

**Development of Chromatin Immunoprecipitation Microarray  
Technology for the Identification of Regulatory Elements in  
the Human Genome**

Shane Dillon  
St. Catharine's College

Thesis submitted for the degree of  
Doctor of Philosophy

University of Cambridge  
2008

## **Disclaimer**

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

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

Shane Dillon

7<sup>th</sup> February 2008

## **Abstract**

### **Development of Chromatin Immunoprecipitation Microarray Technology for the Identification of Regulatory Elements in the Human Genome**

The recent development of methods which use chromatin immunoprecipitation in combination with genomic microarrays (ChIP-chip) have transformed the way in which the dynamics of chromatin and the regulation of gene expression are studied genome-wide. The aim of this thesis was to use both conventional and improved ChIP-chip approaches to characterize a variety of regulatory elements (promoters, enhancers, and putative insulators) across selected regions of the human genome in a number of cell types. Firstly, histone modifications which define promoter and enhancer elements have been used to map and characterise these elements in the K562 cell line. In parallel, the transcription factor CTCF and its known binding partners USF1, USF2, and mSin3a have been used to characterize putative insulator element. The findings of this work are discussed. Secondly, whilst ChIP-chip assays have been successfully used to map numerous DNA-protein interactions, there are limitations restricting their use in the study of cell populations that are rare and/or of limited availability. Therefore, this thesis describes the development of ChIP-chip assays that allow as few as 10,000 cells to be used per ChIP condition. These experiments do not necessitate the need to perform the ChIP reactions in the presence of carrier chromatin or the need to amplify the ChIP material prior to hybridization onto the microarray. The distribution of histone methyl and acetyl modifications with material derived from 10 000 cells was detected with a similar efficiency as that obtained from more conventional ChIP-chip approaches. This method has been applied in the study of a range of histone modifications across regions of the human genome using limited numbers of human monocytes and human embryonic stem cells. The results presented in this thesis demonstrate that developments in ChIP-chip technology can be used to accelerate our understanding of the general principles of gene regulation in the human genome.

## Acknowledgements

First and foremost I would like to thank my supervisor Dave Vetrie for his help, advice, and support throughout my studies. His commitment to my project kept me going through difficult stages! I am also grateful to the Wellcome Trust for providing me with the opportunity to study at the Sanger Institute.

I would also like to thank a number of individuals at the Sanger Institute who made valuable contributions to my studies. In particular I would like to thank Pawandeep Dhami who taught me how to perform ChIP-chip. Thanks to Christoph Koch, Gayle Clelland, Sarah Wilcox and Ian Dunham, who developed the Sanger Institute ENCODE microarray. Thanks to the Microarray facility for printing fantastic arrays; in particular I would like to thank Cordelia Langford and Peter Allis. Former team members Alex Bruce, Philippe Couttet, Jonathan Cooper, Amanda Hall, and Johanna Jim helped in many ways and were a pleasure to work with. Thanks to the media preparation team, glassware team, the library staff, and Bee Ling for flow-sorting. Rob Andrews, Keith James, and Gregory Lefebvre provided great help with computational analysis of data. I would like to thank Stephan Beck and Steve Jackson for their input during thesis committee meetings.

Thanks to Enrique Millan and Roger Pedersen (Cambridge Institute for Medical Research) for providing human embryonic stem cells. Nicola Foad and Nick Watkins at the Department of Haematology, University of Cambridge, provided monocyte cells and monocyte microarray gene expression data respectively. Thanks also to Ulas Karaoz (Boston University) for performing ROC analysis of ChIP-chip data.

Finally, a huge thanks to my family for their constant support and encouragement during my studies. Carol and Ella, this thesis is for you. Now I can spend more time with you, something you both deserve!



# Table of Contents

<b>Chapter 1 Introduction</b> .....	1
1.1. Coding regions in the human genome .....	1
1.2. Non-coding elements in the human genome .....	2
1.3. Regulation of gene expression.....	4
1.4. Types of non-coding regulatory elements involved in controlling gene expression.....	4
1.4.1. Core and Proximal Promoters .....	5
1.4.2. Enhancers.....	7
1.4.3. Silencers/repressors .....	8
1.4.4. Insulators .....	9
1.4.4.1 Enhancer blocking insulators .....	11
1.4.4.2 Barrier Insulators .....	12
1.4.4.3 CTCF is a multifunctional protein .....	13
1.4.5. Locus control regions.....	15
1.4.6. Matrix Attachment regions (MARs).....	15
1.5. Proteins involved in transcriptional regulation.....	16
1.5.1. RNA polymerase complexes .....	16
1.5.2. Sequence-specific transcription factors .....	17
1.5.3. Proteins involved in chromatin remodelling and modification.....	19
1.5.3.1. ATP-dependent chromatin remodelling enzymes.....	19
1.5.3.2. Histone modifying enzymes.....	20
1.6. Epigenetic regulation of transcription.....	22
1.6.1. The Nucleosome – the fundamental repeating unit of chromatin.....	22
1.6.2. Histone modifications and their functions .....	26
1.6.2.1 Acetylation.....	27
1.6.2.2 Methylation .....	28
1.6.2.3. Phosphorylation.....	31
1.6.2.4. Ubiquitination.....	31
1.6.2.5. Sumoylation.....	32
1.6.2.6. ADP-ribosylation.....	32
1.6.3. The histone code hypothesis .....	32
1.6.4. DNA methylation.....	34
1.6.5. Regulation of embryonic stem cell pluripotency .....	35
1.6.5.1. Transcription factors involved in regulating pluripotency.....	35
1.6.5.2. Epigenetic regulation of pluripotency.....	36
1.7. Identification and characterisation of non-coding regulatory elements.....	37
1.7.1. Classical or low-throughput methods.....	38
1.7.1.1. DNA footprinting.....	38
1.7.1.2. DNase I hypersensitive site mapping .....	38
1.7.1.3. Electromobility shift assays.....	39
1.7.1.4. PCR-based methods for detecting DNA-protein interactions .....	39
1.7.1.5. Reporter-gene assays which confer function .....	40
1.7.2. Computational detection of regulatory elements .....	43
1.7.2.1. Promoter location prediction .....	43
1.7.2.2. Prediction of transcription factor binding sites .....	44

1.7.2.3.	Comparative sequence analysis.....	44
1.7.3.	Applications of genomic DNA microarrays to identify regulatory elements.....	46
1.7.3.1.	Transcript profiling.....	47
1.7.3.2.	Replication timing.....	48
1.7.3.3.	DNase I hypersensitive site microarrays.....	49
1.7.3.4.	Matrix attachment regions microarrays.....	50
1.7.3.5.	Chromatin immunoprecipitation microarrays (ChIP-chip).....	50
1.7.4.	Other ChIP-based methods.....	54
1.8.	Genomic microarray platforms used in this study.....	55
1.8.1.	The SCL genomic tiling path microarray.....	55
1.8.2.	The ENCODE project.....	58
1.8.2.1.	A summary of the pilot phase findings.....	58
1.8.2.2.	The Sanger Institute ENCODE Microarray.....	60
1.9.	Aims of this thesis.....	60
<b>Chapter 2 Materials and Methods</b> .....		61
2.1.	Composition of solutions.....	61
2.2.	Reagents.....	68
2.3.	Cells and cell lines.....	69
2.4.	Tissue culture.....	70
2.4.1.	Culturing of cell lines.....	70
2.4.2.	Cell cryopreservation.....	71
2.4.3.	Chromatin immunoprecipitation (ChIP).....	71
2.5.	Construction of the SCL tiling path microarray and the ENCODE tiling path microarray....	77
2.5.1.	Generation of tiling path amplicons.....	77
2.5.2.	Microarray slide printing and processing.....	77
2.6.	Hybridisation of microarrays.....	78
2.6.1.	Labelling by random priming of DNA samples.....	78
2.6.2.	Purification of labelled DNA samples.....	79
2.6.3.	Array hybridisation set-up using the Tecan HS 4800 hybridisation station.....	79
2.6.4.	Scanning and processing of ChIP-chip data.....	83
2.7.	Analysis of ChIP-chip data with respect to genomic features.....	84
2.8.	Statistical analysis.....	84
2.9.	Gene expression analysis.....	85
2.10.	Receiver operator characteristic (ROC) curve analysis.....	86
2.11.	DNA sequence motif analysis.....	86
2.12.	Real-time PCR.....	87
2.13.	Western blotting procedure.....	88
<b>Chapter 3 Applying ChIP-chip to study regulatory elements in 1% of the human genome</b> .....		92
3.1.	Introduction.....	92
3.2.	Aims of this chapter.....	96
3.3.	Overall strategy.....	97
3.4.	Criteria used for performing chromatin immunoprecipitation assays and microarray hybridisation.....	99
3.5.	Creating histone modification profiles across the ENCODE regions in K562 cells.....	104
3.5.1.	Assessing the performance of the ENCODE array.....	104
3.5.2.	Constructing histone modification and nucleosome density profiles.....	108

3.6.	Analysing the distribution of histone modifications .....	110
3.6.1.	Defining statistically enriched regions.....	110
3.6.2.	The distribution of histone modifications with respect to gene features .....	111
3.6.3.	Analysis of the combinatorial nature of histone modifications .....	112
3.7.	Distinct histone modification signatures define transcription start sites and distal elements in the ENCODE regions .....	114
3.8.	Histone modifications and transcription .....	118
3.8.1.	The relationship between histone modifications and transcription status .....	118
3.8.2.	Histone modifications and predicting expression status.....	121
3.8.3.	Nucleosome density at transcription start sites and distal elements .....	123
3.9.	Discussion .....	125
3.9.1.	Histone modification signatures associated with promoters and distal elements .....	126
3.9.2.	Histone modifications and transcriptional activity .....	127
3.9.3.	Nucleosome depletion at regulatory elements .....	129
3.9.4.	Analysis and interpretation of CHIP-chip data .....	129
3.9.5.	Conclusions .....	131
<b>Chapter 4 Identification and characterisation of binding sites of the insulator binding protein CTCF in the human genome .....</b>		<b>131</b>
4.1.	Introduction.....	132
4.2.	Aims of this chapter.....	137
4.3.	Overall strategy .....	137
4.4.	Assessing the specificity of transcription factor antibodies in western blotting assays.....	138
4.5.	Developing a CHIP-chip assay to detect putative insulators at the SCL locus .....	139
4.6.	Mapping and characterising CTCF binding sites in the ENCODE regions .....	142
4.6.1.	Implementation and validation of the CTCF CHIP-chip assay .....	142
4.6.2.	Distribution of CTCF-binding sites in the ENCODE regions.....	147
4.6.3.	CTCF sites at individual genes and gene clusters .....	150
4.7.	A comparative and sequence based analysis of CTCF sites in different cell types .....	153
4.8.	Analysis of other transcription factors implicated in CTCF or insulator function.....	157
4.8.1.	Developing assays using the SCL locus as model system .....	157
4.8.2.	Mapping the distribution of mSin3a, USF1, and USF2 binding sites in the ENCODE regions.....	159
4.8.3.	Analysing interactions between CTCF and mSin3a, USF1, and USF2.....	161
4.8.4.	Transcription factor binding and gene expression status .....	165
4.8.5.	Motifs analysis of CTCF, mSin3a, USF1 and USF2 binding sites.....	167
4.9.	Chromatin structure at insulators .....	172
4.9.1.	CTCF binding sites are located in accessible chromatin domains .....	172
4.9.2.	CTCF is located at the boundary between active and inactive chromatin domains.....	175
4.10.	Discussion .....	181
4.10.1.	Widespread distribution of CTCF binding sites in the human genome .....	181
4.10.2.	Cross-study comparison of CTCF binding sites .....	182
4.10.3.	CTCF co-localisation with mSin3a, USF1 and USF2 binding sites.....	185
4.10.4.	mSin3a interaction at promoters is associated with active gene expression .....	185
4.10.5.	Identification of transcription factor consensus binding motifs .....	186
4.10.6.	CTCF and chromatin structure .....	187

<b>Chapter 5 Optimisation of conditions for ChIP-chip when using cell types that are limiting in number</b> .....	190
5.1. Introduction.....	190
5.2. Aims of this Chapter .....	198
5.3. Overall strategy .....	199
5.4. The effect of titrating cell numbers in ChIP-chip and its effect on the detection of regulatory elements at the human SCL tiling locus .....	199
5.5. The effect of titrating antibody levels on low cell number ChIP .....	204
5.5.1. Histone H3 K9/K14 acetylation experiments .....	204
5.5.2. Histone H3 K4 trimethylation experiments .....	211
5.6. Examining the effect of protein-G agarose concentration on ChIP-chip experiments performed with low cell numbers .....	220
5.7. Assessing reproducibility of the limited cell numbers ChIP-chip method.....	222
5.8. Histone H3K4me3 microarray data correlates with real-time quantitative PCR of ChIP material derived from 10 <sup>5</sup> and 10 <sup>4</sup> cells.....	223
5.9. Assessing the impact of normalising ChIP-chip data with normal rabbit IgG data .....	225
5.10. Testing of other histone modification antibodies with the modified ChIP-chip method.....	231
5.11. Detection of transcription factor interactions using low cell number ChIP-chip .....	238
5.12. Discussion .....	240
5.12.1. The development of a ChIP-chip method applicable to small cell populations .....	241
5.12.2. Further optimisation of the modified ChIP-chip method .....	242
5.12.3. Comparisons with other methods.....	243
5.13. Conclusions.....	244
<b>Chapter 6 Histone modification profiles of human embryonic stem cells and lineage-committed human monocytes</b> .....	245
6.1. Introduction.....	245
6.2. Aims of this Chapter .....	249
6.3. Creating chromatin maps in pluripotent and lineage committed cells.....	249
6.4. Analysis of chromatin state in pluripotent hESCs and lineage committed monocytes .....	250
6.4.1. Promoter chromatin state in hESCs and monocytes.....	250
6.4.2. The chromatin state of promoters reflects developmental potential.....	253
6.4.3. Bivalent promoters are associated with developmental genes in hESCs and monocytes .....	256
6.4.4. The Polycomb Repressive Complex 2 (PRC2) subunit SUZ12 co-localises with bivalent developmental genes in human embryonic stem cells .....	258
6.4.5. Histone H3K36 trimethylation is associated with primary protein-coding transcripts and non-coding RNAs in hESCs and monocytes.....	262
6.4.6. Histone H3K79 trimethylation: Possible link to transcription elongation in hESCs and monocytes .....	267
6.5. A detailed analysis of histone acetylation and methylation modifications in human monocytes .....	269
6.5.1. Histone acetylation modifications are associated with active gene expression .....	270
6.5.2. Histone H3 lysine methylation patterns in human monocytes .....	272
6.5.2.1. Histone H3K4 methylation patterns at active and inactive genes .....	272
6.5.2.2. The histone signature of distal enhancer/repressor elements .....	273
6.5.2.3. Histone H3K9 methylation is implicated at actively transcribed genes.....	275
6.5.2.4. Histone H3K27 methylation states are found at active and inactive genes .....	277

6.5.2.5. Analysis of histone H3K36 methylation states .....	278
6.5.2.6. Histone H3K79 methylation states are associated with highly transcribed genes..	280
6.6. Discussion .....	281
6.6.1. Bivalent chromatin structures are present in pluripotent hESCs and lineage committed monocytes.....	282
6.6.2. Histone H3K36me3 and H3K79me3 are associated with different stages of transcription in hESCs and monocytes .....	284
6.6.3. Identification of a consensus histone code for active and inactive genes in human monocytes .....	285
<b>Chapter 7 Summary and future work .....</b>	<b>289</b>
7.1. Summary of the work presented in this thesis .....	289
7.1.1. Application of microarray technology for large-scale characterisation of promoter and distal enhancer/repressor elements in the human genome (Chapter 3).....	289
7.1.2. ChIP-chip analysis of the insulator binding protein CTCF and its associated transcription factors (Chapter 4) .....	290
7.1.3. Development of a modified ChIP-chip method for use with fewer cells (Chapter 5).....	291
7.1.4. Application of the modified ChIP-chip method to study chromatin regulation during differentiation (Chapter 6).....	292
7.1.5. Identification of a consensus histone code in human monocytes (Chapter 6) .....	293
7.2. Future Work.....	293
7.2.1. Further development of ChIP-chip assays for characterising regulatory elements and the role of histone modifications in gene regulation .....	294
7.2.2. Applications of microarrays for characterising other features of the human genome ...	296
7.2.3. Testing the proposed histone code.....	298
7.2.4. Further characterisation of putative CTCF insulator elements .....	299
7.3. Final thoughts.....	300
<b>References .....</b>	<b>301</b>
<b>Appendix 1 .....</b>	<b>338</b>

## List of Figures

Figure 1.1: Genome organisation of the major model organisms .....	3
Figure 1.2: The typical regulatory features involved in the regulation of gene expression .....	5
Figure 1.3: REST-mediated repression of target genes.....	9
Figure 1.4: The two types of insulator activity .....	10
Figure 1.5: Genomic organisation of the chicken $\beta$ -globin locus.....	13
Figure 1.6: Sites of histone modifications.....	26
Figure 1.7: Functional plasmid-based reporter gene assays for the identification of regulatory elements .....	42
Figure 1.8: The basic principles of two-colour competitive hybridisation .....	47
Figure 1.9: Outline of the carrier CHIP (CCHIP) method.....	53
Figure 1.10: Microarray surface chemistry .....	55
Figure 1.11: The genomic region included on the human SCL tiling path microarray.....	56
Figure 1.12: Histone acetylation and methylation states define regulatory elements at the SCL locus .....	57
Figure 3.1: Schematic representation of the overall strategy used to map regulatory interactions across 1% of the human genome.....	98
Figure 3.2: Analysis of cell growth by flow-cytometry.....	100
Figure 3.3: Electrophoresis of CHIP DNAs .....	101
Figure 3.4: Labeling of DNA by a random priming method .....	102
Figure 3.5: Electrophoresis of fluorescently labeled DNAs.....	103
Figure 3.6: A composite image of the Sanger Institute ENCODE array.....	106
Figure 3.7: Assessing the performance of the ENCODE microarray in independent CHIP-chip assays.....	107
Figure 3.8: An example of the histone modification and nucleosome density maps generated across the ENCODE regions .....	109
Figure 3.9: Distribution of histone modifications in K562 cells.....	111
Figure 3.10: The relative occurrence of overlapping histone modification sites in K562 .....	112
Figure 3.11: The extent of overlapping H3K4 methylation sites in K562. The pie-chart illustrates the occurrence of the seven H3K4 methylation combinations. ....	113
Figure 3.12: The distribution of histone H3K4 methylation combinations relative to the nearest transcriptional start site (TSS) .....	115
Figure 3.13: Histone modification profiles at TSSs and distal sites.....	117
Figure 3.14: Histone modification profiles for non-expressed and expressed genes in K562 cells .....	119
Figure 3.15: Profiles showing H3K4me3 to H3K4me1 ratios at the promoters of active and inactive genes in K562 cells .....	120
Figure 3.16: Receiver operating characteristic (ROC) curves illustrating the predictive power of histone modifications in K562 cells.....	122
Figure 3.17: Nucleosome occupancy at transcription start sites and distal elements in K562 cells .....	124
Figure 3.18: Histone modification levels relative to nucleosome occupancy in K562 cells.....	125
Figure 4.1: CTCF enhancer-blocking insulator at the Igf2/H19 imprinted region.....	134
Figure 4.2: A 20-mer motif is recognised by CTCF .....	136
Figure 4.3: Western blot analysis of CTCF, mSin3a, USF1 and USF2 in K562 cells .....	139

Figure 4.4: ChIP-chip profile of CTCF binding across the SCL locus in K562 before and after Goat IgG normalisation.....	141
Figure 4.5: CTCF binding sites in the IGF2/H19 locus.....	143
Figure 4.6: CTCF binding sites at the $\beta$ -globin locus.....	144
Figure 4.7: Comparison of CTCF binding sites in the K562 cell with binding sites identified in primary human fibroblast, IMR90 cells.....	146
Figure 4.8: The binding intervals and distribution of CTCF sites relative to transcription start sites (TSSs) in K562.....	147
Figure 4.9: Correlation of CTCF binding events with gene density in the ENCODE regions.....	148
Figure 4.10: The distribution of CTCF binding sites.....	149
Figure 4.11: CTCF binding intervals indicate multiple CTCF sites can bind in close proximity...	150
Figure 4.12: CTCF binding sites flank a cluster of genes in ENCODE region Enm006.....	151
Figure 4.13: Genes can contain several CTCF binding sites.....	152
Figure 4.14: A comparison of CTCF binding sites in different cell types and between studies ...	156
Figure 4.15: ChIP-chip profile of mSin3a interactions across the SCL locus in K562 before and after normal rabbit IgG normalisation.....	158
Figure 4.16: The distribution of USF1, USF2, and mSin3a sites of interaction in the ENCODE regions in K562 cells.....	159
Figure 4.17: Distribution of mSin3a, USF1, USF2 interactions.....	161
Figure 4.18: CTCF binding sites that overlap with binding sites of mSin3a, USF1, and USF2 are found in diverse locations.....	163
Figure 4.19: Predictive power of transcription factor binding for gene expression in K562.....	166
Figure 4.20: Nucleosome density and FAIRE profile at CTCF binding sites in the ENCODE regions.....	173
Figure 4.21: CTCF binding sites co-localise with sites of DNase I hypersensitivity at the $\alpha$ - and $\beta$ -globin loci.....	175
Figure 4.22: ChIP-chip profile of Histone H3K27me3 across the SCL locus in K562.....	176
Figure 4.23: CTCF binding sites demarcate the boundary between chromatin regions associated with active and inactive histone modifications.....	178
Figure 5.1 Principles of ChIP.....	192
Figure 5.2: ChIP-chip profiles of H4 K9/K14 diacetylation across the human SCL locus in K562 for a range of cell numbers.....	201
Figure 5.3: Titration of H3K9/K14 diacetylation antibody in ChIP assays performed with $10^7$ K562 cells.....	205
Figure 5.4: Titration of H3K9/K14 diacetylation antibody in ChIP assays performed with $10^6$ K562 cells.....	206
Figure 5.5: Titration of H3K9/K14 diacetylation antibody in ChIP assays performed with $10^5$ K562 cells.....	207
Figure 5.6: Titration of H3K9/K14 diacetylation antibody in ChIP assays performed with $10^4$ K562 cells.....	208
Figure 5.7: Titration of H3K4me3 antibody in ChIP assays performed with $10^7$ K562 cells.....	212
Figure 5.8: Titration of H3K4me3 antibody in ChIP assays performed with $10^6$ K562 cells.....	213
Figure 5.9: Titration of H3K4me3 antibody in ChIP assays performed with $10^5$ K562 cells.....	214
Figure 5.10: Titration of H3K4me3 antibody in ChIP assays performed with $10^4$ K562 cells.....	215
Figure 5.11: A comparison of the optimal antibody concentrations for titration assays performed with H3 acetylation and H3K4me3 antibodies.....	219
Figure 5.12: Titration of protein-G agarose in ChIP assays performed with $10^5$ K562 cells and 0.5 $\mu$ g of H3 acetylation antibody.....	221

Figure 5.13: Biological replicates for ChIP-chip assays performed with $10^4$ cells.....	222
Figure 5.14: SYBR green real-time quantitative PCR of H3K4me3 ChIP material from reduced cell numbers .....	224
Figure 5.15: How normalisation with a normal rabbit IgG affects the spread of data for a ChIP-chip experiment performed with $10^5$ cells.....	226
Figure 5.16: Rabbit IgG normalisation of a ChIP-chip experiment performed with $10^5$ cells .....	227
Figure 5.17: Rabbit IgG normalisation of data derived from ChIP-chip experiments performed with $10^4$ cells .....	229
Figure 5.18: Testing the modified ChIP-chip method to detect regions associated with H3K27me3 .....	233
Figure 5.19: Testing the modified ChIP-chip method to detect regions associated with H3K36me3 .....	235
Figure 5.20: Testing the modified ChIP-chip method to detect regions associated with H3K79me3 .....	237
Figure 5.21: Detection of GATA-1 enrichments in reduced numbers of K562 cells.....	239
Figure 6.1: The relationship between SSEA3+ and SSEA3- H9 hESCs and CD14+ monocyte cells.....	248
Figure 6.2: Chromatin modification patterns of promoters in human embryonic stem cells and monocyte cells. ....	251
Figure 6.3: Distinct histone modification profiles at high and low CpG content promoters .....	253
Figure 6.4: Promoter chromatin state and developmental potential.....	254
Figure 6.5: Conserved and cell-type specific bivalent promoters.....	258
Figure 6.6: Representative examples of genes associated with a bivalent promoter structure and SUZ12 binding in human embryonic stem cells .....	262
Figure 6.7: H3K36me3 annotates transcribed genes in hESCs and monocytes .....	264
Figure 6.8: Histone H3K36me3 is associated with non-coding RNAs and putative novel transcripts .....	267
Figure 6.9: H3K79me3 association with the immediate transcribed region of active genes in hESCs and monocytes .....	269
Figure 6.10: Histone acetylation patterns at active and inactive genes .....	271
Figure 6.11: Histone H3K4 methylation patterns at active and inactive genes .....	273
Figure 6.12: Histone acetylation and methylation patterns at distal enhancer/repressor elements .....	274
Figure 6.13: Histone H3K9 methylation patterns at active and inactive genes .....	275
Figure 6.14: Histone H3K27 methylation states at active and inactive genes .....	277
Figure 6.15: Histone H3K36 methylation patterns at active and inactive genes .....	278
Figure 6.16: Histone H3K79 methylation patterns at active and inactive genes .....	280
Figure 6.17: A consensus histone code for regulatory and transcribed regions in human monocytes .....	285



## List of Tables

Table 1.1: The structure of metazoan core promoters .....	7
Table 1.2: Histone variants and their functions .....	25
Table 2.1: Summary of chromatin, antibody and buffer amounts used in reduced cell number ChIP experiments .....	74
Table 2.2: Wash steps for hybridisations performed on the Tecan .....	82
Table 3.1: Description of the ENCODE regions .....	95
Table 3.2: Descriptive statistics for Chipotle sites for K562 ChIP-chip data .....	110
Table 4.1: Gene clusters flanked by CTCF binding sites. 10 gene clusters (containing five or more genes) were flanked by CTCF sites .....	151
Table 4.2: ENCODE Genes containing multiple CTCF binding sites in K562 are membrane components .....	153
Table 4.3: Correspondence between ChIP-chip defined and predicted CTCF sites.....	154
Table 4.4: Overlapping combinations of CTCF, mSin3a, USF1, and USF2 binding sites in K562 .....	162
Table 4.5: Known motifs associated with CTCF binding sites .....	168
Table 4.6: De novo CTCF motif discovery .....	169
Table 4.7: Discovery of novel motifs associated with mSin3a interactions .....	170
Table 4.8: Searching for known motifs in USF1 and USF2 binding sites.....	171
Table 4.9: CTCF binding sites in K562 located at the boundary between active and inactive genomic regions defined by enriched levels of H3K27me1 and H3K27me3 respectively .....	180
Table 5.1: A summary of the performance of H3 K9/K14 diacetylation ChIP-chip assays performed with a standard range of cell numbers and 10 µg of antibody .....	203
Table 5.2: Statistical assessment of assay performance for the six antibody concentrations used in ChIP-chip assays performed with 10 <sup>7</sup> , 10 <sup>6</sup> , 10 <sup>5</sup> , and 10 <sup>4</sup> cells and H3 K9/K14 diacetylation antibody .....	210
Table 5.3: Statistical assessment of assay performance for the six antibody concentrations used in ChIP-chip assays performed with 10 <sup>7</sup> , 10 <sup>6</sup> , 10 <sup>5</sup> , and 10 <sup>4</sup> cells and H3K4me3 antibody.....	217
Table 5.4: Comparisons of 10 <sup>5</sup> and 10 <sup>4</sup> ChIP-chip data before and after rabbit IgG normalisation .....	230
Table 5.5: Assessing the impact of normalisation on the average titration series signal, average series standard deviation and average series signal:noise values for experiments performed with 10 <sup>5</sup> and 10 <sup>4</sup> cells.....	231
Table 6.1: Gene ontology analysis of genes associated with bivalent promoters.....	257
Table 6.2: SUZ12 is associated with bivalent genes in human embryonic stem cells .....	259
Table 6.3: Non-coding RNAs in the ENCODE regions are associated with H3K36me3 .....	265

## Chapter 1

### Introduction

A genome consists of all the genetic material contained in a cell of an organism and contains all of the information necessary for life. In general, this inherited information is encoded in three types of elements - genes, regulatory elements and maintenance elements. Genes contain information to code for proteins while regulatory elements control the spatial and temporal production of proteins by genes. These regulatory elements include promoters, enhancers, insulators and other elements such as non-coding regulatory RNAs. In addition, a further level of gene regulation is achieved by methylation of DNA and modification of chromatin. Maintenance elements contain information for DNA repair, replication and recombination. These elements include centromeres, telomeres, origins of replication and recombination hotspots. The sequencing of the human genome was completed in 2004 (IHGSC, 2004) and the challenge in the post-genome era lies in understanding how the information encoded in the genome sequence regulates both normal development and disease states. Identifying and annotating the human genome for all of the elements mentioned above and characterising the complex events that are associated with these elements will give us an unprecedented understanding of human biology.

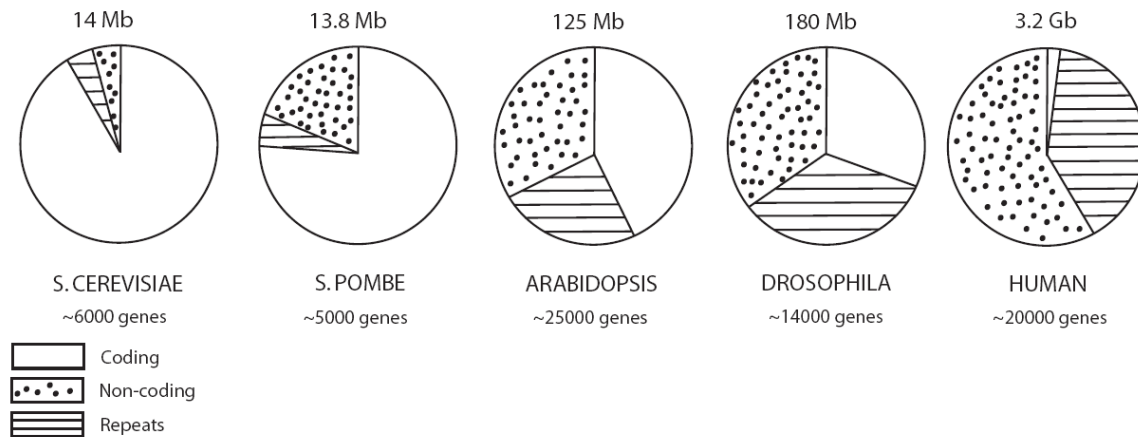
### 1.1. Coding regions in the human genome

In essence, a gene is a genomic sequence directly encoding functional product molecules, either RNA or protein (Gerstein *et al.*, 2007). In eukaryotes, genes are typically composed of alternating exon and intron sequences. One of the important goals in the post-genome era is to produce a definitive catalogue of genes in the human genome and many computational methods such as GeneWise (Birney and Durbin, 2000), GenomeScan (Yeh *et al.*, 2001), GENSCAN (Burge and Karlin, 1997), and TWINSKAN (Flicek *et al.*, 2003) have been developed to identify genes by searching for features such as splice site signals and comparing protein-coding regions in different organisms. A small percentage of the human genome sequence is currently known to code for functional products, mainly protein-coding genes (~2.2%) (Frith *et al.*, 2005) and a

limited number of structural and regulatory RNAs, such as snoRNAs and microRNAs (Mattick, 2007). The current estimate of protein-coding genes is ~20,000 (IHGSC, 2004; Goodstadt and Ponting 2006), but it is possible that the human genome contains a higher number of genes as microarray- based approaches have revealed the existence of many more transcribed sequences of unknown function (Cawley *et al.*, 2004; Rinn *et al.*, 2003; Kapranov *et al.*, 2007; Birney *et al.*, 2007).

## **1.2. Non-coding elements in the human genome**

The current estimate of protein-coding genes in the human genome is surprisingly similar to the number of genes present in the nematode worm (~19,000) (Stein *et al.*, 2003), despite the large difference in developmental complexity and genome sizes. Furthermore, a comparison of genome organisation between the major model organisms showed that there was more than a 300 fold increase in genome size between the yeast *S. cerevisiae* and *S. pombe* and humans but only a 4-fold increase in gene number (Figure 1.1). This large increase in genome size and developmental potential is associated with a dramatic increase in non-coding and repetitive sequences (Taft *et al.*, 2007). The genomes of unicellular yeast contain little non-coding DNA compared with the genomes of multi-cellular eukaryotes. The human genome, in particular, contains large amounts of repetitive and non-coding DNA as protein-coding sequences account for only around 2% of the genome sequence. Understanding the function of the remaining 98% of the genome sequence is an important goal. Comparative analysis of the human and mouse genomes established that approximately 5% of the genomic sequences are highly conserved regions of 50-100 base pairs (bp), which is much higher than can be accounted for by protein-coding sequences alone (Waterson *et al.*, 2002). There are also shorter and weaker homologous elements found in the two genomes, some of which contain binding sites for known transcription factors and regulatory proteins, while others have as of yet no known function (Kondrashov and Shabalina, 2002). In addition, there are species-specific functional sequence elements which are not conserved. All of this, taken together, suggests that the percentage of the human genome which contains functional non-coding elements may be even higher than 5%.



**Figure 1.1: Genome organisation of the major model organisms.** The genome sizes of the major eukaryotic model organisms are indicated above each pie-chart and the approximate number of protein-coding genes is indicated below each pie-chart. The emergence of complex multi-cellular organisms is associated with an increase in genome size. This increase in genome size is due to a large expansion in the amount of non-coding (intronic and intergenic sequences) and repeat DNA (satellite, LINE and SINE elements) sequences.

Non-protein coding sequences include regulatory and maintenance elements. A detailed discussion of regulatory elements controlling gene expression will be presented in later sections of this chapter. Maintenance elements mediate genome structure and dynamics and determine how the genome is passed on to the next generation via repair, replication and recombination. This class of “maintenance” elements includes centromeres and telomeres, origins of DNA replication, and recombination hot spots. Centromeres and telomeres are specialised structures which are involved in replication and are well characterised in terms of sequence content (Eichler and Sankoff, 2003). It is believed that the order of replication is regulated as some regions of the genome replicate earlier than others (Schubeler *et al.*, 2002; Woodfine *et al.*, 2004) and while the process of DNA replication is well documented, replication origins are not understood at the sequence level (Gilbert *et al.*, 2004). A small number of recombination hotspots have been precisely located and as of yet there is no confirmation of sequence similarity (Jeffreys *et al.*, 2001; Yauk *et al.*, 2003). Thus, in order to fully understand key biological processes

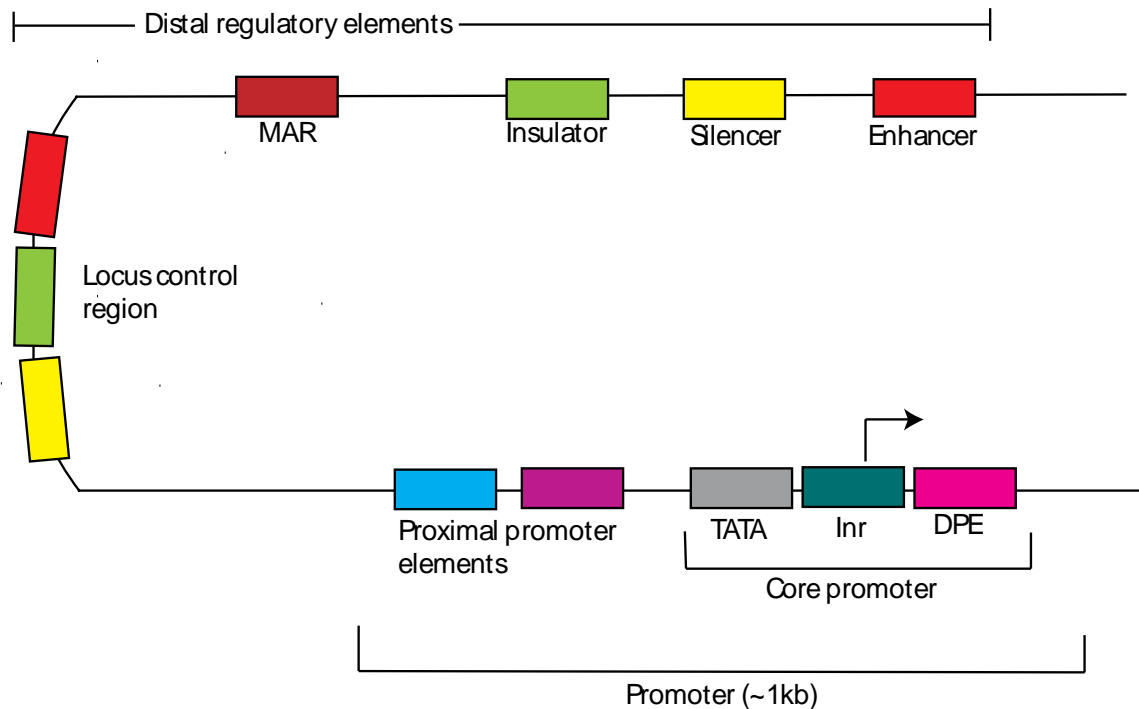
as well as developmental complexity, it is crucial to annotate and characterise the coding and non-coding regions of the human genome.

### **1.3. Regulation of gene expression**

Multi-cellular organisms such as humans are composed of a multitude of different cell types which all develop from the same genetic template. The development of these numerous cell types is dependent upon the appropriate gene expression patterns being established within individual cells in response to external and internal cues. Some genes are expressed constantly in all cell types, while other genes are only expressed as a cell enters a particular developmental pathway. In order to understand why a particular gene is expressed or not expressed, one must be aware of what features are involved in regulating its activity. Each gene has its own associated *cis*-acting elements, which are sequences that regulate gene expression levels. *Trans*-acting factors, which include transcription factors and other proteins involved in regulating gene expression, are encoded by other genes and bind to *cis*-acting elements to control gene expression. This system allows for the cell to use multiple combinations of regulatory features in order to control gene expression. For example, a *trans*-acting factor can act on *cis*-elements of multiple genes or it can form a complex with another *trans*-acting factor to act on an individual gene or multiple genes.

### **1.4. Types of non-coding regulatory elements involved in controlling gene expression**

Protein-coding genes which are transcribed by eukaryotic RNA polymerase II (Pol II) contain two types of *cis*-acting non-coding regulatory sequences which regulate transcription initiation: (1) a promoter composed of a core promoter and proximal regulatory elements, and (2) distal regulatory sequences, which include enhancers, insulators, silencers/repressors, locus control regions and matrix attachment regions (MARs) (Figure 1.2). These regulatory elements act in a co-ordinated manner and can be located over distances which range from several tens to several hundreds of kilobases. These elements are described below.



**Figure 1.2: The typical regulatory features involved in the regulation of gene expression.** The promoter is composed of a core promoter which generally contains a TATA box, initiator element (INR) and a downstream promoter element (DPE). Proximal promoter elements are usually located nearby and typically a promoter element spans less than 1 kb. A gene is also associated with several distal elements such as enhancers, silencers/repressors, insulators, and locus control regions (which are composed of several types of regulatory elements). Each element may contain several binding sites for different sequence-specific transcription factors, allowing for combinatorial control of regulation, which increases the number of possible expression patterns. Enhancers, silencers and insulators controlling the expression of a single gene can be located over several tens to several hundreds of kilobases in the human genome, making the identification of all elements controlling the expression of a particular gene a difficult task. DNA methylation at promoter regions can inhibit transcription, histone modifications at promoters, enhancers and within the gene itself can be involved in gene activation and repression. MARs anchor chromatin loops to the nuclear matrix and can shield genes from position effects. Figure adapted from Maston *et al.*, 2006.

#### 1.4.1. Core and Proximal Promoters

The core RNA polymerase II promoter is a region surrounding a transcription start site (TSS) and is associated with a set of DNA sequence elements (Table 1.1). Combinations of core promoter elements play a crucial role in regulating gene expression patterns (Ren

and Maniatis, 1998). The TATA box is an A/T-rich sequence located approximately 25 bps upstream of the TSS. TATA-binding protein (TBP) recognizes this sequence and begins pre-initiation complex formation. The initiator (Inr) is a pyrimidine rich sequence, which can direct transcription alone or in combination with a TATA box. The downstream promoter element (DPE) lies 28-34 bp downstream of the TSS in TATA-less promoters - it is believed to have a similar function to the TATA box in directing the pre-initiation complex to the TSS (Kadonaga, 2002). Motif ten elements (MTE) function with the Inr to enhance Pol II transcription. The downstream core element (DCE) contains three sub-elements: SI, SII, and SIII. The TF<sub>II</sub>B recognition element (BRE) can be found upstream (BREu) or downstream (BREd) of the TATA box and can either decrease or increase transcription (Sandelin *et al.*, 2007).

Analysis of the eukaryotic promoter database (EPD) and the database of human transcriptional start sites (DBTSS) found that 22% of human genes contain a TATA box. Of these TATA-containing promoters, 62% contained an Inr, 24% contain a DPE and 12% have a BREu (Gershenzon and Ioshikhes, 2005). Seventy-eight percent (78%) of human promoters contain no TATA box, 45% of which contain an Inr, 28% have a BREu, and 25% possess a DPE. The promoters of housekeeping genes, growth factors and transcription factors often have no TATA box (Zhou and Chiang, 2001). The proximal promoter region is located approximately 200 bp upstream of the core promoter and contains several binding sites for activators. Approximately 60% of human promoters are found near a CpG island (Venter *et al.*, 2001), which are formally defined as genomic regions of at least 200 bp with a GC percentage greater than 50% and with an observed/expected CpG ratio greater than 60% (Gardiner-Garden and Frommer, 1987). Methylation of cytosine bases in these islands inhibits transcription (Jones *et al.*, 1998). It has been suggested that proximal promoter elements may block the local promoter region from being inappropriately methylated (Maston *et al.*, 2006).

Core promoter element	Position	Consensus sequence (5'-3')	Interacting Protein
BREu	-38 to -32	(G/C)(G/C)(G/A)CGCC	TF <sub>II</sub> B
TATA	-31 to -24	TATA(A/T)A(A/T)(A/G)	TBP
BREd	-23 to -17	(G/A)T(T/G/A)(T/G)(G/T)(T/G)(T/G)	TF <sub>II</sub> B
Inr	-2 to +5	PyPyAN(T/A)PyPy	TAF1/TAF2
MTE	+18 to +29	C(G/C)A(A/G)C(G/C)(G/C)AACG(G/C)	n/a
DPE	+28 to +34	(A/G)G(A/T)CGTG	TAF6/TAF9
DCE	three sub-elements: +6 to +11 +16 to +21 +30 to +34	Core sequence: S <sub>I</sub> CTTC S <sub>II</sub> CTGT S <sub>III</sub> AGC	TAF1

**Table 1.1: The structure of metazoan core promoters.** Core promoters are composed of a number of distinct elements, which include and upstream and downstream TF<sub>II</sub>B-recognition element (BREu and BREd respectively), a TATA box, an Initiator element (Inr), a motif ten element (MTE), a downstream promoter element (DPE), and a downstream core element (DCE). The DCE does not occur with an MTE and DPE. With the exception of BRE motifs, all other core promoter elements are recognized by TFIID complex members. Table adapted from Thomas and Chiang, 2006.

#### 1.4.2. Enhancers

An enhancer element is defined as a DNA sequence which functions in an orientation and distance independent manner to enhance expression of a gene. The first enhancer element, which could dramatically increase gene transcription from a human  $\beta$ -globin gene, was a genomic region of the SV40 virus (Banerji *et al.*, 1981). The first human enhancer to be identified was located downstream of the immunoglobulin heavy chain locus and was found to have cell-type specific enhancer activity (Banerji *et al.*, 1983). Enhancers may facilitate transcription within a specific tissue or cell type, through the recruitment of tissue-specific activators, which in turn recruit general transcription factors (Szutorisz *et al.*, 2005). General transcription factor complexes aid in the binding and function of Pol II at core promoters (see section 1.5.1). Enhancers and proximal promoter elements are very similar in that they both bind activators to enhance transcription but



enhancers can be located several hundred kilobases upstream, downstream or even within an intron of their target gene(s). Two models have been proposed to explain how enhancers communicate with promoters to achieve the desired level of gene expression:

i) **Looping/ direct contact model.** This model proposes that enhancers and promoters directly interact. For example, Johnson *et al.* suggest that in the case of the mouse  $\beta$ -globin locus, the tissue specific activators NF-E2 and GATA-1 function in the transfer of Pol II from the enhancer to the  $\beta$ -globin promoter during blood development (Johnson *et al.*, 2001; 2002). This transfer occurs by direct physical contact between the enhancer and promoter and occurs over a 50 kb region. Therefore the intervening region of chromatin must be looped out to allow enhancers to directly interact with the promoter. This model is supported by studies carried out at the human and mouse  $\beta$ -globin loci in which promoters and distal enhancers were shown to co-localise within chromatin hubs which also contain RNA polymerase II (Patrinos *et al.*, 2004; Osborne *et al.*, 2004).

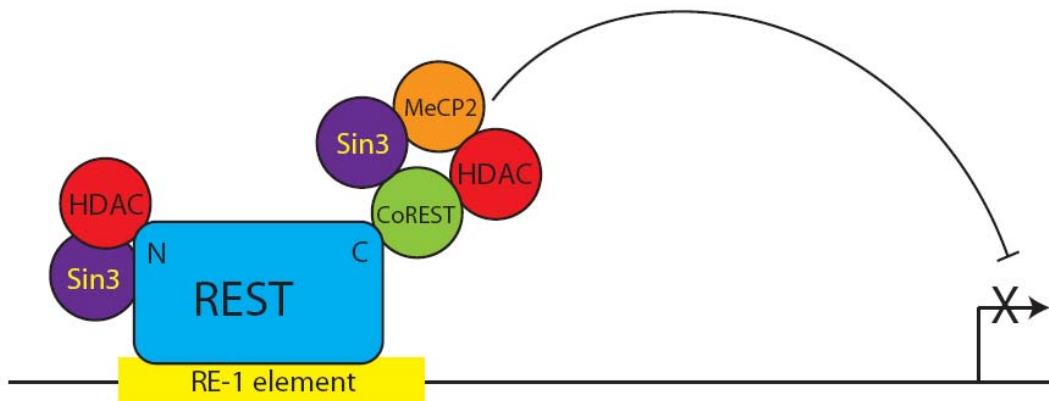
ii) **Tracking/non-contact model.** This model proposes that enhancers act as entry sites for factors that ultimately interact with the promoter (Bulger and Groudine, 1999). The transfer of general transcription factors and Pol II to the promoter occurs by a continuous linear tracking of the complex from the enhancer to the core promoter along the length of the intervening DNA sequence (Kim and Dean, 2004; Ling *et al.*, 2004).

### 1.4.3. Silencers/repressors

The opposite of an enhancer is a silencer/ repressor element, which is involved in repressing transcription. In common with enhancers, the majority of identified silencers can function independently of direction and distance. Silencer elements can be located at the proximal promoter of a target gene, within an intron, 3' untranslated region, or as part of a distal enhancer (Ogbourne and Antalis, 1998).

Silencers function as binding sites for sequence-specific transcription factors known as repressors. For example, a 21-bp DNA repressor/silencer element known as the repressor element 1 (RE-1) is found at approximately two thousand sites in the human genome (Bruce *et al.*, 2004). The repressor element-1 silencing transcription factor (REST) binds to RE-1 sites and acts as a transcriptional repressor by blocking the expression of many neuronal RE-1 containing genes in non-neuronal cells. REST recruits co-repressors via its

repression domains at the amino and carboxy termini (Figure 1.3). These co-factors then facilitate the creation of a repressive chromatin state through histone deacetylation (Huang *et al.*, 1999), chromatin remodelling (Battaglioli *et al.*, 2002) and DNA methylation (Lunyak *et al.*, 2002).



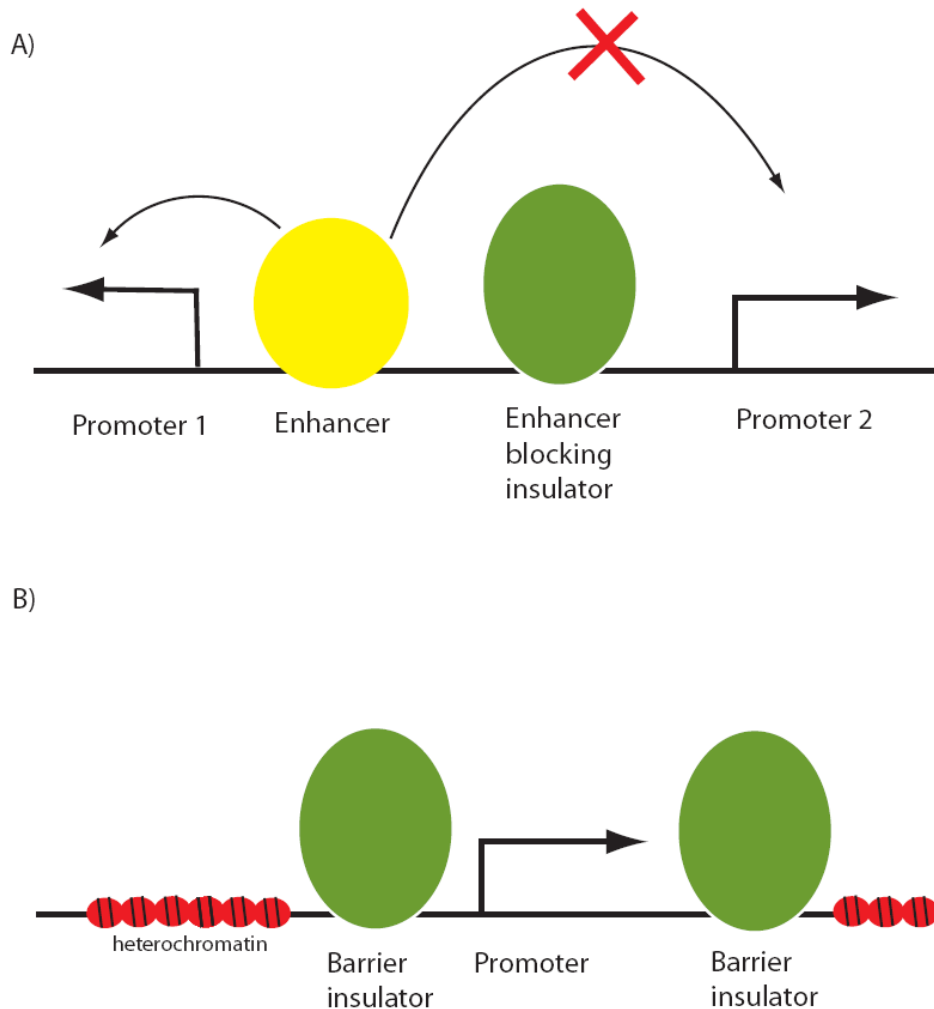
**Figure 1.3: REST-mediated repression of target genes.** The amino terminus (N) of REST recruits a Sin3-histone deacetylase (HDAC) complex to repress neuronal genes in non-neuronal cells. The carboxy terminus (C) recruits many complexes via the co-repressor CoREST, which can include Sin3-HDAC, the methyl-DNA binding protein MeCP2, SWI/SNF chromatin remodelling complex (not shown), histone H3K9 methyltransferase (not shown) and histone demethylase LSD1 (not shown) to mediate epigenetic silencing.

Other repressors work by different mechanisms, some function by blocking the binding of a nearby activator (Harris *et al.*, 2005), by directly competing for the same binding site (Li *et al.*, 2004) or by inhibiting Pol II pre-initiation complex formation (Kim *et al.*, 2003).

#### 1.4.4. Insulators

There are many examples of regions in the human genome (and other eukaryotic genomes) which contain a cluster of genes that are actively expressed in a particular cell type and in close proximity to another group of genes which are not expressed in that cell type (Sproul *et al.*, 2005). Alternatively, an expressed gene may be located in a region of constitutively silent chromatin. In both situations, the maintenance of appropriate

expression patterns relies upon a class of DNA sequence elements known as insulators (Gaszner and Felsenfeld, 2006). Insulators that prevent inappropriate promoter activation by enhancers have been termed enhancer blocking insulators while those which prevent the spread of silent chromatin are known as barrier insulators (Sun and Elgin, 1999) (Figure 1.4).



**Figure 1.4: The two types of insulator activity.** A) An enhancer-blocking insulator interferes with enhancer-promoter communication only when positioned between an enhancer and a promoter as is the case for promoter 2. In other situations (i.e. promoter 1), enhancer-promoter communication is not blocked. B) A barrier insulator prevents the spread of heterochromatin into a euchromatin region when placed at a junction between the two regions.

#### 1.4.4.1 Enhancer blocking insulators

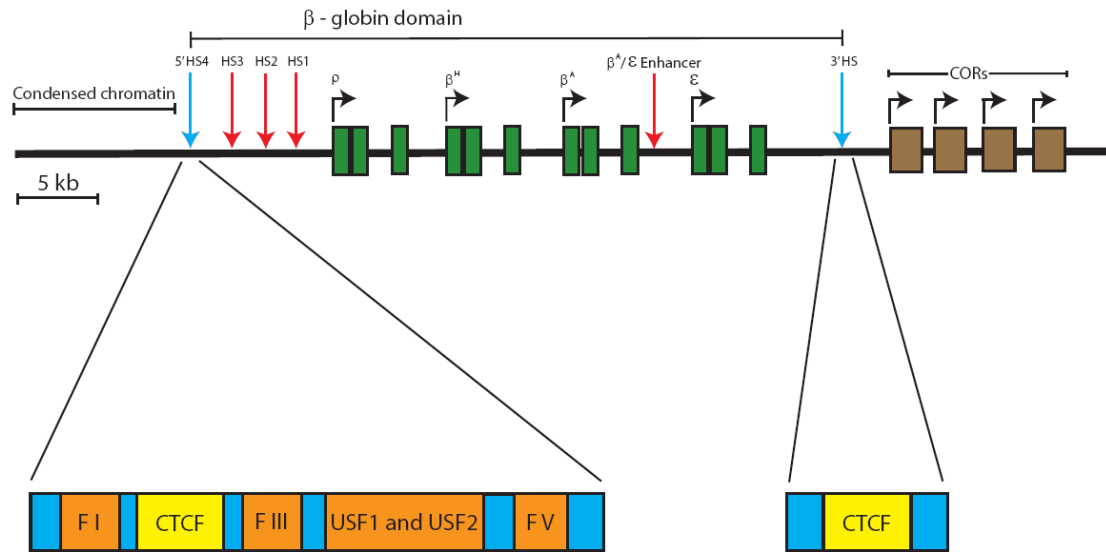
Enhancer blocking elements interfere with enhancer-promoter interactions only when located between these two elements. They function to prevent an enhancer from incorrectly activating a promoter. It has been suggested that insulators perform this function by tethering the chromatin fibre to the nuclear matrix. This may result in the formation of chromatin loops which prevent an enhancer and a promoter from communicating with each other due to their location in separate loop domains (Gaszner and Felsenfeld, 2006).

The initial work in identifying enhancer blocking elements was performed in *Drosophila* with the gypsy retro-transposon (Geyer *et al.*, 1986). Insertion of this element within enhancers at the yellow locus blocked the action of those enhancers located distal to the insertion site but did not affect enhancers located proximal to the promoter (Geyer *et al.*, 1986). The enhancer blocking activity of gypsy was mapped to 12 binding sites for Suppressor of Hairy Wing (Su(Hw)), which interacts with Topoisomerase-I-interacting protein (Topors). Topors can then bind the nuclear lamina (Capelson and Corces, 2005), which is consistent with the idea that chromatin loop formation is the important event in mediating enhancer-blocking activity. The first vertebrate enhancer blocking insulator to be identified, 5'HS4, was located at the 5' end of the chicken  $\beta$ -globin locus (Figure 1.5) (Chung *et al.*, 1993; Pikaart *et al.*, 1998). This complex element possessed both enhancer blocking and barrier insulator activity (discussed in section 1.4.4.2). The enhancer blocking activity was attributed to the binding of the 11 zinc finger transcription factor CTCF (Bell *et al.*, 1999). CTCF can interact with itself and the nucleolar protein nucleophosmin, which may lead to the formation of discrete chromatin domains (Yusufzai *et al.*, 2004). Furthermore, CTCF has been shown to mediate long range chromatin looping at the mouse  $\beta$ -globin locus (Splinter *et al.*, 2006) and inter-chromosomal co-localisation between the Igf2/H19 and Wsb1/Nf1 loci (Ling *et al.*, 2006). A more detailed discussion on CTCF is presented in section 1.4.4.3 and Chapter 4 (section 4.1).

#### 1.4.4.2 Barrier Insulators

Barrier insulators guard against position-effect variegation (PEV), which is the silencing of a euchromatic gene due to the spread of heterochromatin. Heterochromatin is more condensed than euchromatin and is associated with methylation of histone H3 lysine 9 and lysine 27 (see section 1.6.2). Heterochromatic DNA is also associated with extensive CpG methylation. Heterochromatin formation begins with the methylation of H3K9 at an initiation site, which then recruits heterochromatin protein 1 (HP1) (Grewal and Jia, 2007). HP1 can in turn recruit histone H3K9 methyltransferase activity and this cycle of events leads to the spread of heterochromatin. Euchromatin represents a less condensed form of chromatin, which is located at transcriptionally active regions of the genome and is associated with a number of histone modifications such as histone H3 acetylation and H3 K4 methylation (see section 1.6.2).

As discussed in the previous section, the chicken 5'HS4 element also displays barrier insulator activity (Pickaart *et al.*, 1998), which was found to be independent of CTCF binding (Recillas-Targa *et al.*, 2002). The 5' HS4 is located between a region of condensed inactive chromatin in chicken erythrocytes and the active  $\beta$ -globin chromatin domain (Figure 1.5). The 5' HS4 is marked by peaks of euchromatic histone modifications (Litt *et al.*, 2001). These are due to the recruitment of histone acetyltransferases (HAT) and histone methyltransferases (HMT) by upstream stimulatory factor 1 (USF1) and USF2 (West *et al.*, 2004; Huang *et al.*, 2007). Disruption of USF1 and USF2 binding abolishes HAT and HMT recruitment along with barrier activity. CTCF binding sites have also been found close to the transition between active and silent chromatin domains in mouse and human cell types (Filippova *et al.*, 2005; Barski *et al.*, 2007). Although it has not been directly shown that CTCF prevents the spread of heterochromatin, it cannot be ruled out that CTCF may also have barrier activity. This hypothesis would reconcile with the chromatin loop model of enhancer blocking discussed above as flanking a gene with a CTCF binding site would provide barrier activity by creating an independent expression domain. Support for barrier insulator's functioning through the formation of chromatin loops has been provided by a study of barriers elements in yeast (Ishii *et al.*, 2002). This study showed that barrier activity was linked to the anchoring of chromatin fibres to the nuclear pore.



**Figure 1.5: Genomic organisation of the chicken  $\beta$ -globin locus.** The 5'HS4 and 3'HS insulator elements (indicated by blue arrows) define the chicken  $\beta$ -globin chromatin domain, which contains the developmentally regulated  $\beta$ -globin gene cluster and its locus control region (LCR), composed of HS1-3 and the  $\beta^A/\epsilon$  enhancers. The  $\beta$ -globin domain is flanked by a region of condensed chromatin at the 5' end and chicken olfactory receptor genes (CORs) at the 3' end. 5' HS4 displays both enhancer-blocking activity (mediated by CTCF) and barrier activity (mediated by USF1, USF2 and as yet uncharacterised footprint I, III and V binding proteins). 3'HS binds CTCF and displays only enhancer-blocking activity. Figure adapted from Gaszner and Felsenfeld, 2006.

#### 1.4.4.3 CTCF is a multifunctional protein

Following the discovery that CTCF was responsible for the enhancer-blocking activity of the chicken  $\beta$ -globin locus insulators, there has been a great deal of interest in identifying other CTCF binding sites and understanding how this factor functions. CTCF enhancer blocking activity was subsequently identified in the human and mouse *Igf2/H19* imprinted locus (Bell and Felsenfeld, 2000; Hark *et al.*, 2000). During embryonic development *Igf2* is expressed from the paternal allele only and is under the control of the imprinted control region (ICR) lying between the enhancer and *Igf2* promoter. The ICR contains CTCF insulator elements, the CpG islands of which are methylated in the paternal allele. Methylation of the CpG islands abolishes CTCF binding and therefore the enhancer blocking activity of the ICR.

In addition to its role as an insulator binding protein, CTCF has been shown to function as a classical transcription factor, capable of functioning as a transcriptional repressor (Filippova *et al.*, 1996; Burcin *et al.*, 1997) and a transcriptional activator (Vostrov and Quitschke, 1997). The diversity of functions that CTCF performs may be attributed to its protein structure. CTCF contains 11 zinc finger domains and different combinations of zinc fingers can engage different binding sequences (Filippova *et al.*, 1996). Two recent genome-wide studies have employed different approaches to identify CTCF binding sites in the human genome. Ren and colleagues used chromatin immunoprecipitation in combination with microarrays (section 1.7.3.5) to identify CTCF binding sites in human fibroblasts (Kim *et al.*, 2007), while Lander and colleagues used a computational approach to identify conserved motifs and one of these motifs was able to bind CTCF (Xie *et al.*, 2007). These studies identified approximately 15,000 CTCF binding sites in the human genome. The experimental identification of a large number of CTCF binding sites by Ren and colleagues allowed for a consensus CTCF binding sequence to be determined. A 20 bp consensus motif was described which refined a previously described motif (Bell and Felsenfeld, 2000) but a number of nucleotides at several sites in the consensus sequence were ambiguous suggesting that variation within the consensus sequence is responsible for the functional versatility of CTCF.

A number of factors have been shown to interact with CTCF (Wallace and Felsenfeld, 2007) and the variety of CTCF functions is likely to reflect the diversity of these interacting partners. CTCF interacts with DNA binding proteins, histone interacting proteins, histone themselves and other regulatory factors. For example the chromodomain helicase family member CHD8 interacts with CTCF and has been implicated in insulator function (Ishihara *et al.*, 2006). CHD8 forms complexes with histone modifying enzymes (Dou *et al.*, 2005) and may recruit these enzymes to CTCF bound insulators.

CTCF function may be regulated through the choice of zinc fingers used in DNA binding. Furthermore, CTCF can be chemically modified which also affects its functional properties. It can be poly (ADP-ribosyl)ated, and inhibiting this modification can impair its insulator function (Yu *et al.*, 2004). Poly (ADP-ribosyl)ation has been implicated in maintaining DNA hypomethylation (Zardo and Caiafa, 1998) in the genome and CTCF binding is sensitive to DNA methylation.

#### **1.4.5. Locus control regions**

Locus control regions (LCR) consist of a cluster of regulatory element involved in regulating an entire locus (Li *et al.*, 2002). The LCR is usually composed of several *cis*-acting elements such as enhancers, insulators, silencers and nuclear matrix attachment regions (MARs) (see section 1.4.6). These elements are bound by various transcription factors and chromatin modifiers, which regulate spatial and temporal gene expression patterns. LCRs can regulate transcription from large distances in a position independent way and can be found upstream of their target locus, within an intron of a target gene (Aronow *et al.*, 1992) or neighbouring gene (Adlam and Siu, 2003) or downstream of their regulated locus (Lang *et al.*, 1991).

The first LCRs to be identified were the human and mouse  $\beta$ -globin LCR (reviewed by Chakalova *et al.*, 2005). The human  $\beta$ -globin LCR is located ~6-25kb upstream of the locus and controls five genes that are expressed during different stages of erythrocyte development. The orientation of the LCR is critical as reversing its direction destroys most of its activity (Tanimoto *et al.*, 1999). LCRs are believed to function in a similar way to the enhancer looping model discussed above as long range interactions between DNase I hypersensitive sites have been observed at the active  $\beta$ -globin locus (Tolhuis *et al.*, 2002), which tethers these sequences into an active chromatin hub. Critically, these long range interactions are only observed when the locus is active.

#### **1.4.6. Matrix Attachment regions (MARs)**

The nuclear matrix refers to a structure obtained from nuclei that is resistant to extraction by high concentrations of NaCl or the non-ionic detergent-like salt lithium diiodosalicylate (LIS) (Mirkovitch *et al.*, 1984). The DNA fragments bound to this residual structure are known as MARs and serve to anchor chromatin loops to the nuclear matrix. While it is thought that there is a structural role for MARs (i.e. they mediate binding of chromatin to the nuclear matrix), biological functions have been proposed for MARs such as transcription activation as a result of chromatin remodelling (Bode *et al.*, 2000) and insulation of genes from position-mediated silencing of transgenes (Allen *et al.*, 2000). MARs have also been shown to antagonise DNA methylation-dependent repression of long range enhancer-promoter interactions (Forrester *et al.*, 1999), and to



extend an enhancer mediated region of open chromatin (Jenuwein *et al.*, 1997). CTCF has been shown to interact with the nuclear matrix (Dunn *et al.*, 2003), suggesting that CTCF may be responsible for anchoring chromatin to the nuclear matrix. Recillas-Targa and colleagues have shown that an insulator 5' upstream of the chicken  $\alpha$ -globin gene domain co-localises with a MAR element, which binds CTCF (Valadez-Graham *et al.*, 2004). The insulator activity of this element is dependent on CTCF binding, linking MARs and CTCF once more with a role in demarcating chromosomal boundaries.

## **1.5. Proteins involved in transcriptional regulation**

The proteins which bind to these regulatory elements discussed above are numerous. There may be as many as 2000 transcription factors in the human genome (Lander *et al.*, 2001) which are divided into three classes of proteins, namely, general members of the RNA polymerase complexes, sequence-specific DNA-binding proteins that mediate activation or repression of transcription and chromatin remodelling and modification complexes.

### **1.5.1. RNA polymerase complexes**

At the promoter, interactions between RNA polymerase II and DNA lead to transcription initiation. Several general transcription factors (GTFs) are necessary for recognition and stable binding at the promoter (Thomas and Chiang, 2006). These GTFs were designated as TF<sub>II</sub> A-H (TF for transcription factor, II represents Pol II transcription and the letter refers to the nuclear extract from which the GTF was isolated). These GTFs function collectively to specify transcription start sites (TSSs). Pre-initiation complex formation begins with the binding of TF<sub>II</sub>D to a core promoter element such as a TATA box or Inr element followed by stepwise recruitment of the other GTFs or alternatively by recruitment of a pre-assembled Pol II holoenzyme.

TF<sub>II</sub>D is a complex of TBP and at least 14 TBP-associated factors (TAFs). TBP and some TAFs bind different core promoter elements, allowing TF<sub>II</sub>D to recognize TATA containing and TATA-less promoters. The TF<sub>II</sub>D complex is also associated with enzymes that post-translationally modify chromatin and other proteins involved in gene regulation. TAF1 in the TF<sub>II</sub>D complex is a histone acetyltransferase (HAT), which

methylates histones H3 and H4 (Mizzen *et al.*, 1996). TF<sub>II</sub>A stabilizes TBP-TATA box interactions through direct contact with TBP (Geiger *et al.*, 1996). The binding of TF<sub>II</sub>B to TBP stabilizes TF<sub>II</sub>D /TBP binding to the promoter (Orphanides *et al.*, 1996). TF<sub>II</sub>B also recruits Pol II/TF<sub>II</sub>F to the promoter bound TF<sub>II</sub>D-TF<sub>II</sub>B complex. TF<sub>II</sub>F enhances the affinity of Pol II for the promoter complex and is necessary for the recruitment of TF<sub>II</sub>E and TF<sub>II</sub>H (Orphanides *et al.*, 1996). TF<sub>II</sub>F also facilitates Pol II promoter escape (Yan *et al.*, 1999). TF<sub>II</sub>E stimulates the ATPase, C-terminal domain (CTD) kinase, and DNA helicase activities of TF<sub>II</sub>H (Lee and Young, 2000), which are essential for transcription initiation and elongation.

Pol II itself is composed of 12 subunits (RPB1-12) (Young, 1991). RPB1 and RPB2 are the key catalytic units and are responsible for phosphodiester bond formation (Hampsey, 1998). The CTD of RPB1 contains tandem heptapeptide repeats of Tyr-Ser-Pro-Thr-Ser-Pro-Ser (YSPTSPS) (Prelich, 2002). The CTD can be phosphorylated at Ser2 and Ser5 and hyperphosphorylated Pol II is associated with transcription elongation (Sims *et al.*, 2004). Ser5 phosphorylation levels peak early during transcript production, whereas Ser2 predominates in the middle and later stages of transcript production (Komarnitsky *et al.*, 2000).

### **1.5.2. Sequence-specific transcription factors**

This class includes the largest and most diverse group of factors responsible for transcription initiation, enhancement and inhibition. As discussed before, as many as 10% of the genes in the human genome - approximately 2000 genes - may encode transcription factors (TFs). Specificity of transcription is achieved by combinatorial binding of these factors to sequence-specific *cis*-regulatory elements. TFs alter transcription by interacting directly or indirectly with RNA polymerase as described in the previous section. Therefore, TFs must possess certain structural features which allow them to modulate gene expression. Currently, more than 100 different DNA-binding domains have been identified in transcription factors (Kummerfeld and Teichmann, 2006), some of which are discussed below.

The homeobox or homeodomain was originally identified in the homeotic genes of *Drosophila* (Affolter *et al.*, 1994) and found to contain a DNA-binding motif called a

helix-turn-helix motif. The homeobox domain is conserved in mammalian Hox gene clusters. The POU family of TFs (named after the transcription factors Pit, Oct and Unc) was subsequently found to contain both a homeobox sequence and a conserved POU-specific domain. Both domains are needed for high-specificity DNA-binding. The TFs containing these domains play important roles in development.

The leucine zipper motif was identified in several different transcription factors such as C/EBP, yeast GCN4 and the oncogenes Myc, Fos and Jun (Bosher *et al.*, 1996). In this motif every seventh amino acid is a leucine, whereby leucine residues occur every two turns on the same side of a helix, allowing dimerization of TFs. Dimerization also correctly positions the nearby basic DNA-binding domains, which is necessary for DNA binding to occur. The basic DNA-binding domain has also been identified in other transcription factors which do not contain a leucine zipper (Prendergast and Ziff, 1989). These TFs include E12 and E47 factors which are important for the development of B lymphocytes. This class of TFs also contain a different type of helix-turn-helix motif, which functions similarly to a leucine zipper. It facilitates dimerization of TFs, which in turn allows for DNA-binding by the basic motif (Jones, 1990).

In addition to these DNA-binding domains, TFs also possess distinct transcriptional activation domains (Stephanou *et al.*, 2002). Analysis of many different activation domains identified three distinct categories, acidic domains, glutamine-rich domains, and proline-rich domains. Acidic activation domains contain a large number of acidic amino acids which produces a strong negative charge. The negative charge allows long range electrostatic interactions with members of the TF<sub>II</sub>D complex (Uesugi *et al.*, 1997). Glutamine-rich activation domains have been found in TFs such as Sp1, Oct-1, and Oct-2. The activation domain of CTF/N1, which binds to the CCAAT box motif, is proline rich and has been found in other TFs such as Jun and AP2. In summary, many different DNA binding and activation motifs exist and reflect the diversity of DNA sequences which TFs bind and the proteins that they interact with.

### **1.5.3. Proteins involved in chromatin remodelling and modification**

Sequence-specific TFs need to gain access to the DNA template to initiate transcription. However, the DNA template is normally folded into a compact chromatin fibre which must be unfolded or remodelled to allow transcription to occur. (De la Serna *et al.*, 2006). A description of chromatin structure is presented in section 1.6. The two classes of proteins which regulate accessibility of the DNA template – ATP-dependent chromatin remodelling enzymes and histone modifying enzymes - are described below.

#### **1.5.3.1. ATP-dependent chromatin remodelling enzymes**

ATP-dependent chromatin remodelling enzymes use energy from ATP hydrolysis to remodel nucleosomes (Narlikar *et al.*, 2002). Remodelling enzymes disrupt histone-DNA interactions, i.e. they promote nucleosome ‘sliding’, allowing transcription factors to access the DNA (Becker, 2002). They can also reposition the DNA so that it is accessible on the surface of the histone octamer (Aoyagi *et al.*, 2002). In addition to these activities, remodelling complexes can transfer a histone octamer from one DNA template to another (Whitehouse *et al.*, 1999) and can cause changes in super-helicity by twisting the DNA which disrupts histone-DNA interactions (Gavin *et al.*, 2001).

There are three classes of ATP-dependent remodellers, the SWI/SNF (Switch/Sucrose non-fermentable), CHD (chromodomain and helicase-like domain) and ISWI (imitation SWI) families (De la Serna *et al.*, 2006). The SWI/SNF family members contain a bromo domain, which binds acetylated histones (Hassan *et al.*, 2002). The CHD family members have two chromodomains, which bind methylated histone tails (Bannister *et al.*, 2001; Lachner *et al.*, 2001; Flanagan *et al.*, 2005). The ISWI family contains a SANT domain, which acts as a histone-binding domain (Boyer *et al.*, 2004). Each class of enzymes forms complexes with other proteins, for example ISWI SNF2H enzymes are found in the ACF (ATP-utilizing chromatin assembly and remodelling factor) complex as well as the RSF (remodelling and spacing factor) complex. SNF2L is found as part of the NuRF (nucleosome remodelling factor) and CERF (CECR2-containing remodelling factor) complexes.

### 1.5.3.2. Histone modifying enzymes

The other class of proteins involved in chromatin remodelling are the histone modifying enzymes. Unlike the ATP-dependent chromatin remodelling enzymes, which expose the underlying DNA by promoting nucleosome movement, histone modifying enzymes influence transcription by covalently modifying amino acid residues located in the N-terminal 'tail' and the core of histones (Kouzarides, 2007). These enzymes modify specific amino acids by adding or removing various chemical groups. The covalent modifications include acetylation, methylation, phosphorylation, ubiquitination, sumoylation and ADP-ribosylation.

Acetylation of histones was first proposed to be involved in activation of transcription over 40 years ago (Allfrey *et al.*, 1964), but it wasn't until 1995 that the first histone acetyltransferase (HAT) was identified (Kleff *et al.*, 1995). Since then, a large number of HATs have been characterised (Reid *et al.*, 2000). HATs are divided into three main families, GNAT, MYST, and CBP/p300 (Sternier and Berger, 2000). HATs function as part of large complexes *in vivo* and different complexes are involved in distinct biological processes (Roth *et al.*, 2001). These different complexes contain specific non-acetyltransferase components which interact with different sequence specific activators, targeting the complex to distinct genes. Consistently, given that histone acetylation can create a more open chromatin structure, many transcriptional co-activators, such as Gcn5/PCAF, CBP/p300 and SRC-1, have been shown to possess intrinsic HAT activity. Similar to transcriptional co-activators possessing HAT activity, many transcriptional co-repressor complexes, such as mSin3a and NURD/Mi-2, contain subunits with HDAC activity (Shahbazian and Grunstein, 2007). However, the Rpd3 small complex (Rpd3S) is an HDAC-containing complex that associates with the actively transcribing, elongating form of RNA polymerase II. Through this association, Rpd3S has been implicated in preventing inappropriate initiation within the protein-coding region of actively transcribed genes (Keogh *et al.*, 2005).

Histone deacetylase (HDACs) complexes remove acetyl groups (Kurdistani and Grunstein, 2003). Deacetylation correlates with transcriptional repression and there are three catalytic groups of HDACs which are conserved from yeast to human - type I, II, and III (Narlikar *et al.*, 2002). The type I family include HDACs 1, 2, 3, and 8, while type

II includes HDACs 4, 5, 6, 7, 9 and 10. These two types of enzymes have a similar mechanism of deacetylation which does not involve a co-factor. The type III family of Sir2 related enzymes require the co-factor nicotinamide adenine dinucleotide (NAD) as part of their catalytic mechanism.

Histone methyltransferases are responsible for catalysing the methylation of lysine and arginine residues in histones. Methylation modifications can result in either activation or repression of transcription (Bannister and Kouzarides, 2005). All methyltransferases which modify lysine residues, with the exception of Dot1, contain a SET domain (Marmorstein, 2003), which was named after the *Drosophila* chromatin proteins Su(var)3-9, Enhancer of zeste [E(z)], and Trithorax, in which it was first identified. Dot1 is responsible for methylating lysine 79 of histone H3, which is located in the core of the histone (Feng *et al.*, 2002). These enzymes are conserved from yeast to man, for example the first H3K4 methyltransferase to be identified was yeast Set1 (Briggs *et al.*, 2001; Roguev *et al.*, 2001) and human MLL is highly related to yeast Set1 and is also a H3K4 specific methyltransferase (Yokoyama *et al.*, 2004). Arginine methylation is catalysed by the protein arginine methyltransferases family 1 (PRMT1) of proteins (Lee *et al.*, 2005).

The first histone lysine demethylase LSD1/BHC110 was identified and characterised recently (Shi *et al.*, 2004; Lee *et al.*, 2005 b). This enzyme specifically demethylates histone H3K4 and subsequently the jumonji C class of demethylases were identified which could demethylate histone H3 K4, H3 K9, H3 K27, and H3 K36 residues (Liang *et al.*, 2007; Klose *et al.*, 2007; Lee *et al.*, 2007; Yamane *et al.*, 2007; Secombe *et al.*, 2007; Christensen *et al.*, 2007; Cloos *et al.*, 2006; Lan *et al.*, 2007; Tsukada *et al.*, 2006; Whetstine *et al.*, 2006; Yamane *et al.*, 2006; Hong *et al.*, 2007) (section 1.6.2). While no enzymes have been identified which can reverse arginine methylation, the human enzyme peptidylarginine deiminase (PAD4/PADI4) can catalyze the conversion of an arginine residue to citrulline, antagonizing the effect of arginine methylation as citrulline prevents arginine residues from being methylated (Wang *et al.*, 2004; Cuthbert *et al.*, 2004).

RSK2 (Sassone-Corsi *et al.*, 1999) and MSK1 (Thomson *et al.*, 1999) have been identified as the mammalian kinases which perform histone H3 phosphorylation. ADP-ribosylation can be mono, catalyzed by mono-ADP-ribosyltransferases (MARTs), or poly, catalyzed by poly-ADP-ribosyltransferases (PARTs) (Hassa *et al.*, 2006). Histones

are mono-ubiquitynated by ubiquitin-conjugating enzymes such as the Bmi/Ring1A protein (Wang *et al.*, 2004).

## **1.6. Epigenetic regulation of transcription**

Conrad Waddington first coined the term ‘epigenetics’ in 1942 to mean “the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being” (Waddington, 1942). This definition has evolved over the years and now epigenetics is defined as changes to gene function that occur in the absence of changes to the underlying DNA sequence. Modern epigenetic research is focused on the study of covalent and non-covalent modifications of histones and DNA and how these modifications influence overall chromatin structure, gene expression and replication. The four core histone proteins which make up nucleosomes (section 1.6.1) can be modified by more than 100 different post-translational modifications (section 1.6.2). These occur mainly at specific amino acids on the N-terminal tail and recent years have seen a great increase in our understanding of these modifications. Vertebrate DNA methylation occurs almost exclusively at CpG dinucleotides (Bird, 2002) (section 1.6.3) and histone modifying proteins may be involved in directing DNA methylation to promoters (Vire *et al.*, 2006).

### **1.6.1. The Nucleosome – the fundamental repeating unit of chromatin**

A typical mammalian nucleus is 11-22  $\mu\text{m}$  in diameter, into which two meters of DNA is packaged. The DNA is packaged in a highly ordered manner; the first level of compaction is achieved by wrapping DNA around the histone core proteins to produce a structure called the nucleosome, the basic unit of the chromatin fibre (Kornberg, 1974). Interactions between individual nucleosomes drive the folding of a nucleosomal array (11nm in diameter) into a secondary fibre, 30nm in diameter, which are then further condensed into large structures that form chromosomes.

Nucleosomes are arranged like ‘beads on string’ along the chromatin fibre (Kornberg and Thomas, 1974) and a typical nucleosome consists of approximately 200 bp of DNA wrapped around a histone octamer. Each histone octamer is composed of two copies of the core histone proteins H2A, H2B, H3 and H4, which wraps 146 bp of DNA in 1.7

superhelical turns (Luger *et al.*, 1997), while approximately 60 bp DNA forms a linker between adjacent octamers. Nucleosomes are synthesized in an ordered manner. Firstly two heterodimers of H3 and H4 are deposited onto the DNA to form a (H3/H4)<sub>2</sub> tetramer. Then two H2A-H2B heterodimers bind on either side of the tetramer to form the octamer. Histone H1 interacts with the linker DNA and is involved in higher order folding of the chromatin fibre (Khorasanizadeh, 2004). In the nucleosome core there are 14 contact points between the histones and DNA (Luger *et al.*, 1997), making nucleosomes one of the most stable protein-DNA complexes known.

Histones have N-terminal tails, which protrude from the octamer and are subject to numerous post-translational modifications, which have been implicated in a number of processes such as transcriptional activation/silencing, DNA replication and repair, and chromatin assembly (see section 1.6.2). Recent genome-wide studies in yeast (*S. cerevisiae*) have reported that nucleosomes are depleted from active regulatory elements (Lee *et al.*, 2004; Pokholok *et al.*, 2005; Yuan *et al.*, 2005) and the chicken  $\beta$ -globin 5'-HS4 has been shown to be depleted of nucleosomes (Zhao *et al.*, 2006). Histone replacement has been reported to mark the boundaries of *cis*-regulatory domains in *D. melanogaster* (Mito *et al.*, 2007) and low nucleosome density has been observed in the vicinity of transcription start sites in human cells (Nishida *et al.*, 2006; Heintzman *et al.*, 2007). These numerous observations lend further support to the idea that nucleosomes are removed or moved along the chromatin fibre by chromatin remodellers to expose the underlying DNA.

The core histone proteins are expressed during the S phase and are involved in the packaging of newly synthesized DNA and were once believed to be the common components of all nucleosomes (Kornberg and Lorch, 1999). However variant forms of these histones have been identified (Kamakaka and Biggins, 2005) (Table 1.2). The chromatin fibre can be modified by the incorporation of these variants, whose expression is not restricted to the S phase. Histone variants are distinguished by amino acid sequence differences and their replication-independent deposition is important for transcriptional regulation and epigenetic maintenance. Histone H2A has the largest number of variants, which include H2A.Z and H2A.X, found in the majority of eukaryotes, and MacroH2A and H2A.Bbd, which are only found in vertebrates. In yeast, H2A.Z prevents the spread



of heterochromatin (Meneghini *et al.*, 2003) and can be incorporated into a nucleosome by ATP-dependent histone exchange (Mizuguchi *et al.*, 2004) or a replication-independent chaperone, Nap1, can facilitate its deposition (Park *et al.*, 2005). H2A.Bbd is excluded from the inactive X chromosome (Chadwick and Willard, 2001) and has been shown to confer lower stability to the nucleosome (Gautier *et al.*, 2004). MacroH2A is concentrated on the inactive X chromosome (Costanzi *et al.*, 2000) and interferes with transcription factor binding and SWI/SNF nucleosome remodelling (Angelov *et al.*, 2003). H2A.X is phosphorylated in response to DNA double strand breaks and is recognized by the proteins involved in DNA repair (Celeste *et al.*, 2003). The H3 variant H3.3 is found in transcriptionally active chromatin (Ahmad and Henikoff, 2002; Chow *et al.*, 2005). It has been suggested that replication-independent deposition and inheritance of H3.3 in regulatory regions preserves transcriptionally active chromatin (Mito *et al.*, 2005). The CenH3 variant is involved in the assembly of centromeric chromatin (Ahmad and Henikoff, 2001).

Histone	Variants	Role(s)	Localization	Structural features	Function(s)
H2A	macroH2A	Inactivation of X chromosome	Inactive X chromosome	C-terminal non-histone-like region responsible for most functions	Repressing transcription initiation, Interferes with acetylation by p300. Blocks sliding by ACF and remodelling by Swi/Snf
	H2A.X	Repression	General Distribution	Conserved C-terminal SQ(E/D) motif is phosphorylated upon DNA damage	
	H2A.Z	Transcription activation/repression	Promoter, Hetero-chromatin boundary	C-terminal $\alpha$ -helix is essential for recognition	Facilitates TBP binding, is evicted upon transcription activation; Prevents elongation associated modification and remodelling at promoter
	H2A.Bbd	Transcription activation	Active X-chromosome and autosomes	Lack of c-terminal; Wraps 118-130 bp DNA around it.	p300 and Gal4-VP16 activated transcription is more robust on H2A.Bbd nucleosomes
H3	H3.3	Activation of transcription	Transcribed regions	Differs from H3 at only four amino acids	Transcription triggers deposition and removal
	CenH3	Organization of centromeric chromatin	Centromeres	Divergent N-terminal tails	

**Table 1.2: Histone variants and their functions.** The incorporation of histone variants into chromatin impacts on transcriptional regulation in various ways as described in the text. Adapted from Li *et al.*, 2007.

### 1.6.2. Histone modifications and their functions

As discussed previously, several different types of post-translational modifications have been identified on histones. There are over 60 different amino acids on histones where modifications have been detected and lysine and arginines can be methylated in one of three different forms, resulting in more than 100 different post-translational modifications. The majority of these modifications occur on the exposed N-terminal ‘tails’, some of which are shown in Figure 1.6. These modifications are discussed in detail below.



**Figure 1.6: Sites of histone modifications.** The amino terminal ‘tails’ of histones H2A, H2B, H3, and H4 host the vast majority of covalent histone modifications. Modifications can also occur in the globular core of histones (indicated by boxed regions). The location of acetylation (Ac), methylation (Me), phosphorylation (P) and ubiquitination (Ub) modifications are indicated above the relevant numbered amino acid residues.

The presence of these covalent modifications alters the arrangement of nucleosomes by means of *cis*- and *trans*-effects. *Cis*-effects are changes in the physical properties of

nucleosomes that are brought about by the presence of a histone modification - for example, the positive charge on lysine residues is neutralised by the addition of an acetyl group which reduces the binding of basic histones to negatively charged DNA, thereby enabling transcription factors to access the DNA (Vettese-Dadey *et al.*, 1996). Phosphorylation adds a net negative charge that is believed to compact nucleosome packaging (Nowak and Corces, 2004). Histone modifications can also elicit *trans*-effects by acting as a docking platform for the recruitment of enzymatic complexes that engage chromatin. For example, methylated lysine residues are recognised by chromo-like domains of the Royal family (chromo, tudor, MBT) and non-related PHD domain containing proteins. These proteins then facilitate downstream chromatin modulating events (Taverna *et al.*, 2007).

#### **1.6.2.1 Acetylation**

The core histones are reversibly acetylated at several lysine (K) residues. There are twelve known modification sites (Figure 1.6), two in histone H2A (K5, K9), two in histone H2B (K12, K15), four in histone H3 (K9, K14, K18, K56) and four in histone H4 (K5, K8, K12, K16). Histone acetylation is strongly associated with active transcription and histone acetylation sites are required for gene activity (Shahbazian and Grunstein, 2007). It is believed that acetylation may affect chromatin structure as it neutralizes the basic charge of lysine, which may affect the interaction of DNA with histones and nucleosome-nucleosome interactions (Tse *et al.*, 1998) and recently it has been shown that acetylation of H4K16 has a negative effect on the formation of the 30 nm chromatin fibre (Shogren-Knaak *et al.*, 2006). Thus, histone acetylation results in a ‘loosening’ of chromatin structure to allow greater access to transcription factors. Furthermore, histone H3 K9 acetylation in promoter regions is associated with low nucleosome density in the vicinity of transcription start sites (Nishida *et al.*, 2006). Histone acetylation can also function as recognition sites for factors that promote transcription (Shahbazian and Grunstein, 2007). For example, the bromodomain in BRG1, which is the catalytic subunit of the SWI/SNF chromatin remodelling complex, binds acetylated H4 K8, while

acetylation of H3 K9 and K14 is critical for the recruitment of TF<sub>II</sub>D (Agalioti *et al.*, 2002).

Recent studies which mapped acetylated histones on a genome-wide level found that acetylation of most lysines in the histone H3 and H4 tails was observed in the 5' end of coding regions and correlated with active transcription (Roth *et al.*, 2001; Kurdistani *et al.*, 2004; Liang *et al.*, 2004; Roh *et al.*, 2004; Schubeler *et al.*, 2004; Liu *et al.*, 2005; Pokholok *et al.*, 2005; Roh *et al.*, 2006; Koch *et al.*, 2007). Furthermore, many inducible genes are marked by histone acetylation even in the inactive state suggesting that the presence of histone acetylation serves to prime these genes for activation at a later stage (Roh *et al.*, 2004; Vogelauer *et al.*, 2000).

#### **1.6.2.2 Methylation**

Histones can be methylated either on lysine (K) or arginine (R) residues (Murray, 1964; Patterson and Davies, 1969). Lysine residues can be mono-, di-, or tri-methylated (Lachner *et al.*, 2003) whereas arginine residues can only be mono- or di-methylated and di-methylation can occur in a symmetrical or asymmetric configuration (Zhang, 2004). Histone H3 can be methylated on a number of lysine sites, which include K4, K9, K27, K36 and K79. Histone H3 can also be methylated on a number of arginine sites - R2, R8, R17 and R26. On histone H4 the main sites of methylation are K20 and R3.

Methylation at H3K4 was first observed in the trout testes (Honda *et al.*, 1975) and recently studies have linked it to active gene expression in numerous eukaryotes (Santos-Rosa *et al.*, 2002; Ng *et al.*, 2003; Schneider *et al.*, 2004; Schubeler *et al.*, 2004; Bernstein *et al.*, 2005; Pokholok *et al.*, 2005). The consensus emerging from these large-scale studies is that high levels of H3K4 trimethylation (H3K4me3) are associated with 5' regions of actively transcribed genes. This modification is also positively correlated with histone acetylation and RNA polymerase II occupancy. However, there are differences in the patterns of H3K4 dimethylation (H3K4me2) between yeast and vertebrate chromatin. In vertebrates, most H3K4me2 co-localizes with H3K4me3 in discrete regions located nearby highly expressed genes (Schneider *et al.*, 2004; Bernstein *et al.*, 2005). In contrast to this H3K4me2 in *S. cerevisiae* is spread throughout genes,

peaking in the middle of the coding region and can be associated with active as well as 'poised' genes (Santos-Rosa *et al.*, 2002; Ng *et al.*, 2003; Pokholok *et al.*, 2005). H3K4 monomethylation (H3K4me1) is most abundant at the 3' end of yeast genes and outside of promoter regions and it has been correlated with functional enhancers in human cell lines (Heintzman *et al.*, 2007; Roh *et al.*, 2007). Clearly the number of methyl groups plays a significant role in the functional consequences of histone methylation.

Heterochromatin in higher eukaryotes is characterised by histone hypoacetylation and H3 K9 methylation (Richards and Elgin, 2002). Heterochromatin protein 1 (HP1) was shown to specifically recognize methylated H3 K9 via its chromodomain (Lachner *et al.*, 2001; Nakayama *et al.*, 2001). This recognition of H3 K9 by HP1 is required for the formation of heterochromatin. The production of short heterochromatic RNA (shRNA) is involved in the targeting of H3 K9 methylation to heterochromatin regions (Grewal and Moazed, 2003). However, H3 K9 methylation and HP1 binding were recently detected on active genes (Vakoc *et al.*, 2005, 2006) suggesting that H3 K9 methylation is not limited to inactive regions of chromatin. Methylation of histone H3 K27 exhibits some similarities to K9 methylation. Both lysines are found within an ARKS sequence on histone H3 and K27 methylation is also associated with transcriptional silencing (Ringrose *et al.*, 2004). In particular, methylation of H3 K27 is characteristic of the inactive X chromosome in female cells (Wang *et al.*, 2001; Mak *et al.*, 2002). H3 K27 methylation facilitates the binding of polycomb via its chromodomain. Polycomb is a component of the polycomb repressive complex 1 (PRC1) and is required for transcriptional silencing by the polycomb group (PcG) complex (Cao *et al.*, 2002).

Methylation of H3K36 is associated with the elongating, serine 2-phosphorylated form of RNA Polymerase II (Xiao *et al.*, 2003; Schaft *et al.*, 2003; Krogan *et al.*, 2003) and is detected across actively transcribed regions, peaking at the 3' end of genes (Bannister *et al.*, 2005 b; Mikkelsen *et al.*, 2007). Processivity of Pol II through coding regions requires histone acetylation. Transcriptional regulation also needs to suppress initiation from cryptic start sites that occur within coding regions. To suppress these initiation events, H3 K36 methylation creates a recognition site for the chromodomain protein Eaf3, which in turns recruits the Rpd3 HDAC complex (Keogh *et al.*, 2005; Carrozza *et*

*al.*, 2005). The HDAC activity of this complex removes histone acetylation associated with transcriptional elongation, thereby suppressing internal initiation of transcription. Methylation of H3 K79 is unusual because this modification lies at the core of the nucleosome rather than on the tail. Global analysis of H3 K79 methylation has shown that this modification primarily associates with the coding region of actively transcribed genes (Miao *et al.*, 2005) but so far no protein has been identified that binds to this modified residue and links it to transcriptional regulation. The only evidence for how H3 K79 methylation functions in transcriptional activation comes from yeast, where it was shown that the presence of H3 K79 methylation in euchromatic regions prevents the Silent information regulator (Sir) proteins from interacting with active chromatin, thus concentrating Sir Complex binding at silent chromatin regions (Ng *et al.*, 2003 b). H3 K79 methylation has also been implicated in DNA repair as the checkpoint protein P53BP1 has been shown to bind methylated H3 K79 (Martin and Zhang, 2005). H4 K20 methylation is connected to transcriptional repression and DNA repair, although very little is known about how it functions in these processes. The lysine demethylase JMJD2A binds to methylated H4 K20 via its tudor domain (Huang *et al.*, 2006) and this interaction may contribute to transcriptional repression. In fission yeast, sites of DNA damage contain methylated H4K20, which is recognized by the checkpoint protein Crb2 (Sanders *et al.*, 2004; Botuyan *et al.*, 2006).

Histone arginine methylation can contribute to active and repressive chromatin states (Strahl *et al.*, 2001; Yu *et al.*, 2006). Methylation of histone H3 at R2, R17, and R26 (Schurter *et al.*, 2001) enhances nuclear receptor-mediated gene activation (Chen *et al.*, 1999) while methylation of histone H4 R3 is involved in nuclear receptor-mediated transcription activation (Wang *et al.*, 2001 b). Histone arginine methylation has recently been implicated in regulating pluripotency in the early mouse embryo (Torres-Padilla *et al.*, 2007). Methylation of arginine residues is enhanced in four cell blastomeres that contribute to the inner cell mass and is minimal in the cells that contribute to the mural trophectoderm, suggesting that this modification could contribute to early cell fate determination. However, it is not understood how arginine modifications contribute to chromatin remodelling and gene activation as no proteins have been identified which bind to methylated arginine residues.

### **1.6.2.3. Phosphorylation**

The core histones are phosphorylated on specific serine and threonine residues (Figure 1.6). Most studies have focused on the role of H3S10 phosphorylation (Johansen and Johansen, 2006). Phosphorylation on this residue was found to occur in tandem with the activation of immediate early genes such as c-jun and c-fos (Mahadevan *et al.*, 1991) and at activated heat shock genes (Nowak and Corces, 2000). Phosphorylation of H3S10 was also observed during chromosome condensation (Wei *et al.*, 1998). Therefore H3S10 phosphorylation is implicated with chromatin states, the ‘open’ chromatin of active genes during interphase and the ‘closed’ condensed chromatin of mitotic chromosomes. H3S10 seems to function by regulating a methylation/phosphorylation switch that inhibits HP1 binding to H3K9me3 (Fischle *et al.*, 2003) and indeed this was shown to be true as phosphorylation of H3S10 is responsible for HP1 dissociation during mitosis (Fischle *et al.*, 2005). These observations suggest a model for how H3S10 phosphorylation functions in the two opposing processes of gene activation and chromosome condensation. During interphase this modification promotes removal of HP1 from specific regions, allowing gene expression. Removal of the phosphorylation mark would therefore promote heterochromatin formation and promote chromatin condensation. Recently DNA methylation has been shown to have a role in targeting H3S10 phosphorylation to pericentromeres (Monier *et al.*, 2007).

### **1.6.2.4. Ubiquitination**

Histones H2A and H2B have been reported to be ubiquitinated, the carboxyl end of ubiquitin is added to K119 in H2A and K120 in H2B in humans. Histone H2A was the first ubiquitinated histone to be identified (Goldknopf *et al.*, 1975) and the majority of this modification is the monoubiquitinated form. H2A ubiquitination has been linked to polycomb silencing and X-chromosome inactivation (Wang *et al.*, 2004; De Napoles *et al.*, 2004). Ubiquitinated H2A at K119 was found on the inactive X-chromosome in females and is correlated with the recruitment of PcG proteins PRC1-like (PRC1-L). The ubiquitin moiety is approximately half the size of a core histone, so it has been suggested that ubiquitination of a nucleosome would impact chromatin folding, thus affecting transcription (Shilatifard, 2006).



#### **1.6.2.5. Sumoylation**

SUMO is a small ubiquitin-related protein of ~100 amino acids, which is capable of being ligated to its target protein. Protein sumoylation is involved in the regulation of transcription factors and components of the transcriptional machinery (Manza *et al.*, 2004, Hannich *et al.*, 2005) and often results in transcriptional repression (Ghioni *et al.*, 2005). Histone H4 sumoylation has been reported in mammalian cells and correlates with transcriptional repressive events such as histone deacetylation and HP1 recruitment (Shiio and Eisenman, 2003). The recent description of histone sumoylation in *S. cerevisiae* and its association with transcriptional repression may address the fact that, unlike vertebrates and *S. pombe*, *S. cerevisiae* has no histone marks associated with transcriptional repression (Nathan *et al.*, 2006). This evolutionarily conserved repressive mark seems to block activating acetylation and ubiquitination marks.

#### **1.6.2.6. ADP-ribosylation**

Mono-ADP ribosylation of histones is linked to DNA repair and cell proliferation (Hassa *et al.*, 2006). Histones are mono-ADP-ribosylated when exposed to DNA damaging agents. This modification has the potential for ‘cross-talking’ with other modifications as mono-ADP-ribosylation on H4 occurs preferentially when H4 is acetylated (Golderer and Grobner, 1991). Poly-ADP-ribosylation has not been confirmed on histones but it may play a role in local chromatin compaction (Hassa *et al.*, 2006).

#### **1.6.3. The histone code hypothesis**

The initial observations that histone acetylation influenced the initiation and elongation phases of transcription led to the suggestion that acetylation of histones provided an epigenetic code for the regulation of transcription (Turner, 2000). Strahl and Allis then provided a histone code model for the function of specific modifications (Strahl and Allis, 2000). The histone code hypothesis predicted that histone modifications on the same or different histones may be interdependent and that different combinations of histone modifications may act synergistically or antagonistically to affect transcription. For example, acetylation of H3K14 by GCN5 is enhanced by H3S10 phosphorylation

(Berger, 2002) while methylation of H3K9 is inhibited by phosphorylation of H3S10 and vice versa (Rea *et al.*, 2000). Furthermore, the code predicted that distinct histone modifications provide a binding site for chromatin-associated 'effector' proteins, which mediate downstream functions. For example, the PHD (plant homeodomain) finger in BPTF, the largest subunit of the nucleosomal remodelling factor (NURF) complex, interacts specifically with peptides modified with H3K4me3 (Wysocka *et al.*, 2006).

Agalioti and colleagues carried out one of the first studies to examine the histone code hypothesis (Agalioti *et al.*, 2002). The human IFN- $\beta$  gene is turned on by three transcription factors which form an enhanceosome at the enhancer region. The enhanceosome facilitates transcription by recruiting HATs, SWI/SNF chromatin remodelling complex and basal transcription factors in an ordered manner. GCN5 acetylated H4K8, which was required for the recruitment of the SWI/SNF complex and acetylation of H3K9 and H3K14 was required for TFIID recruitment. The authors proposed that distinct acetylation marks were required for the recruitment of transcription complexes and that this constituted a histone code. A number of large scale studies have also been carried out in budding yeast, fission yeast, *D. melanogaster*, mouse and human, which have examined a range of histone modifications (Roh *et al.*, 2004, 2005, 2006; Kurdistani *et al.*, 2004; Dion *et al.*, 2005; Liu *et al.*, 2005; Pokholok *et al.*, 2005; Millar *et al.*, 2006; Rao *et al.*, 2005; Wiren *et al.*, 2005; Sinha *et al.*, 2006; Schubeler *et al.*, 2004; Bernstein *et al.*, 2005, 2006; Boyer *et al.*, 2006; Heintzman *et al.*, 2007; Barski *et al.*, 2007; Koch *et al.*, 2007). Some of these studies reported a clear coordination between various histone modifications and gene activity, namely histone H3 and H4 acetylation and H3K4 methylation states are found at the 5' regions of active genes (Pokholok *et al.*, 2005; Bernstein *et al.*, 2005; Schubeler *et al.*, 2004; Roh *et al.*, 2005, 2006; Koch *et al.*, 2007), whereas elevated H3K27 methylation correlates with gene repression (Boyer *et al.*, 2006; Roh *et al.*, 2006). In addition, H3 acetylation and H3K4me1 modifications outside of promoter regions has been correlated with enhancer elements (Heintzman *et al.*, 2007; Roh *et al.*, 2005). In contrast, other studies did not find a correlation between histone acetylation and gene activity (Kurdistani *et al.*, 2004; Liu *et al.*, 2005) and therefore dispute the existence of a histone code. Recent discoveries have complicated the simplified view that specific histone modifications underlie either an

active or inactive chromatin status - for example the co-localization of the apparently contradictory modifications H3K4me3 and H3K27me3 in bivalent domains (Bernstein *et al.*, 2006; Mikkelsen *et al.*, 2007), while H3K9me3 has been shown to be enriched at a number of active promoters (Vakoc *et al.*, 2005). In addition, the Sin3-HDAC complex has been shown to bind to H3K4me3 via the PHD (plant homeodomain) domain of the Ing2 protein to repress gene expression (Shi *et al.*, 2006). This suggests that the presence of multiple modifications may be required to elicit a specific biological output and Ruthenburg and colleagues (2007) have proposed that in many cases one histone modification is not sufficient to recruit a given complex, rather multiple histone modifications all contribute to the recruitment and stabilization of chromatin complexes and dictate functional outcomes.

#### **1.6.4. DNA methylation**

In eukaryotes, DNA methylation is confined to cytosine bases at CpG dinucleotides and is associated with gene repression (Klose and Bird, 2006). CpGs often cluster into CpG islands and approximately 60% of human promoters are associated with CpG islands (Bird, 2002). It has been suggested that the majority of CpG islands are always unmethylated but some are methylated in a tissue-specific manner. DNA methylation functions to inhibit gene expression by two mechanisms, firstly, DNA methylation can inhibit transcription factors from binding to their DNA recognition sequence (Watt and Molloy, 1988). Secondly, proteins which recognize methyl-CpG, methyl-CpG binding proteins (MBPs) (Hendrich and Bird, 1998), can recruit transcriptional co-repressors such as histone methyltransferases and deacetylases to modify chromatin and mediate gene silencing (Jones *et al.*, 1998; Nan *et al.*, 1998; Sarraf and Stancheva, 2004). In these cases, DNA methylation is coupled with repressive histone modifications. Furthermore, DNA methylation within the body of a gene has been shown to alter chromatin structure and gene expression by affecting Pol II elongation efficiency (Lorincz *et al.*, 2004).

There are two classes of mammalian DNA methyltransferases (DNMTs) - *de novo* and maintenance DNMTs. DNMT3a and DNMT3b are members of the *de novo* class, as they are responsible for methylating at previously unmethylated CpGs. DNMT1 is a maintenance enzyme as it copies methylation patterns onto newly synthesized DNA

strands. *De novo* methylation is important during embryogenesis as the paternal genome is actively demethylated following fertilization and the maternal genome is demethylated passively as a result of DNA replication. Many CpG sites are re-methylated in the blastocyst resulting in the patterns observed in the adult (Reik *et al.*, 2001). Alterations in DNA methylation patterns have been implicated in numerous diseases including cancer. The promoters of tumour suppressor genes are often hypermethylated, which silences their expression (Jones and Baylin, 2007) and it has recently been proposed that cancer may evolve from stem cells which carry epigenetic alterations in addition to other genomic alterations (Feinberg *et al.*, 2006), but their exact origins remain controversial (Bjerkvig *et al.*, 2005).

### **1.6.5. Regulation of embryonic stem cell pluripotency**

One of the most important discoveries in the field of molecular biology was the development of methods during the 1980s to generate pluripotent embryonic stem (ES) cells from the inner cell mass of pre-implantation embryos (Evans and Kaufman, 1981; Martin, 1981). These cells could be grown indefinitely in culture and microinjected back into mouse blastocytes where they contributed to the formation of various cell lineages in the adult mouse. ES cells represent an undifferentiated cell type that can in theory generate daughter cells capable of differentiating into every cell type found in the adult organism, a property known as pluripotency. It is becoming clear that epigenetic mechanisms play a role in maintaining ES cells in a pluripotent state (Azuara *et al.*, 2006, Mikkelsen *et al.*, 2007) and that the maintenance of ES cell pluripotency depends on the transcriptional expression and silencing of a number of genes (Boyer *et al.*, 2006). The key factors involved in regulating pluripotency are discussed below.

#### **1.6.5.1. Transcription factors involved in regulating pluripotency**

Several pluripotency-sustaining transcription factors such as NANOG and OCT4 are expressed in ES cells and are silenced upon differentiation (Ramalho-Santos *et al.*, 2002; Ivanova *et al.*, 2002). Two recent genome-wide studies identified targets of NANOG, OCT4, and SOX2 in ES cells (Boyer *et al.*, 2005; Loh *et al.*, 2006) and demonstrated that

these proteins bind to several hundred genes which can be transcriptionally active or silent. These loci are often involved in developmental processes and included ES cell specific genes and repressed tissue-specific transcription factors.

PcG proteins are required for ES cell pluripotency and are dramatically down-regulated upon differentiation (Cao and Zhang, 2004). Genome-wide studies investigating PcG protein binding have been carried out in human and mouse ES cells (Lee *et al.*, 2006; Bracken *et al.*, 2006; Boyer *et al.*, 2006; Schwartz *et al.*, 2006; Negre *et al.*, 2006; Tolhuis *et al.*, 2006). Genes that are required during differentiation and development, such as members of the Hox and Pax transcription families, are repressed in human and mouse ES cells by the PcG machinery and are often found in bivalent chromatin regions (Azuara *et al.*, 2006; Bernstein *et al.*, 2006) (see section 1.6.5.2). The majority of the genes cooperatively regulated by NANOG, OCT4, and SOX2 were repressed and overlapped with PcG protein binding sites. This suggests that these transcription factors may maintain pluripotency by coordinating the recruitment of PcG complexes to repress tissue specific genes.

