that the amount of work required to identify all of the non-coding regulatory sequences by experimental approaches would be arduous. Thus, computational approaches could be employed to rapidly examine whole genomes for such features.

1.6.2.1 Comparative genomic sequence analysis

This is a powerful approach for the identification of new unknown regulatory elements. This approach involves comparative analysis of sequences from two or more genomes to identify the genome-wide extent of similarity of various features. With the completion of genome sequences of many organisms such as mouse (Waterston et al. 2002), rat (Gibbs et al. 2004), fugu (Aparicio et al. 2002), chicken and human (IHGSC, 2004), it has now become easier to compare the genome sequences to look for regions of high sequence conservation.

Initially, such comparative techniques were primarily applied to the coding regions of the genomes, to identify genes or exon-intron boundaries (Batzoglou et al. 2000). However, comparative analysis of human sequence with closely related species such as primates and mouse, as well as with distantly related non-mammalian species such as chicken and fish have greatly aided the identification of *cis*-regulatory elements based on evolutionary conservation (Boffelli et al. 2003; Lee et al. 2004; Loots et al. 2000; Lien et al. 2002; Bagheri-Fam et al. 2001). The reasoning behind the approach is that, just like coding sequences, regulatory elements are functionally important and are under evolutionary selection, so they should have evolved much more slowly than other non-

coding sequences. An example where comparative sequence analysis has aided in the identification of regulatory element has been the analysis of the SCL gene. Comparative sequence analysis of the genomic DNA containing the SCL locus from human, mouse and chicken identified several peaks of sequence homology which corresponded to a subset of the known enhancers of the SCL gene. However, one homology peak at +23 region (23 kb downstream of the SCL promoter 1a), which did not correspond to any known enhancers was tested in a transgenic reporter assay and was found to have enhancer activity (Gottgens et al. 2000).

In order to perform cross-species comparisons, several algorithms and software programs have been developed to detect conservation between species. For example, local alignment tools, such as BLAST (Altschul et al. 1990) is useful for comparing a short query sequence against a large sequence database to identify sequence similarities between the two. Software programs such as VISTA (Mayor et al. 2000), LAGAN (Brudno et al. 2003) and PIPMaker (Schwartz et al. 2000) are alignment and visualization tools that have been developed to compare sequence alignments across larger regions. VISTA (visualization tool for alignment) and LAGAN combine a global alignment program with a graphical tool for analysing alignments that allows the identification of conserved coding and non-coding sequences between species. PIPMaker, on the other hand, generates a percentage identity plot (PIP) after local sequence alignments indicate regions of similarity based on the percentage identity of each gap-free segment of the alignment (the number of matches in the region divided by the length of the region). For multiple sequence alignments, VISTA and Multi-PIPMaker (Schwartz et al. 2003), generate pairwise plots comparing various species in an alignment with a reference sequence. However, another web-based tool, SynPlot, allows all the information from a multiple sequence alignment to be considered simultaneously, rather than requiring a reference sequence (Chapman et al. 2004). Whilst VISTA and PIPMaker tools are useful for comparing and visualizing very large-scale alignments, SynPlot is more useful to examine multiple alignments of single gene loci (Chapman et al. 2004).

1.6.2.2 Identification of TF binding motifs

To fully understand the regulation of a gene, it is important to know the regulatory elements associated with the gene as well as all of the transcription factors (TFs) that bind to the elements. Most binding sites are very short sequences of 6-12 base pairs and are degenerate in nature; usually only 4-6 bases within each binding site are fully conserved and these sites are often present in clusters (Maniatis et al. 1987). The availability of consensus binding sites for many of the known TFs has been used to

construct databases that can be searched to identify potential TF binding sites in a DNA sequence. TRANSFAC (Wingender et al. 2000), TRRD (Transcription Regulatory Region database) and COMPEL (a database of composite regulatory elements) (Heinemeyer et al. 1998) are three online transcription factor databases which provide a catalogue of experimentally determined transcription factors and their binding sites. These databases can be searched to identify putative transcription factor binding sites in a given DNA sequence.

In databases, transcription factor binding sites are usually classified in one or more of the following ways:

i) by using a single unambiguous sequence to categorize a specific binding site (for example, TATAA)

ii) by incorporating ambiguous positions in the consensus binding site (for example, TARAA, where R=A or G)

iii) by using position-weighted matrices to assign a score for each base at each position in the transcription factor binding site (Pennacchio and Rubin 2001).

Due to the short length and degenerate nature of transcription factor binding sites, the output from the transcription factor database searches yields a large number of false-positive predictions. It is possible to reduce these false-positive predictions by detecting clustered or composite binding sites (Pennacchio and Rubin 2001). An equally important problem is the large number of binding sites that could be missed in such searches (false-negatives) as the list of TFs and their binding sites in these databases is not exhaustive. In such instances, comparative sequence analysis could be useful by identifying the presence of conserved binding sites that might not have been predicted using sequence from a single species. It is believed that conserved bindings sites identified in cross-species analysis are far more likely to be real than those found only in single species. The term ‘phylogenetic footprint’ has been used to refer to these short orthologous sequences that are conserved over 6 bp or more (Tagle et al. 1988; Pennacchio and Rubin 2001). ConSite (http://phylofoot.org/) is an interactive, web-based computational platform which allows users to do their own phylogenetic footprinting (Lenhard et al. 2003).

1.6.3. Microarray-based methods

It is obvious that with the sequencing of the human genome and that of other organisms, there is an ever-increasing need to be able to survey complete or near-complete genomes, their transcriptomes and proteomes as a whole. DNA microarrays

has emerged as a popular and important technology which allows the profiling of global patterns involved in gene expression and regulation.

Typically, a microarray is a collection of a large set of DNA sequences in a genome that can be generated by PCR or by using oligonucleotides. Genomic microarrays comprise large insert genomic clones (BACs, PACs and cosmids), sequence-defined PCR products, or short oligonucleotides tiled across a genomic region (Fiegler et al. 2003; Albertson and Pinkel 2003; Mantripragada et al. 2004; Lucito et el. 2000). Expression microarrays contain cDNA clone inserts or oligonucleotides representing the genes (Duggan et al. 1999; Lipshutz et al. 1999). The fabrication of the arrays can be achieved using a robotic device to mechanically spot the PCR products, oligos or cloned DNA fragments on glass slides (Schena et al. 1995). In these cases, the glass slides are coated with one of a variety of reactive molecular groups (i.e., poly-L-lysine, epoxy, amino-reactive silane etc.) in order to bind the DNA to the solid support. High-density oligonucleotide arrays can also be synthesized directly on the surface of the arrays by using photolithography, an ink-jet device or programmable optical mirrors (Lipshutz et al. 1999; Hughes et al. 2001; Singh-Gasson et al. 1999).

For spotted arrays, the target material (RNA or genomic DNA or ChIP material, discussed later in this chapter) is labelled with fluorescent dyes (usually Cy3 and Cy5) and a competitive hybridisation is performed on the array. The DNA sequences included on the arrays may contain repeat sequences, which are suppressed by using C_ot1 DNA. For some array platforms, the arrays are hybridised using only a single target sample (e.g. Affymetrix arrays) and the data is generated only in a single channel. During the hybridisation process, the DNA sequences of the immobilized probes bind to their complementary sequences in the labelled target. The fluorescent signal for each array element is quantitated in two channels (or single channel, see above) for the hybridised samples and analysis programs, which are appropriate to the tested samples, are used to analyse the datasets.

1.6.4. Applications of DNA microarrays

1.6.4.1 Expression studies

Expression profiling was one of the first applications of DNA microarrays and the initial studies using microarray-based method for monitoring gene expression was reported in *Arabidopsis thaliana* (Schena et al. 1995). However, since then, the analysis of global expression profiling has extended to a wide variety of genomes and experimental systems including the study of complex biological pathways such as metabolism (DeRisi et al. 1997), early development (White et al. 1999), elucidation of gene function

(Holstege et al. 1998, Hughes et al. 2000), the study of disease (DeRisi et al. 1996, Ramaswamy et al. 2001, Ship et al. 2002), and drug target validation (Marton et al. 1998) among many others.

Expression microarrays allow one to view the expression profiles of normal and the perturbed systems, and genes with similar expression patterns can be clustered together using computational methods (Page 1996). It is believed that co-expression of genes reflects commonality of regulatory activities by transcription factors which bind to conserved sites on the regulatory elements of their target genes. Altering the activity of a transcription factor, or in other words, when a transcription activator or repressor is inappropriately expressed, results in major changes in the expression patterns of its downstream target genes. Accordingly, genes that are co-regulated will respond similarly due to their shared regulatory motifs that require the binding of specific transcription factor (Hughes et al. 2000). Thus, microarray analysis and subsequent sequence analysis of the subset of genes that are co-expressed aids in identifying the binding motifs for various transcription factors (Chu et al. 1998). One limitation of this strategy is that the changes which are observed in the gene expression patterns could in fact be due to secondary effects. Or in other words, some of the genes which exhibit co-expression could be the indirect targets of the transcription factor being studied.

1.6.4.2 Array Comparative Genomic Hybridisation (array CGH)

The development of array-based comparative genomic hybridisation (array CGH) allows screening of the human genome for genomic copy number changes. Array CGH offers the sensitivity and dynamic range to quantitatively measure single copy number losses or gains of genomic DNA sequences associated with various diseases. Initial studies used genomic array platforms containing large genomic insert clones as array elements, which allowed higher resolution detection of copy number changes as compared to CGH on metaphase chromosomes (Albertson and Pinkel 2003; Fiegler et al. 2003). Since then, other array platforms containing sequence-defined PCR products, oligonucleotides, and even cDNA clones have been used for investigation of genomic copy number changes (Mantripragada et al. 2003; Lucito et al. 2000; Pollack et al. 1999). Until recently, the array CGH resolution was limited to ~40-50 kb (Albertson and Pinkel 2003, Mantripragada et al 2004). However, it has now been shown that it is possible to accurately measure copy number changes at a much higher resolution i.e., at the level of individual exons in the human genome (Dhami et al. 2005).

1.6.4.3 Chromatin immunoprecipitation and microarrays ("ChIP-chip")

A recent application of microarray technology has been to study chromatin structure and function. It is obvious that chromatin organization, its modifications and the interactions of DNA-protein complexes play important roles in all of the processes that are fundamental to the biology of any organism. DNA microarrays combined with methods such as chromatin immunoprecipitation (ChIP) have been used to investigate the *in vivo* interactions of the genomic DNA with transcription factors or other regulatory protein complexes associated with chromatin structure. The use of chromatin immunoprecipitation (ChIP) with arrays has been termed "ChIP-chip" or "ChIP-on-chip" (see also chapter 4, section 4.1).

Chromatin immunoprecipitation is one of the most powerful methods by which the occupancy of a binding site of a given transcription factor *in vivo* can be probed by cross-linking the DNA-protein interactions in the native chromatin environment. The pioneering studies using this method were published in the 1980s and, since then, the method has been used for various organisms, ranging from yeast to mammalian cells, with surprisingly little variation in the methodologies employed (Solomon and Varshavsky 1985). The ChIP procedure itself is relatively simple (see Figure 1.3). The cells (or tissues) grown under the desired experimental conditions are fixed with formaldehyde to cross-link the DNA-protein complexes *in vivo*. The cells and nuclei are lysed to extract the cross-linked chromatin. This chromatin is sonicated to generate sheared fragments of approximately 300 bp to 1 kb in size. Following sonication, the DNA fragments that are cross-linked to the protein of interest are immunoprecipitated with an antibody specific to the protein of interest. The cross-links of the immunoprecipitated fragments are reversed and the DNA is purified. The ChIP sample can then be further analysed on microarrays to identify the genomic sequences that are enriched in the ChIP sample: the ChIP sample and a reference sample (usually the total genomic sheared DNA, see Figure 1.3) are labelled with fluorescent dyes (Cy5 and Cy3) and hybridised on to the microarray. Most of the studies using ChIP-chip either use ChIP DNA from multiple experiments to perform a single hybridisation or amplify the ChIP DNA material prior to labeling (see section 4.1, chapter 4).

