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.

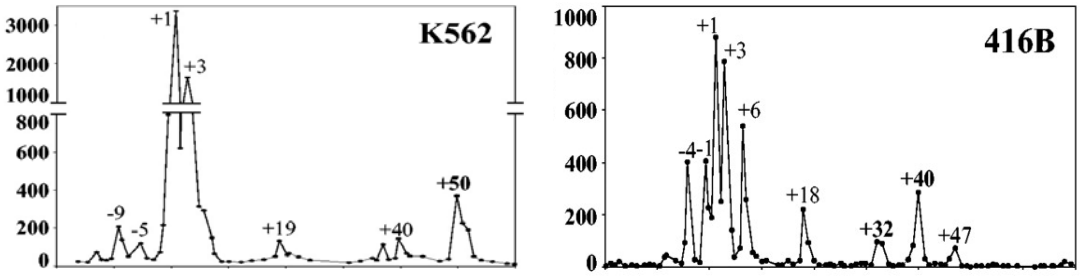


Figure 4.1: ChIP-PCR based histone H3 acetylation profiles in K562 and 416B cell lines. Approximately 90 kb region across the SCL locus was mapped for histone H3 and H4 acetylation using ChIP-PCR based assays in these cell lines. The peaks in the profiles represent regions showing enrichment for H3 acetylation. 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.

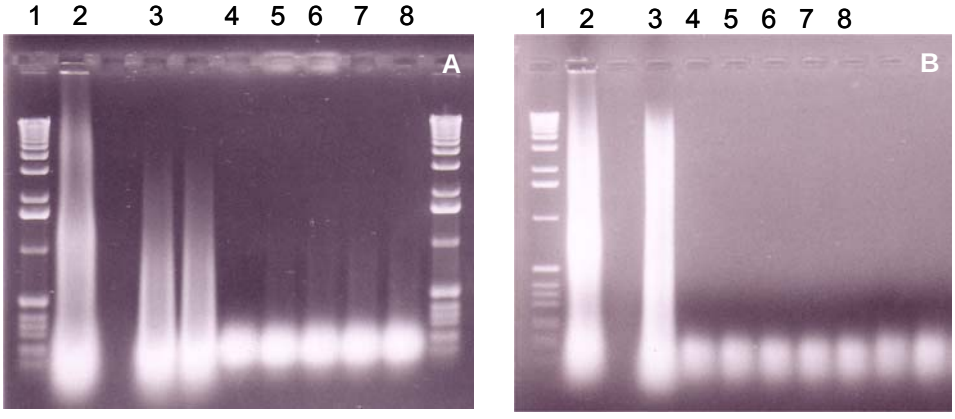


Figure 4.2: The electrophoretic analysis of ChIP DNAs. Panel A shows ChIP assay performed for histone H3 acetylation and panel B for GATA-1 transcription factor. In both panels, lane 1= 1 kb ladder; lane 2= the sheared cross-linked material; lane 3= the input DNA (sheared purified total genomic DNA); 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.

