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 immuno-precipitation 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.

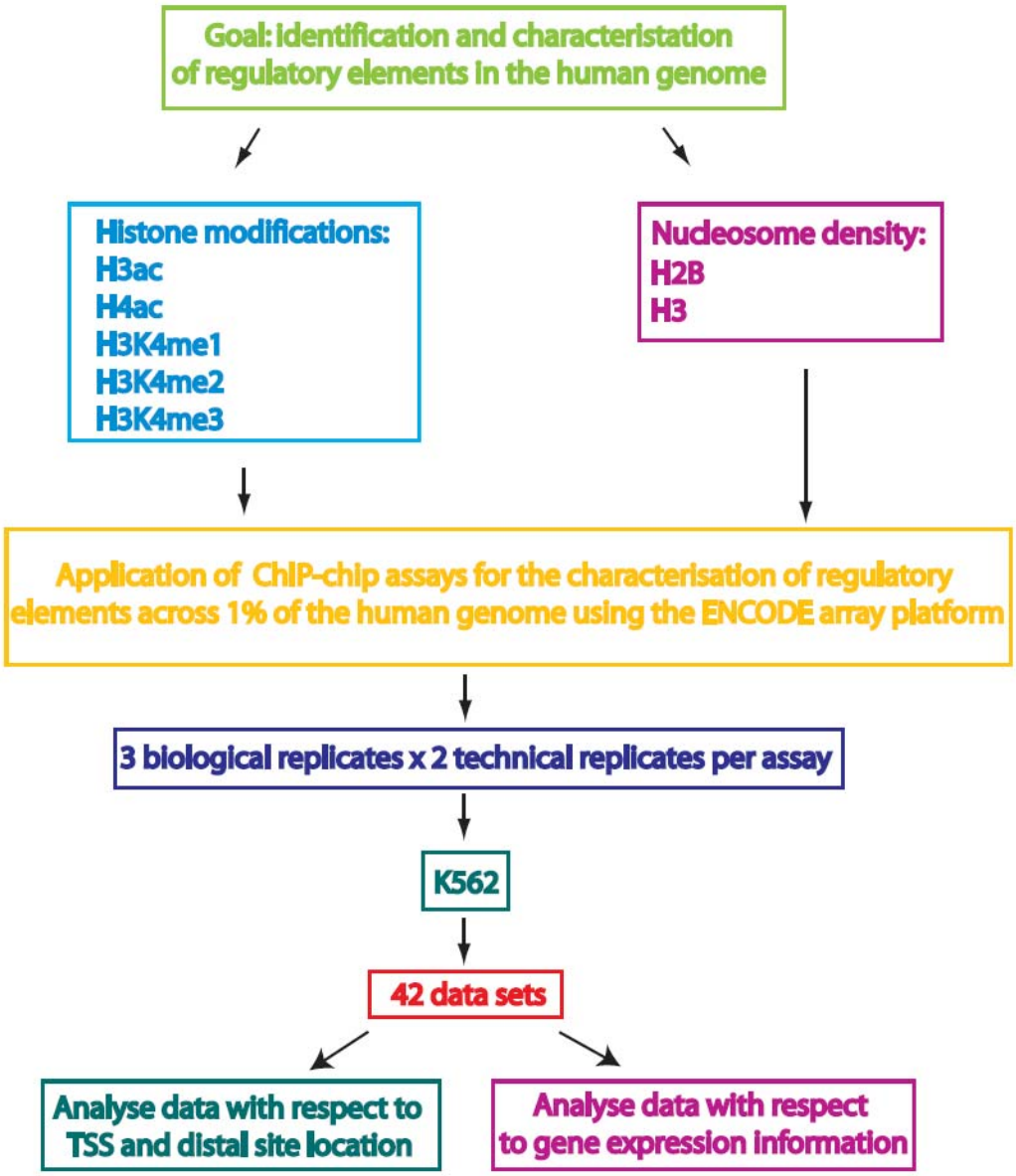

Figure 3.1: Schematic representation of the overall strategy used to map regulatory interactions across 1% of the human genome. This flow diagram illustrates the strategy used to identify and characterise regulatory elements in the human genome. Definitions of biological and technical replicates are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

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.