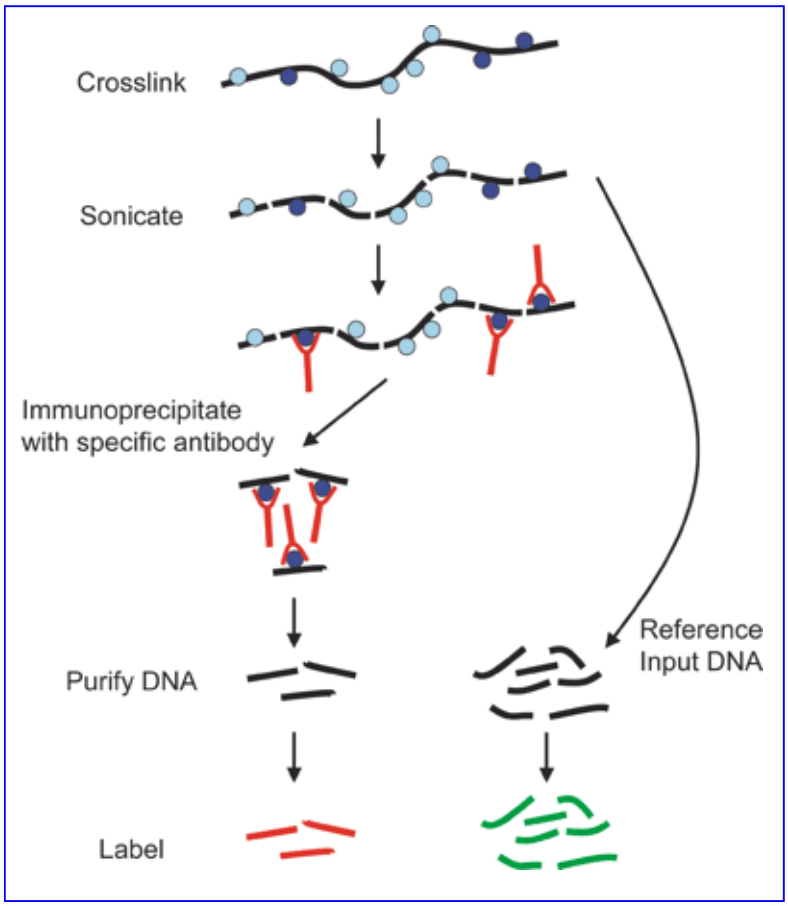

Figure 1.3: Schematic diagram of ChIP assay. The description of the method is provided in the text. The basic steps involved in the method are cross-linking the DNA-protein complexes, sonication to generate sheared fragments, immunoprecipitation with a specific antibody, and DNA purification. For ChIP-chip, the reference sample, obtained by extracting DNA from total sheared genomic DNA (called input DNA and this step is shown by the long black arrow on the right), and the ChIP sample are then differentially labelled and hybridised onto DNA microarrays. Figure taken from Carter and Vetrie (2004) with permission.

The ChIP-chip technique was first applied successfully to identify binding sites for individual transcription factors in *Saccharomyces cerevisiae* (Iyer et al. 2001; Ren et al. 2000). In yeast, the ChIP-chip method has also been used in other applications including study of DNA replication (Wyrick et al. 2001), recombination (Gerton et al. 2000), and chromatin structure (Robyr et al. 2002; Nagy et al. 2003; Bernstein et al. 2002). The microarrays used in these studies contained intergenic regions, open reading frames (ORFs) or both. Since then, numerous studies using ChIP-chip method to elucidate DNA-protein interactions, and genomic regulatory sequences in the genomes of other organisms, including mammals, have been published. In the analysis of the human genome (and others) using ChIP-chip, two main approaches have been used:

(i) Biased approach: The arrays contain only sub-sets of regulatory sequences from across the genome such as promoter regions or CpG islands.

A number of studies have been published using ChIP coupled with CpG-island microarrays to identify novel *in vivo* targets of E2F1, E2F4, E2F6, pRb and c-myc (Weinmann et al. 2002; Wells et al. 2003; Oberley et al. 2003; Mao et al. 2003). Similarly, other studies have been performed using microarrays which contained PCR products representing only the promoter regions of human genes. These arrays were used to map binding interactions of c-myc, E2F and HNF (HNF1α, HNF4α, and HNF6) transcription factors (Ren et al. 2002; Odom et al. 2004; Li et al. 2003).

The biased approach is immensely helpful in building a framework of interactions between transcription factors, DNA sequences and the genes they control. A major disadvantage, however, is that binding sites for TFs and other proteins which regulate transcription by binding to enhancers and other elements, will be missed - since they invariably would not be included on promoter or CpG island arrays.

(ii) Unbiased approach: These arrays represent entire genomic regions (often as the non-repetitive DNA portion) in the form of tiling paths of DNA sequences.

Tiling arrays were first used in mammalian systems to map the binding distribution of GATA-1 across the human β-globin locus (Horak et al. 2002). Similarly, binding sites for p65, NfkB, CREB (cyclic AMP-responsive element-binding protein), Sp1, c-myc, and p53 were mapped across chromosomes 21 and 22 in an unbiased manner (Rinn et al. 2003; Martone et al. 2003; Cawley et al. 2004; Euskirchen et al. 2004). The results from the above mentioned studies revealed that only a small proportion of TF binding sites were identified near the 5' ends of the genes. A large number of TF binding sites mapped to the 3' ends of the genes, within introns or previously unannotated regions. Based on these results, it is evident that such TF binding sites would be missed by using the biased approach. Thus, tiling microarrays across larger genomic regions are extremely useful for mapping the binding sites of transcription factors in a comprehensive manner.

Similarly, tiling arrays have proved to be extremely useful in profiling specific genomic regions, whole chromosomes or whole genomes for histone modifications (Pokholok et al. 2005; Bernstein et al. 2005; Kurdistani et al. 2004). More recently, a genome-wide map of active promoters in human fibroblast cells has been constructed using tiling arrays representing the non-repetitive genomic sequence of the entire human genome (Kim et al. 2005). The use of genomic tiling arrays has also been used to profile human and mouse transcriptomes. This analysis has shown that a huge number of transcripts – about half of the total – do not encode proteins and may represent functional non-coding RNAs (Rinn et al. 2003; Kapranov et al. 2002; Cawley et al. 2004). The results from these studies suggest that tiling array across any genomic region would be immensely

helpful in a comprehensive elucidation of RNA- and DNA-protein interactions across that region.

1.6.5. Other ChIP-related methods

Two new methods to determine transcription factor binding sites in an unbiased manner have been reported recently. The first method, called STAGE, is based on high-throughput sequencing of concatemerized tags derived from DNA enriched by ChIP. In this method, the immunoprecipitated chromatin is either directly cloned and then sequenced or turned into small tags, concatemerized, cloned and sequenced (Kim et al. 2005). The second method, called DamID, is based on creating a fusion protein consisting of *E. coli* DNA-adenine methyltransferase (Dam) and the transcription factor of interest. Dam methylates adenine residues in the sequence GATC. Upon expression of the fusion protein in cultured cells or in an intact organism such as Drosophila, Dam is targeted to the native binding sites of the transcription factor which results in local methylation of the adenine residues. Thus, the sequences near the binding sites of the transcription factor will be marked with a unique methylation tag, which can be tested using a southern blot, PCR and microarray based assays that take advantage of restriction enzymes that are methylation-sensitive (Orian et al. 2003). To date, this method has not been used in mammalian systems.

1.7. Chromatin, epigenetics and transcriptional regulation

A typical eukaryotic cell contains over two meters of DNA which is packaged into a nucleus of about 5-20 μm in diameter. The packaging of DNA molecules is achieved in a highly ordered process. The first level of compaction is achieved by the winding of double-helical DNA around the histone protein-core to produce a “bead-like” structure called a nucleosome, which forms the basic unit of the chromatin fibre. The chromatin fibre (11 nm in diameter) is further folded together to form a compact fibre of about 30 nm in diameter. The final level of compaction is achieved when the 30 nm chromatin fibre is folded and organized into a series of looped domains, which subsequently condenses to form chromosomes. The structure and function of chromatin is intimately related to the transcriptional regulation of genes, as discussed below.

1.7.1. Nucleosome: the structural and functional subunit of chromatin

Nucleosomes, which are arranged like “beads-on-a-string” along the length of the DNA, form the basic structural and functional units of chromatin. A typical nucleosome consists of about 200 bp of DNA wrapped around a histone octamer that contains two copies of the four histone proteins H2A, H2B, H3 and H4. However, the nucleosome

core particle consists of about 146 bp of DNA wrapped in 1-3/4 turns around the histone octamer; with the remaining DNA forming a linker (8-114 bp long) to the next core. A fifth type of histone, H1, interacts with the linker DNA outside of the core particle, and is believed to be involved in higher-order folding of the nucleosomes (reviewed in Khorasanizadeh 2004).

The histones are highly conserved, small basic proteins consisting of a globular domain and a more flexible and charged N-terminus (the histone tail). The four core histones, H2A, H2B, H3 and H4 share a common structural motif, termed the histone fold (reviewed in Ramakrishnan 1997) which consists of a long central helix flanked on both sides by a loop and a short helix (helix-strand-helix motif). The histone fold mediates histone-histone and histone-DNA interactions (Luger et al. 1997). In order to form the core histone octamer, each of the core histone forms dimers, (H2A/H2B) and (H3/H4), in which the two monomers are intimately associated in a head to tail manner in a so-called handshake motif (Ramakrishnan 1997). The two (H3/H4) dimers associate to form a tetramer $(H3/H4)_2$ (Eickbush and Moudrianakis 1978) which is followed by the dimers (H2A/H2B) binding on each side of the tetramer to form the histone octamer (Figure 1.4). The binding of H2A/H2B dimers to the opposite sides of the tetramer results in a tripartite structure of the octamer. Linker histone, H1, interacts stably with nucleosomal DNA only after it is wrapped around the core histones.

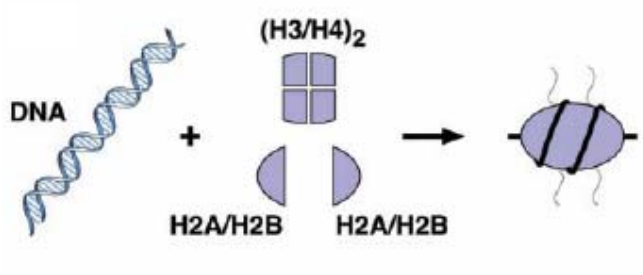

Figure 1.4: Diagrammatic representation of the formation of a nucleosome. Two (H3/H4) dimers associate to form a tetramer $(H3/H4)_2$. The (H2A/H2B) dimers then bind on each side of the tetramer to form the histone octamer. 146 bp of genomic DNA is then wrapped twice around the highly conserved histone octamer to form a nucleosome. The panel to the right of the black arrow shows one nucleosome around which the wrapped DNA is shown by the black lines. The histone N-terminal tails can be see protruding from the nucleosome. This image has been adapted from Grewal and Moazed 2003.

Owing to the super-coiling of DNA around the histone octamer, the nucleosomes can also bring two regulatory elements into close proximity. This role has been observed in the case of heat-shock elements in the hsp26 promoter in *Drosophila melanogaster* (Lu et al. 1995) and in the vitellogenin (Schild et al. 1993) B1 gene. In both cases, a nucleosome positioned between two regulatory elements is important for transcription.

