Chapter 6

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

6.1 Introduction

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

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

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

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

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

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

6.2 Aims of this chapter

The aims of the work reported in this chapter were:

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

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

Results

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

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

A

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

B

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

C

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

D

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

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

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

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

A

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

B

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

C

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

D

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

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

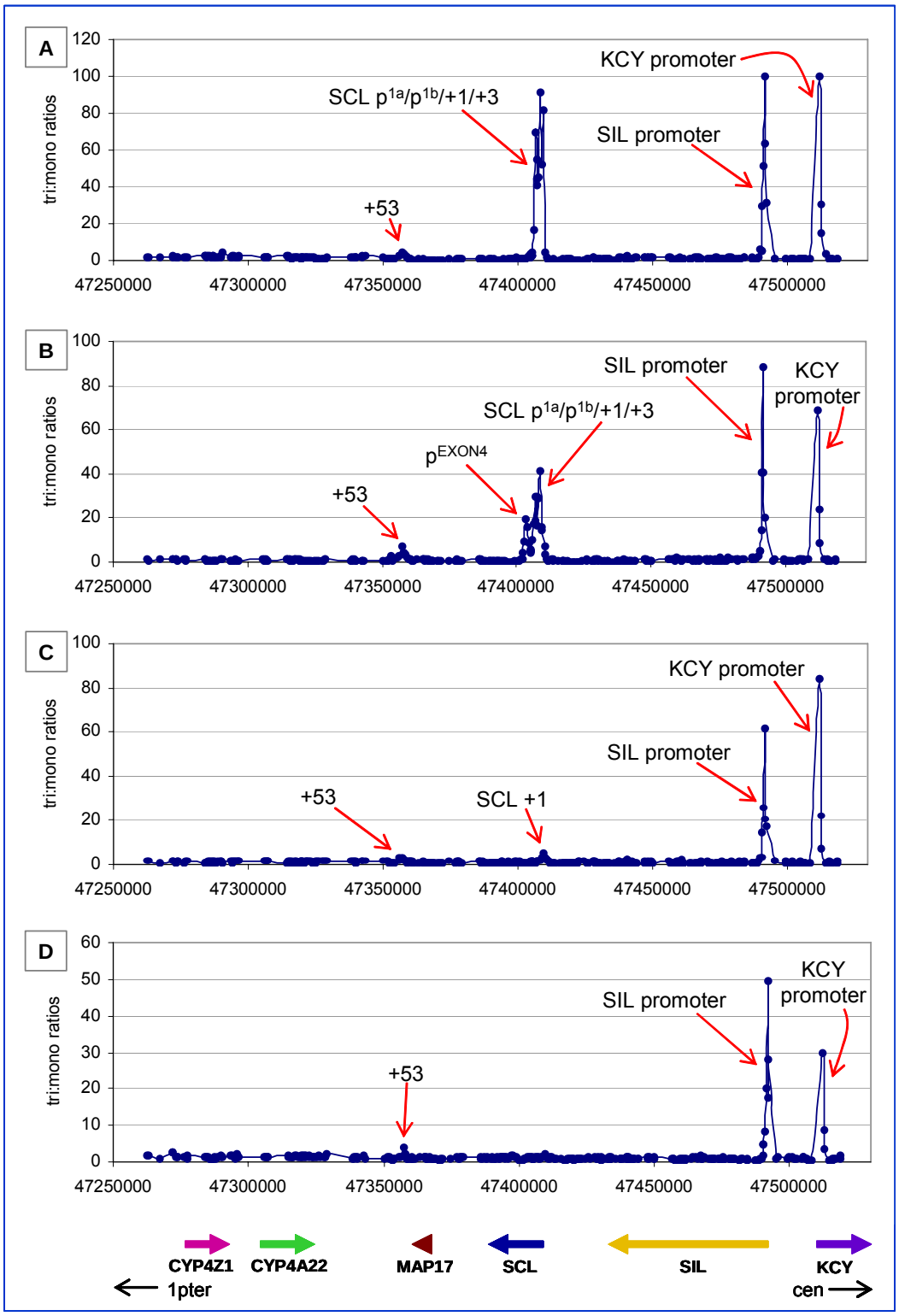
6.5 Relative levels of monomethylation to trimethylation of H3 lysine 4 correlate with transcriptional activity

