The SCL Gene
and
Transcriptional Control of Haematopoiesis

Pawandeep Dhami
Darwin College

Thesis submitted for the degree of
Doctor of Philosophy

University of Cambridge
2005

Disclaimer

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text.

This Dissertation does not exceed the word limit prescribed by the Biology Degree Committee

Pawandeep Dhami

Wednesday, 30 November 2005

This thesis is dedicated to my mum and dad

Abstract

The SCL Gene and Transcriptional Control of Haematopoiesis

Understanding the events which occur as stem cells differentiate into committed cell lineages is a fundamental issue in cell biology. It has been shown that the SCL transcription factor, also known as TAL1, is central to the mechanisms whereby pluripotent stem cells differentiate into haematopoietic stem cells (HSCs) that ultimately give rise to the various blood lineages. While this process is thought to be tightly regulated at the level of gene expression, the exact ways in which SCL helps direct this process is not well understood. To further understand the biology of SCL and the key regulatory interactions it is involved in during blood development, the powerful techniques of genomic microarray resources in combination with chromatin immunoprecipitation (ChIP-chip) were used. High resolution (400-500 bp) genomic tiling path microarrays spanning the human and mouse SCL loci were constructed. ChIP-chip experiments using a large battery of antibodies raised against various histone modifications, transcription factors, and other regulatory proteins were performed in a number of SCL expressing and non-expressing cell lines. The ChIP material used in the array experiments was not amplified prior to hybridization and resulted in ChIP-chip assays which were reproducible, robust and as sensitive as real-time PCR. Based on the ChIP-chip data that was generated, relationships between transcriptional regulatory events and the underlying DNA sequence were studied across the SCL locus. The results described in this thesis will greatly accelerate our understanding of important biological events which are essential for the expression of SCL, as well as provide insights into mechanisms of mammalian gene regulation likely to be widely applicable.

Acknowledgements

I would like to thank my supervisor Dave Vetrie for all the help, guidance, advice and support provided throughout the course and most importantly being my biggest source of inspiration, especially during difficult times.

I would also like to thank our team of collaborators at the Department of Haematology, University of Cambridge – Tony Green and Bertie Gottgens for their help and advice during the course of this study. Thanks to Eric Delabesse who taught me how to do ChIP, and Ian Donaldson's help in performing the sequence analysis is greatly appreciated.

A number of people at The Sanger Institute have made invaluable contributions to the work presented in this thesis. Special thanks to Philippe Couttet for help with the real-time PCR and expression analysis, Jonathan Cooper for doing all the enhancer trap experiments, and Shane Dillon for performing the CTCF ChIP experiment. Rob Andrews provided great help with designing the primers for the SCL tiling array and Oliver Dovey for printing such beautiful arrays. Heike Fiegler and Philippa Carr helped with learning how to perform array hybridisations. Thanks to Rebecca Curley and Susan Gribble for performing the array-CGH experiments, Hazel Arbery and Ruth Bennett for the DNA extractions and Bee Ling Ng for flow-sorting.

Special thanks to all the support teams at The Sanger Institute particularly, the Media preparation team, the glassware team, the Library team, the Sequencing team, and the Systems support team.

Thanks to the three voices of sanity - Sheila, Cords and Susan, for keeping my spirits up over endless cups of tea and gossip.

My friends Anu and Bharti deserve special thanks. Anu for keeping the telephone lines across the Atlantic very busy and listening to my woes at all times. Bharti has been a constant support for the past one year and shared my highs and lows.

A big thanks to my family in India, who despite being so far away, showered me with lots of love, encouragement and prayers for me to finish this successfully. They were keen for me to finish this just so that I can spend more time with them!

I do not have words to thank you, Arvind. I could not have done this without you.

Table of Contents

List of Figures

List of Tables

Glossary of Abbreviations

ATP	adenosine triphosphate
bHLH	basic helix-loop-helix
bp	base pair
CBP	CREB binding protein
CGH	comparative genomic hybridisation
ChIP	chromatin immunoprecipitation
chip	microarray
$C_o t$	concentration x time (DNA reassociation kinetics)
CREB	cyclic AMP-responsive element-binding protein
CTD	C-terminal domain
CV	coefficient of variation
Cy3	cyanine-3
Cy5	cyanine-5
dATP	2' deoxyadenosine 5'-triphosphate
dCTP	2' deoxycytidine 5'-triphosphate
dGTP	2' deoxyguanosine 5'-triphosphate
DNA	deoxyribonucleic acid
dTTP	2' deoxythymidine 5'-triphosphate
ES cells	embryonic stem cells
ESTs	expressed sequence tags
FISH	fluorescence *in situ* hybridization
HAT	histone-acetyltransferase
HDAC	histone deacetylase
HMT	histone methyltransferase
HS	hypersensitive site
HSC	haematopoietic stem cell
IHGSC	International Human Genome Sequence Consortium
kb	kilobase (DNA)
MAP17	membrane associated protein
Mb	megabase (DNA)
mRNA	messenger ribonucleic acid

nc-RNA	non coding RNA
ORF	open reading frame
PCR	polymerase chain reaction
PIC	pre-initiation complex
RNA	ribonucleic acid
RNA Pol II	ribonucleic acid polymerase II
RT-PCR	real-time polymerase chain reaction
SCL	stem cell leukaemia
SD	standard deviation
SIL	SCL interrupting locus
TAF	TFIID associating factor
T-ALL	T-cell acute lymphoblastic leukaemia
TBP	TATA-binding protein
TCR	T-cell receptor
TF	transcription factor
tRNA	transfer RNA
UTR	untranslated region

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.

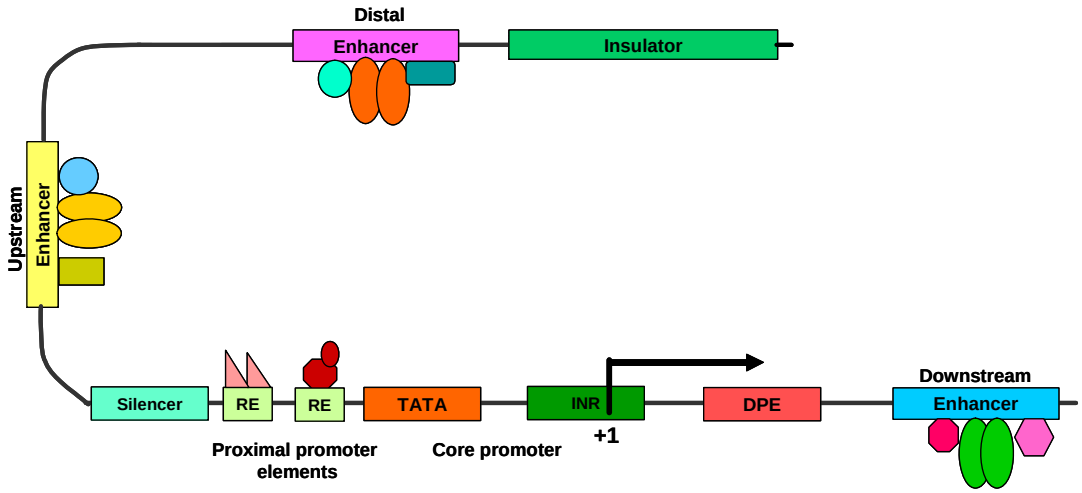


Figure 1.1: A schematic representation of the non-coding regulatory elements involved in the regulation of gene expression. The core promoter of the gene illustrated above contains a TATA box, an initiator element (INR) and a downstream promoter element (DPE), which can recruit the TATA binding protein (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).

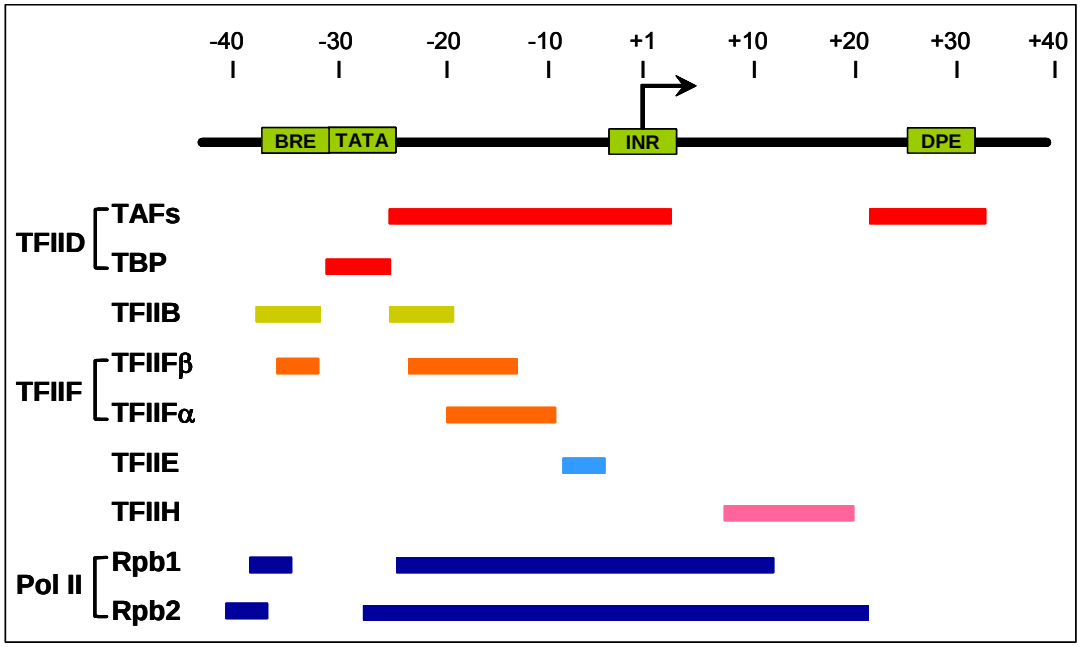


Figure 1.2: Summary of the human general transcription factor interactions at a promoter. The figure shows those proteins which interact with RNA pol II. The black, thick line at the top represents promoter DNA with the position of functional elements (scale at the top) indicated by the green boxes with 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).

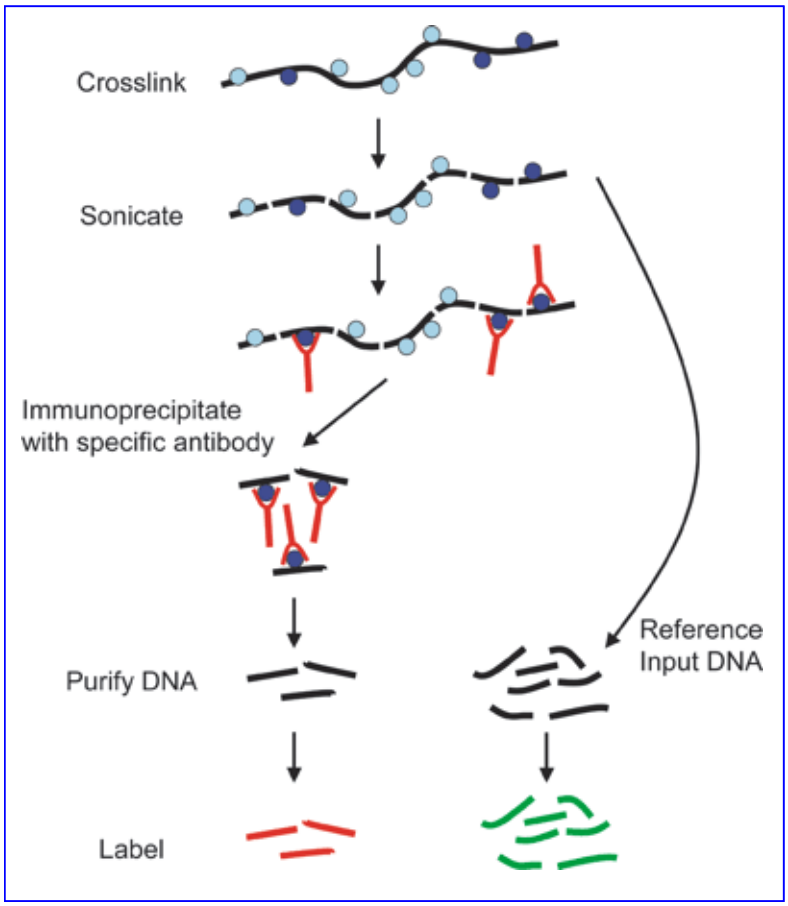


Figure 1.3: Schematic diagram of ChIP assay. The description of the method is provided in the text. The basic steps involved in the method are cross-linking the DNA-protein complexes, sonication to generate sheared fragments, immunoprecipitation with a specific antibody, and DNA purification. For ChIP-chip, 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.

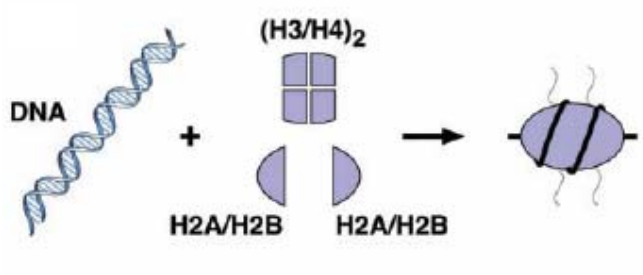


Figure 1.4: Diagrammatic representation of the formation of a nucleosome. Two (H3/H4) dimers associate to form a tetramer $(H3/H4)_2$. The (H2A/H2B) dimers then bind on each side of the tetramer to form the histone octamer. 146 bp of genomic DNA is then wrapped twice around the highly conserved histone 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).

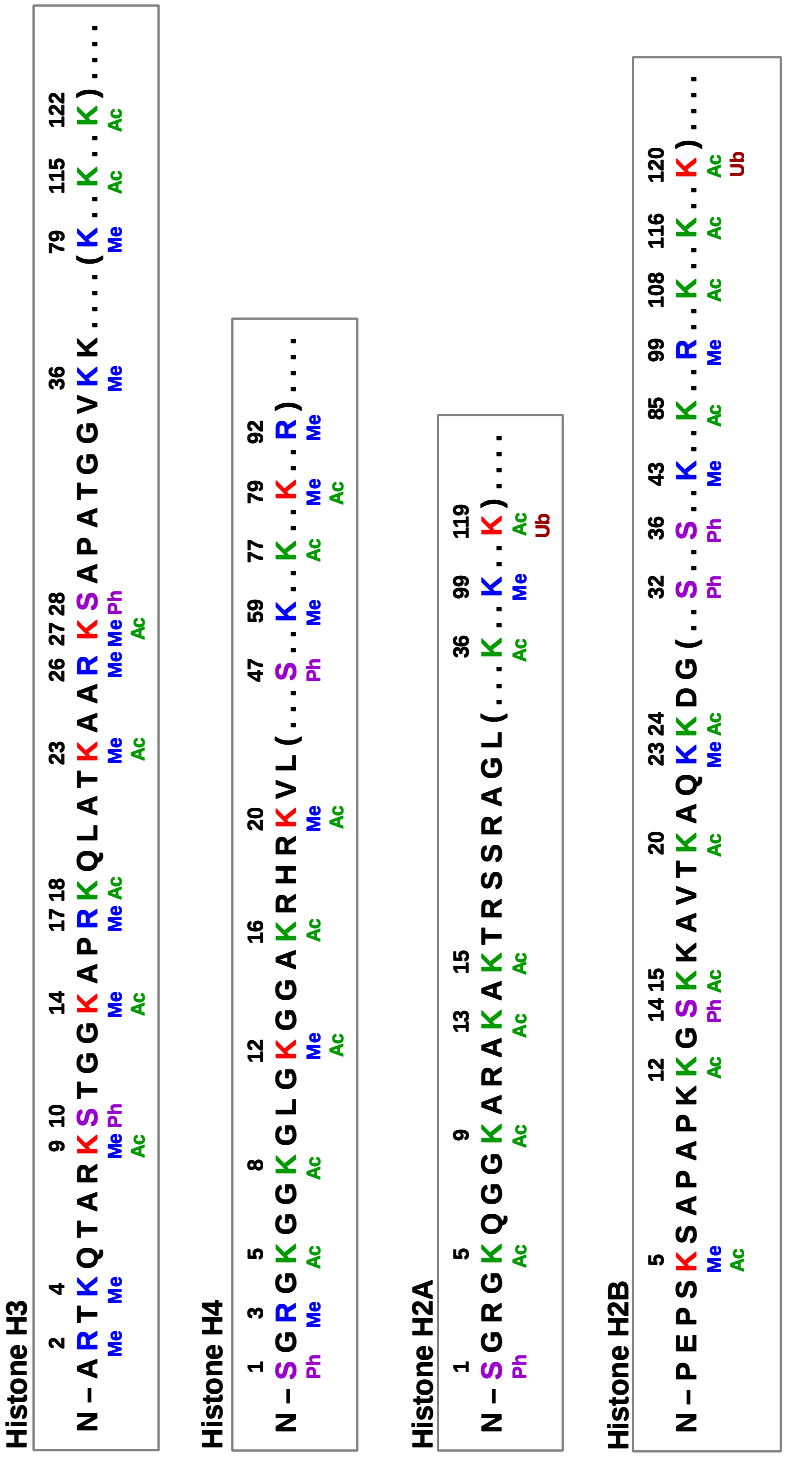


Figure 1.5: Histone modification map of mammalian histone N-terminal tails. The N-terminal tails of histones H3, H4, H2A and H2B contain highly conserved lysines. The residues that can be covalently modified are shown in different colours. R = arginine, K = lysine, S = serine. The residues in blue colour 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.

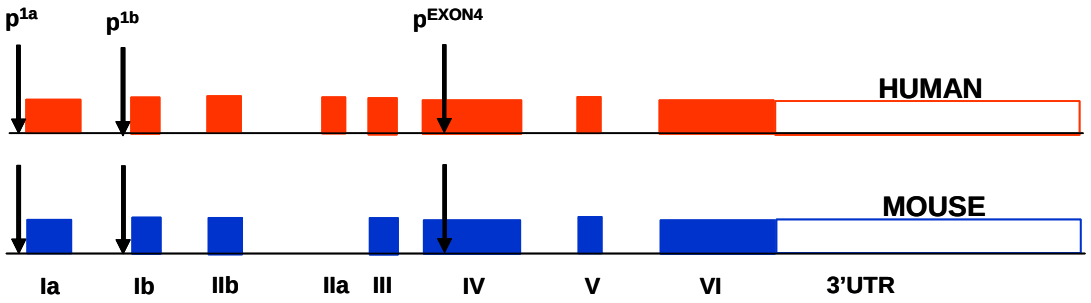


Figure 1.6: Schematic of the human and mouse SCL loci. The human and mouse SCL loci shown in the figure are not made to an exact scale. The human SCL locus is shown at the top (orange) and mouse at the bottom (blue) of the figure. Each box (orange and blue colour) represents an exon. The 3' UTRs are 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).

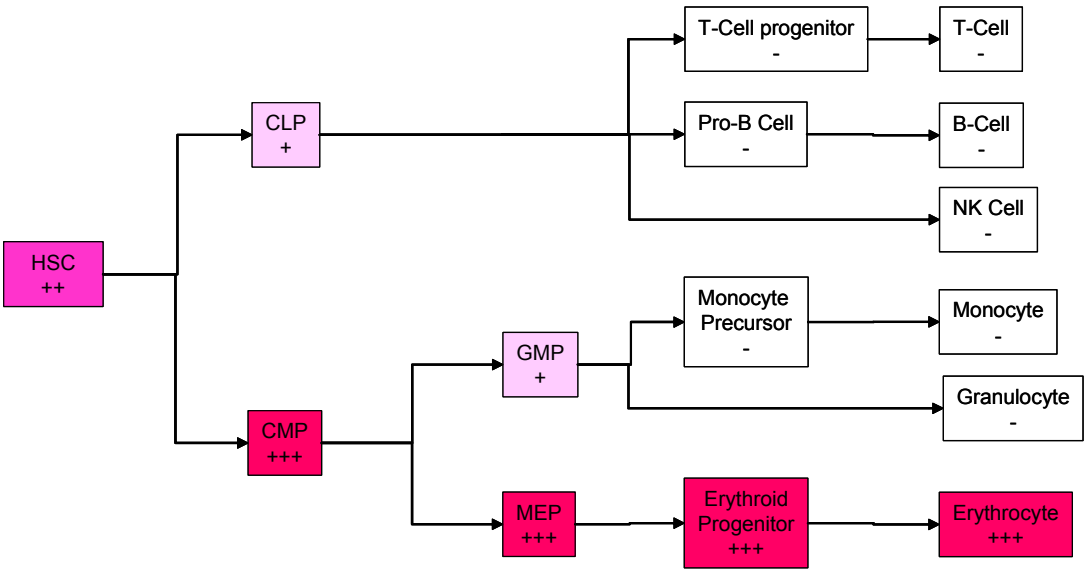


Figure 1.7: SCL mRNA expression during human haematopoietic differentiation. A model of human haematopoietic differentiation is depicted. Increased levels of SCL expression are shown by the colour intensities in the boxes and symbols (light pink, -, +, ++, +++, dark pink). SCL expression levels are highest 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).

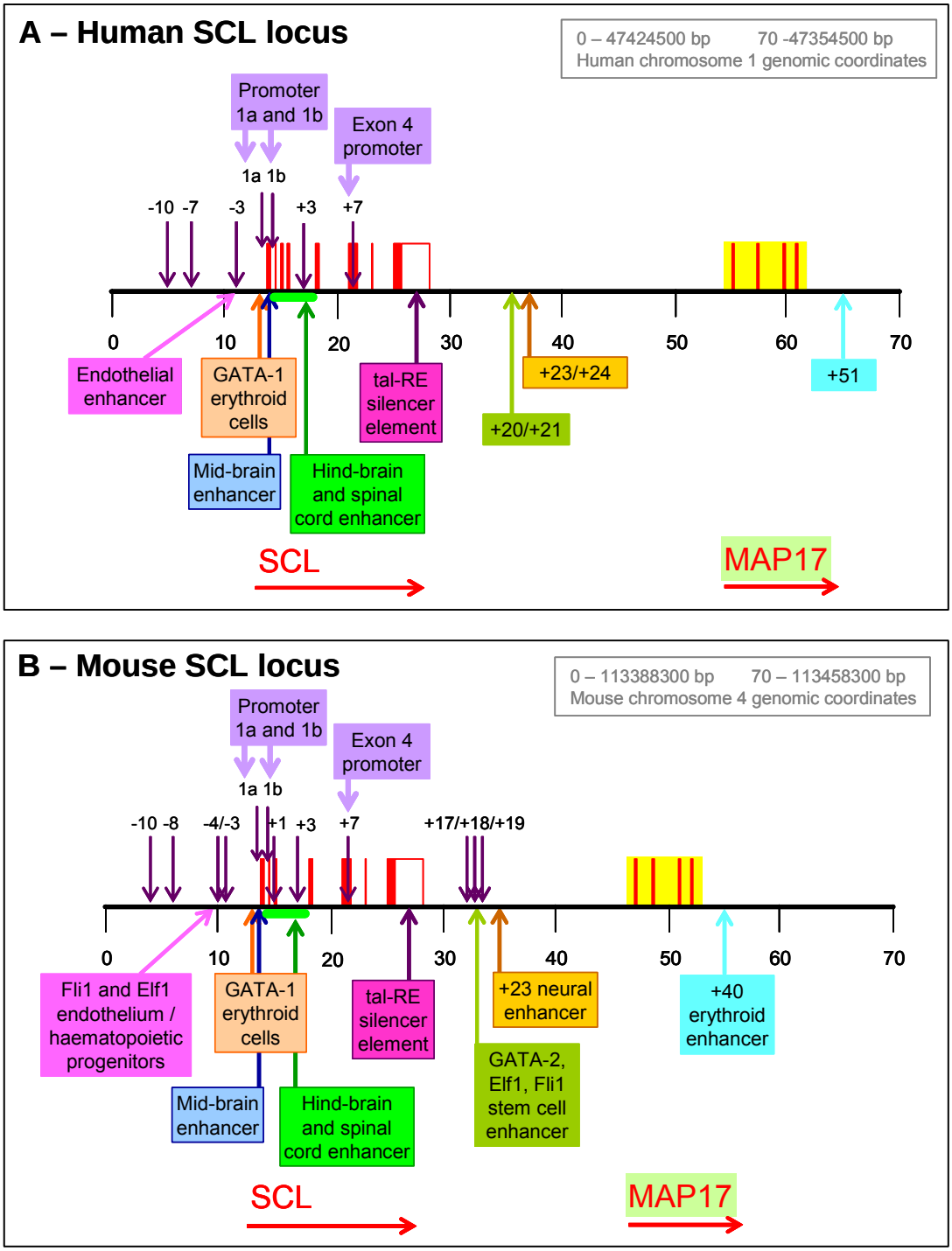


Figure 1.8: A schematic representation of the known regulatory regions at the human and mouse SCL loci. Panel A: human and panel B: mouse. The description of these regulatory elements has been

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.

Chapter 2

Materials and Methods

Materials

2.1. Composition of Solutions

Note: Only sterile filtered HPLC water was used to prepare solutions.

1 X SET

- 10 mM Tris-Cl pH 7.5
- 100 mM NaCl
- 1 mM EDTA

10 X PCR buffer

- 500 mM KCl
- 50 mM Tris pH 8.5
- 25 mM $MgCl_2$

10 mM dNTP mix for PCR

- 10 mM each dNTP (dCTP, dGTP, dATP, dTTP)

4 X Spotting buffer

- 1 M sodium phosphate buffer pH 8.5
- 0.001% sarkosyl

Lysis Buffer for DNA preps

- 50 mM glucose
- 10 mM EDTA
- 25 mM Tris pH 8.0

10 X dNTP mix used in DNA labeling

(a) The following mix was used in the labeling reactions that were used with array hybridisations set up manually.

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

(b) The following mix was used in the labeling reactions that were used with array hybridisations set-up using Tecan

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

Hyb 3 Buffer

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

Tecan-hyb Buffer

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

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

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

- 7.33 g NaCl
- 2.36 g Na_2HPO_4
- 1.52 g $NaH_2PO_4H_2O$
- 500 µl Tween 20

Cell lysis buffer (CLB)

- 10 mM Tris-HCl pH 8.0
- 10 mM NaCl
- 0.2% Igepal
- 10mM Sodium butyrate
- 50 µg/ml PMSF
- 1 µg/ml Leupeptin

Nuclei lysis buffer (NLB)

- 50 mM Tris-HCl pH 8.1
- 10 mM EDTA
- 1% SDS
- 10 mM Sodium butyrate
- 50 µg/ml PMSF
- 1 µg/ml Leupeptin

IP dilution buffer (IPDB)

- 20 mM Tris-HCl pH 8.1
- 150 mM NaCl
- 2 mM EDTA
- 1% Triton X-100
- 0.01% SDS
- 10 mM Sodium butyrate
- 50 µg/ml PMSF
- 1 µg/ml Leupeptin

IP wash buffer 1 (IPWB1)

- 20 mM Tris-HCl pH 8.1
- 50 mM NaCl
- 2 mM EDTA
- 1% Triton X-100
- 0.1% SDS

IP wash buffer 2 (IPWB2)

- 10 mM Tris-HCl pH 8.1
- 250 mM LiCl
- 1 mM EDTA
- 1% Igepal
- 1% Deoxycholic acid

IP elution buffer (IPEB)

- 100 mM $NaHCO_3$

- 1% SDS

TE (pH 8.0)

- 10 mM Tris base (pH 8.0)
- 1 mM EDTA

1 X PBS (for ChIP assays)

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

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

2.2. Reagents

Antibodies

- a complete list of all the antibodies with the company name and catalogue numbers, used for ChIP assays in this study, is included in chapter 5
- Avidin-Texas Red (Molecular Probes)
- Goat anti-mouse FITC (Sigma)
- Mouse anti-digoxigenin FITC (Sigma)

Enzymes

- Proteinase K, ≥ 20 units/mg (GibcoBRL)
- RNase A, ≥ 50 Kunitz units/mg (ICN Biochemicals)
- Gold AmpliTaq, 5 units/µl (Perkin Elmer-Cetus)
- Taq polymerase, 5 units/µl (Perkin Elmer-Cetus)
- Klenow fragment, 40 units/µl (Invitrogen)
- DNase I, 80 Kunitz units/ml (Sigma)
- DNA polymerase I, 10 units/µl (Sigma)

Fluorophores

- Cy3-dUTP (Amersham)
- Cy5-dUTP (Amersham)

- Biotin-16-dUTP (Boehringer)
- Digoxigenin-11-dUTP (Boehringer)

Primer pairs

- Lists and sequences of human and mouse primer pairs used to construct the SCL array or perform real-time PCR are provided in Appendices 1, 2, 3, 4 and 5. The appendix numbers are referenced in the relevant sections of this thesis.

Other reagents

- Human C_ot 1 DNA (Invitrogen)
- Mouse C_ot 1 DNA (Invitrogen)
- Herring sperm DNA (Sigma)
- Trizol (GibcoBRL)

2.3. Cells and Cell lines

Human haematopoietic cell lines i.e. K562 (Lozzio and Lozzio 1977), Jurkat (Schneider et al. 1977), HL-60 (Collins et al. 1977) and HPB-ALL (Morikawa et al. 1978) were gifts from Dr. Tony Green, The Department of Haematology, Cambridge.

Mouse haematopoietic cell line 416B (Dexter et al. 1979) was also a gift from Dr. Tony Green, The Department of Haematology, Cambridge.

Mouse E14 embryonic stem cell line (Hooper et al. 1987) was obtained from Dr. Peri Tate, The Sanger Institute, Cambridge.

The normal human male and normal human female lymphoblastoid cell lines “HRC 575” and “HRC 160” respectively were kindly provided by Dr. Nigel Carter, The Sanger Institute, Cambridge.

2.4. Bacterial clones

The human PAC RP1-18D14 and BAC RP11-332M15 clones used for PCR amplifications for the array construction and FISH analysis were identified using the ENSEMBL Cytoview (www.ensembl.org/Homo_sapiens/cytoview). Similarly, the mouse BAC clones, RP23-242O20, RP23-246H17 and RP23-32K12, used for the array construction were identified using the ENSEMBL Cytoview (www.ensembl.org/Mus_musculus/cytoview). The human PAC clone (RP1-18D14) was a gift from Dr. Tony Green, The Department of Haematology, Cambridge. The other human BAC clone as well the mouse BAC clones were picked individually from the clone

archives established at the Sanger Institute. All the clones were phage tested and subsequently prepped.

Methods

2.5. Tissue Culture

2.5.1. Culturing of all Cell lines

All cell lines were cultured in suspension in 50 ml of media (Sigma) with the appropriate amount of fetal bovine serum (GibcoBRL) and other supplements (Sigma) in 75 cm^2 tissue culture flasks with vented caps (Corning). Table 2.1 provides the information about the media and supplements used for each cell line. Note: Mouse E14 ES cells were cultured by Dr. Peri Tate (The Sanger Institute, Cambridge).

Cell line	Media	Fetal bovine serum	Supplements
K562	DMEM	9% v/v	1% penicillin-streptomycin solution
Jurkat	RPMI 1640	9% v/v	1% penicillin-streptomycin solution, 2 mM L-glutamine
HL60	RPMI 1640	9% v/v	1% penicillin-streptomycin solution, 2 mM L-glutamine
HPB-ALL	RPMI 1640	16% v/v	1% penicillin-streptomycin solution, 2 mM L-glutamine
416B	RPMI 1640	16% v/v	1% penicillin-streptomycin solution, 2 mM L-glutamine
HRC 575	RPMI 1640	16% v/v	1% penicillin-streptomycin solution, 2 mM L-glutamine
HRC 160	RPMI 1640	16% v/v	1% penicillin-streptomycin solution, 2 mM L-glutamine

Table 2.1: List of all the cell lines used for the study presented in this thesis. Note: All cell types were cultured under 5% CO_2 at 37°C.

Once confluent, sub-culturing was carried out as follows:

1. 25 ml of fresh media was added to each flask and any clumps of cells were gently broken up using a syringe.
2. The culture was then distributed between three new 75 cm^2 flasks (Corning) and further 25 ml of fresh media was added to each flask, effecting a 1/3 dilution of the confluent starting culture.
3. The number of cells needed for chromatin immunoprecipitation (ChIP) experiments was quite high. Therefore, culture volumes to obtain the required number of cells per flask were suitably scaled up in 175 cm^2 culture flasks with vented caps (Corning).

2.5.2. Cell cryopreservation

For frozen storage, cells were pelleted at 259 g for 5 to 8 minutes, and resuspended at approximately $1X10^7$ cells/ml in 10% (v/v) DMSO in FBS (GibcoBRL). The resulting cell mixture was transferred into polypropylene cryotubes which were cooled overnight to -70°C. The cryotubes were then transferred to the gas phase of a liquid nitrogen vessel (approximately -180°C) for permanent storage. To reconstitute cultures, cells were thawed rapidly at 37°C, washed once with fresh media and finally resuspended in 10 ml of fresh media.

2.6. DNA Preparation

2.6.1. Preparation of Total Genomic DNA

Total genomic DNA was extracted from all the human cell lines listed in Table 2.1. Fresh cultures were grown in the appropriate media with added supplements (Table 2.1).

1. When confluent, the cells were spun down at 259 g for 5-8 minutes and the pellets were washed with PBS and spun down again.
2. The pellets were resuspended in 4.5 ml 1XSET, 250 µl of 10% SDS and 100 µl of 10 mg/ml Proteinase K (GibcoBRL) and the resulting solutions were incubated overnight at 37ºC.
3. 5 ml of Phenol (saturated with 0.1M Tris, Rathburn Chemicals) was added to each sample and the samples were rolled for 10 minutes at room temperature.
4. Samples were then centrifuged at 1620 g for 5 min at room temperature and the aqueous (top) layer was collected very gently using a P1000 tip (cut off at the tip to make a wider bore) and transferred to a fresh 50 ml Falcon tube. An equal volume of phenol/iso-amyl alcohol/chloroform (25:1:24, pH 6.6-8.0, BDH) was added to the aqueous layer.
5. 5 ml of 1XSET was added to the organic (bottom) layer and rolled for 10-15 minutes at room temperature.
6. The two layers were separated as before by centrifuging at 1620 g for 5 minutes at room temperature. The aqueous layer was again removed very gently using a cut-off P1000 tip and added to the aqueous layer collected in the earlier step containing the phenol/iso-amyl alcohol/chloroform.
7. The aqueous layer/isoamyl-alcohol/chloroform mixture was rolled for 10-15 minutes and then centrifuged at 1620 g for 5 minutes, both at room temperature. As in previous steps, the two layers were separated and the aqueous (top) layer was

added to a fresh 50 ml Falcon tube containing an equal volume of iso-amyl alcohol/chloroform (1:24) mixture. The organic (bottom) layer was discarded.

8. The aqueous layer/isoamyl alcohol/chloroform mixture was rolled for 10-15 minutes and then centrifuged at 1620 g for 5 minutes, both at room temperature. The aqueous layer was removed and placed in a fresh, empty 50 ml Falcon tube.
9. The DNA was precipitated by adding 1/10th the volume, i.e. 1 ml, of 3M NaAc (pH 6.0) and 2.5 X volume, i.e. 25 ml, absolute ethanol to the aqueous layer and incubated at –20ºC overnight.
10. The precipitated DNA was spooled out using sterile glass hooks and then dipped in 70% ethanol to rinse the DNA. The DNA was air dried on the hooks and then very gently dissolved in 400 µl 1XTE in 1.5 ml microfuges. The DNA was kept at 4ºC to dissolve. The concentration of the DNA samples was determined by measuring the optical density at 260 nm.

2.6.2. Preparation of DNA from BAC and PAC clones

DNA from human and mouse BAC, PAC clones was prepared by the following alkaline lysis “mini-prep” method.

1. All the clones (section 2.3) were grown in 10 ml cultures with 2 X TY media at 37ºC overnight supplemented with 25 µl of 10 mg/ml kanamycin (Sigma) for PAC and 10 µl chloramphenicol (Sigma) for the BACs.
2. The cultures were centrifuged the next day at 2080 g for 10 minutes at room temperature. The supernatant was poured off and 200 µl of lysis buffer was added to the pellets to lyse the cells.
3. The solution was transferred to 1.5 ml microfuges and 400 µl of freshly made 0.2M NaOH/1% SDS solution was added to each sample. The samples were incubated on ice for 10 minutes.
4. 300 µl of 3M NaAc (pH 5.2) was added to each sample and incubated on ice for 10-30 minutes.
5. The samples were centrifuged at 18000 g for 5 minutes at room temperature and the supernatant was transferred to fresh 1.5 ml microfuges and the precipitates were discarded. This step was repeated 3-4 times till the supernatant was clear.
6. 600 µl Isopropanol was added to the clear supernatant and incubated at –70ºC for 10 minutes.
7. The tubes were centrifuged at 18000 g for 5 minutes and the pellets were resuspended in 200 µl of 0.3M NaAc, pH 7.0.

8. 200 µl of water saturated phenol/chloroform[a] was added to the samples and vortexed briefly. Samples were centrifuged at 18000 g for 3 minutes. The aqueous (top) layer was collected in fresh 1.5 ml eppendorf tubes. This step was repeated twice.
9. The DNA was precipitated from the aqueous layer by adding 200 µl isopropanol and incubating the solutions at –70ºC for 10 minutes.
10. The solutions were centrifuged at 18000 g for 5 minutes and the pellets obtained were washed with 70% ice-cold ethanol.
11. The pellets were air dried at 37ºC and finally resuspended in 50 µl of T0.1E containing 200 µg/ml RNaseA. After a quick vortex and a quick centrifuge, the tubes were incubated at 55ºC for 15 minutes in a waterbath.
12. 1 µl of each sample was analyzed on a 1% agarose 1 X TBE gel and stained with ethidium bromide for visualization. The DNA samples were stored at –20ºC.

[a]Safety Note: All the steps with phenol/chloroform were performed in a fume cabinet.

2.6.3. Chromatin Immunoprecipitation (ChIP)

K562, Jurkat, HL60, HPB-ALL, 416B and ES cell line E14 were used for chromatin immunoprecipitation. Fresh cultures were grown (Table 2.1) for each cell line and cells were harvested for ChIP. Aliquots of cells used for each ChIP experiment were subjected to flow-sorting (Cytomation MoFlo High Performance Cell Sorter, Dako Cytomation) in parallel for cell-cycle analysis. For this, ~5 ml of cell culture was washed with 10 ml of PBS. The cells were then fixed in 5 ml of 70% ethanol.

Approximately 1 X 10^8 cells were harvested for each ChIP procedure which were then used to set-up ten immunoprecipitation (IP) conditions as described below.

Fixation

1. The cells were collected by centrifuging at 259 g for 8 minutes at room temperature and resuspended in 50 ml of serum free media in a glass flask.
2. DNA-protein and protein-protein interactions were cross-linked by adding formaldehyde (37%, BDH AnalaR). 500 µl, 1010 µl or 1355 µl formaldehyde was added drop-wise to a final concentration of 0.37%, 0.75% for histone modifications and 1% for transcription factors respectively.
3. The cross-linking was carried out at room temperature with constant but gentle stirring for 10 minutes (for histone modifications) or 15 minutes (for transcription factors).
4. 3.15 ml, 3.41 ml or 3.425 ml (for 0.37%, 0.75% or 1% formaldehyde concentration respectively) of ice-cold 2M glycine was added to a final concentration of 0.125M with

constant but gentle stirring for 5 minutes at room temperature to stop the cross-linking reaction.

5. Cells were transferred to 50 ml falcon tubes and kept on ice whenever possible. The cells were pelleted by centrifuging at 259 g for 6-8 minutes at 4°C and washed with 1.5 ml of ice-cold PBS.
6. After washing, the cells were pelleted at 720 g at 4°C for 5 minutes and the supernatant was removed.

Cell and Nuclei Lysis

7. Cells were lysed by adding 1.5 X pellet volumes of ice-cold cell lysis buffer (CLB). The cell pellets were gently resuspended and incubated on ice for 10 minutes.
8. The nuclei were recovered by centrifuging the samples at 1125 g for 5 minutes at 4°C.
9. After carefully removing the supernatant, the nuclei were lysed by resuspending the pellet in 1.2 ml of nuclei lysis buffer (NLB) and incubating on ice for 10 minutes.

Sonication

10. 720 µl of IP dilution buffer (IPDB) was added and the samples were transferred to 5 ml glass falcon tubes (Falcon 2058).
11. The chromatin was sonicated to reduce the DNA length to an average size of 600 bp using the Sanyo/MES Soniprep sonicator. The tip of the probe was dipped to reach approximately halfway down the total level of the liquid sample and the tube was kept constantly on ice (Conditions for sonication like number of bursts, length of bursts and power setting depend on the sonicator tip used). The settings used for the sonicator were:
    - Amplitude: 14 microns
    - Number of bursts: 8
    - Length of bursts: 30 seconds

The samples were allowed to cool on ice for 1 minute between each pulse (5 µl of the sheared chromatin was run on an agarose gel to check sonication, see step 32).

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

Immunoprecipitation

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

14. The chromatin was precleared by adding 100 µl of normal rabbit IgG (Upstate Biotechnology). The samples were incubated for I hour at 4°C on a rotating wheel.
15. 200 µl of homogeneous protein G-agarose suspension (Roche) was added to the precleared chromatin and the samples were incubated for 3-5 hours at 4°C on a rotating wheel.
16. The samples were centrifuged at 1620 g for 2 minutes at 4°C to pellet the protein G-agarose beads and the supernatant was used to set up various immunoprecipitation (IP) conditions in 2 ml microfuge tubes. An aliquot of 270 µl of chromatin was stored at -20°C to be used as input sample for array hybridisations and real-time PCR. An NLB:IPDB buffer at the ratio of 1:4 was prepared to set-up IP conditions as follows:
    - No chromatin – 1350 µl NLB:IPDB buffer
    - No antibody – 675 µl chromatin + 675 µl NLB:IPDB buffer
    - Normal Rabbit IgG – 675 µl chromatin + 675 µl NLB:IPDB buffer + 10 µl rabbit IgG (Upstate Biotechnology)
    - Test IP conditions – 675 µl chromatin + 675 µl NLB:IPDB buffer + 5-20 µg* of test antibody

(*5-20 µg for antibodies raised against histone modifications and 10 µg for the antibodies raised against specific transcription factors). A complete list of antibodies that were tested for this study is provided in chapter 5.

17. The samples were incubated at 4°C overnight on a rotating wheel.
18. The samples were centrifuged at 18000 g for 5 minutes at 4ºC and the lysate/Ab samples were transferred to fresh 2 ml microfuge tubes. 100 µl of homogeneous protein G-agarose suspension was added to each sample and the samples were incubated at 4°C for at least 3 hours on a rotating wheel.
19. The samples were centrifuged at 6800 g for 30 seconds at 4°C to pellet the protein G-agarose beads.
20. The supernatant was removed and the protein G-agarose beads were carefully washed. For each wash, the wash buffer was added, the samples were vortexed briefly, were centrifuged at 6800 g for 2 minutes at 4°C and left to stand on ice for 1 minute before removing the supernatant. The washes were carried out in the following sequence:
    a) The beads were washed twice with 750 µl of cold IP wash buffer 1. The beads were transferred to a 1.5 ml microfuge tube after the first wash.
    b) The beads were washed once with 750 µl of cold IP wash buffer 2.

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

Elution

21. DNA-protein-antibody complexes were eluted from the protein G-agarose beads by adding 225 µl of IP elution buffer (IPEB). The bead pellets were resuspended in IPEB, briefly vortexed and centrifuged at 6800 g for 2 minutes at room temperature.
22. The supernatant was collected in fresh 1.5 ml microfuge tubes. The bead pellets in the original tubes were resuspended in 225 µl of IPEB again, briefly vortexed and centrifuged at 6800 g for 2 minutes. Both the elutions were combined in the same tube.

Reversal of cross-links

23. The reversal of cross-links step was carried out on the Input sample which was stored at -20°C previously. 0.1 µl of RNase A (10 mg/ml, 50 Kunitz units/mg[*], ICN Biochemicals) and 16.2 µl of 5M NaCl (to the final concentration of 0.3 M) was added to the Input DNA sample.
24. Similarly, 0.2 µl of RNase A (10 mg/ml, 50 Kunitz units/mg[*]) and 27 µl of 5M NaCl (to a final concentration of 0.3 M) was added to each of the IP test samples. All the samples including the Input DNA sample were incubated at 65ºC for 6 hours to reverse the cross-links.
25. 9 µl of Proteinase K (10 mg/ml, 20 U/mg, GibcoBRL) was added to each sample and incubated at 45ºC overnight[b].

Extraction of DNA

26. 2 µl of yeast tRNA (5 mg/ml, Invitrogen) was added to each sample just before adding 250 µl of phenol (Sigma) and 250 µl of chloroform[c].
27. The samples were vortexed and centrifuged at 18000 g for 5 minutes at room temperature. The aqueous layer (top layer) was collected in fresh 1.5 ml microfuge tubes and 500 µl of chloroform was added to each sample.
28. The samples were vortexed and centrifuged at 18000 g for 5 minutes at room temperature. The aqueous layer was transferred to a fresh 2.0 ml microfuge tubes.
29. 5 µg of glycogen (5 mg/ml, Roche), 1 µl of yeast tRNA (5 mg/ml, Invitrogen) and 50 µl of 3M NaAc (pH 5.2) was added to each sample and mixed well. The DNA was precipitated with 1375 µl of 100% ethanol and incubating at –70ºC for 30 minutes (or -20°C overnight).
30. The samples were centrifuged at 20800 g for 20 minutes at 4ºC. The DNA pellets were washed with 500 µl of ice-cold 70% ethanol and air-dried for 10-15 minutes.

31. The DNA pellets of the IP samples were resuspended in 50 µl of sterile filtered HPLC water and 100 µl for the Input DNA samples.
32. 5 µl of each sample was run on a 1% agarose 1XTBE gel and visualised with ethidium bromide to check DNA size. Samples were stored at –20°C.

[a]The sheared chromatin can be snap frozen in liquid nitrogen at this stage and the frozen samples should be stored at -70°C. When needed, the samples should be thawed on ice and the experiment carried on as per the protocol.

[b]The samples can be stored at -20°C after the step no. 22. When needed the samples can be thawed at room temperature and the DNA extracted as per the protocol.

[c]Safety Note – The phenol/chloroform steps were carried out in a fume cabinet.

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

2.7. Construction of the SCL Genomic Tiling Path Microarray

2.7.1. Generation of Human and Mouse SCL Genomic Tiling Path Amplicons

A. Primer Design

1. Primers pairs used to amplify PCR products for the human and mouse SCL tiling array were designed from the relevant genomic sequence for the human and mouse SCL. All the sequences were first analysed for their repetitive content using RepeatMasker. Primers were then designed using the masked sequences and Oligo 6 Software (Molecular Biology, Insights) and Primer3 software and website (Whitehead Institute, http//Frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi) (Rozen and Skaletsky 2000).
2. Primer pair sequences were compared against the entire human or mouse genome sequence using e-PCR (Schuler 1997). The primer sequences for human and mouse amplicons are shown in Appendices 1 and 2 respectively.

B. First Round Amplification

1. PCR reactions were set up in 96 well thin-walled Thermowell plates (Costar, 6511) by mixing the following on ice:

1 µl	DNA template (50 ng/µl)
7 µl	1mM dNTP
3.2 µl	10Xtaq Gold buffer II
3.2 µl	DMSO

2.1 µl	$MgCl_2$
0.2 µl	5 U/µl Gold AmpliTaq
1 µl	Forward primer with 5' adaptor sequence
1 µl	Reverse primer
13.3 µl	Sterile ddH_2O
32.0 µl	

2. The cycling conditions for the reaction were as follows: 95ºC for 8 min; then 35 cycles of: 95ºC for 30 sec, 56ºC for 45 sec, 72ºC for 45 sec; followed by 72ºC for 10 min.
3. The PCR products were analyzed on 2.5% agarose 1 X TBE slab gels and stained with ethidium bromide for visualisation.
4. Depending on the intensity of the band, 1-4 µl of the first round PCR products were transferred to new 96-well plates in 100 µl TE. These served as the source of template for the second round reaction and were stored at –20ºC.

C. First Round Amplification (alternative method)

First round PCR was also performed using DNA from bacterial PAC and BAC clones as follows:

5. PCR reactions were set up in 96 well thin-walled Thermowell plates (Costar, 6511) by mixing the following on ice:

5 µl	DNA Template (1/100 dilution of miniprep DNA)
2 µl	10 X PCR buffer
1 µl	10mM dNTP mix
0.5 µl	gene-specific primer (sense) with 5' adaptor sequence
0.5 µl	gene-specific primer (antisense)
0.125 µl	5 U/µl Taq polymerase (Perkin Elmer-Cetus)
4 µl	5M Betaine (Sigma)
6.875 µl	Sterile ddH_2O
20.0 µl	

6. The following thermal cyclic conditions were used with varying number of cycles: 95ºC for 5 min; then 30/35 cycles of: 95ºC for 1 min, 60/55ºC for 1 min 30 sec, 72ºC for 1 min 30 sec; followed by 72ºC for 5 min.
7. The PCR products were analyzed on 2.5% agarose 1 X TBE slab gels and stained with ethidium bromide for visualization.

8. The single gel bands were picked up as gel stabs using P20 filter tips and transferred to deep well plates and sterile water was added to each and left overnight at 4ºC for the DNA to diffuse into the water. This DNA was used as a template for the second round PCR and was stored at –20ºC.

D. Second Round Amplification

1. PCR reactions were set up in 96 well thin-walled Thermowell plates (Costar, 6511) by mixing the following on ice:

15 µl	DNA template as prepared in the first round
6 µl	10 X PCR buffer
3 µl	10mM dNTP mix
1.5 µl	5'-aminolink universal primer (sense) 200 ng/µl
1.5 µl	Reverse primer 20 µM (antisense)
0.375 µl	5 U/µl Taq polymerase (Perkin Elmer-Cetus)
32.625 µl	Sterile ddH_2O
60.0 µl	

2. The thermal cyclic conditions used were: 95ºC for 5 min; then 25 cycles of 95ºC for 1 min, 45ºC for 1 min 30 sec, 72ºC for 1 min 30 sec; followed by 72ºC for 5 min.
3. The PCR products were analyzed on 2.5% agarose 1 X TBE slab gels and stained with ethidium bromide for visualization.
4. 50 µl of the PCR products were filtered through into new 96 well plates (Falcon, Microtest U-Bottom Tissue culture plates) along with 15 µl of 4 X Spotting buffer and stored at –20ºC until required for array printing.

2.7.2. Microarray Slide Printing and Processing

1. The DNA elements were arrayed onto amino binding slides (Motorola) at 20-25ºC, 40-50% relative humidity using the MicroGrid II arrayer (Biorobotics/Apogent Discoveries).
2. The array elements were spotted in a 16 block format with spots in quadruplicate or duplicate in the first and second generation arrays respectively.
3. The slides were transferred into a microscope slide rack and placed in a humid chamber (NaCl saturated with water in an air-tight container) and incubated for 24-72 hours at room temperature.
4. The slides were removed from the humid chamber and immersed in a 1% (w/v) solution of ammonium hydroxide and incubated for 5 minutes with gentle shaking.

5. The slides were then transferred to a solution of 0.1% (w/v) sodium dodecyl sulphate and incubated for 5 minutes with gentle shaking.
6. The slides were briefly rinsed in Milli-Q ddH_2O (Miili-Q plus 185 purification system) at room temperature and then placed in 95ºC Milli-Q ddH_2O for 2 minutes to completely denature the bound DNA elements.
7. The slides were transferred to ice-cold Milli-Q ddH_2O and then briefly rinsed two times in Milli-Q ddH_2O at room temperature.
8. The slides were dried by spinning at 180 g for 5 minutes.
9. The slides were stored in a slide box and kept at room temperature in a cool dry place until used.

2.8. Hybridisation of the array

2.8.1. Random Labeling of DNA samples

The SCL genomic tiling path array hybridisations using total genomic DNA and/or DNA obtained by chromatin immunoprecipitation were set up either manually or using the Tecan HS 4800 hybridisation station (an automated hyb-station). The array-CGH hybridisations using the 1 Mb BAC genomic microarrays were set up using Tecan automated hyb-station. For both hybridisation set-ups, the DNA was labeled using BioPrime Random Labeling Kit (Invitrogen) but with different labeling reactions as described below.

A. Labeling method used with manual hybridisation set-up

1. The following reagents were mixed on ice in a 1.5 ml microfuge:
    - 40 µl 2.5 X Random primer solution
    - x µl DNA *
    - (44-x) µl sterile H_2O

* The DNA amount that was labeled was different for each sample. For example, 600 ng of total genomic DNA was labeled per reaction for array-CGH hybridisations whereas the amount of DNA labeled in the ChIP-chip hybridisations was ~200 ng.

2. This mixture was heated at 100°C for 10 minutes to denature the DNA and then snap-chilled on ice. The following reagents were added to the tube on ice:
    - 10 µl 10 X dNTP mix (a)
    - 4 µl 1 mM Cy3/Cy5 labeled dCTP (1 mM Cy3-dCTP, 1 mM Cy5-dCTP, Amersham)
    - 2 µl Klenow fragment (40 U/µl)

The final volume per labeling reaction was 100 µl.

3. The reagents were mixed gently but thoroughly and incubated at 37°C overnight.
4. 10 µl stop buffer was added to the reaction mix to terminate the reaction.

B. Labeling method used with the hybridisation set-up using Tecan

1. The following reagents were mixed on ice in a 1.5 ml microfuge:
    - 60 µl 2.5 X Random primer solution
    - x µl DNA *
    - (70.5-x) µl sterile H_2O

* The DNA amount labeled was different for each sample. For example, 450 ng of total genomic DNA was labeled per reaction for array-CGH hybridisations whereas the amount of DNA labeled in the ChIP-chip hybridisations was ~200 ng.

2. This mixture was heated at 100°C for 10 minutes to denature the DNA and then snap-chilled on ice. The following reagents were added to the tubes on ice:
    - 15 µl 10 X dNTP mix (b)
    - 1.5 µl 1 mM Cy3/Cy5 labeled dCTP (1 mM Cy3-dCTP, 1 mM Cy5-dCTP, Amersham)
    - 3 µl Klenow fragment (40 U/µl)

The final volume per labeling reaction was 150 µl.

3. The reagents were mixed gently but thoroughly and incubated at 37°C overnight.
4. 15 µl stop buffer was added to the reaction mix to terminate the reaction.

Note: Only one labeling reaction was required for the SCL genomic tiling path array but two labeling reactions were used for the 1 Mb BAC genomic array.

2.8.2. Purification of labeled DNA Samples

The DNA samples labeled using the two labeling reactions described above were purified using the same protocol as described below.

1. Micro-spin G50 columns (Pharmacia Amersham) were used to remove the unlabeled nucleotides from the labeled DNA samples.
2. Two columns were used for each of the 100 µl labeling reactions and three columns were used for each of the 150 µl labeling reactions (100 µl and 150 µl being the final volumes of the labeling reactions).
3. The resin was resuspended in the columns by vortexing gently. The caps were loosened one-fourth and the bottom of the tubes snapped off. The columns were placed in 2.0 ml microfuge tubes and centrifuged at 1700 g for 1 minute.

4. 50 µl of sterile filtered HPLC water was applied to the resin-bed and the columns were centrifuged at 1700 g for 1 minute.
5. The columns were placed in fresh 1.5 ml microfuge tubes and the labeled DNA samples were carefully applied to the resin-bed. The columns were then centrifuged at 1700 g for 2 minutes.
6. The purified DNA samples were collected in the 1.5 ml microfuge tubes and the samples from the same labeled reaction were pooled together. The final volumes for the labeled DNA samples were approximately 120 µl (from 100 µl reactions) and 180 µl (from 150 µl reactions).
7. 5 µl of each labeled DNA was analyzed on a 1% agarose 1 X TBE gel and stained with ethidium bromide for visualization. The samples were used for hybridisation and stored in an opaque box at –20ºC.

2.8.3. Manual set-up of array hybridisations

The grid size of the SCL genomic tiling path array was 2 X 2 cm and the "open-well" method described below was used for the hybridisations that were set-up manually.

Preparation of the slide and the hybridisation chambers

1. The slide was prepared for hybridisation as follows:

   Based on the position of the spotted array grid on the slide, four layers of rubber cement were applied carefully around the grid to make a well. All the corners of the well were sealed properly to avoid leaking during hybridisation. The rubber cement was allowed to dry completely before setting up the hybridisations.
2. Two humidified slide chambers were prepared for the pre-hybridisation and hybridisation steps. The pre-hybridisation chamber was made from a square petri-dish and two cocktails sticks were glued to the floor of the dish. 2 X SSC/40% formamide (Fluka) solution was applied to Whatman paper strips lining the floor of the slide chamber.
3. Similarly, hybridisation chambers were prepared by lining the top and bottom of a slide mailer (R. A. Lamb) with whatman paper strips soaked in 2 X SSC/20% formamide (Fluka) to keep it humid.

Preparation of pre-hybridisation and hybridisation solutions

4. Two 2.0 ml microfuge tubes were set up for each slide - one with the pre-hybridisation mix and the second with the hybridisation mix. The following reagents were mixed together on ice and kept in the dark as much as possible:

Tube 1 (Pre-hybridisation mix)

- 400 µg Herring sperm DNA (10 mg/ml, Sigma)
- 67.5 µl Human Cot 1 DNA (Invitrogen)
- 25 µl 3M NaAc (pH 5.2)
- 800 µl 100% ice-cold ethanol

Tube 2 (Hybridisation mix)

- 120 µl Cy3-labeled DNA
- 120 µl Cy5-labeled DNA
- 90 µl Human Cot 1 DNA (Invitrogen)
- 35 µl 3M NaAc (pH 5.2)
- 1100 µl 100% ice-cold ethanol

5. The DNA samples in both reactions were precipitated by incubating at –70ºC for 30 minutes (or -20°C overnight). The DNA samples were centrifuged at 18000 g for 20 minutes at room temperature. The pellets were washed with 500 µl of ice-cold 80% ethanol and air dried.
6. The pre-hybridisation DNA pellet was resuspended in 75 µl of Hyb-3 buffer and the hybridisation DNA pellet in 30µl of Hyb-3 buffer and 3 µl of yeast tRNA (100 µg/µl, Invitrogen). Heating the solutions at 70°C helps to resuspend the pellets properly.

Array hybridisation and washing

7. Both the pre-hybridisation and the hybridisation DNA mixes were denatured for 10 minutes at 100ºC and quickly centrifuged.
8. The hybridisation mix was snap-chilled on ice and then kept at 37ºC for an hour for pre-annealing.
9. The denatured herring sperm/Cot 1 pre-hybridisation mix was applied very carefully to the slide covering the whole array area.
10. The slide was transferred to the humidified slide chamber (see step 2 above) and incubated at 37ºC in a rocking incubator at 5 rpm for an hour.
11. After the incubation, the slide was removed from the slide chamber and held at an angle to let the pre-hybridisation solution run into a corner to remove as much solution as possible.
12. The hybridisation mix was applied very carefully covering the array area and made sure that there were no bubbles within the array area.

13. The slide was transferred into a humidified slide chamber (see step 2 above) and the slide chamber was tightly sealed with parafilm to minimize evaporation and thus avoid the slide from drying out during incubation.
14. The slide was incubated for 48 hours at 37ºC in a gently rocking incubator (5 rpm).
15. The slide was removed from the incubator and the rubber cement was removed very carefully.
16. The slide was given a quick rinse in PBS/0.05% Tween 20 (BDH) to remove excess hybridisation solution.
17. The slide was first washed in PBS/0.05% Tween 20 (BDH) for 10 minutes at room temperature on a shaking platform.
18. The second wash was performed in 50% Formamide/2 X SSC at 42ºC for 30 minutes in a rocking incubator (maximum speed).
19. The third and final wash was performed in PBS/0.05% Tween 20 for 10 minutes at room temperature on a shaking platform.
20. The slides were then centrifuged at 180 g for 5 minutes to dry and were ready to be scanned. The slides were always stored in a light-proof slide box.

2.8.4. Array hybridisation set-up using the Tecan hybridisation station

Tecan HS 4800 hybridisation station (automated hyb-station) was used to set-up array-CGH hybridisations (on SCL tiling path array and 1 Mb BAC genomic array) and some of the ChIP-chip hybridisations (on the SCL genomic tiling path array). Tecan is a fully automated hyb station where the microarray slides are loaded on the Tecan's slide holder and the hybridisation mix is agitated to ensure even hybridisation. The SCL array area was 2 X 2 cm and consequently the smaller chambers were used on the Tecan to set up the hybridisations. However, larger chambers were required for the 1 Mb BAC genomic arrays

Preparation of the hyb station

1. The slide holders and the slide chambers were carefully cleaned and the slides were loaded on to the slide holder.
2. The wash solutions were prepared (described in section 2.1) and poured into the wash bottles of the Tecan and the hyb station was primed to remove any air bubbles in the liquid channels.

Preparation of pre-hybridisation and hybridisation solutions

Note: The following volumes were doubled for the larger chambers for the 1 Mb BAC genomic arrays.

3. Two 2.0 ml microfuge tubes were set up for each slide - one with the pre-hybridisation mix and the second with the hybridisation mix. The following reagents were mixed together on ice and kept in the dark as much as possible:

Tube 1 (Pre-hybridisation mix)

- 40 µl Herring sperm DNA (10 mg/ml, Sigma)
- 67.5 µl Human Cot 1 DNA (Invitrogen)
- 12 µl 3M NaAc (pH 5.2)
- 300 µl 100% ice-cold ethanol

Tube 2 (Hybridisation mix)

- 180 µl Cy3-labeled DNA
- 180 µl Cy5-labeled DNA
- 135 µl Human Cot 1 DNA (Invitrogen)
- 55 µl 3M NaAc (pH 5.2)
- 1100 µl 100% ice-cold ethanol

4. The DNA samples in both reactions were precipitated by incubating at –70ºC for 30 minutes (or -20°C overnight). The DNA samples were centrifuged at 18000 g for 20 minutes at room temperature. The pellets were washed with 500 µl of ice-cold 80% ethanol and air dried.
5. The pre-hybridisation and hybridisation DNA pellets were resuspended in 120 µl of Tecan-hyb buffer each and 3 µl of yeast tRNA (10̲0̲ µg/µl, Invitrogen) was added to the hybridisation mix. Heating the solutions at 70ºC helps to resuspend the pellets properly.

Array hybridisation and washing

6. The pre-hybridisation and hybridisation solutions were denatured at 100°C for 10 minutes. The hybridisation solution was snap-chilled on ice and then pre-annealed at 37°C for 1 hour. The pre-hybridisation solution was kept at 70°C until applied to the slide.
7. After vortexing, using a displacement pipette, 100 µl of the pre-hybridisation solution was injected onto the slide very slowly and carefully to avoid any air bubbles.
8. The pre-hybridisation step was performed at 37°C for 1 hour. The slides were washed once with PBS/0.05% Tween and dried with short blasts of nitrogen gas.

9. 100 µl of the hybridisation solution was injected slowly onto the slide using a displacement pipette. The hybridisation step was performed at 37°C for 48 hours (the 1 Mb BAC array was hybridised for 24 hours).
10. The slide washing was carried out on the Tecan which was programmed to perform the washes in the sequence listed in Table 2.2.

Steps	Wash Solutions	Temperature	No. of Washes	Wash Duration	
				Wash time	Soak time
1	PBS/0.05% Tween	37°C	10	1 min	30 secs
2	0.1 X SSC	52°C	5	1 min	2 min
3	PBS/0.05% Tween	R/T	10	1 min	30 secs
4	HPLC water	R/T	2	30 secs	

Table 2.2: Wash steps for hybridisations performed on the Tecan. The solutions were prepared in advance (section 2.1) using HPLC water. R/T stands for room temperature.

11. The slides were dried on the Tecan with short blasts of nitrogen gas and stored in the dark until scanning.

2.8.5. Scanning and Data Analysis for ChIP-chip experiments

After the images were scanned, two softwares – Quantarray and Scanarray Express (Perkin Elmer) were used to analyse the scanned images as described below. The analysis output from both the softwares were analysed using an excel spreadsheet.

1. Cy3 and Cy5 images at 5 µm resolution were acquired using the Scanarray 4000 confocal laser-based scanner (Perkin-Elmer) using a laser power of 100% and a photo multiplier tube (PMT) value of between 80%-85%.
2. Fluorescent intensities of each spot on the array were quantitated using Quantarray (version 3.0) (Perkin-Elmer) using the fixed circle quantitation methods. Using this software, each spot has to be aligned manually to obtain the mean signal intensity values (intensity-background) for each spot. Data was normalized to the median ratio for each array element.
3. The other software, ScanArray Express (Perkin Elmer) was used to quantitate the fluorescent intensities of the spots using the fixed circle quantitation and the TOTAL normalization methods. This software can automatically locate the spot position on the scanned image of the array to obtain the signal intensity values. Mean intensity ratios (intensity-background) were reported for each spot representing an array element.

4. Further analysis of the ChIP-chip data was carried out in a Microsoft Excel spreadsheet in an identical manner irrespective of the software used to quantitate the signal intensities. In the spreadsheet, each array element was associated with its genomic sequence position information.
5. The array elements which did not perform reliably in the validation experiments or sequencing of the products were removed from the final data-set (as described in chapter 3). In total, 360 array elements in human and 411 array elements in mouse were used for the final analysis.
6. Mean ratios, standard deviations (SD) and coefficients of variation (CV) were calculated for the replicate spots representing each array element (quadruplicate spots in the 1st generation array and duplicate spots in the final array, described in chapter 3).
7. The data was visualized by plotting the mean ratios of all array elements along the Y-axis and the respective genomic positions along the X-axis (see chapters 4, 5 and Appendices 6 to 11 for the ChIP-chip profiles generated for various assays performed for this study).
8. The baseline levels of each plot was normalised to a value of one, so that all the experiments could be directly compared from this baseline value. This was done by determining the median ratio for each experiment and dividing all the ratios (obtained in that experiment) by this value.

2.9. Sequence Analysis

A. Comparative genomic sequence analysis of mouse and human SCL loci

1. The genomic sequences for the human and mouse SCL loci represented on the tiles were aligned using the mLAGAN web server (www.lagan.stanford.edu/lagan_web/index) (Brudno et al. 2003).
2. The alignment also included the corresponding rat locus, thus providing a three-species alignment which was of better quality than an alignment of only the human and mouse syntenic regions.
3. The sequence co-ordinates used for all three species were as follows: human (chr1: 47,262,288-47,518,922, NCBI build 35), mouse (chr4:113,872,305-114,130,000, NCBI build 34) and rat (chr5:135,282,000-135,540,000, RGSC 3.4).
4. Based on this alignment, sequence homology peaks between human and mouse were determined as percentage identity across 100 bp windows of sequence. Peaks

of non-coding sequence conservation were then associated with histone acetylation and methylation data (in chapters 4 and 6).

5. The mean sequence conservation of non-coding DNA for each array element was determined and the array elements were then grouped by level of conservation (in 5% intervals). From this, the average level of acetylation or methylation for each group was calculated.
6. Simple regression analysis was performed based on the acetylation levels of array elements and their percentage identity of non-coding sequence conservation in these groupings.

B. Local comparative genomic sequence alignments of SCL regulatory regions

Local comparative genomic sequence alignments of SCL regulatory regions found in human, chimp, mouse, rat and dog were obtained from the UCSC Genome Browser (http://www.genome.ucsc.edu/). These included sequences representing the human regulatory regions known as SCL -12, -7, p^{1a}, +3 and +51 (see chapters 4 and 7). The genomic coordinates of these sequences are shown in the relevant sections.

C. Identification of conserved transcription factor (TF) sites

TF binding sites were identified using TFSEARCH (http://www.cbrc.jp/research/db/TFSEARCH.html) (Heinemeyer et al. 1998) and TESS (http://www.cbil.upenn.edu/tess/) (Schug and Overton 1997) web servers.

2.10. Statistical analysis

z-tests and simple regression analyses were performed using the Microsoft Excel macro (chapters 4 and 6).

2.11. Cluster analysis of histone modification data

1. Cluster and TreeView programs were used to functionally cluster and visualize the histone modification data (Eisen et al. 1998).
2. The ratios obtained in the ChIP-chip experiments were log transformed (i.e. all the data values, x, were replaced by $\log_2(x)$.
3. Hierarchical clustering was performed (Cluster by Eisen et al. 1998) by using “average linkage clustering” with the correlation (centred) similarity metric and no weightings calculated for the individual array elements.
4. The output files were then visualized in TreeView.

2.12. Real-time PCR

A. Primer design

1. Primer pairs for all the real-time PCR assays, performed for this study, were designed by using the Primer Express software version 2.0 (Applied Biosystems).
2. Primer pair sequences were compared against the entire human genome using e-PCR (Schuler 1997). The amplicons generated by these primer pairs were between 70 bp to 150 bp in length.
3. Standard curves were generated for the primer pairs used in the ChIP verification assays. From these, the PCR yields were calculated for each of the tested primer pairs.

The complete lists of all the primer pair sequences, used in the real-time PCR assays, are provided in Appendices 3, 4 and 5 and referenced in the relevant sections of this thesis.

B. Real-time PCR amplification

The chromatin immunoprecipitated (ChIP) DNA samples were used to set-up quantitative real-time PCR as follows:

1. The ChIP DNA samples were diluted to 1 in 10 dilution i.e. 5 µl of the sample was resuspended in 45 µl of sterile filtered HPLC water.
2. The SYBR green PCR was set-up in a 96-well plate (Applied Biosystems) in a 25 µl reaction, in duplicate or triplicate for each sample, by mixing the following reagents on ice:
    - 2.5 µl Water
    - 5 µl 1.5 µM forward and reverse primer mix
    - µl SYBR green PCR mix (Applied Biosystems)
    - 5 µl ChIP DNA samples
3. PCR was performed on a 7700 sequence detection system (Applied Biosystems).The following thermal cyclic conditions were used: 50ºC for 2 min; 95ºC for 10 min; then 40 cycles of: 95ºC for 15 sec and 60ºC for 1 min.
4. C_T values were extracted using Sequence Detector 1.7a (Applied Biosystems) with the same threshold and the ΔC_T values were determined as follows:

   $\Delta C_T = C_T$ input - C_T ChIP sample
5. Fold enrichments were calculated by using the following formula:

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

Mean fold enrichments were calculated for each assay.

2.13. Procedures for fluorescence in-situ hybridisation (FISH)

2.13.1. Preparation of metaphase chromosomes

Fixed metaphase preparations were prepared from human haematopoietic cell lines K562, Jurkat, HL-60 and HPB-ALL by the following method:

1. Fresh cultures for each cell line were grown to confluency and sub-cultured to get approximately 50% confluent cultures and left overnight.
2. 40 µl of BrdU (3 mg/ml) was added to 20 ml of culture and mixed well. The culture was incubated for 3 hours at 37°C.
3. 20 µl ethidium bromide (10 mg/ml) was added to the 20 ml culture (final concentration: 10 µg/ml). At the same time, 40 µl colcemid (10 µg/ml) was also added to a final concentration of 0.02 µg/ml and the cells were incubated for 2 hours.
4. At the time of harvest, the contents of the culture flask were transferred to a 50 ml falcon tube and the cells were pelleted by centrifuging at 259 g for 8-10 minutes.
5. The supernatant was discarded to leave the pellet as dry as possible. The cell pellet was loosened gently but thoroughly by flicking the base of the tube.
6. 10 ml of pre-warmed (to 37°C) 75 mM KCl was added and the cells were gently resuspended to break any clumps of cells. The cell suspension was incubated at 37°C for 12-15 minutes.
7. 1 ml of fresh ice-cold 3:1 methanol:glacial acetic acid (BDH) fix was added to the cell-suspension and was immediately mixed thoroughly by gentle swirling. The cells were pelleted by centrifugation at 259 g for 10 minutes.
8. The supernatant was removed and the cell pellet was loosened thoroughly. 10 ml of fixative was added with constant but gentle mixing. The cell suspension was centrifuged at 259 g for 10 minutes.
9. The fixation step no. 8 was repeated a further 3 times and after the last centrifugation step, the supernatant was removed to leave the cell pellet as dry as possible.
10. The fixed cell pellet was resuspended in a small volume (1 to 2 ml) of 3:1 fix so that the solution remained cloudy.
11. The fixed metaphase preparation was assessed by dropping a small aliquot onto a microscope slide from a pipette tip. Metaphase spreads can be detected under phase contrast using a light microscope.
12. The fixed metaphase preparations for all cell lines were stored at -20°C until required.

2.13.2. Probe labelling by nick translation

PAC (RP1-18D14) and BAC (RP11-332M15) DNA probes for use in FISH were labeled by nick translation by the following method:

1. 1 µg of DNA (5 µl) was labeled in a 25 µl reaction by adding the following to a 1.5 ml microfuge tube on ice:
    - 2.5 µl 10 X nick translation buffer
    - 1.9 µl nick translation dNTP mix
    - 0.7 µl Biotin-16-dUTP or Digoxigenin-11-dUTP (Boehringer)
    - 0.5 µl DNase I (1 µg/ml working solution)
    - 1 µl DNA polymerase I (10 U/µl, Sigma)
    - 5 µl DNA (1 µg)
    - x µl sterile ddH_2O (to make the final volume to 25 µl)
2. The contents were mixed by lightly flicking the tube. The tube was spun microfuged briefly to bring down the solution.
3. The solution was incubated at 14°C for 90 minutes.
4. To stop the reaction, 2.5 µl of 0.5M EDTA was added to the tube. The contents were mixed thoroughly and the tube was transferred to ice.
5. 2.5 µl of the sample was analysed by electrophoresis on a 1% 1 X TBE agarose gel and visualized by staining with ethidium bromide.
6. Meanwhile, the labeled DNA sample was precipitated by adding 2.5 µl of 3M NaAc (pH 7.0) followed by 1 ml of ice-cold 100% ethanol. The solution was mixed well by inverting and incubated at -70°C for 30 minutes (or overnight at -20°C).
7. The DNA was pelleted by spinning in a microfuge at 18000 g for 10 minutes. The supernatant was discarded; the pellet was washed with 500 µl of ice-cold 70% ethanol and microfuged at 18000 g for 5 minutes.
8. The supernatant was removed to leave the pellet as dry as possible. The pellet was air-dried by keeping open tubes at 37°C for 25 to 30 minutes.
9. The DNA was resuspended in 10 µl of $T_{0.1}E$ buffer by flicking the tube and the DNA sample was stored at -20°C until required.

2.13.3. Metaphase spread slide preparation

Fixed metaphase spreads (of K562, Jurkat, HL-60 and HPB-ALL) were prepared on clean glass microscope slides by the following procedure:

1. The fixed metaphase suspension of a cell line was brought to room temperature (initially stored at -20°C).
2. The cell-suspension was mixed by gently flicking the tube or, if necessary, using a pasteur pipette.
3. Using a fine-tip pasteur pipette, a single drop of the cell-suspension was dropped onto a clean glass microscope slide kept on a horizontal surface, immediately followed with 1 drop of fresh 3:1 fix.
4. The slide was left to air-dry in the horizontal position. Once dry, the slide was examined for metaphase spreads under a phase contrast light microscope.
5. The area of spread cells on the slide was marked with a diamond pen.
6. The slides were fixed by incubating in fresh 3:1 fix at room temperature for 30-60 minutes.
7. After air-drying, the slides were dehydrated by passing through a fresh 70%, 70%, 90%, 90% and 100% ethanol series for 1 minute each.
8. After air-drying, the slides were incubated in acetone for 10 minutes at room temperature.
9. The slides were air-dried and stored in a sealed box containing desiccant until used (usually within the week they were made).

2.13.4. Hybridisation of metaphase spreads

Human PAC and BAC DNA probes were hybridised onto K562, Jurkat, HL-60 and HPB-ALL metaphase spreads in two-colour hybridisations by the following method:

1. A coplin jar of 70% formamide (70% formamide/2 X SSC) was pre-warmed to 65°C in a waterbath.
2. The following reagents were added to a 1.5 ml microfuge tube on ice:
    - 0.5 µl labeled DNA each from the two probes (30-50 ng)
    - 2 µl Human C_ot-1 DNA (Invitrogen)
    - 0.2 µl placental DNA
    - 9.8 µl Hyb-3 buffer
3. The contents of the tube were mixed thoroughly by vortexing and spun briefly in a microfuge to bring down the solution.
4. The probe mix was denatured by incubating at 65°C for 10 minutes.
5. The probe mix was then pre-annealed by incubating at 37°C for 20 minutes to 3 hours.

6. Meanwhile, the metaphase-spread slides were denatured in 70% formamide at 65°C for 1 minute 45 seconds (the time duration of slide denaturation is crucial to obtain good morphology chromosomes).
7. The denatured slides were quenched in 70% ice-cold ethanol for 1-2 minutes.
8. The slides were then dehydrated by passing through a fresh 70%, 70%, 90%, 90%, 100% ethanol series, for 1 minute in each jar. The slides were then air-dried.
9. The pre-annealed probe mix was pipetted onto the slide and the metaphase-spread spot was covered with a clean 22 X 22 mm coverslip.
10. The edges of the coverslip were sealed with rubber cement.
11. The slides were incubated overnight at 37°C in a humid atmosphere.

2.13.5. Detection of labeled probes

After hybridisation, biotin- and digoxigenin-labeled probes were detected immunochemically by the following three layer detection method carried out on an automated wash-station, Cadenza immunostainer (Shannon).

Pre-Cadenza steps:

1. 2 coplin jars of 2 X SSC and 2 coplin jars of 50% formamide (50% formamide/1 X SSC) were pre-warmed to 42°C in a waterbath.
2. The dried rubber cement sealing from the slides was removed by soaking the slides in 2 X SSC at room temperature.
3. After the coverslips came off the slides, the slides were passed through 4 stringency washes. The first wash was carried out by incubating the slides in 2 X SSC at 42°C for 5 minutes.
4. The slides were then washed twice by incubating in 50% formamide at 42° for 5 minutes each.
5. The slides were finally washed by incubating in 2 X SSC at 42°C for 5 minutes.

Cadenza washes:

6. The slides were then loaded very carefully onto the Cadenza.
7. The antibody solutions were prepared in blocking buffer and 100 µl of the prepared solution was used per slide.
8. Three layer antibody detection was performed on the Cadenza in the following order:
    - The first detection layer was 1:333 dilution of 1 mg/ml avidin conjugated to Texas Red (Molecular Probes) and 1:500 dilution of mouse anti-digoxigenin conjugated to FITC (Sigma).

- The second detection layer was 1:250 dilution of 1 mg/ml biotinylated anti-avidin (Vector Labs) and 1:250 dilution of goat anti-mouse FITC (Sigma).
- The third detection layer was 1:333 dilution of 1 mg/ml avidin conjugated to Texas Red (Molecular Probes).

9. Between detection layers, the slides were washed with 4 X SSC/0.05% Tween 20 (BDH).
10. The final wash on the Cadenza after the three layer detection was with blocking buffer.

Post-Cadenza steps:

11. The slides were taken off the Cadenza and rinsed briefly in 2 X SSC.
12. The slides were then stained for 2-3 minutes in 0.8 µg/ml 4,6-Diaminidine-2-phenylindole dichloride (DAPI, Boehringer) solution prepared in 2 X SSC.
13. After staining, the slides were again rinsed briefly in 2 X SSC.
14. The slides were then dehydrated by passing through a fresh 70%, 70%, 90%, 90% and 100% ethanol series allowing 1 minute in each coplin jar.
15. The slides were air-dried.
16. 20 µl of anti-fade mountant (Citiflour AF1) was applied to each slide and 20mm X 50mm clean coverslip was placed on the slide carefully to avoid air-bubbles.
17. The edges of the coverslip were sealed by applying clear nail-varnish. The slides were stored at 4°C.

2.13.6. Acquisition of FISH images

1. Slides were imaged using an Axioplan 2 microscope (Zeiss) with a CoolSNAP HQ camera (Photometrics) and narrow band pass filters (Chroma).
2. The slides were scanned to locate metaphase spreads at X20 magnification using a DAPI filter and probes signals were detected using a FITC and a Texas Red filter.
3. Metaphase spreads were captured at X100 magnification using SmartCapture 2 (Digital Scientific) imaging software.

2.14. RNA Extraction

Fresh cultures of cell lines K562, Jurkat, HL-60 and HPB-ALL were grown in the appropriate media (section 2.4.1). When the cultures were confluent, RNA was extracted from each of the cell line by the following procedure:

1. Cultured cells were counted using a haemocytometer (to get approximately $1X10^7$ cells). The cells were transferred to 50 ml falcon tubes and spun down by centrifuging at 259 g for 10 minutes.
2. The cell pellet was washed with 10 ml of PBS and the cells were spun down again by centrifuging at 259 g for 10 minutes.
3. 1 ml of Trizol reagent (GibcoBRL) was added per $1X10^7$ cells and mixed well and the sample was incubated at room temperature for 5 minutes.
4. 0.2 ml of chloroform (Sigma) was added per 1 ml of Trizol reagent used.
5. The sample was mixed vigorously for 15 seconds and incubated at room temperature for 2-3 minutes.
6. After incubation, the sample was centrifuged at 18000 g for 15 minutes at 2-8°C.
7. The aqueous phase (upper layer) of the centrifuged sample was transferred to a new 2 ml microfuge tube and 0.5 ml of isopropanol (BDH AnalaR) was added per 1 ml of the Trizol reagent used. The solutions were mixed by inverting the tube.
8. The sample was incubated at room temperature for 10 minutes and centrifuged at 18000 g for 15 minutes at 2-8°C. The RNA should now be visible as a pellet at the bottom of the tube.
9. The supernatant was removed very carefully and the pellet was washed once with 75% ethanol (BDH AnalaR). At least 1 ml of 75% ethanol was used per 1 ml of Trizol reagent used.
10. The sample was vortexed and centrifuged at 18000 g for 5 minutes at 2-8°C.
11. The supernatant was removed and the pellet was air-dried.
12. The pellet was resuspended in 50-100 µl DEPC water (0.1% diethyl pyrocarbonate dissolved in ddH_2O and autoclaved) and incubated at 55-60°C until the pellet was completely dissolved.
13. The total RNA extracted was quantitated using spectrophotometer.
14. The RNA quality was assessed by electrophoresis of 2 µg of the sample on 1% agarose 1 X TBE minigel (made up with DEPC water).
15. 3 X volumes of 100% ethanol (BDH AnalaR) was added to the aqueous sample and the samples were then stored at -70°C.

Chapter 3

Construction and Validation of the SCL Genomic Tiling Path Microarrays and Characterization of SCL Expressing and Non-expressing Cell Lines

3.1. Introduction

In recent years, genomic microarrays have emerged as a preferred platform for genomic analysis including annotating the transcriptome, elucidating DNA-protein interactions, and for comparative genomic hybridisations to detect genomic copy number changes among many other applications. The development and use of chromatin-immunoprecipitation coupled with genomic microarrays (ChIP-chip) has greatly facilitated the elucidation and annotation of functional elements and site-specific DNA-protein interactions in the genomes of various organisms (discussed in chapter 1, section 1.6.4) including mammalian genomes.

Genomic tiling arrays (representing non-repetitive DNA across a genomic region) were first used in the mammalian system to map the binding distribution of the haematopoietic transcription factor GATA-1 across ~75 kb of the human β-globin locus in K562 cells (Horak et al. 2002). Since then, tiling arrays have been used extensively to map DNA-protein binding sites across regions of interest, whole chromosomes or in some cases across whole genomes (Martienssen et al. 2005; Martone et al.2003; Euskirchen et al. 2004; Cawley et al. 2004; Pokholok et al. 2005; Schübeler et al. 2004; Kim et al. 2005). Although complete tiling arrays of the human genome studies would prove very useful in elucidating genome-wide DNA-protein interactions, a major limitation is the huge costs involved to carry out such studies. To this end, it is more realistic to use tiling path arrays for a discrete genomic region which would be an ideal resource from which to identify novel regulatory elements and interactions in their genomic context. Thus, a tiling array across the SCL locus, would greatly aid in the understanding of the complex regulation of SCL expression.

The SCL locus is well-characterized as a number of regulatory regions directing SCL expression in distinct compartments during embryonic development are already known (see chapter 1, section 1.9.5). Several studies have identified regulatory sequences located distant from SCL, for example, the erythroid enhancer located 40 kb downstream of SCL promoter 1a in mouse (Delabesse et al. 2005). The identification of this enhancer, suggests that in order to fully understand the regulation of SCL and identify all

of its regulatory elements, it is necessary to examine sequences which are located distant from SCL – both upstream and downstream from the gene itself.

Long range sequence comparisons between the human and mouse loci revealed that the genomic regions surrounding the SCL gene on either side are structurally similar in the two species (Gottgens et al. 2000). The upstream region of SCL contains the SIL gene which is an immediate-early gene and ubiquitously expressed in proliferating cells (Izraeli et al. 1997). The downstream region contains the MAP17 gene which is expressed at significant levels in the epithelial cells of human adult kidney (Kocher et al. 1996). However, downstream of the 3' end of the MAP17 gene is the cytochrome p450 gene cluster - the genes in this cluster are not orthologous in human and mouse (Gottgens et al. 2001) and are mainly expressed in liver and kidney (Henderson et al. 1994).

The majority of the studies undertaken to understand the regulation of SCL expression have been carried out in mice using mouse cell lines, transfection assays and transgenic models (Gottgens et al. 1997; Sanchez et al. 1999; Sinclair et al. 1999). Similar studies in humans have utilized human haematopoietic cell lines and tissues to understand biological function of SCL in blood development (Bernard et al. 1992; Leroy-Viard et al. 1994). Since many of the human cell lines were originally established from patients with a variety of myeloid and lymphoid leukaemias, detailed characterizations have reported a number of genomic imbalances involving various chromosomes with amplifications, deletions and translocations (see Table 3.2). Despite having various structural and numerical genomic imbalances, these cell lines have served as excellent models to study functions of genes such as SCL, that are important in blood development and haematopoietic differentiation (Green et al. 1991; Leroy-Viard et al. 1994; Shimamoto et al. 1995).

Some of the common human cell lines used in widespread haematopoietic studies include K562, Jurkat, HL60, HPB-ALL, U937 and HEL (Delabesse et al. 2005). Based on their SCL expression, the cell lines can be broadly classified as described below.

1. **SCL expressing cell lines***:* Cell lines exhibiting either normal or inappropriate SCL expression include:

   a) K562: This is a myelogenous cell line in which the predominant cell-type present has been described as a highly undifferentiated granulocytic cell (Lozzio and Lozzio.1977). The cell line can be induced to differentiate down the granulocytic or erythroid lineages and therefore has been used to study the biological role of SCL in cell-proliferation and differentiation in these lineages (Green et al. 1991; Green et al. 1993).

   b) Jurkat: This is a lymphoblastic T-cell line (Schneider et al. 1977). SCL expression is down-regulated in common lymphoid progenitors and is subsequently silenced in

progenitors and terminally differentiated cells of T- and B- lineages. Although Jurkat is a T-cell line, SCL is inappropriately expressed and the activation of SCL has not been linked to any genomic rearrangements at the SCL locus (Leroy-Viard et al. 1994).

2. **SCL non-expressing cell lines:** These include:

c) HL60: This cell line is promyelocytic (Collins et al. 1977) and can be induced to differentiate down the monocytic lineage. SCL expression is turned off in monocyte precursors and, thus, SCL is not expressed in HL60.

d) HPB-ALL is another T-cell line (Morikawa et al. 1978) which, unlike Jurkat, does not exhibit SCL expression.

Identification of chromosomal imbalances in cells using conventional cytogenetic analysis, fluorescence in situ hybridisation (FISH) and comparative genomic hybridisation (CGH) on metaphase chromosomes have a limited resolution of 3 to 5 Mb that is defined by the use of metaphase chromosomes. Development of genomic microarrays, which can measure quantitative changes in genomic copy number, has helped to resolve this problem to a great extent by using cloned DNA segments or PCR-generated sequences instead of metaphase chromosomes as targets for hybridisation (Solinas-Toldo et al. 1997; Pinkel et al. 1998; Albertson et al. 2000; Fiegler et al. 2003). These arrays can detect chromosomal anomalies which are undetectable by FISH, including some involving DNA sequences 30-50 kb in length (Albertson and Pinkel 2003; Mantripragada et al. 2004). A recent study reported the development of a very high resolution array-CGH platform that is highly sensitive and can measure copy-number changes accurately at the resolution of single exons (Dhami et al. 2005). However, even array-CGH methods are not without their limitations, for example array-CGH, at any resolution, will not detect balanced translocations and thus karyotype analysis is required in such cases (Shaffer and Bejjani 2004). Therefore, ideally, the combination of conventional cytogenetic techniques and high resolution array-CGH methods could lead to the identification of the molecular basis of many chromosomal imbalances involved in disease and in cell lines used as experimental models.

3.2. Aims of this chapter

One of the overall aims of the study presented in this thesis was to develop a robust and reproducible array-based platform which would be as sensitive as real-time PCR and could be used in ChIP-chip studies across the SCL locus. To this end, the aims of the work described in this chapter were:

1. To construct and validate sensitive genomic tiling path microarrays containing the human and mouse SCL loci at a resolution of approximately 400-500 bp.
2. To characterize the human haematopoietic cell lines that were selected to be used in ChIP-chip assays across the SCL locus. This would determine the genomic integrity of the SCL locus, which would aid in the interpretation of the data obtained from the ChIP-chip studies presented in subsequent chapters of this thesis.

3.3. Construction of the SCL Genomic Tiling Path Microarray

3.3.1. The array chemistry

As discussed in section 3.1, studies have demonstrated that SCL regulatory elements are located quite distant to SCL, thereby suggesting that it was important to interrogate larger genomic regions to elucidate all key regulatory interactions that could play a role in SCL regulation (Gottgens et al. 2000; Delabesse et al. 2005). Therefore, in order to increase the likelihood that, any novel regulatory elements distant from SCL could be identified, the SCL tiling path arrays were constructed across a larger genomic region than was previously analyzed (90 kb in Delabesse et al. 2005). The construction of such a sensitive array platform for the SCL region was made possible by using the 5'-aminolink array surface chemistry developed at the Sanger Institute, which allows single-strands of DNA derived from double-stranded PCR products to be retained on the surface of the microarray slide (Dhami et al. 2005). This involves incorporation of a 5'-(C6) amino-link modification at the end of one strand of a double-stranded PCR product.

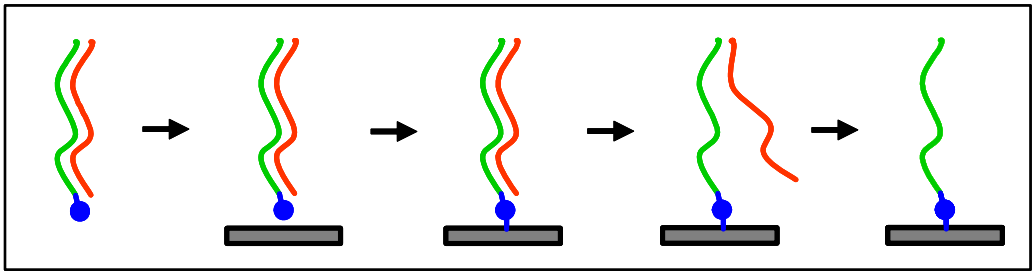


Figure 3.1: Schematic diagram shows the approach adopted to make single-stranded PCR products. Double-stranded PCR products (red/green denote strands) containing a 5'-(C6) amino-linked modification on one strand (blue circle on the green strand) are arrayed onto the surface of the slide (grey bar). Covalent attachment occurs via the 5' amino-link (blue line) and the slide surface. Denaturation of the PCR product renders them single-stranded.

This 5'-aminolink modification drives a covalent attachment between the modified strand and the surface of the slide (Figure 3.1). Upon slide processing, which involves physical and chemical denaturation, the strand attached to the slide is retained, whereas, the unmodified strand is removed. As a result, the single-stranded DNA molecules attached

at one end to the surface of the slide provide an ideal substrate to hybridise with the labeled DNA sample. The sensitivity of this array platform has already been proved by testing its ability to accurately report copy-number changes of individual exons (from 571 bp to 139 bp) in the human genome (Dhami et al. 2005). Thus, utilizing this type of array platform would be ideal to construct a sensitive, quantitative and high resolution genomic tiling path microarray representing the SCL loci in human and mouse.

3.3.2. Defining the genomic regions to be represented on the SCL arrays

The SCL gene in human is flanked upstream by the human SIL and KCY genes and downstream by the MAP17, CYP4A22 and CYP4Z1 genes. Similarly in mouse, SCL is flanked upstream by the mouse SIL and KCY genes and downstream by the MAP17 and Cyp4x1 genes. Figure 3.2 shows a schematic of the human and mouse SCL loci and the genomic regions contained on the SCL tiling arrays.

The construction of the SCL genomic tiling path array was carried out in two phases (1st generation array and 2nd generation or final array) owing to the availability of the finished genome sequence at the human and mouse SCL loci at the time. In the 1st generation array, approximately 193 kb and 199 kb of the genomic regions were chosen in human and mouse respectively to generate contiguous, non-overlapping amplicons. However, upon the release of a subsequent build of the finished human genome sequence (NCBI build 35), it was apparent that regions of the first tiled array were no longer contiguous and the genomic region across the SIL gene had been reorganised in the finished sequence. Thus, it was necessary to fill all of the gaps which resulted in extending the human array to include the 5' end of the KCY gene; the genomic region encompassed on the final human array was approximately 256 kb. In the case of final mouse array, the genomic region was extended further to approximately 207 kb but did not include the 5' end of the KCY gene.

3.3.3. Primer design and PCR amplification

The primers for the human and mouse arrays were designed as described in chapter 2 to generate amplicons that were not more than 600 bp in size; some amplicons contained short interspersed and low complexity repeats. All human and mouse primer pair sequences and their respective genomic coordinates are listed in Appendix 1 and Appendix 2 respectively. Table 3.1 summarizes the characteristics of the human and mouse tiling arrays with respect to their construction.

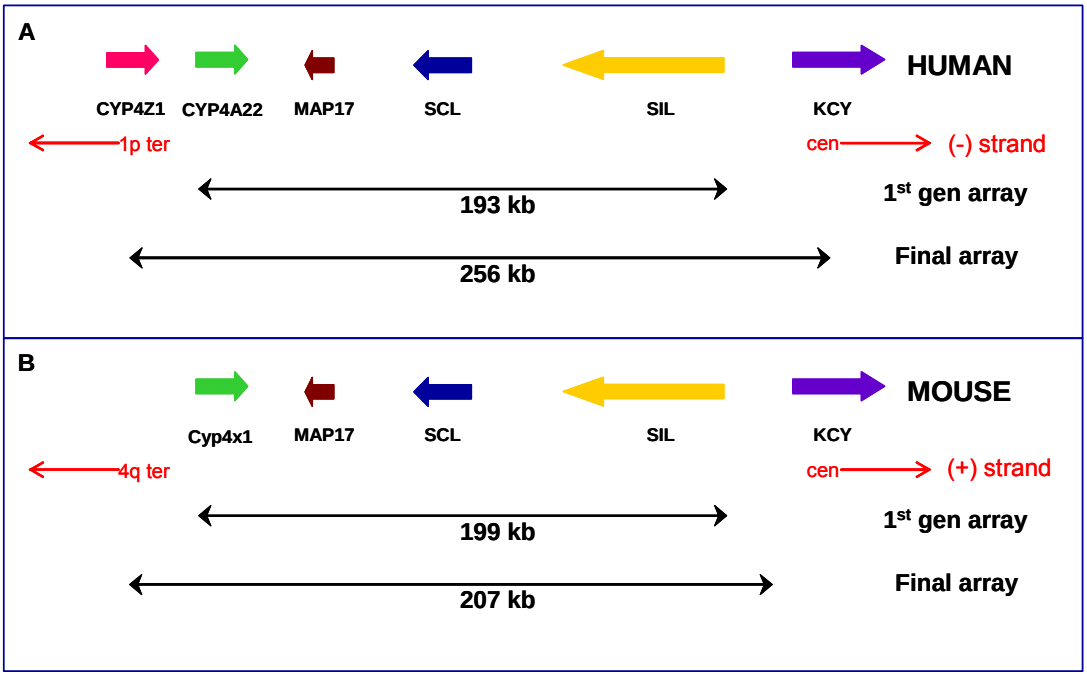


Figure 3.2: A schematic of the genomic regions across the human and mouse SCL loci. The size of the genomic regions included in the 1st generation array and the final array in human and mouse are shown by the black double-headed arrows. The thick, coloured arrows represent the genes. The gene order and the direction of transcription in human and mouse are shown as annotated on the chromosomes 1 and 4 in the respective genomes. The upstream SIL gene and downstream MAP17 gene are highly conserved in both species. The human SCL locus is annotated on the (-) strand of human chromosome 1 whereas the mouse SCL locus is annotated on the (+) strand of mouse chromosome 4. The orientation of the loci with respect to the centromere (cen) and telomere (ter) is shown at the bottom with red arrows.

		HUMAN		MOUSE	
		1st generation array	Final array	1st generation array	Final array
Genomic region included on the array (in kb)		193	256	199	207
Total no. of amplicons designed		171	419	283	530
No. of amplicons passed 1st round PCR		153	407	265	492
No. of amplicons passed 2nd round PCR	Products with correct bands	143	372	250	438
Validation of amplicons by electrophoresis	Missing bands	0	3	1	18
	Multiple bands	4	17	6	21
	Weak bands	6	15	8	12
	Wrong sized products	0	0	0	3
Validation of amplicons by sequencing (amplicons with >90% identity to the expected sequence were considered good matches)	No. of amplicons sequenced	na	180	na	na
	Good matches	na	145	na	na
	Not good matches	na	5*	na	na
	No sequence reads	na	30	na	na
In silico BLAST analysis	Not annotated/homology/repeat	5	5	23	27
Final number of amplicons "passed"		138	367	227	411
Average pass rate (%)		80.7	87.6	80.2	77.5
Number of amplicons in the final array			367		411
Average product size (bp)		420	458	400	443

Table 3.1: Summary of results of the construction of the SCL genomic tiling path array. The tiling path array was constructed in human and mouse in two phases (1st generation and final array). The table lists the sizes of the genomic regions chosen to construct the tiles for human and mouse arrays, the total number of

amplicons designed and the total number of amplicons that finally "passed" after two rounds of PCR amplification. The human and mouse SCL array 2nd round PCR amplicons were electrophoresed on agarose gels and the bands were scored visually. The array elements which were missing, multiple, weak bands or were the wrong-sized products were excluded from the final data sets in all the analyses. In total, 180 randomly selected amplicons from the final human SCL array were sequenced; 145 of the amplicons were confirmed to originate from the expected genomic region; 30 amplicons did not generate successful sequence reads and 5 amplicons were considered not to be a good match to the expected sequences. These 5 (shown with an asterisk) were also products which gave multiple bands as visualised on agarose gels. na = not analysed. The genomic sequences of the human and mouse array elements were mapped against the respective genome sequences using BLASTN (Altschul et al. 1997). Elements showing sequence homology, not mapping to the expected region, not annotated in NCBI build 35 of the human genome or mapping extensively within repeat regions were also excluded.

The array amplicons were generated using a two round PCR amplification procedure (see chapter 2). For the 1st generation array, human and mouse amplicons were successfully amplified in the first round of PCR using total human or total mouse genomic DNA as a template (hereafter called the genomic set). Subsequently, the whole tile for the final array in human and mouse was generated using DNA derived from bacterial PAC and BAC clones as a template in the first round PCR (hereafter called the clone set) in order to increase the success rate of primer pairs. BAC and PAC clones spanning the SCL locus in human (PAC RP1-18D14 and BAC RP11-332M15) and mouse (BACs RP23-242O20, RP23-32K12 and RP-23-246H17) were identified using ENSEMBL Cytoview (http://www.ensembl.org) and used to amplify the appropriate amplicon set.

3.3.4. Controls included on the arrays

A set of PCR products corresponding to human X-linked and autosomal sequences was included as control array elements on the 1st generation SCL tiling path array. The performance of these PCR products was previously tested in competitive hybridisation experiments in order to obtain accurate measurement of copy-number changes (Dhami et al. 2005). Therefore, these PCR products provided an ideal resource to be included as control array elements to test the performance of the SCL array at reporting accurate quantitative measurements of DNA copy number.

Given that the SCL tiling path arrays would be used primarily for ChIP-chip studies (see chapters 4 and 5), sets of sequences were included on the final arrays which had previously been shown to be associated with histone modifications and/or bind sequence-specific transcription factors. The SCL locus had been studied previously using histone modifications (Delabesse et al. 2005) and therefore the SCL tiled amplicons themselves served as a good source of internal control regions for ChIP-chip assays for histone modifications. For ChIP-chip assays for sequence-specific transcription factors, genomic sequences containing known GATA-1 binding sites were

spotted onto the array. In addition to the SCL promoter 1a, which is known to bind GATA-1 in erythroid cells (Bockamp et al. 1995), three array elements (Hb/9BG, Hb/32BG-1, and Hb/32BG-2) containing GATA-1 binding sites at the β-globin locus (Horak et al. 2002) were also included on the array; collectively, these amplicons served as positive controls on the SCL tiling path array – at least for the GATA-1 ChIP-chip experiments described in chapter 4. The primer sequences used to amplify these controls are listed in Appendix 3A.

3.4. Validation of the SCL genomic tiling path microarrays

3.4.1. Assessment of human and mouse SCL array amplicons by electrophoresis

The human and mouse array elements generated after two rounds of PCR amplification were analysed by agarose gel electrophoresis. Each amplicon was assessed visually for the band produced on the gel.

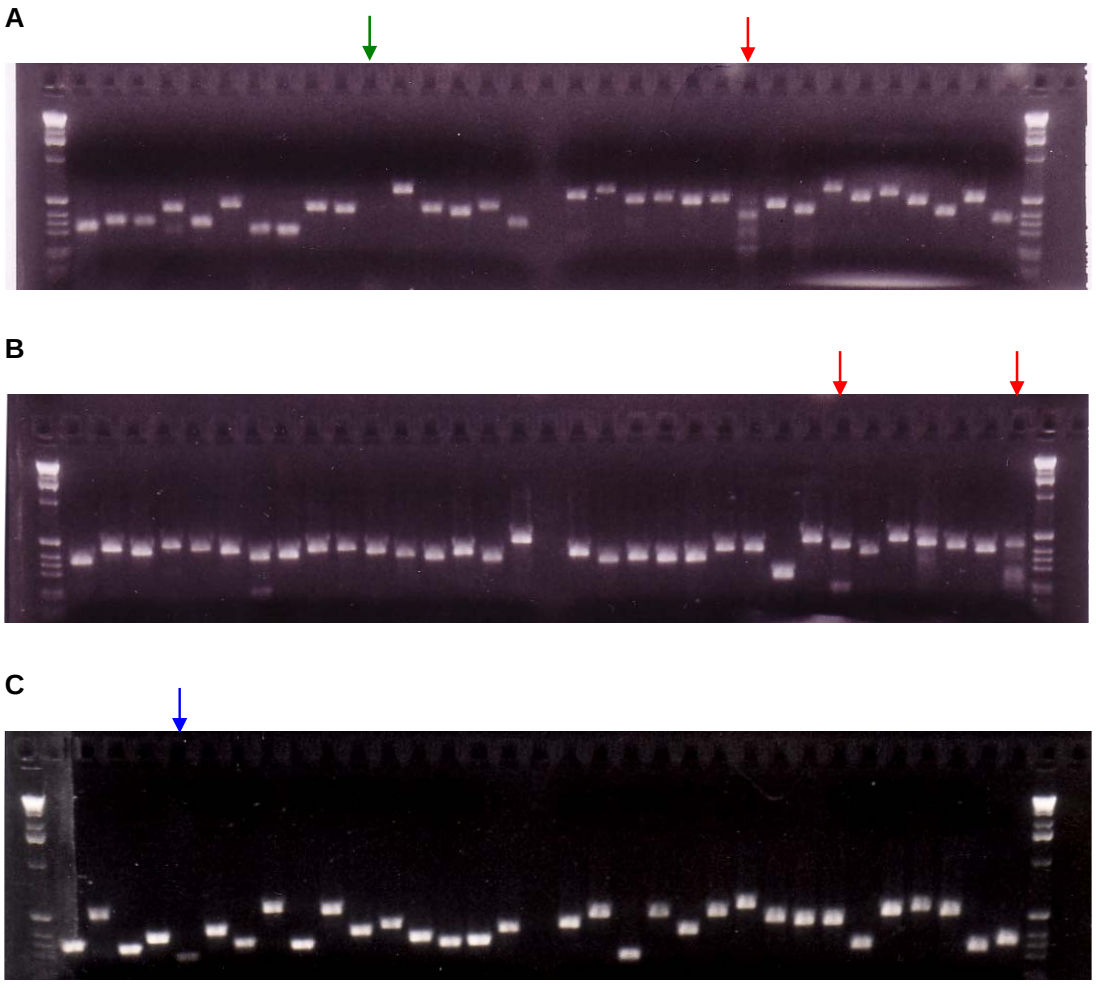


Figure 3.3: Images of agarose gels after electrophoresis of second round PCR amplicons of SCL tiling path array. Panels A and B show PCR amplicons for human and panel C shows PCR amplicons for

mouse. The entire sets of human and mouse array elements were electrophoresed on agarose gels and their bands were scored visually. Lane marked with green arrow highlights a missing band, red arrows highlight multiple bands and the blue arrow highlights a weak band. Products were electrophoresed on 2.5% agarose 1 X TBE gels and visualised with ethidium bromide.

Array elements which did not yield clearly visible amplification products of the correct size were scored as (i) missing bands, (ii) multiple bands, (iii) weak bands or (iv) PCR products of the wrong size (see Figure 3.3, panels A, B and C for examples). The results of this analysis for the human and mouse array amplicons generated for the 1st generation and final arrays are summarized in Table 3.1.

3.4.2. Sequence analysis of the SCL amplicons

Human and mouse SCL array amplicons were subjected to a BLASTN *in silico* sequence analysis (Altschul et al. 1997) to determine whether the amplicons had sequence homology with other sequences in their respective genomes (including repeat sequences). These results are summarised in Table 3.1. Furthermore, in order to determine that the human amplicons were actually derived from the expected genomic regions, a significant proportion of these were verified by sequencing. In total, 180 human amplicons represented on the final array were selected at random and submitted for sequencing. The sequences obtained were compared against the expected sequences; sequences showing more than ninety percent identity with the expected sequences were considered as good matches. The results of the sequence analysis of human amplicons are summarized in Table 3.1.

3.4.3. Validation of the human array elements in genomic array assays

The performance of the human array elements was assessed in genomic microarray assays to determine whether each element reported quantitative measures of genomic copy number. Although the tiling arrays were generated for both human and mouse SCL loci, only the human array elements were tested for their performance in the validation assay; the human SCL arrays serve the basis for the majority of the work described in this thesis. The hybridisation conditions were optimised and standardized for the human array elements so as to obtain accurate copy number measurements. A series of hybridisations were performed on both 1st generation and final SCL tiling arrays using male (XY) versus female (XX) genomic DNA comparisons.

3.4.3.1 Validation of the 1st generation SCL genomic tiling path array

At least six male/female DNA hybridisations were performed on the SCL 1st generation genomic tiling path array. Genomic DNA extracted from normal human male (HRC 575) and normal female (HRC 160) lymphoblastoid cell lines (see chapter 2), was differentially

labeled with Cy5 and Cy3 respectively and hybridised onto the array. Mean Cy5:Cy3 ratios, standard deviations (SDs) and coefficients of variation (CVs) were calculated for each array element spotted in quadruplicate. Figure 3.4 shows the data set obtained from one representative male/female DNA hybridisation.

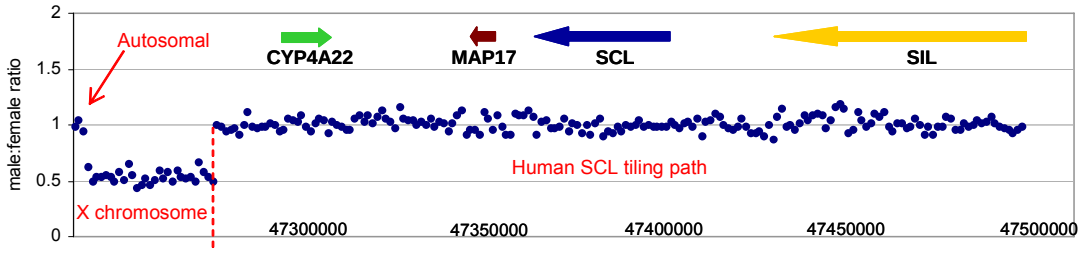


Figure 3.4: A histogram plot of a male versus female hybridisation on the 1st generation human SCL tiling path array. The X chromosome, autosomal and SCL array elements are labelled in red. A representation of the genomic coordinates across the SCL locus is shown along the x-axis, at the bottom of the plotted histogram. The gene order and the direction of transcription are shown at the top of histogram above the SCL array elements. The dotted red line represents a demarcation of the X chromosome array elements from the SCL array elements. The mean male to female ratios for each array element are represented along the y-axis.

The expected male:female ratio for all the human X chromosome array elements is 0.5. The human X chromosome array elements performed consistently and reported accurate copy-number measurements with a mean ratio of 0.56 across the six male/female hybridisations. In the six hybridisations, an average of 96% (28 out of 29) of the X chromosome array elements reported accurate copy number values for analysis and the mean standard deviation (SD) was 0.06 (mean CV = 11%). The three control autosomal array elements (excluding the SCL array elements) reported a mean ratio of 1.02 across the six hybridisations with a mean SD of 0.07 (mean CV = 7%). The SCL array elements collectively reported a mean ratio of 1.01 across the six hybridisations carried out to test their performance and the mean SD was 0.06 (6%). These results demonstrated that the array assays could report accurate copy number changes and that the human SCL array elements were able to report accurate copy number measurements in a genomic array assay. These results were consistent with the assessment of the single-stranded array platform reported previously (Dhami et al. 2005).

3.4.3.2 Validation of the final SCL genomic tiling path array

Three human male/female DNA hybridisations were performed using normal human male and female genomic DNAs. The data analysed for each experiment only included the clone set, as only the clone set of amplicons (i.e., those derived from BAC and PAC templates) included the complete set of array elements representing the human SCL

locus in the final array. In total, 367 human SCL array elements (i.e., the total number of "passed" array elements determined from the analyses summarized in Table 3.1) were included in the final data set. Mean ratios, SDs and CVs were calculated for each array element spotted in duplicate.

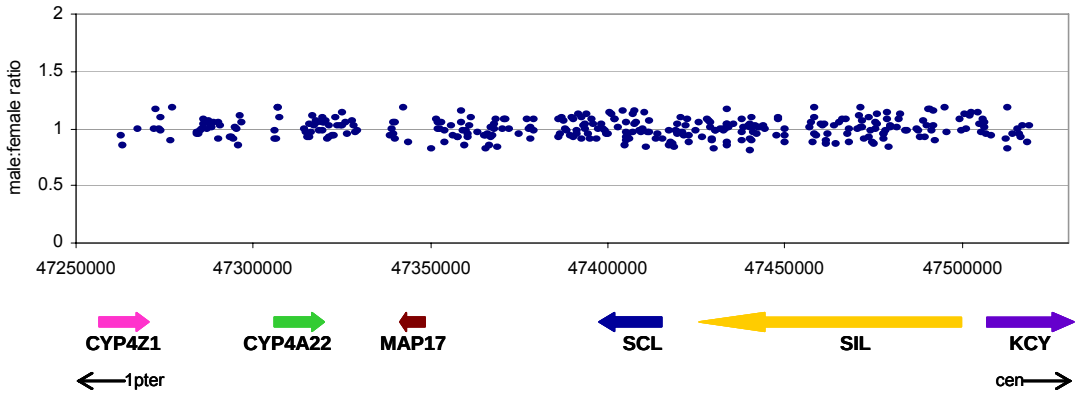


Figure 3.5: A histogram plot of a male versus female hybridisation on the final human SCL tiling path array. The array elements representing only the clone set, i.e. the set which was amplified using DNA derived from bacterial clones, was used for the final analysis as this set included the complete set of array elements representing the human SCL locus. The array elements reporting deviant ratios are not shown on the plot (see text). The array elements are plotted based on their genomic positions and x-axis represents the genomic coordinates along human chromosome 1. The y-axis represents male to female ratio for each array element on the array. The thick coloured arrows represent the gene order and the direction of transcription as shown at the bottom of the histogram plot.

Figure 3.5 shows the final SCL data set of mean ratios for each array element from one representative male/female DNA hybridisation. The mean ratio of the human SCL array elements was 0.99 (the expected ratio was 1) with a mean SD of 0.08 (mean CV = 8%). The final data set reported ratios within 0.2 copy-number units of the theoretical value of 1. Seven array elements (including HSSCL/M93B, HSSCL/M93B3, HSSCL/M93B5, HSSIL/M18B, HSTAL.89, HSTAL.226 and HSTAL.227) consistently reported deviant ratios (either below 0.8 or above 1.2) in the three hybridisation experiments and were excluded from further analyses. Furthermore, no strong correlation was found between the unexpected ratios reported by these elements and their G+C content, length, or repeat content. A lack of correlation between the performance of array elements and sequence features has been reported previously in the detection of copy-number changes at the level of individual exons in the human genome (Dhami et al. 2005). The array elements representing the GATA-1 control amplicons (Hb/BG09, Hb/BG032-1, and Hb/BG032-2) reported mean ratios of 0.91 with mean SD of 0.11 (mean CV = 12%).

Based on the results from all the above described validation experiments, 360 and 411 array elements representing the human and mouse SCL loci respectively were included

in all the array analyses described in this thesis (see Appendix 1 and 2 for the complete lists in human and mouse respectively).

3.5. Characterization of human SCL expressing and non-expressing cell lines

As stated in section 3.1, numerous human and mouse haematopoietic cell lines have been used to understand the regulation of SCL (Bernard et al. 1992; Leroy-Viard et al. 1994; Fordham et al. 1999; Gottgens et al. 1997; Gottgens et al. 2002). Four human haematopoietic cell lines, K562, Jurkat, HL60 and HPB-ALL (also see section 3.1) and two mouse cell lines 416B and E14 ES cell line were selected to perform ChIP-chip experiments across the SCL locus. Between them, these cell lines cover a range of SCL expression patterns from normal (K562, 416B cell lines), inappropriate (Jurkat cell line) to no expression (HL60, HPB-ALL, mouse E14 ES cell lines) and provided excellent experimental systems to understand the transcriptional regulation of SCL. The human cell lines are described in section 3.1 of this chapter. The mouse 416B cell line is a bipotential cell line with majority of the cells identified as undifferentiated blast cells (Dexter et al. 1979). About 95% of the cells possess a diploid chromosome complement, while the remaining cells being true tetraploid or octaploid cells (Dexter et al. 1979). The mouse E14 ES cell line was derived from strain 129/Ola mouse blastocysts and the cells have the potential to differentiate into multiple cell types (Hooper et al. 1987).

Since the four human haematopoietic cell lines were originally derived from leukaemic patients, detailed characterizations using G-banding, M-FISH and conventional CGH analysis on metaphase chromosomes had reported numerous structural and numerical genomic imbalances in these cell lines (see Table 3.2). From all of the reported studies so far, it is evident that different versions of the same cell lines can exist with varied cytogenetic and cellular characteristics. Furthermore, it is widely believed that *in vitro* culturing of cell lines can also introduce additional chromosomal rearrangements, changes in gene expression patterns and changes in DNA methylation (Drexler et al. 2000). To establish the level of SCL expression, extent of genomic imbalances and any chromosomal rearrangements that may affect the SCL locus, the four human cell lines were fully evaluated using four different approaches as detailed below.

Cell-line	Cell-type	SCL expression	Cytogenetic characteristics	Chromosomes involved in genomic imbalances	Previously unreported
K562[a,b]	myelogenous	Yes[f]	Diploid (46, XX) Hyperdiploid (50-52, XX) Near-triploid (69-73, XX) or (62-69, XX) Near-tetraploid (90-96, XX)	1, 3, 5, 6, 7, 9, 10, 13, 14, 17, 18, 20, 21, 22	4, 8, 12, 15, 16
Jurkat[c,d]	T-cell	Yes[f]	Diploid (45-48, XY) Hypotetraploid	2, 3, 4, 5, 8, 9, 18, 20, X, Y	10, 15, 22
HL60[d]	Promyelocytic	No[f]	Diploid (44-47, XX)	5, 8, 9, 11, 13, 14, 15, 18	4, 6, 7, 10, 16,17, 22
HPB-ALL[e]	T-cell	No[f]	Diploid (46, XY) Tetraploid (94, XY)	1, 2, 3, 5, 14, 16, 21	7, 8, 10, 17

Table 3.2: Cytogenetic characteristics and SCL expression patterns in the cell lines: K562, Jurkat, HL60 and HPB-ALL. Note: The cytogenetic and expression data reported in this table has been compiled from other published studies: [a] Gribble et al. 2000; [b] Naumann et al. 2001; [c] Snow et al. 1987; [d] Cottier et al. 2004; [e] MacLeod et al. 2003; [f] Delabesse et al. 2005. The last column (previously unreported) shows the results obtained with 1 Mb array-CGH performed in the study presented for this thesis.

3.5.1. Real-time PCR analysis of the SCL expression in the four cell lines

To investigate the levels of SCL expression in the four cell lines used in this study, quantitative real-time PCR (RT-PCR) using SYBR Green was performed using a 7000 sequence detection system (Applied Biosystems). The SIL gene (upstream of the SCL gene) is ubiquitously expressed in all four cell lines and thus represented a positive control in the experiment. Primer pairs mapping to the 3'-end of the SCL (within exon 6), SIL and β-actin transcripts were designed and are listed in Appendix 3B. The expression level of β-actin was used as an endogenous reference gene against which SCL and SIL expression levels were normalised.

Figure 3.6 illustrates the relative expression levels of the SCL and SIL genes in K562, HL60, HPB-ALL, and Jurkat. The reported results were as expected with the previously published results; SCL gene was expressed in K562 and Jurkat but was not expressed in HL60 and HPB-ALL. As expected, SIL expression was observed in all of the cell lines.

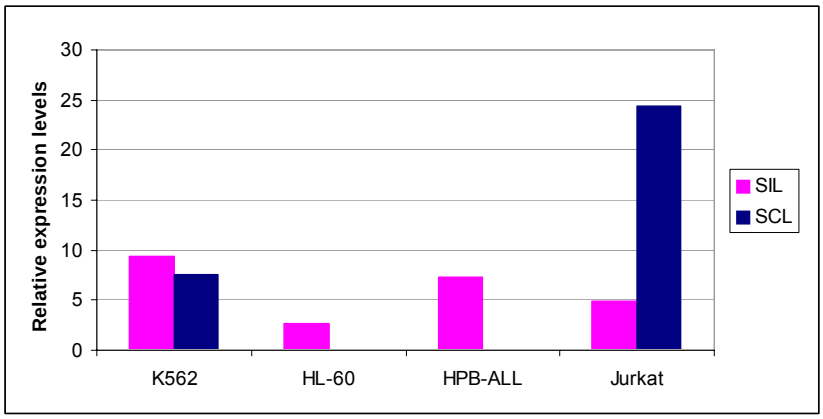


Figure 3.6: Analysis of SCL expression by real-time PCR. Relative levels of SCL and SIL expression in the four cell lines as measured by quantitative real-time PCR. SIL gene is ubiquitously expressed in all cell lines and served as a positive control. The expression level of β-actin was used as an endogenous reference gene against which SCL and SIL expression levels were normalised. The cell lines are represented at the bottom of the plot and the y-axis represents the relative levels of SIL and SCL transcripts.