Recently, in yeast (*S. cerevisiae*), nucleosome depletion was observed in the promoters of active genes which contained multiple conserved motifs for specific transcription factors or which recruited Rap1 (Rap1 has previously been described to have a role in opening chromatin and altering nucleosome positioning) (Bernstein et al. 2004). Similarly, other genome-wide studies also reported that nucleosomes were depleted from active regulatory elements throughout the yeast (*S. cerevisiae*) genome *in vivo* (Lee et al. 2004; Pokholok et al. 2005; Yuan et al. 2005). This depletion has also been shown to be a feature of active genes in *D. melanogaster* (Mito et al. 2005) but has not been reported in the mammalian genome. The observations in yeast and *Drosophila* further support the idea that the nucleosomes are either evicted or moved along the DNA strand to expose the underlying chromatin (see section 1.5.3.1).

Until recently, not much attention had been paid to the variant histones which are the separately encoded forms of canonical histones distinguished by amino acid sequence differences (Malik and Henikoff 2003). However, recently published studies have revealed that the variant histones, and the way they are deposited in chromatin, are important for chromatin differentiation and epigenetic maintenance. The bulk of the histones are expressed during S-phase (Osley 1991) and deposited on nucleosomes during DNA replication. The histone variants, however, are expressed at low levels throughout the cell-cycle and their deposition is replication-independent. For instance, mammalian centromere protein A (CENP-A), which is a histone H3 variant, participates in the assembly of centromeric nucleosomes together with H4, H2A, and H2B (Ahmad and Henikoff 2001). Another H3 variant, H3.3, is very similar in sequence to H3 and deposited at transcriptionally active loci without a requirement for DNA replication (Ahmad and Henikoff 2002; Chow et al. 2005). It has now been proposed that replication-independent deposition and inheritance of actively modified H3.3 in regulatory regions maintains transcriptionally active chromatin (Mito at el. 2005). In the case of histone H2A variants – H2AZ and H2AX, are important for the regulation of silencing for a subset of genes (Suto et al. 2000) and DNA repair by non-homologous end joining (Rogaku et al. 1998).

1.7.2. Histone modifications

Several lines of evidence suggest that covalent post-translational modifications of histones play crucial roles in generating a chromatin structure which is conducive for regulation of gene transcription. These modifications could be generated transiently in response to changes in physiological or environmental stimuli or could be more permanent inheritable marks specifying gene expression patterns to the next cellular

generation. Such inheritable changes which do not involve changes in DNA sequence are commonly referred to as 'epigenetic'.

1.7.2.1 Acetylation

The core histones are reversibly acetylated at specific lysine residues located in the N-terminal tails of the histones. The sites of acetylation include at least four conserved sites in histone H4 (K5, K8, K12 and K16), five in histone H3 (K9, K14, K18, K23 and K27), as well as less conserved sites in histones H2A and H2B (see Figure 1.5). It is believed that histone acetylation may alter the folding properties of the chromatin fibre. Addition of an acetyl group to the lysine residue of the histone tails results in neutralizing the positive charge which could lead to disrupting the histone-DNA interactions. This may impede the folding of the N-terminal tails and thus, destabilize higher order chromatin organization (Hansen et al. 1998). The increased acetylation of histones has long been proposed to facilitate transcription (Tse et al. 1998), whereas decreased acetylation correlates with transcriptionally repressed state. Thus, the heterochromatic state is associated with hypoacetylation of histones.

Histone acetylation is thought to affect the interaction of transcription factors or other non-histone proteins in at least two ways. First, histone acetylation facilitates the interaction of transcription factors with the nucleosomal DNA (Workman and Kingston 1998). Secondly, acetylation may modulate the interactions of the proteins which interact with the N-terminal domains. For example, acetylation disrupts interactions between the tail domain and the repressor Tup1 (Edmondson 1996).

Another notably interesting point is that acetylation can activate or repress transcription. For example, the activity of the IFNβ enhanceosome is partly regulated by acetylation. The enhanceosome consists of NF-κB, IRF1, ATF2/c-Jun and HMG-I(Y) and once assembled, the complex recruits CREB binding protein (CBP) which acetylates H3 and H4, ultimately resulting in the transcription of the gene. However, CBP can also acetylate HMG-I(Y) at a DNA binding site which results in the disruption of the enhanceosome and turning off of IFNβ gene expression (Munshi et al. 1998).

In order to gain further insight into the roles played by acetylation, chromatin immunoprecipitation in combination with PCR or microarrays has been used to study acetylation levels across genomic loci. Acetylated histones have been mapped along the chicken β-globin locus and it was observed that acetylated histones were associated with the erythrocyte DNase I-sensitive and transcriptionally active regions (Hebbes et al 1994). Most of studies reported, also found that the promoter region of transcriptionally active genes were associated with highly acetylated H3 and/or H4, while the coding

regions and regions upstream of the promoter were depleted in highly acetylated histones (Roth et al. 2001; Roh et al. 2004; Kurdistani et al. 2004; Liang et al. 2004; Pokholok et al. 2005; Liu et al. 2005; Schübeler et al. 2004).

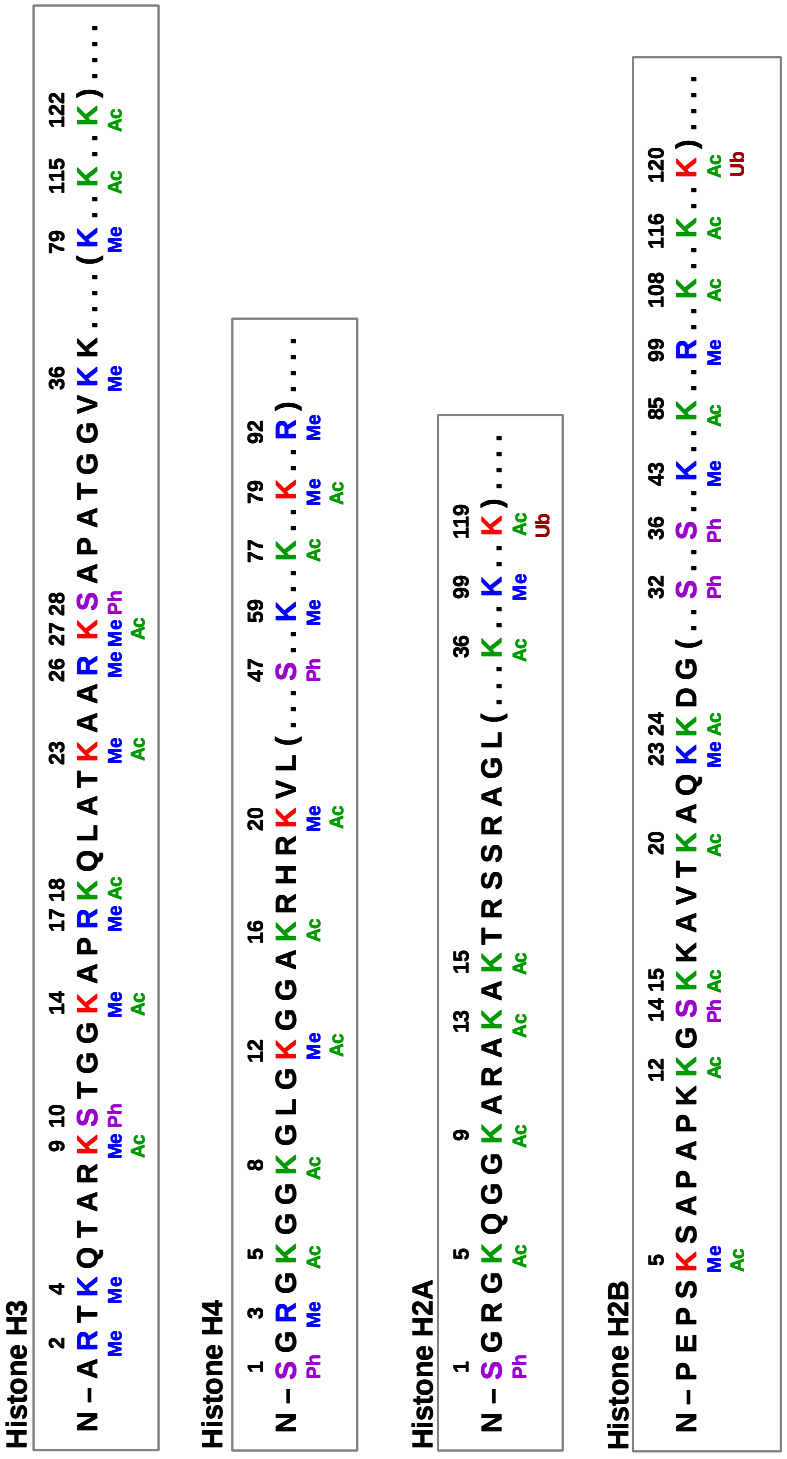

Figure 1.5: Histone modification map of mammalian histone N-terminal tails. The N-terminal tails of histones H3, H4, H2A and H2B contain highly conserved lysines. The residues that can be covalently modified are shown in different colours. R = arginine, K = lysine, S = serine. The residues in blue colour can be methylated (Me), residues in red can be methylated or acetylated (Me/Ac), residues in lilac can be phosphorylated (Ph). The positions of these residues within the N-terminal tails are shown at the top of each residue. The dotted lines represent continuation of each tail. Ub represents ubiquitination of the respective residue in histones H2A and H2B. This image has been adapted from www.upstate.com.

1.7.2.2 Methylation

Histone methylation is a relatively stable modification and the core histones can be modified at lysine residues and at arginine residues (see Figure 1.5). H3 may be methylated at a number of residues including K4, K9, K27, K36 (K-lysine) and R2, R17, R26 (R-arginine), whereas the histone H4 methylation sites are mainly K20 or R3 residues. Additionally, residues in the globular domain of the histone H3 can also be methylated, such as H3 K79 (Feng et al. 2002; Ng et al. 2002; van Leeuwen et al. 2002). Lysine residues can be mono-, di- or tri-methylated *in vivo* (Lachner and Jenuwein 2003) whereas arginine residues can only be mono- or di-methylated (Zhang 2004). Unlike acetylation, methylation does not alter the overall charge on the N-terminal tails of histones. This implies that there must be another mechanism to modify the chromatin structure other than affecting compaction due to charge neutralization. Histone methylation was reported to serve as a mark for chromatin associated proteins such as HP1, which binds to H3 K9 methylated residues to establish heterochromatic regions (Lachner et al. 2001).

Histones methylated at different residues have been associated with different functions. It has been suggested that H3 K4 methylation is linked to active genes, whereas H3 K9 methylation is linked to inactive genes (Lachner and Jenuwein 2002). In order to define the role of H3 K4 methylation in both activation and repression, it has been reported that tri-methylation is specific for the active state of transcription, whereas di-methylation of H3 K4 is found in both active and repressed genes (Santa-Rosa 2002; Ng et al. 2003; Schneider et al. 2004). It has also been reported that monomethylation of H3 K4 is linked with repressed chromatin (van Dijk et al. 2005). These results suggest that the number of methyl groups at a specific residue appear to play an important role in the functional consequences of histone methylation. Bernstein et al demonstrated correlation of di-methyl H3 K4 in coding regions with transcriptional activity and suggested that Set1 facilitated transcription, in part, by protecting active coding regions from deacetylation (Bernstein et al. 2002).

In order to study the role of H3 K4 methylation in human genes, Kim et al investigated the changes in methylation following gene activation and repression of the human prostate specific antigen (PSA) gene (Kim et al. 2003). Upon induced activation of the gene, decreased di- and tri-methylation of H3 K4 were observed at the enhancer and promoter but increased levels were seen in the coding regions. Conversely, di- and tri-methylation of H3 K4 increased at the enhancer and promoter after the gene was repressed.