#### **1.6.5.2. Epigenetic regulation of pluripotency**

In addition to transcription factors, chromatin structure and epigenetic modifications play a key role in the maintenance of pluripotency (Spivakov and Fisher, 2007; Reik, 2007). Evidence suggests that ES cell chromatin may be less compact and more transcriptionally 'permissive' than differentiated cells (Bernstein *et al.*, 2007). For example, differentiated and undifferentiated human and mouse ES cells showed dramatic differences in the nuclear organization of centromeric heterochromatin and regions involved in pluripotency (Wiblin *et al.*, 2005). In differentiated cells, many inactive genes are positioned close to centromeric heterochromatin (Brown *et al.*, 1999), but this phenomenon has not been observed in ES cells. Furthermore, chromatin proteins are more loosely associated in ES cells when compared with differentiated cells, indicating that the chromatin of ES cells is more accessible (Meshorer *et al.*, 2006).

Recent studies of histone modifications in pluripotent and differentiated cells have also advanced our understanding of the chromatin properties important for initiating and

maintaining a pluripotent state (Bernstein *et al.*, 2006; Azuara *et al.*, 2006; Chambeyron *et al.*, 2005; Szutorisz *et al.*, 2005 b; Mikkelsen *et al.*, 2007). These studies have shown that inactive genes in ES cells can be associated with high levels of H3K4me2, H3K4me3 and acetylated histones H3 and H4. Some of these genes were also enriched for the repressive H3K27me3 modification. Given that histone acetylation, H3K4me2 and H3K4me3 are normally associated with expressed genes while H3K27me3 is associated with non-expressed genes, the presence of these ‘contradictory’ modifications on non-expressed genes in ES cells was intriguing. Consecutive or sequential ChIP reactions (re-ChIP) confirmed that these contradictory modifications were present on the same or neighbouring nucleosome (Azuara *et al.*, 2006, Bernstein *et al.*, 2006). These ‘bivalent’ domains were found to mainly overlay developmental regulator genes, the majority of which were not expressed in ES cells. In differentiated cell types, H3K27me3 alone marked several inactive developmental genes whereas H3K4me3 marked the active ones. This suggests that transcription factors involved in tissue specific development are both primed for expression and ‘held back’ at the same time so that their expression can be tightly regulated in ES cells.

### **1.7. Identification and characterisation of non-coding regulatory elements**

It is evident that gene expression in eukaryotes is a highly controlled process requiring regulation at many different levels. The non-coding regulatory elements, in combination with the proteins that interact with them, are crucial in determining gene expression patterns. In order to understand the regulatory networks which control gene expression patterns, it is important to identify and characterise the regulatory elements associated with genes. Over the past four decades, various assays have been used to identify regulatory elements in the human genome, the majority of which have been low-throughput processes. The completion of the human genome sequence is allowing for the development of high-throughput experimental and computational methods which should ensure that these elements are identified in a more efficient manner. Discussed below are many of the low-throughput and high-throughput methods which have been used to identify non-coding regulatory elements in a variety of species.

## **1.7.1. Classical or low-throughput methods**

### **1.7.1.1. DNA footprinting**

DNA footprinting is used to identify binding sites of proteins that bind to DNA. It works on the principle that when a protein binds to DNA it ‘protects’ the underlying DNA from cleavage by DNase I when compared with unbound DNA (Galas and Schmitz, 1978). Usually DNA fragments 200-300 bp in length are used as targets and are radiolabelled at one end. These fragments are then incubated in the presence or absence of a protein extract and then exposed very briefly to low concentrations of DNase I. Digested products are size fractionated on denaturing polyacrylamide gels and then autoradiographed. Control samples show a series of bands of different lengths due to random digestion by DNase I, whereas a test sample contains gaps where no fragments are observed (footprints). The gaps indicate the sites where protein is bound. While DNA footprinting can identify sites of DNA-protein interaction, it does not confer any functionality to the site. Therefore, it is often used in combination with other techniques such as electromobility shift assays (section 1.7.1.3) to gain a greater insight into the function of the regulatory element.

### **1.7.1.2. DNase I hypersensitive site mapping**

This technique is based on the observation that regions of chromatin which are accessible to transcription factors are more sensitive to DNase I digestion than condensed chromatin. Regions identified by their DNase I hypersensitivity are therefore likely to be involved in transcriptional regulation. This approach has the advantage of identifying most if not all *cis*-regulatory elements, but does not directly confer functionality to identified elements. DNase I hypersensitive sites (HSs) are identified by exposing chromatin to low amounts of DNase I for a short period of time, followed by DNA purification of cleaved DNA, restriction enzyme digestion, gel electrophoresis, southern blotting and hybridization with a radioactive DNA probe. This technique has been widely used to identify *cis*-regulatory elements such as promoters, enhancers, repressors, insulators and locus control regions in many cell types (Weintraub and Groudine, 1976;

Wu, 1980; Gross and Garrard, 1988). More recently, high-throughput approaches have been used to identify HSs (see section 1.7.3.3).

#### **1.7.1.3. Electromobility shift assays**

An electromobility shift assay (EMSA), also known as a gel-shift or gel retardation assay, is another technique used for studying DNA-protein interactions. It relies on the principle that a DNA sequence bound by a protein(s) migrates slower than unbound DNA during electrophoresis (Garner and Revzin, 1981). In a typical experiment, radiolabelled DNA fragments suspected of containing a regulatory sequence are incubated with or without protein extracts and then size-fractionated by polyacrylamide gel electrophoresis. DNA fragments which bind protein are identified by their lower mobility band on a gel. This type of assay can be used to quickly identify the presence of a specific DNA-protein interaction but with the caveat that *in vitro* identified binding sites do not always reflect *in vivo* binding sites (Lieb *et al.*, 2001). Both EMSA and footprinting assays can often detect unintended DNA-protein interactions as result of non-specific proteins, such as DNA repair proteins, binding to the end of DNA probes (Klug, 1997).

#### **1.7.1.4. PCR-based methods for detecting DNA-protein interactions**

PCR can be used to amplify DNA fragments that bind proteins. Several PCR-based techniques have been developed for this purpose, namely **s**ystematic **e**volution of **l**igands by **e**xponential enrichment (SELEX) (Tuerk and Gold, 1990), **s**election **a**nd **a**mplification **b**inding site (SAAB), **c**yclic **a**mplification **s**election **t**argets (CASTing) (Wright *et al.*, 1991), and **m**ultiplex **s**election **t**arget (MUST) (Nallur *et al.*, 1996). In these methods DNA-protein complexes are isolated by techniques such as immunoprecipitation and binding to affinity columns. The DNA is then recovered, PCR amplified, then mixed with fresh proteins and after many rounds of this DNA fragments interacting with a specific protein are enriched. Quantitative PCR in combination with chromatin immunoprecipitation can also be used to identify DNA-protein interactions (section 1.7.3.5).



### 1.7.1.5. Reporter-gene assays which confer function

There are many variations to the reporter-gene assay, which allow it to be used for identifying the majority of regulatory elements (Matson *et al.*, 2006) (See Figure 1.7). The 'test' genomic region is cloned into a plasmid upstream of a reporter gene such as the luciferase, chloramphenicol acetyltransferase (CAT),  $\beta$ -galactosidase, green fluorescent protein (GFP), or G418 resistance gene. The construct is then transfected transiently or stably into cultured cells and the output of the reporter gene is assayed to determine if the test sequence has functional activity.

The arrangement of the construct depends on the regulatory activity being tested for (Fig 1.7), for example if a DNA segment is being tested for core promoter activity it is cloned immediately upstream of a reporter gene which lacks a promoter. Proximal promoters are assayed by cloning them upstream of a reporter gene whose expression is driven by a weak heterologous core promoter. This system can also be used to test for gene enhancer and silencer activity, by cloning these putative elements in a reporter construct whose expression is driven by a weak or strong promoter. An increase/decrease in reporter gene expression is used to determine enhancer/silencer activity. These approaches have been used to characterise these elements in a number of genes including c-Myc and SCL (Mautner *et al.*, 1995; Gottgens *et al.*, 1997).

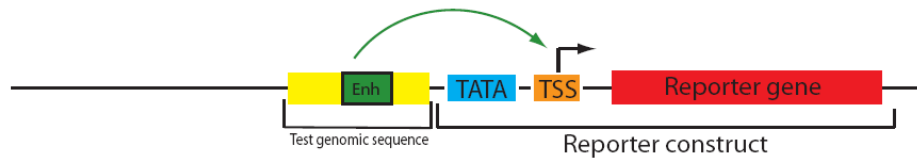
Insulators can be measured for enhancer-blocking (Bell *et al.*, 1999) or heterochromatin barrier activity (Recillas-Targa *et al.*, 2002). In enhancer-blocking assays, the putative insulator sequence is cloned between an enhancer and promoter that are known to interact. If the element has enhancer-blocking activity, then it will interfere with enhancer-promoter communication and reporter gene expression will be reduced. Assaying for heterochromatin barrier activity requires an assay in which the reporter construct is stably integrated into the genome. Barrier elements flanking the reporter gene would shield it from position effects which cause transgene silencing. Thus barrier elements allow for position-independent expression of the reporter gene. Conferring LCR activity requires the identification of a genomic segment that can overcome position effects to confer temporal and tissue specific expression of a reporter gene (Grosveld *et al.*, 1987).

There are several disadvantages associated with using reporter assays to identify regulatory elements. Firstly, the location of these elements is often not known and they can be found close to and far from a gene. Furthermore, regulatory elements can be different sizes and knowing what size segment to test can be a problem. Secondly, chromatin context plays a key role in regulating gene expression patterns and these reporter constructs do not reflect the correct context. Thirdly, if the cell culture system used does not match developmental conditions under which the regulatory element is normally active, then the activity may not be detected. Despite these disadvantages, reporter genes assays are still the most accurate way of conferring functionality upon a putative regulatory element.

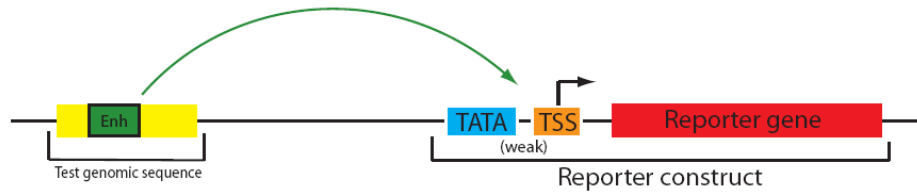
### A) Core promoter



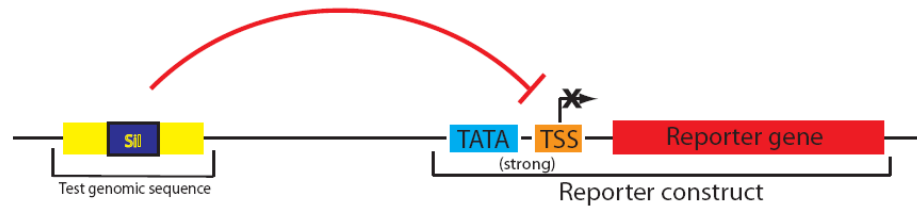
### B) Proximal Promoter



### C) Enhancer

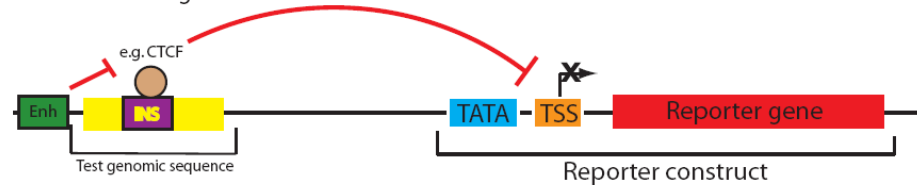


### D) Silencer

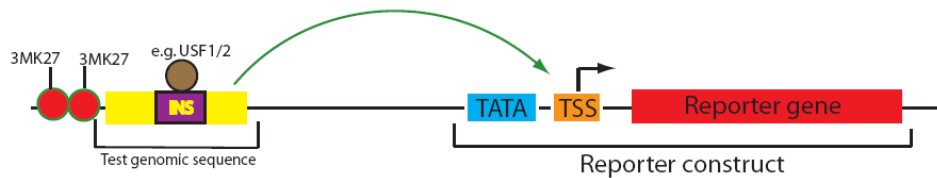


### E) Insulator

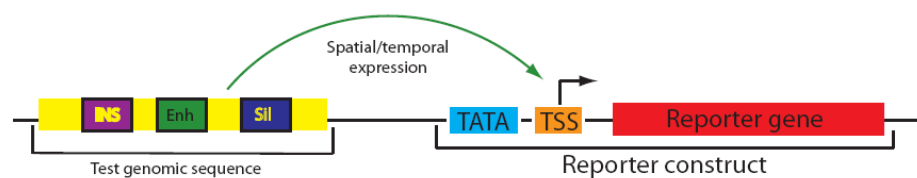
#### 1. Enhancer blocking element



#### 2. Barrier element



### F) Locus control region



**Figure 1.7: Functional plasmid-based reporter gene assays for the identification of regulatory elements.** A) A genomic sequence (yellow box) is tested for core promoter activity by cloning it immediately upstream of a reporter gene which lacks an endogenous promoter. B-D) Proximal promoters, enhancers and silencers can be assayed for by cloning a genomic segment upstream of a reporter gene driven by the appropriate strength promoter. E) Enhancer-blocking insulator elements interfere with enhancer-promoter communication, thereby down-regulating reporter gene expression. Barrier insulators shield a reporter gene from heterochromatin-mediated gene inactivation. F) Locus control regions can overcome position effects and confer correct expression patterns in reporter assays. Enh: enhancer; INS: insulator; Sil: silencer.

### 1.7.2. Computational detection of regulatory elements

A complete computational approach to studying transcriptional regulatory elements (Elnitski *et al.*, 2005) often requires diverse data sets in order to determine (1) promoter location (2) predicted and verified transcription factor binding sites (3) gene expression profiles and (3) sequence conservation. The availability of such data sets has allowed rigorous computational prediction of regulatory elements in vertebrate genomes as discussed below (Prakash and Tompa, 2005; Cora *et al.*, 2005; Hallikas *et al.*, 2006; Prabhakar *et al.*, 2006).

#### 1.7.2.1. Promoter location prediction

Identifying the promoter of a particular gene can be a difficult task as core promoter sequences can be located a large distance from the first coding exon due to 5'-untranslated regions and introns (Maston *et al.*, 2006). As discussed in section 1.4.1, promoters contain various combinations of core promoter motifs (Gershenson *et al.*, 2005) and searching for the co-occurrence of these motifs has had limited success in predicting promoter locations (Fickett and Hatzigeorgiou, 1997). Promoter prediction programs based on the analysis of known core promoters have been most successful and include PromoterInspector (Scherf *et al.*, 2000), First EF (Davuluri *et al.*, 2001) and Eponine (Down and Hubbard, 2002). However, the sensitivity and specificity of these programs is limited by the number of known core promoters and are limited to finding new promoters that are similar to ones in the training data. Approximately 60% of human genes lie near CpG islands and a comparison of promoter-prediction programs found that promoters associated with CpG islands are predicted well. However, prediction of the

other 40% of promoters is much less reliable (Bajic *et al.*, 2004). Therefore, the availability of more experimental data in the form of novel transcripts (Carninci *et al.*, 2005), more precise mapping of 5' ends of transcripts and ChIP-chip data for factors which bind to promoters (Kim *et al.*, 2005) will aid the training of these programs.

#### **1.7.2.2. Prediction of transcription factor binding sites**

Genome sequences can be scanned for sequence motifs which match experimentally verified transcription factor binding sites (TFBSs). Most TFBSs are very short sequences and are often degenerate, usually only 4-6 bp within each TFBS are fully conserved and these sites often occur in clusters (Maniatis *et al.*, 1987). Experimental data on the location of transcription factor binding sites has been compiled in databases such as TRANSFAC (Wingender *et al.*, 2000) and information on various TFBSs can be used to build a position-specific scoring matrix (PSSM) for a particular TF (Vavouri and Elgar, 2005). Programs such as MATCH (Kel *et al.*, 2003) can scan an input sequence for matches to all PSSMs in TRANSFAC. However programs based on PSSMs often identify a high number of false positives due to the quality of data used to build a matrix (Fogel *et al.*, 2005). To overcome this problem, more sophisticated statistical models have been used to predict TFBSs such as the program JASPAR (Sandelin *et al.*, 2004). Hallikas and colleagues have recently used the well defined binding specificities of several transcription factors involved in the Hedgehog, Wnt and Ras/MAPK signalling pathways to develop a computational tool which could identify mammalian enhancer elements regulated by these pathways (Hallikas *et al.*, 2006). The authors made use of the observation that TF binding sites tend to be found in clusters when forming a tissue-specific enhancer element, to produce a computational tool capable of accurately identifying enhancers. However the problem still exists that the majority of TF binding specificities have not been determined.

#### **1.7.2.3. Comparative sequence analysis**

With the completion of genome sequences for a number of organisms, it has become easier to compare genomic sequences and identify regions of high sequence conservation across species. The use of comparative sequence analysis or comparative genomics to

identify non-coding regulatory elements has recently become very popular. The rationale behind this approach is that, just like coding sequences, regulatory sequences are under evolutionary selective pressure and so should have evolved at a slower rate than other non-coding sequences. A number of programs have been developed to identify sequences which have been conserved through evolution such as PhastCons (Siepel et al., 2005), Footprinter (Blanchette and Tompa, 2003), SynPlot (Gottgens *et al.*, 2001) and VISTA (Visel *et al.*, 2007). Comparative sequence analysis has been used in many cases to identify *bona fide* regulatory elements (Gottgens *et al.*, 2000; Martin *et al.*, 2004; Nobrega *et al.*, 2003; Woolfe *et al.*, 2005; Pennacchio *et al.*, 2006, 2007). For example, whole genome comparisons between humans and the pufferfish, *Fugu rubripes*, identified 1400 highly conserved non-coding sequences (Woolfe *et al.*, 2005). Many of the sequences displayed tissue-specific enhancer activity when tested in functional assays. Pennacchio and colleagues (2006) used a similar approach to expand the number of characterized human enhancers and the data derived from such studies will be extremely useful in the training of enhancer prediction programs.

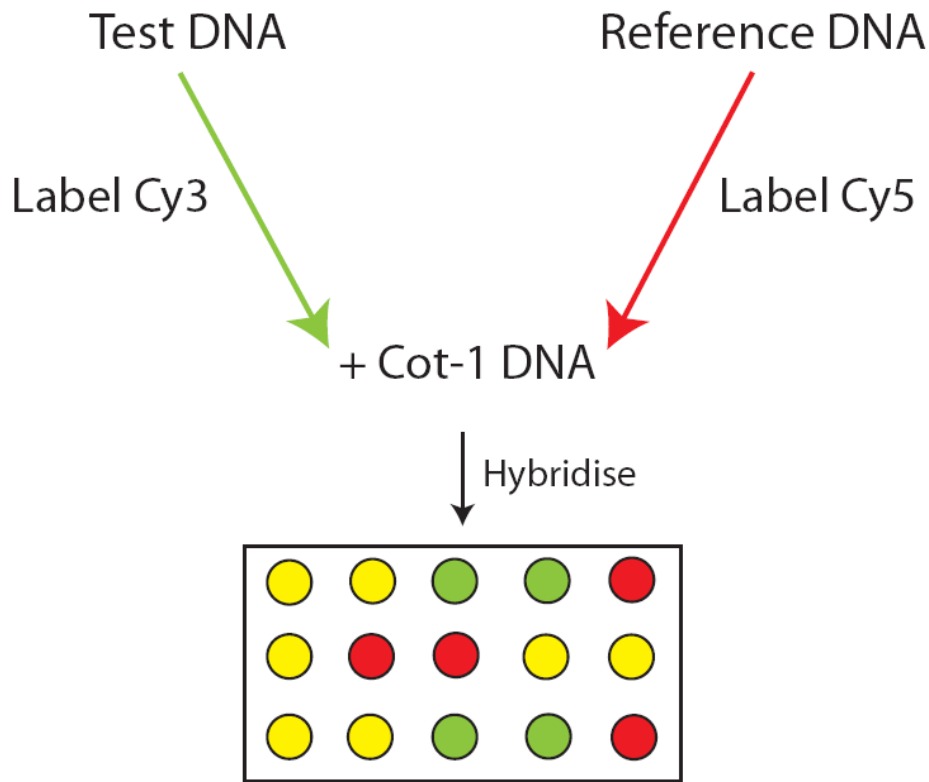
However, conserved sequence elements do not always correspond to functional regulatory regions (Balhoff and Wray, 2005). It has been suggested that there is a high rate of evolutionary turnover of mammalian TFBSs (Dermitzakis and Clark, 2002). This may be due to degeneracy of TFBSs or a specific regulatory element may not be conserved (Follows *et al.*, 2003). A recent study by Odom and colleagues, utilized ChIP-chip to map the binding sites of conserved tissue-specific transcription factors (FOXA2, HNF1A, HNF4A, and HNF6A) in human and mouse hepatocytes (Odom *et al.*, 2007). It was observed that many of the binding events for the transcription factors were species-specific. This analysis also showed that for many TF binding events in human, the orthologous gene in mouse is bound but not at the conserved sequence element. Therefore it seems that many of the transcriptional regulatory elements relevant to human development may not be highly functionally conserved between evolutionarily distant species. Furthermore, sequence comparisons showed that 21,855 of the sequences associated with histone acetylation islands are not conserved between the human and mouse genomes and random sampling showed that 50% of these non-conserved sequences function as enhancers (Roh *et al.*, 2007). Therefore, histone modification data

can be used to identify species-specific regulatory elements that would otherwise be missed by comparative sequence analysis alone. Phylogenetic shadowing (Boffelli *et al.*, 2003), which analyses sequences from closely related species such as primates, may be required to identify human specific regulatory elements (Prabhakar *et al.*, 2006).

### **1.7.3. Applications of genomic DNA microarrays to identify regulatory elements**

With the completion of many genome sequences, DNA microarray technology has emerged as an important technology for investigating global gene regulation events as described below (Hoheisel, 2006). Typically, DNA microarrays consist of a large collection of DNA sequences that are fixed to the surface of a glass slide. The DNA sequences can be comprised of large genomic clones (BACs, PACs and cosmids), cDNA clones, primer specific PCR products, or short oligonucleotides (Fiegler *et al.*, 2003; Duggan *et al.*, 1999; Dhimi *et al.*, 2005; Lipshutz *et al.*, 1999). High-density oligonucleotide microarrays are manufactured by the direct synthesis of oligonucleotides on the slide surface. Photolithography, optical mirrors or ink-jets can be used to synthesize oligonucleotides (Lipshutz *et al.*, 1999; Singh-Gasson *et al.*, 1999; Hughes *et al.*, 2001). Spotted arrays rely on robotic devices to spot clone fragments, PCR products, or oligonucleotides (Schena *et al.*, 1995). The glass slides are coated with reactive molecular groups such as poly-L-lysine, which allows the DNA probes to bind to the slide.

For spotted microarrays, the test sample (RNA or DNA) and a reference sample are normally fluorescently labelled with nucleotide derivatives, usually containing Cy3 and Cy5 (Figure 1.8). DNA sequences on the microarray may contain repetitive sequences so the binding of repetitive DNA elements is suppressed by using Cot1 DNA in a competitive hybridization. During hybridization, the labelled samples bind to their complementary immobilized probes sequences and the fluorescent signal is calculated in the two channels to determine which sequences are enriched in the test sample relative to the reference sample.



**Figure 1.8: The basic principles of two-colour competitive hybridisation.** The test DNA sample is labelled with Cy3 nucleotide derivatives while the reference DNA sample is labelled with Cy5 nucleotide derivatives. The two DNA samples are mixed together with Cot-1 DNA and are competitively hybridised to a microarray. Green spots represent genomic regions that are enriched in the test sample relative to the reference sample, red spots represent DNA sequences depleted in the test sample relative to the reference and yellow spots represented DNA sequences that are present in equal amounts in the test and reference samples.

### 1.7.3.1. Transcript profiling

Microarray-based monitoring of gene expression was first reported for *Arabidopsis thaliana* in 1995 (Schena *et al.*, 1995). Since then transcriptional profiling has become the most widely used application of microarray technology, and has been used to study gene expression in a number of different organisms in numerous experimental systems. For example, it has been used to study gene expression patterns during normal development (White *et al.*, 1999) and disease (Shipp *et al.*, 2002). Transcriptional



profiling can also be used to identify genes which display significant changes in gene expression upon inactivation of a particular transcription factor (Young, 2000). Once a list of perturbed genes is compiled, it allows you to search for sequence motifs present in their upstream/downstream regions and may result in the discovery of novel TFBSs. Several programs are available for motif discovery such as MEME (Bailey and Elkan, 1995), AlignACE (Hughes *et al.*, 2000), and NestedMICA (Down and Hubbard, 2005). However, the number of direct targets identified by this method can vary depending on TF redundancy, and developmental stage being studied. Furthermore, the expression of many genes may be perturbed due to secondary effects observed upon inactivating a particular factor.

#### **1.7.3.2. Replication timing**

It has been known for some time that different parts of the genome replicate at different times during S phase. Actively transcribed regions replicate early during S phase while inactive regions replicate late in S phase. Microarray-based assays have been used to analyse replication timing in yeast (Raghuraman *et al.*, 2001), *Drosophila* (Schubeler *et al.*, 2002) and humans (Woodfine *et al.*, 2004; White *et al.*, 2004). Woodfine *et al.* adapted the technique of comparative genomic hybridization to assess replication timing in human cells. S-phase cells were isolated from asynchronously growing cells; the DNA was extracted and labelled. This was hybridized with DNA isolated from G1-phase cells and relative fluorescence at each array spot can be used to infer replication timing. The earlier a locus replicates the more DNA content it will have relative to the same locus in a G1 cell. In theory, an early replicating locus will have a copy number ratio of 2:1, whereas a late replicating locus will have a 1:1 ratio and intermediate replicating loci will be in between. A correlation between early replication and high gene density, high GC content, low SINE repeat content and high transcriptional activity was observed.

White and colleagues used a similar method to analyse replication timing in two different cell lines, but instead of comparing S1:G1 ratios, they isolated DNA from early and late S phase and compared these ratios. They concluded that early replication and transcriptional activity are often correlated but found that genes differentially transcribed

between the two cell lines were replicated at the same time. This implies that cell-specific transcription does not alter replication timing but global changes in chromatin architecture may be needed. A study by Gilbert and colleagues (2004) used random fragmentation of chromatin followed by sucrose centrifugation to separate ‘open’ and ‘closed’ chromatin fragments based on mass and density. Replication timing was also assessed and the authors suggested that there was a link between open chromatin and early replication.

### **1.7.3.3. DNase I hypersensitive site microarrays**

Traditional methods used for mapping of DNase I hypersensitive sites (HSs) rely on laborious techniques (see section 1.7.1.2) and can only be applied to study small genomic regions in a single experiment (Cockerill, 2000). To circumvent these problems, several protocols have recently been developed which allow DNase I HSs to be mapped in a high-throughput manner (Crawford *et al.*, 2004, 2006, 2006 b; Sabo *et al.*, 2004, 2006; Follows *et al.*, 2006). A number of studies have used cloning of DNase I HSs coupled with large-scale sequencing (Crawford *et al.*, 2004, 2006 b; Sabo *et al.*, 2004). Crawford and colleagues developed a method in which nuclei that had been digested with DNase I were blunted ended by T4 DNA polymerase. Blunted ended DNA was digested with restriction enzymes and cloned into a vector for sequencing. Sabo and colleagues (2004) attached a biotinylated linker to DNA that had been exposed to DNase I and then cut with a restriction enzyme. The DNase I cut fragments were then captured using streptavidin beads, a second linker was attached and the DNA fragments were amplified and cloned for sequencing. This method was then used in combination with tiled microarrays to identify DNase I HSs (Crawford *et al.*, 2006, Follows *et al.*, 2006). Sabo and colleagues developed a novel method to isolate DNase I HSs which were then mapped using microarrays (Sabo *et al.*, 2006). This method isolated DNA fragments associated with two DNase I cuts that occurred in close proximity (<1200 bp). These short fragments were isolated by size fractionation on a sucrose gradient. Chromatin and equally sized non-chromatin fragments were then labelled and hybridized to a microarray to identify DNase I HSs.

#### **1.7.3.4. Matrix attachment regions microarrays**

Eukaryotic chromatin is organized into loops by attachment to a chromosome scaffold or matrix (Mirkovitch et al., 1984). The DNA and proteins associated with this nuclear scaffold/matrix can be isolated by extraction of histones with high salt or mild detergent followed by restriction enzyme treatment to remove all DNA except matrix attached DNA. The AT-rich DNA segments that mediate attachment of chromatin to the nuclear matrix are known as matrix attachment regions (MARs) and occur on average every 50-200kb in the human genome (Bode, 2000). PML and SATB1 have recently been identified as MAR-binding proteins that regulate transcription by orchestrating chromatin loop formation (Kumar *et al.*, 2007). Sumer and colleagues isolated MAR DNA and hybridized it to a BAC/PAC array to define a 2.5 Mb region of MAR enriched chromatin at a human neocentromere (Sumer *et al.*, 2003). Ioudinkova and colleagues have also used arrays to map MARs at the chicken  $\alpha$ -globin domain (Ioudinkova *et al.*, 2005) suggesting that high-resolution microarrays could be used to map MAR sites throughout the human genome.

#### **1.7.3.5. Chromatin immunoprecipitation microarrays (ChIP-chip)**

Another recent application of microarray technology has been to study chromatin structure and function. DNA microarrays in combination with chromatin immunoprecipitation (ChIP) have been used to investigate the *in vivo* interactions of transcription factors or other regulatory complexes with genomic DNA. The use of ChIP in combination with microarrays has been termed ChIP-on-chip or ChIP-chip.

ChIP is one of the most powerful and widely used techniques for investigating *in vivo* DNA-protein interactions as these events are cross-linked in the native chromatin environment. Solomon and Varshavsky (1985) pioneered the development of a ChIP procedure and since then ChIP has been used in organisms ranging from yeast to human cells. The ChIP procedure is typically performed by cross-linking DNA-protein interactions using formaldehyde. The chromatin is then extracted by lysing the cells and nuclei. The chromatin is then sonicated into sheared fragments of approximately 300 bp to 1000 bp in size. The cross-linked protein-DNA complexes of interest are then

immunoprecipitated with a specific antibody, the cross-links are reversed and the enriched ChIP DNA is recovered. DNA that has not been immunoprecipitated or immunoprecipitated with a mock-antibody is used as a reference. Both the ChIP DNA and reference DNA can be fluorescently labelled and hybridized to a DNA microarray to identify the *in vivo* interactions of regulatory proteins with DNA.

ChIP-chip was first pioneered for the study of yeast transcription factors (Ren *et al.*, 2000; Iyer *et al.*, 2001; Wyrick *et al.*, 2001; Damelin *et al.*, 2002). The ChIP-chip method was subsequently used to study chromatin structure and function in yeast (Robyr *et al.*, 2002, Bernstein *et al.*, 2002; Nagy *et al.*, 2003; Robert *et al.*, 2004; Kurdistani *et al.*, 2004; Pokholok *et al.*, 2005; Liu *et al.*, 2005; Lee *et al.*, 2004; Bernstein *et al.*, 2004; Yuan *et al.*, 2005) and has also been applied to study DNA-protein interactions in other genomes, including the human genome. When analysing larger genomes, two main approaches have been taken:

**Biased approach:** This approach uses arrays containing sub-sets of regulatory elements from across the genome such as promoter regions or CpG islands. Promoter arrays, have been used to identify E2F (Ren *et al.*, 2002), c-Myc (Li *et al.*, 2003), and HNF transcription factors (Odom *et al.*, 2004) binding sites in human cells. CpG island microarrays have also been used to identify c-Myc and E2F target genes (Mao *et al.*, 2003; Weinmann *et al.*, 2002; Wells *et al.*, 2002). However the disadvantage of these types of microarrays is that they are inherently biased for the regions of the genome selected to study. Promoter or CpG islands arrays represent a particular set of regulatory elements, so their use in ChIP-chip is restricted to associating function with these elements.

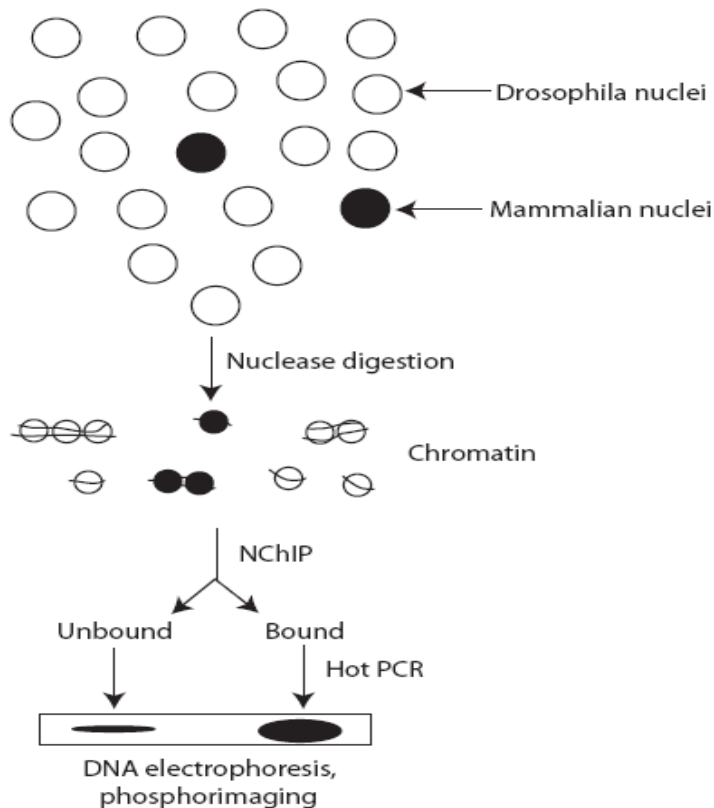
**Unbiased approach:** The unbiased approach uses arrays containing entire genomic regions in the form of tiling paths of DNA sequences. Tiling arrays were first used in a mammalian system to map GATA-1 binding sites across the human  $\beta$ -globin locus (Horak *et al.*, 2002). Entire chromosomal tiling paths of chromosome 21 and 22 (Martone *et al.*, 2003; Euskirchen *et al.*, 2004; Cawley *et al.*, 2004) have also been constructed, which allowed for the unbiased identification of NF $\kappa$ B (Martone *et al.*, 2003), CREB (Euskirchen *et al.*, 2004), Sp1, c-myc, and p53 binding sites (Cawley *et al.*, 2004), and estrogen receptor targets (Carroll *et al.*, 2005). Many of the binding sites mapped to 3'

ends of genes and within introns, which would have been missed by promoter or CpG arrays. Similarly, tiling microarrays covering the entire human genome have been used to identify active core promoters across the entire human genome in human fibroblast cells (Kim *et al.*, 2005). Genome-wide tiling arrays have also been used to identify target sites for the transcription factor p63 (Yang *et al.*, 2006), estrogen receptor binding sites (Carroll *et al.*, 2006) and the insulator binding protein CTCF (Kim *et al.*, 2007), amongst many other examples. Thus, tiling path microarrays can be used to comprehensively map DNA-protein interactions across genomes in an unbiased way.

Despite the rapid advances in generating large datasets, there are several disadvantages associated with current ChIP-chip methods, which limit its application. Firstly, efficiency of the ChIP reaction depends on antibody quality and epitope accessibility and formaldehyde fixation may introduce biases by ‘masking’ epitopes of chromatin proteins. Alternate techniques such as N-ChIP, biotin-tag affinity purification, or DamID can overcome these problems (Mito *et al.*, 2005; O’Neill and Turner, 2003; Van Steensel *et al.*, 2000). The N-ChIP method uses native or uncross-linked chromatin and offers a major advantage in terms of antibody specificity as epitopes that are recognised by antibodies can be disrupted by formaldehyde cross-linking (O’Neill and Turner, 2003). However, N-ChIP can only be used to investigate histone proteins as the majority of non-histone proteins are not retained on the DNA during nuclease digestion. Biotin-tag affinity purification has been used to map histone variants by fusing a biotin ligase recognition peptide to the histone H3.3 protein and streptavidin pull-down achieves high-specificity (Mito *et al.*, 2005). DamID maps DNA binding proteins by fusing a protein of interest to DNA adenine methylase, which then methylates adenine bases at binding sites. These sites are then identified by digestion with adenine methylation sensitive restriction enzymes (Van Steensel *et al.*, 2000).

Secondly, because of the small DNA yields obtained after a ChIP reaction, immunoprecipitated DNAs are usually PCR amplified (Horak *et al.*, 2002). This may result in amplification bias and an increase in false positives and false negatives. Alternatively, many sample DNAs are pooled (Weinmann *et al.*, 2002), before being labelled with fluorescent cyanine-conjugated dyes and hybridized to DNA microarrays. Thirdly, the number of cells needed for a ChIP-chip assay is somewhere between  $10^7$  and

$10^8$  for a single assay. This constraint prevents the analysis of cell populations where cell numbers are limited or rare - for example, cells found in the early stages of embryonic development. A modified ChIP method has recently been developed, which allows histone modifications to be studied from as few as 100 mouse embryonic stem cells (O'Neill *et al.*, 2006). This carrier ChIP (CChIP) procedure involves mixing a large number ( $5 \times 10^7$ ) of *Drosophila* cells with a small number ( $10^2 - 10^3$ ) of mammalian cells before preparing nuclei and chromatin (Figure 1.9). Native chromatin fragments were prepared by nuclease digestion and immunoprecipitated with an antibody to a histone modification. Mammalian DNA fragments were quantified by radioactive PCR, electrophoresis and phosphorimaging. The presence of a large excess of *Drosophila* DNA in the ChIP DNA samples may prevent this method from being used in combination with microarray hybridization to identify interactions in a high-throughput manner as fluorescently labelled nucleotides would be preferentially incorporated into *Drosophila* DNA samples during the labelling process.



**Figure 1.9: Outline of the carrier ChIP (CChIP) method.** In the CChIP method,  $5 \times 10^7$  *Drosophila* SL2 cells are mixed with a small number of mammalian cells ( $10^2$ - $10^3$ ) and nuclei are prepared. Nuclease digested chromatin is then prepared and immunoprecipitated with an antibody to a specific histone modification. Mammalian antibody bound and unbound fractions are then quantified by radioactive PCR, electrophoresis and phosphorimaging. Bound/unbound ratios are used to represent histone modification levels.

#### 1.7.4. Other ChIP-based methods

ChIP has also been combined with sequencing to determine the location of protein-DNA interactions. Sequencing can be performed from individually cloned ChIP fragments (Weinmann *et al.*, 2001), from cloned concatenations of single tags, where each tag is a signature from a ChIP DNA (ChIP-STAGE). ChIP-STAGE has been used to map several histone modifications (Roh *et al.*, 2005, 2006) and transcription factor binding sites (Impey *et al.*, 2004). ChIP in combination with sequencing of concatenated paired-end ditags (ChIP-PET) has also been used to map transcription factor target sites in the human genome (Wei *et al.*, 2006; Loh *et al.*, 2006). In both ChIP-PET and ChIP-STAGE methods, relative tag representation is used to calculate binding enrichment.

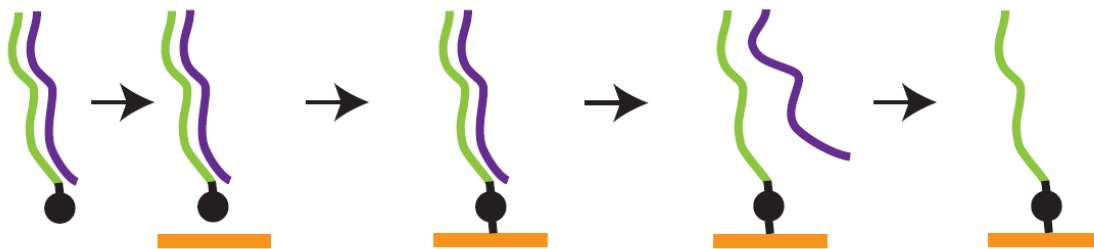
Advances in sequencing technology have seen ChIP combined with massively parallel sequencing (ChIP-Seq) to identify various protein-DNA interactions across the entire human genome. Unlike ChIP-STAGE or ChIP-PET, the ChIP-Seq method does not involve plasmid library construction as the very large number of short sequence reads produced by sequencing allows for the direct quantification of all DNA sequences present in a ChIP sample. Barski and colleagues recently generated high-resolution maps for the genome-wide distribution of 20 histone lysine and arginine methylation states as well as histone variant H2A.Z, RNA polymerase II, and CTCF across the human genome using massively parallel Solexa sequencing technology (Barski *et al.*, 2007). This technology attaches randomly fragmented ChIP DNAs to an optically transparent surface, followed by solid-phase amplification of DNAs to create more than 10 million clusters which are then sequenced using four colour sequencing by synthesis technology. Short sequence reads are then aligned against the reference genome sequence to calculate relative enrichment levels in a ChIP sample. This method has also been used to map REST

(Johnson *et al.*, 2007) and STAT1 (Robertson *et al.*, 2007) transcription factor binding sites in the human genome.

## 1.8. Genomic microarray platforms used in this study

### 1.8.1. The SCL genomic tiling path microarray

The development of ChIP-chip technology was used at the Sanger Institute to investigate regulatory elements at the SCL locus (Dhami, PhD Thesis, University of Cambridge, 2005; Dhami, submitted). The Stem Cell Leukemia (SCL) gene (also known as TAL1) is a basic helix-loop-helix transcription factor (TF) that is considered to be a master regulator of haematopoiesis (Robb and Begley, 1997). Over-expression of the SCL gene is the most common molecular abnormality found in human acute T-cell leukaemia and this TF is required for the normal development of all adult haematopoietic lineages. A tiling-path microarray was constructed to understand the regulation of SCL during haematopoiesis. The construction of a sensitive array platform for the SCL locus was made possible by using the 5'-aminolink array surface chemistry developed at the Sanger Institute. This surface chemistry allowed for single-stranded DNA molecules (derived from double-stranded PCR products) to be retained on the surface of a glass slide (Dhami *et al.*, 2005). A 5'-(C6) amino-link modification is incorporated at the end of one strand of DNA, which allows the modified strand to be covalently attached to the surface of the slide (Figure 1.10). During slide processing, chemical and physical denaturation removes the unmodified strand, while the strand attached to the slide is preserved. The single-stranded DNA molecules provide an ideal hybridisation target for a labelled DNA sample.



**Figure 1.10: Microarray surface chemistry.** Double-stranded PCR products (denoted by green and purple strands) containing a 5'-(C6) amino linker on one strand (black circle) are arrayed onto the surface



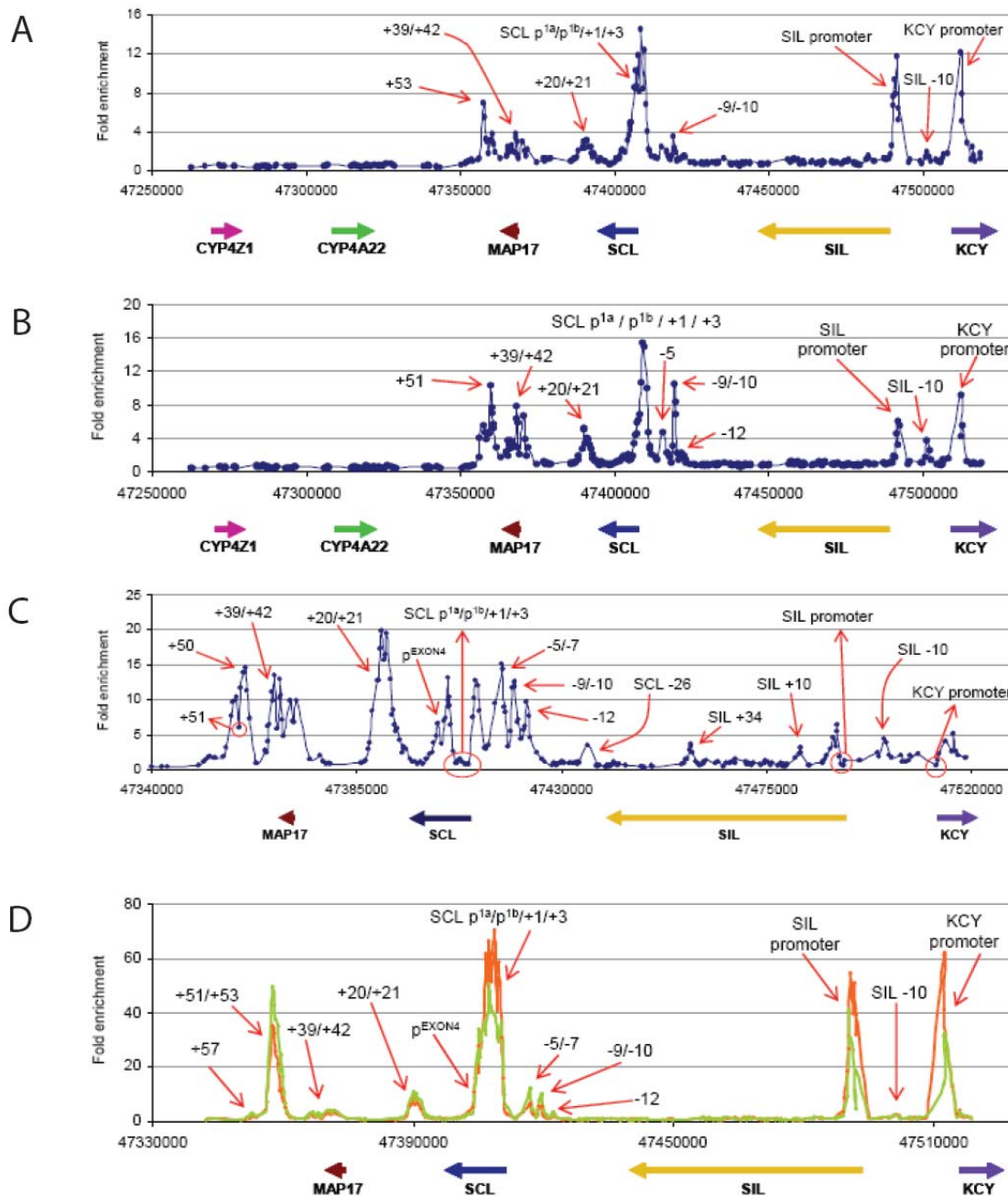
of the slide (orange rectangle). Covalent attachment of the PCR products occurs via the 5'-amino-link and the surface of the amine coated slide. Denaturation removes the strand that is not covalently attached to the slide surface (purple strand), resulting in a single-stranded DNA probe.

The genomic regions represented on the SCL tiling path array included the SCL gene, flanked upstream by SIL and KCY and downstream by MAP17, CYP4A22 and CYP4Z1 genes (Figure 1.11). The tiling path array covered 256 kb of human chromosome 1, with 419 PCR amplicons designed at an average product size of 458 bp.



**Figure 1.11: The genomic region included on the human SCL tiling path microarray.** The genomic region contained on the SCL tiling path microarray is indicated by double-headed arrows. This 256 kb region contains 6 genes, represented by coloured arrows. The tiling path covers the entire length of CYP4A22, MAP17, SCL, and SIL genes, while CYP4Z1 and KCY genes are partially covered. The gene order and direction of transcription is indicated by coloured arrows. The region is annotated on the negative strand of chromosome 1, and the orientation, with respect to the 1p telomere and centromere, is indicated by black arrows.

The SCL tiling path array and ChIP assays were used to identify a number of protein-DNA interactions which defined regulatory elements in the locus (Dhami, PhD thesis, University of Cambridge, 2005; Dhami, submitted). ChIP-chip assays were used to detect regions enriched for H3 K9/K14 diacetylation (H3 acetylation) and H4 K5/8/12/16 tetra-acetylation (H4 acetylation) at the SCL locus (Figure 1.12). It was observed that the most prominent enrichments for H3 acetylation and H4 acetylation were located at known and novel promoters. Lower enrichments were also detected at known and novel enhancer elements. The methylation status of histone H3K4 was also investigated. H3K4me1 was found to be enriched at known and novel enhancers (Figure 1.12). It was also shown that H3K4me2 and H3K4me3 occurred at the promoters of transcriptionally active genes across the SCL locus.



**Figure 1.12. Histone acetylation and methylation states define regulatory elements at the SCL locus.** Panels A and B show the histone H3 acetylation and H4 acetylation profiles across the SCL locus in K562 cells. Panel C shows the H3K4me1 profile across the MAP17, SCL, SIL and KCY genes in K562 cells. Panel D shows the H3K4me2 (green) and H3K4me3 (orange) profiles across the MAP17, SCL, SIL and KCY genes in K562 cells. The location of promoter and other regulatory elements (many of which were already known) are indicated by red arrows. The x-axes represent genomic coordinates along human

chromosome 1 and fold enrichments are displayed on the y-axes. The regulatory elements are denoted based on their distance upstream (-) or downstream (+) in kilobases from the SCL promoter 1a. The coloured arrows below each panel represent the gene order and direction of transcription. Figure from Dhimi 2005 thesis.

Those previously known regulatory elements detected by these ChIP-chip assays included (names based on distances upstream (-) or downstream (+) in kilobases from SCL promoter 1a):

- (i) **Promoters:** SCL has three promoters p1a, p1b (Aplan *et al.*, 1990) and pEXON4 (Bernard *et al.*, 1992). p1a and p1b are active in erythroid and megakaryocytic lineages while pEXON4 is active in leukaemic T-cells.
- (ii) **Stem cell enhancer:** The stem cell enhancer is located at +20/+21 and directs SCL expression to most haematopoietic progenitors and endothelium (Sanchez *et al.*, 1999; Sanchez *et al.*, 2001; Gottgens *et al.*, 2001; Pimanda *et al.*, 2006).
- (iii) **Neural regulators:** Regulatory elements located at SCL p1a, +1, and +3 direct SCL expression to regions within the brain and spinal cord (Sinclair *et al.*, 1999).
- (iv) **Erythroid enhancer:** The +51 erythroid enhancer targets SCL expression to primitive erythroblasts (Delabesse *et al.*, 2005).
- (v) **-9/-10 enhancer:** The -9/-10 region shows enhancer activity in reporter assays (Gottgens *et al.*, 1997).

## 1.8.2. The ENCODE project

### 1.8.2.1. A summary of the pilot phase findings

In 2003, an international consortium of research groups established a pilot study to evaluate a number of experimental approaches to catalogue all functional elements in 30 Mb (1%) of the human genome, comprised of 44 distinct genomic regions (The ENCODE Project Consortium, 2004) (see Chapter 3). The goal of this project was to develop efficient approaches for the large-scale identification and characterisation of regulatory elements, with the expectation of adopting these methods to analyse the whole genome. Thirty of the 44 regions were randomly picked by the ENCODE consortium to satisfy various non-exonic conservation and gene-density rates across the genome. The

remaining 14 regions were manually chosen because of their important biological or pathological role - and extensive regulatory information already exists for some of these regions. These regions include the CFTR locus, the interleukin cluster, the  $\alpha$ - and  $\beta$ -globin loci, the HOXA cluster and the IGF2/H19 imprinted region. The inclusion of these regions allowed for data obtained from the ENCODE project to be validated with respect to previously characterised regulatory elements. Methodologies used by groups in the consortium included using tiling microarrays to identify transcribed regions (Emanuelsson *et al.*, 2007), high-throughput mapping of DNase I hypersensitive sites (Sabo *et al.*, 2006; Crawford *et al.*, 2006), comparative sequence analysis (Margulies *et al.*, 2007; King *et al.*, 2007), computational analysis (Greenbaum *et al.*, 2007; Bajic *et al.*, 2006; Zheng *et al.*, 2007) replication timing assays (Karnani *et al.*, 2007), and ChIP-chip assays to detect histone modifications (Koch *et al.*, 2007; Rada-Iglesias *et al.*, 2007) and sequence-specific transcription factors (Bieda *et al.*, 2006). Over 200 data sets were generated by the consortium members and analysed (Birney *et al.*, 2007). The principle findings of this analysis are summarised as follows:

- (i) The majority of the human genome sequence is transcribed.
- (ii) Many non-coding transcripts were identified, many of which overlapped with coding regions.
- (iii) Many novel transcription start sites were identified, many of which were associated with a chromatin structure similar to well characterised promoters.
- (iv) Chromatin accessibility and histone modification patterns can be used to accurately predict the location and activity of transcription start sites.
- (v) Distal sites are associated with a characteristic histone modification pattern
- (vi) Replication timing correlates with chromatin structure.
- (vii) The majority of evolutionarily constrained sequences are associated with an experimentally determined function while many other functional elements are not under evolutionary constraint.

### **1.8.2.2. The Sanger Institute ENCODE Microarray**

Identification of regulatory elements in the ENCODE regions at the Sanger Institute focused on using ChIP-chip assays to detect a wide range of DNA-protein interactions. An array containing the 44 regions was constructed at the Sanger Institute (Koch *et al.*, 2007). Double-stranded PCR products were spotted on microarrays using the same 5'-aminolink array surface chemistry used to construct the SCL tiling path array, which was then processed to generate single-stranded DNA probes. The Sanger Institute ENCODE microarray consisted of 24,005 PCR fragments with an average size of 1024 bp (average non-overlapping tile length = 992 bp). The array covered approximately 80% of the targeted regions and over 90% of non-repetitive regions. The Sanger Institute ENCODE array provides a new resource for investigators interested in identifying functional elements, and formed the basis for much of the work presented in this thesis.

## **1.9. Aims of this thesis**

At the time this PhD project was initiated, there was relatively little information known about non-coding regulatory elements across the human genome. Furthermore, given some of the limitations of high-throughput approaches such as ChIP-chip (discussed in section 1.7.3.5), it was necessary to improve existing methods in order to identify and characterise non-coding elements in a systematic way. Therefore, with these views in mind, the aims of this thesis were as follows:

1. To use existing ChIP-chip approaches to characterise a variety of types of regulatory elements (promoters, enhancers and insulators) across selected regions of the human genome.
2. To develop further existing ChIP-chip approaches in order to improve sensitivity of the method when using cell types which are limiting in number.
3. Having improved ChIP-chip for aim 2, to then apply these methods to study cell types which are limiting in number.
4. To analyse the ChIP-chip data obtained for non-coding regulatory elements in the human genome and thereby understand fundamental principles of gene regulation.

## Chapter 2

### Materials and Methods

#### Materials

##### 2.1. Composition of solutions

Sterile HPLC-grade water (BDH) was used to prepare all solutions.

##### 10 X PCR buffer

- 500 mM KCl
- 50 mM Tris pH 8.5
- 25 mM MgCl<sub>2</sub>

##### 10 mM dNTP mix for PCR

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

##### 10 X dNTP mix used in DNA labeling

The following mix was used in the labeling reactions that were used with microarray hybridizations set-up using a Tecan HS 4800 hybridization station

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

##### 20 X SSC

The following were dissolved in 800 ml water:

- 175.3 g NaCl
- 88.2 g Sodium citrate

The volume was adjusted to 1000 ml and the pH adjusted to 7.0

##### Tecan HS 4800 hybridization station buffer

- 50% formamide (Fluka)

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

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

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

- 7.33 g NaCl
- 2.36 g Na<sub>2</sub>HPO<sub>4</sub>
- 1.52 g NaH<sub>2</sub>PO<sub>4</sub>H<sub>2</sub>O
- 500 µl Tween 20

**Solutions for ChIP:**

Cell lysis buffer (CLB)

- 10 mM Tris-HCl pH 8.0
- 10 mM NaCl
- 0.2% NP-40 (Igepal)
- 10mM Sodium butyrate
- 50 µg/ml PMSF
- 1 µg/ml leupeptin

Nuclear lysis buffer (NLB)

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

- 1 µg/ml leupeptin

#### IP dilution buffer (IPDB)

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

#### IP wash buffer 1 (IPWB1)

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

#### IP wash buffer 2 (IPWB2)

- 10 mM Tris-HCl pH 8.1
- 250 mM LiCl
- 1 mM EDTA
- 1% NP-40
- 1% deoxycholic acid

#### IP elution buffer (IPEB)

- 100 mM NaHCO<sub>3</sub>
- 1% SDS



### TE (pH 8.0)

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

### 1 X PBS

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

- 8 g NaCl
- 0.2 g KCl
- 1.44 g Na<sub>2</sub>PO<sub>4</sub>
- 0.24 g KH<sub>2</sub>PO<sub>4</sub>

### **Solutions for western blotting procedure:**

#### Nuclear protein extraction buffer

- 20mM HEPES pH7.9
- 0.2M EDTA
- 25% Glycerol
- 1.5mM MgCl<sub>2</sub>
- 0.42M NaCl

#### 4 x NuPAGE® LDS sample buffer

5 ml of 4 x NuPAGE® LDS sample buffer was prepared by dissolving the following in 4 ml of water:

0.003 g EDTA

2 g glycerol

0.4 g LDS

0.33 g Tris HCl

0.34 g Tris base

0.375 ml 1% SERVA Blue G250 solution

0.125 ml 1% phenol red solution

The volume was adjusted to 5 ml with water and stored at 4°C.

#### 20 X NuPAGE® MOPS SDS running buffer

A 20X stock solution was prepared by dissolving the following in 800 ml of water:

- 209.2 g MOPS
- 121.2 g Tris base
- 20 g SDS
- 6.0 g EDTA

The volume was adjusted to 1000 ml with water and stored at 4°C. This buffer was diluted to 1 X with water for electrophoresis.

#### 20 X NuPAGE® transfer buffer

250 ml of 20 X transfer buffer was prepared by dissolving the following in 200 ml water:

- 20.4 g bicine
- 26.2 g Bis-Tris
- 1.5 g EDTA

The volume was adjusted to 250 ml with water and stored at 4°C. This buffer was diluted to 1 X with water for western transfer.

1 litre of 1 X NuPAGE transfer buffer was prepared as follows and stored at 4°C:

- 50 ml 20 X NuPAGE transfer buffer
- 850 ml HPLC water
- 100 ml methanol

#### Blocking buffer

The following were dissolved in 90 ml water and stored at 4°C.

- 10 ml 10 X TBS
- 5 g non-fat dry milk
- 100 µl Tween-20

### 10 X TBS buffer

The following were dissolved in 1 litre of HPLC water, the pH was adjusted to 7.6 and the solution was stored at 4°C.

- 24.4 g Tris base
- 80 g NaCl

### 1 X TBST buffer

The following were dissolved in 900 ml of HPLC water:

- 100 ml 10 X TBS
- 1 ml Tween-20

## 2.2. Reagents

### Antibodies

Factor	Supplier	Catalogue number
H2B unmodified	Abcam	ab1790
H3 unmodified	Abcam	ab1791
H3ac	Upstate Biotechnology	06-599
H3K4me1	Abcam	ab8895
H3K4me2	Abcam	ab7766
H3K4me3	Abcam	ab8580
H3K9ac	Upstate Biotechnology	07-352
H3K9me1	Abcam	ab9045
H3K9me2	Upstate Biotechnology	07-212
H3K9me3	Upstate Biotechnology	07-523
H3K9ac	Upstate Biotechnology	07-352
H3K18ac	Upstate Biotechnology	07-354
H3K27ac	Upstate Biotechnology	07-360
H3K27me1	Upstate Biotechnology	07-448
H3K27me2	Abcam	ab24684
H3K27me3	Upstate Biotechnology	07-449
H3K36me1	Abcam	ab9048
H3K36me2	Upstate Biotechnology	07-274
H3K36me3	Abcam	ab9050
H3K79me1	Abcam	ab2886
H3K79me2	Abcam	ab3594
H3K79me3	Abcam	ab2621
H4K5ac	Abcam	ab1758
H4K8ac	Abcam	ab1760
H4K16ac	Abcam	ab1762
CTCF	Santa Cruz Biotechnology	sc15914x
mSin3a	Abcam	ab3479
USF1	Santa Cruz Biotechnology	sc229x
USF2	Santa Cruz Biotechnology	sc862x

### Enzymes

- Proteinase K,  $\geq 20$  units/mg (GibcoBRL)
- RNase A,  $\geq 50$  Kunitz units/mg (ICN Biochemicals)
- Taq polymerase, 5 units/ $\mu$ l (Perkin Elmer-Cetus)
- Klenow fragment, 40 units/ $\mu$ l (Invitrogen)

### Fluorophores

- Cy3-dCTP (GE Healthcare)
- Cy5-dCTP (GE Healthcare)

### Primer pairs

- Sequences of primer pairs used to construct the SCL tiling path microarray are available in the PhD thesis of Dr. Pawandeep Dhani (University of Cambridge, 2005). Primer pairs used to amplify PCR products for the ENCODE array are available at <ftp://ftp.sanger.ac.uk/pub/encode/microarrays/>
- Sequences of primer pairs used for quantitative SYBR green PCR are available in Appendix 1.

### Other reagents

- Human C<sub>0</sub>t 1 DNA (Invitrogen)
- Herring sperm DNA (Sigma)
- Phenol solution (Sigma)
- Chloroform solution (BDH)

## **2.3. Cells and cell lines**

The human cell lines K562 (Lozzio and Lozzio, 1979) and U937 (Sundstrom and Nilsson, 1976) were obtained from Dr. Pawandeep Dhani (Wellcome Trust Sanger Institute). K562 cells were established from a patient with myelogenous leukaemia in which myelo-proliferation was of the erythrocyte precursors (erythroleukemia). The U937 cell line was established from a patient with diffuse histiocytic lymphoma and

displays many monocytic characteristics. Human CD14<sup>+</sup> monocytes were isolated from peripheral blood mononuclear cell samples using Stem Cell Technologies RoboSep CD14 beads by Nicola Foad and obtained courtesy of Dr. Willem Ouwehand (Department of Haematology, University of Cambridge). H9 Human Embryonic Stem Cells (Thomson *et al.*, 1998) were grown and sorted for expression of SSEA3 by Dr. Enrique Millan as described in section 2.4.1 and obtained courtesy of Prof. Roger Pedersen (Cambridge Institute for Medical Research).

## **Methods**

### **2.4. Tissue culture**

#### **2.4.1. Culturing of cell lines**

K562 and U937 cell lines were cultured in suspension in 50 ml of media (Sigma) with the appropriate amount of fetal bovine serum (GibcoBRL) and other supplements (Sigma) in 75 cm<sup>2</sup> tissue culture flasks with vented caps (Corning). K562 cells were cultured in DMEM, supplemented with 10% fetal calf serum, 1% penicillin-streptomycin, and 2mM L-glutamine, while U937 cells were cultured in RPMI, supplemented with 10% fetal calf serum, 1% penicillin-streptomycin, and 2mM L-glutamine. K562 and U937 cells were cultured under 5% CO<sub>2</sub> at 37°C. Note: H9 Human ES cells were cultured in chemically defined medium as described previously (Vallier *et al.*, 2007). Differentiation of human embryonic stem cells was induced by embryoid body formation as described (Vallier *et al.*, 2007). Cells positive and negative for expression of stage specific embryonic antigen 3 (SSEA3) were isolated by fluorescence-activated cell sorting as described (Henderson *et al.*, 2002).

Once K562 or U937 cells reached confluency (i.e. 7-8 x 10<sup>5</sup> cells/ml of culture media), media was replenished and sub-culturing was carried out as follows:

1. 25 ml of fresh media was added to each flask and any clumps of cells were gently broken up using a syringe.
2. The culture was then distributed between three new 75 cm<sup>2</sup> flasks (Corning) and further 25 ml of fresh media was added to each flask, effecting a 1/3 dilution of the confluent starting culture.

3. The number of cells needed for standard chromatin immunoprecipitation (ChIP) experiments was 100 million. Therefore, culture volumes to obtain the required number of cells per flask were suitably scaled up in 175 cm<sup>2</sup> culture flasks with vented caps (Corning).

#### **2.4.2. Cell cryopreservation**

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

#### **2.4.3. Chromatin immunoprecipitation (ChIP)**

K562, U937, human monocytes and human ES cell line H9 were used for chromatin immunoprecipitation. Fresh K562 and U937 cultures were grown and used to prepare chromatin. Aliquots of K562 and U937 cells used for each ChIP experiment were subjected to flow-sorting (Cytomation MoFlo High Performance Cell Sorter, Dako Cytomation) in parallel for cell-cycle analysis (Performed by Bee Ling, Wellcome Trust Sanger Institute). For this,  $3 \times 10^6$  cells were harvested, washed with 10 ml of PBS and then fixed in 5 ml of 70% ethanol. An equal volume of Hoechst 33342 dye (2mg/ml) was added to samples, incubated at 37°C for 15 minutes and analysis of fluorescence was used to determine DNA content of the cells. This step was performed to ensure that the percentage of actively dividing cells was consistent for the preparation of each batch of chromatin.

Approximately  $1 \times 10^8$  K562 and U937 cells were harvested for each ChIP procedure. The number of human monocytes obtained from donors varied across three biological samples obtained (Table 2.1).  $3.49 \times 10^6$  SSEA3+ and  $2.89 \times 10^6$  SSEA3- Human embryonic stem cells were obtained for ChIP.



### Fixation

1. The cells were collected by centrifuging at 259 g for 8 minutes at room temperature and resuspended in 50 ml of serum free media in a glass flask.
2. DNA-protein and protein-protein interactions were cross-linked by adding formaldehyde (37%, BDH AnalaR). 500  $\mu$ l, 1010  $\mu$ l or 1355  $\mu$ l formaldehyde was added drop-wise to a final concentration of 0.37%, 0.75% for histone modifications and 1% for transcription factors respectively.
3. The cross-linking was carried out at room temperature with constant but gentle stirring for 10 minutes (for histone modifications) or 15 minutes (for transcription factors).
4. 3.15 ml, 3.41 ml or 3.425 ml (for 0.37%, 0.75% or 1% formaldehyde concentration respectively) of ice-cold 2M glycine was added to a final concentration of 0.125M with constant but gentle stirring for 5 minutes at room temperature to stop the cross-linking reaction.
5. Cells were transferred to 50 ml falcon tubes and kept on ice whenever possible. The cells were pelleted by centrifuging at 259 g for 6-8 minutes at 4°C and washed with 1.5 ml of ice-cold PBS.
6. After washing, the cells were pelleted at 720 g at 4°C for 5 minutes and the supernatant was removed.

### Cell and nuclei lysis

7. Cells were lysed by adding 1.5 x pellet volume of ice-cold cell lysis buffer (CLB). The cell pellets were gently resuspended and incubated on ice for 10 minutes.
8. The nuclei were recovered by centrifuging the samples at 1125 g for 5 minutes at 4°C.
9. After carefully removing the supernatant, the nuclei were lysed by resuspending the pellet in 1.2 ml of nuclei lysis buffer (NLB) and incubating on ice for 10 minutes. Note: 0.6 ml of NLB was used when preparing chromatin from human embryonic stem cells as pellet volumes were much smaller.

### Sonication

10. 0.72 ml of IP dilution buffer (IPDB) was added and the samples were transferred to 5 ml glass falcon tubes (Falcon 2058).
11. The chromatin was sonicated to reduce the DNA length to an average size of 600 bp using the Sanyo/MES Soniprep sonicator. The tip of the probe was dipped to reach approximately halfway down the total level of the liquid sample and the tube was kept constantly on ice (Conditions for sonication like number of bursts, length of bursts and power setting depend on the sonicator tip used). The settings used for the sonicator were:
  - Amplitude: 14 microns
  - Number of bursts: 8
  - Length of bursts: 30 seconds

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

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

### Immunoprecipitation

13. The supernatant was transferred to a 15 ml falcon tube and 4.1 ml of IPDB (NLB:IPDB ratio is 1:4) was added<sup>a</sup>. Note: 1.68 ml of IPDB was added to human embryonic stem cell chromatin preparations to maintain a 1:4 ratio of NLB:IPDB.
14. The chromatin was precleared by adding 100  $\mu$ l of normal rabbit IgG (Upstate Biotechnology). The samples were incubated for 1 hour at 4°C on a rotating wheel.
15. 200  $\mu$ l of homogeneous protein G-agarose suspension (Roche) was added to the precleared chromatin and the samples were incubated for 3-5 hours at 4°C on a rotating wheel.
16. The samples were centrifuged at 1620 g for 2 minutes at 4°C to pellet the protein G-agarose beads and the supernatant was used to set up various immunoprecipitation (IP) conditions in 2 ml microfuge tubes. [Note: the following applies when setting-up standard ChIP conditions with K562 and U937 cells (i.e.  $10^7$  cells and 5-10  $\mu$ g of

antibody are used)]. An aliquot of 270  $\mu\text{l}$  of chromatin was stored at  $-20^{\circ}\text{C}$  to be used as input sample for array hybridisations. An NLB:IPDB buffer at the ratio of 1:4 was freshly prepared and used to ensure that the final volume of all ChIP conditions was 1350  $\mu\text{l}$ . Experimental and control ChIP conditions were set up as follows:

- Normal species specific IgG control (matching the species from which antibody used in ChIP condition(s) was derived) – 675  $\mu\text{l}$  chromatin + 675  $\mu\text{l}$  NLB:IPDB buffer + 10  $\mu\text{g}$  species specific IgG
- No chromatin control – 1350 NLB:IPDB buffer
- No antibody control - 675  $\mu\text{l}$  chromatin + 675  $\mu\text{l}$  NLB:IPDB buffer
- ChIP conditions – 675  $\mu\text{l}$  chromatin + 675  $\mu\text{l}$  NLB:IPDB buffer + 5-10  $\mu\text{g}$ \* of antibody

(\*5-10  $\mu\text{g}$  for antibodies raised against histone modifications and 10  $\mu\text{g}$  for the antibodies raised against specific transcription factors).

16a. See Table 2.1 for alterations when setting up ChIP conditions with fewer numbers of cells.

Sample	Number of cells used in chromatin preparation	Final volume of chromatin preparation ( $\mu\text{l}$ )	Number of cells used per ChIP condition	Volume of chromatin used in ChIP ( $\mu\text{l}$ )	Volume of NLB:IPDB buffer ( $\mu\text{l}$ )	Antibody amount ( $\mu\text{g}$ )	Volume of chromatin used to prepare Input DNA ( $\mu\text{l}$ )
K562	$10^8$	6000	$10^6$	60	1290	0.1 - 10	270
K562	$10^8$	6000	$10^5$	6	1344	0.1 - 10	270
K562	$10^8$	6000	$10^4$	0.6	1350	0.1 - 10	270
Monocytes (BR 1)	$5.6 \times 10^7$	6000	$10^6$	100	1250	1	540
Monocytes (BR2)	$2.5 \times 10^7$	6000	$10^6$	200	1150	1	1000
Monocytes (BR3)	$5.0 \times 10^7$	6000	$10^6$	125	1225	1	540
hESCs: SSEA3+	$3.49 \times 10^6$	3000	$10^5$	86	1264	0.5	1000
hESCs: SSEA3-	$2.89 \times 10^6$	3000	$10^5$	104	1246	0.5	1000

**Table 2.1: Summary of chromatin, antibody and buffer amounts used in reduced cell number ChIP experiments.** BR= biological replicate, hESCs = human embryonic stem cells, SSEA3 = stage specific embryonic antigen 3.

17. The samples were incubated at 4°C overnight on a rotating wheel.
18. The samples were centrifuged at 18000 g for 5 minutes at 4°C and the samples were transferred to fresh 2 ml microfuge tubes. 50 µl of homogeneous protein G-agarose suspension was added to each sample and the samples were incubated at 4°C for at least 3 hours on a rotating wheel.
19. The samples were centrifuged at 6800 g for 30 seconds at 4°C to pellet the protein G-agarose beads.
20. The supernatant was removed and the protein G-agarose beads were carefully washed. For each wash, the wash buffer was added, the samples were vortexed briefly, were centrifuged at 6800 g for 2 minutes at 4°C and left to stand on ice for 1 minute before removing the supernatant. The washes were carried out in the following sequence:
  - a) The beads were washed twice with 750 µl of cold IP wash buffer 1. The beads were transferred to a 1.5 ml microfuge tube after the first wash.
  - b) The beads were washed once with 750 µl of cold IP wash buffer 2.
  - c) The beads were washed twice with 750 µl of cold TE pH 8.0.

#### Elution

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

### Reversal of cross-links

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

### Extraction of DNA

26.  $2\ \mu\text{l}$  of yeast tRNA (5 mg/ml, Invitrogen) was added to each sample just before adding  $250\ \mu\text{l}$  of phenol (Sigma) and  $250\ \mu\text{l}$  of chloroform<sup>c</sup>.
27. The samples were vortexed and centrifuged at 18000 g for 5 minutes at room temperature. The aqueous layer (top layer) was collected in fresh 1.5 ml microfuge tubes and  $500\ \mu\text{l}$  of chloroform was added to each sample.
28. The samples were vortexed and centrifuged at 18000 g for 5 minutes at room temperature. The aqueous layer was transferred to a fresh 2.0 ml microfuge tubes.
29.  $5\ \mu\text{g}$  of glycogen (5 mg/ml, Roche),  $1\ \mu\text{l}$  of yeast tRNA (5 mg/ml, Invitrogen) and  $50\ \mu\text{l}$  of 3M NaAc (pH 5.2) was added to each sample and mixed well. The DNA was precipitated with  $1375\ \mu\text{l}$  of 100% ethanol and incubating at  $-70^{\circ}\text{C}$  for 30 minutes (or  $-20^{\circ}\text{C}$  overnight).
30. The samples were centrifuged at 20800 g for 20 minutes at  $4^{\circ}\text{C}$ . The DNA pellets were washed with  $500\ \mu\text{l}$  of ice-cold 70% ethanol and air-dried for 10-15 minutes.
31. The DNA pellets of the IP samples were resuspended in  $50\ \mu\text{l}$  of sterile filtered HPLC water and  $100\ \mu\text{l}$  for the Input DNA samples.
32.  $5\ \mu\text{l}$  of each sample was run on a 1% agarose 1XTBE gel and visualised with ethidium bromide to check DNA size. Samples were stored at  $-20^{\circ}\text{C}$ .

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

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

<sup>c</sup>Safety Note – The phenol/chloroform steps were carried out in a fume cabinet.

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

## **2.5. Construction of the SCL tiling path microarray and the ENCODE tiling path microarray**

### **2.5.1. Generation of tiling path amplicons**

Primers pairs used to amplify PCR products for the SCL tiling array were designed from the relevant genomic sequence of chromosome 1 and PCR amplicons were generated as described (Dhami, PhD thesis, University of Cambridge, 2005). A 5'-(C6) amino-link was added to all forward primers to allow binding to Codelink slides (GE) (Dhami *et al.*, 2005; Dhami, submitted). Similarly primer pairs used to amplify PCR products for the ENCODE array were designed using Primer3 and PCR amplicons were generated as described (Koch *et al.*, 2007).

### **2.5.2. Microarray slide printing and processing**

Spotting buffer was added at final concentration of 0.25M sodium phosphate pH 8.5 and 0.00025% sodium sarkosyl (BDH) to PCR products prior to arraying. The PCR products were then filtered through multiscreen-GV 96 well filter plates (Millipore) and aliquotted into 384 well plates (Genetix). PCR products were then spotted onto Codelink slides in a 16 block format for the SCL array and a 48 block format for the ENCODE array using a Microgrid II robotic arrayer (Biorobotics/Genomics solutions). In order to generate microarrays containing single-stranded DNA elements, all PCR amplicons were printed on arrays and processed as described ([www.sanger.ac.uk/Projects/Microarrays/arraylab/methods.shtml](http://www.sanger.ac.uk/Projects/Microarrays/arraylab/methods.shtml)). Arrays were subject to quality control analysis by hybridization of random oligonucleotide sequence coupled to

Alexa-647 (Panomer-9, Invitrogen) according to manufacturer's instructions and visualised by scanning with a Scanarray 4000 confocal laser-based scanner (Perkin-Elmer). Arrays that passed this quality control step were stored at room temperature in a desiccated environment until ready for hybridisation.

## **2.6. Hybridisation of microarrays**

### **2.6.1. Labelling by random priming of DNA samples**

The SCL genomic tiling path array and ENCODE array hybridisations using DNA obtained by chromatin immunoprecipitation were set up using the Tecan HS 4800 hybridisation station (an automated hyb-station). The DNA was labeled using BioPrime Random Labeling Kit (Invitrogen) as described below.

#### **Labelling method used for Tecan HS 4800 hybridisation set-up**

1. The following reagents were mixed on ice in a 1.5 ml microfuge:

- 60 µl 2.5 X Random primer solution
- x µl DNA \*
- (70.5-x) µl sterile H<sub>2</sub>O

\* The DNA amount labeled was different for input and ChIP samples. The amount of DNA labeled was 40% of unamplified ChIP DNA and approximately 200 ng of Input DNA (2-5% of recovered Input DNA).

2. This mixture was heated at 100°C for 10 minutes to denature the DNA and then snap-chilled on ice. The following reagents were added to the tubes on ice:

- 15 µl 10 X dNTP mix
- 1.5 µl 1 mM Cy3/Cy5 labeled dCTP (1 mM Cy3-dCTP, 1 mM Cy5-dCTP, GE Healthcare). Input samples were labeled with Cy5 dCTP and ChIP samples were labeled with Cy3 dCTP
- 3 µl Klenow fragment (40 U/µl)

The final volume per labeling reaction was 150 µl.

3. The reagents were mixed gently but thoroughly and incubated at 37°C overnight.

4. 15 µl stop buffer was added to the reaction mix to terminate the reaction.

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

### **2.6.2. Purification of labelled DNA samples**

Labelled DNA samples were purified using the protocol described below.

1. Micro-spin G50 columns (GE Healthcare) were used to remove the unlabeled nucleotides from the labelled DNA samples.
2. Three columns were used for each of the 150  $\mu$ l labelling reactions.
3. The resin was resuspended in the columns by vortexing gently. The caps were loosened and the bottom of the tubes snapped off. The columns were placed in 2.0 ml microfuge tubes and centrifuged at 1700 g for 1 minute.
4. 50  $\mu$ l of sterile filtered HPLC water was applied to the resin-bed and the columns were centrifuged at 1700 g for 1 minute.
5. The columns were placed in fresh 1.5 ml microfuge tubes and the labelled DNA samples were carefully applied to the resin-bed. The columns were then centrifuged at 1700 g for 2 minutes.
6. The purified DNA samples were collected in the 1.5 ml microfuge tubes and the samples from the same labeled reaction were pooled together. The final volume for the labelled DNA samples was approximately 180  $\mu$ l.
7. 5  $\mu$ l of each labelled DNA was analyzed on a 1% agarose 1 X TBE gel and stained with ethidium bromide for visualization. The samples were used for hybridisation and stored in the dark at  $-20^{\circ}\text{C}$ .

### **2.6.3. Array hybridisation set-up using the Tecan HS 4800 hybridisation station**

Tecan HS 4800 hybridisation station (automated hyb-station) was used to set-up ChIP-chip hybridisations (on the SCL genomic tiling path array and ENCODE array). The Tecan HS 4800 is a fully automated hyb station where the microarray slides are loaded on the Tecan's slide holder and the hybridisation mix is agitated to ensure even hybridisation. The SCL array area was 2 x 2 cm and consequently the smaller chambers



were used on the Tecan to set up the hybridisations. However, larger chambers were required for the ENCODE microarrays

### **Preparation of the hyb station**

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

### **Preparation of pre-hybridisation and hybridisation solutions**

Note: The following volumes are for hybridisation of the SCL tiling path array.

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

#### **Tube 1 (Pre-hybridisation mix)**

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

#### **Tube 2 (Hybridisation mix)**

- 180 µl Cy3-labeled DNA
- 180 µl Cy5-labeled DNA
- 135 µl Human Cot 1 DNA (Invitrogen)
- 55 µl 3M NaAc (pH 5.2)
- 1200 µl 100% ice-cold ethanol

Note: The following volumes are for hybridisation of the ENCODE array.

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

**Tube 1 (Pre-hybridisation mix)**

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

**Tube 2 (Hybridisation mix x 2)**

- 180 µl Cy3-labeled DNA
- 180 µl Cy5-labeled DNA
- 135 µl Human Cot 1 DNA (Invitrogen)
- 55 µl 3M NaAc (pH 5.2)
- 1200 µl 100% ice-cold ethanol

5. The DNA samples were precipitated by incubating at  $-70^{\circ}\text{C}$  for 60 minutes (or  $-20^{\circ}\text{C}$  overnight) and then centrifuged at 18000 g for 20 minutes at room temperature. The pellets were washed with 500 µl of 80% ethanol and air dried.
6. The pre-hybridisation and hybridisation DNA pellets were resuspended in 120 µl of Tecan hyb buffer each (for hybridisation on the SCL array) and 3 µl of yeast tRNA (100 µg/µl, Invitrogen) was added to the hybridisation mix.
7. For hybridising the ENCODE arrays, the pre-hybridisation DNA pellet was resuspended in 180 µl of Tecan-hyb buffer. Each hybridisation pellet was resuspended in 90 µl of Tecan-hyb buffer and then combined to give a total volume of 180 µl. 3 µl of yeast tRNA (100 µg/µl, Invitrogen) was added to the hybridisation mix. Heating the solutions at  $70^{\circ}\text{C}$  helps to resuspend the pellets properly.

### **Array hybridisation and washing**

8. The pre-hybridisation and hybridisation solutions were denatured at 100°C for 10 minutes. The hybridisation solution was snap-chilled on ice and then pre-annealed at 37°C for 1 hour. The pre-hybridisation solution was kept at 70°C until applied to the slide.
9. After vortexing, using a positive displacement pipette, 100 µl of the pre-hybridisation solution (for SCL array) or 160 µl of the pre-hybridisation solution (for ENCODE array) was injected onto the slide very slowly and carefully to avoid any air bubbles.
10. The pre-hybridisation step was performed at 37°C for 1 hour. The slides were washed once with PBS/0.05% Tween and dried with short blasts of nitrogen gas.
11. 100 µl or 160 µl of the hybridisation solution was injected slowly onto the slide using a displacement pipette. The hybridisation step was performed at 37°C for 48 hours.
12. The slide washing was carried out on the Tecan which was programmed to perform the washes in the sequence listed in Table 2.2.

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

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

13. The slides were dried on the Tecan using nitrogen gas and stored in a dark, low temperature, low humidity environment until ready for scanning to prevent loss of fluorescent signal.

#### 2.6.4. Scanning and processing of ChIP-chip data

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

1. Cy3 and Cy5 images at 5  $\mu\text{m}$  resolution were acquired using the Scanarray 4000 confocal laser-based scanner (Perkin-Elmer) using a laser power of 100% and a photo multiplier tube (PMT) value of between 70%-85%.
2. ScanArray Express (Perkin Elmer) was used to quantitate the fluorescent intensities of the spots using the adaptive circle quantitation and the TOTAL normalisation methods. This software can automatically locate the spot position on the scanned image of the array to obtain the signal intensity values. Mean intensity ratios (intensity-background) were reported for each spot representing an array element. Those spots identified as not found were removed from the data set.
3. Further analysis of the ChIP-chip data was carried out in a Microsoft Excel spreadsheet in which each array element was associated with its genomic sequence position information.
4. The SCL microarray data was visualised by plotting the mean ratios of all array elements along the Y-axis and the respective genomic positions along the X-axis. ENCODE data was visualised in the UCSC genome browser (Kuhn *et al.*, 2007) by uploading a 'wiggle' file which contained chromosome start and end coordinates and fold enrichment ratios or a 'bed' file which contained the location of ChIPOTle peaks only.
5. The median ratio of technical or biological replicates performed using the ENCODE array were calculated using a script written in the R programming language (<http://www.sanger.ac.uk/PostGenomics/encode/data-access.shtml>).
6. The baseline levels of each SCL or ENCODE data set was normalised to a value of one, so that all the experiments could be directly compared from this baseline value. This was done by calculating the median ratio for each experiment and dividing all the ratios (obtained in that experiment) by this number.

## 2.7. Analysis of ChIP-chip data with respect to genomic features

The ChIPOTle program (Buck *et al.*, 2005) was used to define peaks of enrichment in ChIP-chip data sets by using a sliding window approach and then estimating the significance of enrichment for a genomic region using a standard Gaussian error function. ChIPOTle assigns a p-value to the average  $\log_2$  ratio within each window and are corrected for multiple comparisons using the Bonferroni correction. The significance p-value cut-off assigned for analysing histone H3 acetylation, H4 acetylation and H3K4 methylation data sets was p0.0005. A p-value cutoff that produces about 50 times more significant regions than significant negative regions was suggested to be a satisfactory cut-off for the majority of applications (Buck *et al.*, 2005). A window size of 2000bp and a step size of 500bp (which should be approximately  $\frac{1}{4}$  the window size) was used analyzing ENCODE histone modification data sets. A p-value cut-off of p0.00001 was used for the analysis of ENCODE transcription factor data sets (analysis performed by Dr. Rob Andrews, Wellcome Trust Sanger Institute).

GENCODE annotations were downloaded from GENCODE genes version 02.2 from [ftp://genome.imim.es/pub/projects/gencode/data/havana-encode/current/44regions/44regions\\_CHR\\_coord.gtf](ftp://genome.imim.es/pub/projects/gencode/data/havana-encode/current/44regions/44regions_CHR_coord.gtf). The distribution of histone modifications and transcription factors with respect to GENCODE annotated features such as transcription start sites was performed using scripts written in the R programming language (performed by Dr. Rob Andrews). All ChIP-chip data was analysed on NCBI human genome build 35 (hg17).

## 2.8. Statistical analysis

Statistical analysis of data was performed using Microsoft Excel. This included the calculation of standard deviations, mean coefficient of variations (CVs), and Pearson correlation coefficients. Standard deviations were calculated using the STDEV function, Pearson correlation coefficients were calculated using the PEARSON function. CVs were calculated by dividing the standard deviation of a set of values by the mean of those values and then expressed as percentage by multiplying by 100.

Significant enrichments in SCL array data were calculated by determining the mean background ratios of two regions which did not contain any known regulatory elements in K562 (Dhami, PhD thesis, Cambridge University, 2005). These regions spanned

47262287 bp to 47343557 bp, and 47424426 bp to 47489321 bp on chromosome 1. The significant enrichment threshold was three standard deviations above the mean background ratios. Significant enrichments in ENCODE array data sets were also calculated by determining the mean ratio of the entire data set and the significant enrichment threshold was three standard deviations above the mean ratio of the data set. Over-representation of ChIPOTle sites with respect to genomic features was performed using a randomisation strategy. Each experimental ChIPOTle dataset was randomised 100 times to generate 100 random data sets within the ENCODE regions, conserving the size and number of ChIPOTle sites. The mean overlap with genomic features was then calculated and compared to experimental values (performed by Dr. Rob Andrews).

## **2.9. Gene expression analysis**

Four replicate K562 Affymetrix U133 plus 2.0 gene expression data sets were provided by Dr. Christoph Koch and Dr. Philippe Couttet (Wellcome Trust Sanger Institute). All Affymetrix analyses were performed by Dr. Rob Andrews using the affy package in Bioconductor with default parameters (Gentleman *et al.*, 2004). Normalised MAS5 data was generated using the affy package, which was then used for Present/Absent calls per probe set. Robust Multichip Average (RMA) data analysis was also performed using the affy package, in which background correction and normalization was performed to give an expression value per probe set. The RMA expression values for ENCODE genes present on the Affymetrix array were ranked in order of expression as high (100-75%), low (75%-50%), indeterminate (50%-25%), and off (25%-0%). The RMA values of these four classes of genes were compared to MAS5 absent and present expression calls. Genes which were called as present by MAS5 and which were in the top 50% of ranked RMA values were considered as expressed genes (either as high or low expression based on the above criteria). Genes which were called by MAS5 as absent and which were in the bottom 50% of ranked RMA values were considered as not expressed. Genes which showed discrepancies between MAS5 and RMA data were classified as indeterminate.

Two replicate human CD14+ monocyte illumina human ref-8 expression beadchip data sets were provided by Dr. Nick Watkins (Department of Haematology, University of Cambridge). Illumina gene expression analyses were performed in Bioconductor and

BeadStudio packages. Present/absent analysis was performed using the lumi package within Bioconductor where present is called when the detection score provided by BeadStudio was greater than 0.95. Transcript abundance analysis was performed using the lumi package using quantile normalization (performed by Dr. Rob Andrews).

### **2.10. Receiver operator characteristic (ROC) curve analysis**

The association of histone modifications and transcriptions factors at TSSs with gene expression state was examined by plotting ROC curves for each of the histone modifications and transcription factors within 1kb around the TSS (analysis performed by Dr. Ulas Karaöz, Boston University). Present and absent MAS5 calls from 238 Affymetrix probe-sets were used to define the on/off state. The ROC curves illustrate the predictive accuracy of histone marks or the presence of a transcription factor on classifying the expression state of genes. A threshold was applied to each histone modification or transcription factor level and a prediction of the on (or off) state of a gene is made if the level is higher (or lower) than the threshold. The true positive rate is plotted against the false positive rate for each threshold and all possible thresholds are applied so that a curve is obtained (with each point of the curve corresponding to a threshold). The best operating point is the point on a ROC plot which lies on a 45 degree line closest to the northwest corner of the ROC plot.

### **2.11. DNA sequence motif analysis**

Motif matrices from JASPAR (Bryne *et al.*, 2007) and TRANSFAC (Matys *et al.*, 2006) databases were used to search for enrichment in transcription factor ChIPOTle sites. The DME program (Smith *et al.*, 2005) was also used to search for novel motifs in ChIPOTle sites. 1000bp centred sequences were extracted from ChIPOTle sites and defined as foreground sequences, i.e. sequences in which a binding motif was believed to be present. 1000 bp sequences flanking ChIPOTle sites were also extracted and defined as background sequences, i.e. sequences in which no specific binding motif was believed to be present. Motifs which distinguished foreground sequences from background sequences with a low relative error rate were identified. The sensitivity (Sn) and specificity (Sp) of these motifs was also calculated. The sensitivity associated with a motif and p-value cut-

off is the proportion of true foreground sequences that were classified as foreground; the specificity is the proportion of true background sequences classified as background sequences. The relative error rate (Error) for a given motif and associated with a particular p-value cut-off was  $1-(Sn+Sp)/2$  (Smith *et al.*, 2005).

## **2.12. Real-time PCR**

### A. Primer design

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

The complete lists of all the primer pair sequences, used in the real-time PCR assays, are provided in Appendix 1.

### B. Real-time PCR amplification

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

1. The ChIP DNA samples were diluted to 1 in 10 dilution i.e. 5  $\mu$ l of the sample was resuspended in 45  $\mu$ l of sterile filtered HPLC water.
2. The SYBR green PCR was set-up in a 96-well plate (Applied Biosystems) in a 25  $\mu$ l reaction, in triplicate for each sample, by mixing the following reagents on ice:
  - 2.5  $\mu$ l Water
  - 5  $\mu$ l 1.5  $\mu$ M forward and reverse primer mix
  - 12.5  $\mu$ l SYBR green PCR mix (Applied Biosystems)
  - 5  $\mu$ l ChIP DNA samples



3. PCR was performed on a 7700 sequence detection system (Applied Biosystems) .The following thermal cyclic conditions were used: 50°C for 2 min; 95°C for 10 min; then 40 cycles of: 95°C for 15 sec and 60°C for 1 min.
4.  $C_T$  values were extracted using Sequence Detector 1.7a (Applied Biosystems) with the same threshold and the  $\Delta C_T$  values were determined as follows:  
$$\Delta C_T = C_T \text{ input} - C_T \text{ ChIP sample}$$
5. Fold enrichments were calculated by using the following formula:  
$$\text{Fold enrichment} = (1 + \text{PCR yield})^{\Delta C_T}$$
  
Mean fold enrichments were calculated for each assay and data sets were normalised to a median of 1.

### **2.13. Western blotting procedure**

#### Nuclear lysate preparation

Note: All buffers were freshly prepared and stored at 4°C. Extractions were carried out on ice.

1.  $10^7$  K562 cells were harvested by centrifuging at 259g for 5 minutes. The supernatant was removed and the cell pellet was washed in PBS.
2. The cells were then spun down as before, the supernatant was removed and the pellet was resuspended in 1 ml of cell lysis buffer containing 10  $\mu$ l protease inhibitor cocktail (Sigma).
3. The cell pellet was transferred to a 1.5ml tube (pre-cooled on ice) and left on ice for 5 minutes to ensure efficient cell lysis.
4. Nuclei were spun down by centrifuging for 1 minute at 11000g at 4°C. The supernatant was then removed and the nuclei were washed in 1 ml of cell lysis buffer containing 10  $\mu$ l of protease inhibitor cocktail.
5. The nuclei were spun down by centrifuging for 1 minute at 11000g at 4°C and the supernatant was removed

6. 1 $\mu$ l 0.1M DTT & 1 $\mu$ l Sigma protease inhibitor cocktail was freshly added to 98 $\mu$ l of nuclear protein extraction buffer. The pelleted nuclei were resuspended in 70 $\mu$ l of this extraction buffer.
7. The samples were placed in a small box of ice on top of a vortexer, taped down and vortexed at the lowest setting possible for 30 minutes to avoid foam formation.
8. The samples were centrifuged for 1 minute at 11000g at 4°C to pellet any debris. The supernatant containing nuclear proteins was transferred to a fresh tube and store at -70°C.

#### Determination of protein concentration

A Bradford assay was performed as follows to determine protein concentration:

9. 0.8 ml of HPLC water was added to seven 1.5 ml spectrophotometer cuvettes (Biorad).
10. 0, 2, 5, 10, 15, 20 $\mu$ l of BSA (1  $\mu$ g/ $\mu$ l) was added to individual cuvettes to generate a standard curve. 2  $\mu$ l of protein extract was added to the final cuvette.
11. 200  $\mu$ l of protein assay dye reagent concentrate (Biorad) was added to each cuvette and mixed well by pipetting.
12. The cuvettes were kept in the dark for 20 minutes and the optical density was measured at 595 nm using a spectrophotometer.

#### Electrophoresis of protein samples

13. NuPAGE™ 4-12% Bis-Tris polyacrylamide gels (Invitrogen) and the XCell Surelock™ Mini-Cell (Invitrogen) were used for the electrophoresis of protein samples.
14. 1 litre of 1 X MOPS SDS running buffer was prepared and placed at 4°C to cool
15. The comb and tape at the bottom of the gel were removed and the wells were washed with 1 X MOPS running buffer using a syringe to remove acrylamide traces.
16. The gel was then slotted into the upper chamber (cathode) of the tank and locked into place. The upper chamber was filled with 200 ml of 1 X MOPS running buffer.

17. The lower buffer chamber (anode) was half-filled by pouring 300 ml of 1 X MOPS running buffer.
18. Protein samples were thawed at 37°C for 5 minutes if stored at -70°C and 20 µg protein samples were diluted in 4 X LDS sample buffer (maximum loading volume for 1.5 mm 10 well gels is 37 µl).
19. Protein samples were heated at 70°C for 10 minutes and then placed back on ice.
20. The Samples were then loaded onto the gel, along with 10 µl of SeeBlue® Plus2 pre-stained standard.
21. The gel tank was then placed in a cold-room (4°C) and the gel was run at 200 V constant, with a starting current of 125 mA, for 90 minutes.

#### Transfer of proteins onto PVDF membrane (Western Blotting)

22. 1 litre of 1 X NuPAGE transfer buffer was prepared while the gel was running and cooled at 4°C.
23. The gel, blotting pads, and filter paper (Whatman, cut to size 7.5 x 8 cm) were equilibrated in 1 X NuPAGE transfer buffer for 5 minutes in a fume hood.
24. PVDF transfer membrane (Sigma, 0.45 µm) was cut to size (7.5 x 8 cm) equilibrated in 100% methanol for 5 minutes and then in 1 x NuPAGE transfer buffer for 5 minutes.
25. The transfer 'sandwich' was then assembled in the blot module as shown:
  - Top (+)
  - 2 x blotting pads
  - Filter paper
  - Transfer membrane
  - Gel
  - Filter paper
  - 2 x blotting pads
  - Bottom (-)

26. Air bubbles in the blotting pads and between the gel and membrane were carefully removed. The blot module was then held together and slid into the guide rails on the lower chamber of the tank.
27. The blot module was filled with 1 x NuPAGE transfer buffer until the gel/membrane 'sandwich' was covered with buffer.
28. The transfer was performed in a cold room at 35 V constant with a starting current of 170 mA for 90 minutes.

#### Blocking the membrane

29. The PVDF transfer membrane was placed in blocking buffer for 60 minutes at room temperature on an orbital shaker and rinsed briefly in TBST.

#### Incubation with primary antibody

30. The membrane was then incubated with the primary antibody diluted 1:2000 in 15 ml blocking buffer. Membranes were incubated at 4°C over-night on an orbital shaker.
31. The membrane was then washed four times with TBST (15 minutes per wash).

#### Incubation with secondary antibody

32. The membrane was incubated with the appropriate Horseradish peroxidase-conjugated secondary antibody diluted 1:20000 in 20 ml blocking buffer at room temperature for 60 minutes on an orbital shaker. The membrane was then washed 4 x 15 minutes with TBST.

#### Immunodetection

33. ECL Plus™ Western Blotting detection reagents (GE healthcare) were used for immunodetection. Solution A (acridan substrate solution containing tris buffer) was mixed with solution B (acridan substrate solution in dioxane and ethanol) in a 40:1 ratio.
34. The detection solution was applied to the membrane for 1 minute (dark room), then placed onto an ECL Hyperfilm™ (GE healthcare) in an X-ray film cassette. The film was exposed for 5-10 seconds and developed using an automated X-ray film developer (Xograph).

## Chapter 3

### Applying ChIP-chip to study regulatory elements in 1% of the human genome

#### 3.1. Introduction

A major challenge in the post-genome era is to identify all of the protein-coding, non-protein-coding transcripts and the regulatory elements which control the expression of each transcript in the human genome. Progress on the definition of the protein coding and non-protein coding transcripts has been significant (International Human Genome Sequencing Consortium, 2004; Katayama *et al.*, 2005; Maeda *et al.*, 2006; Mattick and Makunin, 2006). However, at the start of this project, our understanding of the location and epigenetic features of *cis*-acting regulatory elements such as promoters and enhancers was limited. Promoters are located at the 5' end of genes immediately adjacent to the transcription start site (TSS) and function to recruit the transcriptional machinery as described in Chapter 1. In contrast, enhancers are typically short- or long-distance transcriptional control elements that can help activate their target genes from positions upstream, downstream or within the target gene or a neighbouring gene.

With the completion of the euchromatic portion of the human genome sequence (International Human Genome Sequencing Consortium, 2004), an important post-genome goal was to identify all of the functional elements contained within the human genome sequence (Collins *et al.*, 2003). With the advances in microarray technology and other high throughput technologies, an international collaborative project was initiated in 2003 with the aim of cataloging all of the functional elements within 1% of the human genome and then deciding which methods could best be used to scale this project to the entire genome sequence (ENCODE project consortium, 2004). One of these methods was the use of chromatin immunoprecipitation (ChIP) in combination with microarrays (ChIP-chip). As discussed in chapter 1, ChIP involves cross-linking DNA-protein interactions in living cells, followed by fragmentation of the chromatin. Specific regulatory DNA sequences associated with a particular protein interaction are then isolated by immunoprecipitation of the DNA-protein complex with an antibody specific to the protein of interest. The DNA-protein cross-links are then reversed and the purified DNA is often

analysed in a high-throughput manner by the use of genomic microarrays. The development of ChIP-chip has greatly enhanced our ability to identify and annotate regulatory elements associated with specific DNA-protein interactions. ChIP-chip methods were initially developed in yeast (Ren *et al.*, 2000; Iyer *et al.*, 2001) and since then it has been successfully applied to study DNA-protein interactions in numerous genomes (Martone *et al.*, 2003; Cawley *et al.*, 2004; Euskirchen *et al.*, 2004; Bernstein *et al.*, 2005; Blais *et al.*, 2005; Boyer *et al.*, 2005; Carroll *et al.*, 2005; Kim *et al.*, 2005; Mito *et al.*, 2005; Pokholok *et al.*, 2005; Heintzman *et al.*, 2007, Kim *et al.*, 2007; Koch *et al.*, 2007; Odom *et al.*, 2007).

In the pilot phase of the ENCODE project, 35 groups were involved in examining 30 Mbs of the human genome in unprecedented detail through the application of a number of high-throughput experimental and computational methods. The features examined by the various groups included transcripts, chromatin structure, the binding of sequence-specific transcription factors, DNA replication and genomic copy number variations (Birney *et al.*, 2007). The 30 Mb of genomic sequence to be studied by the ENCODE groups were divided amongst 44 genomic regions, 15 Mbs of which were located in 14 regions of biological or disease importance, many of which were also characterized to varying degrees with respect to non-coding functional elements. The remaining 15 Mbs were located in thirty 500 kb regions chosen by a stratified random sampling method based on gene density and level of non-exonic conservation (Table 3.1). This ensured that a wide sample of genomic regions varying in the content of genes and non-coding regulatory elements were chosen for study. This sampling method divided the human genome into three parts, the top 20%, middle 30%, and the bottom 50% based on gene density and level of non-exonic sequence conservation when compared to the orthologous mouse sequence. This resulted in nine strata, for which three random regions were chosen. For those three strata under-represented by the manual regions, a fourth region was chosen, giving a total of 30 computationally defined regions.

Region	Description	Size (Mb)	Human genomic coordinates	Non-exonic conservation (%)	Gene density
Manually selected regions					
ENm001	CFTR	1.9	Chr7:115365024-117242449	N/A	N/A
ENm002	Interleukin	1	Chr5:131332631-132332630	N/A	N/A
ENm003	Apo cluster	0.5	Chr11:115994758-116494757	N/A	N/A
ENm004	Chr22 pick	1.7	Chr22:30128508-31828507	N/A	N/A
ENm005	Chr21 pick	1.7	Chr21:32666763-34362747	N/A	N/A
ENm006	ChrX pick	1.2	chrX:151582202-152832201	N/A	N/A
ENm007	Chr19 pick	1	Chr19:59023585-60024460	N/A	N/A
ENm008	Alpha globin	0.5	Chr16:1-500000	N/A	N/A
ENm009	Beta globin	1	Chr11:4738729-5740320	N/A	N/A
ENm010	HOXA cluster	0.5	Chr7:26699793-27199792	N/A	N/A
ENm011	IGF2/H19	0.6	Chr11:1707725-2313772	N/A	N/A
ENm012	FOXP2	1	Chr7:113487636-114487635	N/A	N/A
ENm013	Manual	1.1	Chr7:89395718-90510141	N/A	N/A
ENm014	Manual	1.2	Chr7:92665828-93636236	N/A	N/A
Strata classification: Low 50% non-exonic conservation (0.0-6.3%); low 50% gene density (0.0-1.9%)					
ENr111	Random	0.5	Chr13:28318016-28818015	2.8	0.5
ENr112	Random	0.5	Chr2:51633239-52133238	3.8	0
ENr113	Random	0.5	Chr4:118705475-119205474	3.9	0
ENr114	Random	0.5	Chr10:54828416-55328415	2.8	1.2
Strata classification: Low 50% non-exonic conservation (0.0-6.3%); middle 30% gene density (1.9-4.2%)					
ENr121	Random	0.5	Chr2:118389719-118889718	6.2	2.3
ENr122	Random	0.5	Chr18:59410290-59910298	3.4	3.4
ENr123	Random	0.5	Chr12:38626477-39126476	1.7	3.1
Strata classification: Low 50% non-exonic conservation (0.0-6.3%); High 20% gene density (4.2-100%)					
ENr131	Random	0.5	Chr2:234778639-235278638	1.3	4.6
ENr132	Random	0.5	Chr13:111238065-111738064	1.1	5.5
ENr133	Random	0.5	Chr21:39242993-39742992	2.3	5.2
Strata classification: Middle 30% non-exonic conservation (6.3-10.6%); low 50% gene density (0.0-1.9%)					
ENr211	Random	0.5	Chr16:25839478-26339477	9.7	0.5
ENr212	Random	0.5	Chr5:141928468-142428467	6.7	1.7
ENr213	Random	0.5	Chr18:23717221-24217220	7.4	0.9
Strata classification: Middle 30% non-exonic conservation (6.3-10.6%); middle 30% gene density (1.9-4.2%)					
ENr221	Random	0.5	Chr5:55851135-56351134	7.9	2.2
ENr222	Random	0.5	Chr6:132157417-132657416	6.9	2.1

ENr223	Random	0.5	Chr6:73728830-74228829	6.4	3.6
Strata classification: Middle 30% non-exonic conservation (6.3-10.6%); high 20% gene density (4.2-100%)					
ENr231	Random	0.5	Chr1:148374643-148874642	10.2	8.4
ENr232	Random	0.5	Chr9:127061347-127561346	8.3	5.9
ENr233	Random	0.5	Chr15:41448853-41948852	9.7	10.6
Strata classification: High 20% non-exonic conservation (10.6-100%); low 50% gene density (0.0-1.9%)					
ENr311	Random	0.5	Chr14:51867634-52367363	14.9	0.1
ENr312	Random	0.5	Chr11:130637240-131137239	13.5	0.3
ENr313	Random	0.5	Chr16:62051662-62551661	15.4	0
Strata classification: High 20% non-exonic conservation (10.6-100%); middle 30% gene density (1.9-4.2%)					
ENr321	Random	0.5	Chr8:118769628-119269627	11.4	3.2
ENr322	Random	0.5	Chr14:97378512-97878511	15.9	2.9
ENr323	Random	0.5	Chr6:108310274-108810273	18.6	2.3
ENr324	Random	0.5	ChrX: 121480070-121980069	10.7	2
Strata classification: High 20% non-exonic conservation (10.6-100%); high 20% gene density (4.2-100%)					
ENr331	Random	0.5	Chr2:220479885-220979884	13.3	9.1
ENr332	Random	0.5	Chr11:63959673-64459672	13.4	9
ENr333	Random	0.5	Chr20:34556944-35056943	11.5	9.2
ENr334	Random	0.5	Chr9:128561347-129061346	11.4	5.4

**Table 3.1: Description of the ENCODE regions.** 14 manually selected regions (Enm001-Enm014) and 30 computationally defined regions were selected from the human genome (Enr111-Enr334) for study in the ENCODE project. Those computationally defined regions had to meet various non-exonic conservation and gene density criteria as defined in the table. The size of each ENCODE region is indicated along with the chromosome coordinates (from hg17 release).

The focus of the work presented in this thesis was the identification and characterisation of three types of *cis*-acting regulatory elements in the ENCODE regions - promoters, enhancers (this Chapter and Chapter 6) and insulators (Chapter 4). ChIP-chip assays would be used to detect histone modifications associated with promoters and enhancers, whilst the insulator binding factor CTCF would be examined to characterize putative insulators in Chapter 4. Furthermore, analysis of promoters in yeast has shown that nucleosome position is important for regulating gene expression as nucleosomes can repress transcription by occluding transcription factor binding sites (Straka and Horz, 1991, Lohr and Lopez, 1995). Genome-wide studies in yeast have shown that active



promoters are often depleted of nucleosomes (Bernstein *et al.*, 2004; Lee *et al.*, 2004; Yuan *et al.*, 2005), presumably facilitating the binding of transcription factors. However, it was not clear if nucleosome depletion was also a feature of *cis*-acting regulatory elements in the human genome. An ENCODE PCR product tiling-path microarray was designed and fabricated at the Sanger Institute (Koch *et al.*, 2007) to investigate all of the above-mentioned regulatory features. Double-stranded PCR products were spotted on microarrays using the 5'-aminolink array surface chemistry described in Chapter 1, which were then processed to generate single-stranded DNA probes. The Sanger Institute ENCODE microarray consisted of 24,005 PCR fragments with an average size of 1024 bp (average non-overlapping tile length = 992 bp). The array covered approximately 80% of the targeted regions and over 90% of non-repetitive regions.

