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.