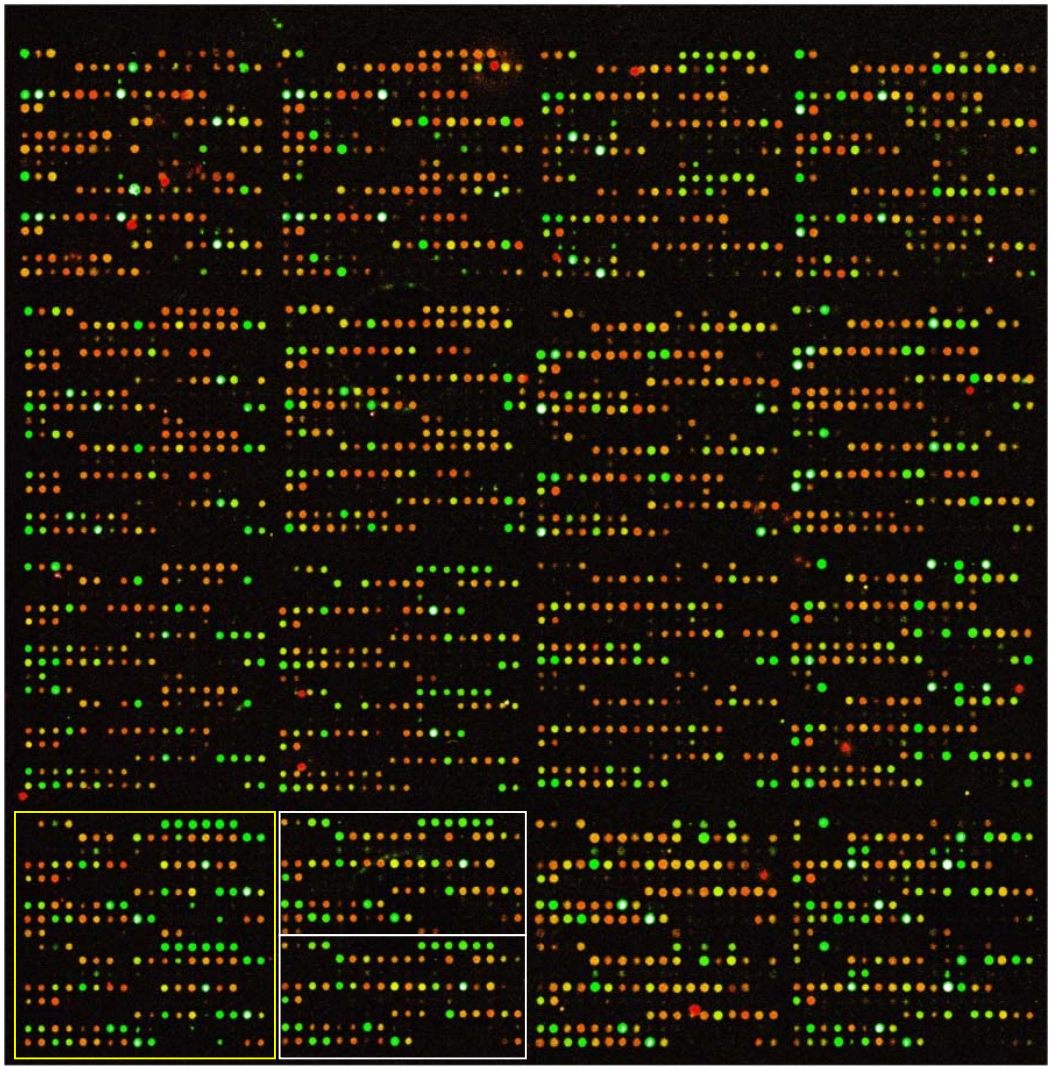


Figure 4.3: A composite image of the human SCL genomic tiling array. The array was hybridised with a ChIP sample for histone H3 K9/14 diacetylation in cell line K562 along with K562 input DNA. Each spot on the array represents an array element. Each array element was spotted in duplicate (a duplicated 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).