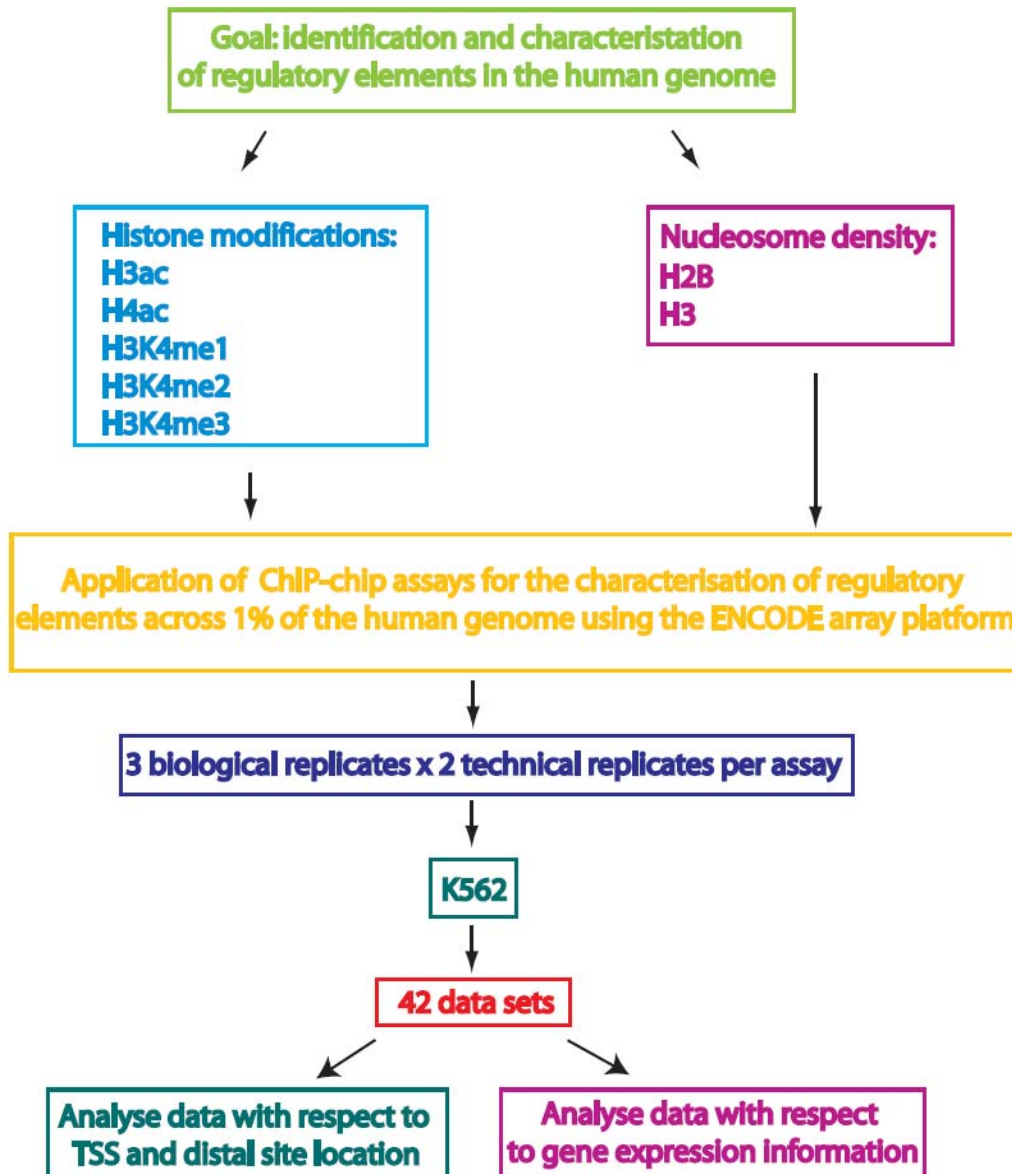
### **3.2. Aims of this chapter**

When this study was initiated relatively little was known about the histone modification state of regulatory elements in the human genome and very few large scale data sets existed which defined the location of these elements in the human genome. Therefore one of the overall aims of this study was to apply ChIP-chip approaches to characterise promoter and enhancer elements in the human genome. To this end, the aims of the work described in this chapter were:

1. To generate high-resolution maps of histone H3 lysine 9/14 di-acetylation (H3ac), histone H4 lysine 5/8/12/16 tetra-acetylation (H4ac), and histone H3 lysine 4 mono-, di-, and tri-methylation (H3K4me1, H3K4me2, H3K4me3, respectively) across the ENCODE regions.
2. To perform a detailed analysis of the distribution of these histone modifications to characterise the chromatin signatures of different types of regulatory elements in K562 cells.
3. To correlate the presence of histone modifications and nucleosome occupancy at promoters with gene expression status in K562

### **3.3. Overall strategy**

As discussed in Chapter 1, the human SCL locus is well characterized with respect to promoter and enhancer function and the histone modifications associated with these elements are known (Pawan Dhami 2005 thesis). A logical extension of this work was to define promoter and enhancer elements across larger sections of the genome. ChIP DNAs were hybridised to the ENCODE array to produce histone modification maps across 30 Mb of the human genome in K562 cells, facilitating the characterization of these regulatory elements (Figure 3.1). In addition histone H2B and H3 ChIP-chip assays were also used to investigate nucleosome occupancy/density in the ENCODE regions. Three biological replicate ChIP assays were performed for each histone modification and core histone protein. Two technical replicates were performed for each biological replicate resulting in a total of six ChIP DNA samples being prepared for each factor. The unamplified ChIP DNAs were hybridised to the PCR-product ENCODE array and the six replicates for each antibody were combined and the median value of the ratio of the ChIP-chip sample fluorescence to input DNA fluorescence was calculated for each array element. Finally, K562 gene expression data was also used to identify histone modifications associated with active and inactive promoters.



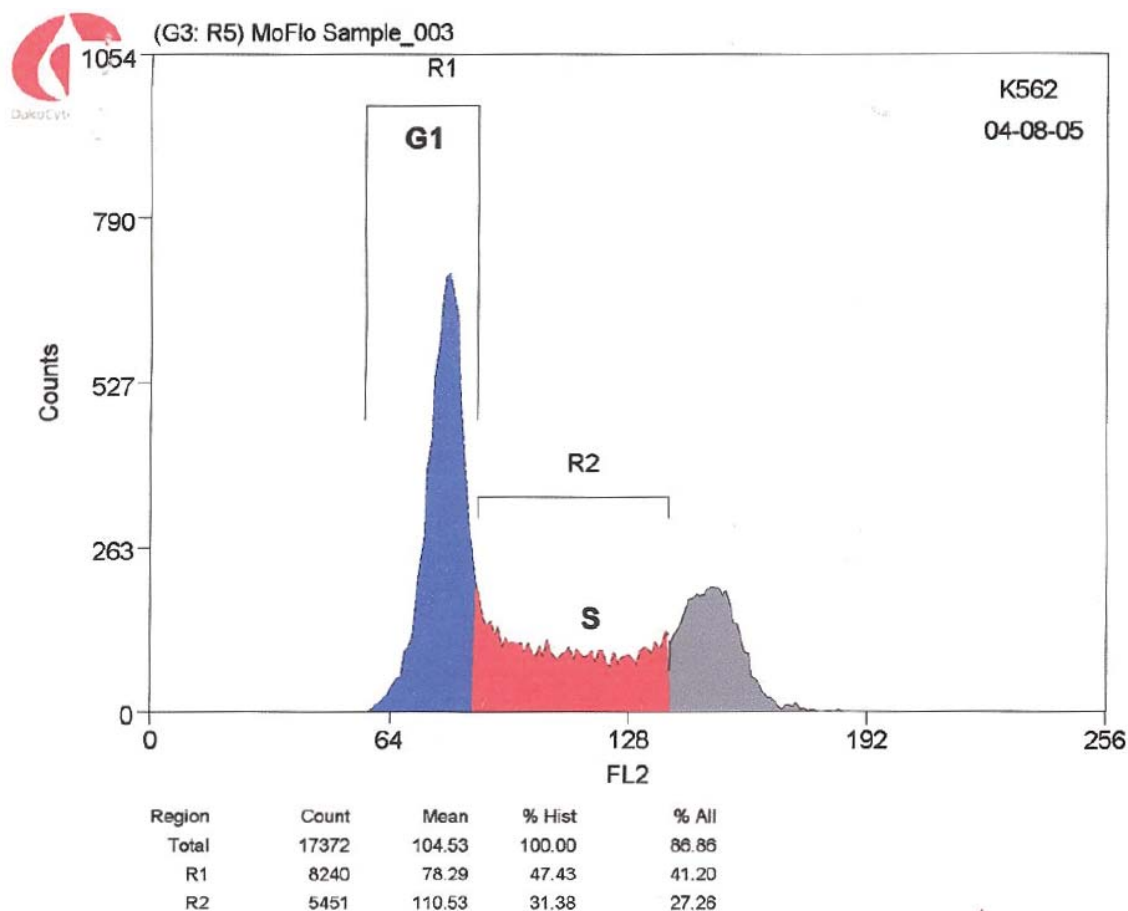
**Figure 3.1:** Schematic representation of the overall strategy used to map regulatory interactions across 1% of the human genome. This flow diagram illustrates the strategy used to identify and characterise regulatory elements in the human genome. Definitions of biological and technical replicates are described in section 3.4.

## **Results**

### **3.4. Criteria used for performing chromatin immunoprecipitation assays and microarray hybridisation**

When performing the ChIP-chip assays described in this thesis a number of experimental criteria were required in order to obtain reproducible ChIP-chip data. These criteria were applied to all ChIP-chip assays and are outlined below.

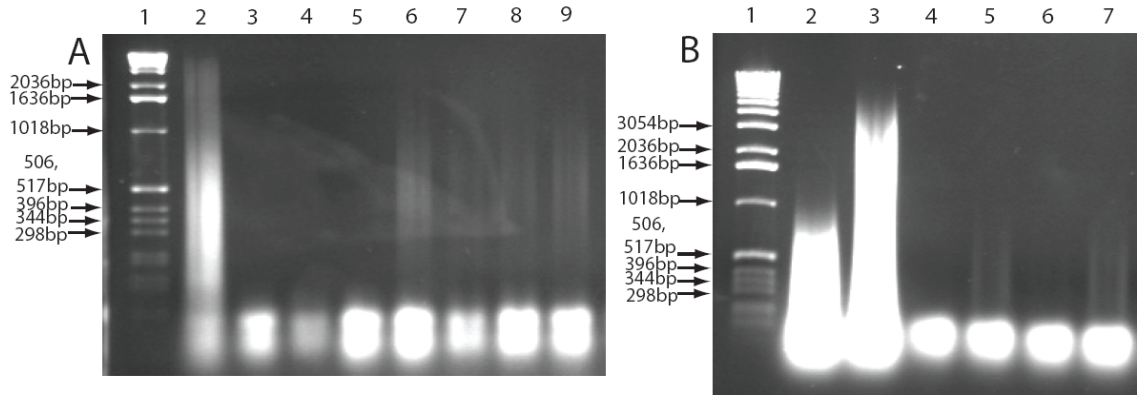
**Culturing of cell lines:** The number of actively dividing K562 cells (and U937 cells used for ChIP-chip experiments described in Chapter 4) were determined prior to the preparation of chromatin (Figure 3.2). This was necessary to ensure that sources of technical and biological variability between batches of cells from the same cell line grown at various times during the project could be reduced. An aliquot of cells were flow-sorted as described in Chapter 2 to determine the DNA content of the cells (i.e., number of cells undergoing DNA replication) – this information was used to determine the percentage of actively dividing cells. Only batches/passages of cells which displayed similar growth patterns between biological replicates were used in ChIP-chip experiments.



**Figure 3.2: Analysis of cell growth by flow-cytometry.** Human cell lines used for ChIP-chip experiments described in this thesis were analysed for their DNA content by staining with the DNA binding dye Hoechst 33342 and the percentage of cells in the G1(indicated by R1) and S (indicated by R2) phase of the cell cycle was determined by calculating fluorescence intensity (performed by Bee Ling, Wellcome Trust Sanger Institute).

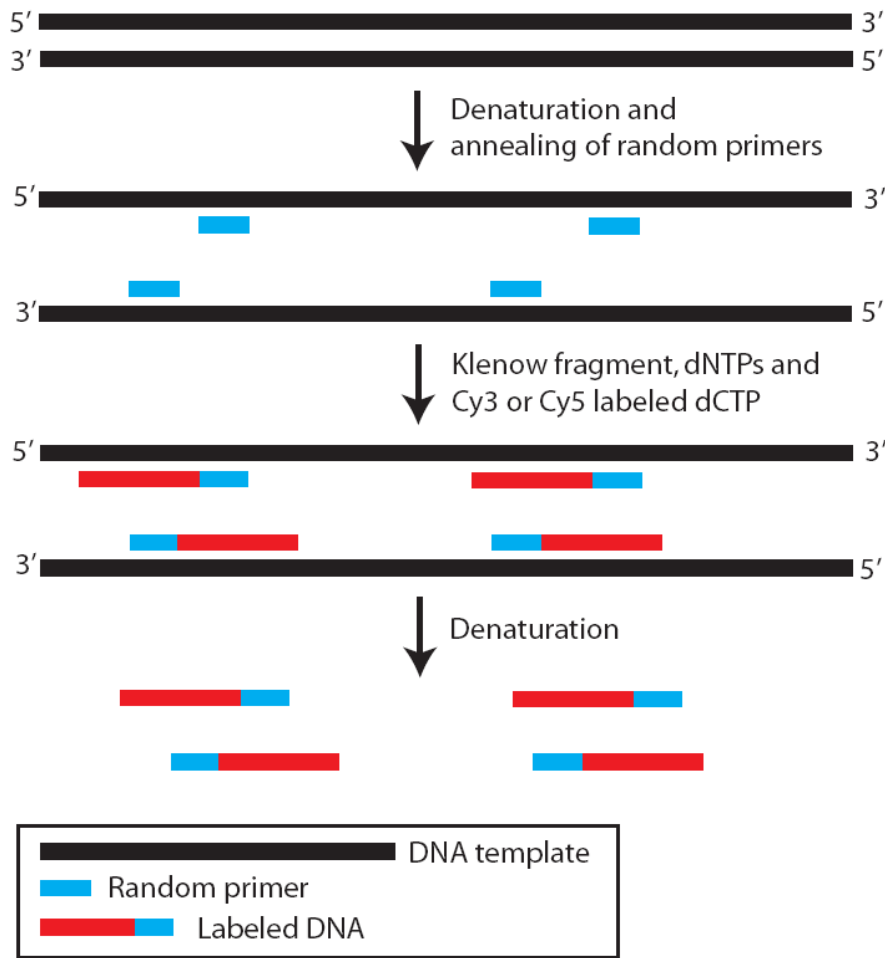
**Preparation of chromatin and ChIP DNAs:** Chromatin immunoprecipitation experiments were performed as described in Chapter 2. The percentage of formaldehyde and the cross-linking time used in the preparation of chromatin samples was crucial for the detection of DNA-protein interactions. A final formaldehyde concentration of 0.37% or 0.75% and a cross-linking time of 10 minutes were sufficient to detect histone modifications. However, a 1% formaldehyde final concentration and 15 minutes of cross-linking was required to detect transcription factor interactions. Chromatin material and input DNA samples were electrophoresed on an agarose gel to examine the effect of

formaldehyde exposure on cross-linking and sonicating efficiency. ChIP DNA samples were electrophoresed prior to labeling to examine the recovery of ChIP DNA (Figure 3.3).



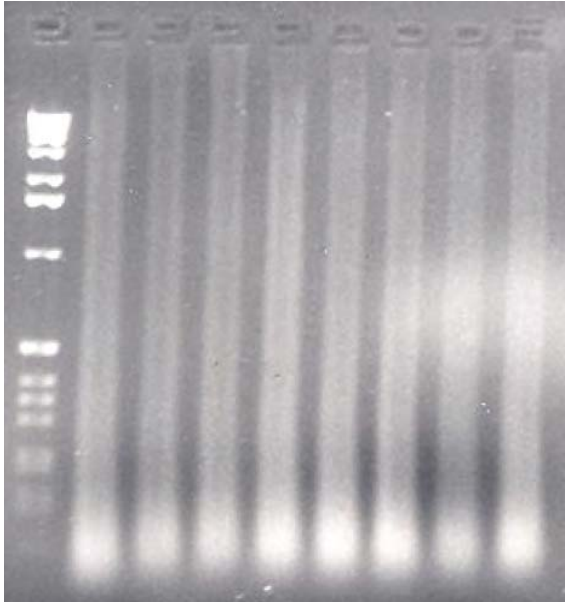
**Figure 3.3: Electrophoresis of ChIP DNAs.** Panel A shows ChIP DNAs performed for various histone modifications and panel B shows the ChIP DNAs for the transcription factor CTCF. In panels A and B, lane 1 = 1 kb DNA marker ladder. In panel A; lane 2 = 0.37% cross-linked and sonicated Input DNA; lane 3 = no antibody control; lane 4 = no chromatin control; lane 5 = rabbit IgG control; lane 6 = H3K4me3 ChIP DNA; lane 7 = H3K4me1 ChIP DNA; lane 8 = H3ac ChIP DNA; lane 9 = H4ac ChIP DNA. In panel B, lane 2 = uncross-linked sonicated human cot-1 DNA; lane 3 = 1% cross-linked and sonicated input DNA; lanes 5 and 7 = CTCF ChIP DNAs. The differences in the average sizes of the cross-linked material (panel A lane 2 and panel B lane 3) reflects the differences in cross-linking time and concentration of formaldehyde (0.37% for 10 minutes in panel A lane 2; 1% for 15 minutes in panel B lane 3). In panel A smears of DNA can be seen for the four histone modification ChIP DNA samples and similarly in panel B a faint smear can be seen for the CTCF ChIP DNA samples. The yeast tRNA used in the precipitation of DNA is observed at the base of each lane in the gel. The samples were electrophoresed on 1% agarose 1 x TBE gels and visualised with ethidium bromide.

**Labeling of input and ChIP DNA samples:** input and ChIP DNA samples were labeled by a random priming method (Figure 3.4). In this method a DNA template is denatured allowing random primers to hybridize to complimentary sequences. The random primers are then extended by the 5'-3' polymerase activity of Klenow resulting in a strand displacement activity with the incorporation of fluorescently labeled nucleotides. This random priming method results in a DNA amplification of at least four-fold over starting amounts (Lieu *et al.*, 2005).



**Figure 3.4: Labeling of DNA by a random priming method.** Input and ChIP DNAs are fluorescently labeled by Klenow mediated incorporation of Cy5 and Cy3 labeled dCTP respectively. Denaturation following this reaction results in single stranded fluorescently labelled DNA, which can then be hybridized to a microarray.

Input and ChIP DNAs which had been fluorescently labeled were electrophoresed on an agarose gel prior to microarray hybridisation (Figure 3.5). Visual inspection of samples determined whether DNA had been generated by labeling and the size distribution. The majority of labeled fragments were in the size range 80-150 bp – however, a smear of fragments extending up to and greater than 12 kb was also evident from this electrophoretic analysis.



**Figure 3.5: Electrophoresis of fluorescently labeled DNAs.** Labeled input and ChIP DNA samples were electrophoresed prior to microarray hybridisation to verify that klenow-mediated labeling reactions had resulted in the generation of single stranded DNA. A large smear of DNA was evident in the samples which had labeled successfully. The samples were electrophoresed on 1% agarose 1 x TBE gels and visualised with ethidium bromide.

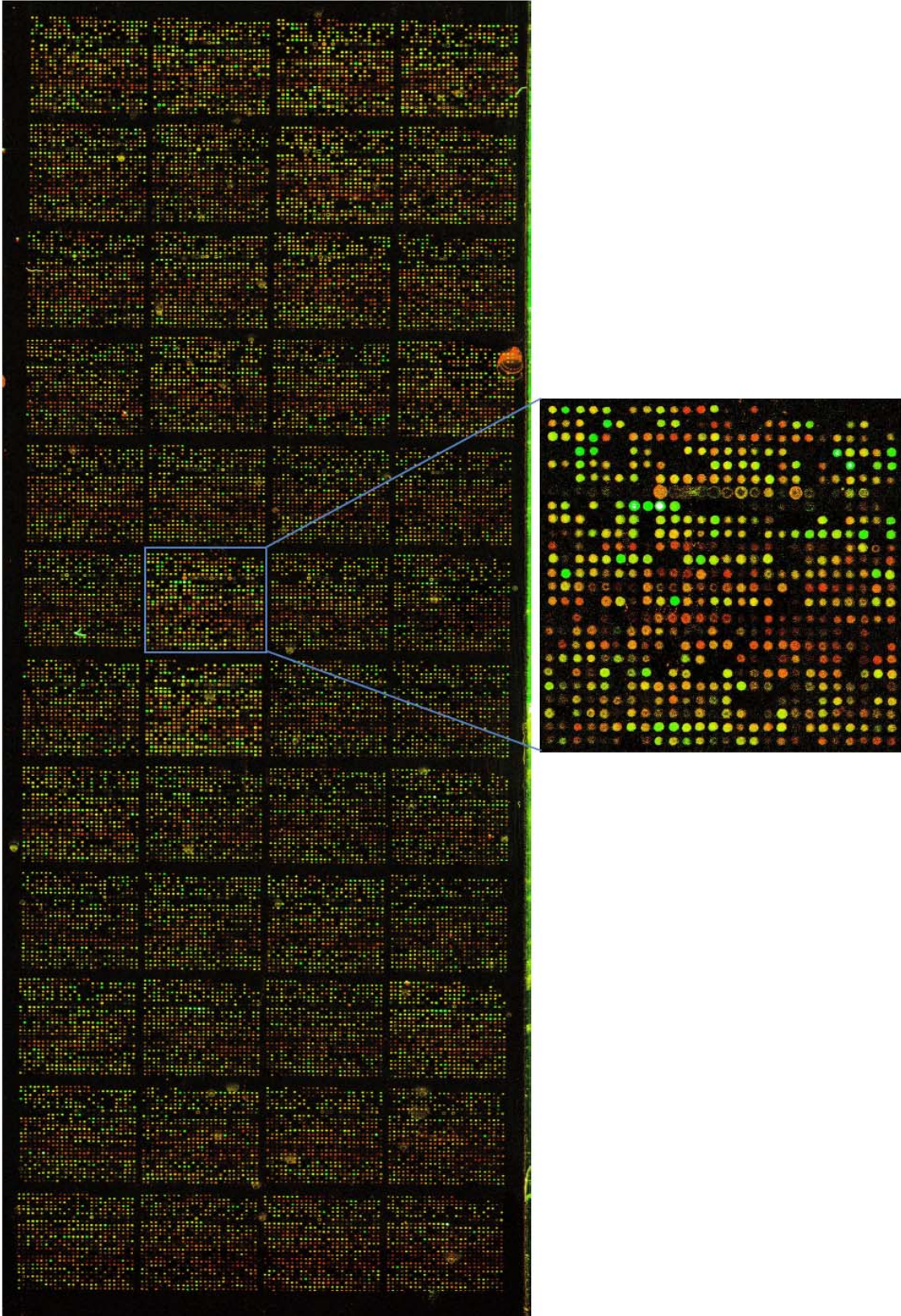
**Assessing biological and technical variation in ChIP-chip assays:** The reproducibility of ChIP-chip assays on the ENCODE arrays was assessed by performing three biological replicates per assay and two technical replicates per biological replicate. Biological replicates constitute hybridisations performed with ChIP DNA samples generated from independent ChIP assays performed with different passages of a cell line or cell type. Performing biological replicates allowed for any growth rate and gene expression differences between culture passages to be accounted for. Technical differences in sample handling, hybridisation conditions, and cyanine dye incorporation could be assessed independently from biological variation by performing technical replicates within each biological replicate.



### **3.5. Creating histone modification profiles across the ENCODE regions in K562 cells**

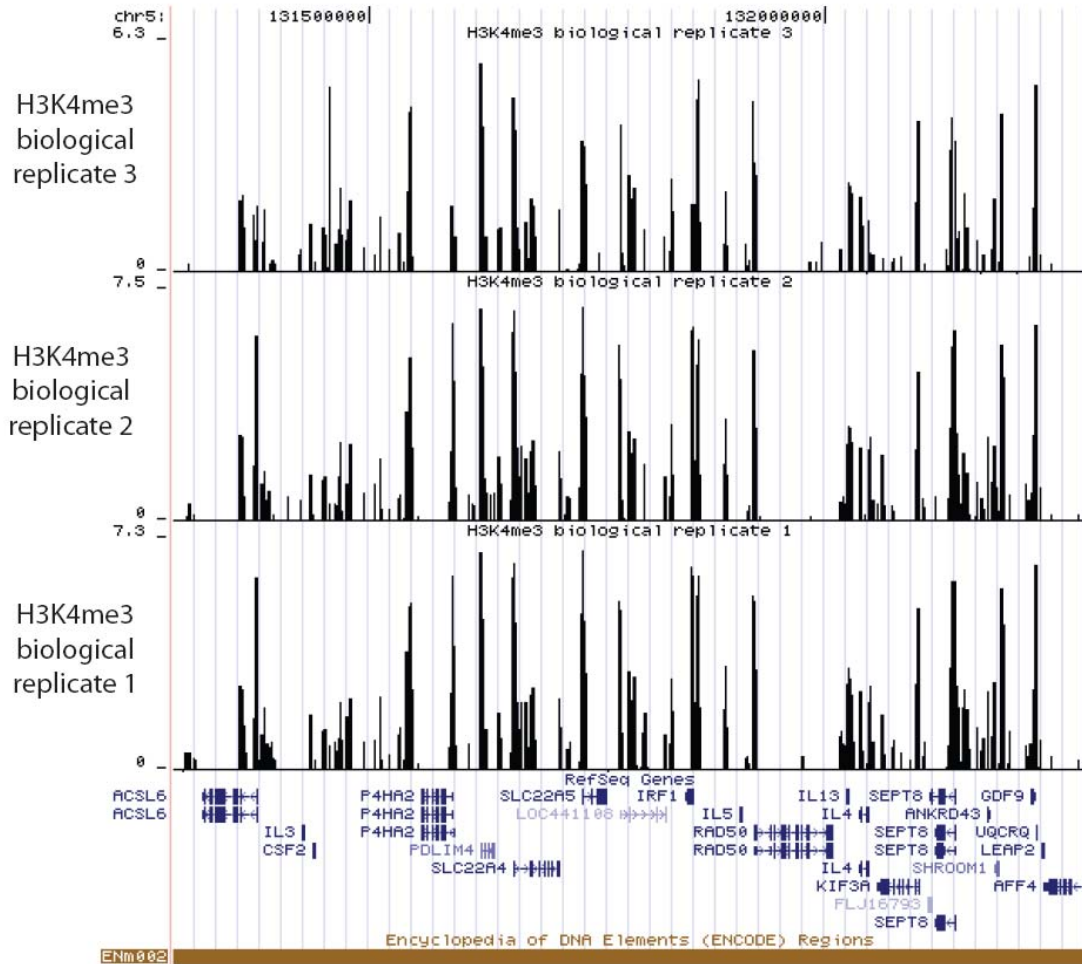
#### **3.5.1. Assessing the performance of the ENCODE array**

H3ac, H4ac, H3K4me1, H3K4me2, and H3K4me3 ChIP assays were tested on the SCL array (data not shown) and the histone modification profiles obtained for the K562 cell line reproduced previous results (Dhami, PhD Thesis, University of Cambridge, 2005), confirming that these assays were performing correctly for this project. The ENCODE microarray was then tested for its ability to visually detect enriched regions in a ChIP DNA sample by performing a hybridisation with a H3K4me3 ChIP DNA and an input DNA sample from the K562 cell line in a competitive hybridisation. This showed that the array was capable of reporting highly enriched regions and that the general signal to background fluorescence was high for the array elements (Figure 3.6).



**Figure 3.6: A composite image of the Sanger Institute ENCODE array.** The array was hybridised with a K562 H3K4me3 ChIP sample along with K562 input DNA. Each spot on the array represents an array element. PCR products were spotted in a 48 sub-grid (12 rows x 4 columns) format and a magnified view of one sub-grid is shown. Green spots represent enrichments in the ChIP sample compared to the input sample. Yellow spots represent equal hybridisation of the ChIP sample and input DNA. Orange/red spots represent regions which are under-represented in the H3K4me3 ChIP sample. White spots reflect saturated enrichments in the H3K4me3 ChIP sample.

The ENCODE array platform was then assessed for its ability to reproducibly detect genomic regions associated with H3K4me3. Three biological replicate experiments were performed and hybridised to the array in order to determine the reproducibility of the array platform. The coefficient of variation (CV) is a measure of dispersion of a probability distribution and can be used to calculate variation between experiments. Each array element was only present once due to space restrictions on the array so the CV of ratios (ratio of ChIP sample fluorescence to input DNA fluorescence) was calculated for corresponding array elements between hybridisations of biological replicate samples as described in Chapter 2. This was expressed as a percentage and the mean CV of ratios was calculated to be 22.37 % between the three experiments. Thus, on average there was 22.37 % variation in the ratio values reported between the three biological replicate experiments. Technical variation was also assessed within an individual biological replicate by performing two technical replicate experiments. The mean CV of ratios was calculated to be 19.33 %. This indicated that the array elements were reporting reproducible ratio values between biological and technical replicate experiments and that greatest variation was observed between biological replicates. An example of the reproducible performance of the array platform in H3K4me3 biological replicate ChIP-chip assays is shown in Figure 3.7.

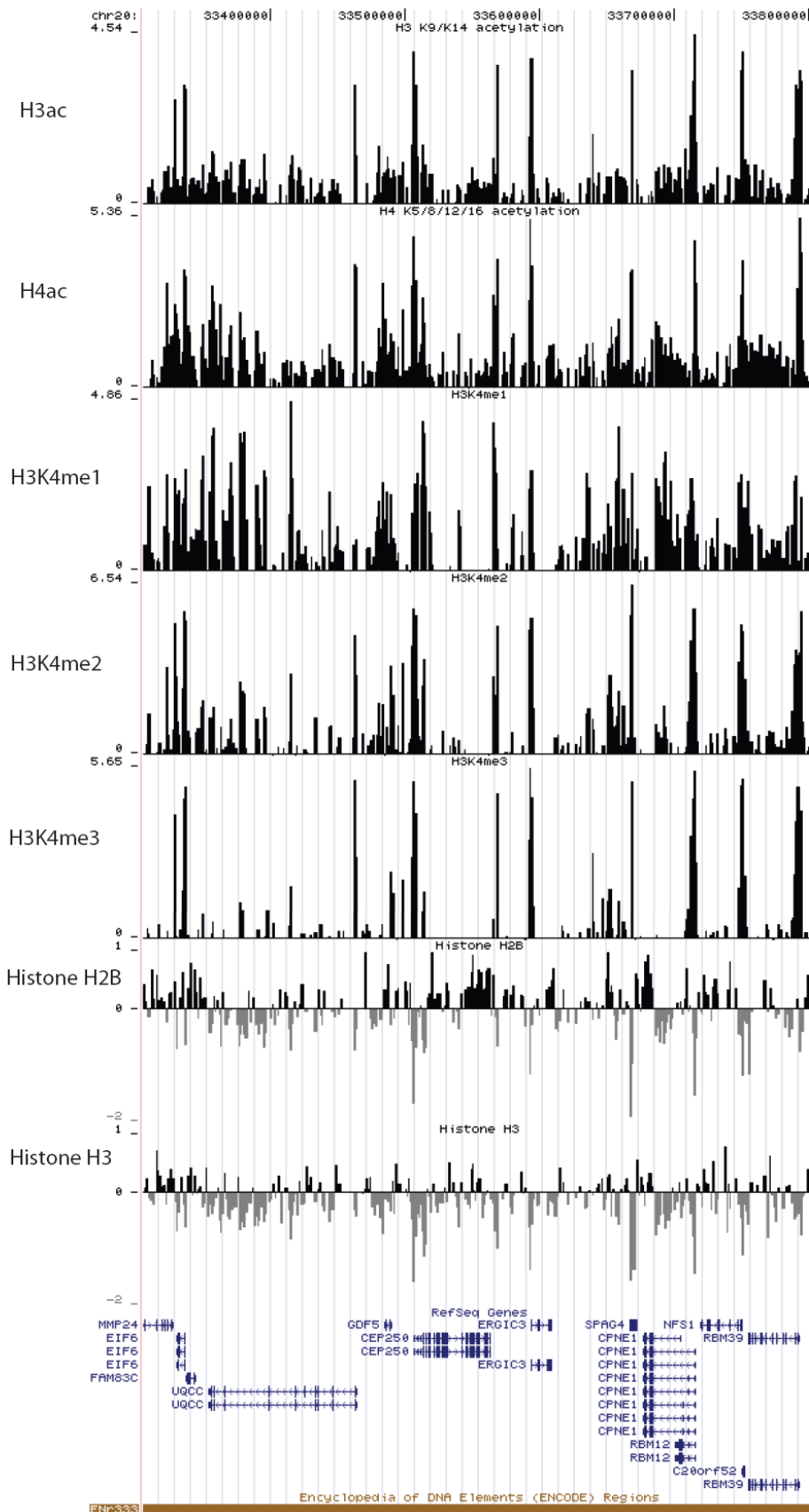


**Figure 3.7: Assessing the performance of the ENCODE microarray in independent ChIP-chip assays.**

A screenshot from the UCSC genome browser (Kuhn *et al.*, 2007) of ENCODE region Enm002 (chr5:131284314 - 132284313) showing ChIP-chip data from three independent H3K4me3 experiments in K562 cells. Reported log<sub>2</sub> fold-enrichments were observed to be very similar between the three experiments indicating that the array platform was performing reproducibly. The scale in base pairs is indicated at the top of the figure. The bottom track shows the Refseq genes (Pruitt *et al.*, 2007) with transcriptional orientation indicated by arrows. The ChIP-chip data is displayed in the three intervening tracks as the ratio of ChIP-chip sample fluorescence to input DNA fluorescence. Each black vertical bar is the enrichment measured at a single array element on the ENCODE microarray with the enrichment represented by the height of the bar.

### **3.5.2. Constructing histone modification and nucleosome density profiles**

The ENCODE microarray was then used to determine the detailed chromatin structure of 1% of the genome in K562 cells. The patterns of H3ac, H4ac, H3K4me1, H3K4me2 and H3K4me3 across the ENCODE regions were investigated along with the core histones H2B and H3. This resulted in the creation of detailed high-resolution histone modification maps across all the ENCODE regions in K562 cells, an example of which is shown in Figure 3.8, visualized as “wiggle” tracks in the UCSC genome browser.



**Figure 3.8: An example of the histone modification and nucleosome density maps generated across the ENCODE regions.** A screenshot from the UCSC genome browser (Kuhn *et al.*, 2007) of ENCODE region ENr333 (human chromosome 20: 33,304,929–33,804,928 bp) showing ChIP-chip data “wiggle” tracks for five antibodies raised to histone modifications H3 K9/K14 acetylation, H4 K5/8/12/16 acetylation, H3K4me1, H3K4me2, H3K4me3, and the core histone proteins H2B and H3. The scale in base pairs is indicated at the top of the figure. The bottom track shows the Refseq genes (Pruitt *et al.*, 2007) with transcriptional orientation indicated by arrows. The ChIP-chip data is displayed in the seven intervening tracks as the median value of the ratio of ChIP-chip sample fluorescence to input DNA fluorescence. Each black vertical bar is the enrichment measured at a single array element on the ENCODE microarray with the enrichment represented by the height of the bar. Regions depleted of histones H2B and H3 are indicated by grey bars below the x-axis in the respective tracks. Note that fold enrichments in the ChIP samples are displayed as  $\log_2$  values for each track and are scaled according to the fold enrichments obtained for each assay.



### 3.6. Analysing the distribution of histone modifications

#### 3.6.1. Defining statistically enriched regions

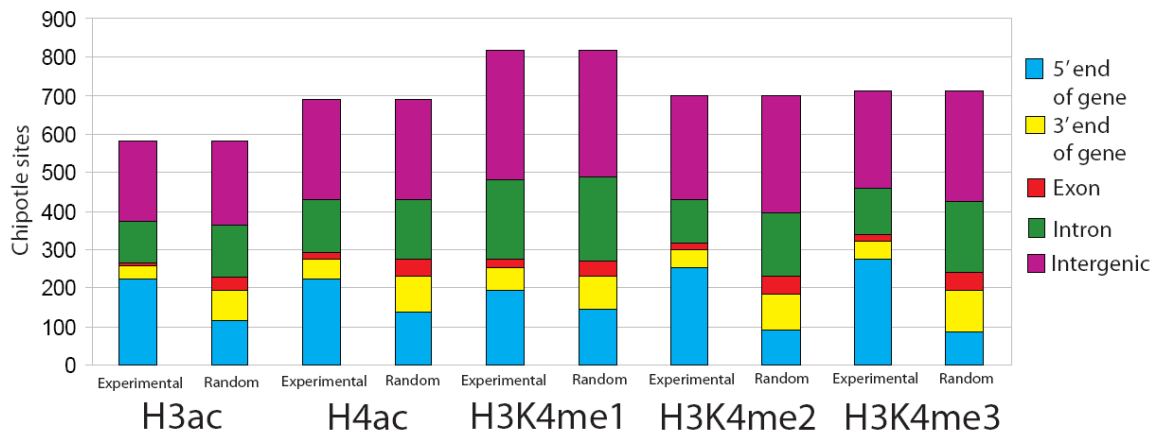
Visual examination of the data from these experiments showed that the largest enrichments for H3K4me2, H3K4me3, H3ac, and H4ac were principally located at the 5' ends of annotated genes. Enrichments for H3K4me1 were also found at the 5' end of genes but this modification seemed to have a more widespread distribution. To more accurately assign significant enrichments to genomic features in this study, the ChIPOTle (Chromatin ImmunoPrecipitation On Tiled arrays) program (Buck *et al.*, 2005) was used to define statistically significant peaks of enrichment in the H3ac, H4ac, H3K4me1, H3K4me2, and H3K4me3 data sets. ChIPOTle uses a sliding window approach to estimate the significance of enrichment for a genomic region using a standard Gaussian error function. It was suggested that a p-value cutoff that produces about 50 times more significant regions than significant negative regions was a satisfactory cut-off for the majority of applications (Buck *et al.*, 2005). The p-value cut-off which came close to this value for this study was determined to be p0.0005 for the five histone modification data sets. On average 56 times more significant regions were identified in the five histone modification data sets when this p-value cut-off was used (data not shown). This cut-off was used to identify numerous sites of histone modifications across the ENCODE regions, many of which were located at TSSs (Table 3.2). ChIPOTle identified significant sites with a median size of 1.5 kb for all five histone modification states.

Antibody	Number of Chipotle calls	Number of Chipotle calls at TSSs (+/- 2.5kb)
H3ac	582	224
H4ac	690	222
H3K4me1	818	192
H3K4me2	697	255
H3K4me3	711	274

**Table 3.2: Descriptive statistics for Chipotle sites for K562 ChIP-chip data.** The number of ChIPOTle sites identified using a p-value cut-off of p0.0005 is indicated for the five histone modifications. Those ChIPOTle calls which overlapped with GENCODE transcription start sites (TSSs) (Harrow *et al.*, 2006) are also indicated.

### 3.6.2. The distribution of histone modifications with respect to gene features

The distribution of histone modification sites defined by ChIPOTle was then examined with respect to the location of gene features. The number of ChIPOTle sites which overlapped with 5' ends of genes, exons, introns, 3' ends and intergenic locations was determined (Figure 3.9). Whilst initial inspection of the histone modification data suggested that 5' ends of genes were associated with large numbers of ChIPOTle hits (Table 3.2) it was necessary to determine whether or not this association occurred by chance. Thus, random simulated data was also produced by generating sites of the same size distribution as the ChIPOTle sites and placing them at random in the ENCODE regions to identify gene features over-represented for histone modifications. This was repeated 100 times and the mean frequencies are plotted in Figure 3.9. By analysing a +/- 2.5kb window surrounding the 5' end of genes it was observed that ChIPOTle sites for all five histone modifications were over-represented at the 5' end of genes compared to the simulated random distribution data. In contrast all five histone modifications were found to be under-represented at the 3' end of genes and in exons. This is consistent with previous reports which described a distinct localization of H3K4 methylation and histone acetylation at promoters in the human genome (Liang *et al.*, 2004; Kim *et al.*, 2005).



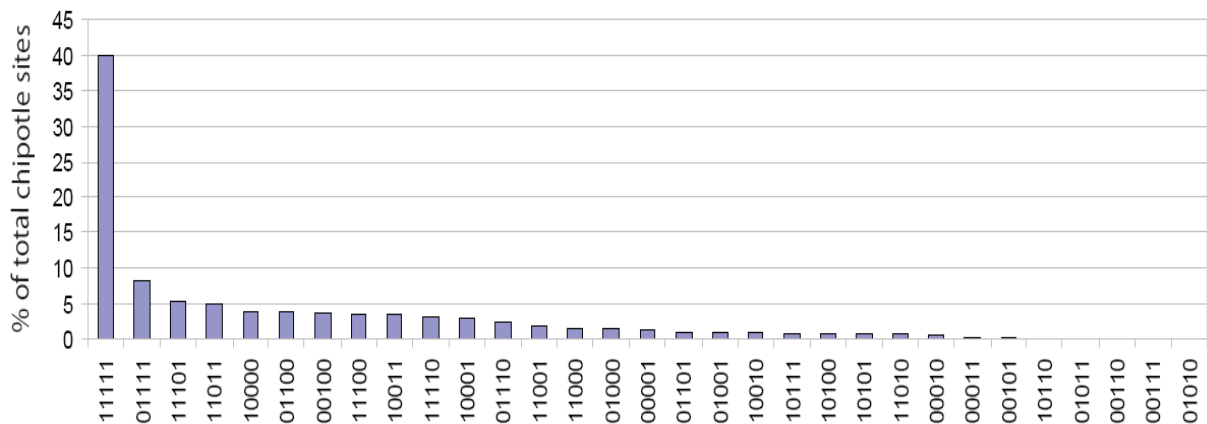
**Figure 3.9: Distribution of histone modifications in K562 cells.** The chart shows the number of H3ac, H4ac, H3K4me1, H3K4me2, and H3K4me3 ChIPOTle sites which overlap with a TSS (blue), 3' end of a gene (yellow), exon (red), intron (green), or intergenic sequence (purple) in ChIP-chip data (experimental) and random simulated data (random). Random data was simulated by generating sites of the same size distribution as the experimental data and placing them at random in the ENCODE regions 100 times. The



mean number of overlapping sites located at the 5' end of genes, the 3' end of genes, within exons, introns and intergenic regions is plotted in the figure.

### 3.6.3. Analysis of the combinatorial nature of histone modifications

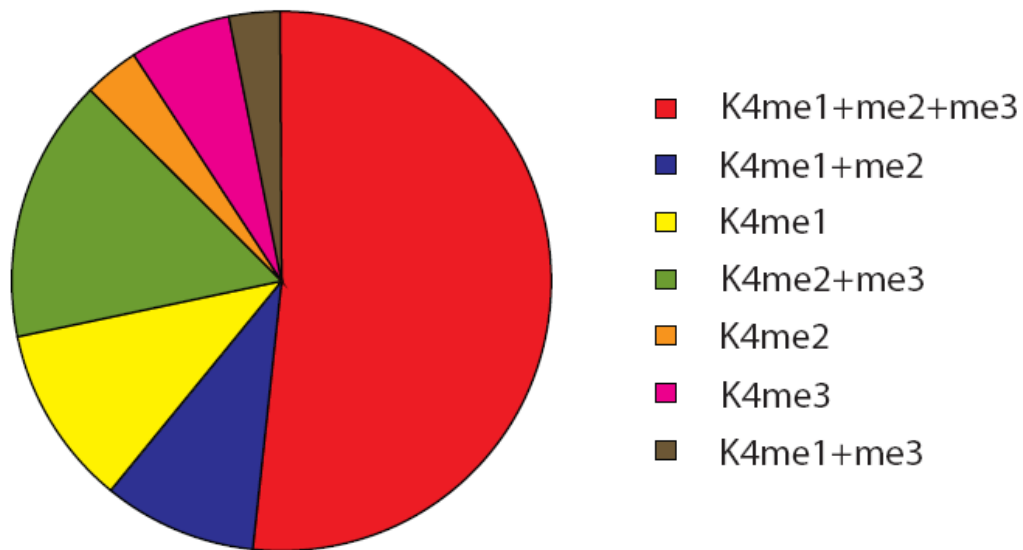
The overlap between individual histone modifications was then examined to determine which combinations occur most frequently together in the ENCODE regions. Using an overlapping window approach ( $\pm 2.5\text{kb}$  from the centre of each ChIPOTle peak), the extent of combinatorial histone modifications was determined (Figure 3.10). A 0, 1 binary code (0 indicating no presence and 1 indicating presence of a histone modification) was used to define sites of overlap between the five histone modification states. The code is presented in the following order: H3K4me1, H3K4me2, H3K4me3, H3ac, H4ac. This approach showed that overlap between all five histone modifications was the most common occurrence (represented as 11111), representing 40% of the total ChIPOTle sites. H3K4me1 (code 10000) was the most frequently observed site (4%) containing only one modification. Further analysis indicates that the many of these H3K4me1 only sites are located at a distance from TSSs (Section 3.7). This analysis also showed that H3K4me3 is usually associated with H3K4me2 (represented as N11NN, in which N can be 0 or 1) which has been described previously (Bernstein *et al.*, 2005).



**Figure 3.10: The relative occurrence of overlapping histone modification sites in K562.** Histone modification sites were deemed to overlap if they were present within a 5kb window centred on a ChIPOTle site, i.e. located  $\pm 2.5\text{kb}$  from the centre of a ChIPOTle site. The relative occurrence of each overlapping combination was calculated and expressed as a percentage of the combined total number of ChIPOTle sites identified for the five histone modifications. Thirty two different permutations existed for the presence or absence of the five histone modifications states at ChIPOTle sites. Overlapping

combinations are presented using a binary code, in which 1 represents the presence of a modification at a site and 0 represents the absence of a modification. The code is presented in the following order H3K4me1:H3K4me2:H3K4me3:H3ac:H4ac.

The binary code of only H3K4 methylation (i.e., excluding H3ac and H4ac data) was also examined and the most common combination of modifications was sites containing all three H3K4 methylation states (Figure 3.11). The next most common combination of histone modifications were sites containing both H3K4me2 and H3K4me3, followed by H3K4me1 only and then sites associated with both H3K4me1 and H3K4me2. Fewer sites containing H3K4me2 only, or H3K4me1 and H3K4me3 together were observed.



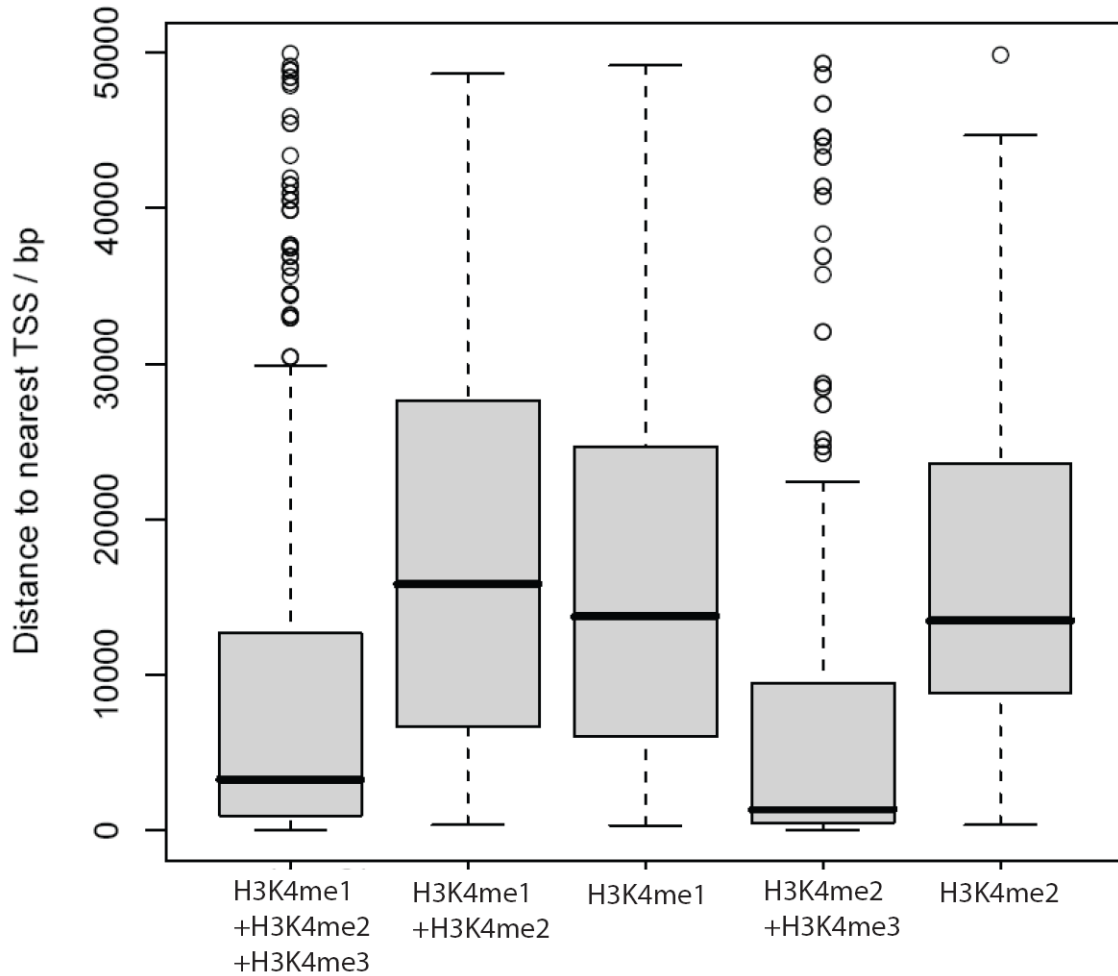
**Figure 3.11: The extent of overlapping H3K4 methylation sites in K562.** The pie-chart illustrates the occurrence of the seven H3K4 methylation combinations.

However, specific combinations of histone modifications cannot be viewed as occurring on the same nucleosome ‘tail’ as the resolution of the array elements ensures that histone modifications are detected across several nucleosomes, representing an average of approximately five nucleosomes per array element. In addition this qualitative analysis does not take into account the enrichment levels of each histone modification at these overlapping sites, but treats all levels of enrichment with equal weighting once identified by ChIPOTle. This issue is addressed in the next section of this Chapter, where the

quantitative enrichment of histone modifications is examined, thus determining quantitative histone modification signatures at these sites.

### **3.7. Distinct histone modification signatures define transcription start sites and distal elements in the ENCODE regions**

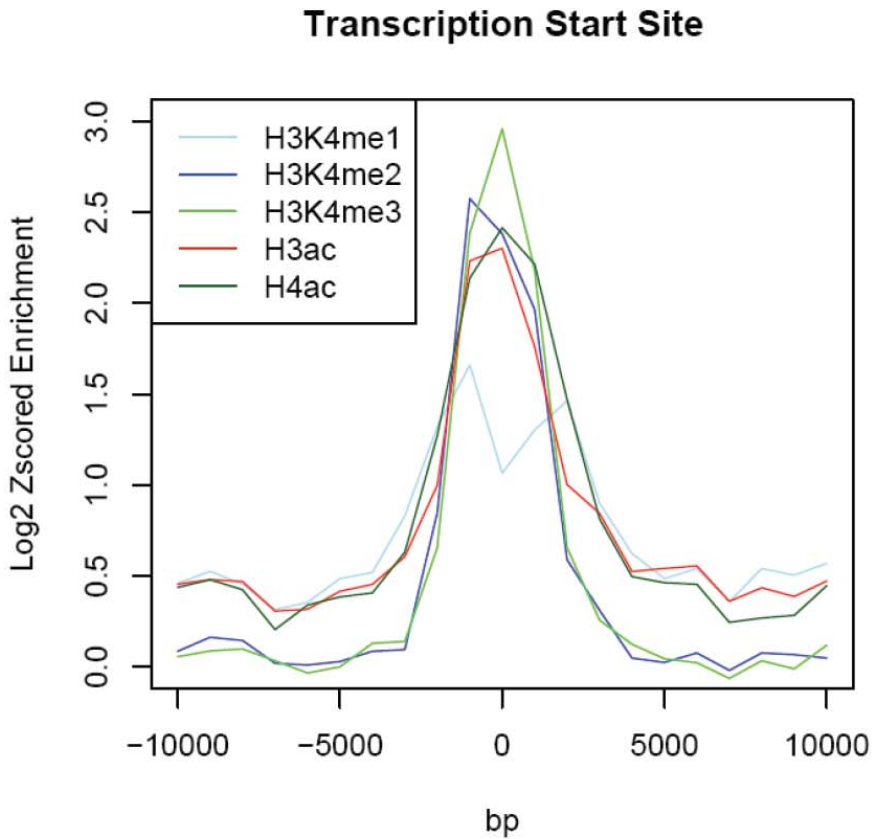
A more detailed analysis of the location of H3K4 methylation modifications was performed by analyzing the distribution of the five main histone H3K4 methylation combinations relative to the location of the nearest TSS (Figure 3.12). The distance to the nearest TSS was calculated for each of these sites. The median distance from TSSs for histone modification sites associated with H3K4me1+H3K4me2+H3K4me3 was 3255 bp and the median distance for sites associated with H3K4me2+H3K4me3 was 1319 bp. The median distance from TSSs for H3K4me1+H3K4me2, H3K4me1, and H3K4me2 sites was 15828 bp, 13747 bp, and 13475 bp respectively. This showed that sites containing H3K4me3 were located closer to TSSs than those which did not contain this modification. Thus this analysis defined the presence of two classes of histone modification sites, those located close to TSSs with H3K4me3 enrichment and a distal class without H3K4me3.



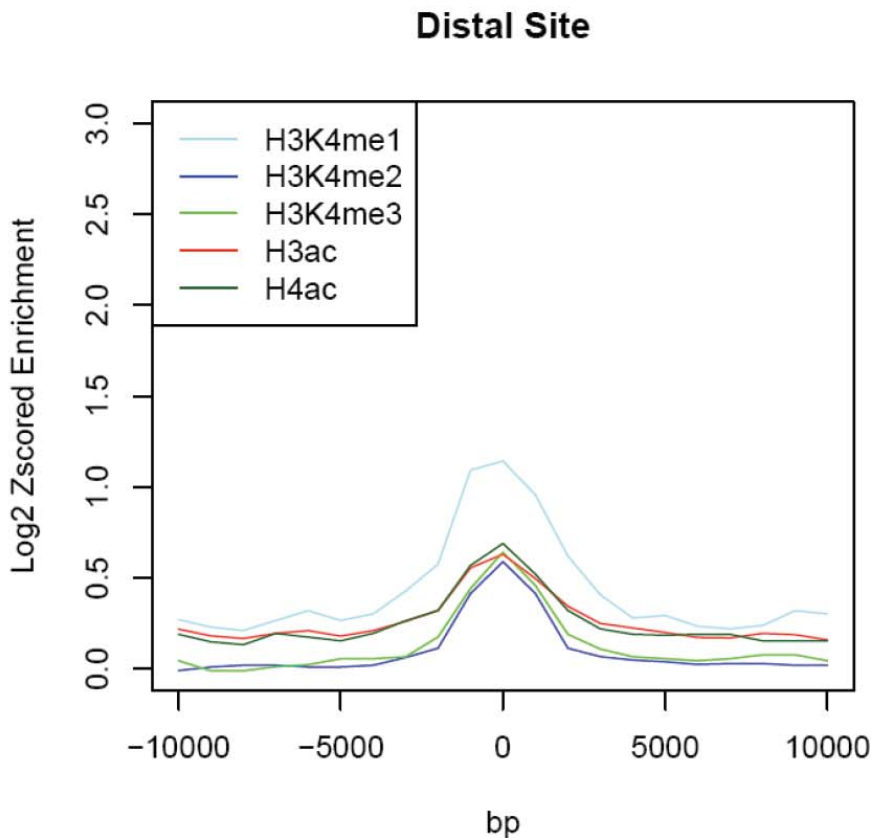
**Figure 3.12: The distribution of histone H3K4 methylation combinations relative to the nearest transcriptional start site (TSS).** Box-plots are presented for the five main histone modification combinations showing their distribution relative to the nearest TSS. The grey boxes show the 25<sup>th</sup> and 75<sup>th</sup> percentiles, with the black line inside the grey box represents the median distance. The whiskers extend to the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentiles while circles represent outliers. Distance to the nearest TSS is presented in base-pairs along the y-axis.

The ChIPOTle sites for each of the five histone modifications were then sub-classified into those located proximal ( $\pm 2.5$ kb) to TSSs and those located distal to TSSs (greater than 2.5kb). The average histone modification profile of the TSS proximal and distal ChIPOTle defined sites were determined and are plotted as the average z-scored  $\log_2$  values in Figure 3.13. The z-score is a dimensionless value derived by subtracting the mean  $\log_2$  value of a data set from an individual  $\log_2$  value and then dividing the

difference by the standard deviation of a data set. The z-score indicates how many standard deviations an observation is above or below the mean. Standardizing of different ChIP-chip data sets by calculating z-scores allows for effective comparisons to be made between different histone modification data sets as various antibodies may have different immunoprecipitation efficiencies. The sites located at TSSs displayed the highest z-scores for H3K4me3, followed closely by H3K4me2, then H4ac and H3ac. H3K4me1 displays the lowest average z-score at TSSs. While H3K4me3, H3ac and H4ac peak directly over TSSs, H3K4me2 peaks approximately 2kb upstream of TSSs. H3K4me1 displays a bimodal pattern as it peaks approximately 2kb upstream of TSSs, then decreased levels of this modification were observed directly over TSS and another peak is observed approximately 2 kb downstream of TSSs. However, this analysis did not take into account the expression status of the genes associated with these TSSs. This is examined in the following section. In contrast to TSSs, distal sites display a different histone modification profile as H3K4me1 was found to be the most prominent modification at these elements, consistent with these sites being enhancer elements (Heintzman *et al.*, 2007). However it cannot be ruled out that some of these distal sites may represent regions in which repressive transcription factors bind and thus act as distal repressors. Distal elements were also associated with intermediate levels of H3K4me2, H3K4me3, H3ac, and H4ac.



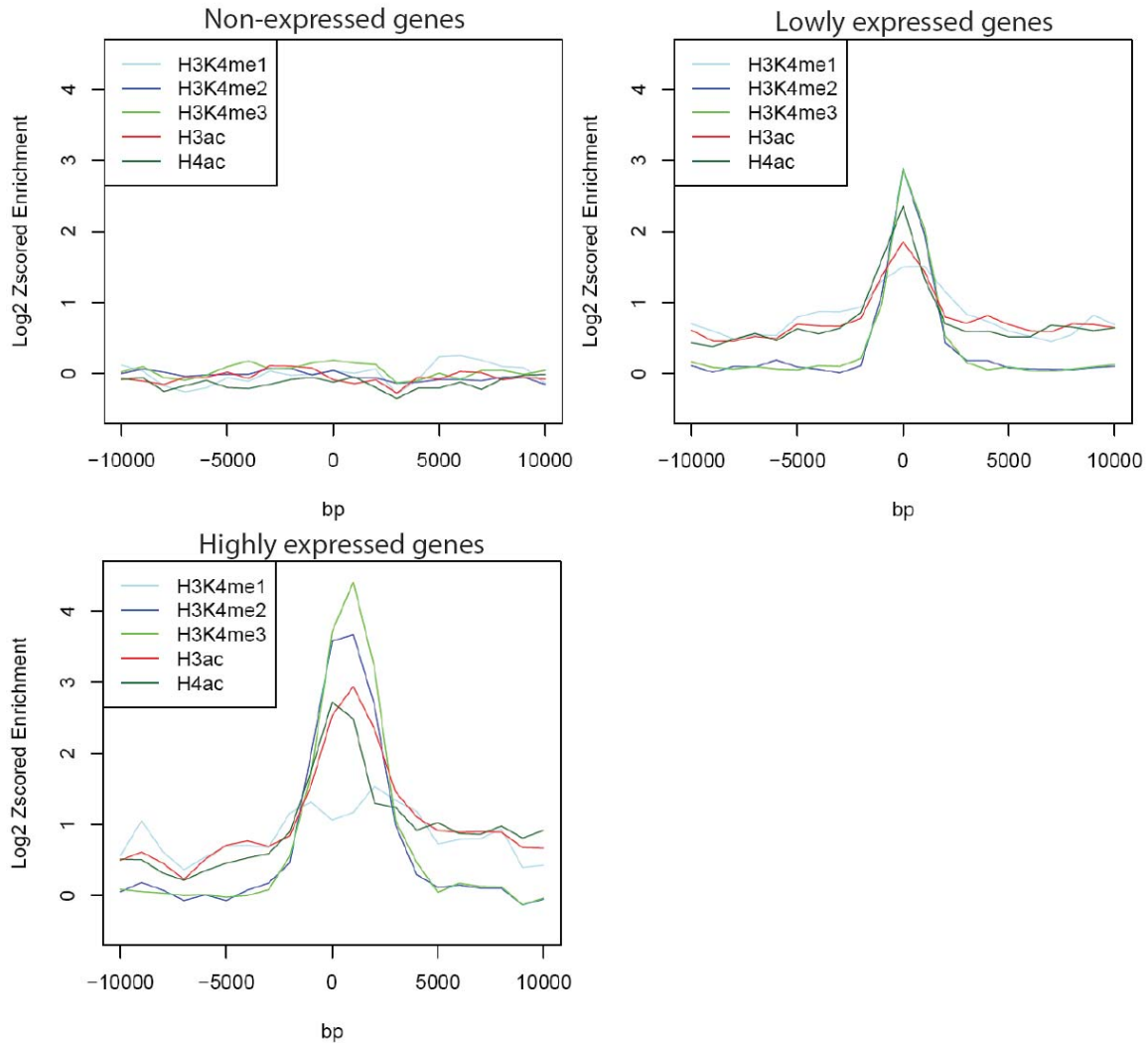
**Figure 3.13: Histone modification profiles at TSSs and distal sites.** The average Log<sub>2</sub> Z-scored histone modification profiles are presented for ChIPOTle defined sites located at TSSs (+/- 2.5kb) and at distal elements (located greater than 2.5kb from TSSs). The average values are plotted +/- 10,000 bp relative to the location of TSSs or distal elements (indicated on the x-axes by 0). Average Log<sub>2</sub> z-scored values are indicated on the y-axes.



### **3.8. Histone modifications and transcription**

#### **3.8.1. The relationship between histone modifications and transcription status**

The relationship between these histone modifications and transcription status was investigated by comparing the TSS histone modification profiles with K562 Affymetrix U133 plus 2.0 gene expression data (obtained courtesy of Dr. Christoph Koch and Dr. Phillippe Couttet, Wellcome Trust Sanger Institute). The robust multichip average (RMA) expression values for ENCODE genes present on the Affymetrix array were ranked in order of expression as high (100-75%), low (75%-50%), indeterminate (50%-25%), and off (25%-0%). The RMA values of these four classes of genes were also compared to MAS5 absent and present expression calls. Genes which were called as present by MAS5 and which were in the top 50% of ranked RMA values were considered as expressed genes (either as high or low expression based on the above criteria). Genes which were called by MAS5 as absent and which were in the bottom 50% of ranked RMA values were considered as not expressed. Extending this logic, genes which showed discrepancies between MAS5 and RMA data were classified as indeterminate and were excluded from further analysis. The reason for excluding these genes was to get an unequivocal answer about histone modifications associated with on and off expression states. The z-scored  $\log_2$  fold enrichments for the five histone modifications were plotted at the TSSs and surrounding regions of highly expressed, lowly expressed, and non-expressed genes (Figure 3.14). Non-expressed genes were associated with very low enrichments for all five histone modifications. In contrast, lowly expressed genes were associated with enrichment for all five modification states. H3K4me3 was the most prominent modification state while H3K4me1 levels were lowest at these promoters. All five methylation states peaked over the TSS of lowly expressed genes. The promoters of highly expressed genes were associated with even higher levels of H3K4me2, H3K4me3, H3ac and H4ac found to be associated with high levels of H3K4me3, H3K4me2, and H3ac which peaked approximately 1.5kb downstream of the TSS. Once again H3K4me3 was the predominant modification at these promoters while a small depletion of H3K4me1 was observed directly over the TSS of these genes. Enrichment for H3K4me1 peaked approximately 2kb downstream from the TSS.

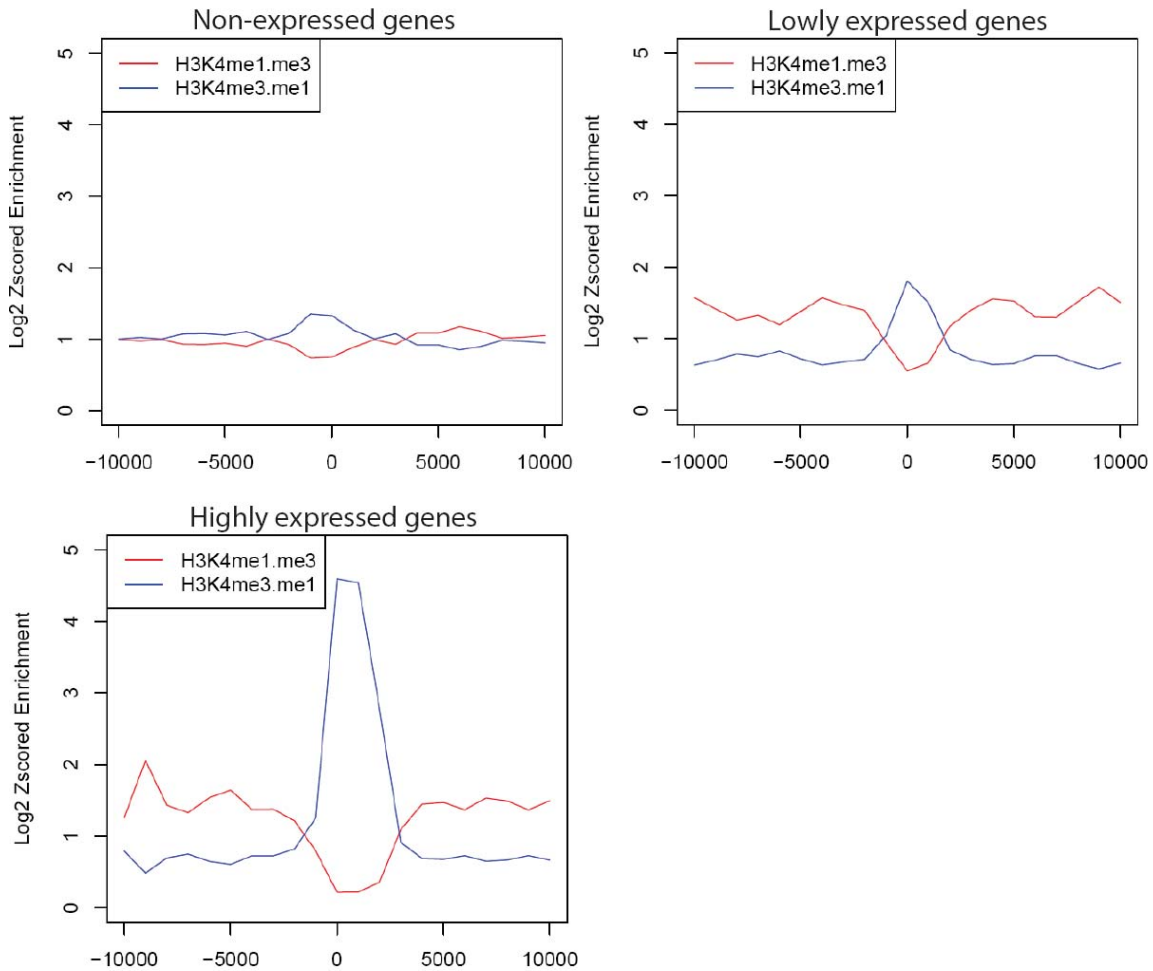


**Figure 3.14: Histone modification profiles for non-expressed and expressed genes in K562 cells.** The average z-scored  $\log_2$  histone modification profiles are presented  $\pm$  10,000 bp surrounding TSSs of non-expressed, lowly expressed, and highly expressed genes.

This analysis suggested that the presence of H3K4me3 and H3K4me2 at promoters was associated with active gene expression and the highly expressed genes were associated with higher levels of this modification compared to lowly expressed genes in K562 cells. In addition, H3K4me1 levels were at much lower levels over the TSSs of highly active genes. Therefore to compare the relative levels of H3K4me1 and H3K4me3 at TSSs, profiles were generated by plotting the average z-scored H3K4me3 to H3K4me1 ratios surrounding the TSSs of non-expressed and expressed genes (Figure 3.15). The promoter regions of non-expressed genes showed a low H3K4me3:H3K4me1 ratio, while lowly



expressed genes were associated with an intermediate H3K4me3:H3K4me1 ratio, and highly expressed genes were associated with a much higher H3K4me3:H3K4me1 ratio. The presence of a low H3K4me3:H3K4me1 ratio at the TSS of inactive genes raises the possibility that the promoters of inactive genes in K562 cells are associated with residual levels of H3K4me3 relative to H3K4me1 which may ‘prime’ them for rapid expression.

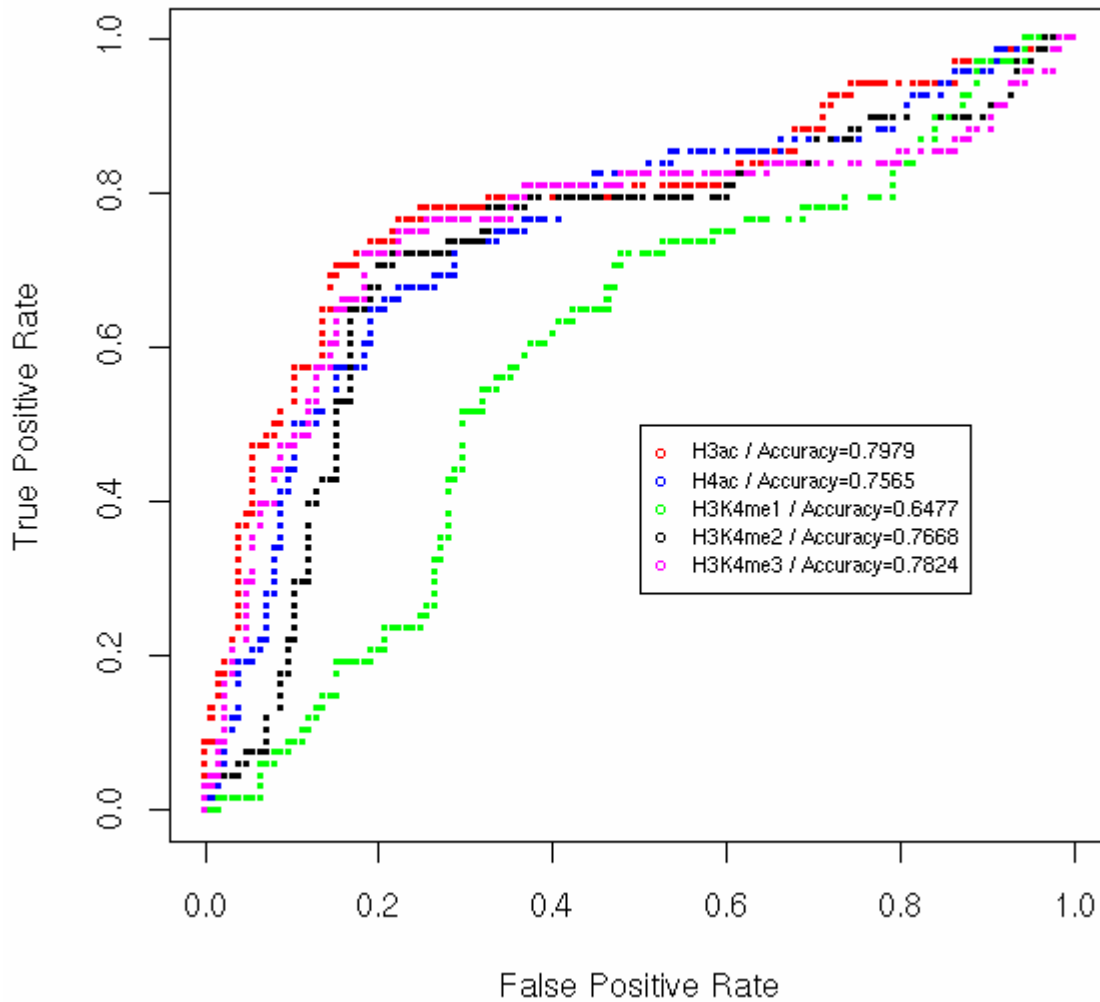


**Figure 3.15: Profiles showing H3K4me3 to H3K4me1 ratios at the promoters of active and inactive genes in K562 cells.** The average z-scored  $\log_2$  H3K4me3 and H3K4me1 values were plotted relative to each other to obtain H3K4me1:H3K4me3 ratios and H3K4me3:H3K4me1 ratios at TSSs associated with non-expressed, lowly expressed and highly expressed genes in K562 cells. Ratios were plotted 10,000 bp upstream and downstream of TSSs.

### **3.8.2. Histone modifications and predicting expression status**

In the previous section it was noted that the relative levels of H3K4me1 and H3K4me3 correlated well with transcript levels in K562 cells. This observation was explored further by investigating if the level of these histone modifications at TSSs could be used to predict the expression status of a gene by plotting receiver operating characteristic (ROC) curves (Figure 3.16) (analysis performed by Dr. Ulas Karaöz, Boston University). The ROC of a classifier shows its performance as a compromise between selectivity and sensitivity. In this case, H3ac, H4ac, H3K4me1, H3K4me2, and H3K4me3 binding at TSSs was used as a classifier of gene expression status as described in Chapter 2. A threshold is applied to each histone modification level and a prediction of the on (or off) state of a gene is made if the level is higher (or lower) than the threshold. A curve of false positive rate versus true positive rate is plotted while the threshold parameter is varied and each point on the curve corresponds to a threshold. The best operating point (maximum accuracy value) gives the best trade off between failing to detect true positives against the cost of detecting false positives. The cost of misclassifying positive and negative cases is assumed to be the same resulting in a line with a slope of 1 (45 degrees). The best operating point is the point on a ROC which lies on a 45 degree line closest to the northwest corner of the ROC plot.

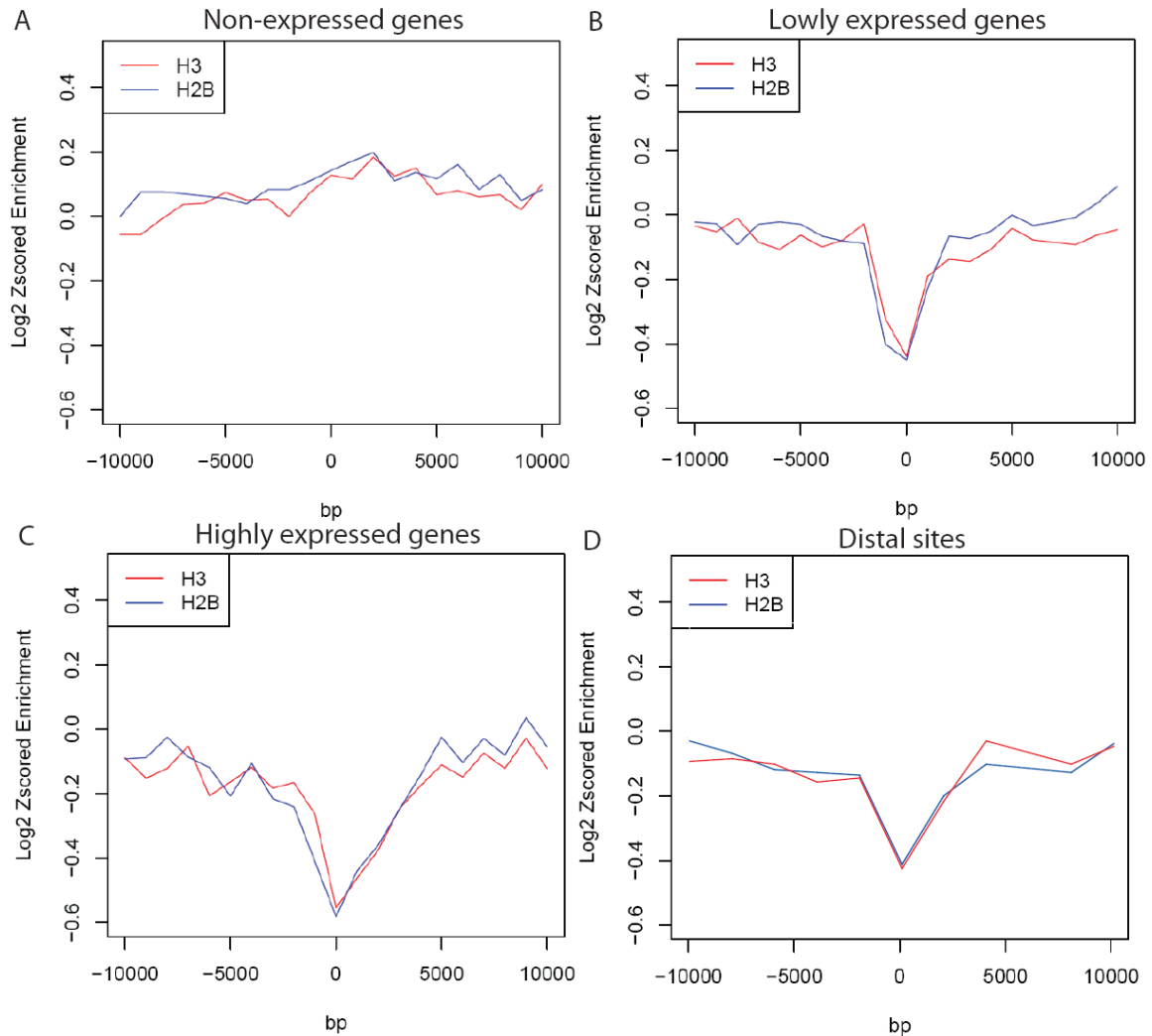
The plots illustrated in Figure 3.16 show sensitivity at all possible specificities and indicate that the presence of H3ac and H3K4me3 at TSSs is highly predictive of active gene expression in K562. The maximum accuracy value is reported for each histone modification and the highest values were calculated to be 0.79 for H3ac and 0.78 for H3K4me3. The presence of H4ac and H3K4me2 at TSSs was also highly accurate at predicting active gene expression. H3K4me1 was the least accurate in terms of predicting active gene expression which is consistent with its role in defining the location of distal enhancer/repressor elements.



**Figure 3.16: Receiver operating characteristic (ROC) curves illustrating the predictive power of histone modifications in K562 cells.** The ROC curve always goes through two points. The first point is 0,0 where a histone modification predicts no true positives (i.e. does not identify any expressed genes) and no false positives (i.e. does not identify any inactive genes as expressed genes). The second point is 1,1 where everything is classified as positive. Here the histone modification correctly predicts all true positive cases but it also wrongly predicts all false positive. A histone modification that is randomly associated with active and inactive gene expression would have a ROC which lies somewhere along the diagonal line connecting 0,0 and 1,1, i.e. when the threshold is raised, an equal number of true and false positives are predicted as positives. A perfect prediction of expressed genes would result in a single point at 0,1. In this case all true positives are found and no false positives are found. The maximum accuracy values for the five histone modifications are indicated.

### 3.8.3. Nucleosome density at transcription start sites and distal elements

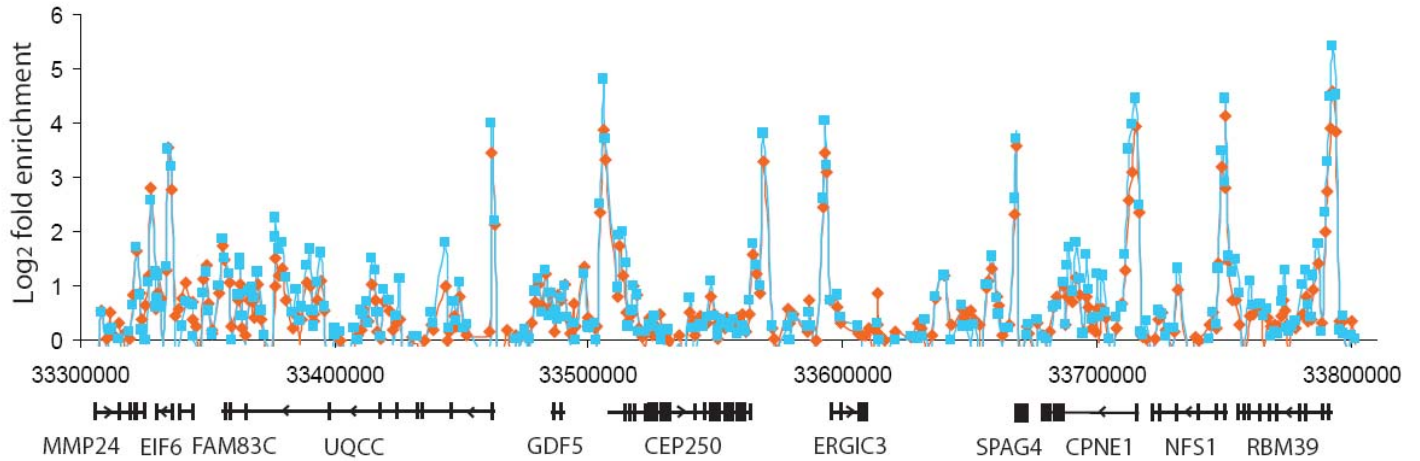
As discussed in the introduction extensive loss of nucleosomes has been observed in the promoters of actively expressed genes in the yeast genome (Lee *et al.*, 2004; Pokholok *et al.*, 2005; Yuan *et al.*, 2005). Nucleosomes consist of a central histone H3-H4 tetramer, which is flanked on either side by H2A-H2B dimers (Luger *et al.*, 1997). In this study, the distribution/density of histones H2B and H3 were examined across the ENCODE regions in K562. Examination of histone H2B and H3 levels at highly expressed, lowly expressed and non-expressed genes showed that relatively lower levels of nucleosomes (i.e. depletion) was a feature of TSSs of expressed genes. The TSSs of highly expressed genes showed the greatest nucleosome depletion, while lowly expressed genes displayed moderate depletion and non-expressed genes showed no nucleosome depletion at the TSS (Figure 3.17). Distal enhancer/repressor elements identified by virtue of their histone modification patterns (see Section 3.7) were also associated with nucleosome depletion.



**Figure 3.17: Nucleosome occupancy at transcription start sites and distal elements in K562 cells.** The average z-scored  $\log_2$  histone H2B and H3 values are presented at TSSs and surrounding regions ( $\pm$  10 kb) of non-expressed (panel A), lowly expressed (panel B) and highly expressed (panel C) genes in K562 cells. No depletion of histone H2B or H3 was observed at the TSS (represented by 0 on the x-axis) of non-expressed genes, while depletion of these two core nucleosome proteins is observed at the TSS of lowly expressed and highly expressed genes. Furthermore, the average histone H2B and H3 values at distal elements (represented by 0) are presented in panel D.

The depletion of nucleosomes at active regulatory elements suggested that the levels of histone modifications observed at these elements may, in part, be a reflection of the density of nucleosomes located therein. Thus, nucleosome density data could be used to normalize histone modification data sets to describe histone modification levels relative to nucleosome density. An example of the effect of normalizing a histone H3 acetylation

data set with an average histone H2B and H3 data set is presented in Figure 3.18. This had the effect of increasing/decreasing H3 acetylation levels relative to regulatory features.



**Figure 3.18: Histone modification levels relative to nucleosome occupancy in K562 cells.** An example of the effect of normalizing histone modification data with nucleosome data is presented for ENCODE region Enr333. Histone H3 acetylation levels are presented before (orange profile) and after (blue profile) nucleosome occupancy normalisation. Fold enrichments are presented as  $\log_2$  values on the y-axis and chromosome 20 coordinates are presented on the x-axis. RefSeq genes located within this region are presented below the x-axis.

### 3.9. Discussion

This Chapter describes the application of ChIP-chip techniques to characterise regulatory elements in the human genome using an array which constituted the 1% of the genome analysed in the pilot phase of the ENCODE project. Five histone modifications were studied which were previously associated with active genes while the core nucleosome proteins, histone H2B and histone H3 were also studied to investigate nucleosome density at regulatory elements. The Sanger Institute ENCODE array, which utilizes single-stranded array technology (Dhami *et al.*, 2005), was shown to be highly reproducible and was used to construct detailed histone modification and histone density maps across 1% of the human genome, with unamplified ChIP DNA used in all assays. The data described in this Chapter provides strong evidence that specific ChIP assays in combination with the ENCODE array can be used to characterise promoter and distal

enhancer/repressor elements on a large-scale. Furthermore, the results obtained from this study provide insights into the regulation of gene expression at the chromatin level.

### **3.9.1. Histone modification signatures associated with promoters and distal elements**

A detailed study of five histone modifications and nucleosome density in approximately 30 Mb of the human genome was performed in this study. Analysis of the distribution of histone modifications in the ENCODE regions revealed the presence of two distinct histone modification signatures - one located at TSS proximal regions and another distinct signature found at locations distal to TSSs. The histone modification signature of TSS proximal regions (i.e. promoters) and distal elements revealed a striking difference in the level of histone modifications. While promoters were associated with elevated levels of H3ac, H4ac, H3K4me2 and H3K4me3 and low levels of H3K4me1, distal regulatory elements were associated with a contrasting signature. The predominant histone modification associated with distal elements was H3K4me1. H3K4me3, H3K4me2, H3ac, and H4ac were also enriched at distal sites but to a lesser extent than H3K4me1. Distal elements identified in this study may represent enhancer elements as the results presented in this thesis are consistent with a recent report in which high levels of H3K4me3 relative to H3K4me1 were found to distinguish the location of promoters while high H3K4me1 levels relative to H3K4me3 were associated with enhancer elements in the HeLa cells (Heintzmann *et al.*, 2007). However, it is also possible that a number of these elements may function as repressors or insulator elements, which are often found at locations distal to TSSs (Bruce *et al.*, 2004; Kim *et al.*, 2007). Further characterization of these insulator elements is performed in Chapter 4.

The identification of characteristic histone modification signatures at promoters and distal enhancer/repressor elements is consistent with the histone code hypothesis which states that “distinct histone modifications, on one or more tails, act sequentially or in combination to form a histone code that is read by other proteins to bring about distinct downstream events” (Strahl and Allis, 2000). H3K4me1 and H3K4me3 (in combination with other modifications) may be responsible for defining the location of active promoter and distal enhancer elements respectively by acting as recognition sites for different

effector proteins which 'read' the histone code. For example chromodomain (CHD) containing proteins, which have been shown to recognize H3K4 methylation, can stimulate gene activation through the recruitment of histone acetyltransferases which create an 'open' local chromatin conformation required for transcription to occur (Flanagan *et al.*, 2005; Pray-Grant *et al.*, 2005). More recently the plant homeodomain (PHD) finger of inhibitor of growth 2 (ING2) and nucleosome remodeling factor (NURF) have been shown to interact specifically with H3K4me3 and modulate gene expression (Shi *et al.*, 2006; Pena *et al.*, 2006; Li *et al.*, 2006; Wysocka *et al.*, 2006). However the histone code is complex as ING2 can recruit both histone acetyltransferase (HAT) and histone deacetylase (HDAC) complexes implicating H3K4me3 in both gene activation and repression. This suggests that the presence of H3K4me3 at promoters alone is not sufficient for active gene expression. Combinations of histone modifications may be important for generating specificity, for example an activating complex may bind H3K4me3 and additional activating histone modifications while a silencing complex may bind H3K4me3 but also simultaneously engage silencing histone modifications. This is consistent with a recent report in which the RNA Polymerase II factor TFIID was shown to bind directly to the H3K4me3 mark via the PHD domain of TAF3 and acetylation of H3K9 and K14 was found to potentiate this interaction (Vermeulen *et al.*, 2007). A protein called L3MBTL1 containing malignant brain tumour (MBT) repeats has been shown to bind mono and di-methyl lysine modifications including H3K4me1 (Li *et al.*, 2007 b), but so far no H3K4me1 specific effector protein has been identified. The identification of factors which recognise the H3K4me1 mark will shed further light on how distal enhancer/repressor elements regulate gene expression. The idea of a strictly deterministic histone code is controversial and the study of other histone modifications is required to understand the possible combinations of modifications which are linked to different downstream events.

### **3.9.2. Histone modifications and transcriptional activity**

Analysis of the histone modification patterns revealed that promoter regions of active genes have a characteristic histone profile which was distinct from inactive genes. H3 acetylation and H3K4me3 were most enriched over the promoter region of highly



expressed genes, while the promoter regions are relatively depleted of nucleosomes. H3K4me2 was found to be most enriched just downstream of the promoter while low levels of H3K4me1 are observed further downstream from the TSS. A similar pattern of H3K4 methylation was observed at the promoters of yeast genes, in which H3K4me3 was found to peak closest to the TSS of active genes, while H3K4me2 and H3K4me1 were located further downstream (Liu *et al.*, 2005). This methylation pattern is also consistent with results obtained from studies performed with human cells (Heintzman *et al.*, 2007; Barski *et al.*, 2007). Moderately expressed genes have a similar histone signature but H3ac, H3K4me2, and H3K4me3 are enriched to a lesser degree suggesting that the level of these modifications correlates with transcriptional activity. In addition ROC analysis of the histone modifications showed that the presence of H3ac and H3K4me3 at TSS was highly predictive of active gene expression. The ratio of H3K4me3:H3K4me1 at TSSs could also be used to predict gene activity; a high ratio indicated active gene expression while inactive genes were associated with no increase in this ratio. The TSSs of inactive genes do not display this histone modification signature but instead are associated with very low levels of H3ac and H3K4me3 while nucleosomes are not depleted at inactive TSSs. The presence of low levels of H3ac and H3K4me3 at inactive TSSs suggests that these genes may be primed for future expression. A recent ChIP-chip study of promoters in human embryonic stem cells and differentiated cells showed that the majority of promoters were associated with H3K4me3, H3ac, and RNA polymerase II irrespective of expression status (Guenther *et al.*, 2007). This suggests that inactive genes may be primed for expression by the presence of low levels of these histone modifications, which could then be up-regulated when future expression is required.

The histone modification signature of active promoters could be used to improve gene annotation by identifying new TSSs for genes that are being expressed when applied to whole-genome studies of different cell types. A recent genome-wide ChIP high-throughput sequencing study has shown that H3K4me3 modification maps in conjunction with H3K36me3 maps could be used to identify novel transcription units in the mouse genome (Mikkelsen *et al.*, 2007). In addition, the ENCODE consortium has shown that it is possible to train a support vector machine to make effective predictions about the

location and activity of TSSs based on the presence of this histone modification data (Birney *et al.*, 2007).

### **3.9.3. Nucleosome depletion at regulatory elements**

To gain further insight into the role of nucleosomes in gene regulation in the human genome, nucleosome density at promoters and distal elements was evaluated by immunoprecipitating histone H2B and H3 DNA and quantifying enrichment with microarrays. The investigation of two histone proteins, which display almost identical occupancy profiles, suggests that the ChIP-chip data sets reflect nucleosome occupancy and are not related to other issues such as epitope accessibility, histone variants or cross-linking efficiency. Promoters of active genes and distal elements were associated with nucleosome depletion, which is consistent with observations in yeast (Bernstein *et al.*, 2004; Lee *et al.*, 2004; Pokholok *et al.*, 2005; Yuan *et al.*, 2005) and more recently in a human cell line (Heintzmann *et al.*, 2007). The loss of nucleosomes at active regulatory elements may be facilitated by a transcription factor binding to its cognate site (Lee *et al.*, 2004), by acetylation of histones (Reinke and Horz, 2003), by ATP-dependent nucleosome remodeling complexes (Lusser and Kadonaga, 2003) or by the initiation of transcription by RNA polymerase.

Furthermore, the lower density of nucleosomes observed at promoters of active genes and distal elements provides a general caveat for examining histone modifications in ChIP-chip studies. The majority of studies assume a homogenous distribution of nucleosomes, which does not seem to be the case and regions which appear to be relatively hypomodified for a histone modification may actually be nucleosome depleted. Normalisation of histone modification levels relative to nucleosome occupancy is important, particularly for studies in which histone modifications are mapped at the resolution of individual nucleosomes.

### **3.9.4. Analysis and interpretation of ChIP-chip data**

The work described in this Chapter illustrates how ChIP-chip assays can be used to characterise the chromatin state of promoters and distal enhancer/repressor across 1% of the human genome. However, in order to identify and characterise these regulatory

elements it was important to have in place appropriate analysis procedures to deal with the large volume of data generated by these experiments. This ChIPOTle algorithm was used to identify histone modification peaks in ENCODE data sets. While the identification of histone modification peaks by ChIPOTle lead to the discovery of histone signatures associated with promoters and distal sites, there may be other histone modifications features which were not picked up by this peak-finding approach – for example enrichments which are low level, and/or encompass large continuous stretches of genomic DNA. In this case other analysis approaches such as hierarchical clustering could be used to identify such features in the future.

In addition to issues of interpreting ChIPOTle peak information; there are a number of other considerations which must be taken into account when interpreting ChIP-chip data:

i) **Heterogeneous cell populations:** The ChIP-chip assays described in this Chapter only provide a snapshot of the histone modifications occurring within a cell population at a particular point in time. In other words ChIP-chip assays for histone modifications provides a survey of the characteristics of an entire cell population and does not necessarily mean that co-localising modifications are present within the same cells.

ii) **Microarray resolution:** Given that the resolution of the array used in this study was approximately 1kb, this suggests that an average of five nucleosomes is sampled per array element. Thus in this study, it is not possible to state categorically that combinations of histone modifications on different residues (for example H3K4me3 and H3K9/K14 acetylation) are present on a single histone tail. Liu and colleagues used micrococcal nuclease digested chromatin and a high resolution array to map histone modifications at the resolution of single nucleosomes in the yeast genome (Liu *et al.*, 2005). However, even with this method it is still not possible to say without doubt that combinations of modifications occur on the same nucleosome tail due to the effect of sampling a population of cells. The co-localisation of two histone modifications in the same nucleosome can be established by performing sequential ChIP (Bernstein *et al.*, 2006) in which an antibody to one histone modification is used to immunoprecipitate micrococcal nuclease digested chromatin fragments, which are then exposed to a second antibody followed by hybridization to a high resolution array.

iii) **Antibody efficiencies:** Although the antibodies used in this study may have been shown to be specific by peptide competition (Koch *et al.*, 2007) and perform well in ChIP-chip assays, it is difficult to compare enrichment levels between assays as a measure of the true level of histone modifications. This is because antibodies may have different efficiencies for their target epitope and may enrich genomic sequences to different levels in the ChIP DNA samples. Z-scored ChIP-chip data helps to correct for different antibody efficiencies by standardizing data sets.

iv) **Allele-specific histone modifications:** allele-specific histone modification patterns may be present in the human genome as was recently demonstrated in the mouse genome (Mikkelsen *et al.*, 2007). In some situations, such as the study of imprinted genes, the analysis of allele-specific histone modification patterns is important and would not be detected using standard ChIP-chip methods as a composite profile of the two alleles is presented in the data. This problem can be overcome by the availability of allele-specific single nucleotide polymorphism (SNP) data and high-throughput sequencing of ChIP DNAs to distinguish allele-specific histone modification patterns. Alternatively, Kadota and colleagues have shown that allele specific histone modification patterns can be investigated by hybridizing ChIP DNAs to an array used for examining SNPs (Kadota *et al.*, 2007).

### 3.9.5. Conclusions

The work described in this Chapter establishes that the ENCODE microarray is a sensitive and reproducible platform which can be used in ChIP-chip assays to characterise promoter and distal enhancer/repressor elements at the histone modification level. This suggests that it could be used to provide further insight into other DNA-protein regulatory interactions in the human genome, in particular interactions associated with insulator elements. This is the subject of the following Chapter.

## Chapter 4

### Identification and Characterisation of Binding Sites of the Insulator Protein CTCF in the Human Genome

#### 4.1. Introduction

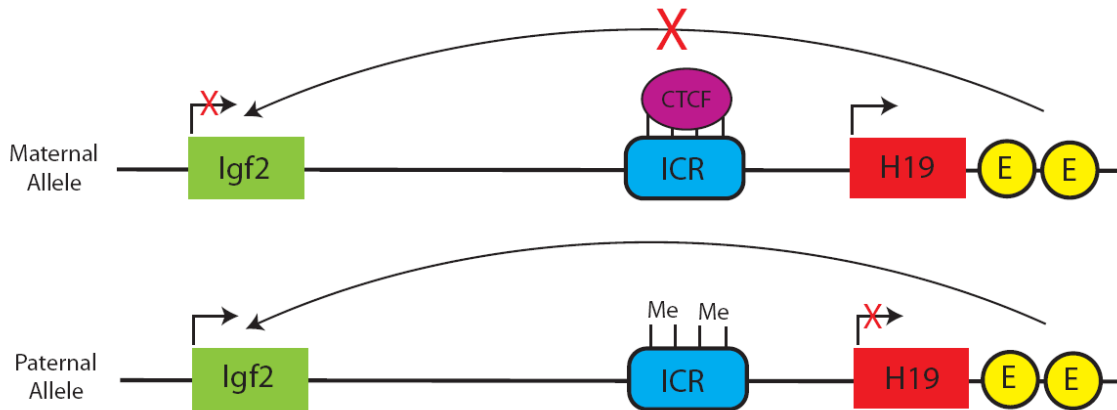
In addition to promoter and enhancer/repressor elements, insulators constitute another major class of regulatory elements found in the human genome. Insulators regulate interactions between promoters and enhancers by preventing inappropriate enhancer/promoter contact as well as acting as boundary elements to prevent the spread of silencing heterochromatin (Burgess-Beusse *et al.*, 2002). They can insulate entire genes or a cluster of genes from the influence of heterochromatin as well as facilitating the establishment of complex cell specific gene expression patterns when individual enhancers are flanked by insulator elements (Brasset and Vaury, 2005). Sequences that prevent inappropriate activation by enhancers have been termed enhancer blocking insulators and those which prevent the spread of heterochromatin have been termed barrier insulators (Sun and Elgin, 1999). Enhancer blocking insulators may function by looping of chromatin into distinct regulatory domains that prevent inappropriate enhancer-promoter communication (Cai and Shen, 2001; Gruzdeva *et al.*, 2005; Kurukuti *et al.*, 2006), while barrier elements may function by recruiting histone modifiers which deposit histone modifications associated with active chromatin, thus preventing the spread of heterochromatin (West *et al.*, 2004).

CCCTC-binding factor (CTCF) is a widely expressed 11 zinc-finger nuclear protein that was first identified by its ability to bind to the promoters of chicken, mouse, and human MYC genes (Filippova *et al.*, 1996; Klenova *et al.*, 1993, Lobanekov *et al.*, 1990). Initial characterisation of CTCF revealed that it could act as both a transcriptional repressor (Burcin *et al.*, 1997) and activator (Vostrov and Quitschke, 1997). Subsequently it was found to bind the HS4 insulator of the chicken  $\beta$ -globin locus (Bell *et al.*, 1999) and since then has been found to bind all known vertebrate insulator elements (Bell *et al.*, 1999; Bell *et al.*, 2001; Mukhopadhyay *et al.*, 2004) as well some insulator elements in *Drosophila* (Moon *et al.*, 2005; Holohan *et al.*, 2007). CTCF was

shown to bind to diverse and long (~50bp) DNA sequences by using different combinations of its individual zinc fingers (Burcin *et al.*, 1997; Filippova *et al.*, 1996). This multiple sequence specificity of CTCF is believed to mediate its classical transcription factor and insulator functions through the formation of distinct CTCF complexes at different CTCF sites (Gaszner and Felsenfeld, 2006; Filippova, 2008).

Binding of CTCF was found to be necessary for enhancer blocking function of the chicken beta-globin locus HS4 insulator (Bell *et al.*, 1999) and was separable from the barrier function which prevented the spread of heterochromatin (Recillas-Targa *et al.*, 2002, West *et al.*, 2002). Two models for insulator activity have been proposed - the chromatin loop domain model and the tracking model. The chromatin loop domain model is based on the ability of CTCF to form chromatin loop domains by interacting with nucleolar structural components (Dunn *et al.*, 2003; Yusufzai *et al.*, 2004) or with other CTCF sites (Kurukuti *et al.*, 2006; Ling *et al.*, 2006). Enhancer-blocking activity is conferred by positioning of promoters and enhancers into separate chromatin loop domains. The tracking model was proposed based on the ability of CTCF to interact with the transcriptional machinery and block the transfer of RNA polymerase II between an enhancer and a promoter (Zhao and Dean, 2004).

The role of CTCF in enhancer blocking is also important for the coordination of gene expression patterns at imprinted gene clusters in mammalian genomes which are regulated by imprinting control regions (ICRs) - also known as differentially methylated domains (DMDs) (Ohlsson *et al.*, 2001; Reik and Walter, 2001). CTCF was found to bind to an ICR upstream of the H19 gene and block access of Igf2 to an enhancer shared with H19 which results in no Igf2 expression from the maternal allele (Figure 4.1) (Hark *et al.*, 2000; Bell *et al.*, 2000; Kanduri *et al.*, 2000). CpG methylation of the ICR on the paternal allele prevents CTCF binding, allowing enhancer-mediated activation of the Igf2 promoter on the paternal allele (Bell *et al.*, 2000). CTCF has also been shown to prevent the spread of DNA methylation thus playing a crucial role in maintaining methylation free regions (Engel *et al.*, 2004; Pant *et al.*, 2004; Filippova *et al.*, 2005). Thus CTCF can also prevent nearby promoters from being epigenetically silenced.



**Figure 4.1: CTCF enhancer-blocking insulator at the *Igf2*/*H19* imprinted region.** CTCF binds to sites in the ICR in the maternal allele and prevents downstream enhancers (E) from activating *Igf2* expression. The ICR is methylated (Me) in the paternal allele which prevents CTCF binding which means that the downstream enhancers are no longer blocked from interacting with the *Igf2* promoter.

The 5' HS4 chicken beta-globin insulator element displays both enhancer blocking and barrier insulator functions. The enhancer blocking and barrier functions were found to be separable activities and CTCF was necessary for the enhancer blocking activity of this insulator but was not required for barrier activity (Recillas-Targa *et al.*, 2002). The binding of USF1 and USF2 proteins was later found to be required for the barrier activity of the HS4 insulator and they were shown to recruit histone modifying complexes responsible for H3K4 and H4R3 methylation and histone acetylation, which prevented the spread of histone modifications associated with nearby condensed chromatin (West *et al.*, 2004, Huang *et al.*, 2007). However it is not clear if USF1 and USF2 are responsible for the barrier activity of other insulators. CTCF may be required for both enhancer blocking and barrier activity in other vertebrates. It has not been directly shown that CTCF prevents the spread of heterochromatin, but CTCF binding sites have been found close to the transition between active and silent chromatin domains (Filippova *et al.*, 2005; Barksy *et al.*, 2007). CTCF-mediated barrier activity would reconcile with the chromatin loop model of enhancer blocking as flanking a gene with a CTCF binding site could also provide barrier activity by creating an independent expression domain.

CTCF can bind to different DNA sequences using various combinations of the 11 zinc finger domains (Filippova *et al.*, 2008). CTCF complexes formed at these different DNA

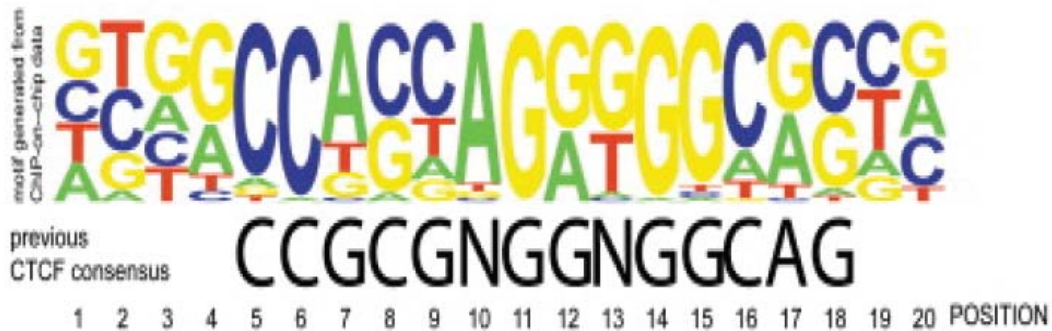
sequences may be functionally different depending on CTCF interaction with other proteins. The use of different combinations of zinc fingers may mean that different combinations of zinc fingers are available for interaction. CTCF has been shown to interact with nucleophosmin (Yusufzai *et al.*, 2004), Kaiso (Defossez *et al.*, 2005), and CHD8 (Ishihara *et al.*, 2006) and these interactions have been linked to the regulation of CTCF insulator function. The transcription factor YY1 has been shown to frequently bind in close proximity to CTCF sites and is an important co-factor in CTCF mediated regulation of X inactivation (Donohoe *et al.*, 2007).

CTCF is also known to interact with the transcription factor mSin3a (Lutz *et al.*, 2000). The mammalian Sin3 proteins, mSin3A and mSin3B were discovered as a result of their interaction with the transcriptional repressors Mad1 and Mxi1 (Ayer *et al.*, 1995). They were later shown to associate with HDAC1 and HDAC2 to form a large multiprotein complex, the Sin3/histone deacetylase (HDAC) co-repressor complex (Silverstein and Ekwall, 2005). CTCF interaction with mSin3a may be responsible for CTCF mediated transcriptional repression via recruitment of HDACs.

Identifying the location of putative insulator elements in the human genome would greatly increase our understanding of how this class of *cis*-acting regulatory elements controls genome structure and function. However, computational prediction of insulator location in the human genome is a difficult task as CTCF can use various combinations of zinc fingers to bind different target sequences (Ohlsson *et al.*, 2001). Mukhopadhyay and colleagues (2004) used ChIP cloning and sequencing to identify 200 CTCF binding sequences with enhancer-blocking activity in the mouse genome but no consensus binding motif was identified by this study. Xie and colleagues used a systematic approach to discover and characterise regulatory motifs within mammalian conserved non-coding elements (CNEs) by searching for long motifs (12-22 nt) with significant enrichment in CNEs (Xie *et al.*, 2007). One of these motifs (CCACTAGATGGCA) was found to at 15,000 conserved locations in the human genome and was found to experimentally bind CTCF. Kim and colleagues performed a genome-wide ChIP-chip analysis of CTCF binding sites in the human genome and identified over 13,000 binding sites (Kim *et al.*, 2007). This large data set was used to define a 20-mer CTCF consensus binding motif, which was found at over 75% of their experimentally determined binding sites in the



human genome and was able to bind recombinant CTCF. This consensus 20 bp motif was similar to a 14bp GC rich sequence previously defined based on a limited number of characterised sites (Bell and Felsenfeld, 2000) but was refined at a number of positions (Figure 4.2). The 13 bp motif identified by Xie and colleagues (2007) was also similar to the core of this 20bp consensus motif.