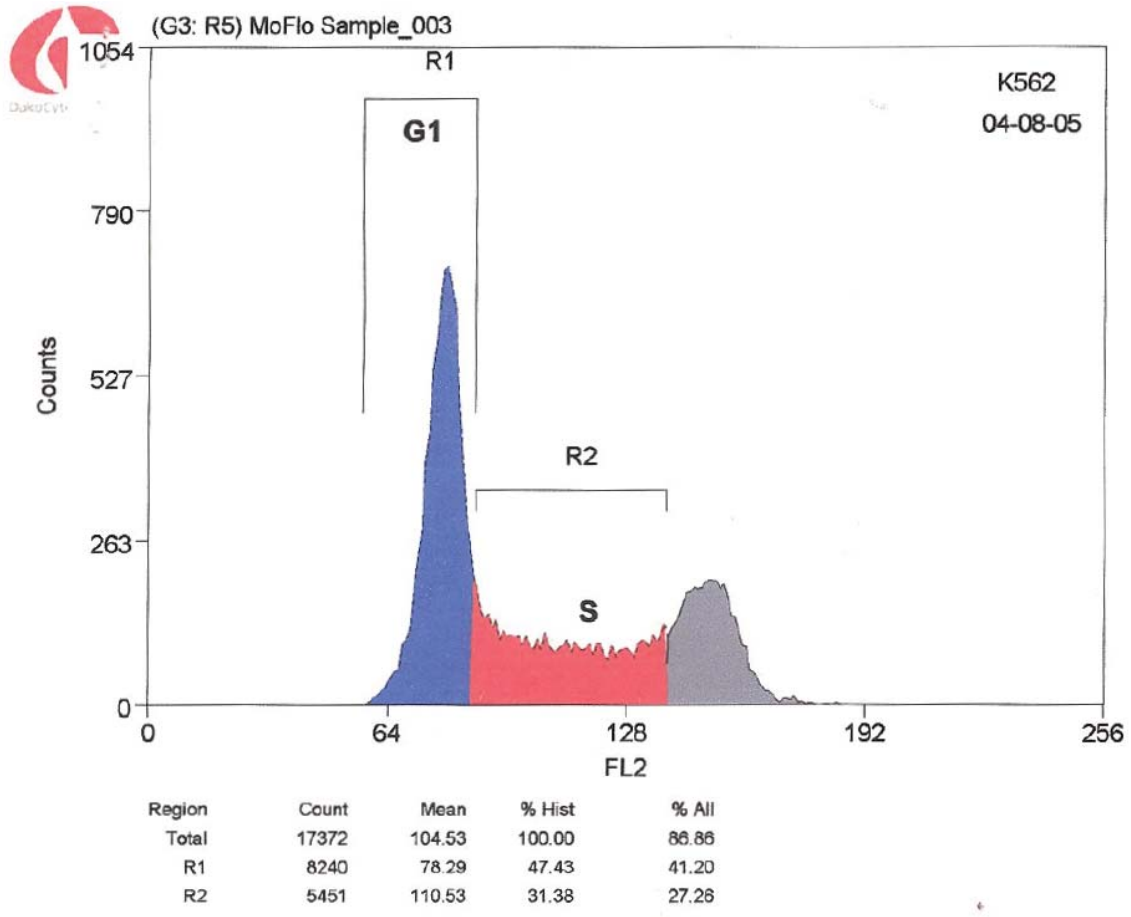

Figure 3.2: Analysis of cell growth by flow-cytometry. Human cell lines used for ChIP-chip experiments described in this thesis were analysed for their DNA content by staining with the DNA binding dye Hoechst 33342 and the percentage of cells in the G1(indicated by R1) and S (indicated by R2) phase of 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).

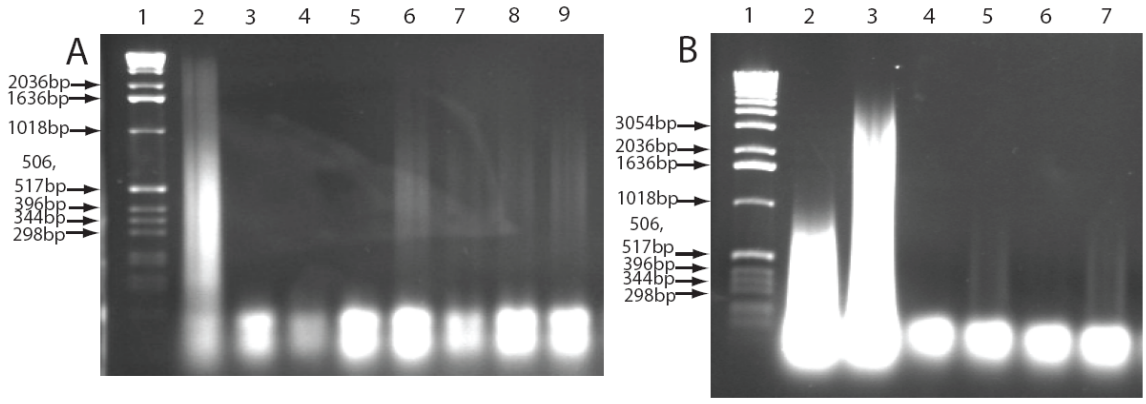

Figure 3.3: Electrophoresis of ChIP DNAs. Panel A shows ChIP DNAs performed for various histone modifications and panel B shows the ChIP DNAs for the transcription factor CTCF. In panels A and B, lane 1 = 1 kb DNA marker ladder. In panel A; lane 2 = 0.37% cross-linked and sonicated Input DNA; lane 3 = 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).

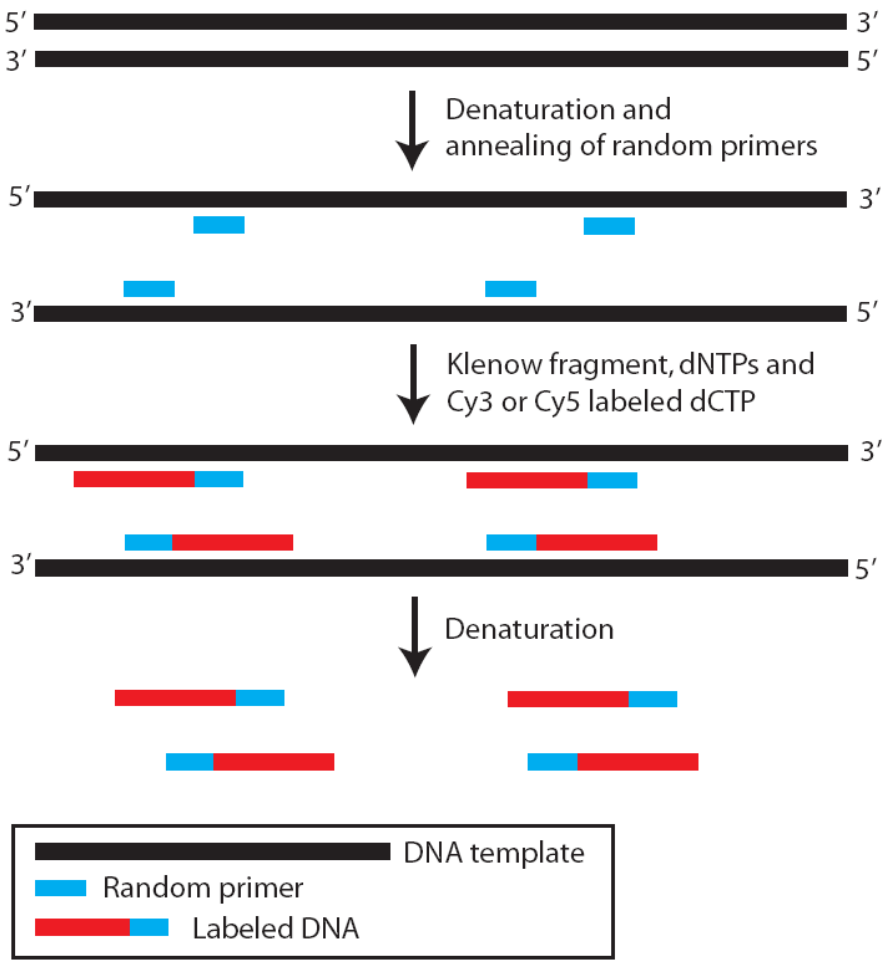

Figure 3.4: Labeling of DNA by a random priming method. Input and ChIP DNAs are fluorescently labeled by Klenow mediated incorporation of Cy5 and Cy3 labeled dCTP respectively. Denaturation following this reaction results in single stranded fluorescently labelled DNA, which can then be hybridized to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

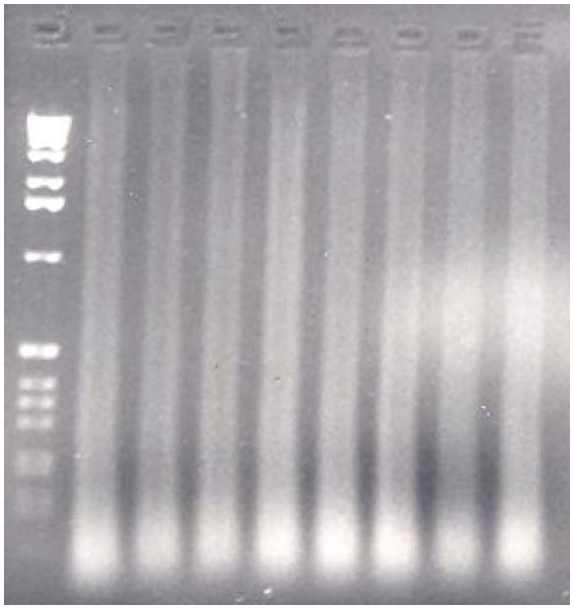

Figure 3.5: Electrophoresis of fluorescently labeled DNAs. Labeled input and ChIP DNA samples were electrophoresed prior to microarray hybridisation to verify that klenow-mediated labeling reactions had resulted in the generation of single stranded DNA. A large smear of DNA was evident in the samples 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).