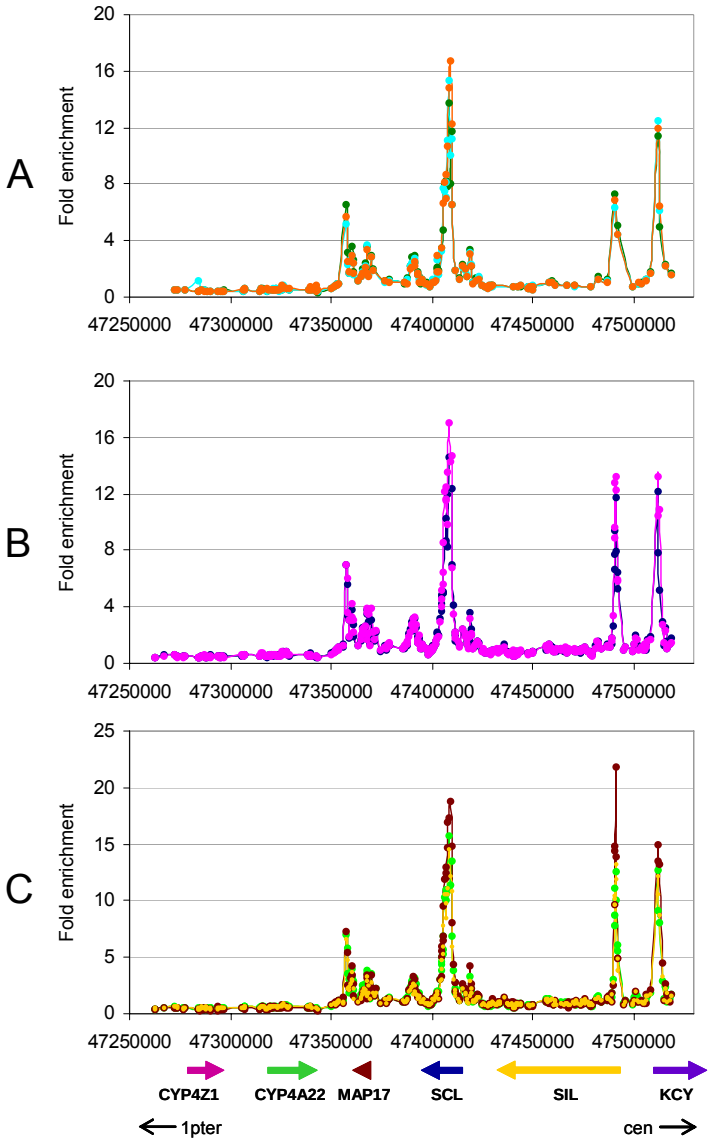


Figure 4.4: Performance of the SCL array platform in ChIP-chip assays for H3 K9/14 diacetylation in K562. Panel A: The orange (genomic), green (genomic) and turquoise (clone) profiles illustrate the reproducibility of the SCL array elements which were amplified from different PCR templates (Chapter 3). 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).

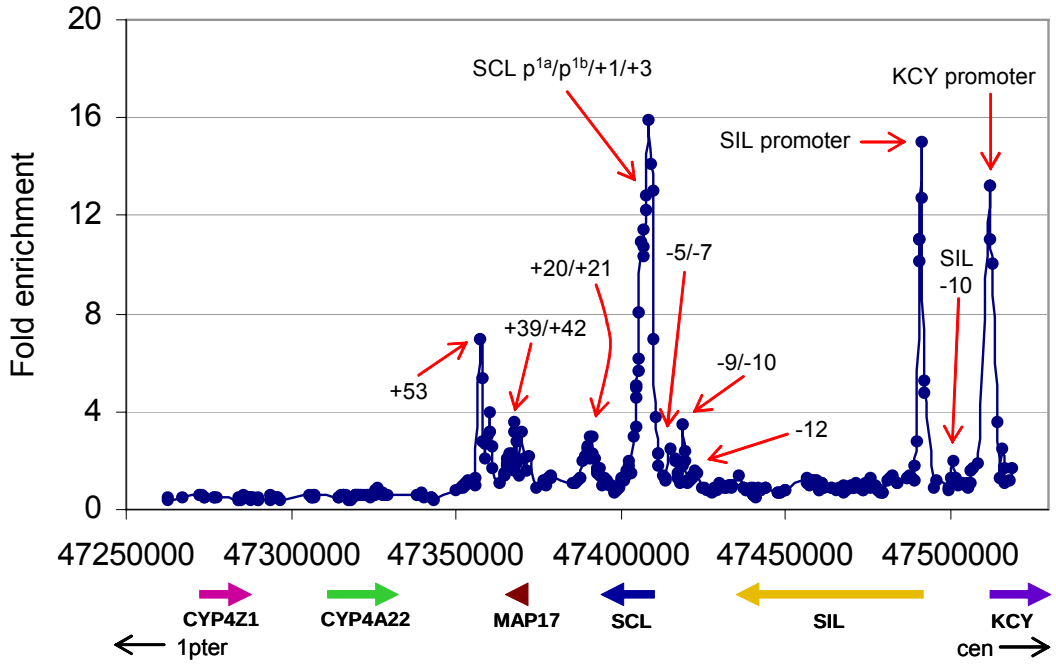


Figure 4.5: ChIP-chip profile of H3 K9/14 diacetylation across the human SCL locus in K562. Genomic regions showing significant enrichments for H3 K9/14 diacetylation are marked with red arrows. The genomic regions denoted by +1, +3, +7 etc. are based on their distance upstream (-) or downstream (+) 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).

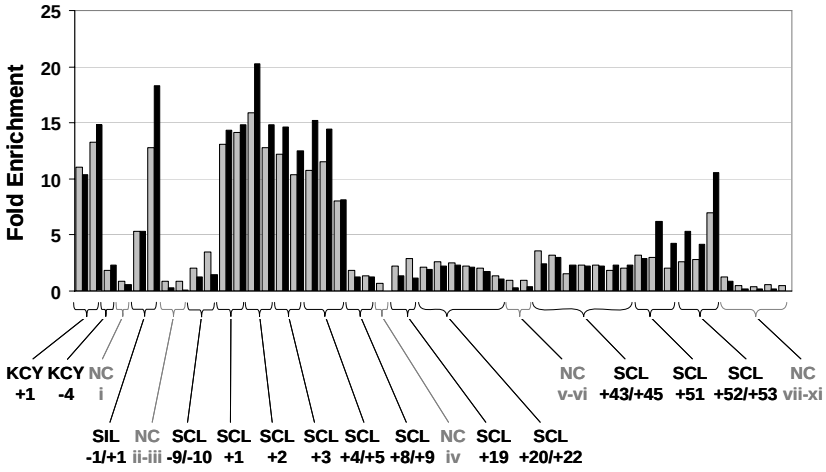


Figure 4.6: SyBr green real-time PCR of ChIP material for H3 acetylation across the human SCL locus in K562. The data points across the locus are shown on the x-axis. Enrichments reported by the array (grey bars) and those reported by real-time PCR (black bars) are shown as pairs for each region tested (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.