**Figure 4.2: A 20-mer motif is recognised by CTCF.** A DNA logo representing the CTCF-binding motif derived from a genome-wide ChIP-chip study (Kim *et al.*, 2007) and the previously reported consensus CTCF-binding sites (Bell and Felsenfeld, 2000) are shown. The relative frequency at which each nucleotide occurs in the motif is indicated by height of letter at each position. The ChIP-chip derived motif refines the previous consensus motif at six nucleotide positions (positions 7, 8, 9, 10, 13, and 17). Figure from Kim *et al.*, 2007

Kim and colleagues used their experimentally determined motif to scan the human genome sequence and a total of 31,905 sites were found to contain this motif, 12,799 of which were conserved in at least one other vertebrate genome. This suggests that the location of a large number of insulators may be conserved between vertebrates.

When work towards this PhD thesis was initiated very little was known about insulators and the binding of CTCF in the human genome. Therefore as a logical step towards understanding where insulators are found in the human genome, a ChIP-chip assay would need to be developed for CTCF. Although the work of Kim and colleagues (2007) greatly aided in the discovery of CTCF binding events in the human genome, understanding how CTCF interacts with other proteins at insulators still remains largely unexplored genome-wide. Therefore, in order to further characterise CTCF binding events, ChIP-chip assays to detect binding sites for the CTCF-associated proteins mSin3a, USF1, and USF2 were

investigated. This information in combination with data on histone modifications associated with active and inactive chromatin domains was used to further characterise CTCF binding events and determine whether there were features, apart from CTCF-binding, which distinguished insulators from the other major classes of regulatory sequences.

#### **4.2. Aims of this chapter**

In order to identify and characterise CTCF binding sites in 1% of the human genome the aims of the work presented in this chapter were as follows:

1. To develop a ChIP-chip assay for the detection of CTCF binding sites using the SCL tiling path array as a model.
2. To apply this assay for the identification of CTCF binding sites in the 1% of the human genome covered by the ENCODE regions.
3. To further characterise CTCF interactions by developing additional ChIP-chip assays to investigate the binding of the known CTCF interacting partner mSin3a and the barrier insulator proteins USF1 and USF2.
4. To investigate the conservation of CTCF binding sites in human cell lines.
5. To investigate the histone modifications associated with CTCF binding events at insulators.

#### **4.3. Overall strategy**

As discussed in Chapter 3, the SCL locus tiling path array was previously used to develop ChIP-chip assays for the detection of histone modification events associated with promoter and enhancer elements (Dhami, submitted). However, because of a lack of information in the literature on histone modifications that help define insulators, it was necessary to develop an assay which is insulator specific. Therefore a ChIP-chip assay was developed using the SCL locus as a model system for the identification of DNA sequences interacting with the insulator binding protein CTCF in K562 cells. The SCL region was used initially to develop the assay and determine whether the CTCF binding data reconciles with what is already known about regulatory features in the region. In order to gain a greater understanding of the location of putative insulator elements

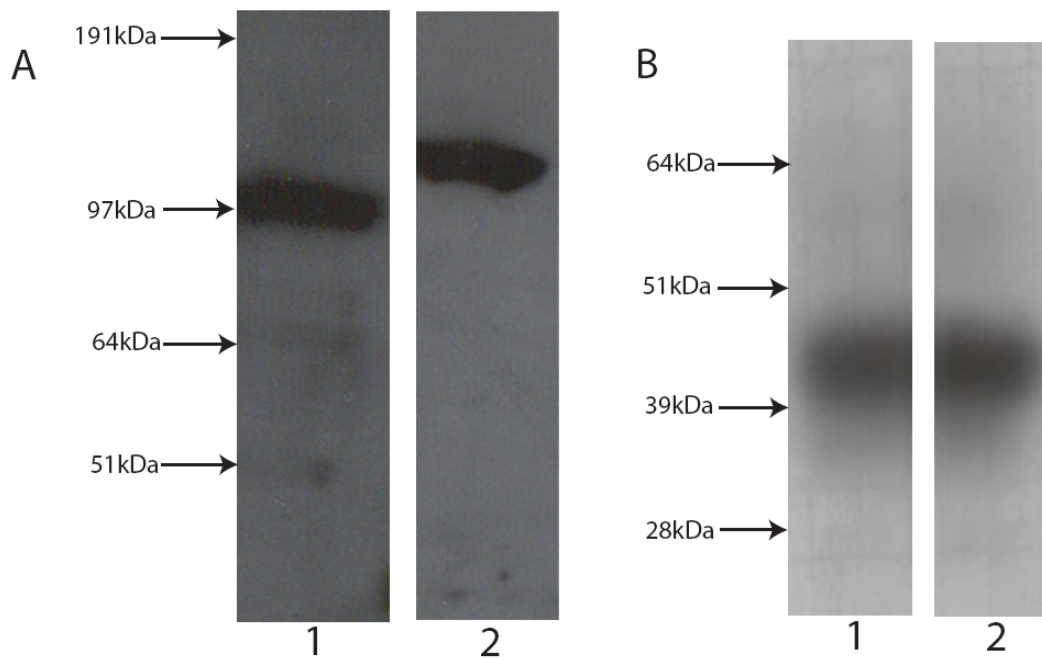
relative to genes and other epigenetic and genomic features, this assay was then applied to detect CTCF interactions in K562 cells in the 1% of the human genome covered by the Sanger Institute ENCODE genomic tiling array. CTCF interactions were further characterised by performing ChIP-chip experiments to identify binding locations of the CTCF interacting partner mSin3a and the barrier proteins USF1 and USF2 (Lutz *et al.*, 2000; West *et al.*, 2004; Huang *et al.*, 2007). CTCF binding sites were also examined by in the human cell line U937 and compared with K562 binding sites to determine if the location of insulators is conserved between different cell types. All of this data was then compared to that generated from other sources, both within our laboratory and in other laboratories, which includes the location of DNase I hypersensitive sites, predicted CTCF sites, histone modifications, formaldehyde assisted isolation of regulatory elements (FAIRE), and nucleosome data.

## **Results**

### **4.4. Assessing the specificity of transcription factor antibodies in western blotting assays**

In order to ensure that ChIP-chip assays could detect *bona fide in vivo* CTCF, mSin3a, USF1 and USF2 binding sites, it was important to verify the specificity of the antibodies to be used in ChIP-chip assays, and to avoid cross-reactivity with proteins which share amino acid sequence similarity. To this end western blotting assays were performed. The mSin3a protein is theoretically 145 kDa in size and the antibody used in this study detected a single band at approximately the theoretical size in K562 cells (Figure 4.3). USF1 encodes a protein of molecular mass 43 kDa and this antibody detected one diffuse band in K562 nuclear extracts which was at the correct molecular weight. USF2 encodes a protein of 44 kDa molecular and a single diffuse band at the predicted molecular weight was also detected with K562 nuclear extracts– this band was of similar size as that detected for USF1. Given that USF1 and USF2 are of a similar molecular weight, it was not possible to determine for certain that the antibodies were not cross-reacting. CTCF encodes a protein of theoretical molecular mass 82 kDa and a single band at approximately 97kDa was detected in K562 cells. CTCF has been found to migrate aberrantly in SDS-PAGE as it has been observed to migrate at 130, 97, 80, 73, 70 and 55

kDa's (Klenova *et al.*, 1997). Furthermore, CTCF is known to be poly-ADP ribosylated (Yu *et al.*, 2004) and this large post-translational modification would affect the observed mass of the protein in western blotting assays. All of this data taken together suggests that the CTCF antibody used in this study was likely to be detecting the *bona fide* CTCF protein.



**Figure 4.3: Western blot analysis of CTCF, mSin3a, USF1 and USF2 in K562 cells.** The CTCF antibody used in this study detected at single band at approximately 97kDa in K562 cells (Panel A, lane 1). Panel A, lane 2 shows that the mSin3a antibody detects a single band at approximately the theoretical molecular mass of 145 kDa. Panel B lanes 1 and 2 show that the USF1 and USF2 antibodies detect single bands at the theoretical molecular masses of 43 and 44kDa's respectively.

#### **4.5. Developing a ChIP-chip assay to detect putative insulators at the SCL locus**

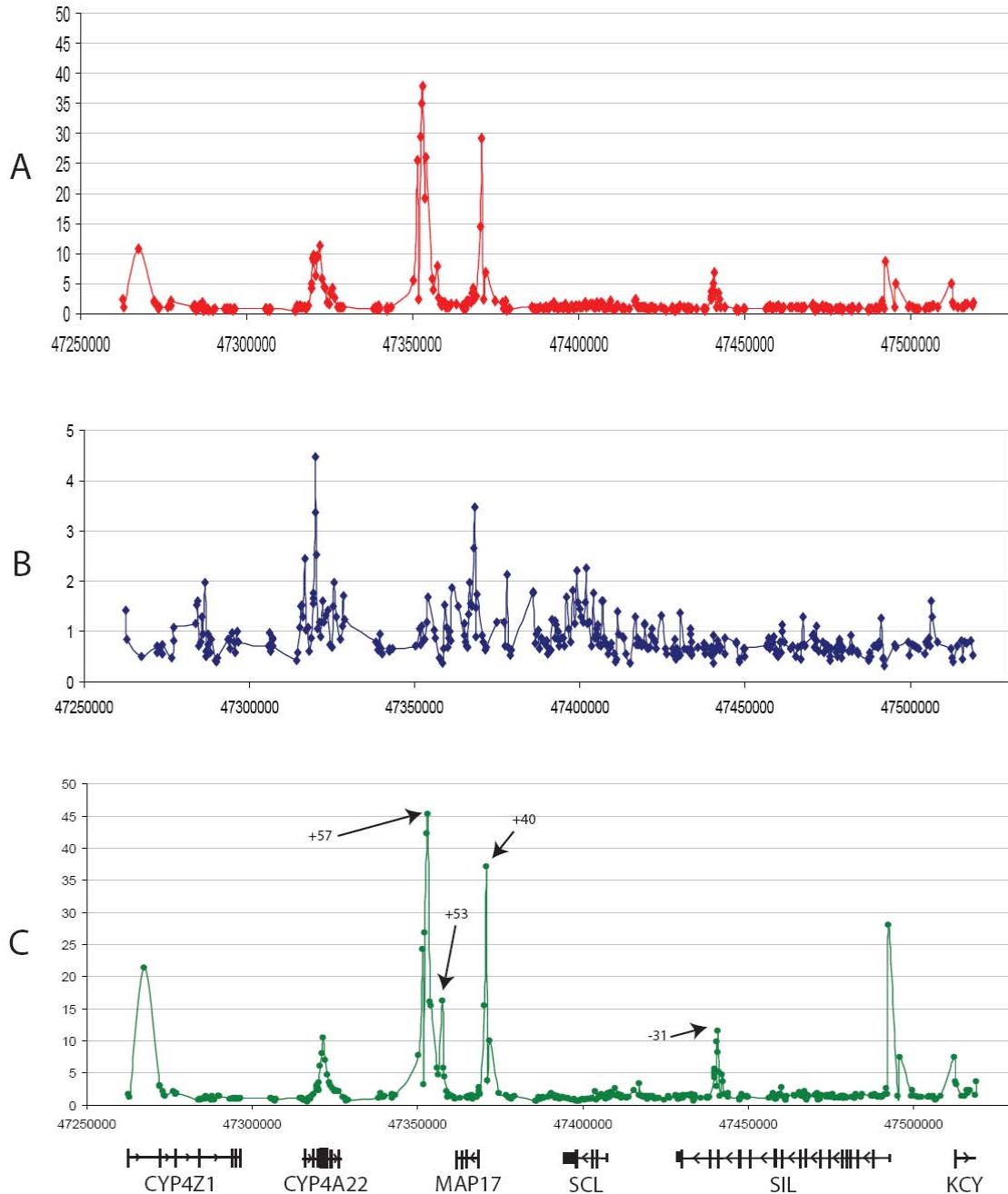
The SCL locus is well characterised in terms of promoter and enhancer elements based on ChIP-chip studies using a tiling array of the SCL locus (Dhami, submitted). However, no CTCF binding-elements (i.e. putative insulators) had been characterised in this region at the start of this PhD project. This same SCL tiling array was used to test for the

presence of insulators because a number of genes present on this array were expressed in a tissue specific manner, for example SCL is expressed in blood, in endothelium, and within specific regions of the central nervous system (Begley and Green, 1999) while the gene for the cytochrome P450 family member CYP4A22 (also tiled on the SCL array) is expressed in liver tissue (Savas *et al.*, 2003). Therefore it was hypothesized that the CYP4A22 and SCL genes may require insulators to regulate their expression patterns – thus, a ChIP-chip assay was developed for studying CTCF interactions in the SCL expressing cell line K562.

Eight regions were identified as significantly enriched (significant enrichments in transcription factor ChIP-chip experiments were considered to be those values that were more than three standard deviations away from the mean ratio of background levels) for CTCF binding across the SCL locus (Figure 4.4, panel A). As the CTCF antibody used in this study was raised in goat, a ChIP-chip ‘mock’ antibody control experiment was also performed with a normal goat IgG antibody to identify any non-specific enrichments associated with performing ChIP-chip experiments with a goat IgG (Figure 4.4, panel B). Data from this mock antibody control experiment was used to normalise for non-specific interactions by dividing the CTCF data set values with the corresponding goat IgG values. This approach ensured that non-specific enrichments observed in both data sets were normalised to background values. Fold enrichments for seven of the eight sites were increased following normalisation with the goat IgG data set (Figure 4.4, panel C) while one region at the CYP4A22 gene remained unchanged indicating that all 8 regions represent *bona fide* sites of CTCF interaction and non-specific interactions by a goat IgG. Significant peaks of enrichment were found at a region within the SIL gene, 31 kb upstream of SCL promoter1a (-31), and at +57 between the SCL erythroid enhancer at +51 and the CYP4A22 gene. Peaks were also found within the CYP4A22 and CYP4AZ1 genes, at the SIL and KCY promoters, at +53 region, and the MAP17 enhancer (+40).

The +53 CTCF region is associated with high levels of H3 acetylation, H3K4me2 and H3K4me3 consistent with promoter activity and novel transcripts that have been identified near this region (Dhami, submitted). In contrast the +40 CTCF region is associated with low levels of H3 acetylation and H3K4me2/H3K4me3 but displays high levels of H3K4me1, consistent with enhancer function (Chapter 3) (Follows *et al.*, 2006).

The +57 and -31 CTCF sites do not display either of these histone modification profiles suggesting that these CTCF binding regions may be functionally distinct. The genomic regions contained within +57 and -31 defines a 98kb regulatory domain containing SCL, MAP17 (thought to be co-regulated with SCL) and all known SCL regulatory elements. None of the nearby genes outside of this domain are thought to share regulatory elements, suggesting that the presence of CTCF at these sites marks the location of insulators.



**Figure 4.4: ChIP-chip profile of CTCF binding across the SCL locus in K562 before and after Goat IgG normalisation.** Panel A: Fold enrichments reported for CTCF interactions before normalisation with normal goat IgG. Panel B: A normal goat IgG ChIP-chip profile showing non-specific enrichment at a number of locations. Panel C: fold enrichments increased at CTCF binding sites after normalisation with goat IgG. The location of -31, +40, +53 and +57 regions are indicated by black arrows. The human chromosome 1 genomic coordinates are indicated along the y-axes, while fold enrichments are indicated on the x-axes. Gene order and direction of transcription is shown below panel C.

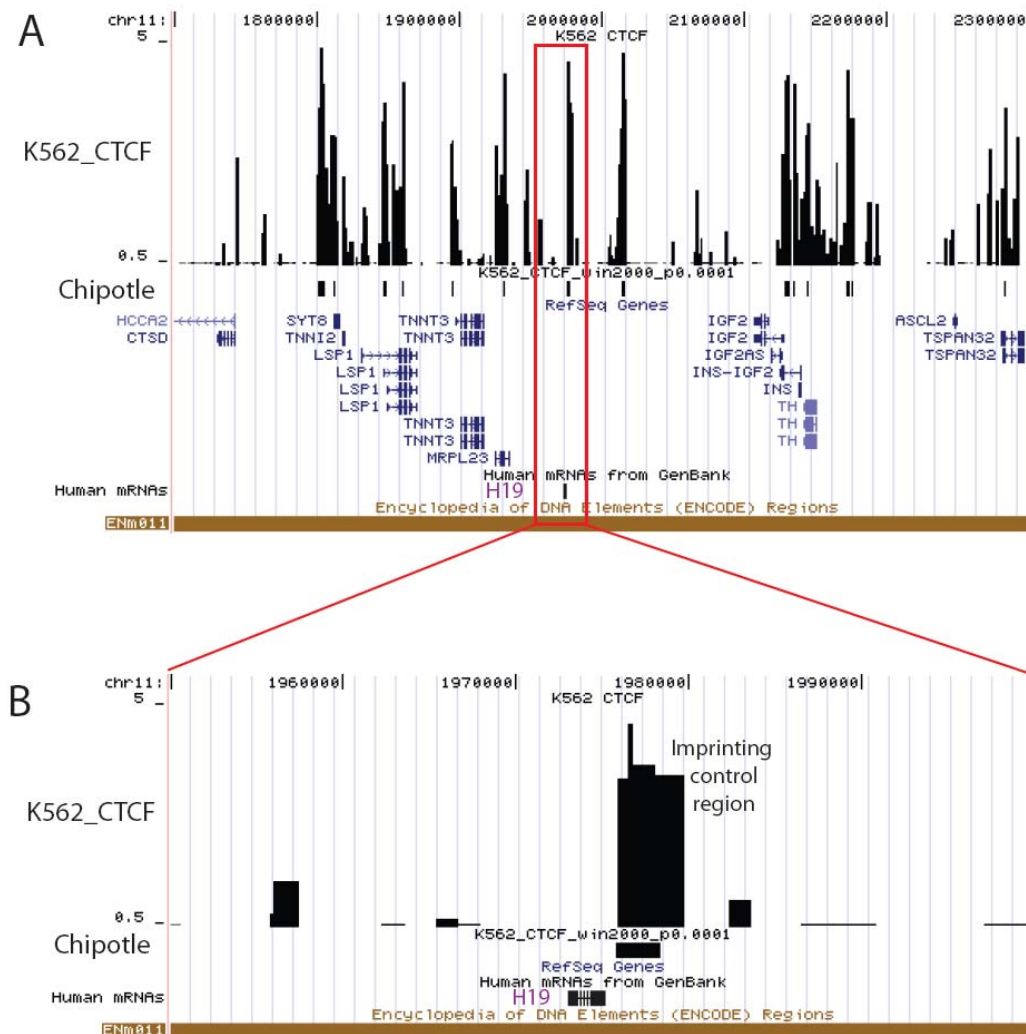
## **4.6. Mapping and characterising CTCF binding sites in the ENCODE regions**

### **4.6.1. Implementation and validation of the CTCF ChIP-chip assay**

Following the identification of a number of CTCF sites across the SCL locus, it was important to gain a more complete understanding of the genome-wide binding patterns of CTCF. The Sanger Institute ENCODE array (chapters 1 and 3) was used to analyse the binding patterns of CTCF in 1% of the human genome. ChIP-chip assays were performed with the CTCF antibody across three biological replicates of K562 chromatin as described previously. The ChIP DNA were hybridised to the ENCODE array and the median values of the three hybridisation experiments were normalised with goat IgG values to eliminate non-specific enrichments. To identify CTCF binding sites, across 1% of the human genome, the ChIP-chip normalised data sets were analysed by CHIPOTle (Buck *et al.*, 2005). As discussed in Chapter 3, this program was developed specifically to analyse ChIP-chip data and uses a sliding windows approach to identify peaks of enrichment and then estimates the significance of enrichment for a region using a Gaussian error function. Using this analysis tool 571 CTCF sites were identified as significantly enriched across the ENCODE regions in K562 cells using a stringent p-value cut-off of p0.0001. As the ENCODE regions represent 1% of the human genome, this figure suggests that there may be more than 50,000 binding sites in the entire genome.

In order to verify the specificity of the ChIP-chip method used in this study, known CTCF binding sites in the ENCODE regions were investigated for binding in K562 cells. Eight closely associated CTCF sites have been characterised at the IGF2/H19 imprinting control region by a number of groups (Bell and Felsenfeld, 2000; Hark *et al.*, 2000, Szabo *et al.*, 2000), spanning a 4 kb region upstream of the H19 locus (coordinates

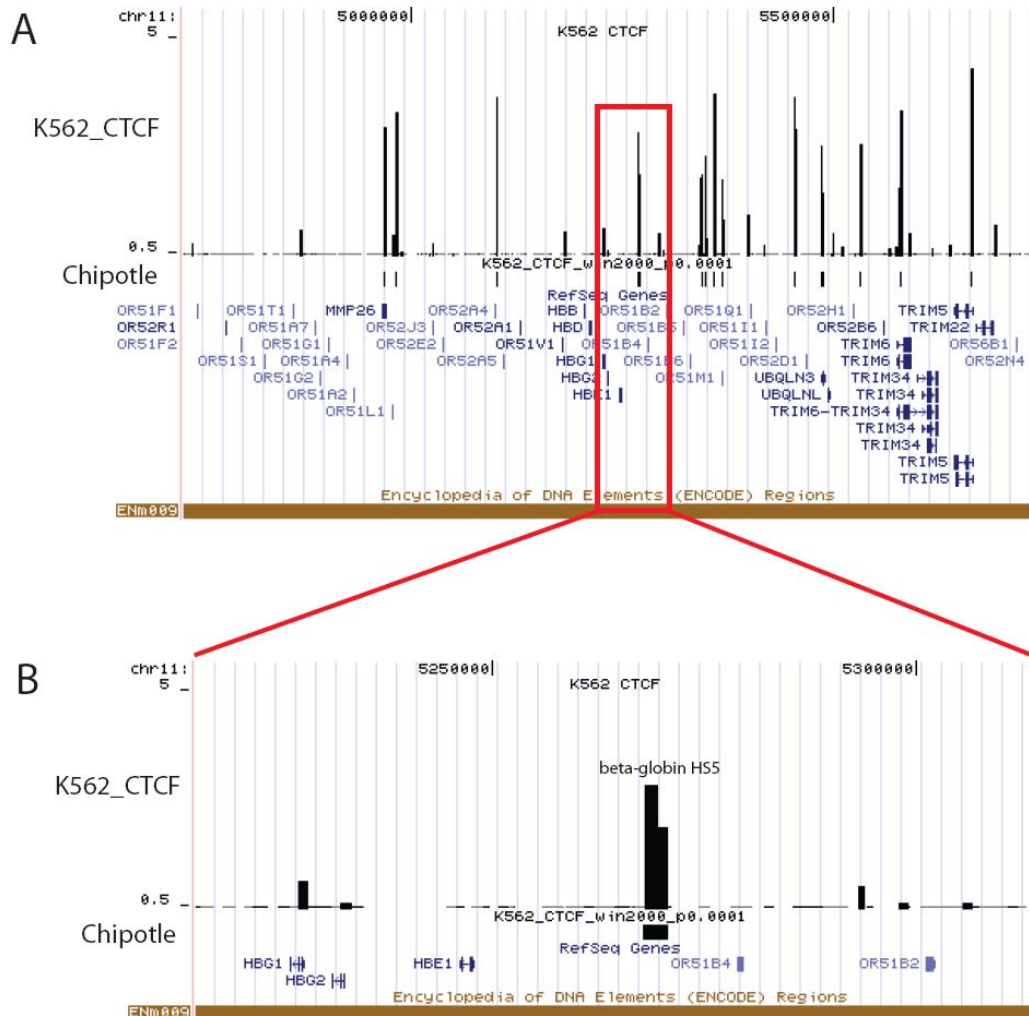
chr11:1976843-1980864). The CTCF binding profile at the IGF2/H19 locus in K562 is shown in Figure 4.5. A region upstream of the H19 gene shows a highly enriched CTCF binding peak identified by ChIPOTle, which correlates with previously characterised CTCF binding sites in the imprinting control region.



**Figure 4.5: CTCF binding sites in the IGF2/H19 locus.** Panel A illustrates the CTCF binding profile in ENCODE region Enm011 (IGF2/H19 locus). Log<sub>2</sub> fold enrichments are represented in the top half of the panel (K562\_CTCF track) and sites identified by ChIPOTle (p0.0001) are indicated below the x-axis (Chipotle track). Known RefSeq genes and the location of the H19 maternally transcribed mRNA are indicated at the bottom of panel A. A magnified view of CTCF binding at the H19 locus is illustrated in panel B. A highly enriched peak of CTCF binding is observed at the imprinting control region.

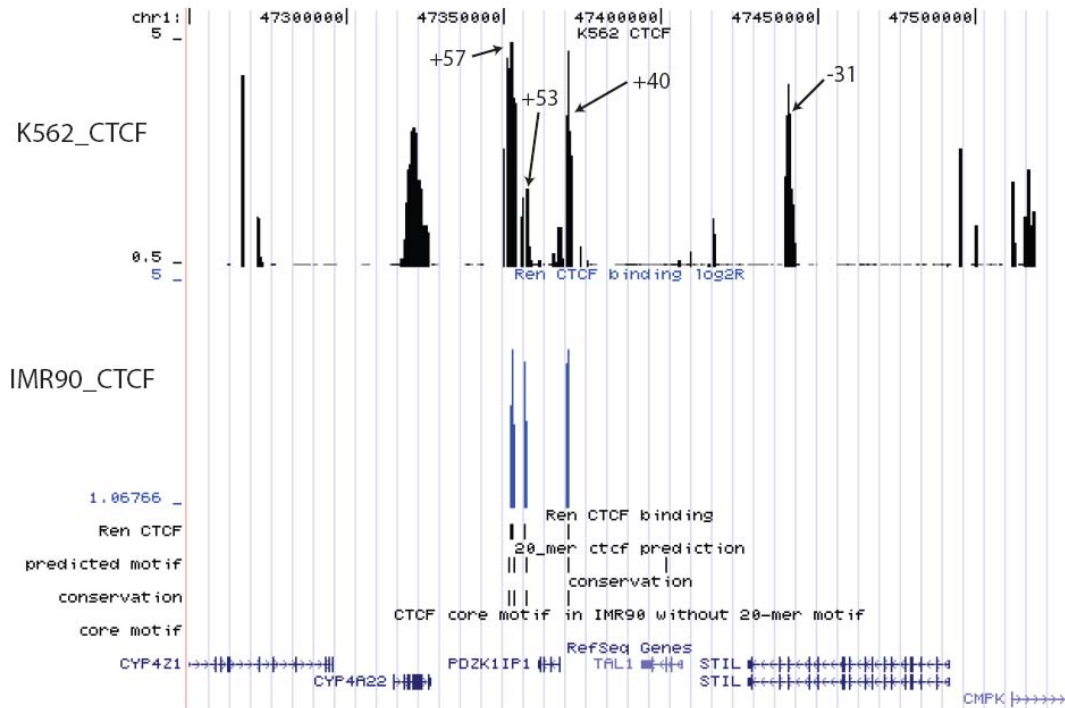


In addition to the IGF2/H19 CTCF binding sites, a CTCF binding site has previously been characterised at the  $\beta$ -globin locus (beta-globin HS5) in K562 cells (genome coordinate's chr11:5269210-5269281) (Farrell *et al.*, 2002). Thirteen CTCF binding regions were identified at the  $\beta$ -globin locus in this study (Figure 4.6), one of which is located at the same coordinates as the  $\beta$ -globin HS5.



**Figure 4.6: CTCF binding sites at the  $\beta$ -globin locus.** Panel A illustrates the CTCF binding events at ENCODE region Enm009 ( $\beta$ -globin locus) in K562 cells. Log<sub>2</sub> fold enrichments are represented in the top half of the panel (K562\_CTCF track) and CHIPOTle sites are indicated below the x-axis (chipotle track). Known genes are indicated at the bottom of panel A. A magnified view of CTCF binding is presented in panel B. A highly enriched peak of CTCF binding is observed at the previously characterised  $\beta$ -globin HS5.

In addition, the CTCF data was compared with that of Kim and colleagues who identified nearly 14,000 CTCF binding sites in IMR90 cells and also predicted the location of over 30,000 CTCF sites based on an experimentally defined consensus sequence (Kim *et al.*, 2007). The location of predicted CTCF sites at the SCL locus was examined to determine the correlation with CTCF sites identified in K562 cells by ChIP-chip. The study by Kim and colleagues (2007) predicted five CTCF binding sites at the SCL locus (Figure 4.7). Four of these predicted sites were associated with CTCF binding in K562, namely the +57 region (associated with two copies of the consensus motif), +53 region and +40 region. These sites also bound CTCF in primary human fibroblast IMR90 cells (Kim *et al.*, 2007). Four of the five predicted CTCF binding sites were conserved at the sequence level in at least one other vertebrate genome and these four sites were bound in both K562 and IMR90 cells suggesting that they represent genuine functional elements. Sites of CTCF binding within or close to the CYP4Z1, CYP4A22, SIL, and KCY genes were not associated with predicted motifs and displayed K562-specific binding. However, Kim and colleagues also acknowledged that the consensus motif does not match all CTCF sites (Kim *et al.*, 2007). A more detailed comparison of the data generated here with that of other laboratories is presented in section 4.7.



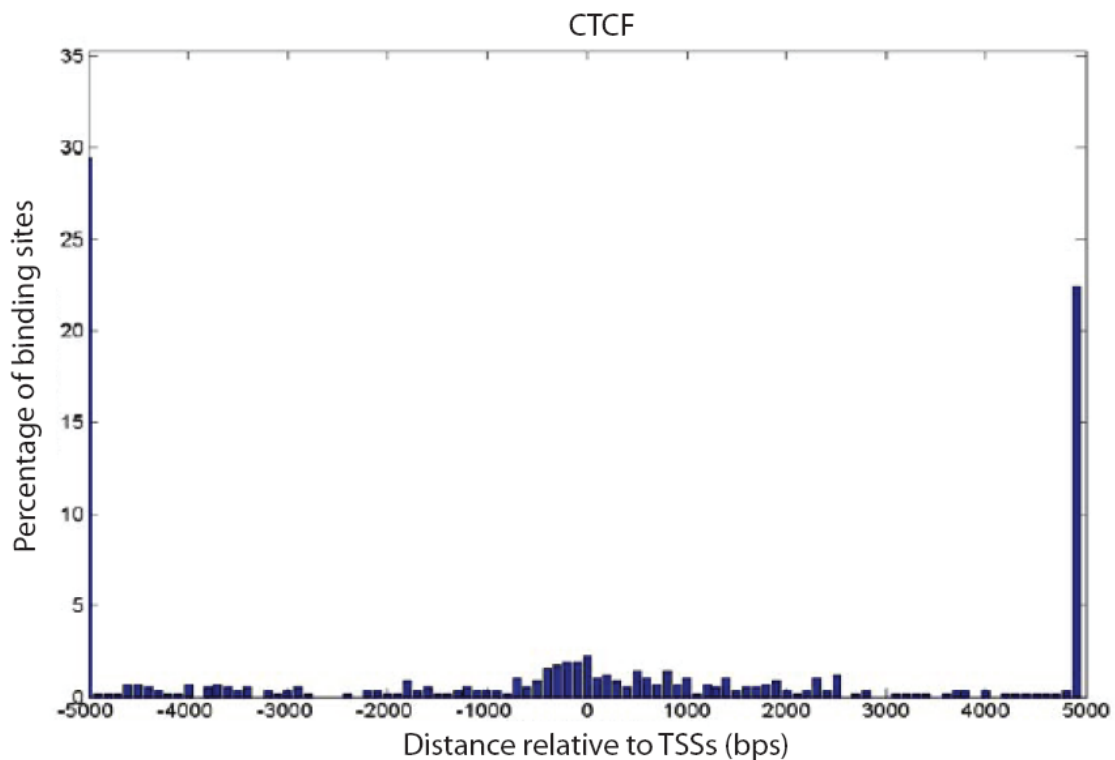
**Figure 4.7: Comparison of CTCF binding sites in the K562 cell with binding sites identified in primary human fibroblast, IMR90 cells.** K562 binding sites are indicated at the top of the figure by black vertical bars (K562\_CTCF track). Binding sites identified in IMR90 cells by Kim and colleagues (2007) are indicated below the K562 data by blue vertical bars (IMR90\_CTCF track). Reported fold enrichments are displayed as  $\text{Log}_2$  values in both profiles. The +57, +53, and +40 binding sites are bound by CTCF in the two cell types. The +57, +53, and +40 regions are associated with the predicted 20-mer binding sequence (indicated by black vertical lines in the predicted motif track). Four of the five predicted peaks are conserved in at least one other vertebrate genome, excluding the chimpanzee genome (indicated by the conservation track). An additional five binding sites were identified in K562 cells which were not associated with the known consensus motif, which were located within the CYP4Z1, CYP4A22, and SIL (also known as STIL) genes, and at the SIL and KCY (also known as CMPK) promoters. The core motif track represents predicted CTCF binding sites based on the previously reported CTCF consensus sequence (Bell and Felsenfeld, 2000). No core motif sites are predicted in the SCL locus. Known RefSeq genes (Pruitt *et al.*, 2007) are indicated at the bottom of the figure and human chromosome 1 coordinates are displayed at the top of the figure. Note: PDZK1IP1, TAL1, STIL, CMPK are also known as MAP17, SCL, SIL, and KCY respectively.

The accurate identification of previously characterised CTCF binding sites at the IGF2/H19 and  $\beta$ -globin loci and overlap of CTCF sites with the predicted consensus CTCF motif at the SCL locus indicated that this ChIP-chip method in combination with

ChIPOTle analysis could be used to accurately map CTCF sites of interaction in a high-throughput manner across 1% of the human genome sequence.

#### 4.6.2. Distribution of CTCF-binding sites in the ENCODE regions

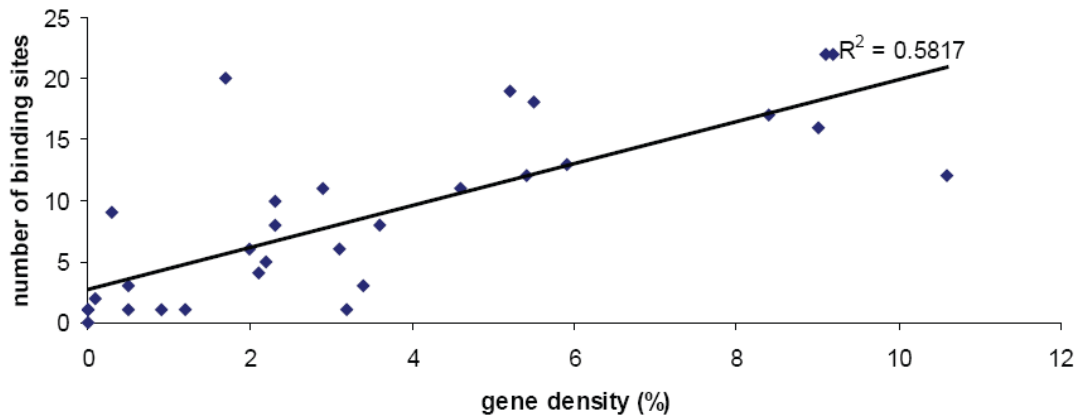
The distribution of CTCF binding sites was investigated with respect to the location of the nearest TSSs of protein-coding genes (Figure 4.8). More than 50% of CTCF sites of interaction were found to be located 5kb or more from TSSs, consistent with CTCF functioning at distant insulator or enhancer/repressor elements.



**Figure 4.8: The binding intervals and distribution of CTCF sites relative to transcription start sites (TSSs) in K562.** Panel B shows the relative distance of CTCF binding sites to the closest TSS of protein-coding genes. Relative distance is indicated in bp's along the x-axis and the percentage of binding sites is shown on the y-axis. Note that the total percentage of CTCF binding sites located 5 kb or more from a TSS is indicated by the blue vertical bars on the extreme left and right of the figure.

Although the CTCF sites tend to be located far from transcription start sites they are not randomly distributed across the ENCODE regions. The distribution of CTCF binding sites was examined by comparing the number of binding events with gene density in the

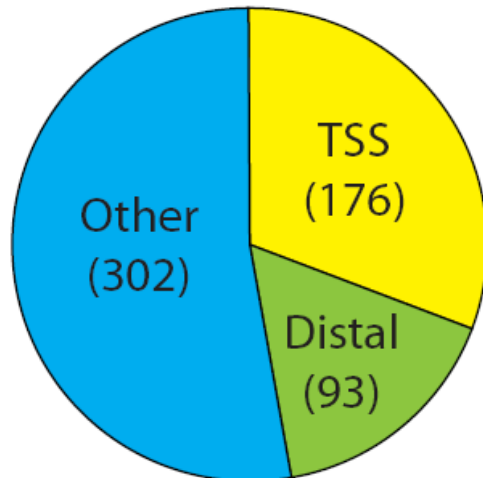
ENCODE regions. The distribution of CTCF sites closely follows the distribution of gene density, with a correlation coefficient of 0.76 (Figure 4.9 shows  $R^2$  value of 0.5817). This is consistent with CTCF regulating gene expression, rather than simply performing a structural role such as the formation of chromatin loops.



**Figure 4.9: Correlation of CTCF binding events with gene density in the ENCODE regions.** Gene density was expressed as a percentage of the total region size for computationally defined ENCODE regions (Chapter 3) and was plotted against the number of binding sites in each region. In general those regions associated with a high gene density also contained a high number of CTCF binding sites ( $R^2=0.5817$ ).

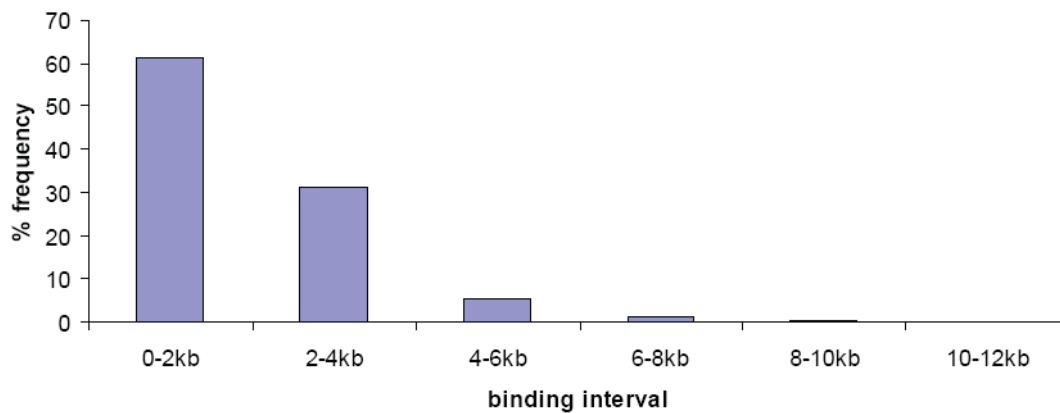
The distribution of CTCF binding sites was further examined by comparing their location with respect to transcription start sites (TSSs), distal enhancer/repressor sites, and other sites (Figure 4.10). Distal enhancer/repressor sites were defined as regions associated with a peak of enrichment for H3K4me1, or H3K4me2, or H3K4me3, and not within 2.5 kb of a TSS (Chapter 3) and other sites can be located anywhere except at TSSs or distal enhancer/repressor sites. 31% of binding events were located at TSSs, consistent with a role as a transcription factor involved in gene repression/activation, while 16% of binding events were located at distal enhancer/repressor sites. Like its role at TSSs, CTCF may function as a classical transcription factor by binding to distal enhancer/repressor sites. There is also the possibility that CTCF binding at distal sites may be related to the enhancer blocking activity of this protein. Sites not at TSSs or at distal enhancers/repressors (shown as “other” in Figure 4.10) account for the largest percentage

(53%) of CTCF sites. This class of CTCF sites is not associated with any promoter or enhancer/repressor function, suggesting that they may define the location of insulators.



**Figure 4.10: The distribution of CTCF binding sites.** The pie-chart shows the distribution of CTCF binding sites mapped to transcription start sites (TSSs), distal enhancer/repressor sites, and other locations within the 44 ENCODE regions.

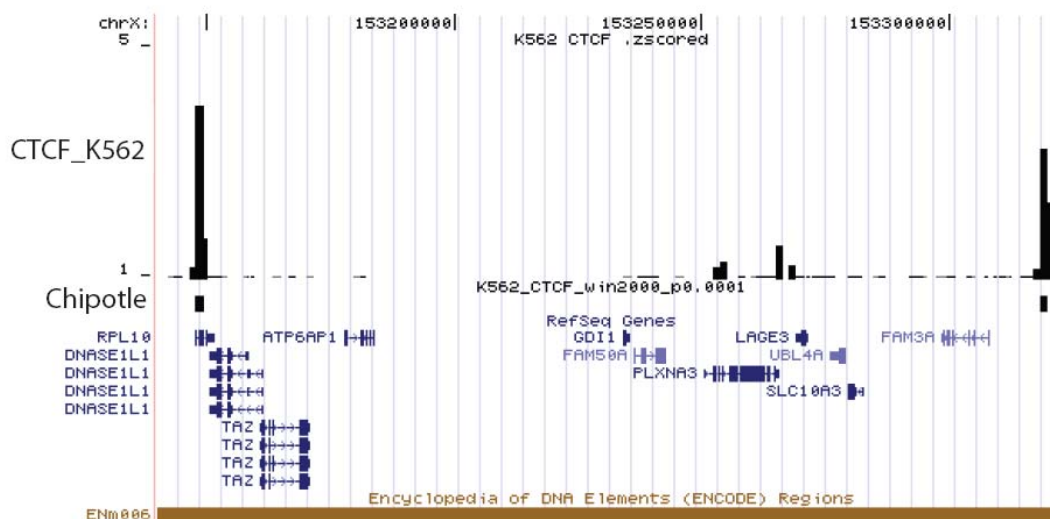
The possibility of multiple CTCF binding sites found in close proximity within ChIPOTle defined regions was investigated by determining the size distribution of CTCF ChIPOTle peaks as a function of 2 kb intervals (Figure 4.11). Its important to note that the size of microarray elements has an important bearing on this determination- as the average resolution of ENCODE array elements was calculated to be 1024 bp, ChIPOTle peaks within the size range of 0-2kb would suggest a single CTCF binding site. The majority (61%) of CTCF binding sites were 2kb or less in size suggesting that most CTCF binding sites were likely to be associated with one CTCF binding event. 39% were found spanning more than 2 kb suggesting that multiple copies of CTCF may be bound in close proximity at these locations, although the biological significance of these multiple close binding events is not known.



**Figure 4.11: CTCF binding intervals indicate multiple CTCF sites can bind in close proximity.** CTCF binding intervals defined by ChIPOTle were found to range in size from less than 2 kb up to 12 kb for one particular site. The 2kb binding intervals are indicated on the x-axis, while the percentage frequency of binding sites in each 2 kb interval is indicated on the y-axis.

#### **4.6.3. CTCF sites at individual genes and gene clusters**

While examining the distribution of CTCF sites it was noted that several individual genes were flanked in their entirety by CTCF sites and clusters of genes were also found to be flanked by two CTCF sites, as opposed to CTCF sites separating each member of the cluster. 307 genes in the ENCODE regions were flanked in their entirety by CTCF binding sites, representing 52% of the genes, while 10 clusters containing 5 or more genes were flanked by CTCF sites (Table 4.1). One cluster on the X chromosome contained 17 genes flanked by CTCF sites (Figure 4.12) and some of the other clusters were the well-characterised  $\alpha$ -globin,  $\beta$ -globin and HOXA loci. The significance of CTCF binding sites at genes clusters is not known but CTCF may form chromatin domains to regulate the expression of co-transcribed genes.



**Figure 4.12: CTCF binding sites flank a cluster of genes in ENCODE region Enm006.** 17 RefSeq genes (which include a number of alternative transcripts) are flanked by two CTCF sites in K562. The top track (CTCF\_K562) indicates the log<sub>2</sub> fold enrichments in this region and the ChIPOTle defined peaks are indicated below the x-axis. The chromosome X coordinates are indicated at the top of the figure and known RefSeq genes are indicated in blue at the bottom of the figure.

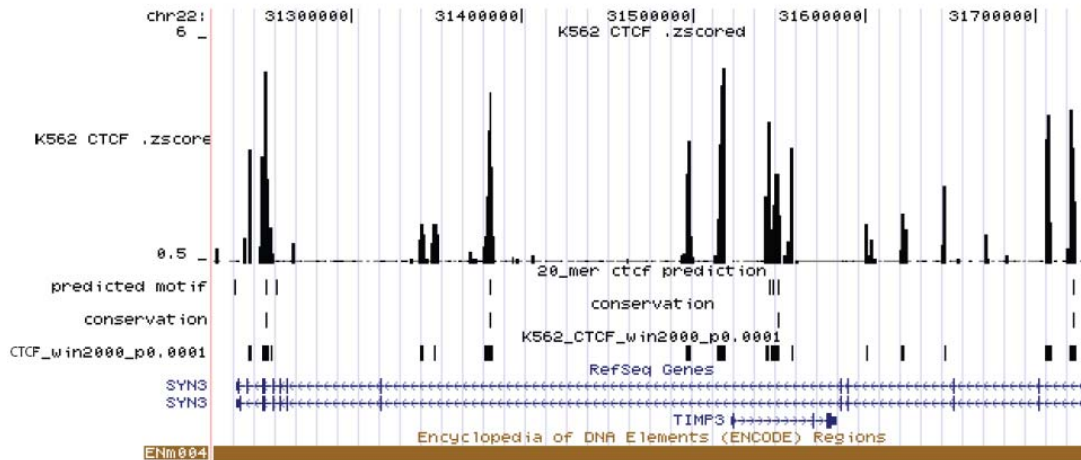
Coordinates	Description	Number of genes Flanked by CTCF
chr21:33,730,000-34,120,000	unrelated	9
chrX:153,140,000-153,323,000	unrelated	17
chrX:153,627,437-153,781,938	F8A genes	5
chr19:59,223,959-59,346,770	unrelated	8
chr19:59,701,626-59,886,317	leukocyte immunoglobulin-like receptors (LILR) cluster	8
chr16:162,168-320,833	$\alpha$ globin region	9
chr11:5,099,780-5,277,643	$\beta$ globin region	8
chr11:5,570,932-5,669,811	tripartite motif (TRIM) cluster	7
chr7:26,852,428-26,953,093	hoxa cluster	7
chr15:41,851,756-41,958,421	unrelated	5

**Table 4.1: Gene clusters flanked by CTCF binding sites.** 10 gene clusters (containing five or more genes) were flanked by CTCF sites. The start and end coordinates (human genome release hg17) of the regions flanked by CTCF binding sites are indicated, along with a description and the number of genes contained within each cluster.

In contrast to CTCF sites flanking clusters of genes, a number of individual genes contained several intergenic CTCF binding sites, such as the SYN3 gene which is



associated with 16 CTCF sites (Figure 4.13). Nine ENCODE genes contained five or more intergenic CTCF binding sites (Table 4.2). Gene Ontology (GO) annotations were examined for these nine genes to determine if any GO terms were enriched in genes containing multiple CTCF sites. 7 of these contained a GO annotation and five were annotated with the integral to membrane GO term (p0.01 or less). Although the significance of multiple CTCF binding sites within genes that code for integral membrane proteins is not known, it suggests that these either these genes have an unusually complex enhancer-blocking mechanism or that CTCF may have a role as a transcription factor at enhancers/repressor elements which are often located within genes.



**Figure 4.13: Genes can contain several CTCF binding sites.** The SYN3 gene encodes a member of the synpasin family of proteins and is associated with 16 CTCF binding sites in K562 cells. The TIMP3 gene encodes a member of the tissue inhibitors of the matrix metalloproteinases family and is located within an intron of this gene. It is transcribed in the opposite direction to SYN3 and a number of the CTCF binding sites may be associated with TIMP3 regulation.

Coordinates	Gene	Description	CTCF sites
chr7:116,187,332-116,464,024	ST7	suppression of tumorigenicity 7	6
chr22:31,233,094-31,727,237	SYN3	synapsin III	16
chr5:142,130,476-142,586,243	ARHGAP26	rho GTPase-activating protein 26	8
chr7:125,672,608-126,477,261	GRM8	glutamate receptor metabotropic 8 precursor	6
chr11:130,745,779-131,710,752	AY358331	member of the IgLON (LAMP, OBCAM, Ntm) family of immunoglobulin (Ig) domain-containing glycosylphosphatidylinositol (GPI)-anchored cell adhesion molecules	8
chr13:112,392,644-112,589,467	ATP11A	integral membrane ATPase, Class VI, type 11A	7
chr2:220,204,557-220,228,997	ACCN4	amiloride-sensitive cation channel 4	5
chr5:141,953,307-142,045,812	FGF1	fibroblast growth factor 1	5
chr11:64,130,222-64,247,236	NRXN2	neurexin 2 isoform alpha-2 precursor	5

**Table 4.2: ENCODE Genes containing multiple CTCF binding sites in K562 are membrane components.** The genes containing multiple CTCF sites are involved in cell signaling or cell adhesion processes. GO cellular component annotation indicates that five of the nine genes are integral to membrane formation.

Thus in summary, the distribution of CTCF binding sites is complex as it can be found flanking individual genes or groups of genes or multiple CTCF sites can be found within individual genes. This complex binding pattern suggests a multi-functional role for CTCF in the regulation of gene expression.

#### **4.7. A comparative and sequence based analysis of CTCF sites in different cell types**

In order to investigate the conservation of CTCF binding in different cell types, the human cell line U937, established from a patient with generalised histiocytic lymphoma (Sundstrom and Nilsson, 1976) and displaying properties of monocytes (Anderson and Abraham, 1980) was used to perform CHIP-chip analysis of CTCF binding sites. Three

biological replicate experiments were performed as described for K562, the median data was calculated and IgG normalised as described previously. 661 U937 CTCF sites were identified in the ENCODE regions by ChIPOTle (p0.0001) compared to 571 CTCF sites in K562. The locations of CTCF sites in K562 and U937 cells were also compared with the data of Kim *et al.* (2007) and Barski *et al.* (2007). Kim *et al.* predicted the location of 412 CTCF binding sites in the ENCODE regions based on a 20-mer consensus motif. 172 of these predicted sites were conserved at the sequence level in at least one other vertebrate genome, excluding the chimpanzee genome. The number of experimentally determined K562 CTCF sites that overlapped with the location of a consensus binding motif was 187 (33% of 571 sites identified), 121 of which were conserved sites (Table 4.5). Therefore 121 of the 172 (70%) predicted and conserved CTCF sites in the ENCODE regions were bound in K562. The number of experimentally determined U937 CTCF sites that overlapped with predicted sites was 195 (29.5% of 661 sites identified), 126 of which were conserved sites (Table 4.3). Therefore 126 of the 172 (73%) predicted and conserved CTCF sites were bound in U937. This suggests that *in silico* predicted CTCF sites that are conserved in at least one other vertebrate genome is a relatively accurate predictor of CTCF binding *in vivo*. While the specificity of this approach is high (70-73% of predicted and conserved sites are bound in K562 and U937 cells respectively), the sensitivity is low as 79% and 81% of experimentally determined binding events in K562 and U937 respectively were not identified by this approach. Therefore the majority of CTCF sites identified in K562 and U937 cells were not associated with the consensus motif defined by Kim *et al.* (2007).

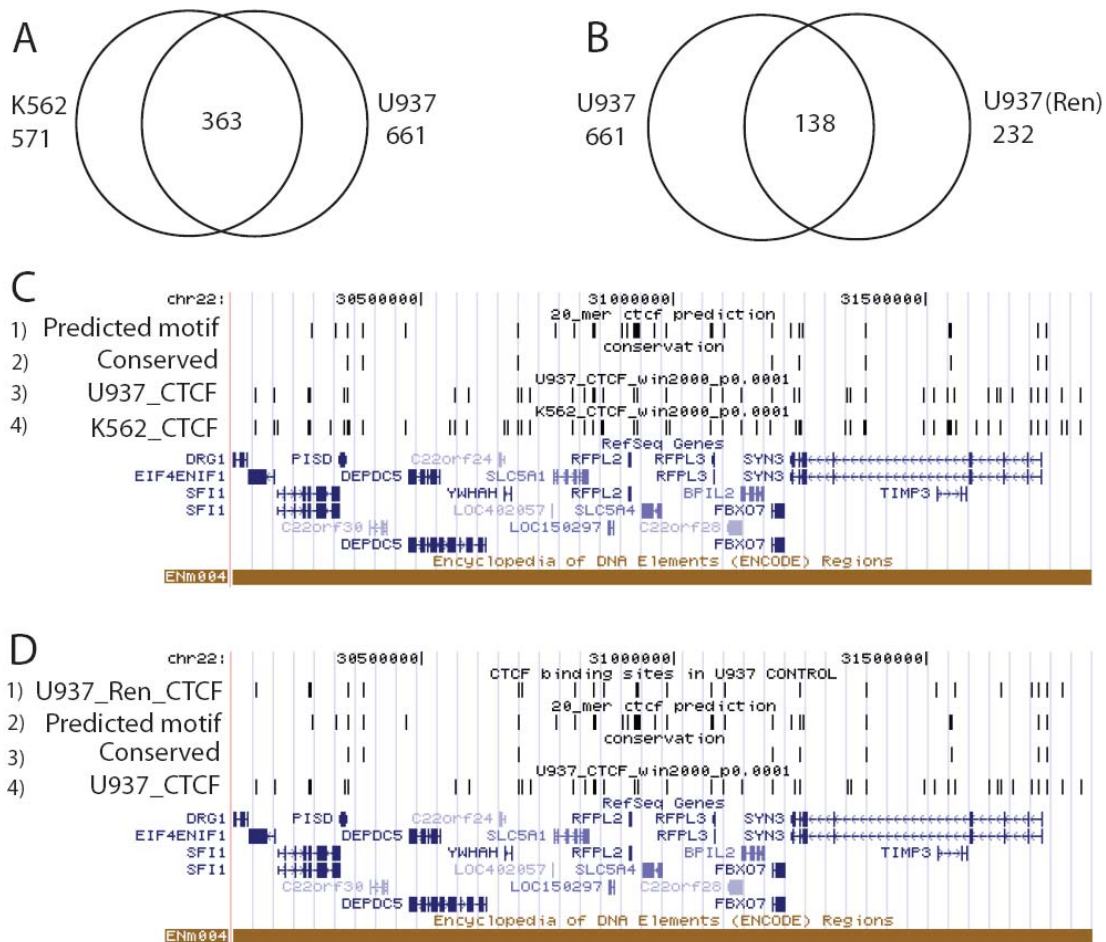
Cell line	Number of experimental CTCF sites	Number of predicted CTCF sites	Number of predicted and conserved CTCF sites	Number of experimental sites located at predicted sites	Number of experimental sites located at conserved sites
K562	571	412	172	187	121
U937	661	412	172	195	126

**Table 4.3: Correspondence between ChIP-chip defined and predicted CTCF sites.** 412 CTCF binding sites were predicted in the ENCODE regions by Kim and colleagues based on the presence of a 20-mer

consensus motif (Kim *et al.*, 2007). 172 of these predicted sites were conserved at the sequence level in at least one other vertebrate genome. 187 and 195 of the CTCF binding sites identified in K562 and U937 cells by ChIP-chip analysis overlapped with the CTCF consensus motif. 121 and 126 of the experimentally identified CTCF sites in K562 and U937 were associated with predicted CTCF sites that were also conserved in at least one other vertebrate genome.

A comparison of K562 and U937 CTCF sites identified in this study was also performed. Of the 571 and 661 binding sites identified in K562 and U937 cells respectively, 393 sites (69%) overlapped between the two cell lines (Figure 4.14). This suggests that almost a third of CTCF sites are involved in cell-type specific regulation. ENCODE region Enm004 contained the greatest number of overlapping sites, 36 of 50 sites (72%) identified in the two cell lines overlapped (Figure 4.14). As described previously, Ren and colleagues had performed a genome-wide study of CTCF binding in IMR90 cells and identified 225 CTCF sites in the ENCODE regions (Kim *et al.*, 2007). 182 (31%) of IMR90 sites overlapped with the 571 K562 binding sites identified as part of this study. Barski and colleagues used ChIP-sequencing to identify 20,262 CTCF binding sites in CD4+ cells (Barski *et al.*, 2007), of which 353 were located in the ENCODE regions. 227 and 232 of these 353 sites overlapped with CTCF sites in K562 and U937 cells respectively, representing an overlap of 40% and 35%. Therefore between 31%-40% of K562 CTCF sites identified in this study overlapped with CTCF sites reported by two other studies.

As the study of Kim *et al.* (2007) had also examined U937 CTCF binding sites in the ENCODE regions, a direct comparison of the two U937 datasets was performed. 138 of the 232 CTCF sites identified by Kim and colleagues overlapped with U937 CTCF sites identified in this study (Figures 4.14 panels B and D). The gene-rich Enm004 region on chromosome 22 contained the highest number of overlapping sites at 21 (Figure 4.14, panel D). However, 39% of U937 sites identified by Kim and colleagues were not identified in this study and almost three times more CTCF binding sites were identified in this study. The possible reasons for the differences in data sets derived from the same cell line are outlined in the discussion of this Chapter.



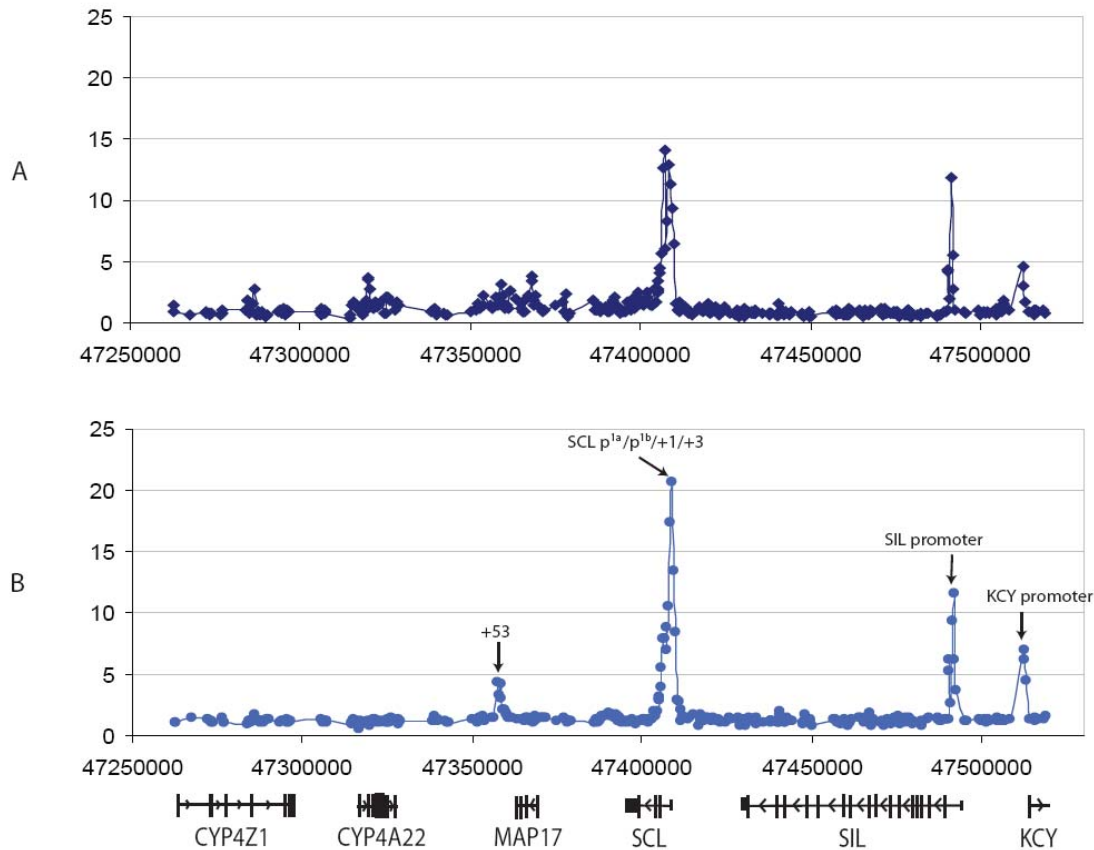
**Figure 4.14: A comparison of CTCF binding sites in different cell types and between studies.** Panel A: 571 and 661 binding sites were identified in K562 and U937 as part of this study. 363 of these sites overlapped in the two different cell types. Panel B: 661 CTCF sites that were identified as part of this study were compared with the 232 CTCF sites identified by Kim and colleagues in U937 cells. 138 sites overlapped between the two studies. Panel C: A UCSC screenshot of ENCODE region enm004, which contained the greatest number of overlapping CTCF sites in K562 and U937 cells (36 sites) identified in this study. Track number 1 (predicted motif) indicates the location of consensus CTCF sites, track 2 (conserved) indicates those consensus sites that are conserved in at least one other vertebrate genome, tracks 3 (U937\_CTCF) and 4 (K562\_CTCF) indicate the location of ChIPOTle defined CTCF binding sites in U937 and K562 cells respectively. Panel D: ENCODE region enm004 also contained the greatest number of overlapping CTCF sites in two independent studies on CTCF binding in U937 cells (21 sites). Track number 1 (U937\_Ren\_CTCF) indicates the location of CTCF sites identified in U937 by Kim and colleagues (2007), track 2 (predicted motif) shows the location of predicted CTCF sites, track 3 (conserved) indicates which predicted sites are conserved in at least one other vertebrate genome, and tracks 4 (U937\_CTCF) indicates the location ChIPOTle defined CTCF binding sites in U937 cells as part of this

study. Enm004 chromosome coordinates are shown at the top of panels C and D, while the RefSeq genes are indicated below the data tracks.

## **4.8. Analysis of other transcription factors implicated in CTCF or insulator function**

### **4.8.1. Developing assays using the SCL locus as model system**

As discussed in the introduction, CTCF associates with a number of proteins and these interactions are important for modulating the function of CTCF. CTCF is known to interact with mSin3a (Lutz *et al.*, 2000) to mediate transcriptional repression. Therefore mapping sites of mSin3a interactions would allow for a more detailed picture of whether this protein is important for CTCF function genome-wide. The SCL microarray was used to develop a ChIP-chip assay to detect mSin3a interactions in K562 cells (Figure 4.15) and the data was normalised with a normal rabbit IgG control. Four peaks of significant enrichment were detected for mSin3a interaction at the SCL locus, namely the +53 region, the SCL promoter region, the SIL promoter and the KCY promoter. Three of the four regions also bound CTCF (the +53 region, the SIL promoter and the KCY promoter) suggesting that CTCF may function as a classical transcription factor at these regions. However, it was a surprising to find mSin3a at the promoter region of three actively transcribed genes (in addition the +53 region also displays bi-directional promoter activity in K562 cells; Dhimi, submitted).



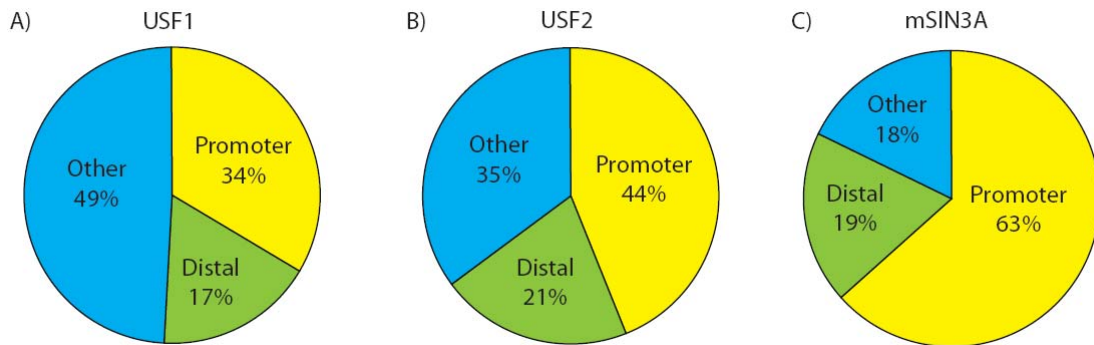
**Figure 4.15: ChIP-chip profile of mSin3a interactions across the SCL locus in K562 before and after normal rabbit IgG normalisation.** Panel A: Fold enrichments reported for mSin3a interactions before normalisation with normal rabbit IgG. Panel B: fold enrichments increased at mSin3a interacting sites after normalisation with normal rabbit IgG. The location of the +53 region, the SCL, SIL, and KCY promoters are indicated by black arrows. The human chromosome 1 genomic coordinates are indicated along the y-axis, while fold enrichments are indicated on the x-axis. Gene order and direction of transcription is shown below panel B.

In contrast to mSin3a, USF1 and USF2 are known to bind in close proximity to CTCF at the chicken  $\beta$ -globin HS4 insulator and are responsible for the chromatin barrier activity of this insulator (West *et al.*, 2004; Huang *et al.*, 2007). Thus, mapping USF1 and USF2 interactions in combination with that of CTCF and histone modification data may allow for chromatin barrier insulators to be identified in the human genome. USF1 and USF2 ChIP-chip assays were tested using the SCL microarray but no significant enrichments were detected (data not shown). However, the ENCODE array had previously been used in ChIP-chip assays to detect a number of USF1 binding sites in HepG2 cells (Rada-

Iglesias *et al.*, 2005). USF1 and USF2 ChIP-chip assays were therefore tested using the ENCODE array and several hundred USF1 and USF2 binding sites were identified in K562 cells using this array as described in the following section.

#### 4.8.2. Mapping the distribution of mSin3a, USF1, and USF2 binding sites in the ENCODE regions

In order to further characterise the 571 K562 ENCODE CTCF sites, ChIP-chip experiments were performed with K562 cells to detect mSin3a, USF1 and USF2 interactions in the ENCODE regions. Microarray experiments were performed and the data normalized with the relevant mock IgG data as described earlier in this Chapter. ChIPOTle was then used to identify peaks of enrichment at a high confidence ( $p < 0.0001$ ) for each transcription factor. 483 binding sites were identified for USF1, 219 sites were identified for USF2, and 310 mSin3a interactions were mapped. This represented a substantial number of peaks in 1% of the genome, suggesting that like CTCF these transcription factors may regulate the expression of a large number of genes across the genome. The distributions of mSin3a, USF1, and USF2 binding sites were determined with respect to the location of promoters, distal enhancer/ repressors, and other sites (Figure 4.16)

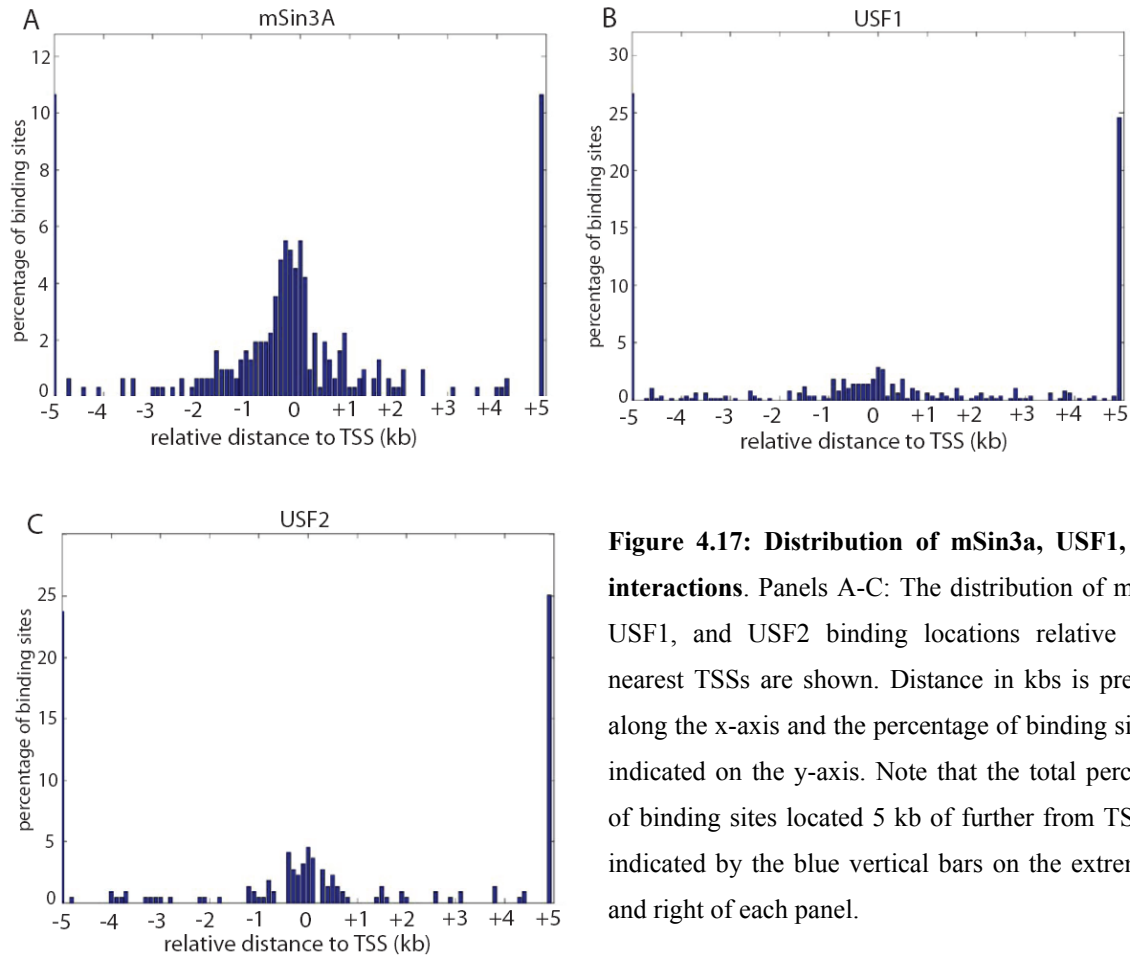


**Figure 4.16: The distribution of USF1, USF2, and mSin3a sites of interaction in the ENCODE regions in K562 cells.** The ChIPOTle sites for USF1 (A), USF2 (B) and mSin3a (C) were mapped to promoters (within 2.5kb of known transcription start sites), distal enhancer/repressors (associated with H3K4me1, H3K4me2, or H3K4me3 and not within 2.5kb of a TSS) and other sites which were not associated with H3K4 methylation or within 2.5kb of promoters. The percentages of sites that map to each genomic feature are indicated.



The distribution of USF1 binding sites mirrored that of CTCF as approximately equal percentages of binding events were identified (31% of CTCF sites were located at promoters, 16% were at distal sites and 53% were located at other sites). USF2 binding was marginally more biased towards promoter elements than USF1 but broadly similar to the pattern of binding observed for USF1 and CTCF. However, in contrast, the distribution of mSin3a interactions was found to be heavily biased towards promoters. This finding is consistent with the known role of mSin3a in binding close to or at promoters to repress transcription by recruiting other proteins such as histone deacetylases (Dannenber *et al.*, 2005).

To more accurately map the locations of mSin3a, USF1, and USF2 binding sites at promoters, their binding patterns were mapped with respect to the location of the nearest TSSs (Figure 4.17). The majority of mSin3a binding sites were located within 2 kb of TSSs and further analysis determined that 82% of the TSSs associated with an mSin3a binding event (less than 1 kb from a TSS) in K562 cells were also associated with CpG islands. In contrast only about 50% of the ENCODE gene promoters are associated with a CpG island. This suggests that mSin3a associates more readily with promoters with CpG islands and may interact with some proximal promoter sequence binding proteins that are specific to CpG-containing promoters. In contrast, as discussed above, the binding pattern of USF1 and USF2 was somewhat similar to CTCF as binding sites were not restricted to promoter regions - approximately half of the sites for these TFs were located 5 kb or more from TSSs.



**Figure 4.17: Distribution of mSin3a, USF1, USF2 interactions.** Panels A-C: The distribution of mSin3a, USF1, and USF2 binding locations relative to the nearest TSSs are shown. Distance in kbs is presented along the x-axis and the percentage of binding sites are indicated on the y-axis. Note that the total percentage of binding sites located 5 kb of further from TSSs are indicated by the blue vertical bars on the extreme left and right of each panel.

### 4.8.3. Analysing interactions between CTCF and mSin3a, USF1, and USF2

CTCF is known to interact with mSin3a (Lutz *et al.*, 2000) while USF1 and USF2 bind in close proximity to CTCF at the HS4 chicken  $\beta$ -globin insulator (West *et al.*, 2004). Therefore, potential interaction or co-localisation of CTCF with mSin3a, USF1, and USF2 was examined by analysing the extent of overlapping ChIPOTle sites. The number of CTCF sites which overlapped with interactions for one or more of the other factors were determined (Table 4.4). The binding events were then categorised into those found at transcription start sites (TSSs), at distal enhancer/repressor sites, and “other” sites. Distal enhancer/repressor sites were defined as regions associated with a peak of enrichment for H3K4me1, or H3K4me2, or H3K4me3, and not within 2.5 kb of a TSS (Chapter 3). Sites defined as “other” are not TSSs or distal enhancer/repressors (according to the definitions used in this study). Overlapping interactions at TSSs were

further sub-categorised into those located at active and inactive TSSs (as determined by available gene expression data for K562).

	CTCF only	CTCF +mSin3a only	CTCF +USF1 only	CTCF +USF2 only	CTCF +mSin3a +USF1	CTCF +mSin3a +USF2	CTCF +USF1 +USF2	CTCF +mSin3a +USF1 +USF2
<b>All sites</b>	384	47	37	0	15	31	31	26
<b>TSSs</b>	73	33	13	0	11	14	14	18
<b>Active TSSs</b>	29	17	3	0	8	7	7	12
<b>Inactive TSSs</b>	11	1	0	0	0	1	1	3
<b>Distal enhancer/repressors</b>	52	9	11	0	4	5	5	7
<b>Other sites</b>	259	5	13	0	0	12	12	1

**Table 4.4: Overlapping combinations of CTCF, mSin3a, USF1, and USF2 binding sites in K562.**

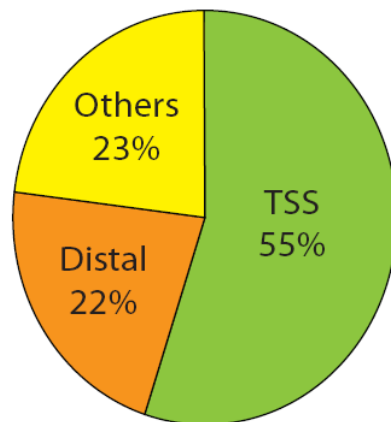
ChIPOTle hits for each of the four transcription factors were analysed and the number of overlapping binding events are shown for the different factor combinations. The data is mutually exclusive for each of the seven combinations, for example binding sites shared by CTCF and mSin3a indicates that USF1 and USF2 binding sites do not overlap. Sites at TSSs are defined as binding events within 2.5kb of transcription start sites (TSSs), Distal enhancer/repressor sites are defined as sites containing either a H3K4me1, or H3K4me2, or H3K4me3 ChIPOTle peak not within 2.5 kb of a TSS. The remaining sites were classified as other sites. Where possible, TSSs associated with binding events were classified as active or inactive using Affymetrix expression microarray data (Chapter 3). Note: binding sites that spanned a TSS and a distal site were not counted twice but were preferentially assigned as a TSS binding event.

A number of observations arose from this analysis, demonstrating that the distribution of CTCF, mSin3a, USF1, and USF2 overlapping sites is complex. Furthermore, given that the number of sites analysed in the groups shown in Table 4.4 were relatively small, it is difficult to determine an accurate picture of the relationship of these four regulators and how they act in combination to regulate gene expression.

The majority (67.3%) of CTCF binding sites (384/571) do not overlap with either mSin3a, USF1 or USF2, whilst the remaining 32.7% of CTCF sites (187/571) do overlap

with sites of interaction for one or more of the three factors. This latter figure was broken down into the following proportions:

- 8.2% of CTCF binding sites overlap with mSin3a binding sites only (CTCF+mSin3a)
- 6.5% of CTCF binding sites overlap USF1 binding sites only (CTCF+USF1)
- 2.6% of CTCF binding sites overlap with both mSin3a and USF1 binding sites (CTCF+mSin3a+USF1)
- 5.4% of CTCF binding sites overlap with both mSin3a and USF2 binding sites (CTCF+mSin3a+USF2)
- 5.4% of CTCF binding sites overlap with both USF1 and USF2 binding sites (CTCF+USF1+USF2)
- 4.5% of CTCF binding sites overlap with all three of mSin3a, USF1 and USF2 binding sites (CTCF+mSin3a+USF1+USF2)



**Figure 4.18: CTCF binding sites that overlap with binding sites of mSin3a, USF1, and USF2 are found in diverse locations. 55% of CTCF sites that overlap with one or more of the other factors are located at TSSs, 22% are at distal sites and 23% are found at other locations.**

These co-localising interactions were located at TSSs, distal sites and “other” sites (Figure 4.18) and implicated CTCF and these other factors in gene activation, gene repression, and/or insulator functions. The majority (55%) of CTCF binding sites that overlapped with mSin3a or USF1 or USF2 were located at TSSs and may be involved in regulating gene expression. 22% of overlapping interactions were located at distal enhancer/repressor sites and may be also involved in regulating gene expression or may act as enhancer-blocking insulators by binding at or in close proximity to enhancers. The remaining 23% of overlapping interactions were located at “other” sites, which represent

good candidates for insulator function as they are not associated with promoter or enhancer/repressor activity.

Upon examination of the distribution of co-localisation sites at TSSs, over two-thirds of CTCF+mSin3a overlapping interactions (33/47) were located at TSS consistent with the finding that CTCF interacts with mSin3a to repress transcription (Lutz *et al.*, 2000). However, data was available for the expression status of 18 of the genes associated with these TSS and 17 were associated with active and only 1 was inactive. This suggests that CTCF together with mSin3a may be involved in activating gene expression rather than repression. This was a surprising finding as mSin3a is best known as a co-repressor protein that recruits histone deacetylases to silence gene expression (Heinzel *et al.*, 1997) although more recent evidence suggests that its yeast homolog can also function in gene activation (De Nadal *et al.*, 2004).

Overall, of those CTCF sites which overlapped with at least one other of the three TFs (mSin3a, USF1, or USF2) at TSSs (103/187 as described above) - 52.4% (54/103) of these sites were associated with active gene expression. Specifically, approximately one-third of CTCF+USF1 overlapping sites (13/37) were at TSSs but no conclusive evidence regarding their association with active gene expression or repression could be determined as data was only available for three of these genes. Nearly three-quarters of CTCF+mSin3a+USF1 overlapping sites were located at TSSs (11/15) and 8/8 (for which expression data was available) were associated with active gene expression. Approximately half of CTCF+mSin3a+USF2 overlapping sites were also located at TSSs (14/31) and 7/8 were associated with active gene expression (as was the case for genes where CTCF+USF1+USF2 co-localised). Finally, approximately 70% of the CTCF+mSin3a+USF1+USF2 overlapping sites (18/26) were also located at TSSs, the majority of which were associated with actively expressed genes (12/15).

Overlapping sites for the four regulators were then examined at distal enhancer/repressor elements. Fewer overlapping sites were observed at these locations. 16-30% of overlapping interactions were located at distal sites but the functional relevance of these interactions is not known as the sample sizes were small. CTCF and these factors may be involved in regulating gene expression binding acting as classical transcription factors

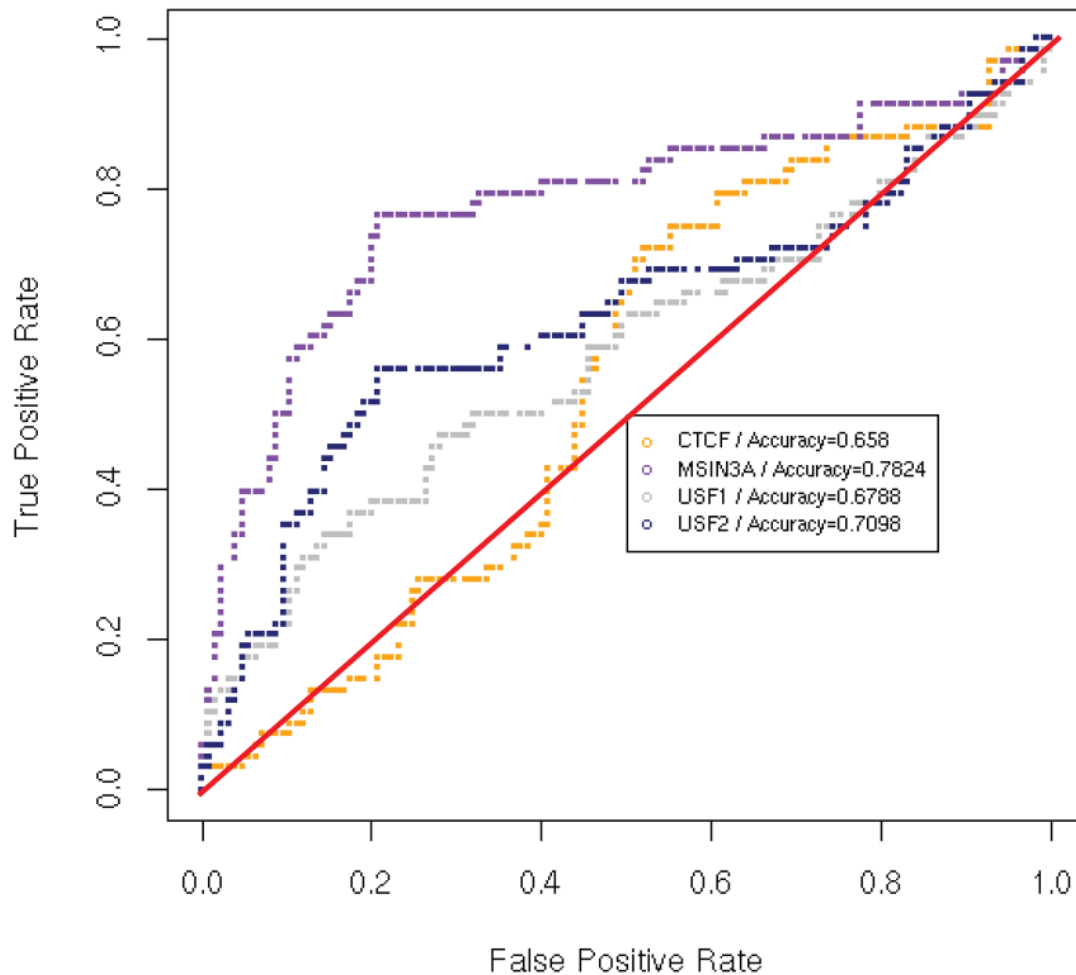
that bind enhancer/repressor elements or alternatively may be acting as enhancer-blocking insulators.

Finally, CTCF overlapping sites were then examined at the “other” sites, which may represent true insulator elements that could be distinguished from promoters or distal enhancers/repressors. Again, very few overlapping CTCF+mSin3a sites were observed at “other” sites (only 5 of 47), consistent with CTCF interacting primarily with mSin3a at promoters to regulate gene expression. However, one-third of CTCF+USF1 overlapping sites (13/37) were located at “other” sites and may be involved in insulator functions. No CTCF+mSin3a+USF1 sites or CTCF+USF2 sites were located at “other” sites. Of CTCF+mSin3a+USF2 sites and CTCF+USF1+USF2 sites, one-third were at “other” locations (12/31 in both cases). Overall, only 14.6% of “other” sites (38/259) showed co-localisation of CTCF with either USF1 or USF2, suggesting that their co-localisation may not be a general paradigm for insulator barrier function in the human genome (see section 4.9).

#### **4.8.4. Transcription factor binding and gene expression status**

In the previous section it was noted that CTCF sites at TSSs that overlapped with mSin3a or USF1 or USF2 sites were predominantly associated with active gene expression. This observation was explored further by investigating whether CTCF, mSin3a, USF1, or USF2 binding at TSSs could be used to predict the expression status of a gene by plotting receiver operating characteristic (ROC) curves (Figure 4.19) (Analysis performed by Dr. Ulas Karaöz, Boston University). The ROC of a classifier shows its performance as a compromise between selectivity and sensitivity. In this case CTCF, mSin3a, USF1, and USF2 binding at TSSs are used as a classifier of gene expression status at 238 K562 Affymetrix probe sets (on/off state based on present or absent MAS5 calls as described in Chapter 2). The plots illustrated in Figure 4.19 show sensitivity at all possible specificities and indicate that mSin3a binding at TSSs is highly predictive of active gene expression. The maximum accuracy value is reported for each TF and a value of 0.7824 was obtained for mSin3a, which is identical to the value obtained for H3K4me3 (Chapter 3). Therefore the binding of mSin3a at TSSs is highly predictive of active gene expression. USF2 shows a relatively high value also of 0.7098 while USF1 is less

predictive (0.6788) and CTCF is the least predictive (0.658). CTCF displays a similar value to H3K4me1 (0.6477) and the CTCF ROC curve is found to be negatively correlated with gene expression at low thresholds as observed by the curve located below the diagonal line. Therefore CTCF binding at TSSs is the least accurate predictor of active transcriptional state. This finding implicates mSin3a in active gene expression which is surprising given its accepted role in gene repression (as mentioned above). However, recent work has implicated the yeast homolog Sin3 in gene activation (De Nadal *et al.*, 2004) and this is discussed in greater detail in section 4.10.



**Figure 4.19: Predictive power of transcription factor binding for gene expression in K562.** Receiver operating characteristic (ROC) curves were plotted for 1kb regions around TSSs to determine the association between transcription factor (TF) binding and gene expression. ROC curves illustrate the predictive accuracy of TF binding on classifying the expression states of genes (on/off). The red diagonal line represents the ROC curve of a TF that is randomly associated with active or inactive gene expression.

TF binding that is positively associated with the gene on state will have a ROC curve above the diagonal line, while a TF associated with the gene off state will have a ROC curve below the diagonal. A TF associated randomly with expression state of genes achieves a ROC score of 0.5 (red line) while a TF that is a perfect predictor of gene expression state receives a ROC score of 1.0. The maximum ROC score is indicated for each factor and the highest score was obtained for mSin3a (0.7824).

#### **4.8.5. Motifs analysis of CTCF, mSin3a, USF1 and USF2 binding sites**

As 384 (67%) and 466 (70%) CTCF sites identified in K562 and U937 cells did not contain the predicted consensus CTCF motif, a computational analysis was performed to determine if any other known motifs were enriched in the CTCF binding sites. Motif matrices from TRANSFAC (Matys *et al.*, 2006) and JASPER (Bryne *et al.*, 2007) databases were used to scan the CTCF sites. 1 kb centred sequences from the ChIPOTle hits were extracted as the foreground sequence, while 1 kb sequences flanking the ChIPOTle sites were defined as background sequences. The best known motif which distinguishes the foreground and background sequences with a relative error rate of 0.381 was motif PF0045 from the JASPAR database (Table 4.5). This motif was previously identified in a study of regulatory motifs identified in human promoters and 3' UTRs by comparative analysis of several mammalian genome sequences (Xie *et al.*, 2005) and more recently was shown to bind CTCF (Xie *et al.*, 2007). This motif also matches the core of the longer 20-mer CTCF motif reported by Ren and colleagues. One other motif from JASPAR (PF0156) and three other motifs from TRANSFAC were also enriched, however, the motif that matched the known CTCF consensus sequence distinguished the foreground and background sequences with the highest sensitivity and specificity coupled with the lowest relative error rate.








Name	Logo	Sn	Sp	Error	pvalue
1. PF0045		0.651	0.586	0.381	0
2. PF0156		0.578	0.554	0.434	0
3. M00325		0.653	0.473	0.437	0
4. M00411		0.511	0.603	0.443	0
5. M01057		0.392	0.714	0.447	0

**Table 4.5: Known motifs associated with CTCF binding sites.** Motif matrices from JASPAR and TRANSFAC databases were used to search for enrichment in CTCF binding sites. The PF0045 motif in the JASPAR database, which matches the known CTCF motif, was found to be the best motif for distinguishing the foreground and background sequences with a relative error rate of 0.381, sensitivity (Sn) of 0.651, and specificity (Sp) of 0.586. The sensitivity associated with a motif and p-value cut-off is the proportion of true foreground sequences that were classified as foreground; the specificity is the proportion of true background sequences classified as background sequences. The relative error rate (Error) for a given motif and associated with a particular p-value cut-off is  $1-(Sn+Sp)/2$ . A good motif has a low error rate and balanced Sn and Sp values (Smith *et al.*, 2005). DNA logos are presented for each motif and the height of each letter indicates relative occurrence of nucleotides in the binding sites.

Because TRANSFAC and JASPAR can only be used to scan sequences for the presence of known motifs, the DME (discriminating matrix enumerator) program (Smith *et al.*, 2005) was used to search for the presence of novel motifs in the CTCF binding sites. The DME motif discovery algorithm calculates motif relative over-representation between two sets of sequences- foreground (identified binding sites) and background (sequences in the vicinity of identified binding sites) - using a strategy of enumerating position-weight matrices. Motifs with length from 8-12 nucleotides were searched for by extracting 1 kb centred sequences from the ChIPOTle sites and defining these as foreground sequences while 1 kb sequences flanking the ChIPOTle sites were defined as background sequences. The top DME motif which distinguished foreground and background sequences with a relative error rate of 0.381 was consistent with the core of the known CTCF consensus motif (Table 4.6) (Note DME retrieves the reverse

compliment of a binding motif). Therefore DME identified no novel motifs in the CTCF binding sites.

Name	Logo	Sn	Sp	Error	pvalue
1. DME0001		0.555	0.683	0.381	0
2. DME0002		0.566	0.574	0.43	0.0002
3. DME0003		0.357	0.78	0.431	0.0003
4. DME0004		0.737	0.399	0.432	0.0003
5. DME0005		0.405	0.728	0.434	0.0008

**Table 4.6: De novo CTCF motif discovery.** The DME program (Smith A PNAS) was used to identify novel motifs in CTCF binding sites. DME0001 motif was found to be the best motif for distinguishing the foreground and background sequences with a relative error rate of 0.381 at a p-value cutoff of 0. This motif was consistent with the core of the known CTCF motif (positions 5-16) as it is identified in reverse compliment by the DME program. The sensitivity (Sn) associated with a motif and p-value cut-off is the proportion of true foreground sequences that were classified as foreground; the specificity is the proportion of true background sequences classified as background sequences. The relative error rate (Error) for a given motif and p-value cut-off is  $1-(Sn+Sp)/2$ . A good motif has a low error rate and balanced Sn and Sp values (Smith *et al.*, 2005). DNA logos are presented for each motif and the height of each letter indicates relative occurrence of nucleotides in the binding sites.

USF1 and USF2 are basic helix-loop-helix TFs and this family of TFs generally binds to a consensus sequence of 5'-CANNTG-3' (where N is any nucleotide) called enhancer box (E-box) motif which is the second most conserved motif in higher eukaryotes (Xie *et al.*, 2005). A large number of USF1 and USF2 binding sites were identified in the course of this study and this data could be used to confirm the presence of E-box motifs in USF sites or identify a novel USF binding motif. Similarly a large number of mSin3a interactions were identified in the ENCODE regions but as mSin3a does not directly interact with DNA itself but instead interacts with sequence specific DNA binding proteins, an analysis of mSin3a interacting sequences may identify a motif associated with its DNA-binding partner. To these ends, the mSin3a, USF1 and USF2 data sets from the K562 cell line were analysed for enrichment of known motifs contained in JASPAR






(Bryne *et al.*, 2007) and TRANSFAC (Matys *et al.*, 2006) databases and for novel motifs using the DME program (Smith *et al.*, 2005) as described previously. No known motif in JASPAR or TRANSFAC displayed high sensitivity and specificity in distinguishing foreground sequences from background sequences in the mSin3a dataset (data not shown). Therefore the identification of novel motifs by the DME program was investigated. Several novel motifs were discovered to be associated with mSin3a interactions and may represent potential binding motifs for an mSin3a interacting factor (Table 4.7). However, it must be noted that many of these motifs have a high GC content and many of the mSin3a binding sites overlap with promoter regions that have CpG islands.

Name	Logo	Sn	Sp	Error	pvalue
1. DME0001		0.626	0.565	0.404	0
2. DME0002		0.432	0.752	0.408	0
3. DME0003		0.371	0.806	0.411	0.0002
4. DME0004		0.619	0.556	0.412	0.0003
5. DME0005		0.565	0.594	0.421	0.0019






**Table 4.7: Discovery of novel motifs associated with mSin3a interactions.** The DME program was used to identify novel motifs in genomic regions associated with mSin3a binding. The GC-rich DME0001 motif was found to be the best motif for distinguishing the foreground and background sequences with a relative error rate of 0.404 at a p-value cutoff of 0. The sensitivity (Sn) associated with a motif and p-value cut-off is the proportion of true foreground sequences that were classified as foreground; the specificity is the proportion of true background sequences classified as background sequences. The relative error rate (Error) for a given motif and p-value cut-off is  $1-(Sn+Sp)/2$ . A good motif has a low error rate and balanced Sn and Sp values (Smith *et al.*, 2005). DNA logos are presented for each motif and the height of each letter indicates relative occurrence of nucleotides in the binding sites.

The E-box motif (MA0093) in the JASPAR database best distinguished foreground sequences from background sequences for the USF1 binding sites (Table 4.8), followed closely by the USF1 motif from TRANSFAC (M00121), which contains an E-box motif at its core. No known USF2 binding motif is present in these databases but this study shows that the USF1 motif (M00121) also best identifies USF2 binding sites. Only 19

USF2 only sites were identified when studying co-localisation with CTCF, mSin3 and USF1 and 172 USF2 sites overlapped with USF1 sites. This suggests that USF1-USF2 heterodimers are more common than USF2 binding alone and is consistent with the same binding motif being identified for the two factors. Alternatively, this could also suggest that the two antibodies used in ChIP-chip for USF1 and USF2 cross-reacted to some degree. Novel motif analysis using the DME program also identified the same E-box motifs for USF1 and USF2 (data not shown). Therefore CTCF and USF1 sites were associated with known binding motifs, while USF2 binding was also associated with an E-box motif and novel GC rich motifs were associated with mSin3a interactions.

A		Name	Logo	Sn	Sp	Error	pvalue
1.	MA0093		0.679	0.602	0.359	0	
2.	M00121		0.565	0.707	0.364	0	
3.	M01029		0.571	0.693	0.368	0	
4.	M00187		0.584	0.673	0.372	0	
5.	MA0058		0.54	0.702	0.379	0	

B		Name	Logo	Sn	Sp	Error	pvalue
1.	M00121		0.484	0.845	0.336	0	
2.	MA0104		0.484	0.813	0.352	0	
3.	M00055		0.457	0.836	0.354	0	
3.	MA0093		0.406	0.886	0.354	0	
5.	M00217		0.484	0.806	0.355	0	

**Table 4.8: Searching for known motifs in USF1 and USF2 binding sites.** Known motif matrices from JASPAR and TRANSFAC databases were enriched in USF1 (panel A) and USF2 (panel B) binding sites. The E-box motif in the JASPAR database (MA0093) and the USF1 motif in TRANSFAC (M00121), which contains the E-box sequence at its core, were found to be the best performing known motifs for distinguish foreground sequences from background sequences in USF1 and USF2 ChIP-chip data sets respectively.

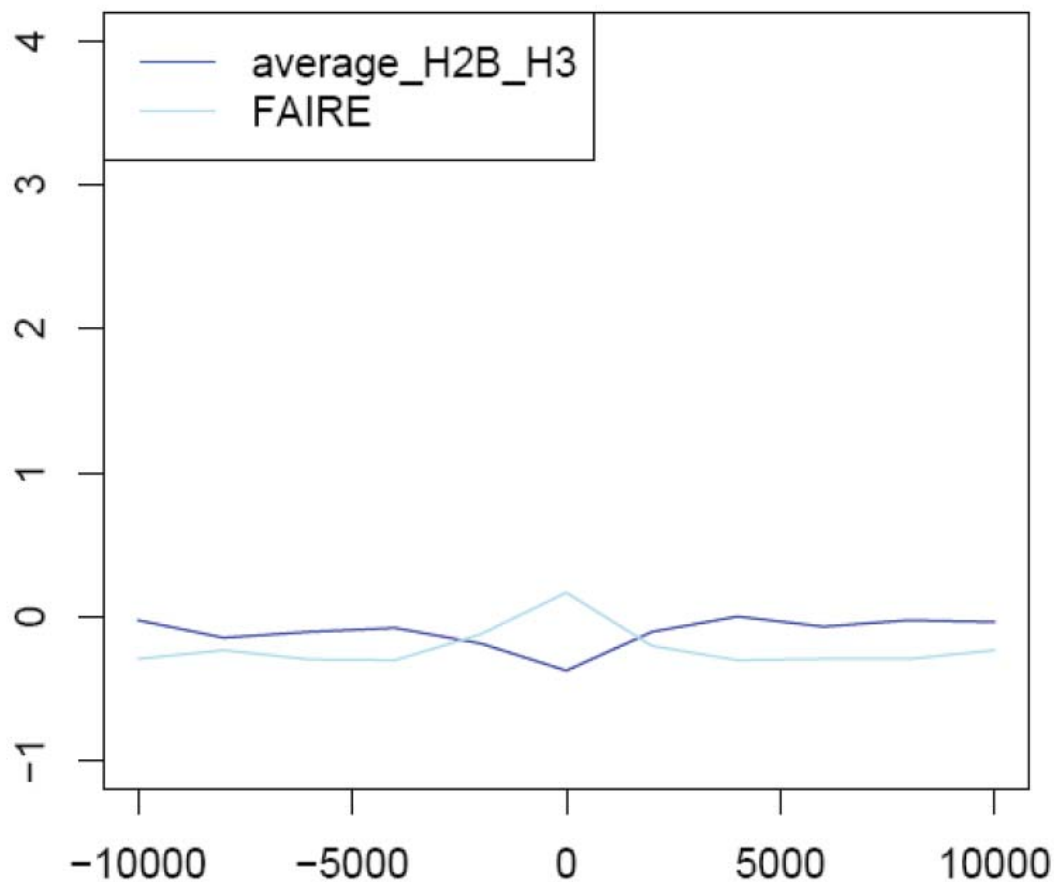
#### **4.9. Chromatin structure at insulators**

CTCF has been implicated in establishing local chromatin structure at a number of loci (Cho *et al.*, 2005; Filippova *et al.*, 2001) and has also been found at the transition regions between chromosomal domains of X inactivation and escape (Filippova *et al.*, 2005) suggesting that CTCF may influence local chromatin conformation to facilitate barrier insulator function. In addition, chromatin structure at the nucleosome level has been shown to be important for the binding of CTCF (Kanduri *et al.*, 2002). Therefore, the chromatin properties of CTCF binding sites were examined to gain further insights into these processes.

##### **4.9.1. CTCF binding sites are located in accessible chromatin domains**

While regulatory elements such as promoters and enhancers are known to be associated with DNase I hypersensitive regions in the human genome (Follows *et al.*, 2006) it is not clear if chromatin accessibility is a general feature of insulators elements in the human genome. Chromatin accessibility at CTCF sites was examined using three different but complementary ENCODE datasets generated using K562 cells – DNase I hypersensitive data (Xi *et al.*, 2007), histone H2B and H3 data (Chapter 3), and formaldehyde assisted isolation of regulatory elements (FAIRE) data (obtained courtesy of Dr Pawan Dhama and Dr. Alex Bruce, Wellcome Trust Sanger Institute). Xi and colleagues recently performed a study of DNase I hypersensitive sites in the ENCODE regions using the DNase-chip method and identified over 1200 hypersensitive sites in K562 cells (Xi *et al.*, 2007). This data was publicly available and the location of DNase I hypersensitive sites were compared with the location of CTCF sites. More than half (296 of 571) of the CTCF sites were found to overlap with one or more hypersensitive sites. Histone density and FAIRE are inversely correlated at regions of open chromatin (Dhama, submitted; Giresi *et al.*, 2007) so the results of both assays were compared to CTCF binding. The averaged H2B/H3 z-scored  $\log_2$  values were calculated across a 20 kb window across all 571 CTCF sites. A depletion of these core nucleosome proteins was observed approximately 2 kb upstream and downstream of CTCF binding sites, with the greatest depletion observed at the centre of the CTCF binding sites (Figure 4.20). In the FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) procedure, formaldehyde

cross-linked and sonicated chromatin was phenol-chloroform extracted to identify nucleosome-depleted DNA (Giresi *et al.*, 2007). Genomic regions depleted of nucleosome proteins were enriched in the aqueous phase following phenol-chloroform extraction and the purified DNA was hybridised to a microarray in a similar fashion to a ChIP experiment. CTCF sites were associated with a peak of enrichment in this assay, further supporting the hypothesis that CTCF binds at regions of open chromatin in the human genome.

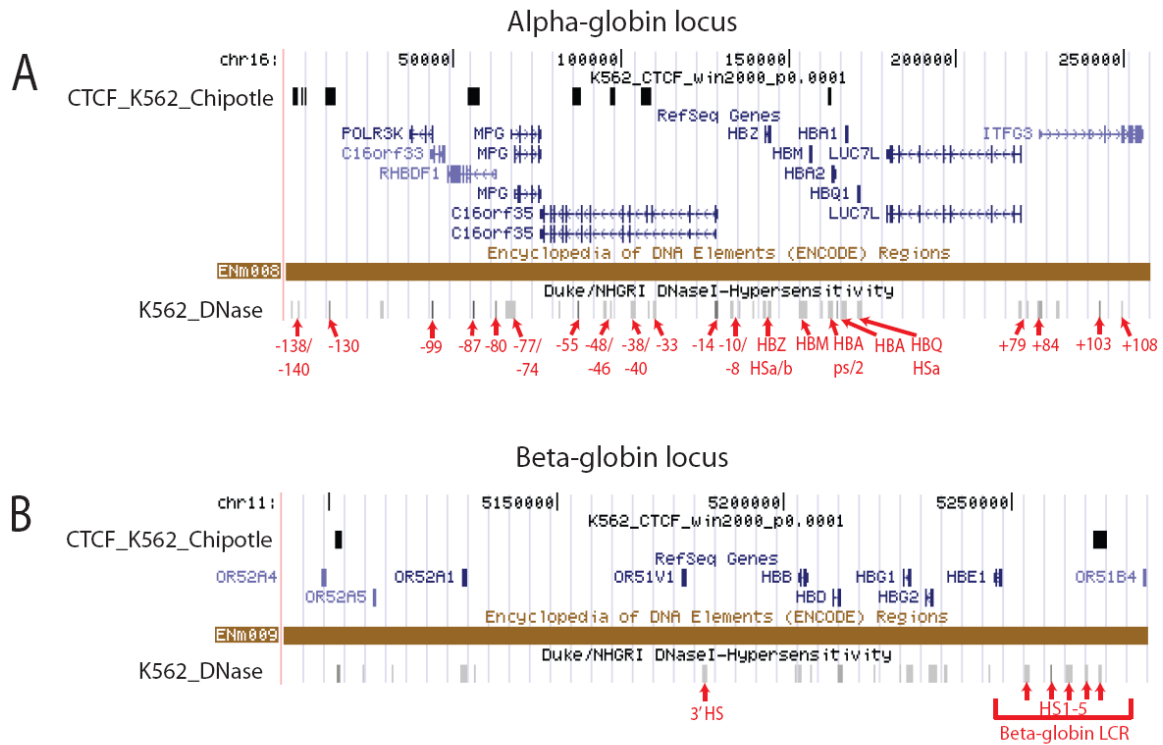


**Figure 4.20: Nucleosome density and FAIRE profile at CTCF binding sites in the ENCODE regions.** CTCF binding sites in K562 are associated with a large region of histone H2B and H3 depletion (approximately 4 kb) as can be seen by the depletion in the average histone H2B and H3 signal (dark blue profile). This depletion in nucleosomes correlates with a peak of FAIRE enrichment (light blue profile). The values along the x-axis represent distance in base pairs up and downstream of CTCF binding sites. The scale on the y-axis represents z-scored log<sub>2</sub> ratios.

Several ENCODE regions have been well-characterised in terms of the location and function of DNase I hypersensitive sites. For example, DNase I hypersensitive sites have been systematically mapped at the human  $\alpha$ -globin locus and functionally tested in a number of studies (Higgs *et al.*, 1990; Jarman *et al.*, 1991; Sharpe *et al.*, 1993; Vyas *et al.*, 1995). CTCF binding sites at the  $\alpha$ -globin locus in K562 cells were compared with the location of DNase I hypersensitive sites and all CTCF sites overlapped with hypersensitive sites in this region (Figure 4.21). Four of the CTCF binding sites were associated with previously identified hypersensitive sites at HBAs/2 located near the HBA2 gene, the -33 region, the -48/-46 region, and the -55 region, all of which are located within the c16orf35 gene. Three other hypersensitive sites located further upstream also contained CTCF binding sites, namely the -87 region, -130 region and the -138/-140 region. The -87 region was located within the RHBDF1 gene, while the -130 and -138/-140 regions were located approximately 30 and 40 kb from the POLR3K promoter. CTCF binding sites have been identified upstream of the chicken alpha-globin locus (Valadez-Graham *et al.*, 2004; Klochkov *et al.*, 2006) and, whilst this region is conserved and located upstream of the human HBZ gene, no binding site was identified in this region in K562 cells.

In erythroid cells, a region upstream of the  $\beta$ -globin genes contains a series of DNase I hypersensitive sites that comprise the locus control region (LCR), which regulate  $\beta$ -globin gene expression in erythroid cells. The chicken  $\beta$ -globin LCR contains hypersensitive site 5' HS4 which binds CTCF while another site outside of the LCR, 3' HS1, also binds CTCF and together these two sites form the boundaries of the  $\beta$ -globin locus in chicken erythrocytes (Bulger *et al.*, 1999). Homologous hypersensitive sites are also observed in the human  $\beta$ -globin locus, HS5 and 3'HS, which also bound CTCF in K562 cells (Farrell *et al.*, 2002). Unlike the chicken  $\beta$ -globin locus, neither element possesses barrier activity and both have been proposed to function as enhancer blocking elements. In this study, HS5 was also found to bind CTCF, but no CTCF binding was identified at the 3'HS (Figure 4.12). However, a novel CTCF binding site was identified upstream of the locus between the olfactory receptor genes OR52A4 and OR52A5, which also correlated with a hyper-sensitive site in K562 cells. The work presented in this thesis

is the first demonstration of CTCF binding in the  $\alpha$ -globin locus in a human erythrocytic cell line (K562) and also confirms binding of CTCF in the  $\beta$ -globin locus. This suggests that CTCF may play an important role in the regulation of haemoglobin synthesis.