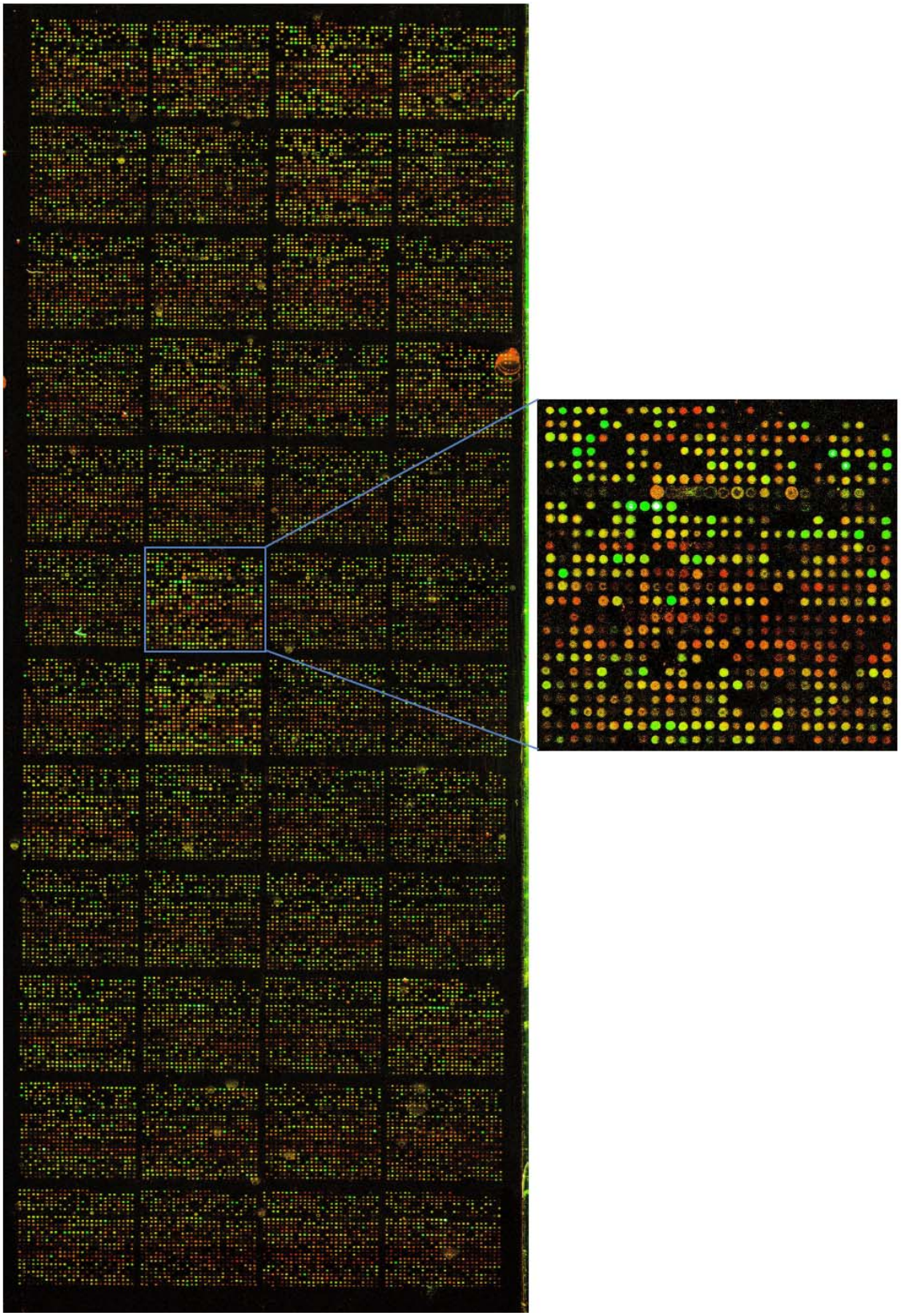

Figure 3.6: A composite image of the Sanger Institute ENCODE array. The array was hybridised with a K562 H3K4me3 ChIP sample along with K562 input DNA. Each spot on the array represents an array element. PCR products were spotted in a 48 sub-grid (12 rows x 4 columns) format and a magnified view of one 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.