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

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

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

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


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

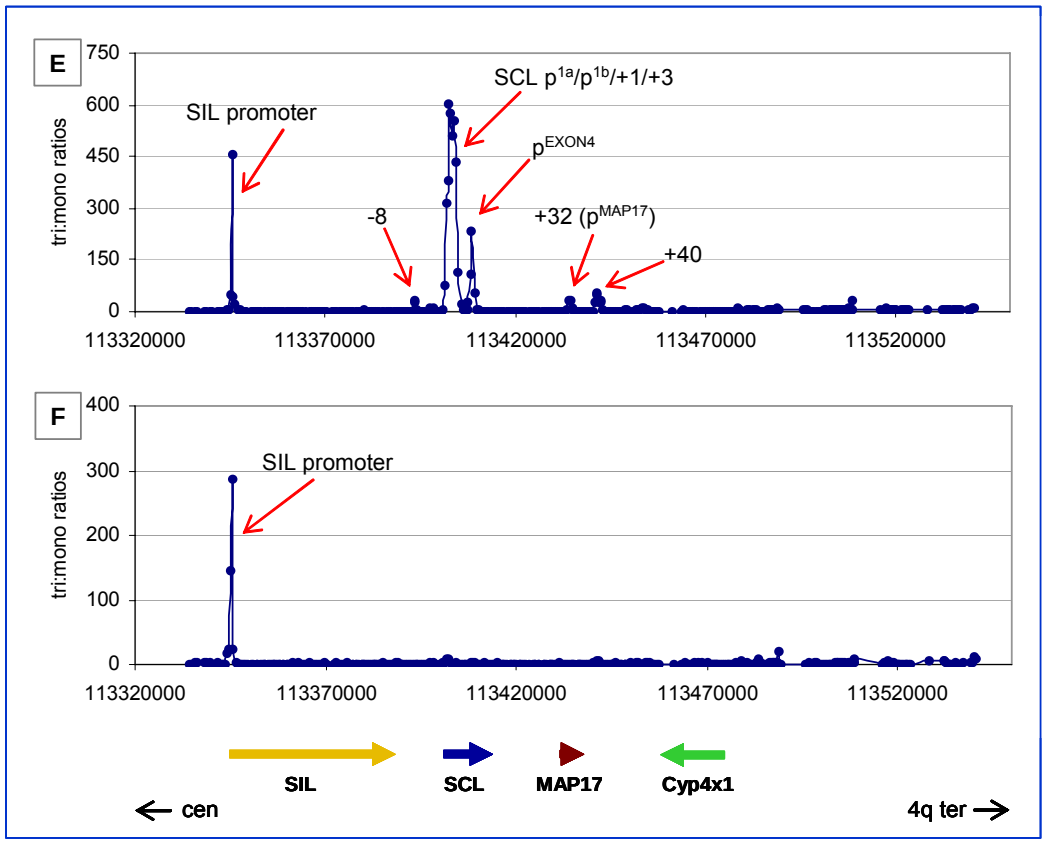


Figure 6.1: Profiles showing trimethylation to monomethylation ratios in the human and mouse cell lines. Panel A: K562, panel B: Jurkat, panel C: HL60, panel D: HPB-ALL, panel E: 416B and panel F: mouse ES cell line. The y-axes in the panels represent the tri- to mono- ratios in the respective cell line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