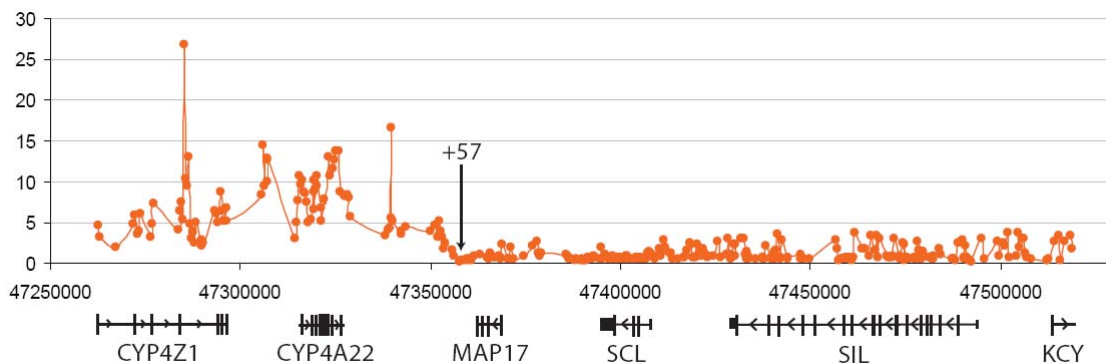
**Figure 4.21: CTCF binding sites co-localise with sites of DNase I hypersensitivity at the  $\alpha$ - and  $\beta$ -globin loci.** Panel A: A 250 kb region of Enm008 containing the  $\alpha$ -globin locus and adjacent genes is shown. K562 CTCF ChIPOTle sites are shown as black bars at the top of the figure and align with DNase I hypersensitive (HS) sites (grey bars at the bottom of the figure) identified in K562 cells. The numbering of HS sites refers to kilobase distances up or downstream of HBZ exon 1 (Follows *et al.*, 2006). Panel B: DNase-chip identified the five HS sites of the human  $\beta$ -globin LCR in K562 in addition to the 3' HS site. HS5 co-localises with a CTCF binding site in K562 cells in this study but no CTCF interaction was observed at the 3'HS. Another HS site in the olfactory receptor cluster was associated with CTCF binding in K562.

#### 4.9.2. CTCF is located at the boundary between active and inactive chromatin domains

The +57 CTCF binding site at the SCL locus represents a transition from a chromatin region containing inactive liver-specific genes (CYP4Z1 and CYP4A22 genes) to a



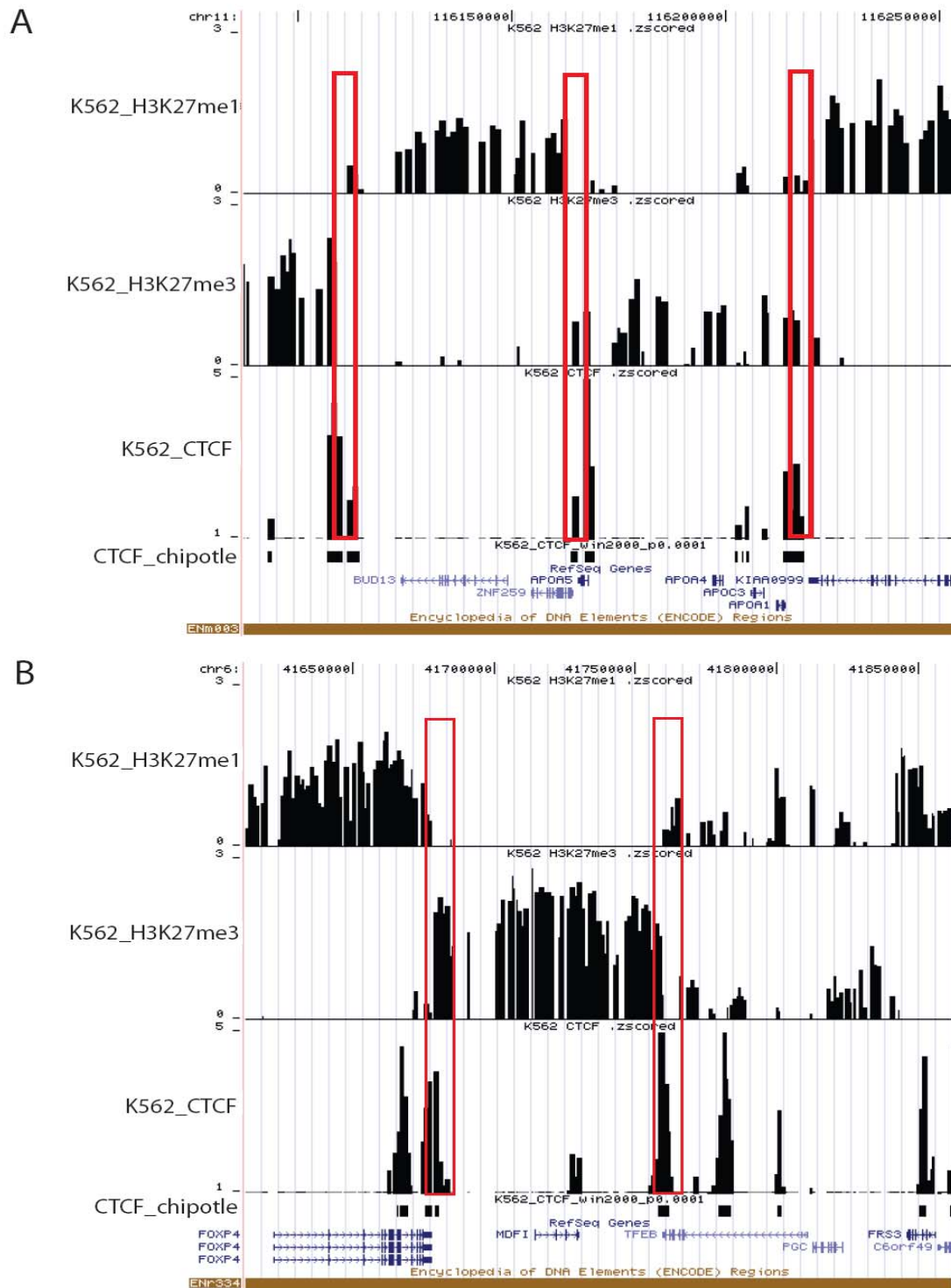
genomic region containing actively expressed genes in K562 that are involved in erythroid development (Dhami, submitted). Therefore it was hypothesized that the +57 region may act as barrier insulator to prevent the spread of silencing histone modifications associated with the inactive CYP4Z1 and CYP4A22 genes. The formation of transcriptionally inactive chromatin domains is an important mechanism for silencing of specific genes during developmental programming (Craig, 2005) and as the H3K27me3 modification has been implicated in the formation and maintenance of silent chromatin regions via the recruitment of Polycomb group proteins (Cao *et al.*, 2002; Lee *et al.*, 2006; Boyer *et al.*, 2006), ChIP-chip experiments were performed with K562 cells to identify regions of the SCL locus associated with elevated levels of H3K27me3 (Dhami, submitted) (Figure 4.22). As can be seen in the Figure, the liver-specific CYP genes are associated with high levels of H3K27me3 in K562 cells, while SCL and other active genes contain low levels of this modification. This is consistent with the formation of an inactive chromatin domain at the CYP region in K562 cells. This modification is enriched over the entire body of both genes, while little enrichment for this modification is observed over the expressed SCL, SIL and KCY genes. There seemed to be a clear transition point at +57 from high levels of H3K27me3 associated with the CYP genes to low levels associated with actively transcribed genes. This data suggest that this CTCF binding site acts as a barrier element to prevent the spread of the silencing H3K27me3 modification associated with the inactive CYP region into the nearby actively expressed MAP17 and SCL genes.



**Figure 4.22: ChIP-chip profile of Histone H3K27me3 across the SCL locus in K562.** The CYP region is associated with high levels of H3K27me3, while the MAP17, SCL, SIL and KCY genes are associated with low levels of H3K27me3. The transition from a high to low H3K27me3 enrichment coincides with the

binding of CTCF at the +57 region. The human chromosome 1 genomic coordinates, gene order and direction of transcription are indicated along the y-axis, while fold enrichments are indicated on the x-axis.

The distribution of CTCF binding sites in the ENCODE regions was compared with K562 histone H3K27 methylation profiles generated by others at the Sanger Institute (courtesy of Drs. Pawan Dhama and Alex Bruce) to determine the location of other putative barrier insulators. H3K27me1 and H3K27me3 have been found to be associated with neighbouring active and inactive chromatin domains respectively in K562 cells (Dhama and Bruce, unpublished data). These domains were therefore examined for binding of CTCF. 38 CTCF binding sites were identified which demarcated active chromatin regions from inactive chromatin regions (Table 4.9), two of which are presented in Figure 4.23. Panel A illustrates the dynamic change in H3K27me3 and H3K27me1 states at ENCODE region Enm003 and a CTCF binding sites were located at chromatin state transition points. This region contains the APO cluster of genes (APOA1, APOA5, APOA4, and APOC3) involved in apolipoprotein metabolism. These genes are mainly expressed in the liver and small intestine and are associated with elevated levels of H3K27me3 in K562. However the neighbouring BUD13, ZNF259 and KIAA0999 genes are associated with elevated levels of H3K27me1. CTCF binding sites are located at boundaries between active and inactive chromatin domains suggesting that CTCF in combination with other factors may prevent the spread of H3K27me3 into active chromatin regions or vice versa. Figure 4.23 panel B illustrates the same phenomenon in another ENCODE region (Enm334). The fork-head box P4 transcription factor (FOXP4) is involved in cancer progression and is expressed in K562. This gene is associated with elevated levels of H3K27me1 while the nearby MDFI gene, a MyoD inhibitor involved in cartilage formation is not expressed in K562 and is associated with high levels of H3K27me3. CTCF once again forms a barrier between an inactive chromatin region associated with high H3K27me3 levels and a neighbouring active chromatin region associated with high H3K27me1 levels.



**Figure 4.23: CTCF binding sites demarcate the boundary between chromatin regions associated with active and inactive histone modifications.** Panel A: the top track (K562\_H3K27me1) represents histone H3K27me1 enrichments in ENCODE region enm003, the middle track (K562\_H3K27me3) represents

H3K27me3 enrichments and the bottom track represents CTCF enrichments and ChIPOTle defined CTCF sites are indicated below this track as black bars. CTCF sites located at the boundary between chromatin regions associated with H3K27me1 or H3K27me3 are indicated by red boxes. Panel B: ENCODE region enr334 contains CTCF binding sites that demarcate active and inactive chromatin regions. RefSeq genes in the two regions are presented below the CTCF ChIPOTle tracks, while chromosome coordinates are presented at the top of each panel. Fold enrichments for histone modifications and CTCF are presented as  $\log_2$  values (scale indicated on the y-axis of each panel).

Twenty one of the 44 ENCODE regions contained at least one putative CTCF chromatin barrier elements while 9 ENCODE regions contained multiple barrier elements (Table 4.10). USF1 and USF2 recruit histone modifying enzymes to the chicken beta-globin barrier insulator (West *et al.*, 2004; Huang *et al.*, 2007) so the pattern of USF1 and USF2 binding was examined at the 38 putative barrier elements (Table 4.10). 6 of the 38 sites were found to co-localise with a USF1 binding site, while 7 other putative barrier insulators overlapped with both USF1 and USF2 binding sites. This suggests that USF factors may play a role in recruiting histone modifying enzymes at 13 of the 38 (34%) putative barrier elements. In total 140 of the 571 CTCF sites co-localised with USF1 and/or USF2 sites, representing 25% of CTCF interactions. As 34% of CTCF sites located at putative barriers co-localised with USF1 or USF2, this suggested that there may be an over-representation of USF binding at putative barrier elements relative to the total percentage of co-localised CTCF and USF sites. However, the vast majority (91%) of CTCF sites which co-localised with USF1 and/or USF2 do not seem to be involved in barrier function.

Region	CTCF boundary coordinates	Overlapping USF1 coordinates	Overlapping USF2 coordinates
Enm001	chr7:115760053-115761552	N/A	N/A
Enm003	chr11:116111357-116114356	N/A	N/A
	chr11:116166857-116169356	N/A	N/A
	chr11:116213357-116218356	chr11:116216857-116218356	N/A
Enm004	chr22:30690594-30692093	N/A	N/A
	chr22:30693094-30694593	chr22:30694094-30694593	N/A
	chr22:30698594-30700093	N/A	N/A
	chr22:31252844-31253343	Chr22:31253094-31258093	Chr22:31253594-31254093
Enm005	chr21:32687032-32688531	N/A	N/A
	chr21:32906532-32907531	chr21:32905998-32907497	chr21:32905998-32907497
	chr21:33492032-33494531	N/A	N/A
Enm006	ChrX:152696278-152697777	ChrX:1562694278-152696277	N/A
	ChrX:153093278-153094777	N/A	N/A
Enm007	Chr19:59406293-59407792	Chr19:59403793-59407792	Chr19:59403793-59406292
Enm009	Chr11:5101319-5102818	N/A	N/A
Enm011	Chr11:1720702-1723701	N/A	N/A
	Chr11:1750702-1751201	N/A	N/A
	Chr11:1914702-1917201	N/A	N/A
	Chr11:1978202-1979701	N/A	N/A
Enm014	Chr7:126629702-126631201	N/A	N/A
Enr121	Chr2:118309511-118312010	N/A	N/A
	Chr2:118286511-118290010	Chr2:118288511-118290010	Chr2:118288511-118290010
Enr131	Chr2:234522984-234524483	Chr2:234521984-234523483	N/A
Enr132	Chr13:112390066-112391565	N/A	N/A
Enr133	Chr21:39444495-39450994	N/A	N/A
Enr223	Chr6:74075593-74077592	N/A	N/A
Enr232	Chr9:129014732-129016231	Chr9:129014732-129015231	N/A
	Chr9:129243732-129244731	N/A	N/A
Enr233	Chr15:41668924-41671423	Chr15:41668527-41676026	Chr15:41668527-41670526
	Chr15:41768424-41770923	Chr15:41768527-41776026	Chr15:41768527-41771026
	Chr15:41981424-41984923	Chr15:41982527-41986026	N/A
Enr322	Chr14:98921983-98931982	N/A	N/A
Enr331	Chr2:220309066-220310565	N/A	N/A
Enr332	Chr11:64267062-64268561	Chr11:64267062-64268561	Chr11:64266812-64267311
Enr333	Chr20:33352934-33358433	N/A	N/A
	Chr20:33499434-33500933	N/A	N/A
	Chr20:33668434-33669933	N/A	N/A
Enr334	Chr6:41678954-41680453	N/A	N/A

**Table 4.9: CTCF binding sites in K562 located at the boundary between active and inactive genomic regions defined by enriched levels of H3K27me1 and H3K27me3 respectively.** The 21 ENCODE regions which contained CTCF sites at chromatin boundaries are indicated in the first column, while the second column contains the genomic coordinates of the 38 proposed CTCF boundary elements. The third

and fourth columns contain USF1 and USF2 binding coordinates which co-localise with these CTCF boundary sites. N/A: not applicable, i.e. no co-localisation of USF1/2.

#### **4.10. Discussion**

Insulator elements, which can have either enhancer-blocking activity or act as barriers between genomic regions of active and inactive chromatin, represent an important class of regulatory element for which little information is available in the literature. Therefore, in this study K562 and U937 cells were used to map the location of the insulator binding protein, CTCF, using a ChIP-chip strategy. The Sanger Institute ENCODE microarray was used to map the location of 571 and 661 CTCF sites in K562 and U937 cells respectively. The distribution of CTCF was examined and motifs associated with CTCF binding sites were analysed. In addition, CTCF binding sites were further characterised by investigating the binding distribution of mSin3a, USF1, and USF2, factors which have been implicated in CTCF transcriptional repression and insulator functions (Lutz *et al.*, 2000, West *et al.*, 2004). As a means of identifying putative barrier insulators, the chromatin structure at CTCF binding sites was also investigated

##### **4.10.1. Widespread distribution of CTCF binding sites in the human genome**

CTCF is known to be a highly versatile transcription factor that can regulate gene expression by different modes of action, such as promoter activation and repression, and constitutive- and methylation-dependent chromatin insulation (Filippova, 2008). In order to gain a greater understanding of insulators, the study presented in this thesis used a ChIP-chip method to map CTCF binding sites in 1% of the human genome and approximately 600 CTCF binding sites were identified in two different cell lines. The distribution of CTCF sites in K562 was examined and while most CTCF binding sites were located far from TSSs, CTCF binding was not randomly distributed across the genome but followed the distribution of genes. A similar correlation was also observed by Xie and colleagues (Xie *et al.*, 2007). Further analysis showed that CTCF sites classified into three categories: promoters bound by CTCF, distal enhancers/repressors bound by CTCF and “other” locations. The majority of CTCF sites were located at “other” locations which is consistent with an insulator function as these sites are not

associated with promoter or enhancer activity. The varied localisation of CTCF binding sites is consistent with previous studies which reported that CTCF can bind at intron, exon, promoter, and intergenic locations (Mukhopadhyay *et al.*, 2004, Kim *et al.*, 2007, Barksy *et al.*, 2007). This study also confirmed the findings of Kim and colleagues (2007) who noted that CTCF binding sites often occurred in what they described as CTCF-paired domains (CPDs). They found that 74% of genes in the human genome are surrounded completely by CTCF binding sites. In this study approximately half of the ENCODE genes were flanked in their entirety by CTCF binding sites, while a number of clusters of related (and unrelated genes) were also observed to be flanked by CTCF in this study. This suggests that CTCF can regulate individual genes or groups of genes. CTCF sites flanking clusters of related or apparently unrelated genes may function to ensure that enhancers associated with a gene cluster cannot inadvertently activate other genes outside the cluster and similarly enhancers associated with genes outside the cluster cannot influence the regulation of the genes within a cluster. Multiple CTCF binding sites can also be located across individual genes and may be used to control the interaction of multiple lineage- or temporal-specific enhancers, or some may be located at enhancers where their function is not insulator activity but transcription factor activity.

#### **4.10.2. Cross-study comparison of CTCF binding sites**

This study has demonstrated the accuracy of the ChIP-chip method in detecting previously characterised CTCF binding sites located in the ENCODE regions. In addition to identifying known CTCF binding sites, approximately 600 novel CTCF interactions were identified in K562 and U937 cells in this study. As the ENCODE regions were chosen to contain features representative of the entire human genome sequence, this suggests that there may be as many as 50,000-60,000 CTCF binding sites in the human genome. Mukhopadhyay and colleagues had previously mapped the location of 200 CTCF binding sites in the mouse genome using a ChIP-chip approach, the majority of which display insulator functions in assays, and estimated that the total number of CTCF binding sites in the mouse genome would be around 4,000 (Mukhopadhyay *et al.*, 2004). More recently Xie and colleagues used sequence conservation analysis to predict the location of 15,000 CTCF binding sites in the human genome (Xie *et al.*, 2007). However,

this number does not take into account CTCF binding sites which are not conserved in other mammalian species. Recent genome-wide ChIP studies derived data sets have identified more CTCF binding sites in the human genome. Barski and colleagues identified approximately 20,000 CTCF binding sites in human CD4<sup>+</sup> T cells using a high-throughput ChIP-sequencing method (Barski *et al.*, 2007) while Kim and colleagues used a whole genome ChIP-chip approach to map CTCF binding sites in human cells and used their experimentally derived consensus motif to predicted the location of approximately 30,000 CTCF binding sites in the human genome (Kim *et al.*, 2007). However the authors also note that 25% of their experimentally defined CTCF sites do not contain the consensus motif indicating that more than 30,000 CTCF binding sites are located in the human genome.

Analysis of CTCF binding sites in the ENCODE regions revealed that Kim *et al.* identified 225 and 232 sites in IMR90 and U937 cells respectively, while Barski and colleagues identified 353 CTCF binding sites in CD4<sup>+</sup> T cells. In contrast this study identified approximately 600 CTCF binding sites in K562 and U937 cells, 500 of which were not detected by Kim and colleagues when a direct comparison of CTCF binding sites in U937 cells was performed. This study and that of Kim and colleagues used similar ChIP protocols suggesting that antibody quality or sensitivity of array platform may be responsible for the large difference in the number of CTCF sites identified between the two ChIP-chip studies. In addition, the possibility that ChIPOTle is over-calling CTCF sites cannot be excluded, although a stringent p value threshold was used to minimize this possibility. The CTCF antibody used by Kim and colleagues was a mixture of monoclonal antibodies while the antibody used in this study was a polyclonal one. Small changes in the structure or accessibility of an epitope upon cross-linking can dramatically affect the function of a monoclonal antibody. In contrast, because polyclonal antibodies recognise a number of antigenic epitopes, the effects of cross-linking are less of a problem. Therefore, the reduced number of CTCF sites identified by Kim and colleagues relative to this study may be the due to the type of antibody used in the ChIP procedure.

Different array platforms may also have affected the number of CTCF sites identified by the two studies. Kim and colleagues used Nimblegen oligonucleotide arrays and



oligonucleotide probes are not as good a hybridisation element as larger PCR product probes. Nimblegen arrays also require ChIP DNAs to be amplified prior to labeling with fluorescent dyes. Any bias during the amplification process could change the representation of the sample and some CTCF interacting sequences may not be detected following microarray hybridisation. Sanger Institute in-house constructed PCR product microarrays do not require amplification of ChIP DNAs prior to hybridisation illustrating that this platform may be more sensitive than commercially-available microarrays.

In addition, different analysis methods were used in the two studies which may have affected how many CTCF sites were identified. Kim *et al.* normalised microarray data by a Lowess (locally weighted regression) normalisation method (Berger *et al.*, 2004.), which removes intensity-dependent effects in  $\log_2$  ratios. Probability statistics were then calculated for each probe based on a single array error model (SAEM) (Li *et al.*, 2003). Kim *et al.* (2007) also used a very stringent statistical threshold ( $p$  0.000001) to obtain a similar number of CTCF binding sites in U937 cells as detected in IMR90 cells. In contrast microarray data obtained in this study was total intensity normalised (Quackenbush, 2002), which assumes that total hybridisation intensities summed over all array elements should be the same for each sample. A normalisation factor is calculated by summing measured intensities in both the Cy3 and Cy5 channels, which adjusts each ratio such that the mean ratio is equal to 1. In addition each data point was then normalised by dividing with the corresponding value obtained from a goat IgG ChIP-chip experiment which can help to identify authentic sites that are subtly enriched prior to normalisation. Goat IgG normalisation can also help to remove non-specific enrichments from a ChIP-chip data set. Binding sites were then detected in normalised data using the ChIPOTle program with a high stringency threshold ( $p$ 0.0001). ChIPOTle has been shown to be more accurate than SAEM in accurately detecting TF binding events using ChIP-chip data (Buck *et al.*, 2005). Therefore a combination of the very high stringency threshold chosen by Kim and colleagues coupled with the use of SAEM and no species-specific IgG normalisation may have resulted in the identification of fewer CTCF sites compared to this study.

#### **4.10.3. CTCF co-localisation with mSin3a, USF1 and USF2 binding sites**

As described in the introduction to this Chapter, CTCF interacts with a diverse number of factors which modulate the activity of this multifunctional protein. In this study, CTCF co-localisation/co-operation with other factors was examined by performing ChIP-chip experiments to identify mSin3a, USF1 and USF2 interactions that overlapped with CTCF sites to gain a greater understanding of CTCF function. CTCF binding sites which overlapped with mSin3a binding sites may be involved in transcriptional regulation while those CTCF sites which overlapped with USF1/2 binding may be barrier elements similar to the chicken  $\beta$ -globin locus. One third of CTCF binding sites co-localised with binding sites for one or more of these transcription factors and these overlapping sites were located at active and inactive TSSs, distal enhancer/repressor sites and “other” locations further demonstrating that that CTCF is a highly versatile factor in terms of binding location, function and potential interacting partners. No clear correlation between CTCF sites overlapping with mSin3a and gene repression was observed – however, surprisingly, CTCF and mSin3a co-localisation at promoters was associated with active gene expression (see section 4.10.4). However, the picture is indeed complex as CTCF and mSin3a binding sites also overlapped at diverse genomic locations including active and inactive TSSs and distal sites. Similarly, USF1 and USF2 binding sites overlapped at a number of CTCF sites, a small percentage of which may function as barrier insulators based on the presence of histone H3K27 methylation modifications. In addition CTCF, mSin3a, USF1 and USF2 overlapping interactions were observed in various combinations suggesting that the modulation of CTCF function is highly complex and that no clear functional categorisation of CTCF binding sites can be performed based solely on co-localisation with mSin3a or USF1 or USF2.

#### **4.10.4. mSin3a interaction at promoters is associated with active gene expression**

While mSin3a co-localisation with CTCF was limited, due to the number of sites analysed in this study, and did not allow for a clear functional categorisation of CTCF sites, an interesting observation arose from this study which suggested that the accepted view of mSin3a acting as co-repressor of gene expression may not be always correct. In

this study the majority of mSin3a binding events at transcription start sites were associated with actively expressed genes and a ROC analysis identified that the presence of mSin3a at TSSs was as accurate an indicator of active gene expression as the presence of H3K4me3 or H3 acetylation (see Chapter 3). Further evidence supporting a role for mSin3a involvement in gene activation comes from studies of the yeast homologue Sin3. De Nadal and colleagues proposed that yeast mitogen-activated protein kinases Hog1 induces gene expression by recruiting the Rpd3-Sin3 histone deacetylase complex complex to the promoters of genes regulated by osmostress (De Nadal *et al.*, 2004). More recently Sharma and colleagues described how the Rpd3-Sin3 complex is required for the activation of DNA damage inducible genes (Sharma *et al.*, 2007) and a similar phenomenon was observed by Sertil and colleagues who described that Rpd3-Sin3 was required for the transcriptional induction of anaerobic genes (Sertil *et al.*, 2007). An emerging theme in yeast studies is that histone deacetylase (HDAC) complexes containing Sin3 are required for resetting promoters in the wake of elongating RNA Polymerase II to prevent spurious transcription initiation (Lee and Shilatifard, 2007). Thus far there are no examples in the literature of mSin3a associating with active promoters in mammalian genomes but the data presented in this thesis suggests that mSin3a is predominantly associated with actively expressed genes and may recruit HDACs to ‘reset’ these promoters following the passage of RNA polymerase II.

#### **4.10.5. Identification of transcription factor consensus binding motifs**

Approximately one third of the CTCF binding sites identified in this study overlapped with the CTCF consensus binding motif identified by Kim and colleagues (2007). This suggested that other CTCF recognition motifs exist, a fact pointed out by Kim and colleagues who noted that approximately 20% of CTCF binding sites did not contain this motif but could bind CTCF when further characterised *in vitro*. In addition a number of previously characterised CTCF binding sites do not contain this motif suggesting that CTCF recognises a number of DNA sequences (Filippova, 2008). *De novo* motif analysis in this study failed to identify a novel motif suggesting that many CTCF sites may not be associated with a consensus motif.

USF1 binding sites identified by ChIP-chip experiments were enriched for an E-box motif present in the TRANSFAC database (Matys *et al.*, 2006). In addition the ChIP-chip approach was used to define an identical consensus sequence for USF2 for which no *in vitro* motif had been previously established. While USF1 and USF2 were not extensively linked to insulators, they are important regulators of genes expression (Di Duca *et al.*, 2006; Pezzolesi *et al.*, 2007) and identifying the genome-wide binding sites of these TFs is important for understanding their function. A number of novel GC rich motifs associated with mSin3a interaction were identified as part of this study. As mSin3a is not known to interact directly with DNA itself, but is recruited by other sequence-specific TFs, this suggests that the mSin3a binding partner(s) may have a preference for GC rich binding sequences.

#### **4.10.6. CTCF and chromatin structure**

CTCF has been implicated in regulating local chromatin domains (reviewed in Filippova 2008) and in this study the local chromatin structure at CTCF binding sites was assessed using nucleosome data, DNase I hypersensitive site data and FAIRE data. CTCF binding sites were depleted of histones H2B and H3 and were enriched in FAIRE assays consistent with being located in ‘open’ chromatin regions like other regulatory elements (Dhami, submitted; Giresi *et al.*, 2007). This is consistent with the observation that binding of CTCF to its target sites may be controlled by nucleosome occupancy as CTCF is unable to interact with a target site if it is occupied with a nucleosome (Kanduri *et al.*, 2002). In addition more than 50% of K562 CTCF binding sites were associated with DNase I hypersensitive sites. A recent study also showed that approximately 70% of the 225 CTCF binding sites identified in IMR90 cells in the ENCODE regions overlapped with DNase I hypersensitive sites suggesting that CTCF preferentially binds in accessible regions of the genome (Xi *et al.*, 2007). A number of well-characterised DNase I hypersensitive sites at the  $\alpha$ - and  $\beta$ -globin loci were associated with CTCF binding in K562 cells. CTCF sites were located at  $\alpha$ -globin sites which included HBAs/2 at the HBA2 promoter, -33, -48/-46 and -55 regions (all within C16orf35). Three uncharacterised hypersensitive sites located further upstream, at -87, -130 and -138/-140 and were also associated with CTCF binding. The HS5 of the  $\beta$ -globin LCR was

associated with CTCF in K562, but unlike a previous report (Farrell CM 2002) the 3'HS was not associated with CTCF in this study. The nearest site to the 5'HS that was associated with CTCF binding in this study was located further upstream between two olfactory receptor genes. However, the presence of CTCF sites at these two loci in a relevant cell type provides evidence that CTCF may be involved in the regulation of haemoglobin synthesis.

CTCF has also been implicated in the formation of chromatin domains that escape X-inactivation during early development (Filippova *et al.*, 2005) and a number of CTCF binding sites have been detected between active and silent chromatin domains in a recent study (Barski *et al.*, 2007). Barski and colleagues described how several large regions of chromatin containing inactive genes were associated with high levels of H3K7me3 and neighbouring chromatin regions containing actively transcribed genes were associated with H3K27me1 (Barski *et al.*, 2007). These active and inactive chromatin domains were separated by CTCF binding sites. A similar phenomenon was observed in K562 cells in this study. Nearly 40 CTCF sites were located at the boundary between regions of active and inactive chromatin associated with H3K27me1 and H3K27me3 respectively. While USF1 and USF2 function as barrier proteins by recruiting a histone methyltransferases and histone acetyltransferases to prevent the spread of heterochromatin at the chicken  $\beta$ -globin insulator (West *et al.*, 2004; Huang *et al.*, 2007), only one third of CTCF boundary sites co-localised with either USF1 or USF2 in a human cell type. This suggests that the chicken  $\beta$ -globin USF-mediated barrier insulator model may not be applicable to the majority of barrier elements in the human genome. Perhaps other proteins are responsible for recruiting chromatin modifying enzymes such as histone H3K27 demethylase enzymes to these barriers to maintain K27 methylation states. There is precedence for this in other eukaryotic genomes as lysine-specific histone demethylase 1 (LSD1), which removes methyl groups from H3K4me1, H3K4me2, H3K9me1, and H3K9me2 (Shi *et al.*, 2004; Metzger *et al.*, 2005), has been implicated in the formation of boundaries between euchromatin and heterochromatin in *S. pombe* and *Drosophila* (Lan *et al.*, 2007 b; Rudolph *et al.*, 2007). In *S. pombe* the Lsd1/2 complex is recruited to boundary elements and limits the formation of heterochromatin perhaps by demethylating H3K9. In contrast *Drosophila* LDS1 homologue SU(VAR)3-3 is required for heterochromatin

formation by demethylating H3K4me1 and H3K4me2 and does not demethylate H3K9me1 or H3K9me2. This prevents the spread of H3K4 methylation into heterochromatin regions. In this study CTCF is associated with boundaries between active and inactive chromatin regions defined by the presence of H3K27me1 and H3K27me3 respectively. The recently identified H3K27 demethylases UTX and JMJD3 (Agger *et al.*, 2007), which are capable of demethylating H3K27me3, may be recruited to these chromatin boundaries, thus preventing the spread of H3K27me3 into active chromatin domains.

## Chapter 5

### Optimisation of conditions for ChIP-chip when using cell types that are limiting in number

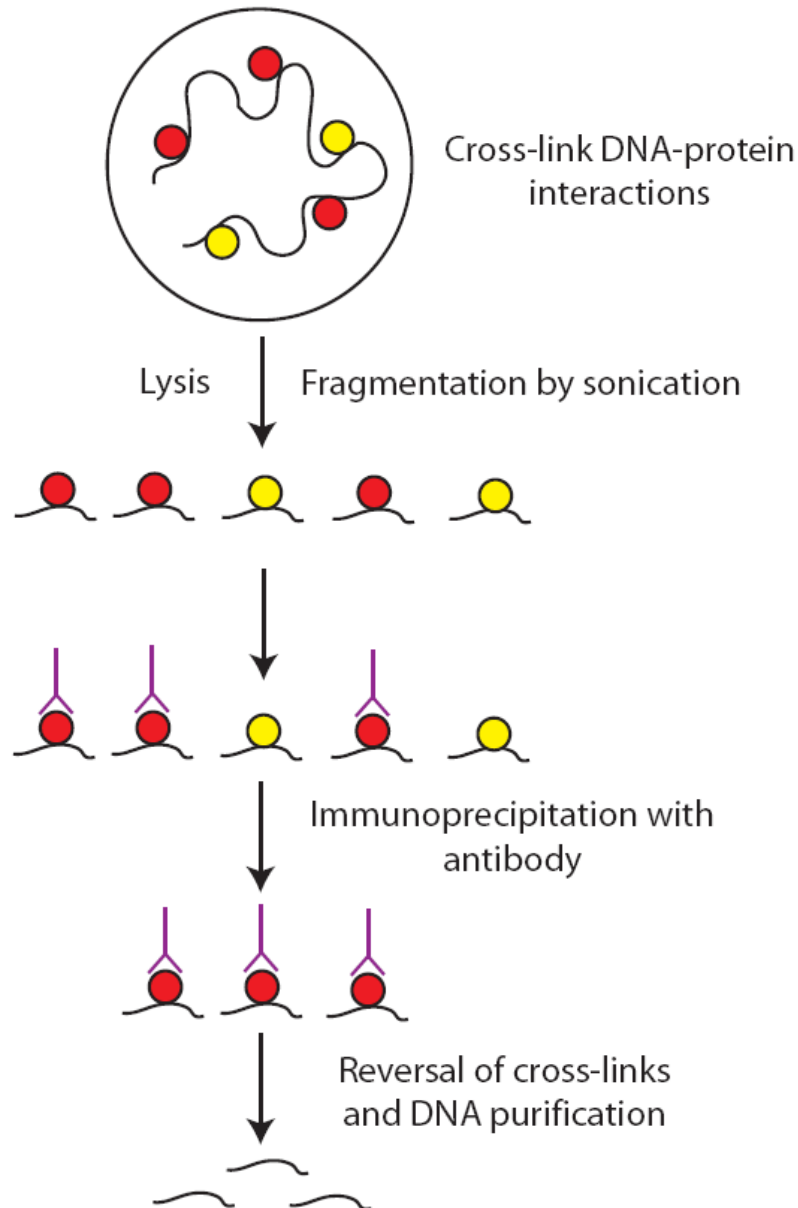
#### 5.1. Introduction

As discussed in Chapters 1 and 3, the recent development of microarray technology coupled with chromatin immunoprecipitation (ChIP-chip) has accelerated our understanding and annotation of DNA-protein interactions. ChIP-chip methods were initially developed in yeast (Ren *et al.*, 2000; Iyer *et al.*, 2001) and subsequently applied to study DNA-protein interactions in other organisms including humans. However, the development of ChIP itself has its origins long before ChIP-chip was developed. In the 1980's a protocol for investigating DNA-protein interactions in living cells was developed (Solomon and Varshavsky, 1985; Solomon *et al.*, 1988) and improved upon through the years to produce a ChIP method that is now widely used (Orlando, 2000). The fundamental steps performed in this ChIP procedure include cross-linking of DNA-protein interactions in living cells, extraction and shearing of chromatin, immuno-affinity 'pulldown' of chromatin bound by a protein of interest, and reversal of cross-links to isolate the interacting DNA (Figure 5.1). The isolated DNA can then be interrogated by microarray hybridisation. However, there are several technical parameters which need to be optimised to ensure the success of individual ChIP-chip experiments. These parameters are discussed below.

**Preparation of cross-linked chromatin:** Cross-linked chromatin can be prepared by the use of UV light or formaldehyde to covalently attach proteins to DNA sites *in vivo* (Orlando and Paro, 1993; Walter *et al.*, 1994; Biggin, 1999). The most commonly used cross-linking agent is formaldehyde, which was first used in the late 1970's to investigate *in vivo* DNA-protein interactions (Varshavsky *et al.*, 1979). It readily permeates the cell and nuclear membrane and generates a covalent cross-link between free amines on DNA, RNA or proteins, producing DNA-protein, RNA-protein, and protein-protein cross-links (Orlando *et al.*, 1997). Other cross-linking agents have been used in combination with formaldehyde to detect indirect associations between DNA and transcription co-factors,

including dimethyl apidimidate (DMA) (Kurdistani *et al.*, 2002), and disuccinimidyl glutarate (DSG) (Tian *et al.*, 2005). The amount of cross-linking reagent and incubation time can affect the accessibility of the antibody to the antigen epitope (Orlando, 2000), the fragmentation of the chromatin (Orlando *et al.*, 1997), and the efficiency by which weaker protein-DNA interactions are cross-linked. In general proteins with weaker DNA interactions require longer cross-linking times but it is important to optimize this as longer cross-linking times will increase the size of sonication fragments that can be obtained. Fragmentation of chromatin by sonication requires optimisation to ensure that the majority of the sonicated DNA is in the 300-1000bp range as larger DNA fragments are not efficiently immunoprecipitated (Orlando *et al.*, 1997). Chromatin can also be prepared without the use of cross-linking reagents and in this method the 'native' chromatin is fragmented through nuclease digestion (NChIP) (O'Neill and Turner, 2003). This method can be used to detect modified histones, but cannot be used to detect the majority of non-histone interactions such as transcription factor binding events. Fragmentation of 'native' chromatin by micrococcal nuclease digestion also requires careful optimisation as over-digestion can result in sub-nucleosomal particles, which are not as efficiently precipitated as oligo-nucleosomes (O'Neill and Turner, 1995).





**Figure 5.1 Principles of ChIP.** The main principles of the ChIP procedure are outlined in the figure, beginning with the cross-linking of DNA-protein interactions in living cells followed by lysis and sonication of DNA. The DNA associated with a protein of interest (red circle) is immunoprecipitated by a specific antibody, the DNA-protein cross-links are reversed and the DNA which interacted with the protein is isolated.

**Antibody specificity, affinity and epitope accessibility:** The specificity and affinity of an antibody is key to the success of any ChIP-chip experiment. While western blotting and competition assays with peptides containing the epitope of interest can highlight the

specificity of an antibody (Suka *et al.*, 2001), these techniques cannot predict which antibodies will bind to the protein of interest *in vivo*. An epitope on a target protein may be masked by formaldehyde cross-linking, meaning that a number of antibodies raised to different epitopes are usually tested in ChIP-chip assays to identify one which can efficiently immunoprecipitate the target protein and its bound DNA (Horak *et al.*, 2002). It is often important to use antibodies generated against C-terminus or the N-terminus epitopes, as these are often exposed on intact proteins and allow greater accessibility compared to epitopes buried within the protein. In addition, antibodies may cross-react with other proteins containing similar epitopes, which is often a problem when investigating members of the same protein family as they are often highly similar at the amino acid sequence level. Limitations in the number of highly specific antibodies for use in ChIP assays have led investigators to adopt alternative approaches for the study of DNA-protein interactions. An epitope tagging approach has been used successfully in yeast to map a number of DNA-protein interactions (Lee *et al.*, 2002). In this approach, the DNA sequence for a defined epitope such as a c-Myc, TAP, His, FLAG, HA, V5 or Pk tag is introduced to a locus whose interactions are to be mapped (Puig *et al.*, 2001). The tagged protein and its associated DNA binding sites can then be isolated using highly specific immuno-affinity approaches. This approach is more difficult in mammalian systems owing to the difficulty in introducing targeted genetic alterations and ensuring that the tagged transcription factor retains its normal DNA binding profile.

**Amplification and labeling of ChIP DNA:** The amount of DNA recovered from a ChIP assay is in a limiting quantity, typically in the range of 10 to 200 hundred nanograms. The limits of sensitivity of most array platforms require microgram amounts of target DNA to be labeled and hybridised to the array. Therefore a number of studies have combined multiple ChIP DNAs (Weinmann *et al.*, 2002) or used DNA amplification to obtain enough DNA for microarray hybridisation. Three amplification methods have been reported: (i) ligation-mediated PCR (Ren *et al.*, 2000) in which a double stranded oligo is ligated to the end of ChIP DNAs, which acts as a template for PCR amplification, (ii) random amplification in which a degenerate oligo sequence randomly anneals to DNA allowing the DNA to be amplified by PCR (Iyer *et al.*, 2001), and (iii) T7-based linear amplification, in which poly dTs are added to the end of DNA by terminal

transferase, followed by a round of DNA synthesis using polyA-T7 sequences as primers (Liu *et al.*, 2003; Bernstein *et al.*, 2005). The synthesized DNA is then transcribed into RNA by T7 RNA polymerase. However, these methods can introduce sequence and length dependent biases during the amplification process, which can directly affect the ChIP-chip data. Following amplification, the DNA is fluorescently labeled by random priming using the Klenow fragment of *E. coli* DNA polymerase. Labeling by random priming can result in an additional amplification of 10-100 fold due to the strand displacement activity of the exonuclease deficient Klenow fragment (Walker *et al.*, 1992). Klenow-mediated labeling of ChIP samples, without prior amplification of ChIP DNA samples, can be used to increase the amount of material in a non-biased way for ChIP-chip experiments (Koch *et al.*, 2007, De Gobbi *et al.*, 2007; Dhami, submitted).

**Array platforms:** A number of array platforms are available for ChIP-chip experiments, each of which has its own advantages and disadvantages. An array platform is judged on a number of features, which include (i) the density or number of features, (ii) the resolution of the array features, (iii) the percentage of genome sequence covered by the array and (iv) the sensitivity and reproducibility at reporting a wide dynamic range of enrichments in ChIP samples. Tiling path arrays which contain probes spanning large continuous regions of the human genome are commonly used for ChIP-chip studies and can be composed of oligonucleotide or PCR product probes (Bernstein *et al.*, 2005; Koch *et al.*, 2007). Microarrays containing oligonucleotide probes are used exclusively for very large scale studies as a much greater number of array features can be accommodated and the resolution is much higher. In addition, oligonucleotide probes can be synthesized directly onto the slide and avoid the necessity of PCR amplification. However, the use of PCR products as microarray probes may result in higher signal:noise ratios due to the length of the probe available for hybridisation. Therefore this type of array platform is often more sensitive and reliable in reporting enrichments compared to array platforms which use much smaller oligonucleotide probes (D. Vetrie, unpublished data).

**Analysis of data:** The analysis of 'raw' ChIP-chip data sets is a crucial step in identifying *bona fide* sites of DNA-protein interaction. Often the first step in the analysis procedure is normalisation. Normalisation is an important process in which any systematic and non-systematic biases in a ChIP-chip dataset can be accounted for and

adjusted accordingly to ensure that only true ChIP enrichments are reported in the dataset. Biases in datasets which can be accounted for by normalisation include low signal, high background in one channel (when performing two-channel hybridisation experiments), uneven signal in the two channels, and non-specific binding of antibodies. Standard array quantitation programs can be used to perform LOWESS normalization which allows for biases in ratio calculations to be accounted for in a signal-dependent manner (Quackenbush, 2002). This is a form of local normalization in which all data-points are not treated equally. On the other hand, total scaling normalization is a form of global normalization in which all datapoints are treated equally. A scaling factor is calculated to ensure that the median ratio of the data set is 1. This normalization method is usually used due to uneven signal intensity in the two microarray channels. Non-specific binding normalization can be used to correct for variations in non-specific binding as measured by a species-specific IgG control experiment (as discussed and performed in Chapters 3 and 4 of this thesis). This form of local normalization does not treat all data points equally; instead a normalized ratio is calculated by dividing the experimental ratio at individual data-points by the corresponding IgG ratio.

Once data sets have been appropriately normalized, a number of approaches can be used to determine which regions correspond to the binding of a protein of interest. A fold enrichment cut-off or a standard deviation cut-off can be used to define significant enrichments. However these methods make a number of assumptions regarding the dataset, namely that there is a normal Gaussian distribution of a dataset around the mean and that both negative and positive regions are present in the data set in similar proportions. These assumptions may not apply to all data sets and could result in false positive regions being identified. Several other more complex statistical analysis methods have been developed to analyse the likelihood of binding, which include single-array error model (Ren *et al.*, 2000), rank-based target identification (Iyer *et al.*, 2001) and linear regression models (Buck *et al.*, 2005; Gibbons *et al.*, 2005; Li *et al.*, 2005). The single-array error model also assumes a normal distribution of the data and calculates a one-sided probability to determine which peaks of enrichments are significant, while rank-based identification does make any assumptions about the distribution of ChIP-chip data, instead it requires multiple replicates to determine whether an enrichment is

significant or not. Linear regression methods, such as ChIPOTle (discussed in Chapter 3), utilise two properties of ChIP-chip data to identify *bona fide* interactions. Namely, that enrichment of DNA fragments adjacent to sites of direct interaction is observed. Also the cluster of enriched elements surrounding a peak forms a distinct shape, with tailing off left and right as you move away from the site of direct interaction. Hidden Markov Models can also be used to identify significant peaks of enrichment in ChIP-chip data by portioning the data into locations that are either consistent or inconsistent with antibody binding and identifies protein interaction sites and centres with their probabilities (Koch *et al.*, 2007). While these methods are useful for analysing discrete interactions, they are often unreliable when it comes to identifying low-level enrichments spread over large distances such as histone modifications. Furthermore, hierarchical clustering (Eisen *et al.*, 1998) may be used to identify continuous features.

**Number of cells required to perform a ChIP reaction:** While a number of the technical considerations discussed above have been overcome in recent years with the rapid development of microarray, antibody, and analysis technology, there is still one major caveat when performing ChIP experiments in combination with microarrays. This is the requirement for large numbers of cells, typically  $10^7$  per ChIP-chip assay. While this is usually not a problem when using cultured cell lines, it can be difficult to obtain enough primary cells for ChIP-chip experiments. This is a particular problem in higher organisms, such as mammals, whose specialised cell types are often rare or difficult to access for experimentation.

There are numerous examples of biological questions which require ChIP-chip based approaches to help further our knowledge of fundamental issues relating to development and disease but which are also limited by the small amounts of tissue available for study. For example the epigenetic analysis of primary cells in many developmental systems and cancer stem cells (Reya *et al.*, 2001) are examples where cell numbers are limiting as typically 3-4 orders of magnitude fewer cells are obtained than currently used in ChIP-chip protocols.

Recently, O'Neill and colleagues described a protocol which they called carrier ChIP (CChIP), in which 'native' or un-crosslinked chromatin immunoprecipitation (NChIP) procedures were used in combination with *Drosophila* 'carrier' chromatin to investigate

histone modifications in small populations of cells (O'Neill *et al.*, 2006). Cultured mouse ES cells were used as a model system to test this new protocol and it was shown that CChIP could be used to analyse the distribution of histone modifications from 1,000 cells. The protocol was shown to be reproducible and was then applied to study histone modifications in samples containing 50-100 cells obtained from the inner cell mass and trophectoderm of mouse embryos. Fold enrichments were determined using radioactive PCR and phosphorimaging and the modest enrichments observed suggest that this procedure may not be applicable to microarray analysis. The presence of carrier chromatin which makes up the vast bulk of the DNA in the ChIP sample will invariably affect labeling efficiency of the material being assayed which makes up a very small proportion of the DNA in the sample. In addition the authors suggest that this CChIP protocol may not be applicable to formaldehyde cross-linked chromatin (XChIP), in which yields tends to be low when immunoprecipitating with antibodies raised to modified histones. This may be due to cross-linking of lysine-rich histone tails to nearby DNA, obscuring antibody access to its epitope (Robyr and Grunstein, 2003; Suka *et al.*, 2001). Less than 1% of bound and unbound fractions are typically recovered in XChIP (compared to 10-20% yields in NChIP), which O'Neill and colleagues suggest may prevent the use of cross-linked chromatin in the analysis of histone modifications in very small cell populations.

More recently two XChIP-based methods have been developed to study histone modifications in limited populations of cells, known as Q<sup>2</sup>ChIP and miniChIP (Dahl and Collas, 2007; Attema *et al.*, 2007). MiniChIP was used to analyse histone modifications in 50,000 cells using quantitative PCR while the Q<sup>2</sup>ChIP method in combination with quantitative PCR showed that it was possible to analyse histone modifications in as few as 100 cells.

Thus, developing a robust ChIP-chip method for use with much fewer cells would allow us to understand global epigenetic events which may be specific to small populations of cells, which would be not be determined when studying mixed populations of cells. In addition, developing a ChIP protocol for use with fewer cells may have cost implications too. It would cost less to do whole-genome studies if lower numbers of cells and antibody

amounts could be used to generate the same data sets as experiments performed with more conventional cell numbers and antibody amounts.

The development of a highly sensitive PCR product-based array platform (Dhami *et al.*, 2005) precludes the need for amplification of ChIP DNA prior to array hybridization (Koch *et al.*, 2007). While this platform is sensitive enough to detect histone and transcription factor interactions from unamplified ChIP DNA derived from assays performed with  $10^7$  cells, it was not clear if this platform could be used to faithfully detect interactions from fewer cells. As the SCL locus tiling path array had been used to characterize in detail a number of histone modifications and transcription factor interactions from ChIP assays performed with  $10^7$  cells (Dhami, PhD Thesis, University of Cambridge, 2005) (Chaper3) it represented an ideal model system in which to develop a ChIP-chip protocol for the detection of DNA-protein interactions from reduced numbers of cells. The success of a low cell numbers ChIP-chip protocol could therefore be determined by comparisons with known profiles of ChIP enrichments across the SCL locus.

## **5.2. Aims of this Chapter**

The aim of this chapter was to further develop existing ChIP-chip approaches in order to improve the sensitivity of the method when using cells which are limiting in number. This method will then be exploited to perform analyses of regulatory interactions in one such cell type in Chapter 6 of this thesis. Therefore, the aims of the work presented in this Chapter were:

1. To determine parameters important for developing low cell number ChIP-chip.
2. To determine the types of DNA interactions (histone modifications and transcription factors) that can be identified with these protocol developments.
3. To evaluate the sensitivity and reproducibility of the method at detecting known ChIP-chip profiles at the SCL locus. Quantitative PCR would be used to determine sensitivity while independent biological replicates would be used to confirm the reproducibility of the method.

### **5.3. Overall strategy**

As discussed in Chapters 3 and 4, the human SCL locus has been extensively characterized for a number of histone modification and transcription factor interactions in the K562 cell line (Dhami, submitted). These results could be used to assess the limit of detection of known regulatory elements using low cell number ChIP-chip. The SCL tiling path array would then be used to develop a ChIP-chip protocol for the detection of histone acetyl and methyl modification patterns from a range of cell numbers. Various parameters were tested to optimize the ChIP protocol for use with low cell numbers and included antibody-chromatin ratio, amount of protein G agarose used to immunoprecipitate DNA-protein-antibody complexes, and the volume in which immunoprecipitations were performed. Statistical analysis would be used to identify the antibody amount that gave the optimal signal:noise ratio based on the analysis of known regulatory elements across the SCL locus. Given that signal:noise ratios may be compromised when performing experiments with low cell numbers, normalization methods would be investigated which improved signal:noise. The reproducibility of the low cell numbers protocol would be tested by performing replicates. Furthermore, histone modification enrichments reported by the SCL array in low cell number assays would be verified by SYBR green real-time PCR. Finally, the detection of transcription factor interactions in reduced cell numbers would also be investigated.

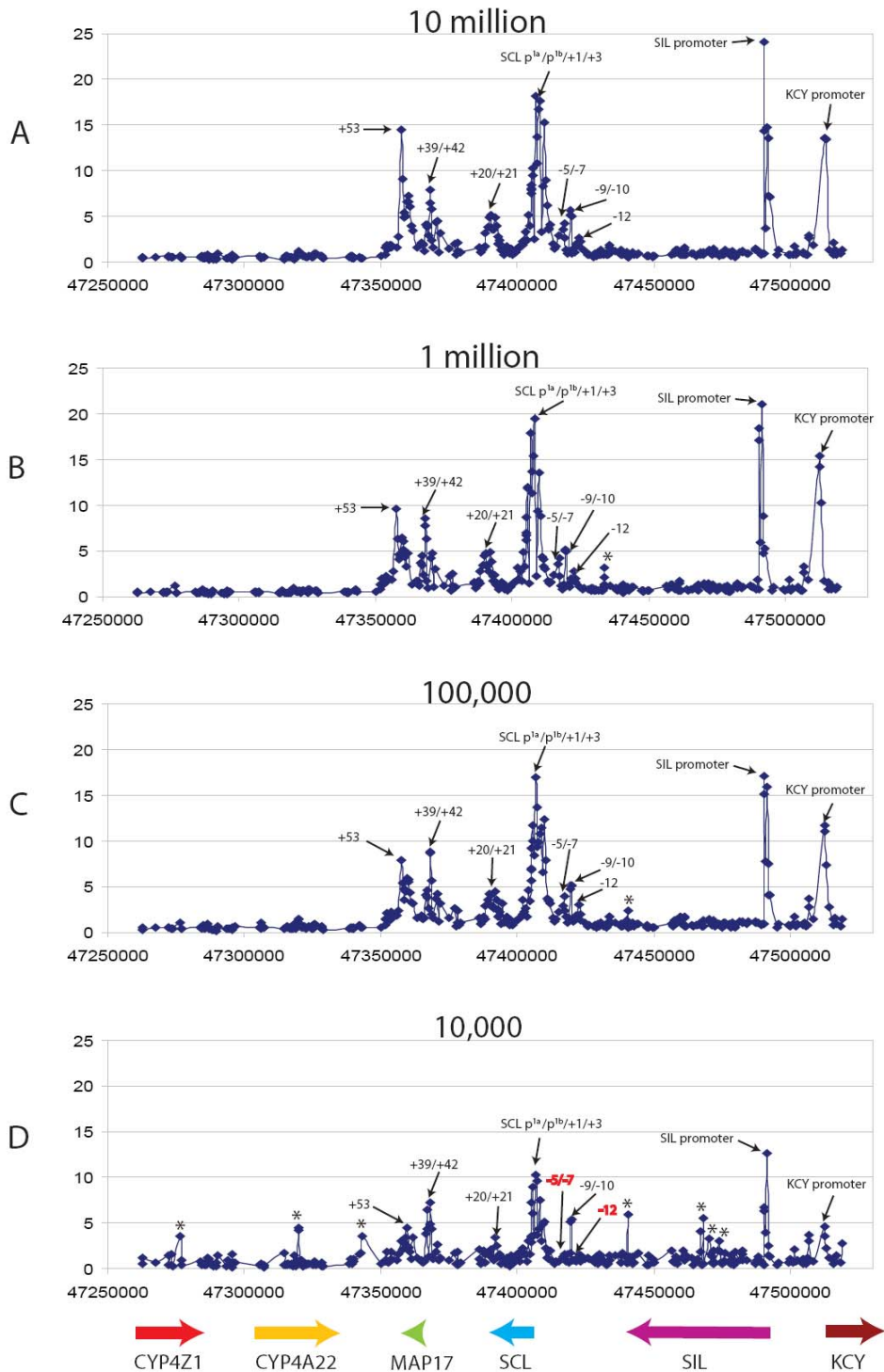
## **Results**

### **5.4. The effect of titrating cell numbers in ChIP-chip and its effect on the detection of regulatory elements at the human SCL tiling locus**

The SCL array had been previously shown to be a highly reproducible platform for the detection of known regulatory elements associated with histone H3 K9/K14 diacetylation (also referred to as histone H3 acetylation) in K562 cells (Dhami, submitted). Histone H3 acetylation is associated with active promoters and enhancers (Chapter 3). Initial experiments aimed at develop a ChIP-chip protocol for lower cell numbers focused on establishing the reproducible detection of known SCL regulatory elements in histone H3 acetylation ChIP assays whilst titrating cell numbers used.



Chromatin was prepared from  $10^8$  K562 cells and aliquots equivalent to  $10^6$ ,  $10^5$ , and  $10^4$  cells were used to perform H3 K9/K14 acetylation ChIP assays as described in Chapter 2. All other conditions were the same as for ChIP assays performed with  $10^7$  cells. The only variable in these experiments was the amount of chromatin (i.e., cell number equivalents) used in each assay. Significant enrichments in ChIP-chip experiments were obtained by determining the mean ratios of two “background” regions represented on the SCL array which did not contain any known regulatory elements in K562 (Dhami, PhD Thesis, University of Cambridge, 2005). These regions spanned 47262287 bp to 47343557 bp, and 47424426 bp to 47489321 bp on chromosome 1. The significant enrichment threshold was three standard deviations above the mean background ratios. All known SCL regulatory regions associated with significant histone H3 acetylation in K562 cells were detected in three of the four assays ( $10^7$ ,  $10^6$  and  $10^5$  cells) with varying degrees of enrichment (Table 5.1 and Figure 5.2). Some of these enrichments were located at or near the SCL, SIL and KCY promoters as these genes are expressed in K562.



**Figure 5.2: ChIP-chip profiles of H4 K9/K14 diacetylation across the human SCL locus in K562 for a range of cell numbers.** Known regulatory elements showing significant enrichment for H3 K9/K14 diacetylation are indicated with black arrows. The nomenclature of regions denoted by +1, +3 etc. are based on their distance upstream (-) or downstream (+) in kilobases from the SCL promoter p<sup>1a</sup>. The chromosome 1 coordinates are shown along the x-axis for each panel and the y-axis indicates the fold-enrichments. The

thick coloured arrows at the bottom of the figure represent the gene order and direction of transcription. Panel A represents data obtained from the ChIP-chip experiment conducted with 10 million cells and 10 µg of antibody. Panel B represents the 1 million cells experiment, panel C the 100,000 cells experiment, and finally panel D represents the data obtained from the experiment conducted with 10,000 cells. In panels B-D, significant enrichments that are not associated with regulatory elements are indicated with an asterisks. In panel D, the -5/-7 and -12 regulatory elements are not detected as significant peaks and are indicated in red lettering.

Histone acetylation at the 5' end of the SCL gene extended 8 kb into the coding region and is found at regions such as the +1 and +3 regions which show DNase I hypersensitivity (Leroy-Viard *et al.*, 1994; Follows *et al.*, 2006). Enrichments were also found at the stem cell enhancer (+20/+21) (Gottgens *et al.*, 2002), the MAP17 promoter (+42), the -9/-10 region (Gottgens *et al.*, 1997) and the SCL erythroid enhancer (from +50 to +53) (Delabesse *et al.*, 2005). In addition, significant enrichments were also observed at the -12, -5/-7 and +39 regions as described previously (Dhami, PhD Thesis, University of Cambridge, 2005). No significant enrichments were observed across the genomic region containing the CYP4Z1 and CYP4A22 genes (which are not expressed in K562) in assays performed with  $10^7$ ,  $10^6$  and  $10^5$  cells. However, both the  $10^6$  and  $10^5$  cell assays contained one significant peak of enrichment not associated with any known regulatory element.

Overall, the fold enrichment ratios reported for the  $10^4$  cell experiment were lower than the other three experiments. Whilst seven SCL regulatory regions were detected, two known elements (-5/-7 and -12) were not significantly enriched in this assay. However, unexpectedly, two significant peaks of enrichment were observed within the CYP4Z1 and CYP4A22 genes in an assay performed with  $10^4$  cells and seven other peaks of significant enrichment were also detected in this assay, which were not associated with known regulatory elements (indicated by asterisks in Figure 5.2).

Number of cells	No. of known regulatory Elements detected	Known regulatory elements detected	No. of elements not associated with known regulatory function	No. of known regulatory elements not detected	Mean CV of ratios (%)
10 <sup>7</sup>	9/9	KCY p, SILp, -12, -9/-10, -5/-7, SCLp, +20/+21, +39/+42, +53	0	0	6.45
10 <sup>6</sup>	9/9	KCY p, SILp, -12, -9/-10, -5/-7, SCLp, +20/+21, +39/+42, +53	1	0	8.01
10 <sup>5</sup>	9/9	KCY p, SILp, -12, -9/-10, -5/-7, SCLp, +20/+21, +39/+42, +53	1	0	8.85
10 <sup>4</sup>	7/9	KCY p, SIL p, -9/-10, SCLp, +20/+21, +39/+42, +53	7	2	22.42

**Table 5.1: A summary of the performance of H3 K9/K14 diacetylation ChIP-chip assays performed with a standard range of cell numbers and 10 µg of antibody.** The SCL tiling array was used to detect regulatory elements associated with H3 K9/K14 diacetylation in assays performed with 10<sup>7</sup>, 10<sup>6</sup>, 10<sup>5</sup>, and 10<sup>4</sup> cells and 10ug of antibody. Assays performed with 10<sup>7</sup>, 10<sup>6</sup>, and 10<sup>5</sup> cells were associated with a low mean CV of ratios and all nine known regulatory elements were detected. Seven out of nine known regulatory elements were detected when this assay was performed with 10<sup>4</sup> cells and a high mean CV of ratios was calculated. The number of significantly enriched elements that were detected and not previously associated with known regulatory function was one for assays performed with 10<sup>6</sup> and 10<sup>5</sup> cells, while this increased to seven for the assay performed with 10<sup>4</sup> cells. Note KCYp , SILp and SCLp refer to the promoters of these genes.

The mean ratios and CVs of array elements spotted in triplicate were also calculated to determine if array elements were reporting reproducible ratios within individual assays. The reproducibility of the reported ratios between triplicate spots was high for assays performed with 10<sup>5</sup> to 10<sup>7</sup> cells as the mean CV of ratios were calculated to be between 6-8% (Table 5.1). These values were very similar to those previously described (Dhami 2005 thesis). However there was a significant increase in the mean CV of ratios for the experiment performed with 10<sup>4</sup> cells (22.4%).

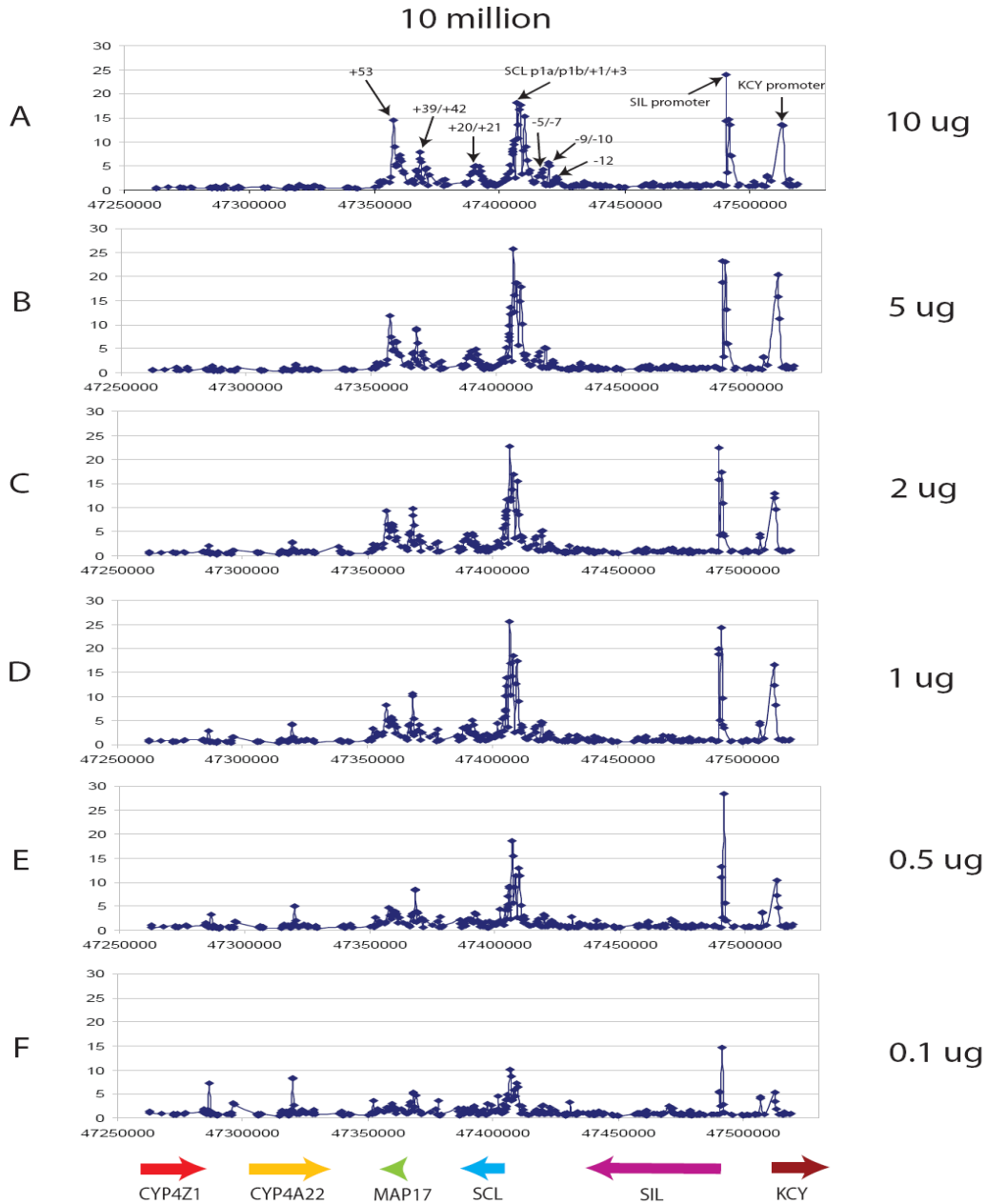
Thus, it appears that as cell numbers are decreased in ChIP-chip assays, there was a concomitant increase in the detection of non-specific interactions, reduction in the detection of known regulatory sequences and decreased reproducibility in the ratios obtained. Factors which may attribute to these observations include (i) limitations in the amount of recovered DNA available for labeling and hybridization resulting in lower signal:noise ratio on the array and (ii) an excess of antibody during the immunoprecipitation step resulted in enrichment of non-specific interactions. Therefore while a ChIP-chip assay performed with  $10^4$  cells could detect the majority of known regulatory elements associated with H3 acetylation at the SCL locus, it was necessary to further optimize the procedure to ensure that the assays were as equally robust as conventional ChIP assays performed with  $10^7$  cells.

## **5.5. The effect of titrating antibody levels on low cell number ChIP**

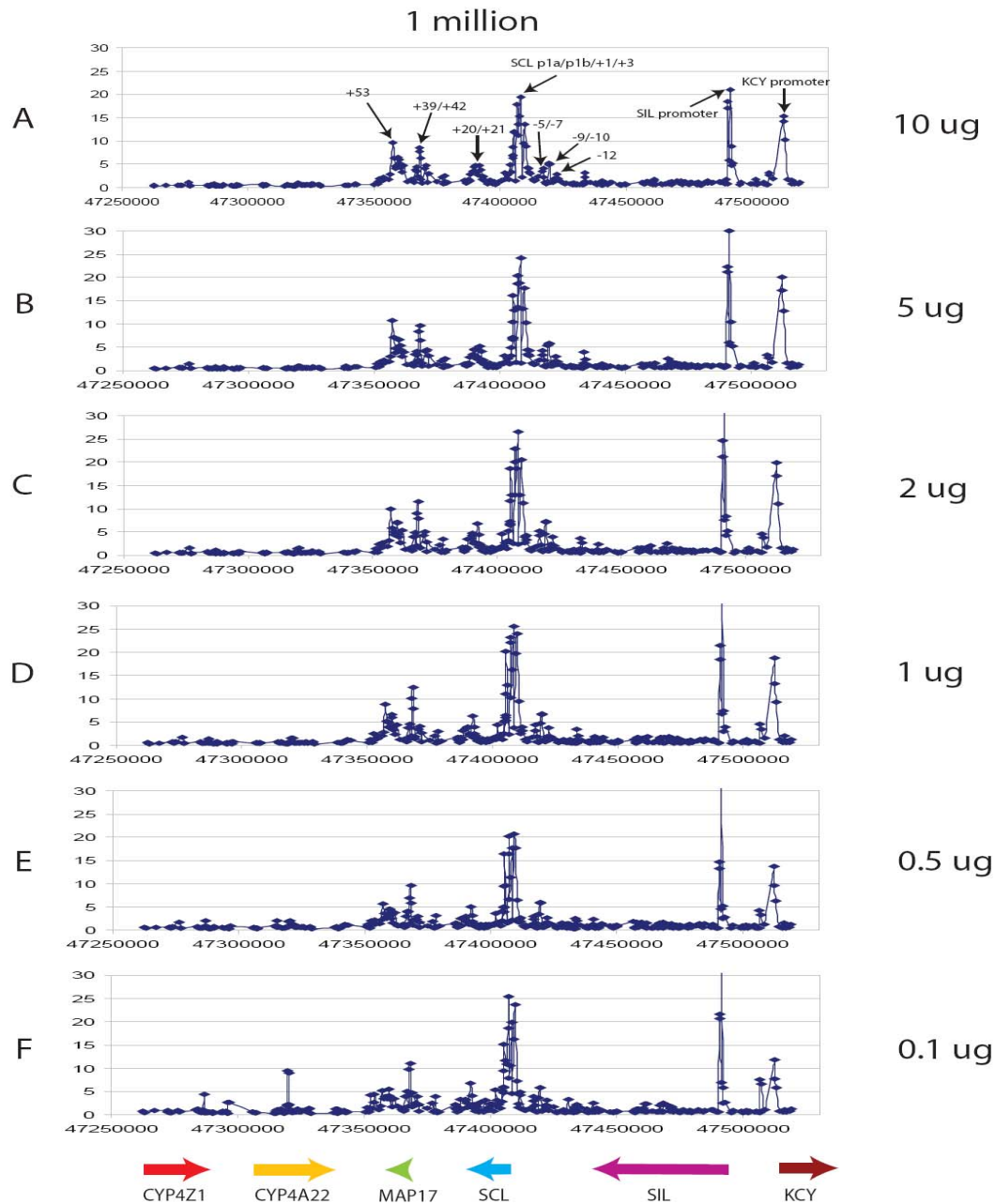
### **5.5.1. Histone H3 K9/K14 acetylation experiments**

As mentioned above, an excess of H3 K9/K14 acetylation antibody in ChIP-chip assays using the chromatin equivalent from  $10^4$  cells may be responsible for the enrichment of DNA sequences not previously identified in assays performed with  $10^7$  cells. Therefore amount of antibody used in ChIP was titrated in series of experiments to identify the optimal antibody:chromatin ratio for a range of cell numbers. In addition, these antibody titration experiments was also performed for an antibody raised to H3K4me3, to ensure that results obtained with the H3 acetylation assays could also be achieved with other ChIP-chip assays.

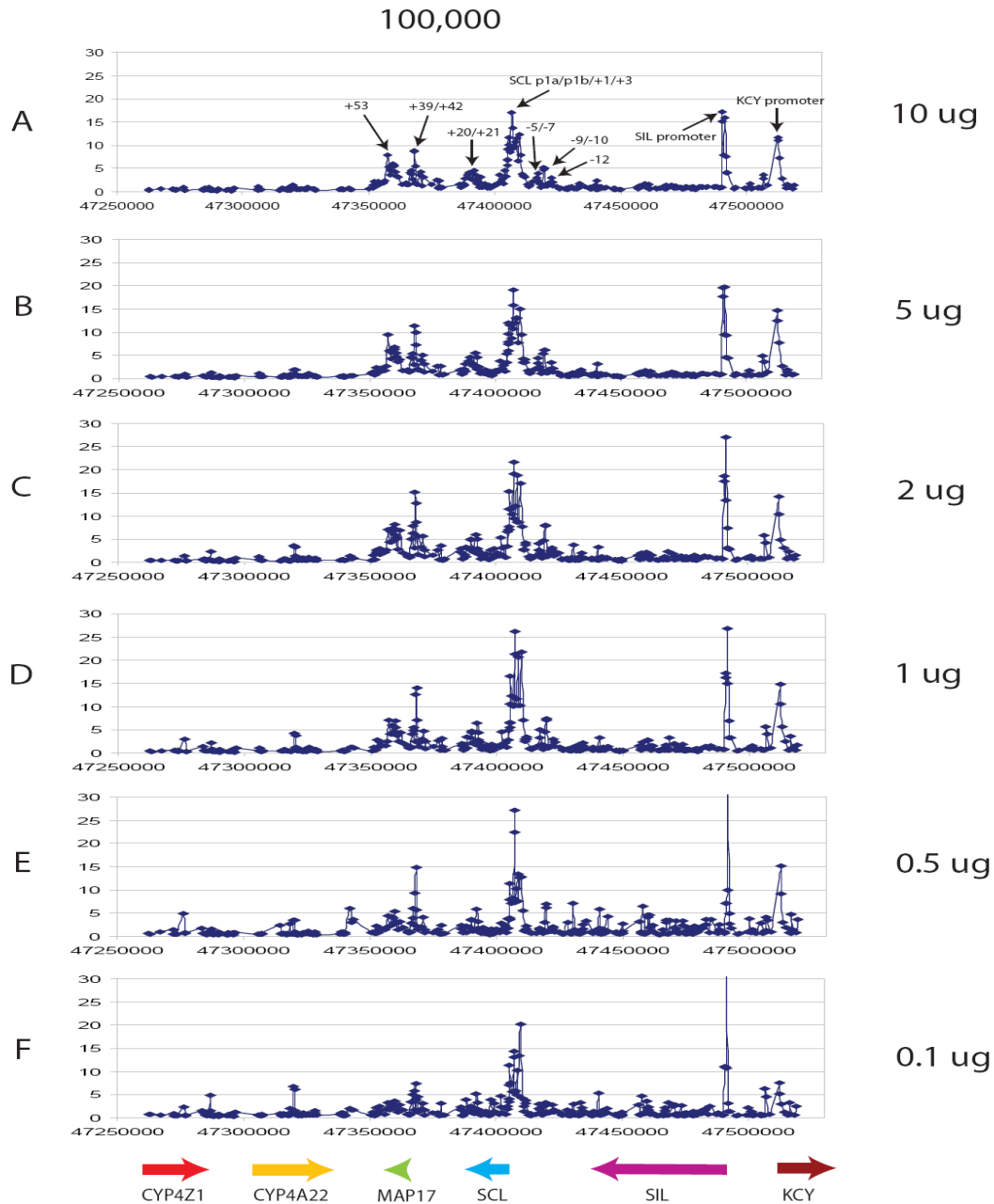
Once again, ChIP experiments were performed using a panel of conditions with varying amounts of chromatin from K562 cells ( $10^7$ ,  $10^6$ ,  $10^5$ , and  $10^4$  cell equivalents). However, in these experiments, the amount of antibody was varied (10  $\mu\text{g}$ , 5  $\mu\text{g}$ , 2  $\mu\text{g}$ , 1  $\mu\text{g}$ , 0.5  $\mu\text{g}$ , and 0.1  $\mu\text{g}$ ) across each constituent of the chromatin panel. In otherwords, for a given chromatin amount, 6 different amounts of antibody were tested in ChIP. This resulted in a set of results across a series of 24 conditions (4 chromatin amounts versus 6 antibody amounts). Figures 5.3 to 5.6 show the results of these ChIP-chip experiments for the detection of regulatory features across the SCL locus.



**Figure 5.3: Titration of H3K9/K14 diacetylation antibody in ChIP assays performed with  $10^7$  K562 cells.** The SCL ChIP-chip profiles for experiments performed with  $10^7$  K562 cells and the six antibody concentrations are shown. The known regulatory elements are indicated by black arrows in panel A. Variations in fold-enrichments are observed for regulatory elements across the six antibody concentrations in panels A-F. The detection of enriched regions not associated with known regulatory elements also varied with antibody concentration. Human chromosome 1 coordinates are indicated on the x-axis, with fold enrichments indicated on the y-axis. Each profile is scaled to 30 on the y-axis to allow for comparisons between titration series to be performed. The antibody amount is indicated to the right of each panel. The gene order and direction of transcription is indicated by thick coloured arrows at the bottom of the figure.

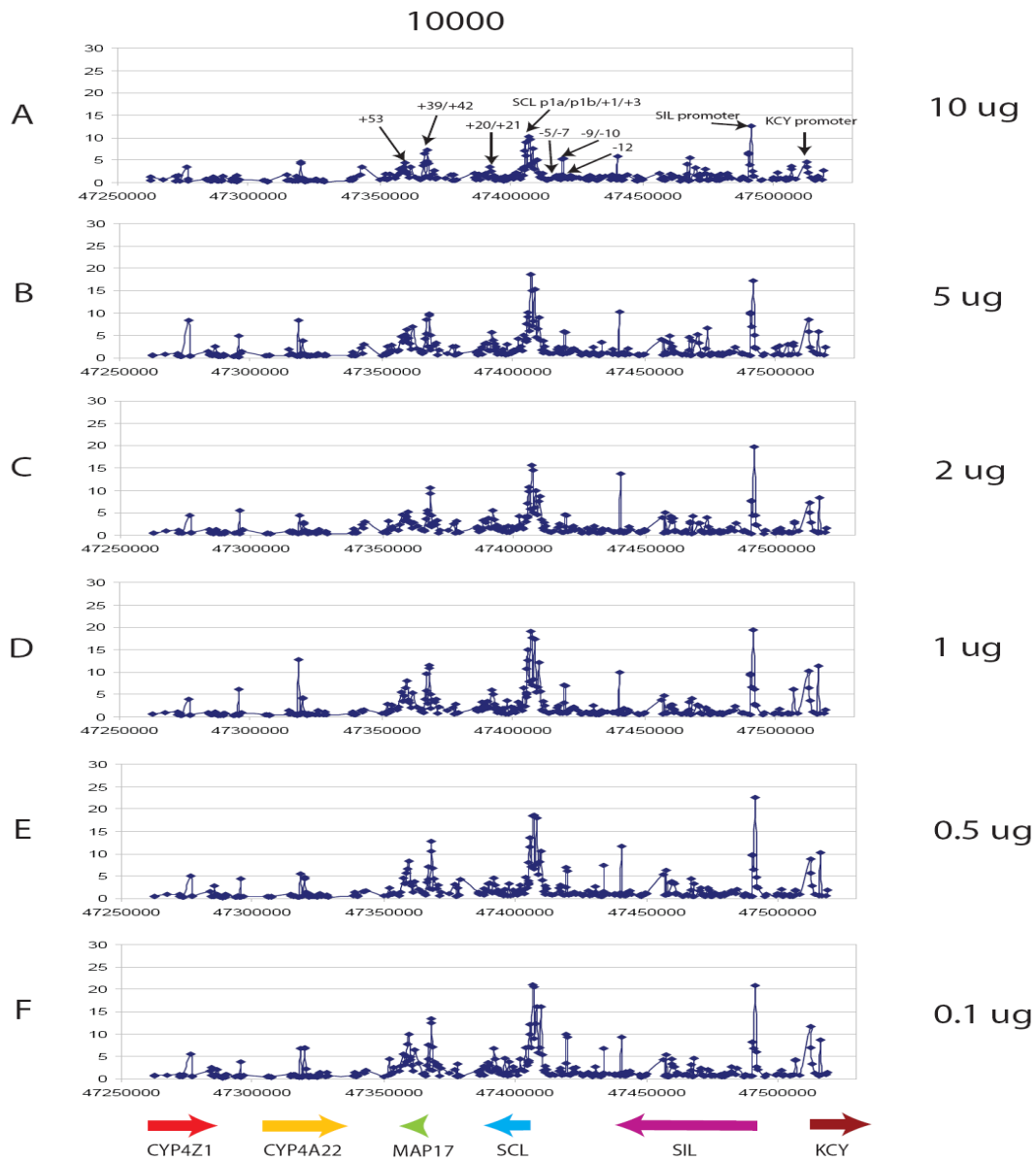


**Figure 5.4: Titration of H3K9/K14 diacetylation antibody in ChIP assays performed with  $10^6$  K562 cells.** The SCL ChIP-chip profiles for experiments performed with  $10^6$  K562 cells and the six antibody concentrations are shown. The known regulatory elements are indicated by black arrows in panel A. Variations in fold-enrichments are observed for regulatory elements across the six antibody concentrations in panels A-F. The detection of enriched regions not associated with known regulatory elements also varied with antibody concentration. Human chromosome 1 coordinates are indicated on the x-axis, with fold enrichments indicated on the y-axis. Each profile is scaled to 30 on the y-axis to allow for comparisons between titration series to be performed. The antibody amount is indicated to the right of each panel. The gene order and direction of transcription is indicated by thick coloured arrows at the bottom of the figure.



**Figure 5.5: Titration of H3K9/K14 diacetylation antibody in ChIP assays performed with  $10^5$  K562 cells.** The SCL ChIP-chip profiles for experiments performed with  $10^5$  K562 cells and the six antibody concentrations are shown. The known regulatory elements are indicated by black arrows in panel A. Variations in fold-enrichments are observed for regulatory elements across the six antibody concentrations in panels A-F. The detection of enriched regions not associated with known regulatory elements also varied with antibody concentration. Human chromosome 1 coordinates are indicated on the x-axis, with fold enrichments indicated on the y-axis. Each profile is scaled to 30 on the y-axis to allow for comparisons between titration series to be performed. The antibody amount is indicated to the right of each panel. The gene order and direction of transcription is indicated by thick coloured arrows at the bottom of the figure.





**Figure 5.6: Titration of H3K9/K14 diacetylation antibody in ChIP assays performed with  $10^4$  K562 cells.** The SCL ChIP-chip profiles for experiments performed with  $10^4$  K562 cells and the six antibody concentrations are shown. The known regulatory elements are indicated by black arrows in panel A. Variations in fold-enrichments are observed for regulatory elements across the six antibody concentrations in panels A-F. The detection of enriched regions not associated with known regulatory elements also varied with antibody concentration. Human chromosome 1 coordinates are indicated on the x-axis, with fold enrichments indicated on the y-axis. Each profile is scaled to 30 on the y-axis to allow for comparisons between titration series to be performed. The antibody amount is indicated to the right of each panel. The gene order and direction of transcription is indicated by thick coloured arrows at the bottom of the figure.

To rigorously assess the results of these 24 experiments, the data were assessed as follows:

1. An optimal signal:noise ratio was determined for the titration series at each cell number. To do this, the data for 60 array elements on the SCL array which encompassed nine known regulatory elements detected in K562 cells were extracted and the average fold enrichment “signal” was calculated. The standard deviation of “background” regions represented on the SCL array (as defined in section 5.4) was determined. This represented a numerical value to describe ‘noise’ which would incorporate information obtained from non-specific enrichments. The average fold enrichment value was then divided by the background standard deviation value to obtain a “signal:noise” ratio (Table 5.2). This ratio would effectively determine how effective an experiment was at detecting *bona fide* regulatory elements across the SCL locus.
2. The total number of known regulatory regions at the SCL locus associated with significant enrichment by ChIP-chip was also assessed (Table 5.2). The significant enrichment threshold was set at three standard deviations above the mean background ratios as described previously.

Using these criteria, assays conducted with  $10^7$ ,  $10^6$  or  $10^5$  cells, showed the highest signal:noise ratios and detection of all known regulatory elements when using 5-10  $\mu\text{g}$  of histone H3 K9/K14 antibody. However, as cell numbers were decreased to  $10^4$ , the highest signal:noise was obtained when 0.1  $\mu\text{g}$  of antibody was used; however, even at this antibody level, the signal:noise in these assays were more than 4-fold lower than those reported in assays with higher cell numbers, and only 7 of 9 known regulatory elements could be detected. These results demonstrate that antibody amount is crucial for increasing signal:noise in low cell number ChIP-chip experiments. However, it was still not possible to achieve the same sensitivity in low cell number assays, as that of conventional assays, for detecting all known regulatory elements across the SCL locus under the experimental conditions performed here.

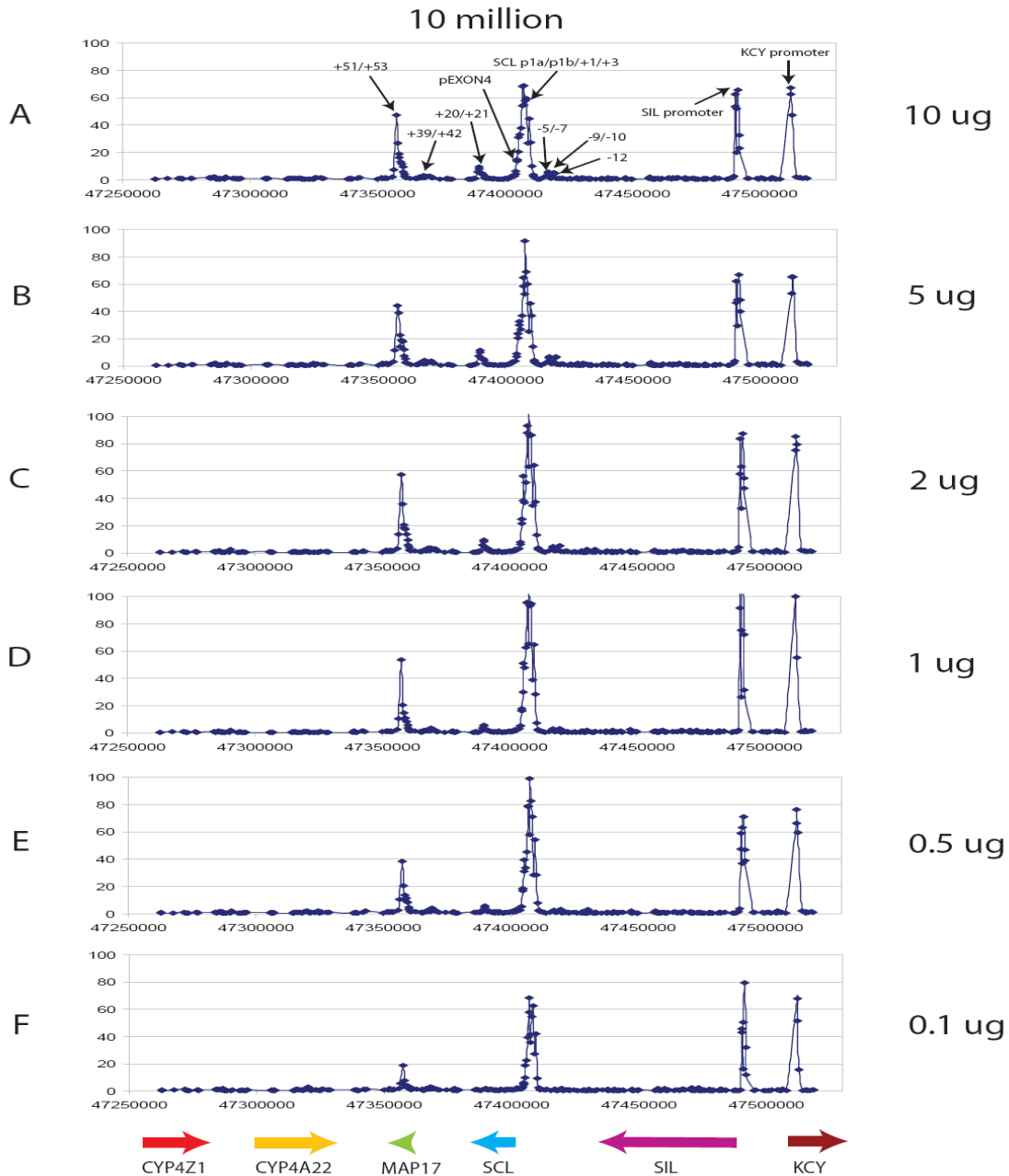
Cell number	Antibody amount (µg)	Average fold enrichments (Signal)	Standard deviation of background regions (noise)	Signal:noise ratio	Significantly enriched known regulatory elements
10 <sup>7</sup>	10	7.82	0.28	27.08	KCYp, SILp, -12, -9/-10, -5/-7, SCLp, +20/+21, +39/+42, +53
	5	8.44	0.29	28.15	KCYp, SILp, -12, -9/-10, -5/-7, SCLp, +20/+21, +39/+42, +53
	2	7.35	0.36	20.26	KCYp, SILp, -12, -9/-10, -5/-7, SCLp, +20/+21, +39/+42, +53
	1	7.60	0.47	15.87	KCYp, SILp, -9/-10, -5/-7, SCLp, +20/+21, +39/+42, +53
	0.5	5.42	0.57	9.46	KCYp, SILp, -9/-10, SCLp, +20/+21, +39/+42, +53
	0.1	3.08	0.96	3.19	KCYp, SILp, SCLp, +39/+42,
10 <sup>6</sup>	10	7.40	0.35	21.08	KCYp, SILp, -12, -9/-10, -5/-7, SCLp, +20/+21, +39/+42, +53
	5	8.55	0.40	20.91	KCYp, SILp, -12, -9/-10, -5/-7, SCLp, +20/+21, +39/+42, +53
	2	8.98	0.43	20.50	KCYp, SILp, -12, -9/-10, -5/-7, SCLp, +20/+21, +39/+42, +53
	1	8.66	0.41	20.70	KCYp, SILp, -12, -9/-10, -5/-7, SCLp, +20/+21, +39/+42, +53
	0.5	6.59	0.41	15.77	KCYp, SILp, -12, -9/-10, -5/-7, SCLp, +20/+21, +39/+42, +53
	0.1	7.45	1.00	7.42	KCYp, SILp, -9/-10, SCLp, +20/+21, +39/+42, +53
10 <sup>5</sup>	10	6.81	0.33	20.32	KCYp, SILp, -12, -9/-10, -5/-7, SCLp, +20/+21, +39/+42, +53
	5	7.81	0.36	21.22	KCYp, SILp, -12, -9/-10, -5/-7, SCLp, +20/+21, +39/+42, +53
	2	8.21	0.57	14.40	KCYp, SILp, -12, -9/-10, -5/-7, SCLp, +20/+21, +39/+42, +53
	1	8.16	0.62	13.05	KCYp, SILp, -12, -9/-10, -5/-7, SCLp, +20/+21, +39/+42, +53
	0.5	6.30	1.22	5.15	KCYp, SILp, -9/-10, SCLp, +20/+21, +39/+42, +53
	0.1	5.17	0.94	5.45	KCYp, SILp, SCLp, +20/+21, +39/+42,
10 <sup>4</sup>	10	3.61	1.02	3.51	KCYp, SILp, -9/-10, SCLp, +20/+21, +39/+42, +53
	5	5.51	1.38	3.97	KCYp, SILp, -9/-10, SCLp, +20/+21, +39/+42, +53
	2	5.03	1.30	3.86	KCYp, SILp, -9/-10, SCLp, +20/+21, +39/+42, +53
	1	6.29	1.33	4.71	KCYp, SILp, -9/-10, SCLp, +20/+21, +39/+42, +53
	0.5	5.92	1.34	4.40	KCYp, SILp, -9/-10, SCLp, +20/+21, +39/+42, +53
	0.1	6.54	1.31	4.98	KCYp, SILp, -9/-10, SCLp, +20/+21, +39/+42, +53

**Table 5.2: Statistical assessment of assay performance for the six antibody concentrations used in ChIP-chip assays performed with  $10^7$ ,  $10^6$ ,  $10^5$ , and  $10^4$  cells and H3 K9/K14 diacetylation antibody.**

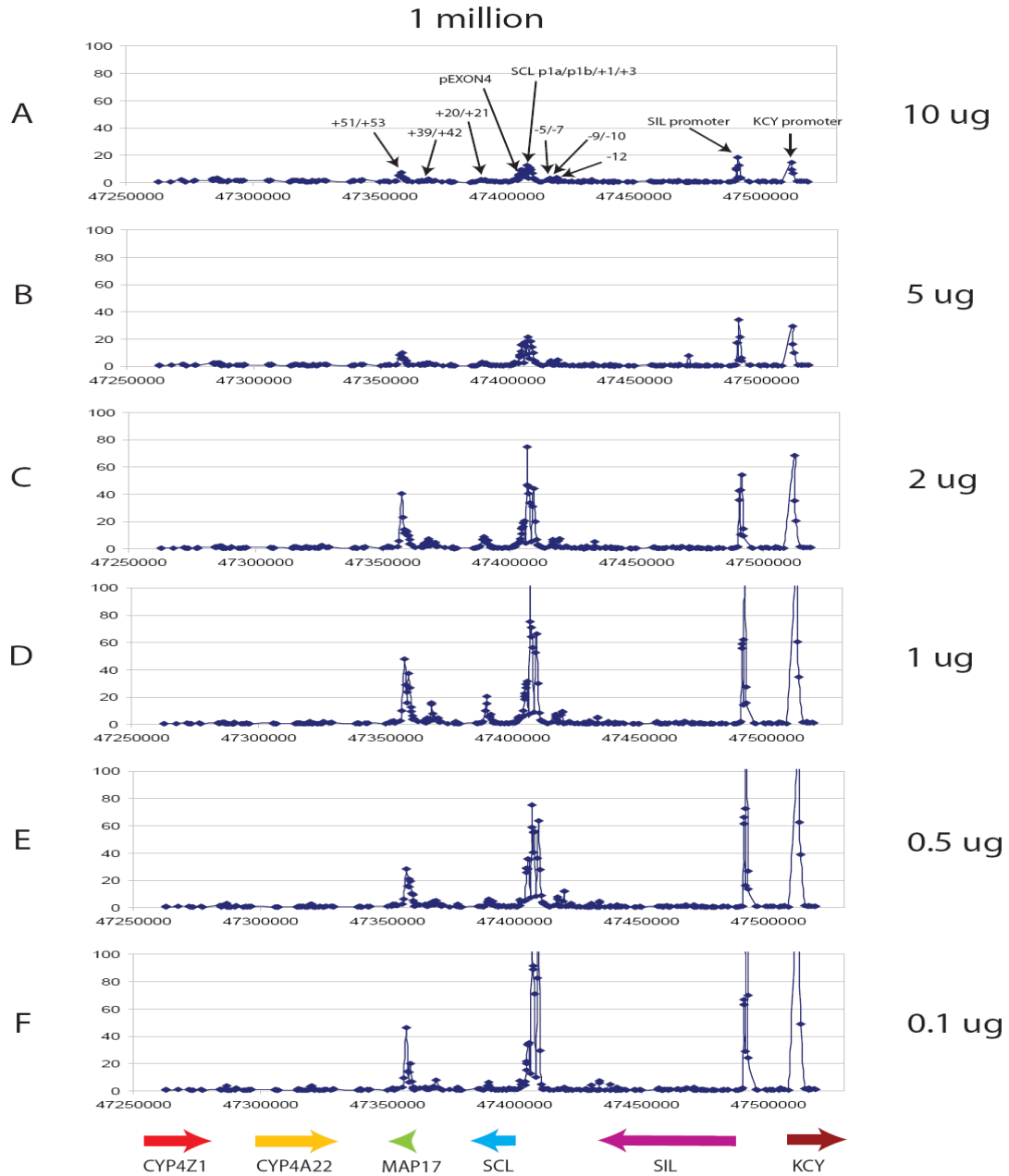
The average fold enrichments at known regulatory elements (signal) were calculated along with the standard deviation of background regions ('noise'). A signal:noise ratio for each antibody and cell amount was determined and the optimal ratio is indicated in red. Known regulatory regions associated with significant enrichments are also indicated for each assay in the titration series.

### **5.5.2. Histone H3 K4 trimethylation experiments**

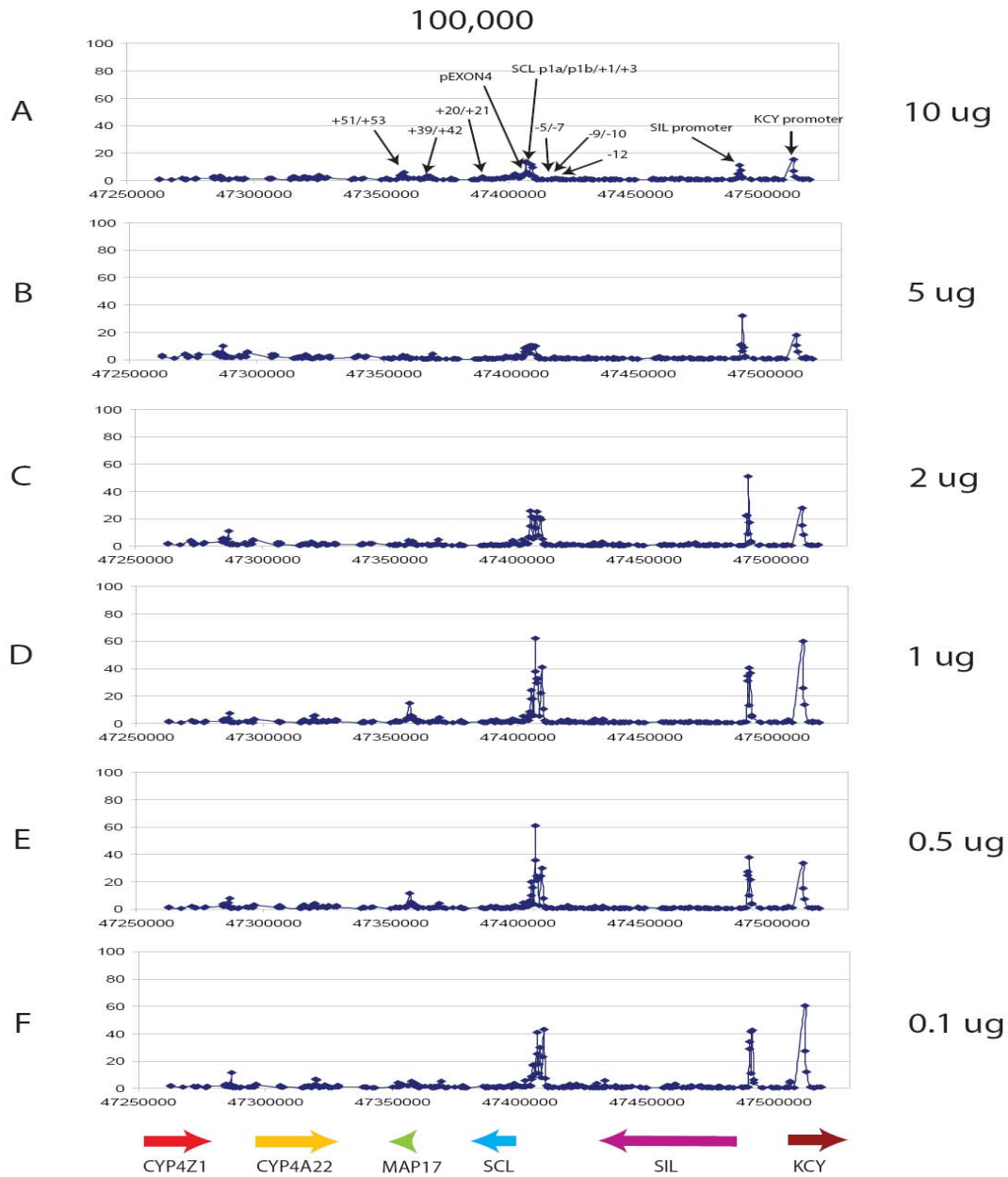
In the previous section the optimal amount of antibody was determined for H3 K9/K14 diacetylation assays performed with a range of cell numbers. However, to ensure that the results obtained for H3 K9/K14 acetylation titration experiments were not related to the efficiency of only the H3 acetylation antibody, ChIP-chip titration experiments with an antibody raised to H3K4me3 were also performed under the same conditions as for H3 K9/K19 acetylation (Figures 5.7 to 5.10). The data was assessed using the same criteria as discussed above (Table 5.3).



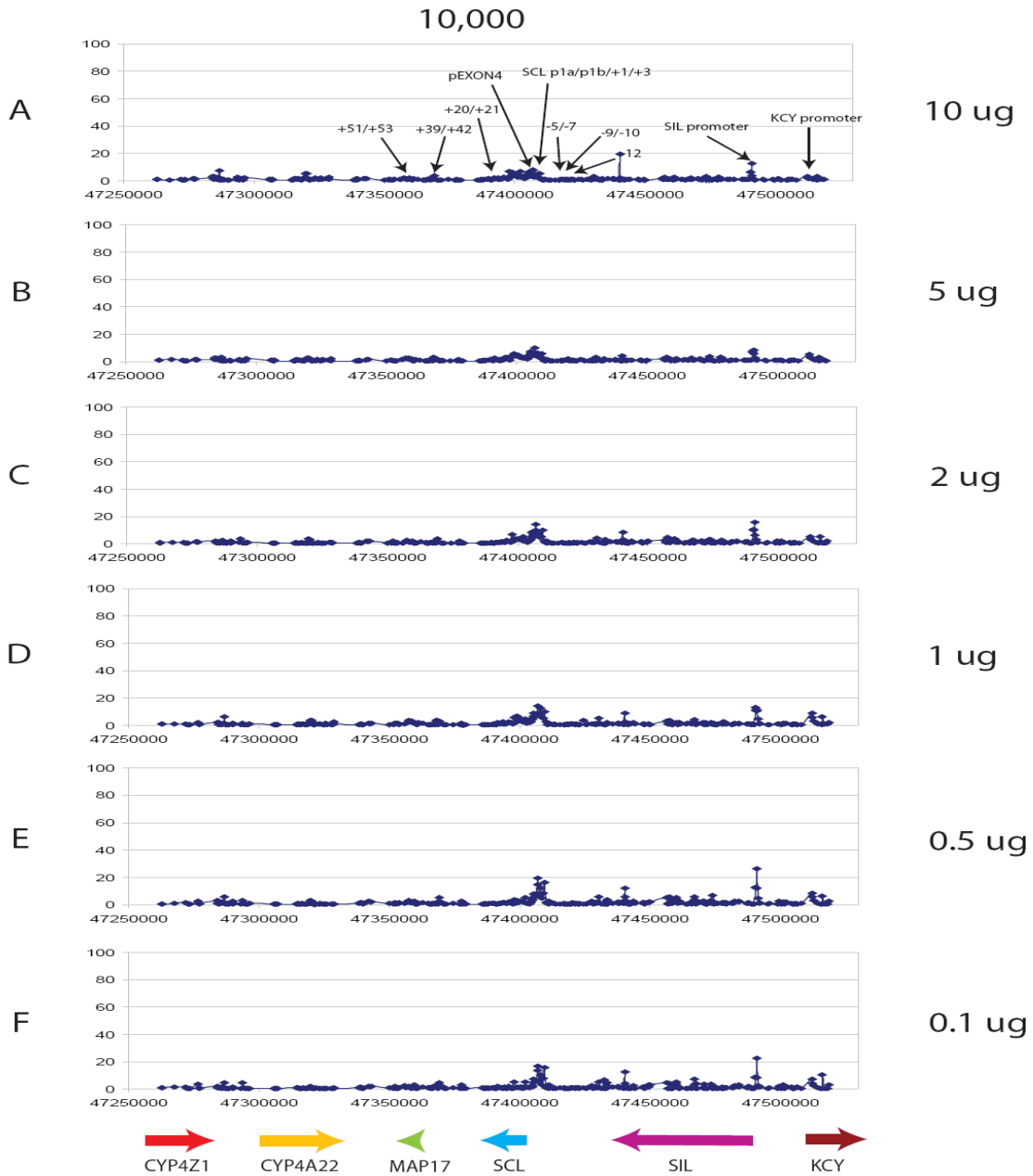
**Figure 5.7: Titration of H3K4me3 antibody in ChIP assays performed with  $10^7$  K562 cells.** The SCL ChIP-chip profiles for experiments performed with  $10^7$  K562 cells and the six antibody concentrations are shown. The known regulatory elements are indicated by black arrows in panel A. Variations in fold-enrichments are observed for regulatory elements across the six antibody concentrations in panels A-F. The detection of enriched regions not associated with known regulatory elements also varied with antibody concentration. Human chromosome 1 coordinates are indicated on the x-axis, with fold enrichments indicated on the y-axis. Each profile is scaled to 100 on the y-axis to allow for comparisons between titration series to be performed. The antibody amount is indicated to the right of each panel. The gene order and direction of transcription is indicated by thick coloured arrows at the bottom of the figure.



**Figure 5.8: Titration of H3K4me3 antibody in ChIP assays performed with  $10^6$  K562 cells.** The SCL ChIP-chip profiles for experiments performed with  $10^6$  K562 cells and the six antibody concentrations are shown. The known regulatory elements are indicated by black arrows in panel A. Variations in fold-enrichments are observed for regulatory elements across the six antibody concentrations in panels A-F. The detection of enriched regions not associated with known regulatory elements also varied with antibody concentration. Human chromosome 1 coordinates are indicated on the x-axis, with fold enrichments indicated on the y-axis. Each profile is scaled to 100 on the y-axis to allow for comparisons between titration series to be performed. The antibody amount is indicated to the right of each panel. The gene order and direction of transcription is indicated by thick coloured arrows at the bottom of the figure.



**Figure 5.9: Titration of H3K4me3 antibody in ChIP assays performed with  $10^5$  K562 cells.** The SCL ChIP-chip profiles for experiments performed with  $10^5$  K562 cells and the six antibody concentrations are shown. The known regulatory elements are indicated by black arrows in panel A. Variations in fold-enrichments are observed for regulatory elements across the six antibody concentrations in panels A-F. The detection of enriched regions not associated with known regulatory elements also varied with antibody concentration. Human chromosome 1 coordinates are indicated on the x-axis, with fold enrichments indicated on the y-axis. Each profile is scaled to 100 on the y-axis to allow for comparisons between titration series to be performed. The antibody amount is indicated to the right of each panel. The gene order and direction of transcription is indicated by thick coloured arrows at the bottom of the figure.



**Figure 5.10: Titration of H3K4me3 antibody in ChIP assays performed with  $10^4$  K562 cells.** The SCL ChIP-chip profiles for experiments performed with  $10^4$  K562 cells and the six antibody concentrations are shown. The known regulatory elements are indicated by black arrows in panel A. Variations in fold-enrichments are observed for regulatory elements across the six antibody concentrations in panels A-F. The detection of enriched regions not associated with known regulatory elements also varied with antibody concentration. Human chromosome 1 coordinates are indicated on the x-axis, with fold enrichments indicated on the y-axis. Each profile is scaled to 100 on the y-axis to allow for comparisons between titration series to be performed. The antibody amount is indicated to the right of each panel. The gene order and direction of transcription is indicated by thick coloured arrows at the bottom of the figure.