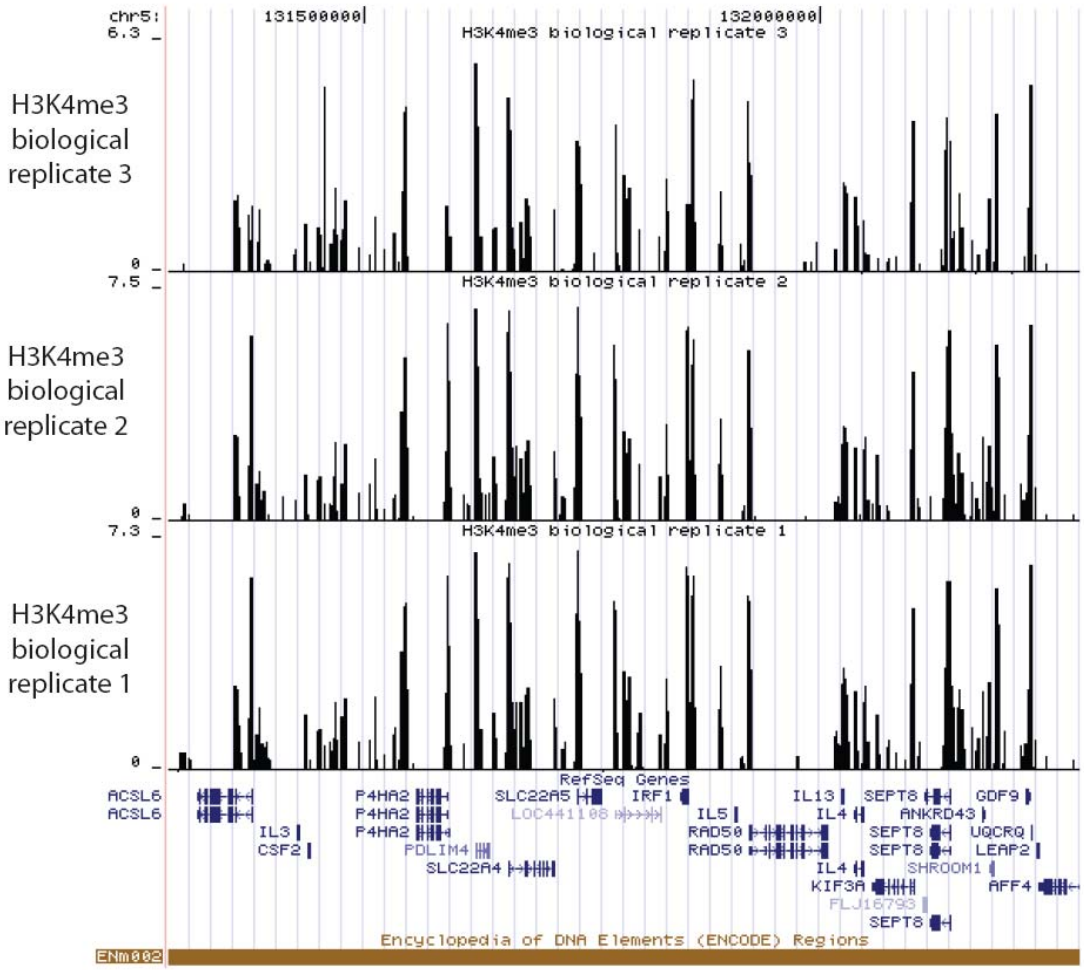

Figure 3.7: Assessing the performance of the ENCODE microarray in independent ChIP-chip assays. A screenshot form the UCSC genome browser (Kuhn *et al.*, 2007) of ENCODE region Enm002 (chr5:131284314 - 132284313) showing ChIP-chip data from three independent H3K4me3 experiments in K562 cells. Reported $\log_2$ 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.

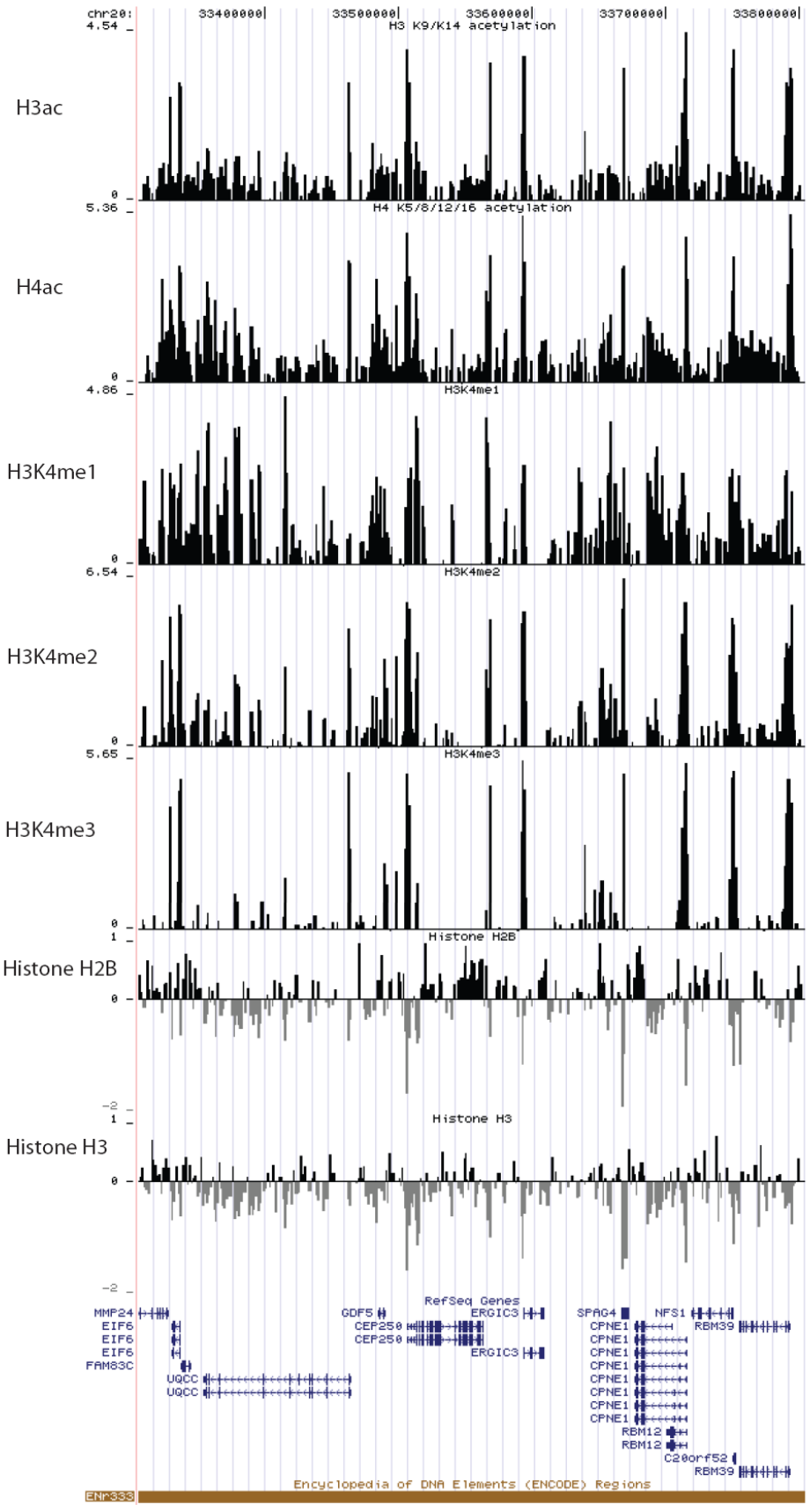

Figure 3.8: An example of the histone modification and nucleosome density maps generated across the ENCODE regions. A screenshot from the UCSC genome browser (Kuhn *et al*., 2007) of ENCODE region ENr333 (human chromosome 20: 33,304,929–33,804,928 bp) showing ChIP-chip data "wiggle" tracks for five antibodies 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).