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

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

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

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

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

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

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

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

A

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

B

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

C

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

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

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

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

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

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

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

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

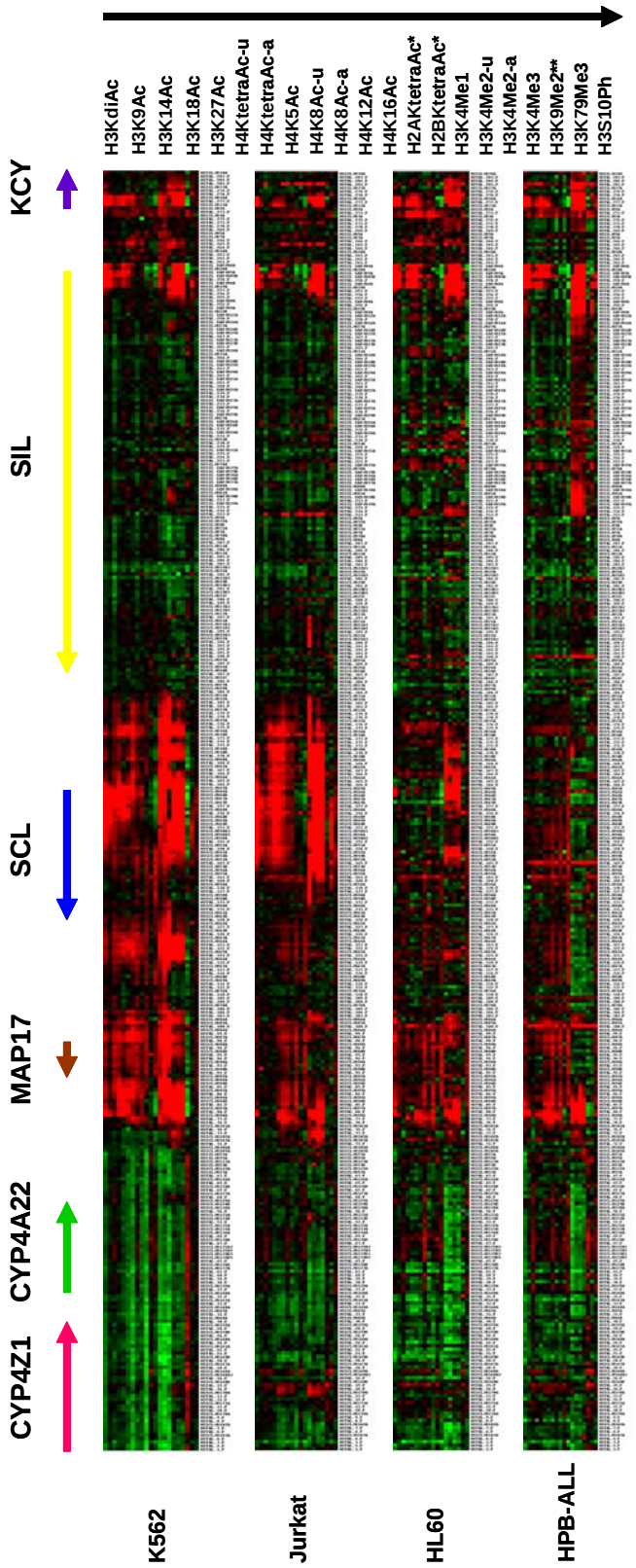


Figure 6.2: A digital output of ChIP-chip profiles across the SCL locus for all the histone modifications in human cell lines. The cell lines are listed at the bottom of the figure. The thick, coloured arrows along the y-axis represent the gene order and the direction of transcription. The array elements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