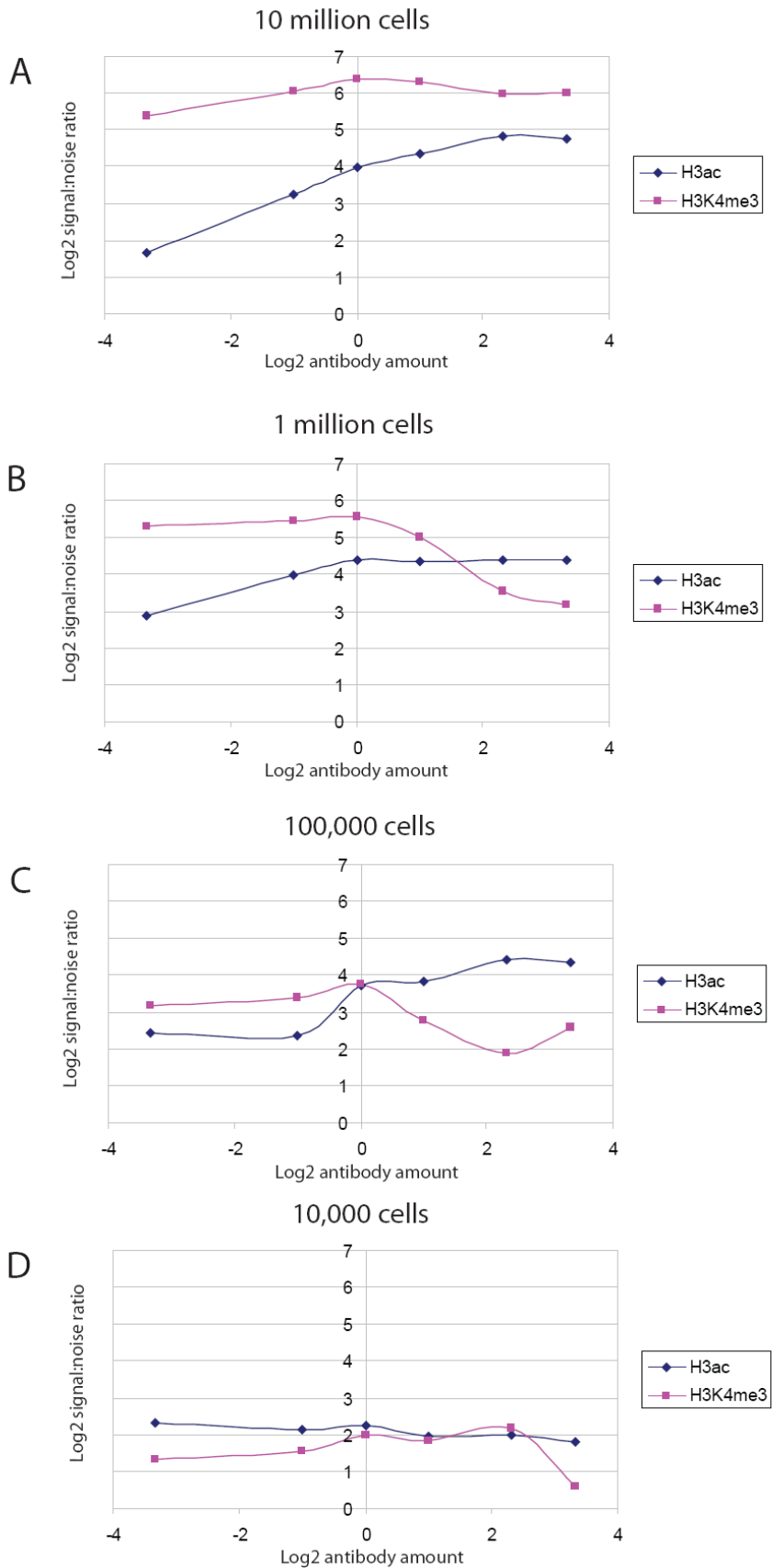
As was found with the histone H3 K9/K14 acetylation assays, there were several significant peaks of enrichment detected at regions not associated with known regulatory for assays performed with  $10^5$  and  $10^4$  cells at all antibody concentration levels. Yet, the overall effect of H3K4me3 antibody concentration on signal:noise across the panel of chromatin from  $10^7$ ,  $10^6$ ,  $10^5$  and  $10^4$  cell equivalents was markedly different than that obtained for the H3 K9/K14 acetylation. The signal:noise levels for experiments which used chromatin from  $10^7$ ,  $10^6$ , and  $10^5$  cells was highest at antibody levels of 1  $\mu\text{g}$ . However, signal:noise decreased more than 6-fold across this range, and resulted in fewer known SCL regulatory regions being detected at  $10^5$  cell levels. Furthermore, the highest signal:noise level for the H3K4me3 ChIP-chip assay with  $10^4$  cells was achieved at an antibody concentration of 5  $\mu\text{g}$  – however, signal:noise was more than 20-fold reduced and only 4 known SCL regulatory elements were detected (when compared to the data obtained for H3K4me3 in conventional ChIP-chip assays with  $10^7$  cells). Thus, as cell numbers were reduced in this series of experiments, the most effective antibody level was higher than for assays with higher cell numbers – this was in marked contrast to the results achieved with the histone H3 K9/K14 titration series.

Cell number	Antibody amount (µg)	Average fold enrichments (Signal)	Standard deviation of background regions (noise)	Signal: noise ratio	Known regulatory elements significantly Enriched
10 <sup>7</sup>	10	23.14	0.36	63.84	KCYp, SILp, SCLp, pExon4, -12, -9/-10, -5/-7, +20/+21, +39/+42, +51/+53
	5	26.00	0.42	61.21	KCYp, SILp, SCLp, pExon4, -12, -9/-10, -5/-7, +20/+21, +39/+42, +51/+53
	2	31.53	0.40	78.64	KCYp, SILp, SCLp, pExon4, -12, -9/-10, -5/-7, +20/+21, +39/+42, +51/+53
	1	32.08	0.38	82.52	KCYp, SILp, SCLp, pExon4, -12, -9/-10, -5/-7, +20/+21, +39/+42, +51/+53
	0.5	25.73	0.38	66.56	KCYp, SILp, SCLp, pExon4, -9/-10, +20/+21, +39/+42, +51/+53
	0.1	16.99	0.41	41.28	KCYp, SILp, SCLp, pExon4, +51/+53
10 <sup>6</sup>	10	4.95	0.55	8.93	KCYp, SILp, SCLp, pExon4, -9/-10, -5/-7, +51/+53
	5	7.93	0.68	11.57	KCYp, SILp, SCLp, pExon4, -9/-10, -5/-7, +51/+53
	2	17.76	0.54	32.30	KCYp, SILp, SCLp, pExon4, -9/-10, -5/-7, +20/+21, +39/+42, +51/+53
	1	27.77	0.58	47.21	KCYp, SILp, SCLp, pExon4, -9/-10, -5/-7, +20/+21, +39/+42, +51/+53
	0.5	24.53	0.55	43.99	KCYp, SILp, SCLp, pExon4, -9/-10, -5/-7, +20/+21, +39/+42, +51/+53
	0.1	33.72	0.84	39.93	KCYp, SILp, SCLp, pExon4, +20/+21, +39/+42, +51/+53
10 <sup>5</sup>	10	3.59	0.59	5.99	KCYp, SILp, SCLp, pExon4, +51/+53
	5	4.40	1.19	3.68	KCYp, SILp, SCLp, pExon4,
	2	7.67	1.13	6.76	KCYp, SILp, SCLp, pExon4,
	1	11.91	0.88	13.48	KCYp, SILp, SCLp, pExon4, +51/+53
	0.5	9.28	0.88	10.46	KCYp, SILp, SCLp, pExon4, +51/+53
	0.1	10.29	1.15	8.89	KCYp, SILp, SCLp, pExon4
10 <sup>4</sup>	10	2.50	1.65	1.51	KCYp, SILp, SCLp, pExon4,
	5	2.97	0.66	4.50	KCYp, SILp, SCLp, pExon4,
	2	3.41	0.94	3.61	KCYp, SILp, SCLp, pExon4,
	1	4.06	1.01	4.02	KCYp, SILp, SCLp, pExon4,
	0.5	4.21	1.42	2.96	KCYp, SILp, SCLp, pExon4,
	0.1	3.65	1.47	2.48	KCYp, SILp, SCLp, pExon4,

**Table 5.3: Statistical assessment of assay performance for the six antibody concentrations used in ChIP-chip assays performed with 10<sup>7</sup>, 10<sup>6</sup>, 10<sup>5</sup>, and 10<sup>4</sup> cells and H3K4me3 antibody.** The average fold enrichments at known regulatory elements (signal) were calculated along with the standard deviation of background regions ('noise'). A signal:noise ratio for each antibody and cell amount was determined and

the optimal ratio is indicated in red. Known regulatory regions associated with significant enrichments are also indicated for each assay in the titration series.

To more clearly describe the differences in the behaviour of the two histone modification assays used in the titration series described above, the relationship between antibody amount and signal:noise was plotted for each panel of cell amounts for both the H3K9/K14 and H3K3me3 assays (Figure 5.11). This figure showed that the general trend between antibody amount and signal:noise displayed different patterns for three of the four cell amounts ( $10^7$ ,  $10^6$  and  $10^5$ ), whilst the pattern for experiments performed with  $10^4$  cells was more similar between the two histone modification assays. In general, signal:noise was greatest for lower H3K4me3 antibody amounts while the reverse was true for experiments performed with the H3 acetylation antibody. Yet, while optimal antibody levels differed between the two antibodies at the various cell equivalents, this analysis showed that it was possible to identify concentrations at which both antibodies performed well. Thus, titration of antibody and cell levels in ChIP-chip experiments are crucial issues that should be addressed when optimizing conditions for low cell numbers.

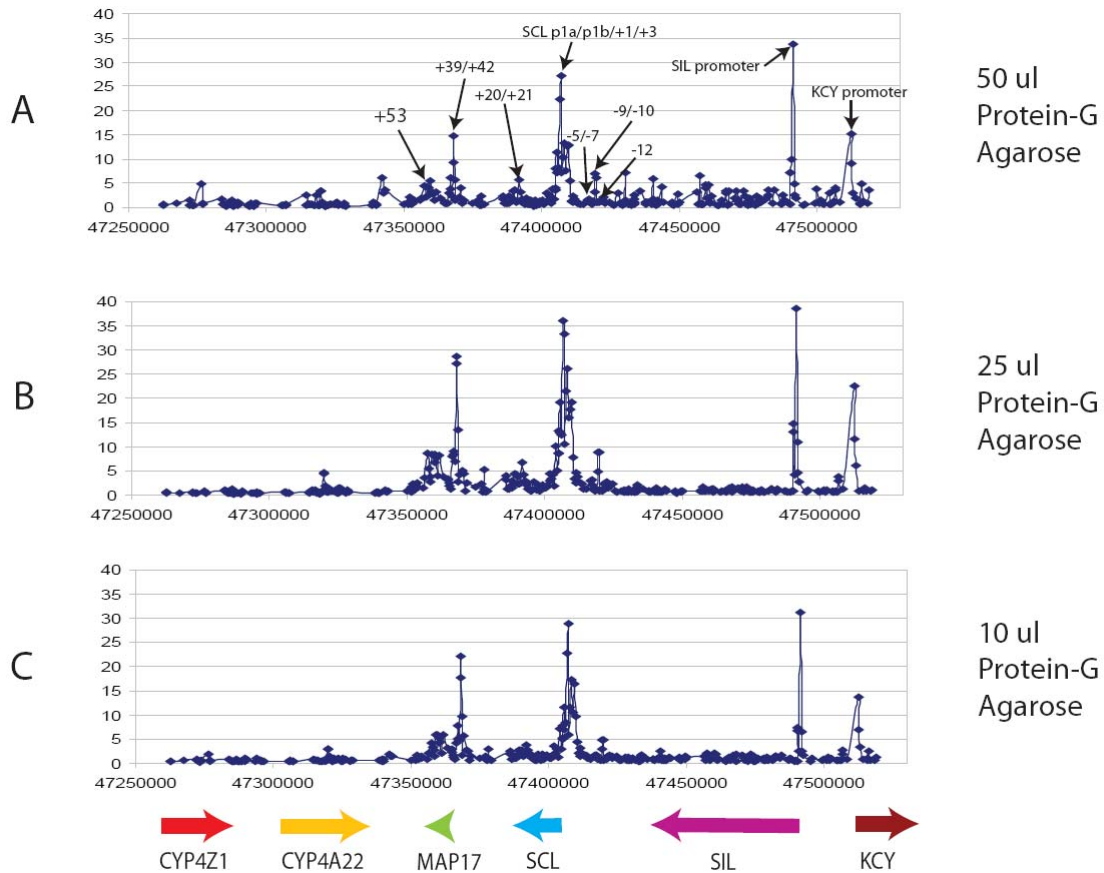


**Figure 5.11: A comparison of the optimal antibody concentrations for titration assays performed with H3 acetylation and H3K4me3 antibodies.** The log<sub>2</sub> values of the six antibody amounts (displayed on the x-axes) were plotted against the log<sub>2</sub> signal:noise ratios (displayed on the y-axes) for assays performed with H3 acetylation and H3K4me3 antibodies across the range of standard cell numbers. The plots in Panels A-D represent the values obtained for antibody titration assays performed with 10<sup>7</sup>, 10<sup>6</sup>, 10<sup>5</sup>, and 10<sup>4</sup> cells respectively with H3ac plots indicated in blue while H3K4me3 plots are in pink.

## **5.6. Examining the effect of protein-G agarose concentration on ChIP-chip experiments performed with low cell numbers**

It was previously reported that low cell number ChIP-chip identified a number of enrichments at regions of the SCL locus (in assays with  $10^5$  and  $10^4$  cell equivalents - see sections, 5.4 and 5.5) for which there was no evidence to support them as being *bona fide* regulatory elements [(such as high levels of non-coding DNA sequence conservation or annotation with features which would normally be associated with regulatory function – i.e., promoters, CpG islands). These regions were considered to be false-positives or regions that enriched non-specifically in ChIP-chip. The presence of these enrichments in assays for histone modifications using low cell numbers precipitated the need to optimize other conditions in the ChIP-chip procedure to remove or account for any technical artifacts of the method.

It was hypothesized that these “false” enrichments may be present in ChIP experiments due to an excess of protein-G agarose. Protein-G agarose contains recombinant protein-G covalently bound to agarose beads and is used in the immunoprecipitation of DNA-protein-antibody complexes as protein-G has a high affinity for IgGs. However, protein-G agarose can also bind non-specifically to DNA during the immunoprecipitation and some protocols use sheared herring or salmon sperm DNA in the immunoprecipitation step to block this non-specific binding (Weinmann *et al.*, 2001). Therefore, the effect of reducing protein-G agarose was examined for the  $10^5$  cell experiment performed with 0.5  $\mu\text{g}$  of H3 acetylation antibody which was associated with the highest standard deviation of background regions (‘noise’) for the  $10^5$  cell titration series (Table 5.2). 50  $\mu\text{l}$  of protein-G agarose was ordinarily used in immunoprecipitation reactions so the effect of using less (25  $\mu\text{l}$  and 10  $\mu\text{l}$ ) protein-G agarose was examined (Figure 5.12).



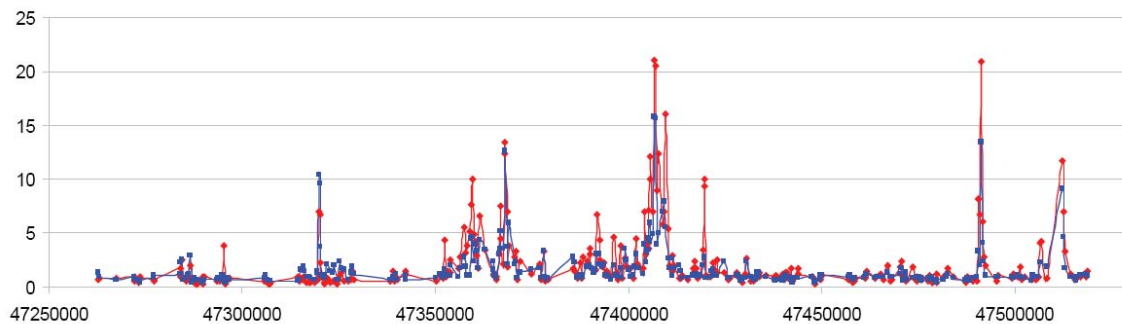
**Figure 5.12: Titration of protein-G agarose in ChIP assays performed with  $10^5$  K562 cells and  $0.5 \mu\text{g}$  of H3 acetylation antibody.** The SCL ChIP-chip profiles for experiments performed with  $10^5$  K562 cells and 50  $\mu\text{l}$ , 25  $\mu\text{l}$ , and 10  $\mu\text{l}$  of protein-G agarose are shown in panels A-C respectively. The known regulatory elements are indicated by black arrows in panel A. Variations in fold-enrichments are observed for regulatory elements across the three protein-G agarose concentrations. The detection of enriched regions not associated with known regulatory elements also varied with protein-G agarose concentration. Human chromosome 1 coordinates are indicated on the x-axis, with fold enrichments indicated on the y-axis. The gene order and direction of transcription is indicated by thick coloured arrows at the bottom of the figure.

It was clear from this targeted titration series that many of the regions associated with no known regulatory function were reduced to background levels when 25  $\mu\text{l}$  or 10  $\mu\text{l}$  of protein-G agarose was used in the immunoprecipitation step instead of 50  $\mu\text{l}$ . This may be because the protein-G agarose was no longer in such excess resulting in less non-specific DNA being immunoprecipitated. The effect of using less protein-G agarose was quantified in terms of average enrichment at known regulatory elements (signal) and standard

deviation of background regions (noise) as described before. The average signal was calculated to be 8.19 for the experiment performed with 25  $\mu\text{l}$  of protein-G agarose and 7.02 for the experiment performed with 10  $\mu\text{l}$  of protein-G agarose. Thus, the use of less protein-G agarose resulted in an increase in average signal when compared to the experiment performed with 50  $\mu\text{l}$  of protein-G agarose (average signal of 6.30). Noise was also reduced in background regions when less protein G-agarose was used, as it was calculated to 0.64 and 0.68 for experiments performed with 25  $\mu\text{l}$  and 10  $\mu\text{l}$  of protein-G agarose respectively (compared to 1.22 for the 50  $\mu\text{l}$  of protein-G agarose experiment). The signal:noise ratios were calculated to be 12.79 and 10.32 for the 25  $\mu\text{l}$  and 10  $\mu\text{l}$  experiments respectively. Thus it seems that 25  $\mu\text{l}$  of protein-G agarose may be a more optimal protein-G agarose concentration for H3 acetylation assays performed with  $10^5$  cells for removing apparent “false” positive enrichments. However, it cannot be ruled out that these “false” enrichments do, in fact, mark the location of *bona fide* regulatory elements which are only identified in ChIP-chip assays with lower cell numbers. This issue is dealt with in more detail in section 5.9 of this Chapter.

### 5.7. Assessing reproducibility of the limited cell numbers ChIP-chip method

With the results of these experiments in mind, the reproducibility of performing ChIP-chip with low cell numbers was assessed. An additional biological replicate (biological replicate refers to a ChIP experiment that was conducted with an independent preparation of chromatin) was performed with the optimal histone H3 K9/K14 acetylation antibody concentration with  $10^4$  cells. Figure 5.13 illustrates the results obtained when comparing two independent biological replicate experiments.



**Figure 5.13: Biological replicates for ChIP-chip assays performed with  $10^4$  cells.** Biological replicate experiments performed with  $10^4$  K562 cells and 0.1  $\mu\text{g}$  of H3 acetylation antibody are presented. The red profile represents the data derived from biological replicate 1, while the blue profile represents the data from biological replicate 2. The human chromosome 1 coordinates are indicated along the x-axis and fold enrichments are indicated on the y-axis. The coloured arrows at the bottom of the figure represent the gene order and direction of transcription.

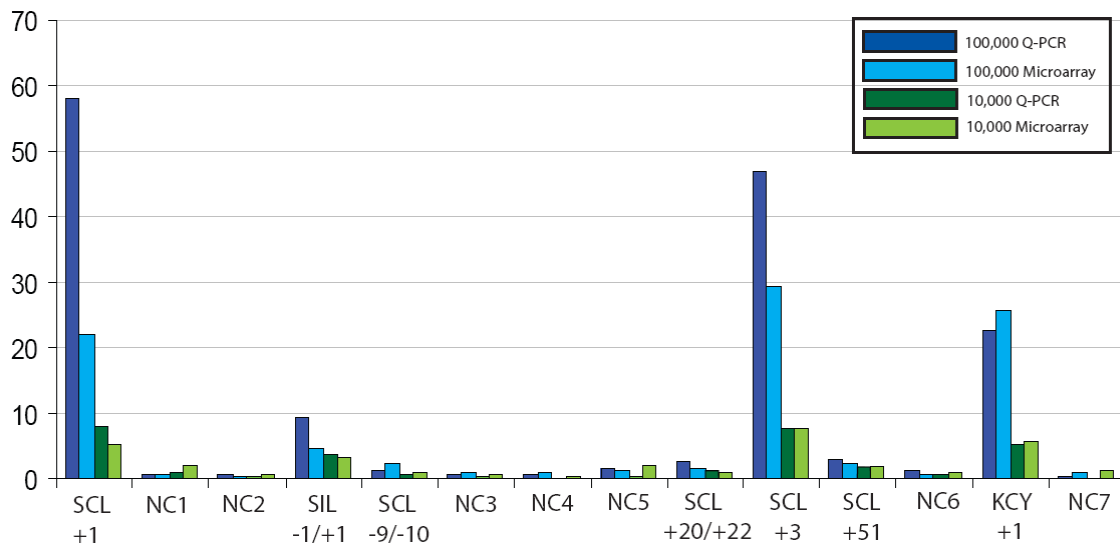
The reproducibility of performing ChIP-chip with limited cell numbers was assessed by calculating the coefficient of variation (CV) across the two biological replicates performed for each assay. This CV was then expressed as a percentage of the mean value for each tile in the SCL tile-path. The mean CV (%) across all tiles for the two H3 acetylation biological replicates was 25.7. While this is a large mean CV, a Pearson correlation coefficient of 0.89 was calculated when corresponding array elements were compared. Pearson's correlation reflects the degree of linear relationship between two variables. A correlation of +1 means that there is a perfect positive linear relationship between two variables, while a correlation of -1 means that there is a perfect negative linear relationship between two variables and a correlation of 0 means there is no linear relationship between the two variables. Therefore a correlation coefficient of 0.89 indicates that there is a high positive linear relationship between fold enrichments reported by the two biological replicates. This suggests that the reduced cell numbers ChIP method could be used to reproducibly identify known regulatory elements when as few as  $10^4$  cells are used per ChIP assay.

### **5.8. Histone H3K4me3 microarray data correlates with real-time quantitative PCR of ChIP material derived from $10^5$ and $10^4$ cells**

In order to verify that enrichments observed in ChIP-chip assays performed with  $10^5$  and  $10^4$  cells were representative of DNA enrichments in the ChIP material, real-time quantitative PCR was performed (See Appendix 1 for sequences of primer pairs). ChIP DNAs obtained from assays performed with  $10^5$  and  $10^4$  cells with 1  $\mu\text{g}$  of H3K4me3 antibody were used for this investigation. A summary of the results are shown in Figure 5.14. The fold enrichments obtained from microarray hybridizations performed with these samples are shown alongside those obtained with quantitative PCR for the 14 SCL



regions investigated. These regions included 7 known regulatory regions which enrich to various levels in ChIP-chip assays and 7 negative controls which show low/no enrichment in ChIP-chip assays. The array elements which showed significant H3K4me3 enrichments in microarray hybridization experiments with  $10^5$  and  $10^4$  cells also showed high levels of enrichment by quantitative PCR. In most cases, the array data and quantitative data were similar within comparable experiments. Array elements corresponding to the SCL+1 and +3 regions and the KCY+1 region reported some of the highest enrichments in ChIP-chip assays and showed the highest enrichments in quantitative PCR of  $10^5$  cell samples. The difference in fold enrichments reported by microarray hybridisation and quantitative PCR can be explained by the fact that quantitation of array elements which show high enrichment are no longer linear when pixel values reach saturation. This is not the case for quantitative PCR as quantitation is linear with DNA copy number in the ChIP sample. Negative control regions were associated with little or no enrichment in quantitative PCR assays performed with material derived from  $10^5$  and  $10^4$  cells.



**Figure 5.14: SYBR green real-time quantitative PCR of H3K4me3 ChIP material from reduced cell numbers.** Enrichments reported by microarray hybridization with material derived from a ChIP performed with  $10^5$  K562 cells and 1  $\mu$ g of H3K4me3 antibody (blue bar) and those reported by quantitative PCR (red bar) are shown side by side for each region. Similarly enrichments reported by microarray hybridization with material derived from a ChIP performed with  $10^4$  K562 cells and 1  $\mu$ g of H3K4me3 antibody (yellow

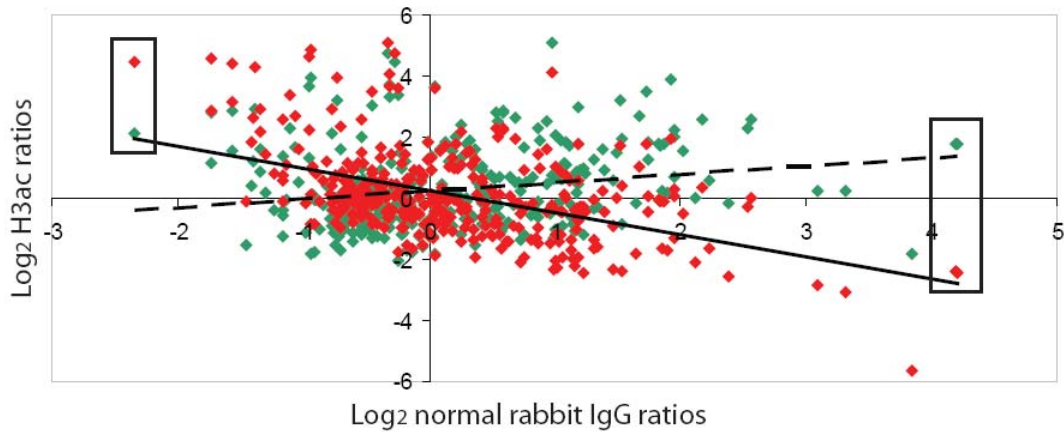
bar) and those reported by quantitative PCR (green bar) are shown side by side for each region. The relevant genomic regions located across the SCL locus are indicated below the x-axis, and fold enrichments are indicated along the y-axis. NC = negative control.

### **5.9. Assessing the impact of normalising ChIP-chip data with normal rabbit IgG data**

As discussed in section 5.5, a number of non-specific enrichments regions were identified as being significantly enriched in ChIP-chip assays performed with  $10^5$  and  $10^4$  cells. Whilst optimizing levels of protein-G agarose can reduce these “false” enrichments (see section 5.6), the use of mock antisera normalization was also investigated with the low cell number ChIP-chip data. It has previously been shown that using mock antisera to normalize histone modification and transcription factor data can effectively remove the effect of non-specific enrichments in conventional ( $10^6$  cell) assays from ChIP-chip data (see Chapters 3 and 4).

In order to further investigate whether these peaks were (i) non novel regulatory elements only observed when ChIP-chip assays were performed with lower cell numbers, or (ii) non-specific interactions/artifact of performing the assays with fewer cells which could be normalised out of the final data sets, a series of ‘mock’ ChIP experiments were performed. As the H3 acetylation and H3K4me3 antibodies were raised in rabbit, mock ChIP experiments were performed with pre-immune antisera from rabbit (normal rabbit IgG). A series of titration experiments were performed with  $10^5$  and  $10^4$  K562 cells with 10  $\mu$ g, 5  $\mu$ g, 2  $\mu$ g, 1  $\mu$ g, 0.5  $\mu$ g, and 0.1  $\mu$ g of pre-immune antisera for the normalisation of corresponding H3 acetylation and H3K4me3 data sets. These experiments would simulate some of the titration conditions used to detect histone modifications in lower cell numbers assays. Thus, by matching up the data for the mock ChIP-chips with their respective experimental datasets, the effect of this type of normalization could be determined. The impact of normalising experimental ChIP-chip data with normal rabbit IgG data was assessed by plotting the  $\log_2$  values for normal rabbit IgG ChIP data against  $\log_2$  values for experimental ChIP data before (green plot) and after normalization (red plot) (Figure 5.15). Data points with low enrichment in the mock ChIP and a high experimental enrichment (known regulatory elements) are located in the top left corner of the scatter plot. Normalisation of these data points with the

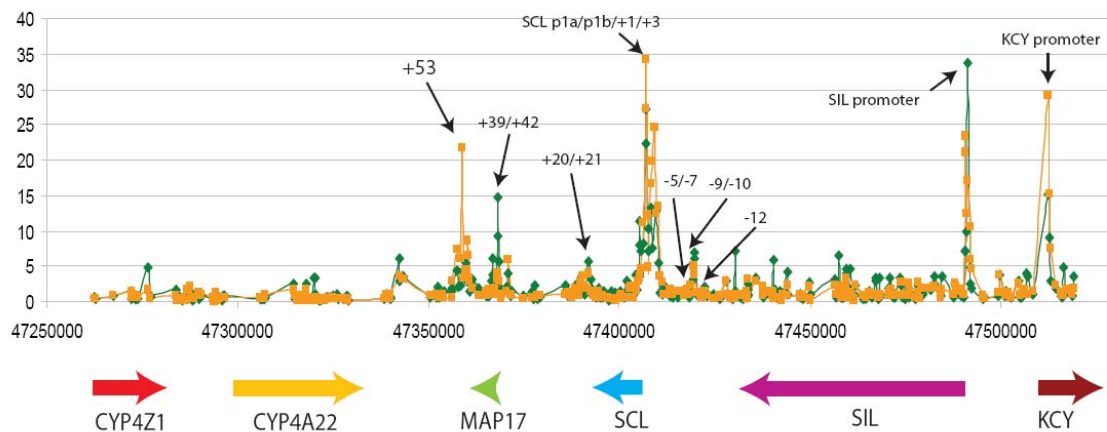
corresponding normal rabbit IgG ChIP value increases the fold enrichment (see boxed data points in top left corner for an example of the data before and after normalisation). Data points with high enrichment in the normal rabbit IgG ChIP and also showing high experimental enrichments (non-specific enrichments) are located in the top right corner of the scatter plot. Normalisation of these data points with the normal rabbit IgG ChIP value decreases fold enrichment values to background levels (see boxed data points on right side of scatter plot for an example of this). Furthermore, no additional peaks of enrichment were introduced in the data by this method of normalisation.



**Figure 5.15: How normalisation with a normal rabbit IgG affects the spread of data for a ChIP-chip experiment performed with  $10^5$  cells.** A scatter plot of  $\text{Log}_2$  mock antibody ChIP chip values versus  $\text{Log}_2$  experimental values before (green data points) and after normalisation (red data points) with normal rabbit IgG data illustrates the impact of normalisation on a data series. The data from a normal rabbit IgG mock ChIP performed with  $10^5$  cells and  $0.5 \mu\text{g}$  of normal rabbit IgG was plotted against the values obtained from a ChIP-chip experiment performed with  $10^5$  cells and  $0.5 \mu\text{g}$  of H3 acetylation antibody. Data points with low enrichment in the mock ChIP and a high experimental enrichment (known regulatory elements) are located in the top left corner of the scatter plot. Normalisation with the corresponding normal rabbit IgG ChIP value increases the fold enrichment for these data points (see boxed data points in top left corner for an example of the data before and after normalisation). Data points with high enrichment in the normal rabbit IgG ChIP and also showing high experimental enrichments (non-specific enrichments) are located in the top right corner of the scatter plot. Normalisation with the normal rabbit IgG ChIP value decreases the fold enrichment for these data points (see boxed data points on right side of scatter plot for an example of this). The slope of the trendline for the unnormalised data (dashed line) and the normalized data (solid

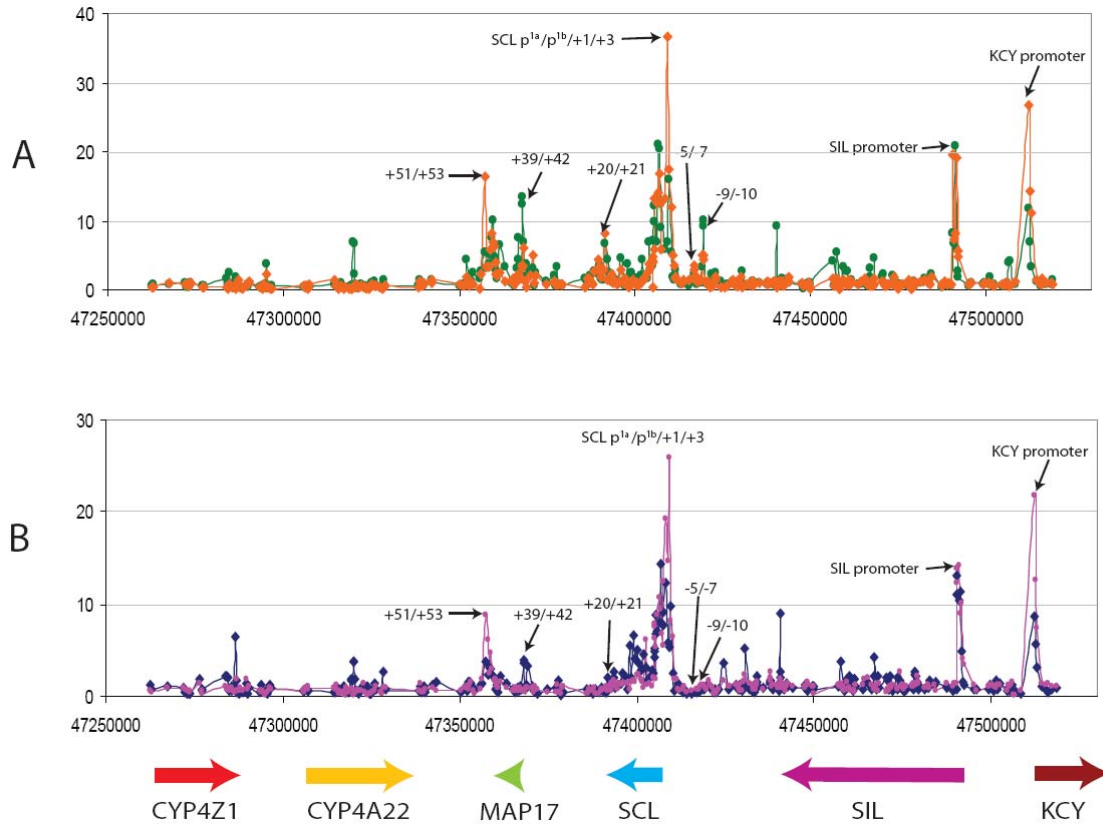
black line) indicate that those data points enriched in the H3ac data set and the normal rabbit IgG data set are reduced following normalisation.

The data from the H3 acetylation experiment performed with  $10^5$  cells and  $0.5 \mu\text{g}$  of antibody was then plotted across the SCL region to examine the impact of normalisation on fold enrichments at known regulatory regions and those enrichments suspected to be non-specific interactions (Figure 5.16). The Figure shows that rabbit IgG normalization had the effect of increasing fold enrichments at the KCY promoter, SCL promoter and the +53 region, while fold-enrichments at other known regulatory elements such as the +39/+42 region, the +20/+21 stem cell enhancer and -9/-10 region were reduced following normalisation. These regions may seem to be more susceptible to non-specific enrichment in rabbit IgG but the reasons for this are not known. More importantly the effect of rabbit IgG normalisation was to reduce to background levels many of the regions associated with significant enrichment that were hypothesized to be non-specific interactions.



**Figure 5.16: Rabbit IgG normalisation of a ChIP-chip experiment performed with  $10^5$  cells.** The data from a ChIP-chip experiment performed with  $10^5$  cells and  $0.5 \mu\text{g}$  of H3K4me3 is presented before and after rabbit IgG normalisation. The green profile represents the unnormalised data and the orange profile represents the rabbit IgG normalised data. Gene order and direction of transcription is indicated by thick coloured arrows at the bottom of the figure. Human chromosome 1 coordinates are indicated on the x-axis and fold enrichments are indicated on the y-axis.

The normal rabbit IgG titration series performed with  $10^4$  cells gave very weak signals in ChIP-chip experiments and it was not possible to quantitate spot enrichments to perform the appropriate normalization of data derived from experiments performed with  $10^4$  cells. However, the effect of using normal rabbit IgG data derived from experiments performed with  $10^5$  cells was examined for normalising H3 acetylation and H3K4me3 experiments performed with  $10^4$  cells. The effect of this normalisation on experiments performed with  $10^4$  cells and 0.1  $\mu\text{g}$  of H3 acetylation antibody and 1  $\mu\text{g}$  of H3K4me3 antibody (both of which gave the highest signal:noise ratios following rabbit IgG normalisation) is presented in Figure 5.17. The data is plotted for each antibody before and after normalisation with normal rabbit IgG data. The peaks of enrichment observed for those interactions not found at known regulatory elements in the unnormalised data sets are reduced to background levels following normalisation. In addition, fold enrichments observed at the majority of known regulatory elements were increased following rabbit IgG normalisation. The data plotted in panel A showed that the fold enrichments at the +53 region, the SCL promoter and the KCY promoter doubled following normalisation of the H3 acetylation data. However, enrichments at lower enriched regions such as the +39/+42 and -9/-10 regulatory elements were reduced following normalisation but remained above the significance threshold. Fold enrichments at the SIL promoter and the +20/+21 enhancer were unaffected by rabbit IgG normalisation. In panel B, fold enrichments associated with the SCL promoter, the KCY promoter and the +53 region were also increased following rabbit IgG normalization of the H3K4me3 data. Fold enrichments associated with the +39/+42 and +20/+21 regions were also reduced following normalisation, while enrichment at the -9/-10 region was found to be significant following normalisation.



**Figure 5.17: Rabbit IgG normalisation of data derived from ChIP-chip experiments performed with  $10^4$  cells.** The data for two ChIP-chip experiments performed with  $10^4$  cells is presented before and after rabbit IgG normalisation. Panel A represents data from a ChIP-chip experiment performed with  $10^4$  K562 cells and 0.1 µg of H3 acetylation antibody. The green profile represents the unnormalised data and the orange profile represents the rabbit IgG normalized data. Panel B represents data from a ChIP-chip experiment performed with  $10^4$  K562 cells and 1 µg of H3K4me3 antibody. The blue profile represents the unnormalised data and the purple profile represents the rabbit IgG normalised data. Gene order and direction of transcription is indicated by thick coloured arrows at the bottom of the figure. Human chromosome 1 coordinates are indicated on the x-axis and fold enrichments are indicated on the y-axis.

In order to empirically define the effect of rabbit IgG on the low cell number data sets, the 24 data sets were analysed to determine the effect on average fold enrichment (signal) and standard deviation of background values (noise) as described previously. The data was then compared with unnormalised data and the findings are summarised in Table 5.4.

Cell number used in ChIP assay	Antibody	Antibody amount ( $\mu\text{g}$ )	Average Signal before and after normalisation		Standard deviation of background regions (noise) before and after normalisation		Signal/noise Before(left) and after Normalization (right)	
$10^5$	H3ac	10	6.81	8.97	0.33	0.46	20.32	19.27
		5	7.81	10.90	0.36	0.42	21.22	25.40
		2	8.21	9.05	0.57	0.37	14.40	24.15
		1	8.16	10.36	0.62	0.36	13.05	28.38
		0.5	6.30	8.00	1.22	0.54	5.15	14.62
		0.1	5.17	7.24	0.94	0.31	5.45	23.23
$10^5$	H3K4me3	10	3.59	4.65	0.59	0.55	5.99	8.33
		5	4.40	6.68	1.19	0.88	3.68	7.52
		2	7.67	9.77	1.13	0.68	6.76	14.17
		1	11.91	17.36	0.88	0.49	13.48	34.81
		0.5	9.28	13.69	0.88	0.64	10.46	21.26
		0.1	10.29	17.75	1.15	0.45	8.89	38.90
$10^4$	H3ac	10	3.61	4.24	1.02	0.42	3.51	9.92
		5	5.51	6.91	1.38	0.47	3.97	14.56
		2	5.03	5.37	1.30	0.64	3.86	8.39
		1	6.29	7.66	1.33	0.42	4.71	17.87
		0.5	5.92	7.68	1.34	0.41	4.40	18.50
		0.1	6.54	8.30	1.31	0.33	4.98	24.76
$10^4$	H3K4me3	10	2.50	2.96	1.65	0.92	1.51	3.20
		5	2.97	3.81	0.66	0.50	4.50	7.61
		2	3.41	3.86	0.94	0.51	3.61	7.45
		1	4.06	5.38	1.01	0.46	4.02	11.46
		0.5	4.21	5.96	1.42	0.97	2.96	6.14
		0.1	3.65	5.13	1.47	1.01	2.48	5.07

**Table 5.4: Comparisons of  $10^5$  and  $10^4$  ChIP-chip data before and after rabbit IgG normalisation.** The average fold enrichment (signal) values associated with known regulatory elements were calculated before and after rabbit IgG normalisation. The standard deviation of known background regions (noise) was also calculated before and after normalization. The optimal so-called signal:noise ratio is indicated in red for each titration series before and after normalisation.

As can be seen from this Table, rabbit IgG normalisation had a profound impact on the signal:noise for the  $10^4$  and  $10^5$  cells data series in two ways. Firstly, three of the four optimal antibody concentrations for maximum enrichments in each series were reduced following normalization compared to those concentrations obtained previously (see section 5.5). Taking normalization into account, the optimal concentration for ChIPs

performed with low cell numbers ( $10^5$  or  $10^4$ ) is reduced somewhere in the range of 10-100 fold when compared to the optimal antibody concentration for conventional ChIPs performed with  $10^7$  cells. Second, the average signal for each of the 24 hybridisations increased following normalisation (by between 22-48% following normalisation), while standard deviations in the background regions were decreased in 22 of the 24 experiments (by between 36-64% following normalization). Overall, this resulted in an increase signal:noise in 23 of the 24 experiments (between 70%-270% following normalisation); the one remaining experiment ( $10^5$  cells and 10  $\mu$ g H3 acetylation antibody) showed a small reduction in the signal:noise ratio following normalisation which was due to a larger increase in the standard deviation of background regions relative to the increase in average signal.

Cell number	Histone modification	Mean series signal before and after normalisation		% increase/decrease after normalisation	Mean series noise before and after normalisation		% increase/decrease after normalisation	Mean series signal:noise before/after normalisation		% increase/decrease after normalisation
$10^5$	H3ac	7.07	9.08	+28.40	0.67	0.41	-39.10	13.625	22.50	+69.68
$10^5$	H3K4me3	7.85	11.65	+48.28	0.97	0.615	-36.59	8.21	20.83	+153.73
$10^4$	H3ac	5.48	6.69	+22.06	1.28	0.44	-64.79	4.23	15.66	+269.64
$10^4$	H3K4me3	3.46	4.51	+30.28	1.19	0.72	-38.88	3.18	6.82	+114.51

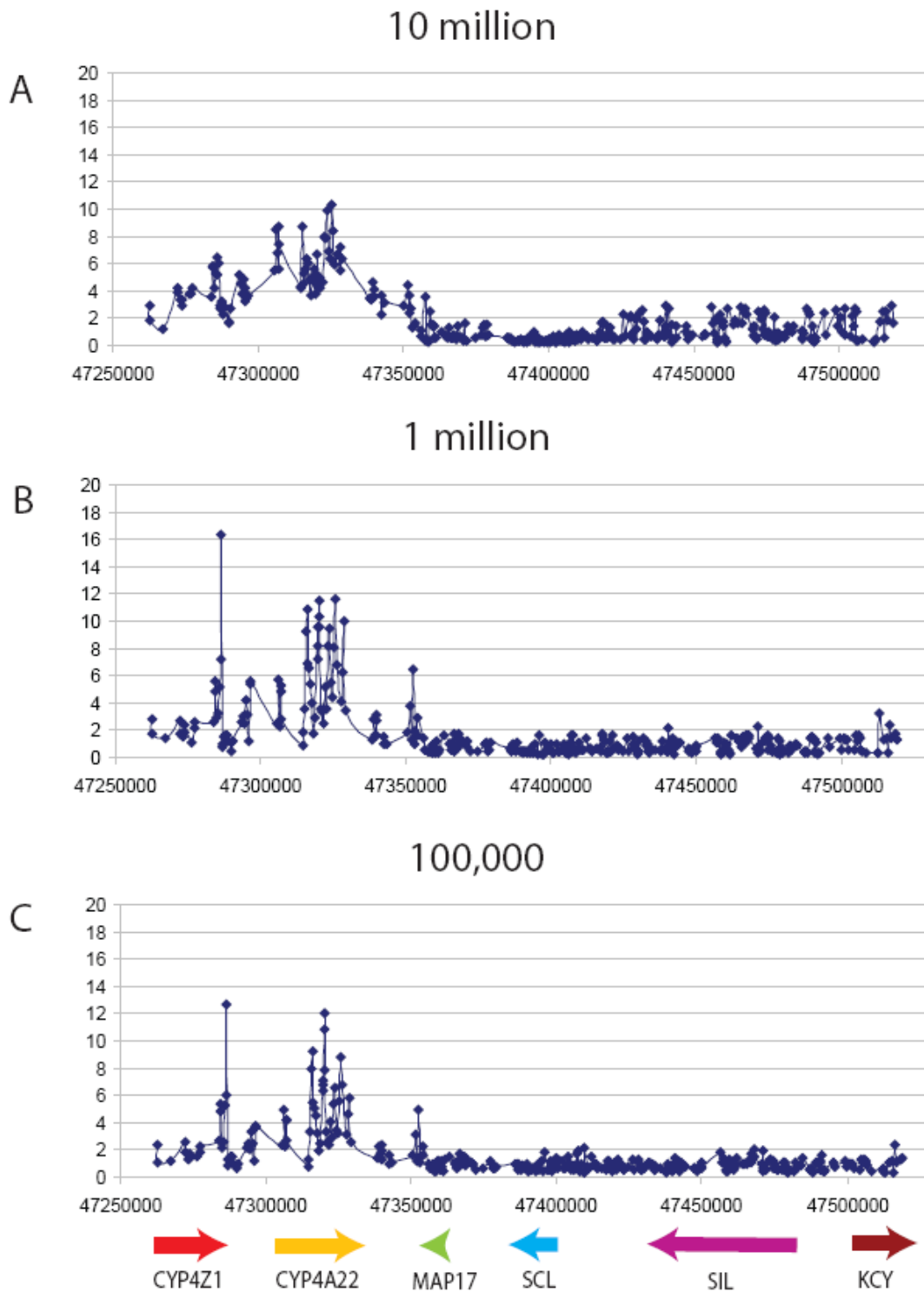
**Table 5.5: Assessing the impact of normalisation on the average titration series signal, average series standard deviation and average series signal:noise values for experiments performed with  $10^5$  and  $10^4$  cells.** The mean titration series ‘signal’ for all six antibody concentrations was calculated for experiments performed with  $10^4$  and  $10^5$  cells with H3 acetylation (H3ac) and H3K4me3 antibodies . Mean signal values were calculated before and after normalisation. The mean titration series ‘noise’ was also calculated before and after normalization. This allowed for a mean series signal:noise ratio to be calculated before and after normalisation and the percentage increase in signal:noise following normalisation is presented in the final column.

### 5.10. Testing of other histone modification antibodies with the modified ChIP-chip method

Finally, to further explore the utility of the lower cell ChIP-chip method to examine other aspects of regulatory function associated with histone modifications. Lower cell number ChIP-chip was validated for three other histone methylation modifications (H3K27me3,



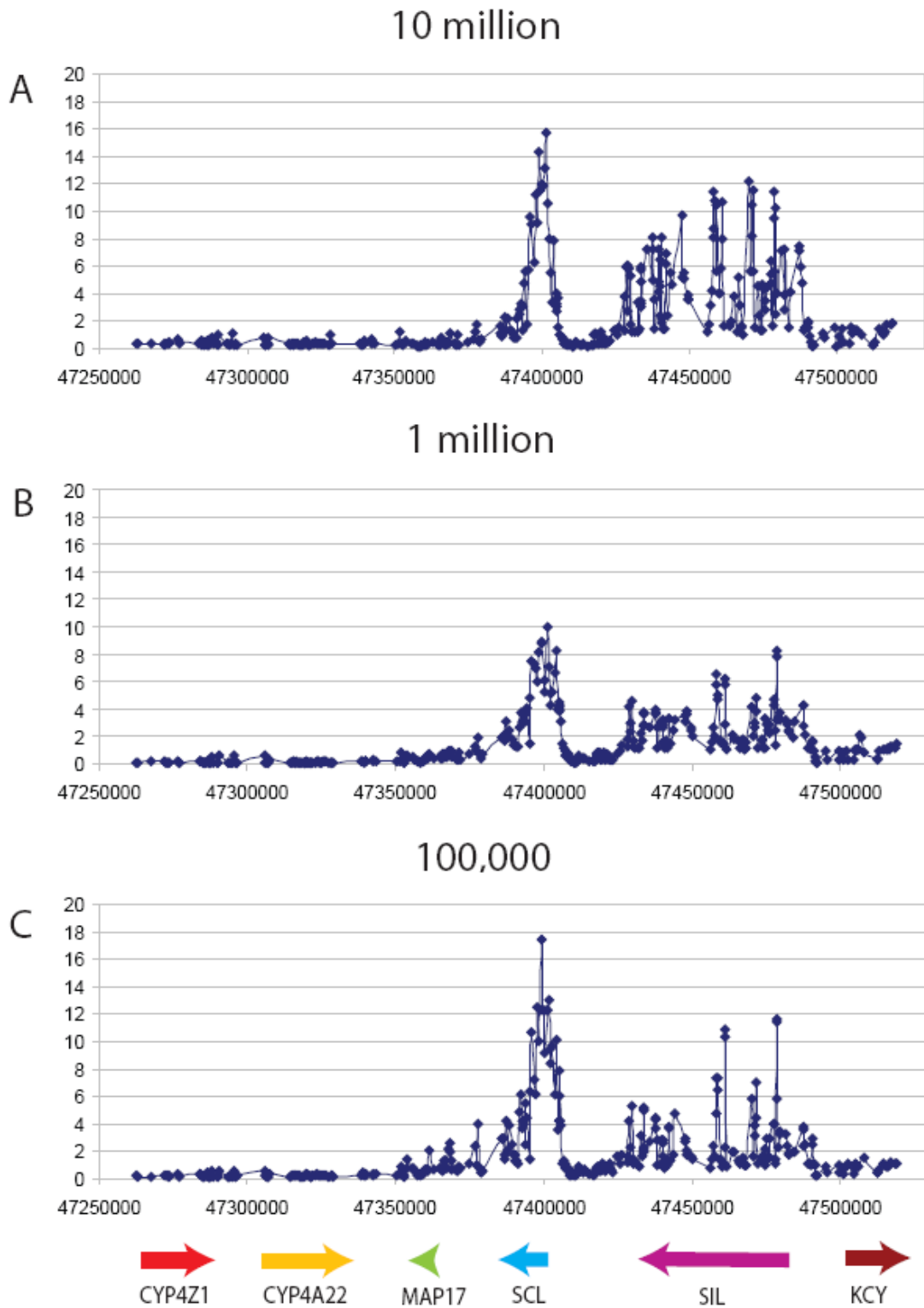
H3K36me3 and H3K79me3). Because of time limitations, comprehensive titration panels could not be performed for these assays. However, based on the results of this Chapter, optimal antibody and cell number conditions were used in these additional assays as those determined for the H3K4me3. The resulting datasets obtained with assays using  $10^6$  and  $10^5$  K562 cells were compared to experiments performed with  $10^7$  K562 cells. Figure 5.18 illustrates the application of the modified ChIP-chip method to detect regions of the SCL locus associated with H3K27me3 in K562 cells. This showed that the CYP4Z1 and CYP4A22 genes (which are not expressed in K562) are associated with elevated levels of H3K27me3 (i.e., H3K72me3 is a mark of repressed gene expression) in assays performed with  $10^6$  and  $10^5$  cells, albeit to slightly different levels as that obtained with conventional  $10^7$  cell assays. Therefore this assay could be used to reliably detect regions enriched for H3K27me3 from as few as  $10^5$  cells.



**Figure 5.18: Testing the modified ChIP-chip method to detect regions associated with H3K27me3.** ChIP-chip data obtained from a experiments performed with a range of K562 cell numbers and an antibody raised to H3K27me3. Panel A: ChIP-chip experiment performed with  $10^7$  cells and  $10\ \mu\text{g}$  H3K27me3.

Panel B: A ChIP-chip experiment performed with  $10^6$  cells and 1  $\mu\text{g}$  H3K27me3, Panel C:  $10^5$  cells, 0.5  $\mu\text{g}$  H3K27me3. Fold enrichment values are presented on the y-axis and the human chromosome 1 coordinates are indicated on the x-axis. The location of genes in the region and direction of transcription is indicated by coloured arrows.

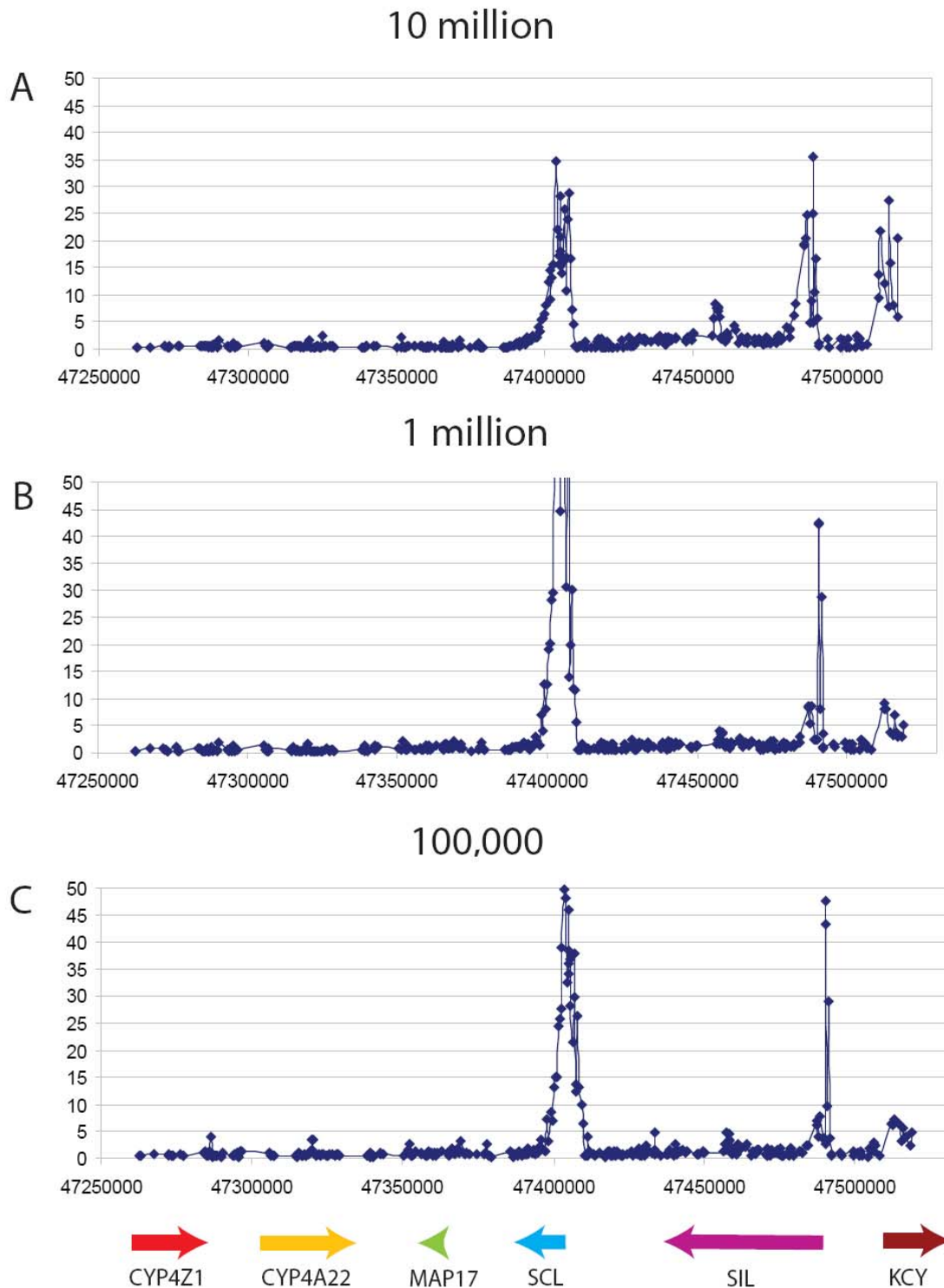
An antibody raised to H3K36me3 was also tested with the modified ChIP-chip method. H3K36me3 is known to be associated with the transcribed portion of active genes (Bannister *et al.*, 2005 b) and was found to be associated with the transcribed portion of the SCL and SIL genes (Figure 5.19) in the K562 cell line. Once again, this assay was found to perform similarly when  $10^7$ ,  $10^6$  or  $10^5$  cells were used.



**Figure 5.19: Testing the modified ChIP-chip method to detect regions associated with H3K36me3.** ChIP-chip data obtained from a experiments performed with a range of K562 cell numbers and an antibody raised to H3K36me3. Panel A: ChIP-chip experiment performed with  $10^7$  cells and  $10\ \mu\text{g}$  H3K36me3.

Panel B: A ChIP-chip experiment performed with  $10^6$  cells and 1  $\mu\text{g}$  H3K36me3, Panel C:  $10^5$  cells, 0.5  $\mu\text{g}$  H3K36me3. Fold enrichment values are presented on the y-axis and the human chromosome 1 coordinates are indicated on the x-axis. The location of genes in the region and direction of transcription is indicated by coloured arrows.

Lastly, ChIP-chip assays for histone H3K79me3 (Figure 5.20) were performed. The role of this modification in mammalian cells is not well understood, but here it was found to associate with the transcribed portion of SCL, SIL and KCY genes immediately downstream of the TSS. All three of these genes are expressed in K562. This pattern of enrichment was again similar across the range of cell numbers studied.

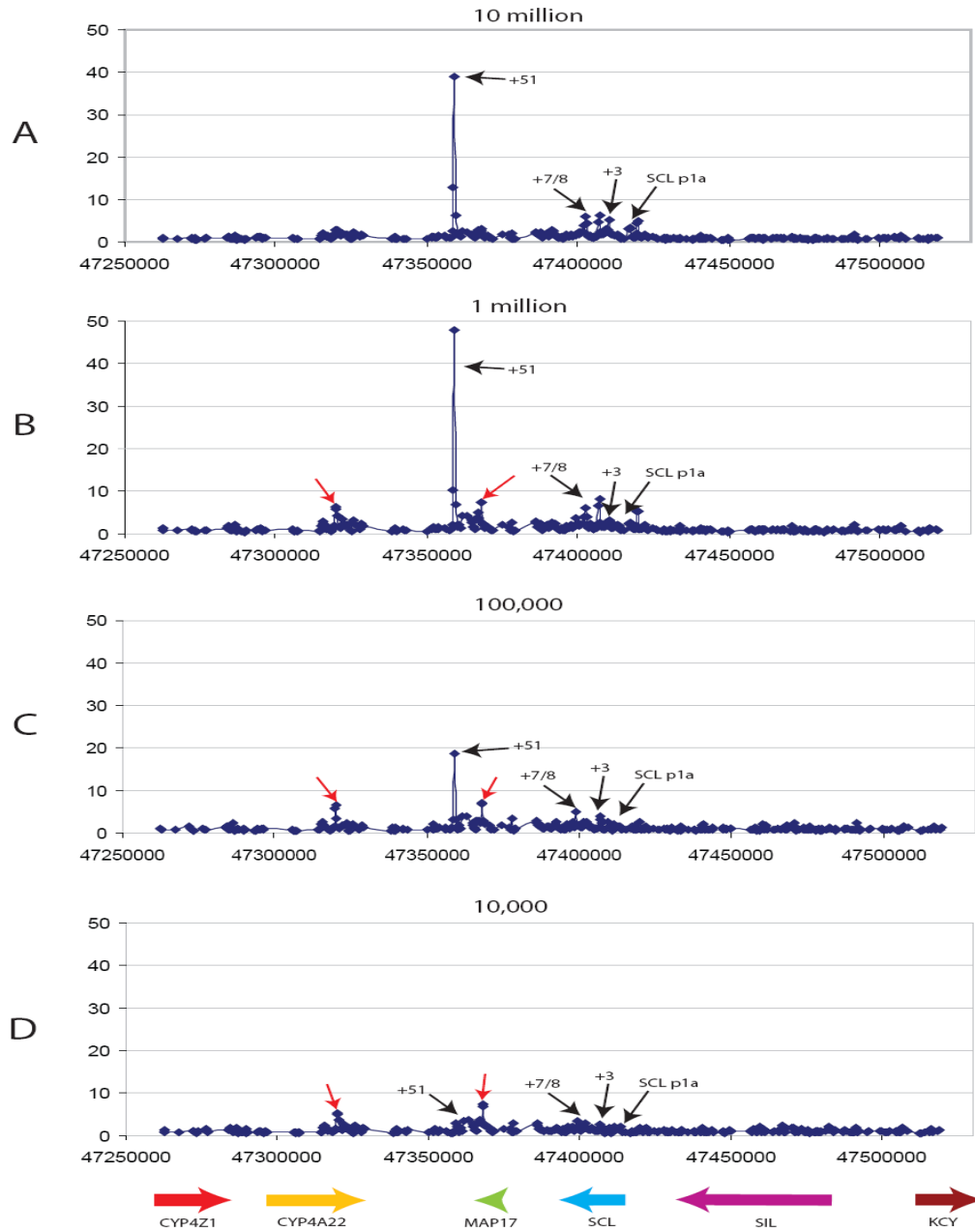


**Figure 5.20: Testing the modified ChIP-chip method to detect regions associated with H3K79me3.** ChIP-chip data obtained from a experiments performed with a range of K562 cell numbers and an antibody raised to H3K79me3. Panel A: ChIP-chip experiment performed with  $10^7$  cells and  $10\ \mu\text{g}$  H3K79me3. Panel B: A ChIP-chip experiment performed with  $10^6$  cells and  $1\ \mu\text{g}$  H3K79me3, Panel C:  $10^5$  cells,  $0.5\ \mu\text{g}$

H3K79me3. Fold enrichment values are presented on the y-axis and the human chromosome 1 coordinates are indicated on the x-axis. The location of genes in the region and direction of transcription is indicated by coloured arrows.

### **5.11. Detection of transcription factor interactions using low cell number ChIP-chip**

While the focus of the modified ChIP-chip method was to investigate histone modifications in low numbers of cells, it was important to determine if ChIP-chip could be used to study transcription factor interactions in low cell numbers. Many transcription factors are expressed at specific stages of development often in discrete populations of cells – therefore the development of a method to study transcription factor interactions in low numbers of cells would be of great benefit. A number of GATA-1 interactions have been previously characterised in K562 cells at the SCL locus (Dhami, submitted). ChIP-chip experiments were performed with  $10^7$ ,  $10^6$ ,  $10^5$  and  $10^4$  cells for the detection of GATA-1 interactions. Because of time limitations, comprehensive titration panels could not be performed for these assays and only one antibody concentration was chosen at each level of cell number. Based on the optimization experiments conducted for histone modifications, experiment performed with  $10^6$  cells used 1 ug of GATA-1 antibody and experiments performed with  $10^5$  cells and  $10^4$  cells used 0.5 ug of GATA-1 antibody. The results of these experiments are presented in Figure 5.21.



**Figure 5.21: Detection of GATA-1 enrichments in reduced numbers of K562 cells.**  $10^7$ ,  $10^6$ ,  $10^5$ , and  $10^4$  cells were used in ChIP-chip experiments for the detection of GATA-1 enrichments. Panel A represents the experiment performed with  $10^7$  cells and 10  $\mu$ g of antibody. Panel B represents the experiment performed with  $10^6$  cells and 1  $\mu$ g of antibody. Panel C represents the experiment performed with  $10^5$  cells and 0.5  $\mu$ g of antibody and panel D represents the experiment performed with  $10^4$  cells and 0.5  $\mu$ g of antibody. The gene order and direction of transcription is indicated below panel D. The human chromosome 1 coordinates are indicated along the x-axes, while the values on the y-axes represent fold enrichments observed in ChIP-chip assays. Known GATA-1 interacting regions are indicated by black arrows while red arrows indicate significant enrichments not associated with known GATA-1 interactions.



In the experiment performed with  $10^7$  cells the genomic regions which displayed significant GATA-1 enrichments are shown. These regions include the SCL erythroid enhancer (+51) which displays the highest fold enrichment, and lower enrichments are observed at the SCL p<sup>1a</sup>, the +3, and +7/+8 regions. These regions are located at or within close proximity to regions of histone H3 acetylation and known regulatory activity (Dhami 2007, submitted). These regions also contain highly conserved GATA family consensus binding sequences, providing further support that they are genuine GATA-1 binding sites in K562. The +51 element can be significantly enriched in assays performed with  $10^6$  cells and  $10^5$  cells, while no enrichment is observed for  $10^4$  cells. While the +51 region is still detected in  $10^5$  cell assays, there is only one array element which is significantly enriched as opposed to three neighbouring array elements which are enriched in assays performed with  $10^7$  and  $10^6$  cells. A number of significantly enriched peaks are observed as cell numbers are reduced (indicated by red arrow-heads), which do not contain a conserved GATA family consensus sequence (data not shown) and may represent an artifact of reducing the number of cells used in a ChIP-chip assay. The subtly enriched +3 region is not detected as significantly enriched in the assays performed with  $10^6$  or  $10^4$  cells, while GATA-1 interaction at SCL p<sup>1a</sup> are not detected as significant when assays were performed with  $10^5$  or  $10^4$  cells. The data suggests that ChIP-chip assays can be used for the detection of highly enriched transcription factor interactions when assays are performed with  $10^7$ - $10^5$  cells. Low enrichments, however, do not reproduce through to low cell number ChIP-chip experiments as the overall level of enrichments decrease and they are lost in the 'noise'.

## **5.12. Discussion**

This Chapter describes experiments aimed at optimizing the conditions necessary to perform ChIP-chip for the detection of regulatory elements in the human genome when using reduced numbers of cells. The SCL genomic tiling path array was used in ChIP-chip to detect statistically significant H3 K9/K14 acetylation and H3K4me3 events associated with known regulatory elements in ChIP assays performed with  $10^4$  -  $10^7$  K562 cells. It was demonstrated that SCL array platform was sensitive enough to detect these

histone modifications from as few as  $10^4$  cells and could reproducibly detect and quantify ChIP enrichments, which were confirmed by quantitative PCR.

#### **5.12.1. The development of a ChIP-chip method applicable to small cell populations**

The main focus of this Chapter was the development of a ChIP-chip method for use with low numbers of cells. Three parameters were investigated for their effect on the efficiency of the ChIP procedure, namely chromatin amount, antibody concentration, and protein-G agarose concentration. It was shown that chromatin aliquots equivalent to  $10^4$  cells could be used to identify known regulatory elements in the SCL locus that were associated with H3 acetylation and H3K4me3. As chromatin amounts were reduced in ChIP assays it became apparent that antibody concentration was an important factor in identifying known regulatory elements. Generally when using less chromatin in a ChIP assay, it was found to be more beneficial to use reduced antibody concentrations, although the optimal antibody to chromatin ratio may be assay-specific. However, that said, the modified procedure was also demonstrated to perform well for H3K27me3, H3K36me3, and H3K79me3 assays, by extrapolating the conditions used for H3K4me3 and H3K9.K14ac. Thus, it may not always be necessary to perform complete titration series for all assays in order to determine a good working range of conditions for low cell number ChIP-chip.

Protein-G agarose concentration in the immunoprecipitation step was also examined for its effect on ChIP efficiency and how it effects non-specific enrichments in the resultant ChIP sample. It was observed that using less protein-G agarose (25 $\mu$ l) may be more beneficial for experiments performed with  $10^5$  cells as too much protein-G agarose may be contributing to ‘noise’ due to non-specific interaction with DNA. This effect was not examined for experiments performed with  $10^4$  cells but it is expected that a similar relationship may also exist for ChIP experiments performed with this number of cells. However, in practical terms, the use of low amounts of protein-G agarose may be limited in low cell ChIP-chip assays, as it is difficult to visualize the protein-G agarose pellets during the ChIP wash and elution steps as it is used in diminishingly small quantities.

The normalization of low cell number ChIP-chip data sets with rabbit “mock” IgG data was also examined and resulted in profound improvements in signal:noise levels across

the SCL locus for the detection of H3K4me3 and H3K9/14ac histone modifications. Many other laboratories routinely publish ChIP-chip data without considering the effect that non-specific interactions can contribute to the overall dataset. However, given that the vast majority of ChIP-chip data is currently performed using conventional ChIP, the contribution that this type of noise has on these datasets is not known. However, as more biological studies move towards low cell number ChIP-chip, the need to consider non-specific noise will become increasingly important.

### **5.12.2. Further optimisation of the modified ChIP-chip method**

In this study, all aliquots of chromatin used for cell equivalent experiments were derived from batches of chromatin prepared from  $10^8$  cells. Therefore, optimising the preparation of chromatin from fewer numbers of cells is an important consideration. For example, when cross-linking low numbers of cells, it may be important to perform the cross-linking step in smaller volumes directly in tubes (as apposed to large flasks) to minimise the loss of material associated with transferring cross-linked cells from flasks to tubes. Cell and nuclei pellets may need to be coloured with a dye when preparing chromatin from low numbers of cells to minimize loss of material. The sonication of smaller numbers of cells may also need to be optimised to generate fragments of the correct size. Finally it may be beneficial to coat tubes in a bovine serum albumin to prevent chromatin proteins adhering to the sides of tubes during the immunoprecipitation step. This may enhance the recovery of precipitated histones. It may also be possible to use the modified ChIP method to investigate histone modifications in fewer cells as random amplification methods (discussed in the introduction to this Chapter) could be used to amplify ChIP DNA samples and provide a quantifiable signal on microarrays. Amplification of ChIP DNA may mean that investigating histone modifications in  $10^3$  or fewer cells may be possible using microarrays.

While the focus of this study was to develop a method for the study of histone modifications in human cell types limiting in number, a ChIP-chip assay for the detection of the transcription factor GATA-1 enrichment in low numbers of cells was also investigated. Proof-of-principle experiments showed that it was possible to detect known transcription factor interactions from as few as  $10^5$  cells. However, given that a number

of parameters (chromatin amount, antibody concentration and protein-G agarose concentration) were shown to be important for detecting histone modifications, the optimal concentrations for detecting transcription may be different to those used for detecting histone modifications and require further detailed investigations.

### **5.12.3. Comparisons with other methods**

A number of ChIP-based methods have been developed recently for investigating histone modifications in small populations of cells as discussed in the introduction of this Chapter (O'Neill *et al.*, 2006; Attema *et al.*, 2007; Dahl and Collas, 2007). The carrier-ChIP method developed by O'Neill and colleagues used *Drosophila* carrier chromatin and PCR amplification of ChIP DNAs to analyse histone modifications in as few as 100 cells (O'Neill *et al.*, 2006). The miniChIP method was developed by Attema and colleagues for the analysis of  $5 \times 10^4$  cells in early haematopoiesis (Attema *et al.*, 2007). The method was developed by altering a number of steps in the ChIP procedure which included the formaldehyde cross-linking step, sonication, pre-clearing and antibody immunoprecipitation conditions. Finally Q<sup>2</sup>ChIP was developed to analyse histone modifications in as few as 100 cells (Dahl and Collas, 2007). Alterations in a number of ChIP steps were used to increase efficiency over conventional ChIP. These included cross-linking cells in suspension rather than in flasks to enhance cell recovery, while a HDAC inhibitor was used during the cross-linking stage as opposed to after cross-linking to maximise the amount of acetylated histones that could be precipitated. However, none of these studies demonstrated that their modified ChIP method could be used for the characterisation of histone modification patterns in a high-throughput manner on microarrays. The carrier ChIP, miniChIP, and Q<sup>2</sup>ChIP methods all used quantitative PCR to determine histone modification enrichments at selected loci and the use of quantitative PCR is limited to investigating smaller genomic regions. In addition it may not be possible to use the carrier ChIP method for microarray analysis of histone modifications as the vast majority of the ChIP DNA is composed of *Drosophila* DNA, which would affect labeling efficiency of the human DNA. Thus the method described in this Chapter is the first known report of microarrays being used to detect histone modifications from as few as  $10^4$  cells.

However, with the recent developments in sequencing technology (Bentley, 2006) such as the Solexa method, it may be possible to sequence ChIP samples from experiments performed with low numbers of cells and provide a direct read-out of DNA sequence enrichment in ChIP samples. In the Solexa method, millions of individual DNA molecules in a ChIP sample can be linked using adapters to the surface of a glass cell and then amplified to create a cluster of DNA molecules with identical tags. The DNA molecules in each cluster are then simultaneously sequenced using the sequencing by synthesis chemistry which can be used to assess abundance of a particular DNA sequence in a ChIP sample (Barski *et al.*, 2007; Mikkelsen *et al.*, 2007). This sequencing method requires only nanogram quantities of starting material and could be used to provide whole-genome epigenetic maps for limited cell populations.

### **5.13. Conclusions**

The work described in this Chapter showed how empirically-tested modifications to the conventional ChIP-chip protocol can be used to reproducibly detect histone modifications in reduced numbers of cells. Optimal antibody concentrations for ChIP assays established here across a range of cell numbers provide a working range of concentrations that may be applicable for the study of different cell types. These optimisations may be used for the study of histone modifications in cell types which previously would not have been amenable to ChIP-chip analysis. To this end, the following Chapter describes the application of these modifications to study several histone modifications in human primary monocytes and human embryonic stem cells to understand chromatin state during differentiation.

## Chapter 6

### **Histone modification profiles of human embryonic stem cells and lineage-committed human monocytes**

#### **6.1. Introduction**

One of the key unanswered questions in biology is the basis of cellular state. Although each human cell contains an identical genome sequence, many types of cells exist each with different properties and functions. The development of these cell types involves a progressive specialisation pathway which begins with the totipotent cells of the human embryo. These cells can give rise to all cell types including the extra embryonic tissues required for foetal development. Totipotent cells then specialise into pluripotent embryonic stem cells which can give rise to any of the differentiated cells of the body. Pluripotent cells then undergo further specialisation into multipotent cells that are committed to a particular cellular lineage. For example, multipotent haematopoietic stem cells can give rise to several specialised cell types which include erythrocytes, leukocytes and thrombocytes (Bhatia, 2007). Evidence suggests that different stages of development may be associated with distinct ‘chromatin states’, i.e. with distinct histone modification and other epigenetic events (Surani *et al.*, 2007). Therefore it is important to construct histone modification maps at different stages of cell commitment if we are to understand the basis of chromatin state during differentiation.

Embryonic stem (ES) cells are derived from the inner cell mass of a developing blastocyst and can be grown indefinitely on a tissue culture dish while still retaining the ability to differentiate into all cell types (Bradley, 1990). The derivation of human embryonic stem cells (hESCs) that can be grown in culture provides an exciting opportunity to study the early stages of human development (Thomson *et al.*, 1998; Reubinoff *et al.*, 2000) and there has been great interest in understanding how chromatin regulates the pluripotency of these cells. A number of recent studies have investigated chromatin structure in human and mouse embryonic stem cells using ChIP in combination with quantitative PCR, microarrays and high-throughput sequencing (Azura *et al.*, 2006; Bernstein *et al.*, 2006; Guenther *et al.*, 2007; Lee *et al.*, 2006; Boyer

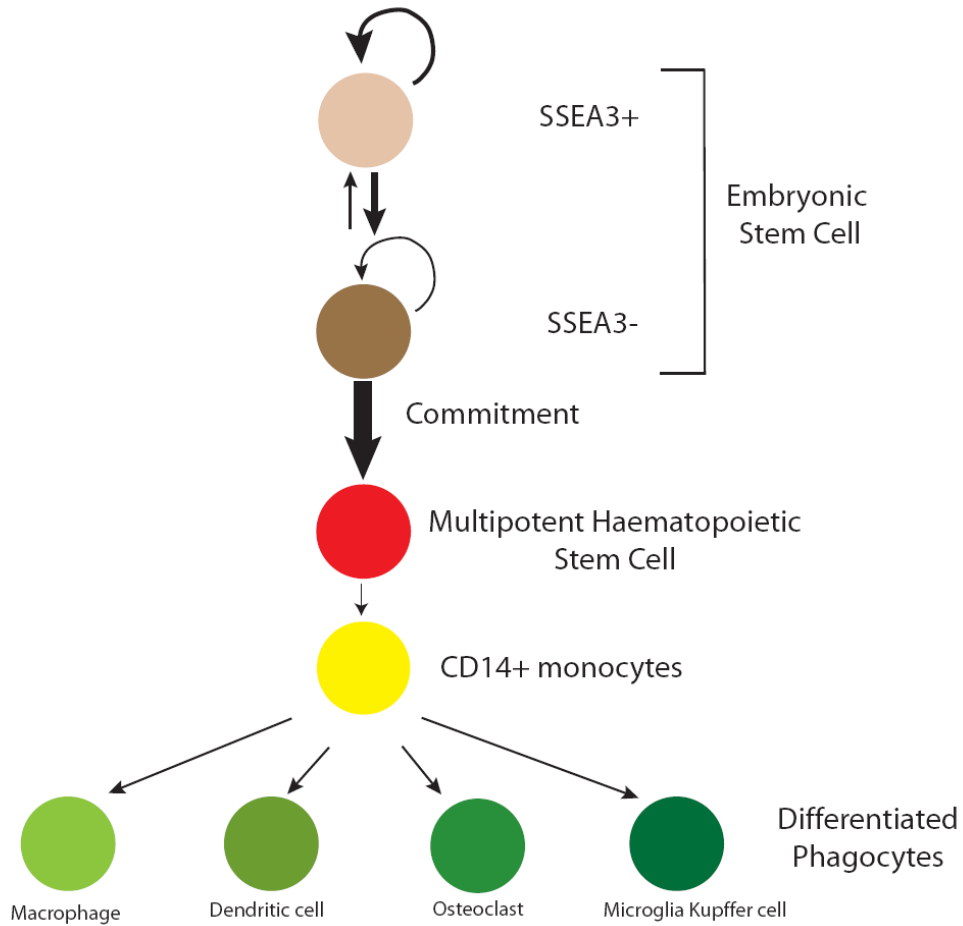
*et al.*, 2006; Mikkelsen *et al.*, 2007; Zhao *et al.*, 2007; Pan *et al.*, 2007). Several of these studies have suggested that histone H3K4me3 and histone H3K27me3 play a particularly important role in regulating pluripotency of human and mouse embryonic stem cells (Azuara *et al.*, 2006; Bernstein *et al.*, 2006; Mikkelsen *et al.*, 2007; Zhao *et al.*, 2007; Pan *et al.*, 2007). H3K4me3 and H3K27me3 are catalysed by trithorax group (TrxG) and Polycomb group (PcG) proteins respectively which have key developmental functions (Ringrose and Paro, 2004). H3K4me3 is associated with 'open' chromatin and active transcription (Sims *et al.*, 2005; Wysocka *et al.*, 2005) while H3K27me3 is associated with compact chromatin and gene repression (Ringrose and Paro, 2004). Azuara (2006) and Bernstein (2006) first proposed that the idea key developmental transcription factors were marked by large expanses of the repressive H3K27me3 modification while at the same time the promoters of these genes contained the H3K4me3 modification. These regions were termed 'bivalent' domains and were proposed to maintain the pluripotent state of embryonic stem cells as H3K27me3 repressed developmental transcription factor expression while the continued presence of H3K4me3 allowed for these genes to be rapidly up-regulated when required during differentiation. Two recent studies showed that H3K27me3 recruits Polycomb repressive complex to transcriptionally repress key developmental regulators to maintain pluripotency in both mouse and human embryonic stem cells (Boyer *et al.*, 2006; Lee *et al.*, 2006). Furthermore, a whole genome ChIP and high-throughput sequencing study of histone modifications in pluripotent mouse embryonic stem cells and lineage committed cells determined that H3K4me3 and H3K27me3 levels discriminated genes that were expressed, poised for expression, or stably repressed during various stages of lineage commitment (Mikkelsen *et al.*, 2007).

In the previous Chapter, improvements to the ChIP-chip procedure was developed for the accurate detection of histone modification enrichments from as few as  $10^4$  K562 cells. This protocol was developed with the aim of applying it to study chromatin state during human lineage commitment, as availability of primary human cells is often a limiting factor when performing ChIP-chip analyses. hESCs and human CD14<sup>+</sup> monocytes were chosen for the study of chromatin state during lineage commitment as hESCs are pluripotent while monocytes represent a multipotent lineage committed progenitor capable of differentiating into several types of phagocytic cells (Seta and Kuwana, 2007)

(Figure 6.1). Studies of cultures of hESCs are regularly hampered because they often contain mixed cell populations of both undifferentiated stem cells and the spontaneously arising differentiated derivatives. This heterogeneity can be addressed by sorting of hESC cultures according to the expression of cell surface markers. Several cell surface antigens have been proposed as markers of undifferentiated hESCs such as the glycolipid antigens stage specific embryonic antigen 3 (SSEA3) and SSEA4 and the keratin sulphate-associated antigens TRA-1-60, TRA-1-81 and GCTM2 (Thomson *et al.*, 1998; Draper JS *et al.*, 2002). Studies of the expression patterns of these antigens in hESCs suggest that SSEA3 in particular might represent a sensitive marker of the most primitive state for hESCs (Reubinoff *et al.*, 2002; Draper *et al.*, 2002; Enver *et al.*, 2005). Thus SSEA3 expression (SSEA3+) is believed to mark pluripotent stem cells and those negative for SSEA3 expression (SSEA3-), are not as primitive as SSEA3+ cells but still retained multilineage differentiation potential (Enver *et al.*, 2005). Obtaining SSEA3+ and SSEA3- sorted hESCs is difficult and often only  $2-3 \times 10^6$  cells can be obtained.

In contrast, monocytes represent a highly specialised and committed myeloid cell type which are formed in the bone marrow and are continuously released into the blood where they circulate for several days before migrating into most tissues, where they mature and differentiate into specialised macrophages (Friedman, 2007). Obtaining donor human blood samples for purification of circulating monocytes is relatively difficult and typically  $2-5 \times 10^7$  cells are obtained from each donor sample which limits the number of ChIP-chip experiments that can be performed using standard protocols.





**Figure 6.1: The relationship between SSEA3+ and SSEA3- H9 hESCs and CD14+ monocyte cells.** SSEA3+ cells represent undifferentiated embryonic stem cells which can self-renew indefinitely in culture. SSEA3+ cells can be induced to spontaneously differentiate in culture and the first cell surface marker to be lost is SSEA3. SSEA3- cells represent a form of ‘differentiated’ embryonic stem cells which have a higher probability of proceeding to a commitment step and subsequently differentiating than reverting to SSEA3+ state. Once SSEA3- cells have proceeded to a commitment step they can then differentiate into various multipotent progenitors which in turn give rise to the numerous terminally differentiated cells found in our bodies. For example SSEA3- cells can, via differentiation into haematopoietic stem cells, give rise to CD14+ circulating monocytes which in turn differentiate into four types of phagocytic cells- macrophages, dendritic cells, osteoclasts and microglia kupffer cells (Seta and Kuwana, 2007).

In Chapter 3, work was described which showed the relationship between gene expression, regulatory function and histone modifications in a cultured erythroleukaemic cell line, K562. This Chapter reports the application of the modified ChIP-chip method (Chapter 5) to further understand these relationships in the *in vitro* developmental

processes associated with hESC differentiation and *in vivo* in monocytes, a primary cell type.

## **6.2. Aims of this Chapter**

Having developed a modified ChIP-chip method for the detection of histone modifications patterns from as few as  $10^4$  K562 cells (Chapter 5), an important goal of this study was to apply this method to study chromatin regulation in human cell types, in which cell numbers limited the number of conventional ChIP-chip assays that could be performed. To this end the aims of the work presented in this chapter were:

1. To apply the modified ChIP-chip method to study a range of histone modifications in hESCs and human CD14+ monocytes across the ENCODE regions. This would allow for the dynamics of chromatin regulation to be studied in cells types representing different stages of cellular differentiation and in the case of monocytes-primary cells.
2. To investigate the presence of a histone code in human primary cells by performing a detailed analysis of 19 histone modifications in human monocytes across the ENCODE regions using the modified ChIP-chip method.

## **Results**

### **6.3. Creating chromatin maps in pluripotent and lineage committed cells**

The number of SSEA3+ and SSEA3- hESCs available (obtained from Dr. Enrique Milan, Cambridge Institute of Medical Research) for study limited the number of ChIP-chip assays that could be performed even with the modified ChIP-chip method. Thus for this project, the focus was on the investigation of four key histone methylation modifications (H3K4me3, H3K27me3, H3K36me3, and H3K79me3). H3K4me3 is associated with 5' ends of active genes and is known to recruit nucleosome remodeling factors which facilitate transcription (Santos-Rosa *et al.*, 2002; Li *et al.*, 2006). In contrast, H3K27me3 is a repressive modification that is recognized by the Polycomb repressive complex 1 (PRC1), which then induces the appropriate changes in chromatin structure (Tolhuis *et al.*, 2006). H3K36me3 has been proposed to be required for efficient elongation of RNA Polymerase II through coding regions (Krogan *et al.*, 2003; Li *et al.*, 2003) and H3K79me3 is also associated with the transcribed region of active genes in yeast

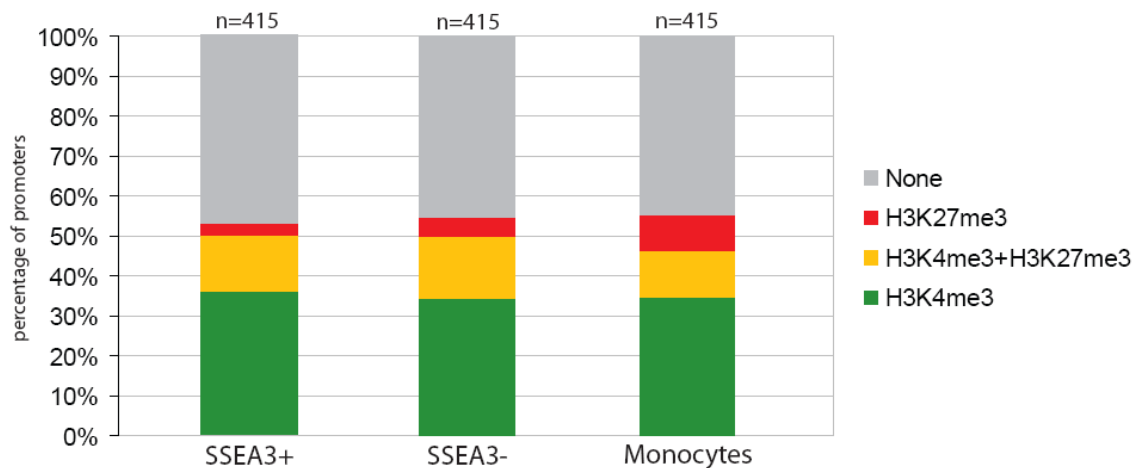
(Pokholok *et al.*, 2005). The modified ChIP-chip method would be used to detailed maps of these histone modifications in undifferentiated H9 human embryonic stem cells (SSEA3+), differentiated H9 human embryonic stem cells (SSEA3-), and human CD14+ monocytes across the ENCODE regions. This would allow for a direct comparison of chromatin state in uncommitted and lineage committed human cells to be performed. An additional 15 histone modifications were also examined in human monocytes and are discussed in section 6.5. Three biological replicates were performed for each monocyte experiment and median Cy5/Cy3 ratio values were used for subsequent analysis as described in Chapter 3. However, only one hESC biological replicate was performed for each histone modification due to limited availability of material.

#### **6.4. Analysis of chromatin state in pluripotent hESCs and lineage committed monocytes**

##### **6.4.1. Promoter chromatin state in hESCs and monocytes**

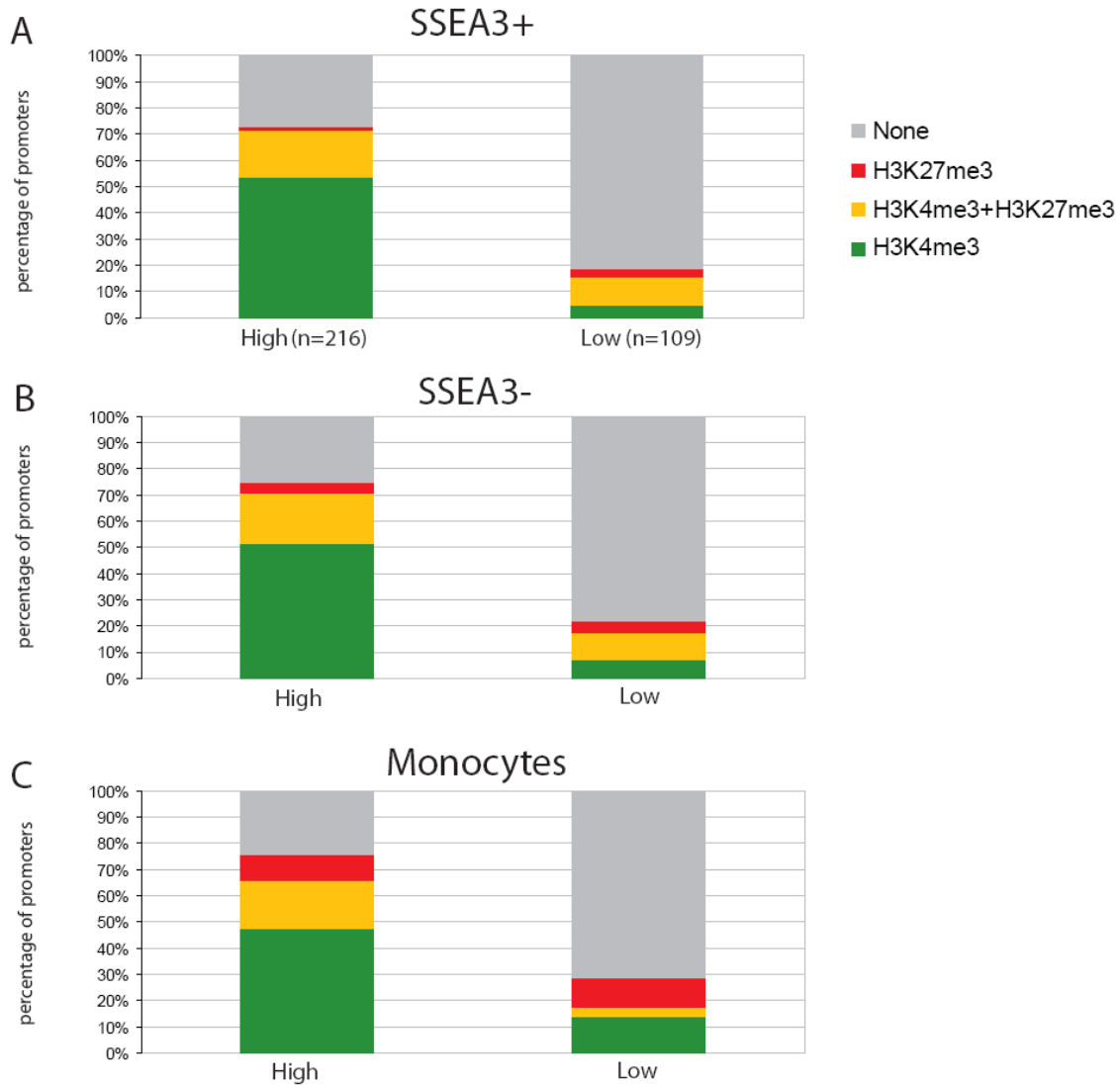
Initial analysis of chromatin state during lineage commitment focused on the study of the histone modifications implicated in the formation of bivalent chromatin domains - H3K4me3 and H3K27me3. It has been suggested that the presence of these bivalent modifications at promoters poises developmental genes for future lineage-specific activation or repression (Bernstein *et al.*, 2006; Azuara *et al.*, 2006). H3K4me3 and H3K27me3 enrichments were examined at the promoters of 415 RefSeq genes (Pruitt *et al.*, 2007) in the ENCODE regions for all three cell types to determine the relative proportions of each of the monovalent and bivalent states for H3K4me3 and H3K27me3. A bivalent promoter was defined as a region in which H3K4me3 and H3K27me3 co-localised in a 5 kb window centred on TSSs (i.e. +/- 2.5 kb from TSSs). Approximately 45% of promoters in these cell types were associated with no enrichment for either H3K4me3 or H3K27me3, whilst the remaining set of promoter regions (+/- 2.5 kb from TSSs) were associated with either the monovalent H3K4me3, the monovalent H3K27me3 or the bivalent state (Figure 6.2). This analysis revealed that approximately 35% of promoters in SSEA3+ hESCs, SSEA3- hESCs and monocytes had monovalent H3K4me3 enrichment. The number of bivalent promoters was greater in hESCs compared to monocytes; 14 and 15% of SSEA3+ and SSEA3- promoters were classified

as bivalent respectively while only 11% of monocyte promoters were bivalent. Furthermore, there was a more striking difference between hESCs and monocytes, with 12 and 38 promoters being H3K27me3 monovalent in SSEA3+ cells and monocytes respectively. This suggests that as cells differentiate from a pluripotent state to a lineage-committed state, the presence of the bivalent state is diminished and the monovalent H3K27me3 state is more common at promoter regions. This is consistent with SSEA3+ cells representing the undifferentiated cell type, while SSEA3- cells represent the early stage of hESC differentiation into multipotent progenitors and circulating CD14+ monocytes represent a specialized progenitor cell with the ability to differentiate into at least four types of phagocytic cells- macrophages, dendritic cells, osteoclasts, and microglia kupffer cells (Seta and Kuwana, 2007). Furthermore, this suggests that lineage commitment is accompanied by a requirement for H3K27me3 gene silencing mediated by recruitment of Polycomb-group proteins (Schuettengruber *et al.*, 2007) to H3K37me3 promoters.



**Figure 6.2: Chromatin modification patterns of promoters in human embryonic stem cells and monocyte cells.** The promoters of 415 well defined RefSeq genes (Pruitt *et al.*, 2007) were examined for H3K4me3 and H3K27me3 modification patterns in SSEA3+ hESCs, SSEA3- hESCs, and CD14+ monocytes. The percentage of promoters which were found to significantly enriched for H3K4me3 alone (green), H3K27me3 alone (red) and both H3K4me3 and H3K27me3 (orange) are indicated. In addition, promoters which were not enriched for H3K4me3 or H3K27me3 are indicated (grey).

H3K4me3 and H3K27me3 modification patterns were also examined in relation to promoter CpG content. Promoters with high CpG content are often associated with both 'housekeeping' genes and genes with complex expression patterns, while promoters with low CpG content are often associated with tissue-specific genes (Saxonov *et al.*, 2006; Weber *et al.*, 2007). Thus, one could expect that as cells become more differentiated, there is a striking difference in the characteristics of HCPs and LCPs with respect to H3K4me3 and H3K27me3 levels. Three hundred and twenty five (325) ENCODE promoters were classified as HCPs or LCPs based on the system of Saxonov and colleagues (Saxonov *et al.*, 2006), 216 of which were HCPs and 109 were LCPs. The presence of H3K4me3 and H3K27me3 was examined at HCPs and LCPs in the three cell types studied here (Figure 6.3). There was a clear distinction at the chromatin level between HCPs and LCPs as the majority of HCPs were associated with H3K4me3 in all three cell types and a minority of LCPs were associated with this modification; 72% of HCPs in SSEA3+ cells were associated with H3K4me3 (monovalent or bivalent) while 71% of HCPs in SSEA3- cells and 66% of HCPs in monocytes were associated with H3K4me3. In contrast only 15-17% of LCPs were associated with H3K4me3 in the three cell types. Mikkelsen and colleagues recently observed that 99% of high CpG content promoters (HCPs) are associated with H3K4me3 in mouse ES cells while less than 10% of low CpG promoters (LCPs) were associated with this modification (Mikkelsen *et al.*, 2007). The data reported here is consistent with, although not as striking, as that reported by Mikkelsen and colleagues.

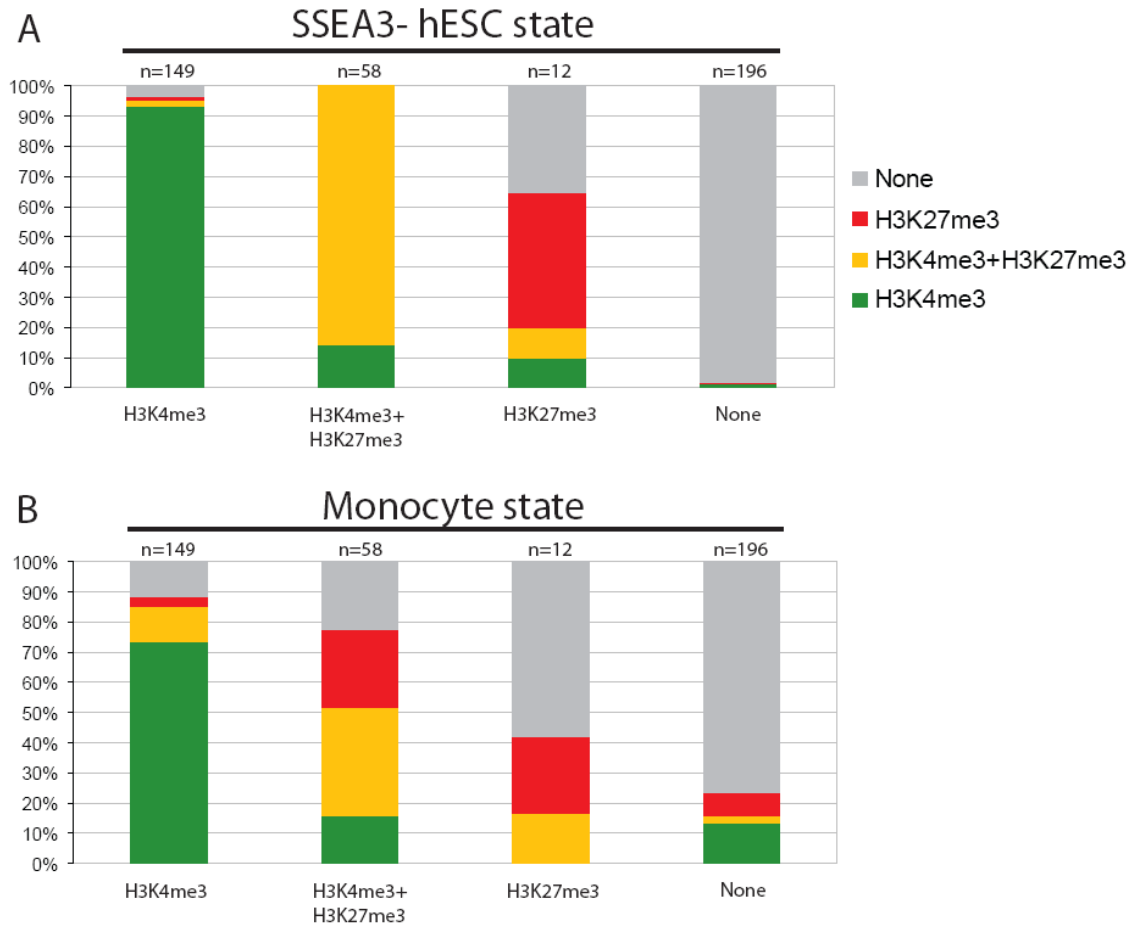


**Figure 6.3: Distinct histone modification profiles at high and low CpG content promoters.** Panels A - C illustrates the chromatin modification pattern of promoters classified as high CpG (n=216) or low CpG content promoters (n=109). The percentage of promoters which were found to significantly enriched for H3K4me3 alone (green), H3K27me3 alone (red) and both H3K4me3 and H3K27me3 (orange) are indicated for SSEA3+ hESCs, SSEA3- hESCs, and CD14+ monocytes. Promoters which were not enriched for H3K4me3 or H3K27me3 are indicated (grey).

#### 6.4.2. The chromatin state of promoters reflects developmental potential

The changes in patterns of H3K4me3 and H3K27me3 modifications at promoters were further examined in hESCs and monocytes to determine whether chromatin state reflects cellular developmental potential. Promoters for all 415 Refseq genes found in the ENCODE regions were previously scored for their H3K4me3 and H3K27me3 histone

modification profiles in SSEA3+ hESCs (Figure 6.2). These promoters were then examined in SSEA3- hESCs and monocytes to identify those promoters which had changed in their histone modification characteristics (Figure 6.4).



**Figure 6.4: Promoter chromatin state and developmental potential.** The chromatin state of 415 promoters had previously been determined for SSEA3+ hESCs resulting in the classification of promoters in this cell type into four groups – monovalent H3K4me3, monovalent H3K27me3, bivalent H3K4me3+H3K27me3 and promoters associated with neither modification. The chromatin state of these four promoter groups was then examined in SSEA3- hESCs (panel A) and CD14+ monocytes (panel B) to establish if promoter state reflects developmental potential. Promoter chromatin state in SSEA3+ hESCs is indicated on the x-axis. The percentage of promoters associated with a particular histone modification or modifications in SSEA3- hESCs and monocytes is indicated by the scale on the y-axis.

As expected, this analysis showed that the chromatin state of promoters in SSEA3+ hESCs was more similar to SSEA3- hESCs than monocytes. The data is summarized below:

- (i) **SSEA3+ monovalent promoters:** Over 90% of promoters found to be H3K4me3 monovalent in SSEA3+ hESCS remained H3K4me3 monovalent in SSEA3- hESCS while only 73% of these promoters were still monovalent H3K4me3 in monocytes. Many of those genes found to be monovalent H3K4me3 in monocytes are involved in roles in the immune system. For example seven members of the leukocyte immunoglobulin (Ig)-like receptors (LILRs) family - LILRA2, LILRA3, LILRA4, LILRA5, LILRB2, LILRB3, and LILRB4 – were associated with monovalent H3K4me3 promoters.
- (ii) **SSEA3+ bivalent promoters:** The vast majority (86%) of bivalent SSEA3+ bivalent promoters remained bivalent in SSEA3- cells and 14% had resolved to monovalent H3K4me3 status. In contrast, only 35% of SSEA3+ bivalent promoters were still bivalent in monocytes, while 16% had resolved to monovalent H3K4me3 status, 27% had resolved to H3K27me3 monovalent status and 22% were no longer associated with either modification. Thus, the chromatin state of bivalent promoters in SSEA3+ cells had changed profoundly in lineage committed monocytes and the majority of these genes were resolved to a state of monovalency or were no longer associated with histone modifications.
- (iii) **Promoters with no histone modifications in SSEA3+:** (iii) Nearly all (99%) of those promoters associated with no H3K4me3 or H3K27me3 in SSEA3+ hESCs were also associated with no modification in SSEA3- cells. However, the chromatin state of these promoters differed in monocytes and a proportion displayed monovalency (12% and 9% respectively).

These data are in general agreement with the known developmental characteristics of the three cell types studied here. SSEA3- cells represent the earliest stage of embryonic stem cell differentiation, but at the same time these cells have not yet committed to a differentiation pathway and can in a small proportion of cases revert to SSEA3+ phenotype (Enver *et al.*, 2005). However subtle changes in chromatin state were already



visible even at this early stage of hESC differentiation. When SSEA3+ cells were compared with CD14+ monocytes, large differences were observed between the cell types. This may be a reflection of the developmental state of CD14+ monocytes as many of the chromatin changes at promoters were associated with activation of genes involved in monocyte functions whilst other genes which may play a role in the development of other cell types become associated with repressive H3K27me3.

### 6.4.3. Bivalent promoters are associated with developmental genes in hESCs and monocytes

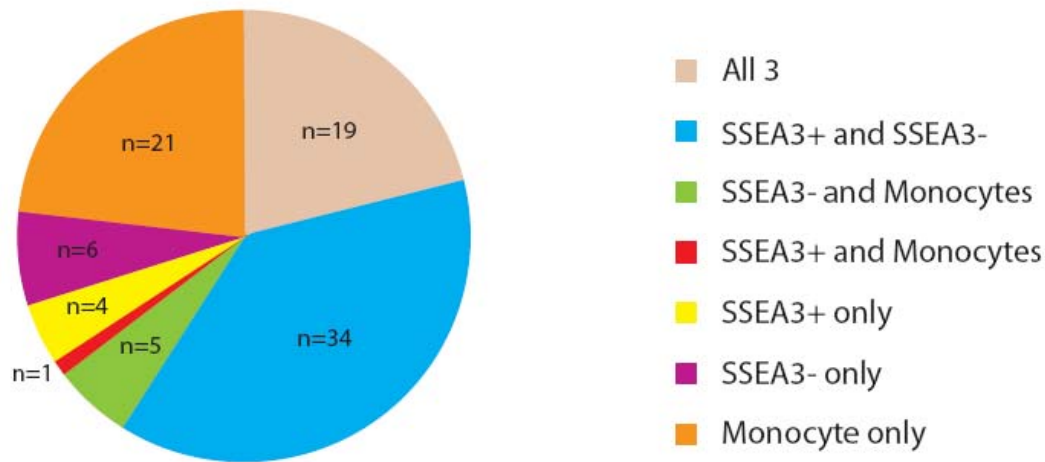
Bivalent chromatin structures have been associated with genes encoding developmental transcription factors in mouse and human embryonic stem cells (Bernstein *et al.*, 2006; Azuara *et al.*, 2006; Mikkelsen *et al.*, 2007; Zhao *et al.*, 2007; Pan *et al.*, 2007). In the present study, ENCODE genes associated with bivalent promoters were analysed in terms of functional gene ontology categories using GOToolBox (Martin *et al.*, 2004). GOToolBox determined over-represented gene ontology categories in the three bivalent gene sets (Table 6.1). Genes involved in developmental processes and transcriptional regulation were over-represented in all three cell types.

Cell type	GO term	No. in ref.	Freq. in ref.	No. in set	Freq. in set	P value
SSEA3+	development	2036	0.0938	14	0.3333	1.34E-05
	regulation of biological process	4698	0.2165	21	0.5	3.48E-05
	regulation of transcription, DNA-dependent	2835	0.1306	15	0.3571	0.000122
	transcription, DNA-dependent	2911	0.1341	15	0.3571	0.0001628
	regulation of transcription	3059	0.1409	15	0.3571	0.0002772
SSEA3-	development	2036	0.0938	17	0.3864	1.56E-07
	transcription, DNA-dependent	2911	0.1341	17	0.3864	2.02E-05
	regulation of biological process	4698	0.2165	22	0.5	2.30E-05
	regulation of transcription	3059	0.1409	17	0.3864	3.81E-05
	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	3094	0.1426	17	0.3864	4.39E-05
Monocytes	development	2036	0.0938	14	0.4667	1.19E-07
	regulation of cellular process	4377	0.2017	16	0.5333	4.59E-05
	transcription, DNA-dependent	2911	0.1341	13	0.4333	4.64E-05
	regulation of transcription	3059	0.1409	13	0.4333	7.74E-05

regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	3094	0.1426	13	0.4333	8.70E-05
---	------	--------	----	--------	----------

**Table 6.1: Gene ontology analysis of genes associated with bivalent promoters.** GOToolBox (Martin *et al.*, 2004) was used to analyse 58, 64, and 46 bivalent genes from SSEA3+ hESCs, SSEA3- hESCs, and monocytes respectively. 42, 44 and 30 of these genes were associated with a GO term annotation in the biological process category. The top five statistically over-represented GO terms are listed for each bivalent gene set. The no. in ref column is the total number of genes in the human genome associated with a particular GO term, which is then compared to the total number of genes in the human genome to give a frequency of a GO-term in the human genome (freq. in ref). The no. in set column refers to the number of bivalent genes associated with a GO term which is compared to the total number of bivalent genes to give a frequency in set value. A p-value <0.01 was considered significant.

The extent of overlap between the bivalent gene lists was then examined. 58, 64 and 46 bivalent promoters had been previously identified in SSEA3+ hESCs, SSEA3- hESCs, and monocytes respectively (Figure 6.4) and comparisons between genes associated with bivalent promoters showed that there was a large amount of overlap in the three cell types (Figure 6.5). A non-redundant list of 90 bivalent genes were identified across the three cell types of which 21% (n=19) were bivalent in all three cell types. This group included genes such as FZD-1 which encodes frizzled-1 receptor involved in WNT signaling, several HOXA genes, MECP2 which encodes methyl-CpG-binding protein 2, and OLIG2 which encodes a transcription factor required for oligodendrocyte and motor neuron specification in the spinal cord. A substantial fraction of genes (38%, n=34) were classified as bivalent only in SSEA3+ hESCs and SSEA3- hESCs and included genes such as CFTR which encodes cystic fibrosis transmembrane conductance regulator, EVX-1 which encodes Homeobox even-skipped homolog protein 1 involved in neuronal specification, and various other HOXA genes. SSEA3+ only or SSEA3- only bivalent genes represented a small fraction of the total number of bivalent genes identified (4 and 7% respectively), while the percentage of monocyte-specific bivalent genes was much greater (23%, n=21). Monocyte-specific bivalent genes included CADH2 which encodes Cadherin-2, a calcium dependent cell adhesion protein, CTGF encoding connective tissue growth factor which is involved in wound-response, and LAIR1/2 which encode Leukocyte-associated immunoglobulin-like receptors 1 and 2.