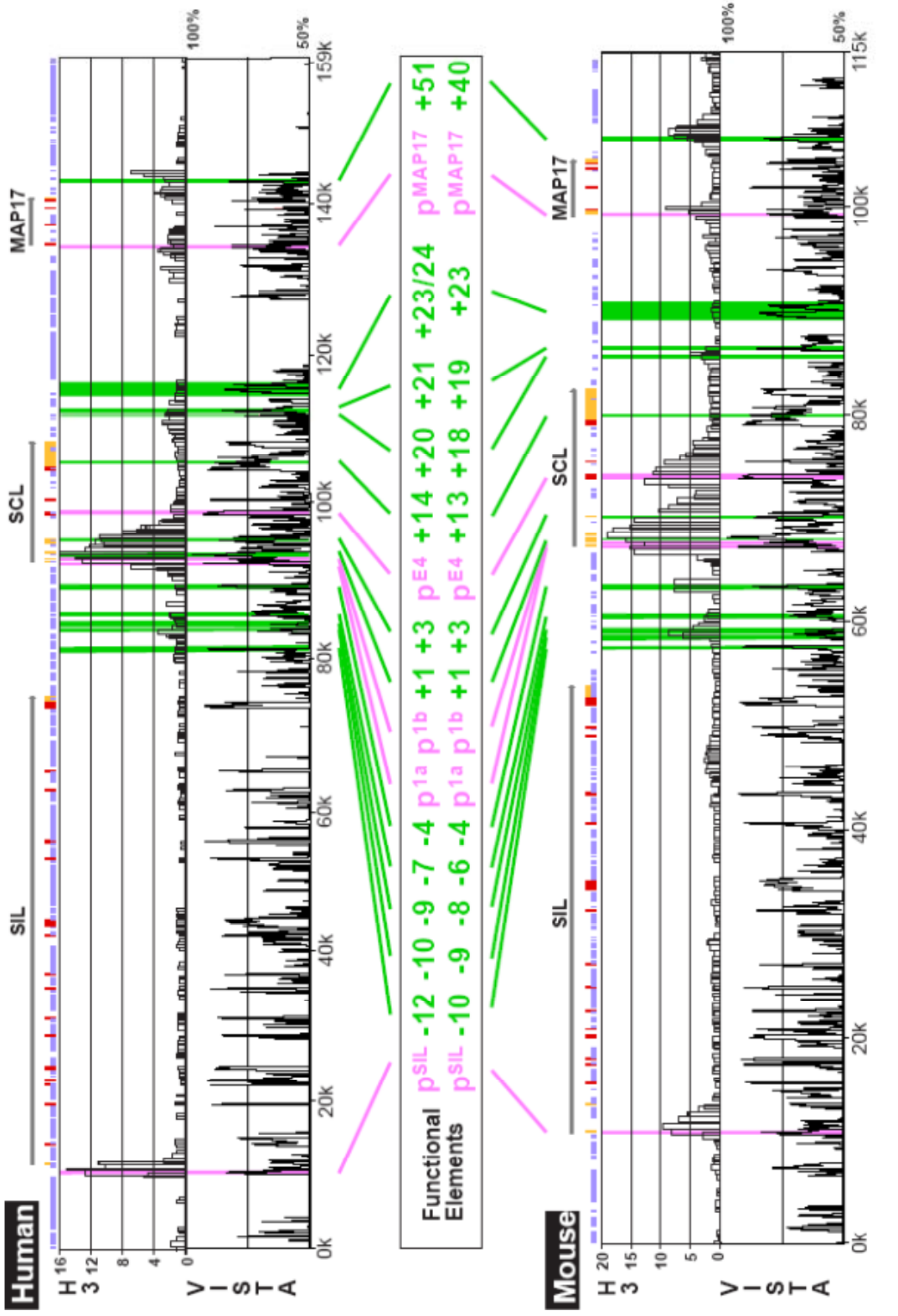


Figure 4.7: Comparison of histone H3 diacetylation profiles across the SCL locus in human K562 and mouse 416B cell lines. Profiles are shown for the genomic region spanning from upstream of the SIL promoter to the MAP17 gene and SCL erythroid enhancer (+51 in human/+40 in mouse). The top panel shows 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.