1.7.2.3 Phosphorylation

The core histones and histone H1 have been shown to undergo phosphorylation on specific serine (see Figure 1.5, for core histones) and threonine residues at their N-terminal tails. Most of the studies published thus far have focussed on the Ser10 residue of histone H3, the phosphorylation of which has a role in a wide range of cellular events. Histone H3 phosphorylation at serine residues was found to be concomitant with the transcriptional activation of immediate early genes such as c-fos and c-jun (Mahadevan et al. 1991). Recent studies have also underlined the essential role of histone phosphorylation during cell-cycle events such as mitosis. In mammals, this modification has been shown to be involved in the initiation of the chromosome condensation process, but not for maintenance of the condensed state of chromatin (Van Hooser et al. 1998). These results suggest that H3 Ser10 plays a dual role, eliciting opposite effects depending on the cellular events - chromatin opening in the case of mitogenically induced gene activation, and chromosome condensation during mitosis.

Phosphorylation of histones H2A, H2B and H4 at specific residues at sites of DNA double strand breaks has been implicated in damage repair in yeast (Cheung et al. 2005; Fernandez-Capetillo et al. 2004). These results suggest a possible link between histone phosphorylation and the damage/repair mechanisms involving specific signalling pathways.

1.7.2.4 Ubiquitination

Histones H2A and H2B have been reported to be reversibly ubiquitinated (see Figure 1.5). The carboxyl end of ubiquitin, a highly conserved 76 amino-acid protein, is attached to the lysine residues in the N-terminal tails of histones - K119 in H2A and K120 in H2B in human. Histone H2A was the first histone identified to be ubiquitinated (Goldknopf et al. 1975) and the majority of ubiquitinated H2A (uH2A) is in the monoubiquitinated form. It has been suggested that as the ubiquitin molecule is about half the size of the core histone, incorporation of ubiquitin into the nucleosome would impact on nucleosome structure and hamper chromatin folding, thus affecting transcription (Zhang 2003).

Several lines of evidence have suggested that histone ubiquitination is most likely to regulate gene transcription in a positive and negative fashion, depending on its genomic and gene location (Nickel et al. 1989). Although the role of histone ubiquitination in chromatin structure needs to be further elucidated, recent studies have indicated a

possible functional link with other histone modifications (Ng et al. 2002; Sun and Allis 2002).

1.7.3. The histone code

It is obvious that post-translational modifications of the N-terminal tails of histones play a pivotal role in modifying chromatin structure and organization, thereby regulating access to the underlying DNA. Jenuwein and Allis proposed the existence of a "histone code", which implies the existence of an epigenetic marking system representing a fundamental regulatory mechanism (Jenuwein and Allis 2001) (also discussed in chapter 6, section 6.1). They suggested that histone proteins and their modifications contribute to this regulatory mechanism that influences chromatin structure, thereby leading to inherited differences in transcriptional "on-off" states. According to the histone code hypothesis, combinations of histone modifications induce chromatin reorganization either directly by changing contacts with underlying DNA or through modification induced recruitment of chromatin-associated proteins. Three observations support the existence of a histone code:

i) acetylated histones in the N-terminal tails interact specifically with a distinct protein domain known as bromodomain, which is found in many chromatin associated complexes (Jacobson et al. 2000). For example, all SWI/SNF-like complexes contain at least one subunit bearing a bromodomain (Marmorstein and Berger 2001).

ii) lysine 9 methylation of histone H3 N-terminal tails is a specific recognition site for another protein module called the chromodomain, which is generally required for heterochromatin formation and gene repression (Jenuwein and Allis 2001). Notably, lysine 9 can also be acetylated which is mutually exclusive with methylation of this residue (Richards and Elgin 2002).

iii) modifications on the same or different histone tails may be interdependent and generate various combinations on any one nucleosome. Given the fact that there are a number of sites within the histones that can be modified, (Figure 1.5), the combinatorial possibilities are very large. Additionally, it has also been suggested that different combinations of histones at different residues may act synergistically or antagonistically to affect gene expression.

A number of studies have been published to support the last observation mentioned above. It has been shown that acetylation of H3 K14 by the GCN5 HAT complex is enhanced by prior phosphorylation of Ser10 in the same tail (Berger 2002). Similarly, it has been suggested that lysine 9 methylation inhibits kinase activity at serine 10, and conversely, serine 10 phosphorylation inhibits methylation at lysine 9 (Rea et al. 2000).

One of the first studies to test the existence of the histone code hypothesis was carried out by Agalioti et al on the human IFN-β gene (Agalioti et al. 2002). The gene is switched on by three transcription factors which form an enhanceosome on the nucleosome-free enhancer DNA. The enhanceosome facilitates initiation of transcription by an ordered recruitment of HATs, SWI/SNF, and basal transcription factors. It was reported that, upon activation of the IFN-β gene, only a small subset of lysines in histones H3 and H4 were acetylated *in vivo* by the GCN5 acetyltransferase. Acetylation of H4 K8 was required for the recruitment of the SWI/SNF complex, whereas acetylation of H3 K9 and H3 K14 was critical for the recruitment of TFIID complex. Based on the results it was proposed that, the distinct pattern of histone acetylation generated at the IFN-β promoter for the recruitment of transcriptional complexes constituted a histone code.

In addition to single-gene studies, it is also important to analyse genome-wide patterns of histone modifications in order to explore the inter-relationships of histone modifications with each other or with the underlying genomic sequences. A number of genome-wide studies have been published in *S. cerevisiae*, *D. melanogaster*, and across large genomic regions in human and mouse (Kurdistani et al. 2004; Dion et al. 2005; Schübeler et al. 2004; Pokholok et al. 2005; Liu et al. 2005; Bernstein et al. 2005) profiling a range of histone modifications (using ChIP-chip). Conflicting opinions, with respect to the existence of the histone code, have emerged based on the results from these studies. A few studies, supporting the existence of a code, reported that there was a clear co-ordination between the various histone modifications, namely acetylation of H3 and H4, and gene activity (Pokholok et al. 2005; Bernstein et al. 2005; Schübeler et al. 2004). Similarly, for methylation at lysine 4 of histone H3, striking but consistent differences were observed for their distribution across the transcribed region of a gene (Pokholok et al. 2005). In contrast, other studies reported no general correlations between acetylation levels at different lysines or between acetylation and transcription and, thus, did not agree with the existence of a histone code (Kurdistani et al. 2004; Liu et al. 2005). Thus, the debate about the existence of histone code is still open and requires further study.

1.8. Haematopoiesis

In order to understand regulatory mechanisms as they occur *in vivo,* it is important to relate these features to well-characterized biological processes. Haematopoiesis has long served as a model for studying complex developmental processes in mammalian biology. Haematopoiesis, the process of blood formation, is a hierarchical process by which haematopoietic stem cells (HSC) give rise to multi- and bipotent progenitor cells which, in turn, differentiate into at least nine lineages of mature and functionally distinct

blood cells. One of the central issues in developmental biology is to understand the molecular mechanisms whereby pluripotent stem cells undergo progressive restriction of lineage potential and acquire the characteristics, defining phenotypes and cell-specific gene expression of mature, terminally differentiated cells.

Haematopoiesis in higher vertebrates occurs in distinct phases and anatomic sites during development. The first wave of blood formation is known as primitive, or embryonic, haematopoiesis. Later in gestation, a second wave of blood development gives rise to definitive, or adult, haematopoiesis. Several tissues within the mammalian embryo serve as reservoirs and/or generators of haematopoietic activity : the yolk sac (Moore and Metcalf 1970), para-aortic splanchnopleura (P-Sp) (Godin et al. 1993; Godin et al. 1995; Cumano et al. 1996), aorta-gonad-mesonephros (AGM) region (Medvinsky et al. 1993; Muller et al. 1994; Medvinsky and Dzierzak 1996), liver (Johnson and Moore 1975; Houssaint 1981), spleen and thymus (Moore and Owen 1967).

Blood development begins in the extra-embryonic yolk sac mesoderm around embryonic day 7 (E7) of gestation in mice (day 15 to18 in humans) where the undifferentiated mesodermal cells, committed to a haematopoietic fate, coalesce to form blood islands (reviewed in Dzierzak et al. 1998). Blood islands initially appear as compact, morphologically identical, cell clusters. As development proceeds, peripheral cells acquire the morphology and markers of endothelial cells, while the inner cells become erythrocytes. These erythrocytes then break free, creating a lumen within the blood island, and, with the establishment of the vascular plexus, circulate. Owing to their temporal appearance, morphology and gene expression pattern, the large nucleated erythrocytes are called embryonic or primitive erythrocytes and express embryonic haemoglobin. The yolk sac is mainly erythropoietic and contains some macrophages (reviewed in Bonifer et al. 1998). Since both endothelial and haematopoietic cells develop from the same clusters of mesoderm - the word angioblast (Sabin 1920), later renamed hemangioblast (Murray 1932), was coined to describe a putative bi-potential precursor cell within the blood island.

Following the first wave of haematopoiesis, the major site of blood development shifts to fetal liver which is colonized by stem cells around E9.5 (Johnson and Moore 1975; Houssaint 1981) in mouse (day 35 to 42 in humans) where most definitive haematopoietic cell types found in the adult animals are produced. By E12, enucleate erythrocytes synthesizing adult globins (fetal globins in humans) together with myeloid cells appear in the embryonic circulation. Around the same time, the fetal thymus becomes active as a lymphopoietic site (Manaia et al. 2000). During late embryogenesis, the haematopoietic site subsequently shifts from fetal liver to spleen

and eventually settles in the bone marrow at day 15-16 of gestation (week 11 of human embryonic development). The bone marrow becomes the major adult haematopoietic organ after birth, throughout the adult life and contains committed precursors for all haematopoietic lineages and is predominantly granulopoietic (reviewed in Bonifer et al. 1998).

Adult haematopoiesis arises from a rare population of haematopoietic stem cells (HSCs) which are characterized by their ability of self-renewal and also maintaining the haematopoietic system throughout adult life (Dzierzak et al. 1998). These cells bear the Sca+, CD34+, c-kit+, lin- phenotype and can be functionally defined by their ability to fully reconstitute haematopoiesis in sub-lethally irradiated adult mice. The origin of HSCs is controversial and, until recently, the development of blood in vertebrates was described as a monophyletic process where a unique organ of haematopoietic cell emergence – the yolk sac – colonized the other organs: first the liver, then the thymus, spleen and finally the bone-marrow (Moore and Metcalf 1970). But in recent years, an intra-embryonic site of haematopoiesis, present prior to the establishment of the fetal liver has been identified, initially in avian (Dieterlen-Lievre 1975), and later in amphibian (Turpen et al. 1997), mouse (Godin et al. 1993, Medvinsky et al. 1993) and human embryos (Tavian et al. 1996). Situated in the caudal half of the embryo, it is called the para-aortic splanchnopleura (P-Sp) in E8.5-9.5 embryos and the aorta-gonad-mesonephros (AGM) region in E10.5-11.5 embryos (Medvinsky and Dzierzak 1996; Cumano et al. 1996). The P-Sp/AGM is not a site of haematopoietic cell maturation but only harbours multipotential, definitive haematopoietic progenitor cells from E7. Before the circulation was established, only the P-Sp gave rise to multipotential haematopoietic progenitors *in vitro*. Whereas, after the circulation connected the yolk sac with the embryo, stem cells with T- and B-lymphoid potential as well as true long-term repopulating activity were detected in both the P-Sp and the yolk sac. It has now been suggested that, independently of the yolk sac, another wave of HSCs arises within the splanchnopleural mesoderm of the embryo between the pre-somitic and liver colonization stage (Dzierzak et al. 1998).

As in other developmental processes, haematopoiesis is likely to be controlled by complex transcriptional cascades regulated by key transcription factors. It is important to bear in mind, though, that different regulatory mechanisms are likely to operate in the different phases of haematopoiesis. The developmental requirements of an early embryo are very different from that of an adult individual, and hence yolk sac and adult haematopoiesis are likely to be regulated differently. The molecular machinery driving haematopoiesis and the ways in which the developmental fate of cells derived from

HSCs differentiate them into different types of blood cells is not yet fully understood. The fate of each multipotential cell depends on the precise combination of transcription factors expressed in that cell.