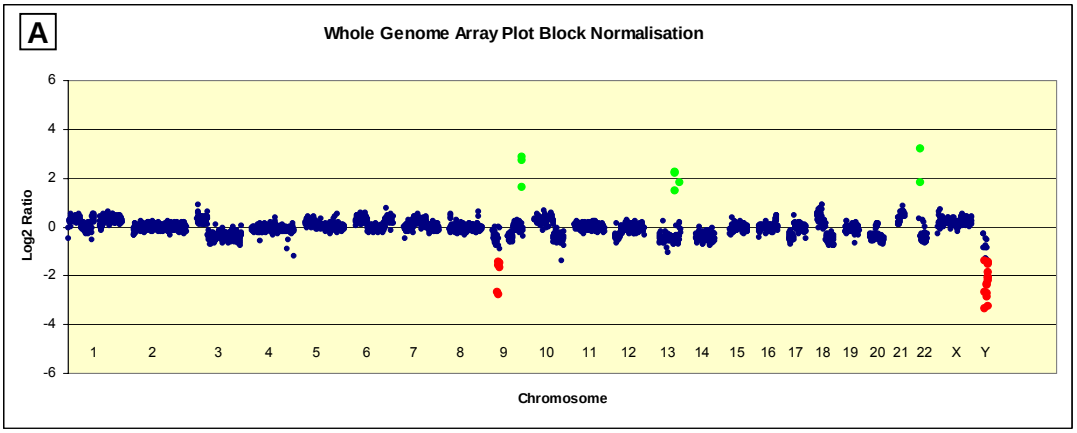
3.5.2. 1 Mb genomic microarray-CGH analysis of K562, Jurkat, HL60 and HPB-ALL

Array CGH hybridisations were performed on a genomic microarray that contains 3500 large insert BAC and PAC clones containing sequences from across the human genome (Fiegler et al. 2003). The purpose of the array-CGH analysis was to identify genomic imbalances in the four cell lines, in addition to the ones previously reported (Table 3.2), and to detect any rearrangements affecting the SCL locus (on chromosome 1) in particular. Since the resolution of the genomic array is one clone per 1 Mb of the human genome, none of the clones represented on the array span the genomic region containing the SCL locus. Two clones, RP11-8J9 at 46.9 Mb and RP11-330M19 at 48.0 Mb positions, map to either sides of the genomic region containing the SCL locus on chromosome 1. A male pool DNA sample that was used as the reference DNA in the array hybridisations was derived from combining peripheral blood DNA from 20 normal males.

The arrays were analysed using a Microsoft Excel spreadsheet specially written for the 1 Mb array. The ratios (median intensity – background) of Cy3:Cy5 intensities for each spot were normalised to the median raw ratio of all the autosomal spots on the array (global normalization). The normalised ratios for each spot in a sub-grid (block) of the array were further normalised to the median ratio of all the spots in that block (block normalization). After the global and block normalizations, the mean ratios, standard deviations and CVs were calculated for duplicate clones. Duplicates reporting difference of more than 20% were excluded from the data set. The mean linear ratios of the accepted clones were converted into $\log_2$ ratios and plotted on a histogram against the

genomic position of the clone in the genome according to build “35” NCBI. In order to detect genomic copy number changes such as gains and deletions, a “significant cut-off” value calculated for individual hybridisations was used. The significant cut-off value was set at five times the calculated standard deviation of 95% ratios in each experiment. This means that any clone reporting a ratio greater or less five times the standard deviation of the majority of the clones was considered significant to indicate copy number gain or deletion respectively. Known polymorphisms in the human genome were also flagged. The results obtained in all the four cell lines are described below.

1. The K562 cell line is known to exhibit multiple/complex genomic imbalances involving various chromosomes with copy number gains (including high-level amplifications), deletions and translocations (Table 3.2). Array-CGH analysis reported imbalances in almost all of the chromosomes (Figure 3.7, A). Copy number gains (including some amplifications) and losses (deletions) were observed in chromosomes 1, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 18, 20, 21 and 22. From this, it was apparent that the K562 cell line used in this study displayed additional genomic rearrangements than those previously reported. Genomic imbalances had not been previously reported in chromosomes 4, 8, 12, 15 and 16 (Table 3.2).



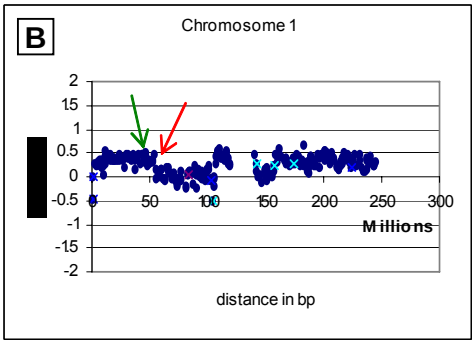
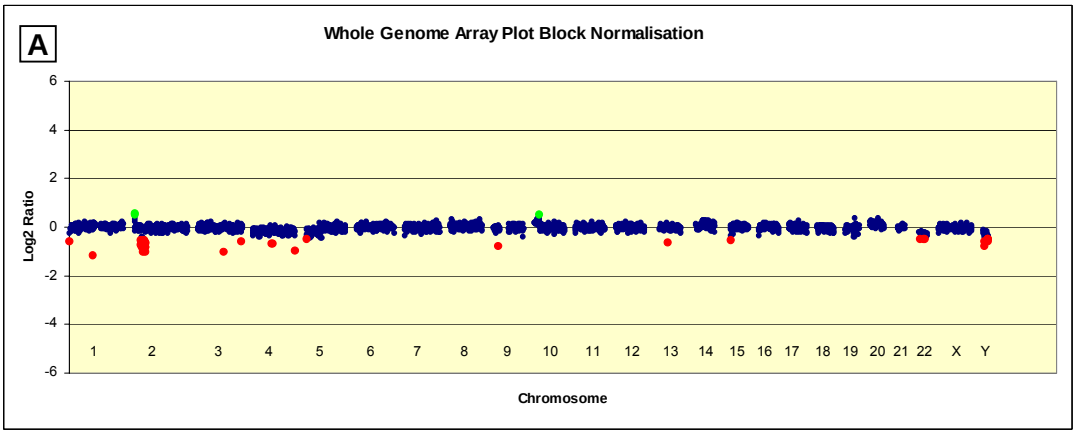


Figure 3.7: Array-CGH analysis of K562 cell line using the 1 Mb genomic array. Panels A and B show the genome plot and the chromosome 1 plot respectively. The red and green dots in the panel A represent some of the clones reporting deletions and copy number gains respectively. The green arrow in the chromosome 1 plot points to the position of the SCL gene and the red arrow points to the chromosomal breakpoint on 1p. The distal arm of 1p containing the SCL locus shows a copy number gain from 12.4 Mb to 53.8 Mb. The clones excluded (based on analysis criteria) in panel B are shown with Xs.

It had been reported previously using FISH analysis that a translocation breakpoint on chromosome 1p (p32-pter) in K562 did not structurally disrupt the SCL locus (Gribble et al. 2000). The array-CGH analysis reported a copy number gain of chromosome 1p from 12.4 Mb to 53.8 Mb which has not been previously reported by conventional CGH analysis on metaphases. This genomic region on the distal arm of chromosome 1p contained the SCL locus (Figure 3.7, B), which did not appear to be structurally disrupted. Clone RP11-243A18 at 53.8 Mb reported a $\log_2$ ratio of 0.44 whereas the clone RP5-1070D5 at 55.6 Mb reported a $\log_2$ ratio of 0.14, suggesting that this may represent a chromosomal breakpoint associated with the gained region. A gain of 1q had previously been reported (Gribble et al. 2000) which was confirmed here by array-CGH.



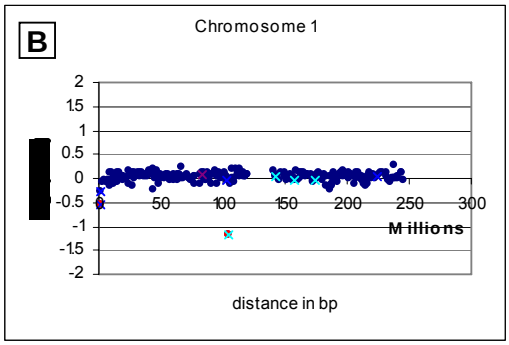


Figure 3.8: Array-CGH analysis of Jurkat using the 1 Mb genomic array. Panels A and B show the genome plot and the chromosome 1 plot respectively. The red and green dots in the panel A represent clones reporting deletions and copy number gains respectively. The chromosomes are listed at the bottom of panel A. No genomic imbalances were reported on chromosome 1p containing the SCL locus. The clones excluded (based on analysis criteria) in panel B are shown with Xs.

2. Array-CGH analysis of the Jurkat cell line determined that various chromosomes were involved in genomic imbalances including some that had previously not been reported (Table 3.2; Figure 3.8, A). A number of clones reported deletions and copy number gains on chromosome 2, deletions on chromosomes 3, 4, 5, 9, 15, 22, and Y, and copy number gains on chromosomes 8, 10, and 21. However, no copy number gains or deletions were seen on chromosome 1 (Figure 3.8, B).

3. Array-CGH analysis of the cell line HL60 did not only detect all the genomic imbalances that had been reported previously, but also detected additional copy number gains and deletions in various chromosomes (Table 3.2; Figure 3.9, A). A number of clones reported significant deletions on regions of chromosomes 4, 5, 9, 10, 14, and 17. Several clones reported copy number gains of regions of chromosomes 6, 7, 8, 13, 18 and 22 including a high-level amplification on the long arm of chromosome 8. Several clones reported significant copy number gains on the short arm of chromosome 16 whereas a number of clones representing the long arm of chromosome 16 reported significant deletions. Clones representing chromosome 1 did not report any significant ratios indicative of copy number gains and/or deletions (Figure 3.9, B).

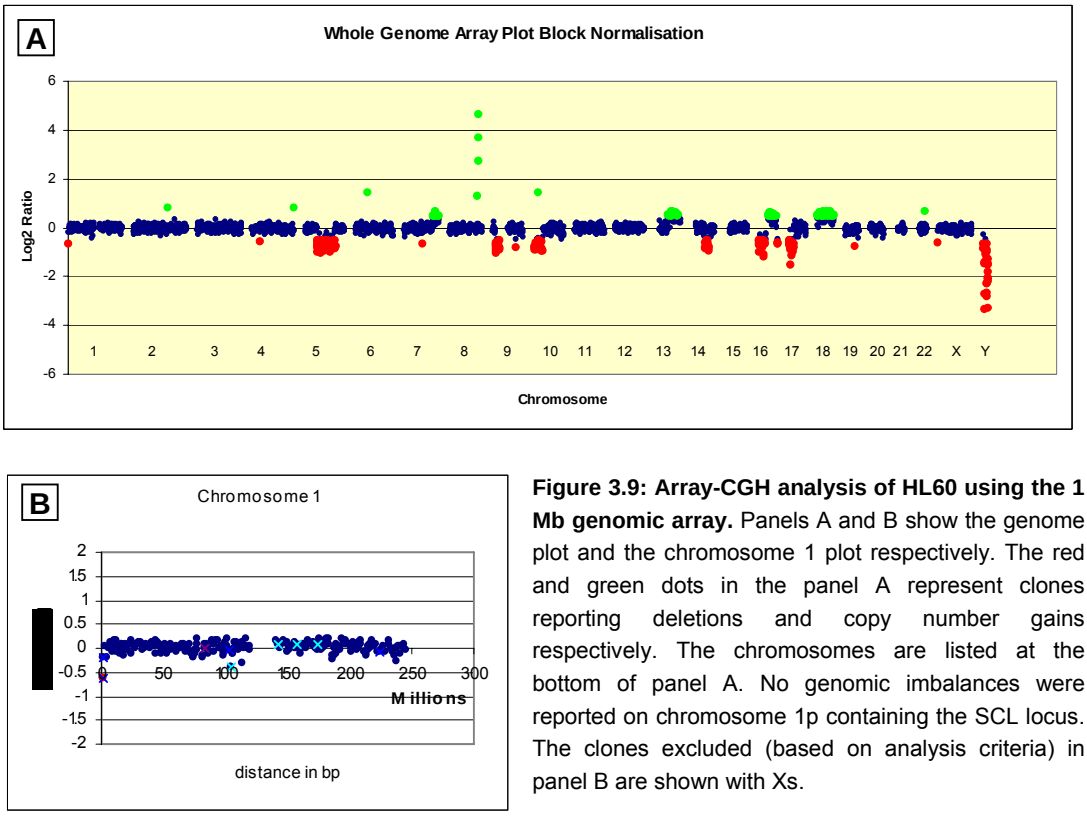
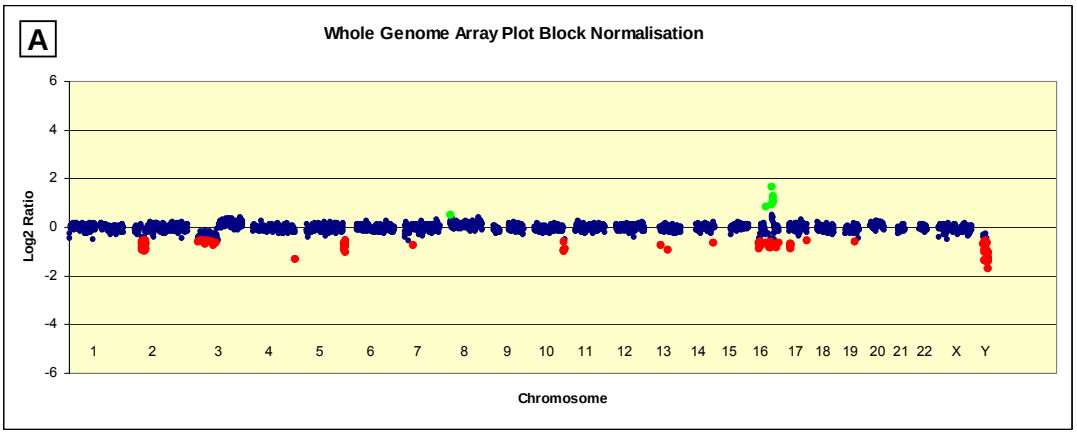


Figure 3.9: Array-CGH analysis of HL60 using the 1 Mb genomic array. Panels A and B show the genome plot and the chromosome 1 plot respectively. The red and green dots in the panel A represent clones reporting deletions and copy number gains respectively. The chromosomes are listed at the bottom of panel A. No genomic imbalances were reported on chromosome 1p containing the SCL locus. The clones excluded (based on analysis criteria) in panel B are shown with Xs.

4. Several chromosomes including chromosome 1 had been reported to be involved in chromosomal translocations and genomic copy number changes in the cell line HPB-ALL (Table 3.2). Array-CGH analysis of HPB-ALL identified additional chromosomal imbalances (Table 3.2; Figure 3.10, A). A number of clones on the chromosomes 2, 3, 5, 7, 10, and 17 reported low ratios indicating deletions whereas one clone on chromosome 8 reported a high ratio, indicating a gain. Several clones on chromosome 16 reported

high levels of amplifications while a number of other clones reported deletions. A clone on chromosome 4 reporting a low ratio was known to be polymorphic.



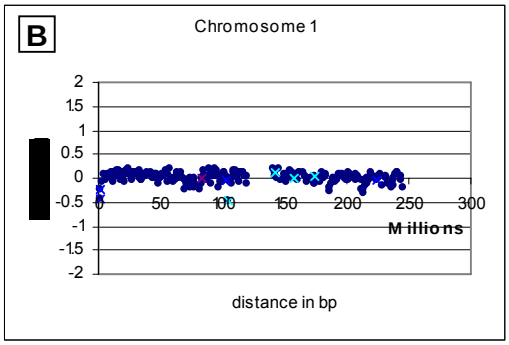


Figure 3.10: Array-CGH analysis of HPB-ALL using the 1 Mb genomic array. A and B represent the genome plot and chromosome 1 plot respectively. The red and green dots in the panel A represent clones reporting deletions and copy number gains respectively. The chromosomes are listed at the bottom of panel A. No genomic imbalances were reported on chromosome 1p containing the SCL locus. The clones excluded (based on analysis criteria) in panel B are shown with Xs.

The involvement of chromosome 1 in a genomic translocation had been reported previously in HPB-ALL (MacLeod et al. 2003). Any chromosomal breakpoint was not obvious with the array-CGH analysis (Figure 3.10, B). It is possible that the translocation could be a balanced rearrangement which would not be detected by array-CGH. However, none of the clones representing chromosome 1 reported any significant deviation from the normal ratios suggesting no obvious copy number gains or deletions were present on chromosome 1.

3.5.3. Array-CGH analysis of K562, Jurkat, HL60 and HPB-ALL using the SCL genomic tiling path array

To investigate the SCL genomic region at a higher resolution (approximately 400-500 bp), array-CGH experiments were performed using the final SCL genomic tiling path array. Hybridisation experiments with DNA samples extracted from the four cell lines against a reference DNA on the tiling path array would report any small gains or losses in the SCL genomic region which could not be detected by array-CGH analysis using the 1 Mb BAC array.

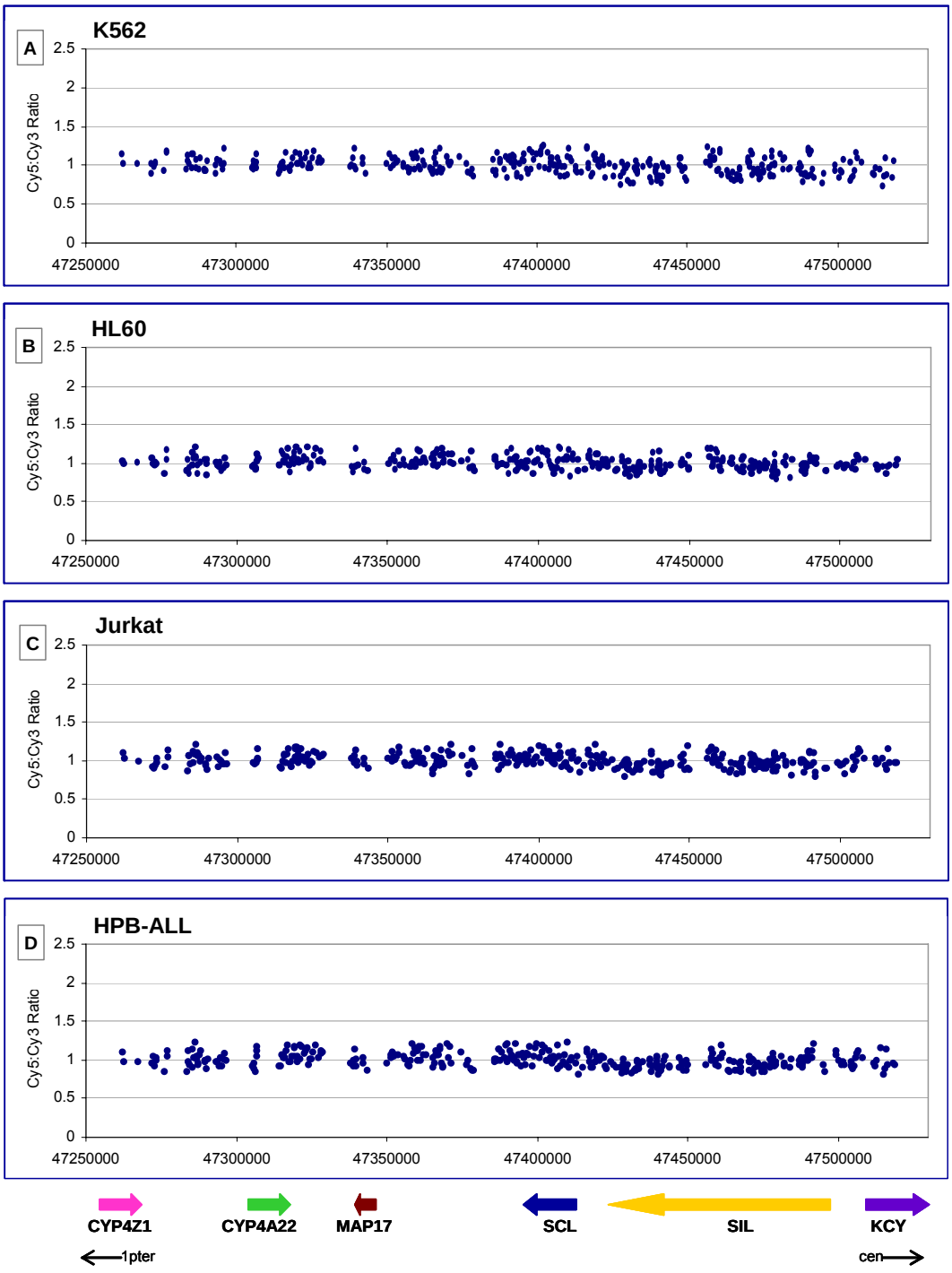


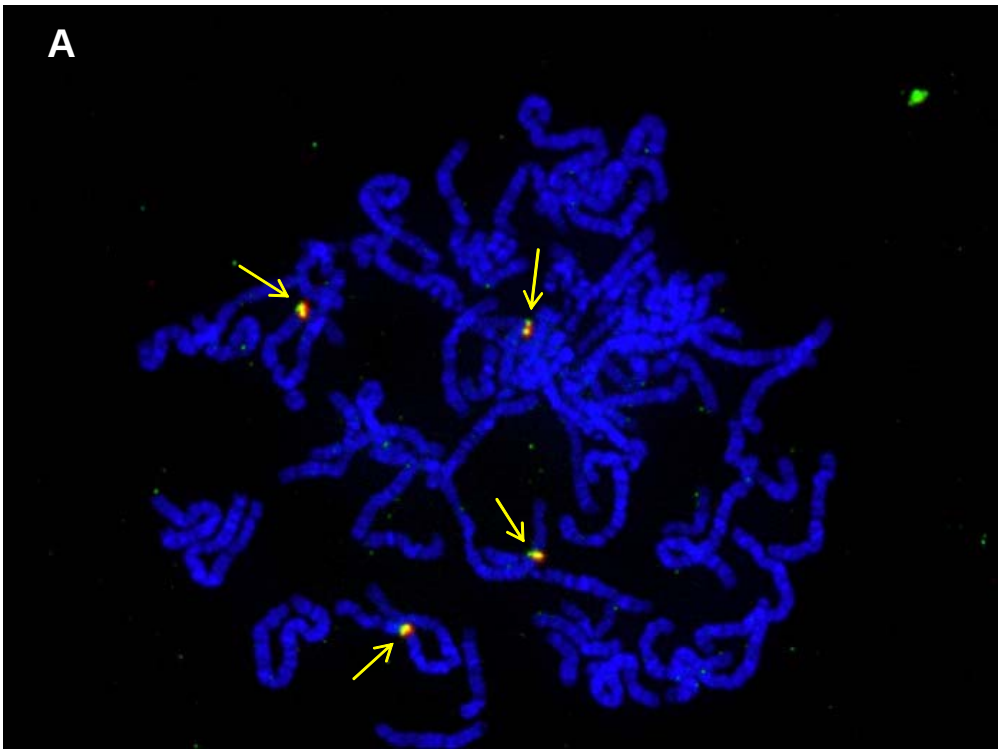
Figure 3.11: Array-CGH analysis of the four cell lines using the high resolution SCL genomic tiling path array. Panels A-D show histogram plots of K562, HL60, Jurkat and HPB-ALL respectively. The Cy5:Cy3 ratios are plotted on the y-axes against genomic position of the array elements along chromosome 1 on the x-axes. Mean SDs calculated for each data set were 0.10 for the K562, 0.08 for the HL60, 0.08 for the Jurkat and 0.09 for HPB-ALL. The thick coloured arrows represent the gene order and the direction of transcription as shown at the bottom of the figure. The orientation of the locus with respect to the centromere (cen) and telomere (ter) on human chromosome 1 is shown by the thin black arrows at the bottom of the figure.

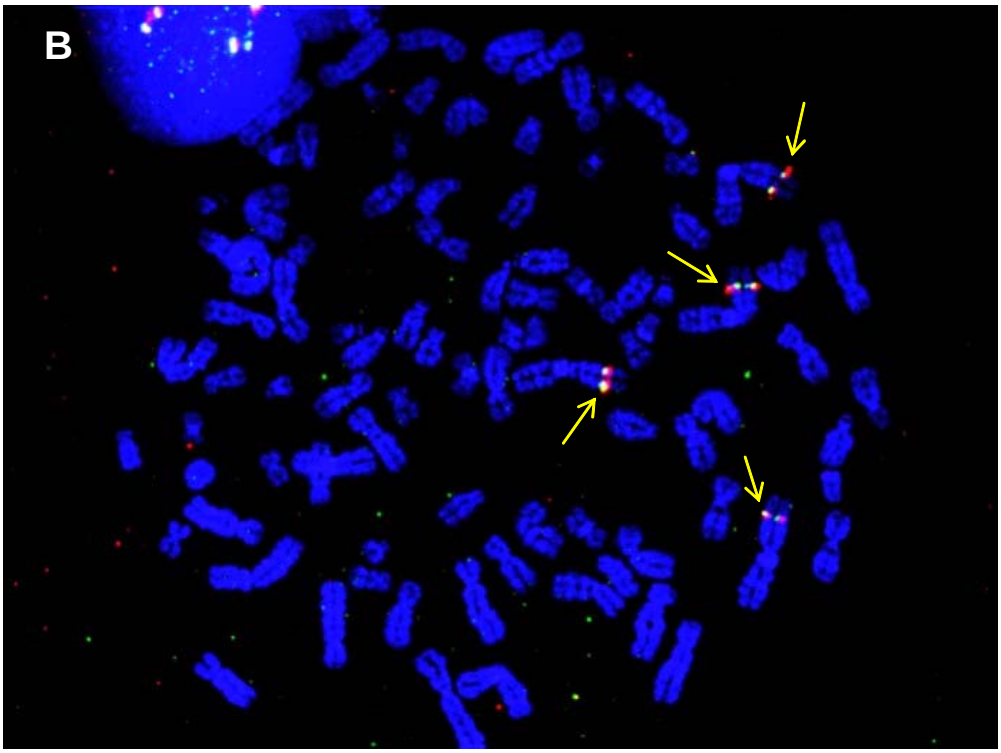
Figure 3.11 shows the histogram plots for high resolution array-CGH analysis across the SCL region for all the four cell lines. The expected Cy5:Cy3 ratio for all the array elements on the SCL tiling path array was 1, as all the array elements correspond to autosomal sequences. No changes in genomic copy number which affected the SCL region were reported by the SCL array elements in any of the four cell lines. This suggested that no genomic imbalances such as copy number gains or deletions lay within the 250 kb genomic region containing the SCL gene which could structurally disrupt the SCL locus, thus affecting its regulation. Although it was known that the cell line K562 had a copy number gain of chromosome 1 encompassing the SCL locus (see Figure 3.7, B), the high resolution SCL array did not detect it, since all of the array elements on the array were affected by this copy number gain in the same way. In other words, copy number differences can only be reported if some of the elements report the normal (modal) values.

3.5.4. FISH analysis of K562, Jurkat, HL60 and HPB-ALL with two DNA clones spanning the human SCL locus

To investigate the possibility of a balanced rearrangement affecting the SCL locus, FISH analysis was performed using two overlapping bacterial clones spanning the SCL locus in human. The clones RP1-18D14 and RP11-332M15 were used (section 3.3.3) for this analysis since these clones contained the genomic region represented on the SCL genomic tile path array. The DNA from both clones was differentially labelled by nick-translation, mixed and hybridised to metaphase chromosome spreads from each cell line (see chapter 2 for methods). After detection, slides were scanned and at least ten metaphase spreads were analysed per cell line. Any translocations affecting the SCL locus would be visible by the appearance of FISH signals for either clone which did not map to the same chromosomal location. The FISH analysis of K562, Jurkat, HL60 and HPB-ALL is shown in Figure 3.12. The SCL locus was not structural disrupted in any of the four cell lines. All four cell lines were mostly polyploid as described in the legend of Figure 3.12.

.


A


B

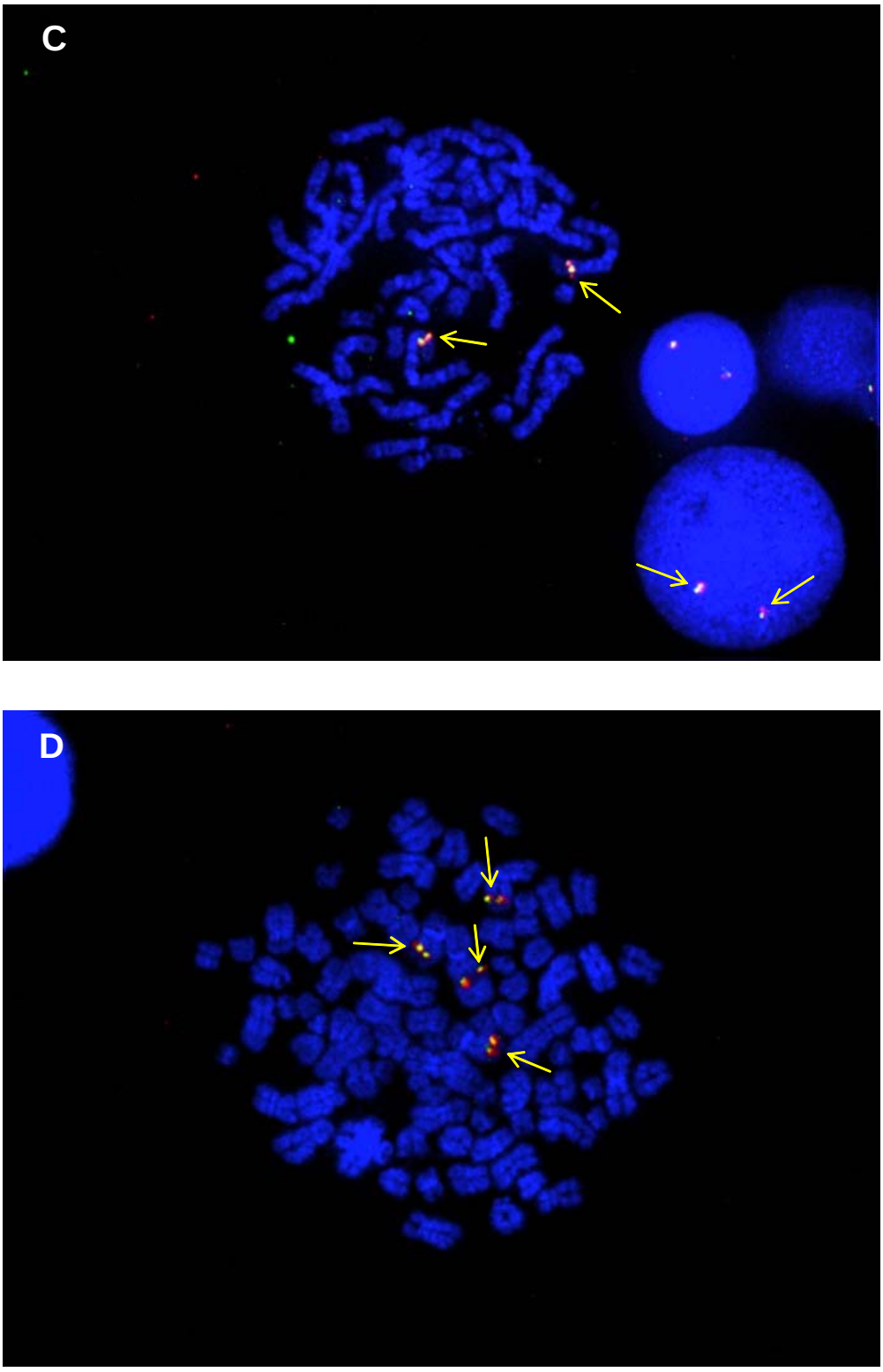


Figure 3.12: FISH images of metaphase spreads from the four human cell lines. A: K562, B: Jurkat, C: HL60 and D: HPB-ALL. FISH analysis was performed by hybridising two clones (RP1-18D14 and RP11-332M15) spanning the SCL locus to metaphase chromosomes prepared from each of the cell lines. Yellow arrows in each image point to the co-localized FISH signals on chromosome 1p. The red signals represent clone RP1-18D14 and the green signals represent clone RP11-332M15. The K562 cell line was observed to

be near-triploid with chromosome number ranging from 62 to 75. In K562, the majority of the cells had four bright co-localized signals overlapping each other on the short arm of chromosome 1 (panel A), while other cells showed two bright signals. The presence of four co-localized signals indicated that the genomic region spanned by the two clones i.e. 1p33 was present in four copies and was not structurally disrupted anywhere in the genome of K562 cell line. 1 Mb array-CGH analysis reported amplification of 1p32.3-1p36.33 region containing the SCL locus (at 1p33) which was also confirmed by FISH analysis. The presence of four co-localized FISH signals in a near-triploid cell line suggested that an extra copy of the chromosome 1p33 region existed in K562. At least, two normal chromosomes 1 could be identified in the majority of the cells; the others were possibly marker chromosomes with genomic material from the short arm of chromosome 1. Similarly, in all the other cell lines, Jurkat, HL60, and HPB-ALL, the signals were always co-localized indicating no disruption in the SCL region. The majority of the cells scanned in Jurkat had four co-localized signals and the cell line was observed to be tetraploid with chromosome number 90 to 92 (panel B). HL60, on the other hand, was observed to be diploid with two co-localized signals per cell (panel C). Four co-localized FISH signals were seen in majority of the cells scanned in HPB-ALL with chromosome number ranging from 80 to 85 making it hypo-tetraploid (panel D).

3.6. Discussion

3.6.1. SCL genomic tiling path array: sensitivity and resolution issues

Genomic microarrays are increasingly being used to unravel and understand aspects of genome biology. Depending on how genomic arrays are used in various applications in various species, the requirements for sensitivity can vary. For example, genomic arrays used for array-CGH are required to report accurate quantitative measurements of single copy number changes; this is even more challenging within highly complex genomes, such as the human genome. The quantification and sensitivity requirements for array-CGH were used as benchmarks to develop the highly sensitive SCL genomic arrays described in this chapter. This work was preceded by the development of an array-CGH platform which had the ability to detect single copy number change at the resolution of a single exon (139-571 bp) in the human genome (Dhami et al. 2005). The development of this “exon” array-CGH platform was shown to be two orders of magnitude more sensitive than all other array-CGH technologies in terms of its ability to detect copy number changes at high resolution in the human genome. Thus, the single-stranded array chemistry which was used to develop the “exon” array-CGH platform was used to construct the SCL arrays presented in this chapter. As a result, the SCL array platform has been shown to be sensitive enough to accurately measure genomic copy number changes within the human genome.

The SCL tiling array platform would be ideal to use in any array-based assay in which accurate quantitative measurements of DNA is required, including mapping DNA-protein interactions at the SCL locus. In mammalian systems, many published studies have made use of genomic tiling arrays to map DNA-protein interactions across genomic region or even genome-wide scale (Horak et al. 2003; Cawley et al. 2004; Martone et al.

2003; Euskirchen et al. 2004; Kim et al. 2005). However, none of the published ChIP-chip studies have tested the ability of array platforms to perform accurately within the dynamic range required for array-CGH. Yet, the ability to report accurate quantitative measurements is highly desirable for ChIP-chip platforms as well - using an array platform which could perform to this level would prove to be a great asset in measuring subtle ChIP enrichments or low-affinity binding sites of transcription factors. Subsequent work in this thesis will demonstrate how this level of sensitivity is an absolute requirement for ChIP-chip studies, in order to detect the full range of regulatory interactions at the SCL locus.

In the published studies mentioned above, genomic tiling arrays were constructed by using either tiled PCR products (Horak et al. 2002; Schübeler et al. 2004; Martone et al. 2003; Euskirchen et al. 2004) or tiled oligonucleotides (Cawley et al. 2004; Pokholok et al. 2005; Kim et al. 2005). The present study used PCR products to construct the SCL arrays. Although using long or short oligonucleotides appears to be an attractive possibility to increase the resolution of mapping DNA-protein interactions, it is notable that oligonucleotides for use in ChIP-chip studies have two potential limitations. Firstly, it is not only the array elements which confer the resolution of the platform, but also the size of the sheared chromatin fragments arising from ChIP which are used in array hybridisations. The average size of these sheared DNA samples is usually in the range of 300-1000 bp, thus lowering the effective array resolution substantially when hybridised onto oligonucleotides. Secondly, analysis of oligonucleotide arrays relies on averaging measurements taken from multiple array elements, which effectively decreases their resolution to approximately 500 bp. Given these issues, it has not yet been empirically determined which type of array platform is better for ChIP-chip studies given that a direct and comprehensive comparison of PCR-spotted arrays or oligonucleotide arrays for ChIP-chip experiments has not yet been published. Furthermore, the effective resolution of oligonucleotide ChIP-chip arrays described above is similar to that for the SCL arrays described in this chapter, suggesting that oligonucleotide arrays would not have conferred a resolution advantage for the work reported later in this thesis.

It should also be noted that the validation experiments of the human SCL array described in this chapter, represent an initial characterization of this platform, but are by no means an exhaustive one. Additional work on the validation of the array for its use in ChIP-chip assays is described in chapter 4 of this thesis. Furthermore, although an SCL tiling path array was also constructed across the mouse SCL locus, the mouse array elements were not assessed for their performance in array-CGH experiments in the same manner as the human array elements. However, the array elements representing

the mouse SCL locus were assessed for their performance in ChIP-chip based assays for histone H3 acetylation as described in chapter 4.

3.6.2. Characterizing human haematopoietic cell lines

All of the human haematopoietic cell lines selected for the study presented for this thesis were derived from leukaemic patients; previous characterizations have reported the existence of numerous genomic imbalances in these cell lines (Gribble et al. 2000; Naumann et al. 2001; Snow et al. 1987; Cottier et al. 2004; Macleod et al. 2003). Performing array CGH analysis using the 1 Mb genomic microarray (Fiegler et al. 2003), identified the previously reported genomic imbalances along with the detection of additional copy number gains and deletions in various chromosomes. The detection of previously unidentified genomic imbalances in these cell lines using array-CGH suggests that (i) these imbalances were undetectable with previous low resolution characterization of these cell lines using CGH on metaphase chromosomes, (ii) the cell lines used in the present study represent outgrowth of cryptic sub-clones already present when these cell lines were originally established (Drexler et al. 2003), or (iii) additional genomic imbalances have occurred during the many years since these cell lines were originally established.

Data presented in this chapter show that the SCL region represented on the human tiling path array is structurally intact in the four human cell lines used for the work of this thesis. Of these four cell lines, only K562 showed chromosomal imbalances on chromosome 1. In a previously published study, a translocation breakpoint was detected on chromosome 1p in K562 which did not structurally affect the SCL gene (Gribble et al. 2000). FISH analysis described here using two overlapping bacterial clones confirmed that the SCL locus was not involved in this translocation. Furthermore, array CGH performed at a resolution of 1 Mb in K562 reported a copy number gain of the distal arm of 1p which encompassed the entirety of SCL, suggesting that the SCL locus was structurally intact. Collectively, these results suggest that SCL is likely to be regulated by its own regulatory elements in K562 cell line, although the effect of very distant chromosome 1p genomic imbalances (i.e., further away than the region contained on the SCL tiling array) on SCL expression is not known.

It is possible that genomic imbalances in other parts of the human genome could impinge on the expression of SCL in any of the cell lines studied here. It is known that transcription factors including GATA-1, GATA-2, Elf-1 and Fli-1 play important roles in the regulation of SCL at various stages of development (Aplan et al. 1990; Lecointe et al. 1994; Gottgens et al. 2002; Gottgens et al. 2004). In K562, the genomic regions

containing GATA-2 and Elf-1 on chromosomes 3 and 13 respectively show single copy deletions which could affect the expression of these transcription factors (Figure 3.7, A). Whether these genomic imbalances affect SCL regulation in K562, is not known.

An interesting feature of the human leukaemic cell lines studied here is that they are mostly polyploid in addition to the existence of genomic imbalances of discrete regions of various chromosomes. For instance, four co-localized FISH signals for the SCL locus were seen in most of the cells analyzed in K562 and Jurkat, both of which exhibit SCL expression. This suggests that at least four copies of the SCL gene are present in these cell lines. SCL expression in Jurkat has previously been found to be mono-allelic (Leroy-Viard et al. 1994), suggesting that all of the other SCL alleles in Jurkat are not expressed. Similarly, it is not known whether SCL is transcribed from all of the copies in K562. Thus, the fact that both these cell lines carry extra copies of the gene should be taken into consideration when analysing ChIP-chip experiments performed in these cell lines. It would be difficult to determine, using ChIP-chip experiments, whether the DNA-protein binding interactions are specific to one locus or represent a composite profile of these interactions at all of the SCL loci in each cell.

Despite the presence of all the genomic imbalances, these haematopoietic cell lines provide excellent cell line based experimental systems to decipher the key regulatory interactions involved in SCL regulation (Aplan et al. 1990; Green et al. 1991; Leroy-Viard et al. 1994; Bernard et al. 1992; Cheng et al. 1993; Delabesse et al. 2005) and are routinely used for such studies. The information gained about regulation of SCL from these is indeed invaluable and the major advantage of using cell lines is the virtually unlimited supply of biological material. However, it has been suggested that some additional cellular characteristics are acquired during establishment of cell lines or during extended *in vitro* culturing (Drexler et al. 2003), thereby, highlighting the pitfalls of using cell lines instead of primary cells or tissues in deducing biological information for living organisms. For ChIP-chip assays, the number of cells required to perform a single assay are quite large – in the range of 10 to 20 million cells (Chapter 4 and 5). Therefore, to obtain such large number of primary cells to carry out an extensive study to elucidate DNA-protein interactions across the SCL locus is not presently feasible. Moreover, under optimal cell culture conditions, haematopoietic cell lines can stably retain the major features of their original cells (Drexler et al. 2003), thus, information about the real biological activities of the cell types can be elucidated.

3.6.3. Conclusions

In conclusion, a high resolution genomic tiling path microarray was constructed across the human and mouse SCL loci. The unique amino-link technology employed for its construction provided the array with the ability to be highly sensitive and quantitative. Furthermore, the biological material i.e. human haematopoietic cell lines selected to perform ChIP-chip assays for subsequent parts of the study were characterized in detail to establish their cytogenetic characteristics and to examine the structural integrity of the SCL locus. Used together, these resources provide an excellent experimental system to obtain detailed profiles of DNA-protein interactions across the SCL locus using ChIP-chip assays.

Chapter 4

Assessment of the SCL Tiling Arrays for use in ChIP-chip Analysis

4.1 Introduction

Interactions between proteins and DNA mediate transcription, DNA replication, recombination and DNA repair i.e. all the processes that are central to the cell biology of every organism. Over the years, several methods have been employed to identify and catalogue regulatory elements and site-specific DNA-protein interactions in order to gain an insight into these complex events. Although, the traditional methods have provided useful information, they are not very high resolution or high throughput methods or do not always reflect real biological events *in vivo* (see chapter 1 for more detail).

The development of ChIP coupled with whole-genome microarrays (ChIP-chip) has made tremendous progress in recent years to allow construction of high throughput genome-wide maps of DNA-protein interactions. These methods were pioneered in yeast (Iyer et al. 2001; Ren et al. 2000; Simon et al. 2001) and have subsequently been successfully used in other organisms (also see chapter 1). The protocols used for ChIP-chip, in various organisms ranging from yeast to mammalian cells, are very similar. Although the ChIP-chip procedure is fairly simple to perform, there are a number of technical issues and concerns which need consideration when performing these assays. These include formaldehyde fixation, epitope accessibility, availability of good quality and highly specific antibodies, the number of cells required to perform an IP condition, low DNA yields in a ChIP sample and the type and sensitivity of the array platform used to perform these experiments. In addition, issues concerning the analysis and normalization of the array data also need to be considered which could affect the interpretation of the ChIP-chip data obtained. Some of these issues are described below.

Cross-linking: Formaldehyde is a very strong and easily reversible cross-linking (and also denaturing) agent that efficiently produces DNA-protein, RNA-protein, and protein-protein cross-links *in vivo* (Orlando et al. 1997). Cross-linking can affect (i) the accessibility of the antigen epitope (to the antibody) (Orlando et al. 2000), (ii) the sonication of the chromatin (Orlando e t al. 1997) and the efficiency by which different types of proteins are cross-linked (Solomon and Varshavsky 1985).

Antibodies: Arguably the most important parameter in a ChIP assay is the availability of high affinity, high specificity antibodies which would reliably bind to the native protein of interest *in vivo*. Antibodies may cross-react with other proteins containing the same or similar epitopes. This is often a particular problem with protein families - whose members are highly similar in protein structure and expressed in the same cell type at the same time. The antibody specificity can be tested in a number of ways such as western blotting, ELISA and competition assays with peptides containing the epitope of interest (Suka et al. 2001). One way to assess the specificity of the DNA-protein interactions in ChIP is to perform the assay using several different antibodies which have been raised to the same epitope (Liu et al. 2005) or by testing antibodies raised to different epitopes (Horak et al. 2002).

Cell Numbers: Another limiting factor in ChIP-based assays is the number of cells required to perform a single IP condition (usually at least 10^7 cells). Although it is easy to obtain large numbers of cells from cultured cell lines, it is quite often difficult to obtain primary cells in large numbers for some cell types, especially in human.

Amount of ChIP DNA: Typically, a ChIP assay results in a low yield of DNA - usually in the order of several tens to several hundreds of nanograms. Several published studies have made use of either DNA amplification or multiple replicate ChIPs (Weinmann et al. 2002) to obtain sufficient amounts of DNA for microarray analysis. A number of DNA amplification methods have been used, including (i) ligation-mediated PCR (Ren et al. 2000; Pokholok et al. 2005), (ii) random priming (Lieb et al. 2001; Iyer et al. 2001) and (iii) T7-based linear amplification (Liu et al. 2003; Liu et al. 2005; Bernstein et al. 2005). However, amplifying the ChIP material may introduce sequence-dependent and length-dependent non-biological biases in the ChIP sample which could directly affect the ChIP-chip data.

Array Platforms and Analysis: Published studies have reported the occurrence of sporadic, non-reproducible enrichments with high standard deviations between ChIP-chip replicate experiments (Horak et al. 2002), suggesting an issue with reliable array quantitation. Similarly, the lack of consensus binding sequences for the relevant transcription factor in the genomic fragments reporting ChIP enrichments have also been reported (Ren et al. 2002; Weinmann et al. 2002; Cawley et al. 2004; Martone et al. 2003), reflecting the degree of non-biological noise with some ChIP-chip platforms. Given that the data obtained from ChIP-chip experiments should reflect real biological events, developing a robust and reproducible ChIP assay coupled with a sensitive and quantitative array platform is of primary importance.

A number of approaches have aided in the identification and characterization of additional DNA elements involved in the transcriptional regulation of SCL (discussed in chapter 1). Most recently, a detailed survey of histone acetylation across approximately 90 kb of the genomic region at the human and mouse SCL loci was carried out using ChIP in combination with real-time PCR (Figure 4.1). This work led to the identification of an additional regulatory element (named the +40 region) which directs SCL expression to primitive erythroblasts (Delabesse et al. 2005). Despite the identification of various regulatory elements, the complex interplay between these *cis*-acting regulatory elements and the *trans*-acting regulatory proteins is poorly understood. It is evident that the identification and characterization of all of the key regulatory interactions at the SCL locus, including sites of histone modifications, transcription factors and other regulatory proteins, would greatly improve the understanding of its complex regulation.

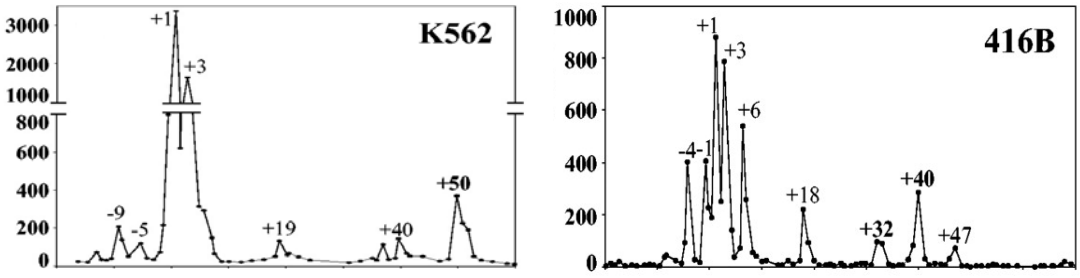


Figure 4.1: ChIP-PCR based histone H3 acetylation profiles in K562 and 416B cell lines. Approximately 90 kb region across the SCL locus was mapped for histone H3 and H4 acetylation using ChIP-PCR based assays in these cell lines. The peaks in the profiles represent regions showing enrichment for H3 acetylation. The numbers denote the distance of these regions upstream (-) or downstream (+) from the start of SCL p1a in human (K562) and mouse (416B). The SCL erythroid enhancer (+50 in human and +40 in mouse) was identified and found to be functional in primitive erythroblasts. (Delabesse et al. 2005) (This figure was taken from Delabesse et al. 2005).

4.2 Aims of the chapter

The aims of the work reported in this chapter were:

1. To further assess the performance of the SCL array platform in terms of reproducibility and sensitivity in ChIP-chip based assays which define different kinds of regulatory features. These include the location of histone modifications, binding sites for transcription factors, and variations in nucleosome levels. The roles of these regulatory features are discussed in chapter 1.
2. To determine whether ChIP enrichments obtained in the above assays reflect true biology by performing various analyses on the data.
3. To further characterize regulatory regions of SCL.

4.3 Overall Strategy

As mentioned in section 4.1, approximately 90 kb region across the SCL locus had been mapped previously for histone H3 K9/14 diacetylation (Delabesse et al. 2005) in a number of haematopoietic cell lines in a ChIP-PCR based assay (see Figure 4.1). These published data sets were used as a benchmark to assess the SCL array platform in ChIP-chip. Thus, ChIP-chip assays were performed for histone H3 K9/14 diacetylation in human K562 and mouse 416B cell lines. The transcription factor GATA-1 is known to bind to the SCL promoter 1a (hereafter called SCL p^{1a}) in erythroid cells (Bockamp et al. 1995) and to regions of the β-globin locus in K562 cells (Horak et al. 2002). Therefore, GATA-1 was chosen to assess the ability of the SCL array platform to identify transcription factor binding sites in K562 at the SCL locus and at regions of the β-globin locus which were spotted onto the SCL array as internal controls. Finally, the ability to detect variations in nucleosome density was assessed using a ChIP-chip assay for histone H3 in K562. The array enrichments reported for histone H3 acetylation and GATA-1 in K562 were verified using SYBR green real-time PCR assay. Based on the results obtained with the ChIP-chip assays performed in the human and mouse cell lines, the regulatory features that were identified were further characterized by sequence analysis and statistical tests to determine whether they reflected real biological events at the SCL locus.

Results

4.4 Establishing criteria for performing chromatin immunoprecipitation and microarray hybridisations

Based on empirical evaluation, a set or criteria were established to aid in obtaining reproducible ChIP-chip assays. These criteria were as follows:

1. Culturing of cell lines: For all ChIP-chip assays described in this thesis, all human and mouse cell lines were cultured according to the protocol described in chapter 2. Prior to harvesting for chromatin immunoprecipitation, aliquots of the cell lines were subjected to flow-sorting (Cytomation MoFlo High Performance Cell Sorter, Dako Cytomation) to determine the DNA content of cells in the population as a means of assessing the proportion of cells which were actively dividing. Only passages of cell lines which showed similar growth characteristics were used in ChIP-chip experiments.

2. Preparation of ChIP DNAs: Chromatin immunoprecipitation, labelling and hybridisation of ChIP DNAs for a wide variety of DNA-protein interactions (see Table 5.1) were performed according to the protocols described in chapter 2. Prior to labelling, the ChIP DNA samples

were routinely electrophoresed on an agarose gel (Figure 4.2). No amplification step was performed prior to the labelling and hybridisation of the ChIP DNA samples.

3. Assessing technical and biological variation for ChIP-chip assays: The SCL array contained two genomic sets and a clone set of array elements (see description of these in chapter 3) which were spotted at different locations on the SCL arrays in order to (i) monitor array position effects and (ii) verify that PCR products derived from different PCR templates would report similarly under identical hybridisation conditions. (Note: only 143 and 250 of the human and mouse SCL array elements respectively were represented in these comparative genomic and clone sets. The complete clone set of 360 and 411 array elements in human and mouse constituted the full complement of array elements found on the final version of the SCL tiling arrays as described in chapter 3). The reproducibility and sensitivity of the ChIP-chip assays were assessed by performing a series of technical and biological replicates (2-3 per assay). Technical replicates included two independent hybridisations derived from using different aliquots of the same ChIP DNA sample. Biological replicates included multiple hybridisations performed from DNA samples which were generated from independent ChIP experiments from different passages of a cell line grown at different times. Performing technical and biological replicates for the ChIP-chip experiments allowed for experimental and biological variations to be taken into account; for example, differences in sample handling and hybridisation conditions, batch to batch variations in cyanine dye incorporation, and differences in culturing conditions and passages of cell lines. Figure 4.3 shows a scanned composite image of the human SCL genomic tiling array as an example of a typical ChIP-chip hybridisation.

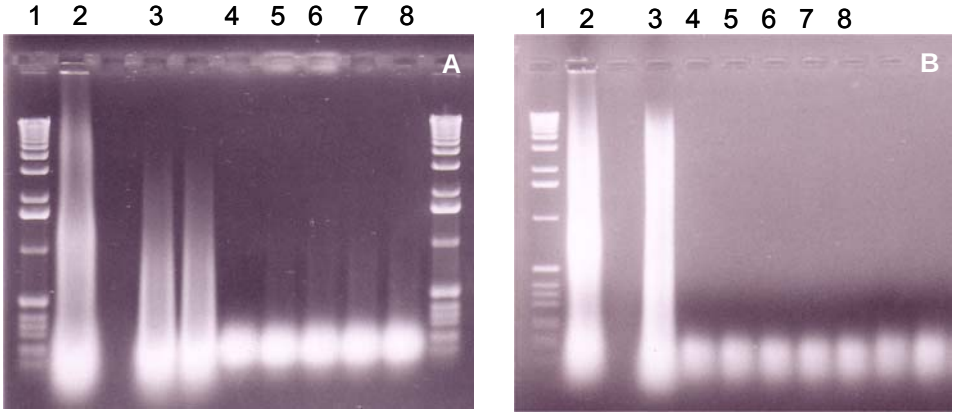


Figure 4.2: The electrophoretic analysis of ChIP DNAs. Panel A shows ChIP assay performed for histone H3 acetylation and panel B for GATA-1 transcription factor. In both panels, lane 1= 1 kb ladder; lane 2= the sheared cross-linked material; lane 3= the input DNA (sheared purified total genomic DNA); lane 4= no antibody control; lanes 5, 6, 7, and 8= ChIP DNAs for H3 acetylation and for GATA-1 in panels A and B respectively. The differences in the average size of the input DNAs in panel A and B (lane 3) reflects

differences in cross-linking time and concentrations of formaldehyde (0.37% for 10 minutes in A; 1% for 15 minutes in B). Faint smears for ChIP DNAs can be seen for H3 K9/14 diacetylation but the DNA samples for transcription factor GATA-1 are not visible suggesting that different amounts of ChIP DNAs are immunoprecipitated in the two assays. The samples were electrophoresed on 1% agarose 1XTBE gels and visualised with ethidium bromide.

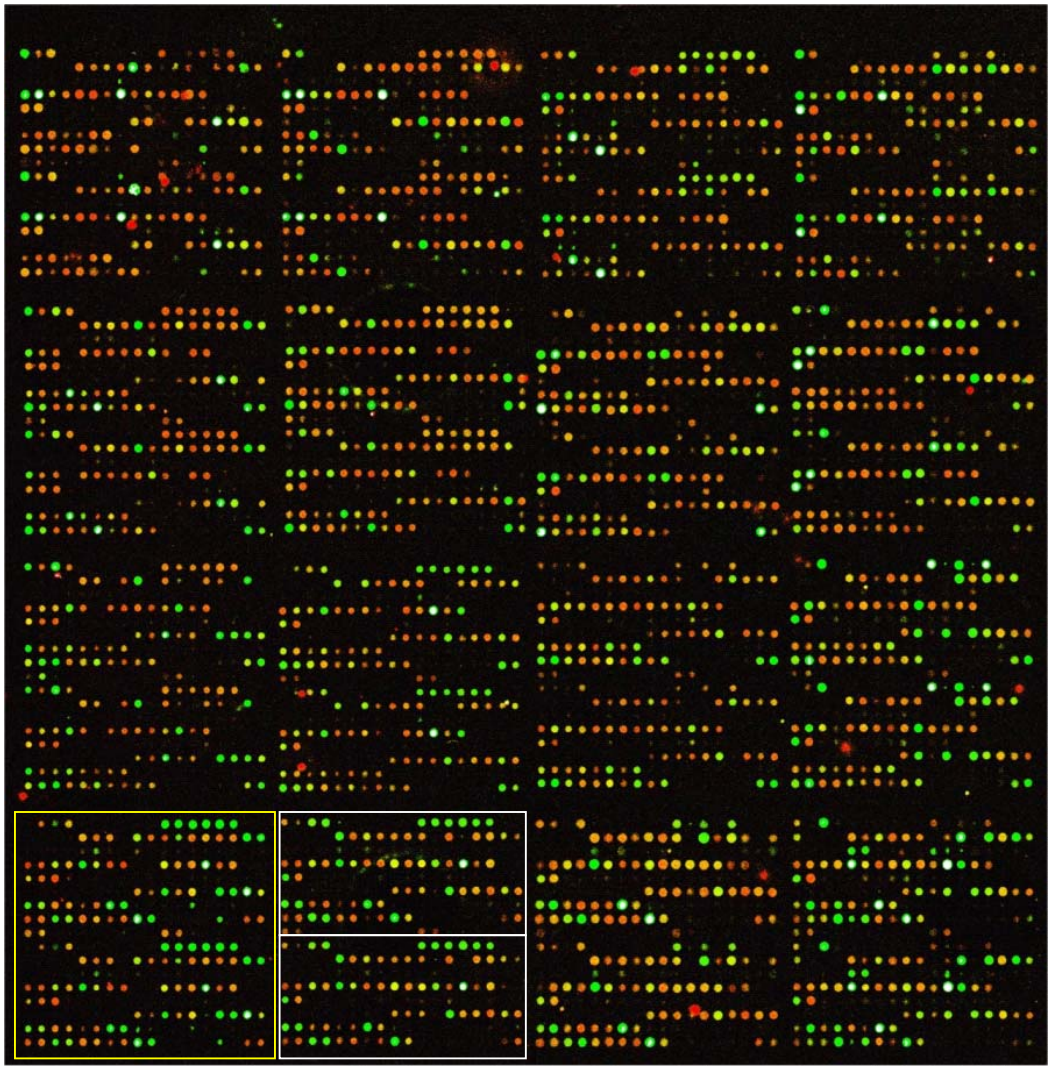


Figure 4.3: A composite image of the human SCL genomic tiling array. The array was hybridised with a ChIP sample for histone H3 K9/14 diacetylation in cell line K562 along with K562 input DNA. Each spot on the array represents an array element. Each array element was spotted in duplicate (a duplicated set of array elements is shown within the two white boxes) in a 16 sub-grid format (a single sub-grid is shown within the yellow box). The green spots on the image show enrichments in the ChIP sample as compared to the input DNA. The yellow spots represent equal hybridisation of the ChIP and input DNA to the spot. Orange/red spots show regions which are under-represented in the ChIP sample. The white spots reflect saturated spots in the ChIP sample.

4. Data analysis: To assess the reproducibility of the array elements within a single hybridisation, mean ratios and coefficients of variation (CVs) were calculated for each array element spotted in duplicate. Similarly, mean ratios and CVs for the clone set of

array elements were calculated for each pair of technical replicate hybridisations. Finally, mean ratios and CVs for the clone set of array elements were calculated across multiple hybridisations representing biological replicates of ChIP assays. Thus, technical and biological reproducibility at the level of the array hybridisation and the ChIP assay could be determined empirically.

Based on the results obtained with the various validation experiments (described in chapter 3), 360 human and 411 mouse array elements derived from the corresponding clone sets were included in the final data sets that were analysed in all the ChIP-chip assays. Significant ChIP enrichments were calculated as follows: two genomic regions representing the cytochrome P450 genes and the coding region of the SIL gene in both human and mouse were chosen to calculate mean background ratios. These regions did not contain any known regulatory regions which were active in the cell lines analysed. In human, the regions spanned from 47262287 bp to 47343557 bp, and from 47424426 bp to 47489321 bp on chromosome 1; in mouse, the regions spanned from 113450516 bp to 113541031 bp, and from 11350226 bp to 113380141 bp on chromosome 4. The significant enrichment threshold was set as three standard deviations (SDs) above the mean background ratios. The ChIP-chip profiles generated for various assays in human and mouse were annotated showing the regions with significant enrichments; the numbering system is based on their distance in kilobases upstream (-) or downstream (+) from the SCL promoter 1a (p^{1a}) in human and mouse respectively, or from the promoter of the nearest gene.

4.5 Performance of the SCL genomic tiling path arrays in detecting regions of histone H3 K9/14 diacetylation in human and mouse

4.5.1 Assessing the performance of SCL array in ChIP-chip assays for H3 K9/14 diacetylation in K562 cell line

It is well established that histone posttranslational modifications play an intrinsic role in transcriptional regulation, with histone H3 acetylation having been shown to be a hallmark of active genes (see chapter 1). Figure 4.4 shows the results of a series of array hybridisations on the human SCL tiling array for H3 K9/14 diacetylation in K562 cells. The features of these profiles will be described in detail in section 4.5.2. From these profiles, it is evident that each of the array elements included in the final data analysis performed in a highly similar manner irrespective of (i) its position on the array, (ii) the template from which it was initially amplified, or (iii) whether the ChIP DNAs were derived from technical or biological replicates. Across a number of these experiments, the reproducibility assessed by calculating mean ratios and CVs of duplicate spots can be summarized as follows:

1. The mean CV of ratios reported by the genomic sets and the clone set of array elements in a typical experiment was approximately 8% within a single hybridisation (assessed for 6 independent hybridisations).

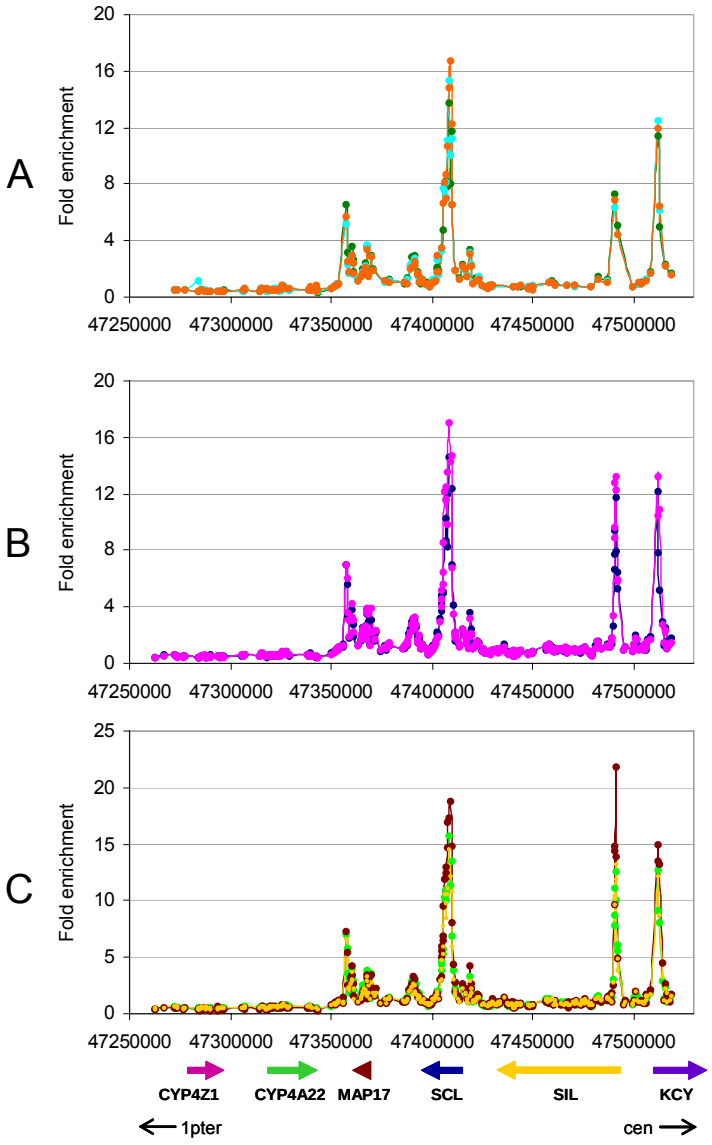


Figure 4.4: Performance of the SCL array platform in ChIP-chip assays for H3 K9/14 diacetylation in K562. Panel A: The orange (genomic), green (genomic) and turquoise (clone) profiles illustrate the reproducibility of the SCL array elements which were amplified from different PCR templates (Chapter 3). Panel B: The pink and blue profiles illustrate the reproducibility of the array elements in technical replicates from two independent hybridisations derived from using different aliquots of the same ChIP DNA. Panel C: The yellow, brown and green profiles illustrate the reproducibility of three independent biological replicates. Each profile in panel C represents the mean ratio of two technical replicates per biological replicate. The gene order and the direction of transcription are shown at the bottom of panel C. The x-axes represent the genomic coordinates along human chromosome 1 and the y-axes represent fold enrichments. The orientation of the locus with regards to centromere (cen) and telomere (ter) is shown at the bottom with black arrows.

2. The mean CV of ratios reported between corresponding array elements in technical replicate hybridisations was less than 6% (assessed for each of three sets of 2 technical replicates).

3. The mean CV of ratios reported between corresponding array elements in three biological replicates was less than 10% (assessed across 3 biological replicates).

Based on the results described above, the greatest source of variation was observed between the biological replicates. Furthermore, the high degree of reproducibility and across multiple ChIP-chip assays for histone H3 K9/14 diacetylation in K562 cell line

suggested that the SCL array platform was robust at reporting consistent measurements. Based on these profiles, a number of genomic regions were identified with significant enrichments for H3 K9/14 diacetylation – these are discussed in the next section.

4.5.2 Histone H3 acetylation coincides with regulatory function at the SCL locus in human

Figure 4.5 shows the ChIP-chip profile for H3 K9/14 diacetylation across the human SCL locus for the cell line K562 based on the mean ratios of array elements derived from the three biological replicates shown in Figure 4.4 (panel C).

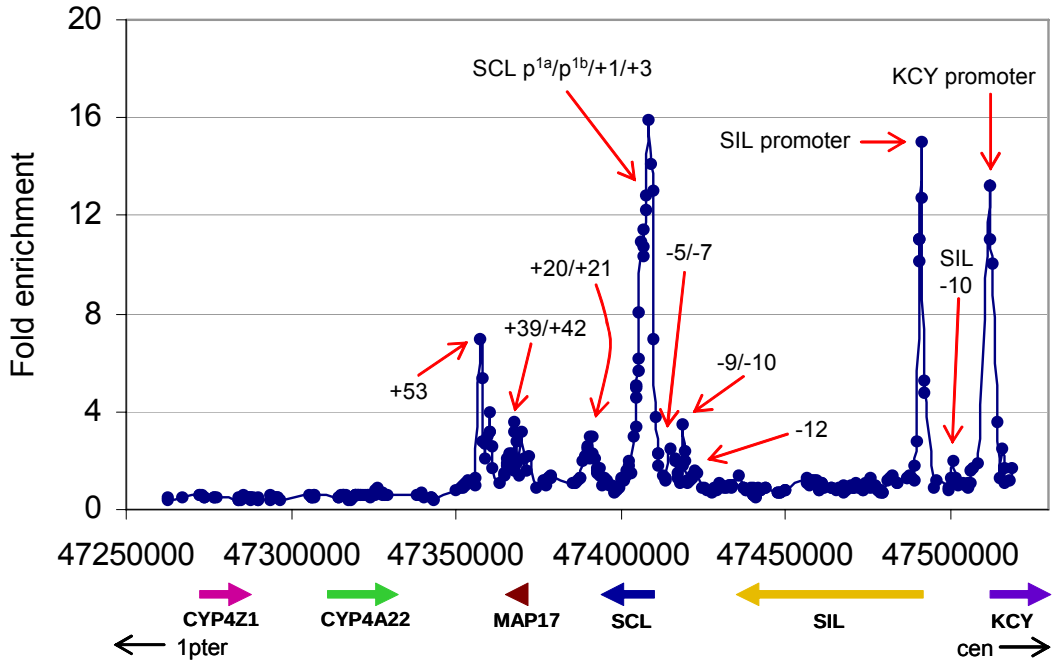


Figure 4.5: ChIP-chip profile of H3 K9/14 diacetylation across the human SCL locus in K562. Genomic regions showing significant enrichments for H3 K9/14 diacetylation are marked with red arrows. The genomic regions denoted by +1, +3, +7 etc. are based on their distance upstream (-) or downstream (+) in kilobases from the human SCL p^{1a}. The x-axis represents the genomic coordinates along human chromosome 1 and y-axis represents fold enrichments. The thick coloured arrows represent the gene order and the direction of transcription as shown at the bottom of the panel. The orientation of the entire locus on chromosome 1p with regards to the centromere (cen) and telomere (ter) is shown at the bottom by thin black arrows.

A number of regions with significant enrichment for H3 K9/14 diacetylation were identified which included those with enrichment ratios as low as 1.54 fold. The most prominent enrichments were found at or near the 5' ends (i.e., promoters) of the KCY, SIL and SCL genes and all of these genes are expressed in K562. The acetylation across the 5' end of SCL extended 8 kb into its coding regions encompassing sequences which show DNase I hypersensitivity, enhancer or promoter activity and include the +1, +3, and p^{EXON4} (+7) regions (Leroy-Viard et al. 1994). Enrichments were also found at the

stem-cell enhancer (+20/+21) (Gottgens et al. 2002), the MAP17 promoter at +42, the -9/-10 region (Gottgens et al. 1997) and the SCL erythroid enhancer at +51 (Delabesse et al. 2005) (which extends from +50 to +53 with the highest peak at +53).

In addition to the known regulatory elements, a number of other genomic regions of significant enrichments for H3 diacetylation were identified within the SIL gene at -10 and the SCL gene at -12, -5/-7 and +39. Two of these, -5/-7 and +39, were also identified with the ChIP-PCR based assay (Delabesse et al. 2005) but no regulatory function for these elements is known. No significant enrichments were seen across the genomic region containing the CYP genes, CYP4A22 and CYP4Z1.

4.5.3 Histone H3 acetylation array data correlates with real-time PCR of H3 acetylation ChIP material

To determine whether the SCL array accurately reported the enrichments found in ChIP material, SYBR green quantitative real-time PCR was performed on aliquots of the same K562 H3 K9/14 diacetylation ChIP material that was used to hybridise to the arrays. The results obtained with the real-time PCR are shown in Figure 4.6 (the sequences of the primer pairs used in real-time PCR are listed in Appendix 4).

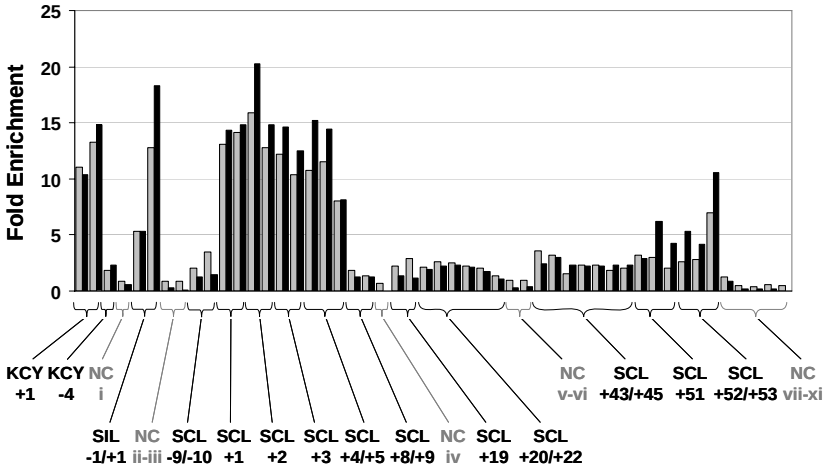


Figure 4.6: SyBr green real-time PCR of ChIP material for H3 acetylation across the human SCL locus in K562. The data points across the locus are shown on the x-axis. Enrichments reported by the array (grey bars) and those reported by real-time PCR (black bars) are shown as pairs for each region tested (along y-axis). Enrichments are ordered with respect to their genomic co-ordinates and bracketed according to the regions they were derived from across the human SCL tiling region. The nomenclature of data points refers to the distance in kilobases the amplicon is located upstream (-) or downstream (+) from the promoter of the closest gene. NC = negative control regions.

The fold enrichments obtained for each region with real-time PCR are shown alongside the enrichments obtained with the array hybridisation. The mean quantitative PCR

values were normalized against the corresponding array data by deriving the median ratio of both datasets and scaling the values accordingly.

For the regions tested, the array elements which showed significant enrichments in the array data also showed acetylation enrichments with real-time PCR. In most instances, the real-time PCR data and array data were similar, particularly for the array elements which showed lower enrichments. However, the array elements which showed the highest enrichments in the array data often showed even higher enrichments with real-time PCR. This latter finding could be explained by the fact that quantitation of high intensity array signals are no longer linear, once pixel values reach saturation. However, this is not the case for real-time PCR as quantitation should be completely linear with DNA copy number in the ChIP sample. Thus, highly enriched sequences will tend to be under-reported on the array and will no longer reflect accurate levels of enrichment in the ChIP sample.

4.5.4 ChIP-chip profiles for Histone H3 K9/14 diacetylation in 416B cell line

ChIP-chip experiments for H3 K9/14 diacetylation were also performed in the mouse blood progenitor cell line 416B. The array data generated using technical replicates was highly reproducible with mean CVs reported to be less than 10% when the corresponding array elements across the technical replicates were compared.

Significant enrichments were obtained across many of the same regulatory regions as those detected in the K562 cell line (based on the functional orthology shown in Figure 4.7). It should be noted that the 5' end of the KCY gene was not present on the mouse SCL tiling path array. Significant enrichments were obtained at or near the 5' ends of SIL and SCL genes, both of which are expressed in 416B cell line. Acetylation across the 5' end of SCL extended up to 12 kb into the coding region encompassing the +1, +3 and SCL p^{EXON4} (+7) regions. Additionally, H3 diacetylation was seen at the -8 region and the endothelial enhancer at -3/-4 region (Gottgens et al. 2004), at the +47 region (Delabesse et al. 2005, Figure 4.1) and at a number of novel regions – including one downstream from SCL p^{1a} at +15, and two upstream within introns of the SIL gene (denoted SIL +18/+19 and SCL -20/-23 from hereafter).

When the mouse 416B and the human K562 data were taken together, it was possible to identify virtually all known SCL regulatory elements in the genomic regions tiled on the SCL arrays using ChIP-chip assays to H3 diacetylation alone. The identified elements included all the SCL regulatory elements known to show enhancer activity in haematopoietic cell types. Only two known SCL regulatory elements [the +23 neural enhancer (Gottgens et al. 2000) and the tal-RE silencer element at +14 (Courtes et al. 2000) were not detected by the ChIP-chip assays across the SCL locus (see Discussion).

The identification of additional regions of significant enrichments of H3 diacetylation in both cell lines, suggests that these sequences may also have regulatory activity.

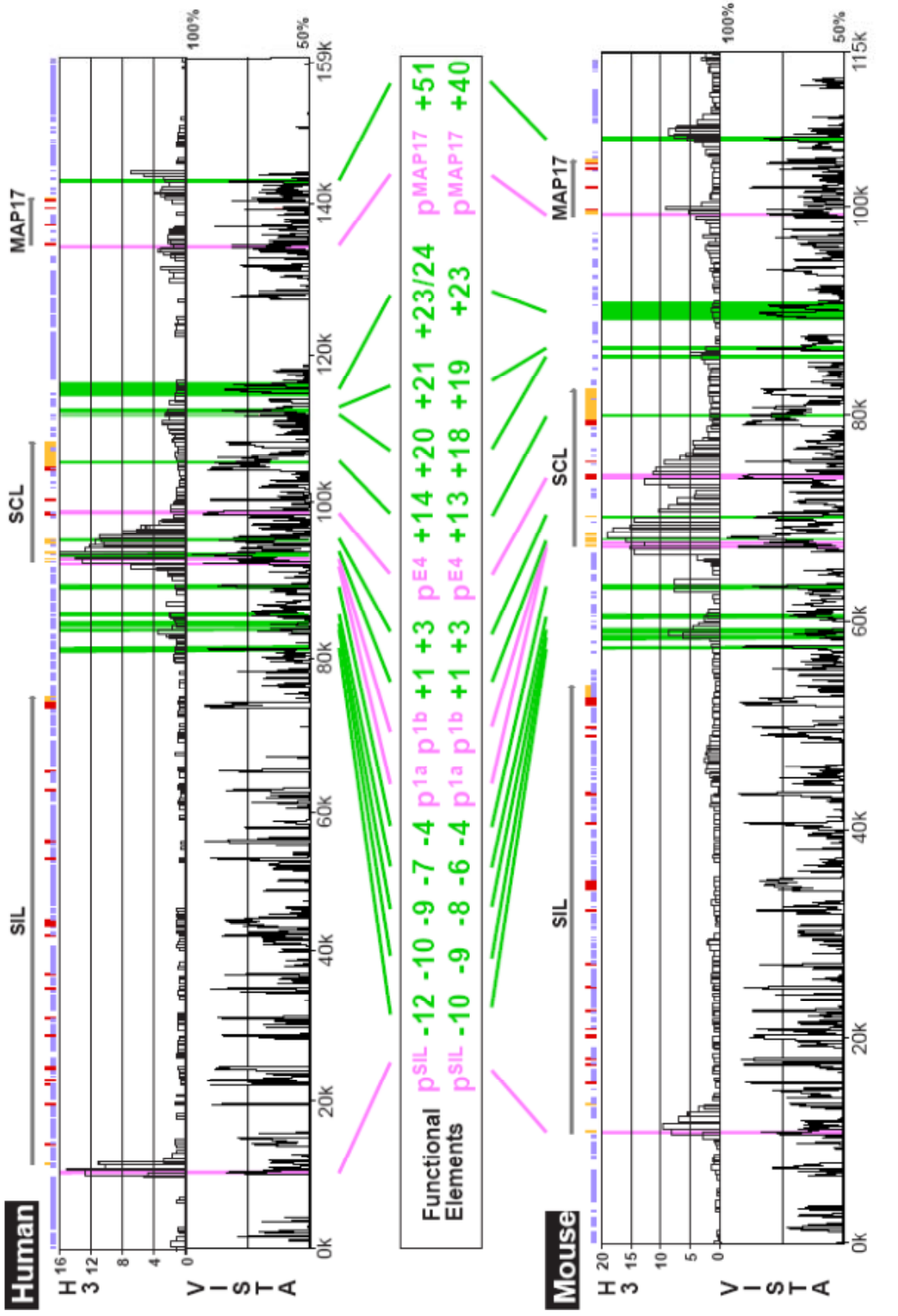


Figure 4.7: Comparison of histone H3 diacetylation profiles across the SCL locus in human K562 and mouse 416B cell lines. Profiles are shown for the genomic region spanning from upstream of the SIL promoter to the MAP17 gene and SCL erythroid enhancer (+51 in human/+40 in mouse). The top panel shows the histone H3 diacetylation profile for K562 and the bottom shows that for 416B. VISTA plots of human-mouse sequence homology (with respect to percentage of sequence identity for 100 base pair windows) are shown adjacent to each profile. The y-axis on the H3 acetylation plots refers to fold enrichments for each amplicon which are shown as vertical bars in order across the tile path. The width of the bars reflects the sizes of the amplicons on the array. Functional elements and their orthologous relationships in human and mouse are shown in the central portion of the figure. Known functional elements (as determined from the work in this thesis or from previous studies) are annotated as the long vertical coloured bars in the top and bottom panels. Promoter sequences are shown in mauve and other regulatory sequences in green. The numbering system for the elements denotes their distance in kilobases upstream (-) or downstream (+) from the start of SCL promoter 1a. The locations of gene coding regions are shown as red (protein coding) and amber (non-protein coding) vertical bars along the x-axis. Repeat sequences are also shown as purple bars. The names and direction of transcription for each gene are shown by the grey arrows.

4.5.5 Sequence conservation at sites of histone H3 acetylation in human and mouse

The above results for H3 diacetylation suggested that the interactions identified using the SCL array platform reflected real regulatory interactions at the SCL locus. Hence, further analyses were performed to understand and explore the relationships between H3 diacetylation and sequence conservation at the SCL locus. In previously reported studies, SCL locus has served as an excellent experimental paradigm for establishing the relationship between non-coding sequence conservation and regulatory function (Gottgens et al. 2000; Gottgens et al. 2001; Gottgens et al. 2002; Chapman et al. 2004). To extend this analysis, the array elements which showed significant enrichments for histone H3 K9/14 diacetylation in K562 and 416B cells were grouped with respect to levels of human-mouse sequence conservation in non-coding DNA. In order to do this, the mean sequence conservation of non-coding DNA for each array element was calculated in human and mouse. Then the array elements were grouped together by their level of sequence conservation; and the average level of acetylation was calculated for each group (Table 4.1, panels A and C). A strong positive correlation was found between the level of enrichment for H3 K9/14 diacetylation and the level of sequence conservation for these groups (R = 0.93 and 0.89 for the human and mouse data respectively for acetylated regions that have sequence conservation with greater than 50% identity, see Appendix 12). These results suggested that there is a relationship between the degree of sequence conservation and the levels of enrichment for histone H3 acetylation in the non-coding DNA at the SCL locus. However, high sequence conservation was not limited to only those regions which were acetylated - non-acetylated non-coding DNA displayed sequence conservation that was similar to the regions that were acetylated (Table 4.1, panels B and D).

A

Sequence identity (%) for non-coding DNA (100 bp windows)	No. of array elements showing significant acetylation that have non-coding sequence identity	Average level of conserved non-coding sequence (%) for array elements	Average enrichment level of acetylation for array elements
95-100	2	98	13.599
80-84.9	3	80	8.099
75-79.9	7	77	5.929
70-74.9	6	73	7.326
65-69.9	17	67	3.996
60-64.9	24	62	4.601
55-59.9	24	57	3.045
50-54.9	6	52	4.049

B

Sequence identity (%) threshold for non-coding DNA (100 bp windows)	No. of conserved non-coding sequence peaks in tiled region	No. of conserved non-coding sequence peaks that are found in array elements showing significant acetylation	Average level of conserved non-coding sequence (%) in array elements showing significant acetylation	No. of conserved non-coding sequence peaks that are not associated with array elements showing acetylation	Average level of conserved non-coding sequence (%) in array elements not associated with acetylation
≥75	27	10	80	17	81
≥70	40	19	77	21	78
≥65	74	34	73	40	72
≥60	143	69	67	74	66
≥55	199	89	64	110	63
≥50	237	100	62	137	60

C

Sequence identity (%) for non-coding DNA (100 bp windows)	No. of array elements showing significant acetylation that have non-coding sequence identity	Average level of conserved non-coding sequence (%) for array elements	Average enrichment level of acetylation for array elements
95-100	2	98	14.312
90-94.9	1	93	15.237
80-84.9	3	82	12.703
75-79.9	5	77	6.059
70-74.9	3	71	4.93
65-69.9	10	67	7.289
60-64.9	19	62	5.731
55-59.9	15	57	6.314
50-54.9	4	53	3.549

D

Sequence identity (%) threshold for non-coding DNA (100 bp windows)	No. of conserved non-coding sequence peaks in tiled region	No. of conserved non-coding sequence peaks that are found in array elements showing significant acetylation	Average level of conserved non-coding sequence (%) in array elements showing significant acetylation	No. of conserved non-coding sequence peaks that are not associated with array elements showing acetylation	Average level of conserved non-coding sequence (%) in array elements not associated with acetylation
≥ 75	34	16	82	18	80
≥70	48	22	78	26	78
≥65	84	32	74	52	72
≥60	150	52	69	98	67
≥55	207	67	65	140	63
≥50	245	77	63	168	61

Table 4.1: The relationship between sequence conservation and histone H3 K9/14 diacetylation at the human and mouse SCL loci. Panel A: H3 K9/14 diacetylation levels in K562 at human non-coding sequence. Panel B: Human non-coding sequence conservation at H3 K9/14 diacetylated and non-acetylated regions in K562. Panel C: H3 K9/14 diacetylation levels in 416B at mouse non-coding sequence. Panel D: Mouse non-coding sequence conservation at H3 K9/14 diacetylated and non-acetylated regions. The non-coding sequence conservation thresholds are shown in the first column of each table.

4.6 Performance of the human SCL genomic tiling path array in detecting binding sites for the GATA-1 transcription factor in K562 cells

In addition to histone modifications, transcription factors mediate important regulatory interactions which control gene expression (see chapter 1). Therefore, it was important to determine whether ChIP-chip analysis using the SCL tiling arrays could reliably report these types of interactions. The transcription factor GATA-1 was chosen to test the performance of the SCL array as it is known to play important roles in the development of some, but not all haematopoietic lineages (reviewed in Ferreira et al. 2005), and it is known to bind to the SCL p^{1a} in erythroid cells (Bockamp et al. 1995).

4.6.1 Assessing reproducibility of the array elements in ChIP-chip assay for GATA-1 transcription factor

Similar to the assessment experiments carried out for histone H3 K9/14 diacetylation (see section 4.5.1), the human SCL tiling arrays were tested for their reproducibility in ChIP-chip assays for GATA-1 in K562 cells.

As was observed with the ChIP-chip assays for H3 K9/14 diacetylation, the human SCL array reported reproducible enrichments for technical and biological replicates in ChIP-chip assays for GATA-1 (Figure 4.8). The description of these profiles with respect to SCL regulatory regions is described in the following section 4.6.2.

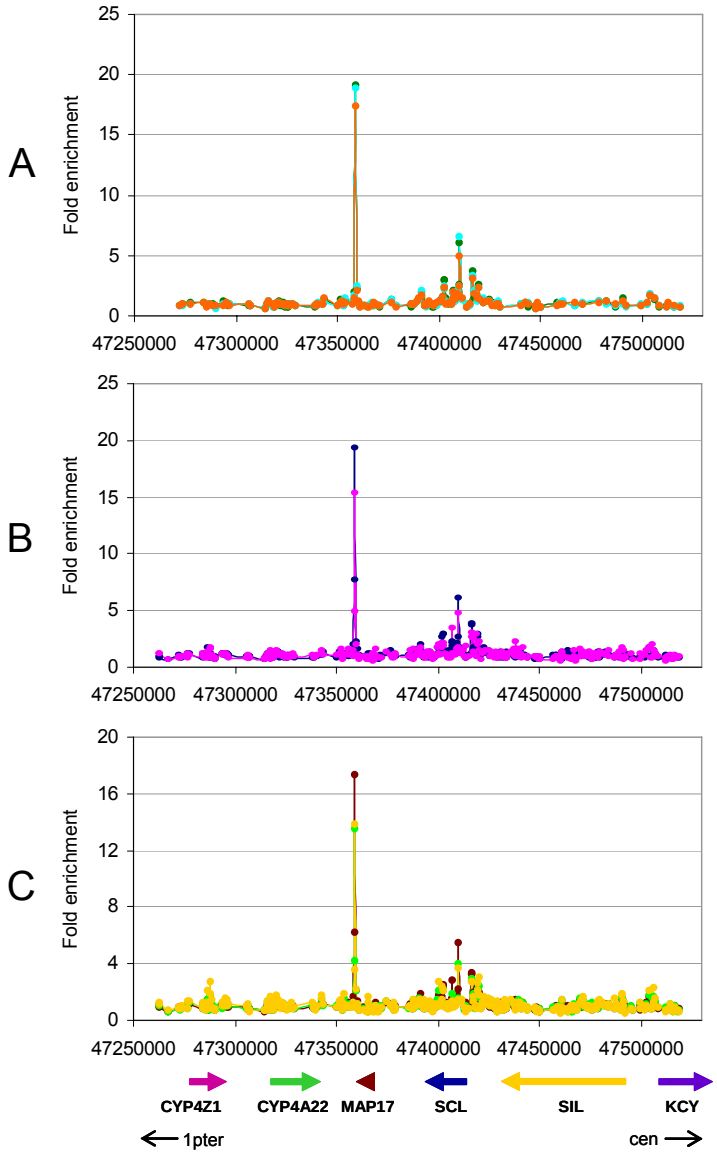


Figure 4.8: Performance of the SCL array platform in ChIP-chip assays for GATA-1 in K562. Panel A: The orange (genomic), green (genomic) and turquoise (clone) profiles illustrate the reproducibility of the SCL array elements which were amplified using different PCR templates (Chapter 3). Panel B: The blue and pink profiles illustrate the reproducibility of the array elements in technical replicates from two independent hybridisations derived from using different aliquots of the same ChIP DNA. Panel C: The yellow, brown and green profiles illustrate the reproducibility of three independent biological replicates. Each profile in panel C is the mean ratio of two technical replicates per biological replicate. The gene order and the direction of transcription are shown at the bottom of panel C. The x-axes represent the genomic coordinates along human chromosome 1 and the y-axes represent fold enrichments. The orientation of the entire SCL locus with regards to the centromere (cen) and telomere (ter) on chromosome 1p is shown at the bottom with black arrows.

The reproducibility of the ChIP-chip assays can be summarized as follows:

1. The mean CV of ratios reported by the genomic sets and the clone set of array elements in a typical experiment was approximately 10% within a single hybridisation (assessed for 6 independent hybridisations).
2. The mean CV of ratios reported between corresponding array elements in technical replicate hybridisations was between 6-10% (assessed for each of three sets of 2 technical replicates).
3. The mean CV of ratios reported between corresponding array elements in three biological replicates was approximately 13% (assessed across 3 biological replicates).

As described in chapter 3 (section 3.3.4), three amplicons (named here and in Horak et al. 2002 as Hb/9BG, Hb/32BG-1 and Hb/32BG-2) spanning known GATA-1 binding regions in K562 at the human β-globin locus (Horak et al. 2002) were also included on the SCL array as positive controls. Horak et al (2002) reported fold enrichments of between 2-4 fold in GATA-1 ChIP-chip using a different array platform. The maximum fold enrichments observed with the SCL array platform were no higher than 2.2 fold for these regions. The product showing the highest enrichment was the same in both studies (Hb/9BG) (data not shown).

4.6.2 Identification of regions enriched for GATA-1 across the SCL locus in K562 cell line

Figure 4.9 shows the ChIP-chip profile for GATA-1 enrichments across the human SCL locus in K562 cells based on the mean ratios obtained from the three biological replicates shown in Figure 4.8 (panel C).

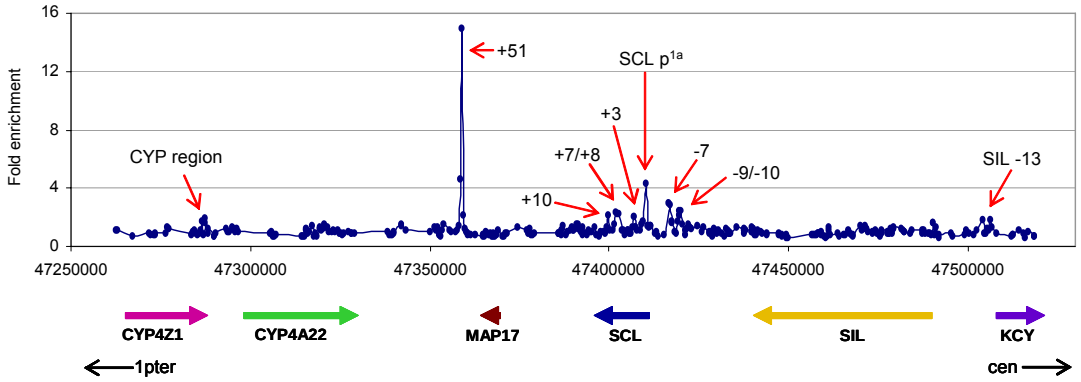


Figure 4.9: ChIP-chip profile for GATA-1 across the human SCL locus in K562. Genomic regions showing significant enrichments for GATA-1 are marked with red arrows. The maximum enrichment was seen at the +51 erythroid enhancer. The genomic regions denoted by +3, -7, -9, etc. are based on their distance in kilobases upstream (-) or downstream (+) from the human SCL p^{1a}. The x-axis represents the genomic coordinates along human chromosome 1 and the y-axis represents fold enrichments. The thick coloured arrows represent the gene order and the direction of transcription as shown at the bottom of the panel. The orientation of the SCL locus on chromosome 1p with regards to the centromere (cen) and telomere (ter) is shown at the bottom of the figure with thin black arrows.

In total, 21 array elements representing nine different genomic regions were identified showing significant enrichments (1.71 to 14.9 fold) for GATA-1. 17 out of the 21 array elements that showed enrichments were at or within 500 bp of regions of histone H3 K9/14 diacetylation and known regulatory activity. Among those most relevant to SCL expression, the highest enrichment was seen at the SCL erythroid enhancer at +51 (Delabesse et al. 2005), and lower but significant enrichments at the SCL p^{1a}, and the -9/-10, -7, +3, and +7/+8 regions (downstream from SCL p^{EXON4}) and at +10. In addition,

significant enrichments were observed at two other regions - one was at SIL -13 and the other was located within the genomic regions containing the cytochrome P450 gene cluster. No regulatory function is known for these regions.

4.6.3 GATA-1 array data correlates with real-time PCR of GATA-1 ChIP material

In order to test whether the enrichments observed in ChIP-chip with GATA-1 were representative of DNA enrichments in the GATA-1 ChIP material, quantitative real-time PCR was performed (sequences of the primer pairs used for the real-time PCR are listed in Appendix 5). Real-time PCR assays were tested for sequences found within 8 of the array elements which were significant in GATA-1 ChIP-chip assays in K562 (Figure 4.10). Seven out of the 8 (not for SIL -13 in Figure 4.10) were found to be above background enrichment levels; this suggested that *bona fide* binding sites for GATA1 had been identified in a high proportion of instances. As seen with the H3 K9/14 diacetylation real-time PCR assay, the ChIP-PCR assay and ChIP-chip assay were highly reproducible for lower enrichments but showed differences for higher level enrichments. This could again be explained in terms of signal saturation of these spots on the array leading to inaccuracies in the reporting of enrichment levels (see section 4.5.3).

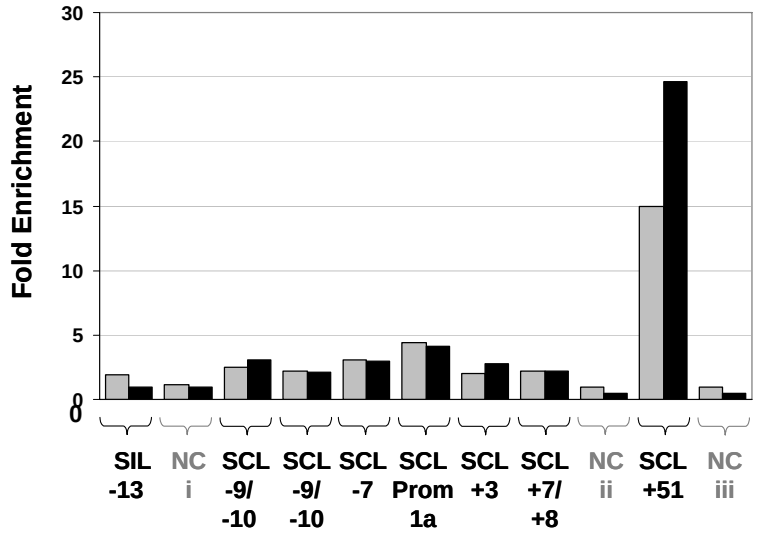
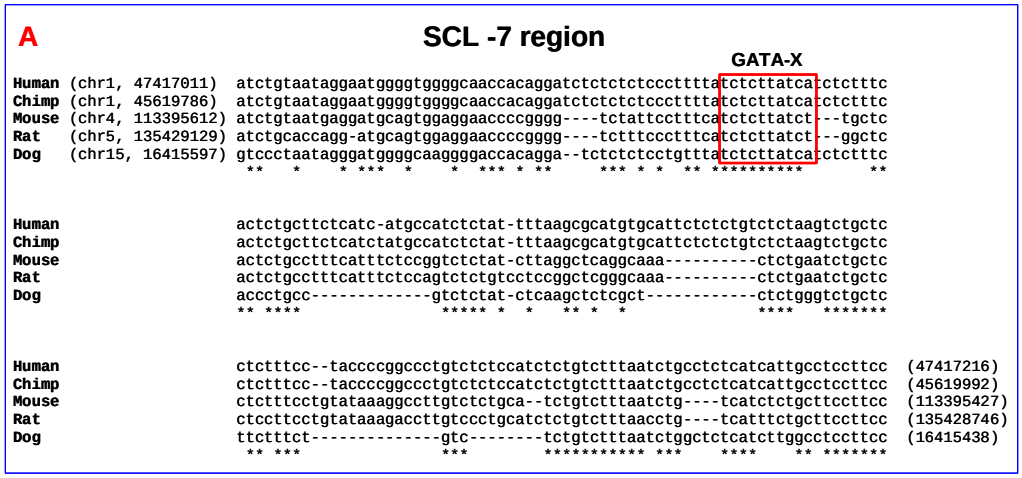


Figure 4.10: SYBR green real-time PCR of ChIP material for GATA-1 across the human SCL locus. The data points across the locus are shown on the x-axis. Enrichments reported by the array (grey bars) and those reported by real-time PCR (black bars) are shown as pairs for each region tested (along y-axis). Enrichments are ordered with respect to their genomic co-ordinates and bracketed according to the regions they were derived from across the human SCL tiling region. The nomenclature of data points refers to the distance the amplicon is located upstream (-) or downstream (+) in kilobases from the promoter of the closest gene. NC = negative control regions.

4.6.4 Conserved sequence analysis of genomic regions showing enrichments with GATA-1 ChIP in K562

To further provide evidence that the GATA-1 enrichments obtained on the human SCL array represented *bona fide* GATA-1 binding sites in K562, each of the regions of enrichments were surveyed for the consensus binding site of the GATA family of transcription factors. These included the SCL p^{1a}, the -9/-10, -7, +3, +7/+8, +10 regions and the +51 SCL erythroid enhancer (Delabesse et al. 2005). GATA family binding sites were identified in all of the regions (data not shown). Comparative sequence alignments performed with the SCL orthologous regions in human, chimp, mouse, rat and dog revealed that the -7 region, SCL p^{1a}, the +3 region and the +51 region contained GATA family binding sites and sites for other transcription factors which were conserved across species (Figure 4.11). Given that multi-species conservation at the sequence level is likely to suggest functional conservation, this data further supports the idea that real GATA-1 interactions in K562 were identified using ChIP-chip assays on the human SCL tiling array.

Of all of the ChIP-chip enrichments observed for GATA-1 in K562, the highest one was found at the SCL +51 region (see Figure 4.9). The +51 erythroid enhancer has been shown to direct expression in the primitive erythroid lineage and contains binding sites for GATA, Ets, bHLH and homeobox families of transcription factors (Delabesse et al. 2005). By analyzing the sequence of the array element which showed the highest enrichment in the ChIP assay, it was found that there was a striking conservation of three GATA and other transcription factor binding sites. One of these conserved GATA sites was found to be contained within a 20 bp sequence which showed the canonical hallmarks of the SCL erythroid DNA-binding complex which includes SCL, E47, GATA-1, LMO2 and LDB1 transcription factors (Wadman et al. 1997) (Figure 4.11, panel D; also see chapter 5).



B

SCL promoter 1a region

```
Human (chr1, 47409975)  tccaccgacgcgctg------------------------taatcccac-tcacgttc----caggcctc
Chimp (chr1, 45612760)  tccaccgacgcgctg------------------------taatcccac-tcacgttc----caggcctc
Mouse (chr4, 113402393) -------gcgcgctc------------------------taatcccgc-tcccgccc----gcggcctc
Rat   (chr5, 135435796) -------gcgcgctc------------------------taatcccgc-tcccgccc----gcggcctc
Dog   (chr15, 16422725) gcgggcgccgcgcgc------------------------taatcccgc-tcccgctc----ccggcctc
                              *****                            ******* * ** **  *      ******

                                                                    GATA-1/X
Human                   gttagcatgggccgaggcgcgccggggccgtgtgcgccgcagagataag-gcactgccgcgggtccgcc-
Chimp                   gttagcatgggccgaggcgcgccggggccgtgtgcgccgcagagataag-gcactgccgcgggtccgcc-
Mouse                   gtta--------------gcgcc--ggccgtgtgcgccgccgagataag-gagccgcctcgggcccgcc-
Rat                     gtta--------------gctcc--ggccgtgtgcgccgcggagataag-gcgccgctgcgggcccgcc-
Dog                     gttagcatgggccgaggcgcgccggggccgcgtgcgccgccgagataag-gcgcagccgcggggccgct-
                        ***               ** **  ***** ********* ******** *  * **  **** ****

                              GATA-1
Human                   cgccccgataagcgcctcgg-ccattatgggccaaatgattcattttaatttggcaatttcaccggagg  (47410151)
Chimp                   cgccccgataagcgcctcgg-ccattatgggccaaatgattcattttaatttggcaatttcaccagagg  (45612936)
Mouse                   cgccccgataagcgcctcgg-ccattatgggccaaatgattcattttaatttggcaatttctccggagg  (113402240)
Rat                     cgtccgcgataagcgcctcgg-ccattatgggccaaatgattcattttaatttggcaatttctccggagg  (135435643)
Dog                     cgccccgataagcgcctcgg-ccattatgggccaaatgattcattttaatttggcaatttcaccggagg  (16422549)
                        ** ** ******** ****** ****************************************** *******
```

C

SCL +3 region

```
                                                GATA-1
Human (chr1, 47406973)  acccgggtctttgctttccccctttttcgctgagaggcctgcagttacg----ctgcggtgtggtcctgg
Chimp (chr1, 45609761)  acccgggtctttgctttccccctttttcgctgagaggcctgcagttacg----ctgcggtgtggtcctgg
Mouse (chr4, 113405231) -cccgaagctggg--tttcccggctctcgctgagagaccagcagttacttagcttgccgtgaggtcctgc
Rat   (chr5, 135438566) -cccggagctgtg-ttttccaggctctcgctgagagaccagcagttacttagcttgccgtgaggtcctac
Dog   (chr15, 16425538) acgggggtctctgctttccccctctctcgctgacagaccagcagctagg----ccaccgtgtggtcctgg
                         *  *   **  *  ** **    * ******* ** ** **** **         * *** ******

                                                GATA-1/3
Human                   gcgatctgccgcgcccaaagcgagtttccaggaaagatagggggtggagagaaagaggcagggcaagaggg
Chimp                   gcgatctgccgcgcccaaagcgagtttccaggaaagatagggggtggagagaaagaggcagggcaagaggg
Mouse                   gcggtctgccgcgcccaaagcgagtttccaggaaagatgggggtggagagaaagagggagggc-------
Rat                     gcggtctgccgcgcccaaagcgagtttccaggaaagatgggggtggagagaaagagggagggc-------
Dog                     gcagtctgccgcgcccaaagccagtttccaggaaagatgggggtggagagaaagagggagggcgagaggg
                        **  ***************** ********** ********** ********** *****


Human                   agggagagagagaaatgggggtcaatggctgggaattacctcctgtccccgaccaaccagtccagggaat
Chimp                   agggagagagagaaatgggggtcaatggctgggaattacctcctgtccccaaccaaccagtccagggaat
Mouse                   ---gagagagaaaatgggggtcaatgactgggaattacctcgtgtgccccgccatc----ccaaagaat
Rat                     ---gagagagaaaatgggggtcaatggctgggaattacctcgtgtgccccgccatc----ccaaagaat
Dog                     agggagagagagaaatgggggtcaatggctgggaattacctcctgt-cctcgccaaccagcccagggaac
                           ******** **  *********** ********** *** **   *** *    ***  ***


Human                   cgcaaacagctttccgcccccaacctgcaggtgtttggagcctttcctctacccccctcctcc---cccca
Chimp                   cgcaaacagctttccgcccccaacctgcaggtgtttggagcctttcctctacccccctcctcc---cccca
Mouse                   agcaaatggctttccgcctccaacttgcaggtgtttggagcc---------caccctccccaaacacat
Rat                     agcaaacggccttccgcctccaacttgcaggtgtttggagc----------ccccctccccaaacacat
Dog                     tgcggagggctttctgcccccacccctgcaggtgttagaggc----------ctccc---------cccaa
                         **  *  ** *** *** *** * ********** *  **          * ***         * *

                                                GATA-1
Human                   cccgcct--taaaa----------------aaccctatacatgaccaatcaggaatagctatttagtcca  (47407297)
Chimp                   cccgccttataaaa----------------aaccctatacatgaccaatcaggaatagctatttagtcca  (45610087)
Mouse                   cccgaat—-taaaaaaagaaaaagaagaagaaccttacacctgaccaatcaggaataactacttagtcca  (113404910)
Rat                     cccgatt--taaaaaaaaaaaaaaatg---aaccctacacctgaccaatcaggaataactacttagtcca  (135438248)
Dog                     accccct--taaaaaag-------------aaccctacacctgaccaatcagaaataactatttactcca  (16425228)
                         **   *  *****               **** ** ** *********** **** *** *** ****
```

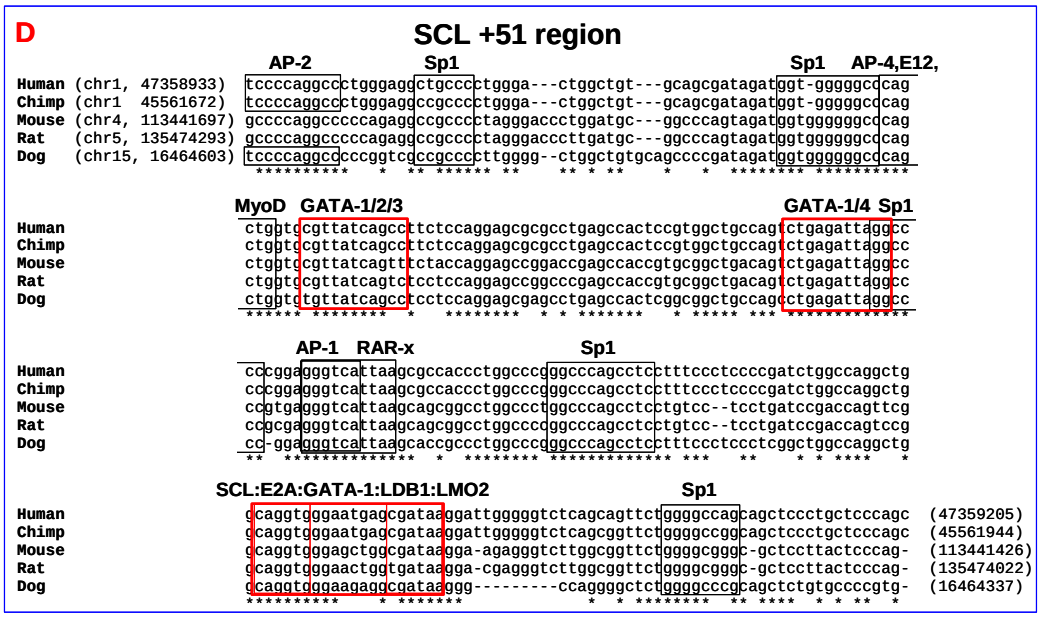


Figure 4.11: Conserved sequence analysis to identify conserved GATA and other transcription factors binding sites at SCL regulatory regions. The sequence alignments are shown for human chimp, mouse, rat and dog. Genomic sequence coordinates for each region of homology are shown in brackets at the beginning and end of each sequence. Bases with sequence identity are denoted with an asterisk (*). Sites for GATA-1 (and other family members) in panels A, B, C and D and the site for the SCL erythroid transcriptional complex (panel D) are boxed in red. Sites for a variety of other transcription factors are boxed in black.

4.7 Performance of the human SCL genomic tiling array in detecting variations in nucleosome density in K562 cells

It has recently been demonstrated that the absence, or decreased levels, of nucleosomes (i.e., nucleosome depletion) may be a general feature of active genes in the yeast genome (Bernstein et al. 2004; Lee et al. 2004; Pokholok et al. 2005; Yuan et al. 2005) (see chapter 1). Given that (i) the average array element on the human SCL array was approximately 400-500 bp in length, (ii) the average nucleosome spacing in mammalian genomes is approximately 200 bp, and (iii) the SCL array platform was capable of reporting very subtle quantitative changes (see chapter 3, and Dhami et al. 2005), the SCL array should be capable of detecting differences in nucleosome density (i.e., a range of 0 to 3 nucleosomes per array amplicon). Therefore, the SCL tiling array was used to examine whether variations in nucleosome density existed at the human SCL locus.

4.7.1 Assessing the reproducibility of the array elements in ChIP-chip assay for histone H3

As a means of assaying for variations in nucleosome levels across the SCL locus, ChIP-chip experiments were performed in K562 cells using an antibody raised to histone H3. As seen with the ChIP-chip assays for histone H3 K9/14 diacetylation and GATA-1, the

SCL array elements showed reproducible variations in histone H3 enrichment across the SCL locus even within the narrow dynamic range required for this assay (Figure 4.12). A detailed description of the key features of this histone H3 profile is found in the subsequent section.

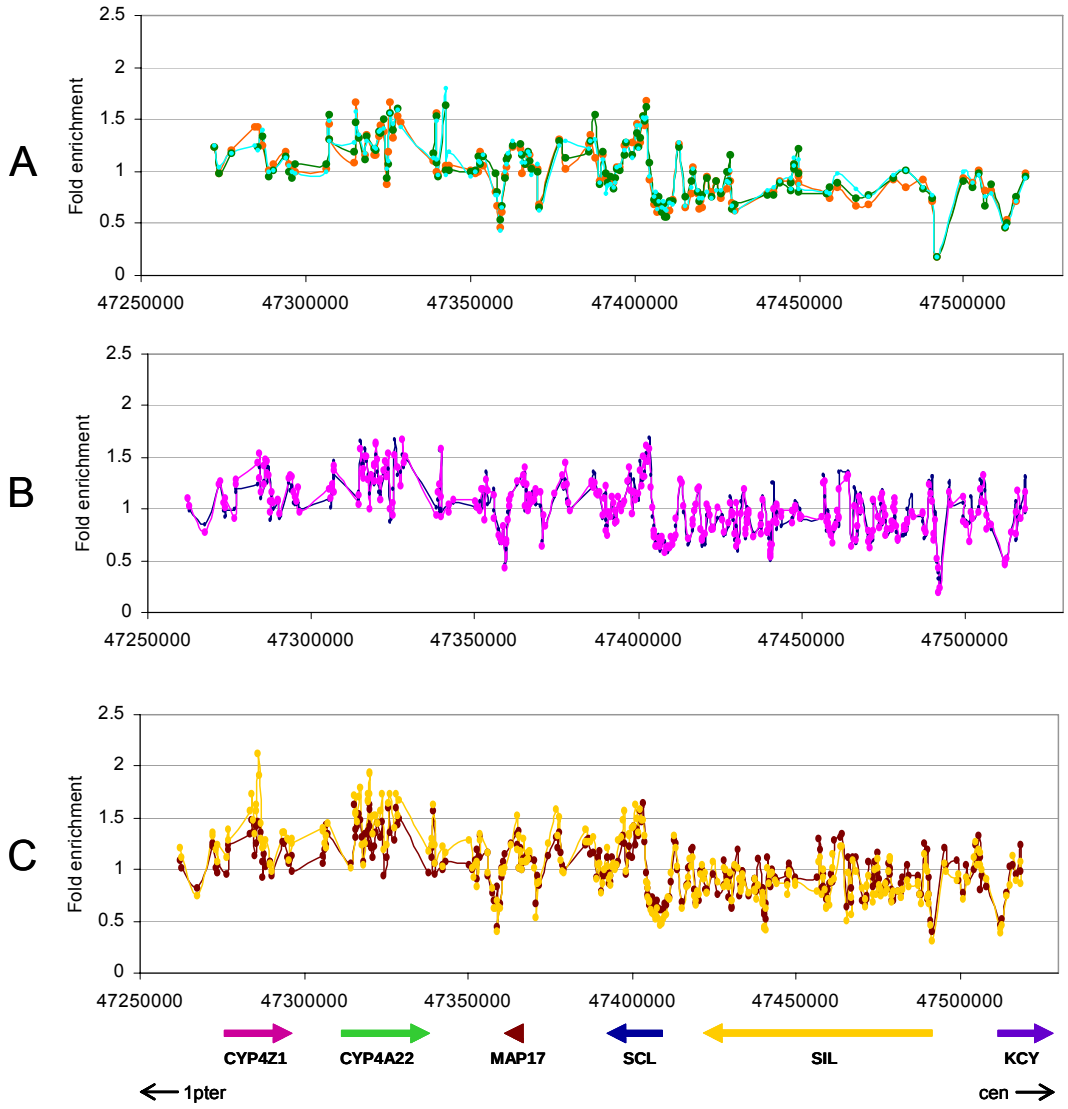


Figure 4.12: Performance of the SCL array platform in ChIP-chip assays for histone H3 in K562. Panel A: The orange (clone), turquoise (genomic) and green (genomic) profiles illustrate the reproducibility of the SCL array elements which were amplified from different PCR templates (Chapter 3). Panel B: The pink and blue profiles illustrate the reproducibility of the array elements in technical replicates from two independent hybridisations derived from using different aliquots of the same ChIP DNA. Panel C: The yellow and brown profiles illustrate the reproducibility of two independent biological replicates. Each profile in panel C represents the mean ratio of two technical replicates per biological replicate. The thick coloured arrows represent the gene order and direction of transcription as shown at the bottom of panel C. The y-axis represents fold enrichments in the ChIP-chip assays and the x-axis represents the genomic coordinates along human chromosome 1.

The reproducibility of these ChIP-chip assays can be summarized as follows:

1. The mean CV of ratios reported by the genomic sets and the clone set of array elements in a typical experiment was approximately 5% within a single hybridisation (assessed for 4 independent hybridisations).
2. The mean CV of ratios reported between corresponding array elements in technical replicate hybridisations was approximately 5% (assessed for each of two sets of 2 technical replicates)
3. The mean CV of ratios reported between corresponding array elements in three biological replicates was approximately 7% (assessed across 2 biological replicates).

4.7.2 Analysis of nucleosome density at the human SCL locus

Figure 4.13 shows a ChIP-chip profile for histone H3 representing the mean ratios of the array elements derived from the two biological replicates. Decreased relative levels of histone H3 were detected at, and close to, the promoter regions of the KCY, SIL and SCL genes and within the coding regions of SCL coincident with the broad peak of histone diacetylation found in this region. Similarly, the erythroid enhancer at +51 showed low levels of histone H3 immediately in the vicinity where substantial ChIP enrichments for GATA-1 had been detected. Decreased levels of histone H3 were also observed across the SIL gene coinciding with the coding regions of the gene (see analysis below) and at the SCL +39 region (the +39 region was also identified with H3 K9/14 diacetylation). In addition, there were noticeably elevated histone H3 levels on either side of some regulatory regions, especially at the SCL and SIL promoters. Such variations in H3 levels were taken to represent variations in nucleosome density across the region.

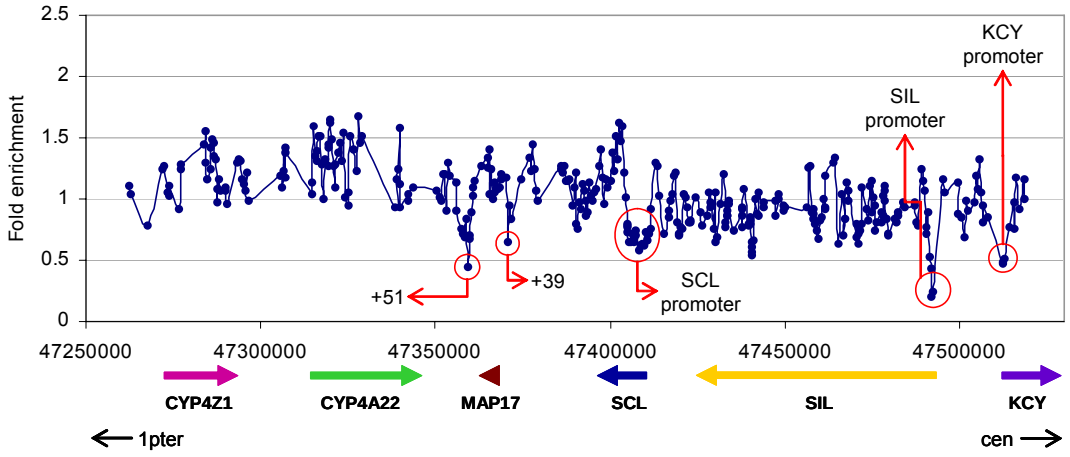


Figure 4.13: ChIP-chip profile for histone H3 across the human SCL locus in K562. Genomic regions showing depletion for histone H3 are marked with red circles. The y-axis represents the fold enrichments and the x-axis represents the genomic coordinates along human chromosome 1. The gene order and the

direction of transcription are shown at the bottom of the panel. The orientation of the SCL locus on chromosome 1p with regards to the centromere (cen) and telomere (ter) is shown at the bottom of the figure with black arrows.

To extend the analysis further, the relative levels of histone H3 across the SCL tiling path were ranked with respect to four types of DNA sequences: Type 1 – acetylated non-coding sequences; Type 2 – acetylated gene coding sequences; Type 3 – non-acetylated gene coding sequences; and Type 4 – neutral sequences (i.e., sequences with no known biological activity) (Table 4.2). It was observed that below the 40th percentile of histone enrichments, there was a noticeable preference for sequences of Type 1-3 (biologically active sequences) which were exclusively from the genes KCY, SIL and SCL (expressed in K562) and their flanking regulatory sequences. Moreover, the genomic region across the CYP4A22 and CYP4Z1 genes (which are not expressed in K562) showed a higher average level of histone H3 (1.31) than the genomic region across KCY, SIL and SCL (0.93). This difference was statistically significant to a 99.9% confidence level in a z test.

			FEATURES OF SEQUENCES FOUND IN ARRAY ELEMENTS				
Percentile Ranking of Array Elements	Relative Histone H3 Level per Array Element	Number of Array Elements	Acetylated Non-Coding Sequence (Type 1)	Acetylated Coding Sequence (Type 2)	Non-acetylated Coding Sequence (Type 3)	Biologically Active Sequences (Type 1-3)	Neutral Sequences
≤ 5th	0.353 – 0.597	17	8	5	4	17	0
≤ 10th	0.353 – 0.665	35	18	7	6	31	4
≤ 20th	0.353 – 0.770	69	29	8	11	48	21
≤ 30th	0.353 – 0.857	104	36	9	20	65	39
≤ 40th	0.353 – 0.918	138	46	9	25	80	58
≤ 50th	0.353 – 1.000	173	53	9	27	89	84

Table 4.2: Sequence features of SCL array elements ranked according to their levels of histone H3 in K562 cells. The table shows the sequence features (described in Results) for array elements ranked in percentile intervals according to their relative levels of histone H3 in ChIP-chip experiments. Only the 173 array elements ranked in the bottom 50% of the dataset are shown (i.e., those with the lowest relative levels of histone H3).

Taken together, this data provides evidence that the nucleosome density is variable across the SCL locus and nucleosome depletion occurs across active genes and their regulatory sequences.

4.7.3 Normalization using nucleosome density

The depletion of nucleosomes at discrete regions of the SCL locus could affect the interpretation of ChIP-chip data, particularly with respect to the location of histone modifications. For example, the absence of histone modifications may be due to

complete or partial absence of the nucleosomes themselves (Reinke and Horz 2003). Thus, the histone H3 K9/14 diacetylation plots which were described in section 4.5 of this chapter, were normalized with the corresponding histone H3 ChIP-chip profile described above to account for variations in nucleosome density (Figure 4.14).