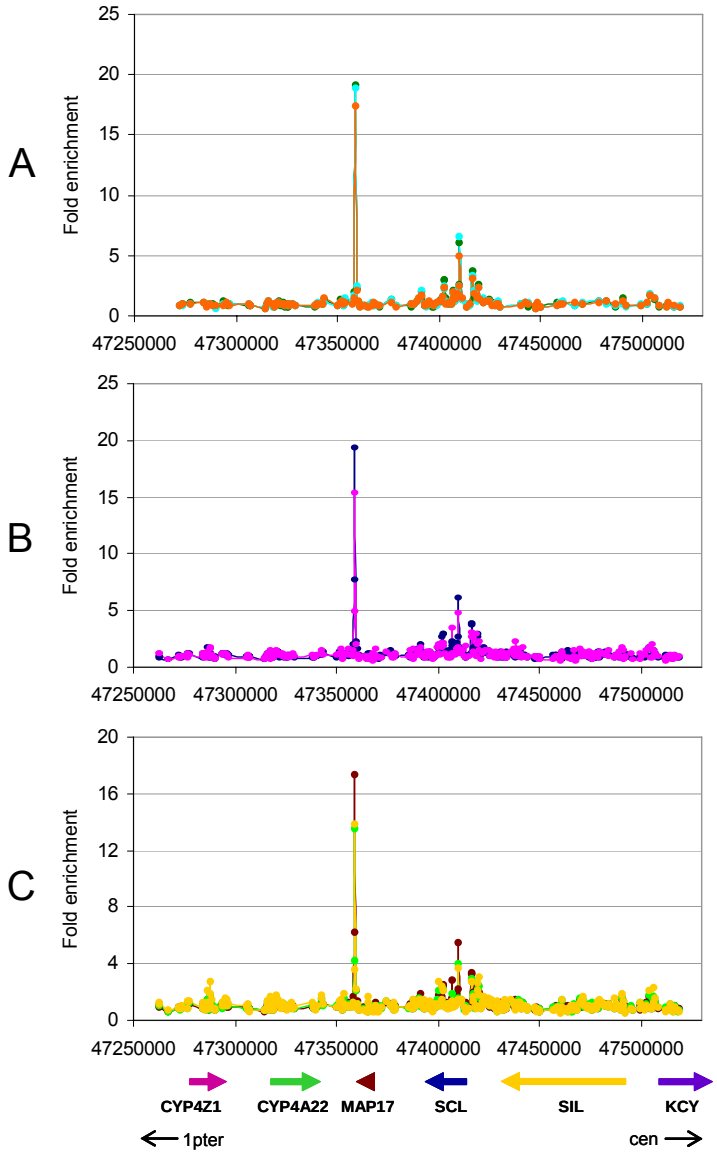


Figure 4.8: Performance of the SCL array platform in ChIP-chip assays for GATA-1 in K562. Panel A: The orange (genomic), green (genomic) and turquoise (clone) profiles illustrate the reproducibility of the SCL array elements which were amplified using different PCR templates (Chapter 3). Panel B: The 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).

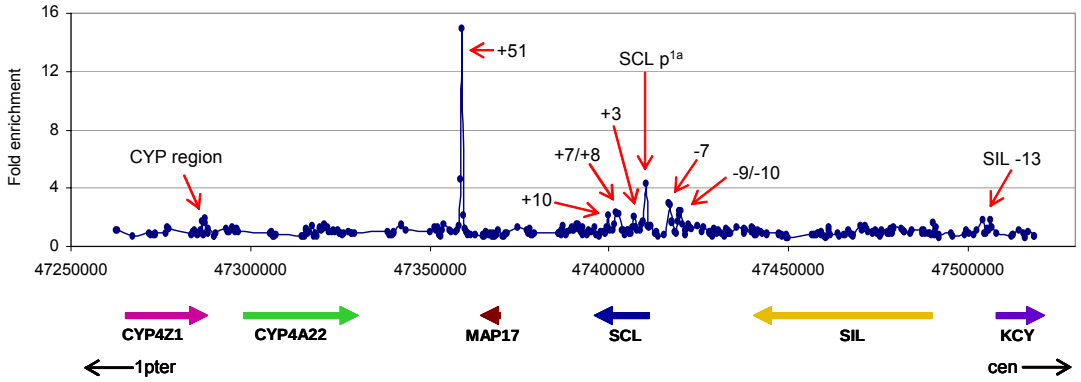


Figure 4.9: ChIP-chip profile for GATA-1 across the human SCL locus in K562. Genomic regions showing significant enrichments for GATA-1 are marked with red arrows. The maximum enrichment was seen at the +51 erythroid enhancer. The genomic regions denoted by +3, -7, -9, etc. are based on their distance 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).