Transcription factors involved in haematopoietic development have been identified from two main sources. First - through the characterization of genes found at translocation breakpoints in patients with leukaemia and, second - through the identification of factors binding to *cis*-regulatory elements in lineage-specific genes. Genes of the first category include the stem cell leukaemia (SCL) gene (see below). Genes of the second category include members of the GATA family (Wall et al. 1988).

1.9. The SCL Gene

SCL was first identified in 1989 when Begley et al (Begley et al. 1989a) cloned and characterized a chromosomal translocation between 1p33 and the T-cell receptor (TCR) δ-chain locus at 14q11 in DU 528 - a multipotential cell line derived from a patient with T-cell acute lymphoblastic leukaemia (T-ALL). This translocation resulted in a fusion transcript between a previously unrecognized gene (which was called SCL for stem cell leukaemia) and part of the TCR δ-chain gene. The same gene was independently reported by other investigators and named Tal1 (Chen et al. 1990) and TCL5 (Finger et al. 1989).

1.9.1. SCL gene structure

The human SCL gene is located on chromosomal band 1p32-33 whereas the murine SCL orthologue has been mapped to the region of chromosome 4 (C7 band) (Begley et al. 1991) known to be syntenic with human chromosome 1p.

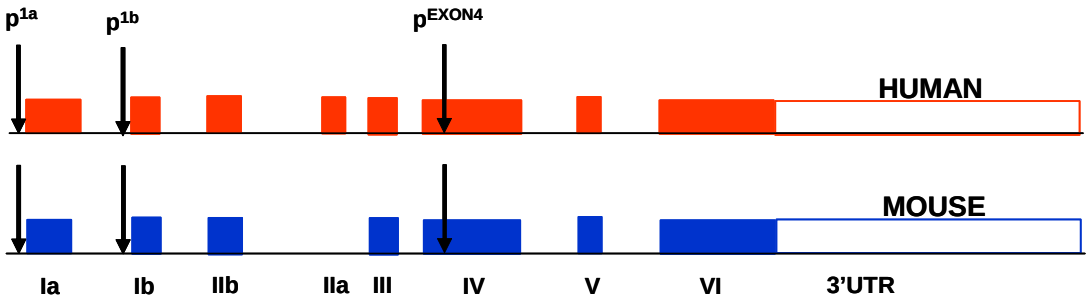

Figure 1.6: Schematic of the human and mouse SCL loci. The human and mouse SCL loci shown in the figure are not made to an exact scale. The human SCL locus is shown at the top (orange) and mouse at the bottom (blue) of the figure. Each box (orange and blue colour) represents an exon. The 3' UTRs are shown by the unshaded boxes. The human SCL gene contains an additional exon as compared to mouse (IIa). Three promoters have been mapped (shown by the black arrows) and are annotated as p^{1a}, p^{1b} and p^{EXON4}.

The human SCL gene is 16 kb in size from the first known transcription start site to the poly(A) site and has eight exons, the first five of which are non-coding (Figure 1.6) (Aplan et al. 1990a). In contrast, the mouse SCL gene consists of seven exons spanning approximately 20 kb of genomic sequence (Begley et al. 1994). There are two alternate promoters – SCL p^{1a} and p^{1b} at the 5' end of the gene. A third promoter, p^{EXON4}, within exon 4 was identified to be active in leukaemic T-cell lines and primary T-ALL samples (Bernard et al. 1992).

The human SCL gene and the murine orthologue share several features:

i) a high degree of sequence homology in the coding regions

ii) highly conserved intronic sequences upstream from exon III

iii) regulatory complexity in both the 5' and 3' UTR (see also section 1.9.2)

iv) a long 3' UTR that is A+T rich

1.9.2. SCL gene expression

Many non-random chromosomal translocations associated with haematological malignancies have led to the identification of novel genes involved in the regulation of haematopoiesis. As SCL was also identified at the breakpoint of a translocation, it was suggested that SCL may play an important role in some aspect of haematopoietic development. However, to fully understand the biological role of a gene, it is vital to know when, where, or to what extent, a gene is expressed. Therefore, a number of studies have been undertaken over the years, to map the expression profile of SCL, within haematopoietic as well as non-haematopoietic tissues and cell-lines, as a first step towards understanding its biological role.

(a) mRNA expression: Characterization of the human SCL gene at the level of genomic DNA and expressed mRNA showed that the 5' end of the SCL gene demonstrates a complex pattern of mRNA splicing producing several different RNA species varying in size from 5 kb, 4.8 kb to 3 kb (Aplan et al. 1990a). At least six alternative splicing events are generated from the 5' UTR, reflecting patterns of alternate 5' exon usage, and all but one RNA species converge on exon III (Aplan et al. 1990a; Bernard et al. 1991). A similar pattern was seen in mouse cells (Begley et al. 1994).

Northern blot analysis in early haematopoietic tissues demonstrated SCL expression in fetal liver and regenerative bone marrow (Begley et al. 1989b; Aplan et al. 1990a; Green et al. 1991b). More refined techniques like *in situ* hybridisation (ISH) and reverse transcription-based polymerase chain reaction (RT-PCR) using purified human haematopoietic cells have shown that expression of SCL in human bone marrow is

further restricted to the erythroid, megakaryocytic, and mast cell lineages (Mouthon et al. 1993). In addition, SCL is expressed in the aorta-associated CD34+ cells - embryonic and fetal haematopoietic progenitors present later in the liver and bone marrow (Labastie et al. 1998). SCL is also expressed by haematopoietic progenitors emerging *in vitro* from ES cells (embryonic stem cells) in the course of their differentiation into embryoid bodies (Elefanty et al. 1997). As expected, both human and mouse leukaemic cell lines, specifically of erythroid, early myeloid and mast lineages (see Figure 1.7), also show robust expression of SCL transcripts and low levels of SCL expression is seen in some pre-B and macrophage lines (Visvader et al. 1991; Green et al. 1992). No appreciable SCL expression has been detected in normal human or mouse thymus (Begley et al. 1989b; Aplan et al. 1990a; Green et al. 1992).

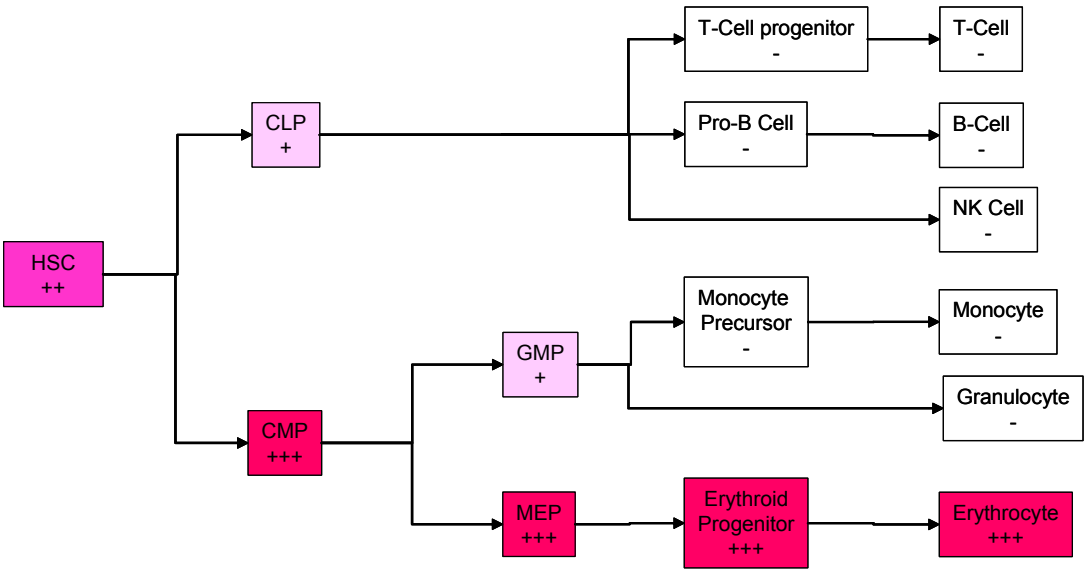

Figure 1.7: SCL mRNA expression during human haematopoietic differentiation. A model of human haematopoietic differentiation is depicted. Increased levels of SCL expression are shown by the colour intensities in the boxes and symbols (light pink, -, +, ++, +++, dark pink). SCL expression levels are highest in pluripotential and early committed erythroid precursors but sequentially extinguish with terminal maturation. In contrast, SCL expression is lower in cells committed to the myeloid lineages and diminishes further with maturation Abbreviations: HSC, haematopoietic stem cell; CLP, common lymphoid progenitor; CMP, common myeloid progenitor; GMP, granulocyte/monocyte progenitor; MEP, megakaryocyte/ erythrocyte progenitor. This model has been adapted from Zhang et al. 2005.

Several investigators have also examined SCL expression in haematopoietic cell lines in response to differentiating stimuli. Initial experiments demonstrated an increase in SCL mRNA as cells were induced to differentiate down the erythroid lineage in response to chemicals or erythropoietin (Green et al. 1992; Cross et al. 1994; Prasad et al. 1995), although further studies of chemically induced erythroid differentiation revealed a late fall in SCL protein levels in the differentiating cells (Murrell et al. 1995). Growth factor induced granulocyte/ monocyte differentiation was accompanied by a marked decrease

in SCL mRNA levels (Green et al. 1993; Cross et al. 1994). These findings concur with the results of experiments to sequentially quantitate SCL gene expression levels as single primitive (CD34+/CD38-) progenitor cells underwent differentiation to become erythrocytes, granulocytes, or monocytes/ macrophages (Cheng et al. 1996; Hoang et al. 1996; Zhang et al. 2005).

(b) Protein expression: While the above mentioned studies provided a strong clue as to which cells within the haematopoietic tissues were likely to contain the SCL protein, immunocytochemical studies confirmed that the pattern of synthesis of the SCL protein corresponds to that of mRNA expression. The protein was detected at both embryonic and extra-embryonic sites, then localized to blood islands of the yolk sac followed by localization to fetal liver and spleen (Elwood et al. 1994; Kallianpur et al. 1994). The SCL protein was detected in erythroblasts in fetal and adult spleen, myeloid cells and megakaryocytes in spleen and bone marrow (Pulford et al. 1995), mast cells in skin, and in rare cells in fetal and adult thymus (Kallianpur et al. 1994; Pulford et al. 1995). Its presence was also detected in both human and mouse erythroid, megakaryocyte, mast cell, early myeloid, some pre-B-cell and T-cell lines (Elwood et al. 1994).

(c) Extra-haematopoietic expression: Outside of the haematopoietic system, SCL mRNA expression was observed in the murine brain in post-mitotic neurons of the metencephalon and the mesencephalic roof; in human medulloblastoma cell-lines (Begley et al. 1989b; Green et al. 1992; Kallianpur et al. 1994). It was also detected in the developing skeletal system, cells of developing cartilage and bone, melanocytes, vascular and visceral smooth muscle cells of the aorta and bladder (Kallianpur et al. 1994) and uterine smooth muscle (Pulford et al. 1995).

Given that blood and endothelial lineages are closely linked, SCL mRNA expression has also been found in endothelial cells, notably in blood vessels in the spleen (Hwang et al. 1993) and also in cultured human endothelial cells. The SCL protein was detected in endothelial progenitors in blood islands, in endothelial cells and angioblasts in a number of organs at times coincident with their vascularization (Kallianpur et al. 1994). The presence of SCL protein in a subpopulation of endothelial cells was also reported in another study (Pulford et al. 1995). Moreover, SCL expression has also been demonstrated within the clusters of cells in the endothelial floor of the ventral aorta (Labastie et al. 1998).