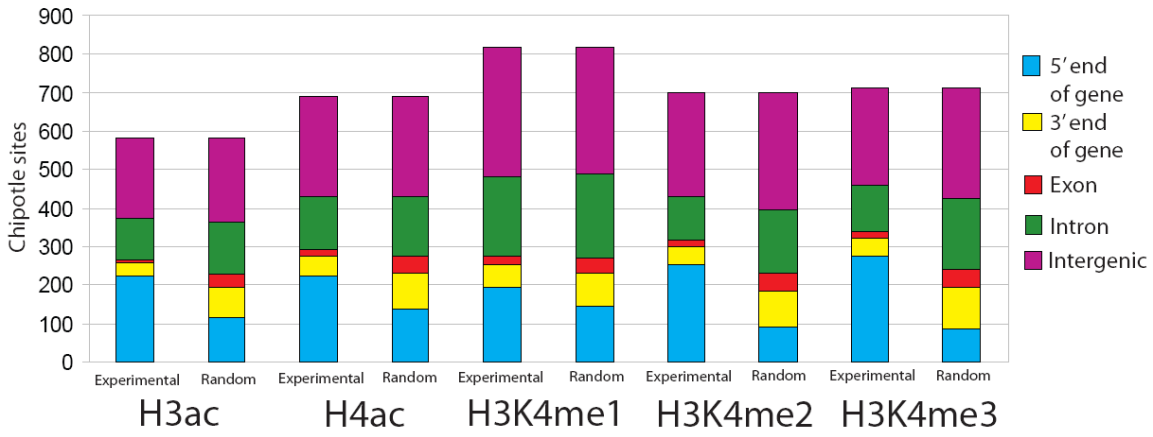

Figure 3.9: Distribution of histone modifications in K562 cells. The chart shows the number of H3ac, H4ac, H3K4me1, H3K4me2, and H3K4me3 ChIPOTle sites which overlap with a TSS (blue), 3' end of a gene (yellow), exon (red), intron (green), or intergenic sequence (purple) in ChIP-chip data (experimental) 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 (+/- 2.5kb 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).

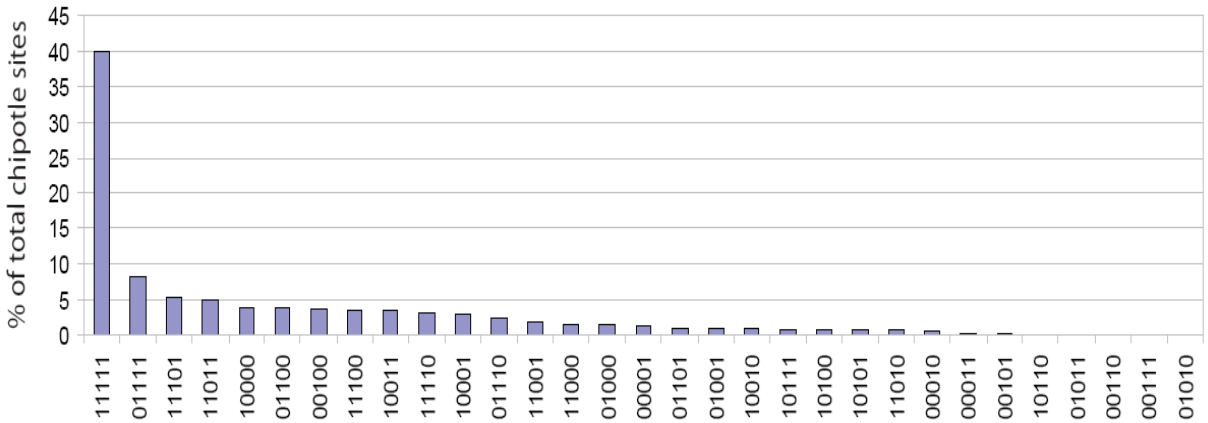

Figure 3.10: The relative occurrence of overlapping histone modification sites in K562. Histone modification sites were deemed to overlap if they were present within a 5kb window centred on a ChIPOTle site, i.e. located +/- 2.5kb from the centre of a ChIPOTle site. The relative occurrence of each overlapping 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.

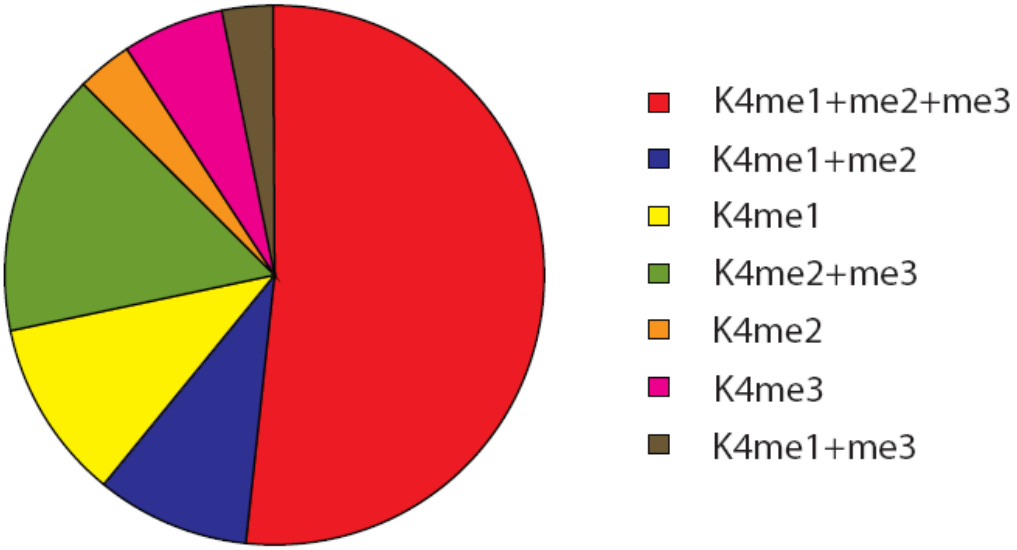

Figure 3.11: The extent of overlapping H3K4 methylation sites in K562. The pie-chart illustrates the occurrence of the seven H3K4 methylation combinations.

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.