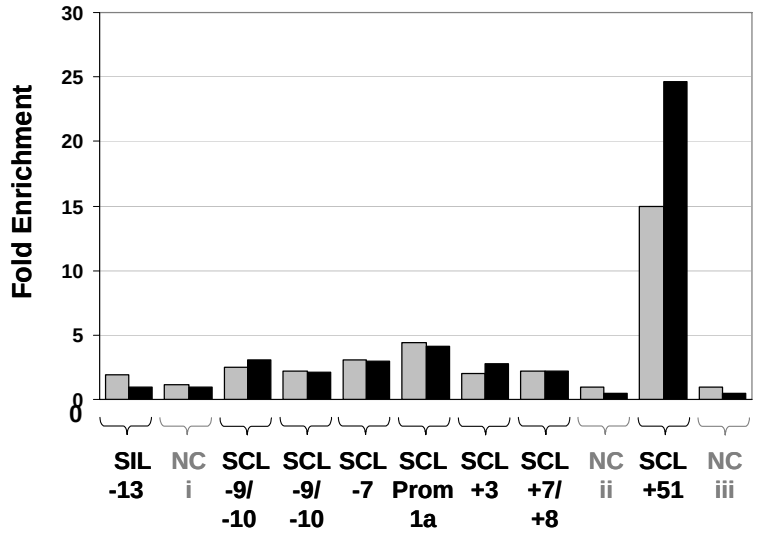
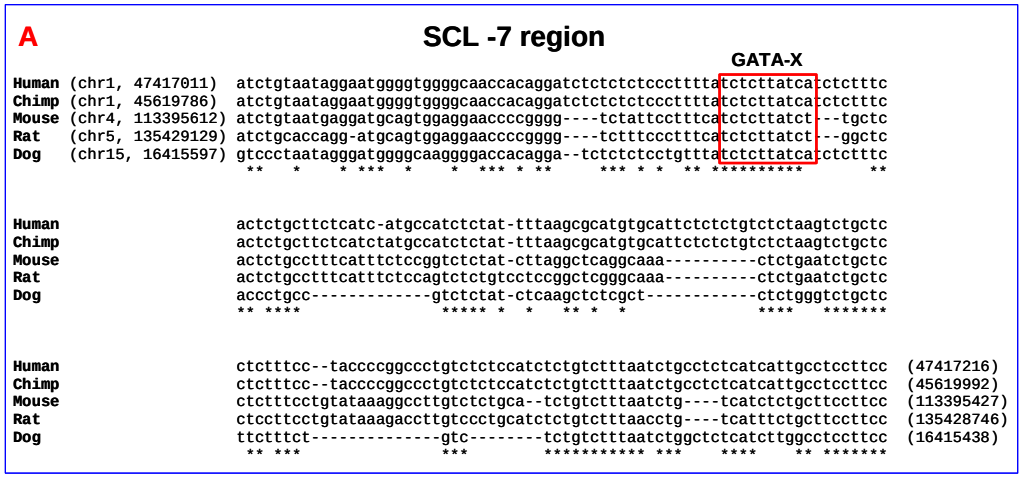