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

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

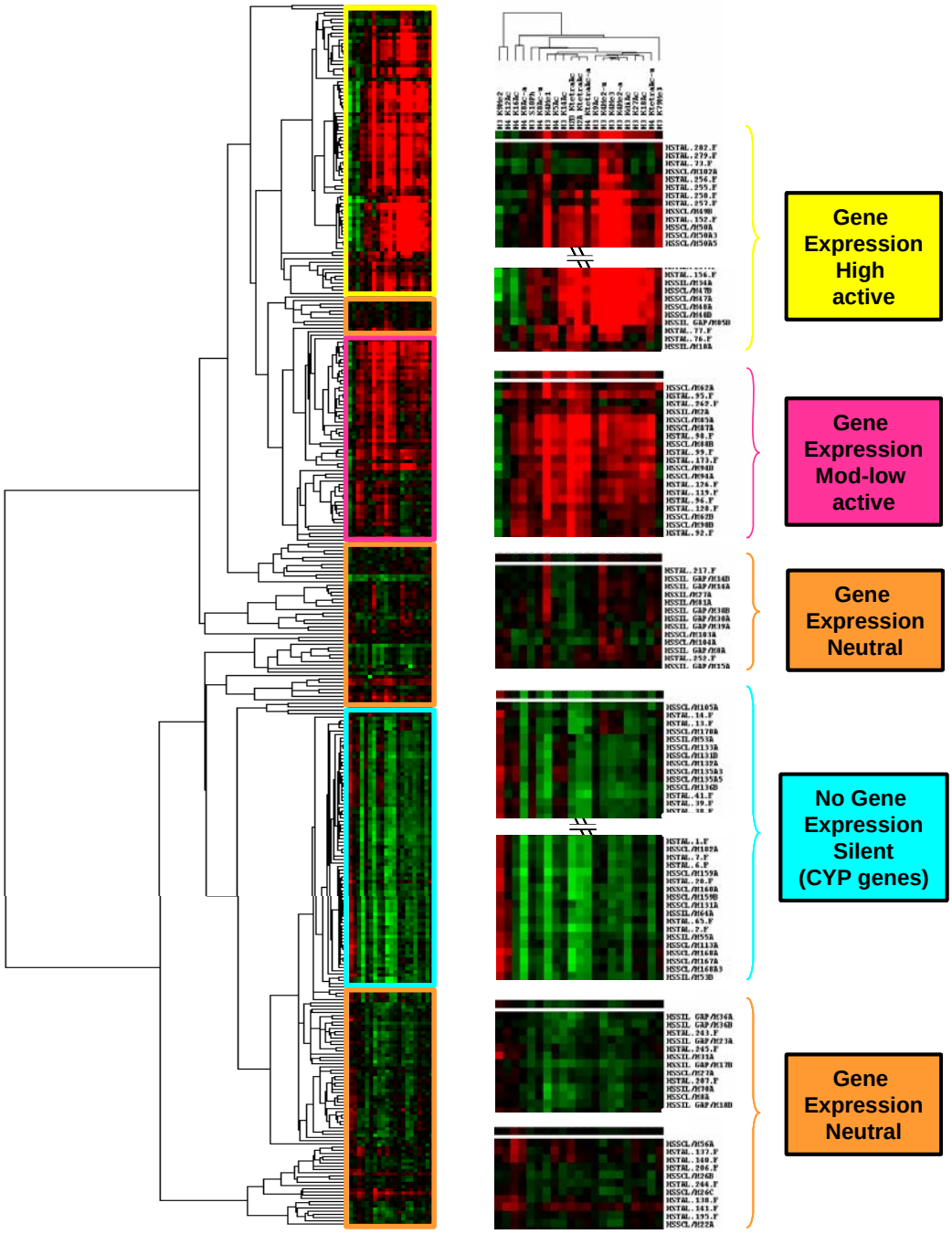


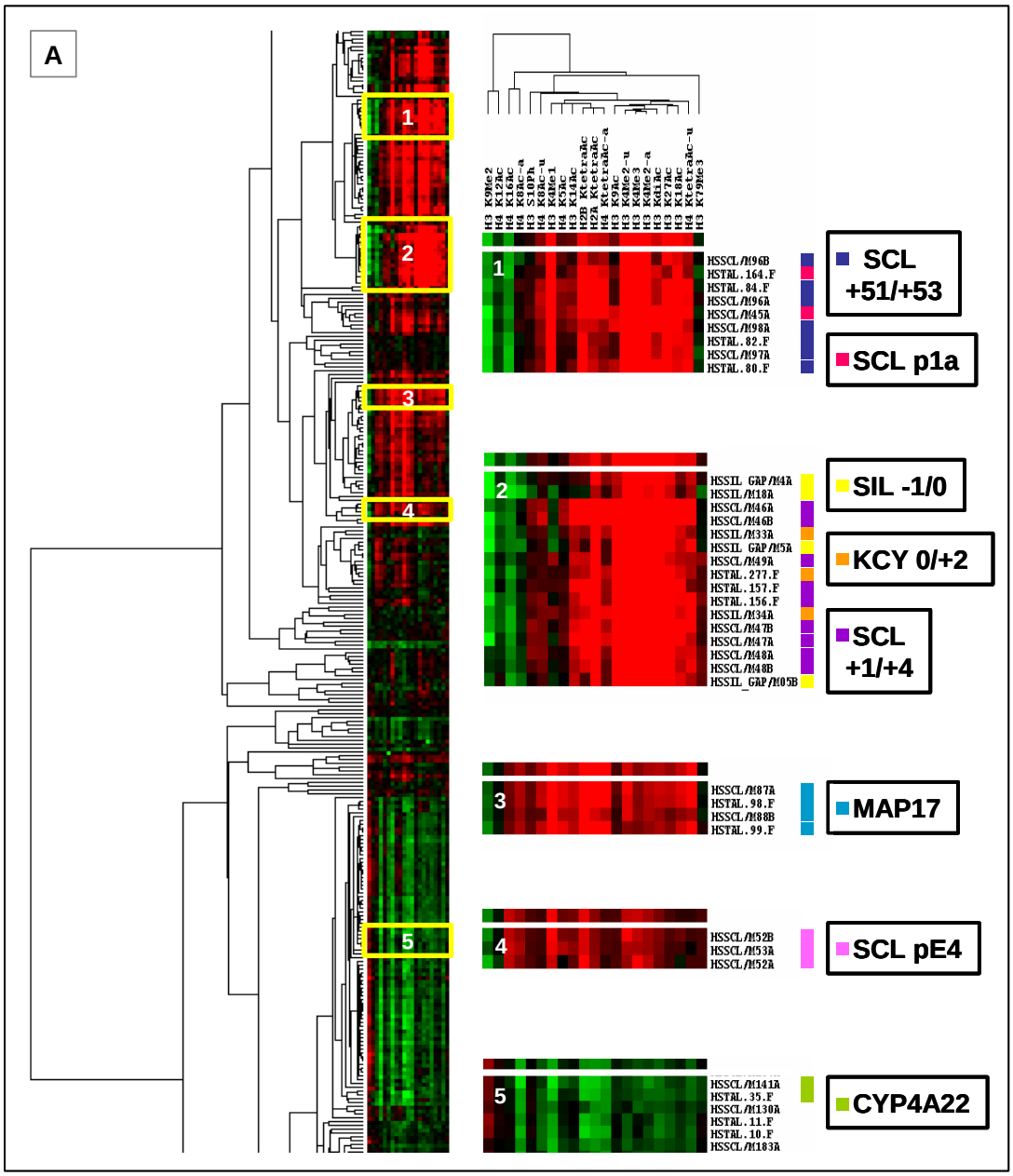
Figure 6.3: Hierarchical clustering defines active and inactive regions of chromatin across the SCL locus in K562. Hierarchical clustering was performed on 21 histone modification data sets in the K562 cell line. The genomic regions identified in the various clusters have been broadly classified into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