(d) Expression in other species: SCL expression is also highly conserved in lower vertebrates. SCL is expressed in avian endothelial cells and in the mesodermal progenitors of endothelial cells, the angioblasts (Drake et al. 1997). SCL expression was

demonstrated by *in situ* analysis at the mesencephalon/diencephalons boundary in developing chicken embryos (Sinclair et al. 1999) closely reflecting the neural pattern of SCL expression seen in mice at an equivalent stage of development.

(e) Co-ordinate expression of SCL and GATA-1: The spectrum of SCL expression in haematopoietic cells is very similar to that of GATA-1 (a zinc-finger transcription factor) shown to be necessary for haematopoietic ontogeny (Ferreira et al. 2005). Expression of both SCL and GATA-1 shows striking restriction to haematopoietic cells of erythroid, megakaryocytic and mast cell lineages. In immature human haematopoietic cells, SCL and GATA-1 genes are co-expressed in committed progenitors cells ($CD34^{+}/CD38^{2+}$), but are not detectable in the most primitive cells ($CD34^{2+}/CD38^{-}$) (Mouthon et al. 1993). SCL and GATA-1 genes were not only expressed in murine extraembryonic mesoderm, from which blood islands are derived but, in addition, SCL expression was detected early followed by GATA-1 (Silver and Palis 1997). The expression of SCL and GATA-1 was noticed to be upregulated to coincide with progenitor cell development, precisely at the initiation of haematopoiesis (Palis et al. 1999).

In several models systems of haematopoietic differentiation, SCL and GATA-1 undergo biphasic co-modulation of expression with virtually identical temporal profiles (Green et al. 1992). HMBA-induced (hexamethylane bisacetamide) erythroid differentiation of murine erythro-leukaemia (MEL) cells caused a biphasic modulation : an early transient fall of SCL and GATA-1 transcripts, returning to normal, and followed by an increase (Green et al. 1992). Conversely, HMBA-induced myeloid differentiation in the human cell-line K562, was accompanied by an early transient fall of SCL and GATA-1 transcripts which initially recovered but was subsequently followed by a persistent down-regulation (Green et al. 1993).

1.9.3. The SCL protein product

The human SCL gene encodes at least two protein products: a full-length species containing amino-acid residues 1-331 (42-49 kDa) and a truncated species containing residues 176-331 (22-26 kDa) (Cheng et al. 1993; Elwood et al. 1994; Bernard et al. 1991). These protein products correspond to the initiation codons at +1 (first transcription start site) and at +176 (within exon 5). In addition, another truncated protein product of ~39 kDa, corresponding to the initiation codon at +26 (within exon 4) has been detected in T-cells (Bernard et al. 1991; Cheng et al. 1993). The full-length SCL protein is phosphorylated on some serine residues (Goldfarb et al. 1992; Cheng et al. 1993). The mouse SCL gene encodes a protein consisting of 329 amino acids and is 34 kDa in size.

The SCL gene encodes a member of the basic helix-loop-helix (bHLH) family of transcription factors (Begley et al. 1989b). There are two important domains within the protein: the helix-loop-helix (HLH) region and the basic region. The HLH region has the DNA-binding and dimerization motif which is common to a large family of proteins (Murre et al. 1989) found in species ranging from mammals to plants. This domain comprises two amphipathic helices separated by an intervening loop (Murre et al. 1989; Chen et al. 1990). The HLH domain is usually preceded by a highly basic motif, of 10-20 amino acids in length that determines DNA-binding specificity. An interesting feature of HLH proteins is their capacity to form either homodimeric or heterodimeric complexes with other members of this family, thereby modulating their DNA binding activity. Transcription factors belonging to the bHLH family control many developmental processes including myogenesis, neurogenesis, sex-determination and cell-lineage determination. Other examples include MyoD, Lyl-1 and myc family of transcription factors (Murre et al. 1989). In addition to the HLH motif, SCL also contains a proline-rich (12 of 44 amino acids; residues 89 to 132), transactivation domain, near its amino terminus, similar to that seen in the activation domain of the transcription factors NF-1, Jun, AP-2, Oct-2 and SRF (Aplan et al. 1990a).

The HLH motif and the upstream hydrophilic region are entirely conserved in the human and mouse proteins (Begley et al. 1991). To bind to DNA, SCL requires interaction with other transcription factors. SCL has been shown to bind in complexes with Lmo2 and Ldb1 along with GATA-1 and E2A (Wadman et al 1997). In these complexes, the SCL proteins dimerizes with the ubiquitously expressed E12 and E47 proteins (products of E2A gene) to form a complex, that binds with higher affinity to their target genes (Visvader and Begley 1991). SCL/E2A or SCL/E47 dimers bind DNA, recognizing ‘E-box’ motifs (CANNTG) (Hsu 1991) with the preferential usage of the sequence AACAGATGGT (Hsu et al. 1994a). The SCL protein can potentially modulate transcription of its target gene in either a positive or negative fashion (Hsu et al. 1994b) (see also section 1.9.6).

1.9.4. SCL function

1.9.4.1 The role of SCL in cell-proliferation and differentiation

The first clue to the normal function of SCL came from a study in which expression of antisense SCL in the human erythroleukaemia cell line, K562, resulted in a decrease in cell proliferation, cell-cycle progression, and self-renewal potential (Green et al. 1991a), though it enhanced spontaneous erythroid differentiation. Similarly, it was demonstrated that enforced expression of SCL in cell lines MEL and K562 enhanced

erythroid differentiation in the absence of added inducer (Aplan et al. 1992a). Enforced SCL expression in myeloid cells inhibited differentiation, whilst enhancing clonogenicity and conferring a growth advantage (Tanigawa et al. 1993; Tanigawa et al. 1995). To further support these results, it was found that SCL acted as a positive regulator of erythroid differentiation and as a negative regulator of monocytic differentiation (Hoang et al. 1996). This suggested a dual role for SCL in haematopoiesis. In addition, it was also demonstrated that enforced expression of SCL in human CD34+ cells resulted in enhanced erythroid and megakaryocytic differentiation (Elwood et al. 1998). Together, these studies provided evidence that SCL may positively influence proliferation, cell-cycling, and self-renewal whilst having a negative effect on apoptosis.

1.9.4.2 The role of SCL in haematopoiesis

An essential function of SCL for haematopoietic development was first revealed following gene ablation and the generation of $SCL^{-/-}$ mice. This resulted in embryonic lethality at E9.5 due to a complete failure of yolk sac haematopoiesis (Robb et al. 1995; Shivdasani et al. 1995). Similarly, knockdown studies in zebrafish demonstrated that SCL is important for haematopoietic development and angiogenesis, highlighting the evolutionary conservation of SCL function (Dooley et al. 2005). Further studies established that when $SCL^{-/-}$ embryonic stem cells (ES) were injected into blastocysts, the $SCL^{-/-}$ cells were unable to contribute to any haematopoietic lineage in mouse chimeras, demonstrating that SCL was required cell-autonomously for the generation of both primitive and definitive haematopoietic cells (Porcher et al. 1996; Robb et al. 1996). These conclusions were also supported by establishing $SCL^{-/-}$ ES cell lines in which the expression of SCL could be induced by tamoxifen-dependent Cre recombinase-loxP site mediated recombination. It was demonstrated, through *in vitro* differentiation studies, that SCL acted in the mesodermal precursors during a critical development “window of opportunity” to specify both the primitive and definitive haematopoietic compartments (Endoh et al. 2002). $SCL^{-/-}$ ES cells were incapable of *in vitro* differentiation into haematopoietic cells, and failed to express haematopoiesis-specific genes (Elefanty et al. 1997). Together, these findings suggest that SCL is one of the earliest-acting regulators of HSC specification.

1.9.4.3 The role of SCL in blood and endothelial development

Even though many of the studies outlined point to a role for SCL in regulating the self-renewal and differentiating capacity of primitive haematopoietic cells, it was vital to know exactly at what stage SCL becomes active in the primitive haematopoietic compartment. In the yolk sac and AGM region, a close relationship exists between the

formation of haematopoietic and endothelial cells, which initially led to the suggestion that both of these lineages derive from a common precursor called the hemangioblast (Risau and Flamme 1995; Orkin and Zon 2002). The existence of hemangioblast is supported by a number of findings. These include:

i) simultaneous emergence of both haematopoietic and endothelial cell types during the formation of the yolk sac blood islands (Murray 1932)

ii) close association of early intra-embryonic sites of haematopoiesis with early blood vessels (Godin et al. 1993; Medvinsky et al. 1993; Godin et al. 1995; Cumano et al. 1996; Tavian et al. 1996)

iii) the haematopoietic and endothelial cells co-express several genes, including CD34, Flk-1, PECAM-1, Tie-2, c-kit, GATA-2, LMO2, Runx1 and SCL (Kallianpur et al. 1994; Risau and Flamme 1995; Watt et al. 1995; Young et al. 1995; Eichmann et al. 1997; Orkin and Zon 2002)

iv) evidence linking Flk-1 with both haematopoiesis and vasculature (Eichmann et al. 1997; Shalaby et al. 1997)

v) the phenotype of the zebrafish mutant *cloche*, which lacks both haematopoietic cells and endocardium (Stainier et al. 1995).

Although these findings are consistent with the notion of a hemangioblast, direct proof that such a progenitor exists was provided by the identification and isolation of a precursor, the blast colony-forming cell (BL-CFC), in differentiating ES cells that can generate both endothelial and haematopoietic progeny in tissue culture (Kennedy et al. 1997; Choi et al. 1998). These observations support the concept of a close ontological relationship between the haematopoietic and endothelial lineages and suggest that both the cell types originate from a common bi-potential precursor cell called the hemangioblast. Although the characterization of the hemangioblast in its *in vivo* context represents a daunting challenge, several key studies have implicated SCL as a potential regulator of this elusive precursor.

It has been shown that in addition to its crucial function in haematopoietic cells, SCL is essential for yolk sac angiogenesis (Visvader et al. 1998; Elefanty et al. 1999). It is expressed in normal endothelium (Hwang et al. 1993; Drake et al. 1997) and is necessary for vitelline vessel formation in SCL knock-out mice (Robb et al. 1995; Visvader et al. 1998). Ectopic SCL expression in zebrafish embryos can enhance the commitment of mesodermal precursors towards the haematopoietic and endothelial cell fate at the expense of other tissues (Gering et al. 1998). Similarly, enforced SCL expression can rescue the formation of haematopoietic and endothelial cells in the

zebrafish mutant cloche, which exhibits a severe defect in hemangioblast function (Stainier et al. 1995; Liao et al. 1997; Liao et al. 1998).

The generation of SCL knock-out mice results in embryonic lethality at E9.5 (as discussed above). In order to explore the potential involvement of SCL outside of the haematopoietic compartment that might have been missed due to early lethality of the $SCL^{-/-}$ embryos, Visvader et al performed a transgenic rescue of the haematopoietic defects in SCL-deficient embryos (Visvader et al. 1998). While haematopoiesis was restored in these mice, angiogenic remodelling of the yolk sac capillary network into larger vessels was deficient, resulting in embryonic lethality. These results established that SCL functions at the interface between the haematopoietic and endothelial lineages, both of which are thought to arise from the hemangioblast.

The identification of the BL-CFCs (blast colony forming cell) from differentiating ES cells, demonstrates that, *in vitro* at least, a hemangioblast-like cell can be isolated (Choi et al. 1998). Recent studies have demonstrated that (i) SCL and flk-1, the receptor for vascular endothelial growth factor (VEGF), are molecular determinants of BL-CFCs; (ii) $SCL^{-/-}$ ES cells are unable to generate blast colonies, the progeny of BL-CFCs (Faloon et al. 2000; Robertson et al. 2000; Chung et al. 2002).