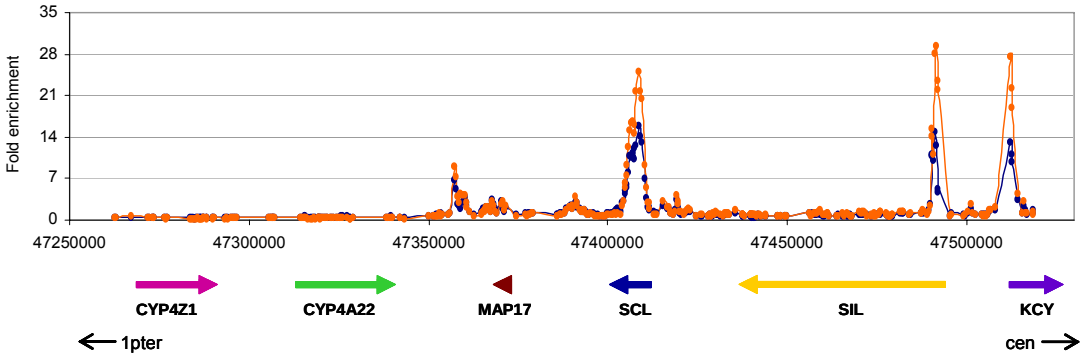


Figure 4.14: Normalization of the array data for H3 K9/14 diacetylation with respect to underlying nucleosome occupancy in K562. The blue and orange profiles represent un-normalized and normalized data respectively. After normalization, the peaks showing significant enrichments for H3 K9/14 diacetylation showed increased fold enrichments. The thick coloured arrows represent the gene order and direction of transcription as shown at the bottom of the panel. The x-axis represents the genomic co-ordinates along human chromosome 1 and the y-axis represents fold enrichments. The orientation of the SCL locus on human chromosome 1p with respect to the centromere (cen) and telomere (ter) is shown at the bottom with black arrows.

As a result of this normalization, no apparent changes were observed in the overall structure of the histone H3 K9/14 diacetylation profile obtained across the SCL locus in K562 cells. However, increased fold enrichments at the 5' ends of the transcriptionally active KCY, SIL and SCL genes were observed (Figure 4.14). The other regulatory regions across the SCL locus, especially the erythroid enhancer at the +51 region also showed increased enrichments for H3 K9/14 diacetylation after normalization. Thus, this novel normalization strategy may help better reflect the true relationship between histone modification levels and regulatory function across the SCL region.

4.8 Discussion

Described in this chapter is an assessment of the performance of the SCL genomic tiling path arrays to characterize the SCL locus with respect to a diverse range of biological activities which underlie fundamental aspects of mammalian gene regulation. The biological activities of histone modifications, transcription factors and nucleosomes density in the regulation of gene transcription are all well known – thus it was important to determine whether the SCL array platform was capable of detecting all three of these different types of DNA-protein interactions using ChIP-chip. The SCL array platform was shown to be highly sensitive and reproducible in these assays and was able to detect

and quantify both high- and low-level ChIP enrichments which could be confirmed with real-time PCR. This was attributed to the development of the highly sensitive single-stranded array platform (Dhami et al. 2005) used in combination with unamplified ChIP DNA in all the assays. The data described in this chapter (and in subsequent chapters in this thesis) provide strong evidence that the SCL genomic tiling path arrays are capable of identifying a wide range of regulatory interactions, and that the data obtained is of biological relevance to understanding the regulation of SCL and its flanking genes.

4.8.1 Histone H3 Acetylation at the SCL Locus

The histone H3 K9/14 diacetylation patterns which were derived for the human and mouse haematopoietic cell lines K562 and 416B respectively were in agreement with those described in a PCR-based analysis of a smaller genomic region across the SCL locus (Delabesse et al. 2005) and are consistent with the location of known regulatory elements for SCL and its flanking genes. In total, all but two known regulatory elements for SCL were identified by the histone H3 diacetylation profiles using the combined data from the two cell lines which were analysed. One of the regions not detected was the neural enhancer at +23 (Gottgens et al. 2000) which is not expressed in haematopoietic cells. The other region which could not be detected was the tal-RE silencer at SCL+14, which is active in K562 (Courtes et al. 2000). The tal-RE silencer, however, may not be detectable by H3 diacetylation profiling, given that repressor sequences are known to recruit deacetylase complexes (Burke and Baniahmad 2000).

The association of histone acetylation and the 5' ends of active genes is well documented (Liang et al. 2004; Bernstein et al. 2005; Pokholok et al. 2005; Roh et al. 2005) and agrees with the data shown here. Prominent enrichments of acetylation were also detected in the coding regions of SCL – which may reflect the involvement of acetylation in the activity of a number of known regulatory elements within the 5' half of the SCL gene. This, in combination with acetylation levels at the various SCL enhancer regions, suggests that acetylation is also a hallmark for other types of regulatory sequences; however, their level of acetylation may generally be less marked than acetylation at the 5' ends of genes.

It was also established that for sequences which are enriched for histone H3 acetylation at the SCL locus, there is a strong correlation between their acetylation level and the level of human-mouse sequence conservation in non-coding DNA. Yet, high sequence conservation *per se* is not a hallmark of only those sequences which are acetylated – an observation previously seen for histone H3 methylation patterns in human cell lines (Bernstein et al. 2005). However, while this may be interpreted that histone modifications

may occur at sites distant from conserved sequences involved in regulatory activity (Bernstein et al. 2005), this would be incongruous with the H3 acetylation patterns described in this chapter which clearly define regions known to have regulatory function. One interpretation is that some highly conserved sequences may define regulatory elements not active, and therefore not acetylated, in the cell types examined here. Alternatively, (i) some highly conserved regulatory sequences may not use histone H3 acetylation as a mechanism to modulate regulatory activity, (ii) not all conserved non-coding sequences may have transcriptional regulatory roles, or (iii) the relationship between H3 acetylation, regulatory function and sequence conservation at the SCL locus may be somewhat unique.

4.8.2 GATA-1 and SCL Regulation

Highly reproducible ChIP-chip profiles on the human SCL array for the GATA-1 transcription factor, in combination with real-time PCR data and comparative sequence analysis, suggests *bona fide* binding events for GATA-1 have been identified at the SCL locus in the human cell line K562. The SCL array also included three genomic array elements for the human β-globin sequences which had previously been shown to bind GATA-1 in K562 (Horak et al. 2002). The highest level of enrichment obtained in the present study was 2.2-fold as compared to 2 to 4-fold in the previous study. The differences in the enrichment for these regions seen in the two studies may be related to a number of factors. For example, the Horak study used a number of different antibodies in addition to the antibody used here. Also, the standard deviations between technical replicates in the Horak study were very large; suggesting that the fold changes reported may not be accurate (Horak et al. 2002).

The identification of GATA-1 binding sites at the SCL locus in K562 provides an insight into the interactions of *trans*-acting factors with the *cis*-acting elements which control SCL expression. The high levels of GATA-1 enrichment at the +51 erythroid enhancer provide the first evidence that GATA-1 may be involved in the activities of this enhancer during erythroid development. Sequence analysis also identified highly conserved binding sites for the erythroid SCL transcriptional complex (Wadman et al. 1997) in the +51 sequence. This finding suggests that GATA-1 may co-operate with SCL itself during erythrocyte development in a self-regulatory role (see chapter 5). The GATA-1 enrichments seen over the SCL promoter 1a is also in complete agreement with previous observations that suggests that one GATA site within the promoter 1a is required for the activity of this promoter in erythroid cells (Bockamp et al. 1995). Similarly, the GATA-1 enrichments seen at +3, -7 and -9/-10 regions are also in complete agreement with known DNase I hypersensitive data in K562 (Leroy-Viard et al. 1994) and the histone H3

acetylation data seen with the ChIP-chip assays. Taken together, these data suggest that these regions are acetylated and therefore in open chromatin which makes these regions accessible to transcription factor binding. However, the levels of enrichment seen over these regions were lower as compared to the +51 erythroid enhancer which may suggest a lower occupancy of GATA-1 at these sites, than at the erythroid enhancer. Whether this relates to, i) different binding affinities of GATA-1 for these sites, ii) epitope accessibility of GATA-1 at these sites, or iii) the proportion of K562 cells which show GATA-1 occupancy at these sites, is not known.

4.8.3 Nucleosome Density at the SCL Locus

ChIP-chip analysis of the SCL locus for histone H3 in K562 provides compelling evidence that nucleosome depletion occurs at the active mammalian genes, particularly at the non-coding regulatory sequences and coding regions of SCL and its flanking genes. Nucleosome depletion was first demonstrated in yeast (Bernstein et al. 2004; Lee et al. 2004; Pokholok et al. 2005; Yuan et al. 2005), and the data presented here suggests that nucleosome depletion may be conserved throughout evolution. Whether nucleosome depletion is a general feature of all active genes in the human genome is not known and previous studies have shown no evidence for nucleosome depletion in other regions of human genome using ChIP-chip assays (Bernstein et al. 2005). However, this latter finding may also reflect sensitivity differences in the array platforms used in this and the previous study.

The mechanism by which nucleosome depletion occurs is still not clear and two models have been proposed. The first invokes a sliding mechanism whereby the histone octamer moves along the DNA resulting in repositioning the nucleosome (Hamiche et al. 1999; Korber et al.2004). The second results in the complete release or eviction of the nucleosome and rebinding at a new location (Lorch et al. 1999). Given that histone H3 levels were elevated on either side of some regulatory regions at the SCL locus, nucleosomes may not be lost completely but displaced. This observation is consistent with the sliding mechanism of nucleosome depletion.

Recent studies in *Drosophila* have suggested that nucleosome depletion is transient and is compensated for by deposition of histone variant H3.3 (Wirbelauer et al. 2005). The depletions presented in this chapter, took into account both H3 and H3.3 variants since the antibody used in this study is raised to the extreme C- terminus of the histone and does not distinguish between them. This suggests that the mechanisms of nucleosome dynamics may be different between mammals and *Drosophila*.

4.8.4 Conclusions

The work described in this chapter established the high reproducibility and sensitivity of the SCL array platform in ChIP-chip assays. Furthermore, neither nucleosome depletion nor relationships between sequence conservation and histone modifications has previously been observed in mammals, suggesting that this ChIP-on-chip system is capable of providing new insights to widely applicable aspects of gene regulation. In the subsequent chapters, the use of the SCL array platform to understand regulatory events at the SCL locus will be further explored.

Chapter 5

Application of ChIP-chip to elucidate a Comprehensive Set of Regulatory Interactions at the SCL locus

5.1 Introduction

In order to fully understand how SCL is regulated in its biological contexts, it is important to not only identify all the key regulatory elements, but also the full complement of regulatory interactions that control their activity *in vivo*. Although a number of key regulatory elements at the SCL locus are already known, it is still far from clear how these elements interact with each other or with other *cis*- and *trans*-acting factors that play roles in regulating their activity. The location, function and known regulators of these elements have been discussed in chapters 1 and 4 of this thesis. To summarize:

i) **Promoters:** SCL has three promoters p^{1a}, p^{1b} (Aplan et al. 1990a) and p^{EXON4} (Bernard et al. 1992) all of which are sites for histone H3 and H4 acetylation (Delabesse et al. 2005; Chapter 4); p^{1a} is active only in erythroid and megakaryocytic lineages; p^{1b} is active in erythroid and megakaryocytic lineages and in leukaemic T-cells (Bernard et al. 1992; Aplan et al. 1990a); p^{EXON4} is active only in leukaemic T cells (Bernard et al. 1992). It has been shown that GATA-1 is involved in the regulation of p^{1a} in erythroid cells in both human and mouse (Bockamp et al. 1995; Lecointe et al. 1994; Chapter 4).

ii) **Endothelial enhancer:** The endothelial enhancer (at -3/-4 in both human and mouse) is functional in vascular endothelium and haematopoietic progenitors, is associated with histone H3 and H4 acetylation (Delabesse et al. 2005; Chapter 4) and is regulated by Fli-1 and Elf-1 (Sinclair et al. 1999; Gottgens et al. 2004).

iii) **Stem cell enhancer:** The stem cell enhancer (at +18/+19 in mouse and at +20/+21 region in human) directs SCL expression to the vast majority of haematopoietic progenitors and endothelium. It is associated with histone H3 and H4 acetylation (Delabesse et al. 2005; Chapter 4) and is regulated by GATA-2, Fli-1 and Elf-1 (Sanchez et al. 1999; Sanchez et al. 2001; Gottgens et al. 2002).

iv) **Neural Regulators:** Regulatory elements located at SCL p^{1a}, +1, +3 and at +23 region in mouse (SCL p^{1a}, +1, +3 and +23/+24 regions in human) direct SCL expression to specific regions within the brain and spinal cord. Most, but not all, of these are known to be associated with histone H3 and H4 acetylation (Delabesse et

al. 2005; Chapter 4). GATA factors may play roles in regulating some of these elements (Sinclair et al. 1999; Gottgens et al. 2000; Chapter 4).

v) **Erythroid enhancer:** The +40 erythroid enhancer (+51 in human) targets primitive erythroblasts in transgenic mice (Delabesse et al. 2005). It is associated with histone H3 and H4 acetylation (Delabesse et al. 2005; Chapter 4) and the transcription factor GATA-1 binds to this enhancer (Chapter 4).

vi) **SCL p^{EXON4} repressor:** A novel silencer element, tal-RE, located at +14 in human (+13 in mouse) represses the activity of SCL p^{EXON4} to ensure its restricted usage in only T-cells (Courtes et al. 2000). The regulators which bind to tal-RE are not known.

vii) **-8/-9 enhancer:** A region at -8/-9 in mouse (-9/-10 in human) shows enhancer activity in reporter assays (Gottgens et al. 1997) and is a site for histone H3 acetylation (Delabesse et al. 2005; Chapter 4). GATA-1 binds to this region in K562 cells (Chapter 4).

viii) **-7 element:** In chapter 4 of this thesis, a putative regulatory element at -7 (-6 in mouse) was described which binds GATA-1 and is a site for histone H3 acetylation in K562 (Chapter 4) and other haematopoietic cell types (Delabesse et al 2005).

It is now becoming increasingly clear that posttranslational modifications of histones regulate chromatin structure which in turn determines how a particular gene is regulated by various regulatory proteins and the transcriptional machinery (also discussed in chapter 1). Histone modifications can be highly reversible, such as acetylation (on lysine residues) and phosphorylation (on serine and threonine residues), or more stable, such as methylation (on lysine and arginine residues) (Fischle et al. 2003; Lachner et al. 2002) (Figure 5.1). The mechanisms by which histone acetylation and deacetylation affect transcription and other DNA-based processes are thought to involve two major pathways: first, histone acetylation may alter the folding properties of the chromatin fibre, thereby modulating the accessibility of DNA through structural changes (Tse et al. 1998); and second, the lysine residues and their modifications also provide specific binding surfaces for the recruitment of repressors and activators of gene activity.

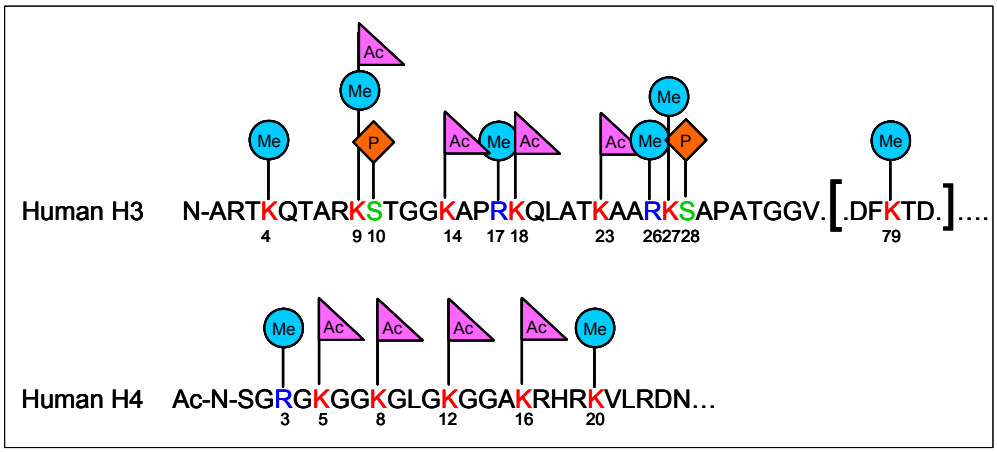


Figure 5.1: The N-terminal tails of histones H3 and H4. The N-terminal tails of H3 and H4 contain highly conserved lysines; K9, K14, K18, K23, K27 and K5, K8, K12 K16 respectively. The blue circles represent methylation, pink triangles represent acetylation and orange diamonds represent phosphorylation marks on various residues of H3 and H4 tails. The lysine residues in the N-terminal tail of H3 that can be methylated include K4, K9, K27 and K36. Lysine K20 can be methylated in H4. In addition, several arginine residues in H3 and H4 can be methylated. Another site of methylation occurs in the histone-fold domain of histone H3 at K79 (Feng et al. 2002; Ng et al. 2002; van Leeuwen et al. 2002). H3 K9 can be acetylated as well as methylated. Lysine residues can be mono-, di- or trimethylated. In addition serine residues at positions 10 and 28 in histone H3 can be phosphorylated.

A number of studies have provided insight into the functions of specific histone modifications. For instance, acetylation at lysine residues is generally associated with transcriptional activity (Wu and Grunstein 2000; Roh et al. 2004). In contrast, methylation of histone tail lysines and arginines has been linked to activation and repression, depending on the residue that is modified (Kouzarides 2002). A number of histone modification enzymes such as acetyltransferases (HATs), deacetylases (HDACs), methyltransferases (HMTs), demethylases and kinases have been identified that exhibit their activity on specific residues of histone H3 or H4. However, the information regarding the patterns of histone modifications at a given gene or entire genome have only begun to be elucidated (Agalioti et al. 2002; Schübeler et al. 2004; Bernstein et al. 2005; Liu et al. 2005). In order to fully comprehend the functional consequences of histone modifications on the transcriptional activity of a gene, it is important to identify the distribution of all of these modifications at a given gene. More specifically, given the development of a highly sensitive ChIP-chip platform for the SCL locus (Chapter 4), a detailed survey of the SCL genomic region to determine the distribution of histone modifications, and binding of transcription factors and other regulatory proteins would be very powerful in answering these questions.

5.2 Aims of this chapter

The aims of the work reported in this chapter were:

1. To identify and test antibodies, raised against various histone H3 and H4 modifications (i.e., acetylation, methylation and phosphorylation) using ChIP-chip assays across the SCL locus in the K562 cell line. Working assays were then used to map regulatory interactions across the SCL locus in SCL expressing and SCL non-expressing cell lines.
2. To further analyse the SCL locus with respect to histone H3 levels to determine the underlying nucleosome occupancy and relationships to gene activity in SCL expressing and SCL non-expressing cell lines.
3. To identify interactions of various transcription factors and other regulatory proteins across the SCL locus in SCL expressing and SCL non-expressing cell lines.

5.3 Overall strategy

This chapter describes the use of ChIP-chip assays to survey DNA-protein interactions across the SCL locus in SCL expressing (K562, Jurkat) and non-expressing (HL60, HPB-ALL) cell lines. The strategy used for these cell lines is outlined in Figure 5.2. The first step was to identify a range of antibodies raised to various histone modifications, transcription factors and other regulatory proteins (Figure 5.2). These were then tested in ChIP-chip experiments to distinguish working assays from non-working ones in K562. The antibodies which reported significant ChIP enrichments in K562 were then further used in the other three cell lines to construct maps of DNA-protein interactions across the SCL locus. Based on this work, it was possible to compare these cell lines with respect to regulatory interactions across the SCL locus, and correlate the findings with the expression of SCL and its flanking genes.

Although, the focus of the study presented in this thesis was to elucidate regulatory interactions at the SCL locus in human haematopoietic cell lines, the availability of the mouse SCL genomic tiling path array and relevant cell lines made it possible to perform some of the assays in mouse. These experiments are also described in this chapter.

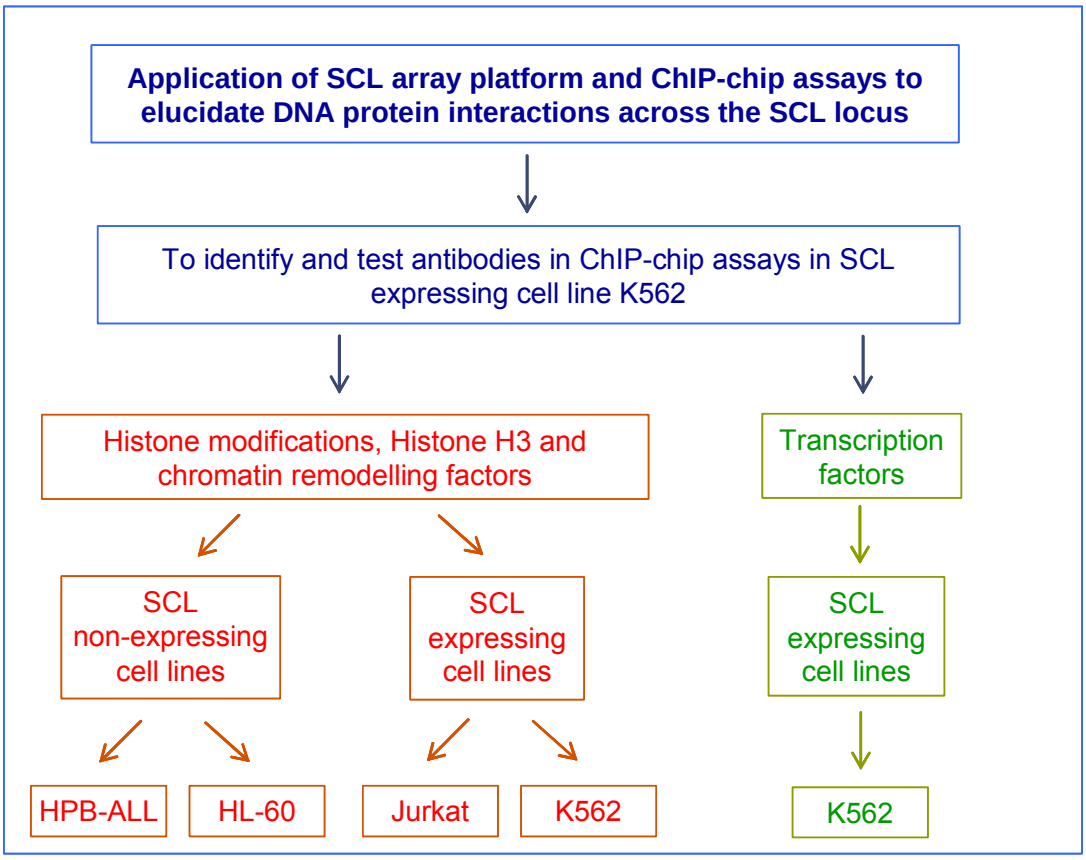


Figure 5.2: Overall strategy employed to map regulatory interactions at the SCL locus using ChIP-chip. This flow diagram shows an overview of the strategy employed to study interactions in four human haematopoietic cell lines. In addition, a sub-set of experiments were performed in mouse cell lines.

Results

5.4 Identifying and testing antibodies to use in ChIP-chip assays

In order to provide a detailed characterization of regulatory interactions at the SCL locus, it was important to identify a host of antibodies which work well in ChIP-chip, the assays for which reflect a range of biological activities involved in transcriptional regulation. The human haematopoietic cell line, K562, was selected to test all the antibodies in the ChIP-chip assays for several reasons. First, K562 had already been used successfully in the ChIP-chip assays described in chapter 4. Second, most of the human SCL regulatory elements studied previously had been characterized using K562 and were shown to be active in this cell line (Aplan et al. 1990a; Leroy-Viard et al. 1994; Courtes et al. 2000). Lastly, it is a well established cell line for the study of haematopoietic development (Horak et al. 2002).

Table 5.1 lists all the antibodies which were identified and tested in ChIP-chip assays across the SCL locus. The antibodies were selected based on a number of criteria as follows:

i) the antibody was not raised to the DNA binding domain of the protein; if possible the antibody was raised to either the N- or C-terminus of the protein

ii) the specificity of the antibody was supported by western blot evidence

iii) the antibody was certified as ChIP grade antibodies by the vendor

iv) the antibody had been shown to perform well in previously published studies

All the antibodies used in the work presented for this thesis had to fulfill at least one of the above-listed criteria. In many instances, antibodies fulfilled all or most of the criteria. To determine whether the antibodies performed well in ChIP-chip experiments in K562, significant ChIP enrichments were calculated for each assay using the statistical analysis described in chapter 4 (section 4.4). Antibodies used in ChIP-chip which gave enrichments which were three standard deviations above background levels were considered as working assays. However, datasets from ChIP-chip assays which measured histone levels were not analysed in this way, given that it was not possible to determine what constituted background levels (i.e., all genomic regions enrich for histones). In these instances, profiles were examined to determine if they differed from profiles obtained from control IgG experiments (using antisera from the appropriate host species in ChIP assays).

Table 5.1. List of antibodies used in ChIP-chip assays. The selected antibodies were raised to various histone modifications, transcription factors and other regulator proteins used in the ChIP-chip experiments. The human and mouse cell lines used in the ChIP-chip experiments are listed in the last 6 columns. All the listed antibodies were first tested in K562 and are colour coded for working (bright pink) and non-working (blue) assays. The criteria for working assays was based on visual assessment of profiles and/or statistical analysis as described in chapter 4 (also see section 5.4). The antibodies that worked successfully in K562 were then further used in other cell lines and are colour coded by light pink.

Classification	Histone	Modification	Function	Company	Cat. No.	Reference	Cell Line					
							HUMAN				MOUSE	
							K562	Jurkat	HL60	HPB-ALL	416B	E14_ES
Acetylation	H3	Lys9/14	gene activation	upstate	06-599	Litt et al. 2001; Roh et al. 2004						
		Lys9	gene activation	upstate	07-352	Litt et al. 2001; Roh et al. 2004						
		Lys14	gene activation	upstate	07-353	Suka et al. 2001; Kurdistani et al. 2004						
		Lys14	gene activation	abcam	ab2381	Suka et al. 2001; Kurdistani et al. 2004						
		Lys18	gene activation	upstate	07-354	Kurdistani et al. 2004						
		Lys23	gene activation	upstate	07-355	Suka et al. 2001; Kurdistani et al. 2004						
		Lys27	gene activation	upstate	07-360	Suka et al. 2001; Kurdistani et al. 2004						
	H4	Lys5/8/12/16	transcriptional regulation	upstate	06-866	Litt et al. 2001; Roh et al. 2004						
		Lys5/8/12/16	transcriptional regulation	abcam	ab76	Litt et al. 2001; Roh et al. 2004						
		Lys5	gene activation	upstate	07-327	Suka et al. 2001; Litt et al. 2001						
		Lys5	gene activation	abcam	ab1758	Suka et al. 2001; Litt et al. 2001						
		Lys8	transcriptional regulation	upstate	07-328	Agalioti et al. 2002; Suka et al.2001						
		Lys8	transcriptional regulation	abcam	ab1760	Agalioti et al. 2002; Suka et al.2001						
		Lys12	transcriptional regulation	upstate	07-323	Agalioti et al. 2002; Suka et al.2001						
		Lys12	transcriptional regulation	abcam	ab1761	Agalioti et al. 2002; Suka et al.2001						
		Lys16	transcriptional regulation	upstate	07-329	Dion et al. 2005; Suka et al. 2001						
		Lys16	transcriptional regulation	abcam	ab1762	Dion et al. 2005; Suka et al. 2001						
	H2A	Lys5/9/13/15	gene activation	upstate	07-376	Suka et al. 2001						
	H2B	Lys5/12/15/20	gene activation	upstate	07-373	Myers et al. 2003						
Methylation	H3	Lys4 mono	gene silencing	abcam	ab8895	Pokholok et al. 2005; Liu et al. 2005						
		Lys4 di	poised chromatin	upstate	07-030	Santos-Rosa et al. 2002; Liu et al. 2005						
		Lys4 di	poised chromatin	abcam	ab7766	Santos-Rosa et al. 2002; Liu et al. 2005						
		Lys4 tri	gene activation	abcam	ab8580	Santos-Rosa et al. 2002; Liu et al. 2005						
		Lys9 di	gene silencing	upstate	07-212	Lachner et al. 2003; Litt et al. 2001						
		Lys27 di	gene silencing	upstate	07-421	Lachner et al. 2003; Cao et al. 2002						
		Lys79 tri	gene activation	abcam	ab2621	Pokholok et al. 2005						
		Arg17 di	transcriptional activation	upstate	07-214	Chen et al. 1999; Daujat et al. 2002						
		Arg26 di	transcriptional regulation	upstate	07-215	Chen et al. 1999						
	H4	Lys20 di	cell-cycle regulation	upstate-a	07-367	Fang et al. 2002.						
		Lys20 di	cell-cycle regulation	upstate-b	07-031	Fang et al. 2002.						
		Arg3 di	gene activation	upstate	07-213	Strahl et al. 2001						
Phosporylation	H3	Ser 10	gene activation/mitotic condensation	upstate	05-598	Mahadevan et al. 1991						
		Ser 10	gene activation/mitotic condensation	abcam	ab4442	Mahadevan et al. 1991						
		Ser 28	cell division/mitotic condensation	upstate	07-145	Loury et al. 2003						

Continued in next page

Classification	Histone	Modification	Function			Reference	Cell Line					
							HUMAN				MOUSE	
							K562	Jurkat	HL60	HPB-ALL	416B	E14_ES
Histones	H3	H3	core histone	santa cruz	sc-8654	Pokholok et al. 2005						
		H3	core histone	abcam	ab1791	Pokholok et al. 2005						
		H3.3	histone H3 variant	abcam	ab4263	Ahmad and Henikoff 2002						
	H4	H4 (C-20)	core histone	santa cruz	sc-8658	Pokholok et al. 2005						
		H4	core histone	abcam	ab2423	Pokholok et al. 2005						
	H2A	H2A (C-19)	core histone	santa cruz	sc8648	Vicent et al. 2004						
	H2B	H2B (C-19)	core histone	santa cruz	sc8651	Vicent et al. 2004						
Chromatin remodellers	p300 (C-20)		HAT activity	santa cruz	sc585	Agalioti et al. 2002						
	p300		HAT activity	abcam	ab2830	Agalioti et al. 2002						
	CBP (A-22)		HAT activity	santa cruz	sc369	Agalioti et al. 2002; Bannister et al. 1996						
	CBP		HAT activity	abcam	ab2832	Agalioti et al. 2002; Bannister et al. 1996						
	GCN5 (C-16)		HAT activity	santa cruz	sc6302	Agalioti et al. 2002						
	PCAF		HAT activity	upstate	07-141	Blanco et al. 1998						
	CGBP (N-21)		Early mammalian development	santa cruz	sc12336	Carlone et al. 2001						
	HDAC1 (C-19)		HDAC activitycell growth/proliferation	santa cruz	sc6298	Lagger et al. 2002						
	HDAC2 (C-19)		HDAC activity	santa cruz	sc6296	De Ruijter et al. 2003						
	MeCP2		gene repression	upstate	07-013	Fan et al. 2005						
	CARM1		arginine methylation	upstate	07-080	Chen et al. 1999						
Preinitiation complex	TAF II p250 (L-20)		Preinitiation complex/HAT activity	santa cruz	sc17134	Ruppert et al. 1995						
	TAF II p250 (6B3)		Preinitiation complex/HAT activity	santa cruz	sc735	Ruppert et al. 1995						
	RNA pol II		Trancription initiation/elongation	abcam	ab5408	Kim et al. 2005						
Trancription factors	Elf-1 (C-20)			santa cruz	sc631	Gottgens et al. 2002; Gottgens et al 2004						
	Fli-1 (C-19)			santa cruz	sc356	Gottgens et al. 2002; Gottgens et al 2004						
	GATA-1 (N-6)			santa cruz	sc265	Wadman et al.1997; Horak et al. 2002						
	GATA-1 (M-20)			santa cruz	sc1234X	Wadman et al.1997; Horak et al. 2002						
	GATA-2 (H116)			santa cruz	sc9008	Gottgens et al. 2002						
	GATA-2 (F-20)			santa cruz	sc16044	Gottgens et al. 2002						
	TAL1			Mathieu lab, France		Wadman et al.1997; Vitelli et al. 2000						
	E2A (YAE)			santa cruz	sc416X	Wadman et al. 1997; Vitelli et al. 2000						
	LMO2 (L-17)			santa cruz	sc10498	Wadman et al. 1997; Vitelli et al. 2000						
	LDB1 (N-18)			santa cruz	sc11198X	Wadman et al. 1997; Vitelli et al. 2000						
	Sp-1 (PEP-2)			santa cruz	sc59X	Lahlil et al. 2004						
	CTCF			abcam	ab10571	Torrano et al. 2005						

5.5 Constructing maps of *in vivo* DNA-protein interactions at the SCL locus

ChIP assays and microarray hybridisations were performed as described in chapter 2. The ChIP DNAs obtained in all the assays described in this chapter were not amplified prior to hybridisation. In order to be able to obtain data for a large number of interactions, only one biological ChIP replicate was performed for each working assay; from the resultant ChIP DNA only one hybridisation was performed. However, given the high reproducibility of the SCL genomic tiling path array in ChIP-chip assays (see chapter 4), it would still be possible to obtain biologically meaningful data using this approach. All the ChIP-chip profiles which were generated for histone modifications in human and mouse cell lines have been compiled in Appendices 6-11. The ChIP-chip profiles described in this chapter and shown in the appendices have been annotated showing the regions with significant enrichments; the numbering system is based on their distance upstream (-) or downstream (+) in kilobases from the SCL promoter 1a (p^{1a}) or from the promoter of the nearest gene.

5.6 Histone H3 acetylation patterns at the SCL locus in SCL expressing and non-expressing cell lines

5.6.1 Histone H3 K9/14 diacetylation correlates with active genes in human cell lines

5.6.1.1 H3 9/14 diacetylation in SCL expressing cell lines

In K562, enrichments for histone H3 K9/14 diacetylation was associated with regions of known regulatory activity across the SCL locus (Chapter 4 and shown in Figure 5.3). Similarly, in Jurkat, the most prominent enrichments were found at or near the 5' ends (i.e. promoters) of KCY, SIL and SCL genes, all of which are expressed in this cell line (Figure 5.3). Acetylation across the 5' end of SCL extended almost 10 kb into its coding regions encompassing SCL p^{1a}, p^{1b}, +1, +3 and SCL p^{EXON4} at the +7 region, the latter of which is known to be active in Jurkat (Bernard et al. 1992). Upstream of SCL, significant enrichments were obtained at -7 and -4 regions, both of which are known to exhibit DNase I hypersensitivity (Leroy-Viard 1994). In addition, significant enrichments were seen at +39 and at p^{MAP17} (+42). Enrichments were low over the +51 erythroid enhancer, but increased over the +53 region. Low level enrichments were also identified at +57. This region was not found to be enriched in K562 and the function of this region is not known. No significant enrichments were observed across CYP4A22 and CYP4Z1 genes.

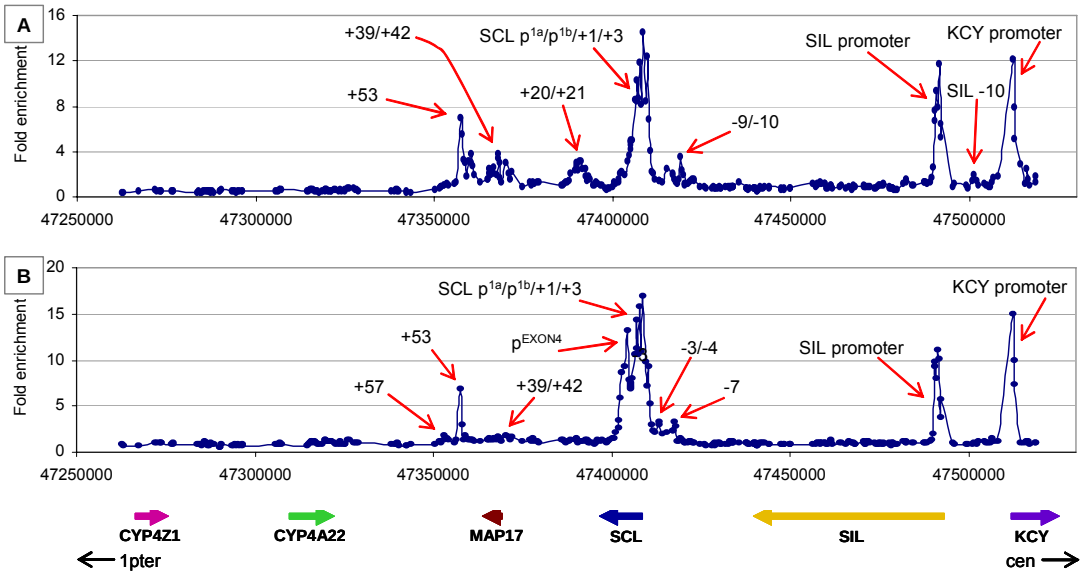


Figure 5.3: ChIP-chip profiles for H3 K9/14 diacetylation across the SCL locus in K562 and Jurkat. Panel A: K562 (as described in chapter 4), Panel B: Jurkat. The x-axes represent the genomic coordinates along human chromosome 1 and the y-axes represent the fold enrichments. Genomic regions showing significant enrichments for H3 acetylation are marked with red arrows. The genomic regions denoted by +1, -3 etc. are based on their distances upstream (-) or downstream (+) in kilobases from the human SCL p^{1a}. The thick, coloured arrows at the bottom of panel B represent the gene order and the direction of transcription. The orientation of the locus with respect to the centromere (cen) and the telomere (ter) on human chromosome 1 is shown by black arrows at the bottom of the figure.

5.6.1.2 H3 K9/14 diacetylation in SCL non-expressing cell lines

Figure 5.4 shows the ChIP-chip profiles generated for H3 K9/14 diacetylation in HL60 and HPB-ALL. As observed in K562 and Jurkat, high levels of enrichments were seen at or near the 5' ends (promoters) of the KCY and SIL genes, both of which are expressed in these two cell lines. At the 5' end of the SCL gene, no significant enrichments were obtained in HPB-ALL, but in HL60, subtle but significant enrichments were present at the +1 region. These absent or greatly reduced levels of acetylation at the 5' end of SCL are in agreement with the absence of SCL expression in these cell lines. Enrichments were also observed at +39 in both cell lines, and at p^{MAP17} in HL60. Furthermore, although neither HPB-ALL nor HL60 express SCL, a region near the +51 erythroid enhancer, at +53, showed high levels of H3 K9/14 diacetylation. This suggests that there may be some genomic sequences in this region, which are active even in the absence of SCL expression. These sequences will be discussed in detail in chapter 7.

In HPB-ALL, a genomic region approximately 4-5 kb downstream of the SIL promoter (SIL +4/+5) also reported significant enrichments. No significant enrichments were observed across the CYP4A22 and CYP4Z1 genes.

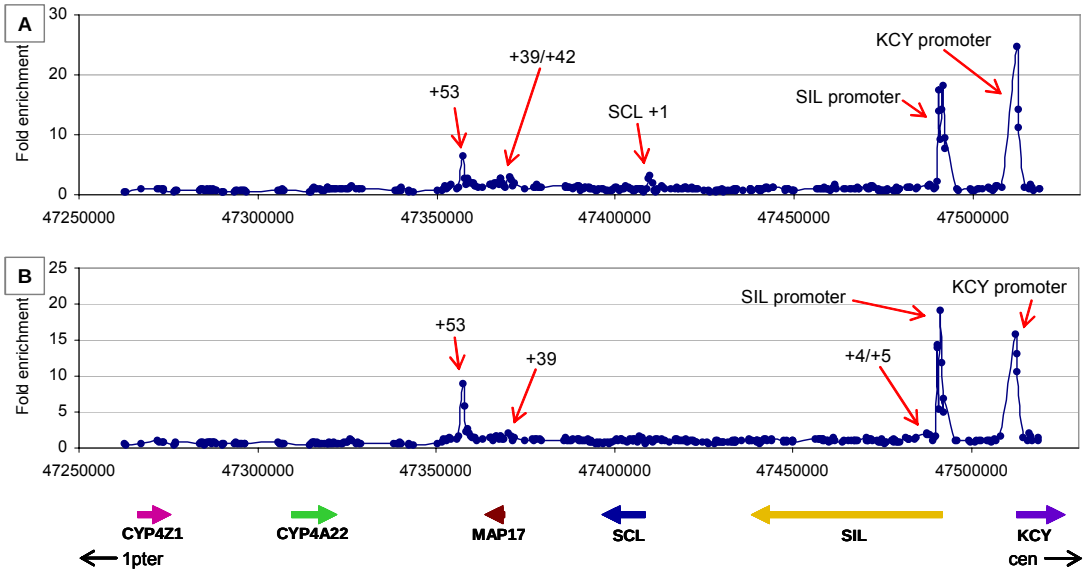


Figure 5.4: ChIP-chip profiles for H3 K9/14 diacetylation across the SCL locus in HL60 and HPB-ALL. Panel A: HL60, Panel B: HPB-ALL. Genomic regions showing significant enrichments for H3 K9/14 diacetylation are marked with red arrows. The x-axes represent the genomic coordinates along human chromosome 1 and the y-axes represent fold enrichments. The genomic regions denoted by +1, +39 etc. are based on their distances downstream (+) in kilobases from the human SCL p^{1a}. The thick, coloured arrows at the bottom of panel B represent the gene order and the direction of transcription. The orientation of the locus with respect to the centromere (cen) and telomere (ter) on human chromosome 1 is shown with black arrows at the bottom of the figure.

5.6.2 Histone H3 K9/14 diacetylation is present at SCL regulatory regions in mouse ES cells which do not express SCL

H3 K9/14 diacetylation profiles were determined for the mouse ES cell line E14 (Hooper et al. 1987) and compared to that obtained for the mouse cell line 416B (Chapter 4) as shown in Figure 5.5. 416B expresses SCL, while E14 does not – thus, analysis of these cell lines provides an interesting parallel to the analysis that was performed on the human SCL-expressing and non-expressing cell lines described in this chapter.

Both cell lines showed significant enrichments at or near the 5' end of SIL gene - which is expressed in both 416B and E14 ES cells. However, both cell lines also showed enrichments for H3 K9/14 diacetylation at SCL regulatory regions irrespective of the fact that the E14 cell line does not express SCL. These acetylated SCL regulatory regions encompassed sequences which are known to show DNase I hypersensitivity, enhancer or promoter activity (+1, +3 and SCL p^{EXON4} at +7). The presence of enrichments for H3 K9/14 acetylation at SCL regulatory regions in E14 ES cells suggests that substantial H3 acetylation can occur at genes in the absence of their expression in ES cell lines.

Although this result was incongruous with previous findings which suggest H3 acetylation is associated with gene activity (Schübeler et al. 2004; Bernstein et al. 2005; Liu et al. 2005), there were differences in the profiles obtained for these two mouse cell lines with respect to SCL regulatory regions. First, levels of enrichments at these regions were higher in the SCL-expressing cell line 416B. Second, in 416B a continuous block of acetylation at the 5' end of the SCL gene spanned the genomic regions starting at -1 and extended almost 12 kb into the coding region. In E14 ES cells, the enrichments over these regions were seen as discrete blocks.

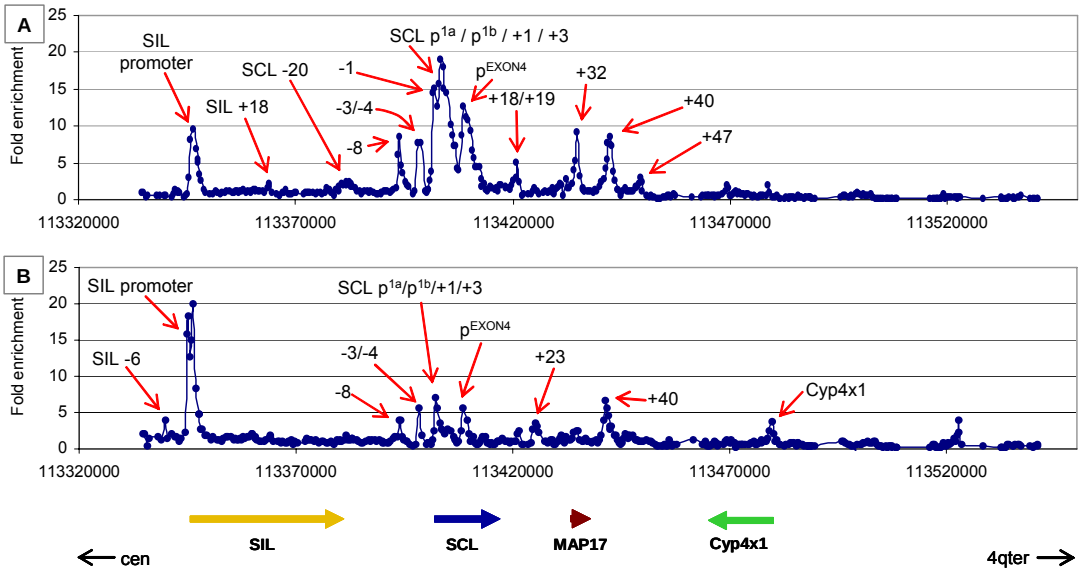


Figure 5.5: ChIP-chip profiles for histone H3 K9/14 diacetylation across SCL locus in 416B and mouse E14 ES cell lines. Panel A: 416B, panel B: mouse E14 ES cell line. Genomic regions showing significant enrichments are marked with red arrows. The x-axes represent genomic coordinates along mouse chromosome 4 and the y-axes represent fold enrichments. The genomic regions denoted by -1, +1 etc. are based on their distance upstream (-) or downstream (+) in kilobases from the mouse SCL p^{1a}. The thick, coloured arrows at the bottom of panel B represent the gene order and the direction of transcription. The orientation of the locus with respect to the centromere (cen) and telomere (ter) on mouse chromosome 4 is shown with black arrows at the bottom of the figure.

Both cell lines also showed significant enrichments at additional known SCL regulatory elements including the -8/-9 region, the endothelial enhancer at -3/-4 region, the stem cell enhancer at +18/+19 region, the MAP17 promoter (p^{MAP17}) at +32 region and the erythroid enhancer at the +40 region. However, some regions showed significant enrichments in one cell line only. Significant enrichments were seen at the neural enhancer at +23/+24 (Gottgens et al. 2000), at a novel region at SIL -6, and in the cytochrome P450 gene cluster in the E14 ES cell line only. Conversely, 416B showed enrichments at SIL +18, SCL -20, SCL +15 and SCL+47 which were not seen in the E14 ES cell line.

5.6.3 Histone H3 modifications at lysines 9, 14, 18 and 27 correlate with transcriptional activity of genes

To gain a greater insight into the residue-specific acetylation events that modulate regulatory activity, ChIP-chip experiments were performed for acetylation at lysine residues 9, 14, 18, and 27 of histone H3 (Figure 5.6 and Table 5.2). A brief summary will now be described for the principle findings of these ChIP-chip experiments across the SCL locus with respect to each of these modifications in turn.

5.6.3.1 Acetylation of histone H3 at lysine 9 (H3 K9)

It has been shown that acetylation of histone H3 at lysine 9 is found at the predicted transcriptional start sites of active genes and correlated with transcriptional rates genome-wide (Pokholok et al. 2005). In K562, Jurkat, HL60 and HPB-ALL, significant enrichments were obtained at the 5' ends of genes that were active in the respective cell lines (Table 5.2 and Figure 5.6, blue profiles in panels A, B, C and D). In addition, significant enrichments were seen at or near the +51 erythroid enhancer in K562, HL60 and HPB-ALL. These results are in agreement with previous studies and demonstrate that H3 K9 acetylation represents a histone modification which delineates primarily the 5' ends of active genes. Interestingly, the relative levels of ChIP enrichments for H3 K9 acetylation were very low in Jurkat as compared to all the other cell lines (Figure 5.6). Although, the experiment was performed twice in Jurkat, similar results were obtained both times. This suggests that H3 K9 acetylation may not mark the 5' ends of active genes in Jurkat in the same way as it does in the other cell lines analyzed.

5.6.3.2 Acetylation of histone H3 at lysine 14 (H3 K14)

The relative levels of enrichments for this modification were very low in all the cell lines (Figure 5.6, pink profiles in panels A, B, C and D). In fact, no significant enrichments were observed for any region in HL60; and in Jurkat and HPB-ALL, no significant enrichments were obtained at the 5' end of KCY gene. These data suggest that this modification may be used in a cell-type specific manner. Furthermore, in K562, the H3 K14 profile was strikingly similar to that seen for histone H3 K9/14 diacetylation, although additional significant enrichments were also present at a novel region 4 to 5 kb upstream of the KCY gene (KCY -4/-5). The function of this region is not known and may represent a novel putative regulatory region.

	SCL expressing cell lines		SCL non-expressing cell lines	
	K562	Jurkat	HL60	HPB-ALL
H3 K9_Ac	5' end of KCY 5' end of SIL 5' end of SCL (p1a, p1b, +1, +3), p^{EXON4} at +7 p^{MAP17} at +42 at +51 region with two peaks on either side at +50 and +53 regions	5' end of KCY 5' end of SIL 5' end of SCL (p1a, p1b, +1, +3), p^{EXON4} at +7	5' end of KCY 5' end of SIL p^{MAP17} at +42 SCL +50, +51, +53 (three peaks with low levels of enrichment)	5' end of KCY 5' end of SIL SCL +53
H3 K14_Ac	5' end of KCY, KCY -4/-5 5' end of SIL 5' end of SCL (p1a, p1b, +1, +3), p^{EXON4} at +7 p^{MAP17} at +42 SCL -12, -9/-10, -5/-7, -3/-4 SCL +20/+21, +39, +50 and +53 regions (very low level enrichments at +51)	5' end of SCL (p^{1a}, p^{1b}, +1, +3), p^{EXON4} at +7 SCL -9/-10		5' end of SIL
H3 K18_Ac	5' end of KCY, KCY -4/-5 5' end of SIL, SIL -10 5' end of SCL (p1a, p1b, +1, +3) p^{MAP17} at +42 SCL -12, -9/-10, -5/-7 SCL +20/+21, +39, +51 region with two peaks at +51 and +53 regions	5' end of KCY, KCY -4/-5 5' end of SIL 5' end of SCL (p1a, p1b, +1, +3), p^{EXON4} at +7 SCL -9/-10, -7 (peak at -7) , -3/-4	5' end of KCY, KCY -4/-5 5' end of SIL SCL -9/-10 SCL +53	5' end of KCY, KCY -4/-5 5' end of SIL SCL +53
H3 K27_Ac	5' end of KCY, KCY -4/-5 5' end of SIL, SIL -10 5' end of SCL (p1a, p1b, +1, +3), p^{EXON4} at +7 p^{MAP17} at +42 SCL -9/-10 SCL +20/+21, +39, peak at +51 (significant enrichments at +50 and +53)	5' end of KCY, KCY -4/-5 5' end of SIL 5' end of SCL (p1a, p1b, +1, +3), p^{EXON4} at +7 SCL -9/-10, -7, -3/-4	5' end of KCY, KCY -4/-5 5' end of SIL	5' end of KCY, KCY -4/-5 5' end of SIL SCL +53

Table 5.2: Genomic regions across the SCL locus with significant ChIP enrichments for acetylation at lysines 9, 14, 18 and 27 of histone H3 in SCL expressing and non-expressing human cell lines. The specific histone modifications are listed in the first column and the cell lines are listed at the top in adjacent columns. The regions in blue showed significant but very low levels of enrichments and the regions in red represent novel putative regulatory regions which had not been identified previously across the SCL locus. The novel regions are named based on their location with respect to the promoter of the nearest gene. The

numbering is based on the distances upstream (-) or downstream (+), in kilobases, of the promoter of the respective gene. For instance, KCY -4/-5 region was located at four to five kb upstream of the KCY promoter. The majority of the novel regions are annotated on Figure 5.6. Note: regions annotated do not include those identified in the cytochrome P450 gene cluster (see Appendix 6).

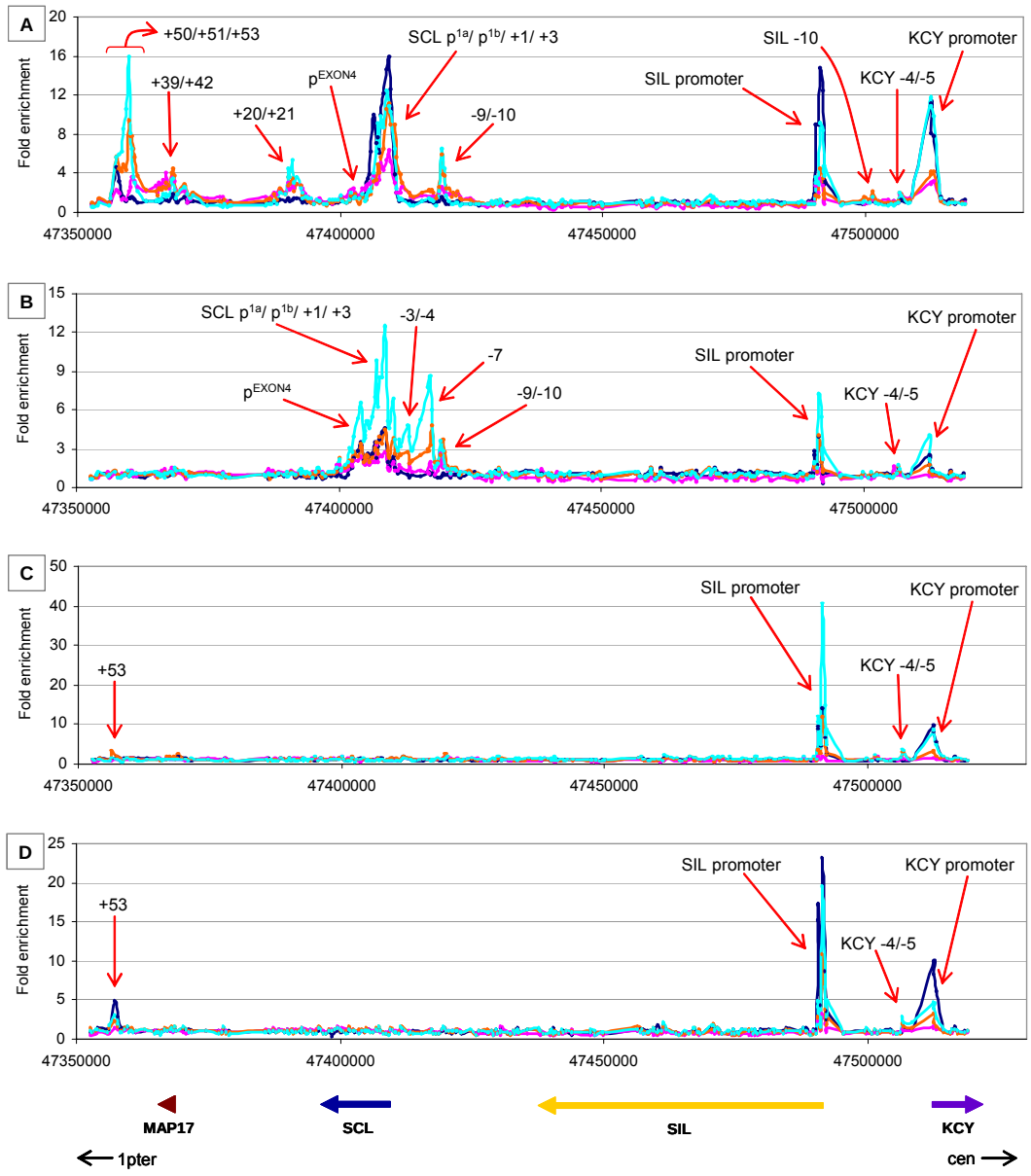


Figure 5.6: Composite ChIP-chip profiles for histone H3 acetylation at lysines 9, 14, 18 and 27 across the SCL locus in the four human cell lines. Panel A: K562, panel B: Jurkat, panel C: HL60 and panel D: HPB-ALL .The above profiles represent only the genomic region across the KCY, SIL, SCL, MAP17 genes and up to the genomic region just downstream of the SCL erythroid enhancer at +51. Blue profile represents H3 K9, pink profile represents H3 K14, orange profile represents H3 K18 and turquoise profile represents H3 K27 acetylation. The genomic regions showing significant enrichments in the ChIP assays for these modifications are marked with red arrows and also listed in Table 5.2. . The x-axes represent the genomic coordinates along human chromosome 1 and the y-axes represent fold enrichments. The thick, coloured arrows at the bottom of the panel D represent the gene order and the direction of transcription. The orientation of the locus with respect to the centromere (cen) and telomere (ter) is shown with black arrows at the bottom of the figure.

5.6.3.3 Acetylation of histone H3 at lysine 18 (H3 K18)

In all four human cell lines, significant enrichments for H3 K18 acetylation were observed primarily at the 5' ends of all the genes that were active in the respective cell lines (Table 5.2 and Figure 5.6, orange profiles in panels A, B, C and D). In K562 and Jurkat, the acetylation for H3 K18 at the 5' end of the SCL gene extended almost 9 kb into the coding region. In addition, both K562 and Jurkat showed significant enrichments at most of the regulatory regions which were also identified by H3 K9/14 diacetylation (Table 5.2) including the novel KCY -4/-5 region. No significant enrichments were observed across the CYP4A22 and CYP4Z1 genes in any of the cell lines apart from Jurkat where significant enrichment was present in intron 8 of the CYP4Z1 gene. This data suggests that H3 K18 acetylation marks virtually all the known regulatory regions across the SCL locus of expressed genes and also marks a few unknown regions which may represent novel putative regulatory regions.

5.6.3.4 Acetylation of histone H3 at lysine 27 (H3 K27)

Similarly, significant enrichments for acetylation at lysine 27 of histone H3 marked the 5' ends of the genes that were active in the respective cell lines (Figure 5.6, turquoise profiles in panels A, B, C and D). In addition, significant enrichments were also present at the novel KCY -4/-5 region in all of the cell lines. In K562 and Jurkat, significant enrichments were obtained at a number of the other SCL regulatory regions (Table 5.2).

5.7 Histones H2A and H2B acetylation patterns at the SCL locus in SCL expressing and non-expressing cell lines

5.7.1 Histone H2A acetylation mimics that of H3 9/14 diacetylation

ChIP-chip experiments for histone H2A tetra-acetylated at lysines 5/9/13/15 were performed in the K562 cells only. The pattern of histone H2A acetylation was very similar to that of histone H3 K9/14 diacetylation (Figure 6.21, Appendix 6). Significant enrichments were obtained at the 5' ends of KCY, SIL and SCL genes. At the SCL gene, significant enrichments at the upstream regulatory regions were obtained at -9/-10 (Gottgens et al. 1997), at -5/-7, and over the 5' end at p^{1a}, p^{1b}, +1, +3 and p^{EXON4}. In addition, significant enrichments were observed at the stem cell enhancer at +20/+21, the +39 region, p^{MAP17} at +42, and near the +51 erythroid enhancer with levels peaking at the +50 and +53 regions.

5.7.2 Acetylation of histone H2B marks tissue-specific genes

It has been shown that acetylation of histone H2B is a feature of actively transcribed tissue-specific genes and not of housekeeping genes (Myers et al. 2003). ChIP experiments for histone H2B tetra-acetylated at lysines 5/12/15/20 were performed only

in the K562 cell line. In agreement with these previous findings, very subtle but significant enrichments were observed at the 5' end of KCY gene (3.8 fold) but not at the 5' end of SIL gene (Figure 5.7).

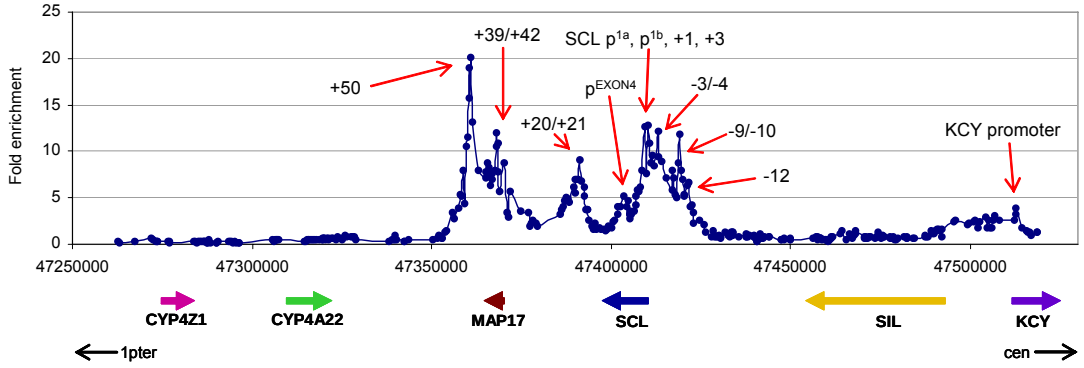


Figure 5.7: ChIP-chip profile of H2B tetra-acetylation at lysines 5/12/15/20 across the SCL locus in K562. Genomic regions with significant enrichments are marked with red arrows. The x-axis represents the genomic coordinates along human chromosome 1 and the y-axis represents fold enrichments. The genomic regions denoted by +1, +3 etc. are based on their distance upstream (-) or downstream (+) in kilobases from the human SCL p^{1a}. The thick, coloured arrows at the bottom of the figure represent the gene order and the direction of transcription. The orientation of the locus with respect to the centromere (cen) and telomere (ter) in human chromosome 1 is shown with black arrows at the bottom of the figure.

However, significant and much higher enrichments were obtained across the SCL gene at upstream regions -12, -9/-10, -5/-7, the endothelial enhancer at -3/-4 and the 5' end of SCL at p^{1a}, p^{1b} extending into the +1, +3 and p^{EXON4} (at +7) regions. In addition, significant enrichments were also observed at the stem cell enhancer at +20/+21, +39 region, p^{MAP17} at +42, and the +51 erythroid enhancer. At the +51 region, the region showing significant enrichments extended from +50 to +53 regions with the highest peak at the +50 region.

5.8 Histone H4 acetylation patterns at the SCL locus in SCL expressing and non-expressing cell lines

5.8.1 H4 acetylation coincides with H3 acetylation across the SCL Locus

It has been shown that, like histone H3 acetylation, histone H4 acetylation correlates with the transcriptional activity and transcriptional rates of genes (Pokholok et al. 2005). The results of this thesis are in agreement with these previous observations. ChIP experiments were performed in human cell lines to histone H4 tetra-acetylated at lysines 5/8/12/16 (Figure 5.8 and Table 5.3). In all the cell lines, significant enrichments were seen at or near the 5' ends of the KCY and SIL genes, and at the 5' end of SCL and its regulatory elements in the SCL expressing cell lines K562 and Jurkat (Table 5.3). Although significant enrichments at the p^{EXON4} were found both in K562 and Jurkat, the relative levels of ChIP enrichments at the p^{EXON4} were much higher in Jurkat as

compared to K562. This is in agreement with previous studies which show that SCL p^{EXON4} is active in T-cell lines (Bernard et al. 1992). In Jurkat and HL60, significant enrichments were observed at the SIL -10 and KCY -4/-5 regions, at SIL -10 in K562 and at KCY -4/-5 in HPB-ALL. These regions were also identified previously with a number of histone H3 acetylation modifications (see section 5.6).

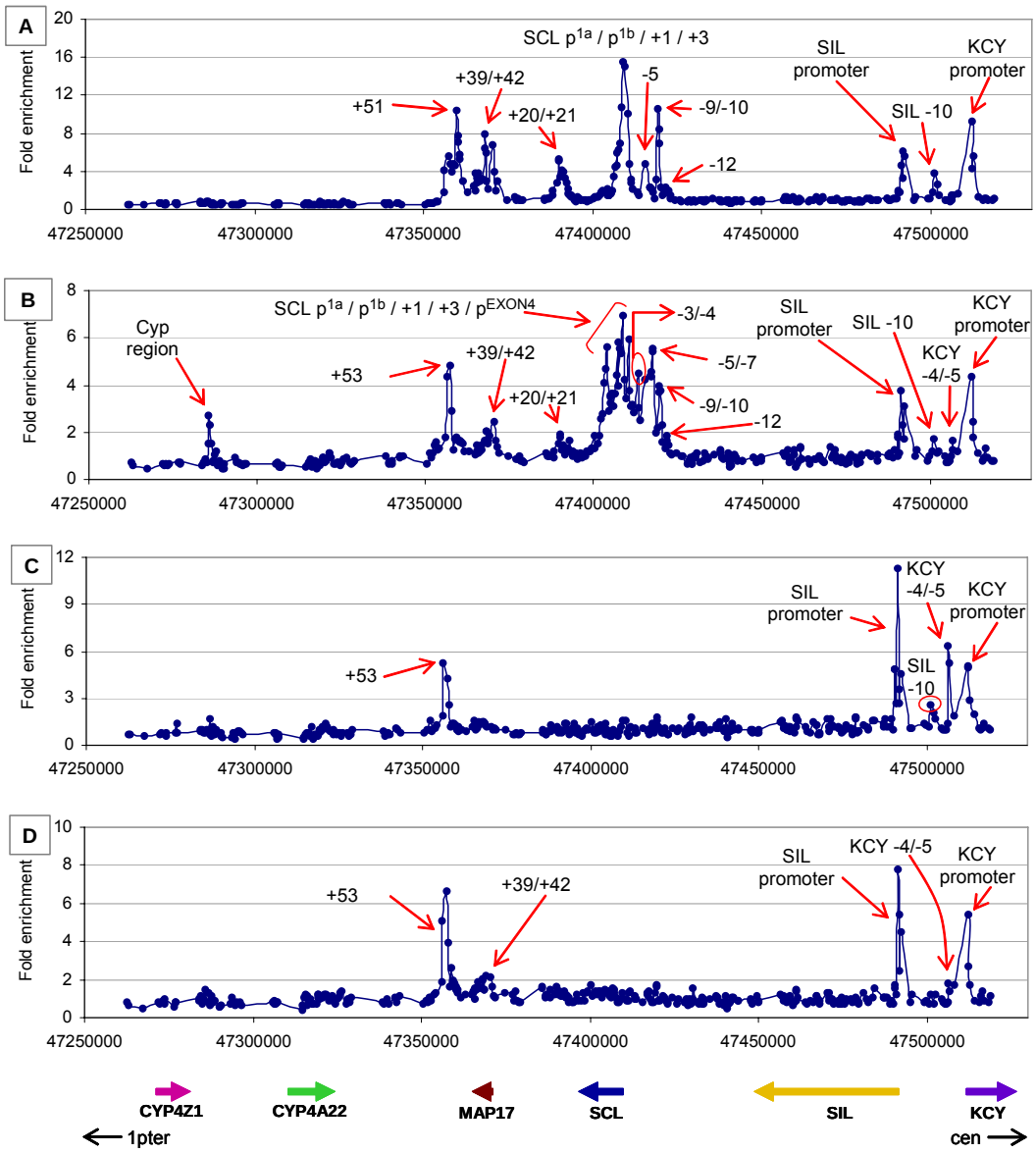


Figure 5.8: ChIP-chip profiles for H4 K5/8/12/16 tetra-acetylation across the SCL locus in the four human cell lines. Panel A: K562, panel B: Jurkat, panel C: HL60 and panel D: HPB-ALL. The genomic regions showing significant enrichments are marked with red arrows. The x-axes represent genomic coordinates along human chromosome 1 and the y-axes represent fold enrichments. The regions denoted by -12, -9/-10 etc. are based on their distance upstream (-) or downstream (+), in kilobases, from SCL p^{1a}. The thick, coloured arrows at the bottom of panel B represent the gene order and the direction of transcription. The orientation of the locus with respect to the centromere (cen) and telomere (ter) on human chromosome 1 is shown by the black arrows at the bottom of the figure.

	SCL expressing cell lines		SCL non-expressing cell lines	
	K562	**Jurkat**	**HL60**	**HPB-ALL**
H4 K5/8/12/16 tetra-acetylated	5' end of KCY 5' end of SIL, SIL -10 5' end of SCL (p^{1a}, p^{1b}, +1, +3) p^{EXON4} at +7 p^{MAP17} at +42 SCL -12, -9/-10, -5/-7 SCL +20/+21, +39 SCL +51 region (from +50 to +53 regions)	5' end of KCY, KCY -4/-5 5' end of SIL, SIL -10 5' end of SCL (p^{1a}, p^{1b}, +1, +3) p^{EXON4} at +7 p^{MAP17} at +42 SCL -12, -9/-10, -5/-7, -3/-4 SCL +20/+21, +39 SCL +51 region with highest peak at +53	5' end of KCY, KCY -4/-5 5' end of SIL, SIL -10 SCL +53	5' end of KCY, KCY -4/-5 5' end of SIL p^{EXON4} at +7 SCL +39 SCL +51 region with highest peak at +53
H4 K5_Ac	KCY -4/-5 SIL -10 5' end of SCL (p^{1a}, p^{1b}, +1) p^{EXON4} at +7 p^{MAP17} at +42 SCL -12, -9/-10, -5/-7, -3/-4 SCL +20/+21, +39, two peaks at +50 and +53 (no significant enrichments at +51)	SIL -10 5' end of SCL (p^{1a}, +1, +3) upstream of p^{EXON4} at +7 p^{MAP17} at +42 SCL -9/-10, -3/-4 SCL +20/+21, +39, +51 (three peaks at +50, +51, +53)	 p^{MAP17} at +42 SCL +51 region with highest peak at +53	5' end of SIL (not over p^{SIL}) downstream region of p^{EXON4} (+8) p^{MAP17} at +42 SCL +39, +51 region with highest peak at +53
H4 K8_Ac	5' end of KCY 5' end of SIL 5' end of SCL (p^{1a}, p^{1b}, +1, +3) SCL +51 SCL -12, -3/-4 p^{EXON4} at +7 SCL +20/+21, p^{MAP17} at +42 SIL -10, KCY -4/-5	5' end of KCY 5' end of SIL 5' end of SCL (p^{1b}, +1, +3) SCL -9/-10, -3/-4 SCL +20/+21, p^{MAP17} at +42 SCL +51 (with three peaks at +50, +51, +53) SIL -10, KCY -4/-5	 5' end of SIL 5' end of SCL (+3) SCL -9/-10 p^{MAP17} at +42 SCL +51 (on either sides at +50, +53 with a peak at +53 region) SIL -10, KCY -4/-5	5' end of KCY 5' end of SIL SCL -9/-10, -7 SCL +3, p^{EXON4} at +7 SCL +20/+21, p^{MAP17} at +42 SCL +51 (and on either sides at +50, +53)
H4 K16_Ac	5' end of KCY 5' end of SIL, SIL -10 5' end of SCL (p^{1a}, p^{1b}, +1, +3) SCL -9/-10, -5/-7 SCL +39, +51 (extending from +50 to +53 region), +57 p^{EXON4} at +7 SCL +23/+24, p^{MAP17} at +42	5' end of KCY 5' end of SIL 5' end of SCL (p^{1a}, p^{1b}, +1, +3) SCL -7 SCL +39, +57 SCL +20/+21, p^{MAP17} at +42, +51 with peak at +50	5' end of KCY 5' end of SIL 5' end of SCL (p^{1a}, p^{1b}, +1) SCL +39	5' end of KCY 5' end of SIL

Table 5.3: Genomic regions across the SCL locus showing significant ChIP enrichments for histone H4 acetylation. The results presented in the above table include tetra-acetylation of H4 at lysines 5/8/12/16 and acetylation at individual lysines at 5, 8 and 16. The histone modifications are listed in the first column and the cell lines are listed at the top of the adjacent columns. The ChIP-chip profiles for H4 K8 and H4 K16 showed significant enrichments as well as decreased levels over regulatory regions across the SCL locus. The regions showing decreased levels are listed in the pink coloured sections. The genomic regions listed in the rest of the table were significantly enriched for the respective histone modifications. The regions in blue showed significant but very low levels of enrichments and the regions in red represent novel putative regulatory regions which were not known previously. The novel regions are named based on their location with respect to the promoter of the nearest gene. The numbering is based on the distances upstream (-) or downstream (+), in kilobases, of the respective gene. The regions collectively identified with histone H4 acetylation included the SCL p^{1a}, p^{1b}, +1, +3 and SCL p^{EXON4} at the +7 region (Aplan et al. 1990a; Leroy-Viard et al. 1994; Bernard et al. 1992; Sinclair et al. 1999), -9/-10 (Gottgens et al. 1997), -7 region (Leroy-Viard et al. 1994), the endothelial enhancer at -3/-4 region (Gottgens et al. 2004), the stem cell enhancer at +20/+21 (Gottgens et al. 2002), the erythroid enhancer at +51 (Delabesse et al. 2005).

5.8.2 Acetylation of histone H4 at specific lysine residues correlates variably with transcriptional activity at the SCL locus

Histone H4 can be acetylated at lysines 5, 8, 12 and 16 (Figure 5.1) and it has been suggested that hyperacetylation of H4 at lysine 5 and 12 and hypoacetylation of lysines 8 and 16 are linked to transcriptional activity (Liu et al. 2005). The genomic regions showing significant enrichments for acetylation at lysines 5, 8 and 16 of histone H4 across the SCL locus in the four human cell lines are listed in Table 5.3. Antibodies assaying for H4 K12 acetylation did not produce any significant enrichment in K562 (Table 5.1). Complete profiles are shown in Appendix 7.

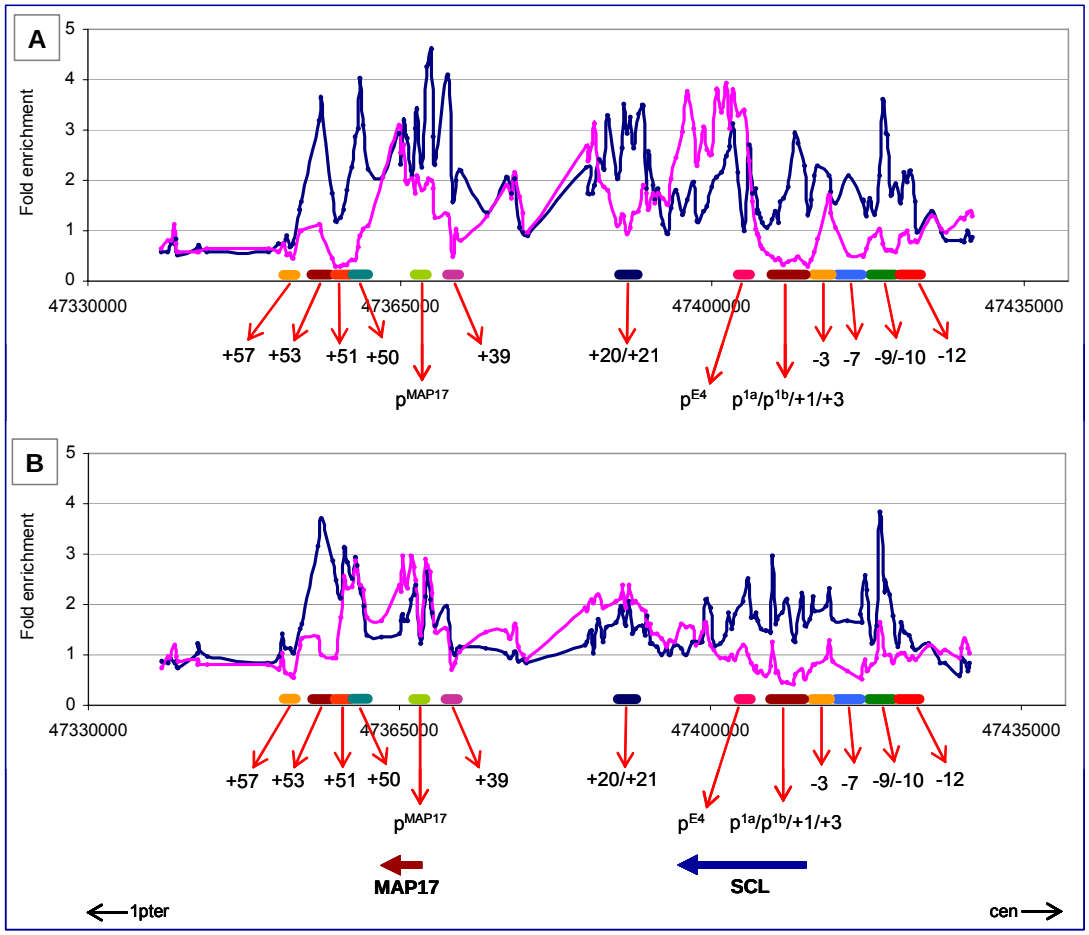


Figure 5.9: ChIP-chip profiles for acetylation at lysines 5 and 16 of histone H4 across the SCL and MAP17 genes. Panel A: K562 and panel B: Jurkat. ChIP-chip profiles for H4 K5 are represented in blue and for H4 K16 in pink across the genomic region containing the SCL and MAP17 genes. The x-axes represent the genomic coordinates along human chromosome 1 and the y-axes represent fold enrichments. The known regulatory regions across the region are represented with coloured bars at the bottom of the plot and annotated with red arrows. The thick, coloured arrows at the bottom of panel B represent the SCL and MAP17 genes and their direction of transcription with respect to the centromere (cen) and telomere (ter) on human chromosome 1. Complete ChIP-chip profiles across the entire SCL locus for all four cell lines are shown in Appendix 7.

5.8.2.1 Acetylation of histone H4 at lysine 5 (H4 K5)

ChIP-chip profiles for H4 K5 acetylation across the SCL and MAP17 genes in K562 and Jurkat are shown in Figure 5.9 (blue profiles in panels A and B). One of the striking observations of the ChIP-chip profiles for H4 K5 acetylation was that the 5' ends (i.e. promoters) of the KCY, SIL and SCL genes were not significantly enriched in any cell lines (Appendix 7). However, a number of the other SCL regulatory regions showed significant ChIP enrichments for H4 K5 acetylation (Table 5.3), although these enrichments were relatively low. This data suggests that H4 K5 acetylation was not linked appreciably to gene activity across the SCL locus and, in particular, is not associated with active promoter regions at the 5' ends of genes.

5.8.2.2 Acetylation of histone H4 at lysine 16 (H4 K16)

Figure 5.9 shows ChIP-chip profiles for H4 K16 acetylation across the SCL and MAP17 genes in K562 and Jurkat cell lines. In all of the cell lines, decreased relative levels were seen at or near the 5' ends (i.e. promoters) of active genes. However, in the SCL non-expressing cell lines, decreased levels were also observed at the 5' end of SCL gene in HL60, while HPB-ALL showed increased levels across this promoter (Appendix 7). The results obtained across all four cell lines suggest that hypo H4 K16 acetylation may be linked to gene activity at the SCL locus. In other words, increases in H4 K16 acetylation may be linked to inactive genes.

5.8.2.3 Acetylation of histone H4 at lysine 8 (H4 K8)

ChIP-chip profiles for H4 K8 acetylation also showed significant enrichments at various regulatory regions across the SCL locus in all of the cell lines (Table 5.3). However, as was the case for H4 K16 acetylation, decreased levels of H4 K8 acetylation were seen at the 5' end of genes that were active in the relevant cell lines. These results are in general agreement with those obtained for H4 K16, which would suggest that the presence of both H4 K8 and K16 acetylation is inversely related to gene activity. This agrees with observations previously reported (Liu et al. 2005).

5.9 Histone H3 lysine methylation patterns at the SCL locus in SCL expressing and non-expressing human and mouse cell-lines

Histone lysine methylation plays an important part in transcriptional regulation and can exist in three states on the lysine residue *in vivo*: mono, di and tri (Strahl et al. 1999). Methylation of lysine 4 of histone H3 has been associated to both the activation and repression status of a gene, depending largely on the degree of methylation of this residue. Trimethylation of lysine 4 of histone H3 has been associated with active genes

and it has been suggested that dimethylation may be linked to "poised" state of chromatin (Santos-Rosa et al. 2002; Ng et al. 2003; Schneider et al. 2004) whereas monomethylation of this residue has been considered to be a mark for repressed or inactive DNA regulatory sequences in chromatin (van Dijk et al. 2005). Similarly, in studies in *Saccharomyces cerevisiae*, the monomethylation of H3 lysine 79, which lies in the central globular domain of the histone H3, has been implicated in gene silencing (Ng et al. 2002; van Leeuwen et al. 2002). Dimethylation of H3 K79 is found to be correlated with active genes (Im et al. 2003) and trimethylation of H3 K79 is associated with coding regions (Pokholok et al. 2005). However, the function and distribution of trimethylation of H3 at lysine 79 in mammalian genomes is not yet clear. Thus, the distribution of these modifications across the SCL locus would provide an insight into how histone lysine methylation states are associated with the SCL regulatory elements and their regulation.

5.9.1 The 5' ends of active genes are hypomethylated for histone H3 lysine monomethylation (H3 K4Me1)

5.9.1.1 Monomethylation of histone H3 lysine 4 in SCL expressing cell lines in human and mouse

As stated earlier, SCL p^{1a} and p^{1b} are transcriptionally active in the erythroid and primitive myeloid cell lines K562 and 416B (Aplan et al. 1990a; Bockamp et al. 1997) whereas, p^{1b} and p^{EXON4} at the +7 region are known to be active in T-cell lines such as Jurkat (Bernard et al. 1992). ChIP-chip profiles for monomethylation at lysine 4 of histone H3 across the SIL, SCL and MAP17 genes in K562, Jurkat and 416B cell lines are shown in Figure 5.10 (profiles across the entire locus are shown in Appendices 9 and 11).

In the human cell lines K562 and Jurkat, lack of enrichments for H3 K4 monomethylation were obtained at or near the 5' ends (i.e. promoters) of the transcriptionally active KCY, SIL and SCL genes. This pattern of decreased levels was also observed for SIL and SCL in 416B cell line (KCY gene is not represented on the mouse SCL array). In the three cell lines, the region at the 5' end of the SCL gene with decreased levels of enrichments extended up to 4 kb into the coding region. The 5' end of the SCL gene showing lack of enrichments for H3 K4 monomethylation encompassed a number of known regulatory regions which had previously been reported to show enhancer activity or DNase I hypersensitivity (see Table 5.4). In Jurkat, decreased H3 K4 monomethylation was observed over the SCL promoter within exon 4 (p^{EXON4}) which is active in this cell line (Bernard et al. 1992).

However, sequences adjacent to active promoter or inactive promoters displayed significant enrichments for H3 K4 monomethylation in all three cell lines (see Table 5.4, yellow coloured section). In K562 and Jurkat, increased levels of enrichments were

observed downstream of both the KCY and SIL promoters and in 416B, enrichments were seen upstream as well downstream of the SIL promoter. At the SCL gene, increased levels were observed at p^{EXON4} (at +7) in K562 correlating with the transcriptional inactivity of this promoter in K562. In Jurkat and 416B, increased levels of enrichment were observed at genomic sequences either side of p^{EXON4}. These data support the idea that decreased levels of H3 K4 monomethylation, in combination with increased levels at either side, are hallmarks of active promoters.

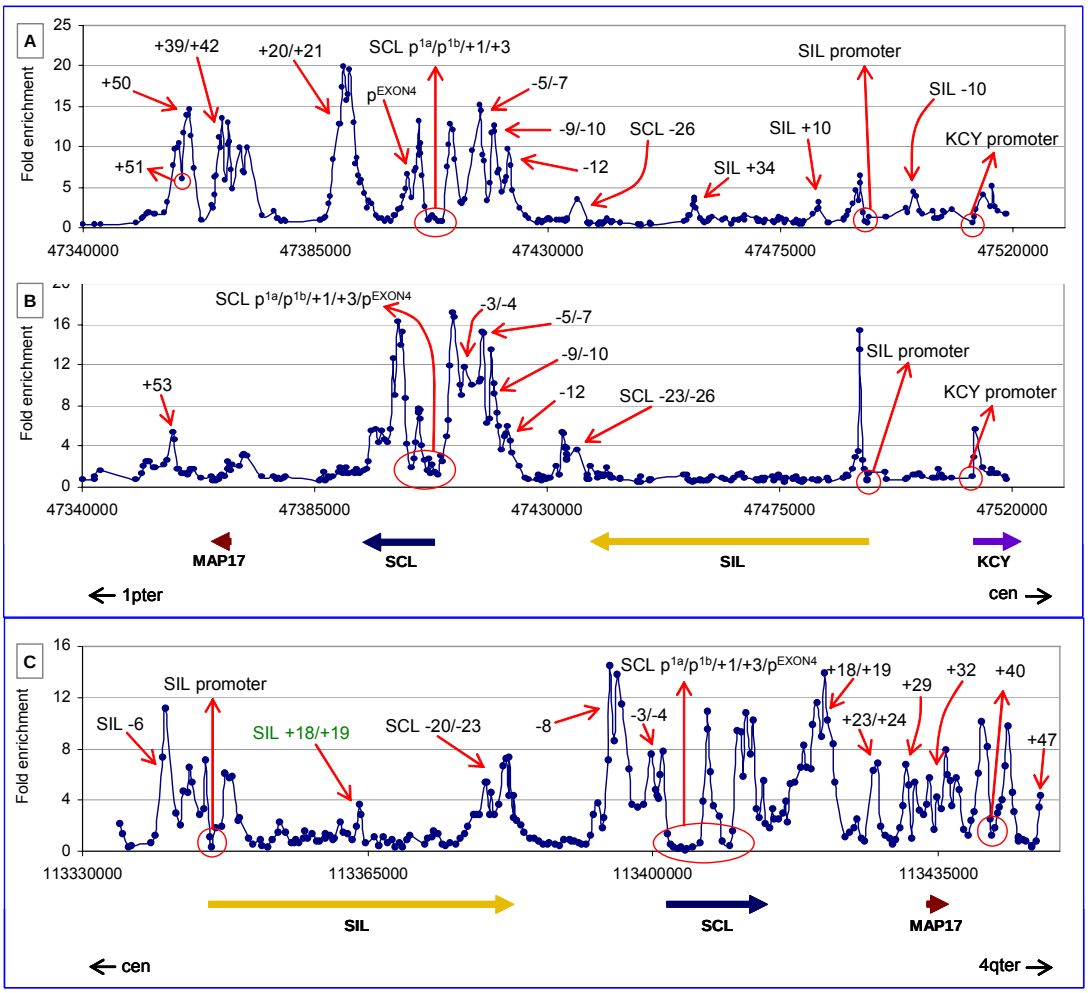


Figure 5.10: ChIP-chip profiles for monomethylation of H3 K4 across the SCL locus in SCL expressing human and mouse cell lines. Panels A and B show the profiles across KCY, SIL, SCL and MAP17 genes in K562 and Jurkat respectively. Panel C represent the profile across the SIL, SCL and MAP17 genes in 416B cell line. The genomic regions showing significant enrichments are marked by red arrows. The genomic regions showing lack of enrichments for monomethylation of H3 K4 are highlighted by red circles. The x-axes in panels A and B represent genomic coordinates for human chromosome 1 and for mouse chromosome 4 in panel C. The y-axes in all three panels represent fold enrichments. The genomic regions are annotated based on their distance (in kilobases) upstream (-) or downstream (+) to the promoter of the nearest gene. For instance, in panel A, SIL +10 and SIL +34 are located 10 and 34 kb downstream of p^{SIL} respectively. Similarly, the SCL -26 region in the same panel is located 26 kb upstream of the SCL p^{1a}. The regions annotated in green were identified within the SIL gene which showed ChIP enrichments but the enrichments were not found to be

significant (see text). The thick, coloured arrows at the bottom of panels B and C represent the gene order and the direction of transcription in human and mouse respectively. The orientation of the SCL loci in human and mouse with respect to the centromere (cen) and telomere (ter) on human chromosome 1 and mouse chromosome 4 are shown by black arrows at the bottom of panels B and C respectively.

Increased H3 K4 monomethylation was associated with the majority of all other known regulatory elements across the SCL loci in human and mouse (Table 5.4; Figure 5.10). In addition, novel genomic regions of unknown function were also identified with significant enrichments for monomethylation in all three cell lines - some of which had been identified previously with a number of histone H3 and H4 modifications (SIL -10 and SCL +39 in human and SCL +47 in mouse) (Table 5.4). Other novel regions identified specifically with monomethylation of H3 K4 are shown in Table 5.4 and Figure 5.10. Many of these H3 K4Me1 novel regions were located within the genomic region of the SIL gene which was used to calculate the background levels to determine significant ChIP enrichments for all assays (Chapter 4). Thus, some of them did not report significant enrichments, since the SDs of background levels in the SIL gene were very high. However, these peaks of enrichment were still highlighted as they were clearly visible by visual inspection and constitute sequences which may have, as yet, an undetermined function in the regulation of the genes across the SCL locus.

5.9.1.2 Monomethylation of histone H3 lysine 4 in SCL non-expressing cell lines in human and mouse

Figure 5.11 shows the ChIP-chip profiles obtained for monomethylation of histone H3 K4 across the SIL, SCL and MAP17 genes in HL60, HPB-ALL and the mouse ES cell line E14 (complete profiles are shown in Appendices 9 and 11). As was observed in SCL expressing cell lines, decreased levels of monomethylation were found at the 5' ends (i.e. promoters) of active genes. In HL60 and HPB-ALL, decreased levels were seen at the promoter regions of KCY and SIL genes, and at the promoter region of the SIL gene in the mouse E14 ES cell line. In E14 ES cells, the promoter region at the 5' end of the SIL gene was bordered by regions of increased levels of monomethylation on either side whereas, in HL60 and HPB-ALL, increased levels of enrichment were present at the downstream regions (Table 5.4). Over the inactive SCL promoters, high levels of monomethylation were observed in both HL60 and mouse E14 ES cells. These findings support the idea that monomethylation levels "drop-out" over active promoters as described in the previous section.

Significant enrichments were also obtained at a number of regulatory regions across the SCL locus (Table 5.4); these were particularly evident in E14 ES cells. Thus, significant monomethylation at H3 K4 can occur at other regulatory sequences across the SCL

gene in the absence of gene expression. Furthermore, E14 ES cells revealed a number of genomic regions within the SIL gene and the cytochrome P450 genes with ChIP enrichments (Figure 5.11 and Appendix 11). Some of these regions were not statistically significant for reasons discussed above.

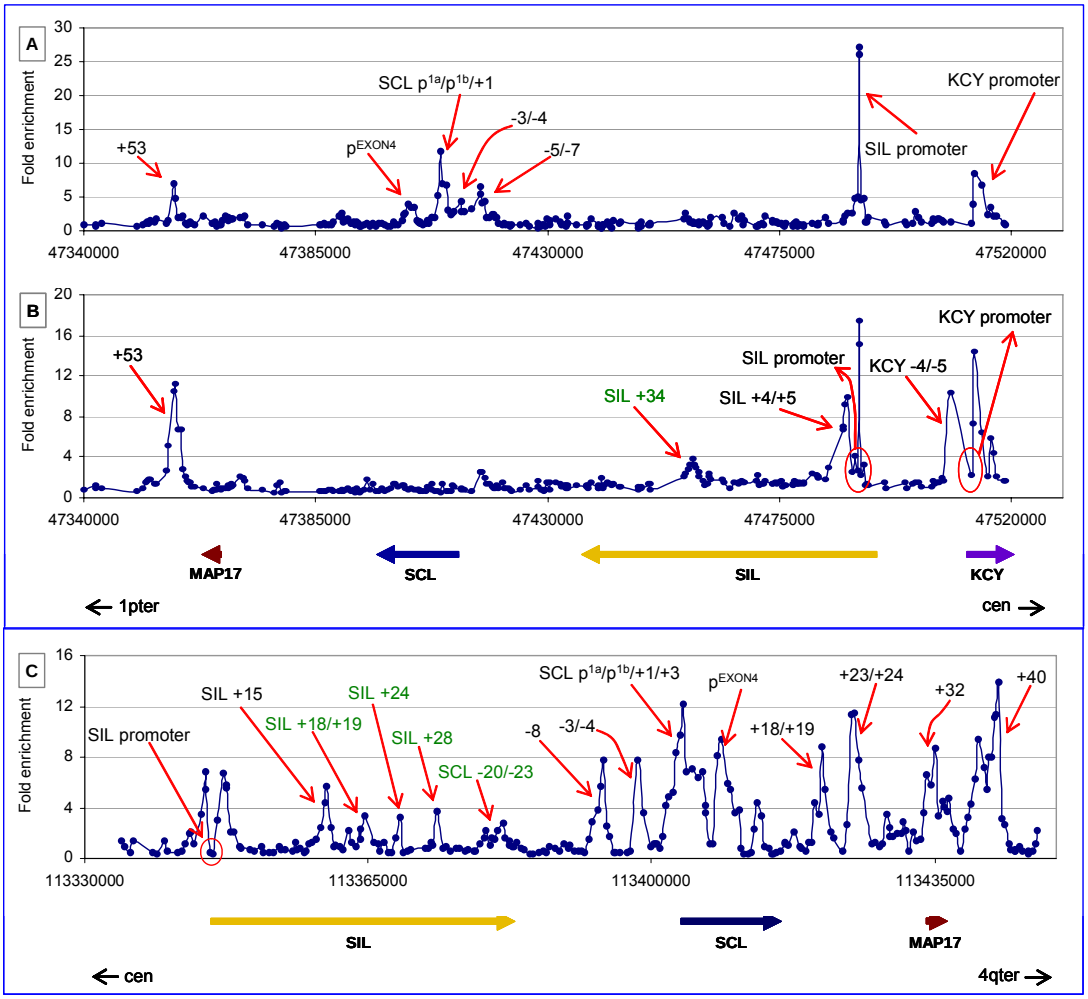


Figure 5.11: ChIP-chip profiles for monomethylation of H3 K4 across the SCL locus in SCL non-expressing human and mouse cell lines. Panels A and B show the profiles across KCY, SIL, SCL and MAP17 genes in HL60 and HPB-ALL respectively. Panel C represents the profile across the SIL, SCL and MAP17 genes in mouse E14 ES cell line. The genomic regions showing significant enrichments are marked by red arrows. The x-axes in panels A and B represent genomic coordinates for human chromosome 1 and in panel C for mouse chromosome 4. The y-axes represent fold enrichments. The genomic regions showing lack of enrichments for monomethylation of H3 K4 are highlighted by red circles. The regions are annotated based on their distance (in kilobases) upstream (-) or downstream (+) to the promoter of the nearest gene. The regions annotated in green text were identified within the SIL gene which showed ChIP enrichments but the enrichments were not found to be significant (for reasons discussed in the text). The thick, coloured arrows at the bottom of panels B and C represent the gene order and the direction of transcription in human and mouse respectively. The orientation of the SCL loci in human and mouse with respect to the centromere (cen) and telomere (ter) on human chromosome 1 and mouse chromosome 4 are shown by black arrows at the bottom of panels B and C respectively.

	SCL expressing cell lines			SCL non-expressing cell lines		
	K562 (Human cell line)	**Jurkat** (Human cell line)	**416B** (Mouse cell line)	**HL60** (Human cell line)	**HPB-ALL** (Human cell line)	**Mouse ES cells** (Mouse cell line)
H3 K4_Me1	5' end of KCY 5' end of SIL 5' end of SCL (p^{1a}, p^{1b}, +1, +3)	5' end of KCY 5' end of SIL 5' end of SCL (p^{1a}, p^{1b}, +1, +3) p^{EXON4} at +7	5' end of SIL 5' end of SCL (p^{1a}, p^{1b}, +1, +3) p^{EXON4} at +7 p^{MAP17} at +32, SCL +40	5' end of KCY 5' end of SIL	5' end of KCY 5' end of SIL	5' end of SIL
	downstream region of p^{KCY} downstream region of p^{SIL} SIL -10, +10, +34 SCL -26 SCL -12, -9/-10, -5/-7 SCL p^{EXON4} at +7 SCL +20/+21, +39, p^{MAP17} at +42 SCL +51 with peaks on either side, highest peak at +50	downstream region of p^{KCY} downstream region of p^{SIL} SCL -23/-26 SCL -12, -9/-10, -5/-7, -3/-4 either side of p^{EXON4} SCL +53	either side of p^{SIL} SIL -6, +18/+19 SCL -20/-23 SCL -8, -3/-4 either side of p^{EXON4} SCL +18/+19, +23/+24, +29, either side of p^{MAP17} SCL +47	downstream region of KCY downstream region of SIL 5' end of SCL (p^{1a}, p^{1b}, +1) SCL -5/-7, -3/-4 SCL p^{EXON4} at +7 SCL +53	downstream region of KCY downstream region of SIL KCY -4/-5, SIL +4/+5, +34 SCL +53	either side of p^{SIL} SIL +15, +18/+19, +24, +28 SCL -20/-23 5' end of SCL (p^{1a}, p^{1b}, +1, +3) SCL -8, -3/-4 SCL p^{EXON4} at +7 SCL +18/+19, +23/+24, p^{MAP17} at +32 SCL +40
H3 K4_Me2	5' end of KCY 5' end of SIL, SIL -10 5' end of SCL (p^{1a}, p^{1b}, +1, +3) SCL p^{EXON4} at +7 SCL -12, -9/-10, -5/-7 SCL +20/+21, +39, pMAP17 at +42 SCL +51 with highest peak at +53, +57	5' end of KCY 5' end of SIL, SIL -10 5' end of SCL (p^{1a}, p^{1b}, +1, +3) SCL p^{EXON4} at +7 SCL -9/-10, -5/-7, -3/-4 SCL +39 SCL +51 with highest peak at +53, +57	5' end of SIL, SIL -6, SIL -2 5' end of SCL (p^{1a}, p^{1b}, +1, +3) SCL p^{EXON4} at +7 SCL -20, -8, -3/-4 SCL +18/+19, +29, p^{MAP17} at +32 SCL +40, +47	5' end of KCY 5' end of SIL 5' end of SCL (p^{1a}, p^{1b}, +1) SCL p^{EXON4} at +7 SCL -5/-7 SCL +51 with highest peak at +53	5' end of KCY 5' end of SIL, SIL +4/+5 SCL +53	5' end of SIL, SIL +24, +28 5' end of SCL (p^{1a}, p^{1b}, +1) SCL p^{EXON4} at +7 SCL -8, -3/-4 SCL +23/+24 SCL +40
H3 K4_Me3	5' end of KCY 5' end of SIL, SIL -10 5' end of SCL (p^{1a}, p^{1b}, +1, +3) SCL p^{EXON4} at +7 SCL -12, -9/-10, -5/-7 SCL +20/+21, +39, +42 (but not over ^{pMAP17}) SCL +51 with highest peak at +53, +57	5' end of KCY 5' end of SIL, SIL -10 5' end of SCL (p^{1a}, p^{1b}, +1, +3) SCL p^{EXON4} at +7 SCL -9/-10, -5/-7, -3/-4 SCL +39 SCL +51 with highest peak at +53, +57	5' end of SIL, SIL -6, SIL -2 5' end of SCL (p^{1a}, p^{1b}, +1, +3) SCL p^{EXON4} at +7 SCL -20, -8, -3/-4 SCL +18/+19, +29, p^{MAP17} at +32 SCL +40, +47	5' end of KCY 5' end of SIL 5' end of SCL (p^{1a}, p^{1b}, +1, +3) SCL p^{EXON4} at +7 SCL -5/-7 SCL +51 with highest peak at +53	5' end of KCY 5' end of SIL, SIL +4/+5 SCL +53	5' end of SIL, SIL +24, +28 5' end of SCL (p^{1a}, p^{1b}, +1) SCL p^{EXON4} at +7 SCL -8, -3/-4 SCL +23/+24 SCL +40

Table 5.4: Genomic regions identified with mono-, di-, trimethylation of histone H3 at lysine 4 in human and mouse cell lines. The histone modifications are listed in the first column and the cell lines are listed at the top in adjacent columns. ChIP-chip assays were performed in human and mouse cell lines. A number of genomic regions for monomethylation showed significant enrichments (shown in the yellow region) as well as lack of enrichments (shown in the pink region). The genomic regions listed for di- and trimethylation showed significant enrichments. The regions in blue text showed very low levels of enrichments, the regions in red text represent novel regions. The regions in green text represent novel regions which showed ChIP enrichments but the enrichments were not found to be significant (see text). The regions are named based on their location with respect to the promoter of the nearest gene. The numbering is based on the distances upstream (-) or downstream (+) in kilobases from the respective genes. The known SCL regulatory regions that were collectively identified with histone H3 methylation included the SCL p^{1a}, p^{1b}, +1, +3 and SCL p^{EXON4} at the +7 region (Aplan et al. 1990a; Leroy-Viard et al. 1994; Bernard et al. 1992; Sinclair et al. 1999), -9/-10 (Gottgens et al. 1997), -7 region (Leroy-Viard et al. 1994), the endothelial enhancer at -3/-4 region (Gottgens et al. 2004), the stem cell enhancer at +20/+21 (Gottgens et al. 2002), the erythroid enhancer at +51 (Delabesse et al. 2005).

5.9.2 Dimethylation and trimethylation at lysine 4 of histone H3 (H3 K4Me2, H3 K4Me3) occur at transcriptionally active genes across the SCL locus in human and mouse

5.9.2.1 Dimethylation and trimethylation at histone H3 lysine 4 in SCL expressing human cell lines

ChIP experiments were performed in human and mouse cell lines using antibodies raised to dimethylated and trimethylated lysine 4 of histone H3 (Table 5.1). Figure 5.12 shows composite ChIP-chip profiles of dimethylation and trimethylation of H3 K4 across the KCY, SIL, SCL and MAP17 genes in K562 and Jurkat cell lines respectively (complete profiles across the entire locus are shown in Appendix 9). The genomic regions which showed significant enrichments for di- and trimethylation of H3 K4 in K562 and Jurkat are listed in Table 5.4.

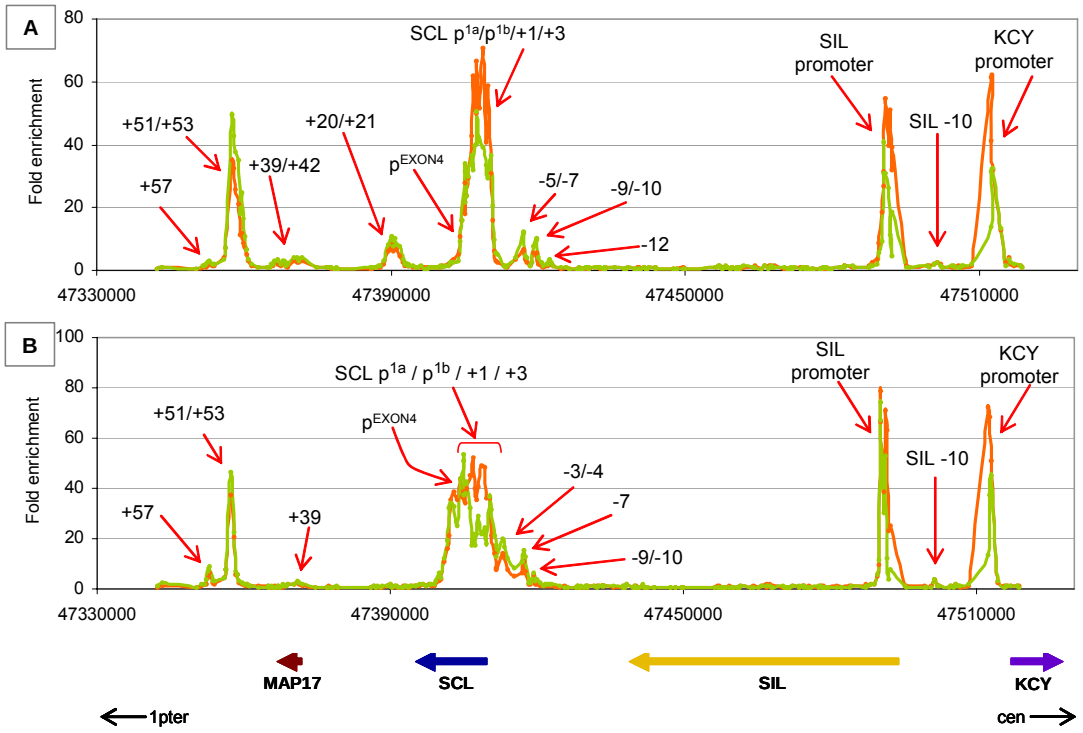


Figure 5.12: ChIP-chip profiles of di- and trimethylated lysine 4 of H3 across the SCL locus in K562 and Jurkat. Panel A: K562 and panel B: Jurkat. The orange profiles in panels A and B represent trimethylation of H3 K4 and the green profiles represent dimethylation of H3 K4. The genomic regions showing enrichments are marked with red arrows. The x-axes represent genomic coordinates along human chromosome 1 and the y-axes represent fold enrichments. The regions are denoted based on their distance upstream (-) or downstream (+) in kilobases from the promoter of the nearest gene. The thick, coloured arrows represent the gene order and the direction of transcription. The orientation of the locus with respect to the centromere (cen) and telomere (ter) on human chromosome 1 is shown with black arrows at the bottom of the figure.

In both cell lines, significant levels of enrichment for di- and trimethylation H3 K4 were obtained at or near the 5' ends of the KCY, SIL and SCL genes, all of which are expressed in these cell lines. The highest peaks of enrichments for trimethylation were observed at the promoter regions while the highest peaks for dimethylation were seen just downstream (3') of the promoter regions for all the active genes. In K562, at the 5' end of the SCL gene, both di- and trimethylation extended almost 9 kb into the coding region; in Jurkat, the significant enrichments extended almost 11 kb into the coding region. In both cell lines, significant enrichments were obtained at SCL p^{EXON4} which is known to be active only in Jurkat (Bernard et al. 1992). However, in K562, the relative levels of enrichments at p^{EXON4} were almost ten fold lower as compared to the active promoter p^{1a} and p^{1b}; in contrast, the relative levels of enrichments at p^{EXON4} were very similar to the other active SCL promoters in Jurkat (Figure 5.12, panel B). These observations suggest that levels of di- and trimethylation may reflect the activity of a promoter.

A number of other regulatory regions across the SCL locus showed significant enrichments for di- and trimethylation (Table 5.4). In K562, significant enrichments at the upstream (5') SCL regulatory regions were seen as discrete regions; in Jurkat, the enrichments were present over a continuous block encompassing all the known upstream SCL regulatory regions. This suggested that the regulation of SCL in K562 and Jurkat may be different with respect to these regulatory regions. Interestingly, significant enrichments for both di- and trimethylation were identified at the +57 region which had previously been shown to report significant enrichments with H3 K9/14 diacetylation (in Jurkat only) (Figure 5.3). The function of this region is not known.

5.9.2.2 Dimethylation and trimethylation at histone H3 lysine 4 in SCL non-expressing human cell lines

Figure 5.13 shows composite ChIP-chip profiles of dimethylation and trimethylation of histone H3 K4 across the KCY, SIL, SCL and MAP17 genes in HL60 and HPB-ALL. Significant enrichments were obtained at the 5' end of KCY, SIL genes, and were in agreement with data from K562 and Jurkat showing the association of transcriptional activity and these modifications. Interestingly, in HL60, relatively low level but significant enrichments were also obtained at the 5' end of the SCL gene although SCL is not expressed in HL60; in HPB-ALL, however, no significant enrichments were seen at the 5' end of SCL. Furthermore, although neither HPB-ALL nor HL60 express SCL, a region near the +51 erythroid enhancer, at +53, showed high levels of di- and trimethylation for H3 K4. This data further supports the findings obtained with H3 acetylation in these cell lines, suggesting that there may be some genomic sequences

in this region, which are active even in the absence of SCL expression. These sequences will be discussed in detail in chapter 7.

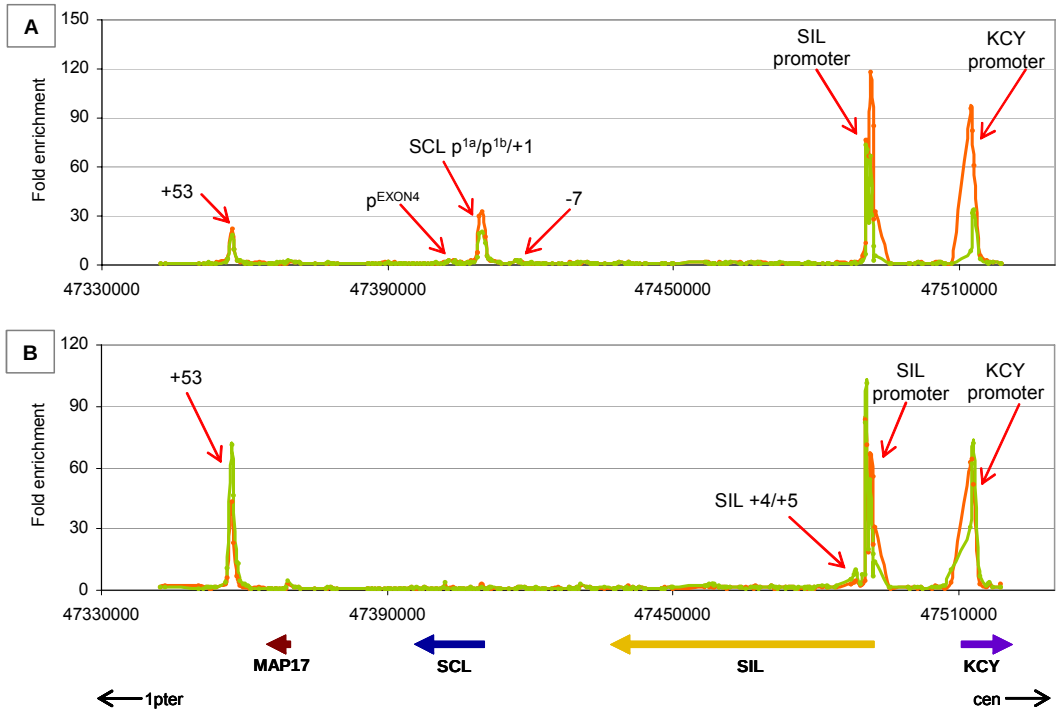


Figure 5.13: ChIP-chip profiles for di- and trimethylated lysine 4 of H3 across the SCL locus in HL60 and HPB-ALL. Panel A: HL60 and panel B: HPB-ALL. The orange profiles in panels A and B represent trimethylation of H3 K4 and the green profiles represent dimethylation of H3 K4. The genomic regions showing enrichments are marked with red arrows. The x-axes represent genomic coordinates along human chromosome 1 and the y-axes represent fold enrichments. The regions are denoted based on their distance upstream (-) or downstream (+), in kilobases, from the promoter of the nearest gene. The thick, coloured arrows represent the gene order and the direction of transcription. The orientation of the locus with respect to the centromere (cen) and telomere (ter) on human chromosome 1 is shown with black arrows at the bottom of the figure.

5.9.2.3 Dimethylation and trimethylation at histone H3 lysine 4 in SCL expressing and non-expressing mouse cell lines

Figure 5.14 shows the composite ChIP-chip profiles for dimethylation and trimethylation of H3 K4 across the SCL locus in 416B and mouse E14 ES cell line. As seen in human cell lines, 416B showed significant enrichments for di- and trimethylation of H3 K4 at the 5' ends of the active SIL and SCL genes - with highest relative levels of enrichments for trimethylation at the promoters, and the highest levels for dimethylation present immediately downstream of the promoters. Similarly, significant enrichments for di- and trimethylation were obtained at the 5' end of the SIL and SCL genes in E14 ES cells although SCL is not expressed in this cell line. This again supports previous results obtained for H3 acetylation, suggesting that modifications normally associated with gene activation (in this instance H3 K4 di- and

tri- methylation) can occur in ES cells, in the absence of gene expression (see section 5.6.2). However, unlike 416B and the human cell lines, the highest enrichments for di- and trimethylation in E14 ES cells were found to be over the same regions at the 5' end of the SCL gene (Figure 5.14, panel B).

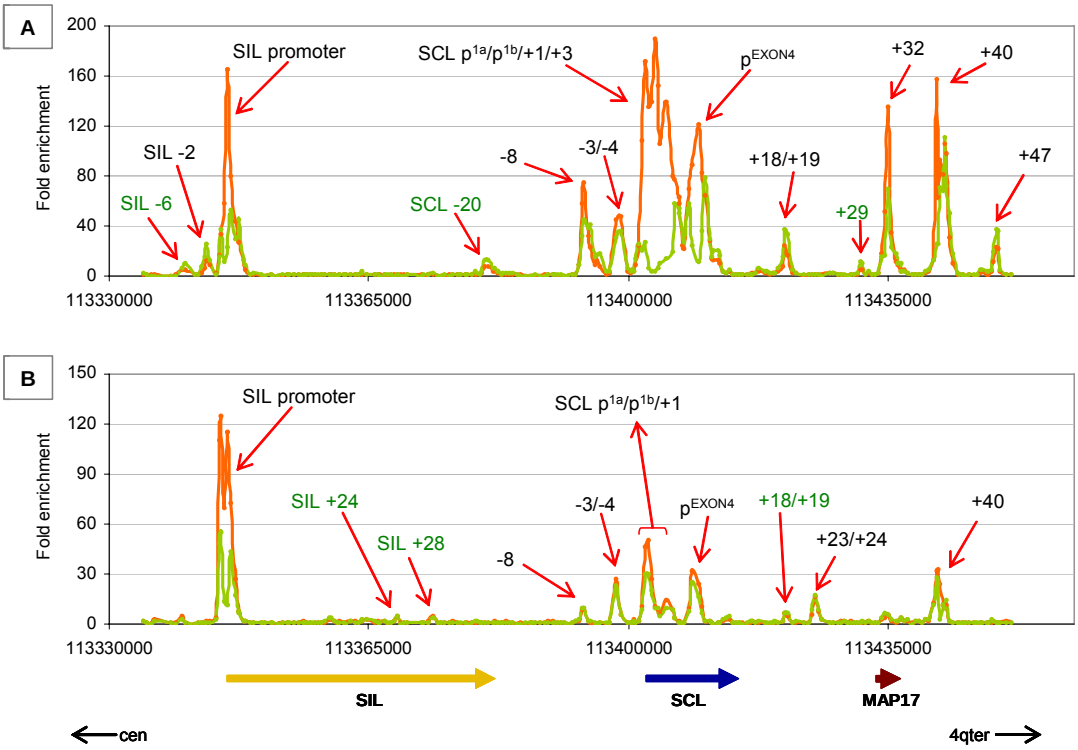


Figure 5.14: ChIP-chip profiles of di- and trimethylated lysine 4 of H3 across the SCL locus in 416B and mouse E14 ES cell line. Panel A: 416B. Panel B: E14 ES cell line. The orange profiles represent trimethylation of H3 K4 and the green profiles represent dimethylation of H3 K4. The genomic regions showing enrichments are marked with red arrows. The x-axes represent genomic coordinates along mouse chromosome 4 and the y-axes represent fold enrichments. The regions are denoted based on their distance upstream (-) or downstream (+), in kilobases, from the promoter of the nearest gene. The regions annotated in green text were identified by visual assessment of the ChIP-chip profiles. The thick, coloured arrows represent the gene order and the direction of transcription. The orientation of the locus with respect to the centromere (cen) and telomere (ter) on mouse chromosome 4 is shown with black arrows at the bottom of the figure.

A number of other known SCL regulatory regions were found to be significantly enriched for H3 K4 di- and trimethylation in 416B and mouse E14 ES cell lines (Table 5.4). Some of these regions were found to be enriched in both cell lines, but others were specifically enriched for one or the other cell line. For instance, the +47 region was significantly enriched in 416B whereas the neural enhancer at +23/+24 showed significant enrichments in E14 ES cells (Figure 5.14). Both cell lines also identified regions within the cytochrome P450 genes (Appendix 11). As previously seen for monomethylation of H3 K4, a number of novel regions were identified with ChIP

enrichments, which could be seen by visual assessment of the profiles but were not statistically significant. These regions included SIL -6, SCL -20 and SCL +29 in 416B cell line and SIL +24, SIL +28 and the stem cell enhancer at +18/+19 (known) in mouse ES cell line. The functions of these regions (apart from the stem cell enhancer) are not known.

5.9.3 Trimethylation of histone H3 lysine 79 (H3 K79Me3) marks the immediate 3' downstream regions of promoters of active genes

Given that the role of histone H3 K79 trimethylation is not known in human cells, ChIP-chip experiments were performed in four human cell lines using an antibody raised to trimethylated lysine 79 of histone H3 (Table 5.1). Figure 5.15 shows ChIP-chip profiles generated from these experiments in all cell lines across SCL, SIL and KCY genes. The ChIP profiles generated across the entire locus from each cell line have been shown in Appendix 9.

In all of the cell lines, significant enrichments were found at the immediate 3' downstream regions of the promoters of active genes. In K562, the SCL gene showed enrichments over a genomic region which spanned from +3 to +6 (with respect to SCL p^{1a}). Similarly, in Jurkat, enrichments were seen at the genomic region spanning the +2 to +6 regions. However, additional enrichments spanning the SCL +10 region were seen in Jurkat, which were downstream of the SCL p^{EXON4} (located at +7). This result correlates with the known active and inactive states of this promoter in Jurkat and K562 respectively. Conversely, HL60 and HPB-ALL, both of which do not express SCL, showed no significant enrichments across the SCL gene. In all four cell lines, significant enrichments were also obtained in the regions downstream of the SIL promoter (from +1 to +3 region seen as two separate peaks in Figure 5.15, panel B) and also the KCY promoter (from +2 to +4). Taken together, these results suggest that the trimethylation at lysine 79 of histone H3 marks the regions immediately 3' downstream of the promoters of active genes.

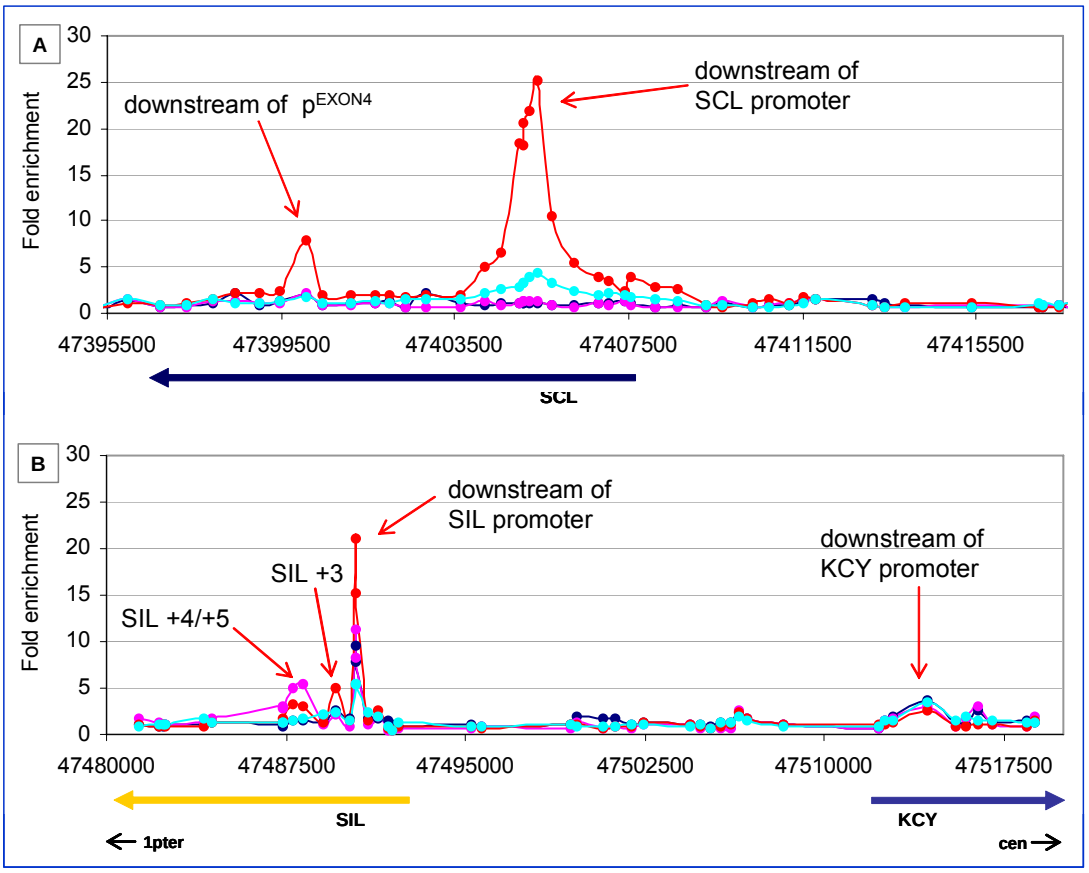


Figure 5.15: ChIP-chip profiles for histone H3 K79 trimethylation across the SCL locus in four human cell lines. The profile in panel A represents only the region at the 5' end of SCL gene and the panel B represents the genomic regions across the 5' ends of KCY and SIL genes. Four colours in the profiles shown above represent the four cell lines: turquoise (K562), red (Jurkat), blue (HL-60) and pink (HPB-ALL). The x-axes represent the genomic coordinates along human chromosome 1 and the y-axes represent fold enrichments. The thick, coloured arrows at the bottom of each panel represent the gene order and the direction of transcription with respect to the centromere (cen) and telomere (ter) on human chromosome 1 is shown with black arrows at the bottom of panel B.