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

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

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

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



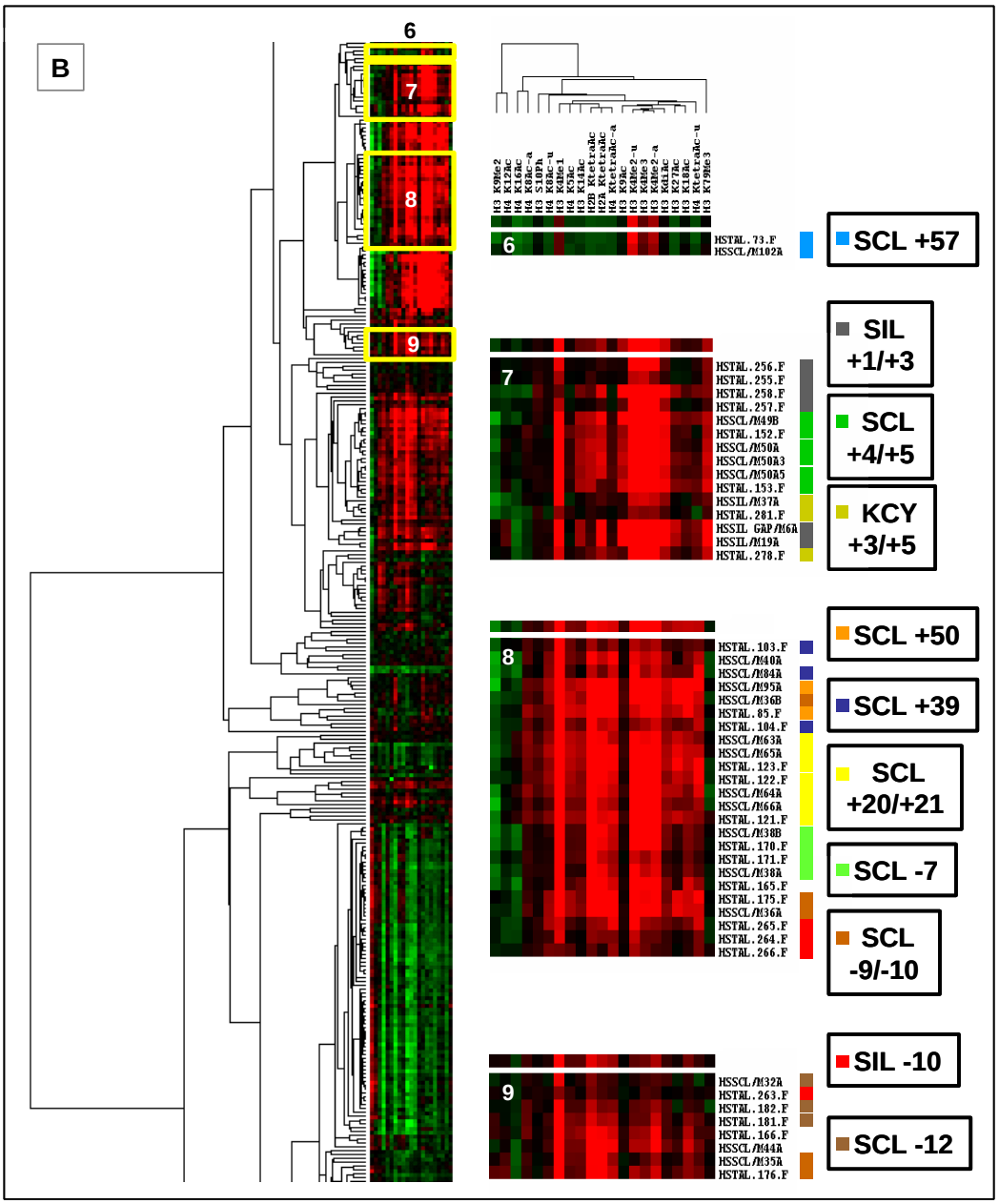


Figure 6.4: Identification and annotation of clusters of regulatory sequences across the SCL locus in K562. The description of the information shown in a cluster diagram is as described in Figure 6.3. In addition, the clusters are named 1 to 9 (boxed in yellow) and the magnified images of each cluster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