Similar to $SCL^{-/-}$ mice, the mutation of flk-1 resulted in complete absence of haematopoietic and endothelial cells *in vivo* (Shalaby et al. 1995), although it was possible to obtain some haematopoietic development from flk-$1^{-/-}$ ES cells *in vitro* (Hidaka et al. 1999; Schuh et al. 1999). In order to test whether enhancing SCL expression in flk-$1^{-/-}$ mice would rescue the loss of endothelial and haematopoietic development, ES cells were generated in which the full length SCL cDNA was knocked into the flk-1 locus, thereby placing SCL expression under control of the flk-1 gene regulatory elements (Ema et al. 2003). These cells were then used to generate mouse chimeras through tetraploid aggregation and were also used to study their hemangioblastic potential *in vitro*. It was seen that expressing SCL from the flk-1 locus was insufficient to restore the haematopoietic and endothelial development *in vivo*. However, targeting SCL to flk-1^{+} cells could clearly enhance their potential to form BL-CFCs and haematopoietic colonies *in vitro*, thereby implicating SCL in hemangioblastic specification. Taken together, these studies support the interpretation that SCL is involved in the function and specification of the hemangioblast.

In a direct contradiction to these findings (that suggest an important role for SCL in the specification of hemangioblast development), another study demonstrated that, SCL was not essential for the development of the BL-CFC, the *in vitro* equivalent of the

hemangioblast, but played a pivotal role in its commitment to the haematopoietic and endothelial lineages (D'Souza et al. 2005). An ES cell line was generated in which lacZ cDNA was targeted to the SCL locus. It was shown that: (i) the majority of the BL-CFCs were detected in the SCL/lacZ- fraction indicating that this progenitor did not express SCL but up-regulation of SCL expression was observed within 24 hours of initiation of blast colony development; (ii) $SCL^{-/-}$ cells initiated colony growth but were unable to generate endothelial and haematopoietic progeny, however, it was possible to rescue this blast colony and haematopoietic potential by retroviral transduction of a wild type SCL gene; and finally (iii) $SCL^{-/-}$ flk-1^{+} cells generated colonies that resembled the early stages of the blast colony development.

Another key study which is in agreement with the concept that SCL is not essential for the hemangioblast specification has been carried out in zebrafish where the expression of SCL was knocked down by utilizing site-directed, anti-sense morpholinos to inhibit SCL mRNA (Dooley et al. 2005) . It was seen that whilst angioblasts were specified normally in the absence of SCL (which means that hemangioblast formation was unaffected), nevertheless, later defects in angiogenesis were evident. Additionally, forced expression of exogenous SCL in wild type embryos caused an expansion of both haematopoietic and endothelial tissues, whereas, forced expression in *cloche* and *spadetail* mutants resulted in expansion of haematopoietic tissue but not endothelial tissue. Based on these findings, it was suggested that SCL played distinct roles in haematopoietic and endothelial development, downstream of hemangioblast development.

It is obvious from the above mentioned studies that, SCL plays an important role in the establishment of the haematopoietic and endothelial lineages. However, its role, prior to and after hemangioblast specification, needs to be further clarified and additional studies would be required to address this issue.

1.9.4.4 The role of SCL in lineage-specification

In addition to its crucial role at the earliest stages of haematopoiesis, SCL is also believed to exert important functions in progenitors and in specific lineages of the definitive haematopoietic compartment. SCL follows a differentiation-dependent pattern: it is expressed in HSC and the most primitive progenitors, it remains expressed in cells differentiating down the erythoid, megakaryocytic and mastocytic lineages (Begley et al. 1989b; Green et al. 1991b; Visvader et al. 1991; Green et al. 1992; Mouthon et al. 1993) but becomes down-regulated as differentiation proceeds into most other lineages (Elefanty et al. 1998) (Figure 1.7). Recently, further evidence

has been provided to show that SCL is essential for erythropoiesis and megakaryopoiesis in adult life (Hall et al. 2003; Mikkola et al. 2003; Curtis et al. 2004). These studies involved generation of mice with loxP sites flanking important intronic sequences within the SCL gene (SCL-loxP) and their subsequent deletion using interferon (or PI-PC) inducible Cre-recombinase (Cre). Such conditional knockouts of SCL help to overcome the early embryonic lethality in $SCL^{-/-}$ mice. It was found that deletion of SCL in adult mice perturbed megakaryopoiesis and erythropoiesis with the loss of early progenitor cells in both lineages, while myeloid precursors were not affected (Hall et al. 2003). Interestingly, immature progenitor cells, such as the CFU-S12, with multilineage capacity were still present after SCL inactivation, but these progenitors had lost the capacity to generate erythoid and megakaryocyte cells, and colonies were composed of only myeloid cells. Another study reported that SCL was dispensable for HSC engraftment, self-renewal and differentiation into myeloid and lymphoid lineages, however, the proper differentiation of erythoid and megakaryocytic precursors was dependent on SCL. Their findings also led to the conclusion that once SCL had specified the formation of HSCs, its continued expression was dispensable for stem cell function (Mikkola et al. 2003). Similarly, Curtis et al presented evidence that SCL was not required for self-renewal of HSCs but was required for the normal function of multipotent short-term HSCs. These findings contrast with lineage choice mechanisms, in which the identity of haematopoietic lineages requires continuous transcription factor expression. A simple explanation for these conflicting observations might just be that while SCL may indeed be non-essential for HSC self-renewal, other mechanisms or multiple levels of regulation may exist that could compensate for SCL loss of function within the stem cell compartment.

1.9.4.5 The role of SCL in T-cell leukaemia

It is known that aberrant expression of the SCL gene plays a key role in haematopoietic neoplasia. Since its original discovery, SCL has been found to be rearranged in up to 30% of cases of T-cell acute lymphoblastic leukaemia (T-ALL) (Brown et al. 1990; Bernard et al. 1991; Aplan et al. 1992b). Most of the reported translocations involve the T-cell receptor delta chain (TCR-δ) locus on chromosome 14 (Begley et al. 1989a; Bernard et al. 1990; Bernard et al. 1991). The translocations mainly disrupt the 5' regulatory elements of the SCL gene, but the coding sequence is unaffected and full length SCL protein can be detected in T-cell blasts. Another translocation breakpoint has also been found downstream of the SCL coding region (Begley et al. 1989a; Finger et al. 1989). In such cases, the transcription of the SCL

gene initiates from a cryptic promoter within exon 4 (Bernard et al. 1992) and a truncated protein is formed (see section 1.9.3).

Another rearrangement at the SCL gene is caused by an interstitial deletion of about 90 kb between the SIL (SCL interrupting locus, located upstream of SCL) gene and the 5' untranslated region of SCL. This disrupts the SCL 5' regulatory region (Aplan et al. 1990b), and as a result, the SCL coding sequences come under the regulation of the SIL promoter which drives the expression of a SIL/SCL fusion transcript. Since the SIL promoter is active in T-cells, it leads to an aberrant expression of the SCL protein in T-cells, resulting in T-cell leukaemia.

One of the interesting features of these translocations and deletions at the SCL gene is that these rearrangements appear to be mediated by the V(D)J recombinase complex, since cryptic heptamer recognition sequences, as well as non-templated N-region nucleotide addition, are present at the breakpoints. These kinds of recombinase mediated gene rearrangements are the hallmarks of the immunoglobulin (Ig) or TCR recombinase system (Aplan et al. 1990b).

The first direct evidence demonstrating that SCL can behave as an oncogene came from a study in which SCL enhanced the tumourigenicity of a v-abl transformed T-cell line (Elwood et al. 1993). It is likely that SCL contributes to leukaemogenesis by multiple mechanisms that include an increase in proliferation, cell-cycling and self-renewal potential with a decrease in cell death (Begley and Green 1999).

It has been suggested that SCL interacts with the LIM domain proteins LMO1 and LMO2 to generate tumours (Larson et al. 1996; Aplan et al. 1997). It is known that SCL forms heterodimers with the products of the E2A gene and related proteins (Hsu et al. 1994; O'Neil et al. 2001). These heterodimers can form part of a large protein complex in which LMO2 acts as a molecular bridge between GATA-1 and SCL/E2A heterodimer (Wadman et al. 1997). In mouse SCL tumours and in Jurkat cells (a human leukaemic T-cell line that expresses SCL), stable SCL/E47 and SCL/HEB heterodimers have been detected (Hsu et al. 1994; O'Neil et al. 2001). It has been postulated that SCL induces leukaemia by interfering with E47 and HEB (O'Neil et al. 2004). This was demonstrated by showing that SCL can sequester E2A proteins, thereby inhibiting the transactivation of genes that normally require E2A proteins (reviewed in Begley and Green 1999). It has recently been reported that mice exhibiting SCL expression in an E2A or HEB heterozygous background showed disease acceleration and perturbed thymocyte development due to repression of E47/HEB target genes (O'Neil et al. 2004). It was suggested that SCL mediated gene repression by depleting the E47/HEB heterodimer and by recruiting the

mSin3A/HDAC1 corepressor complex to the target loci. Further understanding of gene regulation by SCL and its partners would be helpful in unravelling the molecular mechanisms underlying normal haematopoietic cell fate determination as well as leukaemogenesis.

1.9.5. Regulation of SCL

The lineage-specific expression of SCL in the haematopoietic compartment is now very well defined. Over the last few years, some of the regulatory mechanisms responsible for the establishment of the complex pattern of SCL expression have been characterized.

During normal development, SCL can be transcribed from two promoters, p^{1a} and p^{1b}, which are highly conserved from chicken to man (Bockamp et al. 1995). There is also a cryptic third promoter in exon 4, p^{EXON4}, which is only active in T-cells associated with leukaemia (Bernard et al. 1992). The promoters, p^{1a} and p^{1b}, are located in the alternate 5' exons of the SCL gene (Figures 1.6 and 1.8) and exhibit lineage-restricted activity in different haematopoietic cell types. The promoter p^{1a} is active in cells of erythroid and megakaryocytic lineages and is regulated by GATA-1 (Aplan et al. 1990a; Bockamp et al. 1995; Lecointe et al.1994). The promoter p^{1b} was found to be active in primitive myeloid cells and mast cell lines through the action of PU-1, Sp1 and Sp3 transcription factors. The p^{1b} was inactive in committed erythroid cells, functioned in a GATA-independent manner and exhibited low-level activity in leukaemic T-cells (Aplan et al. 1990a; Bernard et al. 1991; Bockamp et al. 1995; Bockamp et al. 1997; Bockamp et al. 1998). However, stable transfections demonstrated that both promoters required additional regulatory elements to exhibit their activity following integration in chromatin (Bockamp et al. 1997).

It has long been established that mapping DNase I hypersensitive (HSs) sites across a genomic region could aid in the identification of putative regulatory elements (see section 1.6.1.1). Therefore, DNase I HSs were mapped across the SCL locus in human and mouse (Gottgens et al. 1997; Fordham et al. 1999; Leroy-Viard et al. 1994) to find putative regulatory elements of SCL. (Note: to aid in the description of these sites, a nomenclature has been adopted that denotes the distance (in kb) of the site upstream (-) or downstream (+) of the SCL p^{1a} of human or mouse). The DNase I HSs identified at the SCL locus in human erythroid/megakaryocytic and leukaemic T-cell lines were -10.3, -7, -3.1, -0.2, +2.8 and +3 (Leroy-Viard 1994). Similarly, distinct combinations of DNase I HSs were identified in different mouse cell lines. Taken together, they mapped to -10, -8.6, -8.1, -4.5, -3, +0.7, +1, +3, +7, +17 and +18. Some of the elements identified by

DNase I HSs were also tested in transient reporter assays to assign functional activity. For instance, the 3' enhancer (containing +17, +18 sites) was found to be active in mast and erythroid cells but inactive in primitive myeloid cells and T-cells (Fordham et al. 1999) and upregulated p^{1a} activity but not p^{1b} (Gottgens et al. 1997).