**Figure 6.5: Conserved and cell-type specific bivalent promoters.** 90 gene promoters were associated with a bivalent chromatin structure in SSEA3+ hESCS, SSEA3- hESCS and monocytes. This pie-chart illustrates that many of these promoters were bivalent in more than one cell type while 19 (21%) of promoters were bivalent in all three cell types.

#### 6.4.4. The Polycomb Repressive Complex 2 (PRC2) subunit SUZ12 co-localises with bivalent developmental genes in human embryonic stem cells

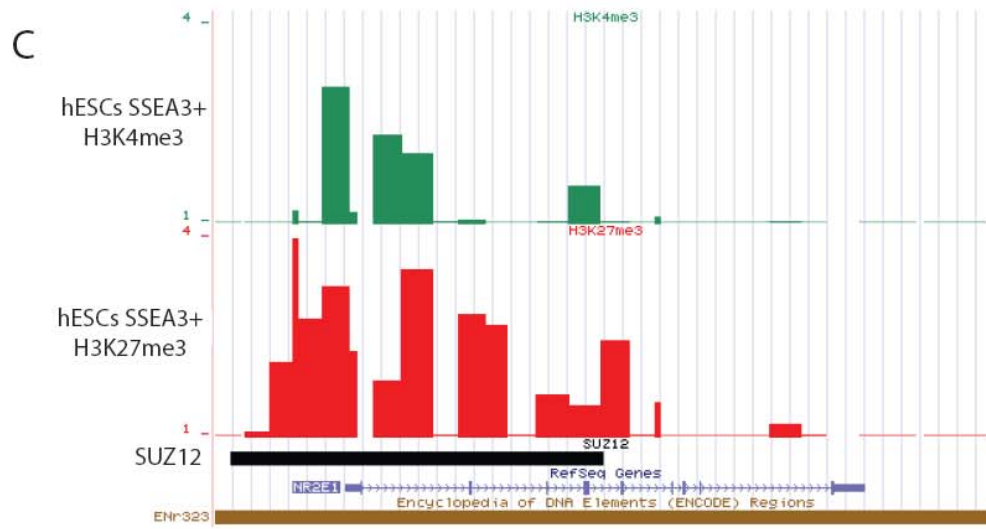
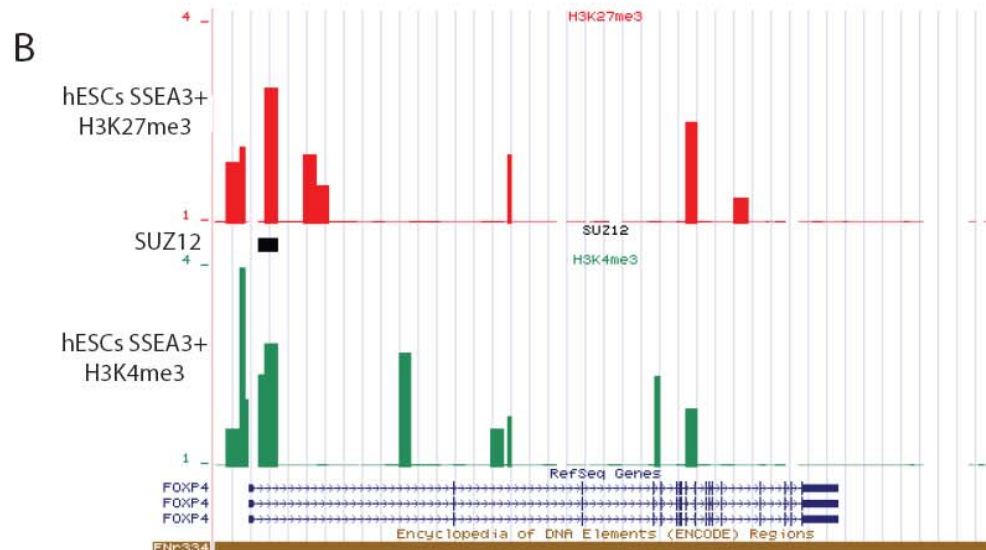
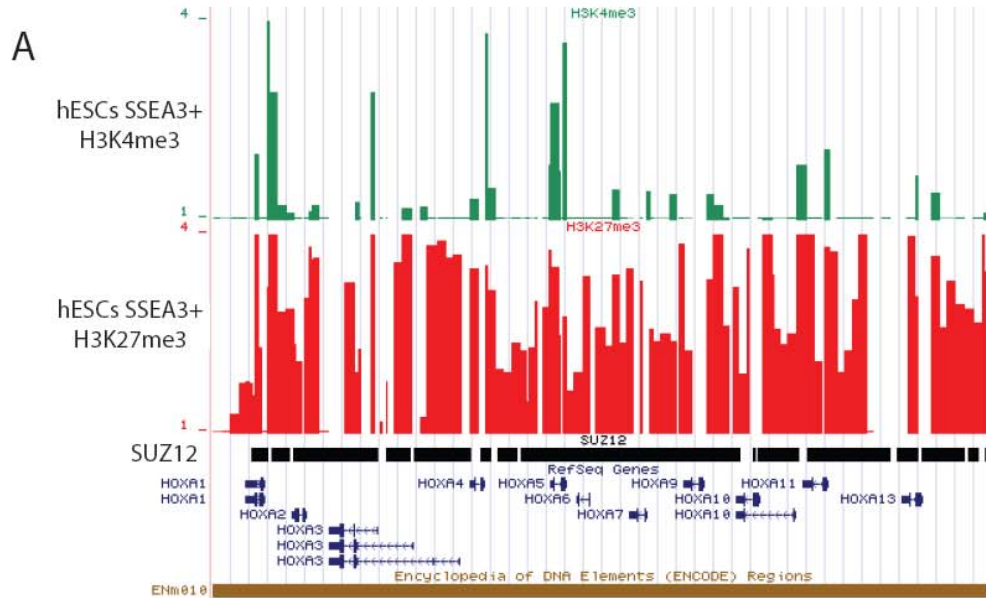
The Polycomb group (PcG) proteins are important regulators of early development and were first identified in *Drosophila* where they repressed the homeotic genes which controlled segmentation (Lewis, 1978; Denell and Frederick, 1983). PcG proteins form multiple Polycomb Repressive Complexes (PRCs) (Birve *et al.*, 2001; Cao *et al.*, 2002; Kuzmichev *et al.*, 2002), which are recruited to sites of gene repression to modify chromatin (Levine *et al.*, 2004, Ringrose and Paro, 2004). PRC2 contains EED, EZH2 and SUZ12 components (Kirmizis *et al.*, 2004; Kuzmichev *et al.*, 2005) and EZH2 functions as a histone H3K27 methyltransferase (Cao *et al.*, 2002; Czermin *et al.*, 2002; Kirmizis *et al.*, 2004). Lee and colleagues recently mapped the genome-wide binding locations of the SUZ12 subunit in H9 hESCS and found that it binds to approximately 1900 promoters, over 200 of which are associated with key developmental regulators (Lee *et al.*, 2006). As many developmental regulators were associated with H3K27me3 in hESCs, it was hypothesized that PRC2 (SUZ12) may also be associated with these genes. The data of Lee and colleagues was used to identify 71 SUZ12 binding sites in the

ENCODE regions. The binding of SUZ12 from that study was compared with the location of H3K27me3 (and H3K4me3) in SSEA3+ H9 hESCs. The majority (52 of 71) of SUZ12 binding sites co-localised with regions of H3K27me3 enrichment, 23 of which were located at bivalent genes (Table 6.2).

Gene	CpG status	Function
WNT2	high	Regulation of cell fate and patterning during embryogenesis
SEPT8	N/A	Septin protein involved in organization of sub-membrane structures
TIMP3	high	Complexes with metalloproteinases (such as collagenases) and irreversibly inactivates them. May form part of a tissue- specific acute response to remodeling stimuli
OLIG2	high	Required for oligodendrocyte and motor neuron specification in the spinal cord, as well as for the development of somatic motor neurons in the hindbrain
HBM	N/A	Haemoglobin mu subunit expressed during erythroblast terminal differentiation
HBA1, HBA2	High	The human alpha-globin cluster contains alpha-1 (HBA1) and alpha-2 (HBA2) genes which form the haemoglobin subunit alpha
HOXA cluster	All high Except HOXA3	Encodes several DNA-binding transcription factors which may regulate gene expression, morphogenesis, and differentiation.
EVX1	High	The encoded protein may play an important role as a transcriptional repressor during embryogenesis.
GRM8	Low	Metabotropic glutamate receptor 8 precursor
NR2E1	High	Acts as transcriptional repressor to maintain neural stem cells in an undifferentiated state
NRXN2	High	Neuronal cell surface protein that may be involved in cell recognition and cell adhesion
FOXP4	High	Transcriptional repressor that represses lung-specific expression

**Table 6.2: SUZ12 is associated with bivalent genes in human embryonic stem cells.** 23 genes associated with bivalent promoters in SSEA3+ H9 hESCS are also associated with SUZ12 binding. The gene names are listed along with a brief description of the encoded protein's function. The CpG content of these promoters is also indicated based on the classification of Saxonov and colleagues (2006). N/A= promoter classification data not available.

The HOXA cluster encodes a family of transcription factors that play a key role in the establishment of cellular identity during embryogenesis (Pearson *et al.*, 2005) and the entire cluster (Figure 6.6) was found to be associated with a large block of H3K27me3 in SSEA3+ and a large domain of SUZ12 binding. Furthermore, a number of the promoters in this block were also associated with H3K4me3. This is consistent with previous reports in which PRC1 and PRC2 were found to be responsible for the repression of developmental regulators including HOX genes in mouse embryonic stem cells (Akasaka *et al.*, 2001; Wang *et al.*, 2002; Cao and Zhang, 2004; Boyer *et al.*, 2006). The vast majority of developmental genes associated with a bivalent chromatin structure and SUZ12 binding had promoters with high CpG content. In summary this study suggests that the combination of bivalent chromatin structure and PRC2 binding may be responsible for the repression of key developmental genes in hESCs, and this mechanism may be used to control the expression of developmental factors in other cell types as bivalent promoters are also a feature of monocyte cells.

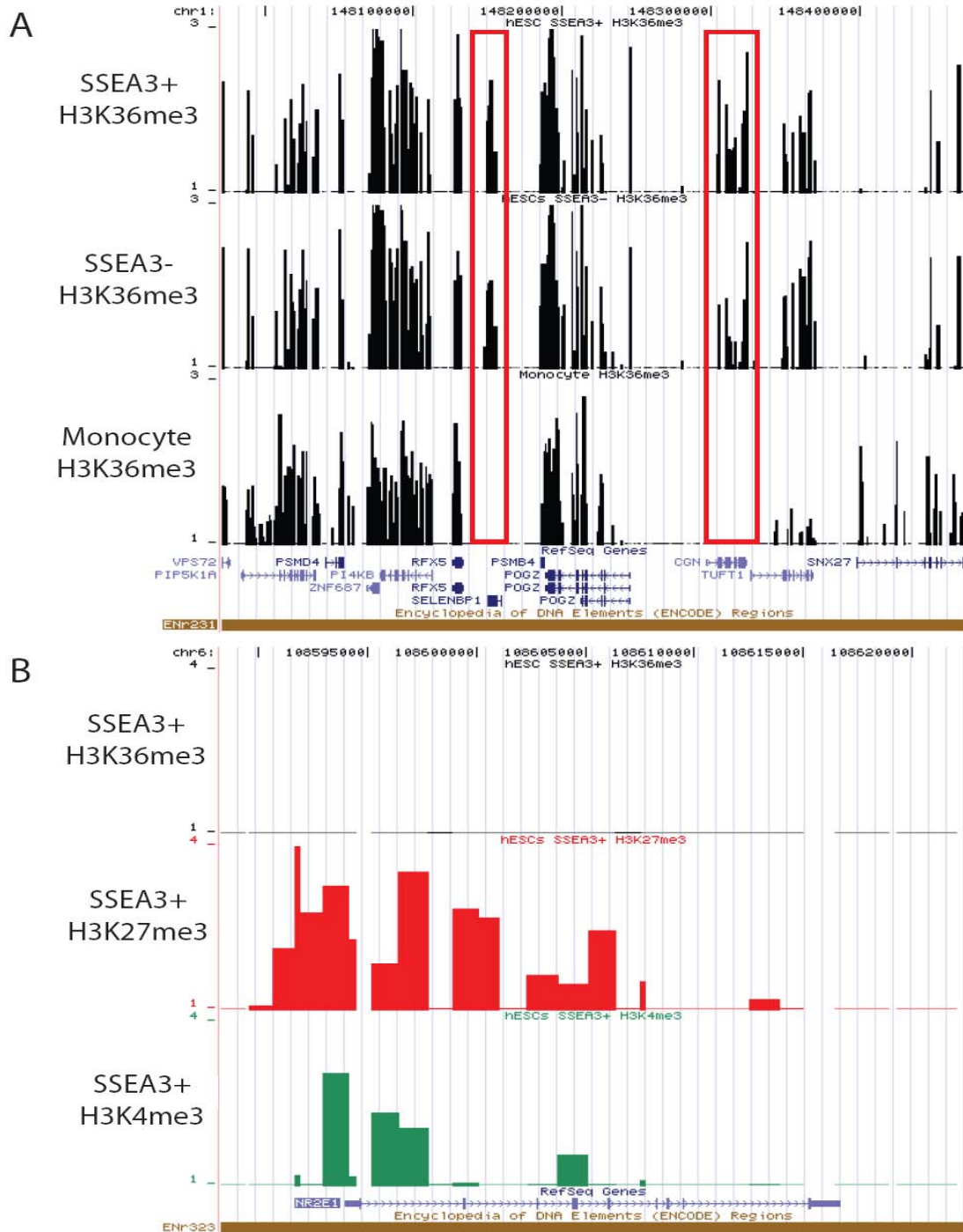


**Figure 6.6: Representative examples of genes associated with a bivalent promoter structure and SUZ12 binding in human embryonic stem cells.** Three screenshots from the UCSC genome browser (Kuhn *et al.*, 2007) showing ChIP-chip data for H3K4me3, H3K27me3, and SUZ12 binding in hESCs. Panel A: The entire HOXA gene cluster is associated with H3K27me3 (red bars), while the promoter regions are associated with H3K4me3 (green bars). The large ‘block’ of H3K27me3 enrichment is coupled with a large domain of SUZ12 genomic binding (black horizontal bar) (Lee *et al.*, 2006) across this 125 kb region. Panel B: H3K27me3 can also be associated with discrete genomic regions as observed at the FOXP4 gene. H3K4me3 is also enriched at the 5’ end of FOXP4 along with SUZ12. In this case the SUZ12 binding domain is small as it covers only 2kb. Panel C: A bivalent chromatin state is observed at the NR2E1 gene and this is associated with a 17kb binding site for SUZ12, which begins upstream of the NR2E1 gene and extends for approximately 11kb into the gene. The scale in base pairs is indicated at the top of the figure. The bottom track shows the Refseq genes (Pruitt *et al.*, 2007) in blue with transcriptional orientation indicated by arrows. The H3K4me3 and H3K27me3 ChIP-chip data is displayed in the intervening tracks as the median value of the ratio of ChIP-chip sample fluorescence to input DNA fluorescence. Each green or red vertical bar is the enrichment measured at a single array element on the ENCODE microarray with the enrichment represented by the height of the bar. Note that fold enrichments in the ChIP samples are displayed as Log<sub>2</sub> values for each track and are scaled 1-4.

#### **6.4.5. Histone H3K36 trimethylation is associated with primary protein-coding transcripts and non-coding RNAs in hESCs and monocytes**

Histone H3K36 trimethylation has been associated with transcriptional elongation in mammalian cells (Bannister *et al.*, 2005; Vakoc *et al.*, 2006; Mikkelsen *et al.*, 2007) and may function to prevent aberrant transcription initiation at cryptic TSSs within genes by blocking the acetylation of H3K36, which is linked to promoters of transcribed genes (Li *et al.*, 2007 c; Morris *et al.*, 2007). Histone H3K36me3 chromatin maps in SSEA3+ hESCs, SSEA3- hESCs and monocytes revealed that H3K36me3 is highly enriched across genes (Figure 6.7). Two replicate human CD14+ monocyte illumina human expression beadchip data sets were provided by Dr. Nick Watkins (Department of Haematology, University of Cambridge) and were used for transcript abundance analysis as described in Chapter 2. Expression levels of ENCODE genes were extracted from these datasets and were ranked in order of expression as high (100-75%), low (75%-50%), indeterminate (50%-25%), and off (25%-0%) as described in Chapters 2 and 3. Analysis showed that high levels of H3K36me3 are associated with the transcribed portion of highly expressed genes and little or no enrichment is present at lowly

expressed and at non-expressed genes (See section 6.5.2.5 for further details). Thus, the presence of H3K36me3 may be used to predict transcriptional status and define the length of primary transcripts. Genes associated with bivalent promoters display little or no H3K36me3 enrichment, consistent with their low/off expression status (Figure 6.7).





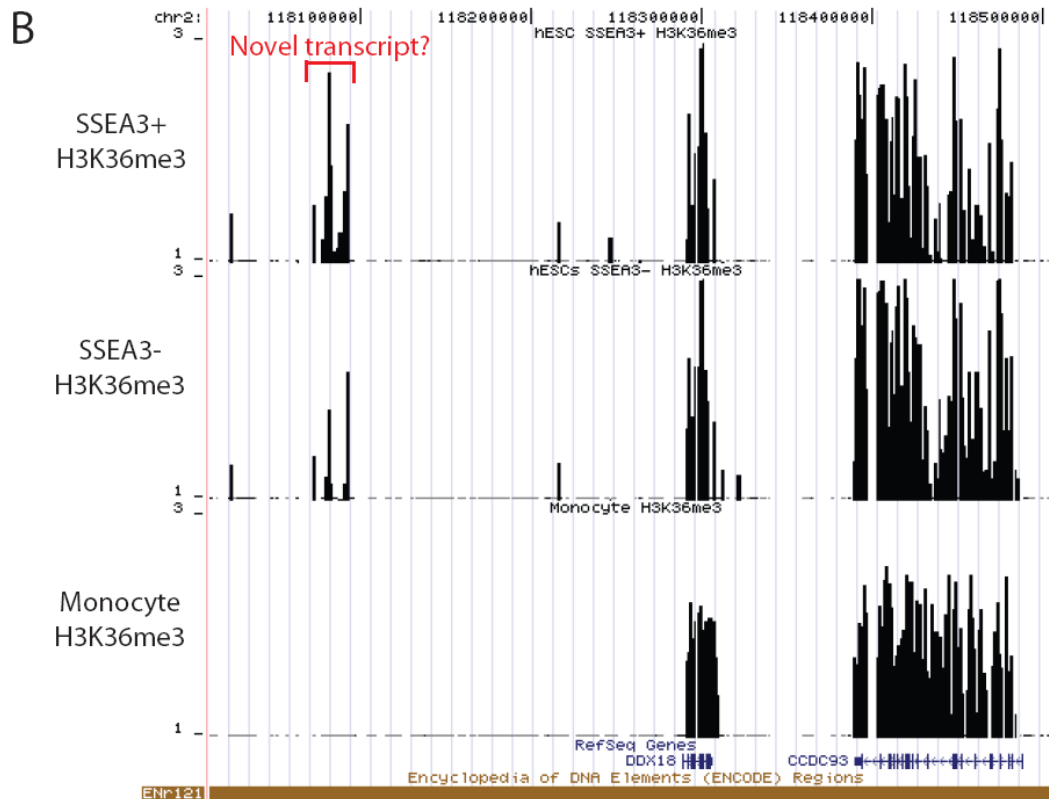
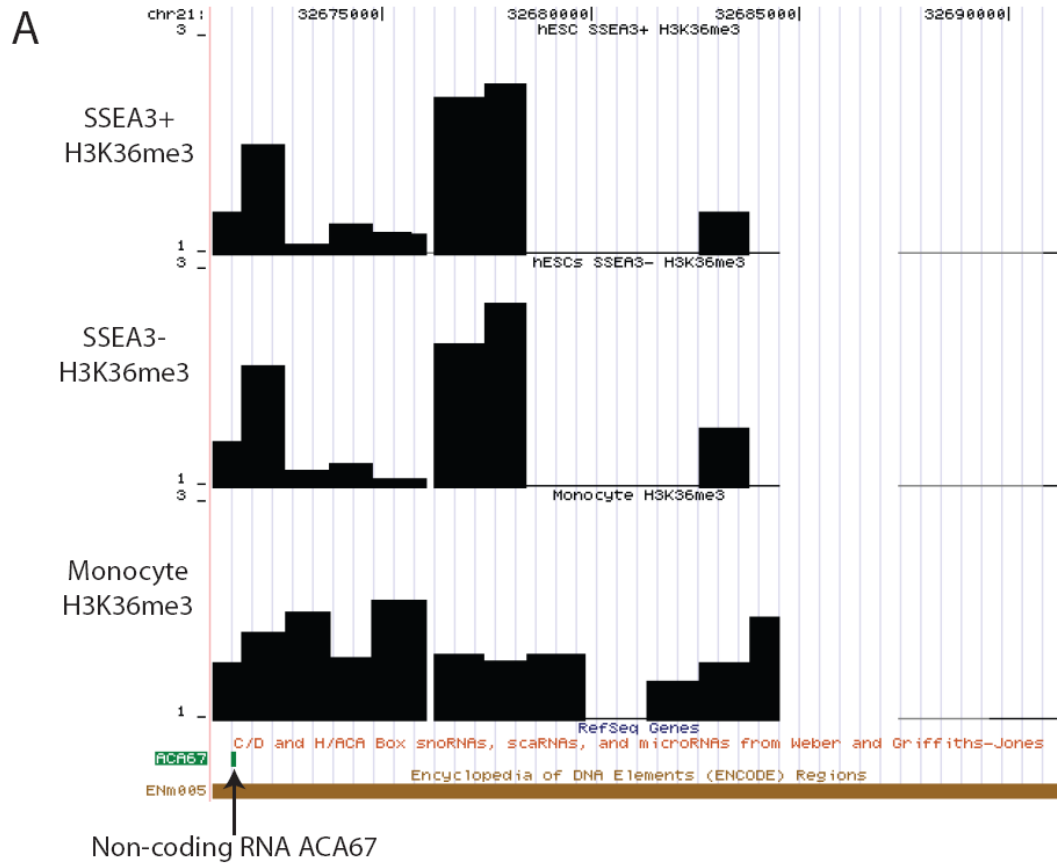
**Figure 6.7: H3K36me3 annotates transcribed genes in hESCs and monocytes.** Two screenshots from the UCSC genome browser (Kuhn *et al.*, 2007) showing the relationship between H3K36me3 and transcriptional activity. Panel A: ENCODE region Enr231 is a gene dense region in which the majority of genes are associated with H3K36me3 in all three cell types. The transcribed region of SELENBP1 and CGN genes are enriched for H3K36me3 in SSEA3+ and SSEA3- hESCs (top and middle tracks respectively) while no enrichment is observed in the bottom monocyte track (highlighted by red boxes) and SELENBP1 and CGN are classified as lowly expressed and off in monocytes respectively. Those other genes in the region associated with high levels of H3K36me3 are classified as highly expressed. Panel B: Genes with a bivalent promoter show little or no H3K36me3 enrichment and the NR2E1 gene is an example of a bivalent gene associated with no H3K36me3 enrichment in SSEA3+ hESCs (top track). The scale in base pairs is indicated at the top of the figure. The bottom track of each panel shows the Refseq genes (Pruitt *et al.*, 2007) in blue with transcriptional orientation indicated by arrows. The H3K36me3 (black bars), H3K4me3 (green bars) and H3K27me3 (red bars) ChIP-chip data is displayed in the intervening tracks. Each vertical bar is the enrichment measured at a single array element on the ENCODE microarray with the enrichment represented by the height of the bar. Note that fold enrichments in the ChIP samples are displayed as  $\text{Log}_2$  values for each track and are scaled 1-3 in panel A and 1-4 in panel B.

While the vast majority of regions enriched for H3K36me3 were located at known protein-coding genes, there were 70 examples of H3K36me3 regions (greater than 2kb in length) not associated with this type of gene. The majority (37) of these regions were detected in SSEA3+ hESCs, while 24 regions were identified in SSEA3- hESCs and 9 in monocytes. Several of these regions of H3K36me3 were associated with non-coding RNAs (Table 6.3). microRNAs from the miRNA registry and small nucleolar RNAs (C/D box and H/ACA box snoRNAs) and Cajal body-specific RNAs (scaRNAs) from snoRNA-LBME-DB (Lestrade and Weber, 2006) were downloaded from the UCSC database (Kuhn *et al.*, 2007). Eight non-coding RNAs were identified in the ENCODE regions. Six of these were associated with H3K36me3 enrichment in hESCs and two were associated with H3K36me3 in monocytes (Table 6.3).

Non-coding RNA	Genomic location	SSEA3+	SSEA3-	Monocyte
U70	chrX:153,149,469-153,149,603	Yes	Yes	No
ACA36	chrX:153,560,507-153,560,638	Yes	Yes	No
ACA56	chrX:153,566,977-153,567,105	Yes	Yes	Yes
ACA67	chr21:32,671,367-32,671,502	Yes	Yes	Yes
hsa-mir-192	chr11:64,415,185-64,415,294	Yes	Yes	No
hsa-mir-194-2	chr11:64,415,403-64,415,487	Yes	Yes	No
hsa-mir-196b	chr7:26,982,339-26,982,422	No	No	No
hsa-mir-483	chr11:2,111,940-2,112,015	No	No	No

**Table 6.3: Non-coding RNAs in the ENCODE regions are associated with H3K36me3.** Six of the eight known non-coding RNAs were associated with H3K36me3 (indicated by yes) in SSEA3+ and SSEA3-hESCs, while two were associated with H3K36me3 in monocytes. The name of the non-coding RNAs and the genomic coordinates are also presented.

While non-coding RNA ACA67 is only 136 bp long it is known that non-coding RNAs are processed from longer precursors (Mattick and Makunin, 2006). Therefore the 13 kb region of H3K36me3 enrichment overlapping with ACA67 may represent the primary transcript from which ACA67 was processed (Figure 6.8). Mapping of H3K36me3 may be useful for classifying primary transcripts which are then processed into smaller non-coding RNAs such as microRNAs. The remaining sites of H3K36me3 enrichment not located at known non-coding RNA transcripts were also often detected specifically in hESCs (Figure 6.8). Thus the mapping of H3K36me3 enriched regions may be useful for the identification of novel transcripts, many of which may be specific to hESCs.

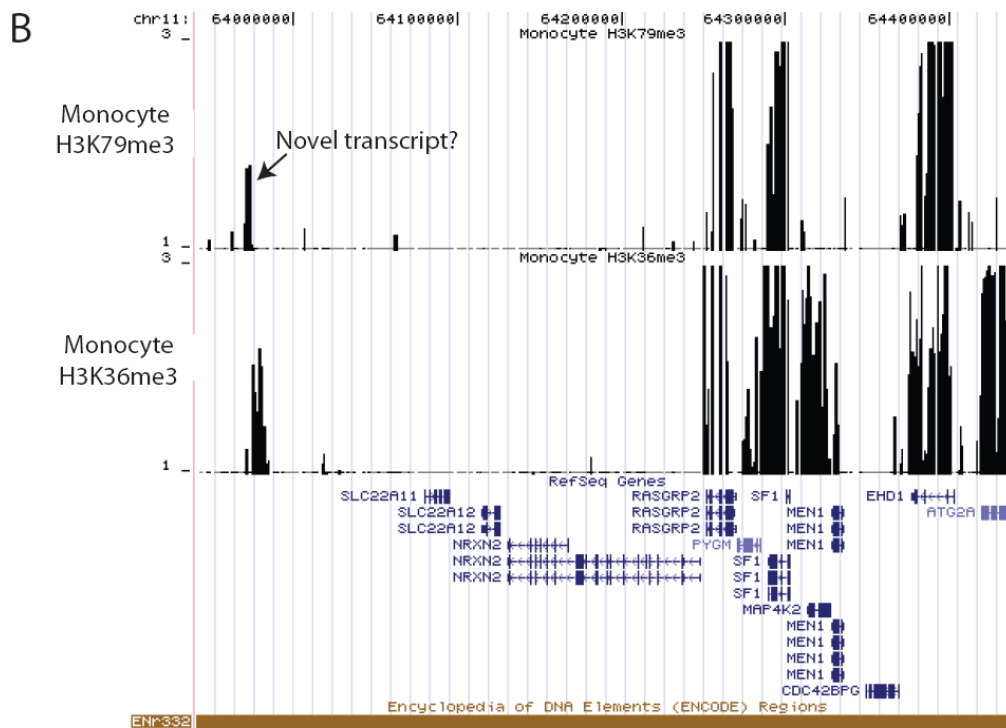
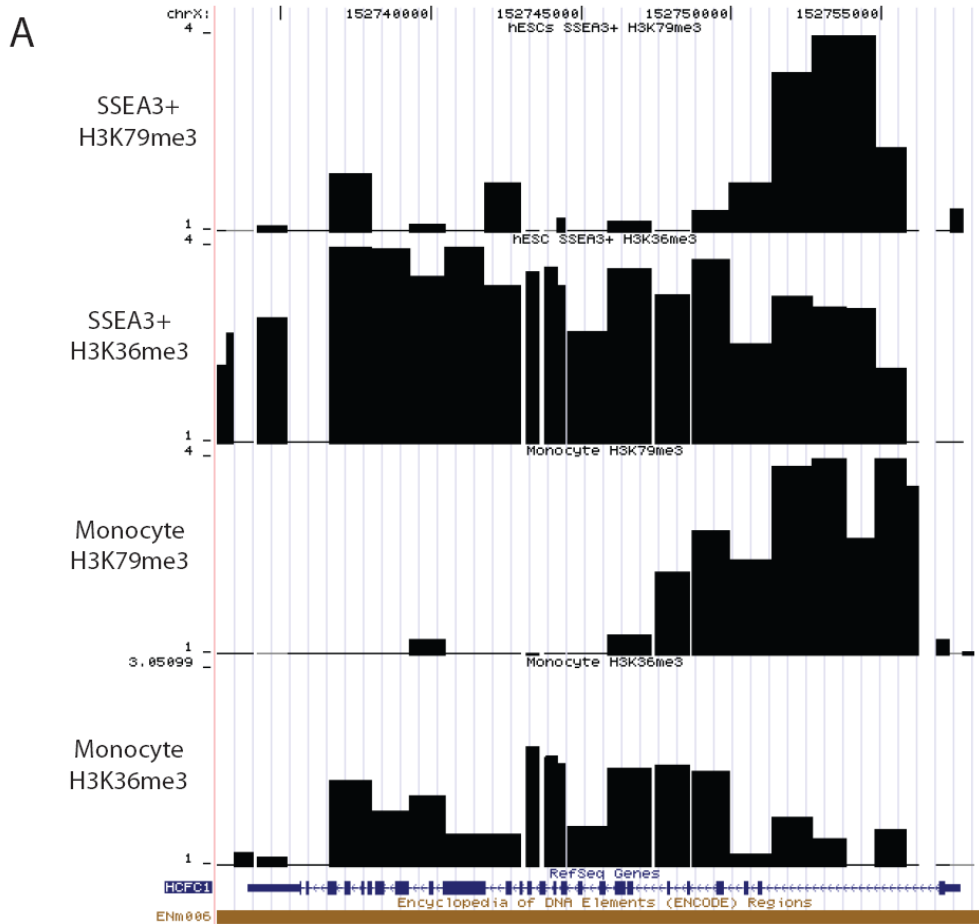


**Figure 6.8: Histone H3K36me3 is associated with non-coding RNAs and putative novel transcripts.**

Two screenshots from the UCSC genome browser (Kuhn *et al.*, 2007) showing H3K36me3 ChIP-chip data for the three cell types Panel A: non-coding RNA ACA67 (indicated by arrow) is associated with H3K36me3 in all three cell types. Panel B: H3K36me3 is enriched in a region not associated with any known transcripts (indicated by bracket). This putative novel transcript may represent a human embryonic stem cell specific transcript. The scale in base pairs is indicated at the top of the two screenshots. The bottom track shows the presence of Refseq genes (Pruitt *et al.*, 2007) with transcriptional orientation indicated by arrows. The location of non-coding RNA ACA67 is indicated in green at the bottom of panel A. The ChIP-chip data is displayed in the three intervening tracks as the median value of the ratio of ChIP-chip sample fluorescence to input DNA fluorescence. The top, middle and bottom tracks represent SSEA3+ hESCs, SSEA3- hESCs and monocytes respectively. Each black vertical bar is the enrichment measured at a single array element on the ENCODE microarray with the enrichment represented by the height of the bar. Note that fold enrichments in the ChIP samples are displayed as  $\text{Log}_2$  values for each track and are scaled 1-3.

**6.4.6. Histone H3K79 trimethylation: Possible link to transcription elongation in hESCs and monocytes**

While methylation of H3K79 has been found in transcribed regions in yeast (Pokholok *et al.*, 2005) the distribution and function of this histone modification in human cells is not well understood. H3K79me3 modification maps were created in hESCs and monocytes to determine the location of this modification relative to ENCODE genes. As was reported in Chapter 5 (section 5.10), H3K79me3 was found to be present in the early transcribed portion of active genes in hESCs and monocytes (Figure 6.9) (See section 6.5.2.6 for further details). H3K79me3, in conjunction with H3K36me3 enrichments, were detected at several regions at which no known gene was present and may represent the location of novel transcripts (Figure 6.9). Consistent with the low expression status of bivalent genes (Mikkelsen *et al.*, 2007), little or no H3K79me3 was detected at bivalent genes.



**Figure 6.9: H3K79me3 association with the immediate transcribed region of active genes in hESCs and monocytes.** Panel A: H3K36me3 is enriched across the entire length of active genes in hESCs and monocytes while H3K79me3 is enriched across the early transcribed portion of genes in hESCs and monocytes. Panel B: H3K79me3 may define the 5' location of novel transcripts (indicated by a black arrow) which is supported by the presence of H3K36me3. The scale in base pairs is indicated at the top of the two panels. The bottom track shows the presence of Refseq gene (Pruitt *et al.*, 2007) with transcriptional orientation indicated by arrows. Each black vertical bar in the four tracks represent the enrichment measured at a single array element on the ENCODE microarray. Note that fold enrichments in the ChIP samples are displayed as  $\log_2$  values for each track.

The presence of H3K79me3 in the early transcribed regions of active genes may facilitate the transition of RNA Polymerase II into productive elongation by maintaining chromatin structure in an open conformation. This is supported by the recent observation that the H3K79 methyltransferase Dot1 is recruited to elongating RNA Polymerase II (Bitoun *et al.*, 2007). In summary, data on the distribution of both H3K36me3 and H3K79me3 could be used to define the 5' and 3' boundaries of both known and novel transcripts in hESCs and monocytes.

### **6.5. A detailed analysis of histone acetylation and methylation modifications in human monocytes**

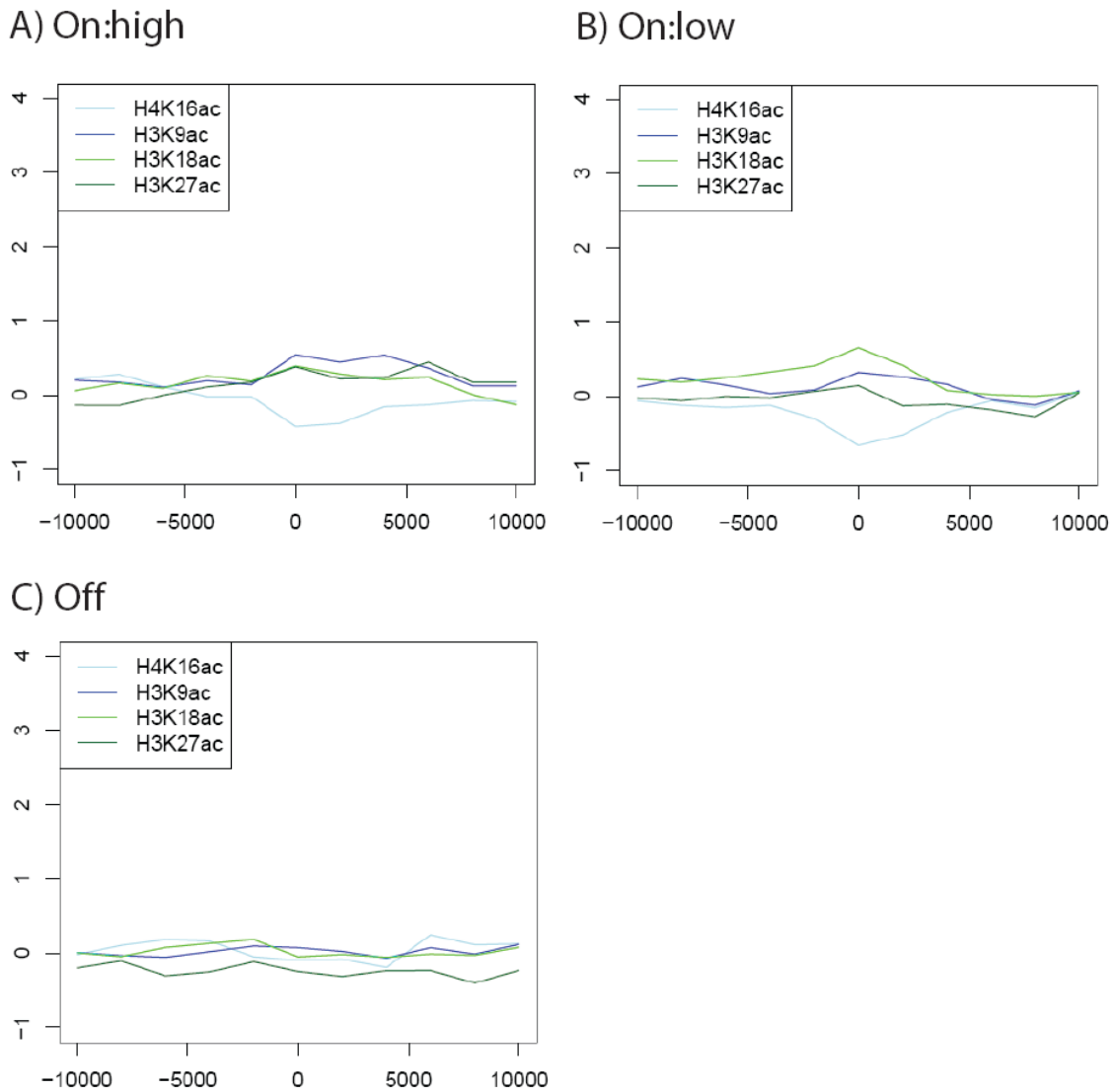
As discussed in section 6.3, the modified ChIP-chip method was used to create chromatin maps for 15 additional histone modifications in human monocytes. Methylation of various lysine residues has been implicated in different gene regulation events as discussed in Chapter 1. In addition, there is the added complexity that mono, di and trimethylation states at the same lysine residue have been implicated in different molecular processes. For example, H3K27me3 methylation has been predominantly linked to gene repression (Boyer *et al.*, 2006; Lee *et al.*, 2006; Roh *et al.*, 2006) while H3K27me1 has been observed in active coding regions (Vakoc *et al.*, 2006). The function and distribution of many of the methylation states are not well understood in human cells. Histone acetylation has been predominantly linked to active gene expression as discussed in Chapter 1, but a detailed analysis of the distribution of specific acetylation modifications in the human genome has been lacking. Therefore an analysis of the mono, di, and trimethylation states of H3K4, K9, K27, K36, K79 along with acetylation of

H3K9, H3K18, H3K27, and H4K16 was performed with human monocytes and the ENCODE microarray. This data was used to create a comprehensive map of histone modifications for a primary human cell type. A detailed analysis of this data was performed to reveal which modifications were associated with active and inactive gene expression patterns in human monocytes as well as distal enhancer/repressor elements. This analysis is described below.

### **6.5.1. Histone acetylation modifications are associated with active gene expression**

It has been shown in Chapter 3 that histone H3 and H4 acetylation at promoters correlates with the transcriptional activity of genes in K562 cells. In order to gain an insight into residue-specific acetylation events associated with gene activity in monocytes, ChIP-chip experiments were performed with the Sanger ENCODE microarray to detect acetylation events at lysine residues 9, 18, and 27 of histone H3 and lysine 16 of histone H4. The presence of these four residue-specific acetylation events was then correlated with transcriptional activity. Analysis of monocyte Illumina gene expression data was performed as described in section 6.4.5 and genes were divided into three category: high expression, low expression and no expression (“off”). 293 of the ENCODE genes were found to be expressed in monocytes while 146 were classified as “off”. Of the 293 expressed genes, 147 were highly expressed and 146 displayed low level expression. Acetylation of H3K9, K18, K27, and H4K16 was examined at the promoter and flanking DNA sequences of highly expressed, lowly expressed, and “off” genes (Figure 6.10). The average  $\log_2$  fold enrichments were plotted 10kb upstream and downstream of the TSSs of these genes. This analysis showed that three of the four acetylation modifications (H3K9ac, H3K18ac, and H3K27ac) were modestly enriched at the 5’ end of highly expressed genes, peaking at the transcription start site. In contrast H4K16 acetylation levels were depleted at promoters of highly expressed genes. Low expressed genes displayed a different acetylation modification profile as H3K18ac was the most prominent modification at these promoters, while H3K9ac was enriched to a lesser extent and H4K16ac depleted. The depletion of H4K16ac at promoters of transcribed genes is interesting as this modification is known to impair mono-nucleosome mobilization by the ACF histone chaperone (Shogren-Knaak *et al.*, 2006), suggesting that its depletion at

promoters of transcribed genes enables mono-nucleosome movement at promoters of transcribed genes. Genes classified as “off” displayed no enrichment for H3K9ac, H3K18ac, H3K27ac or H4K16ac indicating that these acetylation modifications are associated with, or noticeably depleted, expressed genes active.



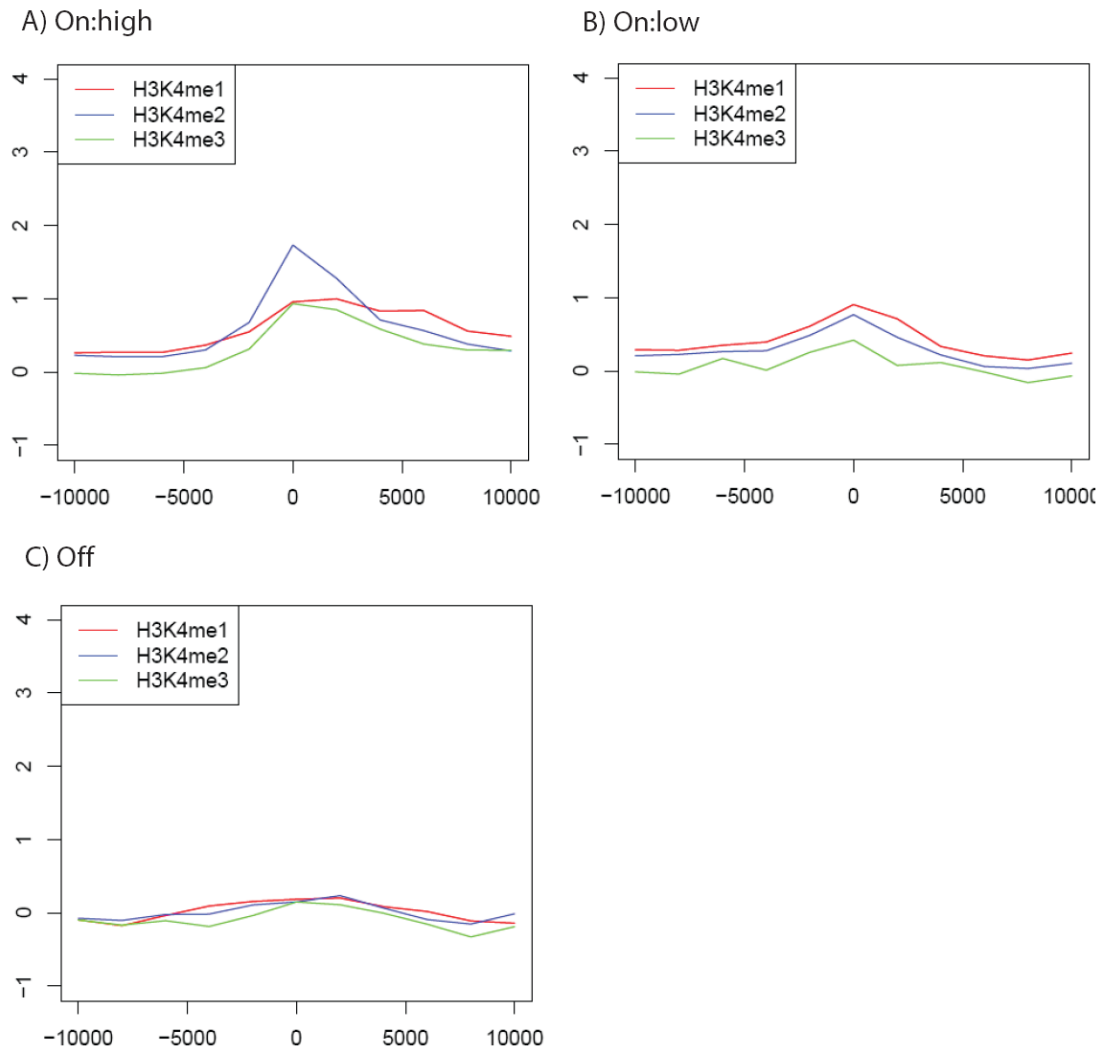
**Figure 6.10: Histone acetylation patterns at active and inactive genes.** H3K9ac, H3K18ac, H3K27ac, and H4K16ac average modification patterns are presented 10 kb upstream and downstream of TSSs associated with highly expressed (panel A), lowly expressed (panel B) and inactive (panel C) genes. The scale on the x-axis of each plot represents distances (bp’s) upstream (negative values) and downstream (positive values) of TSSs (represented by 0). Log<sub>2</sub> fold-enrichment values are presented on the y-axis



## **6.5.2. Histone H3 lysine methylation patterns in human monocytes**

### **6.5.2.1. Histone H3K4 methylation patterns at active and inactive genes**

All three states of histone H3K4 methylation were enriched at the 5' ends of highly active genes and extended 3-4kb upstream, and 4-5Kb downstream, of the TSS (Figure 6.11). H3K4me2 was most prominent at the TSSs of highly active genes, while H3K4me3 and H3K4me1 were also enriched but to a lesser extent. This detection of all three methylation states at promoters of transcribed genes is consistent with previous results obtained with K562 cells; however H3K4me3 was the most prominent modification at promoters of transcribed genes in K562 cells. Lowly expressed genes were associated with a smaller peak of enrichment for all three methylation states, with H3K4me1 displaying the highest enrichment at the promoters of lowly expressed genes. All three modification states were observed to be unenriched at promoters of “off” genes.

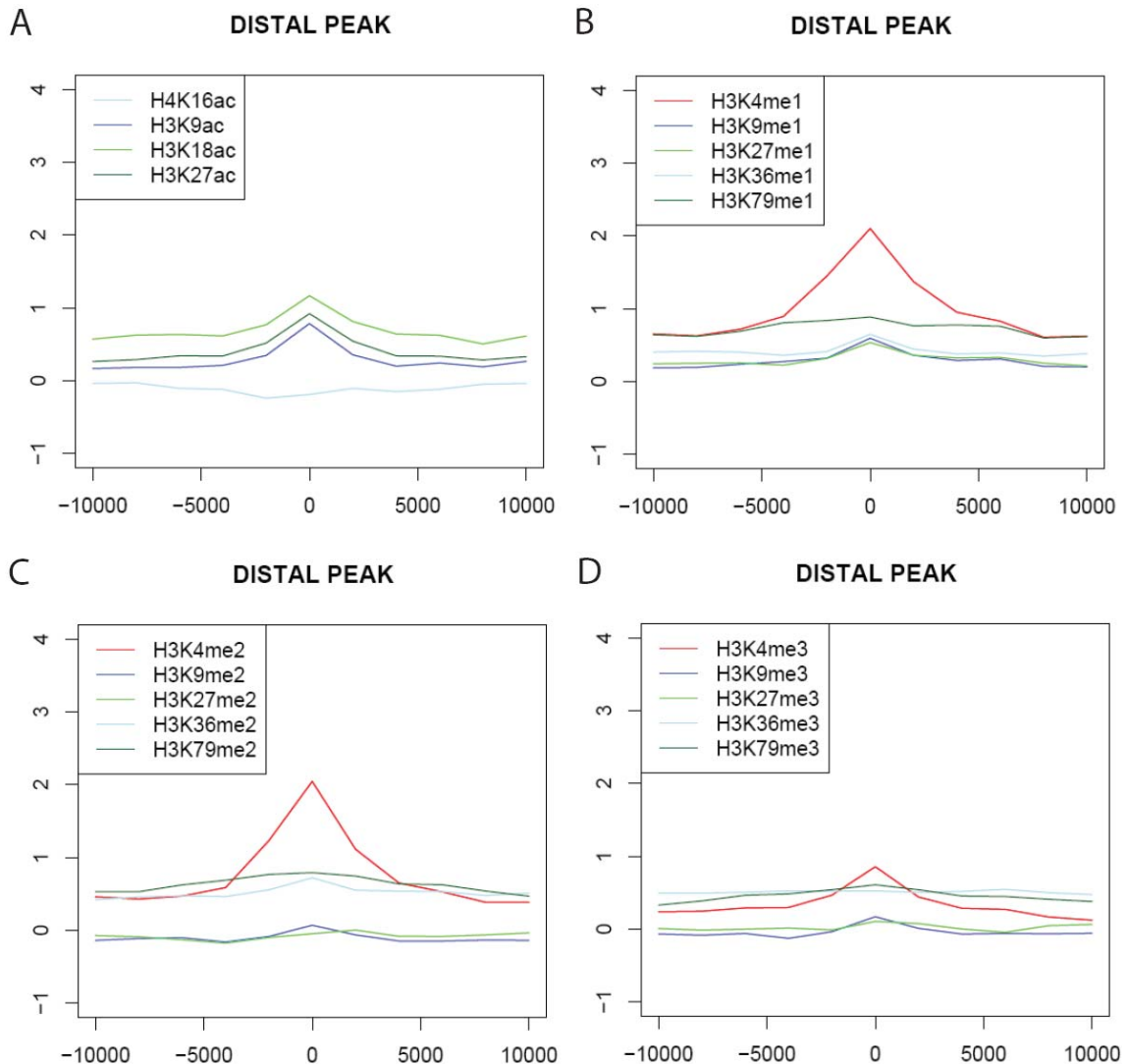


**Figure 6.11: Histone H3K4 methylation patterns at active and inactive genes.** H3K4me1, H3K4me2, and H3K4me3 average modification patterns are presented 10 kb upstream and downstream of promoters associated with highly expressed (panel A), lowly expressed (panel B) and inactive (panel C) genes. The scale on the x-axis of each plot represents distances (bp's) upstream (negative values) and downstream (positive values) of known TSSs and z-scored log<sub>2</sub> fold-enrichment values are presented on the y-axis.

### 6.5.2.2. The histone signature of distal enhancer/repressor elements

The ChIPOTle program (Buck *et al.*, 2005) was used to identify H3K4me1, H3K4me2, and H3K4me3 peaks of enrichment as described in Chapter 2. This resulted in the identification of 676 sites which contained a peak of enrichment for one or more of the three modifications. Of these, 270 peaks of H3K4me1, H3K4me2, or H3K4me3 enrichment were located within 2.5 kb of an annotated TSS while the remaining 406 sites

were classified as putative distal enhancer/repressor sites. Distal sites were predominantly associated with H3K4me1 and H3K4me2 (Figure 6.12), consistent with the hallmarks of distal sites identified in K562 cells (Chapter 3). The presence of H3K9, H3K27, H3K36 and H3K79 methylation was examined at these distal sites to identify a more detailed signature associated with distal enhancers/repressors (Figure 6.12). Distal elements were associated with peaks of enrichment for H3K9ac, H3K18ac, and H3K27ac while no enrichment for H4K16ac was observed (see section 6.5.1). Distal elements were also associated with noticeable peaks of H3K9me1, H3K27me1, H3K36me1 and H3K36me2 enrichment. The other modifications assayed were neither enriched nor depleted at distal sites.

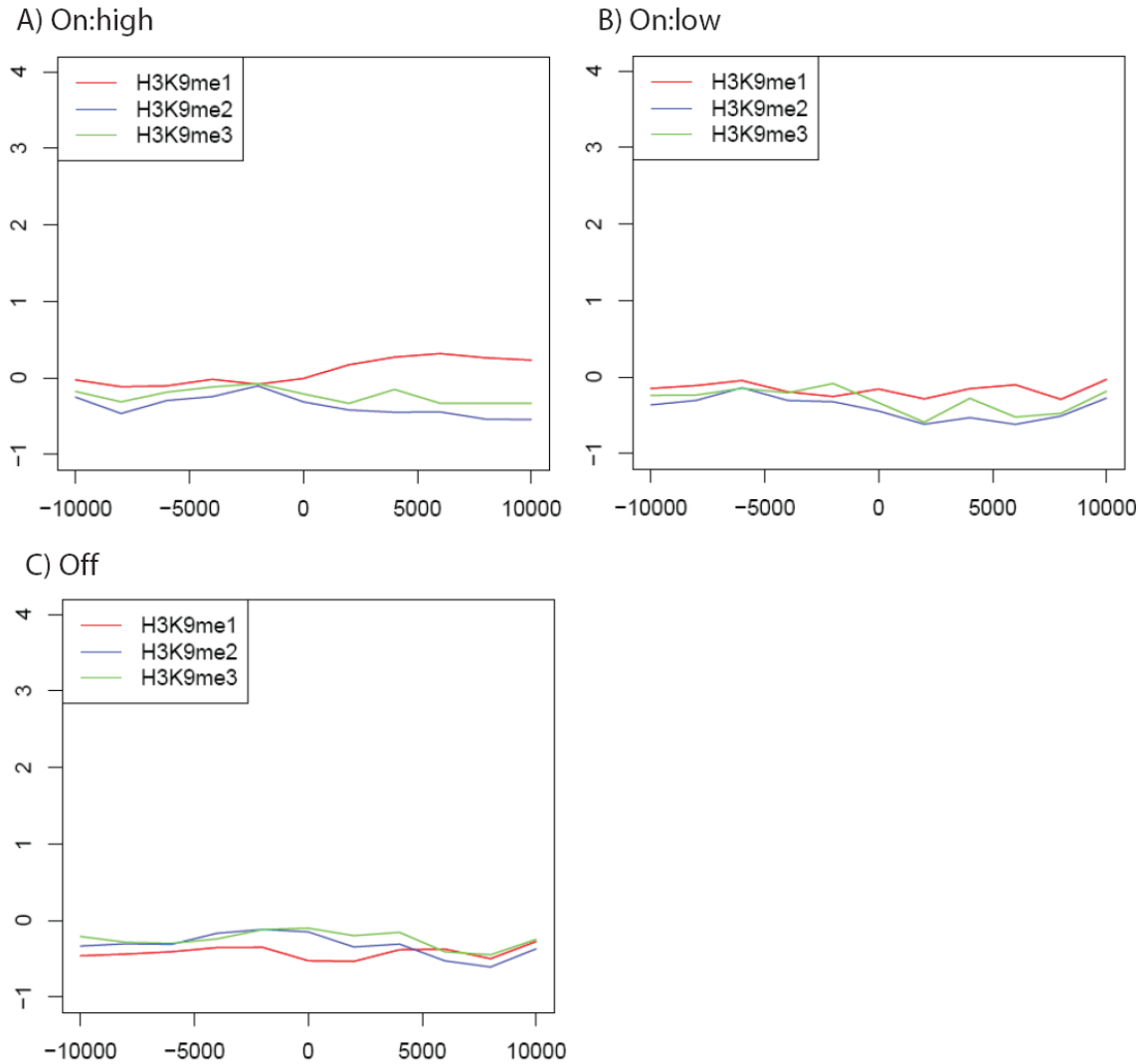


**Figure 6.12: Histone acetylation and methylation patterns at distal enhancer/repressor elements.**

Histone acetylation and methylation patterns are presented 10 kb upstream and downstream of distal elements. The average pattern of H3K9, H3K18, H3K27 and H4K16 acetylation at distal elements is presented in panel A. The average enrichments for the mono-methylation states of H3K4, H3K9, H3K27, H3K36, and H3K79 is presented in panel B while di and tri-methylation states are presented in panels C and D respectively. The scale on the x-axis of each plot represents distance (bp's) upstream (negative values) and downstream (positive values) of distal elements and z-scored  $\log_2$  values are presented on the y-axis.

**6.5.2.3. Histone H3K9 methylation is implicated at actively transcribed genes**

Given that conflicting reports regarding the role of H3K9 methylation in gene expression in human cells have been documented (as described in Chapter 1), all three methylation states were investigated for their presence or absence at expressed and not expressed “off” genes. All three H3K9 methylation states were examined at the promoter regions of these genes (Figure 6.13).



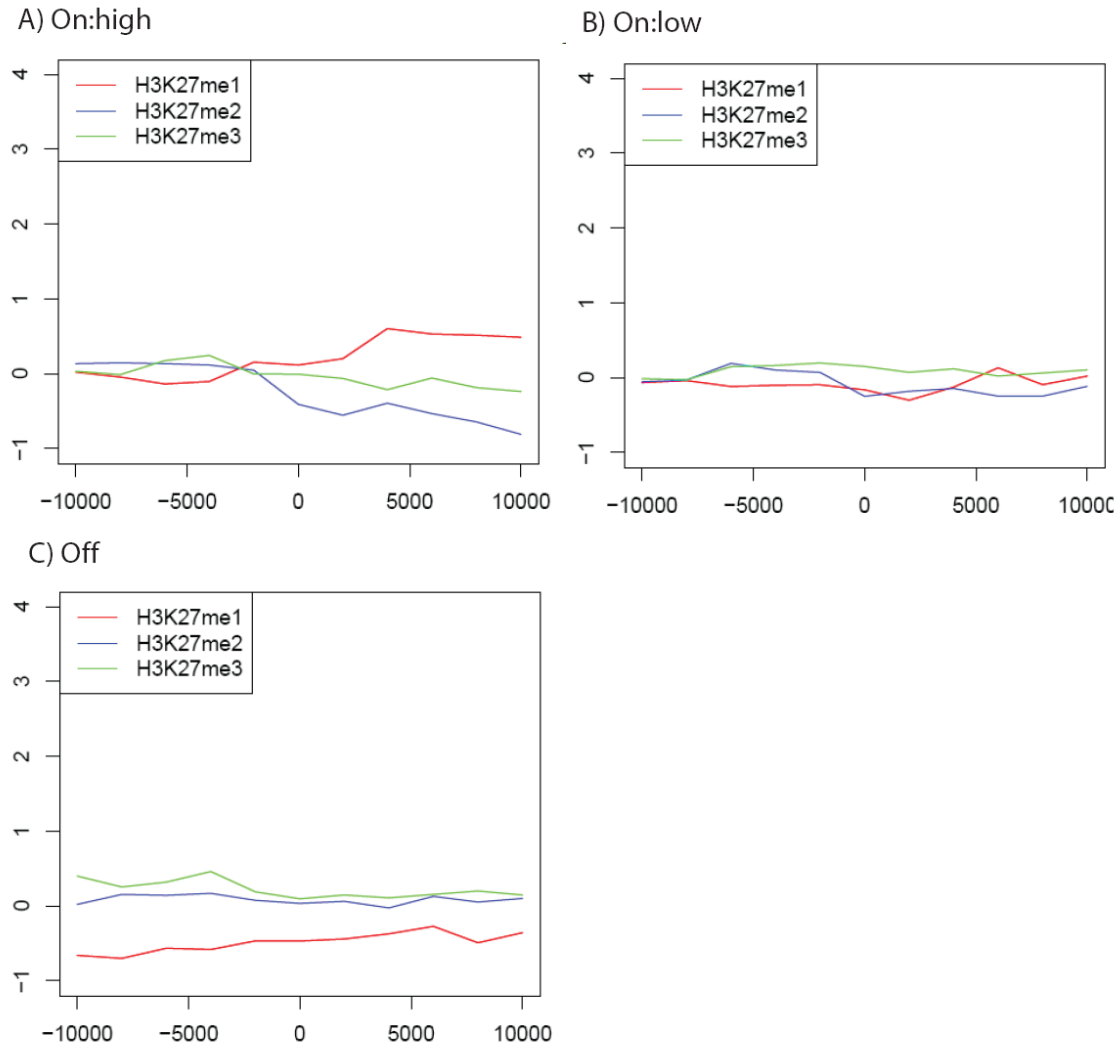
**Figure 6.13: Histone H3K9 methylation patterns at active and inactive genes.** H3K9me1, H3K9me2, and H3K9me3 average modification patterns are presented 10 kb upstream and downstream of TSSs associated with highly expressed (panel A), lowly expressed (panel B) and inactive (panel C) genes. The scale on the x-axis of each plot represents distances (bp's) upstream (negative values) and downstream (positive values) of known TSSs and z-scored log<sub>2</sub> values are presented on the y-axis

H3K9me2 and H3K9me3 levels were depleted at the promoter and transcribed portion of highly expressed genes while H3K9me1 levels increased over the transcribed region of highly expressed genes. Lowly expressed genes were also associated with low levels of H3K9me2 and H3K9me3; while H3K9me1 levels were greater than the other two modification states for these genes, there were at much lower than that associated with

highly expressed genes. This data suggests that H3K9me1 is linked to active gene expression and is consistent with a recent report which noted that elevated levels of H3K9me1 were detected surrounding the TSSs of expressed genes (Barski *et al.*, 2007). Inactive or “off” genes displayed low level enrichment for all three H3K9 methylation states; with H3K9me3 being most enriched of the three at the promoter and gene body regions. However, the average H3K9me3 enrichment observed at inactive genes was low. This may be because the H3K9me3 antibody may not work as efficiently as other antibodies or that H3K9 methylation may have gene-specific modes of action which are not detected when examining overall patterns.

#### **6.5.2.4. Histone H3K27 methylation states are found at active and inactive genes**

H3K27me1, H3K27me2, and H3K27me3 modification profiles across genes were also examined with respect to transcriptional activity (Figure 6.14). An important finding of this study was the observation that H3K27me1 levels were elevated across the transcribed portion of highly expressed genes. H3K27me2 levels were depleted at promoters of transcriptionally inactive genes and their coding regions while H3K27me3 levels were also low at active genes. Lowly expressed genes were associated with a modest increase in H3K27me1 enrichment downstream of the promoter region while low-level H3K27me3 enrichment was observed across lowly expressed genes. While H3K27me3 has been implicated in Polycomb mediated gene silencing (Cao *et al.*, 2002), “off” genes in human monocytes are associated with modest increases in H3K27me3, peaking approximately 4 kb upstream of inactive TSSs. H3K27me1 levels are very low across inactive genes suggesting that the presence of H3K27me1 may be an indicator of gene transcription.

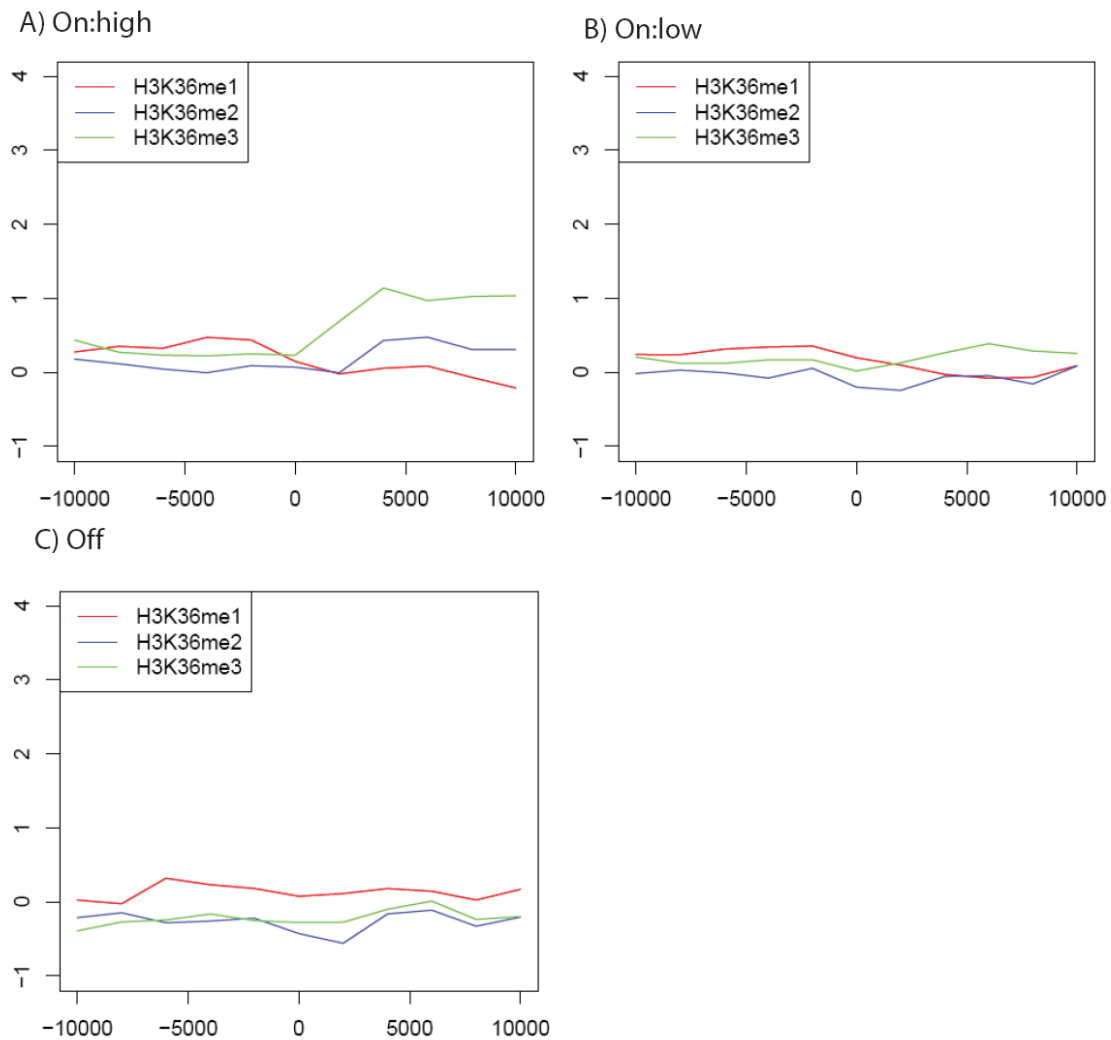


**Figure 6.14: Histone H3K27 methylation states at active and inactive genes.** Average H3K27me1, H3K27me2, and H3K27me3 modification patterns are presented 10 kb upstream and downstream of TSSs associated with highly expressed (panel A), lowly expressed (panel B) and inactive (panel C) genes. The scale on the x-axis of each plot represents distances (bp's) upstream (negative values) and downstream (positive values) of known TSSs and z-scored log<sub>2</sub> values are presented on the y-axis.

### 6.5.2.5. Analysis of histone H3K36 methylation states

As described in section 6.4.5, H3K36me3 was found to be associated with actively transcribed regions in monocytes. Histone H3K36me3 levels were found to increase sharply just downstream of the TSS of highly expressed genes and H3K36me2 enrichment was also associated with active genes but at a lower level than H3K36me3

and began to increase further downstream of active TSSs (Figure 6.15). In contrast, H3K36me1 levels were low across actively transcribed regions. Lowly expressed genes were associated with a lower level of H3K36me3 enrichment compared to highly expressed genes. Silent genes were associated with depleted levels of H3K36me2 and H3K36me3 levels while modest levels of H3K36me1 were detected across inactive genes.

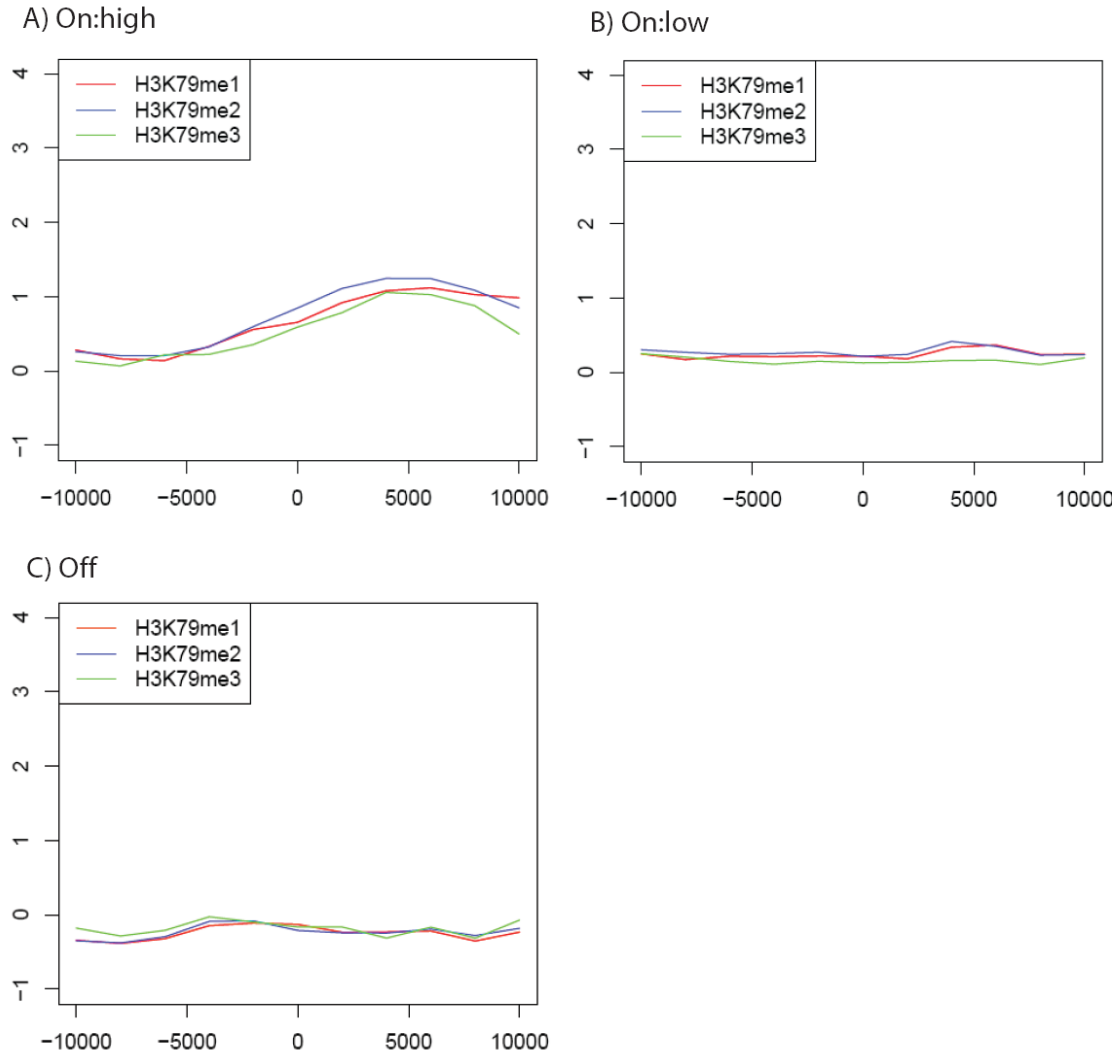


**Figure 6.15: Histone H3K36 methylation patterns at active and inactive genes.** H3K36me1, H3K36me2, and H3K36me3 average modification patterns are presented 10 kb upstream and downstream of TSSs associated with highly expressed (panel A), lowly expressed (panel B) and inactive (panel C) genes. The scale on the x-axis of each plot represents distances (bp's) upstream (negative values) and downstream (positive values) of known TSSs and z-scored log<sub>2</sub> values are presented on the y-axis



#### **6.5.2.6. Histone H3K79 methylation states are associated with highly transcribed genes**

While H3K79me3 has been implicated in both transcriptional activation and silencing in *Saccharomyces cerevisiae* (Ng *et al.*, 2002, 2003 b; Van Leeuwen *et al.*, 2002), there is no consensus on the distribution and role of this modification in mammalian cell types (Im *et al.*, 2003; Barski *et al.*, 2007). Even less is known about the distribution and role of H3K79me1 and H3K79me2 in mammalian cells. In this study of human monocytes all three H3K79 methylation states were observed to be prominently enriched in the transcribed region of highly active genes (Figure 6.16). All three states were enriched to a similar level with enrichment levels peaking approximately 5kb downstream of TSSs (consistent with earlier observations that H3K79me3 is associated with early transcribed regions in hESCs and monocytes). In contrast lowly expressed genes were associated with a subtle increase in H3K79me1 and H3K79me2 downstream of the TSS, while H3K79me3 levels remained low. No enrichment for any of the three modification states was observed at inactive genes. Taken together these results implicate all three states of H3K79 methylation in active gene expression, with H3K79me1 and H3K79me2 allow low level gene expression to occur while the presence of H3K79me3 facilitates higher transcription rates. These findings are consistent with the recent report that the H3K79 methyltransferase, DOT1, is recruited to the elongating RNA Polymerase II complex in human cells to facilitate chromatin remodeling during transcription (Bitoun *et al.*, 2007).



**Figure 6.16: Histone H3K79 methylation patterns at active and inactive genes.** H3K79me1, H3K79me2, and H3K79me3 average modification patterns are presented 10 kb upstream and downstream of TSSs associated with highly expressed (panel A), lowly expressed (panel B) and inactive (panel C) genes. The scale on the x-axis of each plot represents distances (bp's) upstream (negative values) and downstream (positive values) of known TSSs and z-scored  $\log_2$  values are presented on the y-axis.

## 6.6. Discussion

The work described in this Chapter involved using the modified ChIP-chip method to study a number of histone modifications in hESCs and lineage committed monocytes in the ENCODE regions. Four histone modifications (H3K4me3, H3K27me3, H3K36me3, and H3K79me3) were used in the analysis of chromatin state in undifferentiated SSEA3+ hESCs and differentiated SSEA3- hESCs. In CD14+ monocytes these four histone

modifications, plus an additional 15 were investigated. This latter analysis revealed a consensus histone code for gene expression and distal regulatory elements in human monocytes. The principal findings of this work are discussed below.

### **6.6.1. Bivalent chromatin structures are present in pluripotent hESCs and lineage committed monocytes**

Until recently, very little was known about the detailed chromatin structure of mammalian ES cell chromatin and how it contributes towards the maintenance of pluripotency or how it alters during differentiation. Recent reports have described how the functional antagonists - H3K4me3 and H3K27me3 - are often located at the same genomic regions in mouse and human ES cells and this bivalent modification pattern may be responsible for maintaining key developmental genes in a 'poised' state until required during differentiation (Bernstein *et al.*, 2006; Azuara *et al.*, 2006; Pan *et al.*, 2007; Zhao *et al.*, 2007). In this study, H3K4me3 and H3K27me3 modification patterns were examined at promoter regions in hESCs and lineage committed monocytes. This analysis suggested that promoters could be grouped into three categories- active, repressed or 'poised' for alternative developmental roles based on their association with H3K4me3 and H3K27me3. The number of monovalent H3K27me3 promoters was found to increase as cells became more differentiated, consistent with increased numbers of genes being repressed as lineage-specific gene expression patterns are established. Paradoxically, a similar percentage of promoters were associated with monovalent H3K4me3 in SSEA3+, SSEA3-, and CD14+ cells, suggesting that similar numbers of genes may be active in monocytes, as in hESCs. A similar percentage of bivalent 'poised' promoters were also observed across all three cell types which conflicted with a report by Bernstein and colleagues who demonstrated that bivalent promoters were predominantly found in pluripotent stem cells (Bernstein *et al.*, 2006). However, a large number of bivalent promoters have also been observed in CD4+ T cells (Barski *et al.*, 2007), suggesting that bivalent promoters are a feature of other committed cell types and not just restricted to pluripotent stem cells.

This study also showed that bivalent promoters in hESCs and monocytes were associated with genes involved in developmental processes, many of which were transcription factors. Bivalent chromatin structures may be responsible for silencing developmental genes in both hESCs and monocytes whilst still preserving their ability to become activated upon initiation of specific differentiation programs. This is consistent with the finding that bivalent genes are associated with little or no H3K36me3. This also suggests that there may be more epigenetic “flexibility” at the promoters of developmental factors in differentiated cells than previously anticipated. Many of the bivalent promoters were common to all three cell types, although more bivalent promoters were common between SSEA3<sup>+</sup> and SSEA3<sup>-</sup> cells which may reflect the similar developmental state of these two cell types. In addition, SSEA3<sup>-</sup> cells are derived from SSEA3<sup>+</sup> cells while the monocytes used in this study were not derived from SSEA3<sup>+</sup> cells and are therefore less likely to have similar characteristics.

While SSEA3<sup>+</sup> only or SSEA3<sup>-</sup> only bivalent promoters were rare, a proportion of bivalent promoters in monocytes were found only in this cell type. These genes were often associated with roles in immune processes suggesting that they were ‘poised’ for expression upon terminal differentiation of monocytes into phagocytic cell types. Analysis of bivalent promoters in SSEA3<sup>+</sup> hESCs, SSEA3<sup>-</sup> hESCs and monocytes showed that bivalent promoters often resolve during differentiation into monovalent H3K4me3 or H3K27me3 promoters or promoters associated with neither modification. It would be interesting to determine whether those bivalent promoters which are found in monocytes resolve following terminal differentiation into phagocytes.

Promoters were classified in terms of CpG content and this revealed a clear distinction in histone modification profiles between the two types of promoters. High CpG content promoters are known to be associated with ‘housekeeping’ genes or complex expression patterns while low CpG content promoters are often associated with tissue-specific promoters (Saxonov *et al.*, 2006). Approximately 70% of high CpG promoters were associated with H3K4me3 in hESCs and monocytes while only approximately 15% of low CpG promoters were associated with H3K4me3 in hESCs or monocytes. This is consistent with the association of trithorax complexes, which methylate H3K4, with CpG-rich DNA (Lee and Skalnik, 2005). Mikkelsen and colleagues also found that 99% of

high CpG promoters were associated with H3K4me3 in mouse ES cells while less than 10% of low CpG promoters were associated with this modification (Mikkelsen *et al.*, 2007). Very few high or low CpG promoters were H3K27me3 monovalent in SSEA3+ cells but this increased during differentiation as more high and low CpG promoters were observed to be H3K27me3 monovalent in monocytes. Differentiation may be accompanied by repression of both tissue-specific genes as well as genes with housekeeping functions or complex expression patterns. This suggests that high and low CpG content promoters may be regulated by different mechanisms. Most high CpG promoters are targeted by TrxG and may therefore be active by default unless actively repressed by PcG activity. In contrast, most low CpG promoters are targeted by neither TrxG nor PcG activity and may be inactive by default unless activated specifically by tissue-specific transcription factors.

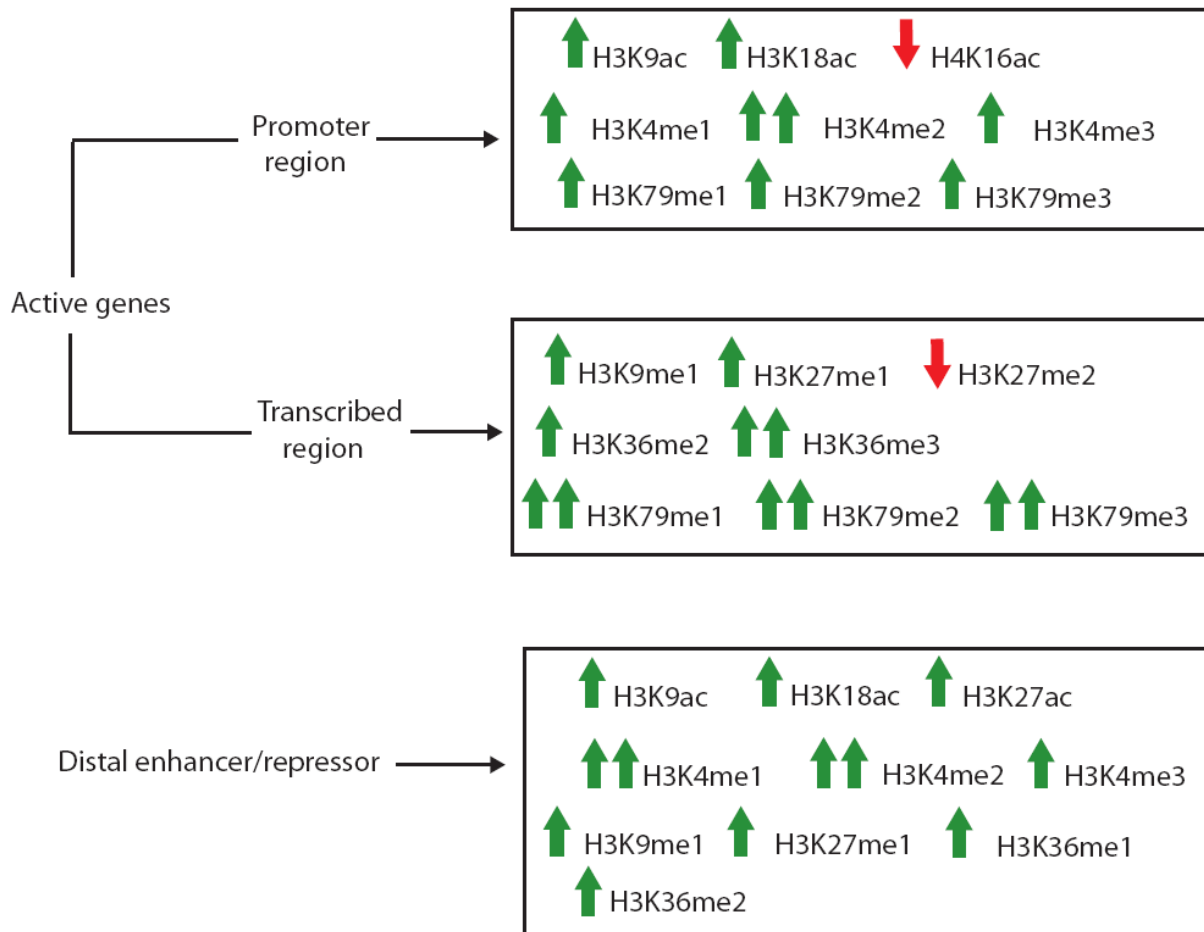
#### **6.6.2. Histone H3K36me3 and H3K79me3 are associated with different stages of transcription in hESCs and monocytes**

In hESCs and monocytes, H3K36me3 and H3K79me3 were strongly associated with active genes. H3K36me3 was enriched across the transcribed regions of active genes, beginning just downstream of the TSS and peaking at the 3' end of genes. This is consistent with previous observations (Bannister *et al.*, 2005; Barski *et al.*, 2007; Mikkelsen *et al.*, 2007). H3K79me3 on the other hand was associated with the early transcribed regions of active genes. Genes with bivalent promoters showed little or no H3K36me3 or H3K79me3. There was little overlap between H3K36me3 or H3K79me3 with H3K27me3 which is consistent with the finding that PcG complexes exclude RNA polymerases (Schuettengruber *et al.*, 2007). A number of putative novel transcripts were also associated with H3K36me3 and H3K79me3 in hESCs. As H3K36me3 and H3K79me3 are associated with 3' and 5' end of transcripts respectively, this information could be used to accurately identify the start and end point of novel transcripts.

### **6.6.3. Identification of a consensus histone code for active and inactive genes in human monocytes**

The work presented in this Chapter is one of the most extensive studies of histone modification states in a primary human cell type. Primary human cells from a healthy donor are not affected by the numerous epigenetic abnormalities associated with immortalized human cell lines (Liu *et al.*, 2005 b) and therefore represent a more accurate reflection of chromatin regulation in human cells. In addition, the CD14<sup>+</sup> monocyte used in this study were obtained from three different normal donors – indicating that the findings of this study are not specific to the genetic constitution of one particular individual. The analysis of 19 histone modifications at a large number of transcriptionally active and inactive genes led to the identification of a set of key histone modifications which defined active gene expression status in human monocytes. The presence and/or absence of these modifications distinguished promoters of transcribed genes from active transcribed regions. Genes not transcribed were associated with the absence of most of these histone modifications and were not readily identifiable by a histone signature. Distal regulatory elements (putative enhancers/repressors) were also associated with a characteristic histone modification pattern. Based on these observations, a histone code is proposed for identifying promoters of transcribed genes and transcribed regions in human monocytes as well as defining the location of enhancer/repressor elements (Figure 6.17). According to this code, promoters of transcribed genes are associated with elevated levels of H3K9ac, H3K18ac in particular and H4K16ac is depleted. Histone acetylation is associated with active gene expression (Kurdistani *et al.*, 2004; Pokholok *et al.*, 2005) as it is known to overcome the inhibitory effects of nucleosome compaction on transcription (Shahbazian and Grunstein, 2007). Furthermore, hypo-acetylation of H4K16 correlates with active gene expression in *S. cerevisiae* as it allows for the activating protein Bdf1 to bind to promoters (Kurdistani *et al.*, 2004). A similar mechanism may also be in operation in human cells. All three states of H3K4 methylation are predicted to be elevated at the promoter region of active genes while H3K4me2 is the most prominent modification. The code predicts H3K9me1 levels to be elevated in the transcribed region of active genes – a finding which has also been observed by Barski and colleagues (2007). The presence of H3K9me1 may allow for active genes to be rapidly ‘switched

off' by the addition of one or two methyl groups. Similarly, the code also predicts H3K27me1 enrichment just downstream of promoters of transcribed genes, extending across transcribed coding regions. This may allow for active genes to be rapidly 'switched off' by the addition of one or two methyl groups. All three K79 methylation states were enriched at promoters of actively transcribed genes while transcribed coding regions were associated with even higher enrichment for all three modifications. The code also predicts H3K36me2 enrichment in transcribed coding regions and high levels of H3K36me3 in the transcribed portion of active genes.



**Figure 6.17: A consensus histone code for regulatory and transcribed regions in human monocytes.** This code was proposed based on observations on the presence/absence of 19 histone acetylation and methylation modifications in monocyte cells. The modifications are shown associated with active promoters, active transcribed regions and distal enhancer/repressor elements. The green arrows facing

upwards represent enrichments for the specific histone modifications while the red arrows facing downwards represent depletions. The increase of a histone modification relative to surrounding regions is also indicated by one or two arrows.

Distal enhancer/repressor elements were also defined by a histone code. Distal elements are associated with acetylation of H3K9, H3K18 and H3K27 which is consistent with the studies of Roh and colleagues (2005, 2007) in which histone H3 acetylation was used to predict functional enhancers in human T cells. According to the code H3K4me1 is the most prominent modification (in conjunction with H3K4me2) at distal enhancer/repressor elements. Lower levels of H3K4me3 are also found at distal elements. A recent study suggested that high levels of H3K4me1 combined with low levels of H3K4me3 can be used to predict the location of enhancers in HeLa cells (Heinz *et al.*, 2007) and a similar pattern was observed in K562 cells (Chapter 3). These histone modifications associated with distal elements were also associated with promoters of transcribed genes - however distal enhancers/repressors could be distinguished from novel promoters by their preference for H3K4me1 rather than H3K4me3. In addition, association of distal elements with H3K9me1, H3K27me1, and H3K36me1 are also distinguishing features as promoters of actively transcribed genes are not associated with these modifications. Thus, distal elements are not defined by the presence of H3K4me1 alone but other mono-methylation modifications also play a role in predicting the location of these elements in the human genome.

This proposed consensus histone code for CD14<sup>+</sup> monocytes, however, does not take into account all the other histone modifications that were not examined by this study. These include methylation of H4K20, which is associated with active (H4K20me1) and repressed chromatin regions respectively (H4K20me2 and H4K20me3) (Vakoc *et al.*, 2006, Barski *et al.*, 2007), and histone arginine methylation modifications which have been linked to hormone nuclear receptor-mediated transcriptional activation (Lee *et al.*, 2002 b, Huang *et al.*, 2005). By developing and testing ChIP-chip assays to detect these and numerous other histone modifications across the ENCODE regions, it may be possible to define an even more comprehensive histone code. Nevertheless this present histone code proposes patterns of histone modifications which may be sufficient to



predict the location and function of promoters of transcribed genes, transcribed regions and associated regulatory elements in the human genome.

## Chapter 7

### Summary and Future work

#### 7.1. Summary of the work presented in this thesis

Understanding how gene expression is regulated is a fundamental issue in molecular biology. It is known that non-coding regulatory elements such as promoters, enhancers, and insulators play a central role in the regulation of gene expression in complex eukaryotic genomes. Therefore a detailed understanding of the location and properties of these elements in the human genome sequence would help further our understanding of gene regulation. However, the identification of these elements in the human genome is a difficult task as enhancers and insulators are often located far from genes in large stretches of non-coding DNA. Therefore, the aim of the work presented in this thesis was to perform large-scale profiling of a range of *in vivo* DNA-protein interactions which could be used to identify the location of regulatory elements in 1% of the human genome. This was achieved by the development of a number of ChIP-chip assays which were used in combination with a highly sensitive and reproducible ENCODE microarray platform. In addition, an improved ChIP-chip method was also developed for the large scale study of DNA-protein interactions in limited cell populations. This method was then applied to investigate the chromatin state of undifferentiated human embryonic stem cells, differentiated human embryonic stem cells, and human CD14<sup>+</sup> monocytes. These cell types represented different stages of cellular differentiation and the dynamics of chromatin-mediated gene regulation during differentiation was analysed. The results and interpretations of this work were presented in Chapters 3, 4, 5, and 6 and are summarised below. Future work which follows on from the work presented in this thesis is also outlined in this concluding chapter.

### **7.1.1. Application of microarray technology for large-scale characterisation of promoter and distal enhancer/repressor elements in the human genome (Chapter 3)**

In Chapter 3, a number of ChIP-chip assays were applied to characterise regulatory elements in the human genome on a large scale using the ENCODE microarray. Those assays included detecting genomic regions associated with H3 acetylation, H4 acetylation, H3K4me1, H2K4me2, and H3K4me3. Histone H2B and H3 density, taken to reflect nucleosome occupancy, was also assessed using this array platform. Un-amplified ChIP DNA material derived from  $10^7$  cultured cells could be used to perform hybridisations with the ENCODE array. The use of un-amplified ChIP DNAs reduced sources of non-biological biases resulting from amplification methods, ensuring that the ENCODE platform performed reproducibly in independently performed assays.

A detailed analysis of H3 acetylation, H4 acetylation, H3K4me1, H3K4me2, and H3K4me3 distribution showed that promoter and distal enhancer/repressor elements could be distinguished in terms of their histone K4 methylation signatures. H3K4me3 was found to be most prominent at promoters of active genes while H3K4me1 was the predominant modification at distal elements. Both promoters of actively transcribed genes and distal regulatory elements associated with histone modifications were shown to be associated with H2B and H3 nucleosome depletion; this feature was not evident at promoters of non-transcribed genes. Taken together these results were the first demonstration that the ENCODE microarray could be used for the large-scale identification and characterisation of regulatory elements in the human genome.

### **7.1.2. ChIP-chip analysis of the insulator binding protein CTCF and its associated transcription factors (Chapter 4)**

Chapter 4 describes the development of ChIP-chip assays for the study of insulator elements in the human genome. A ChIP-chip assay was developed to detect genomic sequences associated with the insulator binding protein CTCF. This assay was developed with the SCL tiling microarray and then applied to study CTCF in the ENCODE regions. The CTCF ChIP-chip assay was validated by comparisons with other experimental defined CTCF binding sites located in the ENCODE regions. Many of the CTCF binding

sites found in this study contained a previously characterised consensus CTCF binding motif. CTCF binding sites were found to be very common in the genome as over 500 binding sites were identified in the ENCODE regions in K562 cells – scaled to the whole genome, this would suggest that there as many as 50,000 CTCF sites genome-wide.

CTCF binding sites were found at promoters, exons, introns and intergenic regions. Over 50% of genes were found to be flanked in their entirety by two CTCF sites and CTCF sites were also found to flank clusters of related or unrelated genes. Multiple CTCF binding sites within genes were also observed. Over half of CTCF binding sites were located far (> 5kb) from TSSs, consistent with a role in insulator function. The remainder was located at TSSs and distal putative enhancer/repressor elements and may be involved in transcriptional activation/repression. Insulator elements may be associated with accessible chromatin domains as CTCF sites were found to co-localise with DNase I hypersensitive sites, FAIRE peaks and regions of histone H2B/H3 depletion. Furthermore, a number of CTCF sites were located at the boundary between active and inactive chromatin domains defined by the presence of H3K27me1 and H3K27me3 respectively, suggestive of a barrier insulator function. CTCF binding sites were found to be highly conserved between K562 and U937 cells indicating that the location of insulators is conserved between cell types.

In addition, known binding partners of CTCF - USF1, USF2, and mSin3a – were studied in K562 cells. Analysis of USF1 and USF2 binding sites revealed a widespread distribution with no bias towards promoter, exonic, intronic and intergenic locations. However, mSin3a interactions were found to be heavily biased towards promoter locations and binding was found to be a good predictor of active gene expression. One third of CTCF sites overlapped with interactions for one or more of the three factors, over half of which were located at TSSs, while 22% and 23% of overlapping interactions were located at distal enhancer/repressor elements and other locations (thought to be insulator elements) respectively. Therefore, the majority of mSin3a, USF1, and USF2 overlapping interactions with CTCF may be involved in regulating gene expression by interacting with promoter and distal elements. However, only a minority of putative insulators were associated with CTCF and either USF1 or USF2, suggesting that their co-localisation may not be a general paradigm for insulator barrier function in the human genome.

### **7.1.3. Development of a modified ChIP-chip method for use with fewer cells (Chapter 5)**

Chapters 3 and 4 demonstrated that ChIP-chip assays could be used to identify various types of regulatory elements using immortalised human cell lines. However, when working with many types of primary human cells and other scarce samples, cell numbers limit the application of ChIP-chip technology. Therefore ChIP assays were developed which could be used in combination with microarray hybridization to detect histone modifications from as few as  $10^4$  K562 cells with similar efficiencies as assays performed with  $10^7$  cells. This protocol was developed using the detection of regions enriched for histone acetylation and H3K4me3 at the SCL locus as a model system. The antibody to chromatin ratio was found to be critical for the reliable detection of known regulatory elements when cell numbers were reduced. Other factors such as the level of protein-G agarose and data normalization procedures to account for non-specific interactions were also assessed as means of improving the detection of *bona fide* regulatory elements. The sensitivity of the method was confirmed by quantitative PCR and further validated with ChIP-chip assays for a number of other histone modifications including H3K27me3, H3K36me3 and H3K79me3.

### **7.1.4. Application of the modified ChIP-chip method to study chromatin regulation during differentiation (Chapter 6)**

In this Chapter the modified ChIP-chip approach was applied to study the patterns of histone modifications at various stages of cellular differentiation. The patterns of H3K4me3, H3K27me3, H3K36me3, and H3K79me3 were examined in SSEA3+ hESCs, SSEA3- hESCs, and CD14+ monocytes. Analysis of H3K4me3 and H3K27me3 at promoters revealed three classes of promoters: H3K4me3 monovalent, H3K4me3/H3K27me3 bivalent promoters, and H3K27me3 monovalent promoters. Bivalent promoters were found to be a feature of all three cell types, many of which were associated with genes involved in the transcriptional regulation of developmental. A number of bivalent promoters were common to the three cell types but a large proportion of monocyte bivalent promoters were specific to this cell type. Analysis of promoter CpG

content showed that the majority of high CpG content promoters (often associated with housekeeping genes) were enriched with H3K4me3 (alone or bivalent) in the three cell types while few low CpG content promoters (often associated with tissue-specific promoters) were enriched for H3K4me3. Comparisons of chromatin state during differentiation showed that promoters in SSEA3<sup>+</sup> and SSEA3<sup>-</sup> hESCs displayed a similar histone modification profile for many of the same genes whilst monocytes had histone modifications associated primarily with different sets of genes. H3K36me3 was found to be associated with the entire length of primary transcripts in hESCs and monocytes. A number of non-coding RNAs were also associated with H3K36me3 in hESCs. H3K79me3 was located at the 5' end of actively transcribed genes in the three cell types.

#### **7.1.5. Identification of a consensus histone code in human monocytes (Chapter 6)**

The modified ChIP-chip method allowed for the detailed study of 19 histone modifications in monocytes. Chromatin modification maps were created for four histone acetylation modifications (H3K9ac, H3K18ac, H3K27ac, and H4K16ac) and the mono-, di-, and tri-methylation states of H3K4, K27, K36, and K79. This resulted in the identification of a detailed consensus histone code for promoters of transcribed genes, transcribed coding regions and distal enhancer/repressor elements. A histone code for inactive genes was not readily identifiable based on the modifications studied in this thesis. Promoters of transcribed genes were associated with H3K9ac, K18ac, and K27ac as well as all three H3K4 methylation states, in particular H3K4me2. Actively transcribed coding regions were associated with high levels of H3K36me3 and all three states of H3K79 methylation. Distal enhancer/repressors were associated with the three histone H3 acetylation events as well as the three H3K4 methylation states. H3K4me1 and H3K4me2 were most prominent at distal elements. H3K9, K27 and K36 mono-methylation states were also associated with distal elements. Taken together these results suggest that the modified ChIP-chip method can be used to provide a detailed characterization of the histone code of regulatory and coding elements in the human genome.

## **7.2. Future Work**

The work described in this thesis has shown how *in vivo* DNA-protein interactions can be assayed across large regions of the genome in numerous cell types using a microarray-based method. Here, the Sanger ENCODE microarray resource was used to detect other DNA-protein interactions associated with regulatory elements in 1% of the human genome, but the data could be extrapolated to suggest mechanisms and functional relationships which represent found across the entire human genome. For example, there are at least 34 amino acid residues in the core histone proteins that are chemically modified in human cells (Garcia *et al.*, 2007) and the distribution and function of many of these modifications is not well understood. In addition, approximately 2600 transcription factors have been identified in the human genome (Babu *et al.*, 2004), and how these factors act individually and collectively to regulate transcription programs is a major challenge in the post-genome era. However, for genomic annotation and identification of all regulatory sequences genome-wide, there is a need to completely interrogation of the entire genome using high-throughput methods such as CHIP-chip. Other techniques which could be used for the high-throughput identification of genome features which influence gene expression are also outlined below.

### **7.2.1. Further development of ChIP-chip assays for characterising regulatory elements and the role of histone modifications in gene regulation**

The work presented in this thesis involved mapping the location of a number of histone modifications to analyse their distribution in 1% of the human genome and explore the relationships between modifications. However, while this represented a detailed study of histone modifications, it was by no means exhaustive for the number of modifications that could be investigated. For example, methylation of H4K20 was not examined in this study and H4K20me3 has been shown to be associated with repressive chromatin while the mono-methylation form has been found at the promoter and coding regions of active genes (Schotta *et al.*, 2004, Talasz *et al.*, 2005, Vakoc *et al.*, 2006). In addition, histone arginine methylation was not examined in this study; histone H3 arginine methylation has been shown to regulate pluripotency in the early mouse embryo (Torres-Padilla *et al.*, 2007) while asymmetrically dimethylated H3R2 has been shown to regulate the

deposition of H3K4me3 in the yeast genome (Kirmizis *et al.*, 2007). The function of arginine methylation is still very much unexplored in the human genome. Phosphorylation of histone residues has been implicated in both transcriptional activation and repression (Nowak and Corces, 2004). Ubiquitylation and sumoylation are two large histone modifications whose roles in regulating gene expression are least understood. Ubiquitylation can be repressive or activating depending on the site modified (Zhu *et al.*, 2005; Wang *et al.*, 2004). Sumoylation is the only repressive modification described in yeast (Shiio and Eisenman, 2003) but its role in mammalian cells is not known.

Furthermore, an understanding of how the various histone modifications interact with other proteins and how this is translated into specific biological outputs is not well understood. Two models have been proposed - the direct and effector mediated models (Jenuwein and Allis, 2001; Strahl and Allis, 2000). In the direct model, histone modifications such as acetylation and phosphorylation directly affect interactions between basic histone proteins and negatively charged DNA (Shogren-Knaak *et al.*, 2006; Ahn *et al.*, 2005). The effector model proposes that histone modifications are 'read' by cognate non-histone binding proteins, known as effectors. Effector proteins can compact chromatin by cross-linking nucleosomes (Francis *et al.*, 2004), enhance the binding of the RNA polymerase complex (Vermeulen *et al.*, 2007) or recruit chromatin remodeling factors (Jenuwein and Allis, 2001). To date there are 11 known protein domain families which 'read' the various histone acetylation, methylation and phosphorylation modifications (Taverna *et al.*, 2007). Effector proteins, such as those which contain bromodomains, are known to interact with lysine acetylation marks (Dhalluin *et al.*, 1999) and are often found in histone acetyltransferases and chromatin remodeling complexes (Zeng and Zhou, 2002). Chromodomains of HP1 and Polycomb are known to recognize di- and tri-methyl H3K9 and H3K27 respectively (Lachner *et al.*, 2001; Min *et al.*, 2003). Double chromodomain, double tudor domain and PHD finger containing proteins have been shown to recognize H3K4me3 (Flanagan *et al.*, 2005; Huang *et al.*, 2006; Pena *et al.*, 2006; Li *et al.*, 2006). The implications of multiple domains which recognize the same modification is not known.

The role of histone modifications in regulating gene expression and other nuclear processes is further complicated by the realization that histone modifications rarely occur



in isolation - mass spectrometry studies are beginning to show that multiple modifications occur on the same histone tail (Cosgrove, 2007). The promiscuity of histone modifications with respect to recruiting effector binding proteins (there are more than 10 effector proteins known to bind H3K4me3) suggests that one histone modification is not sufficient to stably recruit a complex (Ruthenburg *et al.*, 2007). The co-existence of multiple modifications within a given tail or chromatin domain may serve to dictate the recruitment of complexes (Ruthenburg *et al.*, 2007). For example, the PHD finger and bromodomain of BPTF could be used to simultaneously bind H3K4me3 and an acetylated lysine residue. It will be important to gain an increased understanding of how histone modification combinations and effector proteins operate and dictate functional outcomes.

Future ChIP-based approaches will be important in determining the co-localisation of modifications and their effector proteins. However, the ability to perform ChIP assays to detect histone modifications and the binding of effector proteins is dependent on the production of antibodies which perform in ChIP assays and recent years has seen a large increase in the number of 'ChIP grade' antibodies becoming available from commercial suppliers. The continued availability of antibodies to novel histone modifications and histone associated proteins will ensure that future ChIP studies can be performed to gain a more complete understanding of the relationship between histone modifications and the proteins which recognize these marks.

### **7.2.2. Applications of microarrays for characterising other features of the human genome**

While nucleosomes represent the fundamental repeating unit of chromatin and play a crucial role in a number of genome functions, higher order chromatin structures also contribute to the regulation of the genome. It has been suggested that chromatin forms 50–200 kb chromosomal loops which are attached to the nuclear matrix (Bode *et al.*, 2000; Heng *et al.*, 2004). The anchor points of the DNA to the nuclear matrix have been termed matrix attachment regions (MARs), which have been implicated in the control DNA replication and gene expression (Jenke *et al.*, 2004; Amati and Gasser, 1990). The DNA and proteins associated with the nuclear matrix can be isolated by extraction of

histones with high salt or mild detergent followed by restriction enzyme or DNase I treatment to remove all DNA except matrix attached DNA. Sumer and colleagues isolated MAR DNA and hybridized it to a BAC/PAC array to define a 2.5 Mb region of MAR enriched DNA at a human neocentromere (Sumer *et al.*, 2003). Ioudinkova and colleagues have also used arrays to map MARs at the chicken  $\alpha$ -globin domain (Ioudinkova *et al.*, 2005). However high-resolution mapping of MARs across a large region of the human genome has yet to be carried out.

Physical interactions between regulatory elements play an important role in gene regulation. Current evidence suggests that chromatin loops allow these elements to physically interact with their target genes to repress or silence transcription (Spilianakis *et al.*, 2005; Splinter *et al.*, 2006). Two genomic regions can be tested for interaction using the chromosome conformation capture (3C) technique (Dekker *et al.*, 2002), in which protein-DNA interactions are cross-linked using formaldehyde, followed by digestion with a restriction enzyme. Interacting fragments are ligated together and can be quantified individually by qPCR. The 3C method has recently been modified for the high-throughput detection of sequences which interact with a particular DNA fragment in a method known as 5C (3C carbon copy) (Dostie *et al.*, 2006). In this method T7 and T3 primers are ligated to interacting fragments and used as primers for PCR amplification. Custom microarrays can then be used to analyse the junction fragments or alternatively high-throughput sequencing can be used to detect interacting fragments. A circular chromosome conformation capture (4C) method has also been developed (Zhao *et al.*, 2006 b), which involves the circularization of 3C products. Inverse PCR from the known test fragment was used to amplify interacting fragments which were then cloned and sequenced. These sequences were then used to create a customized microarray. Another 4C method (3C-on-chip) was developed by Simonis and colleagues (2006) in which circular inverse PCR was used to amplify interacting sequences which were then identified directly using a custom designed array. These methods could be used to identify which genes are controlled by particular regulatory elements, which is often difficult to decipher based purely on their proximity to the gene which they are thought to control.

While chromatin modifying enzymes are known to co-localize to DNA replication sites (McNairn and Gilbert, 2003), the relationship between replication timing, chromatin and the regulation of gene expression is not fully understood. Microarrays can be used to assess replication timing in the human genome (Woodfine *et al.*, 2004; Jeon *et al.*, 2005) and could also be correlated with information on the location of DNA replication origins and chromatin modifications. The replication of DNA requires the formation of origin recognition complexes (ORCs) (Gilbert, 2007). ChIP-chip analysis of ORC proteins has been used to identify replication origins in the yeast genome (Wyrick *et al.*, 2001; Xu *et al.*, 2006) but this approach has not been applied to identify replication origins in the human genome. ORC complexes may be only formed during the S-phase of the cell cycle, meaning that human cells may need to be synchronized to identify ORC-DNA interactions.

### **7.2.3. Testing the proposed histone code**

The analysis of 19 histone modifications in CD14+ monocytes led to model for a histone code, in which a number of key histone modifications were associated with various regulatory elements and gene activity. This histone code is consistent for the genes contained on the ENCODE array in human monocytes, however it remains to be seen if this code is consistent for all the genes in the human genome. This would be an extremely large undertaking considering the number of histone modifications analysed in this study. Furthermore, a complete understanding of the histone code will be much more complex to decipher than the genetic code given the large number of known histone modifications. Multiple histone modifications can also be located on the same histone tail and different histone modifying enzymes are often responsible for catalysing the same modification. Different effector proteins may be recruited by various histone modification combinations, adding further complexity to understanding the code. Methods such as sequential ChIP and mass spectrometry will allow for histone modification and modification/effector combinations to be determined.