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

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

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

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

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

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

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

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

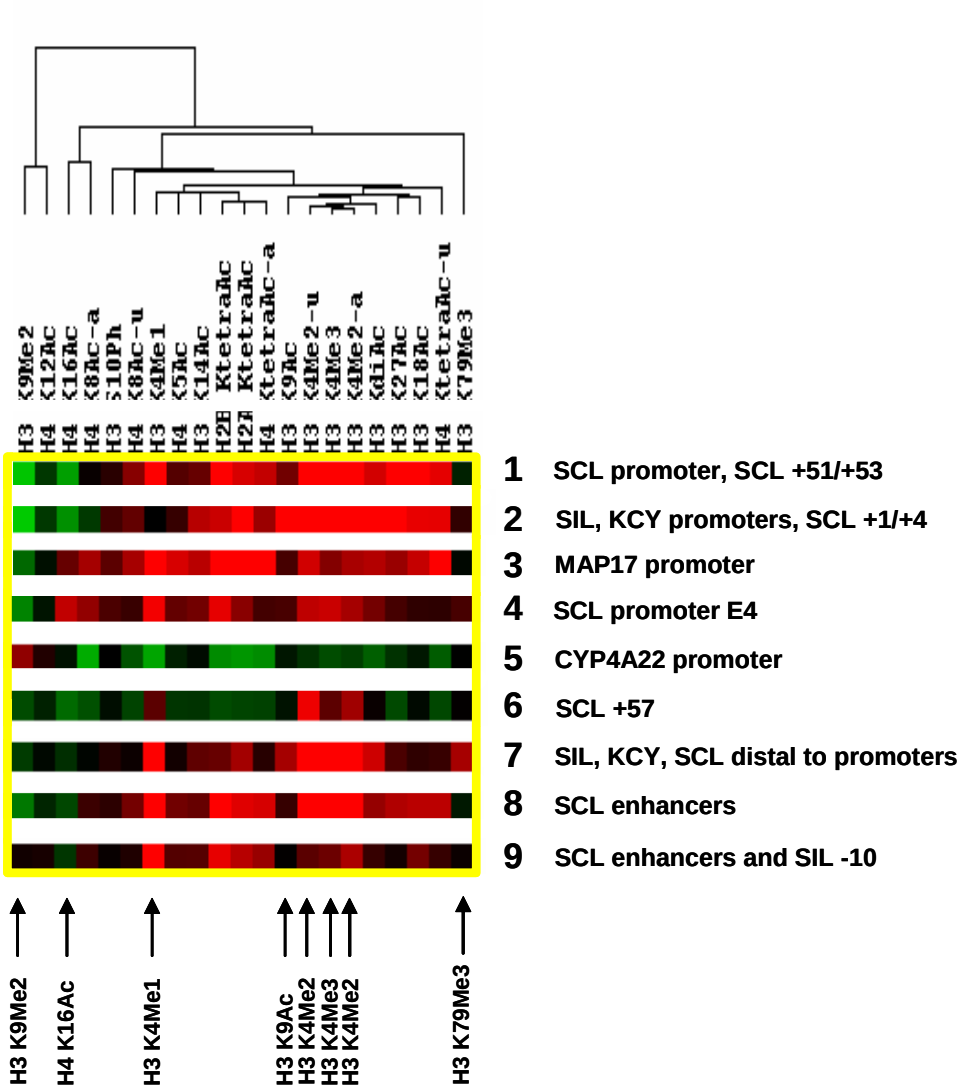


Figure 6.5: Identifying a minimal set of histone modifications for the SCL locus by comparing the "consensus" histone modifications of clusters 1 to 9 in K562. The clusters identified using hierarchical clustering of ChIP-chip data in K562 (clusters 1-9 in Figure 6.4) are represented by a "consensus"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

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

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

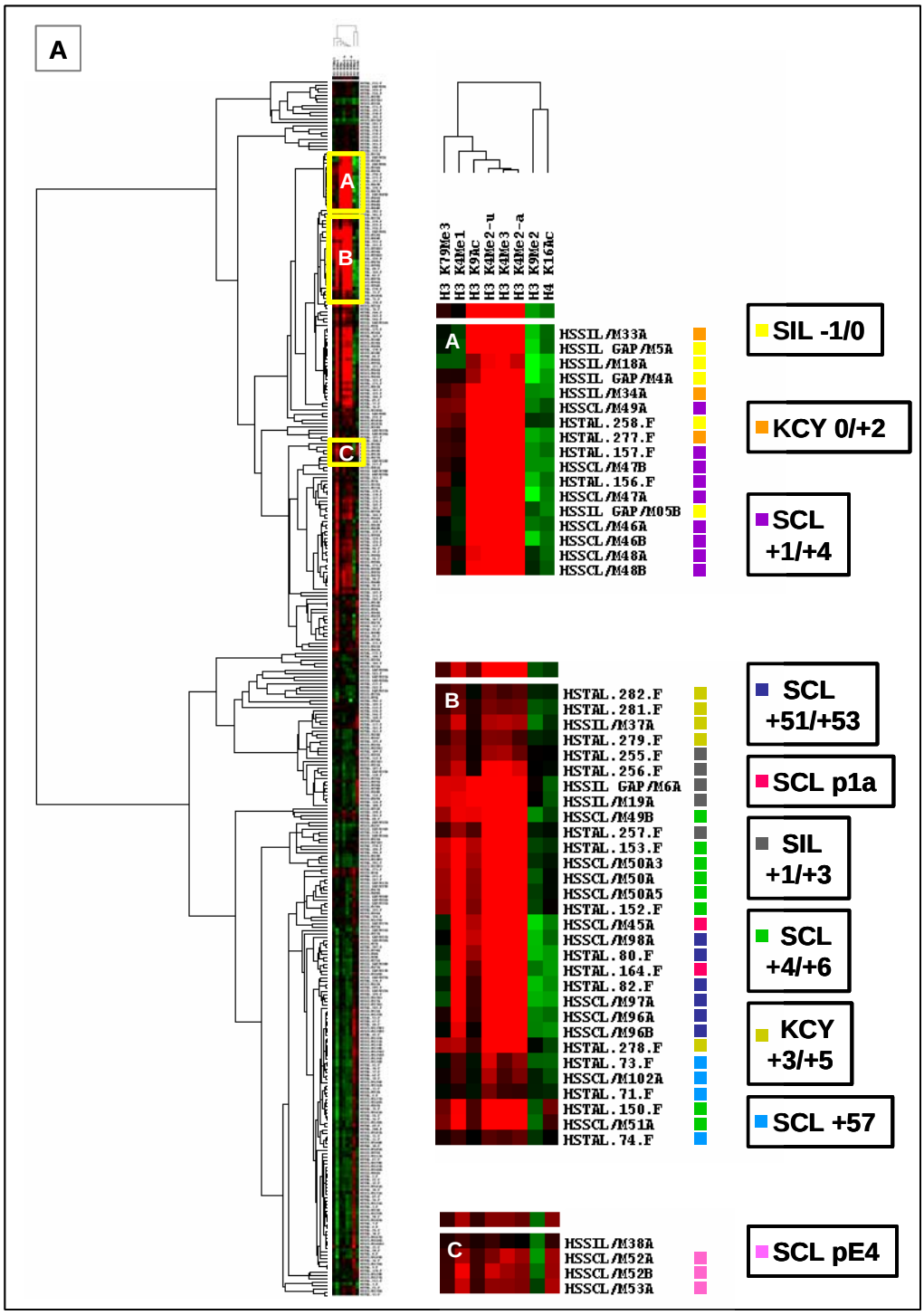
Monomethylation of H3 K4 is known to be a repressive mark (van Dijk et al. 2005) and was found to be depleted only at 5' ends of genes at or near promoters (cluster 2), whereas other regulatory regions including the downstream regions of active promoters (cluster 7), inactive promoters (cluster 4), low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