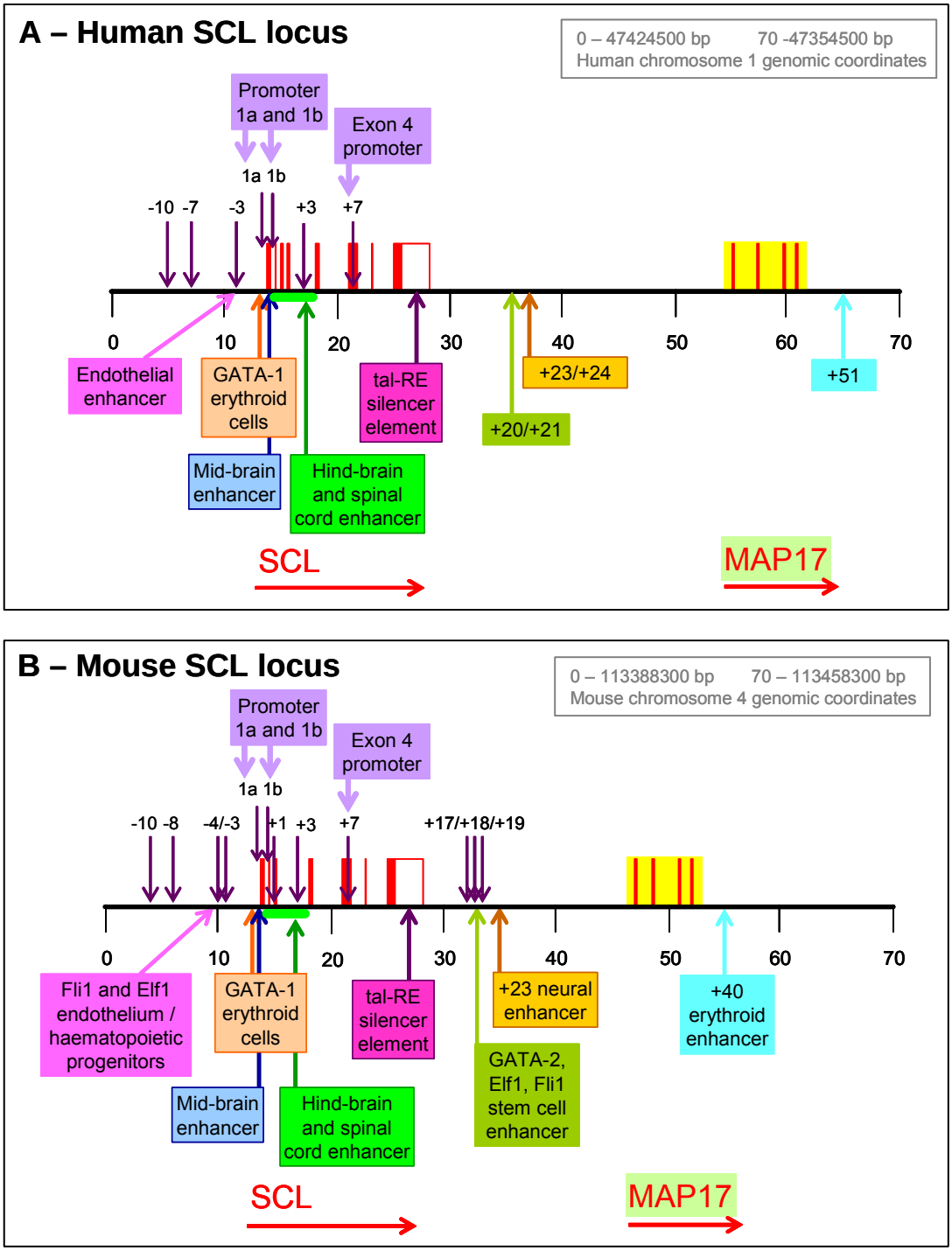

Figure 1.8: A schematic representation of the known regulatory regions at the human and mouse SCL loci. Panel A: human and panel B: mouse. The description of these regulatory elements has been

provided in the text of this chapter and in chapter 5 (section 5.1). The genomic region shown in the figure encompasses the SCL and MAP17 (downstream of SCL) genes in human and mouse, genomic co-ordinates along the respective chromosomes is shown in the grey box in each panel. The red arrows below the gene names show the direction of transcription of these genes. Each of the regions is colour-coded in human and mouse to show their orthologous relationship. The exons are represented by red boxes, lilac arrows pointing downwards represent the mapped DNase I HSs. The TFs which are known to bind some of these regulatory regions have also been shown in the respective boxes. (Image not to scale).

1.9.5.1 The 5' regulatory regions

Three spatially distinct regulatory modules in the 5' region of SCL have been identified, each of which are necessary and sufficient to direct reporter gene expression *in vivo* to three different regions within the normal SCL expression domain (Sinclair et al. 1999). A 940 bp fragment including p^{1a} was both necessary and sufficient to direct *lacZ* expression in developing midbrain and the GATA factor binding sites within this fragment were necessary for neural expression of SCL. Similarly, a 2.9 kb region containing p^{1b}, +1 and +3 HSs directed expression to hindbrain and spinal cord. It was also demonstrated that a 6.1 kb fragment (containing both -4 and -3 HSs) directed expression to endothelial cells and a small minority of haematopoietic cells (Sinclair et al. 1999). Detailed analysis of the 6.1 kb fragment revealed that the -3.8 region was capable of directing expression to haematopoietic progenitors and endothelial cells *in vitro* and *in vivo*, suggesting that this element was bi-functional. This element also contained conserved ETS sites, which were found to be critical for the activity of the enhancer element. Fli-1 and Elf-1 were found to be the transcription factors bound to these sites (Gottgens et al. 2004).

1.9.5.2 The 3' regulatory regions

The genomic region, encompassing the DNase I HSs +17, +18 and +19, directed lacZ expression in transgenic mice to extra-embryonic mesoderm and subsequently, to both endothelial cells and a subset of blood cells at multiple sites of embryonic haematopoiesis (the yolk sac, para-aortic splanchnopleura and AGM region). This 3' enhancer also targeted expression to haematopoietic progenitors in both foetal liver and adult bone marrow (Sanchez et al. 1999) and was termed, the SCL stem cell enhancer. It was further demonstrated that expression of an exogenous SCL driven by this enhancer in $SCL^{-/-}$ embryos rescued the formation of early haematopoietic progenitors and normal yolk sac angiogenesis (Sanchez et al. 2001).

Detailed characterization of the stem cell enhancer, using biochemical and transgenic analyses, defined a 641 bp conserved core element in the +19 region which was responsible for targeting expression to blood progenitors and endothelial cells in transgenic mice. This core enhancer contained three critical transcription factor binding

sites, each of which was essential for enhancer function *in vitro* and *in vivo*. Gel-shift, supershift as well as chromatin immunoprecipitation assays suggested that these sites bound GATA-2, Fli-1 and Elf-1 *in vitro* and *in vivo*. It was postulated that these transcription factors were key components of an enhanceosome responsible for activating SCL transcription and establishing the transcriptional program required for HSC formation (Gottgens et al. 2002). Using a knockout approach, it was discovered however, that despite being sufficient to direct reporter gene expression to hemangioblasts and haematopoietic progenitors, the stem cell enhancer was not necessary within the endogenous SCL locus for blood cell formation or SCL transcription (Gottgens et al. 2004). As a result of this, it was evident that there must be other regulatory elements which are functional in the haematopoietic compartment.

Comparative sequence analysis of the human, mouse and chicken DNA sequences at the SCL loci identified a peak of homology at +23 region, which did not correspond to any of the known enhancers that had been identified previously. A transgenic Xenopus reporter assay demonstrated that the +23 region contained a new neural enhancer directing expression to hindbrain (Gottgens et al. 2000).

In an attempt to identify additional regulatory elements, Delabesse et al systematically mapped acetylated regions (H3 and H4 acetylation) across 90 kb region of the SCL locus in different haematopoietic cell types (acetylated regions indicate regions with putative regulatory activity, see sections 1.7.2.1, 1.7.3). This extended region also included the SIL (SCL interrupting locus) gene upstream and MAP17 (membrane associated protein) gene downstream of the SCL gene. The identified acetylation peaks across the SCL locus corresponded to the known enhancers and an additional peak of acetylation was identified at the +40 region (40 kb downstream of mouse SCL p^{1a}) (see Figure 4.1 in chapter 4, section 4.1). This region was found to be conserved in human-dog-mouse-rat sequence comparisons, functioned as an erythroid-restricted enhancer *in vitro*, and directed expression to primitive but not definitive erythroid cells *in vivo* (Delabesse et al. 2005). The +40 region (homologous to the +50/+51 region in human and has been termed as the erythroid enhancer) contains potential binding sites for multiple transcription factors, such as GATA, ETS, bHLH and homeobox families (Delabesse et al. 2005).

1.9.6. Physiological downstream targets of SCL

While much is known about the regulation of SCL since its discovery 16 years ago, very few genes have been identified which are direct targets of the SCL protein. One of the first SCL targets identified was the c-kit gene (Lecuyer et al. 2002). c-kit encodes

an essential tyrosine kinase receptor that is required for the maintenance of HSC and progenitors. The expression pattern of SCL in primary haematopoietic cells parallels that of c-kit. An initial study linking the two genes demonstrated that antisense and dominant negative SCL constructs reduced levels of c-kit (Krosl et al. 1998). Molecular analyses revealed that SCL assembles a multiprotein complex containing E47, LMO2, Ldb-1, GATA-1/-2, and Sp1 on the c-kit promoter (Lecuyer et al. 2002) and co-expression of all the partners of the complex led to synergistic transactivation.

In contrast to the above finding, it was shown that the retinoblastoma protein (pRb) forms a pentameric complex with LMO2, Ldb-1 and SCL/E12 heterodimer and this complex was found to inhibit c-kit promoter activity (Vitelli et al. 2000) in erythroid cells. In order to explain these discrepant results, it was suggested that depending on the SCL/E2A transcriptional partners, the heterodimer might act positively or negatively even on the same gene at different stages of erythropoiesis.

The erythroid cell-specific glycophorin A gene (GPA) was identified as a target of SCL. The GPA promoter activation was dependent on the assembly of a multifactorial complex containing SCL, E47, Sp1, Ldb-1, LMO2 and GATA-1 (Lahlil et al. 2004). The GPA promoter was activated more strongly by complexes containing GATA-1 compared to those with GATA-2, whereas in c-kit, the complexes containing GATA-2 were more efficient activators (Lecuyer et al. 2002).

Additionally, detailed characterization of the Flk-1 (receptor tyrosine kinase) intron enhancer identified a minimal sequence and two SCL motifs contained in the sequence, which were necessary and sufficient for endothelium-specific reporter gene expression. Analysis of protein-DNA interactions on the enhancer demonstrated a specific binding of SCL and of a GATA factor to the two critical motifs. These results suggested that SCL and GATA factors act upstream of Flk-1 to regulate haematopoietic and vascular development (Kappel et al. 2000).

1.10. Aims of this thesis

To increase our knowledge of the regulation of SCL during haematopoiesis, the aims of this work presented in this thesis were:

1. To construct and validate a high resolution (400-500 bp) genomic tiling path microarray, spanning approximately 250 kb across the human and mouse SCL loci.
2. To map various histone modifications using ChIP-chip in well-characterized SCL expressing and non-expressing cell lines using a large range of antibodies raised against various histone modifications.

3. To elucidate DNA-protein interactions of various transcription factors involved in haematopoiesis or SCL multiprotein complexes and other regulatory proteins, across the SCL locus using the ChIP-chip technique.
4. To further understand the relationships between histone modifications, gene expression and the underlying genomic DNA sequence at the SCL locus.
5. To further characterize novel non-coding regulatory regions relevant to SCL expression which were identified from experiments performed in this study.