5.10 Histone H3 phosphorylation at the SCL locus in SCL expressing and non-expressing human cell lines

Histone H3 phosphorylated at serine 10 and 28 occurs during cell division and is also related to the regulation of transcription (Mahadevan et al. 1991; Sassone-Corsi et al. 1999). Figure 5.16 shows the ChIP-chip profiles for phosphorylation of H3 serine 10 across the 5' end of the SCL gene in all four human cell lines (complete ChIP-chip profiles across the SCL locus are shown in Appendix 10).

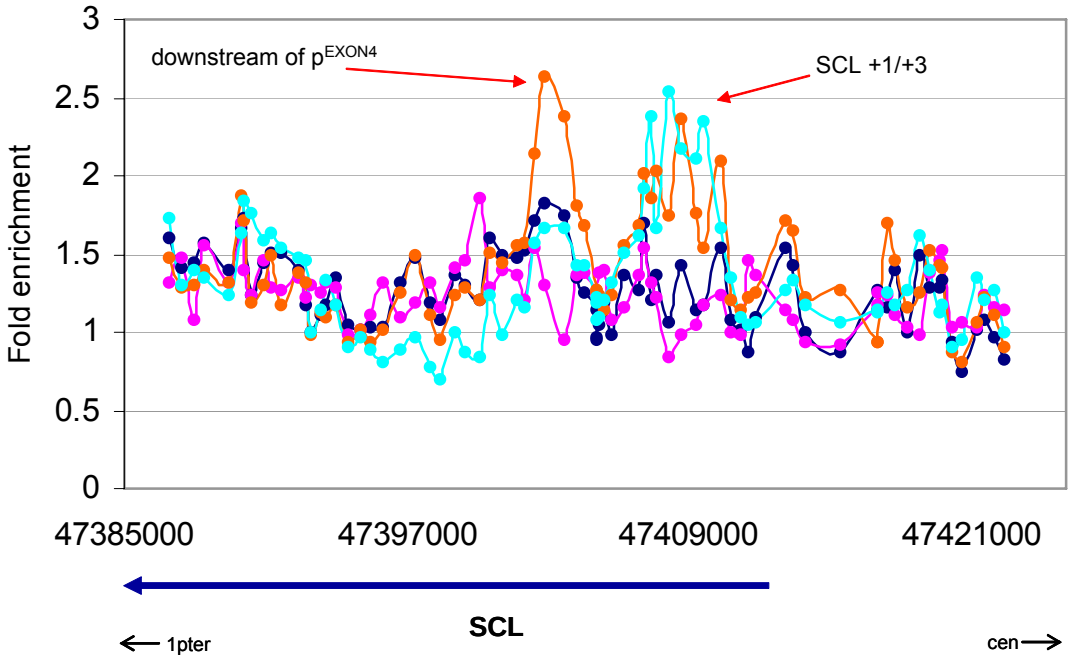


Figure 5.16: ChIP-chip profiles for H3 serine 10 phosphorylation across the 5' end of the SCL gene in four human cell lines. The figure shows ChIP-chip profiles across a small genomic region at the 5' end of the SCL gene only. The four coloured lines in the profile shown above represent the four cell lines: turquoise (K562), orange (Jurkat), blue (HL-60) and pink (HPB-ALL). The x-axis represents the genomic coordinates along human chromosome 1 and the y-axis represents fold enrichments. The thick, coloured arrow at the bottom of the figure represents the SCL gene and the direction of its transcription. The orientation of the gene with respect to the centromere (cen) and telomere (ter) on human chromosome 1 is shown by black arrows at the bottom of the figure.

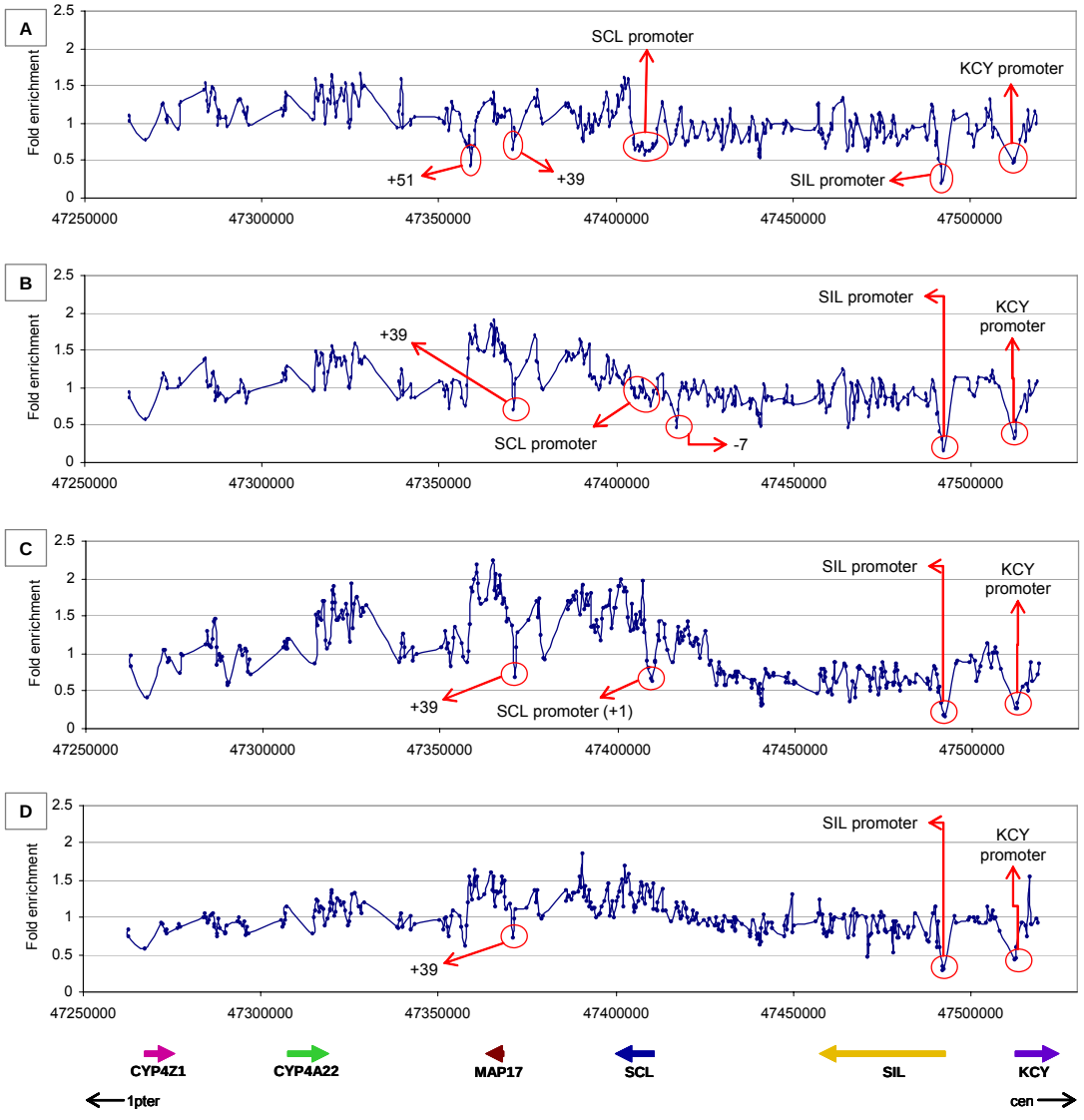
At the 5' end of the SCL gene in K562 and Jurkat, increased levels of enrichments for phosphorylation were observed at +1 and +3 regions located downstream of p^{1a} and p^{1b}, whereas such increased levels were not observed for HL60 and HPB-ALL. In addition, increased levels of enrichment were also observed at the downstream region of p^{EXON4} only in Jurkat and not in the other cell lines. Taken together, it could be interpreted that phosphorylation at serine 10 of H3 marked the downstream regions of active promoters in these cell lines. However, similar pattern of increased levels of enrichments were not observed at the 5' ends of SIL and KCY genes (Appendix 10) suggesting that phosphorylation of H3 S10 downstream of promoters may not be a widespread mechanism applicable to all genes.

5.11 Analysis of nucleosome density at the human SCL locus in Jurkat, HL60 and HPB-ALL cell lines

In chapter 4, the results obtained in ChIP-chip for histone H3 levels across the human SCL locus in K562 (see section 4.7.2, chapter 4) demonstrated that decreased levels of nucleosomes coincide with the regulatory and coding regions of active genes. Based on

these results, the levels of histone H3 were also analysed in the other three human cell lines to determine whether this relationship was present in other cell types. Figure 5.17 shows the ChIP-chip profiles across the SCL locus in K562, Jurkat, HL60 and HPB-ALL cell lines. Panel E of the figure represents the composite ChIP-chip profiles at the 5' end of SCL gene and the +51 erythroid enhancer in the four cell lines.

Decreased relative levels of histone H3 were detected at, and close to the promoter regions of the KCY and SIL genes in all the three cell lines (results for K562 have been discussed in chapter 4). At the 5' end of the SCL gene, decreased relative levels of histone H3 were observed in Jurkat at the -7 region, the promoter region p^{1b}, +1 and +3 regions and also at the +7 region encompassing p^{EXON4} (Figure 5.17, panel E). It has been shown that the -7 region exhibited strong DNase I sensitivity in Jurkat (Leroy-Viard et al. 1994) and that p^{EXON4} is active in Jurkat (Bernard et al. 1992).



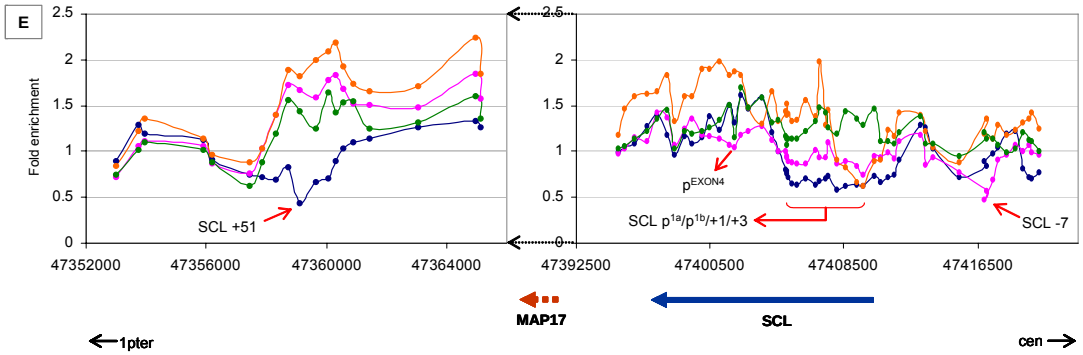


Figure 5.17: ChIP-chip profiles for histone H3 levels across the SCL locus in four human cell lines. Panels A: K562, panel B: Jurkat, panel C: HL60 and panel D: HPB-ALL. The genomic regions with decreased levels for histone H3 are marked with red circles. Panel E shows the ChIP-chip profiles of selective regions across the SCL locus; the panel shows a small genomic region at the 5' end of the SCL gene (right panel) and the genomic region across the erythroid enhancer at +51 (left panel). The four colours of the profiles in panel E represent the four cell lines: blue (K562), (pink) Jurkat, (orange) HL60, (green) HPB-ALL. The genomic regions with decreased levels of histone H3 in these regions are marked with red arrows. The x-axes represent genomic coordinates along the human chromosome 1 and the y-axes represent fold enrichment. The two plots of panel E are joined by dotted arrows in the middle to indicate the continuity between the two profiles. The thick, coloured arrows at the bottom of the panels D and E and the brown coloured dotted arrow in panel E represent the gene order and the direction of transcription. The orientation of the locus with respect to the centromere (cen) and telomere (ter) on human chromosome 1 is shown with black arrows at the bottom of panels D and E.

In HL60, decreased levels were detected only at the +1 region coincident with the region of H3 K9/14 diacetylation seen in this region (see section 5.6.1.2). In HPB-ALL, decreased relative levels were not detected at these regions. In addition, decreased relative levels for histone H3 were also detected at the +39 region in all cell lines which was also seen in K562 (Figure 5.17). In Jurkat, HL60 and HPB-ALL cell lines, decreased levels of histone H3 were not observed at the erythroid enhancer at +51 (Delabesse et al. 2005) which was previously seen in K562 (Figure 5.17, panel E). This observation correlates with the activity of this enhancer only in K562.

Interestingly, it was noted that there were noticeably elevated ratio on either side of the regulatory regions which reported a drop in histone H3 levels. Taken together, these results suggested that, as previously seen for K562, depletion of nucleosomes (reflected by decreased levels of histone H3) occurred at active regulatory regions across the SCL locus in Jurkat, HL60 and HPB-ALL cell lines. The results of these experiments are analyzed further in chapter 6.

5.12 SCL itself binds to the erythroid enhancer along with GATA-1 and Ldb-1 in K562: evidence for the involvement of the SCL multi-protein complex in SCL regulation.

As described in chapter 4, *in vivo* interactions of GATA-1 transcription factor across the SCL locus were identified at nine genomic regions, all of which had conserved and/or non-conserved GATA binding sites. In addition, one array element, at the +51 erythroid enhancer (homologous to mouse +40 region), which showed the highest ChIP enrichments for GATA-1, contained three highly conserved GATA sites and also binding sites for other transcription factors (chapter 4, section 4.6.4). One of these GATA sites was found to be contained within a 20 bp sequence containing a GATA/E-box composite site which showed the canonical hallmarks of the SCL erythroid DNA-binding complex which includes the SCL, E47, GATA-1, LMO2 and Ldb-1 transcription factors (Wadman et al. 1997). Based on this information, ChIP-chip experiments were performed in K562 using antibodies raised to all the proteins involved in the SCL multi-protein complex (Table 5.1). Although ChIP experiments were performed using antibodies raised to all of the above mentioned transcription factors, however, significant enrichments were obtained only with anti-SCL and anti-Ldb-1 antibodies.

Figure 5.18 shows the ChIP-chip profiles generated for GATA-1 (the results have already been discussed in chapter 4, section 4.6.2), SCL and Ldb-1 transcription factors across the SCL locus in K562 cell line. For both SCL and Ldb-1, significant and the highest enrichments were obtained at the +51 region and at the same array elements which were also enriched for GATA-1. These data suggested that at least three of the components of the SCL erythroid DNA binding complex (Wadman et al. 1997) bind to the +51 region. More importantly, this suggests that SCL, in a complex with other transcription factors, regulates the activity of its own expression.

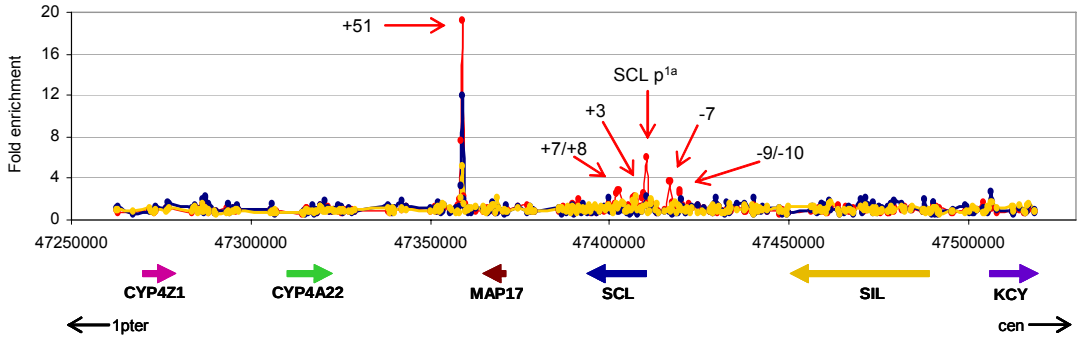


Figure 5.18: Composite ChIP-chip profiles for GATA-1, SCL and Ldb-1 across the SCL locus in K562. Enrichments for GATA-1, SCL and Ldb-1 are represented by red, yellow and blue profiles respectively. The genomic regions enriched for these transcription factors are marked by red arrows and discussed in the text. The x-axis represents the genomic coordinates along human chromosome 1 and the y-axis represents fold

enrichments. The thick, coloured arrows at the bottom of the figure represent the gene order and the direction of transcription. The orientation of the locus with respect to the centromere (cen) and telomere (ter) on human chromosome 1 is shown with black arrows at the bottom of the figure.

5.13 Mapping interactions of proteins with HAT and HDAC activities across the SCL locus in K562

Proteins that catalyse reversible acetylation of histones; histone acetyltransferases (HATs) and histone deacetylases (HDACs) exist as multi-protein complexes that have co-activator and co-repressor activities respectively. Some HATs are also able to acetylate non-histone transcription factors involved in haematopoiesis (Blobel et al. 2000; Huang et al. 1999; Huang et al. 2000). A number of ChIP-chip assays were tested to identify the binding of these proteins at the SCL locus (Table 5.1). Although, ChIP experiments using antibodies raised to these proteins were performed in all cell lines, many of these antibodies did not exhibit significant ChIP enrichments across the SCL locus. However, significant ChIP enrichments were obtained for some of these factors in K562 cell line and the results obtained with p300, CBP and HDAC2 are presented below (Figure 5.19). HDAC2 is a histone deacetylase whereas p300 and CBP possess intrinsic HAT activity. p300 and CBP are also known to be important cofactors for a number of transcription factors both within and outside the haematopoietic system (such as GATA-1 and SCL) (Blobel et al. 2000) and share high sequence homology which suggests that the antibodies raised to these proteins may be cross-reactive.

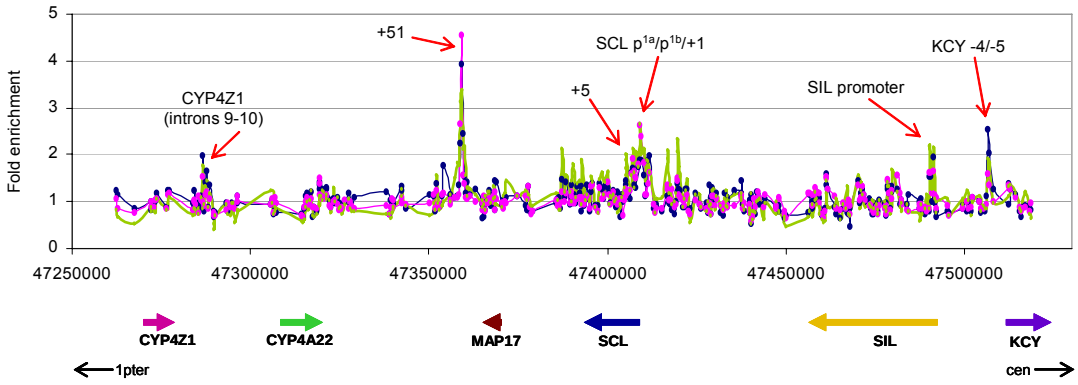


Figure 5.19: Composite ChIP-chip profiles for p300, CBP and HDAC2 across the SCL locus in K562. ChIP enrichments for p300, CBP and HDAC2 are represented by blue, pink and green profiles respectively. The genomic regions showing enrichments for these regulatory proteins are marked by red arrows. The x-axis represents genomic coordinates along human chromosome 1 and the y-axis represents fold enrichments. The thick, coloured arrows at the bottom of the figure represent the gene order and the direction of transcription. The orientation of the locus with respect to the centromere (cen) and telomere (ter) on human chromosome 1 is shown by the black arrows at the bottom of the figure.

Significant enrichments for all three factors were obtained at the +51 erythroid enhancer and at the 5' end of the SCL gene at -1, p^{1a}, p^{1b}, +1 and +3 regions. Additionally,

significant enrichments for all three factors were obtained at the 5' end of the SIL gene but not at the 5' end of the KCY gene. However, significant enrichments for all three factors were obtained at KCY -4/-5 region. The presence of factors with HAT and HDAC activity at the same regions is consistent with their role in maintaining a dynamic acetylation state (Kurdistani and Grunstein 2003).

For HDAC2, significant enrichments were also obtained at the +5 region which were not observed with p300 or CBP. Across the SCL gene, it was observed that the ChIP enrichments for HDAC2 (green profile in figure 5.19) showed several peaks at the downstream as well as upstream region of the 5' end of the gene and the enrichments obtained at these regions were above the significant threshold value. Most of these regions with significant enrichments mapped to known regulatory elements of SCL at -9/-10, -7, and +20/+21 (stem cell enhancer). A region with significant enrichments was also identified within the genomic region containing the CYP4Z1 gene (between introns 9-10) for all three factors but this region did not map to any region with known regulatory function.

5.14 Mapping interactions of factor involved in the pre-initiation complex across the SCL locus in K562

General transcription factor TFIID and RNA pol II are both involved in the assembly of the pre-initiation complex at the promoters (Burley and Roeder 1996) and, of all the components, TFIID is the only component that is capable of binding specifically to the core promoters (also see chapter 1). $Taf_{II}250$ which is a subunit of TFIID is an important cofactor also shows HAT activity *in vitro* (Mizzen et al. 1996). ChIP-chip experiments were performed in human cell lines to map the interactions of RNA pol II and $Taf_{II}250$ at the active promoters. However, no significant ChIP enrichments were obtained for RNA pol II in any cell line tested. For $Taf_{II}250$, significant enrichments were observed only in the K562 cell line.

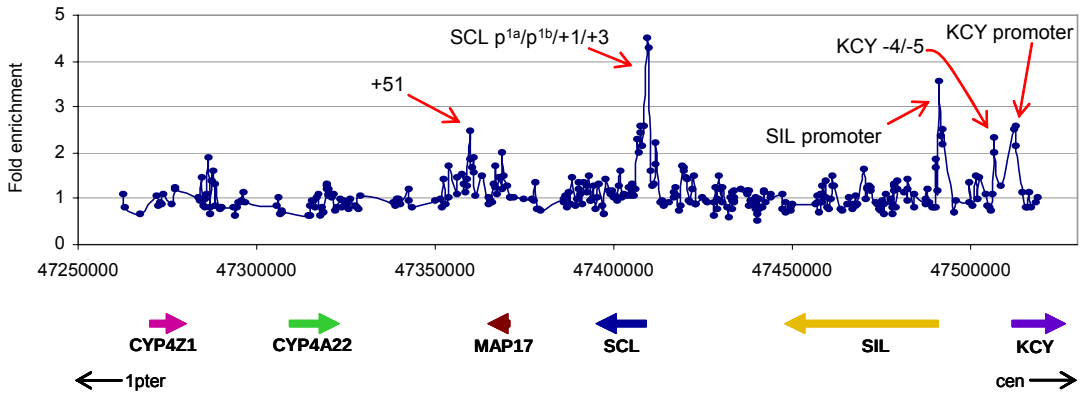


Figure 5.20: ChIP-chip profile for $Taf_{II}250$ across the SCL locus in K562. The genomic regions with significant enrichments are marked with red arrows. The x-axis represents genomic coordinates along human chromosome 1 and the y-axis represents fold enrichments. The thick, coloured arrows at the bottom of the figure represent gene the order and the direction of transcription. The orientation of the locus with respect to the centromere (cen) and telomere (ter) on human chromosome 1 is shown by the black arrows at the bottom of the figure.

In K562, significant enrichments for $Taf_{II}250$ were obtained at or near the 5' ends, i.e. promoters, of KCY, SIL and SCL genes (Figure 5.20). The enrichments at the 5' end of SCL encompassed p^{1a}, p^{1b}, +1 and +3 regions. In addition, significant enrichments were obtained at the +51 erythroid enhancer. However, the highest enrichments were not observed on the array element which showed highest enrichments for GATA-1, SCL and Ldb-1 but on the adjacent array element (500 bp proximal to GATA-binding towards the SCL gene). At the 5' end of the SIL gene, significant enrichments for $Taf_{II}250$ were observed at two distinct regions, one at the promoter and another one approximately 2 kb downstream of the SIL promoter (SIL +2). In the ENSEMBL data base (http://www.ensembl.org), three transcripts for the SIL gene have been annotated and the 5' end of one of these transcripts maps to a region approximately 2 kb downstream of the known SIL promoter, raising the possibility of an alternate promoter for the SIL gene (see also chapter 7). In addition, significant enrichments were also obtained at the KCY -4/-5 region which had previously been identified with various histone modifications but no regulatory function for this region is known.

5.15 Elf-1 binds to the SIL -1 region in K562, Jurkat and HPB-ALL

Elf-1 is mainly a lymphoid specific Ets transcription factor (Wang et al. 1993) involved in the regulation of the SCL stem cell enhancer (+20/+21 region) (Gottgens et al. 2002) and the endothelial enhancer (-3/-4 region) (Gottgens et al. 2004). It was therefore tested in ChIP-chip assays in all cell lines. Figure 5.21 shows a composite ChIP-chip profile for Elf-1 in K562, Jurkat and HPB-ALL.

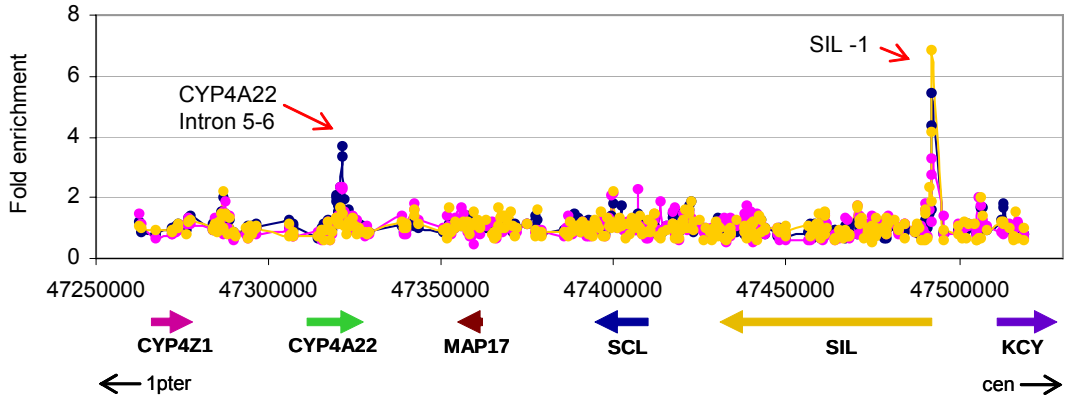


Figure 5.21: Composite ChIP-chip profile for Elf-1 across the SCL locus in K562, Jurkat and HPB-ALL. ChIP enrichments for Elf-1 in K562, Jurkat and HPB-ALL are represented by blue, pink and yellow profiles respectively. The genomic regions showing enrichments are marked with red arrows. The x-axis represents genomic coordinates along human chromosome 1 and the y-axis represents fold enrichments. The thick, coloured arrows at the bottom of the figure represent the gene order and the direction of transcription. The orientation of the locus with respect to the centromere (cen) and telomere (ter) on human chromosome 1 is shown by the black arrows at the bottom of the figure.

Significant enrichments for Elf-1 were obtained at the 5' end of the SIL gene with the highest enrichments at the SIL -1 region in all three cell lines, and additional low level enrichments were observed at SIL +1 region only in HPB-ALL. In addition, significant enrichments were also seen in a region mapping to the CYP4A22 gene (introns 5-6) in K562 and Jurkat, but not in HPB-ALL.

5.16 CTCF binding at the SCL locus in K562 and Jurkat

The MAP17 gene is located between the SCL erythroid enhancer at +51 and the SCL gene itself. Furthermore, the +51 region is closer to the CYP4A22 gene than it is to SCL. Thus, in order to ensure appropriate expression of the SCL gene by this distal enhancer, a mechanism must exist to ensure that the erythroid enhancer interacts with the SCL promoter and does not inappropriately enhance the expression of MAP17 or CYP4A22. Insulators are regulatory sequences that are known to possess dual roles: they form boundaries between regions of actively and inactively expressed genes, as well as modulating interactions between enhancers and promoters. The zinc-finger DNA-binding protein, CTCF, is known to bind to a number of mammalian insulators characterized to date (Bell et al. 1999). In addition to its enhancer blocking function (Bell et al. 1999), CTCF also plays roles in transcriptional regulation (Klenova et al. 1993), and gene imprinting (Bell and Felsenfeld 2000). Based on all of this information, ChIP-chip experiments were performed in two SCL expressing cell lines, K562 and Jurkat, to determine whether CTCF binding was involved in regulating the activities of the SCL erythroid enhancer.

In K562, significant and highest enrichments for CTCF were obtained at the +57 region (Figure 5.22, panel A). In addition, the +39 region and p^{MAP17} also showed significant enrichments for CTCF. A genomic region in the CYP4A22 gene (introns 5-6) was also significantly enriched. In Jurkat, CTCF interactions were identified only at the +57 region (Figure 5.22, panel B). In addition, the array elements exhibiting highest enrichments for CTCF in K562 and Jurkat were different. Another genomic region at the CYP4Z1 gene (introns 8-9) was also found to be significantly enriched in Jurkat, the function for which is not known. The +57 region had been identified previously with significant enrichments for di- and trimethylation at lysine 4 of histone H3 and decreased levels of H3 K16

acetylation, in K562 and Jurkat. Additionally, the +39 region was also found to be enriched for various histone modifications in both the cell lines. The +39 region also displayed nucleosome depletion. Given the facts that (i) the genomic region downstream of +57 contains the CYP4A22 and CYP4Z1 genes (both of which are silent), and (ii) the MAP17 promoter and +39 region is bound with CTCF, it is likely that CTCF binding may contribute to the regulation of the various regulatory sequences in this region as well as serve to demarcate an insulator element at +57.

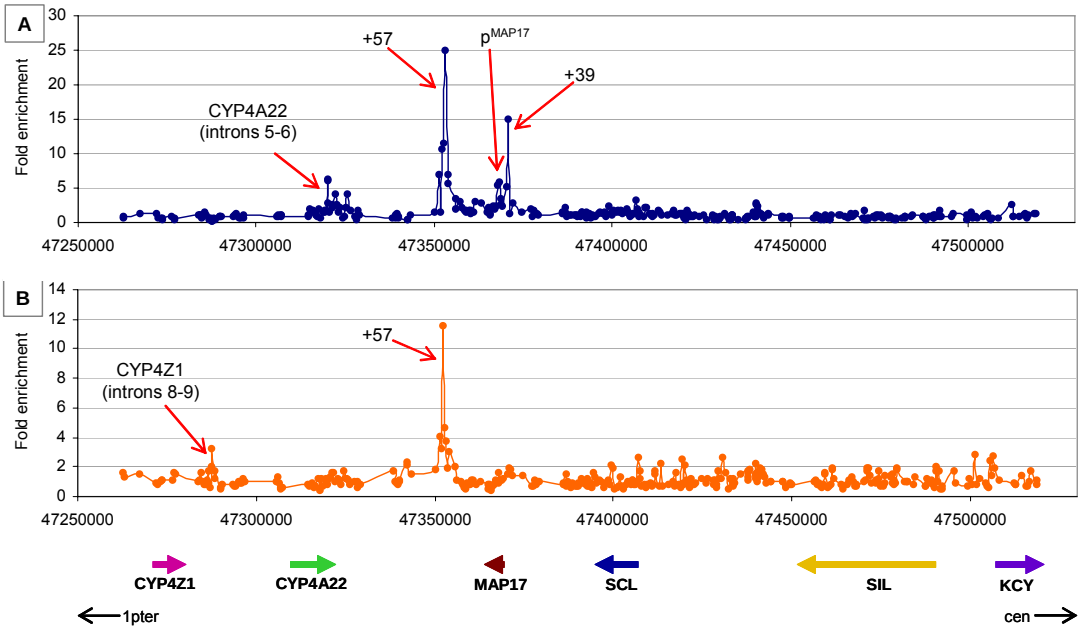


Figure 5.22: ChIP-chip profile for CTCF binding across the SCL locus in K562 and Jurkat. Panel A: K562, panel B: Jurkat. The genomic regions showing significant enrichments for CTCF are marked with red arrows. The x-axes represent the genomic coordinates along human chromosome 1 and the y-axes represent fold enrichments. The thick, coloured arrows at the bottom of panel B represent the gene order and the direction of transcription. The orientation of the locus with respect to the centromere (cen) and telomere (ter) on human chromosome 1 is shown at the bottom of the figure.

5.17 Discussion

5.17.1 Patterns of histone acetylation across the human and mouse SCL locus

The histone H3 and H4 acetylation patterns (using assays for H3 K9/14 diacetylation and H4 K5/8/12/16 tetra-acetylation) that were obtained for the four human haematopoietic cell lines and the mouse 416B line were in complete agreement with those described in a PCR-based analysis of a smaller genomic region across the SCL locus (Delabesse et al. 2005). Furthermore, the distribution of histone H3 and H4 acetylation were consistent with the location of known regulatory elements for SCL and its flanking genes in all of the respective cell lines. It has been reported that the acetylation patterns for histone H3 and H4 are very similar (Roh et al. 2004; Litt et al. 2001) and correlated with transcriptional

activity of a gene; the data obtained across the SCL locus were in agreement with these previous observations.

Hyperacetylation of all specific lysine residues of histone H3 (H3 K9, H3 K14, H3 K18 and H3 K27) was highly correlated with the transcriptional status of a gene which is consistent with similar observations in yeast (Kurdistani et al. 2004; Liu et al. 2005). These modifications mostly followed the pattern of H3 K9/14 diacetylation and marked virtually all the known and active regulatory regions in the human cell lines. In contrast, correlation of acetylation of specific lysine residues of H4 (H4 K5, H4 K8, H4 K12 and H4 K16) with the transcriptional status of the gene was variable. The data obtained in this study suggested that hyperacetylation at lysine 5 of H4, and hypoacetylation at lysines 8 and 16 of H4 were correlated with the transcriptional activity of a gene. Similar results have been observed in yeast with the existence of a hypoacetylated domain for these modifications (H4 K8 and H4 K16) adjacent to the transcription start sites (Liu et al. 2005). It has also been shown by mutating yeast H4 lysines 5, 8, 12 and 16 that only lysine 16 mutations had specific transcriptional consequences independent of the mutational state of the other lysines. This suggests that out of all the lysine modifications of histone H4 N-terminal tail, only the lysine 16 modification is functionally non-redundant (Dion et al. 2005).

Results obtained in the present study for tetra-acetylation of H2A at lysines 5/9/13/15 and H2B at lysines 5/12/15/20 in K562 were in agreement with the observations at the chicken β-globin locus (Myers et al. 2003), which suggested that H2B specifically marks tissue-specific genes. H2A tetra-acetylation pattern was similar to that of histone H3 K9/14 diacetylation for all genes in K562, whereas H2B tetra-acetylation specifically marked the SCL gene and not the SIL gene which is a ubiquitously expressed gene. However, with H2B tetra-acetylation, low level but significant enrichments were observed at the 5' end of KCY gene which is also a ubiquitously expressed gene. This suggests that H2B tetra-acetylation may also have a role in the regulation of housekeeping genes.

5.17.2 Patterns of histone H3 methylation across the human and mouse SCL locus

Patterns of histone H3 K4 methylation at the 5' ends of genes also correlated with the transcriptional activity of genes in the human and mouse cell lines studied here. Di- and trimethylation enrichment levels of H3 K4 were highest at the 5' ends of active genes, i.e., at or near the active promoters. In addition, enrichments for trimethylation were highest over the promoters, but for dimethylation, the enrichments were highest just downstream of the promoter region. Conversely, at active promoters, monomethylation enrichment levels were very low, but rose on either side (see below). These observations

are consistent with the idea that mono-methyl lysine 4 residues are converted to di- and then trimethyl residues to accompany gene activation at promoter regions.

Similar associations of trimethylation of H3 K4 with the promoter regions of active genes have been well documented (Santos-Rosa et al. 2002; Schneider et al. 2004; Liu et al. 2005; Schübeler et al. 2004; Pokholok et al. 2005). However, the presence of dimethylation has not been linked clearly to transcriptional activity. It has been reported that dimethylation of histone H3 K4 is enriched at active and inactive regions (Santos-Rosa et al. 2002; Schneider et al. 2004) with a preferential enrichment over the coding regions of active genes (Bernstein et al. 2002; Liu et al. 2005; Pokholok et al. 2005). The data obtained in this study demonstrated that both di- and trimethylation were enriched over similar regions, with higher levels of enrichment at the 5' ends and with both modifications extending into the coding region to the same extent.

While lack of enrichments for histone H3 K4 monomethylation was found to be associated with the active SCL promoters in K562, Jurkat and 416B (see above section), increased levels of enrichments were found at the SCL promoters in the HL60 and mouse ES E14 cell lines – neither of which express SCL. It has been suggested that when genes are repressed, dimethylation is present but trimethylation is absent and when genes become active, trimethylation appears but dimethylation persists (Santos-Rosa et al. 2002). Thus, relative increase of trimethylation versus dimethylation discriminates the active from the inactive genes (Schneider et al. 2004). However, based on the data obtained in the present study, this suggestion can be extended further by including the relative levels of enrichments for monomethylation which means that instead of di- and tri-, it is possibly the relative enrichments for monomethylation which discriminates active from inactive promoters (see chapter 6).

Increased enrichment levels of monomethylation were found to be associated with other types of regulatory regions including active enhancers. Furthermore, increased levels of enrichments for monomethylation have been associated with the 3' ends of the active genes in yeast (Liu et al. 2005; Pokholok et al. 2005), although this was not apparent across the SCL locus. Overall, the ChIP-chip data obtained for histone H3 lysine 4 methylation in the present study suggests that the distribution of these modifications at the SCL locus are more complex than that previously reported in yeast. Thus, extrapolations of H3 K4 methylation data from yeast to human may not be accurate, given the increased complexity of regulatory interactions found in mammalian genomes.

The data obtained with trimethylation of H3 K79 suggested that this histone mark is linked to transcriptional activity which was not observed in previous studies (Pokholok et

al. 2005). However, the presence of this histone mark at the downstream regions of active promoters suggested that this modification of H3 may be linked to transcriptional elongation. It has been suggested that the Paf1 complex which is involved in transcriptional elongation is also required for methylation at lysine 79 of H3 (Krogan et al. 2003). Taken together, these data provide support for trimethylation of H3 K79 to possibly play a role in transcriptional elongation in mammalian cells.

5.17.3 Histone H3 acetylation and methylation modifications can also be associated with inactive genes

Surprisingly, histone modifications associated with gene activation were found at the inactive SCL locus in HL60 and E14 ES cells. In HL60, subtle, but significant, ChIP enrichments for H3 K9/14 diacetylation were found at the 5' end of the SCL gene. Histone H3 K9/14 diacetylation patterns in E14 ES cells were found to have similarities with those of HL60, although in E14, almost all of the SCL regulatory elements were acetylated albeit at low levels. In addition to the acetylation marks, the SCL promoter also showed significant enrichments for mono-, di- and trimethylation at lysine 4 of histone H3 in HL60 and in E14 ES cells. These data are not consistent with the established relationships between H3 acetylation, H3 K4 methylation and gene activity. However, the relationship between histone H3 K4 methylation and gene activity will be further explored in chapter 6.

The presence of "activating" histone modifications (H3 acetylation and H3 K4 di- and trimethylation), at the SCL gene when it is not expressed, suggest two interesting hypotheses which both have supporting evidence from other studies found in the literature as described below.

HL60 is a promyelocytic cell line which lies downstream of the granulo/myelopoietic progenitors (GMP) in the hierarchy of blood development and SCL is known to be expressed in GMP cells. It is possible that HL60 cells may still be carrying histone acetylation and methylation marks although the cells have differentiated to monocyte precursors and SCL expression has been turned off. This suggests that the silencing of the SCL is not completely dependent on its modification status and other factors are involved in turning its expression off. Similar low levels of acetylation have also been reported over regulatory elements and hypersensitive sites across the β-globin locus in a cell line where the gene was not expressed (Litt et al. 2001).

These results also suggest that regulatory regions in mouse ES cells may already be primed to specify the activation of SCL gene when pluripotent cells become blood progenitors. Similarly, histone H3 acetylation and histone H3 lysine 4 methylation at a *cis*-acting element at the mouse λ5-VpreB1 locus (involved in B-cell development) has

been shown in mouse E14 ES cells prior to the gene becoming active (Szutorisz et al. 2005); this element was then actively involved in the recruitment of regulatory proteins during the differentiation of ES cells through to the pro-B cell stage. Therefore, it seems likely that the marking of regulatory regions for histone marks may be fundamental to the recruitment of transcription factors and the assembly of pre-initiation complex at the SCL locus when pluripotent cells differentiate into haematopoietic progenitors and SCL is turned on. It is interesting to note that the neural enhancer at the +23 region (Gottgens et al. 2000) was also detected with H3 K9/14 diacetylation in ES cells which implies that not only the haematopoietic regulatory elements may be primed, but all elements which may participate in the regulation of SCL in distinct compartments of its normal expression are primed for subsequent expression of SCL. Furthermore, in E14 cells, ChIP-chip assays for H3 K4 methylation identified a number of novel regions across the SCL locus, many of which were located within the SIL gene and were specifically enriched for monomethylation of H3 K4. Based on this observation, it could be suggested that a number of unknown regions were also marked in ES cells and which may represent novel putative regulatory regions which become active at later stages of development. Although the function of these regions is not clear, it is possible that these regions are associated with the regulation of SCL or SIL.

5.17.4 Nucleosome depletion occurs across the SCL locus in all human cell lines

The experiments performed to determine the levels of histone H3 across the SCL locus in SCL expressing and non-expressing human cell lines provide additional evidence that nucleosome depletion occurs at active genes at the SCL locus. A description of results obtained from further analytical approaches which support this evidence is provided in chapter 6.

By observing H3 histone profiles in SCL-expressing and non-expressing cell lines, differences in nucleosome levels could be correlated to the activities of various SCL regulatory regions. For example, only in K562 were depletions at the +51 erythroid enhancer observed. Thus nucleosome differences across the SCL locus between the cell lines studied here are likely to be due to real differences in gene activity rather than just general features of the different cell lines (and not related gene expression). To verify that nucleosome dynamics are responsive to gene activity, it would be necessary to observe the changes in nucleosome density which accompanies the repression or activation of genes in an appropriate cell system. This has been demonstrated in yeast (Lee et al. 2004), but not in human systems to date. However, it is still not clear how this depletion of nucleosomes occurs to make the underlying chromatin available for binding

of the RNA polymerase complex, transcription factors and other regulatory proteins. Mechanisms including sliding or complete eviction of nucleosomes have been suggested (Hamiche et al. 1999; Lorch et al. 1999; Korber et al. 2004). The results in K562 support the idea of nucleosome sliding (see chapter 4 discussion) and the analysis of the three other human cell lines described in this chapter supports this idea.

5.17.5 SCL may be self-regulated in a positive feedback loop

Significant levels of GATA-1, SCL and Ldb-1 enrichments at the +51 erythroid enhancer provided the first evidence that these three transcription factors are involved in the activity of this enhancer during erythroid development. In HSCs, activity of the SCL stem cell enhancer (+20/+21 in human) is independent of SCL expression (Sanchez et al. 1999). The data described here is therefore consistent with a model where once SCL expression has been established in HSCs, maintenance of SCL during differentiation towards erythroid cells is at least partly achieved through a positive feedback loop acting through the +51 erythroid enhancer. However, given that these interactions were shown in the cultured cell line K562, support for this model and the functional relationship between GATA-1, SCL and Ldb-1 at the erythroid enhancer would need to be confirmed in primary primitive erythroblasts, where SCL is under the control of the erythroid enhancer (Delabesse et al. 2005). The presence of the highly conserved binding site for the composite GATA/E-box of the erythroid SCL complex (Wadman et al. 1997) suggests that GATA-1, SCL and Ldb-1 may act together at this site in a multi-protein complex. To further support this evidence, it will also be important to confirm that the transcription factors E2A and LMO-2 bind to this region, once appropriate antibodies are identified which work well in ChIP.

The ChIP-chip enrichments obtained for p300/CBP and HDAC2 at the +51 region and at the SCL promoter region, in addition to the binding events described for SCL, Ldb-1 (this chapter) and GATA-1 (Chapter 4), provide evidence that the regulatory events at these regions may be very complex. GATA-1 interacts with CBP *in vivo* and *in vitro* (Weiss et al. 1997) and SCL has been found to interact with p300 (Huang et al. 1999) and HDAC1 (Huang et al. 2000). Therefore, the presence of p300/CBP and HDAC2 at sites where SCL and GATA-1 are also present is consistent with these previous observations.

5.17.6 Binding interactions of other transcription factors across the SCL locus

5.17.6.1 Elf-1 and the regulation of the SIL gene

Elf-1 is mainly a lymphoid-specific transcription factor but is also known to regulate SCL expression through its stem cell and endothelial enhancers (Gottgens et al. 2002;

Gottgens et al. 2004). ChIP-chip enrichments for Elf-1 one kilobase upstream of the SIL promoter (SIL -1 region) in both lymphoid (Jurkat and HPB-ALL) and myeloid (K562) cell types were unexpected given that the SIL gene is ubiquitously expressed. This suggests that Elf-1 may exert a haematopoietic-specific control on SIL expression, although it cannot be excluded that this interaction is found in cultured cells but not in primary cells *in vivo*. Given that there may be a number of putative novel regulatory regions located in the SIL gene which could control either SIL or SCL expression (as evidenced by regions showing various histone modifications), Elf-1 may be required to modulate SIL expression to ensure appropriate activity of these regulatory regions. The interactions of Elf-1 at the SIL -1 region needs to be further studied in haematopoietic cell types to determine whether this interaction is biologically relevant and whether these putative novel regulatory regions control SIL or SCL expression.

5.17.6.2 CTCF and its role in insulator function at the SCL locus

It was discussed in section 5.16 that based on the location of the +51 erythroid enhancer, the appropriate regulation of SCL, MAP17 and the cytochrome P450 gene cluster may require the involvement of insulator elements. CTCF is known to bind at most known mammalian insulators (Bell et al. 1999) and ChIP-chip enrichments for this transcription factor at the +57 region in both K562 and Jurkat cell lines suggest that this region may represent an insulator element at the SCL locus. It has been suggested that insulator elements possess roles as both enhancer-blockers and as barriers between chromatin domains (West et al. 2002). Whether CTCF performs both functions at the +57 region is not known. Support for its enhancer-blocking function at the SCL locus is evidenced by the fact it also binds to the MAP17 promoter in K562. These interactions indicate that the activity of the MAP17 promoter may be modulated through looping interactions with the +57 region. Alternatively, CTCF may bind directly to the promoter to control MAP17 expression. The binding of CTCF upstream from the MAP17 promoter at the +39 region in K562 suggests that +39 itself could be an insulator which partitions the MAP17 gene from the +51 SCL enhancer, on one side, and the SCL gene and its more proximal regulatory sequences, on the other side. Interestingly, CTCF interactions in Jurkat were not seen at the MAP17 promoter or at +39 - although both K562 and Jurkat express SCL and MAP17 (Delabesse et al. 2005). However, given that SCL expression in Jurkat is inappropriate, the regulation of SCL and MAP17 may be different in the two cell lines.

5.17.7 Issues in interpreting ChIP-chip data

The work described in this chapter illustrates how ChIP-chip can be used to characterize regulatory features at the SCL locus. Although ChIP-chip provides a powerful and high-

throughput method of mapping DNA-protein interactions in this way, it is worth noting at this point, some issues in interpreting the data derived from these experiments. These issues are relevant even if ChIP-quality antibodies are used in ChIP-chip assays which are specific for detecting the appropriate interactions.

i) **Heterogeneity in cell populations:** A ChIP-chip assay provides only a snapshot of *in vivo* DNA-protein interactions within a cell population at a particular point in time or under specific conditions. In other words, ChIP-chip profiles are a composite survey of an entire cell population (which are often growing in culture asynchronously) and the regulatory features observed could be occurring in different cells at any given time. This could also mean that binding events occurring in a small proportion of cells may go unnoticed because they are being masked by events in the rest of the cell population. Thus, multiple interactions occurring at the same sequence, as observed in ChIP-chip assays, does not necessarily mean that these proteins are acting concertedly in a complex within the same cells at the same time.

ii) **Resolution:** Taking this last point further, even when multiple interactions are occurring in the same cells at the same time, it can be difficult to determine whether they bind to the same regulatory element. This is a question of array resolution, and by developing an array of sufficiently high resolution, the location of binding sites can be pinpointed more accurately. Yet, even with a high resolution array as that described in this thesis, proteins can bind in close proximity to a given sequence without the factors existing in the same complex (Metivier et al. 2003).

iii) **Allelic variations in expression:** Similarly, binding interactions at specific gene alleles cannot be inferred from ChIP-chip profiles since the composite profiles report the interactions occurring at all alleles within the genome. For example, in the present study, both K562 and Jurkat showed multiple copies of the SCL locus. Although it is known that SCL expression is mono-allelic in Jurkat (Leroy-Viard et al. 1994), even the allele which does not express SCL may show regulatory events which differ from that at the expressing locus. However, all of these interactions will appear in the composite profile for all of the alleles in a given cell.

iv) **Indirect DNA-protein interactions:** It is possible that enrichments obtained in ChIP-chip are due to regulatory proteins (such as transcription factors) bound to one genomic sequence (for example an enhancer) interacting to a second sequence (for example a promoter) via chromatin looping. This is due to the fact that cross-linking of DNA-protein interactions can physically bind proteins of interest to other sequences with which they interact. These secondary sequences obtained through

such indirect DNA-protein interactions may even lack *bona fide* binding sites for the proteins of interest. This is often suggested as a reason why not all regions which show enrichments in ChIP-chip assays show the requisite binding site for the transcription factors of interest.

v) **Biological and Non-biological Noise:** In this chapter and in chapter 4, it was shown that nucleosome density is variable across the SCL locus, and that this information could be used to help interpret histone acetylation levels (section 4.7.3). The degree to which noise affects ChIP-chip results and the best strategies for data normalization to account for this is a controversial topic and one on which few laboratories agree (Lee et al. 2004; Bernstein et al 2004; Pokholok et al. 2005).

vi) **Antibody efficiencies:** Although antibodies may perform well in ChIP-chip, it is difficult to compare levels of enrichments across assays as a measure of the true levels of histone modifications, or binding stoichiometries of regulatory proteins. This is because antibodies may have variable efficiencies in different assays for enriching for the relevant genomic sequences. It is possible, however, to relate enrichments within an assay to aid in deducing relative binding efficiencies of proteins and relative levels of histone modifications.

5.17.8 Conclusions

The work presented in chapter 5 represents the first ever detailed elucidation of DNA-protein interactions at the SCL locus and one of the most extensive characterization of regulatory interactions of any locus in a mammalian genome. The data obtained from these ChIP-chip experiments provides a comprehensive dataset from which to further explore the relationships of histone modifications with each other and with the underlying DNA sequences. Furthermore, the identification of a number of novel putative regulatory regions and binding interactions of transcription factors at the SCL locus provide opportunities to further characterize the SCL locus to understand its regulation. Chapters 6 and 7 of this thesis show how information from these ChIP-chip profiles is used in these ways.

Chapter 6

Analytical Approaches to Interpreting ChIP-chip Data at the SCL Locus

6.1 Introduction

In chapter 5, ChIP-chip assays were used to construct extensive maps of DNA-protein interactions across the SCL locus with respect to modifications of histones H3 and H4 in SCL expressing and non-expressing cell lines. It is clear that these histone modifications have specific functions related to gene expression. However, in order to fully understand the roles of histone modifications in this way, either at a single gene locus or genome-wide, it is important to map their distribution and determine the relationships they have with each other and with respect to the genomic sequences at which they are found.

In recent years, Jeneuwin and Allis proposed the hypothesis of a histone code (also see chapter 1) which suggests that combinations of histone modifications specify a code that regulates the expression of genes and other chromatin-related activities such as mitosis and replication (Jeneuwin and Allis. 2001). Given the number of sites and the variety of possible modifications, the combinatorial possibilities are numerous; additionally it has been suggested that different combinations of histone modifications at different residues may act synergistically or antagonistically to affect gene expression. For example, *in vivo* evidence has suggested that phosphorylation at H3 serine 10 promotes acetylation (Cheung et al. 2000). In contrast, H3 lysine 9 methylation inhibits kinase activity at serine 10 and serine 10 phosphorylation inhibits methylation at lysine 9 (Rea et al. 2000).

It has been shown in human, yeast and *Drosophila*, that there is a strong correlation between acetylation of H3, H4, methylation of H4 and transcriptional activity (Pokholok et al. 2005; Bernstein et al. 2005; Schübeler et al. 2004), whereas no such correlations were found between acetylation levels at different lysines or between acetylation and transcription (Kurdistani et al. 2004). Other studies have shown that switching acetylation to methylation contributes to euchromatin gene silencing in many organisms (Lachner and Jeneuwin. 2002). Studies involving the mutagenesis of residues on histone tails have demonstrated that not all histone modifications may play distinctive roles and only a few can be considered non-redundant (Dion et al. 2005). Recently, a number of studies have been published utilizing various approaches including ChIP-chip to assess distribution of gene-specific modifications across single genes or genome-wide in various organisms. These results are varied and conclude that (i) there is evidence for a histone code (Pokholok et al. 2005; Bernstein et al. 2005; Schübeler et al. 2004; Agalioti et al.

2002), (ii) if such a code does exist, it is a simple one (Dion et al. 2005) or (iii) the histone code does not exist (Kurdistani et al. 2004; Liu et al. 2005).

Researchers have argued that the discrepancies with respect to histone modification data obtained using ChIP-chip are a result of a number of factors including (i) the use of separate microarrays containing ORFs and intergenic regions in yeast to report expression and histone modifications respectively (Kurdistani et al. 2004), (ii) the use of control experiments to account for sources of noise and nucleosome depletion (Pokholok et al. 2005), (iii) and the resolution and sensitivity of the array systems (Liu et al. 2005). Despite the debate, the use of ChIP-chip approaches still represents the best opportunity to determine whether the histone code exists for specific genes, whether the code is a general feature of all genes found in a cell, or whether there are variations in the code with respect to cell type and developmental states.

Furthermore, most of the studies mentioned above profiled only a small subset of modifications and therefore the combinatorial aspects of any apparent code may have been missed. Similarly, given that much of the work was performed in yeast, whether such a code is consistent between species is not known. The availability of detailed maps for a large range of histone modifications across the SCL locus in different cell types provides an excellent opportunity to analyse the relationships of these modifications with each other and with the numerous sequence features found in known regulatory regions at the SCL locus. In chapter 4, an initial analysis of regulatory interactions was performed to determine whether the SCL genomic tiling path arrays could detect biologically relevant information. The results of this chapter illustrate how the systematic characterization of large numbers of ChIP-chip profiles from the SCL locus using various analytical methods, has revealed novel relationships between histone modifications, DNA sequence, chromatin structure and gene expression.

6.2 Aims of this chapter

The aims of the work reported in this chapter were:

1. To perform further analyses on the SCL ChIP-chip data with respect to relationships between DNA sequence and relative levels of enrichments for various histone modifications in the human and mouse cell lines.
2. To develop an approach to interpret histone modification data with respect to the histone code hypothesis.

3. To perform further analysis of variations in nucleosome density across the SCL locus in the cell lines Jurkat, HL60 and HPB-ALL and relate these variations to gene expression.

Results

6.3 Sequence conservation at sites of histone H3 K4 dimethylation at the SCL locus in K562 and 416B cell lines

In chapter 4, regions showing enrichment for histone H3 K9/14 diacetylation across the SCL locus in both K562 and 416B cell lines were analysed with respect to human-mouse sequence conservation (section 4.5.5). A direct relationship between H3 K9/14 diacetylation levels and non-coding sequence conservation was observed. In a similar way, regions enriched for dimethylation of histone H3 at lysine 4 across the SCL locus in K562 and 416B cell lines were analysed to determine whether relationships existed between human-mouse sequence conservation and regions enriched for this histone modification. The array elements which showed significant enrichments for H3 K4 dimethylation were grouped with respect to levels of human-mouse sequence conservation in non-coding DNA. For each of these groupings, the average level of histone H3 K4 dimethylation was calculated and the data evaluated using simple regression analysis. In K562, the H3 K4 dimethylated regions that had non-coding sequence conservation with greater than 50% identity (Table 6.1, panel A), showed a positive correlation between the level of enrichment and the level of non-coding sequence conservation ($R = 0.744$). In 416B, however, (Table 6.1, panel C), this correlation was weaker and, in fact, negative ($R = -0.512$). The correlation plots for both K562 and 416B are shown in Appendix 12. These results suggest that, at least in human, there is a relationship between the degree of non-coding sequence conservation at the SCL locus and the levels of enrichment for H3 K4 dimethylation at these sequences. In addition, it was also found that highly conserved regions of non-coding DNA which were not associated with H3 K4 methylation showed similar levels of conservation as regions which were enriched for H3 K4 dimethylation in both K562 and 416B (see Table 6.1, panels B and D). These findings were similar as those obtained for H3 9/14 diacetylation (section 4.5.5); however, levels of the latter showed a much higher correlation with non-coding sequence conservation.

A

Sequence identity (%) for non-coding DNA (100 bp windows)	No. of array elements showing significant dimethylation that have non-coding sequence identity	Average level of conserved non-coding sequence (%) for array elements	Average enrichment level of dimethylation for array elements
95-100	2	98	34.226
80-84.9	4	80	46.598
75-79.9	6	77	15.629
70-74.9	8	73	22.517
65-69.9	17	67	16.914
60-64.9	22	62	15.25
55-59.9	18	57	12.047
50-54.9	5	52	11.148

B

Sequence identity (%) threshold for non-coding DNA (100 bp windows)	No. of conserved non-coding sequence peaks in tiled region	No. of conserved non-coding sequence peaks that are found in array elements showing significant dimethylation	Average level of conserved non-coding sequence (%) in array elements showing significant dimethylation	No. of conserved non-coding sequence peaks that are not associated with array elements showing dimethylation	Average level of conserved non-coding sequence (%) in array elements not associated with dimethylation
≥75	27	12	80	15	81
≥70	40	21	77	19	78
≥65	74	33	73	41	72
≥60	143	63	68	80	66
≥55	199	80	65	119	62
≥50	237	90	63	147	60

C

Sequence identity (%) for non-coding DNA (100 bp windows)	No. of array elements showing significant dimethylation that have non-coding sequence identity	Average level of conserved non-coding sequence (%) for array elements	Average enrichment level of dimethylation for array elements
95-100	2	98	5.249
90-94.9	1	93	26.557
80-84.9	2	82	9.139
75-79.9	5	77	23.913
70-74.9	4	71	21.315
65-69.9	13	67	21.491
60-64.9	27	62	23.789
55-59.9	23	57	22.493
50-54.9	4	52	22.714

D

Sequence identity (%) threshold for non-coding DNA (100 bp windows)	No. of conserved non-coding sequence peaks in tiled region	No. of conserved non-coding sequence peaks that are found in array elements showing significant dimethylation	Average level of conserved non-coding sequence (%) in array elements showing significant dimethylation	No. of conserved non-coding sequence peaks that are not associated with array elements showing dimethylation	Average level of conserved non-coding sequence (%) in array elements not associated with dimethylation
≥75	34	17	81	17	80
≥70	48	23	78	25	78
≥65	84	37	73	47	73
≥60	150	64	68	86	67
≥55	207	91	64	116	64
≥50	245	104	62	141	61

Table 6.1: The relationship between sequence conservation and histone H3 K4 dimethylation at the SCL locus in K562 and 416B. Panel A: H3 K4 dimethylation levels in K562 at conserved non-coding sequence. Panel B: Non-coding sequence conservation at H3 K4 dimethylated and non-dimethylated regions in K562. Panel C: H3 K4 dimethylation levels in 416B at conserved non-coding sequence. Panel D: Non-coding sequence conservation at H3 K4 dimethylated and non-dimethylated regions in 416B. The non-coding sequence conservation thresholds are shown in the first column of each table.

6.4 Sequence conservation at sites of histone H3 K4 trimethylation at the SCL locus in K562 and 416B cell lines

The ChIP-chip data for trimethylation of histone H3 K4 across the SCL locus in K562 and 416B cell lines was also further analysed with respect to human-mouse sequence conservation. In K562 and 416B, for the H3 K4 trimethylated regions that had sequence conservation with greater than 50% identity (Table 6.2, panels A and C), a strong positive correlation was found between the levels of enrichment and levels of sequence conservation ($R = 0.855$ for K562 and $R = 0.849$ for 416B). The correlation plots for both K562 and 416B are shown in Appendix 12. Taken together, the results presented in this thesis suggest that levels of H3 K4 di- and trimethylation as well as H3 9/14 diacetylation are correlated with levels of non-coding DNA sequences. Of these, the strongest correlation was found for H3 K9/14 diacetylation. Similarly, as seen with H3 K9/14 diacetylation and H3 K4 dimethylation, high levels of non-coding sequence conservation were not limited to only those regions that were H3 K4 trimethylated (Table 6.2 panels B and D).

A

Sequence identity (%) for non-coding DNA (100 bp windows)	No. of array elements showing significant trimethylation that have non-coding sequence identity	Average level of conserved non-coding sequence (%) for array elements	Average enrichment level of trimethylation for array elements
95-100	2	98	49.826
80-84.9	4	80	48.826
75-79.9	6	77	22.078
70-74.9	6	73	34.047
65-69.9	17	67	14.986
60-64.9	24	62	13.25
55-59.9	24	57	8.536
50-54.9	5	52	17.13

B

Sequence identity (%) threshold for non-coding DNA (100 bp windows)	No. of conserved non-coding sequence peaks in tiled region	No. of conserved non-coding sequence peaks that are found in array elements showing significant trimethylation	Average level of conserved non-coding sequence (%) in array elements showing significant trimethylation	No. of conserved non-coding sequence peaks that are not associated with array elements showing trimethylation	Average level of conserved non-coding sequence (%) in array elements not associated with trimethylation
≥75	27	11	79	16	81
≥70	40	20	77	20	78
≥65	74	33	73	41	72
≥60	143	68	67	75	66
≥55	199	89	64	110	63
≥50	237	102	62	135	60

C

Sequence identity (%) for non-coding DNA (100 bp windows)	No. of array elements showing significant trimethylation that have non-coding sequence identity	Average level of conserved non-coding sequence (%) for array elements	Average enrichment level of trimethylation for array elements
95-100	2	98	137.009
90-94.9	1	93	171.868
80-84.9	2	82	164.326
75-79.9	5	77	51.889
70-74.9	3	70	77.394
65-69.9	8	68	89.0129
60-64.9	21	62	37.487
55-59.9	17	57	39.907
50-54.9	3	52	37.091

D

Sequence identity (%) threshold for non-coding DNA (100 bp windows)	No. of conserved non-coding sequence peaks in tiled region	No. of conserved non-coding sequence peaks that are found in array elements showing significant trimethylation	Average level of conserved non-coding sequence (%) in array elements showing significant trimethylation	No. of conserved non-coding sequence peaks that are not associated with array elements showing trimethylation	Average level of conserved non-coding sequence (%) in array elements not associated with trimethylation
≥75	34	14	82	20	80
≥70	48	19	78	29	78
≥65	84	30	74	54	72
≥60	150	48	69	102	67
≥55	207	67	65	140	63
≥50	245	73	63	172	61

Table 6.2: The relationship between sequence conservation and histone H3 K4 trimethylation at the SCL locus in K562 and 416B cell lines. Panel A: Histone H3 K4 trimethylation levels in K562 at conserved non-coding sequence. Panel B: Non-coding sequence conservation at H3 K4 trimethylated and non-trimethylated regions in K562. Panel C: Histone H3 K4 trimethylation levels in 416B at conserved non-coding sequence. Panel D: Non-coding sequence conservation at H3 K4 trimethylated and non-trimethylated regions in 416B. The non-coding sequence conservation thresholds are shown in the first column of each table.

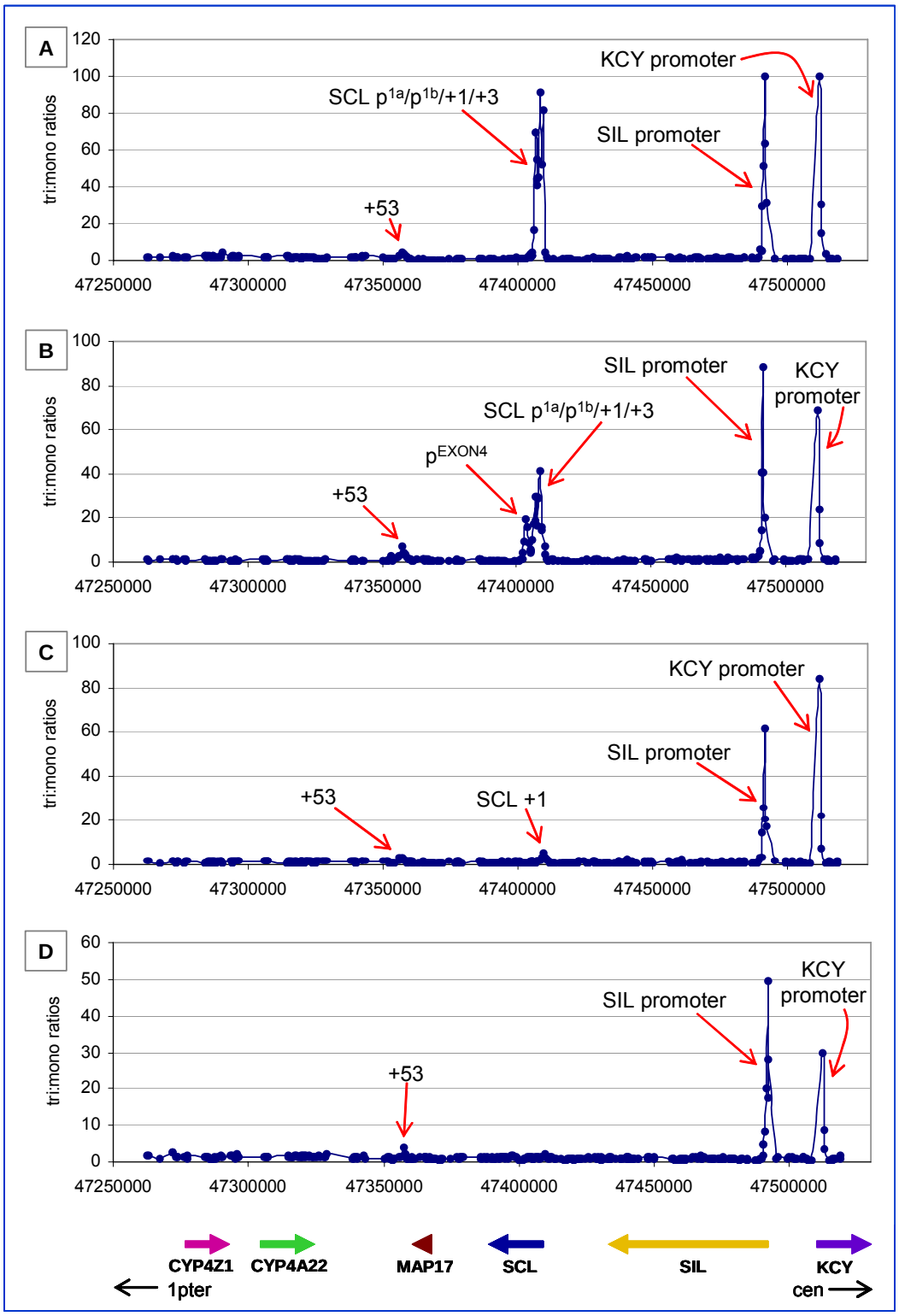
6.5 Relative levels of monomethylation to trimethylation of H3 lysine 4 correlate with transcriptional activity

From the ChIP-chip profiles generated for mono- and trimethylation at lysine 4 of histone H3 in SCL expressing and non-expressing human and mouse cell lines (see chapter 5), a few interesting features were observed, including:

1) The presence of both repressive (monomethylation) and activating (trimethylation) marks at or near the same regions (both promoters and enhancers). Significant peaks of enrichments were found at the inactive promoters and the active promoters showed lack of enrichments for monomethylation, whereas highest peaks of enrichments for trimethylation were seen at both active and inactive promoters (see Figures 5.10 to 5.14 in chapter 5).

2) The presence of methylation marks at the SCL gene in HL60 and mouse ES cell line, both of which are SCL non-expressing cell lines (see Figures 5.11, 5.13 and 5.14 in chapter 5).

The presence of histone marks for dimethylation of H3 K4 at the same regions as mono- and tri- could be attributed to the fact that dimethylation at lysine 4 of H3 is known to occur at both active and inactive regulatory regions (Santos-Rosa et al. 2002). To compare the relative levels of mono- and trimethylation at the regulatory regions across the SCL locus, profiles were generated by plotting trimethylation to monomethylation ratios (for each array element) against the genomic coordinates in SCL expressing and non-expressing human and mouse cell lines (Figure 6.1).


A
tri:mono ratios
KCY promoter
SCL p1a/p1b/+1/+3
SIL promoter
+53
B
tri:mono ratios
SIL promoter
KCY promoter
SCL p1a/p1b/+1/+3
pEXON4
+53
C
tri:mono ratios
KCY promoter
SIL promoter
+53
SCL +1
D
tri:mono ratios
SIL promoter
KCY promoter
+53
47250000
47300000
47350000
47400000
47450000
47500000
CYP4Z1
CYP4A22
MAP17
SCL
SIL
KCY
1pter
cen

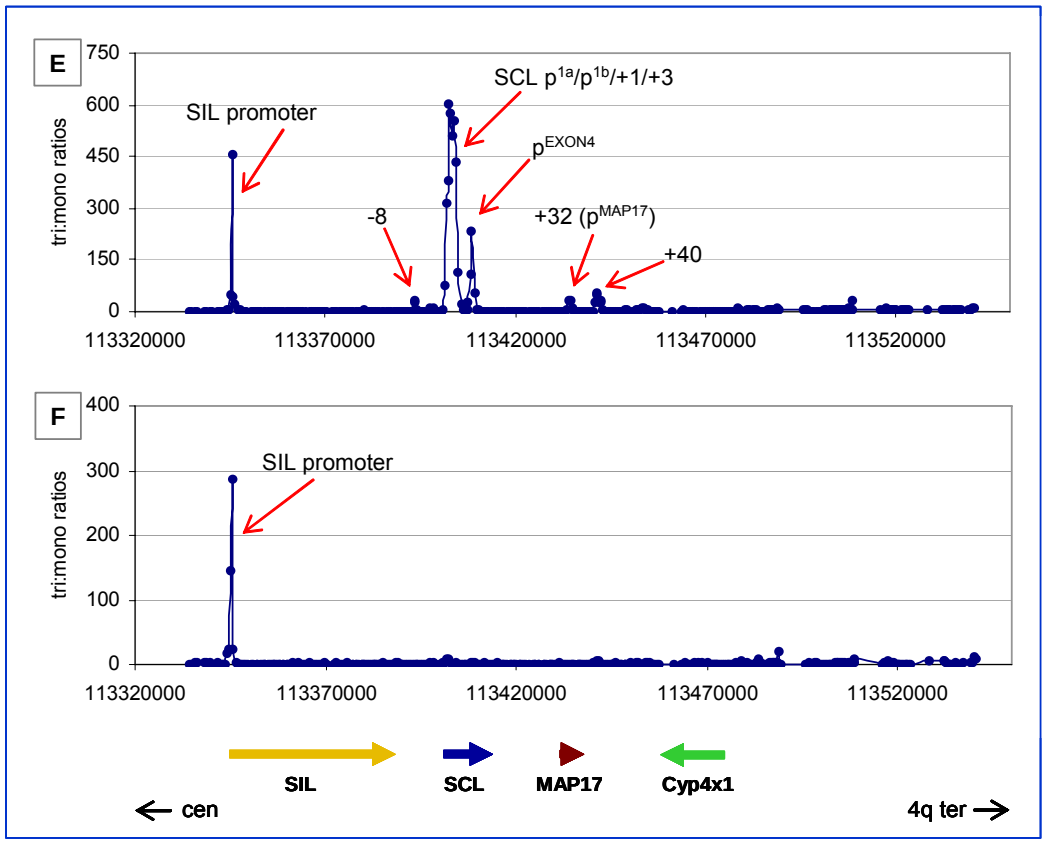


Figure 6.1: Profiles showing trimethylation to monomethylation ratios in the human and mouse cell lines. Panel A: K562, panel B: Jurkat, panel C: HL60, panel D: HPB-ALL, panel E: 416B and panel F: mouse ES cell line. The y-axes in the panels represent the tri- to mono- ratios in the respective cell line and the x-axes represent the genomic coordinates along the human chromosome 1 (in A, B, C, D) and mouse chromosome 4 (in E, F). The genomic regions showing high tri- to mono- ratios are marked by red arrows. The highest tri- to mono- ratios were seen at the active promoters in each cell line. In 416B cell line, low tri- to mono- ratios were also seen at SCL -8 and the +40 erythroid enhancer (Delabesse et al. 2005). The thick, coloured arrows at the bottom of panels D and F represent the genes, their order and the direction of transcription in human and mouse respectively. The orientation of the locus with respect to the centromere (cen) and telomere (ter) on human chromosome 1 and mouse chromosome 4 is shown by black arrows at the bottom of panels D and F.

The 5' ends i.e. the promoter regions of KCY and SIL showed high tri- to mono- ratios in all the human cell lines. These genes are expressed in all the cell lines. Similarly, the promoter region of the SIL gene, which is expressed in 416B and mouse ES cell lines, reported high tri- to mono- ratios. At the 5' end of the SCL gene, high tri- to mono- ratios were seen in K562, Jurkat and 416B cell lines, all of which express SCL. In addition, promoter p^{EXON4} at the +7 region also showed high tri- to mono- ratios in Jurkat and 416B cell lines. The endogenous promoter p^{EXON4} is known to be active in Jurkat (Bernard et al. 1992; Leroy-Viard et al. 1994) but its activity in 416B cell line is not known. In HL60, very low tri- to mono- ratios were found at the 5' end of SCL gene, though, SCL is not expressed in this cell line (ratio of 4.8 in HL60 as compared to ratios of 91.5, 41.5 and

601 in K562, Jurkat and 416B respectively, at the SCL promoter region). The promoter region of the SCL gene did not show increased tri- to mono- ratios in HPB-ALL and mouse ES cell line, both of which do not express SCL.

Interestingly, high tri- to mono- ratios were not found at the promoter region of MAP17 in K562 and Jurkat, although it has been shown in a previous study that this gene is expressed in these cell lines (Delabesse et al. 2005). In 416B, the MAP17 promoter showed very low tri- to mono- ratios which suggested that MAP17 gene is expressed in 416B (Delabesse et al. 2005). A highly unexpected observation, made from the results obtained for tri- to mono- ratios, was the presence of relatively higher ratios at the +53 region in K562, Jurkat and HPB-ALL (as high ratios in these cell lines were found only at the active promoters). This region had no known regulatory function apart from its location in the vicinity of the +51 erythroid enhancer (Delabesse et al. 2005). The +53 region had also been identified previously with significant enrichments for H3 and H4 acetylation and various other histone modifications (see Tables 5.2, 5.3 and 5.4 in chapter 5).

Taken together, the above described results suggested that the ratio of trimethylation to monomethylation more accurately depicts the transcriptional activity of genes than by analysing the profiles individually. However, these results also highlighted the lack of high tri- to mono- ratios on the MAP17 promoter, thus, not correlating with its known activity. Additionally, the presence of high ratios at the +53 region suggested that this region may have a regulatory function separated from the +51 erythroid enhancer (see chapter 7 for detailed discussion on these regions). Similar analysis was also performed for di- to mono- ratios in SCL expressing and non-expressing human and mouse cell lines, but the ratios obtained did not correlate with the transcriptional activity of the genes at the SCL locus (data not shown).

6.6 Analysis of nucleosome density at the SCL locus Jurkat, HL60 and HPB-ALL

In chapters 4 and 5, decreased levels of histone H3 were observed at regions coincident with the active regulatory regions across the SCL locus in four human cell lines. Furthermore, the relative levels of histone H3 in K562 were further analysed to show that the genomic regions covering the expressed and non-expressed genes were shown to have statistically different levels of histone H3 and that both non-coding regulatory and coding regions of expressed genes also showed decreased levels of histone H3 (section 4.7.2).

Further analysis of the results obtained for histone H3 levels in Jurkat, HL60 and HPB-ALL was performed as described for K562 in chapter 4. The relative levels of histone H3 in the three cell lines were ranked with respect to four types of DNA sequences – Type 1- acetylated non-coding sequences, Type 2 – acetylated and gene coding sequences,

Type 3 – non-acetylated and gene coding sequences, and finally Type 4 – neutral sequences, i.e. sequences with no known biological activity (Table 6.3, panels A-C). It was observed that in all three cell lines, below the 10th percentile of histone H3 enrichments, there was a noticeable preference for sequences of Type 1-3 (biologically active sequences). Although, the number of biologically active sequences below the 10th percentile was same in all three cell lines, however, the regions these sequences came from were notably different. In Jurkat, this group included sequences from KCY, SIL and the SCL -7 region, whereas in HL60 and HPB-ALL, these sequences came exclusively from the KCY and SIL genes. In Jurkat, the other regulatory sequences flanking the SCL gene were ranked in the group below 20th percentile and in HL60 only the sequences from the 5' end of SCL were included in this group. In HPB-ALL, all the regulatory sequences flanking the SCL gene were included in the group in the 50th percentile with median ratios at 1. These data support that the lowest levels of histone H3 across the SCL locus in each of the cell lines are associated with features of active genes and is consistent with observations made in K562.

A

			FEATURES OF SEQUENCES FOUND IN ARRAY ELEMENTS				
Percentile Ranking of Array Elements	Relative Histone H3 Level per Array Element	No of Array Elements	Acetylated Non-Coding Sequence (Type 1)	Acetylated Coding Sequence (Type 2)	Non-acetylated Coding Sequence (Type 3)	Biologically Active Sequences (Types 1-3)	Neutral Sequences
≤ 5th	0.167-0.637	17	7	1	7	15	2
≤ 10th	0.167-0.729	35	11	3	8	22	13
≤ 20th	0.167-0.830	70	13	4	18	35	35
≤ 30th	0.167-0.886	105	20	4	24	48	57
≤ 40th	0.167-0.942	140	26	6	32	64	76
≤ 50th	0.167-1.000	175	32	6	33	71	104

B

			FEATURES OF SEQUENCES FOUND IN ARRAY ELEMENTS				
Percentile Ranking of Array Elements	Relative Histone H3 Level per Array Element	No of Array Elements	Acetylated Non-Coding Sequence (Type 1)	Acetylated Coding Sequence (Type 2)	Non-acetylated Coding Sequence (Type 3)	Biologically Active Sequences (Types 1-3)	Neutral Sequences
≤ 5th	0.188-0.480	17	4	1	7	12	5
≤ 10th	0.188-0.567	35	6	3	13	22	13
≤ 20th	0.188-0.654	70	7	4	21	32	38
≤ 30th	0.188-0.752	105	10	4	25	39	66
≤ 40th	0.188-0.879	140	11	4	31	46	94
≤ 50th	0.188-1.000	175	13	4	36	53	122

C

			FEATURES OF SEQUENCES FOUND IN ARRAY ELEMENTS				
Percentile Ranking of Array Elements	Relative Histone H3 Level per Array	No of Array Elements	Acetylated Non-Coding Sequence (Type 1)	Acetylated Coding Sequence (Type 2)	Non-acetylated Coding Sequence (Type 3)	Biologically Active Sequences (Types 1-3)	Neutral Sequences
≤ 5th	0.310-0.680	17	7	3	3	13	4
≤ 10th	0.310-0.761	35	10	3	9	22	13
≤ 20th	0.310-0.827	70	12	3	18	33	37
≤ 30th	0.310-0.898	105	13	3	25	41	64
≤ 40th	0.310-0.949	140	13	3	34	50	90
≤ 50th	0.310-1.000	175	13	4	37	54	121

Table 6.3: Sequence features of SCL array elements ranked according to their relative levels of histone H3 in the human cell lines. Panel A: Jurkat, panel B: HL60, panel C: HPB-ALL. The tables show the sequence features for array elements ranked in percentile intervals according to their relative levels of histone H3 in ChIP-chip experiments.

To further analyse the histone H3 levels across the SCL locus, the relative levels of histone H3 in active and inactive regions in all the cell lines were determined and z-tests were performed. Any differences between active and inactive regions with respect to histone H3 levels could therefore be determined statistically. In HL60 and HPB-ALL, the inactive region included the SCL, CYP4A22 and CYP4Z1 genes (all these genes are not expressed in these cell lines) and the active region included KCY and SIL genes. In Jurkat, the SCL gene was included in the active region along with KCY and SIL. For consistency, the entire SCL locus, including regions covering all of its regulatory elements were included in the active or inactive regions accordingly. The results indicated that in HL60 and HPB-ALL, the average level of histone H3 was higher in the inactive region (reported to be 1.33 for HL60 and 1.11 for HPB-ALL) as compared to the active region (reported to be 0.66 for HL60 and 0.87 for HPB-ALL). These differences were statistically significant to 99.9% confidence level in a z-test. In Jurkat, the average level of histone H3 in the inactive and active regions was reported to be 1.12 and 1.01 respectively and these were also found to be significant to 99.9% confidence level in a z-test.

Taken together, these data provided additional evidence that histone H3 levels and thus nucleosome density is variable across the SCL locus in all the cell lines examined; furthermore, nucleosome depletion and generally lower levels of nucleosomes occur across active genes and their associated regulatory sequences.

6.7 Visualization of ChIP-chip profiles for all histone modifications in SCL expressing and non-expressing cell lines

In order to further understand the roles of histone modifications in the transcriptional regulation of SCL and its flanking genes, it was necessary to develop more sophisticated approaches to analyse the large datasets presented in chapter 5. Therefore, a microarray visualization tool, TreeView (developed by Mike Eisen's Laboratory) (Eisen et al. 1998), originally designed for microarray expression data, was used to allow all ChIP-chip profiles in a given cell line to be viewed simultaneously as a digital output across the SCL locus. For each profile, fold enrichments ratios for each array element were converted to $\log_2$ ratios to reflect degrees of enrichment with respect to the baseline value of 1 (0 in $\log_2$ scale). In TreeView, values above zero (i.e., enrichments) appear as red blocks, values below zero (i.e., depletions) appear as green blocks and values at or very near 0 (i.e., neither enrichments nor depletions), appear as black blocks.

These values for each ChIP-chip profile were then visualised in TreeView with respect to the genomic positions of the array elements. Figure 6.2 shows the TreeView visualizations for all the histone modification datasets obtained for each cell line. In total, ChIP-chip datasets were available for 21 histone modifications in K562, 19 modifications in Jurkat and HPB-ALL and 18 modifications in HL60. The ChIP-chip datasets for acetylation of H4 K12 and dimethylation of H3 K9 were also included; although they were considered as non-working ChIP-chip assays based on the criteria described in chapter 5 (see later sections and the discussion).

By visualising the ChIP-chip data in this manner, it was easy to distinguish regions that were enriched or depleted for various histone modifications. The presence of intense red blocks for a number of histone modifications at the 5' ends of the KCY and SIL genes in all of the cell lines illustrates that these genes are expressed in these cell lines. Similarly, the genomic sequences covering the SCL gene are also highly enriched for a number of histone modifications in K562 and Jurkat cell lines. HL60 shows enrichment for some of the histone modifications at the SCL locus but, in contrast, HPB-ALL does not show significant enrichments as illustrated by the lack of large groups of red blocks across the SCL gene. In K562, the +51 region (erythroid enhancer) which is downstream of the MAP17 gene is also highlighted as intense red blocks for various histone modifications. Blocks of red were also seen in each of the other cell lines in the vicinity of the +51 region, although they were not as widespread as in K562.

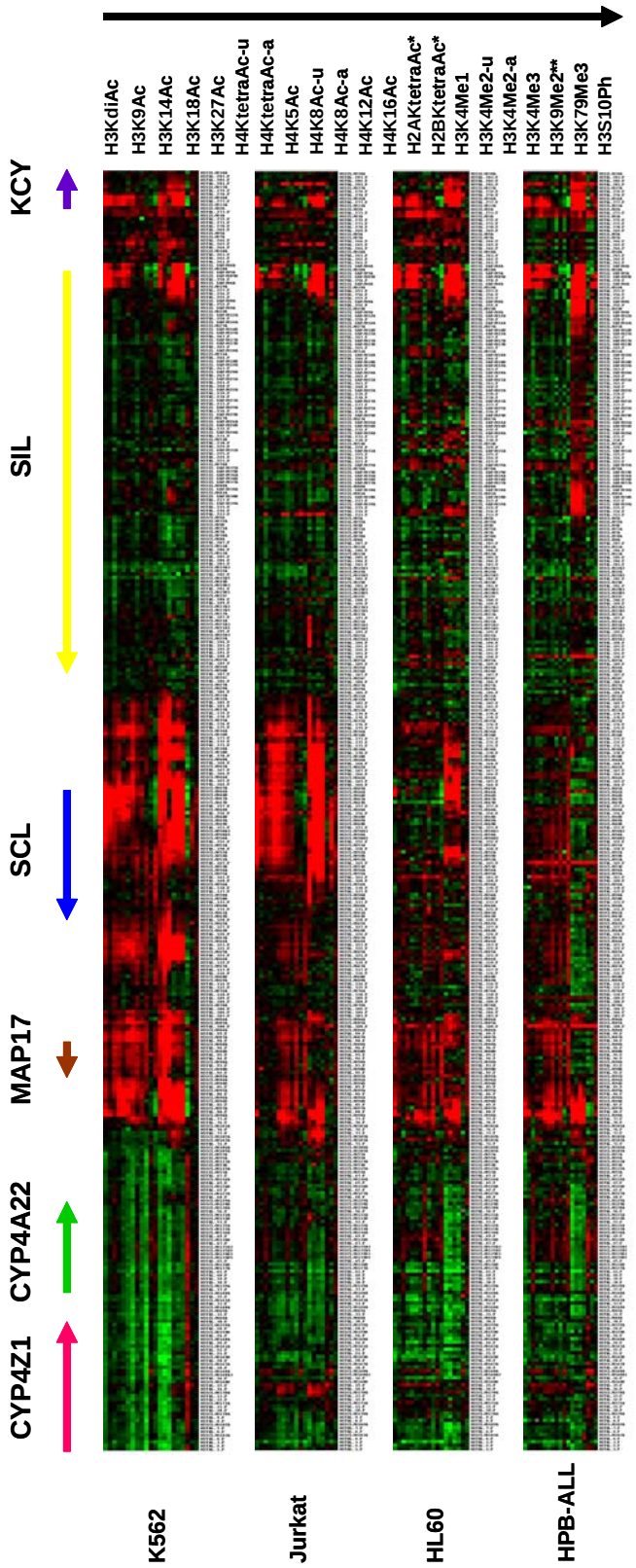


Figure 6.2: A digital output of ChIP-chip profiles across the SCL locus for all the histone modifications in human cell lines. The cell lines are listed at the bottom of the figure. The thick, coloured arrows along the y-axis represent the gene order and the direction of transcription. The array elements in each data set are listed along the y-axis (of each TreeView profile) in order across the SCL tile. The histone modifications and their order of appearance in each profile are shown at the top of the figure. In each profile, red colour indicates enrichment and green colour indicates depletion for the respective histone modification. The intensity of the colour reflects degrees of depletion or enrichment from the baseline value of 1 (value of 0 in $\log_2$ scale). The data sets for acetylated H2A and H2B were only available for K562 and are marked with single asterisk (*). The data sets for H3 lysine 9 dimethylation which are marked with double asterisks (**) were only available for K562, Jurkat and HPB-ALL. For H4 K5/8/12/16 tetra-acetylation, H4 K8 acetylation and H3 K4 dimethylation, two datasets per histone modification were included for each cell line.

6.8 Hierarchical clustering of histone modifications distinguishes types of regulatory sequences and their activity in K562

In order to further analyze the ChIP-chip data to explore possible relationships of the histone modifications with each other and with the underlying regulatory sequences, clustering of the datasets for various histone modifications was performed using a hierarchical clustering algorithm (Cluster, Eisen Laboratory) (Eisen et al. 1998) and visualized in TreeView. Hierarchical clustering performed by Cluster relates the data in two-dimensional space with respect to: (i) the relationships of the genomic sequences (i.e., array elements) in one dimension, and (ii) relationships between the function of the histone modifications in the other dimension. Given that K562 was central to the work of this thesis and represented the most well characterized human cell line with respect to SCL regulation, all the ChIP-chip datasets available for this cell line were used in clustering analysis. Although, the ChIP-chip datasets for H4 K12 acetylation and H3 K9 dimethylation did not report significant enrichments (as calculated for all the ChIP-chip datasets in the present study, see chapter 4), these were also included in this analysis to see if a functional relationship with other histone modifications or with genomic sequences could be seen. Figure 6.3 shows the clustering tree diagram obtained for K562.

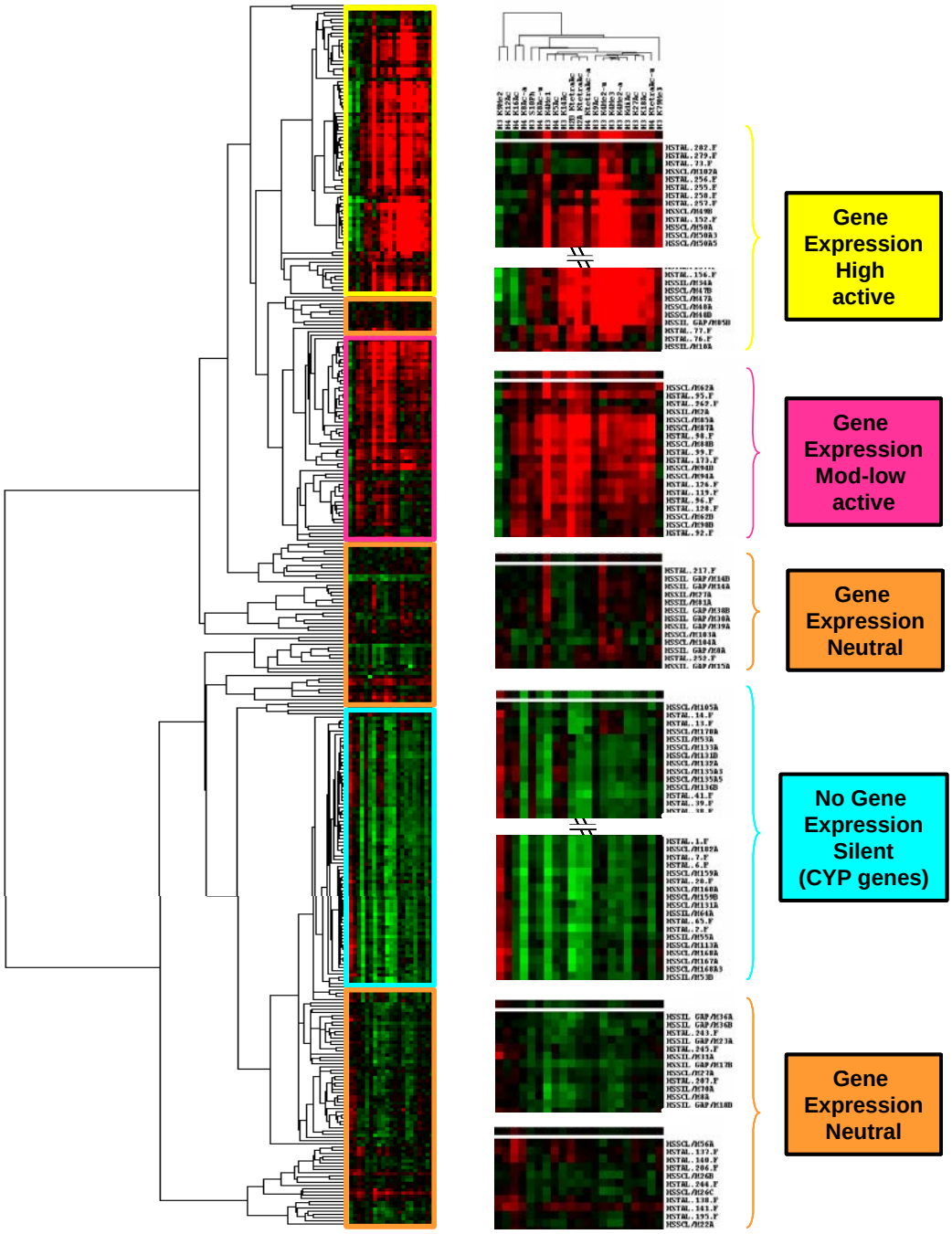


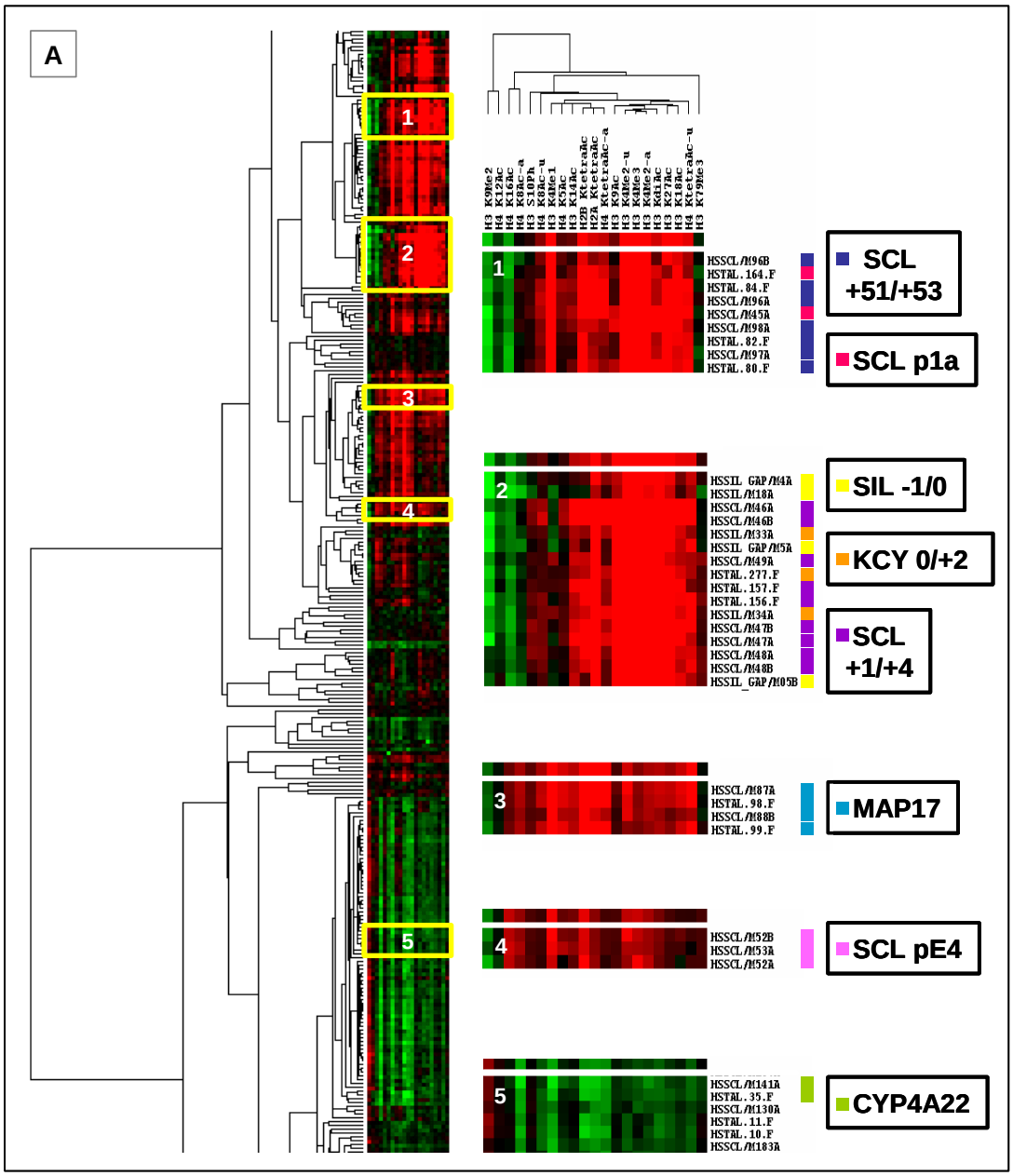
Figure 6.3: Hierarchical clustering defines active and inactive regions of chromatin across the SCL locus in K562. Hierarchical clustering was performed on 21 histone modification data sets in the K562 cell line. The genomic regions identified in the various clusters have been broadly classified into four regions (boxed in yellow, pink, blue and orange) and the description is provided in the text. A representative profile for each of these four regions is shown magnified in the center of the figure. The tree shown in the figure is clustered in two dimensions: based on the functional similarities between genomic regions (array elements) along the y-axis and based on the functional similarities between the various histone modifications along the x-axis (listed at the top of the figure). The array elements representing the genomic regions are listed along the y-axis (to the right of the magnified profiles). Each box corresponds to an array element for a given modification and the presence of red in a box indicates enrichment and the presence of green indicates depletion for that histone modification.

From this tree, the array elements were organised into four broad classifications based on their association with histone modifications, regulatory function and gene expression. The region boxed in yellow contained sequences from the expressed KCY, SIL and SCL genes, which were known to be highly enriched for histone modifications and were active in gene regulation. The region boxed in blue contained sequences from the non-expressed cytochrome P450 genes which represent silent regions. The region boxed in pink contained sequences from the MAP17 gene which is expressed at low levels in K562 (Delabesse et al. 2005) as well as other known SCL regulatory elements with moderate to low levels of histone modifications. Finally, the regions boxed in orange contained sequences coinciding with coding regions which have no known regulatory function and therefore were annotated as neutral. This analysis demonstrated that genomic regions involved in transcriptional regulation could be clustered based on the histone modifications associated with them. Each cluster was then further analyzed to identify and annotate the known regulatory elements of all the genes represented on the tiling path array at the SCL locus. Figure 6.4 shows the same clustering tree diagram for K562 as seen in the previous figure but with specific clusters highlighting the various regulatory elements contained within them.

Clusters 1 to 5 in Figure 6.4 contained the 5' ends (i.e. promoter regions) of all the genes (KCY, SIL, SCL, MAP17 and CYP4A22) on the SCL array. The 5' ends of the KCY, SIL and SCL genes were represented in clusters 1 and 2, the 5' end of MAP17 was shown in cluster 3, and the 5' end of CYP4A22 was found in cluster 5. Given the relative known activities of these promoters, clustering accurately placed these regions in the highly active, moderately active and silent regions respectively (refer to Figure 6.3). SCL p^{EXON4}, which is known to be inactive in K562, clustered away from the active promoters in cluster 4. An interesting and unexpected finding from these clusters was the presence of the SCL +51/+53 region, the known erythroid enhancer, with the SCL p^{1a} and its upstream region in cluster 1; however, data supporting the transcriptional activity of this region was shown in section 6.5.

The CYP4A22 promoter region, shown in cluster 5, noticeably showed enrichment for H3 K9 dimethylation which was not seen in clusters 1 to 4. H3 K9 dimethylation is known to be a hallmark of silent chromatin (Lachner and Jenuwein 2002). Furthermore, enrichments for H3 K9 dimethylation were seen across the entire cytochrome P450 gene cluster and were generally absent from regions containing the KCY, SIL and SCL genes (refer to Figures 6.3 and 6.4). Even though the ChIP-chip assay for this histone modification was considered as a "non-working" assay based on the criteria discussed in chapters 4 and 5 (sections 4.4 and 5.4), the ChIP-chip data for this assay clearly

demarcates active regions of chromatin from inactive ones. This suggests that the criteria used to assess “working” ChIP-chip assays described in this thesis selected against at least one assay which reported interesting biological information.



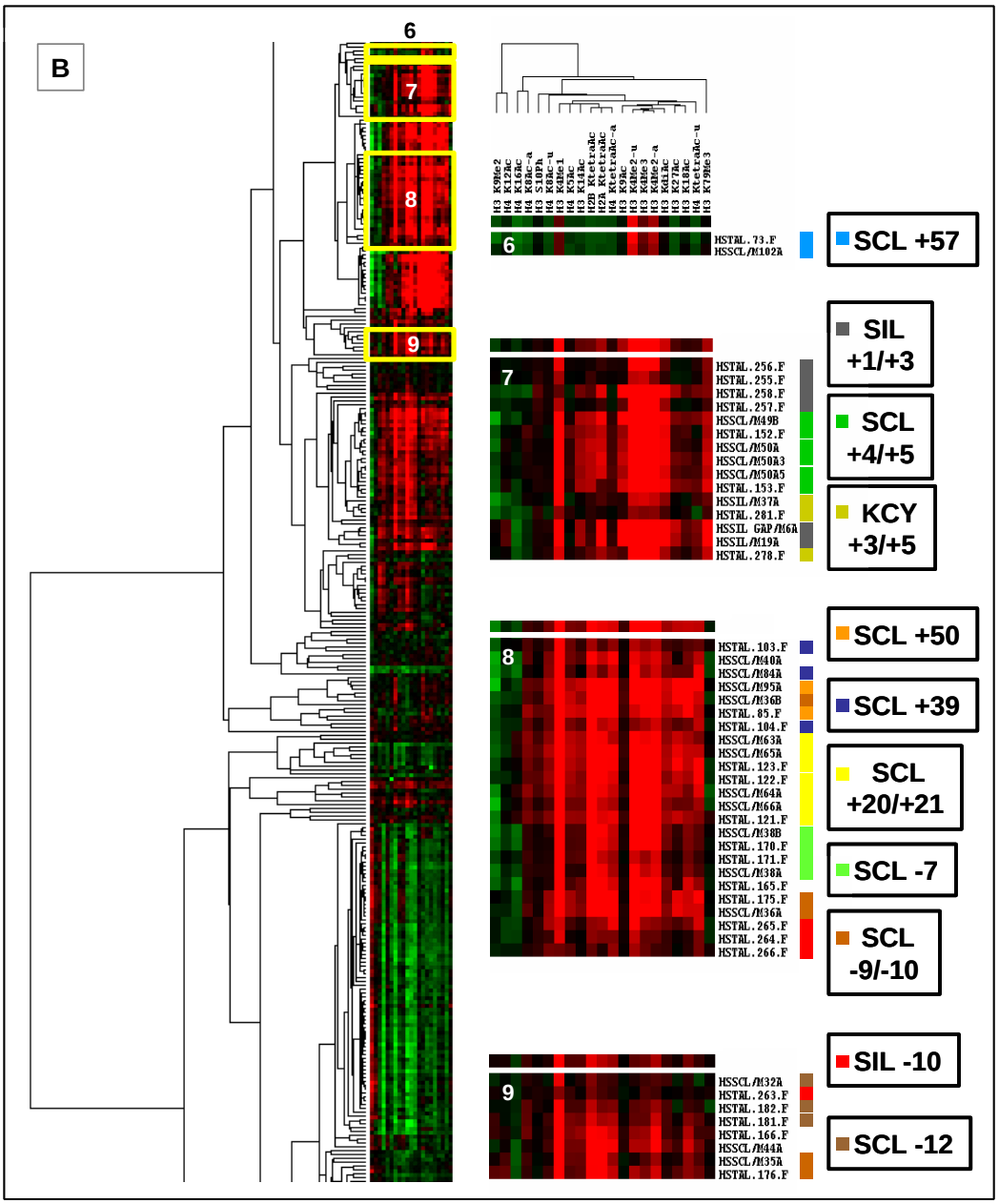


Figure 6.4: Identification and annotation of clusters of regulatory sequences across the SCL locus in K562. The description of the information shown in a cluster diagram is as described in Figure 6.3. In addition, the clusters are named 1 to 9 (boxed in yellow) and the magnified images of each cluster are shown in the centre of the figure. Each row within a cluster represents an array element (names on the right of each row refer to the array element nomenclature). The clusters are annotated with respect to known regulatory function as shown at the right of the figure; the colour coded boxes refer to the array elements which are found in each of the known regulatory elements. Panel A shows clusters 1-5 containing the promoter regions of all the genes on the array and panel B shows clusters 6-9 containing other regulatory regions across the SCL locus.

Clusters 6 to 9 contained other known regulatory regions across the SCL locus (enhancers), along with additional regions identified with enrichments for various histone

modifications which had no known regulatory function (Chapters 4 and 5). The +57 region, which binds CTCF and may represent a putative insulator element at the SCL locus (see chapter 5, section 5.16) was annotated in cluster 6. Cluster 7 contained regions downstream of all the active promoters (i.e. KCY, SIL and SCL) in K562. Cluster 8 contained the stem cell enhancer (+20/+21), the SCL -9/-10 region, sequences immediately upstream from the +51 erythroid enhancer (at +50), along with the novel regions at SCL +39 and at SCL -7 which bind CTCF and GATA-1 respectively (Chapter 5 section 5.16; Chapter 4, section 4.6.2). The SCL -12 and SIL -10 regions previously identified with enrichments for various histone modifications were found in cluster 9 along with some sequences from SCL -9/-10.

Based on all of the above observations, it was evident that hierarchical clustering of histone modifications was able to classify sequences according to known regulatory functions.

6.9 Identification of a minimal set of histone modifications which delineates regulatory function at the SCL locus.

In order to further distinguish differences between the clusters with respect to histone modifications, the consensus level of ChIP-chip enrichments for each modification were determined for each of the 9 clusters described above. This consensus is shown as a single row at the top of each cluster in Figure 6.4, in which each square represents the average enrichment of all the array elements included in that cluster for the respective histone modifications. Subsequently, these consensus values for each cluster were compared, to explore the relationships of histone modifications with the associated regulatory regions.

Figure 6.5 shows the consensus modification profiles for each of the 9 clusters. As previously noted, the CYP4A22 promoter region (cluster 5) was noticeably enriched for dimethylation of H3 K9 (shown in red); furthermore, it was depleted for all other histone modifications (shown in green or black). It was evident from this observation that H3 K9 dimethylation and lack of enrichments for the other modifications tested, clearly distinguishes silent chromatin from the active chromatin at the SCL locus.

Another striking difference was seen in cluster 7, which represents the regions immediately downstream of the KCY, SIL and SCL promoters (all these promoters are active in K562). These regions showed enrichment for H3 K79 trimethylation which was not found in any other region. This suggests that this histone modification clearly marked the immediate downstream regions of highly active promoters and distinguishes them from the inactive promoters and other regulatory regions at the SCL locus.

Of all the acetylation modifications of histone H3, the acetylation of H3 K9 was unique to highly active promoters and their downstream regions (clusters 2 and 7 and to a lesser degree cluster 1). The remaining acetylation marks were present on all the promoters and most of the enhancers and could not really distinguish these clusters from each other. This suggests that the acetylation of H3 K9 marks the 5' ends (i.e. promoters) of active genes and distinguishes them from other types of regulatory regions.

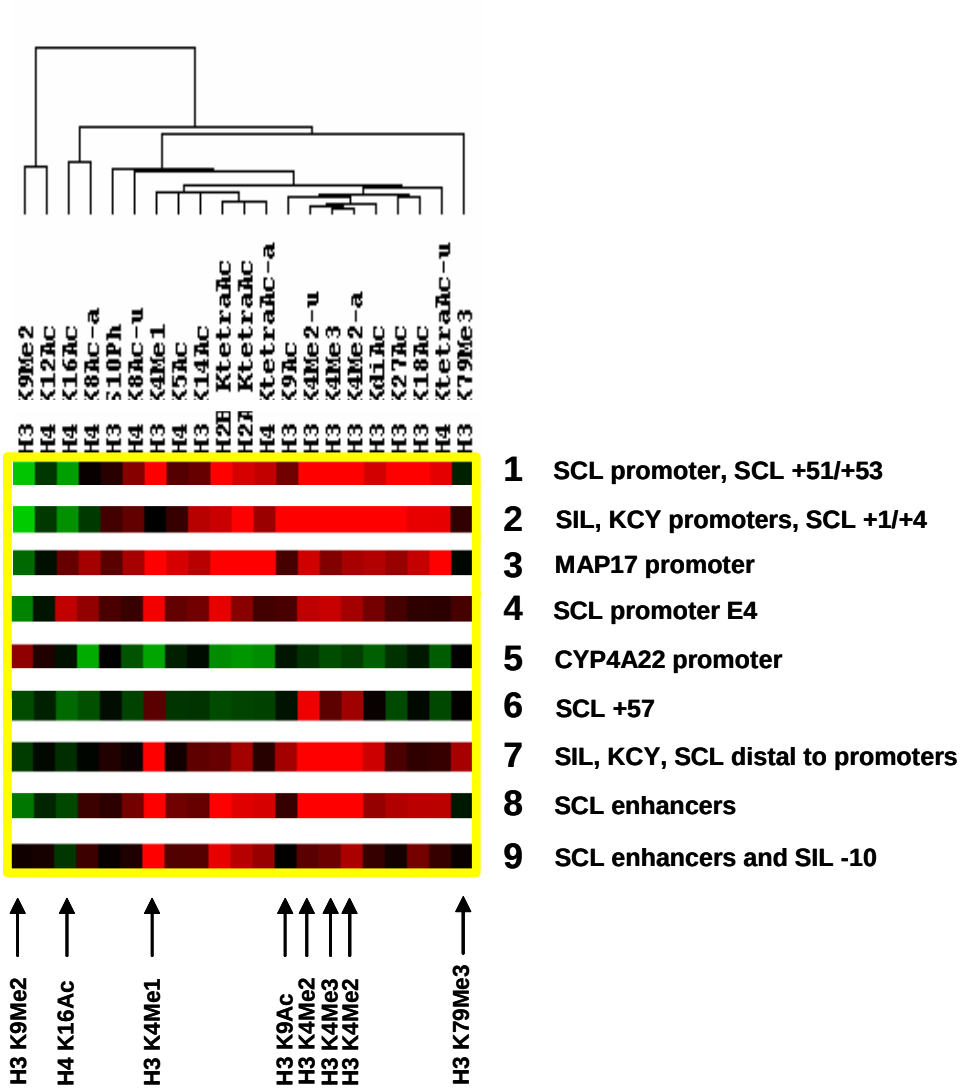


Figure 6.5: Identifying a minimal set of histone modifications for the SCL locus by comparing the "consensus" histone modifications of clusters 1 to 9 in K562. The clusters identified using hierarchical clustering of ChIP-chip data in K562 (clusters 1-9 in Figure 6.4) are represented by a "consensus" for histone modification levels as shown in this figure. Rows 1 to 9 in this figure correspond to the consensus profiles for clusters 1 to 9 respectively. Each of the rows shown in the above figure is annotated corresponding to the regulatory regions associated with the cluster as shown in Figure 6.4 and described in the text. The x-axis along the top of the figure shows the clustering of histone modifications based on their functional similarities/dissimilarities. Each box in a row represents the average value of the respective histone modification in that cluster. The red/black/green colour of each box represents enrichment/depletion for the

corresponding histone modification. The key minimal set of histone modifications that distinguished the various clusters from each other are shown by arrows at the bottom of the figure but for dimethylation of H3 K4, two datasets were included in the clustering analysis (ChIP-chip assays using two different antibodies for dimethyl H3 K4). Both assays performed similarly and are both indicated by arrows within the minimal set.

Hyperacetylation of H4 K16 was noticed only at the SCL p^{EXON4}, MAP17 promoter (clusters 3 and 4) and was depleted at all the remaining clusters. It has been shown in previous studies that (i) SCL p^{EXON4} is inactive in K562 (Bernard et al. 1992) and (ii) in yeast, active promoters are hypoacetylated for H4 K16 (Liu et al. 2005). Taken together, this indicated that acetylation of H4 K16 could be one of the key modifications which distinguished inactive from active regulatory elements. It is known that MAP17 exhibits very low level expression in K562 (Delabesse et al. 2005) suggesting that the promoter is active; however, the presence of hyperacetylation for H4 K16 at the promoter was not consistent with its known activity. This may suggest that the H4 K16 acetylation mark is depleted in highly active promoters and enriched in low activity promoters.

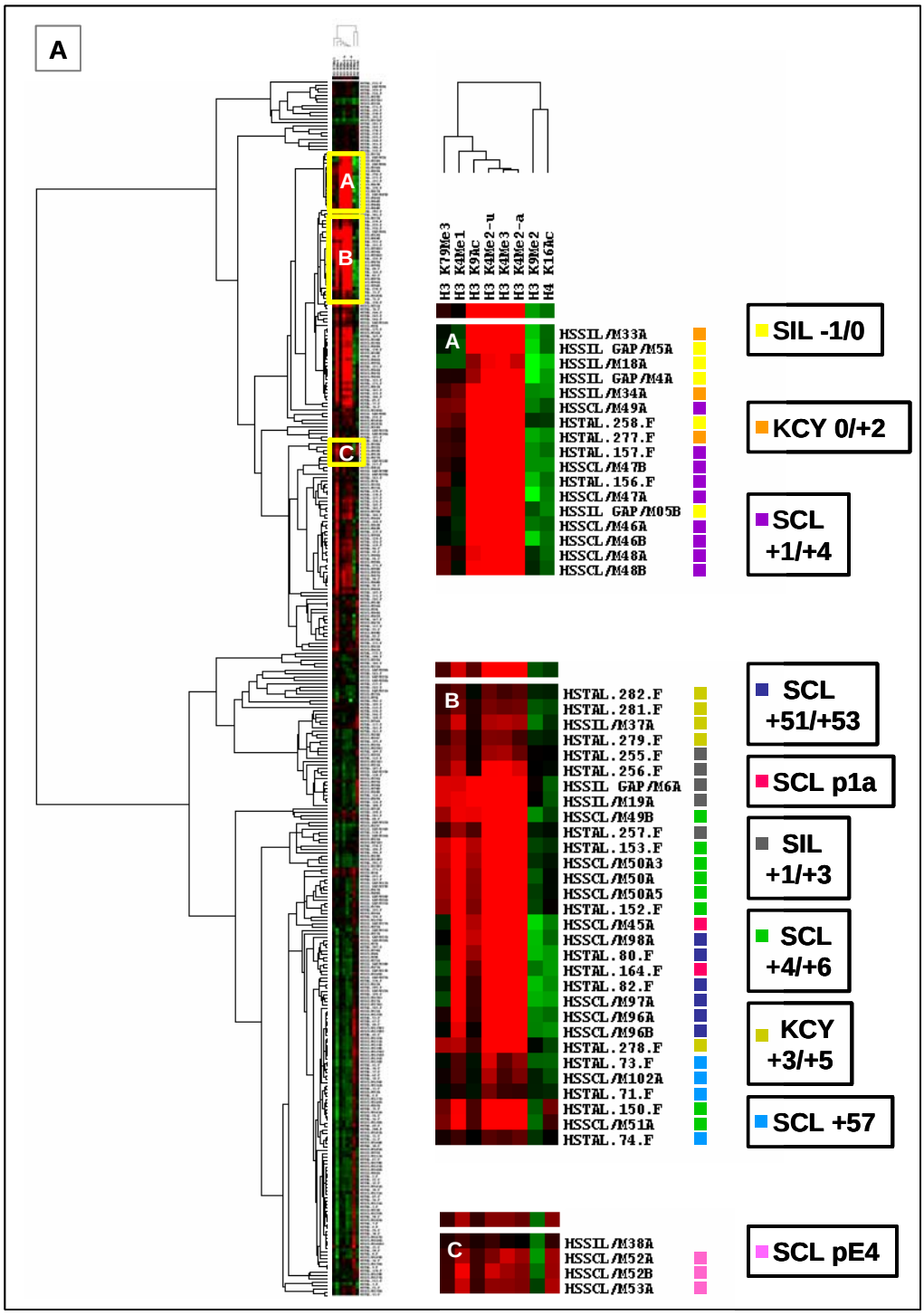
Monomethylation of H3 K4 is known to be a repressive mark (van Dijk et al. 2005) and was found to be depleted only at 5' ends of genes at or near promoters (cluster 2), whereas other regulatory regions including the downstream regions of active promoters (cluster 7), inactive promoters (cluster 4), low activity promoters (cluster 3), known and novel regulatory regions such as enhancers (clusters 6, 8, 9) were enriched for this histone modification. Cluster 1 which contained the SCL promoter, showed increased levels of monomethylation; however data shown in chapter 5 showed depletion of this modification at the SCL promoter (section 5.9.1.1). This discrepancy may have arisen from the "consensus" for cluster 1 not accurately representing the SCL promoter (given that there were sequences from the known enhancer at +51 in this cluster). All of this data taken together suggests that the depletion of monomethylation for H3 K4 is unique to active promoters.

Di- and trimethylation of H3 K4 was present at all the known regulatory regions across the SCL locus where genes were actively expressed (i.e. KCY. SIL and SCL): the highest enrichments, seen by the intensity of the red colour, were found at the 5' ends of active genes (clusters 1, 2 and 7) and SCL enhancers (cluster 8). Additionally, these marks were the only histone modifications present at the +57 region. Based on the binding interactions of CTCF detected at this region (see chapter 5), the +57 region may represent a putative insulator element at the SCL locus. This suggests that the presence of only di- and trimethylation of H3 K4 may be unique to certain regulatory regions exhibiting special functions (such as insulator activity).

Based on all of these observation, seven key histone modifications (shown with arrows at the bottom of the Figure 6.5) were found to be important for not only in distinguishing various clusters of regulatory regions from each other but also were indicative of whether the associated regions were active or not in K562. If these seven modifications were indeed key to specifying a "code" of regulatory activities and functions at the SCL locus, they should be sufficient to define this "code". Therefore, ChIP-chip profiles for only these seven histone modifications were used in a second iteration of hierarchical clustering. The cluster diagram and the relevant clusters identified using this minimal set are shown in Figure 6.6. Indeed, the genomic regions which had clustered together with the full complement of 21 histone modifications remained clustered with the minimal set, although the nine clusters reported previously, were represented by 7 clusters in this analysis. For simplicity, and to avoid confusion with the previous clustering analysis, the clusters from this analysis are labelled A to G in Figure 6.6 and in the text below.

The 5' ends (i.e. promoters) of all the active genes KCY, SIL and SCL were again clustered together in cluster A. The array elements in this cluster showed hypo dimethylation of H3 K9 and hypoacetylation of H4 K16. Depletion for these marks was found to be associated with active regions. Enrichments for H3 K9 acetylation, H3 K4 di- and trimethylation were also seen in cluster A. The array elements included in this cluster also represented the regions immediately downstream of the promoters (i.e. KCY +2 and SCL +1/+4). Thus, relatively increased enrichments were seen for H3 K79 trimethylation. This cluster was also characterized by depleted or low levels of H3 K4 monomethylation, considered to be a hallmark of active promoters.

As seen previously, the +51/+53 erythroid enhancer region was clustered along with the SCL p^{1a} in cluster B. In addition, cluster B also included the regions immediately downstream of the active promoters of the KCY, SIL and SCL genes and the +57 region. A number of array elements in this cluster representing the immediate 3' regions of the active promoters showed hyperacetylation of H3 K79; hyper monomethylation of H3 K4 was found at these regions. As this cluster contained known active regulatory regions, it also showed pronounced hypoacetylation of H4 K16 for some of the members of the cluster. Hyper di- and trimethylation of H3 K4 was also a hallmark of this cluster.