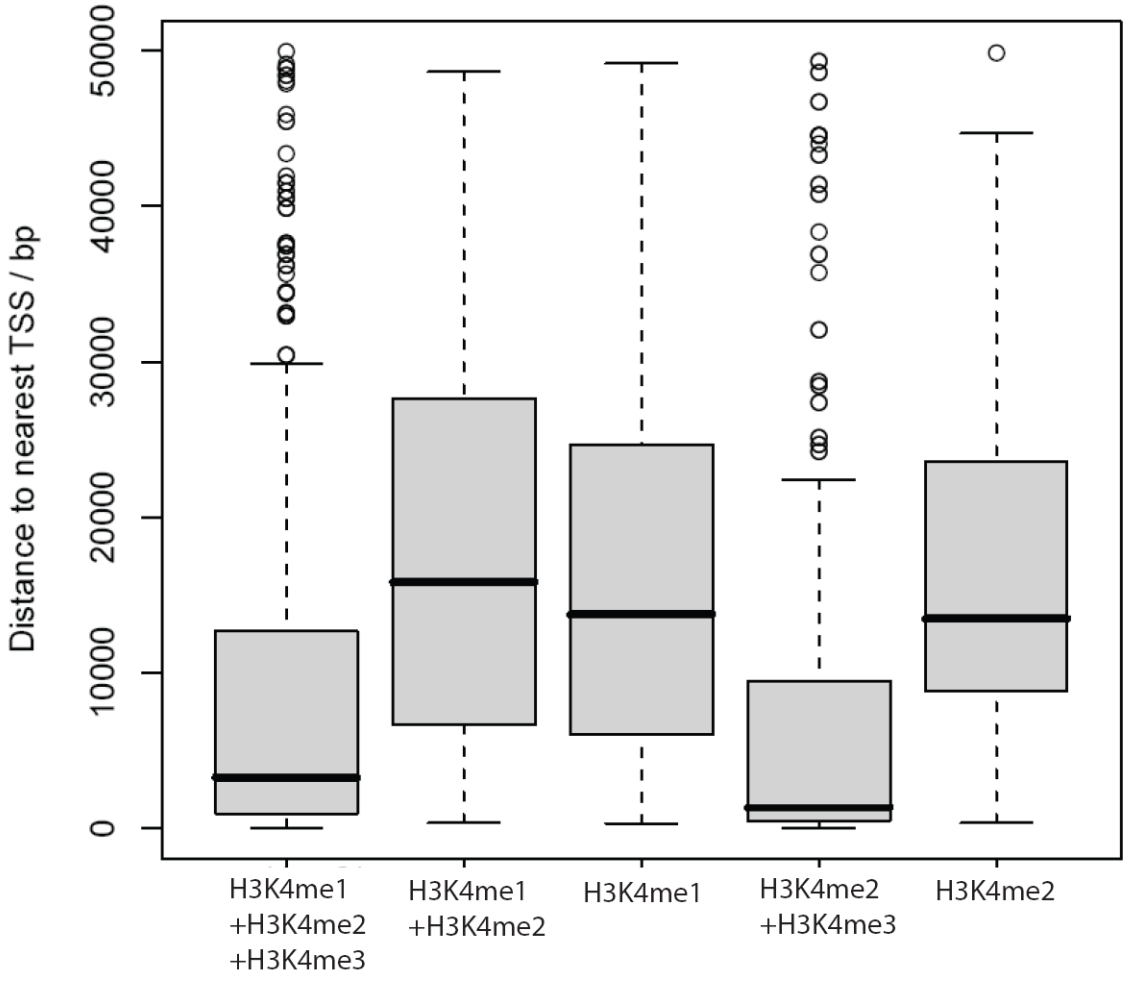

Figure 3.12: The distribution of histone H3K4 methylation combinations relative to the nearest transcriptional start site (TSS). Box-plots are presented for the five main histone modification combinations showing their distribution relative to the nearest TSS. The grey boxes show the 25th and 75th percentiles, with the black line inside the grey box represents the median distance. The whiskers extend to the 2.5th and 97.5th 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 (+/- 2.5kb) 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.

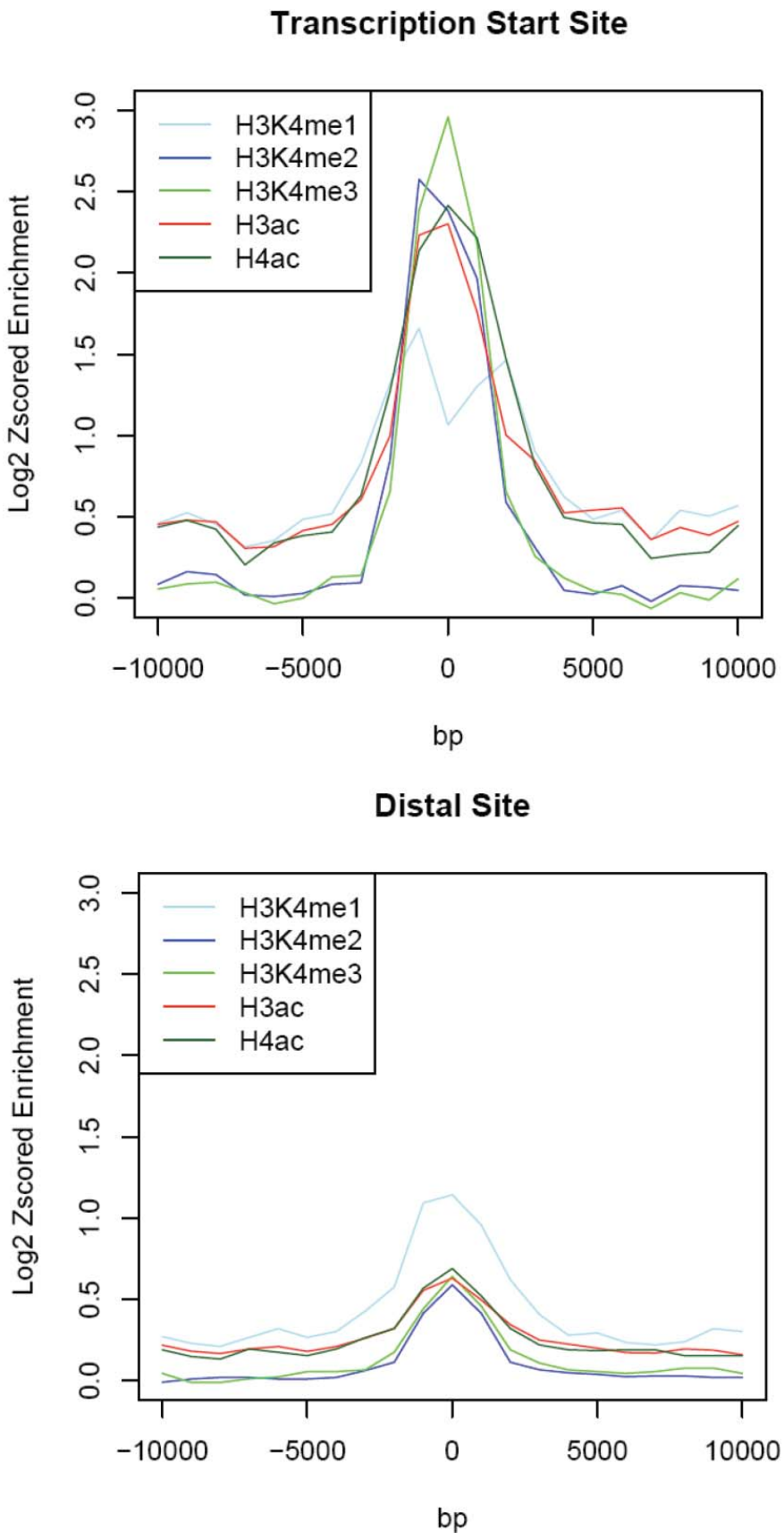

Figure 3.13: Histone modification profiles at TSSs and distal sites. The average Log_2 Z-scored histone modification profiles are presented for ChIPOTle defined sites located at TSSs (+/- 2.5kb) and at distal elements (located greater than 2.5kb from TSSs). The average values are plotted +/- 10,000 bp relative to the location of TSSs or distal elements (indicated on the x-axes by 0). Average Log_2 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 investigating 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.