Figure 4.10: SYBR green real-time PCR of ChIP material for GATA-1 across the human SCL locus. The data points across the locus are shown on the x-axis. Enrichments reported by the array (grey bars) and those reported by real-time PCR (black bars) are shown as pairs for each region tested (along y-axis). 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)
                         **   *  *****               **** ** ** *********** **** *** *** ****
```

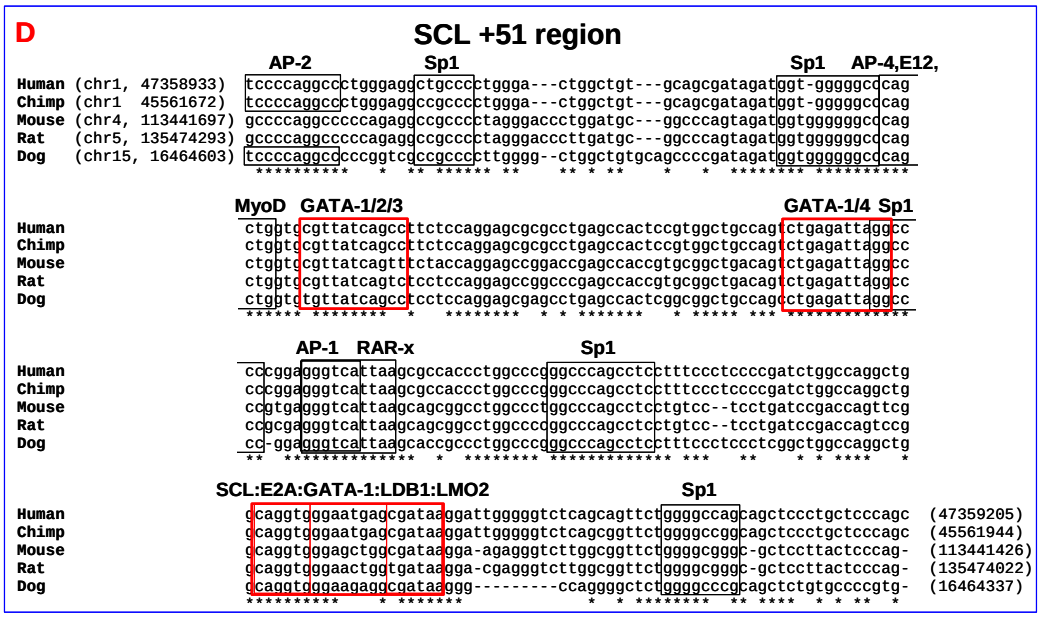


Figure 4.11: Conserved sequence analysis to identify conserved GATA and other transcription factors binding sites at SCL regulatory regions. The sequence alignments are shown for human chimp, mouse, rat and dog. Genomic sequence coordinates for each region of homology are shown in brackets at the beginning 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.

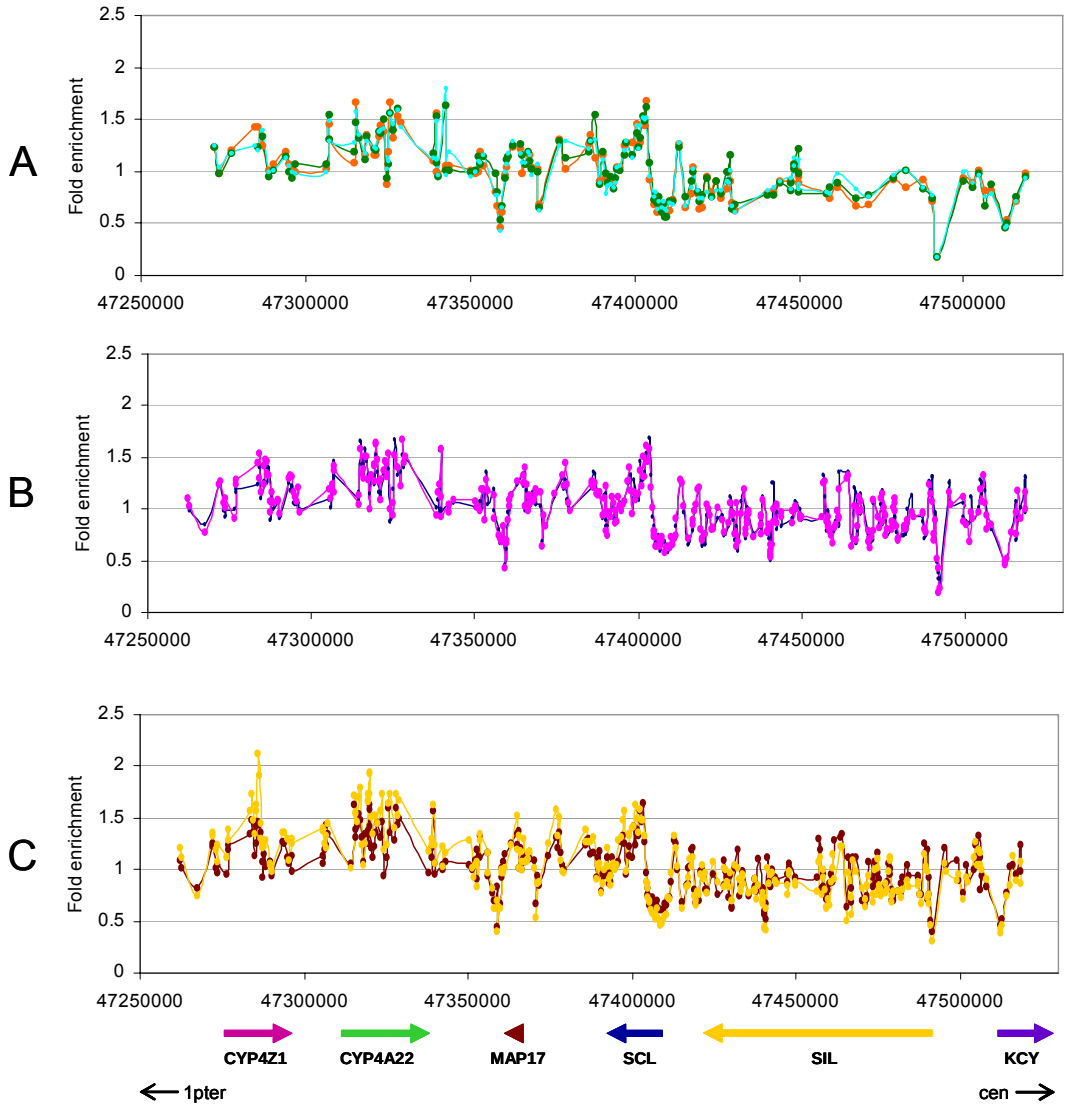