A
A
B
C
H3 K79Me3
H3 K4Me1
H3 K9Ac
H3 K4Me2-u
H3 K4Me3
H3 K4Me2-a
H3 K9Me2
H4 K16Ac
A
HSSIL/M33A
HSSIL GAP/M5A
HSSIL/M18A
HSSIL GAP/M4A
HSSIL/M34A
HSSCL/M49A
HSTAL.258.F
HSTAL.277.F
HSTAL.157.F
HSSCL/M47B
HSTAL.156.F
HSSCL/M47A
HSSIL GAP/M05B
HSSCL/M46A
HSSCL/M46B
HSSCL/M48A
HSSCL/M48B
B
HSTAL.282.F
HSTAL.281.F
HSSIL/M37A
HSTAL.279.F
HSTAL.255.F
HSTAL.256.F
HSSIL GAP/M6A
HSSIL/M19A
HSSCL/M49B
HSTAL.257.F
HSTAL.153.F
HSSCL/M50A3
HSSCL/M50A
HSSCL/M50A5
HSTAL.152.F
HSSCL/M45A
HSSCL/M98A
HSTAL.80.F
HSTAL.164.F
HSTAL.82.F
HSSCL/M97A
HSSCL/M96A
HSSCL/M96B
HSTAL.278.F
HSTAL.73.F
HSSCL/M102A
HSTAL.71.F
HSTAL.150.F
HSSCL/M51A
HSTAL.74.F
C
HSSIL/M38A
HSSCL/M52A
HSSCL/M52B
HSSCL/M53A
SIL -1/0
KCY 0/+2
SCL +1/+4
SCL +51/+53
SCL p1a
SIL +1/+3
SCL +4/+6
KCY +3/+5
SCL +57
SCL pE4

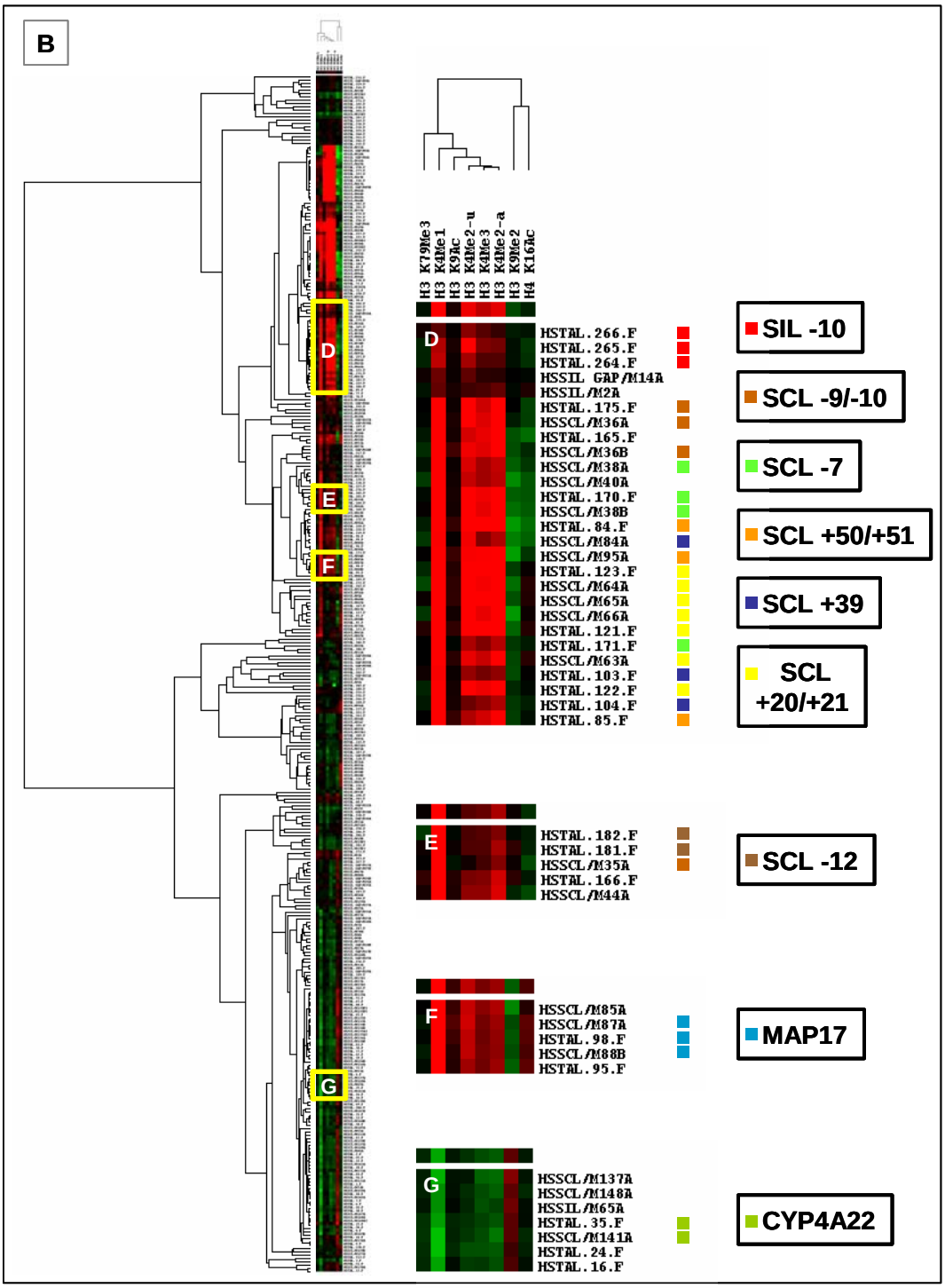


Figure 6.6: Hierarchical clustering of ChIP-chip data across the SCL locus in K562 using the minimal set of seven key histone modifications. Hierarchical clustering was performed using eight ChIP-chip datasets representing seven histone modifications across the SCL locus in K562 (two datasets were included for H3 K4 dimethylation). The description of the information shown in a cluster diagram is as described in Figure 6.4. The clusters are named A to G and magnified images of each cluster are shown in the middle of the figure. Each row within a cluster represents an array element (names on the right of each row refer to the array element nomenclature). The clusters are annotated with respect to known regulatory function as shown at the right of the figure; the colour coded boxes refer to the array elements which are

found in each of the known regulatory elements. Panel (A) shows clusters A, B and C containing the promoter regions of KCY, SIL and SCL genes and panel (B) shows clusters D to G containing other regulatory regions across the SCL locus.

SCL p^{EXON4} which is known to be inactive in K562 (Bernard et al. 1992) clustered away from the other active promoters in cluster C. The most noticeable histone mark in this cluster was hyperacetylation of H4 K16, which was considered to be associated with inactive regions.

SCL regulatory regions including the stem cell enhancer at +20/+21, a segment of the the +50/+51 erythroid enhancer region, and the -9/-10 region clustered along with a number of novel regions identified in the present study including the SCL +39 and -7 regions and the SIL -10 regions (cluster D). Interestingly, the array elements representing the +20/+21 stem cell enhancer showed hyperacetylation of H4 K16 suggesting that despite the presence of other histone marks, this region may be inactive in K562, whereas the active enhancers such as the +50/+51 region and -7 region were hypoacetylated for H4 K16. Hyper mono-, di- and trimethylation of H3 K4 was also a hallmark of this cluster.

Cluster E contained the SCL -12 region and an array element from the SCL -9/-10 region. The SCL -12 region was identified as a putative novel regulatory region in chapter 4 (section 4.5.2). This cluster showed hyper monomethylation of H3 K4, relatively higher enrichments for di- and trimethylation of H3 K4, and hypoacetylation of H4 K16. The activity of this region is not known.

The promoter of MAP17 gene was contained in cluster F. In contrast to the other active promoters, the MAP17 promoter region showed substantial enrichments for monomethylation of H3 K4, and hyperacetylation of H4 K16. Both of these marks are associated with inactive regulatory regions suggesting that the MAP17 promoter is primarily inactive in K562 (see chapter 7).

Finally, the promoter region of the CYP4A22 gene (cluster G) was most noticeably enriched for hyper dimethylation of H3 K9 and depleted for all of the remaining 6 histone modifications. This is consistent with this region being in the silent chromatin.

In summary, this analysis suggested that by using these seven histone modifications, it may be possible to functionally classify (i) silent from active chromatin, (ii) regulatory regions within the active chromatin and (iii) active from inactive regulatory regions within the active chromatin. This data also suggested that at least for the SCL locus, the combination of histone modifications define a “code” of regulatory activities and function, consistent with the histone code hypothesis (see discussion).

6.10 Discussion

The work described in this chapter involved further analysis of the ChIP-chip datasets for various histone modifications generated in chapter 5. These analyses were performed in order to understand the functional roles of these histone modifications with respect to each other, and the underlying sequence. The principal findings of this analysis are discussed below.

6.10.1 Sequence conservation at sites of di- and trimethylation of H3 K4

While there has been intensive interest in understanding the functions of histone modifications with respect to transcriptional regulation, there has been little interest in understanding the relationships between the location of these modifications and features of genomic DNA sequence. In this study, analysis of ChIP-chip data for histone H3 acetylation and methylation (Chapter 4 and this chapter) has shown that such relationships do exist at the SCL locus in both human and mouse. A strong correlation exists between levels of H3 K9/14 diacetylation and levels of human-mouse sequence conservation in non-coding DNA (Chapter 4). Results in this chapter established a strong relationship between the levels of H3 K4 trimethylation and the levels of human-mouse sequence conservation in non-coding DNA. H3 K4 dimethylation levels are related to human-mouse sequence conservation in non-coding DNA, but evidence for this relationship is less strong. However, it was also found that high levels of non-coding sequence conservation were not associated with histone H3 methylation. Studies have reported similar observations for H3 methylation for other regions of the human and mouse genomes, with sequence conservation at sites of histone H3 methylation being no higher than background levels (Bernstein et al. 2005). The data presented in chapter 4 for H3 K9/14 diacetylation is also in agreement with these findings and the reasons for this are discussed in chapter 4 (section 4.8.1).

The difference in the correlations between sequence conservation and H3 K4 di- and trimethylation is explained by the fact that the highest levels of human-mouse sequence conservation are seen at promoters – and enrichments levels for H3 K4 trimethylation were at their highest at promoters, whereas highest enrichments for dimethylation were present at the regions immediately downstream of promoters within transcribed regions (Chapter 5, section 5.9.2). Biochemically, this difference could be explained by the conversion of di- to trimethylation at H3 K4 over the active promoter, thus resulting in lower levels of dimethylation over these regions. Thus, the highest levels of histone H3 dimethylation may skew the analysis performed here because (i) they are located in coding sequences (and therefore not included in non-coding regions used in this

analysis) or (ii) they have lower levels of sequence conservation where histone enrichments go into intronic sequences. In contrast, the highest levels of H3 K4 trimethylation and H3 K9/14 diacetylation were both present at promoter regions and therefore both showed strong correlations with non-coding sequence conservation. However, given that the relationship with H3 diacetylation was marginally stronger than that for H3 H4 trimethylation, this may also be explained by slight variations in their relative distributions with respect to the promoter sequences. Moreover, it should also be noted, that the sample sizes used for these analyses across the SCL locus was very small (containing only three active genes in human and only two in mouse). Therefore, the true relationships between sequence conservation and histone modifications awaits larger scale studies which sample a greater number of regulatory features associated with larger sets of genes.

6.10.2 Relative levels of mono- and trimethylation correlate with transcriptional activity

The promoter regions of active genes were associated with lack of enrichments for monomethylation and highly enriched for di- and trimethylation. It has been suggested that the relative increase of trimethylation versus dimethylation discriminates the active from inactive genes (Schneider et al. 2004). However, the data obtained in the present study suggests that relative enrichments of trimethylation to monomethylation are key indicators of active and inactive genes. It was established that although all of the H3 K4 methylation marks were present at the 5' end of the SCL gene in HL60 and mouse ES cells, both of which do not express SCL, the tri- to mono- ratios at the SCL promoter region in these cell lines were very low as compared to the ratios obtained for highly active genes (for instance, a ratio of 4.8 at the SCL promoter as compared to 62 and 82 for SIL and KCY respectively in HL60; similarly, a ratio of 286 for SIL as compared to 9.6 for SCL in the mouse ES cell line). These results could further explain the presence of H3 diacetylation at the SCL gene in SCL non-expressing cell lines - when it is evident by the tri- to mono- ratios, that the presence of monomethylation, and not H3 diacetylation or trimethylation, represents one of the key marks to reflect the activity of a promoter.

6.10.3 Nucleosome depletion occurs across the SCL locus in all human cell lines

ChIP-chip profiles for histone H3 in all the human cell lines showed decreased levels at a number of regulatory regions across the SCL locus (Chapter 5). Further analysis of the ChIP-chip data established that the sequences showing decreased levels of histone H3 corresponded to the regulatory regions associated with the genes that were expressed in the respective cell lines. These data suggest that nucleosome depletion occurs at the

active genes and is particularly evident at their associated non-coding regulatory sequences and coding regions, across the SCL locus. These results were in agreement with the data obtained in the K562 cell line (also see discussion in chapter 4) and with observations in yeast (Bernstein et al. 2004; Lee et al. 2004; Pokholok et al. 2005; Yuan et al. 2005).

Statistical analysis of the ChIP-chip data for histone H3 also established that the relative levels of histone H3 at transcriptionally active and inactive regions were significantly different in all the human cell lines, albeit this difference was much smaller in Jurkat (see results of z-test in section 6.6). For consistency in comparing the relative levels between the active and inactive regions in the SCL expressing cell lines K562 and Jurkat, sequences up to and including the +51 erythroid enhancer were considered to be in the active regions in both the cell lines. There is no evidence to suggest that this enhancer is active in Jurkat - thus, the genomic region included as active in this cell line may be biased with inactive regions being included. Indeed, the genomic region covering the +51 region in Jurkat showed high relative levels of histone H3 and, thus, the average level across the regions considered as active would be increased. This was not an issue for the SCL non-expressing cell lines because the entire SCL locus was included as inactive in these statistical analyses.

6.10.4 Identification of a consensus histone code at the SCL locus

Recently, various studies have been published either supporting (Pokholok et al. 2005; Dion et al. 2005; Schübeler et al. 2004) or not supporting (Liu et al. 2005; Kurdistani et al. 2004) the hypothesis of a histone code. It has been argued that it is the patterns rather than specific acetylation modifications that define regulatory regions (Kurdistani et al. 2004), whereas other have suggested that specific histone marks are distinctly associated with regulatory elements or transcriptional states (Pokholok et al. 2005; Dion et al. 2005).

The study presented for this thesis was one of the most extensive characterizations for histone modifications performed so far across a well characterized locus where most of the regulatory regions and their functions are well known. Analysis of all the ChIP-chip datasets led to the identification of a set of key histone modifications, the patterns of which are consistent at the known regulatory regions at the SCL locus in the K562 cell line. Interestingly, dimethylation of H3 K9 was found to be one of these key modifications although the ChIP-chip data obtained for this modification did not fulfil the “working assays” criteria. This indicated that in some cases, although the histone modifications may not be significantly enriched at a genomic region in ChIP-chip assays according to

statistical criteria, they may still be functionally important and the ChIP-chip profiles may reveal interesting biological information. The association of H3 K9 dimethylation with silent chromatin has also been observed previously (Litt et al. 2001).

The presence and/or absence of these key modifications allowed for a clear distinction of:

i) active from silent chromatin domains – the active region containing SCL and its flanking genes from the silent region containing the CYP4A22 and CYP4Z1 genes

ii) active from inactive promoters within the active domain

iii) promoters from enhancer regions and enhancers from other regulatory elements such as insulators

Thus, based on these observations and the totality of the histone modification data presented in chapter 5, a consensus histone code for the SCL region is proposed here and shown in Figure 6.7. This code was found to be highly consistent in distinguishing regulatory regions, their function and activity at the SCL locus in the K562 cell line.

According to this consensus histone code, the silent and active chromatin domains can be distinguished by the presence of H3 K9 dimethylation at silent regions and its generalized absence in active domains. Within the active domain, the presence of a few key marks is able to discriminate different functions of regulatory sequences. Promoters are distinguished from other types of regulatory sequences (such as enhancers) by their association with H3 K9 acetylation. Regions downstream from promoters, which may participate in transcriptional elongation (Krogan et al. 2003) are defined through association with H3 K79 trimethylation. In addition, several modifications can confer an active or inactive status to regulatory sequences. These include H4 K16 acetylation which is depleted at active regulatory elements and enriched at inactive elements, and H3 K4 monomethylation which is depleted at active promoters only, enriched at inactive promoters and enriched at other types of regulatory sequences regardless of their activity. Thus, H3 K4 monomethylation has a role both in defining regulatory activity and regulatory function along with H3 K9 acetylation. Furthermore, H3 K9 dimethylation also plays a role in defining regulatory activity in active chromatin (in addition to its role in distinguishing silent from active chromatin domains), by showing obvious depletions at active regulatory regions. Furthermore, the relative levels of histone H3 mono-, di- and tri-methylation can also confer the activity status of a regulatory element, as well as help define insulators (the “other regulatory element” category shown in the figure below), which are associated primarily with H3 K4 methylation. Thus, this code shows that

histone modifications have cellular roles which have a “defining function” component, and a “defining activity” component.

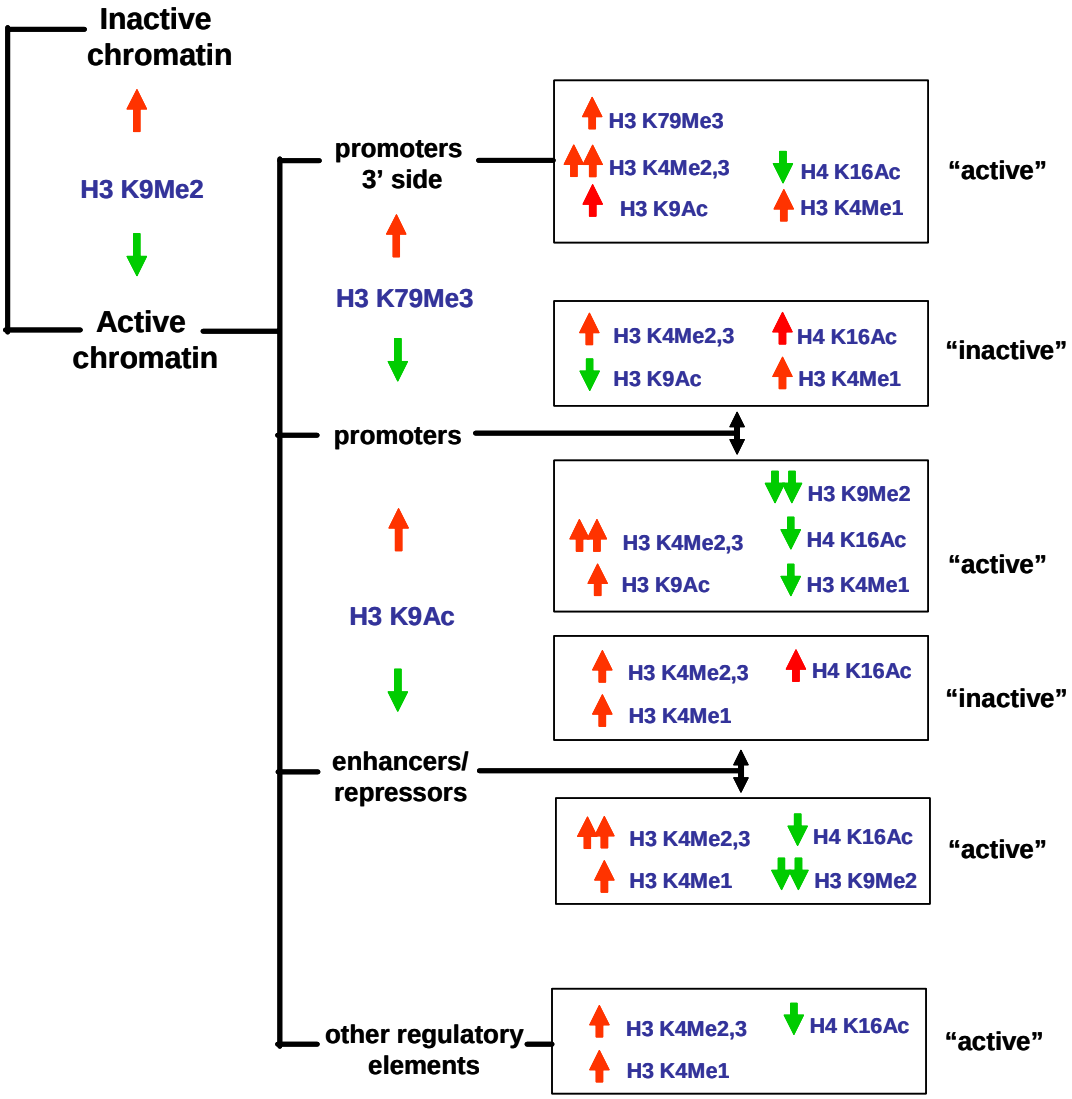


Figure 6.7: A consensus histone code for the regulatory regions across the SCL locus. This code was proposed based on the observations for the presence and/or absence of seven key histone modifications and other data presented in chapter 5. These histone modifications are: dimethylation of H3 K9 (H3 K9Me2), trimethylation of H3 K79 (H3 K79Me3), acetylation of H3 K9 (H3 K9Ac), dimethylation of H3 K4 (H3 K4Me2), trimethylation of H3 K4 (H3 K4Me3), monomethylation of H3 K4 (H3 K4Me1) and acetylation of H4 K16 (H4 K16Ac).The modifications shown are associated with active or inactive regulatory regions such as promoters, immediate regions 3’ of promoter, enhancers, and other regulatory elements such as insulators. The red arrows facing upwards represent enrichments for the respective histone modifications and the green arrows facing downwards represent depletions. The relative increase or decrease in the levels of enrichments or depletion for a given histone modification between the active and inactive states is also shown (two arrows versus one arrow). The description of the key features shown in this figure is provided in the text.

Of all the key histone modifications described above, a few have been specifically linked to the transcriptional activity of genes or genomic regions in previous studies. For instance, the association of H3 K9 dimethylation with silent chromatin has been observed previously (Lachner and Jeneuwin 2002). Dion et al have recently shown through mutagenesis of histone residues that H4 lysine 16 has a vital role in transcriptional activity (Dion et al. 2005). In addition, hyper trimethylation of H3 K79 has been linked to the distal regions of active promoters (Pokholok et al. 2005).

The proposed consensus histone code, however, does not take into account all the other histone modifications that were either not tested or the other assays that did not work. By testing and including addditional histone marks in ChIP-chip studies across the SCL locus, it may be possible to deduce a comprehensive histone code which would provide an even deeper insight into how the various regulatory regions are regulated and how these *cis*-elements then regulate the genes they are associated with. The presence of the code also suggests that patterns of histone modifications may be sufficient to identify the function and activity of regulatory features *de novo* (see chapter 7).

6.10.5 Conclusions

In conclusion, the data presented in this chapter suggests that the inter-relationships of histone modifications and their relationships with DNA sequence help define the functional roles of these modifications at regulatory regions across the SCL locus. In chapter 7, interpretations of this code in a number of human cell lines, will aid in the characterization of all the regulatory regions which have been annotated across the SCL locus.

Chapter 7

Further Characterization of Regulatory Regions From Across the SCL Locus

7.1 Introduction

In chapter 6, a consensus histone code was proposed for the SCL region. This code suggested that regulatory regions across the SCL locus were associated with a set of histone modifications which could define their function and activity in K562 cell line. In chapter 5, ChIP-chip datasets were also described for various histone modifications in three other cell lines: Jurkat, HL60 and HPB-ALL. The purpose of surveying the distribution of histone modifications in these cell lines was to elucidate differences and/or similarities and relate these results with the expression and regulation of SCL accordingly.

Understanding differences in the regulation of SCL in various cell types is important for understanding SCL biology in both normal and disease states. This is particularly important where inappropriate expression of SCL in T-cells results in T-cell acute lymphoblastic leukaemia (T-ALL). In cases of T-ALL, the inappropriate expression of SCL has most often been linked to chromosomal rearrangements (deletions or translocations) at the 5' or 3' end of the SCL gene (Bernard et al. 1991; Brown et al. 1990; Xia et al. 1992). It has also been reported that several established leukaemic T-cell lines, such as Jurkat, exhibit high levels of SCL expression in the absence of any apparent SCL genomic alteration (Bernard et al. 1992; Leroy-Viard et al. 1994). In Jurkat, transcription is mainly initiated from SCL p^{1b}, and to a much lesser extent from p^{EXON4}. The expression was found to be mono-allelic (Leroy-Viard et al. 1994), thus suggesting that a *cis*-acting mechanism might be responsible for activating the single allele. However, to date, the mechanism of SCL activation in Jurkat is not known.

Of the two DNase I hypersensitive sites identified in Jurkat at the SCL locus (Leroy-Viard et al. 1994), the SCL -7 region exhibits a strong DNase I sensitivity and was found to be located near one of the SCL breakpoints reported in SIL-SCL deletions resulting in T-ALL (Breit et al. 1993); this suggests that this genomic region might be implicated in the *cis*-activation of SCL transcription (Leroy-Viard et al. 1994). However, it is also possible that in Jurkat (or other T-cell lines with no genomic rearrangements at the SCL locus),

other regulatory regions at the SCL locus, may be involved in the inappropriate expression of SCL.

This further emphasizes the fact that, in order to fully understand the regulation of SCL in different cell types, it is important to elucidate and characterize the full complement of regulatory regions across the SCL locus - with respect to their activity, function, role in transcriptional regulation of SCL and their associated DNA-protein interactions. To this end, the results presented in this chapter provide additional clues to the biological activity of these regulatory elements.

7.2 Aims of this chapter

The aims of the work reported in this chapter were:

To provide characterizations of known regulatory regions across the SCL locus with respect to data obtained from this study and from other sources where appropriate.

i) To test the proposed histone code with respect to the known regulatory regions, with known activities and functions across the SCL locus in SCL expressing and non-expressing cell lines (i.e. K562, Jurkat, HL60 and HPB-ALL).

ii) To further characterize novel regions identified at the SCL locus.

iii) To identify regulatory elements involved in the *cis*-activation of the SCL gene in Jurkat.

Results

7.3 Generating TreeView visualization profiles for the four cell lines

ChIP-chip datasets for the seven key histone modifications of the SCL histone code (Chapter 6) were used to generate TreeView (Eisen et al. 1998) visualization profiles for the four human cell lines K562, Jurkat, HL60 and HPB-ALL (similar to those shown in Figure 6.2). For the purposes of the work presented in this chapter, these TreeView profiles depicted the various histone modification with respect to their genomic distribution across the SCL locus and were not clustered. In summary, the key histone modifications included were:

i) H3 K79 trimethylation (H3 K79Me3),

ii) H3 K4 monomethylation (H3 K4Me1),

iii) H3 K4 dimethylation (H3 K4Me2) (two datasets were included for this histone modification),

iv) H3 K4 trimethylation (H3 K4Me3),

v) H3 K9 dimethylation (H3 K9Me2) (this dataset was not available for HL60),

vi) H4 K16 acetylation (H4 K16Ac), and

vii) H3 K9 acetylation (H3 K9Ac).

In chapter 5, the assay for H3 K9 acetylation was shown not to yield significant enrichments in Jurkat, suggesting that there may be a cell type specific difference with respect to this histone modification (see section 5.6.3.1). Similarly, H3 K9 dimethylation did not show enrichments across the cytochrome P450 genes to indicate the location of regions of silent chromatin (data not shown). Given that the other three cell lines analyzed here all showed H3 K9 acetylation and dimethylation profiles consistent with the proposed histone code, this suggests that Jurkat may not utilise modifications at lysine 9 residue of histone H3 in the same way as other cell types. However, for consistency, data for both of these ChIP-chip assays are included in the Treeview profiles shown in this chapter. These visualizations are described with respect to each of the regulatory regions across the SCL locus in the subsequent sections of this chapter.

7.4 The SIL promoter and the SIL -10 region

In Figure 7.1, TreeView profiles at the 5' end of the SIL gene and the upstream SIL -10 region are shown for the four human cell lines. In the figure, the known and well characterized SIL promoter (Colaizzo-Anas and Aplan 2003) has been annotated as p^{SILa}.

According to the consensus histone code, the SIL p^{SILa} promoter shows features of an active promoter in all the four cell lines. These features include hyper-acetylation of H3 K9, hyper di- and trimethylation of H3 K4, hypo-acetylation of H4 K16, and hypo monomethylation of H3 K4. In addition, trimethylation of H3 K79 was seen at the region immediately downstream of the promoter region in all cell lines. It should be noted that hypo dimethylation of H3 K9 and hyper-acetylation of H3 K9 were not very apparent in Jurkat for reasons discussed above (see section 7.3). In the K562 cell line, binding interactions for p300, CBP, HDAC2 (all three cofactors have HAT and HDAC activity) and $Taf_{II}250$ (component of the pre-initiation complex) were seen at the SIL p^{SILa}, which provided further support for the activity of this promoter in this cell line (see chapter 5, sections 5.13, 5.14).

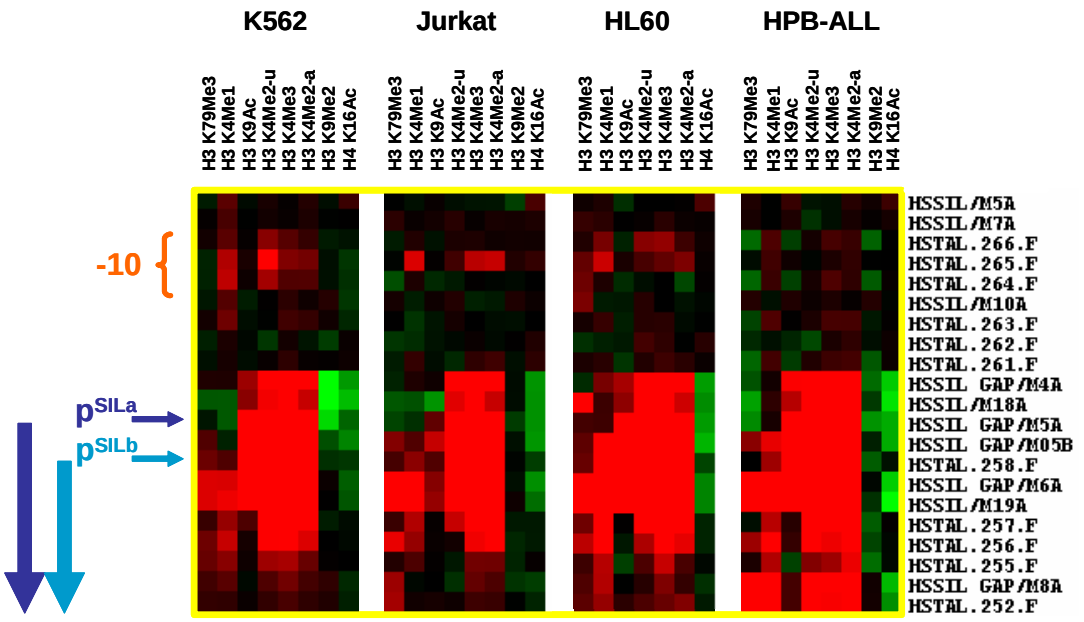


Figure 7.1: TreeView profiles at the 5' end of the SIL gene and the SIL -10 region in four human cell lines. The cell line corresponding to each profile is shown at the top. The well known and characterized SIL promoter (Colaizzo-Anas and Aplan 2003) is annotated as p^{SILa} and another region with potential promoter function has been annotated as p^{SILb}. The thin blue horizontal arrows represent the genomic positions of p^{SILa} and p^{SILb} with respect to the array elements of the human SCL tiling array. The thick blue vertical arrows at the left represent the direction of transcripts originating from p^{SILa} and p^{SILb} (see text). The location of the SIL -10 region is also shown in red at the left of the figure. The profiles were generated using only the key seven histone modifications which are listed along the x-axis at the top of the figure. Two datasets were included for H3 K4 dimethylation in each cell line. The names of the array elements representing the genomic regions are listed along the y-axis (extreme right). Each box in the profiles represents an array element for the corresponding histone modification and the red/black/green colour indicates enrichment/baseline levels/depletion for that modification respectively.

The 5'end of the SIL gene showed enrichments or depletions for various modifications across a region spanning the upstream -1 region to almost +3/+4 region. From the ChIP-chip profiles for various histone modifications, three distinct peaks of enrichments were observed in all four cell lines at -1, p^{SILa} and +1/+2 (the last of which is annotated as p^{SILb}) and are shown in Figure 7.2. The p^{SILb} region was also observed in the TreeView profiles, as a distinct region downstream of p^{SILa}, and exhibited some of the features of being an active regulatory element. These included hyper di- and trimethylation and hypo monomethylation of H3 K4, hypo dimethylation and hyper acetylation of H3 K9, and hypo acetylation of H4 K16. This data suggests that this region may be an alternative SIL promoter. On further investigation, it was found that in the ENSEMBL database (www.ensembl.org), three transcripts for SIL have been annotated - the 5' end of one of these transcripts (annotated as 'SIL' in Figure 7.2, panel E) maps to the p^{SILb} region.

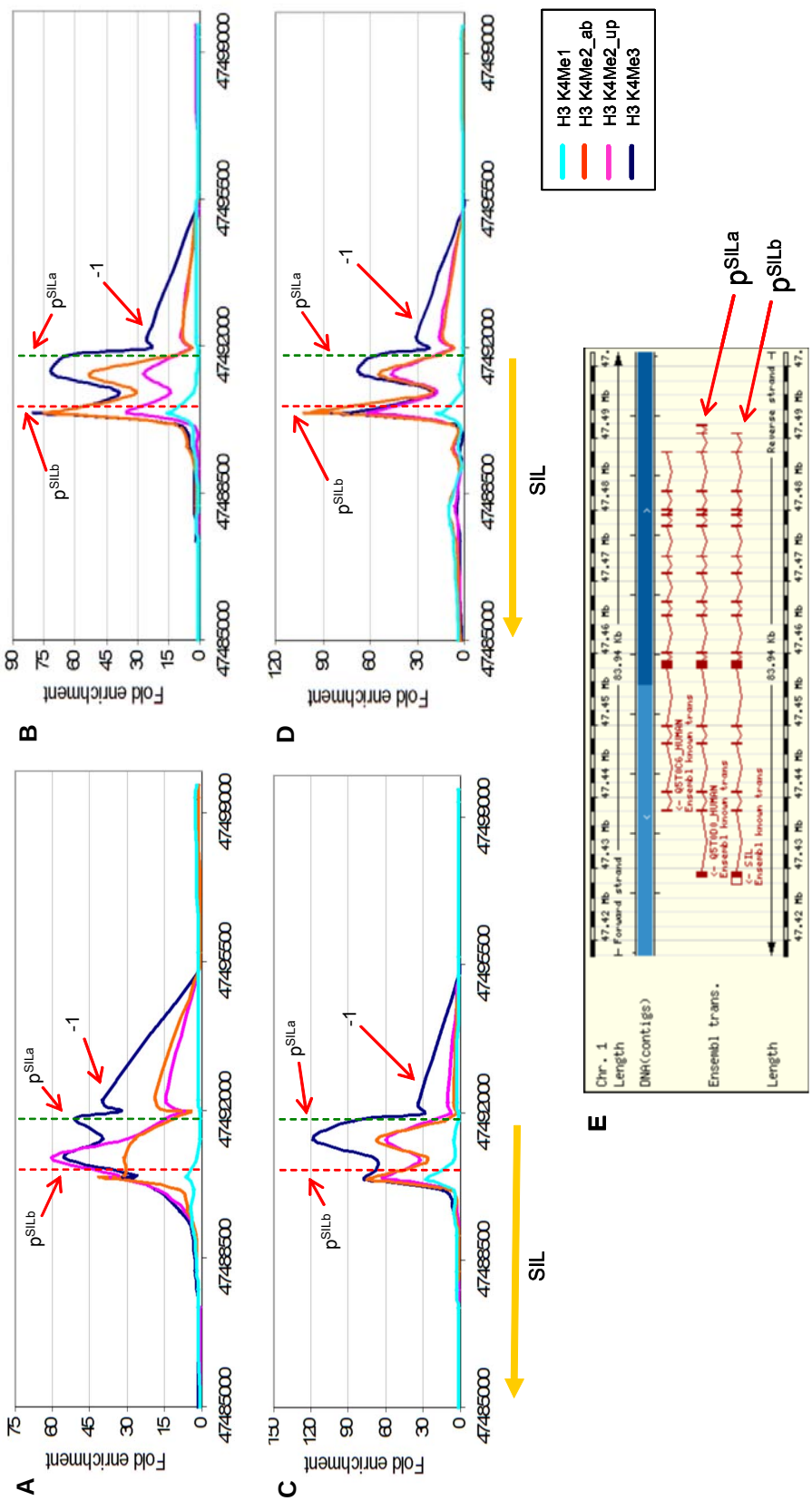


Figure 7.2: Composite ChIP-chip profiles across the 5' end of the SIL gene showing three distinct peaks. Panel A:K562, panel B: Jurkat, panel C: HL60 and panel D: HPB-ALL. The list of histone modifications is shown in a box at the bottom right of the figure. These include mono-, di- and trimethylation of H3K4 and are colour coded as shown in the box. In the profiles, the genomic positions of p^{SILa} and p^{SILb} are marked by green and red dotted lines respectively. It can be clearly seen in the ChIP-chip profiles (panels A-D, turquoise profiles) that the relative enrichments for monomethylation of H3 K4 are lower at p^{SILb} than its immediate downstream region. The yellow coloured arrow at the bottom of profiles B and D represent the SIL gene. The y-axes in the profiles show the fold enrichments and the x-axes represent genomic coordinates along human chromosome 1. Panel E shows the annotated SIL transcripts in ENSEMBL database and are marked with red arrows to show the mapping of p^{SILa} and p^{SILb} as shown in Figures 6.1 and 6.2.

Additionally, in the K562 cell line, binding interactions for CBP, HDAC2 (cofactors with HAT and HDAC activity) and $Taf_{II}250$ (component of pre-initiation complex) were also seen at this region (Chapter 5, sections 5.13, 514). These data provide support for the idea that p^{SILb} may represent an alternative SIL promoter. The histone code data suggests that this promoter may be active in K562. Taken together, these data suggested that the 5' end of the SIL gene contains two promoters, p^{SILa} and p^{SILb}, both of which show hallmarks of being active promoters which are consistent with respect to the consensus histone code in all the four human cell lines.

In addition, a third distinct peak at the SIL -1 region was observed as shown in Figure 7.2. This region is known to bind the transcription factor Elf-1 in K562, Jurkat and HPB-ALL (see chapter 5, section 5.15). This region also showed features of an active regulatory element including hypo-acetylation of H4 K16, hyper di- and trimethylation of H3 K4. Whether this represents a part of the SIL p^{SILa} promoter or a distinct regulatory element is not known, although it does have a histone code of that of a promoter region.

A region approximately 9 to 10 kb upstream of the 5' end of the SIL gene (annotated as SIL -10 and previously identified in chapter 5), showed histone modification marks in K562, Jurkat and HL60 but not in HPB-ALL (Figure 7.1). In the hierarchical clustering analysis of histone modifications across the SCL locus in K562 (Chapter 6, sections 6.8, 6.9), the SIL -10 region clustered with the known enhancers of SCL. The key histone marks displayed by this region, in K562, were consistent with the consensus histone code of an active enhancer element. These marks included hypo-acetylation for H4 K16, and hyper mono-, di- and trimethylation for H3 K4. However, in Jurkat, HL60 and HPB-ALL, this region is likely to be inactive according to its pattern of histone modifications in these three cell lines. These data suggest that a putative novel enhancer region is located at the SIL -10 region.

7.5 The KCY promoter

Figure 7.3 shows Treeview profiles across the 5' end and downstream regions of the KCY gene in K562, Jurkat, HL60 and HPB-ALL. The KCY gene is an ubiquitously expressed gene and thus, the promoter is expected to be active in the cell lines being tested. The profile showed that the genomic region displaying enrichments for various histone modifications extended almost 4 kb into the transcribed sequences. The histone marks present at the promoter region suggested that the KCY promoter is active in all the four cell lines. These marks included hypo monomethylation of H3 K4, hypo-acetylation of H4 K16, hyper di- and trimethylation of H3 K4 and hyper-acetylation of H3 K9. In addition, hyper trimethylation of H3 K79 was present at the regions immediately

downstream of the KCY promoter region. These data for the KCY promoter are in complete agreement with the proposed consensus histone code for the SCL locus. In addition, binding interactions for $Taf_{II}250$ were detected at the KCY promoter region (Chapter 5, section 5.14) in K562, thus providing further data to support that the KCY promoter is active in this cell line.

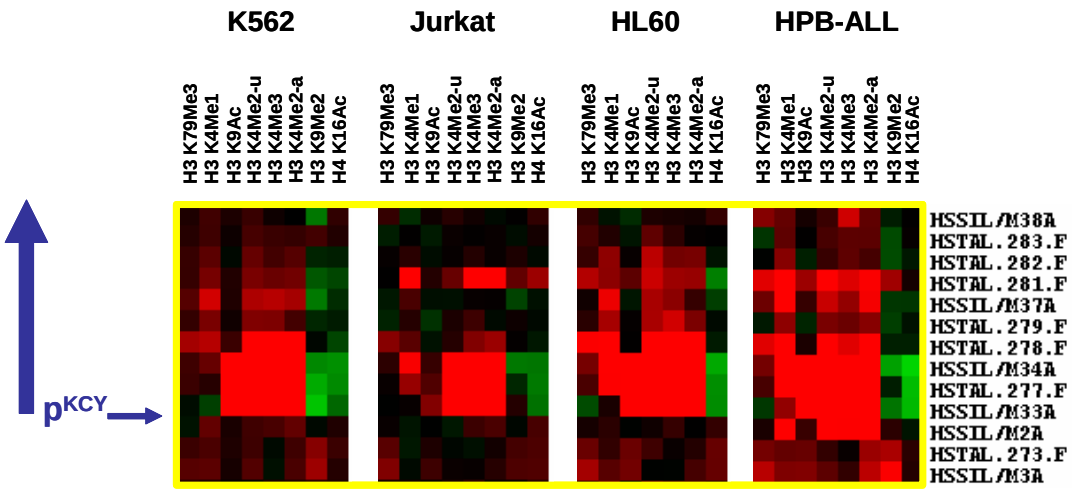


Figure 7.3: TreeView profiles at the 5' end of the KCY gene in four human cell lines. The cell line corresponding to each profile is shown at the top. The genomic position of the KCY promoter p^{KCY} is marked by the blue thin horizontal arrow and the thick blue vertical arrow represents the direction of transcription of the gene. The profiles were generated using only the key seven histone modifications which are listed along the x-axis at the top of the figure. Two datasets were included for H3 K4 dimethylation in each cell line. The array elements representing the shown genomic regions are listed along the y-axis at the extreme right of the figure. The names of the array elements representing the genomic regions are listed along the y-axis (extreme right). Each box in the profiles represents an array element for the corresponding histone modification and the red/black/green colour indicates enrichment/baseline levels/depletion for that modification respectively.

7.6 The SCL promoters and their immediate 3' flanking regions

Figure 7.4 shows the TreeView profiles across the 5' end of SCL gene encompassing the three known promoters (p^{1a}, p^{1b}, and p^{EXON4}) and the regulatory +1 and +3 regions. It is known that in K562, transcription is initiated at both p^{1a} and p^{1b}, whereas in Jurkat, transcription is initiated mostly at p^{1b} with some transcripts originating at p^{EXON4} (Aplan et al. 1990a; Begley et al. 1994; Bernard et al. 1992). Similarly, +1 and +3 are known enhancers which were first identified owing to their hypersensitivity to DNase I and activity in transient and stable reporter assays (Leroy-Viard et al. 1994; Gottgens et al. 1997; Fordham et al. 1999; Sinclair et al. 1999).

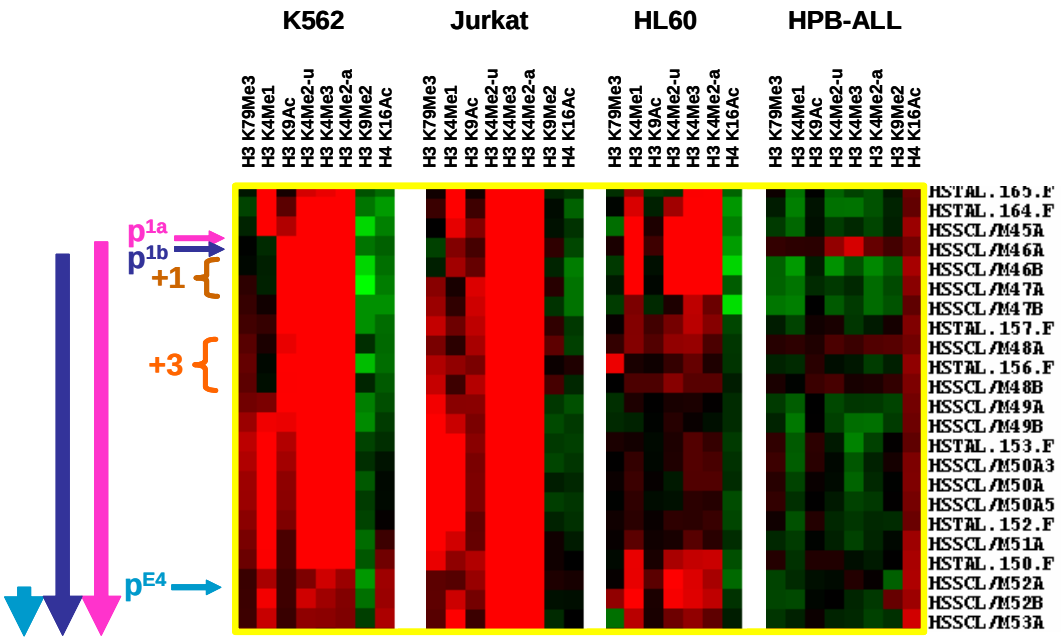


Figure 7.4: TreeView profiles at the 5' end of the SCL gene in four human cell lines. The cell line corresponding to each profile is shown at the top. The promoters, p^{1a}, p^{1b} and p^{EXON4} are marked with thin horizontal coloured arrows (colour coded with respect to the promoters) and the thick coloured vertical arrows represent the transcripts originating at these promoters and the direction of transcription. The +1 and +3 regions are also marked. The profiles were generated using only the key seven histone modifications which are listed along the x-axis at the top of the figure. Two datasets for H3 K9 dimethylation were included for each cell line. The names of the array elements representing the genomic regions are listed along the y-axis (extreme right). Each box in the profiles represents an array element for the corresponding histone modification and the red/black/green colour indicates enrichment/baseline levels/depletion for that modification respectively.

In K562, both p^{1a} and p^{1b} displayed hallmarks of active promoters, whereas the p^{EXON4} displayed histone marks associated with inactive promoters. The histone marks at p^{1a} and p^{1b} include hypo monomethylation of H3 K4, hypo-acetylation of H4 K16, hyper-acetylation of H3 K9 and hyper di- and trimethylation of H3 K4. The genomic region displaying these features extended up to the +3 region. Hyper trimethylation of H3 K79 was seen at the regions immediately downstream of p^{1a} and p^{1b} which is also a feature of active promoters. However, p^{EXON4} region showed hyper monomethylation of H3 K4, hypo-acetylation of H3 K9 and hyper-acetylation of H4 K16. In addition, the relative levels of di- and trimethylation of H3 K4 were very low as compared to that of p^{1a} and p^{1b} (Figure 7.5). It has already been demonstrated in chapter 6, that the tri- to monomethylation ratios for H3 K4 which depict the activity of a promoter more accurately, were found to be very low at p^{EXON4} in K562 (Chapter 6, section 6.5). Thus, the pattern of histone marks present at p^{EXON4} suggests that this promoter is inactive in K562. Binding interactions for GATA-1 were detected at the promoter and +3 regions of SCL in the K562 cell line (Chapter 4, section 4.6.2). GATA-1 binding at the SCL

promoter agrees with the previously known data (Aplan et al. 1990; Bockamp et al. 1995). The cofactors p300, CBP and HDAC2 also showed binding interactions at the SCL promoter in K562 along with $Taf_{II}250$ (Chapter 5, section 5.13, 5.14). All of these data support that SCL p^{1a} and p^{1b} are active but p^{EXON4} is inactive in the K562 cell line.

In Jurkat, p^{1b} and p^{EXON4} displayed hallmarks of active promoters including hypo monomethylation of H3 K4 (as compared to their immediate upstream and downstream regions) hyper di- and trimethylation of H3 K4, and hypo- or basal levels of H4 K16 acetylation (Figure 7.5). In addition, hyper trimethylation of H3 K79 was also present at the immediate downstream region of p^{1b}. Subtle enrichments of H3 K9 acetylation were also apparent (the reasons for this are discussed in section 7.3).

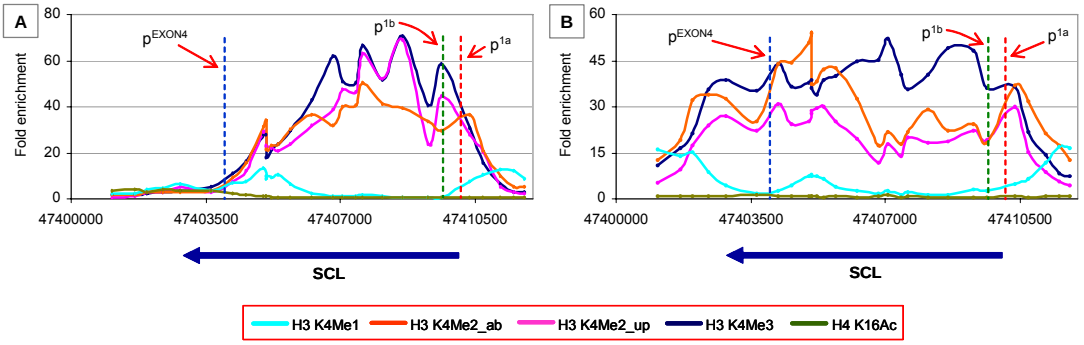


Figure 7.5: Composite ChIP-chip profiles across the 5' end of the SCL gene. Panel A: K562 and panel B: Jurkat. The histone modification shown in the above profiles include mono-, di- and trimethylation of H3 K4 and H4 K16 acetylation (colour coded as shown in the red box at the bottom of the figure). SCL p^{1a}, p^{1b} and p^{EXON4} are marked by red, green and blue dotted lines respectively. The thick blue arrows at the bottom of both panels represent the SCL gene and the direction of transcription. The y-axes represent fold enrichments and the x-axes represent genomic coordinates along the human chromosome 1.

In SCL non-expressing cell lines, the SCL p^{1a}, p^{1b}, and p^{EXON4} displayed hyper di- and trimethylation of H3 K4 and hypo-acetylation of H4 K16 in HL60 – these marks are linked to active regions (Figure 7.4). However, two features associated with inactive promoters were also present - hyper monomethylation of H3 K4 and hypo-acetylation of H3 K9. As shown in chapter 6 (section 6.5), despite the presence of all the methylation marks of H3 K4 at the SCL promoters in HL60, tri- to mono- ratios obtained were very low and reflect the inactivity of the SCL promoters (see chapter 6, section 6.5). Furthermore, the H3 K4 methylated regions in HL60 did not extend beyond the +1 region, whereas in Jurkat and K562 they extended across this region and included the +3 region - suggesting that the +3 region may be important in SCL activation. In HPB-ALL, the inactivity of the promoter was evident by the apparent lack of active histone marks and the presence of hyper-acetylation of H4 K16 – the histone mark associated with inactive regulatory elements.

7.7 Regulatory elements upstream of the SCL promoter 1a

The known SCL regulatory regions located upstream of the 5' end of the gene include the endothelial enhancer at -3/-4 (Gottgens et al. 2004), the -7 and -9/-10 regions (Leroy-Viard et al. 1994; Gottgens et al. 1997) were further analysed and compared for the presence and/or absence of histone marks in the four cell lines (Figure 7.6). A novel region identified in the present study at -12 (see chapter 4, section 4.5.2) was also characterized further.

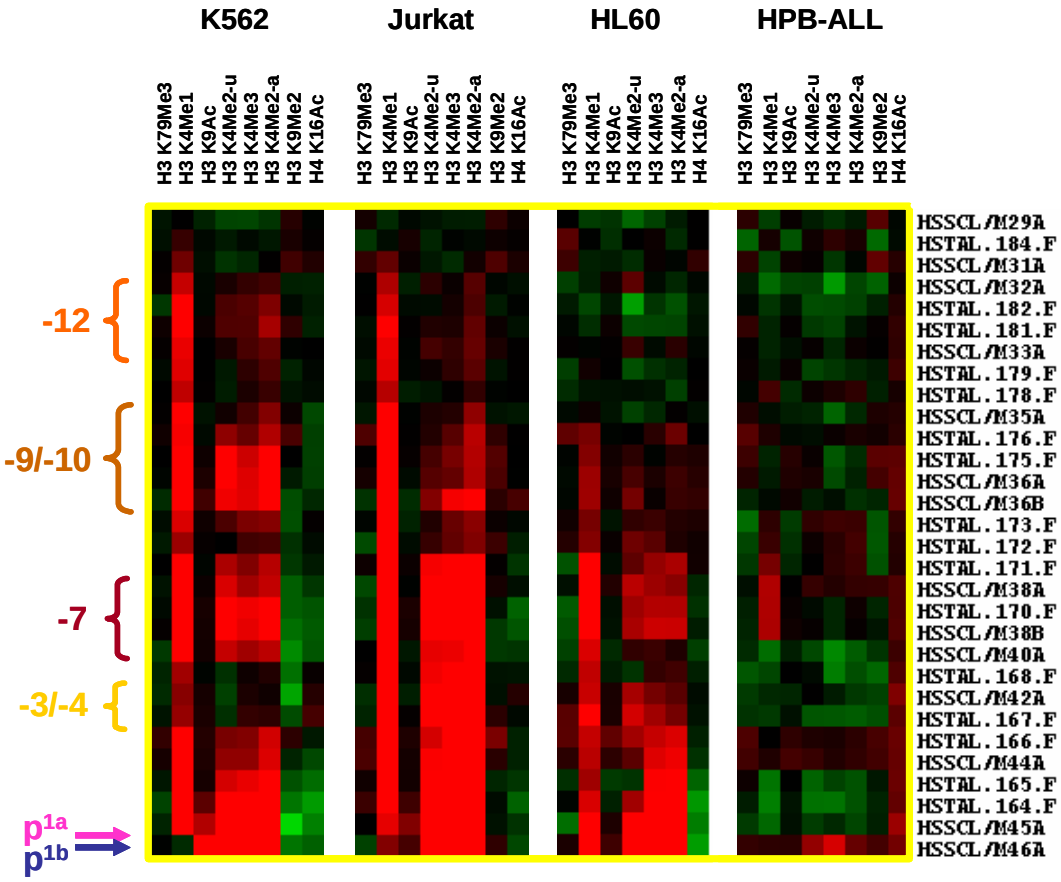


Figure 7.6: TreeView profiles across the genomic regions upstream of the SCL promoter in the four cell lines. The cell line corresponding to each profile is shown at the top. The SCL promoters, p^{1a}, p^{1b} are marked with small coloured horizontal (blue and pink) arrows. The SCL regulatory regions located upstream of the SCL promoter are marked, including the -3/-4 endothelial enhancer, and the -7, -9/-10 and -12 regions. The characterization of the above genomic regions with respect to the histone marks, function and activity is described in the text. The profiles were generated using only the key seven histone modifications which are listed along the x-axis at the top of the figure. Two datasets were included for H3 K4 dimethylation in each cell line. The names of the array elements representing the genomic regions are listed along the y-axis (extreme right). Each box in the profiles represents an array element for the corresponding histone modification and the red/black/green colour indicates enrichment/baseline levels/depletion for that modification respectively.

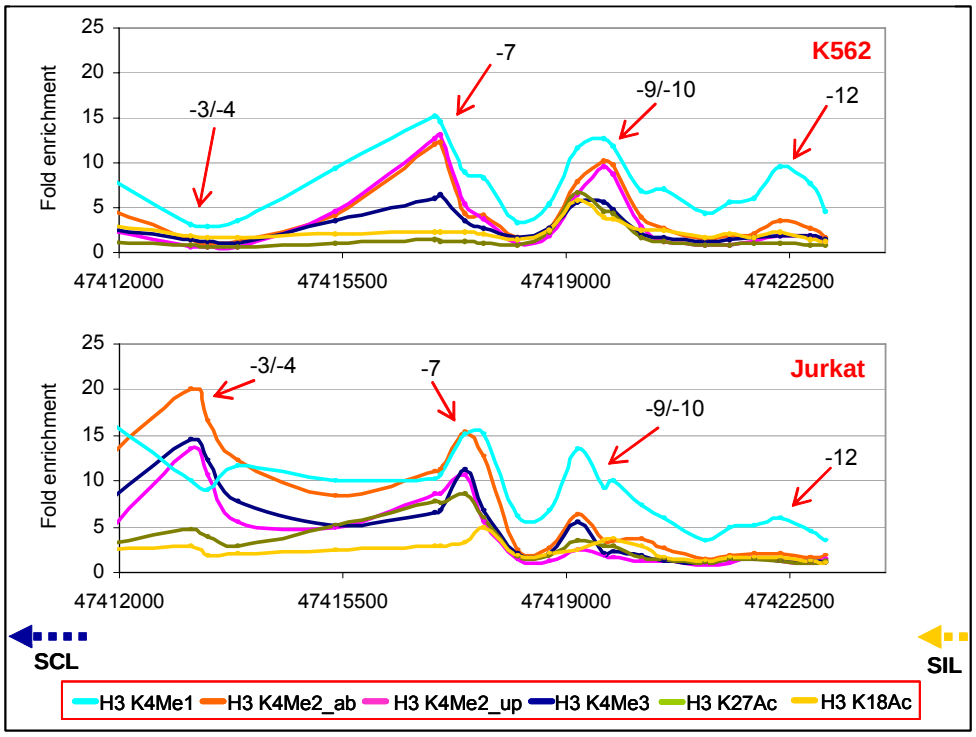


Figure 7.7: Composite ChIP-chip profiles across the SCL regulatory region located upstream of the SCL promoter. Top panel: K562 and bottom panel: Jurkat. The histone modifications shown in the above profiles include H3 acetylation at K18 and K27, mono-, di- and trimethylation of H3 K4 (colour coded as shown in the red box at the bottom of the figure). SCL regulatory regions located upstream of the SCL promoter at -12, -9/-10, -7, and -3/-4 are marked with red arrows. The thick, dotted arrows (yellow: SIL and blue: SCL) at the bottom of the figure indicate the genes on either side of the above shown genomic region and the direction of transcription of these genes. The y-axes represent fold enrichments and the x-axes represent genomic coordinates along the human chromosome 1.

7.7.1 The -9/-10 region

The -9/-10 (-8/-9 in mouse) has been shown to have enhancer activity (Gottgens at el. 1997), DNase I hypersensitivity (Leroy-Viard et al. 1994) and GATA-1 binding (Chapter 4, section 4.6.2) in K562. This region exhibited hyper mono-, di- and trimethylation of H3 K4 and hypo-acetylation of H4 K16 (Figure 7.6) in K562, further suggesting that this region may be active. This was also supported by the presence of histone H3 acetylation marks (Figure 7.7, top panel). In Jurkat, hyper mono-, di- and trimethylation of H3 K4 were present but the relative enrichments for di- and trimethylation H3 K4 and histone H3 acetylation marks were low as compared to K562 (Figure 7.6; Figure 7.7, bottom panel); increased levels of H3 K9 dimethylation were also observed and depletion of H4 K16 acetylation was seen on only one of the array elements at the -9/-10 region. These data suggest that this region may be less active or inactive in Jurkat. The absence of DNase I hypersensitivity at this region in Jurkat (Leroy-Viard et al. 1994) further supports these interpretations.

In HL60 and HPB-ALL, the -9/-10 region showed histone marks associated with inactive regions, such as, hyper-acetylation of H4 K16, hypo di- and trimethylation of H3 K4. These results are consistent with this region being inactive in SCL non-expressing cell lines.

7.7.2 The -7 region: evidence for an active enhancer in K562

The -7 region had previously been shown to exhibit DNAse I hypersensitivity in K562 (Leroy-Viard et al. 1994) and binding of GATA-1 in K562 (Chapter 4, section 4.6.2). However, its function in gene regulation is not known. The histone marks present at this region included hyper mono-, di- and trimethylation of H3 K4, and hypo acetylation of H4 K16 (Figure 7.6) which are all linked to active enhancer regions. Furthermore, the -7 region clustered along with the other known SCL enhancers in the functional clustering analysis (see Figure 6.5, chapter 6).

This region was further tested in transient luciferase reporter assays in K562 along with the -12 region (see section 7.7.4) and the +51 erythroid enhancer which acted as a positive control (Figure 7.8). The transient assays report the expression of the firefly luciferase gene under the control of the SV40 promoter. In the presence of an enhancer sequence cloned into the SV40 promoter-luciferase construct, increased levels of luciferase expression are observed. Both the -7 and +51 regions showed increased luciferase expression, suggesting that the -7 region had enhancer activity in K562. This provides evidence that the -7 region is a novel enhancer which is likely to regulate SCL expression in K562.

This novel enhancer displayed the same histone marks in Jurkat as was seen in K562. Noticeable nucleosome depletion was also seen at the -7 region in Jurkat (Chapter 5, section 5.11) as well as extreme hypersensitivity to DNase I (Leroy-Viard et al. 1994). It was further noticed that although significant enrichments were reported at the -7 region for H3 acetylation in both K562 and Jurkat, the relative enrichments were much higher in Jurkat as compared to K562 (see Figure 7.7). All of this, taken together, suggests that the -7 enhancer is active in Jurkat. In addition, it has been previously reported that the -7 region was found near one of the SCL breakpoints reported in SIL-SCL deletions (Breit et al 1993). This further alludes to the -7 enhancer being involved in the *cis*-activation of SCL in Jurkat.

This region displayed the presence of similar histone marks in HL60 as seen in K562, however, the relative enrichment levels for di- and trimethylation of H3 K4 were found to be very low in HL60 as compared to K562. Hyper-acetylation of H4 K16 and hypo di- and

trimethylation of H3 K4 was seen at the -7 region in HPB-ALL. These results are consistent with this region being inactive in SCL non-expressing cell lines

7.7.3 The endothelial enhancer at the -3/-4 region

The -3/-4 region, also known as the endothelial enhancer, directed expression to haematopoietic progenitors and endothelium and is expressed in blood progenitors and endothelial cells. This enhancer was found to be regulated by the transcription factors Elf-1 and Fli-1 (Gottgens et al. 2004).

In K562, this region exhibited hypo di- and trimethylation of H3 K4 and hyper-acetylation of H4 K16, all of which are histone marks for inactive regulatory region (Figure 7.6). Hypo- K18 and K27 acetylation was also a hallmark of this region in K562 (Figure 7.7). Conversely, in Jurkat, substantial hyper mono-, di- and trimethylation of H3 K4 were present at this region, and H4 K16 acetylation levels were lower than in K562. Significant enrichments for acetylation of histone H3 at K18 and K27 were also seen at this region in Jurkat (Figure 7.7). In a previous study, a DNase I hypersensitive site was identified at -3/-4 region in Jurkat only and not in K562 (Leroy-Viard et al. 1994). These data suggest that the -3/-4 region may not be active in K562 but is possibly active in Jurkat and may also represent a region which could be involved in *cis*-activation of SCL in Jurkat.

In HL60, significant but relatively low level enrichments for active histone marks such as di- and tri-methylation of H3 K4 were observed, but this was not accompanied by hypo-acetylation of H4 K16. In HPB-ALL hyper-acetylation of H4 K16 was observed. These results are consistent with this region being inactive in SCL non-expressing cell lines.

7.7.4 The -12 region: a putative enhancer of SCL regulation

The -12 region was identified through its association with significant levels of H3 K9/14 acetylation in K562 (Chapter 4, section 4.5.2). In K562, it also showed significant enrichments for H3 K18 acetylation (Figure 7.7), along with mono-, di- and trimethylation of H3 K4 and hypo-acetylation of H4 K16. This suggests that this region could represent a putative regulatory element which may be active in K562. In Jurkat, although hyper monomethylation was present, the enrichments for di- and trimethylation of H3 K4 were very low (Figures 7.6 and 7.7) suggesting that this region was inactive. In HL60 and HPB-ALL, the presence of hypo mono-, di- and trimethylation of H3 K4 and hyper-acetylation of H4 K16 were consistent with this region being inactive in SCL non-expressing cell lines (Figure 7.6).

To elucidate the function of the -12 region as a regulatory element, genomic sequences from this region were tested in transient luciferase reporter assays in the K562 cell line

(Figure 7.8). Two different constructs across the -12 region (denoted -12b and -12s in Figure 7.8) were used in these assays along with the -7 region (discussed in section 7.7.2) and the +51 erythroid enhancer which acted as a positive control. The -12 constructs decreased expression of the reporter suggesting that this sequence acted as a silencer/repressor of transcriptional activity (see Appendix 13 for the sequences of the primer pairs used to generate the constructs).

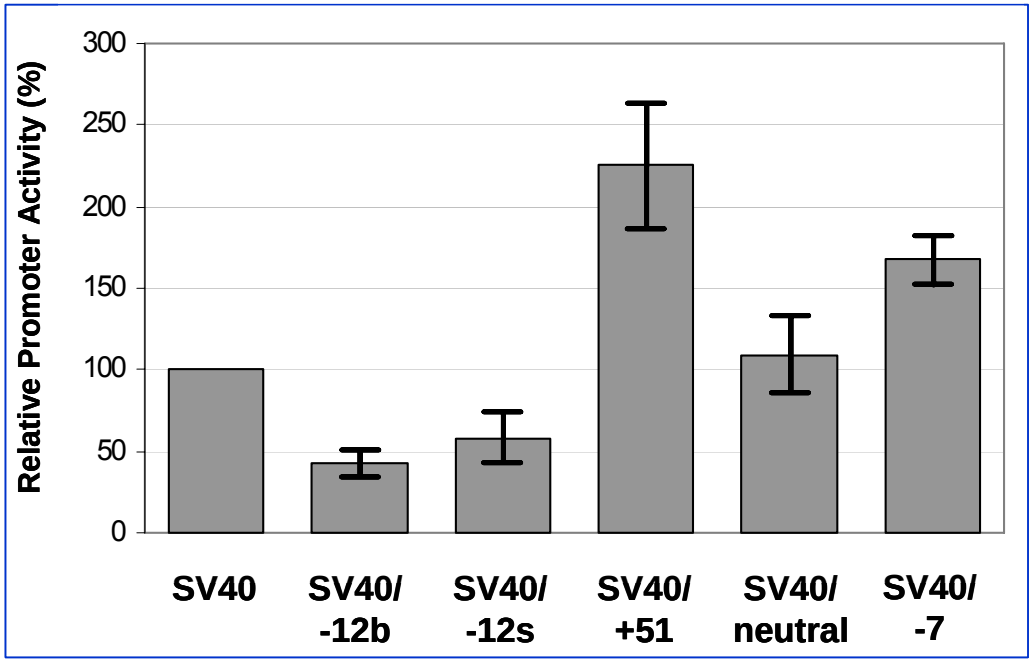


Figure 7.8: Transient luciferase reporter assays in K562 cells. The histogram shows luciferase activities for five constructs containing genomic DNA inserts from the human SCL locus. The y-axis is the luciferase activity expressed as a percentage of the SV40 promoter activity. Standard deviation bars are shown for each construct apart from SV40, to which all data was normalised. Labels on the x-axis are as follows: SV40 (promoter only), SV40/-12b (2.2 kb from the human -12 region), SV40/-12s (453 bp from the human -12 region), SV40/+51 (936 bp from the SCL erythroid enhancer at +51), SV40/neutral (951 bp from the CYP4Z1 gene used as a human DNA negative control), SV40/ -7 (952 bp from the human SCL -7 region).

The -12 region showed stretches of high non-coding sequence conservation in a number of mammalian species which suggests that this putative silencer/repressor may have an evolutionarily conserved function. In order to further elucidate the regulatory interactions which could occur at the -12 region, transcription factor binding sites which were conserved for the orthologous sequences in human, chimp, mouse, rat and dog were identified including an 11 bp sequence for the ETV6/7 (TEL1/2) proteins (Figure 7.9).

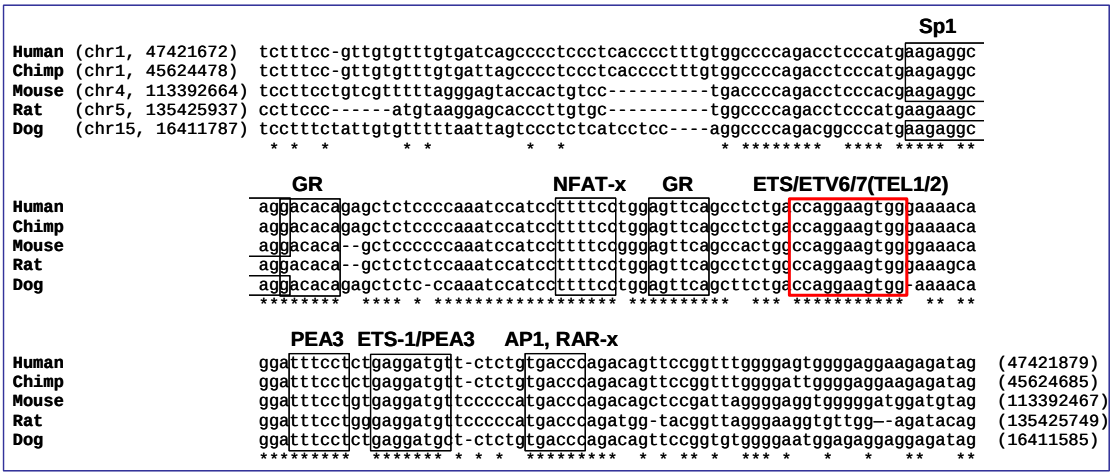


Figure 7.9: Conserved transcription factor binding sites found at the SCL -12 region. Sequence alignments are shown for human, chimp, mouse, rat and dog. Genomic sequence co-ordinates for each region of homology are shown in brackets at the beginning and the end of each sequence. Bases of sequence identity are denoted with an asterisk (*). The transcription factor binding site for ETV6/7 (TEL1/2) is boxed in red. Sites are shown for a variety of other transcription factors including Sp1, PEA3, ETS-1, GR (glucocorticoid receptors), RAR-x (retinoic acid receptors), AP-1,2,4 (activator proteins), and NFAT-x (nuclear factors of activated T cells), all boxed in black.

ETV6 (TEL1) and ETV7 (TEL2) are members of the Ets family of transcription factors, are expressed during haematopoiesis and have been shown to exhibit strong repressor activity *in vivo* (Chakrabarti and Nucifora 1999; Lopez et al. 1999; Poirel et al. 2000; Gu et al. 2001). Therefore, these proteins represent good candidates which may mediate the repressor activity of the -12 element.

7.8 Regulatory regions downstream of the SCL gene

7.8.1 The stem cell enhancer at the +20/+21 region

It has been shown that the SCL stem cell enhancer at +20/+21 (+18/+19 in mouse) is regulated by GATA-2, Elf-1, Fli-1 and targets a vast majority of haematopoietic progenitors (Gottgens et al. 2002). Thus, this enhancer is active at an early stage of blood development. However, SCL expression is not dependent on the activity of this enhancer as initiation of SCL transcription or for formation of haematopoietic cells can occur without the +20/+21 region being functional (Gottgens et al. 2004).

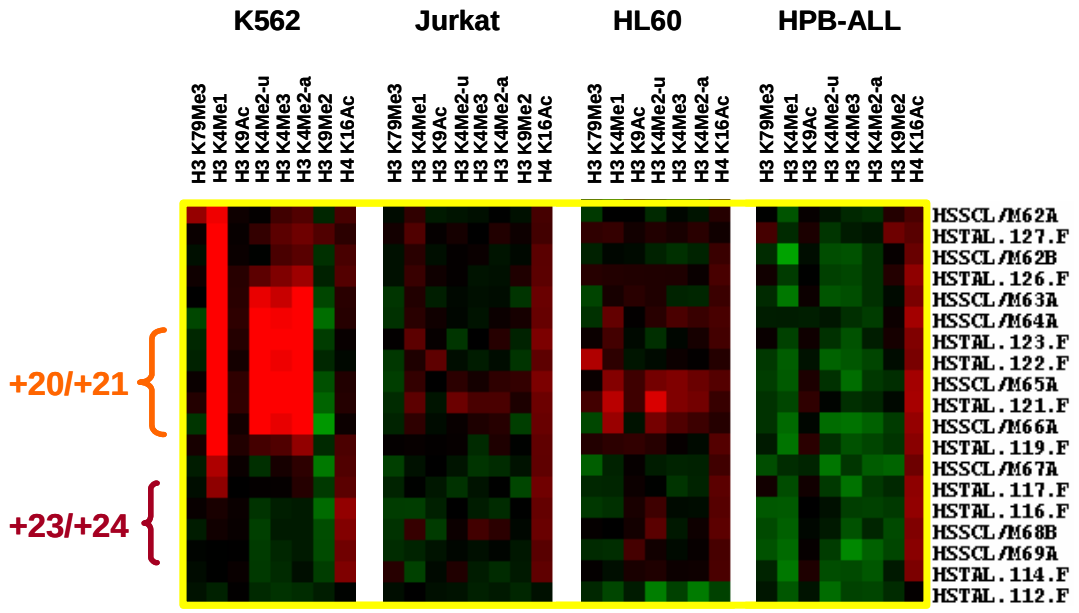


Figure 7.10: TreeView profiles across the stem cell enhancer and the neural enhancer in the four human cell lines. The cell line corresponding to each profile is shown at the top. The genomic regions encompassing the stem cell enhancer at +20/+21 (homologous to +18/+19 in mouse) and the neural enhancer at +23/+24 (homologous to +23/+24 in mouse) are marked on the left. The characterization of the above genomic regions with respect to the histone marks, function and activity is described in the text. The profiles were generated using only the key seven histone modifications which are listed along the x-axis at the top of the figure. Two datasets were included for H3 K4 dimethylation in each cell line. The names of the array elements representing the genomic regions are listed along the y-axis (extreme right). Each box in the profiles represents an array element for the corresponding histone modification and the red/black/green colour indicates enrichment/baseline levels/depletion for that modification respectively.

In K562, the +20/+21 region displayed hyper mono-, di- and trimethylation of H3 K4 which are all active histone marks (Figure 7.10). The presence of hypo/hyper acetylation of H4 K16 at this region is not pronounced and levels are near background levels. However, in comparison with H4 K16 levels at other SCL regulatory regions, the levels at the +20/+21 region are markedly higher (see chapter 5, Figure 5.9), suggesting a reduced activity for this region in K562. However, no binding interactions for Elf-1 and Fli-1 were detected at this region in K562, suggesting that although the activating histone marks may be in place, the region is not active (or other transcription factors may be recruited in K562). Further characterization of this region would be required to deduce its activity and function in the K562 cell line.

In Jurkat, HL60 and HPB-ALL, the interpretation was far easier for the +20/+21 region. It exhibited histone marks that are linked with inactive regions such as hypo mono-, di- and

trimethylation of H3 K4 and hyper-acetylation of H4 K16 suggesting that the +20/+21 region was inactive in these cell lines.

7.8.2 The neural enhancer at the +23/+24 region

The +23/+24 region was originally identified using comparative sequence analysis with the syntenic SCL regions in human, mouse and chicken. Transgenic reporter assays established that this enhancer directed reporter expression to hindbrain and spinal cord (Gottgens et al. 2000). In K562, Jurkat, HL60 and HPB-ALL, the +23/+24 region exhibited hypo mono-, di- and trimethylation of H3 K4 and hyper-acetylation of H4 K16, all of which suggest that this region is inactive in these cell lines (Figure 7.10). This is consistent with the known biology of this enhancer having activity in neural development but not during haematopoiesis.

7.8.3 The +39 region

The +39 region was identified in this study by virtue of its association with various histone modification assays in four human cell lines (See chapter 5). This region, exhibited all the features of an enhancer element based on the hierarchical clustering analysis presented in chapter 6 (section 6.8; 6.9).

This region displayed hyper mono-, di- and trimethylation of H3 K4, and hypo-acetylation of H4 K16 in all the cell lines (see Figure 7.11, top panel), although the relative levels of methylation were quite low in HPB-ALL as compared to other cell lines. In addition, decreased levels of nucleosomes were also observed at this region (Figure 7.11, bottom panel) in all four human cell lines. This data, taken together suggests that +39 is active in all of the cell lines analysed here. In K562, binding interactions for the transcription factor CTCF were detected at the +39 region implicating it in enhancer-blocking activities of insulators, or possibly suggesting that +39 was an insulator itself (Chapter 5, section 5.17.6.2). As the +39 region lies close to and upstream of MAP17 gene (3 kb upstream), it is not known whether this region is associated with regulation of MAP17 or whether it represents a 3' distant element associated with expression of SCL. Therefore, the function of +39 needs further characterization.

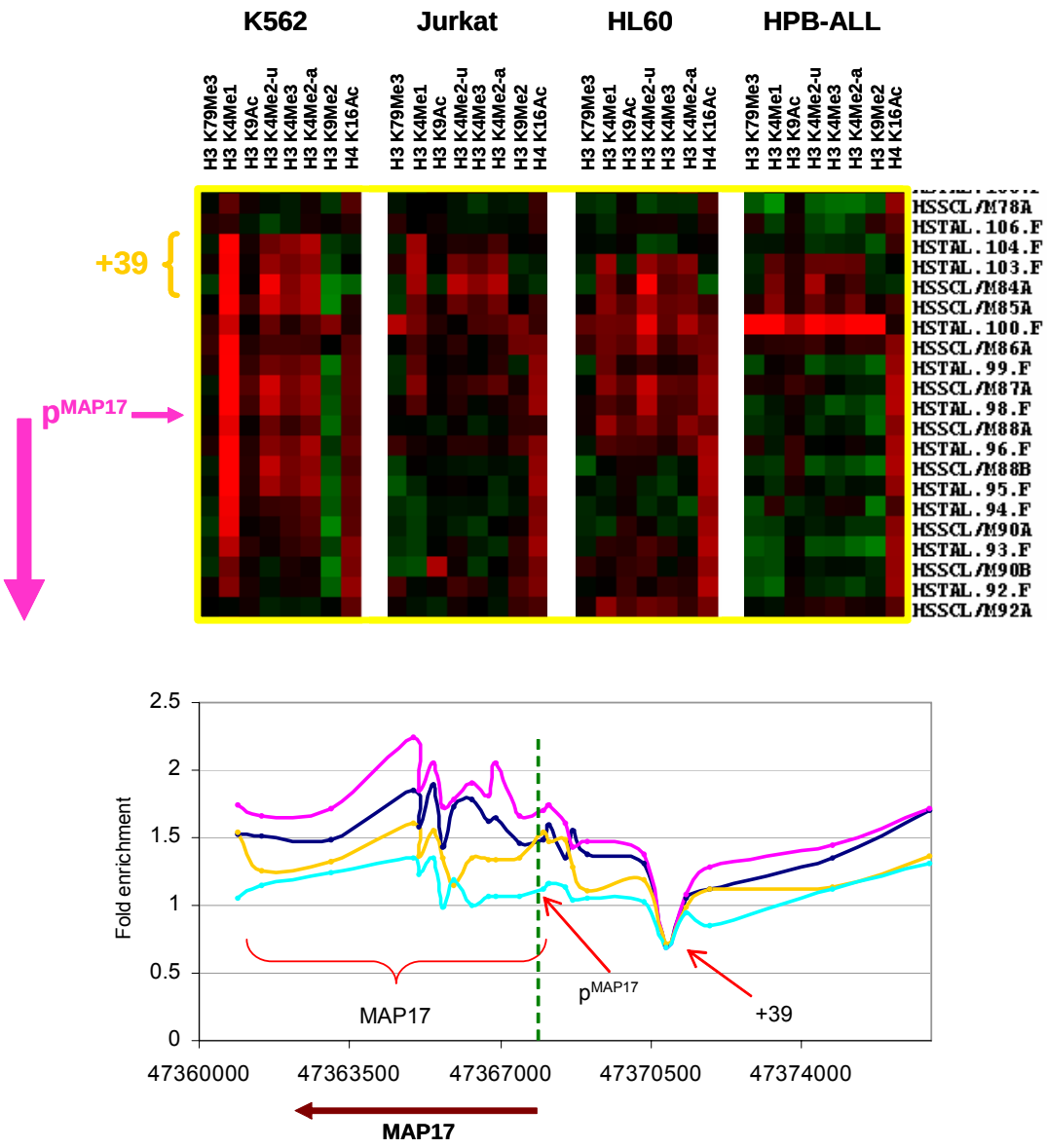


Figure 7.11: TreeView profiles and composite ChIP-chip profiles across the +39 region and MAP17 gene in four human cell lines. The top panel shows the TreeView profiles. The cell line corresponding to each profile is shown at the top. The +39 region is marked on the left. The MAP17 promoter, p^{MAP17} is shown with the thinner pink coloured horizontal arrow and the thick vertical arrow represents the direction of transcription of the MAP17 gene. The characterization of the above genomic regions with respect to the histone marks, function and activity is described in the text. The profiles were generated using only the key seven histone modifications which are listed along the x-axis at the top of the figure. Two datasets were included for H3 K4 dimethylation in each cell line. The names of the array elements representing the genomic regions are listed along the y-axis (extreme right). Each box in the profiles represents an array element for the corresponding histone modification and the red/black/green colour indicates enrichment/baseline levels/depletion for that modification respectively. The bottom panel shows the composite ChIP-chip profiles for histone H3 in the four human cell lines. The regions are marked by red arrows and the genomic position of the p^{MAP17} is shown by the dotted green line. K562: turquoise, Jurkat: blue, HL60: pink and HPB-ALL: yellow profiles. The thick arrow at the bottom represents the MAP17 gene and the direction of transcription. The y-axis in the bottom panel represents fold enrichments and the x-axis represents genomic coordinates along human chromosome 1.

7.8.4 The MAP 17 promoter

In a previously reported study, it was demonstrated that the cell lines which express SCL also co-express MAP17, albeit at low levels (Delabesse et al. 2005). Thus, it was proposed that the co-expression of these genes, was due to their co-regulation by shared regulatory sequences (Delabesse et al. 2005). Thus, the MAP 17 promoter would be expected to represent an active region in K562 and Jurkat, and would be inactive in HL60 and HPB-ALL.

In K562, the promoter region of MAP17 exhibited hyper monomethylation of H3 K4, hyper-acetylation of H4 K16, hypo trimethylation of H3 K4 and hypo-acetylation of H3 K9 - all these histone marks are associated with inactive promoter regions. This suggested that the promoter of MAP17 gene is inactive in K562 (Figure 7.11, top panel). Similarly, in Jurkat, the presence of hypo-acetylation of H3 K9, hypo trimethylation of H3 K4 and hyper dimethylation of H3 K9 suggested that, the MAP17 promoter was also inactive in Jurkat. Additionally, hyper trimethylation of H3 K79 downstream of active promoters was not seen at the region downstream of the MAP17 promoter in K562 or Jurkat.

Additional sources of evidence also support the idea that MAP17 is not active in the SCL-expressing cell lines studied here:

i) It was shown in chapter 6, that the tri- to monomethylation of H3 K4 were very low for the MAP17 promoter as compared to all the other active promoters across the SCL locus in K562 and Jurkat (section 6.5).

ii) The MAP17 promoter region did not display relatively low levels of nucleosomes (Figure 7.11, bottom panel).

iii) In ChIP-chip assay for $Taf_{II}250$ (which binds to active promoters), no significant enrichments were reported at the MAP17 promoter (Chapter 5, section 5.14).

iv) Binding interactions for CTCF transcription factor, which is known for its function as an enhancer-blocker (Bell et al. 1999), were also detected at the MAP17 promoter in the K562 cell line (Chapter 5, section 5.16). This interaction suggests that the promoter of MAP17 is possibly being blocked to modulate its expression (Chapter 5, section 5.17.6.2).

v) Histone modification profiles in HL60 and HPB-ALL, looked similar to those found in K562 and particularly to those in Jurkat; histone modification profiles linked with inactive regions were present at the MAP17 promoter which was in complete agreement with the known inactivity of this promoter in SCL non-expressing cell lines (Figure 7.11).

7.8.5 The erythroid enhancer at the +51/+53 region: evidence for sites of transcription downstream of the +51 region.

The +51 region (homologous to +40 in mouse) was identified as an erythroid enhancer which directed reporter expression to primitive erythroblasts (Delabesse et al. 2005). In the present study, binding interactions for GATA-1, SCL and Ldb-1, p300, CBP and HDAC2 were detected at the +51 region in K562 (see chapters 4 and 5). The highest ChIP enrichments in all the above mentioned ChIP assays were observed at the same array element – named HSSCL/M96B in Figure 7.12.

In K562, the +51 erythroid enhancer (centered around the HSSCL/96B array element) exhibited hyper mono-, di- and trimethylation of H3 K4 and hypo-acetylation of H4 K16 - the presence of these marks is in complete agreement with the known activity of this enhancer in K562 (Figure 7.12). In Jurkat and HL60, lower levels of mono-, di- and trimethylation of H3 K4 and hyper-acetylation of H4 K16 histone marks suggested this enhancer may not be active in these cell lines. Similarly, the presence of hyper-acetylation of H4 K16 in HPB-ALL suggested that this enhancer may not be active in this cell line.

However, there was a noticeable pattern of histone modifications across the genomic region spanning from +50 to +53 in K562 - covering the +51 enhancer and extending in either direction from it. This whole 4 kb region exhibited significant enrichments for histone modifications including acetylation of histone H3 at K9/14, K18, K27 and mono-, di- trimethylation of H3K4 (see Figures 7.12 and 7.13). The relative levels of enrichments for H3 K18 acetylation were higher upstream of +51 (towards the MAP17 gene), whereas enrichments for H3 K27 acetylation were higher at the downstream region of the +51 enhancer at the +53 region (Figure 7.13). This suggests that a different set of modifications marks are present on either side of the +51 enhancer. This was further supported by the presence of a peak of enrichment for H3 K4 di- and trimethylation at +53, and a smaller peak upstream of +51 (Figure 7.13).

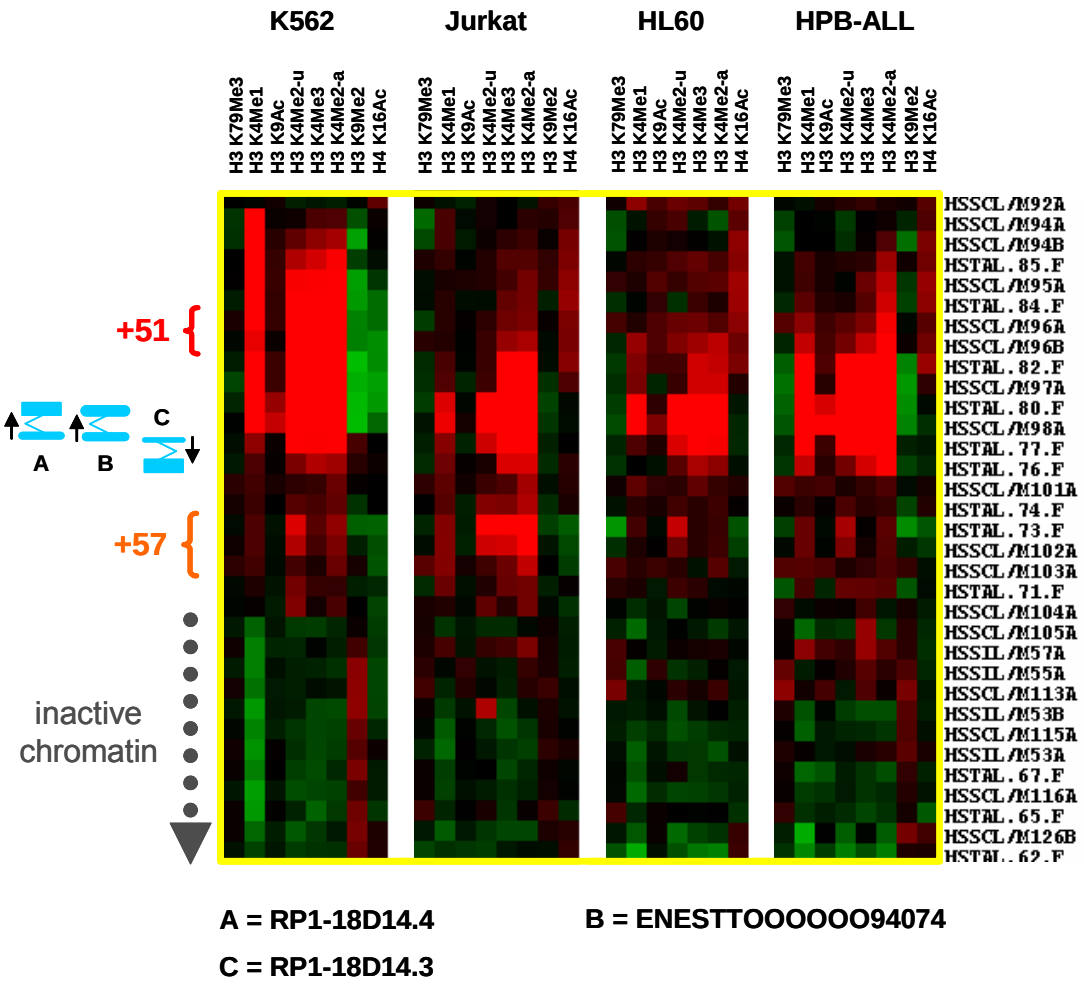


Figure 7.12: TreeView profiles across the erythroid enhancer and the +57 region in four human cell lines. The cell line corresponding to each profile is shown at the top. The genomic regions encompassing the erythroid enhancer at +51 region and the +57 regions are marked on the left. The characterization of the above genomic regions with respect to the histone marks, function and activity is described in the text. The three blue small transcripts (shown on the left) are annotated in the ENSEMBL database and named A, B and C (shown with their ENSEMBL IDs at the bottom of the figure). The small black arrows next to the blue transcripts represent the direction of transcription for these transcripts according to ENSEMBL. The dotted grey vertical arrow shown at the downstream region of +57 represents the beginning of the silent chromatin encompassing the CYP4A22 and CYP4Z1 genes. The above profiles were generated using only the key seven histone modifications which are listed along the x-axis at the top of the figure. Two datasets were included for H3 K4 dimethylation in each cell line. The names of the array elements representing the genomic regions are listed along the y-axis (extreme right). Each box in the profiles represents an array element for the corresponding histone modification and the red/black/green colour indicates enrichment/baseline levels/depletion for that modification respectively.

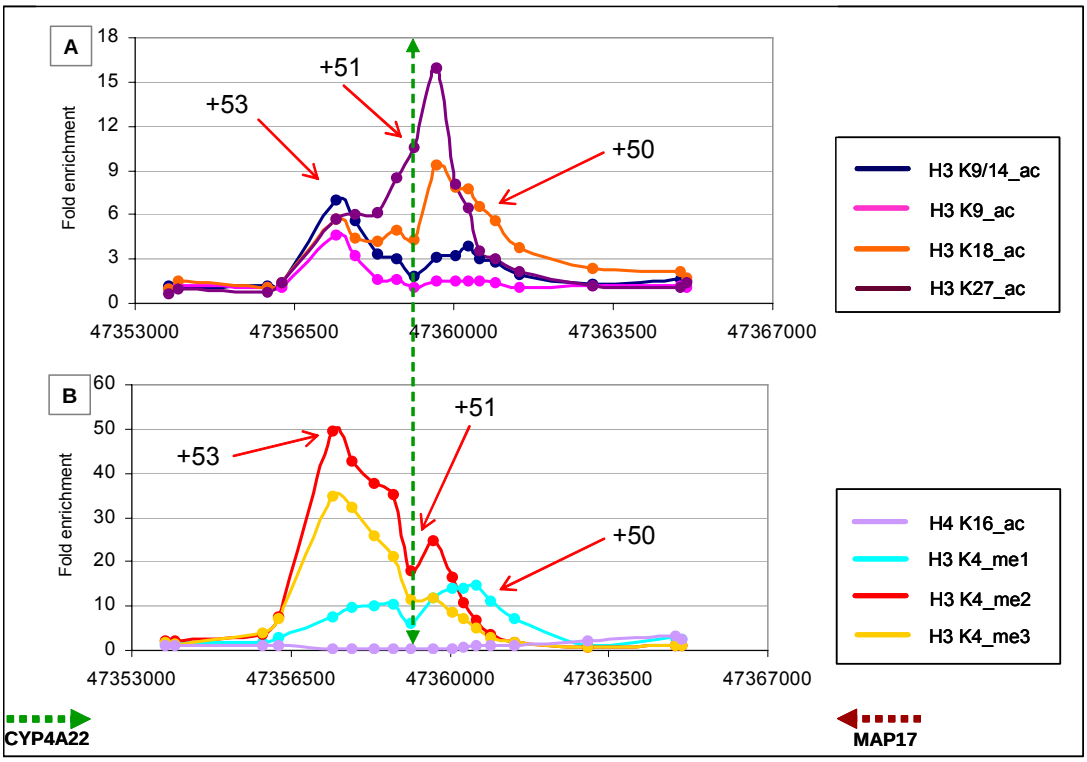


Figure 7.13: Composite ChIP-chip profiles for histone modifications at the +51/+53 regions in K562. The histone modification used in generating the above profiles included H3 K9/14 diacetylation, acetylation of histone H3 at K9, K18 and K27 (panel A) and H4 K16 acetylation, mono-, di- and trimethylation of H3 K4 (panel B). The colour codes of the profiles are colour coded as shown in the boxes on the right. The location of the array element (HSCL/96B) which showed transcription factor binding is represented by the green dotted arrow in panels A and B of the figure. The +50, +51 and +53 regions are marked with red arrows. The dotted thick arrows at the bottom of the figure represent the genes (and the direction of transcription) located on either side of the +50 to +53 region. The y-axes represent fold enrichments and the x-axes represent the genomic coordinates along human chromosome 1.

Another striking observation in the TreeView profiles shown in Figure 7.12 was that the cell lines Jurkat, HL60 and HBP-ALL only showed marked histone modifications on the downstream (+53 side) (Figure 7.12) part of this region distal to the location of array element HSCL/96B. This suggested that the region distal to +51 may have a regulatory function which is distinct from that of the erythroid enhancer, and which was active in all of the cell lines. In the hierarchical clustering analysis performed for histone modifications in K562, the +51/+53 region clustered together with the SCL p^{1a}, whereas the +50/+51 region clustered with the known SCL enhancers (Chapter 6, sections 6.8 and 6.9). These data suggest that the +50 to +53 region may exhibit features of both an enhancer (i.e., the known erythroid enhancer) and an active promoter, delineated on either side of the +51 region. Evidence to further support the presence of a promoter near +53 is as follows:

i) The +53 region showed hyper-acetylation of H3 K9, hyper di- and trimethylation of H3 K4, hypo dimethylation of H3 K9 and hypo-acetylation of H4 K16 in all four cell lines - all these are hallmarks of active promoters (Figure 7.12 and Figure 7.13).

ii) The ratios for tri- to monomethylation for H3 K4 across the +53 region were high (in K562, Jurkat and HPB-ALL), which is indicative of the transcriptional activity of a promoter region (Chapter 6, section 6.5).

iii) Three transcripts annotated in the ENSEMBL database (www.ensembl.org) were found to map to the +53 region. The three transcripts - A, B and C, are shown in Figure 7.12. All three transcripts are only two exons long; two of them are transcribed in one orientation and the third one in the opposite direction. All three, have their first exon beginning within the +53 genomic sequence.

iv) The ENSEMBL database shows a FirstEF promoter prediction (Davuluri et al. 2001) at the +52 region (data not shown).

Taken together, the above presented results suggest that the genomic region encompassing the +51 erythroid enhancer region seems to be associated with two distinct functions – that of an enhancer (+50/+51 region) and of a putative promoter (+53 region). More importantly, the histone code for this region predicts its enhancer and promoter activity, suggesting that the consensus histone code established in chapter 6 is not only consistent for the known regulatory regions across the SCL locus, but it could also be used to identify novel regulatory regions.

7.8.6 The putative insulator element at the +57 region

In K562 and Jurkat, the +57 region showed binding of CTCF in ChIP-chip, a protein known to bind to mammalian insulators (Bell et al. 1999). In K562 and Jurkat, the histone marks displayed by this region included hyper di- and trimethylation of H3 K4, hypo-acetylation of H4 K16 (Figure 7.12). The relative enrichments for di- and trimethylation of H3 K4 in Jurkat were higher as compared to K562 (Figure 7.12). The presence of these histone marks suggests that this region is possibly active in both the cell lines, and this is supported by CTCF binding. In Jurkat, significant enrichments were also seen for H3 K9/14 diacetylation (Chapter 5, section 5.6.1.1). In HL60 and HPB-ALL, the enrichments for di- and trimethylation of H3 K4 histone marks were not significant, although hypo-acetylation of H4 K16 was present. CTCF ChIP-chip was not performed in HL60 and HPB-ALL, therefore the activity of this region in those cell lines is not certain.

7.9 Discussion

The work described in this chapter involved a detailed characterization of a number of regulatory regions, known and novel, across the SCL locus in SCL expressing and non-expressing cell lines. These regions were characterized with respect to the proposed consensus histone code to test the consistency of the code in predicting the function and activity of these regulatory regions in SCL expressing and non-expressing cell lines. Based on the various ChIP-chip analyses performed in the previous chapters of this thesis, regulatory elements were also characterized for their function, role in transcriptional regulation of SCL and their associated DNA-protein interactions.

7.9.1 Identification and characterization of putative novel regulatory elements

The various ChIP-chip analyses performed for this study in SCL expressing and non-expressing cell lines in human and mouse identified a number of DNA-protein interactions across the SCL locus – the majority of which were known previously but some of which were novel. The identification of putative novel regulatory elements at the SCL locus further increases the complexity of regulatory elements which may control expression of SCL or its flanking genes. Table 7.1 summarizes all the known regulatory regions and the novel regions identified in K562 in the present study (data for the other human and mouse cell lines are in Tables 5.2, 5.3 and 5.4 in chapter 5). The binding interactions of transcription factors or other regulatory proteins, either known or identified in the present study, at various regulatory regions across the SCL locus are also listed in the table.

Known regulatory regions	Known *trans*-factor interactions	Novel regions	Novel *trans*-factor interactions	Potential function
KCY promoter			$Taf_{II}250$	
		KCY -4/-5	p300, CBP, HDAC2, $Taf_{II}250$	
SIL promoter, p^{SILa}			p300, CBP, HDAC2, $Taf_{II}250$	
		SIL promoter, p^{SILb}	CBP, HDAC2, $Taf_{II}250$	alternative SIL promoter
		SIL -1	Elf-1	
		SIL -10, +10, +34		
SCL p^{1a}, p^{1b}, +1, +3	GATA-1*		GATA-1*, p300, CBP, HDAC2, $Taf_{II}250$	
SCL p^{EXON4} at +7			GATA-1 at +7/+8 (downstream of p^{EXON4})	
SCL endothelial enhancer at -3/-4	Elf-1, Fli-1			
		SCL -23/-26, -12		SCL -12, repressor element
SCL DNase I HSs at -7, -9/-10			GATA-1, SCL -7	SCL -7, enhancer element
SCL stem cell enhancer at +20/+21	GATA-2, Elf-1, Fli-1			
SCL neural enhancer at +23/+24				
		SCL +39	CTCF	putative enhancer
MAP17 promoter			CTCF	
SCL +51 erythroid enhancer			GATA-1, SCL, Ldb-1, p300, CBP, HDAC2, $Taf_{II}250$ (not the same array element as the other TFs)	
		SCL +53		putative promoter
		SCL +57	CTCF	putative insulator element

Table 7.1: A summary of the known regulatory regions and novel putative regulatory regions at the SCL locus in K562. The list of novel regions and *trans*-factor binding interactions at the SCL locus, shown in

the 3rd and 4th column, has been compiled mostly based on the data generated in the present study in the human K562 cell line. The 5th column shows the proposed functions of a number of the regulatory region at the SCL locus. The data shown in the 1st and 2nd column has been compiled from previously published studies across the SCL loci in human and mouse. Please note: this list is by no means an exhaustive list of all the novel regions identified across the SCL locus. For other novel regions that were identified in the other human and mouse cell lines, see Tables 5.2, 5.3 and 5.4 and the text in chapter 5. The Elf-1 interaction, coloured pink, was identified in K562, Jurkat and HPB-ALL. The CTCF interaction, coloured yellow, was identified in the K562 and Jurkat cell lines. The GATA-1 interaction, marked with a red asterisk (*), was already known from previous studies and was confirmed in the present study (detailed discussions with respect to each regulatory region have been provided in the relevant sections in this chapter).

Notably, a region was identified upstream of the SIL gene at SIL -10, which was enriched for various histone modifications and was detected most notably in only SCL expressing cell lines. Whether this region represents a novel regulatory element which could be involved in the regulation of SCL needs to be further investigated. Elf-1 binding at the SIL -1 region in K562, Jurkat and HPB-ALL may suggest a role for this region in the regulation of the SIL and/or SCL gene which also requires further studies (also see discussion in chapter 5). Similarly, a number of other regions within the SIL gene, for example SIL +10, SIL +34 and SCL -23/-26 region, or the other regions identified in the mouse cell lines (see Table 5.4 in chapter 5) may represent additional regulatory elements associated with the regulation of SCL gene. These regions need further characterization to understand their function and role at the SCL locus.

Based on the analyses of ChIP-chip results for histone modifications and CTCF obtained at the +57 region in the SCL expressing cell lines K562 and Jurkat, this region may represent a putative insulator element. The role of the +57 region in insulator activity is also discussed in chapter 5; section 5.17.6.2). Whether it performs the dual-roles proposed for insulators (West et al. 2002), which include enhancer blocking activity and acting as a barrier to separate silent chromatin from active chromatin, is not known. CTCF is known for its enhancer-blocking functions at the HS4 insulator element at the chicken β-globin locus (Bell et al. 1999), where a separate protein, USF, is responsible for the barrier function (West et al. 2004). For the chicken β-globin locus, it was shown that the histone modification patterns at the HS4 insulator element (with the barrier activity) and the 3' HS element (with only enhancer-blocking activity) were different (Litt et al. 2001a; Litt et al. 2001b). The HS4 element was found to be hyper-acetylated for H3 K9/14 and hyper-methylated for H3 K4, whereas, these modifications were not seen at the 3' HS element. The data obtained in the present study showed higher enrichments for di- and trimethylation of H3 K4 at the +57 region in K562 and Jurkat. In Jurkat, significant enrichments were also seen for H3 K9/14 diacetylation. This may suggest that

the +57 region may indeed function as a barrier element (as well as enhancer blocking functions discussed in chapter 5, section 5.17.6.2).

In addition to the +57 region, two putative regulatory regions were characterised further in the present study. At the SCL gene, the -7 region identified in K562 showed significant levels of acetylation and GATA-1 binding and showed enhancer activity in reporter assays. The -12 region, showed H3 acetylation and repressor activity in reporter assays, but no transcription factors were found to bind to this region, although TEL1 or TEL2 make appealing candidates based on sequence conservation of their binding sites and their known role as repressors in haematopoiesis.

Another intriguing region at the SCL locus is the +51 erythroid enhancer and its adjacent regions at +50 and +53. The data obtained from various analyses in the present study suggest that, this region possibly has two separate functions – enhancer function at the +50/+51 region and a potential promoter function at the +52/+53 region. The erythroid enhancer at the +51 region (homologous to the mouse +40 region) which targets expression to primitive erythroblasts is already known (Delabesse et al. 2005). The proposed histone code at the SCL locus suggests that the +53 region represents a putative promoter, and this idea is further supported by (i) three annotated small transcripts mapping to this region, and (ii) a FirstEF promoter prediction at this region. These transcripts, however, have not been shown to have significant open reading frames which suggest they encode protein products. Recently published studies have reported that a huge number of transcripts in the human and mouse genomes – about half of total – do not appear to encode proteins (Kapronov et al. 2002; Bertone et al. 2004; Okazaki et al. 2002). It is being suggested that these non-coding RNAs (or ncRNAs) may provide an added layer of complexity to the regulation of complex genomes. Many of the ncRNAs appear to be developmentally regulated (Cawley et al. 2004). Functional analysis revealed the involvement of ncRNAs in pathways such as Hedgehog signalling and nuclear trafficking of transcription factor, NFAT (Willingham et al. 2005). It has also been suggested that several of the ncRNAs might also be antisense to known genes (Cawley et al. 2004) and antisense transcription plays important functions in eukaryotic cells (Yelin et al. 2003). Many of the ncRNAs are associated with nearby transcription factor binding sites (TFBS) (Cawley et al. 2004).

Based on the above mentioned studies, and the important role played by the ncRNAs in the regulation of protein-coding genes, suggests a potential role for the +53 region at the SCL locus. It is possible that the novel transcripts mapping to the +53 region represent ncRNAs which may have a functional role in the regulation of SCL itself or other genes.

The previously suggested association of ncRNAs with nearby transcription factor binding sites (TFBS) (Cawley et al. 2004) may mean that the transcription factors binding at the +51 erythroid enhancer may be involved in regulating the expression of these transcripts. However, all these hypotheses need to be further studied in detail. Thus, the +51/+53 region makes an interesting candidate region to be characterized further to understand its role at the SCL locus.

7.9.2 Transcriptional activation of SCL in Jurkat

Jurkat (a leukaemic T-cell line) exhibits inappropriate expression of SCL but the mechanism of its activation in this cell line is not known. Given the fact that no genomic rearrangements at the SCL locus have been found and SCL expression in Jurkat is mono-allelic (Leroy-Viard et al. 1994), this suggests that a *cis*-acting mechanism is involved in activating the single allele. The presence of significant acetylation at the -3/-4 endothelial enhancer (Gottgens et al. 2004) and at the -7 region in Jurkat suggests an open chromatin structure for these regions. This means that the increased positive charges due to acetylation weakens the histone-DNA contacts and promotes nucleosome mobility (Hansen 1998) which would make the underlying chromatin susceptible to nuclease digestion. Both these regions encompass known DNase I HSs in Jurkat (Leroy-Viard et al. 1994); the data obtained here was in complete agreement with these previous observations. The presence of histone marks, considered the hallmarks of active promoters, at SCL p^{1b} and p^{EXON4} were also consistent with the known activity of these two promoters in Jurkat (Bernard et al. 1992). Based on the data obtained, the -3/-4 region and the -7 region are likely candidates to be characterized further for their involvement in SCL activation in Jurkat as both these elements exhibit hallmarks of active regulatory regions. It is also possible that a genomic region more distant to the SCL gene (some other, yet, unknown region) is also involved in mediating the transcriptional activity of SCL, a mechanism by which Myc expression is known to be activated (Lazo et al. 1990). Interestingly, the SCL locus in Jurkat did not appear to be significantly enriched for H3 K9 modifications, acetylation as well as dimethylation. Whether this is specific cell-type preference for certain histone marks or represents a particular feature of the SCL locus in Jurkat is not known.

7.9.3 MAP17 expression in K562 and Jurkat

It has been previously reported that MAP17 is expressed, albeit at low levels, in all the cell types that express SCL, suggesting that the two genes share regulatory elements which coordinate their expression (Delabesse et al. 2005). Several lines of evidence from the present study suggest that this hypothesis is not true. The data obtained, for the

presence and/or absence of characteristic histone marks for active promoters suggest that MAP17 promoter may not be active in K562 or Jurkat. In addition, other data including the average levels of nucleosomes at the MAP17 gene in both these cell lines supports this idea. Another interesting observation in K562 is the binding interactions of CTCF at the MAP17 promoter region, which indicated that activity of the promoter of this gene may be blocked through interactions with an insulator or that CTCF binds directly to the promoter to control its activity. In Jurkat, CTCF interactions were not seen at the MAP17 promoter although both K562 and Jurkat express SCL as well as MAP17. This could be due to the reason that SCL is expressed inappropriately in Jurkat, and thus the regulation of SCL and therefore MAP17 may be different in K562 and Jurkat. The +39 region which is located upstream of the MAP17 gene, may also represent a putative novel regulatory element. The location of this element with respect to the MAP17 gene suggests that this element may be associated with the regulation of this gene. However, whether this element is associated with MAP17 or SCL, whether SCL and MAP17 expressions are linked or the mechanism of MAP17 regulation in haematopoietic cells is not known and requires further investigation.

7.9.4 Conclusions

The detailed characterization of regulatory regions with respect to the proposed consensus histone code suggests that the code is consistent at the regulatory regions across the SCL locus in SCL expressing and non-expressing cell lines. It was possible to infer the function and activity of novel regions using the code which was also confirmed by *in vitro* assays in some cases. The data presented suggests that the complexity of regulatory elements and interactions at the SCL locus represents an excellent model by which to study transcriptional regulation and identify features which may be applicable to mammalian genomes in general.

Chapter 8

Summary and Future work

8.1. Summary of the work presented in this thesis

Understanding the events which occur as stem cells differentiate into committed cell lineages is a fundamental issue in cell biology. It has been shown that the SCL transcription factor (TF), also known as TAL1, is central to the mechanisms whereby pluripotent stem cells differentiate into haematopoietic stem cells (HSCs) that ultimately give rise to the various blood lineages. While this process is thought to be tightly regulated at the level of gene expression, the exact ways in which SCL helps direct this process is not well understood. Therefore, the aim of the work presented in this thesis was to perform large scale profiling of a range of *in vivo* DNA-protein interactions involved in transcriptional regulation across the SCL locus. This was achieved by developing a highly sensitive, robust and reproducible array platform in combination with a diverse range of ChIP assays. The results and interpretations of this work which are presented in chapters 3, 4, 5, 6, and 7 are summarised below. Future work which follows on from the work presented in this thesis is also outlined in subsequent sections.

8.1.1. Construction and validation of the SCL genomic tiling path arrays (Chapter 3)

In chapter 3, the construction of the SCL genomic tiling path arrays was presented in detail. These arrays, fabricated for both the human and mouse SCL loci, were constructed using a unique amino-link chemistry which allowed single-strands of DNA derived from double-stranded PCR products to be retained on the surface of the microarray slide (Dhami et al. 2005). Array validation experiments were performed to ensure that each PCR product of the autosomal SCL region on the array reported accurate genomic copy number change. The array also included a set of PCR products from chromosome X (Dhami et al. 2005) to act as controls for these validation experiments. A series of male versus female hybridizations were performed and it was found that the SCL array elements reliably reported quantitative measures of genomic copy-number. Such quantitative measurements would greatly facilitate the detection of a broad range of DNA-protein interactions at the SCL locus.

8.1.2. Characterization of human haematopoietic cell lines (Chapter 3)

To perform ChIP-chip assays, SCL expressing and non-expressing haematopoietic cell lines were chosen which included human cell lines K562, Jurkat, HL60, and HPB-ALL. In addition, mouse haematopoietic cell line 416B (SCL expressing) and mouse E14 ES cell

line (SCL non-expressing) were also used in some ChIP-chip assays. The human haematopoietic cell lines were characterized for SCL expression, genomic rearrangements, and structural rearrangements within the SCL locus by performing real-time PCR, array CGH at low and high resolutions and fluorescence *in situ* hybridisations (FISH) respectively. Genomic rearrangements in these cell lines were numerous. However, the SCL locus was not disrupted across the region studied in this thesis, although the entire genomic region was contained in some cell line amplifications. This data provided evidence to suggest that the *cis*-regulatory events identified across the SCL locus in the cell lines analysed here would not be due to alterations of the locus itself, although structural rearrangements distant from SCL could not be ruled out as affecting SCL regulation.

8.1.3. Validation of the SCL arrays for ChIP-chip assays (Chapter 4)

As a means of validating the SCL arrays for ChIP-chip analysis, ChIP-chip assays were developed and optimized in order to detect and characterize known regulatory sequences and *in vivo* DNA-protein interactions across the SCL locus in human and mouse. Typically, ChIP DNA material obtained from 10 million cultured cells (this cell number was optimized) was sufficient to perform multiple hybridizations. An important feature of the SCL array platform was to be able to use unamplified ChIP DNA in all of the assays, thereby, reducing sources of non-biological biases resulting from amplification. The ability of the array platform to allow use of unamplified material was attributed to high signal to background ratios obtained owing to the single-stranded approach used in the construction of the array.

Analysis of H3 K9/14 diacetylation and GATA-1 binding allowed the detection of virtually all known regulatory sequences and DNA-protein interactions involved in SCL expression. To assess the reproducibility of the array-based system, multiple biological and technical replicates were performed for each assay in order to assess the performance of each array element. It was observed that using this array-based system, it was possible to reproducibly identify and quantify a wide range of both high- and low-level ChIP enrichments which were independently verified by real-time PCR. Further analysis of levels of histone H3 K9/14 diacetylation at the SCL locus and sequence conservation at non-coding sequences revealed a strong positive correlation - a relationship which had not been observed previously in mammals. Analysis of GATA-1 interactions led to the identification of GATA-1 binding at the +51 erythroid enhancer and other SCL regulatory regions. Conserved sequence analysis revealed conserved GATA sites in each of the regions which reported significant ChIP enrichments for GATA-1 which indicated that the identified interactions of GATA-1 at the SCL locus were real.

Furthermore, by analysing the levels of histone H3 at the SCL locus in K562, it was found that nucleosome levels were variable across the entire locus. Nucleosome depletion occurred at coding and regulatory regions of SCL and its neighbouring genes in the K562 cell line in which they were expressed. These findings were novel for the SCL locus and had not been seen before in mammalian genomes.

Taken together these results provided evidence that the array-based system could reliably detect meaningful *in vivo* DNA-protein interactions that reflected a range of biological activities underlying the regulation of SCL.

8.1.4. Applications and analysis of ChIP-chip assays across the SCL locus (Chapters 5, 6 & 7)

In chapter 5, the work described involved further application of the ChIP-chip assays to carry out a large scale survey of a wide range of *in vivo* DNA-protein interactions at the SCL locus in SCL expressing and non-expressing cell lines in human and mouse. Experiments were performed to map distribution of a broad range of histone modifications, transcription factor binding and interactions of other regulatory proteins at the SCL locus.

The specific modifications of H3 acetylation i.e. at lysines 9, 14, 18 and 27 showed significant enrichments at almost all regulatory regions which were found to be associated with H3 K9/14 diacetylation. The patterns of specific modifications of H4 at lysines 5, 8, and 16 were more variable and H4 K8 and K16 were found to be hypoacetylated at active regulatory regions. Patterns of methylation of H3 at lysine 4 correlated highly with the transcriptional activity of genes at the SCL locus. Di- and trimethylation were highly enriched at the 5' ends of active genes whereas decreased levels of monomethylation were prominent over active promoters. Trimethylation of H3 at lysine 79 was found to be associated with active genes and was located immediately downstream of active promoters. *In vivo* transcription factor interactions for GATA-1, SCL, Ldb-1 and other regulatory proteins were identified at the +51 erythroid enhancer which provided the first evidence that SCL may be self-regulated by a positive feedback loop. CTCF was shown to bind to a putative insulator element at the SCL +57 region.

In chapter 6, the relationships of histone modifications with respect to sequence conservation and with each other was analysed across the SCL locus. It was established that the relative levels of tri- to monomethylation could discriminate between active and inactive promoters. In addition, as seen with H3 acetylation, strong correlations were found between histone H3 K4 methylation levels and sequence conservation at non-coding regions. Furthermore, hierarchical clustering demonstrated that histone

modification data could be used to group DNA sequences according to their regulatory function. Analysis of the ChIP-chip data in this way led to the identification of seven key histone modifications which could distinguish the type and activity of regulatory regions. These key modifications formed a consensus histone code which was found to be consistent on all the regulatory regions in all the human cell lines analysed in this thesis (Chapter 7). Analysis of histone H3 levels in all the human cell lines revealed that decreased levels of nucleosomes were observed at active regulatory regions and coding regions of genes which were expressed in these cell lines.

In chapter 7, a detailed analysis of each of the known regulatory regions across the the SCL locus was described with respect to data generated in this thesis, as well as data obtained from other sources. In particular, evidence was provided that the Jurkat cell line showed histone modifications at the SCL -3/-4 and -7 regions which could potentially be involved in the inappropriate activation of SCL in this cell line. A number of putative novel regulatory regions which were identified in chapters 4, 5, and 6 were further characterized. These included, among others, the SCL -12 region which showed repressor activity in K562, the SCL -7 region which showed an enhancer activity in K562, and the SCL +57 region which was found to bind CTCF (an insulator-binding protein), thereby identifying this sequence as an insulator element. Other regions were also identified which could be considered to be good candidates as regulatory sequences requiring further investigation.

8.2. Future work

The work described in this thesis has shown how *in vivo* DNA-protein interactions at the SCL locus can be assayed using a highly sensitive and robust array-based method. There is tremendous scope to use this resource to further detect the full complement of regulatory interactions at the SCL locus using both ChIP-chip assays and other methods. In addition, this array-based system represents an ideal resource to examine the SCL locus for other epigenetic features as well as in diagnostic studies.

8.2.1. Diagnostic studies

It is known that most patients with T-cell acute lymphoblastic leukaemia (T-ALL) show SCL expression in the leukaemic cells (Bash et al. 1995). Thirty percent of the patients show a genomic rearrangement affecting the SCL locus, which leads to the inappropriate expression of SCL through the involvement of regulatory elements of other genes (Carroll et al. 1990; Brown et al. 1990; Bash et al. 1993; Breit et al. 1993). Thus, SCL

rearrangement is one of the most common genetic defects known to be associated with T-ALL.

The SCL genomic tiling path array constructed as part of this study has proved to be highly quantitative and can accurately report copy-number changes (Chapter 3). It could therefore be used diagnostically to perform array CGH experiments with DNA from T-ALL patients to detect DNA rearrangements at the SCL locus. Because of the high resolution of the array, even small deletions or amplifications could be detected, since this array technology has been shown to detect deletions/amplifications at high resolution in patients with a range of constitutional genomic rearrangements (Dhami et al., 2005). Furthermore, such studies could also be extended to include ChIP-chip assays of cell lines from patients as well, thus providing further evidence that these re-arrangements alter the normal regulatory features of the SCL locus in T-ALL individuals. For example, this thesis reports the identification of regions of histone modifications found in the T-ALL cell line Jurkat which could be involved in the *cis*-activation of SCL (Chapter 7).

8.2.2. Further profiling of DNA-protein interactions at the SCL locus

The work shown in this thesis involved comprehensive mapping of histone modifications to analyse their distribution at the SCL locus and to explore their inter-relationships. However, while the work presented here provides one of the most comprehensive analyses of DNA-protein interactions for any region of a mammalian genome, the analysis is by no means exhaustive for the types of histone modifications which are involved in gene regulation, and there are still a large number of modifications and histone variants which were not assayed for the present project. In a few instances, antibodies for some of these modifications were tested in ChIP-chip assays, but they did not provide significant enrichments (see chapter 5, Table 5.1). Examples of additional modifications which could be studied include modifications of arginine residues of histone H3 and H4, mono- and trimethylation of H3 lysine 9, specific modifications of histones H2A and H2B and also specific modifications within the globular domain of the nucleosomes. The function of these modifications at the SCL locus or their relationship with other histone marks or with the underlying chromatin is still unexplored. Of course, the ability to carry out further analysis of histone modifications by ChIP-chip depends on the availability of antibodies which will yield significant enrichments in ChIP-chip studies. However, given that new antibodies raised to these modifications are being generated by commercial companies at an ever-increasing pace, the availability of antibodies should not limit such studies across the SCL locus. Presently, a further set of antibodies

for as many as 20 additional histone modifications are being tested at the Sanger Institute to facilitate this work (D. Vetrie, A. Bruce, work in progress).