Finally, elucidating the role of these epigenetic events in normal development and disease is the ultimate goal. Having developed a modified ChIP method for the study of histone modifications in limited cell populations, there is scope for employing this method to

study the histone code in rare cell populations at different stages of development. This could be used to gain a greater understanding of the chromatin basis of cellular differentiation and pluripotency. The development of the new generation sequencing technologies, which may be more cost effective and experimentally-efficient than whole genome microarray studies, will also help to realize the goal of annotating regulatory elements in the human genome during normal development and disease. Multiple layers of ChIP information will be overlaid onto the genetic code – an undertaking much more complex project than even the sequencing of the the human genome. It will also be important to understand how chromatin state is altered during abnormal development and disease (De Gobbi *et al.*, 2006). Such chromatin-based studies will be logical extensions of the sequencing of human genomes associated with pathological states. Once this has been done, the relationships between specific genomic sequence elements, their sequence variants, chromatin function and disease will be routinely determined.

However, when complete, this information will give us an unprecedented understanding of our genome. There is also the possibility of extending our understanding of the histone code to other mammalian genome in order to determine the conservation of the code amongst species. A more complex histone code may provide a distinction between higher and lower eukaryotes and there is already evidence for more elaborate patterns of histone modifications in metazoans (Garcia *et al.*, 2007).

#### **7.2.4. Further characterisation of putative CTCF insulator elements**

CTCF binding sites identified in this study could be tested for enhancer-blocking activity or barrier activity using functional plasmid-based assays developed by Felsenfeld and colleagues (Chung *et al.*, 1993; Pikaart *et al.*, 1998). Enhancer-blocking activity is tested by cloning a CTCF binding site between an erythroid-specific enhancer and promoter driving expression of G418 resistance. The construct is then stably transfected into K562 cells and a CTCF binding site which displays enhancer-blocking activity gives a reduced number of colonies relative to the control plasmid when grown on agar. A CTCF binding site can also be tested for barrier activity by flanking the interleukin 2 receptor (IL-2R) expression cassette with CTCF binding sites. The construct is then stably transfected into the human genome and expression of IL-2R is monitored by flow cytometry. Expression

of any transgene stably integrated into the genome is normally silenced after a prolonged period in culture due to the spread of inactivating histone modifications. However, when an IL-2R transgene is flanked by CTCF sites associated with barrier activity constant expression is maintained. However, these assays are not suitable for high-throughput screening of insulator function. Mukhopadhyay and colleagues developed a technique for the high-throughput screening of insulator function (Mukhopadhyay *et al.*, 2004), in which sequences from a mouse CTCF target site library were inserted into a plasmid to interfere with SV40 enhancement of toxin-A reporter gene activity. The number of cell clones increased dramatically when the toxin-A gene was insulated from the SV40 enhancer. Total DNA and DNA from the emerging clones could be differentially labeled and hybridized to a microarray. An increase in microarray signal would be observed for functionally selected sequences.

### **7.3. Final thoughts**

The findings presented in this thesis illustrate how microarray technology is being applied to study global gene regulatory events. Given the recent advances in microarray and sequencing technology, the post-genome goal of annotating the complete repertoire of functional elements in the human genome sequence will surely be realized. The findings of this thesis will make an initial contribution towards this goal.

## References

- (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431(7011): 931-945.
- (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306(5696): 636-640.
- Adlam M, Siu G (2003) Hierarchical interactions control CD4 gene expression during thymocyte development. *Immunity* 18(2): 173-184.
- Affolter M, Montagne J, Walldorf U, Groppe J, Kloter U et al. (1994) The Drosophila SRF homolog is expressed in a subset of tracheal cells and maps within a genomic region required for tracheal development. *Development* 120(4): 743-753.
- Agalioti T, Chen G, Thanos D (2002) Deciphering the transcriptional histone acetylation code for a human gene. *Cell* 111(3): 381-392.
- Agger K, Cloos PA, Christensen J, Pasini D, Rose S et al. (2007) UTX and JMJD3 are histone H3K27 demethylases involved in HOX gene regulation and development. *Nature* 449(7163): 731-734.
- Ahmad K, Henikoff S (2001) Centromeres are specialized replication domains in heterochromatin. *J Cell Biol* 153(1): 101-110.
- Ahmad K, Henikoff S (2002) The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. *Mol Cell* 9(6): 1191-1200.
- Ahn SH, Cheung WL, Hsu JY, Diaz RL, Smith MM et al. (2005) Sterile 20 kinase phosphorylates histone H2B at serine 10 during hydrogen peroxide-induced apoptosis in *S. cerevisiae*. *Cell* 120(1): 25-36.
- Akasaka T, van Lohuizen M, van der Lugt N, Mizutani-Koseki Y, Kanno M et al. (2001) Mice doubly deficient for the Polycomb Group genes *Mei18* and *Bmi1* reveal synergy and requirement for maintenance but not initiation of Hox gene expression. *Development* 128(9): 1587-1597.
- Allen GC, Spiker S, Thompson WF (2000) Use of matrix attachment regions (MARs) to minimize transgene silencing. *Plant Mol Biol* 43(2-3): 361-376.
- Allfrey VG, Faulkner R, Mirsky AE (1964) Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis. *Proc Natl Acad Sci U S A* 51: 786-794.
- Amati B, Gasser SM (1990) Drosophila scaffold-attached regions bind nuclear scaffolds and can function as ARS elements in both budding and fission yeasts. *Mol Cell Biol* 10(10): 5442-5454.
- Anderson CL, Abraham GN (1980) Characterization of the Fc receptor for IgG on a human macrophage cell line, U937. *J Immunol* 125(6): 2735-2741.
- Angelov D, Molla A, Perche PY, Hans F, Cote J et al. (2003) The histone variant macroH2A interferes with transcription factor binding and SWI/SNF nucleosome remodeling. *Mol Cell* 11(4): 1033-1041.
- Aoyagi S, Narlikar G, Zheng C, Sif S, Kingston RE et al. (2002) Nucleosome remodeling by the human SWI/SNF complex requires transient global disruption of histone-DNA interactions. *Mol Cell Biol* 22(11): 3653-3662.
- Aplan PD, Begley CG, Bertness V, Nussmeier M, Ezquerro A et al. (1990) The SCL gene is formed from a transcriptionally complex locus. *Mol Cell Biol* 10(12): 6426-6435.

- Aronow BJ, Silbiger RN, Dusing MR, Stock JL, Yager KL et al. (1992) Functional analysis of the human adenosine deaminase gene thymic regulatory region and its ability to generate position-independent transgene expression. *Mol Cell Biol* 12(9): 4170-4185.
- Attema JL, Papathanasiou P, Forsberg EC, Xu J, Smale ST et al. (2007) Epigenetic characterization of hematopoietic stem cell differentiation using miniChIP and bisulfite sequencing analysis. *Proc Natl Acad Sci U S A* 104(30): 12371-12376.
- Ayer DE, Lawrence QA, Eisenman RN (1995) Mad-Max transcriptional repression is mediated by ternary complex formation with mammalian homologs of yeast repressor Sin3. *Cell* 80(5): 767-776.
- Azuara V, Perry P, Sauer S, Spivakov M, Jorgensen HF et al. (2006) Chromatin signatures of pluripotent cell lines. *Nat Cell Biol* 8(5): 532-538.
- Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA (2004) Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol* 14(3): 283-291.
- Bailey TL, Elkan C (1995) The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol* 3: 21-29.
- Bajic VB, Tan SL, Suzuki Y, Sugano S (2004) Promoter prediction analysis on the whole human genome. *Nat Biotechnol* 22(11): 1467-1473.
- Bajic VB, Brent MR, Brown RH, Frankish A, Harrow J et al. (2006) Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment. *Genome Biol* 7 Suppl 1: S3 1-13.
- Balhoff JP, Wray GA (2005) Evolutionary analysis of the well characterized endo16 promoter reveals substantial variation within functional sites. *Proc Natl Acad Sci U S A* 102(24): 8591-8596.
- Banerji J, Rusconi S, Schaffner W (1981) Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* 27(2 Pt 1): 299-308.
- Banerji J, Olson L, Schaffner W (1983) A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* 33(3): 729-740.
- Bannister AJ, Zegerman P, Partridge JF, Miska EA, Thomas JO et al. (2001) Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* 410(6824): 120-124.
- Bannister AJ, Kouzarides T (2005) Reversing histone methylation. *Nature* 436(7054): 1103-1106.
- Bannister AJ, Schneider R, Myers FA, Thorne AW, Crane-Robinson C et al. (2005) Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. *J Biol Chem* 280(18): 17732-17736.
- Barski A, Cuddapah S, Cui K, Roh TY, Schones DE et al. (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129(4): 823-837.
- Battaglioli E, Andres ME, Rose DW, Chenoweth JG, Rosenfeld MG et al. (2002) REST repression of neuronal genes requires components of the hSWI.SNF complex. *J Biol Chem* 277(43): 41038-41045.
- Becker PB (2002) Nucleosome sliding: facts and fiction. *EMBO J* 21(18): 4749-4753.
- Begley CG, Green AR (1999) The SCL gene: from case report to critical hematopoietic regulator. *Blood* 93(9): 2760-2770.

- Bell AC, West AG, Felsenfeld G (1999) The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* 98(3): 387-396.
- Bell AC, Felsenfeld G (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the *Igf2* gene. *Nature* 405(6785): 482-485.
- Bell AC, West AG, Felsenfeld G (2001) Insulators and boundaries: versatile regulatory elements in the eukaryotic. *Science* 291(5503): 447-450.
- Bentley DR (2006) Whole-genome re-sequencing. *Curr Opin Genet Dev* 16(6): 545-552.
- Berger SL (2002) Histone modifications in transcriptional regulation. *Curr Opin Genet Dev* 12(2): 142-148.
- Berger JA, Hautaniemi S, Jarvinen AK, Edgren H, Mitra SK et al. (2004) Optimized LOWESS normalization parameter selection for DNA microarray data. *BMC Bioinformatics* 5: 194.
- Bernard O, Azogui O, Lecointe N, Mugneret F, Berger R et al. (1992) A third *tal-1* promoter is specifically used in human T cell leukemias. *J Exp Med* 176(4): 919-925.
- Bernstein BE, Humphrey EL, Erlich RL, Schneider R, Bouman P et al. (2002) Methylation of histone H3 Lys 4 in coding regions of active genes. *Proc Natl Acad Sci U S A* 99(13): 8695-8700.
- Bernstein BE, Liu CL, Humphrey EL, Perlstein EO, Schreiber SL (2004) Global nucleosome occupancy in yeast. *Genome Biol* 5(9): R62.
- Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK et al. (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* 120(2): 169-181.
- Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* 125(2): 315-326.
- Bernstein BE, Meissner A, Lander ES (2007) The mammalian epigenome. *Cell* 128(4): 669-681.
- Bhatia M (2007) Hematopoiesis from human embryonic stem cells. *Ann N Y Acad Sci* 1106: 219-222.
- Bieda M, Xu X, Singer MA, Green R, Farnham PJ (2006) Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Res* 16(5): 595-605.
- Biggin MD (1999) Ultraviolet cross-linking assay to measure sequence-specific DNA binding in vivo. *Methods Enzymol* 304: 496-515.
- Bird A (2002) DNA methylation patterns and epigenetic memory. *Genes Dev* 16(1): 6-21.
- Birney E, Durbin R (2000) Using GeneWise in the *Drosophila* annotation experiment. *Genome Res* 10(4): 547-548.
- Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447(7146): 799-816.
- Birve A, Sengupta AK, Beuchle D, Larsson J, Kennison JA et al. (2001) *Su(z)12*, a novel *Drosophila* Polycomb group gene that is conserved in vertebrates and plants. *Development* 128(17): 3371-3379.



- Bitoun E, Oliver PL, Davies KE (2007) The mixed-lineage leukemia fusion partner AF4 stimulates RNA polymerase II transcriptional elongation and mediates coordinated chromatin remodeling. *Hum Mol Genet* 16(1): 92-106.
- Bjerkvig R, Tysnes BB, Aboody KS, Najbauer J, Terzis AJ (2005) Opinion: the origin of the cancer stem cell: current controversies and new insights. *Nat Rev Cancer* 5(11): 899-904.
- Blais A, Tsikitis M, Acosta-Alvear D, Sharan R, Kluger Y et al. (2005) An initial blueprint for myogenic differentiation. *Genes Dev* 19(5): 553-569.
- Blanchette M, Tompa M (2003) FootPrinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res* 31(13): 3840-3842.
- Bode J, Benham C, Knopp A, Mielke C (2000) Transcriptional augmentation: modulation of gene expression by scaffold/matrix-attached regions (S/MAR elements). *Crit Rev Eukaryot Gene Expr* 10(1): 73-90.
- Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I et al. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* 299(5611): 1391-1394.
- Bosher JM, Totty NF, Hsuan JJ, Williams T, Hurst HC (1996) A family of AP-2 proteins regulates c-erbB-2 expression in mammary carcinoma. *Oncogene* 13(8): 1701-1707.
- Botuyan MV, Lee J, Ward IM, Kim JE, Thompson JR et al. (2006) Structural basis for the methylation state-specific recognition of histone H4-K20 by 53BP1 and Crb2 in DNA repair. *Cell* 127(7): 1361-1373.
- Boyer LA, Latek RR, Peterson CL (2004) The SANT domain: a unique histone-tail-binding module? *Nat Rev Mol Cell Biol* 5(2): 158-163.
- Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS et al. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 122(6): 947-956.
- Boyer LA, Plath K, Zeitlinger J, Brambrink T, Medeiros LA et al. (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* 441(7091): 349-353.
- Bracken AP, Dietrich N, Pasini D, Hansen KH, Helin K (2006) Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. *Genes Dev* 20(9): 1123-1136.
- Bradley A (1990) Embryonic stem cells: proliferation and differentiation. *Curr Opin Cell Biol* 2(6): 1013-1017.
- Brasset E, Vaury C (2005) Insulators are fundamental components of the eukaryotic genomes. *Heredity* 94(6): 571-576.
- Briggs SD, Bryk M, Strahl BD, Cheung WL, Davie JK et al. (2001) Histone H3 lysine 4 methylation is mediated by Set1 and required for cell growth and rDNA silencing in *Saccharomyces cerevisiae*. *Genes Dev* 15(24): 3286-3295.
- Brown KE, Baxter J, Graf D, Merckenschlager M, Fisher AG (1999) Dynamic repositioning of genes in the nucleus of lymphocytes preparing for cell division. *Mol Cell* 3(2): 207-217.

- Bruce AW, Donaldson IJ, Wood IC, Yerbury SA, Sadowski MI et al. (2004) Genome-wide analysis of repressor element 1 silencing transcription factor/neuron-restrictive silencing factor (REST/NRSF) target genes. *Proc Natl Acad Sci U S A* 101(28): 10458-10463.
- Bryne JC, Valen E, Tang MH, Marstrand T, Winther O et al. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res* 36(Database issue): D102-106.
- Buck MJ, Nobel AB, Lieb JD (2005) ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. *Genome Biol* 6(11): R97.
- Bulger M, Groudine M (1999) Looping versus linking: toward a model for long-distance gene activation. *Genes Dev* 13(19): 2465-2477.
- Bulger M, van Doorninck JH, Saitoh N, Telling A, Farrell C et al. (1999) Conservation of sequence and structure flanking the mouse and human beta-globin loci: the beta-globin genes are embedded within an array of odorant receptor genes. *Proc Natl Acad Sci U S A* 96(9): 5129-5134.
- Burcin M, Arnold R, Lutz M, Kaiser B, Runge D et al. (1997) Negative protein 1, which is required for function of the chicken lysozyme gene silencer in conjunction with hormone receptors, is identical to the multivalent zinc finger repressor CTCF. *Mol Cell Biol* 17(3): 1281-1288.
- Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 268(1): 78-94.
- Burgess-Beusse B, Farrell C, Gaszner M, Litt M, Mutskov V et al. (2002) The insulation of genes from external enhancers and silencing chromatin. *Proc Natl Acad Sci U S A* 99 Suppl 4: 16433-16437.
- Cai HN, Shen P (2001) Effects of cis arrangement of chromatin insulators on enhancer-blocking activity. *Science* 291(5503): 493-495.
- Cao R, Wang L, Wang H, Xia L, Erdjument-Bromage H et al. (2002) Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science* 298(5595): 1039-1043.
- Cao R, Zhang Y (2004) The functions of E(Z)/EZH2-mediated methylation of lysine 27 in histone H3. *Curr Opin Genet Dev* 14(2): 155-164.
- Capelson M, Corces VG (2005) The ubiquitin ligase dTopors directs the nuclear organization of a chromatin insulator. *Mol Cell* 20(1): 105-116.
- Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC et al. (2005) The transcriptional landscape of the mammalian genome. *Science* 309(5740): 1559-1563.
- Carroll JS, Liu XS, Brodsky AS, Li W, Meyer CA et al. (2005) Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell* 122(1): 33-43.
- Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR et al. (2006) Genome-wide analysis of estrogen receptor binding sites. *Nat Genet* 38(11): 1289-1297.

- Carrozza MJ, Li B, Florens L, Suganuma T, Swanson SK et al. (2005) Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell* 123(4): 581-592.
- Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA et al. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 116(4): 499-509.
- Celeste A, Fernandez-Capetillo O, Kruhlak MJ, Pilch DR, Staudt DW et al. (2003) Histone H2AX phosphorylation is dispensable for the initial recognition of DNA breaks. *Nat Cell Biol* 5(7): 675-679.
- Chadwick BP, Willard HF (2001) Histone H2A variants and the inactive X chromosome: identification of a second macroH2A variant. *Hum Mol Genet* 10(10): 1101-1113.
- Chakalova L, Carter D, Debrand E, Goyenechea B, Horton A et al. (2005) Developmental regulation of the beta-globin gene locus. *Prog Mol Subcell Biol* 38: 183-206.
- Chambeyron S, Da Silva NR, Lawson KA, Bickmore WA (2005) Nuclear re-organisation of the Hoxb complex during mouse embryonic development. *Development* 132(9): 2215-2223.
- Chen D, Ma H, Hong H, Koh SS, Huang SM et al. (1999) Regulation of transcription by a protein methyltransferase. *Science* 284(5423): 2174-2177.
- Cho DH, Thienes CP, Mahoney SE, Analau E, Filippova GN et al. (2005) Antisense transcription and heterochromatin at the DM1 CTG repeats are constrained by CTCF. *Mol Cell* 20(3): 483-489.
- Chow CM, Georgiou A, Szutorisz H, Maia e Silva A, Pombo A et al. (2005) Variant histone H3.3 marks promoters of transcriptionally active genes during mammalian cell division. *EMBO Rep* 6(4): 354-360.
- Christensen J, Agger K, Cloos PA, Pasini D, Rose S et al. (2007) RBP2 belongs to a family of demethylases, specific for tri- and dimethylated lysine 4 on histone 3. *Cell* 128(6): 1063-1076.
- Chung JH, Whiteley M, Felsenfeld G (1993) A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in *Drosophila*. *Cell* 74(3): 505-514.
- Cloos PA, Christensen J, Agger K, Maiolica A, Rappsilber J et al. (2006) The putative oncogene GASC1 demethylates tri- and dimethylated lysine 9 on histone H3. *Nature* 442(7100): 307-311.
- Cockerill PN (2000) Identification of DNaseI hypersensitive sites within nuclei. *Methods Mol Biol* 130: 29-46.
- Collins FS, Green ED, Guttmacher AE, Guyer MS (2003) A vision for the future of genomics research. *Nature* 422(6934): 835-847.
- Cora D, Herrmann C, Dieterich C, Di Cunto F, Provero P et al. (2005) Ab initio identification of putative human transcription factor binding sites by comparative genomics. *BMC Bioinformatics* 6: 110.
- Cosgrove MS (2007) Histone proteomics and the epigenetic regulation of nucleosome mobility. *Expert Rev Proteomics* 4(4): 465-478.

- Costanzi C, Stein P, Worrada DM, Schultz RM, Pehrson JR (2000) Histone macroH2A1 is concentrated in the inactive X chromosome of female preimplantation mouse embryos. *Development* 127(11): 2283-2289.
- Craig JM (2005) Heterochromatin--many flavours, common themes. *Bioessays* 27(1): 17-28.
- Crawford GE, Holt IE, Mullikin JC, Tai D, Blakesley R et al. (2004) Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proc Natl Acad Sci U S A* 101(4): 992-997.
- Crawford GE, Davis S, Scacheri PC, Renaud G, Halawi MJ et al. (2006) DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat Methods* 3(7): 503-509.
- Crawford GE, Holt IE, Whittle J, Webb BD, Tai D et al. (2006 b) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res* 16(1): 123-131.
- Cuthbert GL, Daujat S, Snowden AW, Erdjument-Bromage H, Hagiwara T et al. (2004) Histone deimination antagonizes arginine methylation. *Cell* 118(5): 545-553.
- Czermin B, Melfi R, McCabe D, Seitz V, Imhof A et al. (2002) Drosophila enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites. *Cell* 111(2): 185-196.
- Dahl JA, Collas P (2007) Q2ChIP, a quick and quantitative chromatin immunoprecipitation assay, unravels epigenetic dynamics of developmentally regulated genes in human carcinoma cells. *Stem Cells* 25(4): 1037-1046.
- Damelin M, Simon I, Moy TI, Wilson B, Komili S et al. (2002) The genome-wide localization of Rsc9, a component of the RSC chromatin-remodeling complex, changes in response to stress. *Mol Cell* 9(3): 563-573.
- Dannenbergh JH, David G, Zhong S, van der Torre J, Wong WH et al. (2005) mSin3A corepressor regulates diverse transcriptional networks governing normal and neoplastic growth and survival. *Genes Dev* 19(13): 1581-1595.
- Davuluri RV, Grosse I, Zhang MQ (2001) Computational identification of promoters and first exons in the human genome. *Nat Genet* 29(4): 412-417.
- De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ et al. (2006) A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* 312(5777): 1215-1217.
- De Gobbi M, Anguita E, Hughes J, Sloane-Stanley JA, Sharpe JA et al. (2007) Tissue-specific histone modification and transcription factor binding in alpha globin gene expression. *Blood* 110(13): 4503-4510.
- De la Serna IL, Ohkawa Y, Imbalzano AN (2006) Chromatin remodelling in mammalian differentiation: lessons from ATP-dependent remodellers. *Nat Rev Genet* 7(6): 461-473.

- De Nadal E, Zapater M, Alepuz PM, Sumoy L, Mas G et al. (2004) The MAPK Hog1 recruits Rpd3 histone deacetylase to activate osmoresponsive genes. *Nature* 427(6972): 370-374.
- De Napoles M, Mermoud JE, Wakao R, Tang YA, Endoh M et al. (2004) Polycomb group proteins Ring1A/B link ubiquitylation of histone H2A to heritable gene silencing and X inactivation. *Dev Cell* 7(5): 663-676.
- Defossez PA, Kelly KF, Filion GJ, Perez-Torrado R, Magdinier F et al. (2005) The human enhancer blocker CTC-binding factor interacts with the transcription factor Kaiso. *J Biol Chem* 280(52): 43017-43023.
- Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing chromosome conformation. *Science* 295(5558): 1306-1311.
- Delabesse E, Ogilvy S, Chapman MA, Piltz SG, Gottgens B et al. (2005) Transcriptional regulation of the SCL locus: identification of an enhancer that targets the primitive erythroid lineage in vivo. *Mol Cell Biol* 25(12): 5215-5225.
- Denell RE, Frederick RD (1983) Homoeosis in *Drosophila*: a description of the Polycomb lethal syndrome. *Dev Biol* 97(1): 34-47.
- Dermitzakis ET, Clark AG (2002) Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol* 19(7): 1114-1121.
- Dhalluin C, Carlson JE, Zeng L, He C, Aggarwal AK et al. (1999) Structure and ligand of a histone acetyltransferase bromodomain. *Nature* 399(6735): 491-496.
- Dhami P (2005) The SCL gene and transcriptional control of haematopoiesis. PhD Thesis, University of Cambridge, UK.
- Dhami P, Coffey AJ, Abbs S, Vermeesch JR, Dumanski JP et al. (2005) Exon array CGH: detection of copy-number changes at the resolution of individual exons in the human genome. *Am J Hum Genet* 76(5): 750-762.
- Dhami P, A.W. Bruce, et al. (2008) Re-defining the regulatory landscape at the human SCL (TAL-1) locus. Submitted.
- Di Duca M, Oleggini R, Sanna-Cherchi S, Pasquali L, Di Donato A et al. (2006) Cis and trans regulatory elements in NPHS2 promoter: implications in proteinuria and progression of renal diseases. *Kidney Int* 70(7): 1332-1341.
- Dion MF, Altschuler SJ, Wu LF, Rando OJ (2005) Genomic characterization reveals a simple histone H4 acetylation code. *Proc Natl Acad Sci U S A* 102(15): 5501-5506.
- Donohoe ME, Zhang LF, Xu N, Shi Y, Lee JT (2007) Identification of a Ctfc cofactor, Yy1, for the X chromosome binary switch. *Mol Cell* 25(1): 43-56.
- Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL et al. (2006) Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* 16(10): 1299-1309.

- Dou Y, Milne TA, Tackett AJ, Smith ER, Fukuda A et al. (2005) Physical association and coordinate function of the H3 K4 methyltransferase MLL1 and the H4 K16 acetyltransferase MOF. *Cell* 121(6): 873-885.
- Down TA, Hubbard TJ (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res* 12(3): 458-461.
- Down TA, Hubbard TJ (2005) NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Res* 33(5): 1445-1453.
- Draper JS, Pigott C, Thomson JA, Andrews PW (2002) Surface antigens of human embryonic stem cells: changes upon differentiation in culture. *J Anat* 200(Pt 3): 249-258.
- Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM (1999) Expression profiling using cDNA microarrays. *Nat Genet* 21(1 Suppl): 10-14.
- Dunn KL, Zhao H, Davie JR (2003) The insulator binding protein CTCF associates with the nuclear matrix. *Exp Cell Res* 288(1): 218-223.
- Eichler EE, Sankoff D (2003) Structural dynamics of eukaryotic chromosome evolution. *Science* 301(5634): 793-797.
- Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95(25): 14863-14868.
- Elnitski L, Giardine B, Shah P, Zhang Y, Riemer C et al. (2005) Improvements to GALA and dbERGE II: databases featuring genomic sequence alignment, annotation and experimental results. *Nucleic Acids Res* 33(Database issue): D466-470.
- Emanuelsson O, Nagalakshmi U, Zheng D, Rozowsky JS, Urban AE et al. (2007) Assessing the performance of different high-density tiling microarray strategies for mapping transcribed regions of the human genome. *Genome Res* 17(6): 886-897.
- Engel N, West AG, Felsenfeld G, Bartolomei MS (2004) Antagonism between DNA hypermethylation and enhancer-blocking activity at the H19 DMD is uncovered by CpG mutations. *Nat Genet* 36(8): 883-888.
- Enver T, Soneji S, Joshi C, Brown J, Iborra F et al. (2005) Cellular differentiation hierarchies in normal and culture-adapted human embryonic stem cells. *Hum Mol Genet* 14(21): 3129-3140.
- Euskirchen G, Royce TE, Bertone P, Martone R, Rinn JL et al. (2004) CREB binds to multiple loci on human chromosome 22. *Mol Cell Biol* 24(9): 3804-3814.
- Evans MJ, Kaufman MH (1981) Establishment in culture of pluripotential cells from mouse embryos. *Nature* 292(5819): 154-156.
- Farrell CM, West AG, Felsenfeld G (2002) Conserved CTCF insulator elements flank the mouse and human beta-globin loci. *Mol Cell Biol* 22(11): 3820-3831.
- Feinberg AP, Ohlsson R, Henikoff S (2006) The epigenetic progenitor origin of human cancer. *Nat Rev Genet* 7(1): 21-33.

- Feng Q, Wang H, Ng HH, Erdjument-Bromage H, Tempst P et al. (2002) Methylation of H3-lysine 79 is mediated by a new family of HMTases without a SET domain. *Curr Biol* 12(12): 1052-1058.
- Fickett JW, Hatzigeorgiou AG (1997) Eukaryotic promoter recognition. *Genome Res* 7(9): 861-878.
- Fiegler H, Carr P, Douglas EJ, Burford DC, Hunt S et al. (2003) DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. *Genes Chromosomes Cancer* 36(4): 361-374.
- Filippova GN, Fagerlie S, Klenova EM, Myers C, Dehner Y et al. (1996) An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Mol Cell Biol* 16(6): 2802-2813.
- Filippova GN, Thienes CP, Penn BH, Cho DH, Hu YJ et al. (2001) CTCF-binding sites flank CTG/CAG repeats and form a methylation-sensitive insulator at the DM1 locus. *Nat Genet* 28(4): 335-343.
- Filippova GN, Cheng MK, Moore JM, Truong JP, Hu YJ et al. (2005) Boundaries between chromosomal domains of X inactivation and escape bind CTCF and lack CpG methylation during early development. *Dev Cell* 8(1): 31-42.
- Filippova GN (2008) Genetics and epigenetics of the multifunctional protein CTCF. *Curr Top Dev Biol* 80: 337-360.
- Fischle W, Wang Y, Allis CD (2003) Binary switches and modification cassettes in histone biology and beyond. *Nature* 425(6957): 475-479.
- Fischle W, Tseng BS, Dormann HL, Ueberheide BM, Garcia BA et al. (2005) Regulation of HP1-chromatin binding by histone H3 methylation and phosphorylation. *Nature* 438(7071): 1116-1122.
- Flanagan JF, Mi LZ, Chruszcz M, Cymborowski M, Clines KL et al. (2005) Double chromodomains cooperate to recognize the methylated histone H3 tail. *Nature* 438(7071): 1181-1185.
- Flicek P, Keibler E, Hu P, Korf I, Brent MR (2003) Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. *Genome Res* 13(1): 46-54.
- Fogel GB, Weekes DG, Varga G, Dow ER, Craven AM et al. (2005) A statistical analysis of the TRANSFAC database. *Biosystems* 81(2): 137-154.
- Follows GA, Tagoh H, Lefevre P, Morgan GJ, Bonifer C (2003) Differential transcription factor occupancy but evolutionarily conserved chromatin features at the human and mouse M-CSF (CSF-1) receptor loci. *Nucleic Acids Res* 31(20): 5805-5816.
- Follows GA, Dhami P, Gottgens B, Bruce AW, Campbell PJ et al. (2006) Identifying gene regulatory elements by genomic microarray mapping of DNaseI hypersensitive sites. *Genome Res* 16(10): 1310-1319.
- Forrester WC, Fernandez LA, Grosschedl R (1999) Nuclear matrix attachment regions antagonize methylation-dependent repression of long-range enhancer-promoter interactions. *Genes Dev* 13(22): 3003-3014.
- Francis NJ, Kingston RE, Woodcock CL (2004) Chromatin compaction by a polycomb group protein complex. *Science* 306(5701): 1574-1577.

- Friedman AD (2007) Transcriptional control of granulocyte and monocyte development. *Oncogene* 26(47): 6816-6828.
- Frith MC, Pheasant M, Mattick JS (2005) The amazing complexity of the human transcriptome. *Eur J Hum Genet* 13(8): 894-897.
- Galas DJ, Schmitz A (1978) DNase footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res* 5(9): 3157-3170.
- Garcia BA, Shabanowitz J, Hunt DF (2007) Characterization of histones and their post-translational modifications by mass spectrometry. *Curr Opin Chem Biol* 11(1): 66-73.
- Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. *J Mol Biol* 196(2): 261-282.
- Garner MM, Revzin A (1981) A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the *Escherichia coli* lactose operon regulatory system. *Nucleic Acids Res* 9(13): 3047-3060.
- Gaszner M, Felsenfeld G (2006) Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet* 7(9): 703-713.
- Gautier T, Abbott DW, Molla A, Verdel A, Ausio J et al. (2004) Histone variant H2ABbd confers lower stability to the nucleosome. *EMBO Rep* 5(7): 715-720.
- Gavin I, Horn PJ, Peterson CL (2001) SWI/SNF chromatin remodeling requires changes in DNA topology. *Mol Cell* 7(1): 97-104.
- Geiger JH, Hahn S, Lee S, Sigler PB (1996) Crystal structure of the yeast TFIIA/TBP/DNA complex. *Science* 272(5263): 830-836.
- Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5(10): R80.
- Gershenson NI, Ioshikhes IP (2005) Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. *Bioinformatics* 21(8): 1295-1300.
- Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J et al. (2007) What is a gene, post-ENCODE? History and updated definition. *Genome Res* 17(6): 669-681.
- Geyer PK, Spana C, Corces VG (1986) On the molecular mechanism of gypsy-induced mutations at the yellow locus of *Drosophila melanogaster*. *EMBO J* 5(10): 2657-2662.
- Ghioni P, D'Alessandra Y, Mansueto G, Jaffray E, Hay RT et al. (2005) The protein stability and transcriptional activity of p63alpha are regulated by SUMO-1 conjugation. *Cell Cycle* 4(1): 183-190.
- Gibbons FD, Proft M, Struhl K, Roth FP (2005) Chipper: discovering transcription-factor targets from chromatin immunoprecipitation microarrays using variance stabilization. *Genome Biol* 6(11): R96.
- Gilbert N, Boyle S, Fiegler H, Woodfine K, Carter NP et al. (2004) Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell* 118(5): 555-566.



- Gilbert DM (2007) Replication origin plasticity, Taylor-made: inhibition vs recruitment of origins under conditions of replication stress. *Chromosoma* 116(4): 341-347.
- Giresi PG, Kim J, McDaniel RM, Iyer VR, Lieb JD (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* 17(6): 877-885.
- Golderer G, Grobner P (1991) ADP-ribosylation of core histones and their acetylated subspecies. *Biochem J* 277 ( Pt 3): 607-610.
- Goldknopf IL, Taylor CW, Baum RM, Yeoman LC, Olson MO et al. (1975) Isolation and characterization of protein A24, a "histone-like" non-histone chromosomal protein. *J Biol Chem* 250(18): 7182-7187.
- Goodstadt L, Ponting CP (2006) Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput Biol* 2(9): e133.
- Gottgens B, McLaughlin F, Bockamp EO, Fordham JL, Begley CG et al. (1997) Transcription of the SCL gene in erythroid and CD34 positive primitive myeloid cells is controlled by a complex network of lineage-restricted chromatin-dependent and chromatin-independent regulatory elements. *Oncogene* 15(20): 2419-2428.
- Gottgens B, Barton LM, Gilbert JG, Bench AJ, Sanchez MJ et al. (2000) Analysis of vertebrate SCL loci identifies conserved enhancers. *Nat Biotechnol* 18(2): 181-186.
- Gottgens B, Gilbert JG, Barton LM, Grafham D, Rogers J et al. (2001) Long-range comparison of human and mouse SCL loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences. *Genome Res* 11(1): 87-97.
- Gottgens B, Nastos A, Kinston S, Piltz S, Delabesse EC et al. (2002) Establishing the transcriptional programme for blood: the SCL stem cell enhancer is regulated by a multiprotein complex containing Ets and GATA factors. *EMBO J* 21(12): 3039-3050.
- Greenbaum JA, Parker SC, Tullius TD (2007) Detection of DNA structural motifs in functional genomic elements. *Genome Res* 17(6): 940-946.
- Grewal SI, Moazed D (2003) Heterochromatin and epigenetic control of gene expression. *Science* 301(5634): 798-802.
- Grewal SI, Jia S (2007) Heterochromatin revisited. *Nat Rev Genet* 8(1): 35-46.
- Gross DS, Garrard WT (1988) Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* 57: 159-197.
- Grosveld F, van Assendelft GB, Greaves DR, Kollias G (1987) Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell* 51(6): 975-985.
- Gruzdeva N, Kyrchanova O, Parshikov A, Kullyev A, Georgiev P (2005) The Mcp element from the bithorax complex contains an insulator that is capable of pairwise interactions and can facilitate enhancer-promoter communication. *Mol Cell Biol* 25(9): 3682-3689.

- Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA (2007) A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* 130(1): 77-88.
- Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J et al. (2006) Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* 124(1): 47-59.
- Hampsey M (1998) Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiol Mol Biol Rev* 62(2): 465-503.
- Hannich JT, Lewis A, Kroetz MB, Li SJ, Heide H et al. (2005) Defining the SUMO-modified proteome by multiple approaches in *Saccharomyces cerevisiae*. *J Biol Chem* 280(6): 4102-4110.
- Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM et al. (2000) CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* 405(6785): 486-489.
- Harris MB, Mostecky J, Rothman PB (2005) Repression of an interleukin-4-responsive promoter requires cooperative BCL-6 function. *J Biol Chem* 280(13): 13114-13121.
- Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK et al. (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol* 7 Suppl 1: S4 1-9.
- Hassa PO, Haenni SS, Elser M, Hottiger MO (2006) Nuclear ADP-ribosylation reactions in mammalian cells: where are we today and where are we going? *Microbiol Mol Biol Rev* 70(3): 789-829.
- Hassan AH, Prochasson P, Neely KE, Galasinski SC, Chandy M et al. (2002) Function and selectivity of bromodomains in anchoring chromatin-modifying complexes to promoter nucleosomes. *Cell* 111(3): 369-379.
- Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW et al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 39(3): 311-318.
- Heinzel T, Lavinsky RM, Mullen TM, Soderstrom M, Laherty CD et al. (1997) A complex containing N-CoR, mSin3 and histone deacetylase mediates transcriptional repression. *Nature* 387(6628): 43-48.
- Henderson JK, Draper JS, Baillie HS, Fishel S, Thomson JA et al. (2002) Preimplantation human embryos and embryonic stem cells show comparable expression of stage-specific embryonic antigens. *Stem Cells* 20(4): 329-337.
- Hendrich B, Bird A (1998) Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol Cell Biol* 18(11): 6538-6547.
- Heng HH, Goetze S, Ye CJ, Liu G, Stevens JB et al. (2004) Chromatin loops are selectively anchored using scaffold/matrix-attachment regions. *J Cell Sci* 117(Pt 7): 999-1008.
- Higgs DR, Wood WG, Jarman AP, Sharpe J, Lida J et al. (1990) A major positive regulatory region located far upstream of the human alpha-globin gene locus. *Genes Dev* 4(9): 1588-1601.
- Hoheisel JD (2006) Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet* 7(3): 200-210.

- Holohan EE, Kwong C, Adryan B, Bartkuhn M, Herold M et al. (2007) CTCF Genomic Binding Sites in *Drosophila* and the Organisation of the Bithorax Complex. *PLoS Genet* 3(7): e112.
- Honda BM, Dixon GH, Candido EP (1975) Sites of in vivo histone methylation in developing trout testis. *J Biol Chem* 250(22): 8681-8685.
- Hong S, Cho YW, Yu LR, Yu H, Veenstra TD et al. (2007) Identification of JmjC domain-containing UTX and JMJD3 as histone H3 lysine 27 demethylases. *Proc Natl Acad Sci U S A* 104(47): 18439-18444.
- Horak CE, Mahajan MC, Luscombe NM, Gerstein M, Weissman SM et al. (2002) GATA-1 binding sites mapped in the beta-globin locus by using mammalian ChIP-chip analysis. *Proc Natl Acad Sci U S A* 99(5): 2924-2929.
- Huang Y, Myers SJ, Dingledine R (1999) Transcriptional repression by REST: recruitment of Sin3A and histone deacetylase to neuronal genes. *Nat Neurosci* 2(10): 867-872.
- Huang S, Litt M, Felsenfeld G (2005) Methylation of histone H4 by arginine methyltransferase PRMT1 is essential in vivo for many subsequent histone modifications. *Genes Dev* 19(16): 1885-1893.
- Huang Y, Fang J, Bedford MT, Zhang Y, Xu RM (2006) Recognition of histone H3 lysine-4 methylation by the double tudor domain of JMJD2A. *Science* 312(5774): 748-751.
- Huang S, Li X, Yusufzai TM, Qiu Y, Felsenfeld G (2007) USF1 recruits histone modification complexes and is critical for maintenance of a chromatin barrier. *Mol Cell Biol* 27(22): 7991-8002.
- Hughes JD, Estep PW, Tavazoie S, Church GM (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol* 296(5): 1205-1214.
- Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ et al. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* 19(4): 342-347.
- Im H, Park C, Feng Q, Johnson KD, Kiekhäfer CM et al. (2003) Dynamic regulation of histone H3 methylated at lysine 79 within a tissue-specific chromatin domain. *J Biol Chem* 278(20): 18346-18352.
- Impey S, McCorkle SR, Cha-Molstad H, Dwyer JM, Yochum GS et al. (2004) Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions. *Cell* 119(7): 1041-1054.
- Ioudinkova E, Petrov A, Razin SV, Vassetzky YS (2005) Mapping long-range chromatin organization within the chicken alpha-globin gene domain using oligonucleotide DNA arrays. *Genomics* 85(1): 143-151.
- Ishihara K, Oshimura M, Nakao M (2006) CTCF-dependent chromatin insulator is linked to epigenetic remodeling. *Mol Cell* 23(5): 733-742.
- Ishii K, Arib G, Lin C, Van Houwe G, Laemmli UK (2002) Chromatin boundaries in budding yeast: the nuclear pore connection. *Cell* 109(5): 551-562.
- Ivanova NB, Dimos JT, Schaniel C, Hackney JA, Moore KA et al. (2002) A stem cell molecular signature. *Science* 298(5593): 601-604.

- Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M et al. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 409(6819): 533-538.
- Jarman AP, Wood WG, Sharpe JA, Gourdon G, Ayyub H et al. (1991) Characterization of the major regulatory element upstream of the human alpha-globin gene cluster. *Mol Cell Biol* 11(9): 4679-4689.
- Jeffreys AJ, Kauppi L, Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* 29(2): 217-222.
- Jenke AC, Stehle IM, Herrmann F, Eisenberger T, Baiker A et al. (2004) Nuclear scaffold/matrix attached region modules linked to a transcription unit are sufficient for replication and maintenance of a mammalian episome. *Proc Natl Acad Sci U S A* 101(31): 11322-11327.
- Jenuwein T, Forrester WC, Fernandez-Herrero LA, Laible G, Dull M et al. (1997) Extension of chromatin accessibility by nuclear matrix attachment regions. *Nature* 385(6613): 269-272.
- Jenuwein T, Allis CD (2001) Translating the histone code. *Science* 293(5532): 1074-1080.
- Jeon Y, Bekiranov S, Karnani N, Kapranov P, Ghosh S et al. (2005) Temporal profile of replication of human chromosomes. *Proc Natl Acad Sci U S A* 102(18): 6419-6424.
- Johansen KM, Johansen J (2006) Regulation of chromatin structure by histone H3S10 phosphorylation. *Chromosome Res* 14(4): 393-404.
- Johnson KD, Christensen HM, Zhao B, Bresnick EH (2001) Distinct mechanisms control RNA polymerase II recruitment to a tissue-specific locus control region and a downstream promoter. *Mol Cell* 8(2): 465-471.
- Johnson KD, Grass JA, Boyer ME, Kiekhäfer CM, Blobel GA et al. (2002) Cooperative activities of hematopoietic regulators recruit RNA polymerase II to a tissue-specific chromatin domain. *Proc Natl Acad Sci U S A* 99(18): 11760-11765.
- Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316(5830): 1497-1502.
- Jones N (1990) Transcriptional regulation by dimerization: two sides to an incestuous relationship. *Cell* 61(1): 9-11.
- Jones PL, Veenstra GJ, Wade PA, Vermaak D, Kass SU et al. (1998) Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat Genet* 19(2): 187-191.
- Jones PA, Baylin SB (2007) The epigenomics of cancer. *Cell* 128(4): 683-692.
- Kadonaga JT (2002) The DPE, a core promoter element for transcription by RNA polymerase II. *Exp Mol Med* 34(4): 259-264.
- Kadota M, Yang HH, Hu N, Wang C, Hu Y et al. (2007) Allele-specific chromatin immunoprecipitation studies show genetic influence on chromatin state in human genome. *PLoS Genet* 3(5): e81.
- Kamakaka RT, Biggins S (2005) Histone variants: deviants? *Genes Dev* 19(3): 295-310.

- Kanduri C, Pant V, Loukinov D, Pugacheva E, Qi CF et al. (2000) Functional association of CTCF with the insulator upstream of the H19 gene is parent of origin-specific and methylation-sensitive. *Curr Biol* 10(14): 853-856.
- Kanduri M, Kanduri C, Mariano P, Vostrov AA, Quitschke W et al. (2002) Multiple nucleosome positioning sites regulate the CTCF-mediated insulator function of the H19 imprinting control region. *Mol Cell Biol* 22(10): 3339-3344.
- Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R et al. (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316(5830): 1484-1488.
- Karnani N, Taylor C, Malhotra A, Dutta A (2007) Pan-S replication patterns and chromosomal domains defined by genome-tiling arrays of ENCODE genomic areas. *Genome Res* 17(6): 865-876.
- Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M et al. (2005) Antisense transcription in the mammalian transcriptome. *Science* 309(5740): 1564-1566.
- Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV et al. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res* 31(13): 3576-3579.
- Keogh MC, Kurdistani SK, Morris SA, Ahn SH, Podolny V et al. (2005) Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex. *Cell* 123(4): 593-605.
- Khorasanizadeh S (2004) The nucleosome: from genomic organization to genomic regulation. *Cell* 116(2): 259-272.
- Kim M, Park CH, Lee MS, Carlson BA, Hatfield DL et al. (2003) A novel TBP-interacting zinc finger protein represses transcription by inhibiting the recruitment of TFIIA and TFIIB. *Biochem Biophys Res Commun* 306(1): 231-238.
- Kim A, Dean A (2004) Developmental stage differences in chromatin subdomains of the beta-globin locus. *Proc Natl Acad Sci U S A* 101(18): 7028-7033.
- Kim TH, Barrera LO, Zheng M, Qu C, Singer MA et al. (2005) A high-resolution map of active promoters in the human genome. *Nature* 436(7052): 876-880.
- Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI et al. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* 128(6): 1231-1245.
- King DC, Taylor J, Zhang Y, Cheng Y, Lawson HA et al. (2007) Finding cis-regulatory elements using comparative genomics: some lessons from ENCODE data. *Genome Res* 17(6): 775-786.
- Kirmizis A, Bartley SM, Kuzmichev A, Margueron R, Reinberg D et al. (2004) Silencing of human polycomb target genes is associated with methylation of histone H3 Lys 27. *Genes Dev* 18(13): 1592-1605.
- Kirmizis A, Santos-Rosa H, Penkett CJ, Singer MA, Vermeulen M et al. (2007) Arginine methylation at histone H3R2 controls deposition of H3K4 trimethylation. *Nature* 449(7164): 928-932.
- Kleff S, Andrulis ED, Anderson CW, Sternglanz R (1995) Identification of a gene encoding a yeast histone H4 acetyltransferase. *J Biol Chem* 270(42): 24674-24677.

- Klenova EM, Nicolas RH, Paterson HF, Carne AF, Heath CM et al. (1993) CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. *Mol Cell Biol* 13(12): 7612-7624.
- Klenova EM, Nicolas RH, U S, Carne AF, Lee RE et al. (1997) Molecular weight abnormalities of the CTCF transcription factor: CTCF migrates aberrantly in SDS-PAGE and the size of the expressed protein is affected by the UTRs and sequences within the coding region of the CTCF gene. *Nucleic Acids Res* 25(3): 466-474.
- Klochkov D, Rincon-Arango H, Ioudinkova ES, Valadez-Graham V, Gavrilov A et al. (2006) A CTCF-dependent silencer located in the differentially methylated area may regulate expression of a housekeeping gene overlapping a tissue-specific gene domain. *Mol Cell Biol* 26(5): 1589-1597.
- Klose RJ, Bird AP (2006) Genomic DNA methylation: the mark and its mediators. *Trends Biochem Sci* 31(2): 89-97.
- Klose RJ, Gardner KE, Liang G, Erdjument-Bromage H, Tempst P et al. (2007) Demethylation of histone H3K36 and H3K9 by Rph1: a vestige of an H3K9 methylation system in *Saccharomyces cerevisiae*? *Mol Cell Biol* 27(11): 3951-3961.
- Klug J (1997) Ku autoantigen is a potential major cause of nonspecific bands in electrophoretic mobility shift assays. *Biotechniques* 22(2): 212-214, 216.
- Koch CM, Andrews RM, Flicek P, Dillon SC, Karaoz U et al. (2007) The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res* 17(6): 691-707.
- Komarnitsky P, Cho EJ, Buratowski S (2000) Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. *Genes Dev* 14(19): 2452-2460.
- Kondrashov AS, Shabalina SA (2002) Classification of common conserved sequences in mammalian intergenic regions. *Hum Mol Genet* 11(6): 669-674.
- Kornberg RD (1974) Chromatin structure: a repeating unit of histones and DNA. *Science* 184(139): 868-871.
- Kornberg RD, Thomas JO (1974) Chromatin structure; oligomers of the histones. *Science* 184(139): 865-868.
- Kornberg RD, Lorch Y (1999) Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* 98(3): 285-294.
- Kouzarides T (2007) Chromatin modifications and their function. *Cell* 128(4): 693-705.
- Krogan NJ, Kim M, Tong A, Golshani A, Cagney G et al. (2003) Methylation of histone H3 by Set2 in *Saccharomyces cerevisiae* is linked to transcriptional elongation by RNA polymerase II. *Mol Cell Biol* 23(12): 4207-4218.
- Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ et al. (2007) The UCSC genome browser database: update 2007. *Nucleic Acids Res* 35(Database issue): D668-673.

- Kumar PP, Bischof O, Purbey PK, Notani D, Urlaub H et al. (2007) Functional interaction between PML and SATB1 regulates chromatin-loop architecture and transcription of the MHC class I locus. *Nat Cell Biol* 9(1): 45-56.
- Kummerfeld SK, Teichmann SA (2006) DBD: a transcription factor prediction database. *Nucleic Acids Res* 34(Database issue): D74-81.
- Kurdistani SK, Robyr D, Tavazoie S, Grunstein M (2002) Genome-wide binding map of the histone deacetylase Rpd3 in yeast. *Nat Genet* 31(3): 248-254.
- Kurdistani SK, Grunstein M (2003) Histone acetylation and deacetylation in yeast. *Nat Rev Mol Cell Biol* 4(4): 276-284.
- Kurdistani SK, Tavazoie S, Grunstein M (2004) Mapping global histone acetylation patterns to gene expression. *Cell* 117(6): 721-733.
- Kurukuti S, Tiwari VK, Tavosidana G, Pugacheva E, Murrell A et al. (2006) CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. *Proc Natl Acad Sci U S A* 103(28): 10684-10689.
- Kuzmichev A, Nishioka K, Erdjument-Bromage H, Tempst P, Reinberg D (2002) Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. *Genes Dev* 16(22): 2893-2905.
- Kuzmichev A, Margueron R, Vaquero A, Preissner TS, Scher M et al. (2005) Composition and histone substrates of polycomb repressive group complexes change during cellular differentiation. *Proc Natl Acad Sci U S A* 102(6): 1859-1864.
- Lachner M, O'Carroll D, Rea S, Mechtler K, Jenuwein T (2001) Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* 410(6824): 116-120.
- Lachner M, O'Sullivan RJ, Jenuwein T (2003) An epigenetic road map for histone lysine methylation. *J Cell Sci* 116(Pt 11): 2117-2124.
- Lan F, Bayliss PE, Rinn JL, Whetstone JR, Wang JK et al. (2007) A histone H3 lysine 27 demethylase regulates animal posterior development. *Nature* 449(7163): 689-694.
- Lan F, Zaratiegui M, Villen J, Vaughn MW, Verdel A et al. (2007 b) *S. pombe* LSD1 homologs regulate heterochromatin propagation and euchromatic gene transcription. *Mol Cell* 26(1): 89-101.
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822): 860-921.
- Lang G, Mamalaki C, Greenberg D, Yannoutsos N, Kioussis D (1991) Deletion analysis of the human CD2 gene locus control region in transgenic mice. *Nucleic Acids Res* 19(21): 5851-5856.
- Lee TI, Young RA (2000) Transcription of eukaryotic protein-coding genes. *Annu Rev Genet* 34: 77-137.
- Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z et al. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298(5594): 799-804.

- Lee YH, Koh SS, Zhang X, Cheng X, Stallcup MR (2002 b) Synergy among nuclear receptor coactivators: selective requirement for protein methyltransferase and acetyltransferase activities. *Mol Cell Biol* 22(11): 3621-3632.
- Lee CK, Shibata Y, Rao B, Strahl BD, Lieb JD (2004) Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat Genet* 36(8): 900-905.
- Lee DY, Teyssier C, Strahl BD, Stallcup MR (2005) Role of protein methylation in regulation of transcription. *Endocr Rev* 26(2): 147-170.
- Lee MG, Wynder C, Cooch N, Shiekhattar R (2005) An essential role for CoREST in nucleosomal histone 3 lysine 4 demethylation. *Nature* 437(7057): 432-435.
- Lee JH, Skalnik DG (2005) CpG-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex. *J Biol Chem* 280(50): 41725-41731.
- Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS et al. (2006) Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* 125(2): 301-313.
- Lee N, Zhang J, Klose RJ, Erdjument-Bromage H, Tempst P et al. (2007) The trithorax-group protein Lid is a histone H3 trimethyl-Lys4 demethylase. *Nat Struct Mol Biol* 14(4): 341-343.
- Lee JS, Shilatifard A (2007) A site to remember: H3K36 methylation a mark for histone deacetylation. *Mutat Res* 618(1-2): 130-134.
- Leroy-Viard K, Vinit MA, Lecoite N, Mathieu-Mahul D, Romeo PH (1994) Distinct DNase-I hypersensitive sites are associated with TAL-1 transcription in erythroid and T-cell lines. *Blood* 84(11): 3819-3827.
- Lestrade L, Weber MJ (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res* 34(Database issue): D158-162.
- Levine SS, King IF, Kingston RE (2004) Division of labor in polycomb group repression. *Trends Biochem Sci* 29(9): 478-485.
- Lewis EB (1978) A gene complex controlling segmentation in *Drosophila*. *Nature* 276(5688): 565-570.
- Li Q, Peterson KR, Fang X, Stamatoyannopoulos G (2002) Locus control regions. *Blood* 100(9): 3077-3086.
- Li Z, Van Calcar S, Qu C, Cavenee WK, Zhang MQ et al. (2003) A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc Natl Acad Sci U S A* 100(14): 8164-8169.
- Li L, He S, Sun JM, Davie JR (2004) Gene regulation by Sp1 and Sp3. *Biochem Cell Biol* 82(4): 460-471.
- Li W, Meyer CA, Liu XS (2005) A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics* 21 Suppl 1: i274-282.
- Li H, Ilin S, Wang W, Duncan EM, Wysocka J et al. (2006) Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature* 442(7098): 91-95.
- Li B, Carey M, Workman JL (2007) The role of chromatin during transcription. *Cell* 128(4): 707-719.



- Li H, Fischle W, Wang W, Duncan EM, Liang L et al. (2007 b) Structural basis for lower lysine methylation state-specific readout by MBT repeats of L3MBTL1 and an engineered PHD finger. *Mol Cell* 28(4): 677-691.
- Li B, Gogol M, Carey M, Lee D, Seidel C et al. (2007 c) Combined action of PHD and chromo domains directs the Rpd3S HDAC to transcribed chromatin. *Science* 316(5827): 1050-1054.
- Liang G, Lin JC, Wei V, Yoo C, Cheng JC et al. (2004) Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. *Proc Natl Acad Sci U S A* 101(19): 7357-7362.
- Liang G, Klose RJ, Gardner KE, Zhang Y (2007) Yeast Jhd2p is a histone H3 Lys4 trimethyl demethylase. *Nat Struct Mol Biol* 14(3): 243-245.
- Lieb M, Rehmat S, Bhagwat AS (2001) Interaction of MutS and Vsr: some dominant-negative mutS mutations that disable methyladenine-directed mismatch repair are active in very-short-patch repair. *J Bacteriol* 183(21): 6487-6490.
- Lieu PT, Jozsi P, Gilles P, Peterson T (2005) Development of a DNA-labeling system for array-based comparative genomic hybridization. *J Biomol Tech* 16(2): 104-111.
- Ling J, Ainol L, Zhang L, Yu X, Pi W et al. (2004) HS2 enhancer function is blocked by a transcriptional terminator inserted between the enhancer and the promoter. *J Biol Chem* 279(49): 51704-51713.
- Ling JQ, Li T, Hu JF, Vu TH, Chen HL et al. (2006) CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1. *Science* 312(5771): 269-272.
- Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ (1999) High density synthetic oligonucleotide arrays. *Nat Genet* 21(1 Suppl): 20-24.
- Litt MD, Simpson M, Recillas-Targa F, Prioleau MN, Felsenfeld G (2001) Transitions in histone acetylation reveal boundaries of three separately regulated neighboring loci. *EMBO J* 20(9): 2224-2235.
- Liu CL, Schreiber SL, Bernstein BE (2003) Development and validation of a T7 based linear amplification for genomic DNA. *BMC Genomics* 4(1): 19.
- Liu CL, Kaplan T, Kim M, Buratowski S, Schreiber SL et al. (2005) Single-nucleosome mapping of histone modifications in *S. cerevisiae*. *PLoS Biol* 3(10): e328.
- Liu L, Zhang J, Bates S, Li JJ, Peehl DM et al. (2005 b) A methylation profile of in vitro immortalized human cell lines. *Int J Oncol* 26(1): 275-285.
- Lobanenkov VV, Nicolas RH, Adler VV, Paterson H, Klenova EM et al. (1990) A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene* 5(12): 1743-1753.
- Loh YH, Wu Q, Chew JL, Vega VB, Zhang W et al. (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet* 38(4): 431-440.
- Lohr D, Lopez J (1995) GAL4/GAL80-dependent nucleosome disruption/deposition on the upstream regions of the yeast GAL1-10 and GAL80 genes. *J Biol Chem* 270(46): 27671-27678.

- Lorincz MC, Dickerson DR, Schmitt M, Groudine M (2004) Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nat Struct Mol Biol* 11(11): 1068-1075.
- Lozzio BB, Lozzio CB (1979) Properties and usefulness of the original K-562 human myelogenous leukemia cell line. *Leuk Res* 3(6): 363-370.
- Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 389(6648): 251-260.
- Lunyak VV, Burgess R, Prefontaine GG, Nelson C, Sze SH et al. (2002) Corepressor-dependent silencing of chromosomal regions encoding neuronal genes. *Science* 298(5599): 1747-1752.
- Lusser A, Kadonaga JT (2003) Chromatin remodeling by ATP-dependent molecular machines. *Bioessays* 25(12): 1192-1200.
- Lutz M, Burke LJ, Barreto G, Goeman F, Greb H et al. (2000) Transcriptional repression by the insulator protein CTCF involves histone deacetylases. *Nucleic Acids Res* 28(8): 1707-1713.
- Maeda N, Kasukawa T, Oyama R, Gough J, Frith M et al. (2006) Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. *PLoS Genet* 2(4): e62.
- Mahadevan LC, Willis AC, Barratt MJ (1991) Rapid histone H3 phosphorylation in response to growth factors, phorbol esters, okadaic acid, and protein synthesis inhibitors. *Cell* 65(5): 775-783.
- Mak W, Baxter J, Silva J, Newall AE, Otte AP et al. (2002) Mitotically stable association of polycomb group proteins *eed* and *enx1* with the inactive X chromosome in trophoblast stem cells. *Curr Biol* 12(12): 1016-1020.
- Maniatis T, Goodbourn S, Fischer JA (1987) Regulation of inducible and tissue-specific gene expression. *Science* 236(4806): 1237-1245.
- Manza LL, Codreanu SG, Stamer SL, Smith DL, Wells KS et al. (2004) Global shifts in protein sumoylation in response to electrophile and oxidative stress. *Chem Res Toxicol* 17(12): 1706-1715.
- Mao DY, Watson JD, Yan PS, Barsyte-Lovejoy D, Khosravi F et al. (2003) Analysis of Myc bound loci identified by CpG island arrays shows that Max is essential for Myc-dependent repression. *Curr Biol* 13(10): 882-886.
- Margulies EH, Cooper GM, Asimenos G, Thomas DJ, Dewey CN et al. (2007) Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res* 17(6): 760-774.
- Marmorstein R (2003) Structure of SET domain proteins: a new twist on histone methylation. *Trends Biochem Sci* 28(2): 59-62.
- Martin GR (1981) Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proc Natl Acad Sci U S A* 78(12): 7634-7638.
- Martin N, Patel S, Segre JA (2004) Long-range comparison of human and mouse *Spr* loci to identify conserved noncoding sequences involved in coordinate regulation. *Genome Res* 14(12): 2430-2438.

- Martin D, Brun C, Remy E, Mouren P, Thieffry D et al. (2004) GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol* 5(12): R101.
- Martin C, Zhang Y (2005) The diverse functions of histone lysine methylation. *Nat Rev Mol Cell Biol* 6(11): 838-849.
- Martone R, Euskirchen G, Bertone P, Hartman S, Royce TE et al. (2003) Distribution of NF-kappaB-binding sites across human chromosome 22. *Proc Natl Acad Sci U S A* 100(21): 12247-12252.
- Maston GA, Evans SK, Green MR (2006) Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* 7: 29-59.
- Mattick JS, Makunin IV (2006) Non-coding RNA. *Hum Mol Genet* 15 Spec No 1: R17-29.
- Mattick JS (2007) A new paradigm for developmental biology. *J Exp Biol* 210(Pt 9): 1526-1547.
- Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S et al. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 34(Database issue): D108-110.
- Mautner J, Joos S, Werner T, Eick D, Bornkamm GW et al. (1995) Identification of two enhancer elements downstream of the human c-myc gene. *Nucleic Acids Res* 23(1): 72-80.
- McNairn AJ, Gilbert DM (2003) Epigenomic replication: linking epigenetics to DNA replication. *Bioessays* 25(7): 647-656.
- Meneghini MD, Wu M, Madhani HD (2003) Conserved histone variant H2A.Z protects euchromatin from the ectopic spread of silent heterochromatin. *Cell* 112(5): 725-736.
- Meshorer E, Yellajoshula D, George E, Scambler PJ, Brown DT et al. (2006) Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. *Dev Cell* 10(1): 105-116.
- Metzger E, Wissmann M, Yin N, Muller JM, Schneider R et al. (2005) LSD1 demethylates repressive histone marks to promote androgen-receptor-dependent transcription. *Nature* 437(7057): 436-439.
- Miao F, Natarajan R (2005) Mapping global histone methylation patterns in the coding regions of human genes. *Mol Cell Biol* 25(11): 4650-4661.
- Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448(7153): 553-560.
- Millar CB, Xu F, Zhang K, Grunstein M (2006) Acetylation of H2AZ Lys 14 is associated with genome-wide gene activity in yeast. *Genes Dev* 20(6): 711-722.
- Min J, Zhang Y, Xu RM (2003) Structural basis for specific binding of Polycomb chromodomain to histone H3 methylated at Lys 27. *Genes Dev* 17(15): 1823-1828.
- Mirkovitch J, Mirault ME, Laemmli UK (1984) Organization of the higher-order chromatin loop: specific DNA attachment sites on nuclear scaffold. *Cell* 39(1): 223-232.
- Mito Y, Henikoff JG, Henikoff S (2005) Genome-scale profiling of histone H3.3 replacement patterns. *Nat Genet* 37(10): 1090-1097.
- Mito Y, Henikoff JG, Henikoff S (2007) Histone replacement marks the boundaries of cis-regulatory domains. *Science* 315(5817): 1408-1411.

- Mizuguchi G, Shen X, Landry J, Wu WH, Sen S et al. (2004) ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. *Science* 303(5656): 343-348.
- Mizzen CA, Yang XJ, Kokubo T, Brownell JE, Bannister AJ et al. (1996) The TAF(II)250 subunit of TFIID has histone acetyltransferase activity. *Cell* 87(7): 1261-1270.
- Monier K, Mouradian S, Sullivan KF (2007) DNA methylation promotes Aurora-B-driven phosphorylation of histone H3 in chromosomal subdomains. *J Cell Sci* 120(Pt 1): 101-114.
- Moon H, Filippova G, Loukinov D, Pugacheva E, Chen Q et al. (2005) CTCF is conserved from *Drosophila* to humans and confers enhancer blocking of the Fab-8 insulator. *EMBO Rep* 6(2): 165-170.
- Morris SA, Rao B, Garcia BA, Hake SB, Diaz RL et al. (2007) Identification of histone H3 lysine 36 acetylation as a highly conserved histone modification. *J Biol Chem* 282(10): 7632-7640.
- Mukhopadhyay R, Yu W, Whitehead J, Xu J, Lezcano M et al. (2004) The binding sites for the chromatin insulator protein CTCF map to DNA methylation-free domains genome-wide. *Genome Res* 14(8): 1594-1602.
- Murray K (1964) The Occurrence of Epsilon-N-Methyl Lysine in Histones. *Biochemistry* 3: 10-15.
- Nagy PL, Cleary ML, Brown PO, Lieb JD (2003) Genomewide demarcation of RNA polymerase II transcription units revealed by physical fractionation of chromatin. *Proc Natl Acad Sci U S A* 100(11): 6364-6369.
- Nakayama J, Rice JC, Strahl BD, Allis CD, Grewal SI (2001) Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science* 292(5514): 110-113.
- Nallur GN, Prakash K, Weissman SM (1996) Multiplex selection technique (MuST): an approach to clone transcription factor binding sites. *Proc Natl Acad Sci U S A* 93(3): 1184-1189.
- Nan X, Ng HH, Johnson CA, Laherty CD, Turner BM et al. (1998) Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* 393(6683): 386-389.
- Narlikar GJ, Fan HY, Kingston RE (2002) Cooperation between complexes that regulate chromatin structure and transcription. *Cell* 108(4): 475-487.
- Nathan D, Ingvarsdottir K, Sterner DE, Bylebyl GR, Dokmanovic M et al. (2006) Histone sumoylation is a negative regulator in *Saccharomyces cerevisiae* and shows dynamic interplay with positive-acting histone modifications. *Genes Dev* 20(8): 966-976.
- Negre N, Hennetin J, Sun LV, Lavrov S, Bellis M et al. (2006) Chromosomal distribution of PcG proteins during *Drosophila* development. *PLoS Biol* 4(6): e170.
- Ng HH, Feng Q, Wang H, Erdjument-Bromage H, Tempst P et al. (2002) Lysine methylation within the globular domain of histone H3 by Dot1 is important for telomeric silencing and Sir protein association. *Genes Dev* 16(12): 1518-1527.

- Ng HH, Robert F, Young RA, Struhl K (2003) Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Mol Cell* 11(3): 709-719.
- Ng HH, Ciccone DN, Morshead KB, Oettinger MA, Struhl K (2003 b) Lysine-79 of histone H3 is hypomethylated at silenced loci in yeast and mammalian cells: a potential mechanism for position-effect variegation. *Proc Natl Acad Sci U S A* 100(4): 1820-1825.
- Nishida H, Suzuki T, Kondo S, Miura H, Fujimura Y et al. (2006) Histone H3 acetylated at lysine 9 in promoter is associated with low nucleosome density in the vicinity of transcription start site in human cell. *Chromosome Res* 14(2): 203-211.
- Nobrega MA, Ovcharenko I, Afzal V, Rubin EM (2003) Scanning human gene deserts for long-range enhancers. *Science* 302(5644): 413.
- Nowak SJ, Corces VG (2000) Phosphorylation of histone H3 correlates with transcriptionally active loci. *Genes Dev* 14(23): 3003-3013.
- Nowak SJ, Corces VG (2004) Phosphorylation of histone H3: a balancing act between chromosome condensation and transcriptional activation. *Trends Genet* 20(4): 214-220.
- Odom DT, Zizlsperger N, Gordon DB, Bell GW, Rinaldi NJ et al. (2004) Control of pancreas and liver gene expression by HNF transcription factors. *Science* 303(5662): 1378-1381.
- Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW et al. (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* 39(6): 730-732.
- Ogbourne S, Antalis TM (1998) Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochem J* 331 ( Pt 1): 1-14.
- Ohlsson R, Renkawitz R, Lobanenko V (2001) CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet* 17(9): 520-527.
- O'Neill LP, Turner BM (1995) Histone H4 acetylation distinguishes coding regions of the human genome from heterochromatin in a differentiation-dependent but transcription-independent manner. *EMBO J* 14(16): 3946-3957.
- O'Neill LP, Turner BM (2003) Immunoprecipitation of native chromatin: NChIP. *Methods* 31(1): 76-82.
- O'Neill LP, VerMilyea MD, Turner BM (2006) Epigenetic characterization of the early embryo with a chromatin immunoprecipitation protocol applicable to small cell populations. *Nat Genet* 38(7): 835-841.
- Orlando V, Paro R (1993) Mapping Polycomb-repressed domains in the bithorax complex using in vivo formaldehyde cross-linked chromatin. *Cell* 75(6): 1187-1198.
- Orlando V, Strutt H, Paro R (1997) Analysis of chromatin structure by in vivo formaldehyde cross-linking. *Methods* 11(2): 205-214.
- Orlando V (2000) Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem Sci* 25(3): 99-104.

- Orphanides G, Lagrange T, Reinberg D (1996) The general transcription factors of RNA polymerase II. *Genes Dev* 10(21): 2657-2683.
- Osborne CS, Chakalova L, Brown KE, Carter D, Horton A et al. (2004) Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat Genet* 36(10): 1065-1071.
- Pan G, S. Tian, et al. (2007) Whole-Genome Analysis of Histone H3 Lysine 4 and Lysine 27 Methylation in Human Embryonic Stem Cells. *Cell Stem Cell* 1: 299-312.
- Pant V, Kurukuti S, Pugacheva E, Shamsuddin S, Mariano P et al. (2004) Mutation of a single CTCF target site within the H19 imprinting control region leads to loss of Igf2 imprinting and complex patterns of de novo methylation upon maternal inheritance. *Mol Cell Biol* 24(8): 3497-3504.
- Park YJ, Chodaparambil JV, Bao Y, McBryant SJ, Luger K (2005) Nucleosome assembly protein 1 exchanges histone H2A-H2B dimers and assists nucleosome sliding. *J Biol Chem* 280(3): 1817-1825.
- Patrinos GP, de Krom M, de Boer E, Langeveld A, Imam AM et al. (2004) Multiple interactions between regulatory regions are required to stabilize an active chromatin hub. *Genes Dev* 18(12): 1495-1509.
- Patterson BD, Davies DD (1969) Specificity of the enzymatic methylation of pea histone. *Biochem Biophys Res Commun* 34(6): 791-794.
- Pearson JC, Lemons D, McGinnis W (2005) Modulating Hox gene functions during animal body patterning. *Nat Rev Genet* 6(12): 893-904.
- Pena PV, Davrazou F, Shi X, Walter KL, Verkhusha VV et al. (2006) Molecular mechanism of histone H3K4me3 recognition by plant homeodomain of ING2. *Nature* 442(7098): 100-103.
- Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA et al. (2006) In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444(7118): 499-502.
- Pennacchio LA, Loots GG, Nobrega MA, Ovcharenko I (2007) Predicting tissue-specific enhancers in the human genome. *Genome Res* 17(2): 201-211.
- Pezzolesi MG, Zbuk KM, Waite KA, Eng C (2007) Comparative genomic and functional analyses reveal a novel cis-acting PTEN regulatory element as a highly conserved functional E-box motif deleted in Cowden syndrome. *Hum Mol Genet* 16(9): 1058-1071.
- Pikaart MJ, Recillas-Targa F, Felsenfeld G (1998) Loss of transcriptional activity of a transgene is accompanied by DNA methylation and histone deacetylation and is prevented by insulators. *Genes Dev* 12(18): 2852-2862.
- Pimanda JE, Silberstein L, Dominici M, Dekel B, Bowen M et al. (2006) Transcriptional link between blood and bone: the stem cell leukemia gene and its +19 stem cell enhancer are active in bone cells. *Mol Cell Biol* 26(7): 2615-2625.
- Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM et al. (2005) Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* 122(4): 517-527.

- Prabhakar S, Poulin F, Shoukry M, Afzal V, Rubin EM et al. (2006) Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res* 16(7): 855-863.
- Prakash A, Tompa M (2005) Discovery of regulatory elements in vertebrates through comparative genomics. *Nat Biotechnol* 23(10): 1249-1256.
- Pray-Grant MG, Daniel JA, Schieltz D, Yates JR, 3rd, Grant PA (2005) Chd1 chromodomain links histone H3 methylation with SAGA- and SLIK-dependent acetylation. *Nature* 433(7024): 434-438.
- Prelich G (2002) RNA polymerase II carboxy-terminal domain kinases: emerging clues to their function. *Eukaryot Cell* 1(2): 153-162.
- Prendergast GC, Ziff EB (1989) DNA-binding motif. *Nature* 341(6241): 392.
- Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35(Database issue): D61-65.
- Puig O, Casparly F, Rigaut G, Rutz B, Bouveret E et al. (2001) The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* 24(3): 218-229.
- Quackenbush J (2002) Microarray data normalization and transformation. *Nat Genet* 32 Suppl: 496-501.
- Rada-Iglesias A, Enroth S, Ameer A, Koch CM, Clelland GK et al. (2007) Butyrate mediates decrease of histone acetylation centered on transcription start sites and down-regulation of associated genes. *Genome Res* 17(6): 708-719.
- Raghuraman MK, Winzeler EA, Collingwood D, Hunt S, Wodicka L et al. (2001) Replication dynamics of the yeast genome. *Science* 294(5540): 115-121.
- Ramalho-Santos M, Yoon S, Matsuzaki Y, Mulligan RC, Melton DA (2002) "Stemness": transcriptional profiling of embryonic and adult stem cells. *Science* 298(5593): 597-600.
- Rao B, Shibata Y, Strahl BD, Lieb JD (2005) Dimethylation of histone H3 at lysine 36 demarcates regulatory and nonregulatory chromatin genome-wide. *Mol Cell Biol* 25(21): 9447-9459.
- Rea S, Eisenhaber F, O'Carroll D, Strahl BD, Sun ZW et al. (2000) Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature* 406(6796): 593-599.
- Recillas-Targa F, Pikaart MJ, Burgess-Beusse B, Bell AC, Litt MD et al. (2002) Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities. *Proc Natl Acad Sci U S A* 99(10): 6883-6888.
- Reid JL, Iyer VR, Brown PO, Struhl K (2000) Coordinate regulation of yeast ribosomal protein genes is associated with targeted recruitment of Esa1 histone acetylase. *Mol Cell* 6(6): 1297-1307.
- Reik W, Dean W, Walter J (2001) Epigenetic reprogramming in mammalian development. *Science* 293(5532): 1089-1093.
- Reik W, Walter J (2001) Genomic imprinting: parental influence on the genome. *Nat Rev Genet* 2(1): 21-32.
- Reik W (2007) Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* 447(7143): 425-432.

- Reinke H, Horz W (2003) Histones are first hyperacetylated and then lose contact with the activated PHO5 promoter. *Mol Cell* 11(6): 1599-1607.
- Ren B, Maniatis T (1998) Regulation of *Drosophila* Adh promoter switching by an initiator-targeted repression mechanism. *EMBO J* 17(4): 1076-1086.
- Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG et al. (2000) Genome-wide location and function of DNA binding proteins. *Science* 290(5500): 2306-2309.
- Ren B, Cam H, Takahashi Y, Volkert T, Terragni J et al. (2002) E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev* 16(2): 245-256.
- Reubinoff BE, Pera MF, Fong CY, Trounson A, Bongso A (2000) Embryonic stem cell lines from human blastocysts: somatic differentiation in vitro. *Nat Biotechnol* 18(4): 399-404.
- Reya T, Morrison SJ, Clarke MF, Weissman IL (2001) Stem cells, cancer, and cancer stem cells. *Nature* 414(6859): 105-111.
- Richards EJ, Elgin SC (2002) Epigenetic codes for heterochromatin formation and silencing: rounding up the usual suspects. *Cell* 108(4): 489-500.
- Ringrose L, Ehret H, Paro R (2004) Distinct contributions of histone H3 lysine 9 and 27 methylation to locus-specific stability of polycomb complexes. *Mol Cell* 16(4): 641-653.
- Ringrose L, Paro R (2004) Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. *Annu Rev Genet* 38: 413-443.
- Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM et al. (2003) The transcriptional activity of human Chromosome 22. *Genes Dev* 17(4): 529-540.
- Robb L, Begley CG (1997) The SCL/TAL1 gene: roles in normal and malignant haematopoiesis. *Bioessays* 19(7): 607-613.
- Robert F, Pokholok DK, Hannett NM, Rinaldi NJ, Chandy M et al. (2004) Global position and recruitment of HATs and HDACs in the yeast genome. *Mol Cell* 16(2): 199-209.
- Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y et al. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 4(8): 651-657.
- Roby D, Suka Y, Xenarios I, Kurdistani SK, Wang A et al. (2002) Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases. *Cell* 109(4): 437-446.
- Roby D, Grunstein M (2003) Genomewide histone acetylation microarrays. *Methods* 31(1): 83-89.
- Roguev A, Schaft D, Shevchenko A, Pijnappel WW, Wilm M et al. (2001) The *Saccharomyces cerevisiae* Set1 complex includes an Ash2 homologue and methylates histone 3 lysine 4. *EMBO J* 20(24): 7137-7148.
- Roh TY, Ngau WC, Cui K, Landsman D, Zhao K (2004) High-resolution genome-wide mapping of histone modifications. *Nat Biotechnol* 22(8): 1013-1016.
- Roh TY, Cuddapah S, Zhao K (2005) Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev* 19(5): 542-552.



- Roh TY, Cuddapah S, Cui K, Zhao K (2006) The genomic landscape of histone modifications in human T cells. *Proc Natl Acad Sci U S A* 103(43): 15782-15787.
- Roh TY, Wei G, Farrell CM, Zhao K (2007) Genome-wide prediction of conserved and nonconserved enhancers by histone acetylation patterns. *Genome Res* 17(1): 74-81.
- Roth SY, Denu JM, Allis CD (2001) Histone acetyltransferases. *Annu Rev Biochem* 70: 81-120.
- Rudolph T, Yonezawa M, Lein S, Heidrich K, Kubicek S et al. (2007) Heterochromatin formation in *Drosophila* is initiated through active removal of H3K4 methylation by the LSD1 homolog SU(VAR)3-3. *Mol Cell* 26(1): 103-115.
- Ruthenburg AJ, Li H, Patel DJ, Allis CD (2007) Multivalent engagement of chromatin modifications by linked binding modules. *Nat Rev Mol Cell Biol* 8(12): 983-994.
- Sabo PJ, Humbert R, Hawrylycz M, Wallace JC, Dorschner MO et al. (2004) Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. *Proc Natl Acad Sci U S A* 101(13): 4537-4542.
- Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM et al. (2006) Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat Methods* 3(7): 511-518.
- Sanchez M, Gottgens B, Sinclair AM, Stanley M, Begley CG et al. (1999) An SCL 3' enhancer targets developing endothelium together with embryonic and adult haematopoietic progenitors. *Development* 126(17): 3891-3904.
- Sanchez MJ, Bockamp EO, Miller J, Gambardella L, Green AR (2001) Selective rescue of early haematopoietic progenitors in Scl(-/-) mice by expressing Scl under the control of a stem cell enhancer. *Development* 128(23): 4815-4827.
- Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 32(Database issue): D91-94.
- Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y et al. (2007) Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet* 8(6): 424-436.
- Sanders SL, Portoso M, Mata J, Bahler J, Allshire RC et al. (2004) Methylation of histone H4 lysine 20 controls recruitment of Crb2 to sites of DNA damage. *Cell* 119(5): 603-614.
- Santos-Rosa H, Schneider R, Bannister AJ, Sherriff J, Bernstein BE et al. (2002) Active genes are trimethylated at K4 of histone H3. *Nature* 419(6905): 407-411.
- Sarraf SA, Stancheva I (2004) Methyl-CpG binding protein MBD1 couples histone H3 methylation at lysine 9 by SETDB1 to DNA replication and chromatin assembly. *Mol Cell* 15(4): 595-605.
- Sassone-Corsi P, Mizzen CA, Cheung P, Crosio C, Monaco L et al. (1999) Requirement of Rsk-2 for epidermal growth factor-activated phosphorylation of histone H3. *Science* 285(5429): 886-891.
- Savas U, Hsu MH, Johnson EF (2003) Differential regulation of human CYP4A genes by peroxisome proliferators and dexamethasone. *Arch Biochem Biophys* 409(1): 212-220.

- Saxonov S, Berg P, Brutlag DL (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* 103(5): 1412-1417.
- Schaft D, Roguev A, Kotovic KM, Shevchenko A, Sarov M et al. (2003) The histone 3 lysine 36 methyltransferase, SET2, is involved in transcriptional elongation. *Nucleic Acids Res* 31(10): 2475-2482.
- Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270(5235): 467-470.
- Scherf M, Klingenhoff A, Werner T (2000) Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J Mol Biol* 297(3): 599-606.
- Schneider R, Bannister AJ, Myers FA, Thorne AW, Crane-Robinson C et al. (2004) Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. *Nat Cell Biol* 6(1): 73-77.
- Schotta G, Lachner M, Sarma K, Ebert A, Sengupta R et al. (2004) A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin. *Genes Dev* 18(11): 1251-1262.
- Schubeler D, Scalzo D, Kooperberg C, van Steensel B, Delrow J et al. (2002) Genome-wide DNA replication profile for *Drosophila melanogaster*: a link between transcription and replication timing. *Nat Genet* 32(3): 438-442.
- Schubeler D, MacAlpine DM, Scalzo D, Wirbelauer C, Kooperberg C et al. (2004) The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev* 18(11): 1263-1271.
- Schuettengruber B, Chourrout D, Vervoort M, Leblanc B, Cavalli G (2007) Genome regulation by polycomb and trithorax proteins. *Cell* 128(4): 735-745.
- Schurter BT, Koh SS, Chen D, Bunick GJ, Harp JM et al. (2001) Methylation of histone H3 by coactivator-associated arginine methyltransferase 1. *Biochemistry* 40(19): 5747-5756.
- Schwartz YB, Kahn TG, Nix DA, Li XY, Bourgon R et al. (2006) Genome-wide analysis of Polycomb targets in *Drosophila melanogaster*. *Nat Genet* 38(6): 700-705.
- Secombe J, Li L, Carlos L, Eisenman RN (2007) The Trithorax group protein Lid is a trimethyl histone H3K4 demethylase required for dMyc-induced cell growth. *Genes Dev* 21(5): 537-551.
- Sertil O, Vemula A, Salmon SL, Morse RH, Lowry CV (2007) Direct role for the Rpd3 complex in transcriptional induction of the anaerobic DAN/TIR genes in yeast. *Mol Cell Biol* 27(6): 2037-2047.
- Seta N, Kuwana M (2007) Human circulating monocytes as multipotential progenitors. *Keio J Med* 56(2): 41-47.
- Shahbazian MD, Grunstein M (2007) Functions of site-specific histone acetylation and deacetylation. *Annu Rev Biochem* 76: 75-100.

- Sharma VM, Tomar RS, Dempsey AE, Reese JC (2007) Histone deacetylases RPD3 and HOS2 regulate the transcriptional activation of DNA damage-inducible genes. *Mol Cell Biol* 27(8): 3199-3210.
- Sharpe JA, Summerhill RJ, Vyas P, Gourdon G, Higgs DR et al. (1993) Role of upstream DNase I hypersensitive sites in the regulation of human alpha globin gene expression. *Blood* 82(5): 1666-1671.
- Shi Y, Lan F, Matson C, Mulligan P, Whetstine JR et al. (2004) Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell* 119(7): 941-953.
- Shi X, Hong T, Walter KL, Ewalt M, Michishita E et al. (2006) ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression. *Nature* 442(7098): 96-99.
- Shiio Y, Eisenman RN (2003) Histone sumoylation is associated with transcriptional repression. *Proc Natl Acad Sci U S A* 100(23): 13225-13230.
- Shiio Y, Eisenman RN (2003) Histone sumoylation is associated with transcriptional repression. *Proc Natl Acad Sci U S A* 100(23): 13225-13230.
- Shilatifard A (2006) Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression. *Annu Rev Biochem* 75: 243-269.
- Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL et al. (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med* 8(1): 68-74.
- Shogren-Knaak M, Ishii H, Sun JM, Pazin MJ, Davie JR et al. (2006) Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science* 311(5762): 844-847.
- Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15(8): 1034-1050.
- Silverstein RA, Ekwall K (2005) Sin3: a flexible regulator of global gene expression and genome stability. *Curr Genet* 47(1): 1-17.
- Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R et al. (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* 38(11): 1348-1354.
- Sims RJ, 3rd, Belotserkovskaya R, Reinberg D (2004) Elongation by RNA polymerase II: the short and long of it. *Genes Dev* 18(20): 2437-2468.
- Sims RJ, 3rd, Chen CF, Santos-Rosa H, Kouzarides T, Patel SS et al. (2005) Human but not yeast CHD1 binds directly and selectively to histone H3 methylated at lysine 4 via its tandem chromodomains. *J Biol Chem* 280(51): 41789-41792.
- Sinclair AM, Gottgens B, Barton LM, Stanley ML, Pardanaud L et al. (1999) Distinct 5' SCL enhancers direct transcription to developing brain, spinal cord, and endothelium: neural expression is mediated by GATA factor binding sites. *Dev Biol* 209(1): 128-142.
- Singh-Gasson S, Green RD, Yue Y, Nelson C, Blattner F et al. (1999) Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat Biotechnol* 17(10): 974-978.

- Sinha I, Wiren M, Ekwall K (2006) Genome-wide patterns of histone modifications in fission yeast. *Chromosome Res* 14(1): 95-105.
- Smith AD, Sumazin P, Das D, Zhang MQ (2005) Mining ChIP-chip data for transcription factor and cofactor binding sites. *Bioinformatics* 21 Suppl 1: i403-412.
- Solomon MJ, Varshavsky A (1985) Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. *Proc Natl Acad Sci U S A* 82(19): 6470-6474.
- Solomon MJ, Larsen PL, Varshavsky A (1988) Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* 53(6): 937-947.
- Spilianakis CG, Lalioti MD, Town T, Lee GR, Flavell RA (2005) Interchromosomal associations between alternatively expressed loci. *Nature* 435(7042): 637-645.
- Spivakov M, Fisher AG (2007) Epigenetic signatures of stem-cell identity. *Nat Rev Genet* 8(4): 263-271.
- Splinter E, Heath H, Kooren J, Palstra RJ, Klous P et al. (2006) CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev* 20(17): 2349-2354.
- Sproul D, Gilbert N, Bickmore WA (2005) The role of chromatin structure in regulating the expression of clustered genes. *Nat Rev Genet* 6(10): 775-781.
- Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR et al. (2003) The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol* 1(2): E45.
- Stephanou A, Scarabelli TM, Townsend PA, Bell R, Yellon D et al. (2002) The carboxyl-terminal activation domain of the STAT-1 transcription factor enhances ischemia/reperfusion-induced apoptosis in cardiac myocytes. *FASEB J* 16(13): 1841-1843.
- Sterner DE, Berger SL (2000) Acetylation of histones and transcription-related factors. *Microbiol Mol Biol Rev* 64(2): 435-459.
- Strahl BD, Allis CD (2000) The language of covalent histone modifications. *Nature* 403(6765): 41-45.
- Strahl BD, Briggs SD, Brame CJ, Caldwell JA, Koh SS et al. (2001) Methylation of histone H4 at arginine 3 occurs in vivo and is mediated by the nuclear receptor coactivator PRMT1. *Curr Biol* 11(12): 996-1000.
- Straka C, Horz W (1991) A functional role for nucleosomes in the repression of a yeast promoter. *EMBO J* 10(2): 361-368.
- Suka N, Suka Y, Carmen AA, Wu J, Grunstein M (2001) Highly specific antibodies determine histone acetylation site usage in yeast heterochromatin and euchromatin. *Mol Cell* 8(2): 473-479.
- Sumer H, Craig JM, Sibson M, Choo KH (2003) A rapid method of genomic array analysis of scaffold/matrix attachment regions (S/MARs) identifies a 2.5-Mb region of enhanced scaffold/matrix attachment at a human neocentromere. *Genome Res* 13(7): 1737-1743.
- Sun FL, Elgin SC (1999) Putting boundaries on silence. *Cell* 99(5): 459-462.
- Sundstrom C, Nilsson K (1976) Establishment and characterization of a human histiocytic lymphoma cell line (U-937). *Int J Cancer* 17(5): 565-577.

- Surani MA, Hayashi K, Hajkova P (2007) Genetic and epigenetic regulators of pluripotency. *Cell* 128(4): 747-762.
- Szabo P, Tang SH, Rentsendorj A, Pfeifer GP, Mann JR (2000) Maternal-specific footprints at putative CTCF sites in the H19 imprinting control region give evidence for insulator function. *Curr Biol* 10(10): 607-610.
- Szutorisz H, Dillon N, Tora L (2005) The role of enhancers as centres for general transcription factor recruitment. *Trends Biochem Sci* 30(11): 593-599.
- Szutorisz H, Canzonetta C, Georgiou A, Chow CM, Tora L et al. (2005 b) Formation of an active tissue-specific chromatin domain initiated by epigenetic marking at the embryonic stem cell stage. *Mol Cell Biol* 25(5): 1804-1820.
- Taft RJ, Pheasant M, Mattick JS (2007) The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* 29(3): 288-299.
- Talasz H, Lindner HH, Sarg B, Helliger W (2005) Histone H4-lysine 20 monomethylation is increased in promoter and coding regions of active genes and correlates with hyperacetylation. *J Biol Chem* 280(46): 38814-38822.
- Tanimoto K, Liu Q, Bungert J, Engel JD (1999) Effects of altered gene order or orientation of the locus control region on human beta-globin gene expression in mice. *Nature* 398(6725): 344-348.
- Taverna SD, Li H, Ruthenburg AJ, Allis CD, Patel DJ (2007) How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nat Struct Mol Biol* 14(11): 1025-1040.
- Thomas MC, Chiang CM (2006) The general transcription machinery and general cofactors. *Crit Rev Biochem Mol Biol* 41(3): 105-178.
- Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ et al. (1998) Embryonic stem cell lines derived from human blastocysts. *Science* 282(5391): 1145-1147.
- Thomson S, Clayton AL, Hazzalin CA, Rose S, Barratt MJ et al. (1999) The nucleosomal response associated with immediate-early gene induction is mediated via alternative MAP kinase cascades: MSK1 as a potential histone H3/HMG-14 kinase. *EMBO J* 18(17): 4779-4793.
- Tian B, Nowak DE, Jamaluddin M, Wang S, Brasier AR (2005) Identification of direct genomic targets downstream of the nuclear factor-kappaB transcription factor mediating tumor necrosis factor signaling. *J Biol Chem* 280(17): 17435-17448.
- Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W (2002) Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell* 10(6): 1453-1465.
- Tolhuis B, de Wit E, Muijers I, Teunissen H, Talhout W et al. (2006) Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in *Drosophila melanogaster*. *Nat Genet* 38(6): 694-699.
- Torres-Padilla ME, Parfitt DE, Kouzarides T, Zernicka-Goetz M (2007) Histone arginine methylation regulates pluripotency in the early mouse embryo. *Nature* 445(7124): 214-218.

- Tse C, Sera T, Wolffe AP, Hansen JC (1998) Disruption of higher-order folding by core histone acetylation dramatically enhances transcription of nucleosomal arrays by RNA polymerase III. *Mol Cell Biol* 18(8): 4629-4638.
- Tsukada Y, Fang J, Erdjument-Bromage H, Warren ME, Borchers CH et al. (2006) Histone demethylation by a family of JmjC domain-containing proteins. *Nature* 439(7078): 811-816.
- Tuerk C, Gold L (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 249(4968): 505-510.
- Turner BM (2000) Histone acetylation and an epigenetic code. *Bioessays* 22(9): 836-845.
- Uesugi M, Nyanguile O, Lu H, Levine AJ, Verdine GL (1997) Induced alpha helix in the VP16 activation domain upon binding to a human TAF. *Science* 277(5330): 1310-1313.
- Vakoc CR, Mandat SA, Olenchock BA, Blobel GA (2005) Histone H3 lysine 9 methylation and HP1gamma are associated with transcription elongation through mammalian chromatin. *Mol Cell* 19(3): 381-391.
- Vakoc CR, Sachdeva MM, Wang H, Blobel GA (2006) Profile of histone lysine methylation across transcribed mammalian chromatin. *Mol Cell Biol* 26(24): 9185-9195.
- Valadez-Graham V, Razin SV, Recillas-Targa F (2004) CTCF-dependent enhancer blockers at the upstream region of the chicken alpha-globin gene domain. *Nucleic Acids Res* 32(4): 1354-1362.
- Vallier L, Alexander M, Pedersen R (2007) Conditional gene expression in human embryonic stem cells. *Stem Cells* 25(6): 1490-1497.
- van Leeuwen F, Gafken PR, Gottschling DE (2002) Dot1p modulates silencing in yeast by methylation of the nucleosome core. *Cell* 109(6): 745-756.
- van Steensel B, Henikoff S (2000) Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat Biotechnol* 18(4): 424-428.
- Varshavsky AJ, Sundin O, Bohn M (1979) A stretch of "late" SV40 viral DNA about 400 bp long which includes the origin of replication is specifically exposed in SV40 minichromosomes. *Cell* 16(2): 453-466.
- Vavouri T, Elgar G (2005) Prediction of cis-regulatory elements using binding site matrices--the successes, the failures and the reasons for both. *Curr Opin Genet Dev* 15(4): 395-402.
- Venter JC, Adams MD, Myers EW, Li PW, Mural RJ et al. (2001) The sequence of the human genome. *Science* 291(5507): 1304-1351.
- Vermeulen M, Mulder KW, Denisov S, Pijnappel WW, van Schaik FM et al. (2007) Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell* 131(1): 58-69.
- Vettese-Dadey M, Grant PA, Hebbes TR, Crane-Robinson C, Allis CD et al. (1996) Acetylation of histone H4 plays a primary role in enhancing transcription factor binding to nucleosomal DNA in vitro. *EMBO J* 15(10): 2508-2518.
- Vire E, Brenner C, Deplus R, Blanchon L, Fraga M et al. (2006) The Polycomb group protein EZH2 directly controls DNA methylation. *Nature* 439(7078): 871-874.

- Visel A, Minovitsky S, Dubchak I, Pennacchio LA (2007) VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res* 35(Database issue): D88-92.
- Vogelauer M, Wu J, Suka N, Grunstein M (2000) Global histone acetylation and deacetylation in yeast. *Nature* 408(6811): 495-498.
- Vostrov AA, Quitschke WW (1997) The zinc finger protein CTCF binds to the APBbeta domain of the amyloid beta-protein precursor promoter. Evidence for a role in transcriptional activation. *J Biol Chem* 272(52): 33353-33359.
- Vyas P, Vickers MA, Picketts DJ, Higgs DR (1995) Conservation of position and sequence of a novel, widely expressed gene containing the major human alpha-globin regulatory element. *Genomics* 29(3): 679-689.
- Waddington C (1942) The epigenotype. *Endeavour* 1: 18-20.
- Walker GT, Fraiser MS, Schram JL, Little MC, Nadeau JG et al. (1992) Strand displacement amplification--an isothermal, in vitro DNA amplification technique. *Nucleic Acids Res* 20(7): 1691-1696.
- Wallace JA, Felsenfeld G (2007) We gather together: insulators and genome organization. *Curr Opin Genet Dev* 17(5): 400-407.
- Walter J, Dever CA, Biggin MD (1994) Two homeo domain proteins bind with similar specificity to a wide range of DNA sites in *Drosophila* embryos. *Genes Dev* 8(14): 1678-1692.
- Wang J, Mager J, Chen Y, Schneider E, Cross JC et al. (2001) Imprinted X inactivation maintained by a mouse Polycomb group gene. *Nat Genet* 28(4): 371-375.
- Wang H, Huang ZQ, Xia L, Feng Q, Erdjument-Bromage H et al. (2001 b) Methylation of histone H4 at arginine 3 facilitating transcriptional activation by nuclear hormone receptor. *Science* 293(5531): 853-857.
- Wang J, Mager J, Schnedier E, Magnuson T (2002) The mouse PcG gene *eed* is required for Hox gene repression and extraembryonic development. *Mamm Genome* 13(9): 493-503.
- Wang Y, Wysocka J, Sayegh J, Lee YH, Perlin JR et al. (2004) Human PAD4 regulates histone arginine methylation levels via demethylimination. *Science* 306(5694): 279-283.
- Wang H, Wang L, Erdjument-Bromage H, Vidal M, Tempst P et al. (2004) Role of histone H2A ubiquitination in Polycomb silencing. *Nature* 431(7010): 873-878.
- Watt F, Molloy PL (1988) Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes Dev* 2(9): 1136-1143.
- Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S et al. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* 39(4): 457-466.
- Wei Y, Mizzen CA, Cook RG, Gorovsky MA, Allis CD (1998) Phosphorylation of histone H3 at serine 10 is correlated with chromosome condensation during mitosis and meiosis in *Tetrahymena*. *Proc Natl Acad Sci U S A* 95(13): 7480-7484.

- Wei CL, Wu Q, Vega VB, Chiu KP, Ng P et al. (2006) A global map of p53 transcription-factor binding sites in the human genome. *Cell* 124(1): 207-219.
- Weinmann AS, Bartley SM, Zhang T, Zhang MQ, Farnham PJ (2001) Use of chromatin immunoprecipitation to clone novel E2F target promoters. *Mol Cell Biol* 21(20): 6820-6832.
- Weinmann AS, Yan PS, Oberley MJ, Huang TH, Farnham PJ (2002) Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev* 16(2): 235-244.
- Weintraub H, Groudine M (1976) Chromosomal subunits in active genes have an altered conformation. *Science* 193(4256): 848-856.
- Wells J, Graveel CR, Bartley SM, Madore SJ, Farnham PJ (2002) The identification of E2F1-specific target genes. *Proc Natl Acad Sci U S A* 99(6): 3890-3895.
- West AG, Huang S, Gaszner M, Litt MD, Felsenfeld G (2004) Recruitment of histone modifications by USF proteins at a vertebrate barrier element. *Mol Cell* 16(3): 453-463.
- Whetstine JR, Nottke A, Lan F, Huarte M, Smolikov S et al. (2006) Reversal of histone lysine trimethylation by the JMJD2 family of histone demethylases. *Cell* 125(3): 467-481.
- White KP, Rifkin SA, Hurban P, Hogness DS (1999) Microarray analysis of *Drosophila* development during metamorphosis. *Science* 286(5447): 2179-2184.
- White EJ, Emanuelsson O, Scalzo D, Royce T, Kosak S et al. (2004) DNA replication-timing analysis of human chromosome 22 at high resolution and different developmental states. *Proc Natl Acad Sci U S A* 101(51): 17771-17776.
- Whitehouse I, Flaus A, Cairns BR, White MF, Workman JL et al. (1999) Nucleosome mobilization catalysed by the yeast SWI/SNF complex. *Nature* 400(6746): 784-787.
- Wiblin AE, Cui W, Clark AJ, Bickmore WA (2005) Distinctive nuclear organisation of centromeres and regions involved in pluripotency in human embryonic stem cells. *J Cell Sci* 118(Pt 17): 3861-3868.
- Wingender E, Chen X, Hehl R, Karas H, Liebich I et al. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* 28(1): 316-319.
- Wiren M, Silverstein RA, Sinha I, Walfridsson J, Lee HM et al. (2005) Genomewide analysis of nucleosome density histone acetylation and HDAC function in fission yeast. *EMBO J* 24(16): 2906-2918.
- Woodfine K, Fiegler H, Beare DM, Collins JE, McCann OT et al. (2004) Replication timing of the human genome. *Hum Mol Genet* 13(2): 191-202.
- Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK et al. (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* 3(1): e7.
- Wright WE, Binder M, Funk W (1991) Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site. *Mol Cell Biol* 11(8): 4104-4110.



- Wu C (1980) The 5' ends of *Drosophila* heat shock genes in chromatin are hypersensitive to DNase I. *Nature* 286(5776): 854-860.
- Wyrick JJ, Aparicio JG, Chen T, Barnett JD, Jennings EG et al. (2001) Genome-wide distribution of ORC and MCM proteins in *S. cerevisiae*: high-resolution mapping of replication origins. *Science* 294(5550): 2357-2360.
- Wysocka J, Swigut T, Milne TA, Dou Y, Zhang X et al. (2005) WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development. *Cell* 121(6): 859-872.
- Wysocka J, Swigut T, Xiao H, Milne TA, Kwon SY et al. (2006) A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature* 442(7098): 86-90.
- Xi H, Shulha HP, Lin JM, Vales TR, Fu Y et al. (2007) Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genet* 3(8): e136.
- Xiao T, Hall H, Kizer KO, Shibata Y, Hall MC et al. (2003) Phosphorylation of RNA polymerase II CTD regulates H3 methylation in yeast. *Genes Dev* 17(5): 654-663.
- Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 434(7031): 338-345.
- Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M et al. (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A* 104(17): 7145-7150.
- Xu W, Aparicio JG, Aparicio OM, Tavare S (2006) Genome-wide mapping of ORC and Mcm2p binding sites on tiling arrays and identification of essential ARS consensus sequences in *S. cerevisiae*. *BMC Genomics* 7: 276.
- Yamane K, Toumazou C, Tsukada Y, Erdjument-Bromage H, Tempst P et al. (2006) JHDM2A, a JmjC-containing H3K9 demethylase, facilitates transcription activation by androgen receptor. *Cell* 125(3): 483-495.
- Yamane K, Tateishi K, Klose RJ, Fang J, Fabrizio LA et al. (2007) PLU-1 is an H3K4 demethylase involved in transcriptional repression and breast cancer cell proliferation. *Mol Cell* 25(6): 801-812.
- Yan Q, Moreland RJ, Conaway JW, Conaway RC (1999) Dual roles for transcription factor IIF in promoter escape by RNA polymerase II. *J Biol Chem* 274(50): 35668-35675.
- Yang A, Zhu Z, Kapranov P, McKeon F, Church GM et al. (2006) Relationships between p63 binding, DNA sequence, transcription activity, and biological function in human cells. *Mol Cell* 24(4): 593-602.
- Yauk CL, Bois PR, Jeffreys AJ (2003) High-resolution sperm typing of meiotic recombination in the mouse MHC Ebeta gene. *EMBO J* 22(6): 1389-1397.

- Yeh RF, Lim LP, Burge CB (2001) Computational inference of homologous gene structures in the human genome. *Genome Res* 11(5): 803-816.
- Yokoyama A, Wang Z, Wysocka J, Sanyal M, Aufiero DJ et al. (2004) Leukemia proto-oncoprotein MLL forms a SET1-like histone methyltransferase complex with menin to regulate Hox gene expression. *Mol Cell Biol* 24(13): 5639-5649.
- Young RA (1991) RNA polymerase II. *Annu Rev Biochem* 60: 689-715.
- Young RA (2000) Biomedical discovery with DNA arrays. *Cell* 102(1): 9-15.
- Yu W, Ginjala V, Pant V, Chernukhin I, Whitehead J et al. (2004) Poly(ADP-ribosyl)ation regulates CTCF-dependent chromatin insulation. *Nat Genet* 36(10): 1105-1110.
- Yu MC, Lamming DW, Eskin JA, Sinclair DA, Silver PA (2006) The role of protein arginine methylation in the formation of silent chromatin. *Genes Dev* 20(23): 3249-3254.
- Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF et al. (2005) Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* 309(5734): 626-630.
- Yusufzai TM, Felsenfeld G (2004) The 5'-HS4 chicken beta-globin insulator is a CTCF-dependent nuclear matrix-associated element. *Proc Natl Acad Sci U S A* 101(23): 8620-8624.
- Zardo G, Caiafa P (1998) The unmethylated state of CpG islands in mouse fibroblasts depends on the poly(ADP-ribosyl)ation process. *J Biol Chem* 273(26): 16517-16520.
- Zeng L, Zhou MM (2002) Bromodomain: an acetyl-lysine binding domain. *FEBS Lett* 513(1): 124-128.
- Zhang Y (2004) Molecular biology: no exception to reversibility. *Nature* 431(7009): 637-639.
- Zhao H, Dean A (2004) An insulator blocks spreading of histone acetylation and interferes with RNA polymerase II transfer between an enhancer and gene. *Nucleic Acids Res* 32(16): 4903-4919.
- Zhao H, Kim A, Song SH, Dean A (2006) Enhancer blocking by chicken beta-globin 5'-HS4: role of enhancer strength and insulator nucleosome depletion. *J Biol Chem* 281(41): 30573-30580.
- Zhao Z, Tavoosidana G, Sjolinder M, Gondor A, Mariano P et al. (2006 b) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet* 38(11): 1341-1347.
- Zhao XD, X. Han, et al. (2007) Whole-Genome Mapping of Histone H3 Lys4 and 27 Trimethylations Reveals Distinct Genomic Compartments in Human Embryonic Stem Cells. *Cell Stem Cell* 1: 286-298.
- Zheng D, Frankish A, Baertsch R, Kapranov P, Reymond A et al. (2007) Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. *Genome Res* 17(6): 839-851.
- Zhou T, Chiang CM (2001) The intronless and TATA-less human TAF(II)55 gene contains a functional initiator and a downstream promoter element. *J Biol Chem* 276(27): 25503-25511.
- Zhu B, Zheng Y, Pham AD, Mandal SS, Erdjument-Bromage H et al. (2005) Monoubiquitination of human histone H2B: the factors involved and their roles in HOX gene regulation. *Mol Cell* 20(4): 601-611.

## Appendix 1

### Sequences of primer pairs used in real-time PCR verification of ChIP-chip data

Region name	Primer 1 (5'-3')	Primer 2 (5'-3')	Amplicon Size (bp)	Chrom 1 Start coordinate
KCY +1	TGGTTGGTTAGCTGCATTGA C	CCTCCTTTCCCAAGCATTCC	75	47512689
NC1	TCTCTTTGAACACAGGGCA ATG	TATTAGTCTAGGTGTACTG GCAGTTG	71	47499807
SIL - 1/+1	CAGTCGCCGACCAATGATC	GCTAGGTAGACGGAGGAGC G	73	47492145
NC2	TGAATGCTTCCCTTGTGATG	GTAATGTTTCCTTACTGGTT AGCAAC	71	47467138
NC3	GTGCCCTTGAGAGCCTAGG G	CCTCAACAGCCTGTCTTATA ATTG	71	47444008 3
SCL -9/- 10	GGCCAGAGTTCAAATCCTG AC	CAAGCGTAAAGTGACATGC CC	71	47419831
SCL +1	TTCCCCCTTTTCCTTACGC	CGCACTCTCACAATCCCAC C	102	47409535
SCL +3	TTTCGAACCCTCCAACTGG	CAACCGGTAGACACCTCC	72	47407338
NC4	CATCACCTGCAAAATGGAG G	TAAGCTGAGGCAGGCATTG TC	103	47398315
SCL +20/+22	CCCAGTGGTCTGACTCCAA AG	GCACAGAAGGCAGTGAATG G	74	47390543
NC5	GGATTGAGGAGAGGGCATG TG	GCACGGCTGTGGAGCTATG	101	47377642
NC6	CAGCAGAGGTCCCAAAGCC	CAGTACTCCCAGCTTGCTTC C	101	47374561
SCL +51	TTAAGCCGAAGCCCAGAGA G	GCTCCAGGCCTATCCTTGC	71	47359910
NC7	TTCTGTACCTGCCAGCCAA G	CCCGACGAGCGTTATGTAA G	71	47355954