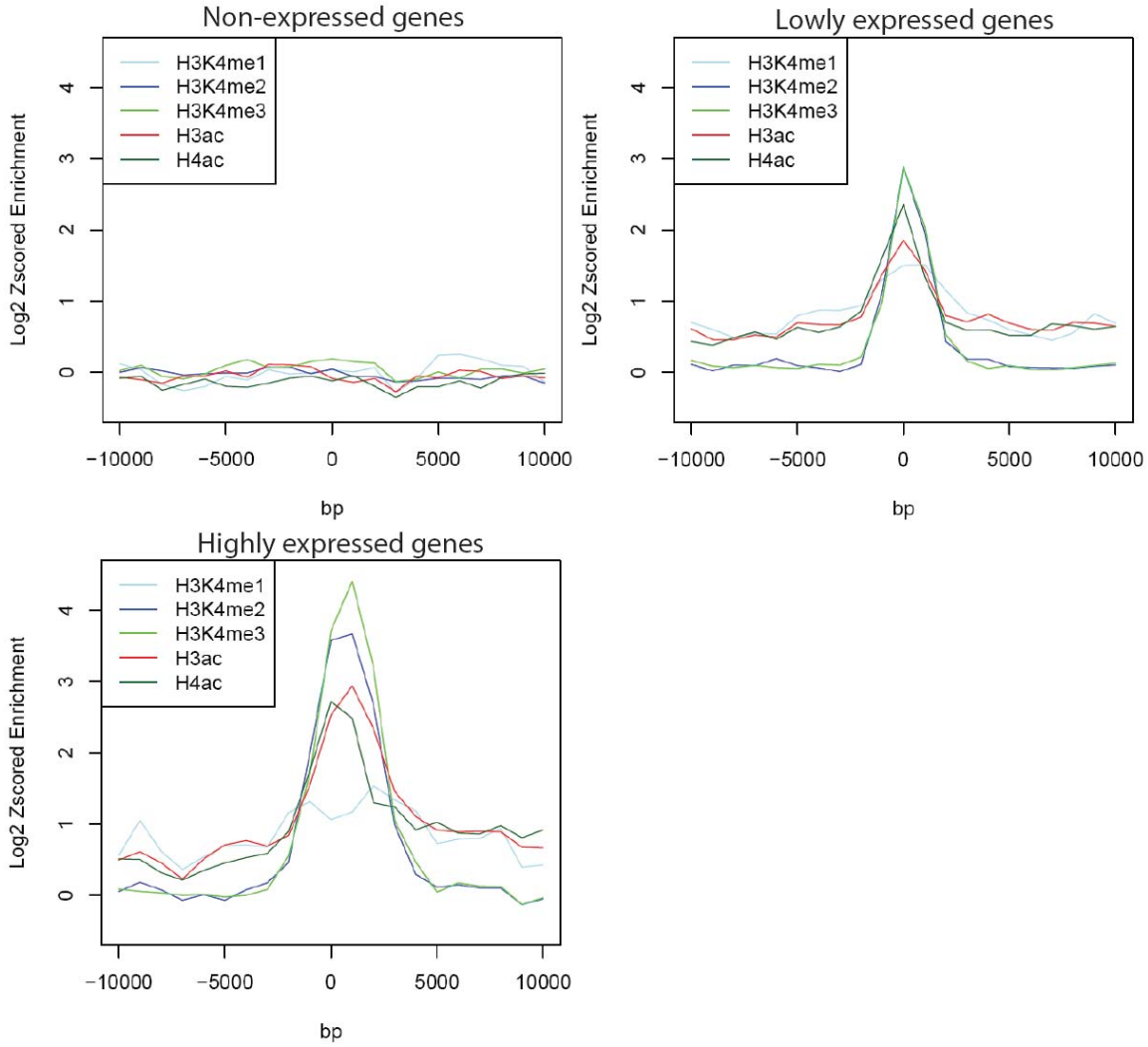

Figure 3.14: Histone modification profiles for non-expressed and expressed genes in K562 cells. The average z-scored log_2 histone modification profiles are presented +/- 10,000 bp surrounding TSSs of non-expressed, lowly expressed, and highly expressed genes.

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.