8.2.3. Testing the histone code

Analysis of the array data for various histone modifications led to the identification of a set of seven key histone modifications which were distinctly associated with different kinds of regulatory elements and their activity. The consensus code for the histone marks was consistent for all the regulatory regions associated with the genes contained on the SCL array, in all the human cell lines analyzed. One issue associated with this histone code hypothesis is whether or not the code is consistent for all genes in the human genome (see also section 8.2.5). Given that analysis of the human genome for these modifications would be a large undertaking, additional studies at the SCL array could be performed on the 5 genes represented on the SCL array to establish further principles which could then be applied to larger regions of the genome. These studies would include testing whether the code is consistent for all of these 5 genes when they are either expressed or not expressed. Therefore, ChIP-chip experiments for the seven key histone modifications (and other modifications as well) could be performed in cell lines in which the MAP17 gene is expressed but not the SCL or CYP genes - for example HEK 293 cell line (a kidney cell line). Similarly, ChIP-chip experiments performed in HepG2, T47-D and MCF-7 cell lines would provide the patterns of histone marks at the regulatory regions of the cytochrome P450 genes but not SCL. Alternatively, developing differentiation systems which follow the expression of these genes through time would also yield information on how the complexity of histone modifications change for a gene and its regulatory elements in situations where (i) the gene is expressed and then turned off, or (ii) the gene is silent and then becomes expressed later in differentiating cell types. For the study of SCL, the differentiation of CD34+ cells provide an excellent experimental model to study the events associated with the silencing of SCL in differentiating blood cells such as erythrocytes or megakaryocytes. In summary, the use of a variety of cell lines and differentiating cell systems for which different combinations of the genes at the SCL locus are expressed, would provide additional evidence that the code is consistent in a variety of biological contexts.

8.2.4. Further characterization of SCL regulatory regions: The +51 erythroid enhancer and +53 promoter

The work presented in this thesis provides data which suggest the identification of novel regulatory sequences as well as the further characterization of known ones. The +51

erythroid enhancer, recently identified as playing a role in erythrocyte differentiation (Delabesse et al., 2005), was further characterized with respect to binding of several transcription factors (Chapters 4, 5). In order to confirm the dependency of *in vivo* interactions of factors such as GATA-1, SCL and Ldb-1 and other TFs at the +51 erythroid enhancer, an approach using RNAi (RNA interference) could be used. In order to do this, siRNAs could be designed against GATA-1, for example, to target its mRNA. Transfecting siRNAs targeted against GATA-1 in K562 cells would result in degradation of the GATA-1 mRNA and knockdown its protein levels. This manipulation could be followed by performing time-course ChIP-chip experiments against other TFs using the transfected cells. If the interaction of a TF at the +51 region is dependent on binding of GATA-1, the absence of GATA-1 may prevent the other TFs from binding. Therefore these siRNA–ChIP-chip experiments would help deduce whether other TFs are able to bind to this region in the absence of GATA-1. Similarly siRNA ChIP-chip experiments could be performed against the other TFs known to bind to the region in order to deduce the temporal relationships between these TFs in binding to the +51 region.

The results from the present study suggested that the SCL erythroid complex might assemble at the erythroid enhancer. However, the ChIP-chip assays for E47 and LMO2, both of which are members of this complex, did not provide significant enrichments in this study. Therefore, new antibodies could be sourced for these proteins and ChIP-chip assays repeated to see if these proteins are also part of the multi-protein complex regulating the erythroid enhancer. Such interactions could also be verified using the siRNA approaches as discussed above.

Another interesting finding at the +51 region was the identification of second regulatory region, most likely a promoter, at the +53 region (Chapter 7). This is further supported by the presence of three annotated novel transcripts being transcribed from this region which may represent non-coding RNAs. The function of non-coding RNAs in the human genome is not yet clear and further experiments to investigate the exact location of the promoter element, its relationship with the enhancer activity of the +51 region, and these novel transcripts might lead to some interesting findings which may or may not be related to SCL regulation. Furthermore, it is not clear whether the binding of SCL and other transcription factors at the +51 region regulates the transcription of these non-coding RNAs. Therefore, siRNA experiments against the relevant transcription factors could also determine whether they have an effect on the expression of these novel transcripts.

8.2.5. ChIP-chip studies: large scale analysis of mammalian genomes

One of the major advantages of array-based systems is that they are completely scalable to study whole genomes. Any genomic regions of interest can be amplified as PCR products and included on arrays to analyse larger fractions of the human and mouse genomes for various regulatory features. A very good example of this is provided by the human ENCODE project (the **Enc**yclopaedia **o**f **D**NA **E**lements) which has been undertaken to identify all functional elements in the human genome sequence. The pilot phase of this project is to test and compare existing methods to identify functional elements and analyze a defined portion which is approximately 1 percent or 30 Mb of the human genome sequence (ENCODE consortium, 2004). The project being undertaken at the Sanger Institute for ENCODE utilizes a similar approach that was used for the SCL tiling arrays to examine DNA-protein interactions, to assess replication timing, to map replication origins and to assess DNA methylation across the specified genomic regions. Similarly, a large scale EU-funded programme has been established to identify functional elements in the mouse genome (the HEROIC programme).

8.2.6. Analysis of other biological features at the SCL locus

8.2.6.1 Replication Timing

Replication of the human genome is reported to be a temporally ordered process which has been linked to chromatin structure, nuclear position and gene expression. Microarray-based assays to observe patterns of replication timing have been used in yeast, *Drosophila* and human (Raguraman et al. 2001; Schubeler et a. 2002; Woodfine et al. 2004). In brief, the assay involves flow-sorting to separate unsynchronized cell nuclei into fractions in S-phase and G1-phase. The extracted DNA is then differentially labeled and hybridized simultaneously to the genomic array to obtain S-phase:G1-phase ratios. By quantitatively measuring genomic copy-number of the represented array elements, the sequences with ratios close to 2:1 represent loci which replicate early in S-phase; conversely, loci with ratios close to 1:1 represent late replicating sequences (Woodfine et al. 2004). This method in combination with the single nucleotide primer extension (SNuPE) assay has also been used to assess replication timing of two differentially replicating alleles (Xiong et al. 1998). The high resolution SCL genomic tiling array and its ability to report quantitatively in a small dynamic range (i.e., for copy number changes in array CGH and to profile nucleosome density using ChIP-chip) would prove to be an excellent platform to carry out replication timing studies at the SCL locus.

8.2.6.2 Mapping replication origins

The initiation of DNA replication depends on the formation of pre-replication complexes and the proteins that are required for establishing these complexes include the origin complex (ORC) proteins (Mendez et al. 2000). Human replication origins and replicons have not been characterized in great detail at the sequence level. Replication origins can be identified in two ways: 1) using ChIP-chip by using antibodies raised to ORC proteins 2) hybridizing short, nascent DNAs representing the DNA sequence surrounding an active origin of replication to genomic arrays. For ChIP-chip studies, antibodies exist for a number of the ORC proteins found in ORC complexes. However, given that these complexes may only be formed during certain periods of the cell cycle (for example, S phase), it may be necessary to synchronize cell populations in order to efficiently identify ORC-DNA interactions using ChIP-chip. For the nascent DNA studies, the cycling cells are pulse labeled with BrDU (bromo deoxy-uridine) for 10-30 minutes; intact nuclei are isolated and the DNA denatured in an alkaline buffer. The resulting solution of denatured nuclear DNA is then applied to an alkaline sucrose gradient and the DNA is separated according to size by centrifugation. Fractions from the gradient are then collected and sized using alkaline gel electrophoresis. Fractions containing short strands (up to 20 kb) are pooled, neutralized and the strands which have incorporated BrDU (nascent strands) are isolated by immunoprecipitation with an anti-BrDU antibody. The DNA is extracted, fluorescently-labelled and hybridized onto an array along with a control DNA which has not been selected on the basis of its size. The origins of replication can be defined by the sequences hybridizing with the nascent DNA fraction. This data, when combined with that of ChIP-chip for ORCs and other assays for markers of gene expression, could provide important insights into the relationship between expression and replication during the cell cycle.

8.2.6.3 Mapping DNase I hypersensitive sites (HSs)

The sensitivity of transcriptional regulatory regions to digestion with DNase I is a classical method that is employed to identify and annotate functional elements (see chapter 1). Developing an array-based system to identify DNase I hypersensitive sites (HSs) in a high throughput manner, as compared to its traditional low-resolution, low-throughput approach, is an attractive proposition which could be validated using the SCL array system as a number of DNase I hypersensitive sites have already been mapped and characterized across the SCL locus (Fordham et al., 1999; Gottgens et al., 1997; Leroy-Viard et al., 1994) Such a method has been developed using the human and mouse SCL array as a test system to map DNase I HSSs across the SCL locus (Follows et al. manuscript submitted and under review). The approach entails isolating

the nuclei, digesting the chromatin with DNase I, extracting DNA, and generating DNA templates by ligation-mediated primer extension reactions. The biotinylated primer extension products are captured with streptavidin beads, labeled and hybridized to the array. Using this approach, all known SCL regulatory regions have been identified (Follows et al., data not shown). Furthermore, this assay could also be used for larger regions of the genome - for example across the ENCODE regions. Given that DNAse I hypersensitivity offers an alternative approach to identifying regulatory regions across the genome, it will be important to compare how this data compared with that of ChIP-chip assays such as histone modifications. However, one advantage of ChIP-chip assays is that they also provide functional information as to regulatory proteins and features associated with the regulatory elements. HSs analysis, on the otherhand, provides only location information.

8.3. Final thoughts: SCL as a model of regulation for mammalian genomes

The above applications, illustrate how the combination of a highly sensitive, high resolution genomic tiling array in combination with biological assays, could be used to further elaborate our knowledge of the SCL locus and understand the full complement of biological events associated with the activity of this important regulator of blood development. Given the data and future studies presented here, SCL will undoubtedly become one of the most well-characterized regions in the human and mouse genomes with respect to the annotation of regulatory information. Thus, it will become an excellent experimental model from which researchers will gain a greater understanding of the features and general principles of mammalian gene regulation. To this end, the development of the ChIP-chip technology described here, will make an important contribution to this area of research.

References

1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**(5396): 2012-2018.

2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**(7011): 931-945.

Adams, M.D. Celniker, S.E. Holt, R.A. Evans, et al. 2000. The genome sequence of Drosophila melanogaster. *Science* **287**(5461): 2185-2195.

Agalioti, T., Chen, G., and Thanos, D. 2002. Deciphering the transcriptional histone acetylation code for a human gene. *Cell* **111**(3): 381-392.

Agalioti, T., Lomvardas, S., Parekh, B., Yie, J., Maniatis, T., and Thanos, D. 2000. Ordered recruitment of chromatin modifying and general transcription factors to the IFN-beta promoter. *Cell* **103**(4): 667-678.

Ahmad, K. and Henikoff, S. 2001. Centromeres are specialized replication domains in heterochromatin. *J Cell Biol* **153**(1): 101-110.

Ahmad, K. and Henikoff, S. 2002. The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. *Mol Cell* **9**(6): 1191-1200.

Alam, J. and Cook, J.L. 1990. Reporter genes: application to the study of mammalian gene transcription. *Anal Biochem* **188**(2): 245-254.

Albertson, D.G. and Pinkel, D. 2003. Genomic microarrays in human genetic disease and cancer. *Hum Mol Genet* **12 Spec No 2**: R145-152.

Albertson, D.G., Ylstra, B., Segraves, R., Collins, C., Dairkee, S.H., Kowbel, D., Kuo, W.L., Gray, J.W., and Pinkel, D. 2000. Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nat Genet* **25**(2): 144-146.

Allfrey, V.G., Faulkner, R., and Mirsky, A.E. 1964. Acetylation And Methylation Of Histones And Their Possible Role In The Regulation Of Rna Synthesis. *Proc Natl Acad Sci U S A* **51**: 786-794.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J Mol Biol* **215**(3): 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**(17): 3389-3402.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N. et al. 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297**(5585): 1301-1310.

Aplan, P.D., Begley, C.G., Bertness, V., Nussmeier, M., Ezquerra, A., Coligan, J., and Kirsch, I.R. 1990a. The SCL gene is formed from a transcriptionally complex locus. *Mol Cell Biol* **10**(12): 6426-6435.

Aplan, P.D., Jones, C.A., Chervinsky, D.S., Zhao, X., Ellsworth, M., Wu, C., McGuire, E.A., and Gross, K.W. 1997. An scl gene product lacking the transactivation domain induces bony abnormalities and cooperates with LMO1 to generate T-cell malignancies in transgenic mice. *Embo J* **16**(9): 2408-2419.

Aplan, P.D., Lombardi, D.P., Ginsberg, A.M., Cossman, J., Bertness, V.L., and Kirsch, I.R. 1990b. Disruption of the human SCL locus by "illegitimate" V-(D)-J recombinase activity. *Science* **250**(4986): 1426-1429.

Aplan, P.D., Nakahara, K., Orkin, S.H., and Kirsch, I.R. 1992a. The SCL gene product: a positive regulator of erythroid differentiation. *Embo J* **11**(11): 4073-4081.

Aplan, P.D., Raimondi, S.C., and Kirsch, I.R. 1992b. Disruption of the SCL gene by a t(1;3) translocation in a patient with T cell acute lymphoblastic leukemia. *J Exp Med* **176**(5): 1303-1310.

Bagheri-Fam, S., Ferraz, C., Demaille, J., Scherer, G., and Pfeifer, D. 2001. Comparative genomics of the SOX9 region in human and Fugu rubripes: conservation of short regulatory sequence elements within large intergenic regions. *Genomics* **78**(1-2): 73-82.

Banerji, J., Olson, L., and Schaffner, W. 1983. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33**(3): 729-740.

Banerji, J., Rusconi, S., and Schaffner, W. 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**(2 Pt 1): 299-308.

Bannister, A.J. and Kouzarides, T. 1996. The CBP co-activator is a histone acetyltransferase. *Nature* **384**(6610): 641-643.

Bannister, A.J., Schneider, R., and Kouzarides, T. 2002. Histone methylation: dynamic or static? *Cell* **109**(7): 801-806.

Barton, L.M., Gottgens, B., and Green, A.R. 1999. The stem cell leukaemia (SCL) gene: a critical regulator of haemopoietic and vascular development. *Int J Biochem Cell Biol* **31**(10): 1193-1207.

Bash, R.O., Crist, W.M., Shuster, J.J., Link, M.P., Amylon, M., Pullen, J., Carroll, A.J., Buchanan, G.R., Smith, R.G., and Baer, R. 1993. Clinical features and outcome of T-cell acute lymphoblastic leukemia in childhood with respect to alterations at the TAL1 locus: a Pediatric Oncology Group study. *Blood* **81**(8): 2110-2117.

Bash, R.O., Hall, S., Timmons, C.F., Crist, W.M., Amylon, M., Smith, R.G., and Baer, R. 1995. Does activation of the TAL1 gene occur in a majority of patients with T-cell acute lymphoblastic leukemia? A pediatric oncology group study. *Blood* **86**(2): 666-676.

Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B., and Lander, E.S. 2000. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res* **10**(7): 950-958.

Begley, C.G., Aplan, P.D., Davey, M.P., Nakahara, K., Tchorz, K., Kurtzberg, J., Hershfield, M.S., Haynes, B.F., Cohen, D.I., Waldmann, T.A., and et al. 1989a. Chromosomal translocation in a human leukemic stem-cell line disrupts the T-cell antigen receptor delta-chain diversity region and results in a previously unreported fusion transcript. *Proc Natl Acad Sci U S A* **86**(6): 2031-2035.

Begley, C.G., Aplan, P.D., Denning, S.M., Haynes, B.F., Waldmann, T.A., and Kirsch, I.R. 1989b. The gene SCL is expressed during early hematopoiesis and encodes a differentiation-related DNA-binding motif. *Proc Natl Acad Sci U S A* **86**(24): 10128-10132.

Begley, C.G. and Green, A.R. 1999. The SCL gene: from case report to critical hematopoietic regulator. *Blood* **93**(9): 2760-2770.

Begley, C.G., Robb, L., Rockman, S., Visvader, J., Bockamp, E.O., Chan, Y.S., and Green, A.R. 1994. Structure of the gene encoding the murine SCL protein. *Gene* **138**(1-2): 93-99.

Begley, C.G., Visvader, J., Green, A.R., Aplan, P.D., Metcalf, D., Kirsch, I.R., and Gough, N.M. 1991. Molecular cloning and chromosomal localization of the murine homolog of the human helix-loop-helix gene SCL. *Proc Natl Acad Sci U S A* **88**(3): 869-873.

Bell, A.C. and Felsenfeld, G. 2000. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* **405**(6785): 482-485.

Bell, A.C., West, A.G., and Felsenfeld, G. 1999. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* **98**(3): 387-396.

Berger, S.L. 2002. Histone modifications in transcriptional regulation. *Curr Opin Genet Dev* **12**(2): 142-148.

Bernard, O., Azogui, O., Lecointe, N., Mugneret, F., Berger, R., Larsen, C.J., and Mathieu-Mahul, D. 1992. A third tal-1 promoter is specifically used in human T cell leukemias. *J Exp Med* **176**(4): 919-925.

Bernard, O., Guglielmi, P., Jonveaux, P., Cherif, D., Gisselbrecht, S., Mauchauffe, M., Berger, R., Larsen, C.J., and Mathieu-Mahul, D. 1990. Two distinct mechanisms for the SCL gene

activation in the t(1;14) translocation of T-cell leukemias. *Genes Chromosomes Cancer* **1**(3): 194-208.

Bernard, O., Lecointe, N., Jonveaux, P., Souyri, M., Mauchauffe, M., Berger, R., Larsen, C.J., and Mathieu-Mahul, D. 1991. Two site-specific deletions and t(1;14) translocation restricted to human T-cell acute leukemias disrupt the 5' part of the tal-1 gene. *Oncogene* **6**(8): 1477-1488.

Bernstein, B.E., Humphrey, E.L., Erlich, R.L., Schneider, R., Bouman, P., Liu, J.S., Kouzarides, T., and Schreiber, S.L. 2002. Methylation of histone H3 Lys 4 in coding regions of active genes. *Proc Natl Acad Sci U S A* **99**(13): 8695-8700.

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas, E.J., 3rd, Gingeras, T.R., Schreiber, S.L., and Lander, E.S. 2005. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120**(2): 169-181.

Bernstein, B.E., Liu, C.L., Humphrey, E.L., Perlstein, E.O., and Schreiber, S.L. 2004. Global nucleosome occupancy in yeast. *Genome Biol* **5**(9): R62.

Birney, E. and Durbin, R. 1997. Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. *Proc Int Conf Intell Syst Mol Biol* **5**: 56-64.

Birney, E. and Durbin, R. 2000. Using GeneWise in the Drosophila annotation experiment. *Genome Res* **10**(4): 547-548.

Blanco, J.C., Minucci, S., Lu, J., Yang, X.J., Walker, K.K., Chen, H., Evans, R.M., Nakatani, Y., and Ozato, K. 1998. The histone acetylase PCAF is a nuclear receptor coactivator. *Genes Dev* **12**(11): 1638-1651.

Blattner, F.R., Plunkett, G., 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., and Shao, Y. 1997. The complete genome sequence of Escherichia coli K-12. *Science* **277**(5331): 1453-1474.

Blobel, G.A. 2000. CREB-binding protein and p300: molecular integrators of hematopoietic transcription. *Blood* **95**(3): 745-755.

Bockamp, E.O., Fordham, J.L., Gottgens, B., Murrell, A.M., Sanchez, M.J., and Green, A.R. 1998. Transcriptional regulation of the stem cell leukemia gene by PU.1 and Elf-1. *J Biol Chem* **273**(44): 29032-29042.

Bockamp, E.O., McLaughlin, F., Gottgens, B., Murrell, A.M., Elefanty, A.G., and Green, A.R. 1997. Distinct mechanisms direct SCL/tal-1 expression in erythroid cells and CD34 positive primitive myeloid cells. *J Biol Chem* **272**(13): 8781-8790.

Bockamp, E.O., McLaughlin, F., Murrell, A.M., Gottgens, B., Robb, L., Begley, C.G., and Green, A.R. 1995. Lineage-restricted regulation of the murine SCL/TAL-1 promoter. *Blood* **86**(4): 1502-1514.

Bode, J., Stengert-Iber, M., Kay, V., Schlake, T., and Dietz-Pfeilstetter, A. 1996. Scaffold/matrix-attached regions: topological switches with multiple regulatory functions. *Crit Rev Eukaryot Gene Expr* **6**(2-3): 115-138.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., and Rubin, E.M. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**(5611): 1391-1394.

Bonifer, C., Faust, N., Geiger, H., and Muller, A.M. 1998. Developmental changes in the differentiation capacity of haematopoietic stem cells. *Immunol Today* **19**(5): 236-241.

Breit, T.M., Mol, E.J., Wolvers-Tettero, I.L., Ludwig, W.D., van Wering, E.R., and van Dongen, J.J. 1993. Site-specific deletions involving the tal-1 and sil genes are restricted to cells of the T cell receptor alpha/beta lineage: T cell receptor delta gene deletion mechanism affects multiple genes. *J Exp Med* **177**(4): 965-977.

Brent, R. and Ptashne, M. 1985. A eukaryotic transcriptional activator bearing the DNA specificity of a prokaryotic repressor. *Cell* **43**(3 Pt 2): 729-736.

Brown, L., Cheng, J.T., Chen, Q., Siciliano, M.J., Crist, W., Buchanan, G., and Baer, R. 1990. Site-specific recombination of the tal-1 gene is a common occurrence in human T cell leukemia. *Embo J* **9**(10): 3343-3351.

Bruce, A.W., Donaldson, I.J., Wood, I.C., Yerbury, S.A., Sadowski, M.I., Chapman, M., Gottgens, B., and Buckley, N.J. 2004. Genome-wide analysis of repressor element 1 silencing transcription factor/neuron-restrictive silencing factor (REST/NRSF) target genes. *Proc Natl Acad Sci U S A* **101**(28): 10458-10463.

Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Green, E.D., Sidow, A., and Batzoglou, S. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* **13**(4): 721-731.

Bulger, M. and Groudine, M. 1999. Looping versus linking: toward a model for long-distance gene activation. *Genes Dev* **13**(19): 2465-2477.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**(1): 78-94.

Burke, L.J. and Baniahmad, A. 2000. Co-repressors 2000. *Faseb J* **14**(13): 1876-1888.

Burley, S.K. and Roeder, R.G. 1996. Biochemistry and structural biology of transcription factor IID (TFIID). *Annu Rev Biochem* **65**: 769-799.

Cao, R., Wang, L., Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R.S., and Zhang, Y. 2002. Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science* **298**(5595): 1039-1043.

Carey, J. 1991. Gel retardation. *Methods Enzymol* **208**: 103-117.

Carlone, D.L. and Skalnik, D.G. 2001. CpG binding protein is crucial for early embryonic development. *Mol Cell Biol* **21**(22): 7601-7606.

Carter, N.P. and Vetrie, D. 2004. Applications of genomic microarrays to explore human chromosome structure and function. *Hum Mol Genet* **13 Spec No 2**: R297-302.

Cawley, S., Bekiranov, S., Ng, H.H., Kapranov, P., Sekinger, E.A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A.J., Wheeler, R., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K., and Gingeras, T.R. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**(4): 499-509.

Chakrabarti, S.R. and Nucifora, G. 1999. The leukemia-associated gene TEL encodes a transcription repressor which associates with SMRT and mSin3A. *Biochem Biophys Res Commun* **264**(3): 871-877.

Chapman, M.A., Donaldson, I.J., Gilbert, J., Grafham, D., Rogers, J., Green, A.R., and Gottgens, B. 2004. Analysis of multiple genomic sequence alignments: a web resource, online tools, and lessons learned from analysis of mammalian SCL loci. *Genome Res* **14**(2): 313-318.

Chen, D., Ma, H., Hong, H., Koh, S.S., Huang, S.M., Schurter, B.T., Aswad, D.W., and Stallcup, M.R. 1999. Regulation of transcription by a protein methyltransferase. *Science* **284**(5423): 2174-2177.

Chen, Q., Cheng, J.T., Tasi, L.H., Schneider, N., Buchanan, G., Carroll, A., Crist, W., Ozanne, B., Siciliano, M.J., and Baer, R. 1990. The tal gene undergoes chromosome translocation in T cell leukemia and potentially encodes a helix-loop-helix protein. *Embo J* **9**(2): 415-424.

Cheng, J.T., Hsu, H.L., Hwang, L.Y., and Baer, R. 1993. Products of the TAL1 oncogene: basic helix-loop-helix proteins phosphorylated at serine residues. *Oncogene* **8**(3): 677-683.

Cheng, T., Shen, H., Giokas, D., Gere, J., Tenen, D.G., and Scadden, D.T. 1996. Temporal mapping of gene expression levels during the differentiation of individual primary hematopoietic cells. *Proc Natl Acad Sci U S A* **93**(23): 13158-13163.

Cheung, P., Tanner, K.G., Cheung, W.L., Sassone-Corsi, P., Denu, J.M., and Allis, C.D. 2000. Synergistic coupling of histone H3 phosphorylation and acetylation in response to epidermal growth factor stimulation. *Mol Cell* **5**(6): 905-915.

Cheung, W.L., Turner, F.B., Krishnamoorthy, T., Wolner, B., Ahn, S.H., Foley, M., Dorsey, J.A., Peterson, C.L., Berger, S.L., and Allis, C.D. 2005. Phosphorylation of histone H4 serine 1 during DNA damage requires casein kinase II in S. cerevisiae. *Curr Biol* **15**(7): 656-660.

Choi, K., Kennedy, M., Kazarov, A., Papadimitriou, J.C., and Keller, G. 1998. A common precursor for hematopoietic and endothelial cells. *Development* **125**(4): 725-732.

Choi, O.R. and Engel, J.D. 1988. Developmental regulation of beta-globin gene switching. *Cell* **55**(1): 17-26.

Chow, C.M., Georgiou, A., Szutorisz, H., Maia e Silva, A., Pombo, A., Barahona, I., Dargelos, E., Canzonetta, C., and Dillon, N. 2005. Variant histone H3.3 marks promoters of transcriptionally active genes during mammalian cell division. *EMBO Rep* **6**(4): 354-360.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O., and Herskowitz, I. 1998. The transcriptional program of sporulation in budding yeast. *Science* **282**(5389): 699-705.

Chung, Y.S., Zhang, W.J., Arentson, E., Kingsley, P.D., Palis, J., and Choi, K. 2002. Lineage analysis of the hemangioblast as defined by FLK1 and SCL expression. *Development* **129**(23): 5511-5520.

Colaizzo-Anas, T. and Aplan, P.D. 2003. Cloning and characterization of the SIL promoter. *Biochim Biophys Acta* **1625**(2): 207-213.

Collins, S.J., Gallo, R.C., and Gallagher, R.E. 1977. Continuous growth and differentiation of human myeloid leukaemic cells in suspension culture. *Nature* **270**(5635): 347-349.

Cottier, M., Tchirkov, A., Perissel, B., Giollant, M., Campos, L., and Vago, P. 2004. Cytogenetic characterization of seven human cancer cell lines by combining G- and R-banding, M-FISH, CGH and chromosome- and locus-specific FISH. *Int J Mol Med* **14**(4): 483-495.

Courey, A.J. and Tjian, R. 1988. Analysis of Sp1 in vivo reveals multiple transcriptional domains, including a novel glutamine-rich activation motif. *Cell* **55**(5): 887-898.

Courtes, C., Lecointe, N., Le Cam, L., Baudoin, F., Sardet, C., and Mathieu-Mahul, D. 2000. Erythroid-specific inhibition of the tal-1 intragenic promoter is due to binding of a repressor to a novel silencer. *J Biol Chem* **275**(2): 949-958.

Crawford, G.E., Holt, I.E., Mullikin, J.C., Tai, D., Blakesley, R., Bouffard, G., Young, A., Masiello, C., Green, E.D., Wolfsberg, T.G., and Collins, F.S. 2004. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proc Natl Acad Sci U S A* **101**(4): 992-997.

Cross, M.A., Heyworth, C.M., Murrell, A.M., Bockamp, E.O., Dexter, T.M., and Green, A.R. 1994. Expression of lineage restricted transcription factors precedes lineage specific differentiation in a multipotent haemopoietic progenitor cell line. *Oncogene* **9**(10): 3013-3016.

Cumano, A., Dieterlen-Lievre, F., and Godin, I. 1996. Lymphoid potential, probed before circulation in mouse, is restricted to caudal intraembryonic splanchnopleura. *Cell* **86**(6): 907-916.

Curtis, D.J., Hall, M.A., Van Stekelenburg, L.J., Robb, L., Jane, S.M., and Begley, C.G. 2004. SCL is required for normal function of short-term repopulating hematopoietic stem cells. *Blood* **103**(9): 3342-3348.

Cuthbert, G.L., Daujat, S., Snowden, A.W., Erdjument-Bromage, H., Hagiwara, T., Yamada, M., Schneider, R., Gregory, P.D., Tempst, P., Bannister, A.J., and Kouzarides, T. 2004. Histone deimination antagonizes arginine methylation. *Cell* **118**(5): 545-553.

D'Souza, S.L., Elefanty, A.G., and Keller, G. 2005. SCL/Tal-1 is essential for hematopoietic commitment of the hemangioblast but not for its development. *Blood* **105**(10): 3862-3870.

Daujat, S., Bauer, U.M., Shah, V., Turner, B., Berger, S., and Kouzarides, T. 2002. Crosstalk between CARM1 methylation and CBP acetylation on histone H3. *Curr Biol* **12**(24): 2090-2097.

Davuluri, R.V., Grosse, I., and Zhang, M.Q. 2001. Computational identification of promoters and first exons in the human genome. *Nat Genet* **29**(4): 412-417.

de Ruijter, A.J., van Gennip, A.H., Caron, H.N., Kemp, S., and van Kuilenburg, A.B. 2003. Histone deacetylases (HDACs): characterization of the classical HDAC family. *Biochem J* **370**(Pt 3): 737-749.

Delabesse, E., Ogilvy, S., Chapman, M.A., Piltz, S.G., Gottgens, B., and Green, A.R. 2005. Transcriptional regulation of the SCL locus: identification of an enhancer that targets the primitive erythroid lineage in vivo. *Mol Cell Biol* **25**(12): 5215-5225.

DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su, Y.A., and Trent, J.M. 1996. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* **14**(4): 457-460.

DeRisi, J.L., Iyer, V.R., and Brown, P.O. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**(5338): 680-686.

Dexter, T.M., Allen, T.D., Scott, D., and Teich, N.M. 1979. Isolation and characterisation of a bipotential haematopoietic cell line. *Nature* **277**(5696): 471-474.

Dhami, P., Coffey, A.J., Abbs, S., Vermeesch, J.R., Dumanski, J.P., Woodward, K.J., Andrews, R.M., Langford, C., and Vetrie, D. 2005. Exon array CGH: detection of copy-number changes at the resolution of individual exons in the human genome. *Am J Hum Genet* **76**(5): 750-762.

Dieterlen-Lievre, F. 1975. On the origin of haemopoietic stem cells in the avian embryo: an experimental approach. *J Embryol Exp Morphol* **33**(3): 607-619.

Dion, M.F., Altschuler, S.J., Wu, L.F., and Rando, O.J. 2005. Genomic characterization reveals a simple histone H4 acetylation code. *Proc Natl Acad Sci U S A* **102**(15): 5501-5506.

Dooley, K.A., Davidson, A.J., and Zon, L.I. 2005. Zebrafish scl functions independently in hematopoietic and endothelial development. *Dev Biol* **277**(2): 522-536.

Dorschner, M.O., Hawrylycz, M., Humbert, R., Wallace, J.C., Shafer, A., Kawamoto, J., Mack, J., Hall, R., Goldy, J., Sabo, P.J., Kohli, A., Li, Q., McArthur, M., and Stamatoyannopoulos, J.A. 2004. High-throughput localization of functional elements by quantitative chromatin profiling. *Nat Methods* **1**(3): 219-225.

Drake, C.J., Brandt, S.J., Trusk, T.C., and Little, C.D. 1997. TAL1/SCL is expressed in endothelial progenitor cells/angioblasts and defines a dorsal-to-ventral gradient of vasculogenesis. *Dev Biol* **192**(1): 17-30.

Drexler, H.G., Dirks, W.G., Matsuo, Y., and MacLeod, R.A. 2003. False leukemia-lymphoma cell lines: an update on over 500 cell lines. *Leukemia* **17**(2): 416-426.

Drexler, H.G., Matsuo, A.Y., and MacLeod, R.A. 2000. Continuous hematopoietic cell lines as model systems for leukemia-lymphoma research. *Leuk Res* **24**(11): 881-911.

Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J.M. 1999. Expression profiling using cDNA microarrays. *Nat Genet* **21**(1 Suppl): 10-14.

Dzierzak, E., Medvinsky, A., and de Bruijn, M. 1998. Qualitative and quantitative aspects of haematopoietic cell development in the mammalian embryo. *Immunol Today* **19**(5): 228-236.

Edmondson, D.G., Smith, M.M., and Roth, S.Y. 1996. Repression domain of the yeast global repressor Tup1 interacts directly with histones H3 and H4. *Genes Dev* **10**(10): 1247-1259.

Eichmann, A., Corbel, C., Nataf, V., Vaigot, P., Breant, C., and Le Douarin, N.M. 1997. Ligand-dependent development of the endothelial and hemopoietic lineages from embryonic mesodermal cells expressing vascular endothelial growth factor receptor 2. *Proc Natl Acad Sci U S A* **94**(10): 5141-5146.

Eickbush, T.H. and Moudrianakis, E.N. 1978. The histone core complex: an octamer assembled by two sets of protein-protein interactions. *Biochemistry* **17**(23): 4955-4964.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**(25): 14863-14868.

Elefanty, A.G., Begley, C.G., Hartley, L., Papaevangeliou, B., and Robb, L. 1999. SCL expression in the mouse embryo detected with a targeted lacZ reporter gene demonstrates its localization to hematopoietic, vascular, and neural tissues. *Blood* **94**(11): 3754-3763.

Elefanty, A.G., Begley, C.G., Metcalf, D., Barnett, L., Kontgen, F., and Robb, L. 1998. Characterization of hematopoietic progenitor cells that express the transcription factor SCL, using a lacZ "knock-in" strategy. *Proc Natl Acad Sci U S A* **95**(20): 11897-11902.

Elefanty, A.G., Robb, L., Birner, R., and Begley, C.G. 1997. Hematopoietic-specific genes are not induced during in vitro differentiation of scl-null embryonic stem cells. *Blood* **90**(4): 1435-1447.

Elwood, N.J., Cook, W.D., Metcalf, D., and Begley, C.G. 1993. SCL, the gene implicated in human T-cell leukaemia, is oncogenic in a murine T-lymphocyte cell line. *Oncogene* **8**(11): 3093-3101.

Elwood, N.J., Green, A.R., Melder, A., Begley, C.G., and Nicola, N. 1994. The SCL protein displays cell-specific heterogeneity in size. *Leukemia* **8**(1): 106-114.

Elwood, N.J., Zogos, H., Pereira, D.S., Dick, J.E., and Begley, C.G. 1998. Enhanced megakaryocyte and erythroid development from normal human CD34(+) cells: consequence of enforced expression of SCL. *Blood* **91**(10): 3756-3765.

Ema, M., Faloon, P., Zhang, W.J., Hirashima, M., Reid, T., Stanford, W.L., Orkin, S., Choi, K., and Rossant, J. 2003. Combinatorial effects of Flk1 and Tal1 on vascular and hematopoietic development in the mouse. *Genes Dev* **17**(3): 380-393.

Endoh, M., Ogawa, M., Orkin, S., and Nishikawa, S. 2002. SCL/tal-1-dependent process determines a competence to select the definitive hematopoietic lineage prior to endothelial differentiation. *Embo J* **21**(24): 6700-6708.

Engel, N., West, A.G., Felsenfeld, G., and Bartolomei, M.S. 2004. Antagonism between DNA hypermethylation and enhancer-blocking activity at the H19 DMD is uncovered by CpG mutations. *Nat Genet* **36**(8): 883-888.

Euskirchen, G., Royce, T.E., Bertone, P., Martone, R., Rinn, J.L., Nelson, F.K., Sayward, F., Luscombe, N.M., Miller, P., Gerstein, M., Weissman, S., and Snyder, M. 2004. CREB binds to multiple loci on human chromosome 22. *Mol Cell Biol* **24**(9): 3804-3814.

Faloon, P., Arentson, E., Kazarov, A., Deng, C.X., Porcher, C., Orkin, S., and Choi, K. 2000. Basic fibroblast growth factor positively regulates hematopoietic development. *Development* **127**(9): 1931-1941.

Fan, G. and Hutnick, L. 2005. Methyl-CpG binding proteins in the nervous system. *Cell Res* **15**(4): 255-261.

Fang, J., Feng, Q., Ketel, C.S., Wang, H., Cao, R., Xia, L., Erdjument-Bromage, H., Tempst, P., Simon, J.A., and Zhang, Y. 2002. Purification and functional characterization of SET8, a nucleosomal histone H4-lysine 20-specific methyltransferase. *Curr Biol* **12**(13): 1086-1099.

Feng, Q., Wang, H., Ng, H.H., Erdjument-Bromage, H., Tempst, P., Struhl, K., and Zhang, Y. 2002. Methylation of H3-lysine 79 is mediated by a new family of HMTases without a SET domain. *Curr Biol* **12**(12): 1052-1058.

Fernandez-Capetillo, O., Allis, C.D., and Nussenzweig, A. 2004. Phosphorylation of histone H2B at DNA double-strand breaks. *J Exp Med* **199**(12): 1671-1677.

Ferreira, R., Ohneda, K., Yamamoto, M., and Philipsen, S. 2005. GATA1 function, a paradigm for transcription factors in hematopoiesis. *Mol Cell Biol* **25**(4): 1215-1227.

Fiegler, H., Carr, P., Douglas, E.J., Burford, D.C., Hunt, S., Scott, C.E., Smith, J., Vetrie, D., Gorman, P., Tomlinson, I.P., and Carter, N.P. 2003. DNA microarrays for comparative

genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. *Genes Chromosomes Cancer* **36**(4): 361-374.

Figueiredo, M.S. and Brownlee, G.G. 1995. cis-acting elements and transcription factors involved in the promoter activity of the human factor VIII gene. *J Biol Chem* **270**(20): 11828-11838.

Finger, L.R., Kagan, J., Christopher, G., Kurtzberg, J., Hershfield, M.S., Nowell, P.C., and Croce, C.M. 1989. Involvement of the TCL5 gene on human chromosome 1 in T-cell leukemia and melanoma. *Proc Natl Acad Sci U S A* **86**(13): 5039-5043.

Fischle, W., Wang, Y., Jacobs, S.A., Kim, Y., Allis, C.D., and Khorasanizadeh, S. 2003. Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains. *Genes Dev* **17**(15): 1870-1881.

Fordham, J.L., Gottgens, B., McLaughlin, F., and Green, A.R. 1999. Chromatin structure and transcriptional regulation of the stem cell leukaemia (SCL) gene in mast cells. *Leukemia* **13**(5): 750-759.

Freiman, R.N., Albright, S.R., Zheng, S., Sha, W.C., Hammer, R.E., and Tjian, R. 2001. Requirement of tissue-selective TBP-associated factor TAFII105 in ovarian development. *Science* **293**(5537): 2084-2087.

Galas, D.J. and Schmitz, A. 1978. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res* **5**(9): 3157-3170.

Garner, M.M. and Revzin, A. 1981. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Res* **9**(13): 3047-3060.

Gavin, I., Horn, P.J., and Peterson, C.L. 2001. SWI/SNF chromatin remodeling requires changes in DNA topology. *Mol Cell* **7**(1): 97-104.

Gering, M., Rodaway, A.R., Gottgens, B., Patient, R.K., and Green, A.R. 1998. The SCL gene specifies haemangioblast development from early mesoderm. *Embo J* **17**(14): 4029-4045.

Gerton, J.L., DeRisi, J., Shroff, R., Lichten, M., Brown, P.O., and Petes, T.D. 2000. Inaugural article: global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **97**(21): 11383-11390.

Gibbs, R.A. Weinstock, G.M. Metzker, M.L. et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**(6982): 493-521.

Godin, I., Dieterlen-Lievre, F., and Cumano, A. 1995. Emergence of multipotent hemopoietic cells in the yolk sac and paraaortic splanchnopleura in mouse embryos, beginning at 8.5 days postcoitus. *Proc Natl Acad Sci U S A* **92**(3): 773-777.

Godin, I.E., Garcia-Porrero, J.A., Coutinho, A., Dieterlen-Lievre, F., and Marcos, M.A. 1993. Para-aortic splanchnopleura from early mouse embryos contains B1a cell progenitors. *Nature* **364**(6432): 67-70.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S.G. 1996. Life with 6000 genes. *Science* **274**(5287): 546, 563-547.

Goldfarb, A.N., Goueli, S., Mickelson, D., and Greenberg, J.M. 1992. T-cell acute lymphoblastic leukemia--the associated gene SCL/tal codes for a 42-Kd nuclear phosphoprotein. *Blood* **80**(11): 2858-2866.

Goldknopf, I.L., Taylor, C.W., Baum, R.M., Yeoman, L.C., Olson, M.O., Prestayko, A.W., and Busch, H. 1975. Isolation and characterization of protein A24, a "histone-like" non-histone chromosomal protein. *J Biol Chem* **250**(18): 7182-7187.

Gonzalez, G.A. and Montminy, M.R. 1989. Cyclic AMP stimulates somatostatin gene transcription by phosphorylation of CREB at serine 133. *Cell* **59**(4): 675-680.

Gottgens, B., Barton, L.M., Chapman, M.A., Sinclair, A.M., Knudsen, B., Grafham, D., Gilbert, J.G., Rogers, J., Bentley, D.R., and Green, A.R. 2002a. Transcriptional regulation of the

stem cell leukemia gene (SCL)--comparative analysis of five vertebrate SCL loci. *Genome Res* **12**(5): 749-759.

Gottgens, B., Barton, L.M., Gilbert, J.G., Bench, A.J., Sanchez, M.J., Bahn, S., Mistry, S., Grafham, D., McMurray, A., Vaudin, M., Amaya, E., Bentley, D.R., Green, A.R., and Sinclair, A.M. 2000. Analysis of vertebrate SCL loci identifies conserved enhancers. *Nat Biotechnol* **18**(2): 181-186.

Gottgens, B., Broccardo, C., Sanchez, M.J., Deveaux, S., Murphy, G., Gothert, J.R., Kotsopoulou, E., Kinston, S., Delaney, L., Piltz, S., Barton, L.M., Knezevic, K., Erber, W.N., Begley, C.G., Frampton, J., and Green, A.R. 2004. The scl +18/19 stem cell enhancer is not required for hematopoiesis: identification of a 5' bifunctional hematopoietic-endothelial enhancer bound by Fli-1 and Elf-1. *Mol Cell Biol* **24**(5): 1870-1883.

Gottgens, B., Gilbert, J.G., Barton, L.M., Grafham, D., Rogers, J., Bentley, D.R., and Green, A.R. 2001. Long-range comparison of human and mouse SCL loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences. *Genome Res* **11**(1): 87-97.

Gottgens, B., McLaughlin, F., Bockamp, E.O., Fordham, J.L., Begley, C.G., Kosmopoulos, K., Elefanty, A.G., and Green, A.R. 1997. Transcription of the SCL gene in erythroid and CD34 positive primitive myeloid cells is controlled by a complex network of lineage-restricted chromatin-dependent and chromatin-independent regulatory elements. *Oncogene* **15**(20): 2419-2428.

Gottgens, B., Nastos, A., Kinston, S., Piltz, S., Delabesse, E.C., Stanley, M., Sanchez, M.J., Ciau-Uitz, A., Patient, R., and Green, A.R. 2002b. Establishing the transcriptional programme for blood: the SCL stem cell enhancer is regulated by a multiprotein complex containing Ets and GATA factors. *Embo J* **21**(12): 3039-3050.

Grant, P.A., Eberharter, A., John, S., Cook, R.G., Turner, B.M., and Workman, J.L. 1999. Expanded lysine acetylation specificity of Gcn5 in native complexes. *J Biol Chem* **274**(9): 5895-5900.

Green, A.R. and Begley, C.G. 1992. SCL and related hemopoietic helix-loop-helix transcription factors. *Int J Cell Cloning* **10**(5): 269-276.

Green, A.R., DeLuca, E., and Begley, C.G. 1991a. Antisense SCL suppresses self-renewal and enhances spontaneous erythroid differentiation of the human leukaemic cell line K562. *Embo J* **10**(13): 4153-4158.

Green, A.R., Rockman, S., DeLuca, E., and Begley, C.G. 1993. Induced myeloid differentiation of K562 cells with downregulation of erythroid and megakaryocytic transcription factors: a novel experimental model for hemopoietic lineage restriction. *Exp Hematol* **21**(4): 525-531.

Green, A.R., Salvaris, E., and Begley, C.G. 1991b. Erythroid expression of the 'helix-loop-helix' gene, SCL. *Oncogene* **6**(3): 475-479.

Grewal, S.I. and Moazed, D. 2003. Heterochromatin and epigenetic control of gene expression. *Science* **301**(5634): 798-802.

Gribble, S.M., Roberts, I., Grace, C., Andrews, K.M., Green, A.R., and Nacheva, E.P. 2000. Cytogenetics of the chronic myeloid leukemia-derived cell line K562: karyotype clarification by multicolor fluorescence in situ hybridization, comparative genomic hybridization, and locus-specific fluorescence in situ hybridization. *Cancer Genet Cytogenet* **118**(1): 1-8.

Gross, D.S. and Garrard, W.T. 1988. Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* **57**: 159-197.

Grosveld, F., van Assendelft, G.B., Greaves, D.R., and Kollias, G. 1987. Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell* **51**(6): 975-985.

Gu, W. and Roeder, R.G. 1997. Activation of p53 sequence-specific DNA binding by acetylation of the p53 C-terminal domain. *Cell* **90**(4): 595-606.

Gu, X., Shin, B.H., Akbarali, Y., Weiss, A., Boltax, J., Oettgen, P., and Libermann, T.A. 2001. Tel-2 is a novel transcriptional repressor related to the Ets factor Tel/ETV-6. *J Biol Chem* **276**(12): 9421-9436.

Hahn, S. 2004. Structure and mechanism of the RNA polymerase II transcription machinery. *Nat Struct Mol Biol* **11**(5): 394-403.

Hall, M.A., Curtis, D.J., Metcalf, D., Elefanty, A.G., Sourris, K., Robb, L., Gothert, J.R., Jane, S.M., and Begley, C.G. 2003. The critical regulator of embryonic hematopoiesis, SCL, is vital in the adult for megakaryopoiesis, erythropoiesis, and lineage choice in CFU-S12. *Proc Natl Acad Sci U S A* **100**(3): 992-997.

Hamiche, A., Sandaltzopoulos, R., Gdula, D.A., and Wu, C. 1999. ATP-dependent histone octamer sliding mediated by the chromatin remodeling complex NURF. *Cell* **97**(7): 833-842.

Hansen, J.C., Tse, C., and Wolffe, A.P. 1998. Structure and function of the core histone N-termini: more than meets the eye. *Biochemistry* **37**(51): 17637-17641.

Havas, K., Flaus, A., Phelan, M., Kingston, R., Wade, P.A., Lilley, D.M., and Owen-Hughes, T. 2000. Generation of superhelical torsion by ATP-dependent chromatin remodeling activities. *Cell* **103**(7): 1133-1142.

Hebbes, T.R., Clayton, A.L., Thorne, A.W., and Crane-Robinson, C. 1994. Core histone hyperacetylation co-maps with generalized DNase I sensitivity in the chicken beta-globin chromosomal domain. *Embo J* **13**(8): 1823-1830.

Heinemeyer, T., Wingender, E., Reuter, I., Hermjakob, H., Kel, A.E., Kel, O.V., Ignatieva, E.V., Ananko, E.A., Podkolodnaya, O.A., Kolpakov, F.A., Podkolodny, N.L., and Kolchanov, N.A. 1998. Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res* **26**(1): 362-367.

Hidaka, M., Stanford, W.L., and Bernstein, A. 1999. Conditional requirement for the Flk-1 receptor in the in vitro generation of early hematopoietic cells. *Proc Natl Acad Sci U S A* **96**(13): 7370-7375.

Hoang, T., Paradis, E., Brady, G., Billia, F., Nakahara, K., Iscove, N.N., and Kirsch, I.R. 1996. Opposing effects of the basic helix-loop-helix transcription factor SCL on erythroid and monocytic differentiation. *Blood* **87**(1): 102-111.

Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S., and Young, R.A. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**(5): 717-728.

Hooper, M., Hardy, K., Handyside, A., Hunter, S., and Monk, M. 1987. HPRT-deficient (Lesch-Nyhan) mouse embryos derived from germline colonization by cultured cells. *Nature* **326**(6110): 292-295.

Hope, I.A., Mahadevan, S., and Struhl, K. 1988. Structural and functional characterization of the short acidic transcriptional activation region of yeast GCN4 protein. *Nature* **333**(6174): 635-640.

Horak, C.E., Mahajan, M.C., Luscombe, N.M., Gerstein, M., Weissman, S.M., and Snyder, M. 2002. GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIp-chip analysis. *Proc Natl Acad Sci U S A* **99**(5): 2924-2929.

Houssaint, E. 1981. Differentiation of the mouse hepatic primordium. II. Extrinsic origin of the haemopoietic cell line. *Cell Differ* **10**(5): 243-252.

Hsu, H.L., Huang, L., Tsan, J.T., Funk, W., Wright, W.E., Hu, J.S., Kingston, R.E., and Baer, R. 1994a. Preferred sequences for DNA recognition by the TAL1 helix-loop-helix proteins. *Mol Cell Biol* **14**(2): 1256-1265.

Hsu, H.L., Wadman, I., and Baer, R. 1994b. Formation of in vivo complexes between the TAL1 and E2A polypeptides of leukemic T cells. *Proc Natl Acad Sci U S A* **91**(8): 3181-3185.

Hsu, H.L., Wadman, I., Tsan, J.T., and Baer, R. 1994c. Positive and negative transcriptional control by the TAL1 helix-loop-helix protein. *Proc Natl Acad Sci U S A* **91**(13): 5947-5951.

Huang, S. and Brandt, S.J. 2000. mSin3A regulates murine erythroleukemia cell differentiation through association with the TAL1 (or SCL) transcription factor. *Mol Cell Biol* **20**(6): 2248-2259.

Huang, S., Qiu, Y., Shi, Y., Xu, Z., and Brandt, S.J. 2000. P/CAF-mediated acetylation regulates the function of the basic helix-loop-helix transcription factor TAL1/SCL. *Embo J* **19**(24): 6792-6803.

Huang, S., Qiu, Y., Stein, R.W., and Brandt, S.J. 1999. p300 functions as a transcriptional coactivator for the TAL1/SCL oncoprotein. *Oncogene* **18**(35): 4958-4967.

Hughes, T.R., Mao, M., Jones, A.R., Burchard, J., Marton, M.J., Shannon, K.W., Lefkowitz, S.M., Ziman, M., Schelter, J.M., Meyer, M.R., Kobayashi, S., Davis, C., Dai, H., He, Y.D., Stephaniants, S.B., Cavet, G., Walker, W.L., West, A., Coffey, E., Shoemaker, D.D., Stoughton, R., Blanchard, A.P., Friend, S.H., and Linsley, P.S. 2001. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* **19**(4): 342-347.

Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., and Friend, S.H. 2000. Functional discovery via a compendium of expression profiles. *Cell* **102**(1): 109-126.

Hwang, L.Y., Siegelman, M., Davis, L., Oppenheimer-Marks, N., and Baer, R. 1993. Expression of the TAL1 proto-oncogene in cultured endothelial cells and blood vessels of the spleen. *Oncogene* **8**(11): 3043-3046.

Im, H., Park, C., Feng, Q., Johnson, K.D., Kiekhaefer, C.M., Choi, K., Zhang, Y., and Bresnick, E.H. 2003. Dynamic regulation of histone H3 methylated at lysine 79 within a tissue-specific chromatin domain. *J Biol Chem* **278**(20): 18346-18352.

Iyer, V.R., Horak, C.E., Scafe, C.S., Botstein, D., Snyder, M., and Brown, P.O. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**(6819): 533-538.

Izraeli, S., Colaizzo-Anas, T., Bertness, V.L., Mani, K., Aplan, P.D., and Kirsch, I.R. 1997. Expression of the SIL gene is correlated with growth induction and cellular proliferation. *Cell Growth Differ* **8**(11): 1171-1179.

Jacobson, R.H., Ladurner, A.G., King, D.S., and Tjian, R. 2000. Structure and function of a human TAFII250 double bromodomain module. *Science* **288**(5470): 1422-1425.

Jenuwein, T. and Allis, C.D. 2001. Translating the histone code. *Science* **293**(5532): 1074-1080.

Johnson, G.R. and Moore, M.A. 1975. Role of stem cell migration in initiation of mouse foetal liver haemopoiesis. *Nature* **258**(5537): 726-728.

Kadonaga, J.T. 2004. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell* **116**(2): 247-257.

Kallianpur, A.R., Jordan, J.E., and Brandt, S.J. 1994. The SCL/TAL-1 gene is expressed in progenitors of both the hematopoietic and vascular systems during embryogenesis. *Blood* **83**(5): 1200-1208.

Kappel, A., Schlaeger, T.M., Flamme, I., Orkin, S.H., Risau, W., and Breier, G. 2000. Role of SCL/Tal-1, GATA, and ets transcription factor binding sites for the regulation of flk-1 expression during murine vascular development. *Blood* **96**(9): 3078-3085.

Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P., and Gingeras, T.R. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**(5569): 916-919.

Kennedy, M., Firpo, M., Choi, K., Wall, C., Robertson, S., Kabrun, N., and Keller, G. 1997. A common precursor for primitive erythropoiesis and definitive haematopoiesis. *Nature* **386**(6624): 488-493.

Khorasanizadeh, S. 2004. The nucleosome: from genomic organization to genomic regulation. *Cell* **116**(2): 259-272.

Kim, J., Bhinge, A.A., Morgan, X.C., and Iyer, V.R. 2005a. Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment. *Nat Methods* **2**(1): 47-53.

Kim, J., Jia, L., Tilley, W.D., and Coetzee, G.A. 2003. Dynamic methylation of histone H3 at lysine 4 in transcriptional regulation by the androgen receptor. *Nucleic Acids Res* **31**(23): 6741-6747.

Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmond, T.A., Wu, Y., Green, R.D., and Ren, B. 2005b. A high-resolution map of active promoters in the human genome. *Nature* **436**(7052): 876-880.

Kingston, R.E. and Narlikar, G.J. 1999. ATP-dependent remodeling and acetylation as regulators of chromatin fluidity. *Genes Dev* **13**(18): 2339-2352.

Klenova, E.M., Nicolas, R.H., Paterson, H.F., Carne, A.F., Heath, C.M., Goodwin, G.H., Neiman, P.E., and Lobanenkov, V.V. 1993. CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. *Mol Cell Biol* **13**(12): 7612-7624.

Kocher, O., Cheresh, P., and Lee, S.W. 1996. Identification and partial characterization of a novel membrane-associated protein (MAP17) up-regulated in human carcinomas and modulating cell replication and tumor growth. *Am J Pathol* **149**(2): 493-500.

Kondo, Y., Shen, L., Yan, P.S., Huang, T.H., and Issa, J.P. 2004. Chromatin immunoprecipitation microarrays for identification of genes silenced by histone H3 lysine 9 methylation. *Proc Natl Acad Sci U S A* **101**(19): 7398-7403.

Korber, P., Luckenbach, T., Blaschke, D., and Horz, W. 2004. Evidence for histone eviction in trans upon induction of the yeast PHO5 promoter. *Mol Cell Biol* **24**(24): 10965-10974.

Kouzarides, T. 2002. Histone methylation in transcriptional control. *Curr Opin Genet Dev* **12**(2): 198-209.

Krogan, N.J., Dover, J., Wood, A., Schneider, J., Heidt, J., Boateng, M.A., Dean, K., Ryan, O.W., Golshani, A., Johnston, M., Greenblatt, J.F., and Shilatifard, A. 2003. The Paf1 complex is required for histone H3 methylation by COMPASS and Dot1p: linking transcriptional elongation to histone methylation. *Mol Cell* **11**(3): 721-729.

Krosl, G., He, G., Lefrancois, M., Charron, F., Romeo, P.H., Jolicoeur, P., Kirsch, I.R., Nemer, M., and Hoang, T. 1998. Transcription factor SCL is required for c-kit expression and c-Kit function in hemopoietic cells. *J Exp Med* **188**(3): 439-450.

Kurdistani, S.K. and Grunstein, M. 2003. In vivo protein-protein and protein-DNA crosslinking for genomewide binding microarray. *Methods* **31**(1): 90-95.

Kurdistani, S.K., Tavazoie, S., and Grunstein, M. 2004. Mapping global histone acetylation patterns to gene expression. *Cell* **117**(6): 721-733.

Labastie, M.C., Cortes, F., Romeo, P.H., Dulac, C., and Peault, B. 1998. Molecular identity of hematopoietic precursor cells emerging in the human embryo. *Blood* **92**(10): 3624-3635.

Lachner, M. and Jenuwein, T. 2002. The many faces of histone lysine methylation. *Curr Opin Cell Biol* **14**(3): 286-298.

Lachner, M., O'Carroll, D., Rea, S., Mechtler, K., and Jenuwein, T. 2001. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410**(6824): 116-120.

Lachner, M., O'Sullivan, R.J., and Jenuwein, T. 2003. An epigenetic road map for histone lysine methylation. *J Cell Sci* **116**(Pt 11): 2117-2124.

Lagger, G., O'Carroll, D., Rembold, M., Khier, H., Tischler, J., Weitzer, G., Schuettengruber, B., Hauser, C., Brunmeir, R., Jenuwein, T., and Seiser, C. 2002. Essential function of histone deacetylase 1 in proliferation control and CDK inhibitor repression. *Embo J* **21**(11): 2672-2681.

Lahlil, R., Lecuyer, E., Herblot, S., and Hoang, T. 2004. SCL assembles a multifactorial complex that determines glycophorin A expression. *Mol Cell Biol* **24**(4): 1439-1452.

Lander, E.S. Linton, L.M. Birren, B. Nusbaum, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Langst, G. and Becker, P.B. 2001. Nucleosome mobilization and positioning by ISWI-containing chromatin-remodeling factors. *J Cell Sci* **114**(Pt 14): 2561-2568.

Larson, R.C., Lavenir, I., Larson, T.A., Baer, R., Warren, A.J., Wadman, I., Nottage, K., and Rabbitts, T.H. 1996. Protein dimerization between Lmo2 (Rbtn2) and Tal1 alters thymocyte development and potentiates T cell tumorigenesis in transgenic mice. *Embo J* **15**(5): 1021-1027.

Lazo, P.A., Lee, J.S., and Tsichlis, P.N. 1990. Long-distance activation of the Myc protooncogene by provirus insertion in Mlvi-1 or Mlvi-4 in rat T-cell lymphomas. *Proc Natl Acad Sci U S A* **87**(1): 170-173.

Lazrak, M., Deleuze, V., Noel, D., Haouzi, D., Chalhoub, E., Dohet, C., Robbins, I., and Mathieu, D. 2004. The bHLH TAL-1/SCL regulates endothelial cell migration and morphogenesis. *J Cell Sci* **117**(Pt 7): 1161-1171.

Lecointe, N., Bernard, O., Naert, K., Joulin, V., Larsen, C.J., Romeo, P.H., and Mathieu-Mahul, D. 1994. GATA-and SP1-binding sites are required for the full activity of the tissue-specific promoter of the tal-1 gene. *Oncogene* **9**(9): 2623-2632.

Lecuyer, E., Herblot, S., Saint-Denis, M., Martin, R., Begley, C.G., Porcher, C., Orkin, S.H., and Hoang, T. 2002. The SCL complex regulates c-kit expression in hematopoietic cells through functional interaction with Sp1. *Blood* **100**(7): 2430-2440.

Lee, C.K., Shibata, Y., Rao, B., Strahl, B.D., and Lieb, J.D. 2004a. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat Genet* **36**(8): 900-905.

Lee, D.U., Avni, O., Chen, L., and Rao, A. 2004b. A distal enhancer in the interferon-gamma (IFN-gamma) locus revealed by genome sequence comparison. *J Biol Chem* **279**(6): 4802-4810.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K., and Young, R.A. 2002. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**(5594): 799-804.

Lenhard, B., Sandelin, A., Mendoza, L., Engstrom, P., Jareborg, N., and Wasserman, W.W. 2003. Identification of conserved regulatory elements by comparative genome analysis. *J Biol* **2**(2): 13.

Leroy-Viard, K., Vinit, M.A., Lecointe, N., Mathieu-Mahul, D., and Romeo, P.H. 1994. Distinct DNase-I hypersensitive sites are associated with TAL-1 transcription in erythroid and T-cell lines. *Blood* **84**(11): 3819-3827.

Leuther, K.K., Salmeron, J.M., and Johnston, S.A. 1993. Genetic evidence that an activation domain of GAL4 does not require acidity and may form a beta sheet. *Cell* **72**(4): 575-585.

Levine, M. and Tjian, R. 2003. Transcription regulation and animal diversity. *Nature* **424**(6945): 147-151.

Liang, G., Lin, J.C., Wei, V., Yoo, C., Cheng, J.C., Nguyen, C.T., Weisenberger, D.J., Egger, G., Takai, D., Gonzales, F.A., and Jones, P.A. 2004. Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. *Proc Natl Acad Sci U S A* **101**(19): 7357-7362.

Liao, E.C., Paw, B.H., Oates, A.C., Pratt, S.J., Postlethwait, J.H., and Zon, L.I. 1998. SCL/Tal-1 transcription factor acts downstream of cloche to specify hematopoietic and vascular progenitors in zebrafish. *Genes Dev* **12**(5): 621-626.

Liao, W., Bisgrove, B.W., Sawyer, H., Hug, B., Bell, B., Peters, K., Grunwald, D.J., and Stainier, D.Y. 1997. The zebrafish gene cloche acts upstream of a flk-1 homologue to regulate endothelial cell differentiation. *Development* **124**(2): 381-389.

Lieb, J.D., Liu, X., Botstein, D., and Brown, P.O. 2001. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat Genet* **28**(4): 327-334.

Lien, C.L., McAnally, J., Richardson, J.A., and Olson, E.N. 2002. Cardiac-specific activity of an Nkx2-5 enhancer requires an evolutionarily conserved Smad binding site. *Dev Biol* **244**(2): 257-266.

Lipshutz, R.J., Fodor, S.P., Gingeras, T.R., and Lockhart, D.J. 1999. High density synthetic oligonucleotide arrays. *Nat Genet* **21**(1 Suppl): 20-24.

Litt, M.D., Simpson, M., Gaszner, M., Allis, C.D., and Felsenfeld, G. 2001a. Correlation between histone lysine methylation and developmental changes at the chicken beta-globin locus. *Science* **293**(5539): 2453-2455.

Litt, M.D., Simpson, M., Recillas-Targa, F., Prioleau, M.N., and Felsenfeld, G. 2001b. Transitions in histone acetylation reveal boundaries of three separately regulated neighboring loci. *Embo J* **20**(9): 2224-2235.

Liu, C.L., Kaplan, T., Kim, M., Buratowski, S., Schreiber, S.L., Friedman, N., and Rando, O.J. 2005. Single-nucleosome mapping of histone modifications in S. cerevisiae. *PLoS Biol* **3**(10): e328.

Liu, C.L., Schreiber, S.L., and Bernstein, B.E. 2003. Development and validation of a T7 based linear amplification for genomic DNA. *BMC Genomics* **4**(1): 19.

Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M., and Frazer, K.A. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**(5463): 136-140.

Lopez, R.G., Carron, C., Oury, C., Gardellin, P., Bernard, O., and Ghysdael, J. 1999. TEL is a sequence-specific transcriptional repressor. *J Biol Chem* **274**(42): 30132-30138.

Lorch, Y., Cairns, B.R., Zhang, M., and Kornberg, R.D. 1998. Activated RSC-nucleosome complex and persistently altered form of the nucleosome. *Cell* **94**(1): 29-34.

Lorch, Y., Zhang, M., and Kornberg, R.D. 1999. Histone octamer transfer by a chromatin-remodeling complex. *Cell* **96**(3): 389-392.

Loury, R. and Sassone-Corsi, P. 2003. Histone phosphorylation: how to proceed. *Methods* **31**(1): 40-48.

Lozzio, B.B. and Lozzio, C.B. 1977. Properties of the K562 cell line derived from a patient with chronic myeloid leukemia. *Int J Cancer* **19**(1): 136.

Lu, Q., Wallrath, L.L., and Elgin, S.C. 1995. The role of a positioned nucleosome at the Drosophila melanogaster hsp26 promoter. *Embo J* **14**(19): 4738-4746.

Lucito, R., West, J., Reiner, A., Alexander, J., Esposito, D., Mishra, B., Powers, S., Norton, L., and Wigler, M. 2000. Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. *Genome Res* **10**(11): 1726-1736.

Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. 1997. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**(6648): 251-260.

Ma, J. and Ptashne, M. 1987. A new class of yeast transcriptional activators. *Cell* **51**(1): 113-119.

MacLeod, R.A., Nagel, S., Kaufmann, M., Janssen, J.W., and Drexler, H.G. 2003. Activation of HOX11L2 by juxtaposition with 3'-BCL11B in an acute lymphoblastic leukemia cell line (HPB-ALL) with t(5;14)(q35;q32.2). *Genes Chromosomes Cancer* **37**(1): 84-91.

Mahadevan, L.C., Willis, A.C., and Barratt, M.J. 1991. Rapid histone H3 phosphorylation in response to growth factors, phorbol esters, okadaic acid, and protein synthesis inhibitors. *Cell* **65**(5): 775-783.

Malik, H.S. and Henikoff, S. 2003. Phylogenomics of the nucleosome. *Nat Struct Biol* **10**(11): 882-891.

Manaia, A., Lemarchandel, V., Klaine, M., Max-Audit, I., Romeo, P., Dieterlen-Lievre, F., and Godin, I. 2000. Lmo2 and GATA-3 associated expression in intraembryonic hemogenic sites. *Development* **127**(3): 643-653.

Maniatis, T., Goodbourn, S., and Fischer, J.A. 1987. Regulation of inducible and tissue-specific gene expression. *Science* **236**(4806): 1237-1245.

Mantripragada, K.K., Buckley, P.G., Benetkiewicz, M., De Bustos, C., Hirvela, C., Jarbo, C., Bruder, C.E., Wensman, H., Mathiesen, T., Nyberg, G., Papi, L., Collins, V.P., Ichimura, K., Evans, G., and Dumanski, J.P. 2003. High-resolution profiling of an 11 Mb segment of human chromosome 22 in sporadic schwannoma using array-CGH. *Int J Oncol* **22**(3): 615-622.

Mantripragada, K.K., Buckley, P.G., de Stahl, T.D., and Dumanski, J.P. 2004. Genomic microarrays in the spotlight. *Trends Genet* **20**(2): 87-94.

Mao, D.Y., Watson, J.D., Yan, P.S., Barsyte-Lovejoy, D., Khosravi, F., Wong, W.W., Farnham, P.J., Huang, T.H., and Penn, L.Z. 2003. Analysis of Myc bound loci identified by CpG island arrays shows that Max is essential for Myc-dependent repression. *Curr Biol* **13**(10): 882-886.

Marmorstein, R. and Berger, S.L. 2001. Structure and function of bromodomains in chromatin-regulating complexes. *Gene* **272**(1-2): 1-9.

Martienssen, R.A., Doerge, R.W., and Colot, V. 2005. Epigenomic mapping in Arabidopsis using tiling microarrays. *Chromosome Res* **13**(3): 299-308.

Marton, M.J., DeRisi, J.L., Bennett, H.A., Iyer, V.R., Meyer, M.R., Roberts, C.J., Stoughton, R., Burchard, J., Slade, D., Dai, H., Bassett, D.E., Jr., Hartwell, L.H., Brown, P.O., and Friend, S.H. 1998. Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat Med* **4**(11): 1293-1301.

Martone, R., Euskirchen, G., Bertone, P., Hartman, S., Royce, T.E., Luscombe, N.M., Rinn, J.L., Nelson, F.K., Miller, P., Gerstein, M., Weissman, S., and Snyder, M. 2003. Distribution of NF-kappaB-binding sites across human chromosome 22. *Proc Natl Acad Sci U S A* **100**(21): 12247-12252.

Mautner, J., Joos, S., Werner, T., Eick, D., Bornkamm, G.W., and Polack, A. 1995. Identification of two enhancer elements downstream of the human c-myc gene. *Nucleic Acids Res* **23**(1): 72-80.

Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A., Pachter, L.S., and Dubchak, I. 2000. VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16**(11): 1046-1047.

Medvinsky, A. and Dzierzak, E. 1996. Definitive hematopoiesis is autonomously initiated by the AGM region. *Cell* **86**(6): 897-906.

Medvinsky, A.L., Samoylina, N.L., Muller, A.M., and Dzierzak, E.A. 1993. An early pre-liver intraembryonic source of CFU-S in the developing mouse. *Nature* **364**(6432): 64-67.

Meersseman, G., Pennings, S., and Bradbury, E.M. 1992. Mobile nucleosomes--a general behavior. *Embo J* **11**(8): 2951-2959.

Mendez, J. and Stillman, B. 2000. Chromatin association of human origin recognition complex, cdc6, and minichromosome maintenance proteins during the cell cycle: assembly of prereplication complexes in late mitosis. *Mol Cell Biol* **20**(22): 8602-8612.

Mermod, N., O'Neill, E.A., Kelly, T.J., and Tjian, R. 1989. The proline-rich transcriptional activator of CTF/NF-I is distinct from the replication and DNA binding domain. *Cell* **58**(4): 741-753.

Metivier, R., Penot, G., Hubner, M.R., Reid, G., Brand, H., Kos, M., and Gannon, F. 2003. Estrogen receptor-alpha directs ordered, cyclical, and combinatorial recruitment of cofactors on a natural target promoter. *Cell* **115**(6): 751-763.

Mielke, C., Kohwi, Y., Kohwi-Shigematsu, T., and Bode, J. 1990. Hierarchical binding of DNA fragments derived from scaffold-attached regions: correlation of properties in vitro and function in vivo. *Biochemistry* **29**(32): 7475-7485.

Mikkola, H.K., Klintman, J., Yang, H., Hock, H., Schlaeger, T.M., Fujiwara, Y., and Orkin, S.H. 2003. Haematopoietic stem cells retain long-term repopulating activity and multipotency in the absence of stem-cell leukaemia SCL/tal-1 gene. *Nature* **421**(6922): 547-551.

Mito, Y., Henikoff, J.G., and Henikoff, S. 2005. Genome-scale profiling of histone H3.3 replacement patterns. *Nat Genet* **37**(10): 1090-1097.

Mizzen, C.A., Yang, X.J., Kokubo, T., Brownell, J.E., Bannister, A.J., Owen-Hughes, T., Workman, J., Wang, L., Berger, S.L., Kouzarides, T., Nakatani, Y., and Allis, C.D. 1996. The TAF(II)250 subunit of TFIID has histone acetyltransferase activity. *Cell* **87**(7): 1261-1270.

Moazed, D. 2001. Common themes in mechanisms of gene silencing. *Mol Cell* **8**(3): 489-498.

Moore, M.A. and Metcalf, D. 1970. Ontogeny of the haemopoietic system: yolk sac origin of in vivo and in vitro colony forming cells in the developing mouse embryo. *Br J Haematol* **18**(3): 279-296.

Moore, M.A. and Owen, J.J. 1967. Experimental studies on the development of the thymus. *J Exp Med* **126**(4): 715-726.

Morikawa, S., Tatsumi, E., Baba, M., Harada, T., and Yasuhira, K. 1978. Two E-rosette-forming lymphoid cell lines. *Int J Cancer* **21**(2): 166-170.

Mouthon, M.A., Bernard, O., Mitjavila, M.T., Romeo, P.H., Vainchenker, W., and Mathieu-Mahul, D. 1993. Expression of tal-1 and GATA-binding proteins during human hematopoiesis. *Blood* **81**(3): 647-655.

Mueller-Storm, H.P., Sogo, J.M., and Schaffner, W. 1989. An enhancer stimulates transcription in trans when attached to the promoter via a protein bridge. *Cell* **58**(4): 767-777.

Muller, A.M., Medvinsky, A., Strouboulis, J., Grosveld, F., and Dzierzak, E. 1994. Development of hematopoietic stem cell activity in the mouse embryo. *Immunity* **1**(4): 291-301.

Munshi, N., Merika, M., Yie, J., Senger, K., Chen, G., and Thanos, D. 1998. Acetylation of HMG I(Y) by CBP turns off IFN beta expression by disrupting the enhanceosome. *Mol Cell* **2**(4): 457-467.

Murray, P. 1932. The development 'in vitro' of blood of the early chick embryo. *Strangeways Res Labor Cambridge*: 497-521.

Murre, C., McCaw, P.S., and Baltimore, D. 1989. A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins. *Cell* **56**(5): 777-783.

Murrell, A.M., Bockamp, E.O., Gottgens, B., Chan, Y.S., Cross, M.A., Heyworth, C.M., and Green, A.R. 1995. Discordant regulation of SCL/TAL-1 mRNA and protein during erythroid differentiation. *Oncogene* **11**(1): 131-139.

Myers, F.A., Chong, W., Evans, D.R., Thorne, A.W., and Crane-Robinson, C. 2003. Acetylation of histone H2B mirrors that of H4 and H3 at the chicken beta-globin locus but not at housekeeping genes. *J Biol Chem* **278**(38): 36315-36322.

Nagy, P.L., Griesenbeck, J., Kornberg, R.D., and Cleary, M.L. 2002. A trithorax-group complex purified from Saccharomyces cerevisiae is required for methylation of histone H3. *Proc Natl Acad Sci U S A* **99**(1): 90-94.

Nan, X., Ng, H.H., Johnson, C.A., Laherty, C.D., Turner, B.M., Eisenman, R.N., and Bird, A. 1998. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* **393**(6683): 386-389.

Narlikar, G.J., Fan, H.Y., and Kingston, R.E. 2002. Cooperation between complexes that regulate chromatin structure and transcription. *Cell* **108**(4): 475-487.

Naumann, S., Reutzel, D., Speicher, M., and Decker, H.J. 2001. Complete karyotype characterization of the K562 cell line by combined application of G-banding, multiplex-fluorescence in situ hybridization, fluorescence in situ hybridization, and comparative genomic hybridization. *Leuk Res* **25**(4): 313-322.

Ng, H.H., Feng, Q., Wang, H., Erdjument-Bromage, H., Tempst, P., Zhang, Y., and Struhl, K. 2002a. Lysine methylation within the globular domain of histone H3 by Dot1 is important for telomeric silencing and Sir protein association. *Genes Dev* **16**(12): 1518-1527.

Ng, H.H., Robert, F., Young, R.A., and Struhl, K. 2003. Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Mol Cell* **11**(3): 709-719.

Ng, H.H., Xu, R.M., Zhang, Y., and Struhl, K. 2002b. Ubiquitination of histone H2B by Rad6 is required for efficient Dot1-mediated methylation of histone H3 lysine 79. *J Biol Chem* **277**(38): 34655-34657.

Nickel, B.E., Allis, C.D., and Davie, J.R. 1989. Ubiquitinated histone H2B is preferentially located in transcriptionally active chromatin. *Biochemistry* **28**(3): 958-963.

O'Neil, J., Billa, M., Oikemus, S., and Kelliher, M. 2001. The DNA binding activity of TAL-1 is not required to induce leukemia/lymphoma in mice. *Oncogene* **20**(29): 3897-3905.

O'Neil, J., Shank, J., Cusson, N., Murre, C., and Kelliher, M. 2004. TAL1/SCL induces leukemia by inhibiting the transcriptional activity of E47/HEB. *Cancer Cell* **5**(6): 587-596.

Oberley, M.J., Inman, D.R., and Farnham, P.J. 2003. E2F6 negatively regulates BRCA1 in human cancer cells without methylation of histone H3 on lysine 9. *J Biol Chem* **278**(43): 42466-42476.

Odom, D.T., Zizlsperger, N., Gordon, D.B., Bell, G.W., Rinaldi, N.J., Murray, H.L., Volkert, T.L., Schreiber, J., Rolfe, P.A., Gifford, D.K., Fraenkel, E., Bell, G.I., and Young, R.A. 2004. Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303**(5662): 1378-1381.

Okazaki, Y. Furuno, M. Kasukawa, T. et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**(6915): 563-573.

Orian, A., van Steensel, B., Delrow, J., Bussemaker, H.J., Li, L., Sawado, T., Williams, E., Loo, L.W., Cowley, S.M., Yost, C., Pierce, S., Edgar, B.A., Parkhurst, S.M., and Eisenman, R.N. 2003. Genomic binding by the Drosophila Myc, Max, Mad/Mnt transcription factor network. *Genes Dev* **17**(9): 1101-1114.

Orkin, S.H. and Zon, L.I. 2002. Hematopoiesis and stem cells: plasticity versus developmental heterogeneity. *Nat Immunol* **3**(4): 323-328.

Orlando, V. 2000. Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem Sci* **25**(3): 99-104.

Orlando, V., Strutt, H., and Paro, R. 1997. Analysis of chromatin structure by in vivo formaldehyde cross-linking. *Methods* **11**(2): 205-214.

Osley, M.A. 1991. The regulation of histone synthesis in the cell cycle. *Annu Rev Biochem* **60**: 827-861.

Page, R.D. 1996. TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci* **12**(4): 357-358.

Palis, J., Robertson, S., Kennedy, M., Wall, C., and Keller, G. 1999. Development of erythroid and myeloid progenitors in the yolk sac and embryo proper of the mouse. *Development* **126**(22): 5073-5084.

Pardy, K. 1994. Reporter enzymes for the study of promoter activity. *Mol Biotechnol* **2**(1): 23-27.

Pennacchio, L.A. and Rubin, E.M. 2001. Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet* **2**(2): 100-109.

Pham, A.D. and Sauer, F. 2000. Ubiquitin-activating/conjugating activity of TAFII250, a mediator of activation of gene expression in Drosophila. *Science* **289**(5488): 2357-2360.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y., Dairkee, S.H., Ljung, B.M., Gray, J.W., and Albertson, D.G. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* **20**(2): 207-211.

Poirel, H., Lopez, R.G., Lacronique, V., Della Valle, V., Mauchauffe, M., Berger, R., Ghysdael, J., and Bernard, O.A. 2000. Characterization of a novel ETS gene, TELB, encoding a protein structurally and functionally related to TEL. *Oncogene* **19**(41): 4802-4806.

Pokholok, D.K., Harbison, C.T., Levine, S., Cole, M., Hannett, N.M., Lee, T.I., Bell, G.W., Walker, K., Rolfe, P.A., Herbolsheimer, E., Zeitlinger, J., Lewitter, F., Gifford, D.K., and Young, R.A. 2005. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **122**(4): 517-527.

Pollack, J.R., Perou, C.M., Alizadeh, A.A., Eisen, M.B., Pergamenschikov, A., Williams, C.F., Jeffrey, S.S., Botstein, D., and Brown, P.O. 1999. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet* **23**(1): 41-46.

Porcher, C., Swat, W., Rockwell, K., Fujiwara, Y., Alt, F.W., and Orkin, S.H. 1996. The T cell leukemia oncoprotein SCL/tal-1 is essential for development of all hematopoietic lineages. *Cell* **86**(1): 47-57.

Prasad, K.S., Jordan, J.E., Koury, M.J., Bondurant, M.C., and Brandt, S.J. 1995. Erythropoietin stimulates transcription of the TAL1/SCL gene and phosphorylation of its protein products. *J Biol Chem* **270**(19): 11603-11611.

Pulford, K., Lecointe, N., Leroy-Viard, K., Jones, M., Mathieu-Mahul, D., and Mason, D.Y. 1995. Expression of TAL-1 proteins in human tissues. *Blood* **85**(3): 675-684.

Raghuraman, M.K., Winzeler, E.A., Collingwood, D., Hunt, S., Wodicka, L., Conway, A., Lockhart, D.J., Davis, R.W., Brewer, B.J., and Fangman, W.L. 2001. Replication dynamics of the yeast genome. *Science* **294**(5540): 115-121.

Ramakrishnan, V. 1997. Histone structure and the organization of the nucleosome. *Annu Rev Biophys Biomol Struct* **26**: 83-112.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P., Poggio, T., Gerald, W., Loda, M., Lander, E.S., and Golub, T.R. 2001. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A* **98**(26): 15149-15154.

Rea, S., Eisenhaber, F., O'Carroll, D., Strahl, B.D., Sun, Z.W., Schmid, M., Opravil, S., Mechtler, K., Ponting, C.P., Allis, C.D., and Jenuwein, T. 2000. Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature* **406**(6796): 593-599.

Reid, J.L., Iyer, V.R., Brown, P.O., and Struhl, K. 2000. Coordinate regulation of yeast ribosomal protein genes is associated with targeted recruitment of Esa1 histone acetylase. *Mol Cell* **6**(6): 1297-1307.

Reinke, H. and Horz, W. 2003. Histones are first hyperacetylated and then lose contact with the activated PHO5 promoter. *Mol Cell* **11**(6): 1599-1607.

Ren, B., Cam, H., Takahashi, Y., Volkert, T., Terragni, J., Young, R.A., and Dynlacht, B.D. 2002. E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev* **16**(2): 245-256.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T.L., Wilson, C.J., Bell, S.P., and Young, R.A. 2000. Genome-wide location and function of DNA binding proteins. *Science* **290**(5500): 2306-2309.

Richards, E.J. and Elgin, S.C. 2002. Epigenetic codes for heterochromatin formation and silencing: rounding up the usual suspects. *Cell* **108**(4): 489-500.

Rinn, J.L., Euskirchen, G., Bertone, P., Martone, R., Luscombe, N.M., Hartman, S., Harrison, P.M., Nelson, F.K., Miller, P., Gerstein, M., Weissman, S., and Snyder, M. 2003. The transcriptional activity of human Chromosome 22. *Genes Dev* **17**(4): 529-540.

Risau, W. and Flamme, I. 1995. Vasculogenesis. *Annu Rev Cell Dev Biol* **11**: 73-91.

Robb, L., Elwood, N.J., Elefanty, A.G., Kontgen, F., Li, R., Barnett, L.D., and Begley, C.G. 1996. The scl gene product is required for the generation of all hematopoietic lineages in the adult mouse. *Embo J* **15**(16): 4123-4129.

Robb, L., Lyons, I., Li, R., Hartley, L., Kontgen, F., Harvey, R.P., Metcalf, D., and Begley, C.G. 1995. Absence of yolk sac hematopoiesis from mice with a targeted disruption of the scl gene. *Proc Natl Acad Sci U S A* **92**(15): 7075-7079.

Robertson, S.M., Kennedy, M., Shannon, J.M., and Keller, G. 2000. A transitional stage in the commitment of mesoderm to hematopoiesis requiring the transcription factor SCL/tal-1. *Development* **127**(11): 2447-2459.

Robyr, D., Suka, Y., Xenarios, I., Kurdistani, S.K., Wang, A., Suka, N., and Grunstein, M. 2002. Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases. *Cell* **109**(4): 437-446.

Rogakou, E.P., Pilch, D.R., Orr, A.H., Ivanova, V.S., and Bonner, W.M. 1998. DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. *J Biol Chem* **273**(10): 5858-5868.

Roh, T.Y., Cuddapah, S., and Zhao, K. 2005. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev* **19**(5): 542-552.

Roh, T.Y., Ngau, W.C., Cui, K., Landsman, D., and Zhao, K. 2004. High-resolution genome-wide mapping of histone modifications. *Nat Biotechnol* **22**(8): 1013-1016.

Roth, S.Y., Denu, J.M., and Allis, C.D. 2001. Histone acetyltransferases. *Annu Rev Biochem* **70**: 81-120.

Ruppert, S. and Tjian, R. 1995. Human TAFII250 interacts with RAP74: implications for RNA polymerase II initiation. *Genes Dev* **9**(22): 2747-2755.

Sabin, F. 1920. Studies on the origin of blood vessels and of red blood corpuscles as seen in the living blastoderm of chicks during the second day of incubation. *Carnegie Inst Wash Pub no 272, Contrib Embryol* **9**: 214-262.

Sanchez, M., Gottgens, B., Sinclair, A.M., Stanley, M., Begley, C.G., Hunter, S., and Green, A.R. 1999. An SCL 3' enhancer targets developing endothelium together with embryonic and adult haematopoietic progenitors. *Development* **126**(17): 3891-3904.

Sanchez, M.J., Bockamp, E.O., Miller, J., Gambardella, L., and Green, A.R. 2001. Selective rescue of early haematopoietic progenitors in Scl(-/-) mice by expressing Scl under the control of a stem cell enhancer. *Development* **128**(23): 4815-4827.

Santos-Rosa, H., Schneider, R., Bannister, A.J., Sherriff, J., Bernstein, B.E., Emre, N.C., Schreiber, S.L., Mellor, J., and Kouzarides, T. 2002. Active genes are tri-methylated at K4 of histone H3. *Nature* **419**(6905): 407-411.

Sassone-Corsi, P., Mizzen, C.A., Cheung, P., Crosio, C., Monaco, L., Jacquot, S., Hanauer, A., and Allis, C.D. 1999. Requirement of Rsk-2 for epidermal growth factor-activated phosphorylation of histone H3. *Science* **285**(5429): 886-891.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**(5235): 467-470.

Schild, C., Claret, F.X., Wahli, W., and Wolffe, A.P. 1993. A nucleosome-dependent static loop potentiates estrogen-regulated transcription from the Xenopus vitellogenin B1 promoter in vitro. *Embo J* **12**(2): 423-433.

Schneider, R., Bannister, A.J., Myers, F.A., Thorne, A.W., Crane-Robinson, C., and Kouzarides, T. 2004. Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. *Nat Cell Biol* **6**(1): 73-77.

Schneider, U., Schwenk, H.U., and Bornkamm, G. 1977. Characterization of EBV-genome negative "null" and "T" cell lines derived from children with acute lymphoblastic leukemia and leukemic transformed non-Hodgkin lymphoma. *Int J Cancer* **19**(5): 621-626.

Schnitzler, G., Sif, S., and Kingston, R.E. 1998. Human SWI/SNF interconverts a nucleosome between its base state and a stable remodeled state. *Cell* **94**(1): 17-27.

Schnitzler, G.R., Cheung, C.L., Hafner, J.H., Saurin, A.J., Kingston, R.E., and Lieber, C.M. 2001. Direct imaging of human SWI/SNF-remodeled mono- and polynucleosomes by atomic force microscopy employing carbon nanotube tips. *Mol Cell Biol* **21**(24): 8504-8511.

Schubeler, D., MacAlpine, D.M., Scalzo, D., Wirbelauer, C., Kooperberg, C., van Leeuwen, F., Gottschling, D.E., O'Neill, L.P., Turner, B.M., Delrow, J., Bell, S.P., and Groudine, M. 2004. The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev* **18**(11): 1263-1271.

Schubeler, D., Scalzo, D., Kooperberg, C., van Steensel, B., Delrow, J., and Groudine, M. 2002. Genome-wide DNA replication profile for Drosophila melanogaster: a link between transcription and replication timing. *Nat Genet* **32**(3): 438-442.

Schug, J. and Overton, G.C. 1997. Modeling transcription factor binding sites with Gibbs Sampling and Minimum Description Length encoding. *Proc Int Conf Intell Syst Mol Biol* **5**: 268-271.

Schuh, A.C., Faloon, P., Hu, Q.L., Bhimani, M., and Choi, K. 1999. In vitro hematopoietic and endothelial potential of flk-1(-/-) embryonic stem cells and embryos. *Proc Natl Acad Sci U S A* **96**(5): 2159-2164.

Schuler, G.D. 1997. Sequence mapping by electronic PCR. *Genome Res* **7**(5): 541-550.

Schwartz, S., Elnitski, L., Li, M., Weirauch, M., Riemer, C., Smit, A., Green, E.D., Hardison, R.C., and Miller, W. 2003. MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res* **31**(13): 3518-3524.

Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. 2000. PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res* **10**(4): 577-586.

Shabalina, S.A. and Spiridonov, N.A. 2004. The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biol* **5**(4): 105.

Shaffer, L.G. and Bejjani, B.A. 2004. A cytogeneticist's perspective on genomic microarrays. *Hum Reprod Update* **10**(3): 221-226.

Shalaby, F., Ho, J., Stanford, W.L., Fischer, K.D., Schuh, A.C., Schwartz, L., Bernstein, A., and Rossant, J. 1997. A requirement for Flk1 in primitive and definitive hematopoiesis and vasculogenesis. *Cell* **89**(6): 981-990.

Shalaby, F., Rossant, J., Yamaguchi, T.P., Gertsenstein, M., Wu, X.F., Breitman, M.L., and Schuh, A.C. 1995. Failure of blood-island formation and vasculogenesis in Flk-1-deficient mice. *Nature* **376**(6535): 62-66.

Shi, Y., Lan, F., Matson, C., Mulligan, P., Whetstine, J.R., Cole, P.A., Casero, R.A., and Shi, Y. 2004. Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell* **119**(7): 941-953.

Shimamoto, T., Ohyashiki, K., Ohyashiki, J.H., Kawakubo, K., Fujimura, T., Iwama, H., Nakazawa, S., and Toyama, K. 1995. The expression pattern of erythrocyte/megakaryocyte-related transcription factors GATA-1 and the stem cell leukemia gene correlates with hematopoietic differentiation and is associated with outcome of acute myeloid leukemia. *Blood* **86**(8): 3173-3180.

Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S., Ray, T.S., Koval, M.A., Last, K.W., Norton, A., Lister, T.A., Mesirov, J., Neuberg, D.S., Lander, E.S., Aster, J.C., and Golub, T.R. 2002. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med* **8**(1): 68-74.

Shivdasani, R.A., Mayer, E.L., and Orkin, S.H. 1995. Absence of blood formation in mice lacking the T-cell leukaemia oncoprotein tal-1/SCL. *Nature* **373**(6513): 432-434.

Silver, L. and Palis, J. 1997. Initiation of murine embryonic erythropoiesis: a spatial analysis. *Blood* **89**(4): 1154-1164.

Simon, I., Barnett, J., Hannett, N., Harbison, C.T., Rinaldi, N.J., Volkert, T.L., Wyrick, J.J., Zeitlinger, J., Gifford, D.K., Jaakkola, T.S., and Young, R.A. 2001. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* **106**(6): 697-708.

Sinclair, A.M., Gottgens, B., Barton, L.M., Stanley, M.L., Pardanaud, L., Klaine, M., Gering, M., Bahn, S., Sanchez, M., Bench, A.J., Fordham, J.L., Bockamp, E., and Green, A.R. 1999. Distinct 5' SCL enhancers direct transcription to developing brain, spinal cord, and endothelium: neural expression is mediated by GATA factor binding sites. *Dev Biol* **209**(1): 128-142.

Singh-Gasson, S., Green, R.D., Yue, Y., Nelson, C., Blattner, F., Sussman, M.R., and Cerrina, F. 1999. Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat Biotechnol* **17**(10): 974-978.

Smale, S.T. and Kadonaga, J.T. 2003. The RNA polymerase II core promoter. *Annu Rev Biochem* **72**: 449-479.

Snow, K. and Judd, W. 1987. Heterogeneity of a human T-lymphoblastoid cell line. *Exp Cell Res* **171**(2): 389-403.

Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Dohner, H., Cremer, T., and Lichter, P. 1997. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer* **20**(4): 399-407.

Solomon, M.J. and Varshavsky, A. 1985. Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. *Proc Natl Acad Sci U S A* **82**(19): 6470-6474.

Stainier, D.Y., Weinstein, B.M., Detrich, H.W., 3rd, Zon, L.I., and Fishman, M.C. 1995. Cloche, an early acting zebrafish gene, is required by both the endothelial and hematopoietic lineages. *Development* **121**(10): 3141-3150.

Strahl, B.D., Briggs, S.D., Brame, C.J., Caldwell, J.A., Koh, S.S., Ma, H., Cook, R.G., Shabanowitz, J., Hunt, D.F., Stallcup, M.R., and Allis, C.D. 2001. Methylation of histone H4 at arginine 3 occurs in vivo and is mediated by the nuclear receptor coactivator PRMT1. *Curr Biol* **11**(12): 996-1000.

Strahl, B.D., Ohba, R., Cook, R.G., and Allis, C.D. 1999. Methylation of histone H3 at lysine 4 is highly conserved and correlates with transcriptionally active nuclei in Tetrahymena. *Proc Natl Acad Sci U S A* **96**(26): 14967-14972.

Studitsky, V.M., Clark, D.J., and Felsenfeld, G. 1994. A histone octamer can step around a transcribing polymerase without leaving the template. *Cell* **76**(2): 371-382.

Suka, N., Suka, Y., Carmen, A.A., Wu, J., and Grunstein, M. 2001. Highly specific antibodies determine histone acetylation site usage in yeast heterochromatin and euchromatin. *Mol Cell* **8**(2): 473-479.

Sun, Z.W. and Allis, C.D. 2002. Ubiquitination of histone H2B regulates H3 methylation and gene silencing in yeast. *Nature* **418**(6893): 104-108.

Suto, R.K., Clarkson, M.J., Tremethick, D.J., and Luger, K. 2000. Crystal structure of a nucleosome core particle containing the variant histone H2A.Z. *Nat Struct Biol* **7**(12): 1121-1124.

Szutorisz, H., Canzonetta, C., Georgiou, A., Chow, C.M., Tora, L., and Dillon, N. 2005. Formation of an active tissue-specific chromatin domain initiated by epigenetic marking at the embryonic stem cell stage. *Mol Cell Biol* **25**(5): 1804-1820.

Tagle, D.A., Koop, B.F., Goodman, M., Slightom, J.L., Hess, D.L., and Jones, R.T. 1988. Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol* **203**(2): 439-455.

Tanigawa, T., Elwood, N., Metcalf, D., Cary, D., DeLuca, E., Nicola, N.A., and Begley, C.G. 1993. The SCL gene product is regulated by and differentially regulates cytokine responses during myeloid leukemic cell differentiation. *Proc Natl Acad Sci U S A* **90**(16): 7864-7868.

Tanigawa, T., Nicola, N., McArthur, G.A., Strasser, A., and Begley, C.G. 1995. Differential regulation of macrophage differentiation in response to leukemia inhibitory factor/oncostatin-M/interleukin-6: the effect of enforced expression of the SCL transcription factor. *Blood* **85**(2): 379-390.

Tavian, M., Coulombel, L., Luton, D., Clemente, H.S., Dieterlen-Lievre, F., and Peault, B. 1996. Aorta-associated CD34+ hematopoietic cells in the early human embryo. *Blood* **87**(1): 67-72.

Thiesen, H.J. and Bach, C. 1990. Target Detection Assay (TDA): a versatile procedure to determine DNA binding sites as demonstrated on SP1 protein. *Nucleic Acids Res* **18**(11): 3203-3209.

Thomson, S., Clayton, A.L., Hazzalin, C.A., Rose, S., Barratt, M.J., and Mahadevan, L.C. 1999. The nucleosomal response associated with immediate-early gene induction is mediated via alternative MAP kinase cascades: MSK1 as a potential histone H3/HMG-14 kinase. *Embo J* **18**(17): 4779-4793.

Torrano, V., Chernukhin, I., Docquier, F., D'Arcy, V., Leon, J., Klenova, E., and Delgado, M.D. 2005. CTCF regulates growth and erythroid differentiation of human myeloid leukemia cells. *J Biol Chem* **280**(30): 28152-28161.

Tse, C., Georgieva, E.I., Ruiz-Garcia, A.B., Sendra, R., and Hansen, J.C. 1998a. Gcn5p, a transcription-related histone acetyltransferase, acetylates nucleosomes and folded nucleosomal arrays in the absence of other protein subunits. *J Biol Chem* **273**(49): 32388-32392.

Tse, C., Sera, T., Wolffe, A.P., and Hansen, J.C. 1998b. Disruption of higher-order folding by core histone acetylation dramatically enhances transcription of nucleosomal arrays by RNA polymerase III. *Mol Cell Biol* **18**(8): 4629-4638.

Tuan, D., Solomon, W., Li, Q., and London, I.M. 1985. The "beta-like-globin" gene domain in human erythroid cells. *Proc Natl Acad Sci U S A* **82**(19): 6384-6388.

Tuerk, C. and Gold, L. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**(4968): 505-510.

Turpen, J.B., Kelley, C.M., Mead, P.E., and Zon, L.I. 1997. Bipotential primitive-definitive hematopoietic progenitors in the vertebrate embryo. *Immunity* **7**(3): 325-334.

van Dijk, K., Marley, K.E., Jeong, B.R., Xu, J., Hesson, J., Cerny, R.L., Waterborg, J.H., and Cerutti, H. 2005. Monomethyl histone H3 lysine 4 as an epigenetic mark for silenced euchromatin in Chlamydomonas. *Plant Cell* **17**(9): 2439-2453.

Van Hooser, A., Goodrich, D.W., Allis, C.D., Brinkley, B.R., and Mancini, M.A. 1998. Histone H3 phosphorylation is required for the initiation, but not maintenance, of mammalian chromosome condensation. *J Cell Sci* **111 (Pt 23)**: 3497-3506.

van Leeuwen, F., Gafken, P.R., and Gottschling, D.E. 2002. Dot1p modulates silencing in yeast by methylation of the nucleosome core. *Cell* **109**(6): 745-756.

Vicent, G.P., Nacht, A.S., Smith, C.L., Peterson, C.L., Dimitrov, S., and Beato, M. 2004. DNA instructed displacement of histones H2A and H2B at an inducible promoter. *Mol Cell* **16**(3): 439-452.

Visvader, J. and Begley, C.G. 1991. Helix-loop-helix genes translocated in lymphoid leukemia. *Trends Biochem Sci* **16**(9): 330-333.

Visvader, J.E., Fujiwara, Y., and Orkin, S.H. 1998. Unsuspected role for the T-cell leukemia protein SCL/tal-1 in vascular development. *Genes Dev* **12**(4): 473-479.

Vitelli, L., Condorelli, G., Lulli, V., Hoang, T., Luchetti, L., Croce, C.M., and Peschle, C. 2000. A pentamer transcriptional complex including tal-1 and retinoblastoma protein downmodulates c-kit expression in normal erythroblasts. *Mol Cell Biol* **20**(14): 5330-5342.

Vogelauer, M., Wu, J., Suka, N., and Grunstein, M. 2000. Global histone acetylation and deacetylation in yeast. *Nature* **408**(6811): 495-498.

Wadman, I.A., Osada, H., Grutz, G.G., Agulnick, A.D., Westphal, H., Forster, A., and Rabbitts, T.H. 1997. The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. *Embo J* **16**(11): 3145-3157.

Wall, L., deBoer, E., and Grosveld, F. 1988. The human beta-globin gene 3' enhancer contains multiple binding sites for an erythroid-specific protein. *Genes Dev* **2**(9): 1089-1100.

Wang, C.Y., Petryniak, B., Thompson, C.B., Kaelin, W.G., and Leiden, J.M. 1993. Regulation of the Ets-related transcription factor Elf-1 by binding to the retinoblastoma protein. *Science* **260**(5112): 1330-1335.

Wang, Y., Wysocka, J., Sayegh, J., Lee, Y.H., Perlin, J.R., Leonelli, L., Sonbuchner, L.S., McDonald, C.H., Cook, R.G., Dou, Y., Roeder, R.G., Clarke, S., Stallcup, M.R., Allis, C.D., and Coonrod, S.A. 2004. Human PAD4 regulates histone arginine methylation levels via demethylimination. *Science* **306**(5694): 279-283.

Waterston, R.H. et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**(6915): 520-562.

Watt, S.M., Gschmeissner, S.E., and Bates, P.A. 1995. PECAM-1: its expression and function as a cell adhesion molecule on hemopoietic and endothelial cells. *Leuk Lymphoma* **17**(3-4): 229-244.

Weber, F., de Villiers, J., and Schaffner, W. 1984. An SV40 "enhancer trap" incorporates exogenous enhancers or generates enhancers from its own sequences. *Cell* **36**(4): 983-992.

Weinmann, A.S., Bartley, S.M., Zhang, T., Zhang, M.Q., and Farnham, P.J. 2001. Use of chromatin immunoprecipitation to clone novel E2F target promoters. *Mol Cell Biol* **21**(20): 6820-6832.

Weinmann, A.S., Yan, P.S., Oberley, M.J., Huang, T.H., and Farnham, P.J. 2002. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev* **16**(2): 235-244.

Wells, J., Graveel, C.R., Bartley, S.M., Madore, S.J., and Farnham, P.J. 2002. The identification of E2F1-specific target genes. *Proc Natl Acad Sci U S A* **99**(6): 3890-3895.

Wells, J., Yan, P.S., Cechvala, M., Huang, T., and Farnham, P.J. 2003. Identification of novel pRb binding sites using CpG microarrays suggests that E2F recruits pRb to specific genomic sites during S phase. *Oncogene* **22**(10): 1445-1460.

West, A.G., Gaszner, M., and Felsenfeld, G. 2002. Insulators: many functions, many mechanisms. *Genes Dev* **16**(3): 271-288.

West, A.G., Huang, S., Gaszner, M., Litt, M.D., and Felsenfeld, G. 2004. Recruitment of histone modifications by USF proteins at a vertebrate barrier element. *Mol Cell* **16**(3): 453-463.

White, E.J., Emanuelsson, O., Scalzo, D., Royce, T., Kosak, S., Oakeley, E.J., Weissman, S., Gerstein, M., Groudine, M., Snyder, M., and Schubeler, D. 2004. DNA replication-timing analysis of human chromosome 22 at high resolution and different developmental states. *Proc Natl Acad Sci U S A* **101**(51): 17771-17776.

White, K.P., Rifkin, S.A., Hurban, P., and Hogness, D.S. 1999. Microarray analysis of Drosophila development during metamorphosis. *Science* **286**(5447): 2179-2184.

Whitehouse, I., Flaus, A., Cairns, B.R., White, M.F., Workman, J.L., and Owen-Hughes, T. 1999. Nucleosome mobilization catalysed by the yeast SWI/SNF complex. *Nature* **400**(6746): 784-787.

Willingham, A.T., Orth, A.P., Batalov, S., Peters, E.C., Wen, B.G., Aza-Blanc, P., Hogenesch, J.B., and Schultz, P.G. 2005. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* **309**(5740): 1570-1573.

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Pruss, M., Reuter, I., and Schacherer, F. 2000. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* **28**(1): 316-319.

Wirbelauer, C., Bell, O., and Schubeler, D. 2005. Variant histone H3.3 is deposited at sites of nucleosomal displacement throughout transcribed genes while active histone modifications show a promoter-proximal bias. *Genes Dev* **19**(15): 1761-1766.

Woodfine, K., Fiegler, H., Beare, D.M., Collins, J.E., McCann, O.T., Young, B.D., Debernardi, S., Mott, R., Dunham, I., and Carter, N.P. 2004. Replication timing of the human genome. *Hum Mol Genet* **13**(2): 191-202.

Workman, J.L. and Kingston, R.E. 1998. Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annu Rev Biochem* **67**: 545-579.

Wright, W.E., Binder, M., and Funk, W. 1991. Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site. *Mol Cell Biol* **11**(8): 4104-4110.

Wu, C., Wong, Y.C., and Elgin, S.C. 1979. The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity. *Cell* **16**(4): 807-814.

Wu, J. and Grunstein, M. 2000. 25 years after the nucleosome model: chromatin modifications. *Trends Biochem Sci* **25**(12): 619-623.

Wyrick, J.J., Aparicio, J.G., Chen, T., Barnett, J.D., Jennings, E.G., Young, R.A., Bell, S.P., and Aparicio, O.M. 2001. Genome-wide distribution of ORC and MCM proteins in S. cerevisiae: high-resolution mapping of replication origins. *Science* **294**(5550): 2357-2360.

Xia, Y., Brown, L., Tsan, J.T., Yang, C.Y., Siciliano, M.J., Crist, W.M., Carroll, A.J., and Baer, R. 1992. The translocation (1;14)(p34;q11) in human T-cell leukemia: chromosome breakage 25 kilobase pairs downstream of the TAL1 protooncogene. *Genes Chromosomes Cancer* **4**(3): 211-216.

Xiong, Z., Tsark, W., Singer-Sam, J., and Riggs, A.D. 1998. Differential replication timing of X-linked genes measured by a novel method using single-nucleotide primer extension. *Nucleic Acids Res* **26**(2): 684-686.

Xu, H.M., Zhang, S., Liu, D.P., Li, X.G., Hao, D.L., and Liang, C.C. 2002. Efficient isolation of regulatory sequences from human genome and BAC DNA. *Biochem Biophys Res Commun* **290**(3): 1079-1083.

Yeh, R.F., Lim, L.P., and Burge, C.B. 2001. Computational inference of homologous gene structures in the human genome. *Genome Res* **11**(5): 803-816.

Yelin, R., Dahary, D., Sorek, R., Levanon, E.Y., Goldstein, O., Shoshan, A., Diber, A., Biton, S., Tamir, Y., Khosravi, R., Nemzer, S., Pinner, E., Walach, S., Bernstein, J., Savitsky, K., and Rotman, G. 2003. Widespread occurrence of antisense transcription in the human genome. *Nat Biotechnol* **21**(4): 379-386.

Young, P.E., Baumhueter, S., and Lasky, L.A. 1995. The sialomucin CD34 is expressed on hematopoietic cells and blood vessels during murine development. *Blood* **85**(1): 96-105.

Yuan, G.C., Liu, Y.J., Dion, M.F., Slack, M.D., Wu, L.F., Altschuler, S.J., and Rando, O.J. 2005. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science* **309**(5734): 626-630.

Zhang, Y. 2003. Transcriptional regulation by histone ubiquitination and deubiquitination. *Genes Dev* **17**(22): 2733-2740.

Zhang, Y. 2004. Molecular biology: no exception to reversibility. *Nature* **431**(7009): 637-639.

Zhang, Y., Payne, K.J., Zhu, Y., Price, M.A., Parrish, Y.K., Zielinska, E., Barsky, L.W., and Crooks, G.M. 2005. SCL expression at critical points in human hematopoietic lineage commitment. *Stem Cells* **23**(6): 852-860.

Zhao, H. and Dean, A. 2004. An insulator blocks spreading of histone acetylation and interferes with RNA polymerase II transfer between an enhancer and gene. *Nucleic Acids Res* **32**(16): 4903-4919.