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

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

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

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


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

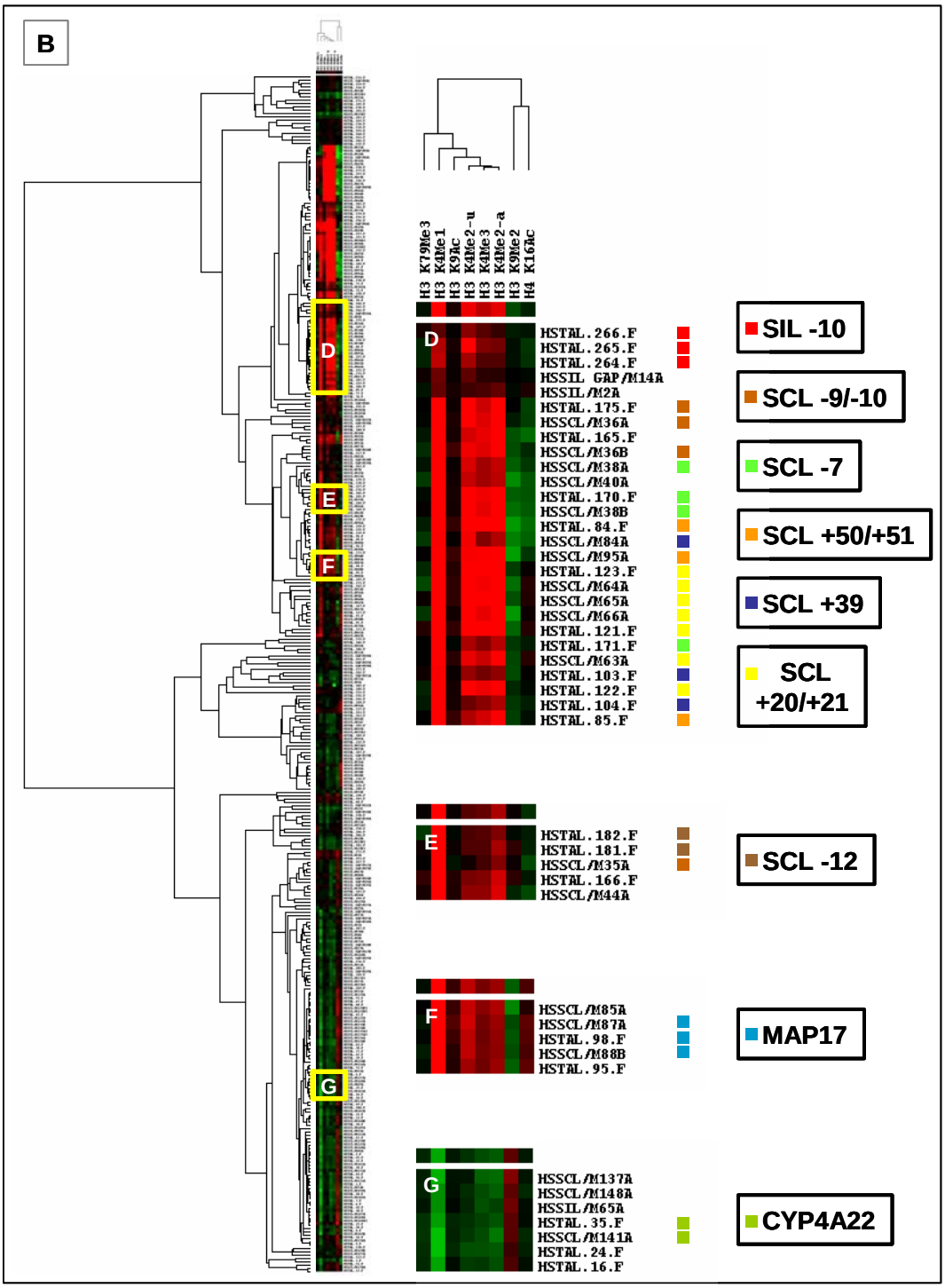


Figure 6.6: Hierarchical clustering of ChIP-chip data across the SCL locus in K562 using the minimal set of seven key histone modifications. Hierarchical clustering was performed using eight ChIP-chip datasets representing seven histone modifications across the SCL locus in K562 (two datasets were included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