Figure 4.12: Performance of the SCL array platform in ChIP-chip assays for histone H3 in K562. Panel A: The orange (clone), turquoise (genomic) and green (genomic) profiles illustrate the reproducibility of the SCL array elements which were amplified from different PCR templates (Chapter 3). Panel B: 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.

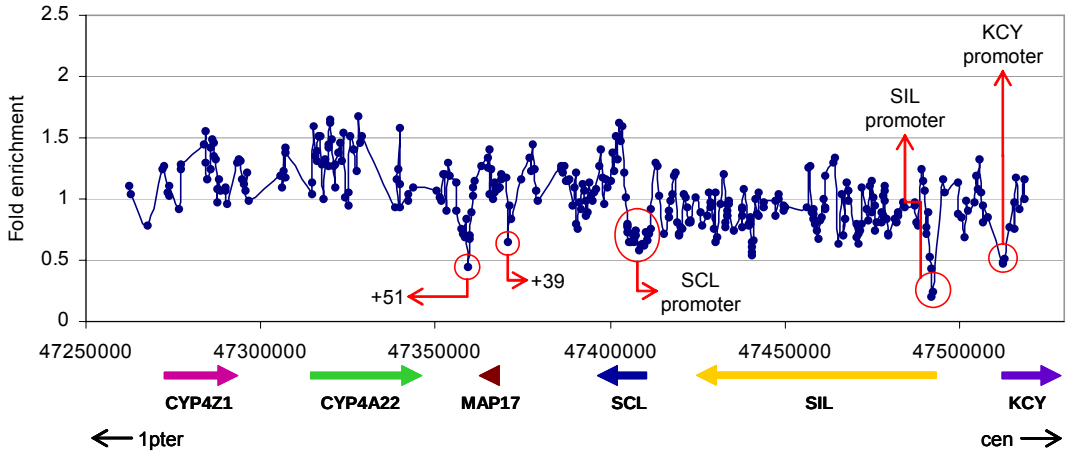


Figure 4.13: ChIP-chip profile for histone H3 across the human SCL locus in K562. Genomic regions showing depletion for histone H3 are marked with red circles. The y-axis represents the fold enrichments and the x-axis represents the genomic coordinates along human chromosome 1. The gene order and the

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).

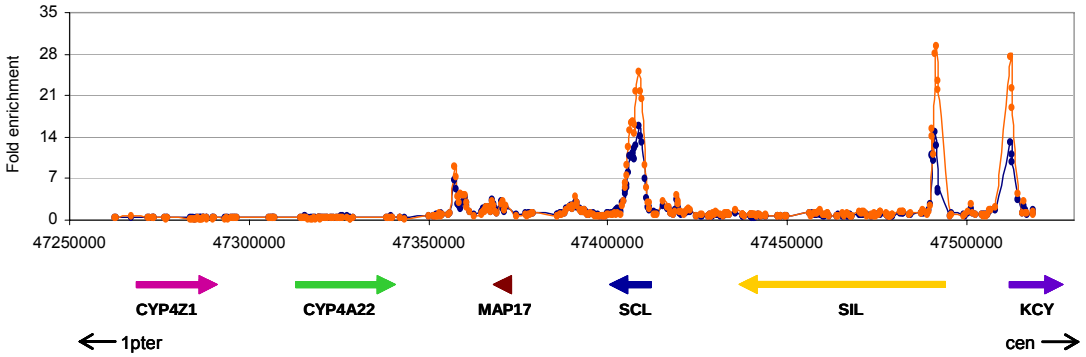


Figure 4.14: Normalization of the array data for H3 K9/14 diacetylation with respect to underlying nucleosome occupancy in K562. The blue and orange profiles represent un-normalized and normalized data respectively. After normalization, the peaks showing significant enrichments for H3 K9/14 diacetylation 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.