Appendix 1

Sequences of Primer Pairs Used to Construct the Human SCL Tiling Path Array

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 1 Co-ordinate Start	Chrom 1 Co-ordinate Finish
HSTAL.1	CATGATATTCATCCCCTGCC	TTCTGAATGATGACCATGCC	499	47262288	47262786
HSTAL.2	TTCAGCTCTCAAGGCCAAAT	TCCTAAGTCCCCAAAAACCC	504	47262613	47263116
HSTAL.3	AATGACAAGTCAAGGGGGAA	AAAAATGGCCTAAGCCTGAA	457	47267138	47267594
HSTAL.4	ATTTGAGCACCAGTCTCTACCA	CCTTCCGGTCCTGGATTACT	459	47271823	47272281
HSSCL/M183A	CCAGCGCCTTTTCTGAGTAGTAT	TACTGTCCCTTACCAACTCAGGAAC	218	47272114	47272331
HSTAL.6	GTCTTTTTCCAAGCCCACAA	AGGTTGAAGATGCTTGGCAC	507	47272825	47273331
HSTAL.7	GCCAAGCATCTTCAACCTTT	ATTTTGCAGTGCCCTGTTCT	447	47273314	47273760
HSSCL/M182A	TTTGCAGTGCCCTGTTCTTAGTA	ATCTCATGCCATTTCCGTTGTAC	284	47273475	47273758
HSTAL.8	CTAAAACTGCCTGCAATCCC	CAACAGAATTCAACTGGGCA	417	47276178	47276594
HSTAL.9	TCTGAAGCAGATCTCCAGGC	GGCAGACTATTCTTTCCGCA	439	47276807	47277245
HSSCL/M178A	GTGATAAGGAAGAGGGGTGTGAAT	AGCAGAGATGCCGAGATGAAAT	488	47276919	47277406
HSTAL.10	GTCACAGGAGCCCTGGAATA	CCTCTTTCTCCCTCTCTGGG	539	47283559	47284097
HSTAL.11	ATCACCTTTCCAGATGGACG	GCTACATCATGGCTCCAGGT	511	47283920	47284430
HSSCL/M171A	TGTTGCCTTTCAATCCATCTTG	TGAGGGGGCCTTCGTGAA	405	47284183	47284587
HSTAL.12	CATTGGTGGAATGCAAGATG	GCAAAGATCCCCATGATAGC	559	47284553	47285111
HSTAL.13	CACATAATATTCCGCCCACC	GCCTAAAACTGAGTCAGCCG	452	47285213	47285664
HSSCL/M170A	GAAAGCTCTGGAGGGCAAGAA	TCCCCCGACCCATACAAAA	399	47285443	47285841
HSTAL.14	CTTTACATTCCTTTCCGCCA	AGCAAATGAAGAGAAGGGCA	561	47285797	47286357
HSTAL.15	CCTTCTCTTCATTTGCTGGC	AAGTGGCAACAGAATCACCC	467	47286341	47286807
HSSCL/M168B	TGCAGCCACCCTCTCATTG	TCTCAGTTATTGCCTTTGTCAATCT	383	47286641	47287023
HSTAL.16	ATTCTGTTGCCACTTGTCCC	TGGACATCCCCTTTTCTCTG	470	47286793	47287262
HSSCL/M168A3	AAAAAAGCATGCACATTATCTAACAAA	AAGGGGATGTCCACTCTGCTG	494	47287249	47287742
HSSCL/M168A	TTATTTAGGGGATTTCATGGAACAC	ACAATGCAGATCCCTTTTCACTAA	625	47287327	47287951
HSTAL.18	TGGCCTAGTATGTGCCATGA	CAGAACAGACTGCCCATTGA	516	47287855	47288370
HSSCL/M167A	AATGGCAACTGGAACACCATAAC	GAGACTCACCCAAAGGAGCTAATAA	447	47288277	47288723
HSTAL.20	TTGCCACTAAAGCAGCATTC	TCCTTTTCAGGGGCTCTACA	439	47289636	47290074
HSSCL/M165B	TTTTCAGGGGCTCTACAATTTTTC	ATCAGTTGGGGAAAGAATAAAAGAG	324	47289748	47290071
HSTAL.21	TGTAGAGCCCCTGAAAAGGA	CAGGGTGGAGATGATATGGG	521	47290055	47290575

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 1 Co-ordinate Start	Chrom 1 Co-ordinate Finish
HSTAL.22	CAGAATCTTCGCTCATGCAC	GAAAGCAGAGAGCAAGGTGG	400	47293173	47293572
HSSCL/M161A	CCCAAAAGCCTACCACTGAAAA	TGGGAAGTAGAGGGGTGTGATT	675	47293506	47294180
HSTAL.24	CGGGTTTACTGTGAAGGGAA	GTCAGCTAACTGAGGGCCTG	495	47294020	47294514
HSSCL/M160A	TAAGCCCCGATCAGAAAGTCC	CATGACAGGCCCAAAGACTACA	426	47294670	47295095
HSTAL.26	CCACAGTTTGCAGGTCTTGA	AAGCGGAGCAGAGTTAATGC	479	47295048	47295526
HSSCL/M159B	AACTATGCTTGAAAGCCACTCTACC	GCAGCATTTTGCCATAATTGAG	437	47295470	47295906
HSTAL.28	GCCACCTTCTGAGTCCCATA	CCCACCCTACAACCTCAAAA	479	47296007	47296485
HSSCL/M159A	AACCTCTGCGTCCTAACTTGTGA	CCCCATGCTGGTGTGTGCT	509	47296301	47296809
HSTAL.30	TCCAGCTGTCACTCATGGAT	GCATGTGGGAAAGTGATAAAAG	583	47305418	47306000
HSSIL/M64A	CACAAGCACACTTTTCTAAGAGTTACA	GCAGCCCCTAAACCTTCCAG	304	47305922	47306225
HSTAL.31	GAGCTGCCCTGGATAGTGAG	TCCCTACTTGGATCAGCACC	561	47306091	47306651
HSSIL/M65A	GGGAGGTTGGGTGGAATGAA	CTTGGGCCATTCAGAGACAGAG	477	47306728	47307204
HSSCL/M148A	GGGCCATTCAGAGACAGAGAAG	CCAGGATTTGAGGCAGAAAGAA	410	47306792	47307201
HSTAL.33	GGATTAGGGAGCAGGAGGAA	GCCATAGAAAACTAAAATGGGG	572	47306900	47307471
HSSCL/M141A	CCAAGCCCTGAATCATTACTTTT	GCCAGGGAGGGTTAATTCATAG	558	47314179	47314736
HSTAL.35	CCAGTCTCCCCAATGGAATA	GGGAGAACAGCTTCATCAGC	448	47314549	47314996
HSSCL/M140A	CTCCCCCACTGAGTTCCTCC	ATGGACAGGGGAATGTGGTAGA	444	47314953	47315396
HSTAL.37	ATCCAGGAGGTAGGGAGGAA	ACTTAGCCAGTTGGTCCCCT	569	47315364	47315932
HSSCL/M139A	ACAGGAAGGGGATGACATGCT	GGAGGGCAATGCCTGATTG	521	47315634	47316154
HSTAL.38	CCTCCAAGTGAGGAGACTGG	GATAGCAGTGCACAGGGGAT	485	47316032	47316516
HSTAL.39	ATCCCCTGTGCACTGCTATC	AAAAGAGTCGGCCCTTCACT	456	47316497	47316952
HSTAL.40	AGTGAAGGGCCGACTCTTTT	TCTAATGTCTGCCCACCTCC	486	47316933	47317418
HSTAL.41	GGAGGTGGGCAGACATTAGA	CATATGGTGCTGGTGTTGGA	500	47317399	47317898
HSSCL/M137A	AGTCAGCCATTGTCTTCATGTTGT	GCAGGCAGCAGGGAGCTCA	448	47317806	47318253
HSSCL/M136B	ATCTGGGGCAATTTACGTTTC	CCTTCACTCCTGCTGCAAAGAC	479	47318407	47318885
HSTAL.45	ACCAAGAACTGATGCTGCCT	TGACTGGTGACAATTTGGGA	557	47319185	47319741
HSSCL/M135B5	GGGCAAATGCAGGAAAAGGA	TCAGACCCGAAATCCCATGG	412	47319258	47319669
HSSCL/M135B3	CTGCACGGACCTAGGGATGA	ACAAATTCCTGGCTCCACGG	476	47319283	47319758
HSSCL/M135A5	AAGGGCCTTGTGTTCACCATG	CTCCCTCCCAAATTGTCACCA	438	47319716	47320153
HSSCL/M135A3	CTGTCCATGCTGCCACTGAAG	CCCAGCCTTCCACAATGACAT	462	47319857	47320318
HSTAL.47	AGAGGGTTCCCCAGAGTTGT	AAGGGGCTAGTATGGGCACT	468	47320067	47320534

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 1 Co-ordinate Start	Chrom 1 Co-ordinate Finish
HSSCL/M134B	GGGGGCATGGACCTTCTTAA	TGCCCAGCACATTTTGATGA	493	47320457	47320949
HSTAL.49	TCTCCTTGATGACCCTGGAC	ATGGCCTGGATGTAGGACTG	465	47321016	47321480
HSSCL/M134A	GGAGTCAGGGCAGGTACTTTCA	ACCGCCACACAAAGGAAGAG	499	47321175	47321673
HSSCL/M133A	CGTGGCCCTCAAACATGAA	AAGAGGAGCCTCCATGAACTGT	470	47321845	47322314
HSSCL/M132A	CCAGCCCTGCACATCTCAAG	GGCCTCCTGGGTGATGGA	529	47322401	47322929
HSTAL.53	ATGGGGTCAGTGCAAAAGAG	CATGATACCTGTGGTGCAGG	592	47322974	47323565
HSSCL/M131B	TCAGAGGGTTGACCAGAGACTTC	TAGGTGTTTTGGGCCTTCTTTT	448	47323461	47323908
HSSCL/M131A	ACCGCAGCCATTCTGGAGT	TCAAATGGGCAGCCTGAATT	665	47324064	47324728
HSTAL.56	TCTTGGAGAGAAAAGCCCAA	ATGTGTGTCTGGAGCACAGG	598	47324602	47325199
HSSCL/M130A	TCCAGCCCACATCCTAGCC	ATCATGTCGCTGCCACTAGAGAA	515	47324950	47325464
HSSCL/M129B	GACAGGGCCAGGGCTCTAAG	GATGCTGTCCCACCTGTGATG	503	47325526	47326028
HSSCL/M129A	GAGTGGGGGTCAGGAAGACAG	GCCAGGATGGGGTAGAAGTGT	401	47326128	47326528
HSTAL.60	TCCCTGAATCTCCACCCAC	GATGACAATCAGGTAGGGGC	591	47327280	47327870
HSSCL/M127A	CTTAGCTCAGCCTTGCTGTGATT	TCAGCCAGACAATCCCTCTATCT	407	47327924	47328330
HSTAL.62	GAGGGAGAAGCTCACACCTG	CTTTGAGGCTTTCCTGCTCA	576	47328237	47328812
HSSCL/M126B	GCCCAGGGTCAGCAGTACC	GAGGCTGCACTGAGCTGAGC	409	47328778	47329186
HSTAL.65	GATCCCTCTCTCCTTGGGTT	GTTGCTGGATGTTTCCTGGT	493	47338055	47338547
HSSCL/M116A	GATGGCAGCCCCTAAACTTTC	CATCATCAACTGGCACAGTCCTC	446	47338682	47339127
HSTAL.67	AGTCCCACAGTTTGCCATTC	TCAGCACCTGTATTTCACGC	593	47338958	47339550
HSSIL/M53A	GGGAGGGAAAGCAAAACTCAC	CAAGGGCACCAATTCAGAAACT	467	47339446	47339912
HSSCL/M115A	GCCACCCAGCAAAGAAGAAA	TGTCCCTACCCTTTGGCTTAGA	439	47339676	47340114
HSSIL/M53B	TTGGTGCCCTTGGCTATGAC	ATGTCACAGTTCAGTTGCTCAAGG	305	47339901	47340205
HSSCL/M113A	GCAGGACCATTTGCCTATTCTT	CTCATTTTCATGTTTGTAGTGGAATAGT	298	47342174	47342471
HSSIL/M55A	ATGATACAATGAAGATTGAAATGTGACA	TTCATCCTAACTGAAGGAGAATGATAAC	325	47342204	47342528
HSSIL/M57A	GTGGCCATGATGGGAAAATTA	TTCCCTTATCTATGCCATATCCATTA	282	47343276	47343557
HSSCL/M105A	TCGGGGTCACTGGGAAATG	CTAAAGAGGGTCAAGTTATCTTCTGG	367	47350069	47350435
HSSCL/M104A	TGATCGTCCCATGAATGTGAAGT	GCCATCAGAGCTTCCATTTTAAT	471	47351208	47351678
HSTAL.71	AGTCCATAACACCAGGGCAG	AAACAGTGTGTGTGGAGGGG	575	47351603	47352177
HSSCL/M103A	GGGGGTAAGGGCACTTGTTT	ATACAGAGAACAGGAACTGGTATTACATAT	391	47352054	47352444
HSSCL/M102A	CGGGTCTCAGGATCTCCTTTT	ACAGCCAGCCAGGATGCTT	418	47352494	47352911
HSTAL.73	GACAATGGCTAGTGGGCAAT	CAGCCATCGTGATTTTCCTT	504	47352766	47353269

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 1 Co-ordinate Start	Chrom 1 Co-ordinate Finish
HSTAL.74	TCCTCTCGCAGATGTGAATG	TTTGTGAGGCAAAGACCACA	508	47353496	47354003
HSSCL/M101A	ACCACCTGCCCTCAGCTCAT	TGCAGGAGGGAGTGGAGCT	271	47353827	47354097
HSTAL.76	GAAAACAGCACCACCACCAT	CAGAGTGAACTGAATTGGATGC	509	47355636	47356144
HSTAL.77	ACATTCTGTACCTGCCAGCC	GAGTTCTGAAACAGCTGGGC	525	47355951	47356475
HSSCL/M98A	TGCCGATTGGTTCAAATTCC	CGGGACAGATGGCTGGAGT	498	47357176	47357673
HSTAL.80	CAATTTTACCGGCAAAGGAA	GAGTTCCGTCTCCAGAGTGC	513	47357585	47358097
HSSCL/M97A	GTGCCCACCTCCTCTTTCTCT	ACGGAACTCCACAGCCTGTTT	491	47358089	47358579
HSTAL.82	AAGGAAGCTTAGGAGGCAGG	CTGGAGAAGGCTGATAACGC	540	47358481	47359020
HSSCL/M96B	GCCCCCTTCTCATTCTGTCTT	CTCCAGCCCCGCCCTAGT	447	47358901	47359347
HSSCL/M96A	CGTGGCTGCCTGTGACTCA	TGCTGGGGAGCAAGGTGA	491	47359379	47359869
HSTAL.84	GAACCTTGGTTGGGTCAAGA	TCACCTCTGTGACTGTTGGC	537	47359783	47360319
HSSCL/M95A	ACGGCAGAGCTGGAGGAGA	AAGCAAGGGGGCTGTGACA	351	47360136	47360486
HSTAL.85	ACAGTCACAGAGGTGAGGGG	CCACTCTACCCTTGCTCCAG	523	47360304	47360826
HSSCL/M94B	GCCAGGGTGAGGGTAAGAGAG	GGGCTACTGGAGGGAGACAGA	485	47360654	47361138
HSSCL/M94A	AGCACCCCGATGTAACCTTCT	AGGGGCTTCCTGCATAATCC	479	47361199	47361677
HSSCL/M92A	TTCAGGGAGAGCCCACACTG	CCCCCTCGTCCAAGCAGA	427	47362829	47363255
HSTAL.92	ACCAAGACTCCAAGAAGCGA	AGGTGGGGTCCTTGAAGAGT	490	47364722	47365211
HSSCL/M90B	CCTGCCTCCCCACAATGTC	CCCACCCTGCCTGCTCAG	296	47364972	47365267
HSTAL.93	ACTCTTCAAGGACCCCACCT	TGTCCTGGTCCACTCCTCTC	515	47365192	47365706
HSSCL/M90A	TGGGGCCTCAACCTAATCTCT	GGCAGCCCAGAGCCTCTTT	427	47365419	47365845
HSTAL.94	GAGAGGAGTGGACCAGGACA	AGAGGAAAGTCCTTTGGGGA	409	47365687	47366095
HSTAL.95	AAGGACTTTCCTCTTCCCCA	ACCTGAGGGAGCAGTCTGAA	470	47366082	47366551
HSSCL/M88B	CAAGGTCTGGCTTTCATTCTACAC	AAGCAGTATTTGGGAGTCATTGTTT	467	47366498	47366964
HSTAL.96	CCCTGTTTCCTCCTAGACCC	GAGAACAGTGCCCAGTGTGA	578	47366602	47367179
HSSCL/M88A	AGACTGAGCCCAGGTTGATTTC	TCACACTGGGCACTGTTCTCTATC	569	47367160	47367728
HSTAL.98	AGTCTATTGGTGTCCGCCTG	AGGCACTGAAGCCGTAGTGT	574	47367688	47368261
HSSCL/M87A	TTCAGGCCCTCTATTGAGATCTTT	GCTGGCATTCGAGGTCATCT	463	47367872	47368334
HSTAL.99	ACACTACGGCTTCAGTGCCT	AGGAACTCTGGCCATTCTCA	536	47368242	47368777
HSSCL/M86A	CCTGGGGACCTTCCATCTCT	TCTTGGCCCCAGATGTGGT	505	47368414	47368918
HSTAL.100	GAGAATGGCCAGAGTTCCTG	GCATCACTCAACTCCAAGCA	485	47368759	47369243
HSSCL/M85A	TGTGGGGCTGCTGAGACCT	CCTCCAACCAGTGCCCTCA	411	47370130	47370540

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 1 Co-ordinate Start	Chrom 1 Co-ordinate Finish
HSSCL/M84A	GCTCTTTGCCTGGATTCCTCT	AACCATCAGGCCCCTTAGACA	457	47370620	47371076
HSTAL.103	GCAGGCAAAGTTAGCCAGAG	TTTAAGGGCTGATGAATGCC	555	47371010	47371564
HSTAL.104	GCAAACTTCTTTGCATGGGT	GTCCTTGCCCCACAGATAGA	581	47371572	47372152
HSTAL.106	GATCAATCAACCCATCCTGC	TTTCCAGTCCCATAGTAGCCA	471	47374482	47374952
HSSCL/M78A	GGCACCTCCTTGATTGTACTCTTT	TGGGTGCAGCCGTGGAG	520	47376719	47377238
HSTAL.108	GCTGCATAGACGCAGAAGTG	TGCTCACTGCACCCACTAAC	593	47377186	47377778
HSTAL.109	GGAGCTCACAGGTTAGTGGG	GACTCCACATGCCCTGAAAT	403	47377748	47378150
HSTAL.110	ATTTCAGGGCATGTGGAGTC	GCTTCTTGACTGAGCAAGGG	445	47378131	47378575
HSSCL/M76A	CTTCTCTCGCCACACTGTGCT	CCCTCGGACCCTTTCACACT	559	47378586	47379144
HSTAL.112	GTAATTGGAGGCTTCCCCTC	GACTCCTTTCCTGGCATTCA	440	47379037	47379476
HSTAL.114	TACAAGTCACATCCCTCCCC	TCCATCCTCTTCTCCCTGAA	485	47385593	47386077
HSSCL/M69A	GGCTGGCAGGCTGTGTGA	GGGCAGGATGGGACCATTAG	634	47385726	47386359
HSSCL/M68B	AGTCCCCAGCCAATCTTTCC	GCCCCAACCCAGCAGATG	461	47386420	47386880
HSTAL.116	GGTGCTTAGGGTGATGGAAA	TATGCCTGGTGACAGCTCAG	586	47386620	47387205
HSTAL.117	TGAGCTGTCACCAGGCATAG	AGAATCCAGGGTAGAGGGGA	512	47387187	47387698
HSSCL/M67A	CACAGCGAGGCTGCTTAGAGA	AGGGAGGGAACAGCAGGACA	482	47387699	47388180
HSTAL.119	ATAGCACCGCAGTCTTGCTT	TGGAGTGCTGCCCTAGAAGT	527	47388080	47388606
HSSCL/M66A	CTGCCTGGGTCTCCCTCTG	CACCAGCAGATCTAGGTTCTAGCTT	414	47389245	47389658
HSTAL.121	CTTCGAACGGATCACATCCT	ATACAGAGCCCTTCCACCCT	421	47389762	47390182
HSSCL/M65A	TGCTTGGGAATGGCAGAAGA	GCAAGTGGGGAGAGGTAGGAA	509	47389839	47390347
HSTAL.122	AGGGTGGAAGGGCTCTGTAT	GTCAGGGCAGTCAATTTGGT	500	47390163	47390662
HSTAL.123	ATTGACTGCCCTGACTGTCC	ATTGGGAGGCTGTTGATTTG	575	47390648	47391222
HSSCL/M64A	GGTGGGAGGCAGAATCATTGT	GAGGGGGAACACTGGCTTCT	334	47391105	47391438
HSSCL/M63A	GTCCTCCAGGGAAAAGAGCTG	CTCCCACCCATCTCCACTCTT	507	47391431	47391937
HSTAL.126	ACCTCATCGTTTCTGCCTTG	ATGGTGAGATTTAGCCGTGG	491	47392128	47392618
HSSCL/M62B	GCCAAGAGGTGAGTTCAAGGAC	CCCTCAAGCCCTTAACCACAG	369	47392495	47392863
HSTAL.127	CCACGGCTAAATCTCACCAT	GGCCATCAAGGTGAAAAAGA	584	47392599	47393182
HSSCL/M62A	TTATGCCTGGCTCTTAAGGATTG	GGTGCCAGTTCCACATGCTG	479	47393085	47393563
HSTAL.128	GTGCCTGCATGACCTACTGA	TCAGCAGACATCCCCTCTTT	404	47393307	47393710
HSSCL/M61B	ATGAATTGCGTTGGTCTCTGC	GGAAGCTCACCCCAACACAA	503	47393673	47394175
HSSCL/M61A	CCAAAACAAGCCTGGTAGGAAA	GCAGGGGGCAAGTCTTCAG	443	47394243	47394685

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 1 Co-ordinate Start	Chrom 1 Co-ordinate Finish
HSTAL.131	CCTGGCCTTGGAACAAGATA	AATCTCAATCCTGCAGTGGG	503	47394753	47395255
HSSCL/M60A	AGCTTGGGGCAACATTGTTC	CAGGCCAAAGCGGTTTACAA	525	47395143	47395667
HSTAL.133	AGGACAAGCCACTAGGCTGA	TGCAGCTTTTGGTCATCAAG	463	47395756	47396218
HSSCL/M58B	TTCGTGAGCCCCATCTTCAC	CGTACGCCCATCAGCACAC	415	47396529	47396943
HSSCL/M58A	GTGCTTTCCCCCAACTCCA	CTCCAGGCCAAAAGCAGTGA	461	47397098	47397558
HSTAL.137	TCCTCCTCCTGGTCATTGAG	GCCTCTTGGCACAAGTGAAT	547	47397662	47398208
HSTAL.138	GCCAAGAGGCCTTACAGCTA	GGCTGTTTCTTCACAGCCTC	462	47398199	47398660
HSSCL/M56A	TCGATGTTAGGGAGAGAAGCAGT	TCCACAGGGCCTTGCAGTAA	475	47398789	47399263
HSTAL.140	CTTGAACTTCTGGGCTGAGG	CCTGCAGTTGTGCTGTGTTT	512	47399219	47399730
HSTAL.141	GACAGAGAGCATCCCTCAGC	CAGACATGTAGGCGAAGGGT	549	47399791	47400339
HSSCL/M55A	GGGGCTGGGATCTGCACA	ACAGGAATCCACATCAAGTCCTAA	549	47400200	47400748
HSSCL/M54A	CGATCCCCCATTCTAGGCAC	CCTTGGGCAGCCTTCCTAC	475	47400853	47401327
HSSCL/M53B	GGTGGGGGGCTCAAGACA	GGGCTAAGCGCCATTATGTG	447	47401442	47401888
HSTAL.145	GGGGAAGGTCTCCTCTTCAC	GTGGTGACCAAGATGCACAC	538	47401719	47402256
HSSCL/M53A	CCGCCTGAGGACTGACCTG	CTCCCGCCCCAGAAACAA	473	47402147	47402619
HSSCL/M52B	GGAGCGCAGGGAGAGGAG	CCTCTGCCACCACCCTCAG	374	47402649	47403022
HSSCL/M52A	CAGGCACCCCTCCCTTCTC	ACGCCGTTCAGCAGGACC	337	47403471	47403807
HSTAL.150	AGGGAAGAGGAGGGAACAAA	TCAGAGTCCAGCTGAGCAGA	446	47403976	47404421
HSSCL/M51A	GGGGTCACAAGGTGGTTTCAC	CGCCCAGCGGATCTTTACT	505	47404306	47404810
HSTAL.152	TCACTTTGCTCAACCCTCCT	CCTGGGTGTCTAGTGGCAAT	493	47404756	47405248
HSSCL/M50A	TGGCATTCCCAAAGAAGGTGT	GTGCTGGGGCTCTTCCATC	443	47404857	47405299
HSSCL/M50A5	TCCCAAAGAAGGTGTACAGCAGTC	CTGGGGCTCTTCCATCACCT	434	47404859	47405292
HSSCL/M50A3	TTCCAGTGGCCTTGGTGTCTT	AAGGGCCAGAGTTTGTGAGTTG	529	47404939	47405467
HSTAL.153	CCAGAAGGGAGAATGGATCA	GCAGTGAGGTAGGGAACCAA	407	47405181	47405587
HSSCL/M49B	TGGGGCATGGACCTCTGG	CAAGGCCACTGGAAATGCAC	505	47405455	47405959
HSSCL/M49A	GGTGGGGAGGCAGGAACA	AAGGAACGCCCGAGAGATTT	447	47406027	47406473
HSSCL/M48B	AGATCGCCCAGGACCACAC	ACAGCCGCCCCAAAGTTAC	462	47406584	47407045
HSTAL.156	CTGTCCTGAGCCTTCCTCAC	GAGGGGTTGTTGTTGCTGTT	515	47406802	47407316
HSSCL/M48A	CCCCTTCCCAACTCCATTTC	CCCCGACCAACCAGTCCA	497	47407155	47407651
HSTAL.157	AACAGCAACAACAACCCCTC	ACCCCGAAAAATCGAAAATC	557	47407297	47407853
HSSCL/M47B	CCGCTACAGCCGTTTTCTAAG	TTCGGGGTTTAGACAGAGTAAGG	497	47407846	47408342

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 1 Co-ordinate Start	Chrom 1 Co-ordinate Finish
HSSCL/M47A	CGCCTGTGCCTAGGTGTTTG	GCTGCGCTCTGGGATGATTA	491	47408395	47408885
HSSCL/M46B	GCGGTTGGGGTGTGTAGATG	CCACAGAAGGGCAGCAAACA	308	47409142	47409449
HSSCL/M46A	GGCCTCCCACCGAATTCTT	CAGGCACACCACACTCGGA	368	47409451	47409818
HSSCL/M45A	CCATGTGGGGAGCAGTGTTC	GCCATTATGGGCCAAATGATT	534	47410103	47410636
HSTAL.164	CCCTTCCCTACCTCTTCACC	AAAGACCCTGGTCTCTGCAA	467	47410476	47410942
HSTAL.165	ACTCATTGCAGAGACCAGGG	CCAGTTCTCAGGCTCCAAAA	512	47410918	47411429
HSSCL/M44A	CCTTATGGTAGGGGGGAACTAGA	TGCCAGGCTGATTTGTCCTT	316	47411366	47411681
HSTAL.166	GGCTTCTCCTTGGTCATCAG	CCCACCTTCAAACCAAACAA	599	47411507	47412105
HSTAL.167	CATTAAATGTCCCCAGCCAG	TGTTCTTGTCCCCTGCTCTT	569	47412831	47413399
HSSCL/M42A	TCGGCTGCTCATCAGAGTGC	GGCGCTGCCTGAAATAGCA	444	47413153	47413596
HSTAL.168	ATGCATGCACTCTGATGAGC	TGGAAGCACAGACGTGACTC	595	47413571	47414165
HSSCL/M40A	AAACCGGGGGATATCACAATG	GCGGAAGACCAAAGCATACAG	556	47415098	47415653
HSSCL/M38B	GAATGCACATGCGCTTAAAATAG	TGGCCCCACATCAATCTTATG	389	47416739	47417127
HSTAL.170	TGGCCCCACATCAATCTTAT	GGGAATGTGGCAAGAGAATC	574	47416739	47417312
HSSCL/M38A	AGGAGAGGGCCACAGCTTTT	CCCCATAGAAGCAAGGCATG	259	47417271	47417529
HSTAL.171	TCACCGTTCAGGAGACACAC	AAGGGATTCTCCTGCCTCTC	568	47417423	47417990
HSTAL.172	GAGAGGCAGGAGAATCCCTT	CCTCTAGCAGGCACACACCT	533	47417971	47418503
HSTAL.173	AAGAATCCATTTGCACTGCC	AGTCCCAGGGCAACCTAACT	588	47418453	47419040
HSSCL/M36B	TTTCCCTGTGGCTGGTTTCTA	GGGGGAAGGGTGAACACCA	425	47418970	47419394
HSSCL/M36A	TGGGAGCAGCAGGTTGAAAT	CCAGCCACAGGGAAATGTGA	401	47419380	47419780
HSTAL.175	GGTTTGGGGATGGGTAGACT	CCCATGCTCTTAACCTGAACA	449	47419510	47419958
HSTAL.176	TGTTCAGGTTAAGAGCATGGG	CATTGGGGATCAGTAGGGTTT	451	47419938	47420388
HSSCL/M35A	TCGGACTCATATGTGCAGAATCAT	CCAATGCGGCAAGCAAATA	320	47420383	47420702
HSTAL.178	ACCAGAGAAAGCGCTATGGA	AATCAAACCCAGGCCTCTCT	536	47420920	47421455
HSTAL.179	TGGAGGCAAAGGGAGTAATG	TCCTGGTCAGAGGCTGAACT	468	47421331	47421798
HSSCL/M33A	GAGTTGGGCTTTGAAAGATTTTG	CCCCAGACCTCCCATGAAGA	421	47421718	47422138
HSTAL.181	CCCAACTCCAGTGCTCTCTC	ACAGAGGCGGTGACATTTTC	422	47422131	47422552
HSTAL.182	GAAAATGTCACCGCCTCTGT	AAGGCAGAACCTTCCCTAGC	577	47422533	47423109
HSSCL/M32A	TGCCCTTTAGCCAAGACCAG	AGTGGCTGCATTGGGCTAGA	510	47422804	47423313
HSSCL/M31A	TCCTCCAGTCCTGCAGCATAC	TCATTCATGTTGGCTTTTCATCA	394	47424426	47424819
HSTAL.184	ACCTTCCAGAGTAGGCTTCG	CACTGGGAGCCCTGTAATGT	555	47425549	47426103

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 1 Co-ordinate Start	Chrom 1 Co-ordinate Finish
HSSCL/M29A	TAAGGCTTTTCTGATGAAGTTAAATTG	ATTACAGGGCTCCCAGTGATAAA	521	47426086	47426606
HSSCL/M27A	CCAGCATGTTGCATCTCTGCA	GACAACAAAGGAAAATAGTCAATCTAGAC	387	47427821	47428207
HSTAL.186	TACACCAAAACCAACCCACC	TCCAGCTCTAAGTCTCAACTGC	587	47427971	47428557
HSSCL/M26C	TAGGCACCTTCTTAGATGTAAAACG	GCAGTTGAGACTTAGAGCTGGATAGTA	345	47428536	47428880
HSTAL.187	AATGATCAGGAGCCTTGTGG	GCTTGGTGGTCAGAAAGAGC	469	47428735	47429203
HSSCL/M26B	CACAGTGGGGCTTAGTTTAATTTC	TCATCTGCTTTAGCGTTTCAGAAG	488	47428896	47429383
HSSCL/M26A	AGGACCATGAACTTAAAGCTGAAGA	GGAGAAGGCTGAATGGGTCAC	328	47429403	47429730
HSTAL.189	TTCAGCTTTAAGTTCATGGTCC	TTGGGGCCTTGTATTTGTTC	519	47429708	47430226
HSSCL/M25A	CCATATCCAGAGACTGGTTGATGTT	AATAATAGCCTAGAAGGTAATCCATGAT	349	47430116	47430464
HSTAL.190	GGCCCCAAATGTAAAAGTCA	TGCCAACATTAGCAAATGACA	485	47430219	47430703
HSTAL.191	ATGGCTTTCTACATCATCCCA	TGGCATCACTGGTTTCAGAG	505	47431421	47431925
HSTAL.192	AAGAAAGCAGGTCTAATTTGGG	TGGCTAGCTTGTGTTAGGCAT	436	47432200	47432635
HSTAL.193	GCCTAACACAAGCTAGCCAA	CCCATAAATTGTTGACATGTGG	414	47432617	47433030
HSTAL.194	TTAGATGTGCAGCCACAAGG	ATGGAGATTGTGGGTTCGAG	512	47432956	47433467
HSSCL/M22A	TCCGCCAGAAGAGTGGCTAA	GTTTTTCAACTACTAAGCAGATAGCCA	288	47433323	47433610
HSSCL/M22A3	AGAATGCCGCAGAGATTAAGAGA	AGCCAGGAAGCAAGGTAAAGAAG	206	47433344	47433549
HSTAL.195	CTCGAACCCACAATCTCCAT	AGGTCATCACCGACGCTTAC	490	47433448	47433937
HSSCL/M21A5	CCGACGCTTACAGTGAAGGTTC	TCTGCGGCATTCTATTAACAGGA	391	47433537	47433927
HSSCL/M21A3	CTTAGTGGGCCTTCAATGAGTCA	TCCCCTGCCTTAGCCACTCT	409	47433581	47433989
HSSCL/M21A	AATCCTCATTGCCTCTACCACTAAC	CTTCTGGCGGAATGATGTGC	415	47433600	47434014
HSTAL.197	AGGGGCCATTTTGATGTAGA	TCCTAACCCTTCCCCTCTTG	568	47435305	47435872
HSSCL/M17A	GCCCATCATTGGCTAGATGAA	CAGAAAATATGCCAGCCTTAATAATAT	492	47437422	47437913
HSSCL/M17A5	TCAGGGCCTACCATATGCATG	CATGACCAGTTGGGCACATTC	290	47437526	47437815
HSSCL/M17A3	ATTGTGGCAATCTCATCAGAAAAGT	TTAGATGGCACGACTGTGTAACTGTA	305	47437573	47437877
HSTAL.199	TATCCAGCCAGGAGTCATCC	CTCCAAGGAAACTGAGCAGC	496	47437707	47438202
HSTAL.200	CTGCTCAGTTTCCTTGGAGG	TTCCCTGGCCTTGTTGTATC	447	47438184	47438630
HSSCL/M15C	TTGTCCATTTGCGAGAAGGC	GTAATTCTCCAAGTATTTTCGAGGAC	311	47439625	47439935
HSSCL/M15B5	AAATGTGATCTGGGACTTAAAGGATAG	AATTAGGCCTTTATGATGTTCACAGAG	286	47439721	47440006
HSTAL.201	AGGCCTTTATGATGTTCACAGA	TGAGAGCCTAGGGGAAACAA	421	47439726	47440146
HSSCL/M15B3	AAGGGGTAATTCTGGTTGGAGG	TTTCTGCATATTTCATTCCCTGAG	249	47439800	47440048
HSSCL/M15B	TTGAAAATTGGAAAGTAGTGGGATAT	GCCTTCTCGCAAATGGACAA	380	47439916	47440295

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 1 Co-ordinate Start	Chrom 1 Co-ordinate Finish
HSSCL/M15A5	TCCTTCCCAGCAGCTCTTCTCT	CCTAGGCTCTCAAGGGCACAG	684	47440134	47440817
HSTAL.202	CTCAAGGGCACAGAGTGTCA	AATGCAACCCTTGCTTCATC	510	47440143	47440652
HSSCL/M15A	ACAACCAGGTGCTGCTTGAGTC	TTTGGTTTGAATTAGGCATAATATCA	581	47440333	47440913
HSSCL/M15A3	TGGGGTGTTTATGTAGCATCATC	ACCTTCTCAAAAAGATGAAATGCT	396	47440556	47440951
HSTAL.203	AGAGAAGAGCTGCTGGGAAG	AGCTTGTTTTCCAGGCTTCA	559	47440797	47441355
HSTAL.204	GAAGCCTGGAAAACAAGCTG	TGCTTCCTAGGTTAGACACCAA	453	47441337	47441789
HSTAL.205	TGGTGTCTAACCTAGGAAGCAA	ATTTGCCTTGCAGCTTCATT	512	47441769	47442280
HSSCL/M13A	CTCTGTCGATCCCAACACATAACTA	AGGTATGCCCAAAGCACAGC	404	47442151	47442554
HSTAL.206	GATCCCAGAGAAGCACCAAA	AATGGGGGTTTTCAGATTGG	571	47442428	47442998
HSSCL/M11A	GGGGCAGGCAGTGGATTACT	ATATGGAAATTACTATGATCCTTACCCT	263	47443770	47444032
HSTAL.207	TCCTTACCCTGCAGGAACTC	TCCCAGGATAATAGCACATGAA	538	47443788	47444325
HSSCL/M8A	TTAACAAAGGTTTGGCAGAGTATTATT	GGATACAGAAAGTGGCTGGTTAAAT	305	47447329	47447633
HSSIL/M70A	AAAGTGGCTGGTTAAATTCTTCTTC	AGGTTTGGCAGAGTATTATTTTTAAAG	290	47447337	47447626
HSSCL/M7A	CCTGCCTGATATTACAGATATTACAGATG	TCAGAAATGCCAGGAAGAGGTAG	313	47448021	47448333
HSSIL/M71A	GCCAGGAAGAGGTAGGTGGAA	ACCTGCCTGATATTACAGATATTACAGA	306	47448029	47448334
HSSCL/M5B	CTTCCCCTGGCTTGCACAT	CCAGAATTGGGTCCAGCCTAG	289	47449530	47449818
HSSIL/M72A	AATTGGGTCCAGCCTAGAAAAGT	GGCGCTTTGTAATTCATTATTAAAC	485	47449534	47450018
HSSCL/M5A	AGTGATAATGGCTTTCTATATTGAATAAGT	CAAGCCAGGGGAAGACTGTG	307	47449805	47450111
HSTAL.213	GACCACGGCCTGTGATCTAT	CGCTTTGGATAGGAACACCT	452	47456154	47456605
HSTAL.214	AGGTGTTCCTATCCAAAGCG	CTGACGTCCTTCCTCCACAT	557	47456586	47457142
HSTAL.215	ATTGGTGCTGGGGAGAAAG	TTTTTCAGAACCTCTGTTGCAG	400	47457079	47457478
HSSIL_GAP/M39A	TTTTTTGCCTCTAATGAAAAGTCTG	GGTTCTGAAAAAAGGGAACATTT	203	47457467	47457669
HSTAL.217	AGCAAACCTGTGCCTGAAGT	CCCCACTGCCATCTTACTGT	450	47457926	47458375
HSSIL_GAP/M38B	ACCCTCCACACTTAATTCTAGGC	GGAGATAGTCCCATCATTCCATTAT	498	47457991	47458488
HSSIL/M81A	TGCAACCCCAGCTTTATGATG	CTGCCCATATCTCCCTGAGGT	625	47458132	47458756
HSSIL_GAP/M38A	TGTTGGATGGCAATTTCATAGAATC	ATATCATGAGATGGCCCATTATGAG	313	47458549	47458861
HSSIL/M80A	CCAACCTGTGGGTCATTTTAGAA	CTCTGCAAAAAATAAAATTGTCACAA	299	47459241	47459539
HSSIL_GAP/M37A	TCACTAGAGCTTGACATTGGTTTTT	AAATTGTCACAAGCAGAATGTTTTC	390	47459255	47459644
HSSIL_GAP/M36B	AAAGAAAACACTGGATTACCAAAGG	GGTCTCTAACCACAAAAGTGAATGC	213	47459993	47460205
HSSIL_GAP/M36A	GCATGCCATCTGTGATACAATG	GGACCAGCCAGCCTTAATAAAC	344	47460327	47460670
HSSIL_GAP/M35B	GGGTGTTGGAGAAGAGCTTTGT	TGCATCAGGCTTTTTTACATCCT	362	47460757	47461118

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 1 Co-ordinate Start	Chrom 1 Co-ordinate Finish
HSSIL_GAP/M35A	GTCAACGGTTTTGTGTGTATATTCATC	AACACCCAAGGCAGCTTGAG	444	47461112	47461555
HSSIL/M78A	AATTCCATGCCACAACTCAGGT	CACCCAAGGCAGCTTGAGATAC	395	47461114	47461508
HSTAL.223	TACCTGAGTTGTGGCATGGA	TGGCCTAAGCTTTAATGAGGA	540	47461486	47462025
HSTAL.224	TGCTGACCCAAGAAGTCACA	TTCCTCTCCTCAAGCTGCAT	579	47463484	47464062
HSTAL.225	ATGCAGCTTGAGGAGAGGAA	GGTCCTGAAATTGATCACCC	462	47464043	47464504
HSSIL_GAP/M31A	GGACAGTTTACAAAAGAGGGTTACAGA	TTCTGCATTTTGGCTTTCAGC	208	47465287	47465494
HSTAL.229	CTGAGTGGACACGTGGTTATTT	TGTGTGGGAATAAGAAATGGTT	598	47465838	47466435
HSTAL.230	TCCCACAAAGTTAACCTTCACA	AGATTGGGAGACTCACATGAAG	556	47466805	47467360
HSSIL/M51B	TTTATGAATGCTTCCCTTGTGATG	CCCACAAAGTTAACCTTCACAAGT	407	47466806	47467212
HSTAL.231	CTTCATGTGAGTCTCCCAATCT	AGATTCACCCACTCCAGGTG	577	47467339	47467915
HSSIL_GAP/M28A	CATCTTTTCATATTGCTTGTGCTAG	AAAGTAGAACTCCAAAAGCATCTGA	276	47467836	47468111
HSTAL.232	CCTGGAGTGGGTGAATCTTT	CTCATCTTCCAGGAACTGCC	577	47467898	47468474
HSSIL_GAP/M26B	CCATTCTCAGAGGCAATCACTGTAA	CCAAGAGGTAATGATCTCCAAGGTA	189	47470259	47470447
HSSIL_GAP/M26A	GAATTCACTTGCCTAACCACAGAA	TCAATTTATTACTATGTGCTTGTACTGCC	438	47470505	47470942
HSSIL/M47A	TACTAAGGCTTTTTAATGTATGGCTG	TCCAAACTCATTTGCAGTCATTTC	551	47470810	47471360
HSSIL_GAP/M25B	CTTTTAGTTGGCTGTCTGGAATTAC	ATTCCTTGGTATTTCCAGGTATACTG	305	47470939	47471243
HSTAL.236	TGGCTTTTATCAGCCATACATT	GGGAACCCATGTAAGGGAAT	508	47471325	47471832
HSSIL_GAP/M25A	TAAAGTGACCAGTATGGAAAGTGAACA	ACAGGTAGTCAGAAAGAGAGAAGGCT	295	47471671	47471965
HSTAL.237	TGTTCACTTTCCATACTGGTCA	CACCCTAAAGAACTCAGGGC	593	47471939	47472531
HSSIL_GAP/M23A	GTGCAAGACAACGTTCATTTGAAT	ACATTTGAACATTTGGCTGATAAGG	257	47473407	47473663
HSTAL.238	TCCCACCAATGGTAGAGAATAA	TATGGCACATGGAAGGTTGA	559	47473459	47474017
HSTAL.239	CCTTCCATGTGCCATATTCC	CATCAGTGGTACTGCCAAGG	564	47474002	47474565
HSSIL_GAP/M22A	AAATTTATCAGATGGGCAAGACATG	GTGATATAAAGCCTTGGCAGTACCA	191	47474535	47474725
HSTAL.240	GCCTTGGCAGTACCACTGAT	GGAAATCATGGAAAGCCTGA	532	47474545	47475076
HSTAL.241	TCAGGCTTTCCATGATTTCC	CCTATTCTCCTTGCCACTTCC	446	47475057	47475502
HSSIL_GAP/M21A	TTTGCCAAAGTACACTCAATTCCA	CGCCCAGCAGTCACAGATG	332	47475546	47475877
HSTAL.242	AAAATCCCATGCTCCACTTG	GGGTGGAGTTTCCTCACAGA	508	47475656	47476163
HSSIL_GAP/M20A	AACCCAAGGGAGGCTAGAGAATT	TGGAGTCACTGGGAAAGGAAAC	385	47475913	47476297
HSTAL.243	TTTTTGTAGGTGGTGGCACA	TCCTTGGACTGTGGTAAGCTG	419	47477396	47477814
HSSIL_GAP/M19A	ATGTAGCAAAGATTCCTTGGACTGT	TGTAGGTGGTGGCACAACTTAAA	428	47477400	47477827
HSSIL_GAP/M18B	TCCCCAATAAGACTGAAGTGCTG	GTGCCCATATCGCCAAAGTT	228	47477887	47478114

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 1 Co-ordinate Start	Chrom 1 Co-ordinate Finish
HSTAL.244	GCCCATATCGCCAAAGTTTA	ATTCTCTGAGGCTCAATCGG	598	47477889	47478486
HSSIL/M31A	TTGTCTGCCAGCCTGATGAAT	TTCCGATTGAGCCTCAGAGAA	322	47478465	47478786
HSSIL_GAP/M18A	TGTCTGCCAGCCTGATGAAT	TTCCGATTGAGCCTCAGAGAATC	321	47478465	47478785
HSTAL.245	CCGATTGAGCCTCAGAGAAT	TGAATTTGTGGGCAGTCAGA	524	47478467	47478990
HSSIL_GAP/M17B	GGGACTTTTTGATTCCATGCA	ACTGCCCACAAATTCAGCAA	373	47478975	47479347
HSSIL_GAP/M17A	ATTCGGCCAACTATTGTATTTGCT	AGTCCCCAGGAAGAGAAGCAGT	403	47479342	47479744
HSTAL.247	GACCAGGATCAAAGCGATCT	TGGTTCTCTGACAGCAGACG	553	47479396	47479948
HSSIL_GAP/M15A	CAAGAAAGAAAGTATTCGGGTTCAA	TTAGATGGCAGCTAAGCACAGC	210	47481287	47481496
HSSIL_GAP/M14B	TTGTGGAACGTTGAGAAGCTTG	GCTAGGCTGACTCAAGGCTCTC	219	47482073	47482291
HSSIL/M27A	CAACGCCAACTGGAGATTTCAT	GCTCTCCCCGCCATAATCTAT	538	47482089	47482626
HSSIL_GAP/M14A	GTACCTTTCCACTTTCCTCCATCA	TCCACAAAGACACCATGCTGAC	385	47482285	47482669
HSTAL.250	ACCAGGAGAATAGGCACTCAAA	TTTGCTTTTCATAACTGTTGGG	548	47483767	47484314
HSSIL_GAP/M12A	TGTGGTTTTCTCGTGTTATACACTCAT	TGAACTGATTTCCAGATAATTGTTAGATAATAC	282	47484261	47484542
HSSIL/M22B	TGGTTGGGTTCCTTCCCTCT	GGCAGGGGGGTAGAACACA	412	47487145	47487556
HSSIL_GAP/M9A	GGTTCCTTCCCTCTTACAAATACTGA	CAGGGGGGTAGAACACAAGG	404	47487147	47487550
HSTAL.252	CCAACCAATTTCCCACTTGT	TTCACCAGACATCATGGAGC	475	47487550	47488024
HSSIL_GAP/M8A	GCCCCAAATAAATATCTAGCCCA	GGGCCGTGCAAAAGGATATA	475	47487982	47488456
HSTAL.255	ATCCTCCACTCCCATTGCTT	AGATGGAACTGTCCAGTGGG	445	47488878	47489322
HSTAL.256	CCCACTGGACAGTTCCATCT	GCACGTTTTTCTCTGTCTTGAA	580	47489303	47489882
HSTAL.257	TTCAAGACAGAGAAAAACGTGC	CCATCTTTTCACCCAAATGAA	541	47489861	47490401
HSSIL/M19A	TAAGGTGGTGGGTATGGAAAGATAA	AGCTTCCACAGAATGCAGGC	409	47490209	47490617
HSSIL_GAP/M6A	TCTTCTAGCAGCATTATGGGGACT	CCACAGAATGCAGGCTTTAAATC	464	47490214	47490677
HSTAL.258	CATACCCACCACCTTAGTTGC	GCATCCTGTGCAGATATTGAAA	600	47490602	47491201
HSSIL_GAP/M5B	ACGGTGGGAATTTCTTGAGGACT	ATGCCTCCCTTAACTTGGAATG	291	47491198	47491488
HSSIL_GAP/M5A	CGCCCGCAGTTCTCCAAG	AGCTCAGATGATACCCAAGGATTTC	497	47491542	47492038
HSSIL/M18A	CGGGTGTCCGCTTCCAGT	CGCGGAGCTGAGGTCTGTT	495	47491737	47492231
HSSIL_GAP/M4A	TGTCACTAAAACAGATGGCTTCCTAG	AGGATCCGAGGATTTAAACCTTT	360	47492059	47492418
HSTAL.261	TTCCTCAACACTCCACTTTTGT	CAGGAAAGCAGAGAATGATGC	509	47495016	47495524
HSTAL.262	GCATCATTCTCTGCTTTCCTG	CATGTAACCAATGGCCGTC	401	47495504	47495904
HSTAL.263	CAAGGTGAGAGCTGAGGCAT	GGGTCAGATCCCCAAAGCTA	600	47499089	47499688
HSSIL/M10A	TCAGTTCCCTTAATCTCTTTGAACA	ACCACTGGACTTCTAAGCAGGATAT	417	47499474	47499890

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 1 Co-ordinate Start	Chrom 1 Co-ordinate Finish
HSTAL.264	GTGCTAGCTGGTATGGGGAG	GCTTTTCTCTGTAAGGGCCA	573	47500449	47501021
HSTAL.265	TGGCCCTTACAGAGAAAAGC	TCATGTGGACAGGGACTGAA	584	47501002	47501585
HSTAL.266	TCACCAGGAACAGGGAAGTT	GGAATTTGGCATAAATTGACAG	558	47501683	47502240
HSSIL/M7A	ACGCCCAGCCCATCTCAA	CAATTTATGCCAAATTCCTTTTAGAC	431	47502223	47502653
HSSIL/M5A	CCCACAAGGCAATAGATGACAA	TAGGCAGCAAAGGGGAGTAGTC	418	47504210	47504627
HSTAL.269	TCATCTATTGCCTTGTGGGG	TACAGCCTTTCCAGTTGGGT	449	47504609	47505057
HSTAL.270	ACCCAACTGGAAAGGCTGTA	TGCTAGATTGTGGGGAGGTC	419	47505038	47505456
HSTAL.271	GACCTCCCCACAATCTAGCA	TGTCTACTATGTGCCAGGCATT	426	47505437	47505862
HSTAL.272	ATGCCTGGCACATAGTAGACA	TTAGTGCTCCTTGGAGGTGG	505	47505842	47506346
HSSIL/M3A	CAAAAGCCCTCTGTGGAACCT	GAAATGCCCCTTTACTGGGTG	512	47506217	47506728
HSTAL.273	TGATCAAATAGGCCCAAAGC	TAGAGACGGGAAGCTGGAAA	566	47506516	47507081
HSSIL/M2A	ACATGGGCAGCCAAGAACAA	ATGGGTGTCTGGGTTTCAGTTTT	475	47508028	47508502
HSSIL/M33A	TTTGGGGACCAGATTCGTCTC	AGGCTGGGAGACAGGAAGTGA	471	47512062	47512532
HSTAL.277	CACTGTCACCGAACGTGTCT	TGGTCGCCTCTTCTGTTTCT	515	47512283	47512797
HSSIL/M34A	GGTTGGTTAGCTGCATTGACACT	AAATACATCTGTCCAACACATAATGACT	342	47512690	47513031
HSTAL.278	GCACACAGAGACAGATGCAGA	TGCATTCAAATATAACACGCAT	484	47514101	47514584
HSTAL.279	AATTGCTATTCCCTAGCTGGC	AATTTCAGGTGCCCTGTGAG	562	47515265	47515826
HSSIL/M37A	CATAAATCCACGGTAAATTCTCCA	AATCCCAAAACCTCCCTCCA	382	47515758	47516139
HSTAL.281	GGTTTTGGGATTGATTTTGC	GAAAGCAATTTGGCATGGTT	595	47516128	47516722
HSTAL.282	TGCTTTCTGTAGCACCCAGA	CCTCTGGGGATCGAATTTTT	591	47516716	47517306
HSTAL.283	TGAGCAGATTTCCAAGCTCC	CTGAAGATCTCCCCTCCTCC	559	47518205	47518763
HSSIL/M38A	ATTAAGGGTGCTTATGACATGGG	TGAGAAAGAGATGGGAAAGCAGTT	220	47518703	47518922

Appendix 2

Sequences of Primer Pairs Used to Construct the Mouse SCL Tiling Path Array

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 4 Co-ordinate Start	Chrom 4 Co-ordinate Finish
MMTAL.3	ATTTGCAGCAGCTGGTAGGT	CATCACCCCCAACTATCCAC	498	113334279	113334776
MMTAL/4ver2	AGGATGGAGCGTGTACCAAG	TAAGGGTTCCACAGGGAATG	641	113334608	113335248
MMTAL/5ver2	TCTGGACAGAGATTGCCCTT	ACTGTGTGGGCTCCAGAGAT	543	113335309	113335851
MMTAL.6	CCAAACCTGGAGACCTGAGA	TTTGTTCTTGCGCTCTGCTA	506	113335809	113336314
MMTAL.10	AGTGTTCAGGGTGTTTTGGC	ATAGCATGCGGTTTCAGTCC	414	113338093	113338506
MMTAL.11	AGCTATTTCCTGGAGGCGTT	CAAGTGCCTTTTAACCCCAA	592	113338592	113339183
MMTAL.13	AGTTCCTCCCATGTCCTCCT	GCCTAGAGGCCCAGTCTCTT	453	113339586	113340038
MMTAL.14	GGCAGAAATGGTTTTGGAGA	CCTACACAAGGCTCAGCACA	459	113339959	113340417
MMTAL.16	GCACATGTTTTTAATCCCCG	GGCATGAAGGCTGTCAAAAT	596	113341178	113341773
MMTAL.17	CATGCCAGCACTTAGGATGA	GGATTGCCTGCCAGATTAAA	514	113341768	113342281
MMSCL/M000A	AAAGAATCGCACAGTATGAAGTCAG	AGCCCAATGGTGCCATTTAG	442	113342233	113342674
MMSCL/M001A	CATTGGGCTGGTTGCATTTAA	CCAAATGGCCCACCTGTAGA	524	113342666	113343189
MMTAL.19	AGATGTGGCATGCAAATGAA	ACCAGAAGAGGGCATAGGGT	539	113342906	113343444
MMSCL/M001B	TACAGGTGGGCCATTTGGG	TCCCCTGCTCTGGTGCATAAT	520	113343172	113343691
MMSCL/M002A	TAGAAGTCCAGTGCCAATAAAATGT	AGGGTCCTGGGTTGAGTGACTAG	464	113344136	113344599
MMTAL.22	GCCTCCCAGAACAGTCTCAG	TGGGAAATTCCTTCTCGTTG	447	113344641	113345087
MMSCL/M003A	CGTTTGCTACCTTCCCCTAACAC	TCAGAGATAGAGGAGGTTAAATTTGC	292	113344861	113345152
MMTAL/23ver2	CACCTCTTCAAAGCTGGGAC	GAGTCAGGATCCAGGTAGCG	611	113345258	113345868
MMTAL.24	CCGTTGACCAAATGAAGGTT	TGAAGTCAATCTGAGGCACG	587	113345544	113346130
MMTAL.25	CATTGTGACGTGCCTCAGAT	CCATCAAGTCACCCTCACCT	574	113346103	113346676
MMTAL.26	GCATGAAAGGAGAGACAGGC	ATCCTCAGACCTCCACATGC	435	113346835	113347269
MMSCL/M005B	TCGGCTCCTTCTGCCAGTAG	GGCTTTCTGGGGATCATGTGT	453	113347282	113347734
MMTAL/27ver2	CTTCTGCCAGTAGGTCCTCG	TTTCCTGTTTGGGCTTTCTG	457	113347289	113347745
MMTAL.28	GGAAGATCTCTGTCCCCACA	GGATGGAGCATTCTTGCCTA	517	113347819	113348335
MMSCL/M006A	ATGAACGGAAGGAAGAGATTTGG	GGAATTACAGGCTGAAGGAACTACA	437	113348124	113348560
MMTAL.30	ATCTCTCCAGGACCCGTTTT	TGCCTTCCACTTAATACGCC	489	113348812	113349300
MMSCL/M007A	TAGCTTGCGAAGGAATACTGGA	CCAGCTCCTCCTCCCTACTCTAC	469	113349110	113349578

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 4 Co-ordinate Start	Chrom 4 Co-ordinate Finish
MMSCL/M008A	GACAGCACTGAACACTTAACAATGAA	ACCCTGGCTTGACCTGAAAAG	427	113350226	113350652
MMSCL/M009A	CCCACTGGTGGAAGCCAAA	GTGGATGCGCAGTAAGCTTCA	325	113350767	113351091
MMTAL.35	TGGACCTCTGGAAGAGCAGT	GGAAACCAGACTGGCCATAA	595	113351530	113352124
MMSCL/M010A	TGCCTGCACCTCCTGAGTTC	AGGAAATAGCCACATTAAACAAAAAC	412	113351859	113352270
MMSCL/M011A	CATTGGTTTCAGCTGGCTATTG	GGGGAAGATGGTGCAATTATGT	513	113352531	113353043
MMSCL/M011B	CCCCTCACGCTCTGAAATTTAT	CCACAACTGATCAAAATGCAGAAA	473	113353074	113353546
MMSCL/M012A	TCTAGCCCCGGCATGACTTC	CCCCTGAGCACTTCCTAGACC	495	113353731	113354225
MMTAL.40	CGTTGTGGTGCTGTGAAAAC	TCTGACTTCAGTGAGCCTGC	477	113353917	113354393
MMSCL/M013A	GGCTGAGTCATAATTGCGTCCTA	GGCTTCACAGGCACACACTTG	485	113354591	113355075
MMTAL.43	GTTGCCTGGAGACTTGCATT	TCCACACTAAGCCGAGAGGT	434	113355399	113355832
MMTAL.44	ACCTCTCGGCTTAGTGTGGA	GCCCTAACACAGAACCAAACA	555	113355813	113356367
MMSCL/M014A	GGTCTCCAGCCTTCTCATTAAATT	GTTTTAAGCCCTAACACAGAACC	239	113356136	113356374
MMSCL/M014B	TCACGAAAAAGAAAAGGTATTTGC	AGCAAGGGCTACCCTGAGAAA	327	113356454	113356780
MMSCL/M015A	AGTTGTGGGCATTTTCTTGTAAGG	ATTCTAGGGTAGCAGGGCTACACA	468	113356886	113357353
MMTAL.46	TTCCTTTCCAGTCCATGCTC	GTATCTCAAGCAGCAAGCCC	486	113357010	113357495
MMSCL/M015B	GTATGGGCTTGCTGCTTGAGA	TGGGCCACACACGGAGTTC	414	113357472	113357885
MMSCL/M016A	GTGTGGCCCAGATTGACTTTG	CCGGGGGTACAGTGAGTGC	469	113357876	113358344
MMSCL/M016B	CTAGCCCCAACCCCAAGGT	TGCCCTCTTTTCCATGAACAC	457	113358447	113358903
MMSCL/M017A	CTGTGGGCGCTGGGTATG	ATCTCAAGAATGCACACACAATCAC	282	113358947	113359228
MMTAL/51ver2	ATACTGGCTCTCTGCCTCCA	CCAGACAACCATGCACAGAC	519	113359480	113359998
MMSCL/M018A	ACTTACCCAGGAGATCTTTCATTATAAC	CCAGGGTAGCCAGGAAGTGAG	434	113359643	113360076
MMTAL.52	GGGCTTCAGCCACTAAGATG	ATGGCTGGGAAACTAAACCC	532	113360129	113360660
MMSCL/M019A	CGTGGGAGCGAGAGATATGG	TGCCTGGCTCTCCTGGAAC	253	113360579	113360831
MMTAL.53	CCAACTGGGTCAGGCTTTTA	TCTAGCACGGGGAAAATGAG	402	113360742	113361143
MMTAL.54	TTATGGATGGTTGTGAGCCA	GGTTTTTGCCAGAAGGTTCA	513	113361325	113361837
MMTAL.55	TGTTCGGATAAATGGGGTTC	TTCAGGGTAGGGGTAAAGCA	416	113361745	113362160
MMTAL.56	TTTTTAAATCATCGTGGCAGG	TTCAGGGTTTGATTCCTGGT	508	113362336	113362843
MMTAL.57	CCAGGAATCAAACCCTGAAC	GGAGAGATGGCTCCAAGGTA	445	113362825	113363269
MMTAL.58	TGGAGCCATCTCTCCAGTTC	CATAACCCTGGCTTTCCTGA	582	113363255	113363836
MMSCL/M022A	AGGAAAGCCAGGGTTATGCA	TGGTGGTGCTGGGATCAAAG	425	113363819	113364243
MMTAL.59	TGCAGAGAAATGCTGTCTGG	GAGAGCCAAAAATGGGTTGA	600	113363835	113364434

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 4 Co-ordinate Start	Chrom 4 Co-ordinate Finish
MMTAL.60	TCAACCCATTTTTGGCTCTC	TCACAACGGTCTTTCTGTGC	548	113364415	113364962
MMSCL/M024A	TGGGAGACAATGACTACAGCTGA	TTGGTTTTCTGTTATGATAGCTACGTG	283	113365630	113365912
MMTAL.63	TTTCTCCGTGTATCCTTGGC	TGCTTTACGTTTCCTGGTCC	579	113366002	113366580
MMSCL/M024B	TGCCTGCTGTGCTGGAATTA	AAGGGCCTGGCTGTGGAA	411	113366240	113366650
MMSCL/M025A	GGGGGCATATCTGTCACACTG	AGGACTAGGAATCGGTCTTGGAG	454	113366759	113367212
MMTAL.65	CGATTCCTAGTCCTCACCCA	GCATCGACTTCTCAAGGGAG	588	113367199	113367786
MMSCL/M026A	TTGAAAGGGACGGAAAATTAATATAG	GGTGGCCTCAATTTTAAACCTTT	426	113367732	113368157
MMTAL.67	TTGAGGCCACCCTGATCTAC	CTGGTCTTCGCTAACCTTCG	469	113368147	113368615
MMTAL.68	AGGTTAGCGAAGACCAGCAA	AACACGAGCGAAAGTTCAGG	568	113368599	113369166
MMSCL/M027A	TTCCTGAACTTTCGCTCGTGT	TGTGGCTGGGCTTGTGGA	421	113369145	113369565
MMSCL/M028A	AGCCCAGCCACACAATCTTG	GCTGCCTGTCTTGCTCTGTGA	485	113369554	113370038
MMTAL.71	ATGGGACTGTCTCCTGATGC	CCAATGCCCTCTTCTGATGT	439	113369987	113370425
MMTAL.74	CCTGTGCCAGAGGAGGTAAA	AGGAACAACATGTAAGGCCG	496	113371813	113372308
MMSCL/M030A	ATGAATTTCCTCCCTCCTTTTTAA	CTCATTATCCTCTTCAGACATCTCAAC	315	113372212	113372526
MMSCL/M031A	ACGCATCGGCTCCCATTAC	ACCTGGGCCTGGGGAAATA	294	113372715	113373008
MMTAL.76	GGTTGCTGGGAACTGAACTC	TGCTGGTACACTGTCTGCAA	567	113372756	113373322
MMTAL.77	TGCAGACAGTGTACCAGCAA	GGACCCATGTACTGGAAGGA	501	113373304	113373804
MMSCL/M032A	CCTTAATGGAGTGGGCAGATGA	CTGCAGTCCCAGGTGATCTAAAC	339	113374027	113374365
MMTAL.79	ATGAGCGACAAAGCAAGGTT	CTGGGGCTTGAGTGACAAGT	600	113374326	113374925
MMSCL/M033A	ACTCAAGCCCCAGGGTATCTAAT	GCTACTAAGGCCCGCTTTCAC	413	113374913	113375325
MMTAL.82	GCAAAACAAATGGAGCTGGT	CCCCTTTCTTTGGTGTGTGT	440	113375591	113376030
MMTAL.83	CACACCAAAGAAAGGGGTTG	TAATTCCTTTTGGAATGGGC	549	113376014	113376562
MMSCL/M035A	TCATGCCTAAAGTTCAAGTGCTG	GGGCTGCACAGAAGGATCAC	443	113376762	113377204
MMSCL/M035B	TGTGCAGCCCCAGATATTTATTC	CAGAGCAGCAAATGAGAAGACAGA	294	113377195	113377488
MMTAL.86	GGTCCTCATGCAAGCAGAAT	TCTTGATCATCACCTGCTGC	588	113377736	113378323
MMSCL/M036A	TTTCCTTTGCCACTGGTTTCTT	TGGCTGATTAACTTCCTCCACA	403	113378096	113378498
MMSCL/M037A	CCCACCCCCATCCATCCT	AAGACCCAACAAGCAGCTGAA	398	113378615	113379012
MMTAL.89	CTATGTATCCTTGGCGCTGC	TGTACTGTCATTTGGAAGCCTG	524	113379132	113379655
MMSCL/M037B	AGCTTAGTGTATTTTCATTGACGTCTTT	ATGCTGTTGTCTTTTTCAAATATAATTG	235	113379447	113379681
MMSCL/M038A	AGGACAACCAGGGCTATACAGAGA	ACAATCGGGAGCTCTGCTTTAC	296	113379846	113380141
MMTAL.91	TGGGCAAAACAAAGTTGACA	CGATTGCTCTCCAAGAGGAC	582	113379902	113380483

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 4 Co-ordinate Start	Chrom 4 Co-ordinate Finish
MMSCL/M038B	CCCAGGTCCTCCAGGTCTCTT	TCAGGGCAGGGTTGTTCTATACA	349	113380450	113380798
MMSCL/M039A	TGGTGGCTTCGGTGCTGTT	GCTCCAACACTGTGCTCAAAAA	377	113380858	113381234
MMSCL/M039B	TGCTGTGGGCTGTTGGTGAT	GCATGGCCTGGCCTATACTG	345	113381494	113381838
MMTAL.94	GCTTTAGGAATCTCCCAGGC	CAAAAACAACAGGTTGCCCT	500	113381670	113382169
MMTAL.95	AGGGCAACCTGTTGTTTTTG	AAGCTGTCCAAACACTGGCT	554	113382150	113382703
MMSCL/M040A	TTGGAGCACAGTGGGAGGAG	GGTGCATGCCTTTAATCCCA	412	113382167	113382578
MMTAL.96	AGGCATGCACCCATTATGTT	CCATCTCTCCAACCTCCAAA	545	113382568	113383112
MMSCL/M041A	GAGGCTGCAGTGAAGTCATGTG	GAGCCAGAAAAAGAACCAGGATT	380	113382769	113383148
MMTAL.97	TTTGGAGGTTGGAGAGATGG	GGAGGAGCAGCACAATAAGC	480	113383093	113383572
MMTAL.98	GCTTATTGTGCTGCTCCTCC	TGCTGTTACTGGCTCCCTCT	545	113383553	113384097
MMSCL/M042B	TGTCCTTTGCAGCAGCTCTTC	TGTCACAAGGAAGGCTAAATATTATCT	531	113383967	113384497
MMSCL/M043A	CCATAGGGCCAAGTGAACCA	TGAGGGAAGGGGAACTTTTGTA	419	113384703	113385121
MMSCL/M043B	GTTCCCCTTCCCTCAAATTCTT	TCGGCAAGCAGTGATAAAGCA	538	113385107	113385644
MMSCL/M044A	AAGCCATCCCAGAAGCCCT	AATGGGAGTCTTAATCATGCCTGTA	359	113385883	113386241
MMSCL/M044B	GGGGTGTTGGGTGTTGGTTT	GGCCAAAACACTGAAAGACACG	415	113386410	113386824
MMSCL/M045A	AGCCCGGAAGCAGTGATCTC	CTGCATGCTGGGTTTCTGTTC	389	113386831	113387219
MMSCL/M045C	CCCAGCCACGCAGATAGTGA	GAAGGTGCCCACAGAATCCA	399	113387164	113387562
MMSCL/M046A	GGCCATGACCTTTGGAATTTC	CCAGACGAGGAAGGCAGAATT	361	113387819	113388179
MMTAL.108	GAACCCAGGTACCTCACACAA	CTAAGTCCAGGACAGCGAGG	585	113388340	113388924
MMTAL.109	CGCTGTCCTGGACTTAGCTC	GGGAAACACCACTGTCACCT	400	113388908	113389307
MMTAL.110	CTTTAAATCCCAGCAGGCAA	CAGGACCTTTGGAAGAGCAG	492	113389501	113389992
MMSCL/M047A	ACTTTGCCACCTGTTATCACATCT	GCTCCGGCCCAAAGATTTAT	349	113389767	113390115
MMSCL/M048A	CCCAACAACCCCATGAAGC	GGGACTGAGCACTGAGAAGGG	426	113390173	113390598
MMSCL/M048B	GCTGGGGGCTCTGCACC	AAGGCCCCTCAATCTGGAAG	446	113390724	113391169
MMSCL/M049A	GAGGGGCCTTTATGTCTGCA	TGCATCACTCCCTCACCAAGA	397	113391160	113391556
MMSCL/M049B	GACGTTGGCACACTGGAGAAA	CATCCTACCTGGCTCACGTCTT	369	113391583	113391951
MMSCL/M050A	GTGCCGAAGAGAGGGAGGAG	CAGGAAGTGGGGAAACAGGAT	417	113392137	113392553
MMSCL/M050B	TTTCCCCACTTCCTGGCC	TTTCCAGGATGTCAGGCTTCA	372	113392539	113392910
MMSCL/M051A	GACCCCCTGCATCATCTCCT	GGCAGGGCAAGCCATACAA	389	113393104	113393492
MMSCL/M051B	CCCCCACCAGCACAAGACT	AGGGAATGGTAAGCACAAAGCA	475	113393491	113393965
MMTAL.118	CAGCTCCATCAGTCTCCCTC	CTCCCAACCTTGACACCAGT	532	113393742	113394273

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 4 Co-ordinate Start	Chrom 4 Co-ordinate Finish
MMTAL.119	GGTGTCAAGGTTGGGAGAAA	AAGTGGGTTTGTGAAGACGG	412	113394257	113394668
MMSCL/M052B	TGGGTTTCTTTGCTTCATCCA	AAAAGCACCATGTCTAGGGCA	347	113394622	113394968
MMSCL/M053A	CTCCTGGCAACCCAAAGCT	ATCTTAGGCTCAGGCAAACTCTG	337	113395186	113395522
MMTAL.122	TCTCAGTGAGCTGGAGCAGA	TTACACCCTCTGTGCCCTTC	484	113395398	113395881
MMTAL.123	GAAGGGCACAGAGGGTGTAA	GAGTTCCAGGACAGCCAGAG	485	113395862	113396346
MMSCL/M055A	CAACAGGGGTGGAGAGGACA	TGTGTCCCCAGTTTGTGTTCC	392	113396860	113397251
MMTAL.126	CTGAGGACAGAAAGCTTGGG	AGCCCCTGCCATTCTAGTTT	479	113397213	113397691
MMSCL/M056A	AGTCCCGACCTTCCCTACGT	GCGCCGCCTGAAATAGCT	456	113398031	113398486
MMTAL.129	GCATAACCTTTTCAAGCCCA	CTTTCGGAGGGGAAACTAGG	457	113398743	113399199
MMSCL/M057C	GACGCCTGGCTTTCCTGG	GAAAGTAGGGGGCAAAGAGATG	376	113399687	113400062
MMTAL.132	TTTGCCCCCTACTTTCCTTT	CAAATCCAGGTCCTCCAGAA	454	113400047	113400500
MMSCL/M058A	ACCTCCTATTACCCCTCATCTCC	GGGGGGGTGTTCATGAGC	332	113400372	113400703
MMTAL.133	TTCTGGAGGACCTGGATTTG	CCTGCCTTGAAGCAGAAGAG	403	113400481	113400883
MMSCL/M058B	GCTGTGAGTTGGTCCTTTTCTGA	CCAATGGGCTGAGGAGTATCTG	340	113400781	113401120
MMTAL.134	AACAGCCAGTTTCCACCATC	GGCCAAGGACTCAATTCTCA	576	113401064	113401639
MMSCL/M059A	TTGAGTCCTTGGCCTTGCAC	CCGACCCATTCCCAACCA	400	113401626	113402025
MMSCL/M060A	CCAAAGCCGCTGAACGAG	ATAAGCGCCTCGGCCATT	242	113402059	113402300
MMSCL/M060B	AACATGGCCACGCACACC	AGACAGCGCAGGGTTTCACA	405	113402419	113402823
MMSCL/M061A	GCGCTGTCTGGGTGTGGAG	CCACCAGCGCCTCGATCT	424	113402815	113403238
MMSCL/M061B	TATTCGGGAGCCAGTGTTGG	AAAACGCGATCCTTCTTGTCC	456	113403295	113403750
MMSCL/M061C	TCGCGTTTTCCCCTTTTACTG	TTGGCGGAAGCTCAGAACTG	406	113403742	113404147
MMTAL.140	GCCTCAAGTAACAACGGGAA	AGATTTTCGACTTTCGGGGT	405	113404010	113404414
MMSCL/M062B	GGCAACCCGCATAGAGACG	CGCCCAAAGCGAGTTTCC	361	113404793	113405153
MMSCL/M063A	CCCGGTTCCGAGCACATT	CCCCAGCCAGCAGGTGAC	421	113405595	113406015
MMSCL/M064A	GCTGGGGCAACTGGAAGG	CCCAGCTTCTAAATGACAAACATAC	422	113406009	113406430
MMTAL.146	CTCAAGGTAGGAGGACAGCG	ACTAACCACCGACATGGAGC	468	113406471	113406938
MMSCL/M064B	AGCAGGGAGGGAAGTCAGGA	ACCACCGACATGGAGCACAC	391	113406544	113406934
MMTAL.147	GCTCCATGTCGGTGGTTAGT	CCCTGAGAAGGGGGTTAAAG	452	113406919	113407370
MMTAL.148	AGGATCATGTGTAGCCCAGG	GAGGGCATCAGATCCCATTA	455	113407311	113407765
MMSCL/M066A	GGAGGGGAGCCCAGATTCT	GCGGGGAATCTGTCCAGTG	361	113408009	113408369
MMSCL/M066B	GCTTGCTCGGGGGATTAGTT	CCGCTCCGTCATCCTGTAAC	385	113408399	113408783

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 4 Co-ordinate Start	Chrom 4 Co-ordinate Finish
MMSCL/M067A	CAGCCGCTCGCCTCACTAG	ACCCTTTGTGGCGAACCTCA	384	113409195	113409578
MMSCL/M067B	GGTGGGTGGACAGGCTGG	CGTGAGGGATGGAAGTGCTAGTA	375	113409601	113409975
MMSCL/M068A	CTTGCCCCTCCCATTTATGTATT	TCGCCTCCAGGGTGAAGACT	434	113410107	113410540
MMTAL.155	CACTTTAGGCTTGCTCCCTG	CCCAGCTCTCTACATGCCTC	491	113410592	113411082
MMSCL/M069A	AGGGGGACTCAGCATTCGTT	AGAAAGAGGCGTGAGAAGCATC	438	113410876	113411313
MMTAL.156	GAACCACAACAGCCCAGAGT	AGAGGGAAGGGAGGAAATCA	529	113411187	113411715
MMSCL/M069B	TGCCCTGTCCACTTAGAAGCTC	AACAGGCAGGCCAAGAGACC	370	113411759	113412128
MMSCL/M070A	GACCCCTGGACCCAGACATT	CACAGCCATGCCTCAGAAAAG	341	113412168	113412508
MMSCL/M070B	TGGGGTTGGGGGACCTAG	CAAAGGCAGGCTGGCAAGTA	328	113412530	113412857
MMTAL.159	CCTTCCTGTGTCTGGATGGT	GGTCACAGATGGTTTTGGCT	596	113412710	113413305
MMTAL.160	GCCAAAACCATCTGTGACCT	TGAGAAGCCAGGGAGCTAAA	573	113413287	113413859
MMSCL/M071A	CTGCATAGCCCATAGGAGGTG	CCTTCCTCCTCCTGGTCATTG	451	113413675	113414125
MMSCL/M072A	TGGCATCCCCCCTGAAGA	GGCTTGGGAAAGGGAGAAGA	428	113414189	113414616
MMTAL.162	GGTCTTCTCCCTTTCCCAAG	AGGCAAGTCTCAGTGAGGGA	436	113414595	113415030
MMSCL/M072B	CTGGCTGGGTGGCTTTTGT	AAACCCAGTGCCCCAAACA	399	113414758	113415156
MMSCL/M073A	CCAGCGCATCCCTGTTTGT	AGGGGGCACTCGTTTCATG	409	113415162	113415570
MMSCL/M073B	TTTCTCCCCTGTGTCAATCCA	CACGCAATGGGAAAGAACCA	447	113415695	113416141
MMTAL.166	CCATTGCGTGGGAGTTAGTT	AAGACCAAAACCCTGGCTCT	597	113416132	113416728
MMSCL/M074A	GCGGTGGACGGACTGTGAT	CTGCCCAGCCCCTTTAGAA	373	113416333	113416705
MMSCL/M074B	AGCCAGGGTTTTGGTCTTTGT	GGCTTGCCCACAGACATGG	438	113416711	113417148
MMSCL/M075A	CTGCGGGCGTGTGACTCTT	ATGCAACCCCCACTGAGTCA	491	113417273	113417763
MMTAL.170	TTTGGTCTCCTGGGAGTTTG	GCTGGGAATTAACGGTGAGA	557	113417981	113418537
MMSCL/M076A	AAAGTGGGCAGATCAACAGGG	CAAGGCCAAATCCTACCATCC	421	113418281	113418701
MMSCL/M076B	TTTGGGGTGGAGGAAGTGG	GGCCTCAGGGCACTTTTCTT	398	113418794	113419191
MMTAL.172	CATTCTCCCTGTCCCAAGAA	TCCCATGGAGTCAGGAAAAC	567	113419111	113419677
MMSCL/M077A	CACCCCACGCCTCACTCTT	GCCAGGCCCAGAGAGAAAAT	427	113419484	113419910
MMSCL/M078A	CTGGGCCTGGCGTCTGAC	CCTGTCCTGGGCCTAAATTTG	491	113419900	113420390
MMTAL.174	AGGAAGGACAGTTGGTGTGG	CCTTCCTATTCCCAGCTTCC	546	113420453	113420998
MMSCL/M079A	AGACGAGGGAAGGACATGAGTG	CCGCTCCACCACTGAAAACT	342	113420838	113421179
MMSCL/M079B	TGCTCCCCCACCTCACTTT	GCTCCGGGCAGAAAACAAC	416	113421248	113421663
MMTAL.177	CCAGCACTAGGTCATTGGGT	CCAAAGCACATCCCATTCTT	506	113421894	113422399

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 4 Co-ordinate Start	Chrom 4 Co-ordinate Finish
MMSCL/M080A	TGCCCTAGACACCAACCAGAAG	AGCGGGCCATCATTCTGTG	436	113422227	113422662
MMSCL/M081A	ATTGGTGGAGGACTTGTTCAGG	GGCCTGCCTAACAAGAAGTACAT	437	113423397	113423833
MMSCL/M082A	TTCCCCACACTGTTTATTCATAGG	GGCAAGGGGGTAGGCAAG	495	113423971	113424465
MMSCL/M082B	CCCCCTTGCCACCCTGT	GGGGATGGGAGTGCCTGTAG	422	113424456	113424877
MMSCL/M083A	CCAGGACCGGCTGAGATCTTA	CAACCCGGCAGATGTGGA	416	113424898	113425313
MMTAL.184	GCCTTTGTCCTGCTCTTGTC	GGATGCTGGGAGTTCAGAAA	404	113425441	113425844
MMSCL/M083B	CCTTCCCTCAGCCAGTGTTG	CAGGTCTTGGGGGTCTATTGAC	341	113425765	113426105
MMTAL.187	ATGCTGGTATGATTGGAGGG	GGTCTCAGCAGGGCAGATAG	585	113426874	113427458
MMSCL/M085A	AGCCATGCCATCCTGAAGC	TGATCCCTCTGCAAACACACG	435	113427369	113427803
MMSCL/M086A	ACCATCTCCCCAAAGCTCACT	GGAGGGGGCAACTGGAAAG	488	113427839	113428326
MMSCL/M086B	TGGCCTGAACCTGCAATGAT	CAGGCCAGCAATGAACCAAG	439	113428384	113428822
MMSCL/M087A	GTCTTGGCCCCTCCATACTTT	GGCAGGTCCGCTCAAGGT	331	113428908	113429238
MMTAL.191	TCTGCCCAGATCATAGTCCC	ATCATCAAGGTGGGAGCATC	517	113429133	113429649
MMSCL/M087B	TGCCCCAAGCTCCCAAAG	CAGGCAGGCGTGGTACAGG	395	113429235	113429629
MMTAL.192	GATGCTCCCACCTTGATGAT	TTGTAAAGCCACAGCCCTCT	477	113429630	113430106
MMSCL/M088A	TGAAGGGCAGCTGAGTGGG	CCCCAGCCCCAAGTCTTCT	419	113430196	113430614
MMSCL/M088B	GGGCTGGGGGAAGGTGA	CACCTCCCAGCCCTCCAG	371	113430606	113430976
MMSCL/M089A	GTGCGGCTCAGTGGTGGA	AACTCACGCCCTGCTTCACA	349	113430974	113431322
MMTAL.195	GGCTAACTTCCCCTCAGCTT	CTCAGGCAGCTGTTCTACCC	474	113431146	113431619
MMTAL.196	GGGTAGAACAGCTGCCTGAG	GATGGGTGAACACTTGGCTT	442	113431600	113432041
MMSCL/M090A	TGACCCGGGAGCTGTTTTCT	GGGATCAAATAAAGGCCTAACATAC	343	113432097	113432439
MMSCL/M090B	TCACCCAATTGCCTAACTCAGA	GGGCCTGTGACCATTCCAC	444	113432545	113432988
MMSCL/M091A	CACAGGCCCACAGCACATG	GTCCCTGTTGCTTGCATGGT	463	113432980	113433442
MMSCL/M091B	GGGGTCTAGCCTGCTCTTGC	GAGCCTGCTTCCCCTATGTGT	340	113433479	113433818
MMSCL/M092A	CTGGTTGGAATGAAAGGAATGAG	GCTGCCCGCTTTCCTTACA	387	113433855	113434241
MMSCL/M092B	CAAGAAGCCCCACCCACTG	CTGGAGGTCTGGGGTGTTTG	398	113434331	113434728
MMSCL/M092C	GTTGGCCTTCAGTCTGCTCGT	AGCCCCCATCCAGAGCAA	398	113434789	113435186
MMSCL/M093A	GGCTAGGGGGTGGTGACATT	GCTCAGGCATTGGGAACAAC	457	113435183	113435639
MMTAL.205	TGTCACCCTTGGAAACATCA	CTGGGGTGTGGAGAGGTAAA	484	113435818	113436301
MMSCL/M094A5	GGGTGGGAATGATTCAGGCT	AGGGCTGGCATCTGGAAGAT	501	113435902	113436402
MMTAL.206	TTTACCTCTCCACACCCCAG	AACACAGCGACAGCAATGAG	553	113436282	113436834

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 4 Co-ordinate Start	Chrom 4 Co-ordinate Finish
MMSCL/M094B	GAGCTGCCCTTTGGTCCCT	CACCCAGCCCTGAGCCTT	442	113436482	113436923
MMSCL/M095A	TTGCCACCATGAGGAAGGAA	CCAGCCCCCAAACCACTC	365	113437003	113437367
MMSCL/M095B	GGGCTGGGAGGGGGTTAA	TGCACCCACATGGAACAGCT	462	113437361	113437822
MMSCL/M096A	GGAGGGGACAGAGGCAGTTC	CAACCCTCCTACCTGATCCACA	476	113437892	113438367
MMSCL/M096B	GCTGGGGGTGTTTTCTGCTT	GGCACCCACTACCTCCTGAAC	374	113438457	113438830
MMSCL/M097A	GGGGAGAAGGGGTTCAATGT	GGGGACTAGGGCCACAGAGA	383	113438833	113439215
MMSCL/M097B	TCAGGCCAGGACTCAGACCA	AACAGCCTCAGCCTCTCCAAC	361	113439324	113439684
MMSCL/M098A	CCTGTGGGTGGCTCTTGATTC	TGGGAAGGGAAGAGAGCTGG	313	113439872	113440184
MMTAL.215	GTACAACACTCATGCCCACG	AGGCAAGCCTTAGACCACAA	547	113440142	113440688
MMSCL/M099A	TTTCCCTCTCTGGTCTTTAGCTTTA	GTGGGGAGAGGGGTTGATTT	382	113440830	113441211
MMSCL/M099B	GGGCCAGGACGGAACCA	CACCGTGCGGCTGACAGT	357	113441235	113441591
MMTAL.218	CAGGGGCTAGGCTGATACTG	GAATCAAGCTTGGGTGGAGA	574	113441360	113441933
MMSCL/M099C	GGAGGGGACCCGGAGCT	CACCTGCCACTCCATTCCAC	409	113441730	113442138
MMTAL.219	GCTGTGGAGTAAACGGTGGT	ACCTTCCCAGGTGCCTTACT	537	113442092	113442628
MMSCL/M100A	ACCAGCCTCAAGACAAGCACA	CCACTGGCGCACTGAAGC	424	113442298	113442721
MMSCL/M100B	AGTGCGCCAGTGGGTTACC	CAGAGCTGCCGATTCCAAAA	294	113442709	113443002
MMTAL.221	CACTCCATGGGAAAGGCTAA	ATTGAAGCTGTGGGGATGAG	484	113443022	113443505
MMSCL/M101A	CATCCCCACAGCTTCAATCC	TGGGTGCTGGGAACTGAATTA	434	113443488	113443921
MMSCL/M102A	AAGAGCCCCTAGAATTGAAATGG	TGGCGGTAGGGGTTTTAACA	415	113444037	113444451
MMTAL.223	TGTCGTGAGATCCACCAGAG	AGCTCCTTCATTGGGAACCT	599	113444109	113444707
MMSCL/M102B5	TGCAGAAGCGGATGCTCAC	GAATCCTTCCCTCCCTGCAC	501	113444646	113445146
MMTAL.224	CAAGGAGCTGAAGGGATTTG	CACCCTTGTAACCTCCTCCA	507	113444720	113445226
MMSCL/M102B	AAGGAGCTGAAGGGATTTGCA	TCCTTCCCTCCCTGCACC	423	113444721	113445143
MMSCL/M103A	TGGGGTGGGGTGAACAGG	ATGGAACGCACGGGATGTAC	388	113445225	113445612
MMSCL/M103B	TCCCCTGAATCCCCTGAATC	CATGGGCCTAGAGGAAAACAGA	459	113445634	113446092
MMTAL.227	TAGTAGTTGGCCACTCCCGT	TGCTGGAATAATCCAAAGGC	592	113446173	113446764
MMSCL/M104A	CAAACCCAGAATAGCACTCAGCA	CCAGCCTCACCATCCTTGTC	427	113446375	113446801
MMSCL/M105A	TGGCACCAGGGAAAAGAATC	GGCTCCAAACCCAAGTCCAG	398	113446805	113447202
MMTAL.229	TTCTGGACTTGGGTTTGGAG	CCCACAGAAGGAAAACCAGA	401	113447181	113447581
MMSCL/M105B	GCCCCTCCAGGTTTTGCTTA	CCCATGCCTAGACCTCACTGAA	455	113447394	113447848
MMTAL.231	TGCTAATCCACATGGCAAAA	TGCCCTCCACGTCCTATTTA	580	113448011	113448590

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 4 Co-ordinate Start	Chrom 4 Co-ordinate Finish
MMSCL/M106A	GGGGCGGAGAGCAATGTG	CCACGTCCTATTTATTTCCTTATTGA	478	113448107	113448584
MMSCL/M106B	GGCACACACCTTTCCTTTTGAT	CCCCCTTCCTTCTCTGCATG	407	113448587	113448993
MMTAL.232	CGCCTTGTGCACTAAACAAA	CACTGCCCCAGATACTGGTT	597	113448801	113449397
MMSCL/M107A	TGTGGGAGGGATGCTGAGTC	AATCCCCTTGCTTCACTGTGTAA	317	113449299	113449615
MMTAL.233	GTATCTGGGGCAGTGCTGTT	CTCTCACCAGCTGTCCATCA	570	113449383	113449952
MMSCL/M108A	AATTTTTGGCTCCTCCTTGCA	TAGATGCTGCTTGCCAAGAGTG	313	113449836	113450148
MMTAL.235	TTTGACACACATCCACACCC	ATGCCCTCTTCTGGTGTGTC	573	113450516	113451088
MMTAL.236	GAGCAGTCGGGTGCTCTTAC	TGCCATGCTTTTTCAGACAG	538	113451157	113451694
MMTAL.237	GCCTCTGGCCTCTATATGCTT	GGCTACTGCCCTGACAGAGT	405	113451726	113452130
MMSCL/M110A	TATGCAGGAAAGAGTCAGGAAGAA	TCGGGGCTACACTGGGATCT	331	113452054	113452384
MMTAL.239	TGGGATTTGAACTCTGGACC	TCCCCAGATATGGCAGTCTC	455	113452522	113452976
MMTAL.240	GAGACTGCCATATCTGGGGA	AGTCCCCCATACCCTACCAC	424	113452957	113453380
MMSCL/M111A	ACGCCAGGGCCAAAAAGT	TGAAAGGGCATGGGGTGTT	428	113453335	113453762
MMSCL/M112A1	TCTTCAGCAAAACAACCTCTCATG	CCCATGGTCCAGGCAAGTG	419	113453853	113454271
MMSCL/M112B	TGGTGAGGTCTGGAGCTACTTGA	GCCCAAACCAGCACAAGTGA	410	113454406	113454815
MMTAL.244	ACCAGCTGTGCTACCCAGAG	CATAAAGGGAAGCAAGGCAG	439	113454707	113455145
MMSCL/M113A	GCTGCCTTGCTTCCCTTTATG	GCTCACGGGCTCAAAGTCAC	397	113455125	113455521
MMSCL/M113B	CACATGGGCCACGAGAGAAA	ATGTGGCCAGCAGAATGTTTTA	307	113455560	113455866
MMTAL.246	CAAGGGGTCAAAGACCTGAA	TGACAACTGAGCAATTTGGC	513	113455738	113456250
MMSCL/M114A	GGGCTTATCTGGCTCTTGCTC	TGGGGTGGAAAGAGTTAGGAGC	446	113456278	113456723
MMSCL/M115A	GGCCCATCAGGTCTAGGTAAAG	AGAGAGCGATGCGAGGTAAGTC	485	113456746	113457230
MMSCL/M115B	AGGCTGCTGTAGTGTGCTCTCTG	GGCAACCCTTCAACCTCTGTAA	471	113457349	113457819
MMTAL.251	CAGTACTGTTCCCCGCCTTA	GGTGTAAGTGCATGTGCAGG	521	113461140	113461660
MMTAL.253	GTCTTAGCATCGATCAGGCA	GTTGCCTCCAAACCTCACAT	417	113463310	113463726
MMSCL/M122A	TGGGGATTAGATTCAGAACAGGC	ACCTCACCAGACTTAATGTATCT	407	113463761	113464167
MMTAL.255	GTCGTTGCCCTTACCACAGT	CCTTAGGCTTGCACTTGGAG	511	113464281	113464791
MMSCL/M123A	GTTCAAAGGCCAAAGTCTACCC	GATGCCGGACAGAGATCAGG	503	113464629	113465131
MMTAL.256	CCAAGTGCAAGCCTAAGGTC	GAACGAGCCTTCTCAGATGC	488	113464774	113465261
MMSCL/M123B	GTCCGGCATCTGTCCTGATG	GCAAATGCTGAGGATAAAGATAACA	404	113465122	113465525
MMSCL/M124A	TGCTATCCACACTTCCAACTTTTC	TAGCTGTCCCCTTTGGTAATGAA	411	113465534	113465944
MMSCL/M124B	CCAAAGGGGACAGCTAAGTCAA	CTTTTAGCTCCTTTAGTTTGATATGTTG	351	113465929	113466279

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 4 Co-ordinate Start	Chrom 4 Co-ordinate Finish
MMSCL/M125A	TGTCTTAACTGAGGTCAATTTTGCT	GGTCCAGTCCCAACAGTAACTTAGT	331	113466392	113466722
MMTAL.260	TGGGTTTAGCTTCTCATCCG	CATTGCATGTGGCTTTCTTG	487	113466510	113466996
MMSCL/M125B	TGCTCCCTTGCTATATATGTATTTGTAA	TTCTTGCAGCACCTATGAGTCTTAC	414	113467016	113467429
MMSCL/M126A	GCAAGAAGAACGTGGAACAGTG	CCATTTGCCAGACGTGAAGTG	419	113467423	113467841
MMSCL/M127A	GAATGAGCCCCCTAAATGGTC	CTTTGGGCAGTTTAGCAGTTAGAC	390	113468293	113468682
MMSCL/M127B	GCCCAAAGTGCCCATCCT	TTCCCTGCTCACATCCTTCCT	434	113468675	113469108
MMSCL/M127C	ACAAGGGCCTGACAAGTGCA	GCACCCCTATTACATGGCTTTC	425	113469109	113469533
MMSCL/M128A	CCTTGCCCAGTCCTTCATACC	ACGATGGCCCACTCTGTGAA	400	113469612	113470011
MMSCL/M128B	GGCCCATCTAAATTCAGGAGTC	AAGGTTGGGTGTTTTGTATGTGTG	315	113470082	113470396
MMTAL.267	AGATGAGGAACACAGGTGGC	TCAGAGAGGGAGGATCAGGA	538	113470159	113470696
MMSCL/M129A	GGCTTCCCCAAACTTCTCCT	GCCAGCGTTCTAAAAACACCA	437	113470661	113471097
MMTAL.268	CGGTTGATCCACACAGTGAC	TTTTCTCTGACGGGCAGTTT	506	113470890	113471395
MMSCL/M130A	TCCTGGTCGAAAATAAACATCAAC	AATTAAAGGTCAACCCAGTGTCAAG	313	113471307	113471619
MMTAL.269	CCTTGACACTGGGTTGACCT	GAGGCAGAGGGAGAAGGAGT	426	113471594	113472019
MMSCL/M130B	GGACTGGGGGTTGAATTTGTG	CTGCCCACTGTTTGCACTGA	460	113471871	113472330
MMTAL.270	ACTCCTTCTCCCTCTGCCTC	TCTGGAAACTTGTCCATCCC	476	113472000	113472475
MMSCL/M131A	TTCCCAGGACCTAAGCTCATATTC	CCAGCAGCAGGGCCAGA	477	113472393	113472869
MMTAL.271	GTTTCCAGAATGGGTGTTGG	GCAAGTATGGAGAGGGGACA	508	113472467	113472974
MMSCL/M131B	TGGGATTTATGGGGTTCTGGA	GACAAACCCCACTGACTCTAATCTC	392	113472876	113473267
MMSCL/M132A	AGGGGGTGAGCTCGTGTCTT	TTACAGCCCCCAGGAGGATAG	393	113473288	113473680
MMSCL/M132B	TGCATCCGTGTGAGTCAGTTG	TGCTATCCTGGGACCTTGACTC	521	113473743	113474263
MMTAL.275	GCCTTGGTCTCTTCTGCCTA	TTTCTGTGCCCTCTGTTTCA	507	113474172	113474678
MMTAL.276	TGAAACAGAGGGCACAGAAA	GGCTCACAACCATCCGTAAT	492	113474659	113475150
MMTAL.277	ATTACGGATGGTTGTGAGCC	GGAAGAATGACTTCCAGGCA	518	113475131	113475648
MMSCL/M134A	TTGGTCAAGGAGTGGCACTATTT	AAGCATTTCTCACAGCATCTTCAA	471	113475540	113476010
MMSCL/M134B	CCGATTGGGCTATGCAGACA	AGGACCCACATGGCAGCTC	329	113476102	113476430
MMTAL.280	GAGCATCAGATCCCCTGAAA	ATCCTGGCACAAATAGGCAC	443	113476372	113476814
MMSCL/M135A	TTGCGCAGTACAGGTTAAAATATTC	GATGTTTTTGCTAGTGGTTCTGATG	463	113476902	113477364
MMTAL.282	ATCTTGACCAACAACCCAGC	TGTATAGTCCTGGATGCCCC	502	113477206	113477707
MMTAL.283	GCCCTTGTTGCTGGGTAATA	GTCCTGCCTGAACGTGTGTA	450	113477581	113478030
MMTAL.284	TACACACGTTCAGGCAGGAC	TGACTCGAGGTCCCTTCAGA	572	113478011	113478582

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 4 Co-ordinate Start	Chrom 4 Co-ordinate Finish
MMSCL/M137A	AATCCCCAGGGCAGACTTTC	CCCCATTGCTCTGCGGAT	345	113478669	113479013
MMSCL/M137B	GCCCCTCCCGCTTTTACA	TGCCGAGGAAGCCAATAGC	394	113479018	113479411
MMSCL/M138A	TCAGGTCGCGCAGCAGTC	CCGCGAACGAAGTCTGTCTG	384	113479489	113479872
MMTAL.288	CAATCGCGGAAGATCTGAAT	GGATGGGGACAGAGAAGTCA	501	113479901	113480401
MMSCL/M139A	TCCCCACTCCTGACTTCTCTGT	TTCCCTCGGCATTCCTATCC	416	113480372	113480787
MMTAL.290	CTGCTCTGGTCTCTGGTTCC	CCTGATCAGCTTCAGGATCA	460	113480674	113481133
MMTAL.291	TGTTTGCATGCACAGAGGAT	GCAGGGTGTGGGTAGAAAAA	441	113481347	113481787
MMSCL/M140A	TCCATCAGGTCCCTTGTCATCT	TGGCAGGGTGTGGGTAGAAA	401	113481389	113481789
MMSCL/M141A	GTCCCTCCTCCTGCTTCTATGA	CAAGACCGGAAACCATCAAATAG	469	113482284	113482752
MMSCL/M141C	GGAGGGTTGGGGTAGGGTAA	AGCAGGAGGCAGAAGGAGTTG	401	113482770	113483170
MMTAL.294	CTGTGTGCAAGCCAACACTT	CGAGCTCTCAGGAACAATCC	589	113483231	113483819
MMSCL/M142A	ATGTCTCGGCTGATTTAATTTGTC	GCCTGCCCCTCTACCTGACT	385	113483612	113483996
MMSCL/M142B	GCAGGCAAGCCATCATTAGG	GCATTGGGAAGCAACTAGCC	379	113483991	113484369
MMTAL.303	GATGGATCCCATTTGTGGAC	CTCTGCCTCTTCCTTGGTTG	551	113484229	113484779
MMSCL/M143A	ACCCATCTCTGCTCCTGTGCT	CGGGAGACTTAAGGATTCCAGA	455	113484552	113485006
MMSCL/M143B	GTCTCCCGTTCTTAATGATCCTTC	GCTGGCCAATCATGAAGACC	342	113484999	113485340
MMSCL/M144A	TGAGGGTGTAGGGCAGTGTTAA	TCATGCTTTCCTGCCCTTCTC	456	113485487	113485942
MMTAL.296	CGTTCAATGAGCTACAGAGGTG	TGTAGCCACTGCCTGTTTTG	497	113485531	113486027
MMTAL.297	AAGGACAAAACAGGCAGTGG	TGGAAGGCATTTGACAAACC	404	113486003	113486406
MMTAL.307	ACGATGCCTACCTCTGGATG	CACCCATAACCAATGGAAGTG	522	113486116	113486637
MMTAL.298	CCTTCCATGCACTGGAGAAT	AAACTGAACCCAAGCCCTCT	471	113486400	113486870
MMTAL.308	AGAGGGCTTGGGTTCAGTTT	ATTTCAGGGAACACCTGCAC	597	113486851	113487447
MMTAL.300	GAAATGAAGCCCATCCTGAA	GTTGTCAAGCTGAGGAAGGC	477	113487443	113487919
MMTAL.309	GAAATGAAGCCCATCCTGAA	GTTGTCAAGCTGAGGAAGGC	477	113487443	113487919
MMTAL.301	GCCTTCCTCAGCTTGACAAC	AGATTGCCCATGAACACACA	478	113487900	113488377
MMTAL.310	GCCTTCCTCAGCTTGACAAC	AGATTGCCCATGAACACACA	478	113487900	113488377
MMTAL.311	GTGTGTTCATGGGCAATCTG	CAAAGAGACATGTGGGGCTT	419	113488359	113488777
MMTAL/302ver2	CTTCGGCTATGGGAAGTCAG	ATCACAGGACAAATGGGAGC	572	113488395	113488966
MMSCL/M147B	TGGGGACAGAATCTTGCTTTTT	GAAACAGGCATAGCAGTCCAAATA	402	113488717	113489118
MMSCL/M148A	ACCTTGCCAGCAGAGTCTCG	TGGGGGGTAGCACTCTCCTT	403	113489296	113489698
MMSCL/M154A	GGGGGCATCCTGATGTCAC	TGGTCCTCCCTTGCTGGTT	273	113495553	113495825

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 4 Co-ordinate Start	Chrom 4 Co-ordinate Finish
MMTAL.317	ACACATGCTGCAACTTGGAA	GGCAGATTCTGTGCAAGTGA	463	113495858	113496320
MMSCL/M155A	CGACGATGCTGGAACCACTT	GTGGCCTCTTCCCGCTATTAT	431	113496375	113496805
MMSCL/M155C	AGGAGGCCAAATAAAAGTGGG	GGTGCCTGATTTGGAGAGTGTC	387	113496989	113497375
MMSCL/M156A	GCCATGGTCCTTGCTGAATC	TGGCAGCTAAAATATGGGAAGTT	414	113497386	113497799
MMSCL/M156B	TTTTAGCTGCCAATCTTCTGTCAT	CCATGCTGCAGTTCCTCACA	373	113497788	113498160
MMTAL.322	CCAGGTCATGGCATAAAGGT	GGGATGGTTTTTCCTCTGGT	440	113498181	113498620
MMSCL/M157A	TCTGCCCAGCTCTTTTGTGTTA	CATCCTGGCTGCTTTCTGTCTT	475	113498685	113499159
MMSCL/M158A	ATGCAGAAGGGATGGAGAGAGA	CTCTGTGATGGTCCCTTGTTCC	518	113499157	113499674
MMSCL/M158B	GGAACAAGGGACCATCACAGAG	CTTTGTCATTCCGAGGTGTTTATATT	436	113499653	113500088
MMTAL.326	CACCTCGGAATGACAAAGGT	TAAGGTGGTGATCCCAGAGC	537	113500071	113500607
MMSCL/M159A	AGCTTTGGCCATTTCTTTGTGA	AATGGCAGGAAAAAATGAGGAA	384	113500489	113500872
MMSCL/M159B	CTCCTTGCGCTCTGACTCCTT	GTTGGGGGACTCACATGGAGA	416	113500971	113501386
MMSCL/M160A	GGGCTGCCTTGTCTGGTCT	GCGCAGCATCAATTCCTTTT	389	113501447	113501835
MMSCL/M160B	GCTGCGCACTGTGAAGGATAT	TGGGAAGGACAGTCTATGGGC	356	113501829	113502184
MMSCL/M161A	GGGGGAAGACGGAGAACAGAT	TTCAGGCAAGGAGATTTTACCAA	275	113502247	113502521
MMTAL.331	AAACTCTGTCTCGACCCCCT	GTCTGATGCTCGACCTGACA	425	113502688	113503112
MMSCL/M161B	TCAACCTGCCCTTACCTCCA	TGGCTCCTGCAATACCTTTCA	430	113502790	113503219
MMSCL/M162A	CAGGGGGTTTTACTCTTGGCA	TTTCCATTCCCGCGTTAGTCT	417	113503307	113503723
MMSCL/M162B	ACTCTTTTTGCCTCATGAACAGC	CATTGATTGGGAAAAGACTAAAGC	409	113503776	113504184
MMSCL/M163A	AGGTAAGGGGGTGACACAAATG	ATGTTTTGCTGCGTATTAGAGTTAGG	314	113504196	113504509
MMTAL.334	GCGTTAACCTTTGCTTCCAA	ATCTCCACCATGGGCTACAG	534	113504305	113504838
MMSCL/M163B	ATGGCCAAAAACAATGTCTCTGA	TTGGCACATGGCATGATCCT	365	113504702	113505066
MMTAL.335	GAACAGGATCATGCCATGTG	TTTGACCTATTGCATACCATCA	597	113505043	113505639
MMSCL/M164A	GCACCACAGCAGCCTAAAACTT	TAAAGTTCTGGGGAGTAATTGTTGC	425	113505700	113506124
MMTAL.337	GGAGCAACAATTACTCCCCA	TGCAAGAGGGCCAAGATACT	588	113506097	113506684
MMSCL/M164C	GCAACAATTACTCCCCAGAACTTTA	AGATCTACCCAGGTGTTGTTCATAAT	402	113506100	113506501
MMSCL/M165A	AAGGCAGGGAGGGGTGAAGT	CTAATGGCGCTCACTTTGGC	364	113506535	113506898
MMSCL/M165B	CAAAGTGAGCGCCATTAGCA	AAGCACAAACCAGCCCAGAA	427	113506881	113507307
MMSCL/M166A	CTGGGCTGGTTTGTGCTTGT	GGCCAAAAATCAAGCACAGC	416	113507290	113507705
MMTAL.339	GCTGGTTTGTGCTTGTTTGA	ATCAAGCACAGCAGGAGAGG	404	113507294	113507697
MMTAL.340	CTGCTGTGCTTGATTTTTGG	TGGACTGATGATGGCTTCAA	542	113507684	113508225

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 4 Co-ordinate Start	Chrom 4 Co-ordinate Finish
MMTAL.341	TTGAAGCCATCATCAGTCCA	TATCTGTCCAGGCTGTGGTG	486	113508206	113508691
MMSCL/M167A	CCATCTGCAAAACGCTCTCTACT	CAGGCTGTGGTGGAGAGAAACT	436	113508248	113508683
MMTAL.347	ATGGACACCTGTCTTCTGGC	GAGAGGAAACACAAAAACCCC	543	113515852	113516394
MMSCL/M175A	TTTTCCCATGTGACGGTGTATC	CAGTCCCACAACACCAGAAGAA	414	113516508	113516921
MMTAL.349	AGGAACTTGTGCCACTGGTC	AAGGCACCACAGGTTTTCAC	585	113516976	113517560
MMSCL/M176A	TGAGGTCCAAGGCAGCTAGG	CTCCCCACACTGCTCAACCT	353	113517314	113517666
MMTAL.350	GTGAAAACCTGTGGTGCCTT	CACTGGAAAGAACACGCTCA	419	113517541	113517959
MMSCL/M176B	GGACGGGGTTTTGCCATT	CCACCACAATCCTCCCTCAA	383	113517860	113518242
MMTAL.351	CGGGATAGAGCAGATCAAGG	TTGCCTCAATGATGATGGTC	586	113517972	113518557
MMSCL/M177A	TGTGGTGGCATTTGAATTTGG	GCTGTGCTTCCTGCCATAATG	482	113518235	113518716
MMSCL/M177A5	CTCAAAGCAGTGCAGGCCA	TGCTGTGCTTCCTGCCATAAT	401	113518317	113518717
MMTAL.352	TCATTATGGCAGGAAGCACA	AAGGCCATTCTTCCTTCTCC	448	113518695	113519142
MMSCL/M177B	CTGGAGGAGCTGAGAGTTCTACAA	GGCTGAACATTTGGCATTTAACA	453	113518738	113519190
MMTAL.353	GGCCTTCTCCAGTAAGAGCA	TTCAGATTTCCCCTCTGGAA	419	113519137	113519555
MMTAL.354	TTGTGGAGCACCCTCATACA	CCCACAGCCATCCTGTTACT	527	113519486	113520012
MMSCL/M178A	TTGGAGTTTGGTTTTGCTTTGA	ACCCCCACCATGAAGAACCT	418	113519839	113520256
MMTAL.355	AGTAACAGGATGGCTGTGGG	ATGTGGCACTGTGCTCTCTG	574	113519993	113520566
MMSCL/M179A	GCAGGCAGAGGGTAAGAGAGG	AGAGGCAGAGCACACTTCCAA	406	113520455	113520860
MMTAL.357	AATGTGAGACTGTCCCCTGG	GGGCTAGTGAGATGGCTCAG	403	113520952	113521354
MMTAL.358	CCCCCTGTTTTTGTTTTCCT	AGCTCCTAGGCTTCTGGGAC	476	113521352	113521827
MMSCL/M180A	TCCCCACTGCTCCAAGGTC	AAGAAGGGAGCCAGGTGTAAACTA	389	113521792	113522180
MMSCL/M181A	GTCCCCACCTTCTTTTGATGTATAC	CAGGATGGCAGCAGAGTTGG	424	113522196	113522619
MMTAL.360	GTTACCCACTGACGCCTTGT	ACTTTGCACCCTACCTCCCT	466	113522361	113522826
MMSCL/M181B	TCGGTCCACCCCAACATCTA	TGGGGCTGGGAGGGAAA	393	113522641	113523033
MMSCL/M181C	GCCTTTCCCCCTTGATCACT	ACTTCTACCTGGTTACCTGTGTTTTC	250	113523170	113523419
MMSCL/M187A	GCCCAGAAACTCAGAATACCCA	AGGGAGCTCAAGGGATACTGGT	426	113528151	113528576
MMTAL.363	TGGAATCTTCTGCACCTGAG	TGCCTGCAGAGAATGTTCAG	585	113532172	113532756
MMSCL/M191A	TTAGGTCTTGGATGGTTTTGTTCA	TCCCAGGTGAAAGAGGCTAACA	397	113532312	113532708
MMSCL/M191B	TGCAGGGAAGGTACACAGAGGT	AACCCTCAAGCTTCTGCCATAG	478	113532768	113533245
MMSCL/M192A	TCACCTTCTCATTTCCTCCAGTCT	ATGGCCCTGTCTTCTTTGCA	409	113533304	113533712
MMSCL/M192B	CAGGGCCATAATCATCAGGG	TGGGGTCATGGTTCAGTTCC	426	113533704	113534129

Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 4 Co-ordinate Start	Chrom 4 Co-ordinate Finish
MMTAL.367	GGTCCAGATTTGGGAACTGA	ATGCCTCCATGAGATCCAAC	474	113534098	113534571
MMSCL/M193C	TCCTGGGCCTTCAATTACAACT	GCCATTGCTGTTTTGATTGTGA	412	113535111	113535522
MMSCL/M193C5	GGCTGCACCTTCACCAATG	AAGATTCTGTAGTTCTGATTAGCTGGG	367	113535135	113535501
MMSCL/M194A	GCAATGGCCCTGAAAAGAGTC	CCAGGAGAGGCAGGCTAGTAAAG	424	113535515	113535938
MMSCL/M194B	GCCTTTCCCCCTTGATCACT	TAAGCTGCCCCTGAGATGGA	420	113536005	113536424
MMSCL/M195A	TCCAAGGGCACAGAACATAACA	TTATCTGACCTCCCTGACTCTGG	440	113536540	113536979
MMSCL/M195B	GTCAGGGAGGTCAGATAACAGCA	TCCTGGATGGTTTTGTTCTCCTA	316	113536962	113537277
MMSCL/M197A	ATTAGCCCAGAAACTTAGGATACCC	GGGTTGCAGATTTCTTTAGCTCC	415	113539004	113539418
MMSCL/M198A	CCCTGGGGATTGAAAGCC	AGATATTTACTGGCTTTTTGAGTTGG	256	113539724	113539979
MMTAL.376	ACCCAACTCAAAAAGCCAGT	AAATACTCTCGGTGGGGTCC	593	113539952	113540544
MMSCL/M199A	GGGACCCCACCGAGAGTATT	GGTGAGTGGTCTTGCTTTCTGTTAG	495	113540524	113541018
MMTAL.377	GGACCCCACCGAGAGTATTT	GGGTTCTGGTGATGGTGAGT	507	113540525	113541031

Appendix 3A

Sequences of Primer Pair used to Amplify PCR Amplicons Containing GATA-1 Binding Sites (included as controls on the SCL tiling path array)

PCR amplicon	**Forward primer sequence (5'→3')**	**Reverse primer sequence (5'→3')**
HB009BG	TGACCATGTGATTCTGCCGCTTCTAGGT	GACCTCCCATAGTCCAAGCA
HB032BG_1	TGACCATGAGCAACAGTTTCAGTGCAGG	CATGAATACTTCCTGCCATGTT
HB032BG_2	TGACCATGGCCACTTAACATGGCAGGAA	GGAACCCAACCAGACTCTCA

Appendix 3B

Sequences of Primer Pairs used to in Real-Time PCR to Analyse SCL Expression in K562, Jurkat, HL-60 and HPB-ALL

Gene	Forward primer sequence **(5'→3')**	Reverse primer sequence **(5'→3')**
β-Actin	AGAAGGAGATCACTGCCCTGG	CACATCTGCTGGAAGGTGGAC
SCL	TTTTGTGAAGACGGCACGG	TGAGAGCTGACAACCCCAGG
SIL	ATGCACATAACGTGGATCACG	TCCATGCTCAAATCCACACC

Appendix 4

Sequences of Primer Pairs used to in the Real-Time PCR Verification of ChIP-chip data for Histone H3 K9/14 diacetylation

Region Name	Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 1 Co-ordinate Start
KCY +1	HSTAL.277q	TGGTTGGTTAGCTGCATTGAC	CCTTCTTTCCCAAGCATTCC	75	47512689
	HSSIL/M33Aq	GCGCAGAGGTTAGCGTGTC	GCCTCTAACCCAAATCCGC	71	47512136
KCY -4	HSSIL/M2Aq	AGTCTTTGTTGTTTCCATGATAGAAC	GGTTTACTAGGAAGTTTCAGACCC	80	47508213
NC i	HSSIL/M10Aq	TCTCTTTGAACACAGGGCAATG	TATTAGTCTAGGTGTACTGGCAGTTG	71	47499807
SIL -1/+1	HSSIL/GAP/M4Aq	CAGTCGCCGACCAATGATC	GCTAGGTAGACGGAGGAGCG	73	47492145
	HSSIL_GAP/M5Aq	GCTCCTACCCTGCAAACAGAC	GGAAACCAGGAGCACAAAGC	71	47491699
NC ii	HSSIL/M51Bq	TGAATGCTTCCCTTGTGATG	GTAATGTTTCCTTACTGGTTAGCAAC	71	47467138
NC iii	HSSCL/M15Bq	GTGCCCTTGAGAGCCTAGGG	CCTCAACAGCCTGTCTTATAATTG	71	47440083
SCL -9/-10	HSTAL.175q	GGCCAGAGTTCAAATCCTGAC	CAAGCGTAAAGTGACATGCCC	71	47419831
	HSSCL/M36Bq	AGAGGAAGGACCTTCAGCTCC	TCCTCAAGGCAGAGAGAGCC	71	47418997
SCL +1	HSSCL/M46Aq	TTCCCCCTTTTCCTTACGC	CGCACTCTCACAATCCCACC	102	47409535
	HSSCL/M46Bq	CCGTTGGTGTTCTCAGCAGG	CACCCAAACACAGTCGCAG	71	47409309
SCL +2	HSSCL/M47Aq	AGGCCCCGAAAGGACAATG	CAGGAGGTGATCCCGAATG	73	47408628
	HSSCL/M47Bq	AGCCGTTTTCTAAGTTGCTGG	CCAAGTCTCTGTGTCCGTGC	71	47408265
SCL +3	HSTAL.157q	TTTCGAACCCTCCAACTGG	CAACCGCGTAGACACCTCC	72	47407338
	HSSCL/M48Aq	ACTGAACCAGACCGATCCCAG	AACAACAACCCCTCCCGAC	71	47407303
SCL +4/+5	HSSCL/M48Bq	ACCGCAGCGTAACTGCAGG	CGAGGAAGAGGATGCACACC	73	47406956
	HSTAL.156q	GAAGCCGAGGAAGAGGATGC	CGTAACTGCAGGCCTCTCAG	71	47406951
	HSSCL/M49Bq	GGAGAGGACATTTGTGGCCAG	CGGTGGATTCCTGAGAGGC	71	47405741
SCL +8/+9	HSSCL/M53Aq	CAGTCAATGAACCTGGCGG	CCTAGCTCTCTGCCCTCACC	73	47402359
	HSSCL/M53Bq	CACCCTGCAGCTGAGCTAGG	TGCAGGACTGTGAGTGTGGTC	71	47401484
NC iv	HSTAL.138q	CATCACCTGCAAAATGGAGG	TAAGCTGAGGCAGGCATTGTC	103	47398315
SCL +19	HSSCL/M63Aq	CACGTGCGATCTATCTCTTCG	GGCAAATGCTGAAAGGAACC	71	47391506
	HSSCL/M64Aq	TCTGCTGTAGCCATGGTCCTG	GCAACAAAAGCAGTGCAAGG	71	47391323
SCL +20/+22	HSTAL.122q	CCCAGTGGTCTGACTCCAAAG	GCACAGAAGGCAGTGAATGG	74	47390543
	HSSCL/M65Aq	GATACAGAGCCCTTCCACCC	CAGAGCAGGATCTCCCGTG	74	47390110
	HSTAL.121q	TTCGAACGGATCACATCCTG	TTGGTCCGAGCTCTGCCTC	76	47389763

Region Name	Amplicon Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 1 Co-ordinate Start
	HSSCL/M66Aq	CCTCACCTCTAGGCAGCCAG	GCTTTGGATCAGACACACGTG	71	47389401
	HSTAL.119q	CTTTGCAGCATTCAAGGCC	GCAGCTGGTAAGGCACCTG	71	47388329
	HSSCL/M67Aq	CGAGGCTGCTTAGAGAGAGGC	GGACCATAGCACCGCAGTC	101	47388075
NC v	HSTAL.108q	GGATTGAGGAGAGGGCATGTG	GCACGGCTGTGGAGCTATG	101	47377642
NC vi	HSTAL.106q	CAGCAGAGGTCCCAAAGCC	CAGTACTCCCAGCTTGCTTCC	101	47374561
SCL +43/+45	HSSCL/M87Aq	CTTTCCCTAGAATCCAGCCCC	GGAGAGTCCCAGCCTCACC	71	47367938
	HSTAL.98q	GCATTCGAGGTCATCTTCCAG	GGACTCTCCCACATCTGCTTG	71	47367876
	HSSCL/M88Aq	GCCTGTCACTTGTTTTCAACG	CAAATCCTGTTCCTCCCTGAG	73	47367324
	HSSCL/M88Bq	TGTCTCCCAGGTCTTGGAAGC	TCTTGCCGTGCTCTGTGAC	71	47366737
	HSTAL.95q	TGGCTCACTCCTGCTCAAATG	GAGAAGAGGTCAGGCCTCCAC	78	47366287
	HSSCL/M90Aq	GGTCCCCAAGACCCAGAGTAG	CAGAGAGAGGAGTGGACCAGG	76	47365683
	HSTAL.93q	GCCCATCTGACCCACTTATGC	TGTTGTTCCTCCAGCTCCC	71	47365279
SCL +51	HSTAL.84q	TTAAGCCGAAGCCCAGAGAG	GCTCCAGGCCTATCCTTGC	71	47359910
	HSSCL/M96Aq	CACATTTCCTCCCAGCTCTGC	GGCTGGTGGAGTGACCTGAC	76	47359505
	HSSCL/96Bq	TGACCTTACAGCCCTTCACCC	AGCTCCCTGCTCCCAGCAC	72	47359190
SCL +52/+53	HSTAL.82q	TGGAGGAGGAGAAAGGCAAAC	CCATCCATCTCTGTCTCCCTG	96	47358590
	HSSCL/M97Aq	GCTCTCAGCCCAGAATGTCC	CAAGGTGCAAGCCCTGTTC	71	47358287
	HSSCL/M98Aq	TCCATGATCAGCGTAGATGCC	GGAGGAAGTGCTGAACCCAG	71	47357425
NC vii	HSTAL.77q	TTCTGTACCTGCCAGCCAAG	CCCGACGAGCGTTATGTAAG	71	47355954
NC viii	HSSIL/M55Aq	TCATGATGATATTTAGCATACTCAGCAAAG	GGAGAATGATAACTTGTGTCAGGC	91	47342423
NC ix	HSSCL/M137Aq	TCTCTGGAAGTCATAAATACAACA	AATCTGCTCATCAAGTAATACG	71	47318153
NC x	HSSCL/M182Aq	TTTGCAGTGCCCTGTTCTTAG	TGTTGGCTACCTTGATCATGTG	71	47273688
NC xi	HSTAL.7q	TCATGCCATTTCCGTTGTAC	TTGAACACTTGGAGATGATGATG	71	47273478

Appendix 5

Sequences of Primer Pairs used to in the Real-Time PCR Verification of ChIP-chip data for GATA-1

Amplicon Name	Alternative Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 1 Co-ordinate Start	Chrom 1 Co-ordinate Finish
SIL -13	HSSIL/M5Aq	CAGGGTATATCTATGTTCCTAGCAC	GATTGATGAATGGTGACAAAGC	71	47504296	47504366
NC i	HSSCL/M15Bq	GTGCCCTTGAGAGCCTAGGG	CCTCAACAGCCTGTCTTATAATTG	71	47440083	47440153
SCL -9/-10a	HSTAL.175q	GGCCAGAGTTCAAATCCTGAC	CAAGCGTAAAGTGACATGCCC	71	47419831	47419901
SCL -9/-10b	HSSCL/M36Aq	GGCCTGTGCTGTGACTCTCC	CCACAGATGGTCTCTGTGCTG	71	47419540	47419610
SCL -7	HSSCL/M38Bq	TTTGTGCCCTGTGTGCCTG	TGATTAGCATACCCTGGAGCC	73	47416835	47416907
SCL Prom 1a	HSSCL/M45Aq	GGATAGGGAGACTGCCCATTG	CACCTCCCAGGGCTTCTTTC	76	47410385	47410460
SCL +3	HSTAL.156q	GAAGCCGAGGAAGAGGATGC	CGTAACTGCAGGCCTCTCAG	71	47406951	47407021
SCL +7/+8	HSSCL/M53Aq	CAGTCAATGAACCTGGCGG	CCTAGCTCTCTGCCCTCACC	73	47402359	47402431
NC ii	HSTAL.108q	GGATTGAGGAGAGGGCATGTG	GCACGGCTGTGGAGCTATG	101	47377642	47377742
SCL +51	HSSCL/M96Bq	TGACCTTACAGCCCTTCACCC	AGCTCCCTGCTCCCAGCAC	72	47359190	47359261
NC iii	HSSCL/M182Aq	TTTGCAGTGCCCTGTTCTTAG	TGTTGGCTACCTTGATCATGTG	71	47273688	47273758