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

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

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

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

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

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

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

6.10 Discussion

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

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

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

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

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

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

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

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

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

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

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

6.10.4 Identification of a consensus histone code at the SCL locus

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

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

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

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

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

ii) active from inactive promoters within the active domain

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

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

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

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

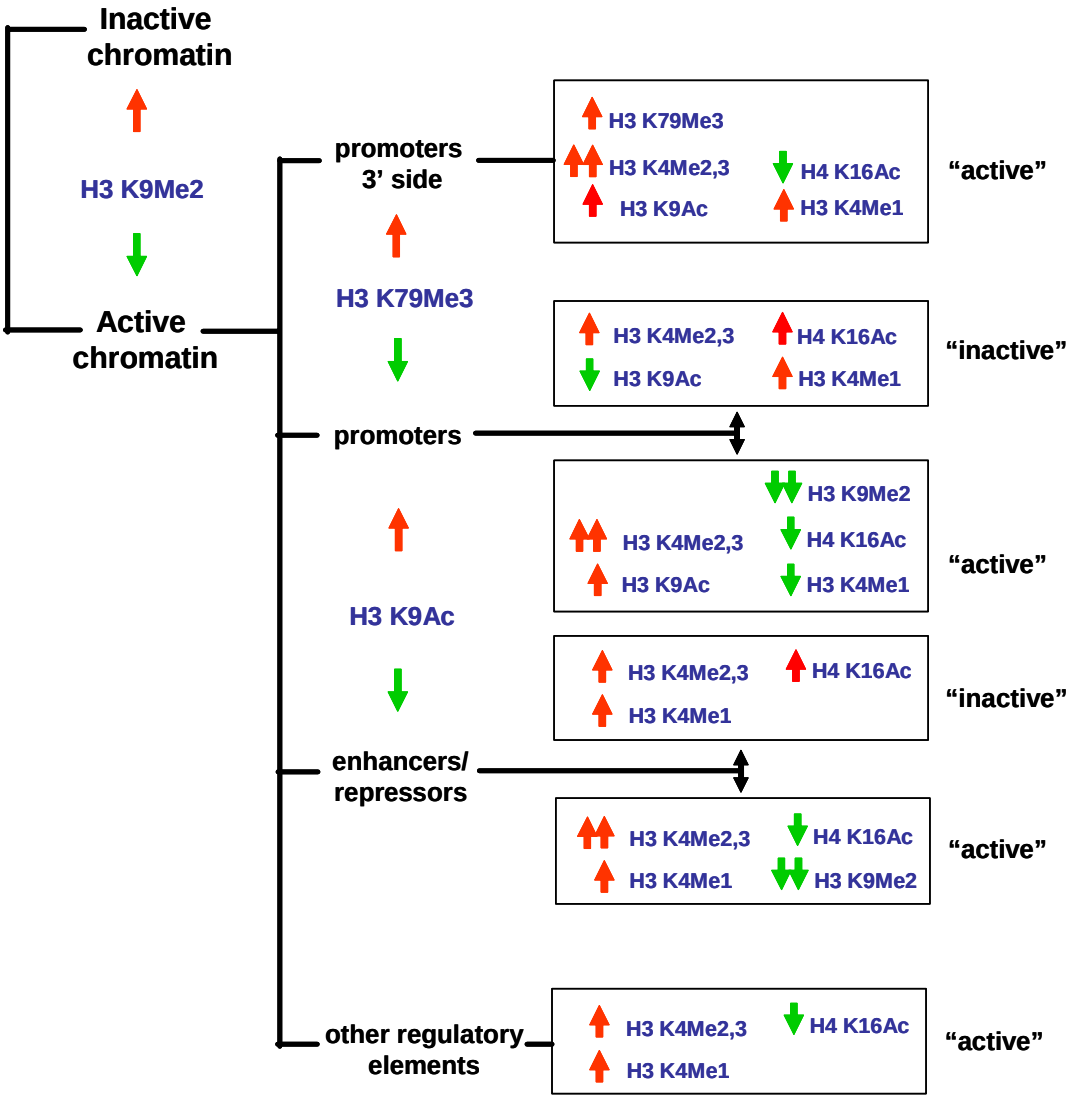


Figure 6.7: A consensus histone code for the regulatory regions across the SCL locus. This code was proposed based on the observations for the presence and/or absence of seven key histone modifications and other data presented in chapter 5. These histone modifications are: dimethylation of H3 K9 (H3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

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

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

6.10.5 Conclusions

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