Appendices 6 to 11

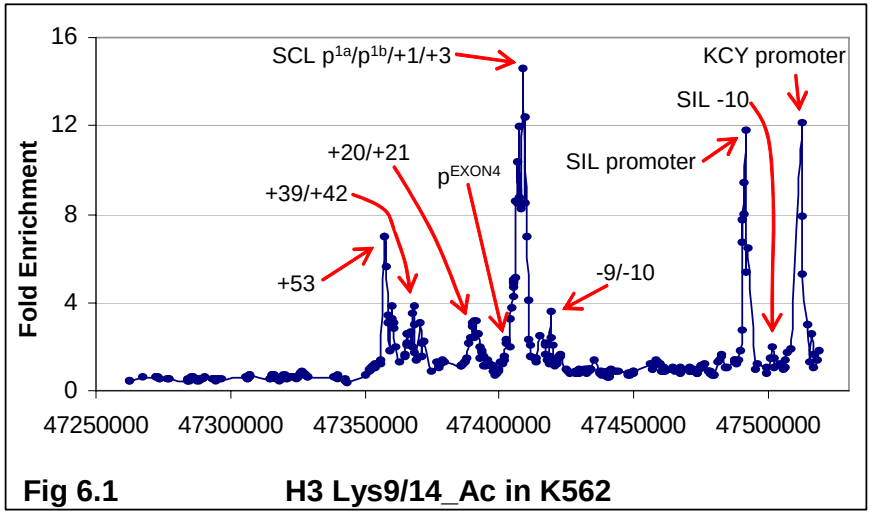


Fig 6.1 **H3 Lys9/14_Ac in K562**

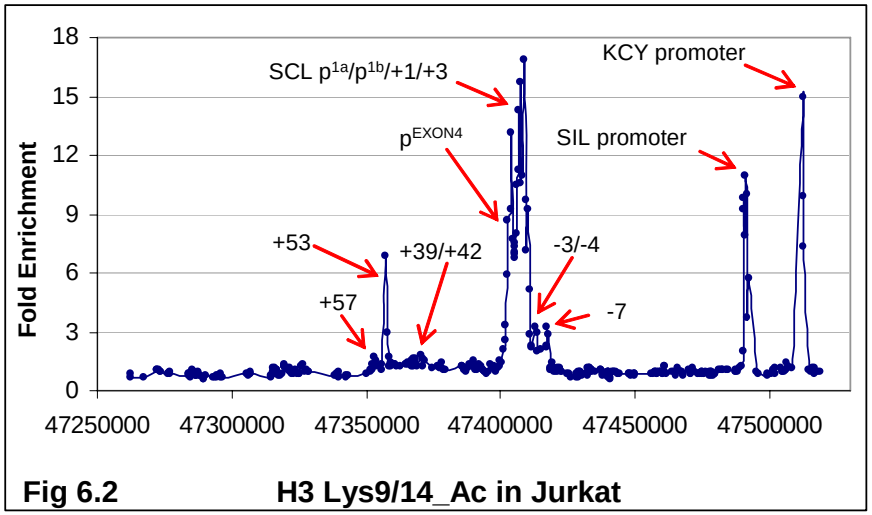


Fig 6.2 **H3 Lys9/14_Ac in Jurkat**

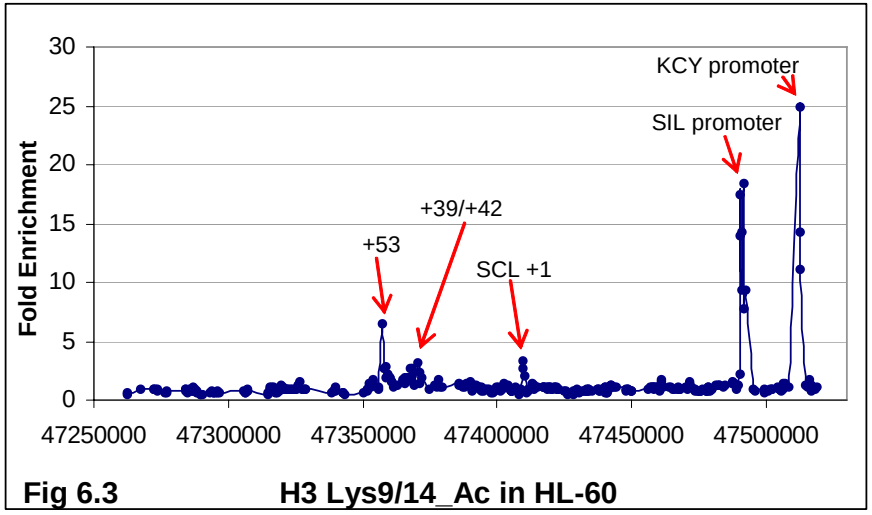


Fig 6.3 **H3 Lys9/14_Ac in HL-60**

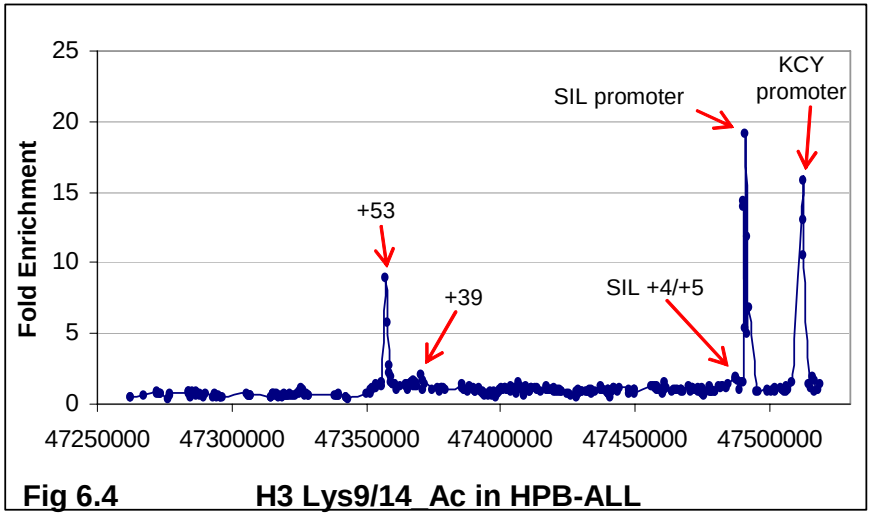


Fig 6.4 **H3 Lys9/14_Ac in HPB-ALL**

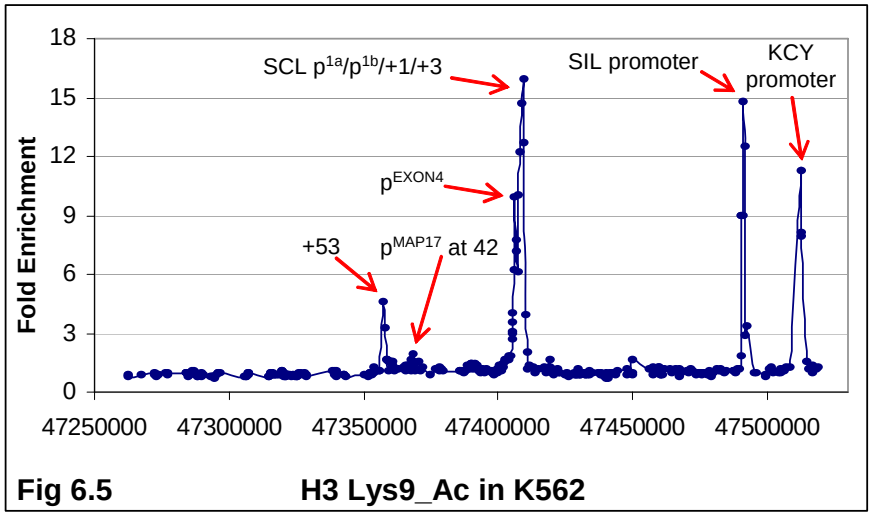


Fig 6.5 **H3 Lys9_Ac in K562**

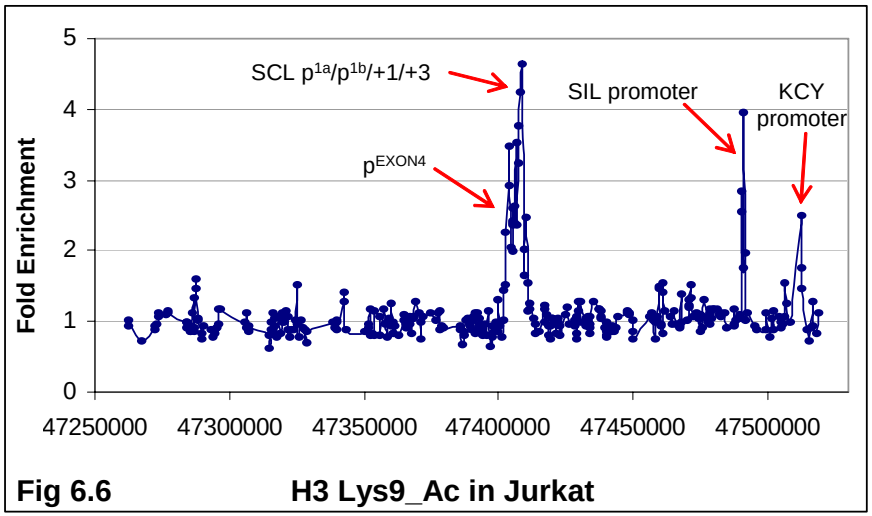


Fig 6.6 **H3 Lys9_Ac in Jurkat**

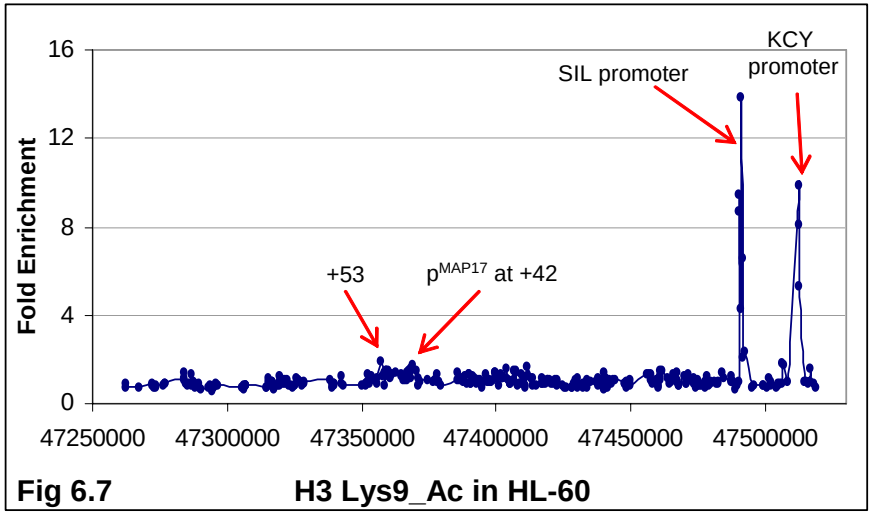


Fig 6.7 **H3 Lys9_Ac in HL-60**

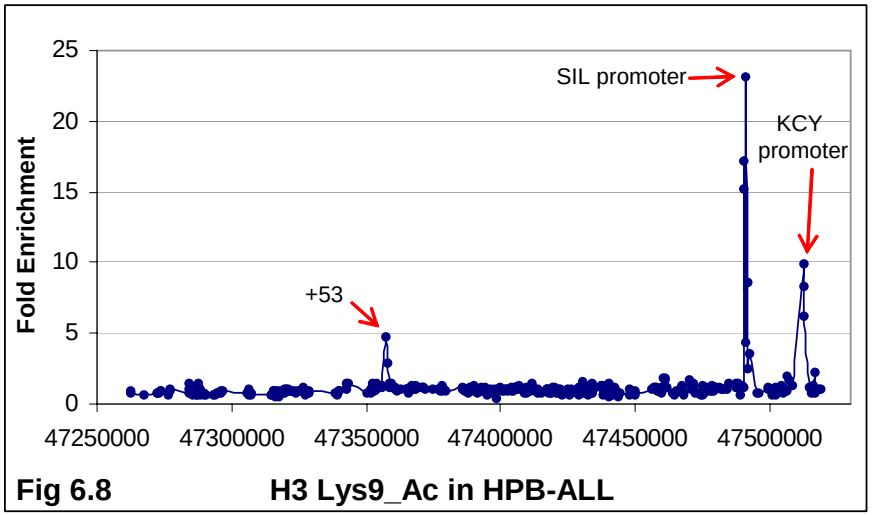


Fig 6.8 **H3 Lys9_Ac in HPB-ALL**

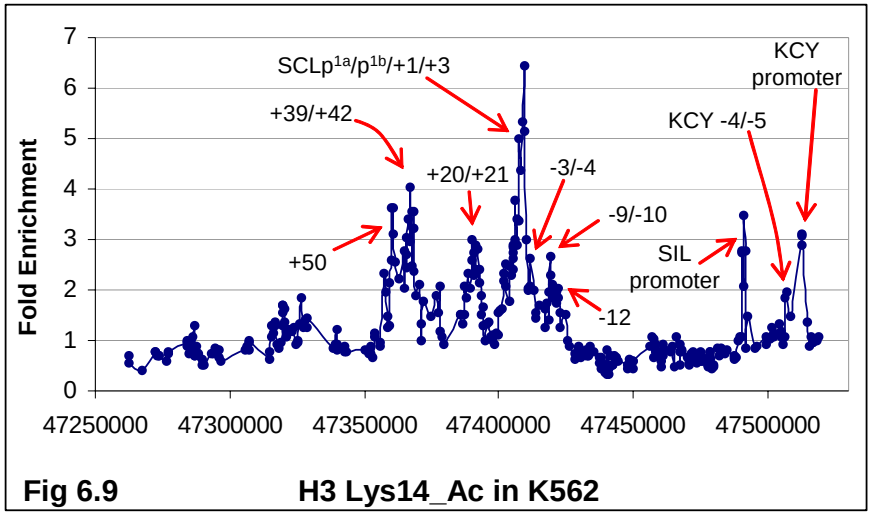


Fig 6.9 **H3 Lys14_Ac in K562**

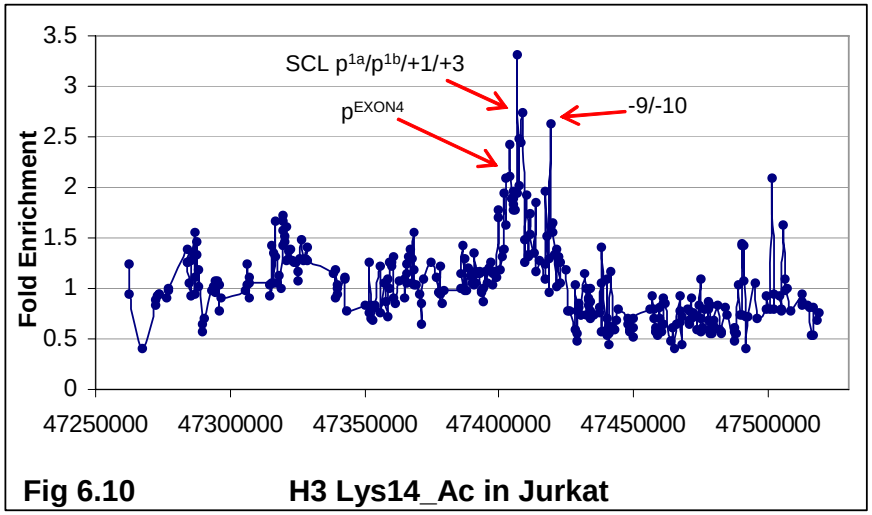


Fig 6.10 **H3 Lys14_Ac in Jurkat**

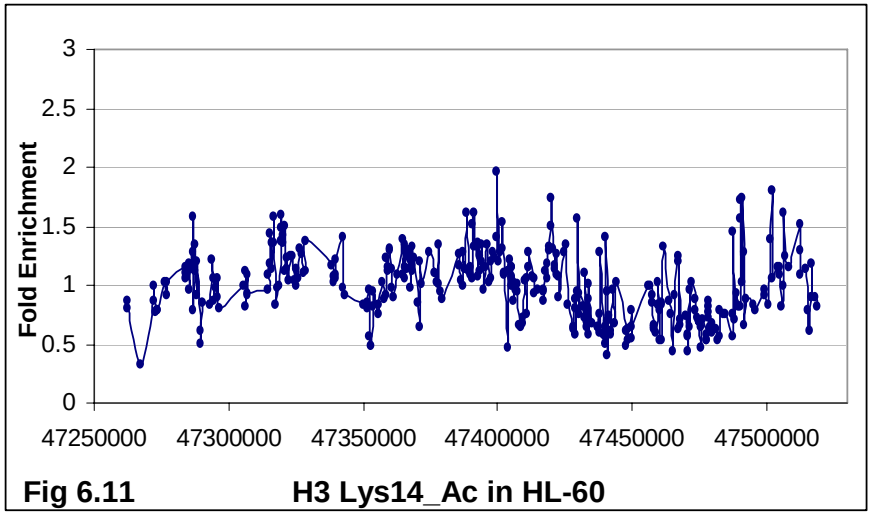


Fig 6.11 **H3 Lys14_Ac in HL-60**

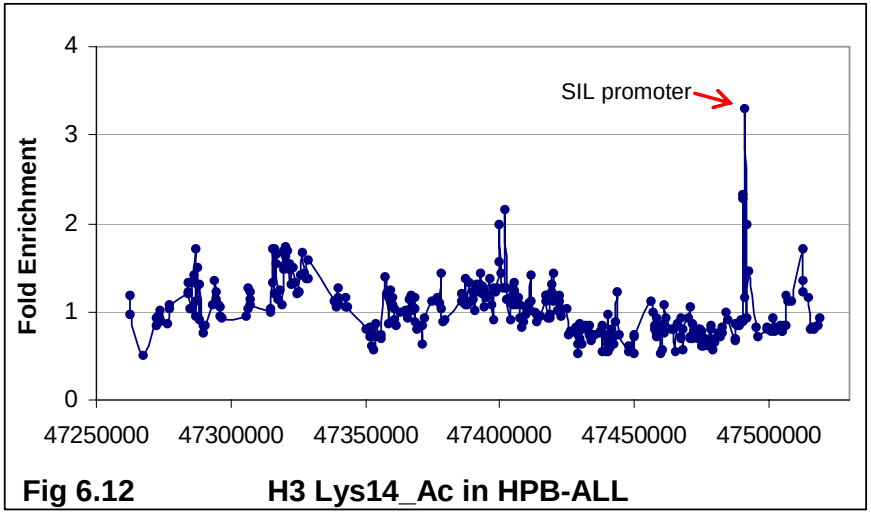


Fig 6.12 **H3 Lys14_Ac in HPB-ALL**

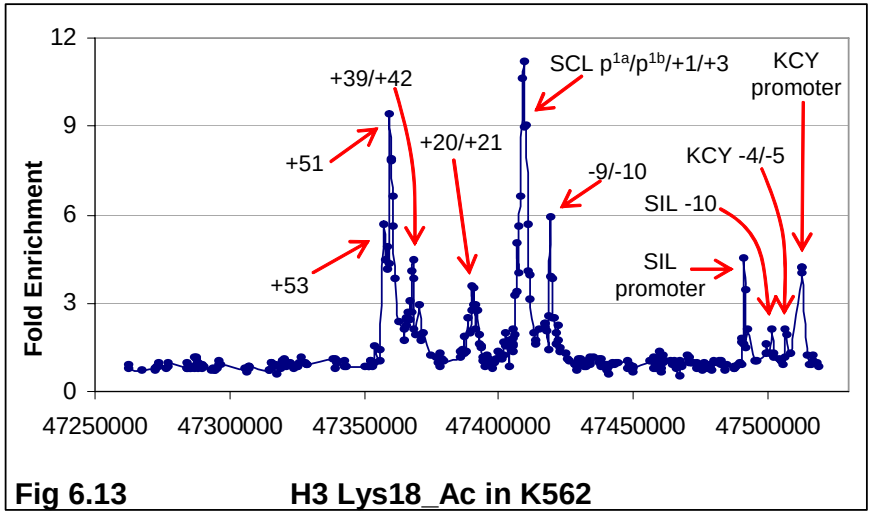


Fig 6.13 H3 Lys18_Ac in K562

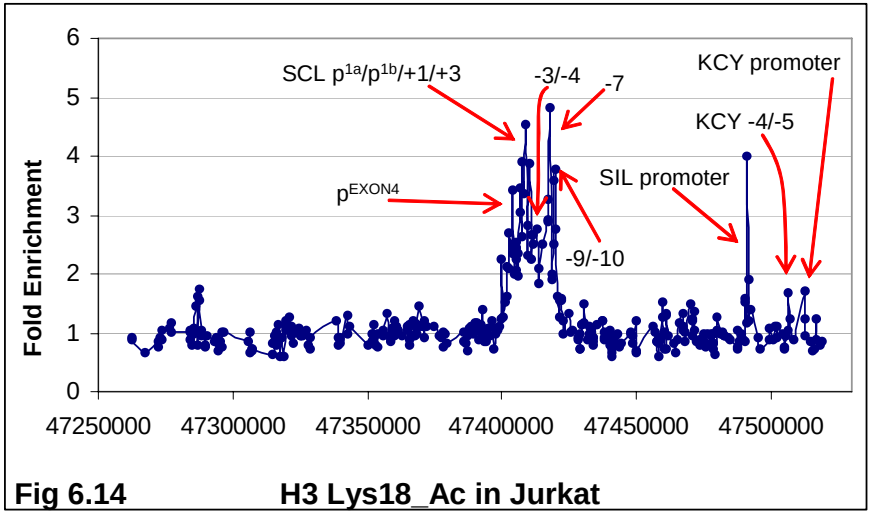


Fig 6.14 H3 Lys18_Ac in Jurkat

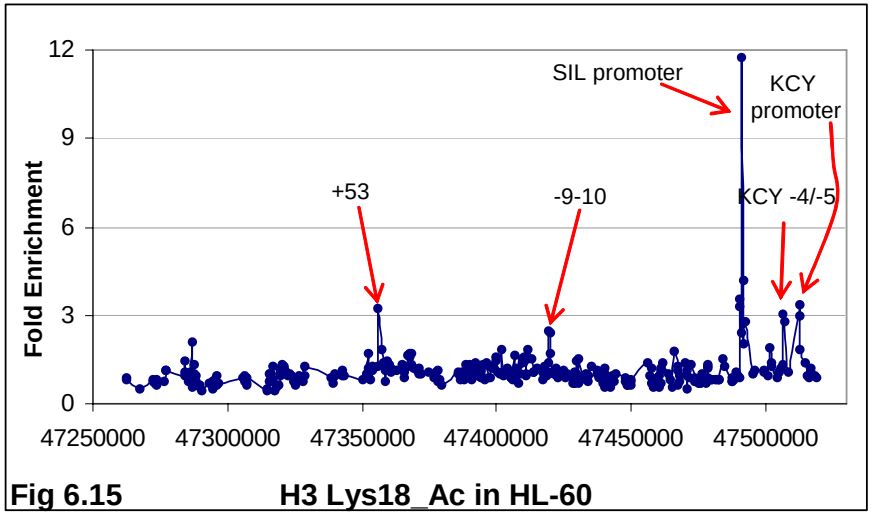


Fig 6.15 H3 Lys18_Ac in HL-60

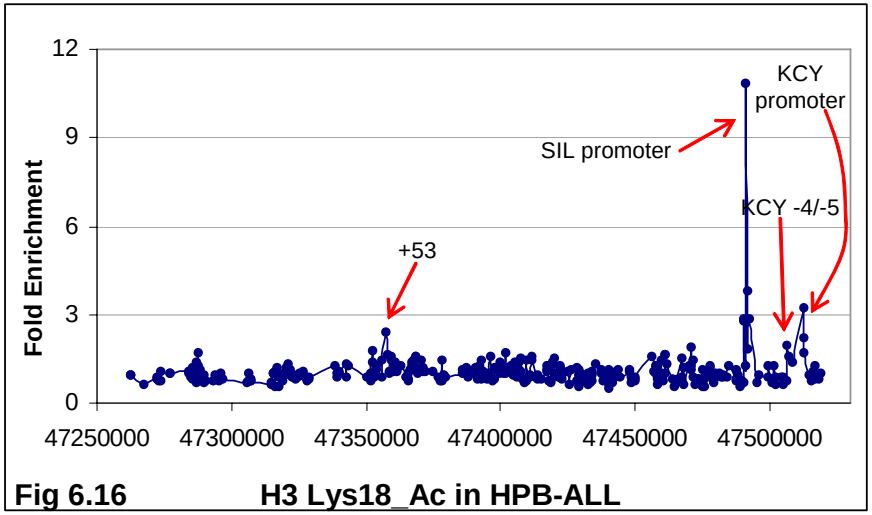


Fig 6.16 H3 Lys18_Ac in HPB-ALL

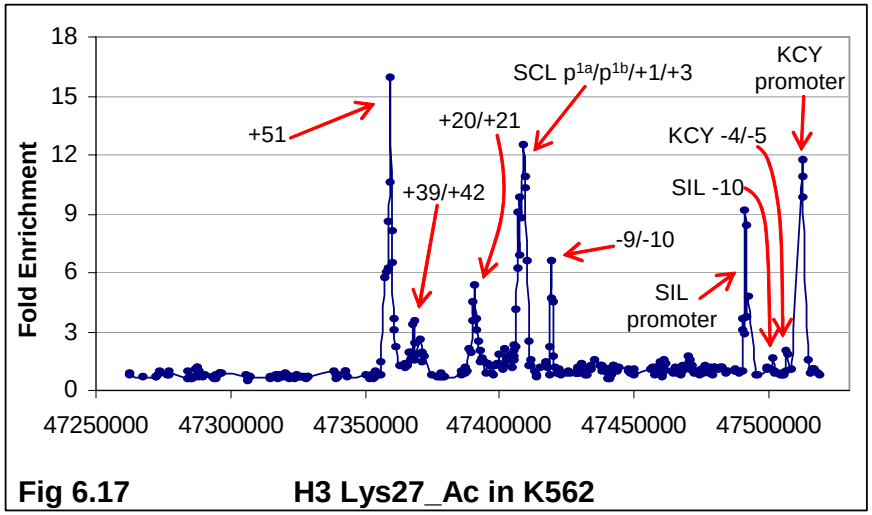


Fig 6.17 **H3 Lys27_Ac in K562**

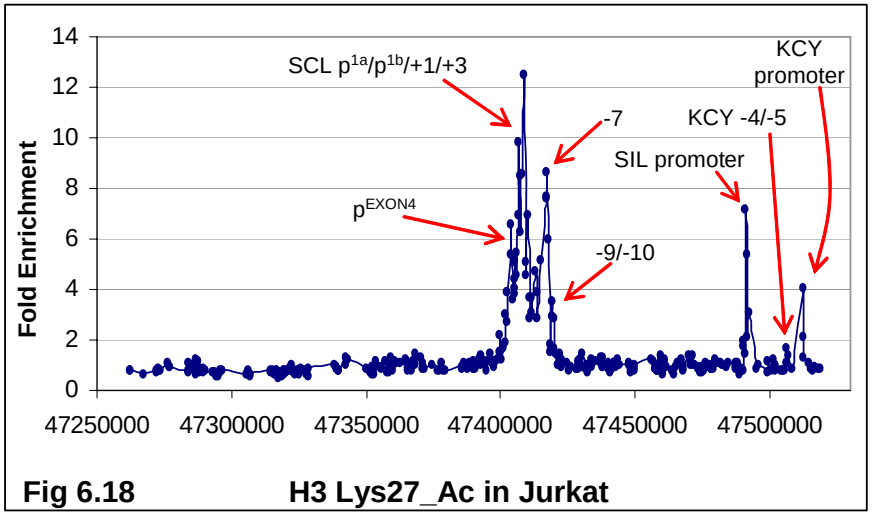


Fig 6.18 **H3 Lys27_Ac in Jurkat**

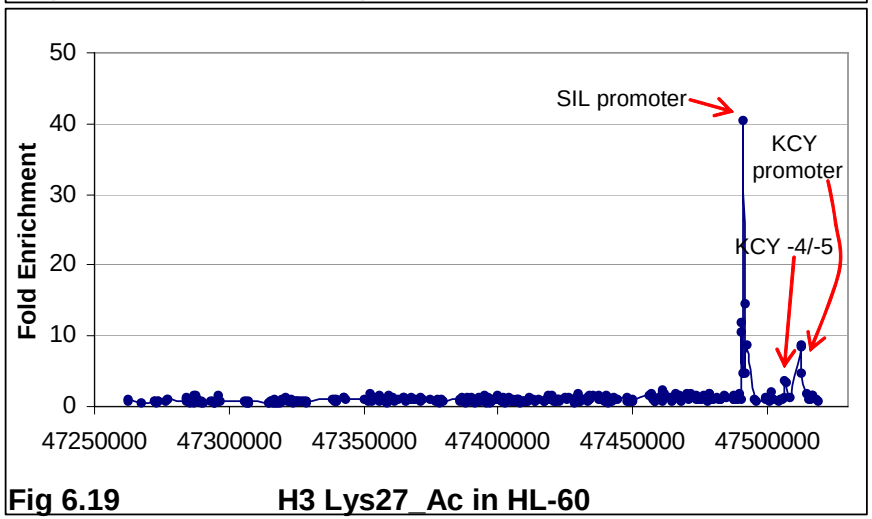


Fig 6.19 **H3 Lys27_Ac in HL-60**

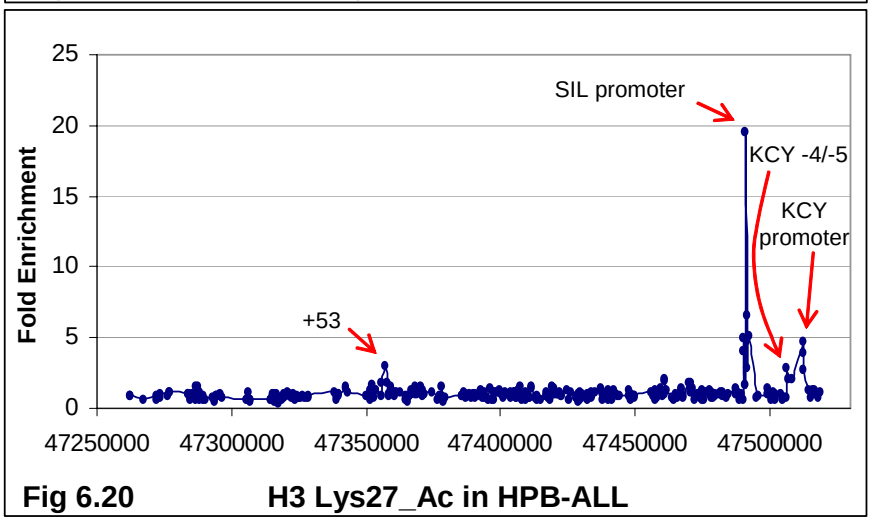


Fig 6.20 **H3 Lys27_Ac in HPB-ALL**

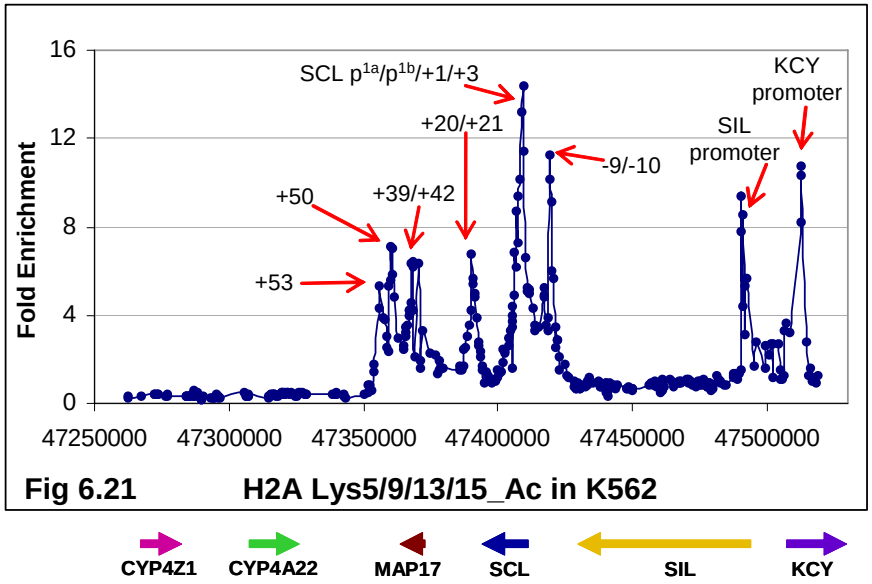


Fig 6.21 **H2A Lys5/9/13/15_Ac in K562**

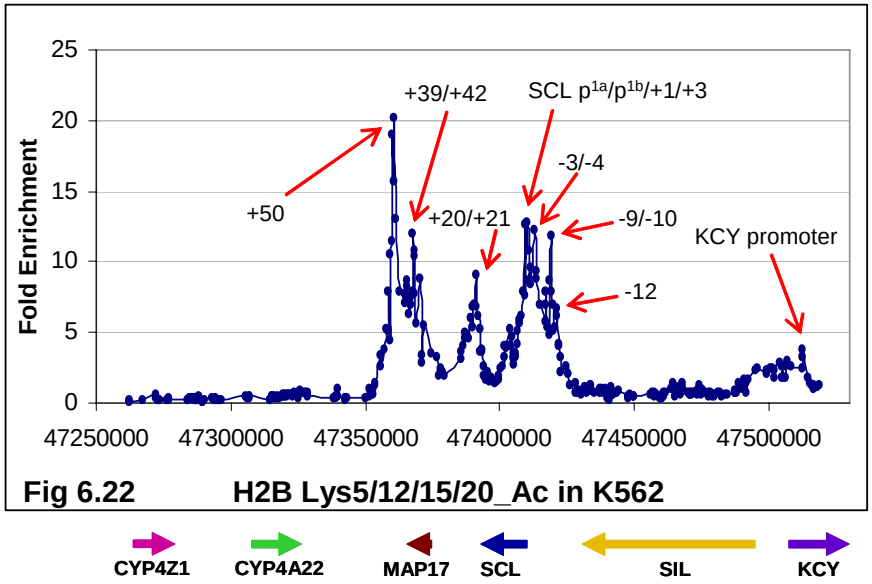


Fig 6.22 **H2B Lys5/12/15/20_Ac in K562**

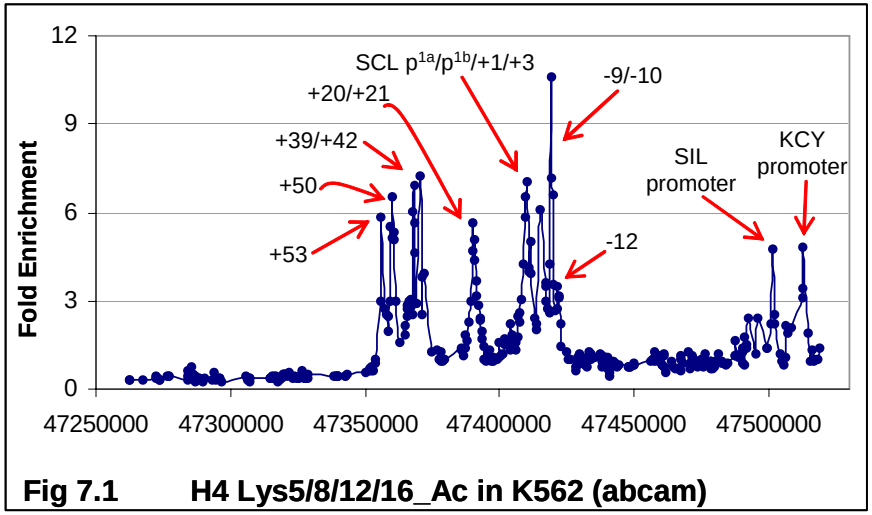


Fig 7.1 H4 Lys5/8/12/16_Ac in K562 (abcam)

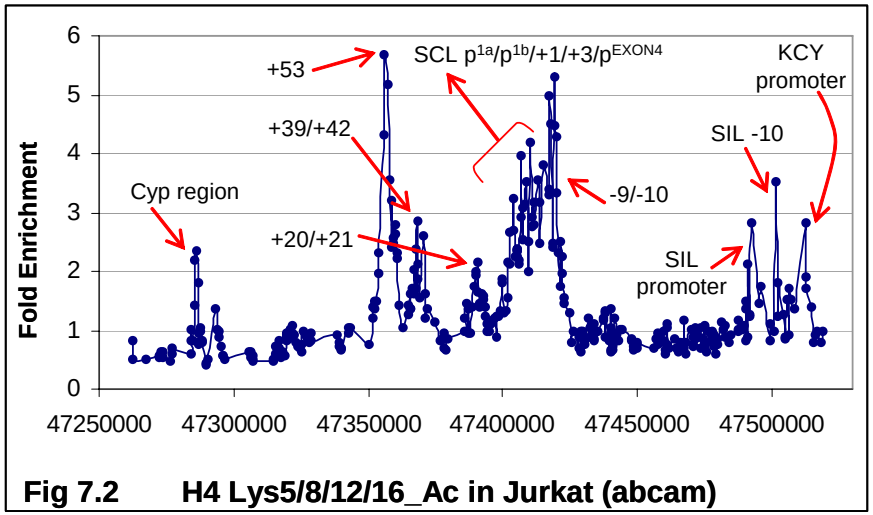


Fig 7.2 H4 Lys5/8/12/16_Ac in Jurkat (abcam)

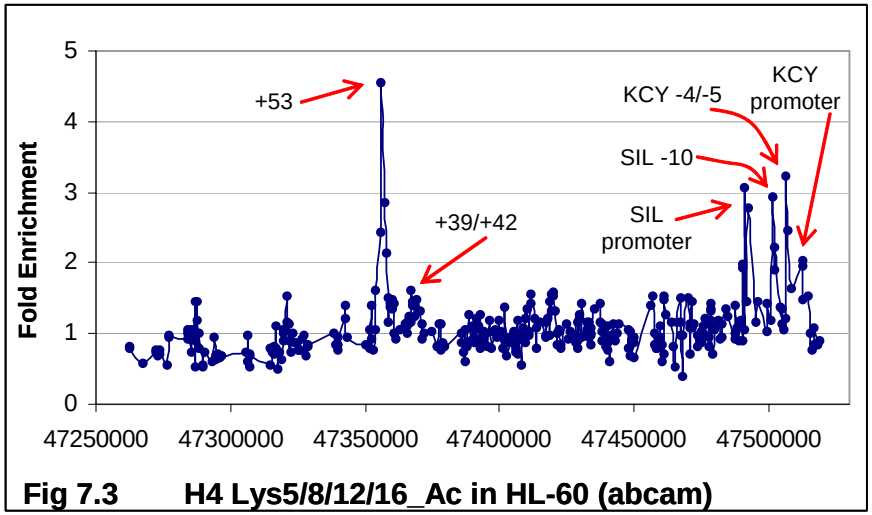


Fig 7.3 H4 Lys5/8/12/16_Ac in HL-60 (abcam)

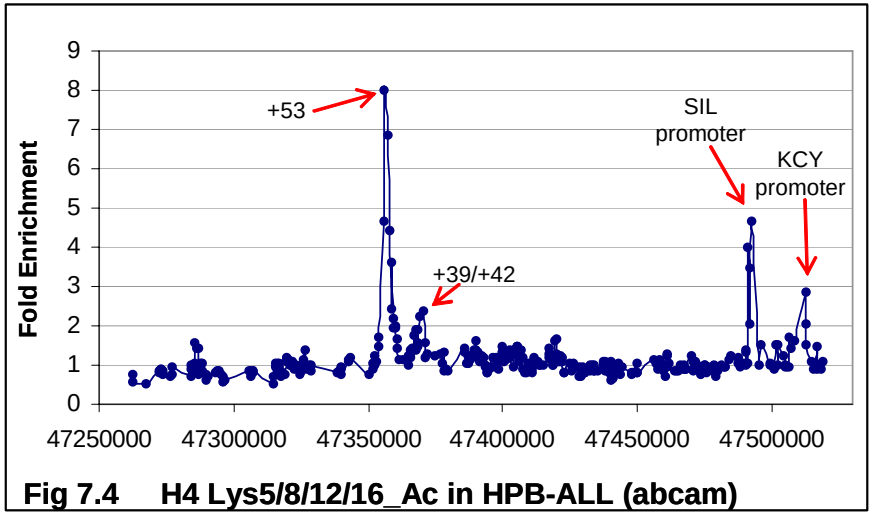


Fig 7.4 H4 Lys5/8/12/16_Ac in HPB-ALL (abcam)

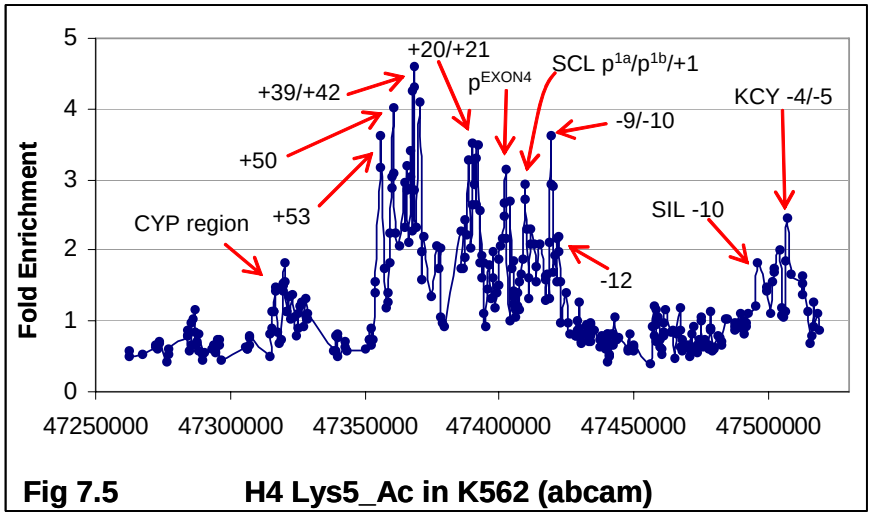


Fig 7.5 **H4 Lys5_Ac in K562 (abcam)**

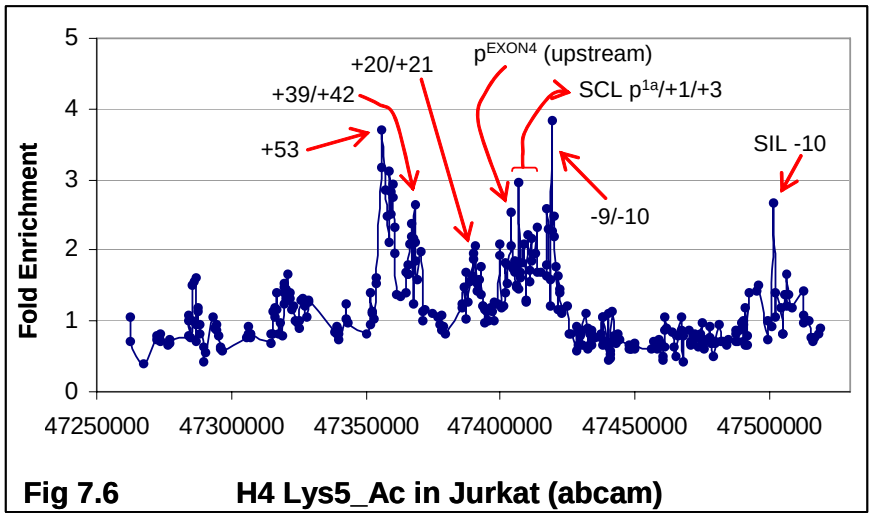


Fig 7.6 **H4 Lys5_Ac in Jurkat (abcam)**

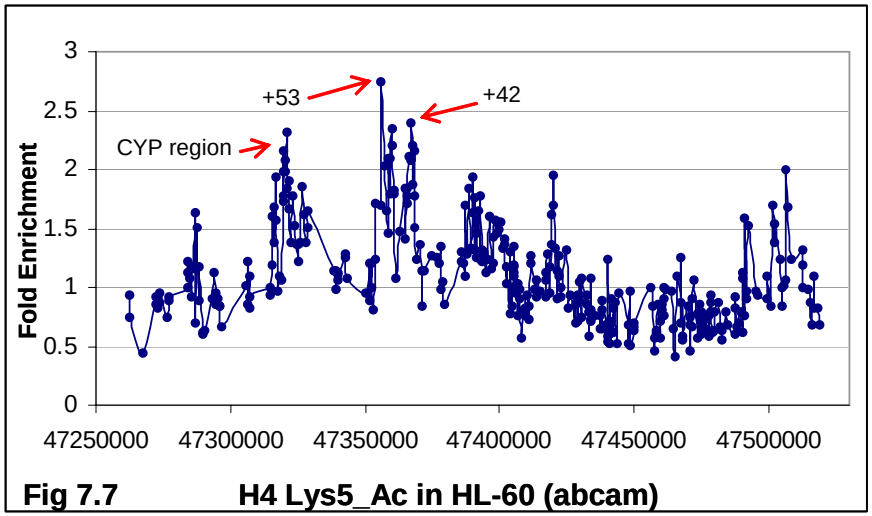


Fig 7.7 **H4 Lys5_Ac in HL-60 (abcam)**

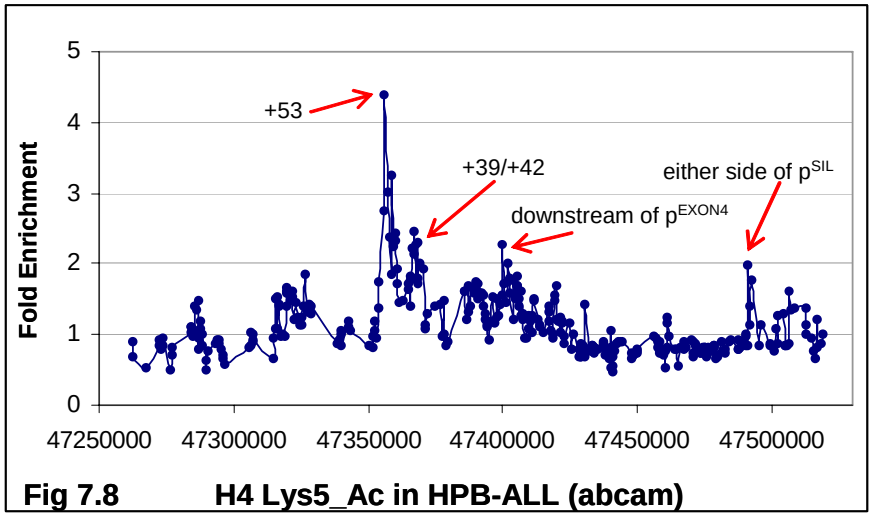


Fig 7.8 **H4 Lys5_Ac in HPB-ALL (abcam)**

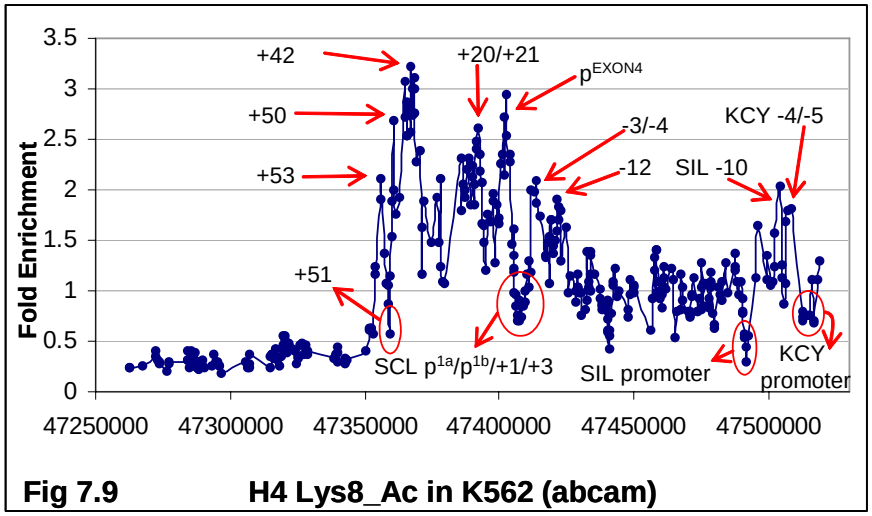


Fig 7.9 **H4 Lys8_Ac in K562 (abcam)**

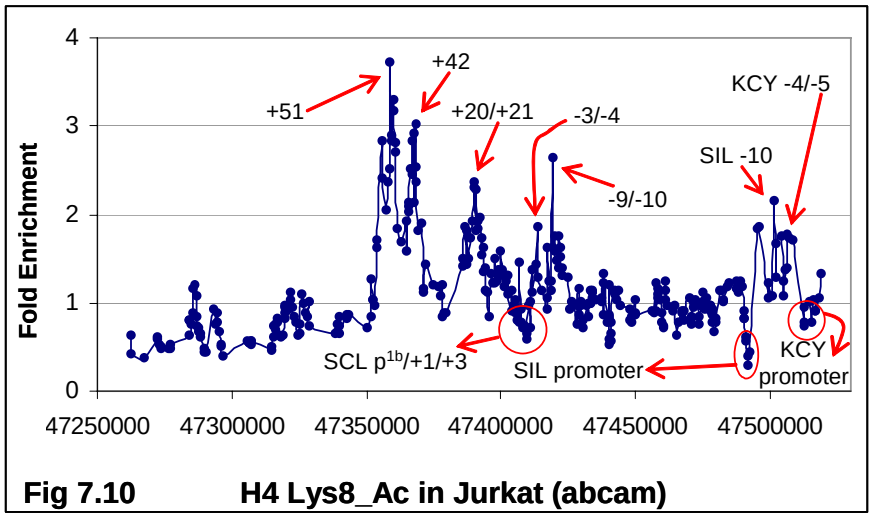


Fig 7.10 **H4 Lys8_Ac in Jurkat (abcam)**

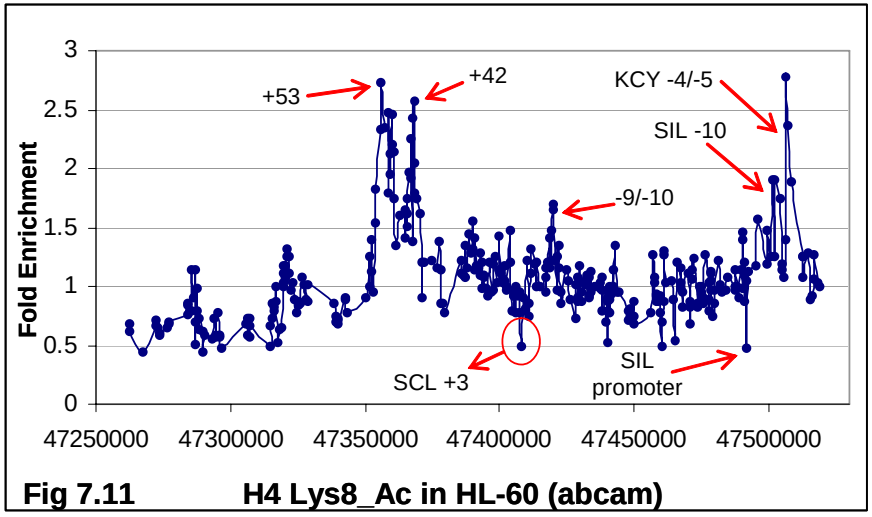


Fig 7.11 **H4 Lys8_Ac in HL-60 (abcam)**

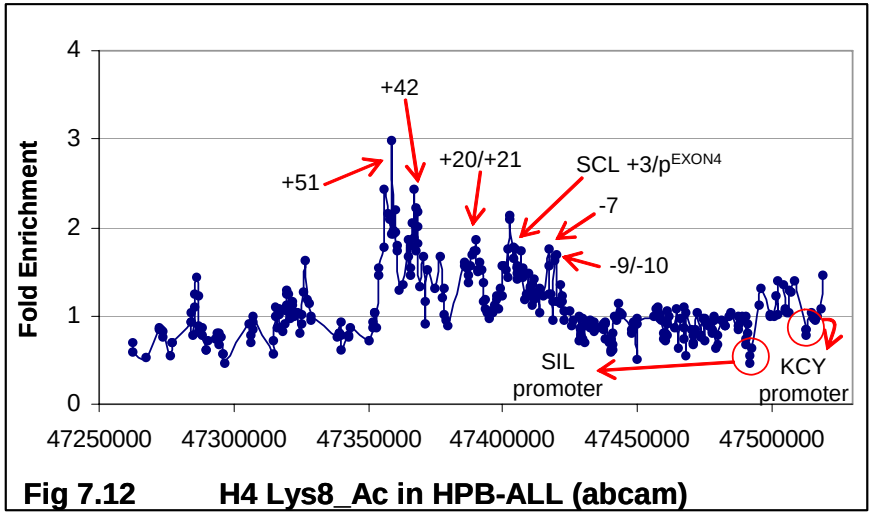


Fig 7.12 **H4 Lys8_Ac in HPB-ALL (abcam)**

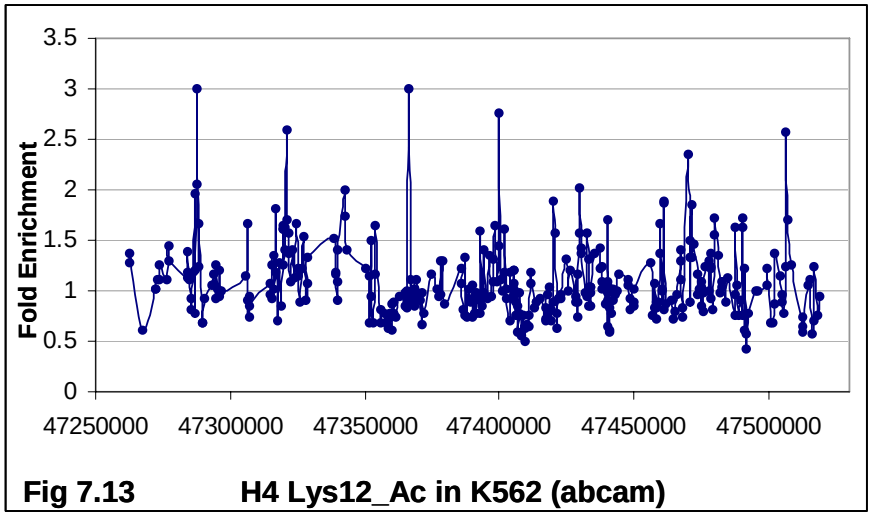


Fig 7.13 H4 Lys12_Ac in K562 (abcam)

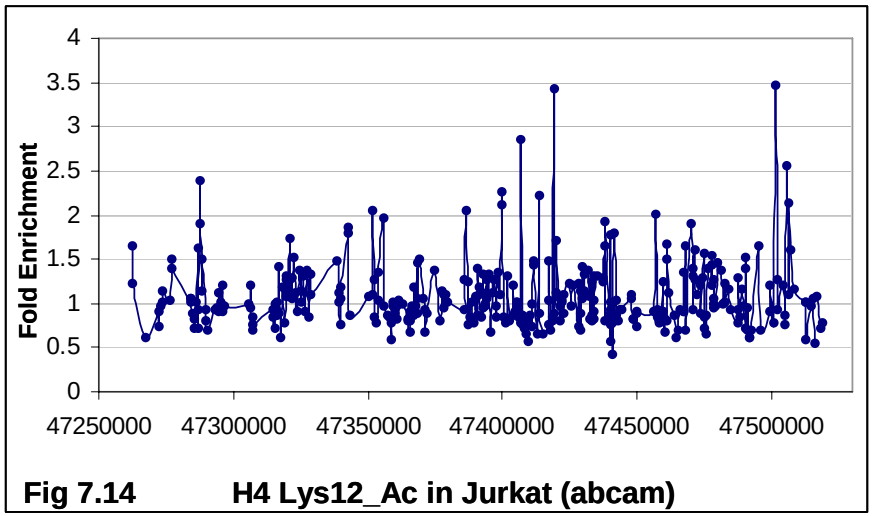


Fig 7.14 H4 Lys12_Ac in Jurkat (abcam)

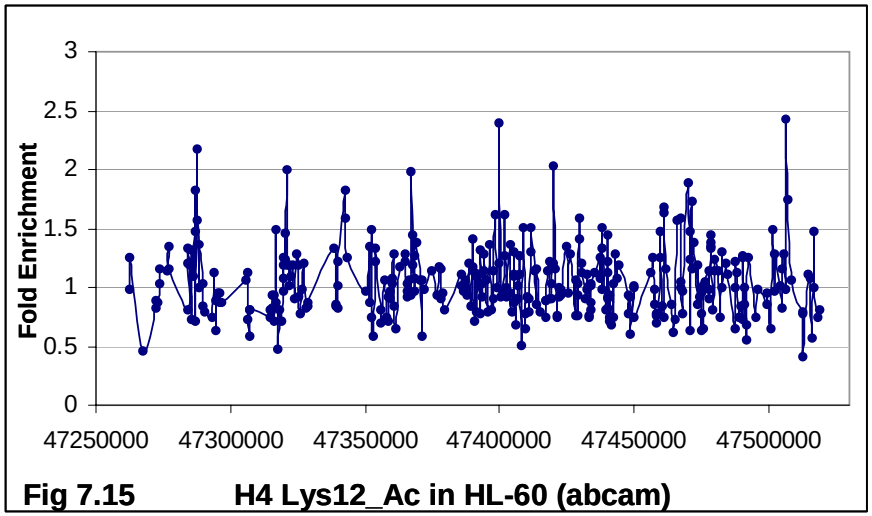


Fig 7.15 H4 Lys12_Ac in HL-60 (abcam)

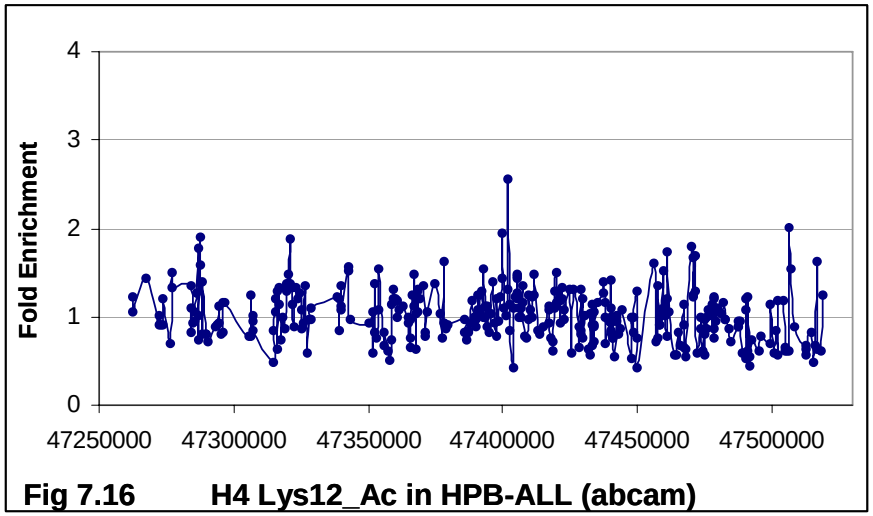


Fig 7.16 H4 Lys12_Ac in HPB-ALL (abcam)

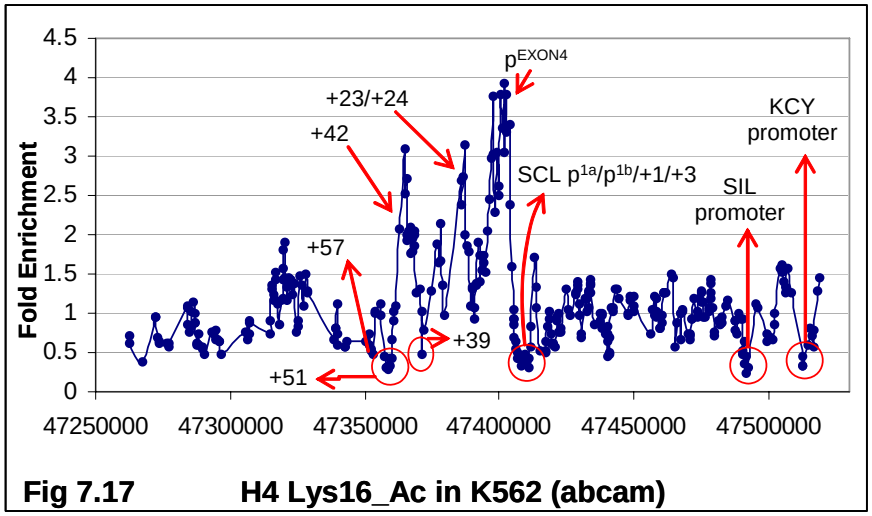


Fig 7.17 **H4 Lys16_Ac in K562 (abcam)**

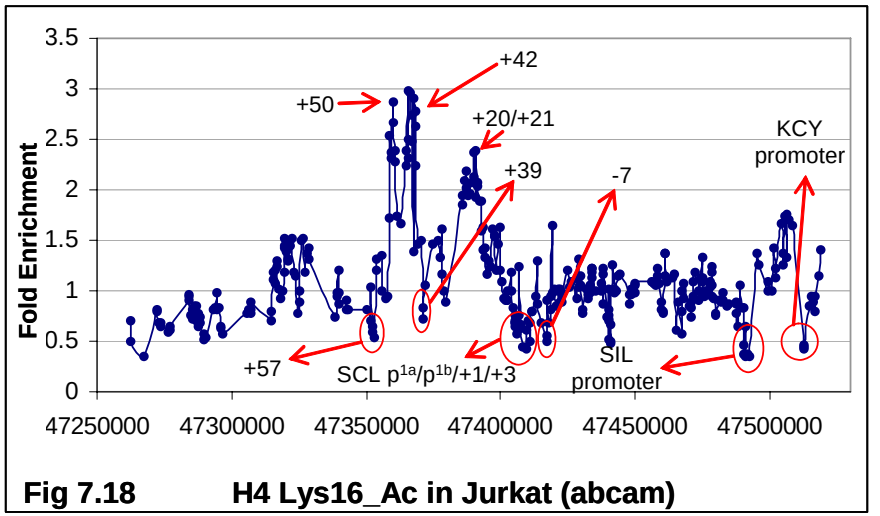


Fig 7.18 **H4 Lys16_Ac in Jurkat (abcam)**

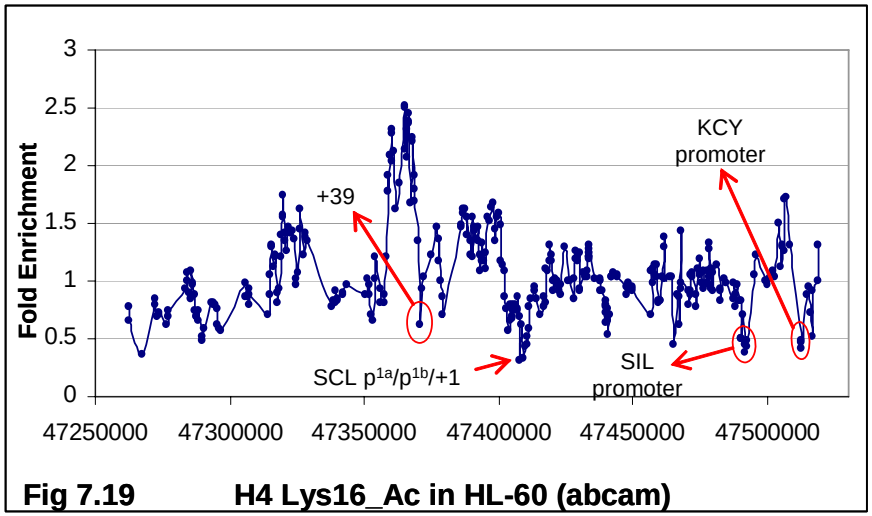


Fig 7.19 **H4 Lys16_Ac in HL-60 (abcam)**

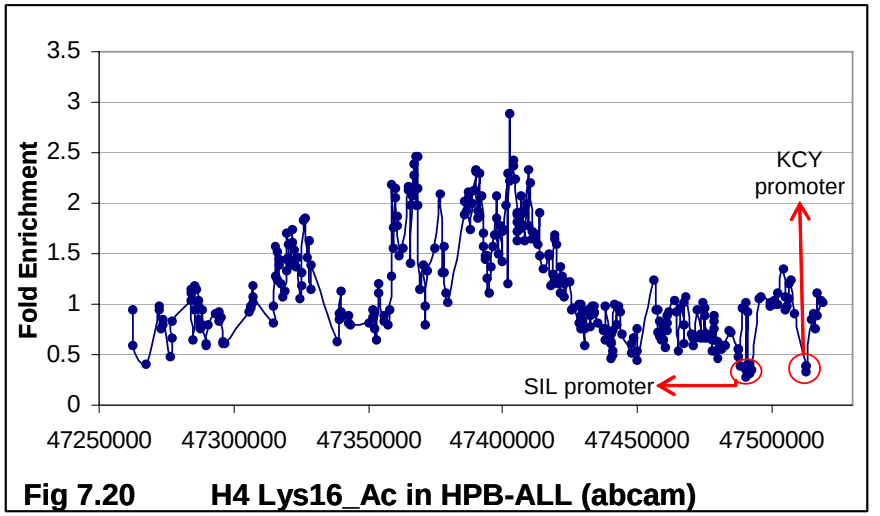


Fig 7.20 **H4 Lys16_Ac in HPB-ALL (abcam)**

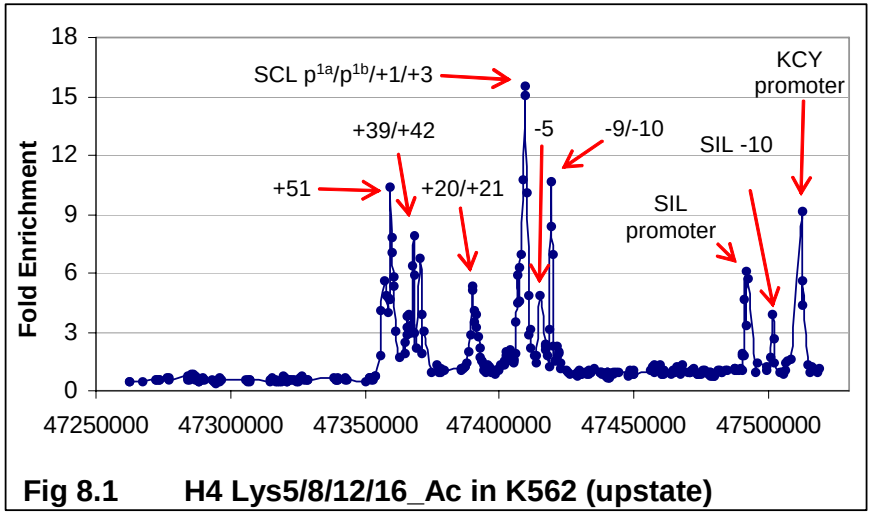


Fig 8.1 H4 Lys5/8/12/16_Ac in K562 (upstate)

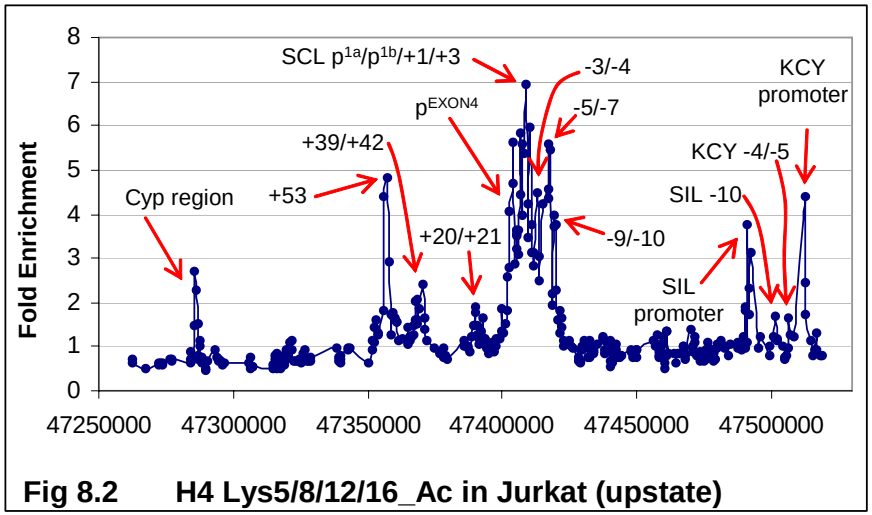


Fig 8.2 H4 Lys5/8/12/16_Ac in Jurkat (upstate)

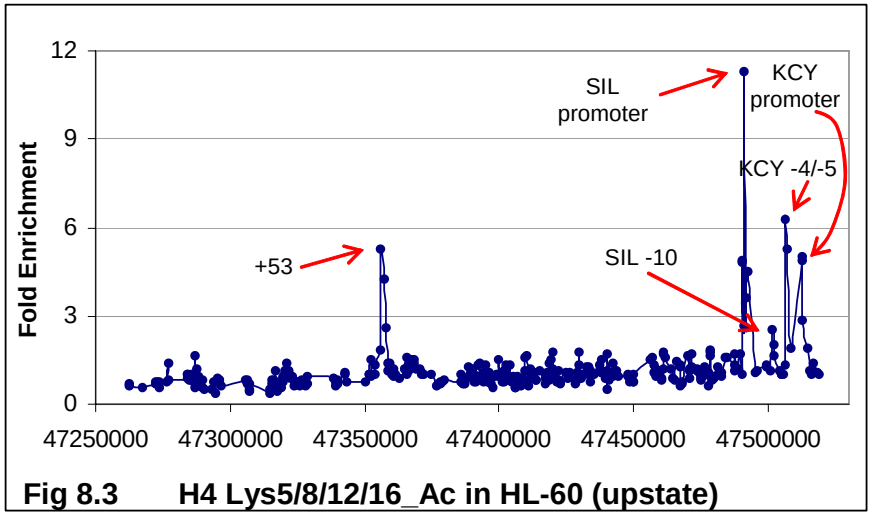


Fig 8.3 H4 Lys5/8/12/16_Ac in HL-60 (upstate)

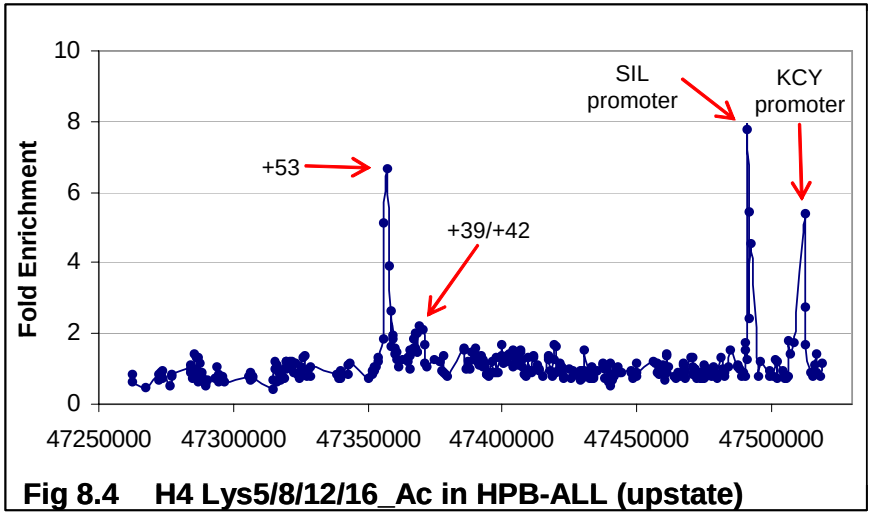


Fig 8.4 H4 Lys5/8/12/16_Ac in HPB-ALL (upstate)

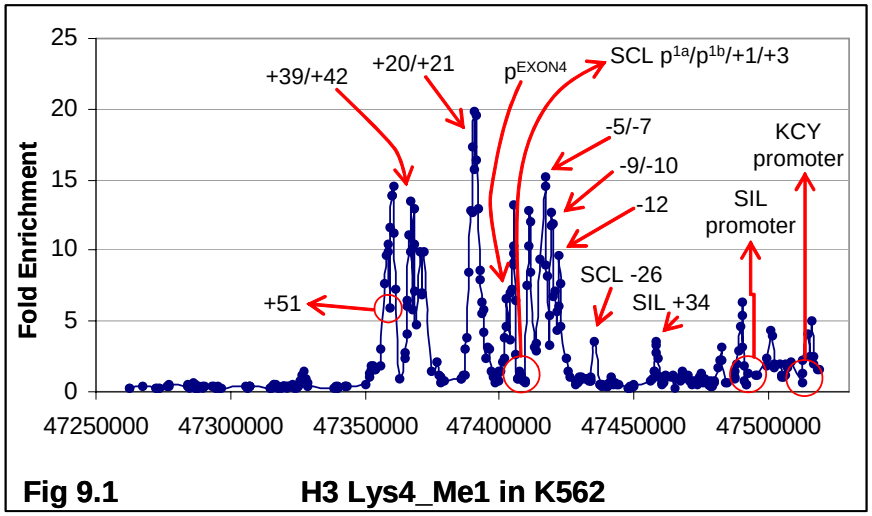


Fig 9.1 **H3 Lys4_Me1 in K562**

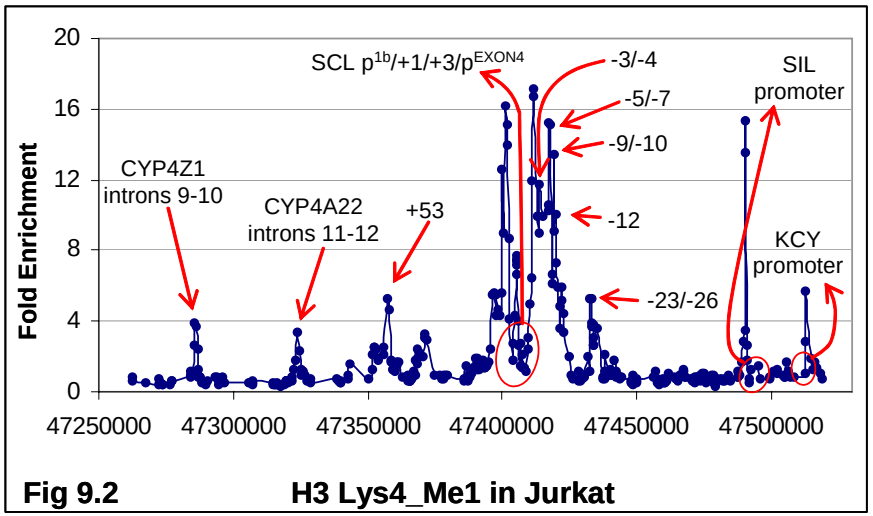


Fig 9.2 **H3 Lys4_Me1 in Jurkat**

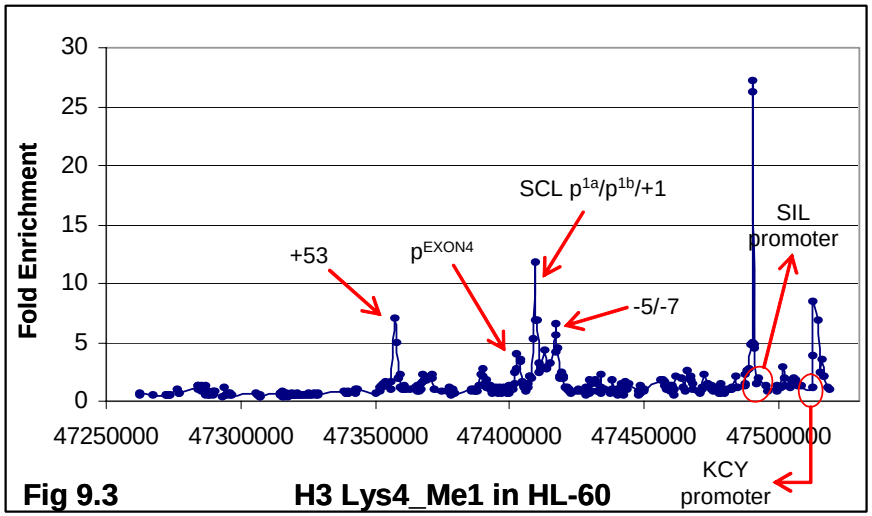


Fig 9.3 **H3 Lys4_Me1 in HL-60**

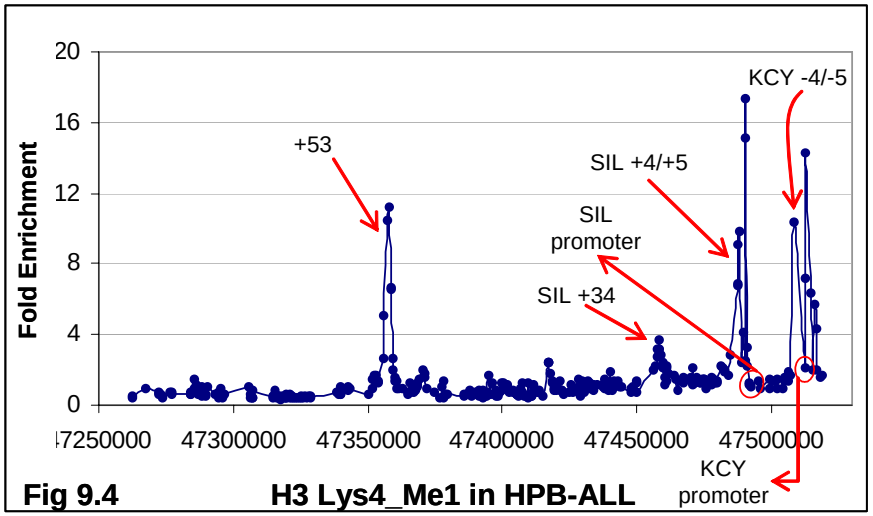


Fig 9.4 **H3 Lys4_Me1 in HPB-ALL**

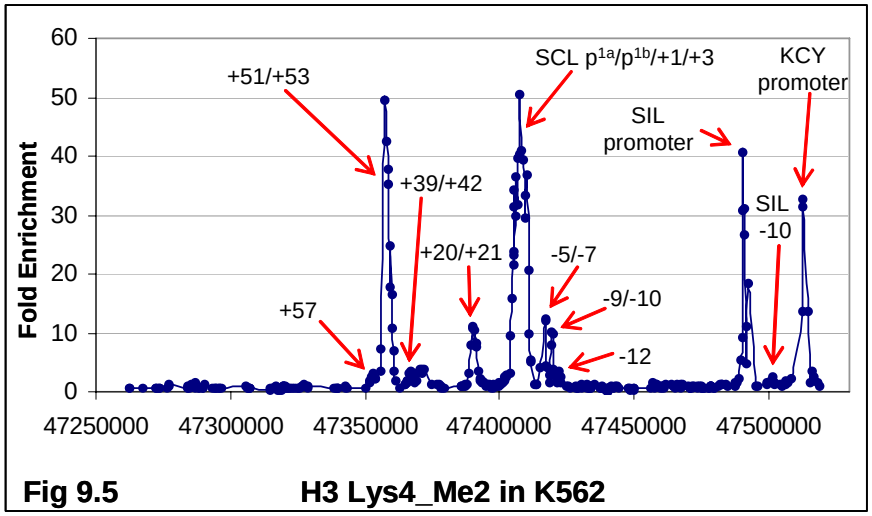


Fig 9.5 **H3 Lys4_Me2 in K562**

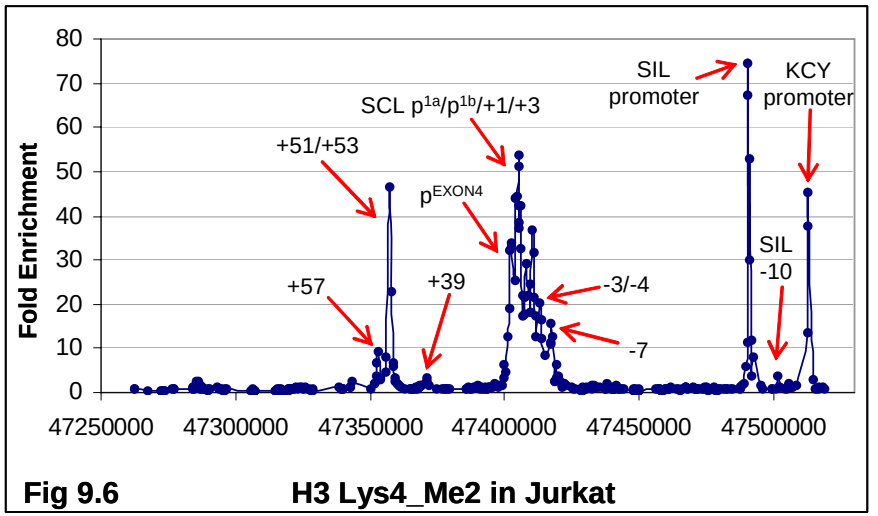


Fig 9.6 **H3 Lys4_Me2 in Jurkat**

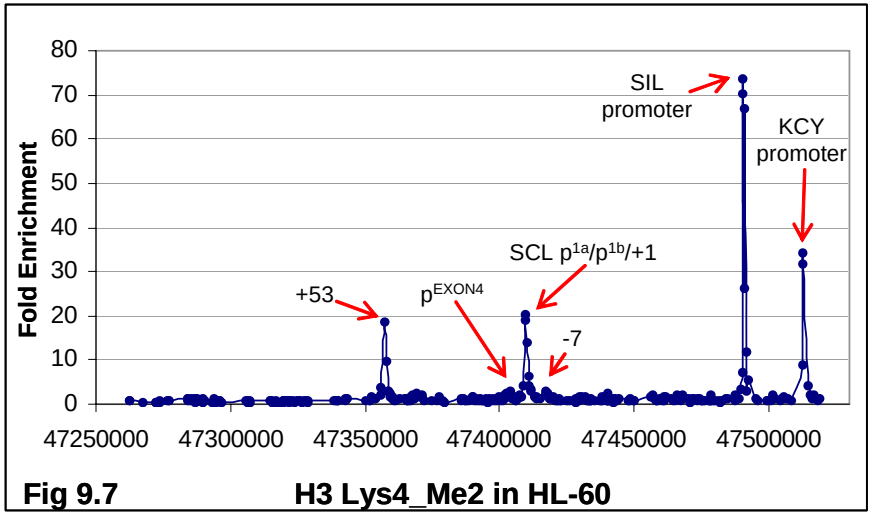


Fig 9.7 **H3 Lys4_Me2 in HL-60**

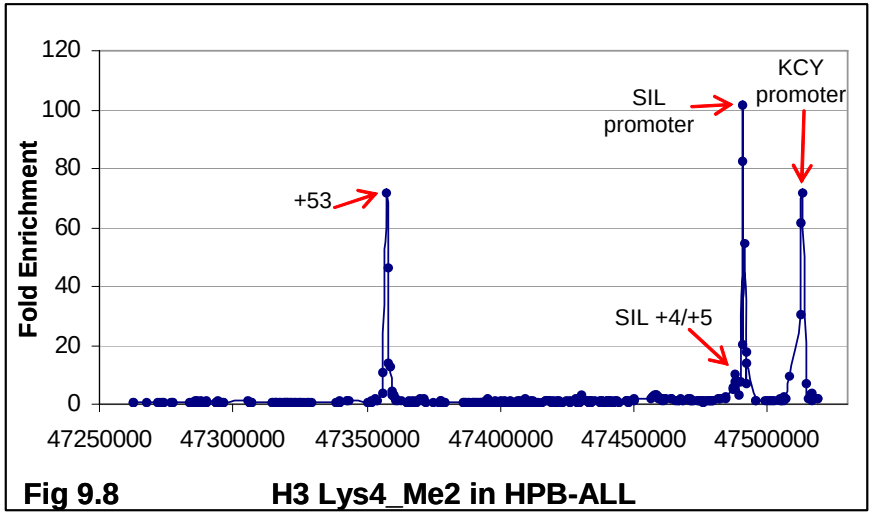


Fig 9.8 **H3 Lys4_Me2 in HPB-ALL**

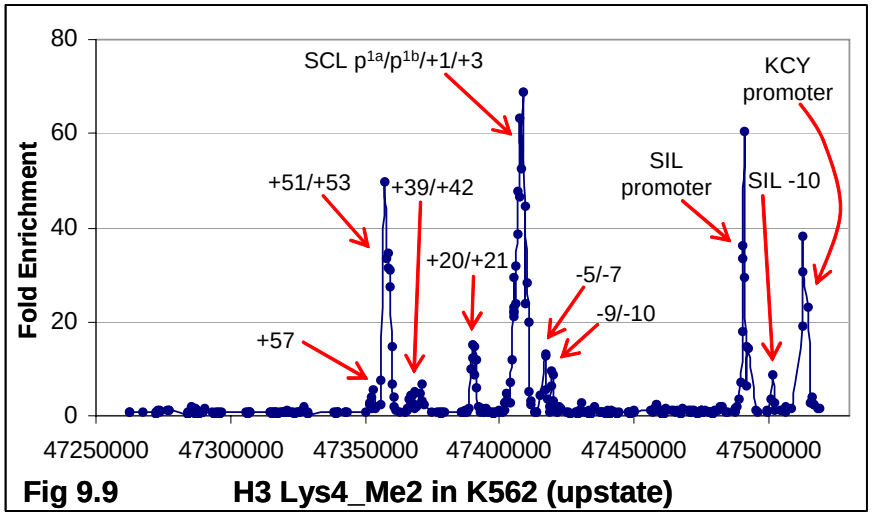


Fig 9.9 **H3 Lys4_Me2 in K562 (upstate)**

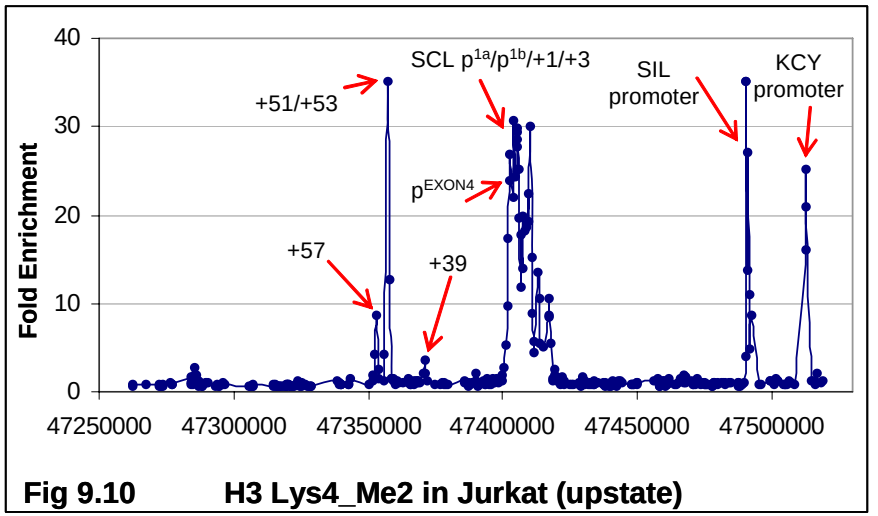


Fig 9.10 **H3 Lys4_Me2 in Jurkat (upstate)**

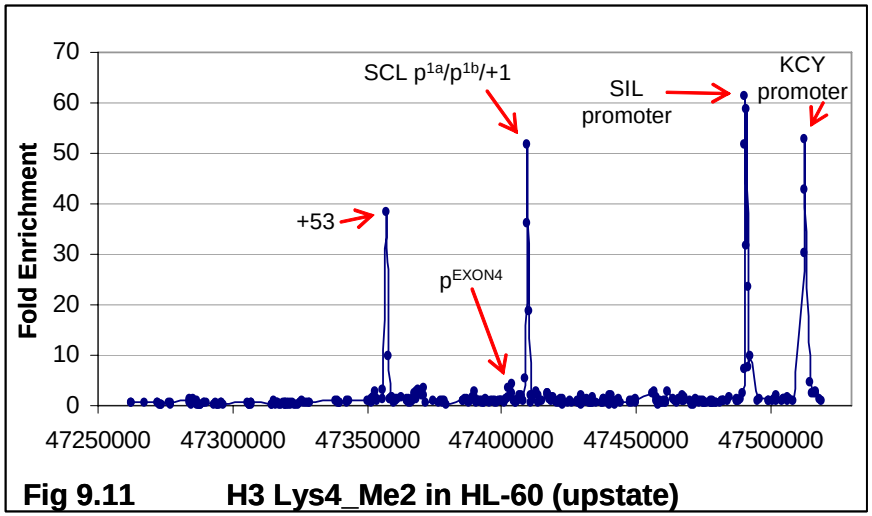


Fig 9.11 **H3 Lys4_Me2 in HL-60 (upstate)**

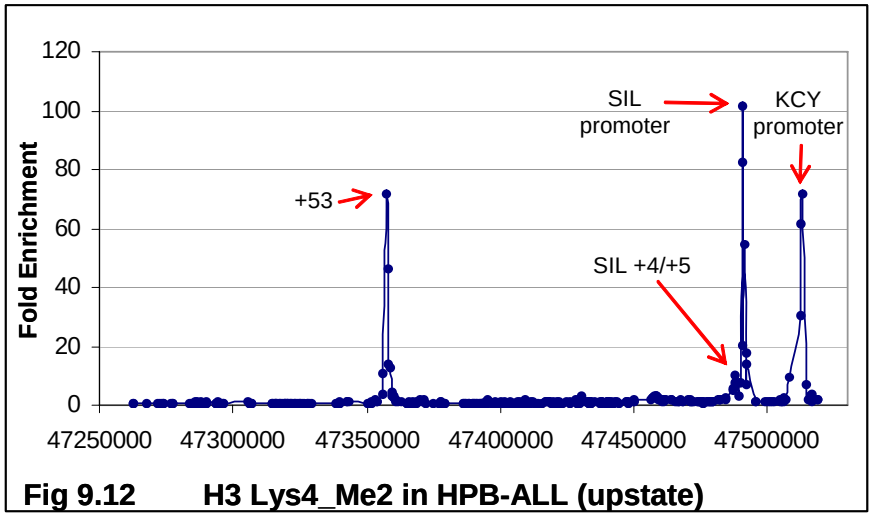


Fig 9.12 **H3 Lys4_Me2 in HPB-ALL (upstate)**

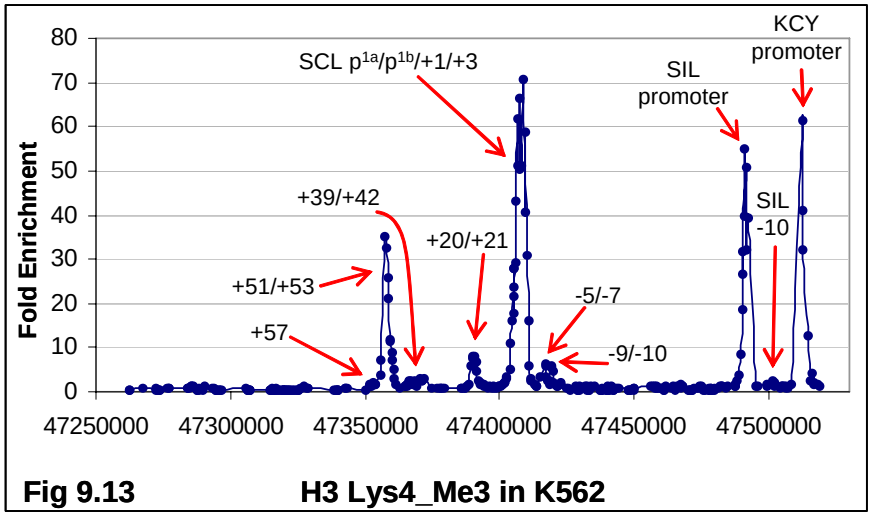


Fig 9.13 **H3 Lys4_Me3 in K562**

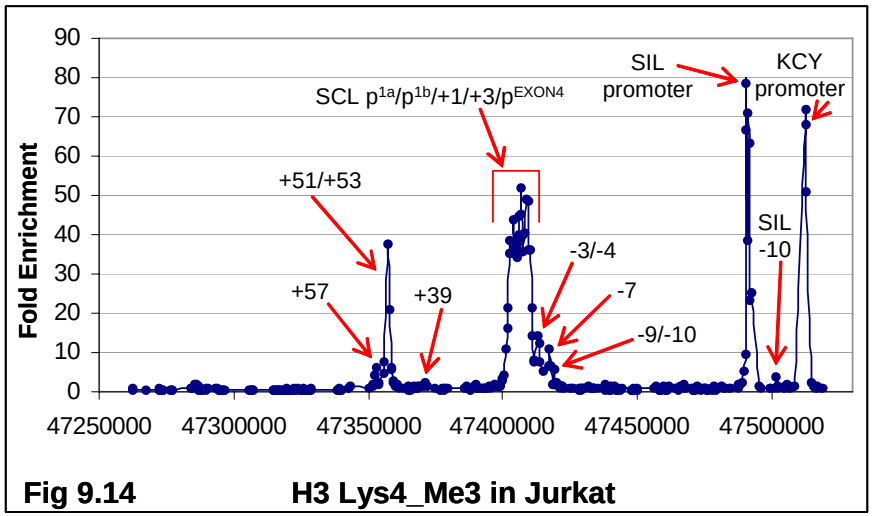


Fig 9.14 **H3 Lys4_Me3 in Jurkat**

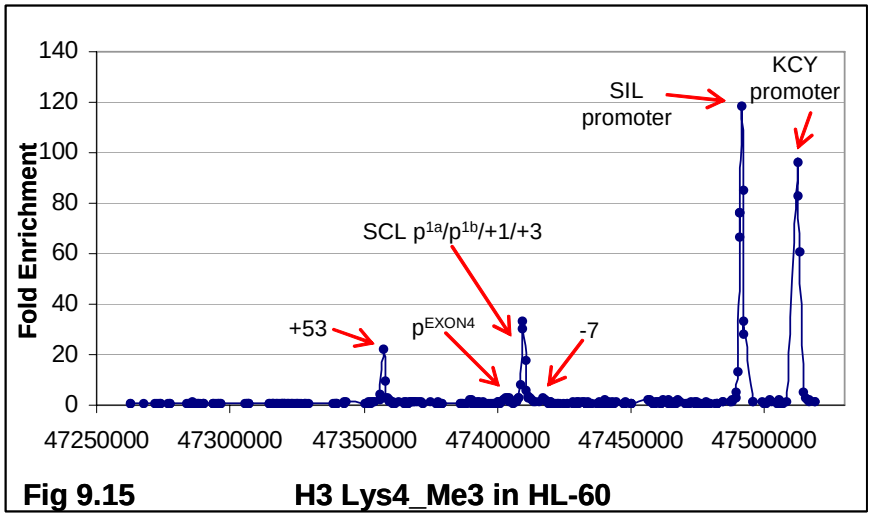


Fig 9.15 **H3 Lys4_Me3 in HL-60**

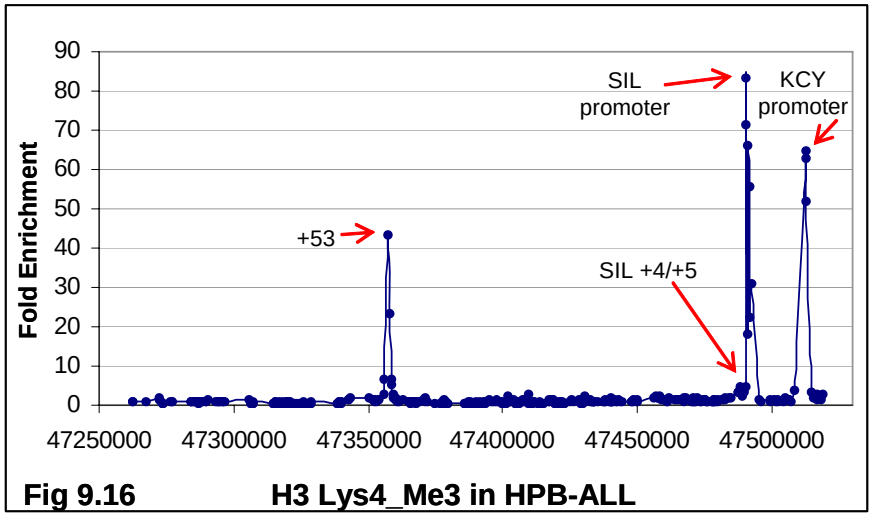


Fig 9.16 **H3 Lys4_Me3 in HPB-ALL**

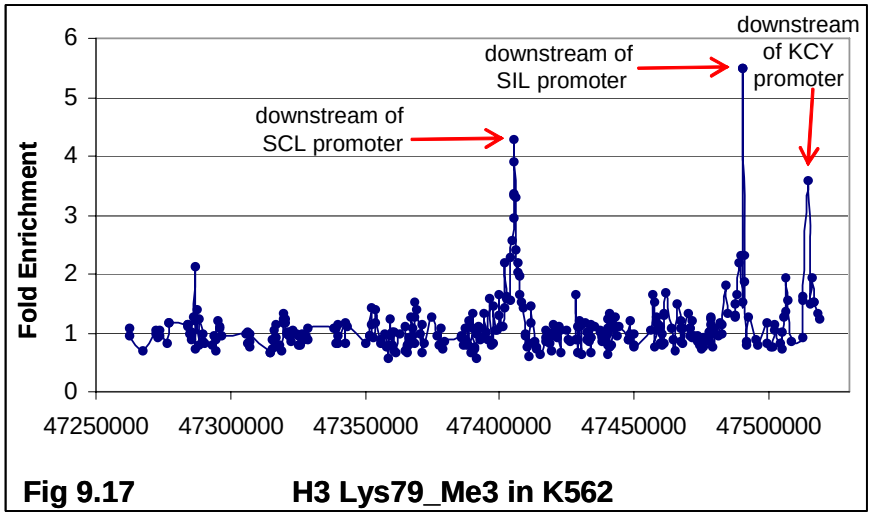


Fig 9.17 **H3 Lys79_Me3 in K562**

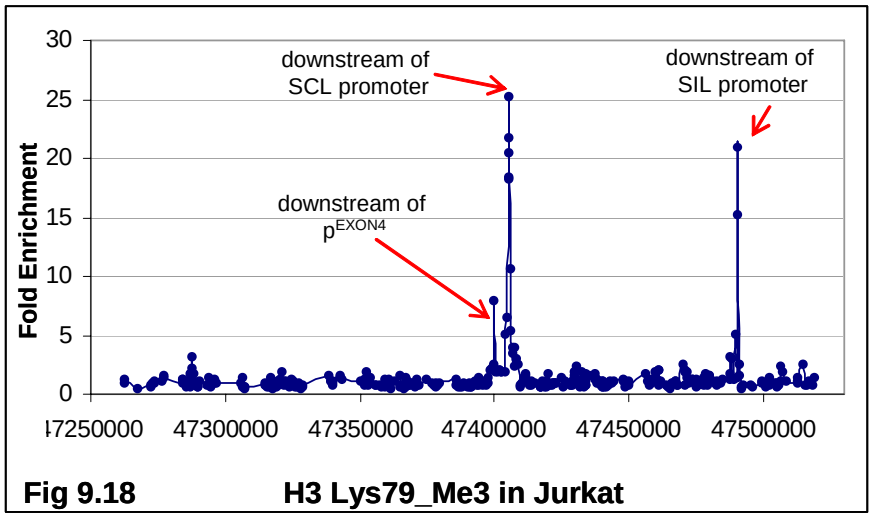


Fig 9.18 **H3 Lys79_Me3 in Jurkat**

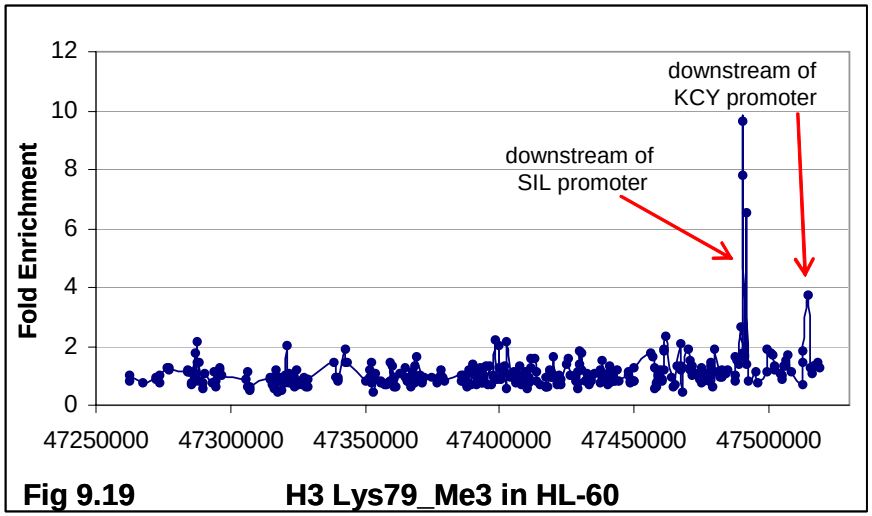


Fig 9.19 **H3 Lys79_Me3 in HL-60**

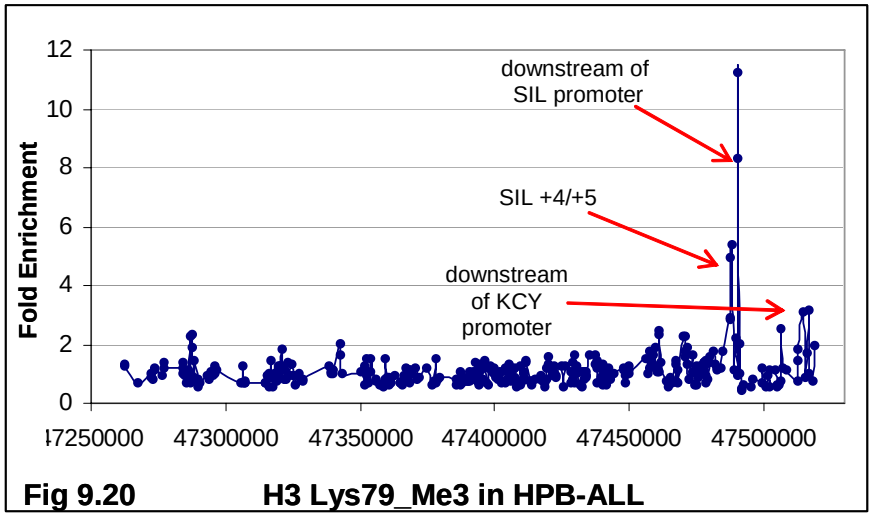


Fig 9.20 **H3 Lys79_Me3 in HPB-ALL**

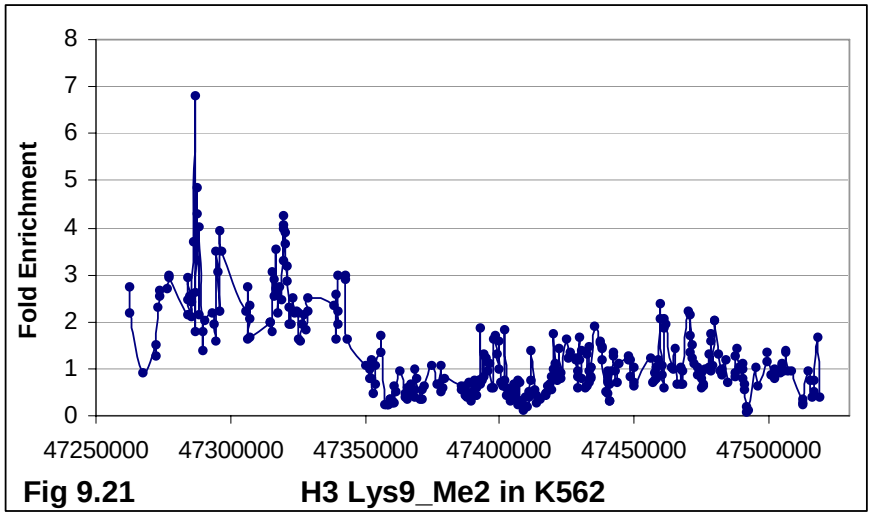


Fig 9.21 **H3 Lys9_Me2 in K562**

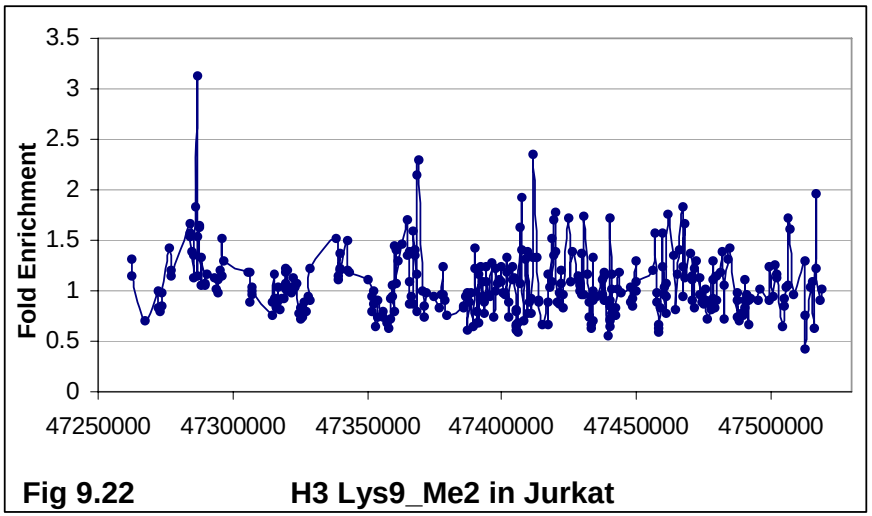


Fig 9.22 **H3 Lys9_Me2 in Jurkat**

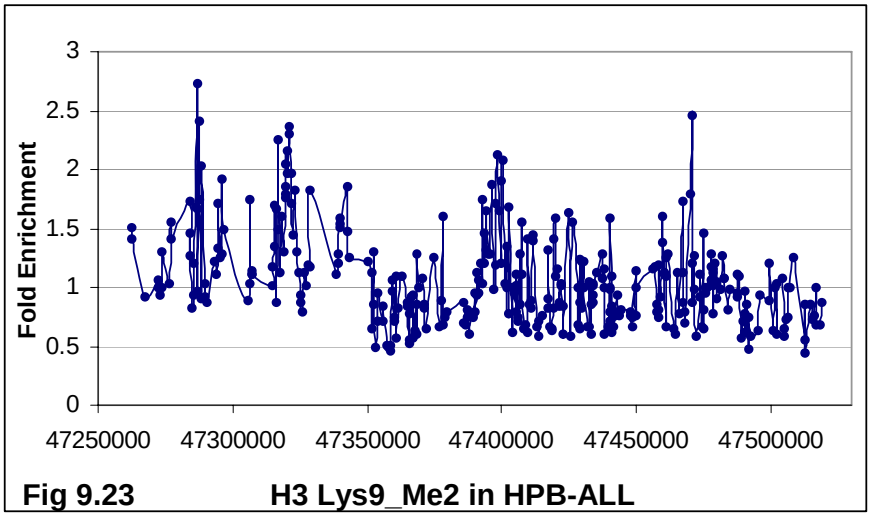


Fig 9.23 **H3 Lys9_Me2 in HPB-ALL**

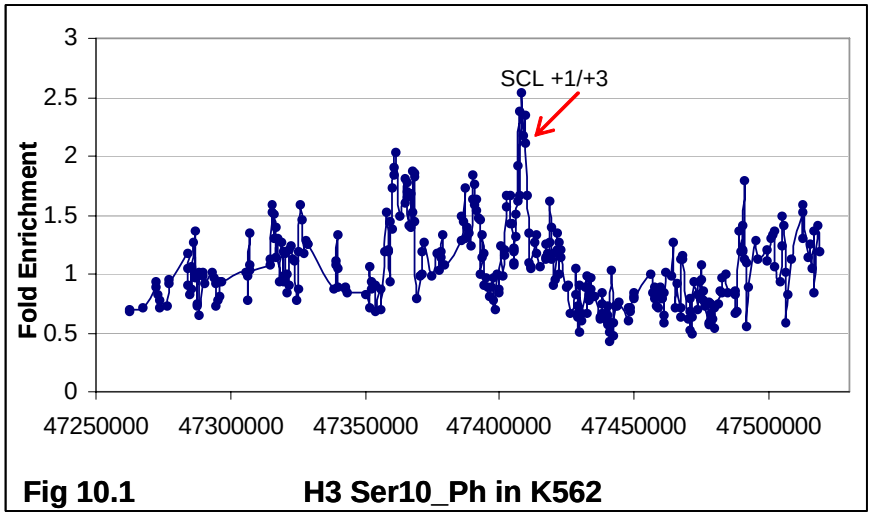


Fig 10.1 **H3 Ser10_Ph in K562**

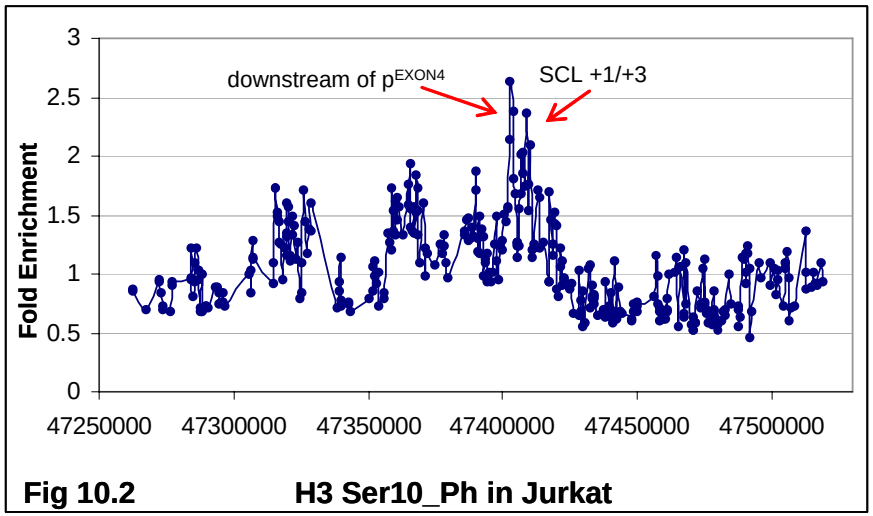


Fig 10.2 **H3 Ser10_Ph in Jurkat**

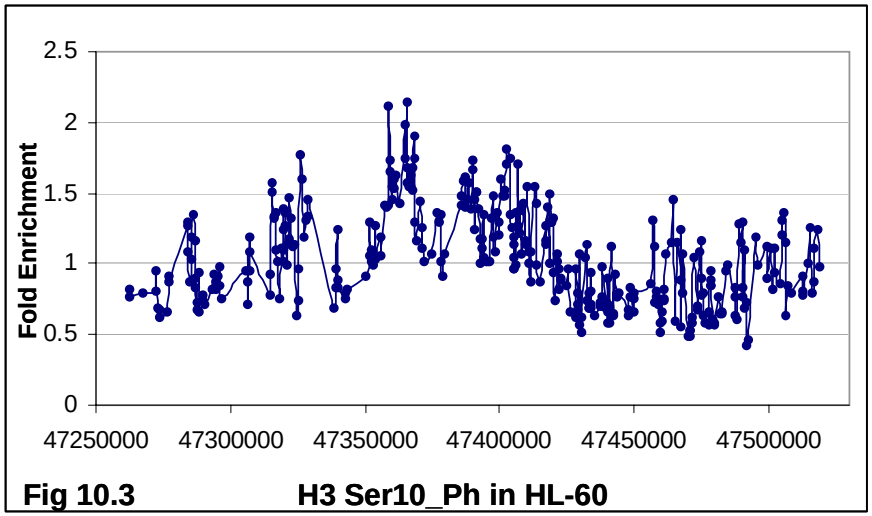


Fig 10.3 **H3 Ser10_Ph in HL-60**

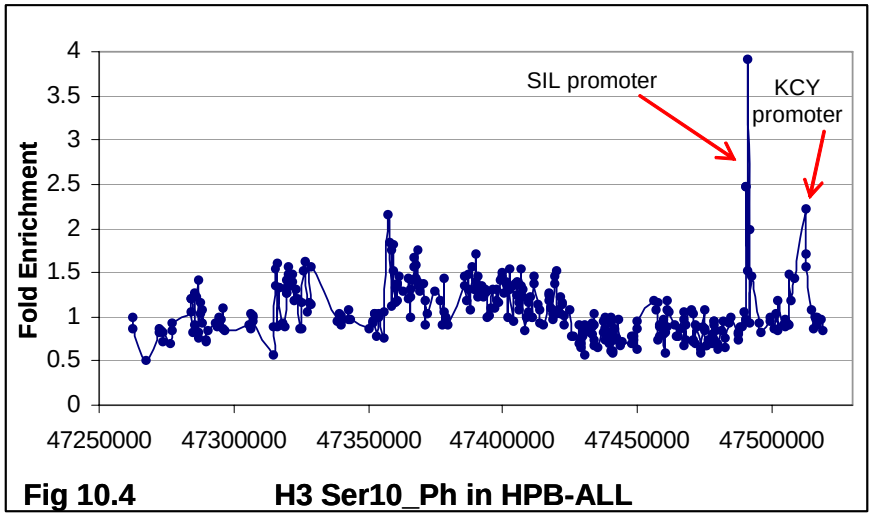


Fig 10.4 **H3 Ser10_Ph in HPB-ALL**

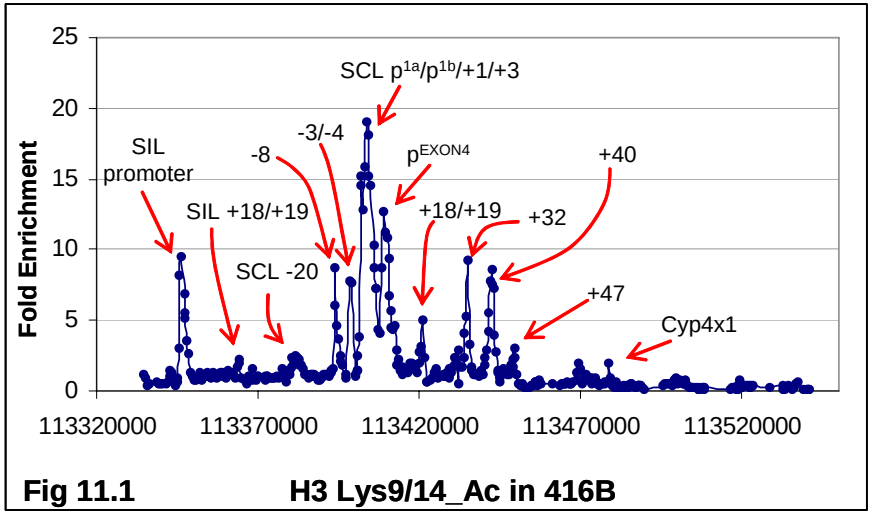


Fig 11.1 **H3 Lys9/14_Ac in 416B**

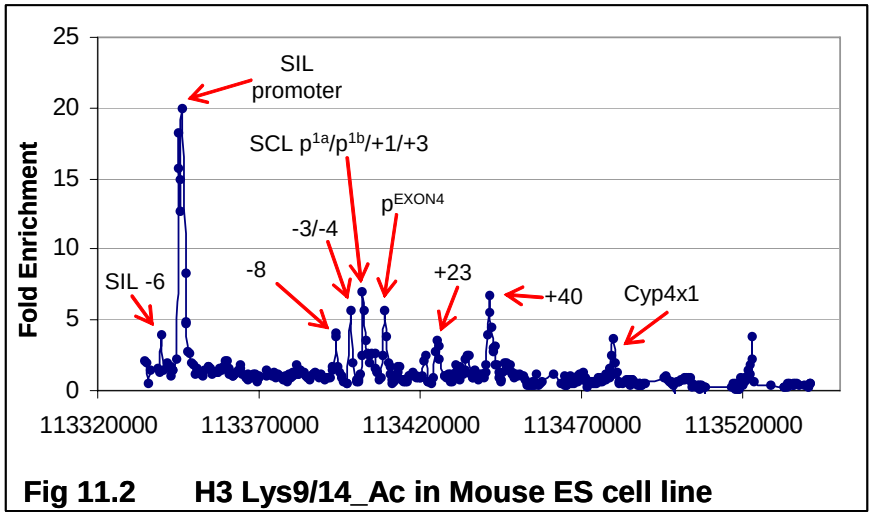


Fig 11.2 **H3 Lys9/14_Ac in Mouse ES cell line**

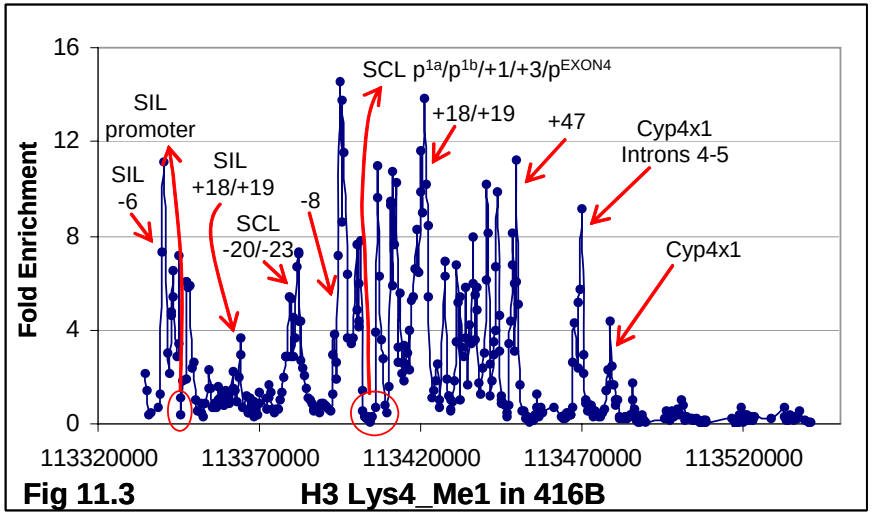


Fig 11.3 **H3 Lys4_Me1 in 416B**

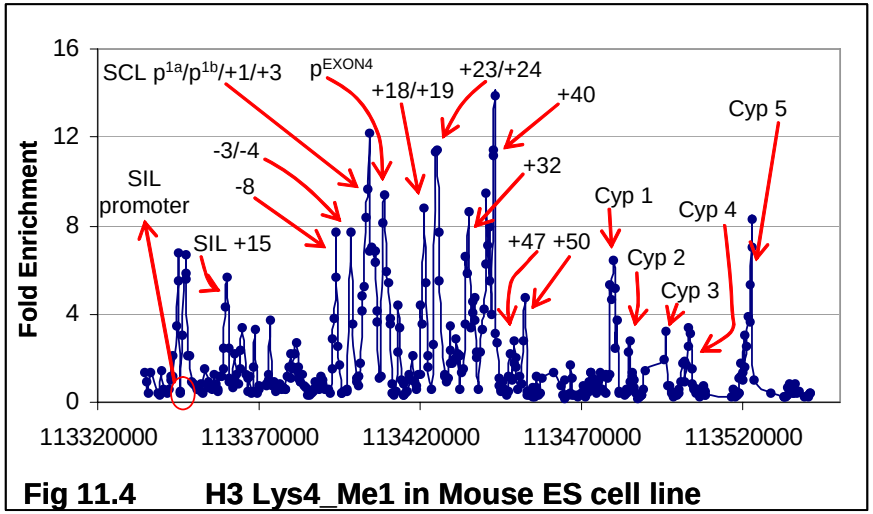
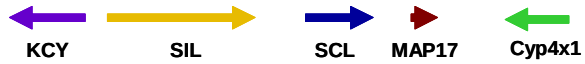


Fig 11.4 **H3 Lys4_Me1 in Mouse ES cell line**



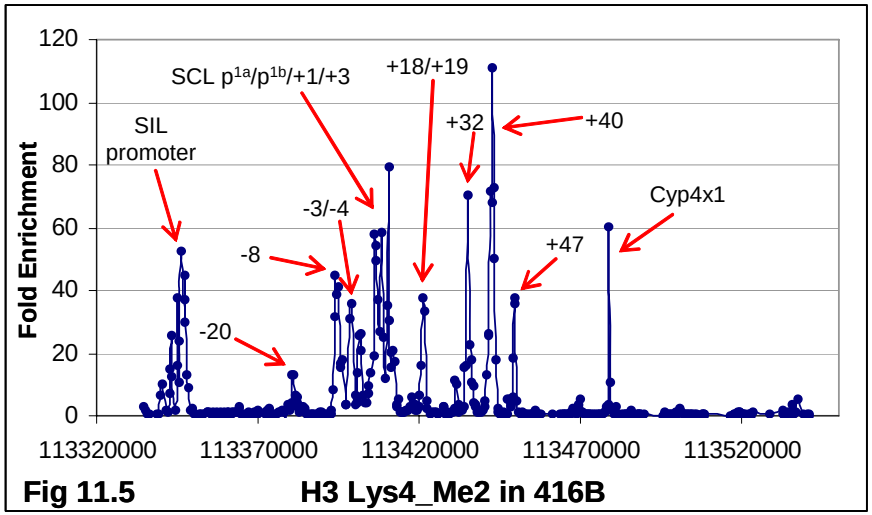


Fig 11.5 **H3 Lys4_Me2 in 416B**

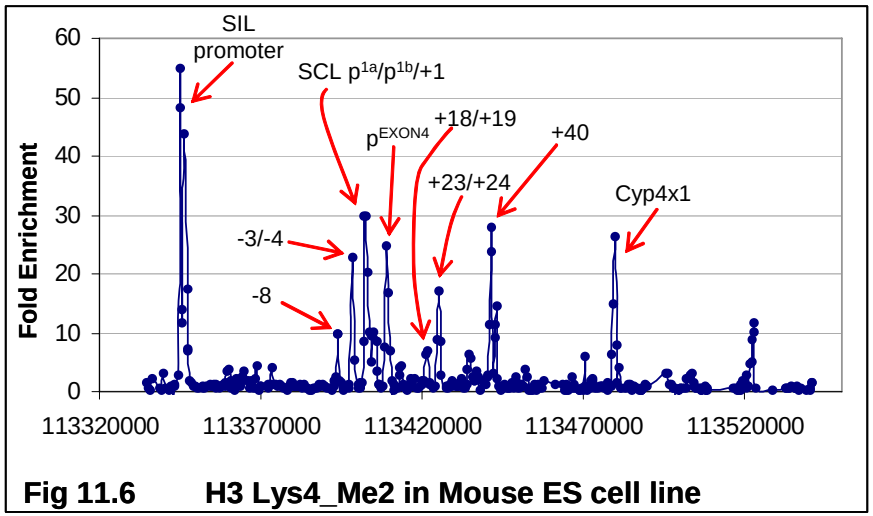


Fig 11.6 **H3 Lys4_Me2 in Mouse ES cell line**

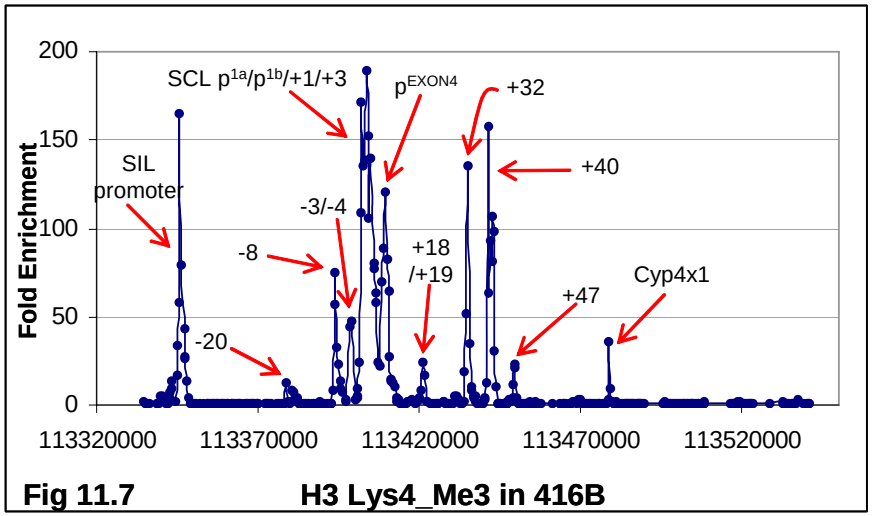


Fig 11.7 **H3 Lys4_Me3 in 416B**

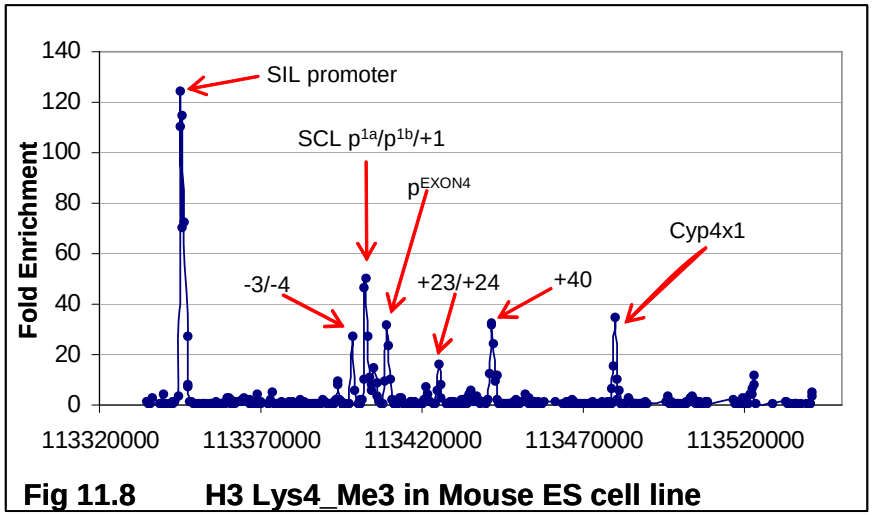
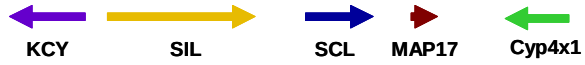
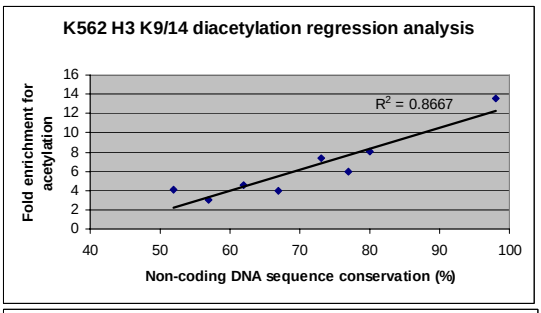


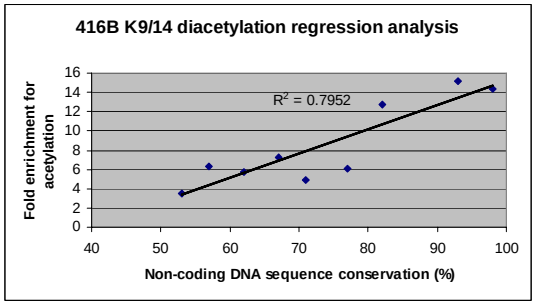
Fig 11.8 **H3 Lys4_Me3 in Mouse ES cell line**

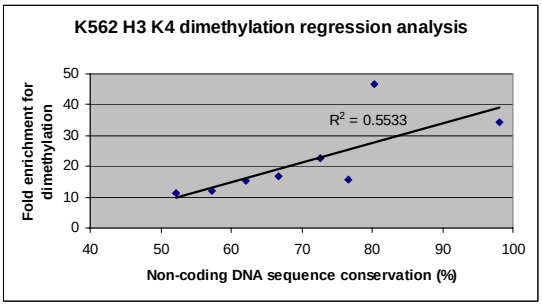


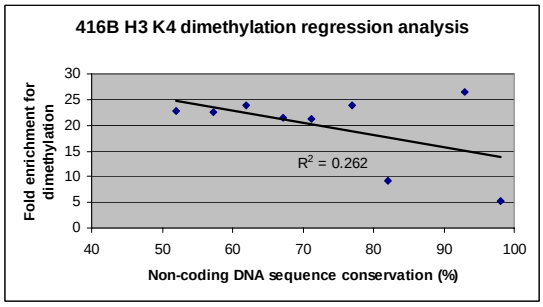
Appendix 12

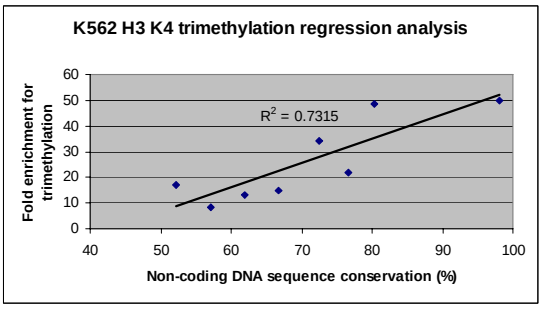
Correlation graphs of non-coding sequence conservation with histone H3 acetylation and methylation

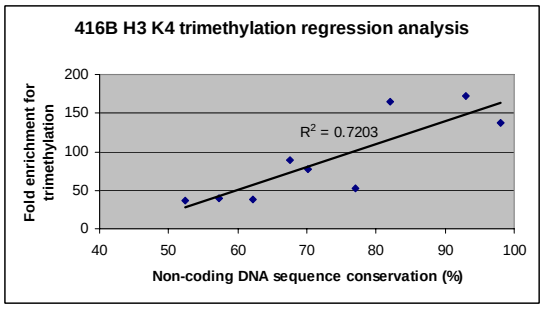












Appendix 13

Sequences of Primer Pairs used to generate constructs for transfection assays

Amplicon Name	Alternative Name	Primer 1 (5'→3')	Primer 2 (5'→3')	Amplicon Size (bp)	Chrom 1 Co-ordinate Start	Chrom 1 Co-ordinate Finish
SV40/-12b	HSSCL/M33Aet-b	(CGGGATCCCG)CAGGAGCATTCAGCAGATCT	(CGGGATCCCG)TGACGTTTGCAGACTCCTTCA	2235	47420799	47423033
SV40/ -12s	HSSCL/M33Aet-s	(CGGGATCCCG)TCTTTCCGTTGTGTTTGTGATCA	(CGGGATCCCG)AAGATTTTGAAGGATTTCCACAGG	453	47421672	47422124
SV40/ -7	HSTAL.170et	(CGGGATCCCG)TTGACCTCAGATGATCCGCC	(CGGGATCCCG)TCATTTCCCCTTCTCCCCAG	952	47416433	47417384
SV40/+51	HSSCL/M96Bet	(CGGGATCCCG)AAAGGTTGGGAAGAGAGCAGG	(CGGGATCCCG)GGGTCAGGCCTCTGCTAAGG	936	47358744	47359679
SV40/neutral	HSTAL.20et	(CGGGATCCCG)ACTTTCCTTTTAAGTACACCAGCAAC	(CGGGATCCCG)TTTCAATGAGCTATTGGATTATGTG	951	47289370	47290320