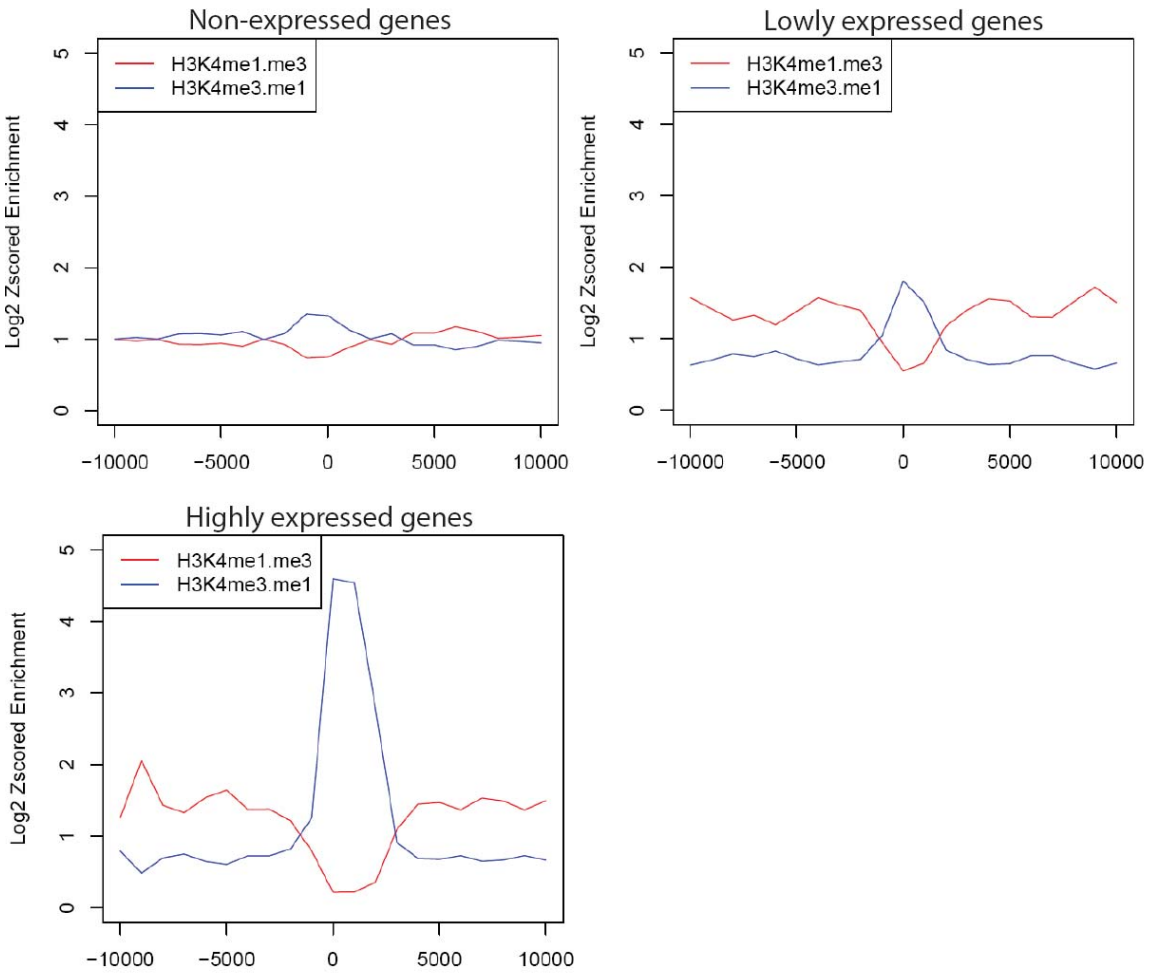

Figure 3.15: Profiles showing H3K4me3 to H3K4me1 ratios at the promoters of active and inactive genes in K562 cells. The average z-scored $\log_2$ H3K4me3 and H3K4me1 values were plotted relative to each other to obtain H3K4me1:H3K4me3 ratios and H3K4me3:H3K4me1 ratios at TSSs associated with non-expressed,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

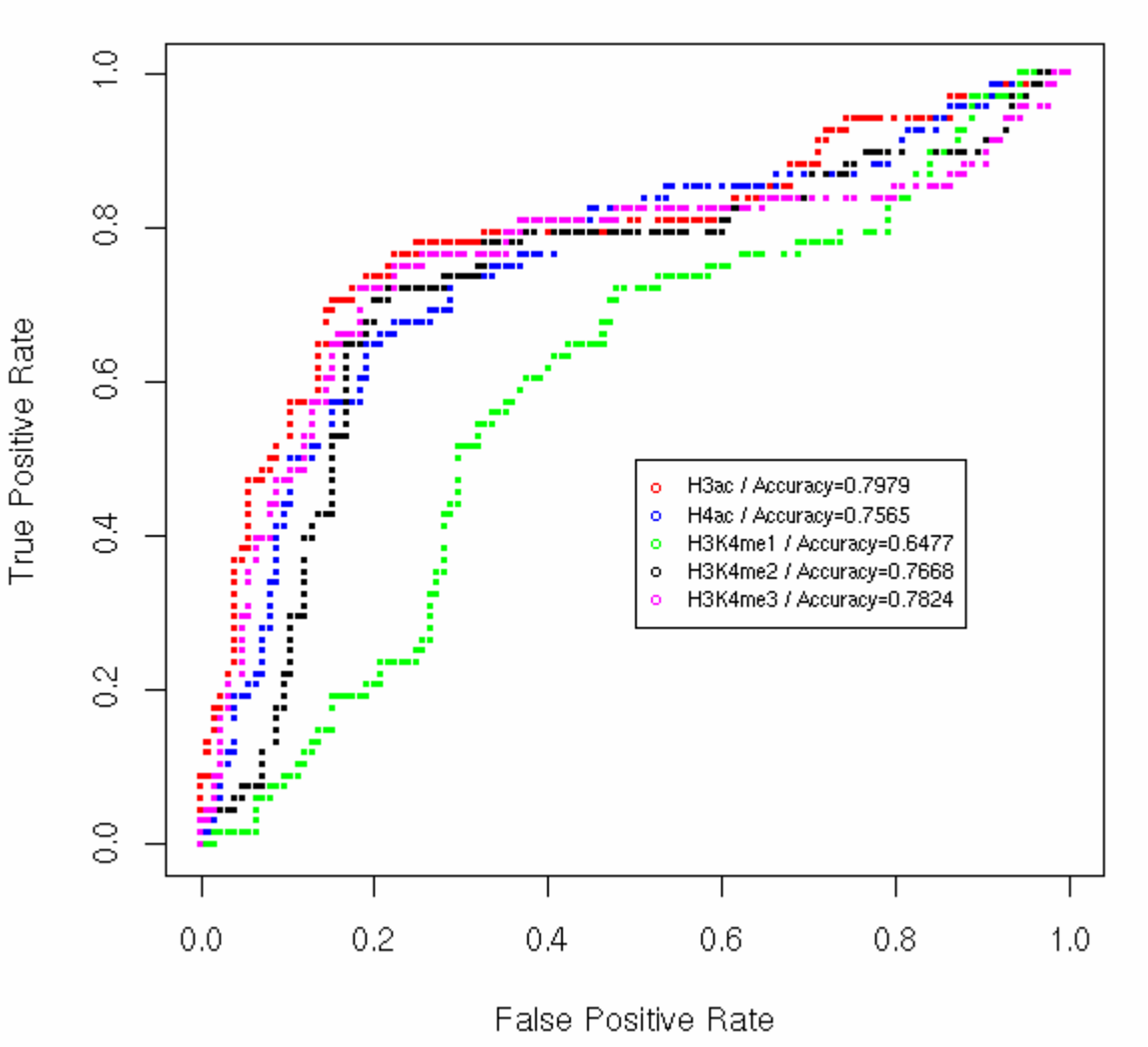

Figure 3.16: Receiver operating characteristic (ROC) curves illustrating the predictive power of histone modifications in K562 cells. The ROC curve always goes through two points. The first point is 0,0 where a histone modification predicts no true positives (i.e. does not identify any expressed genes) 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 predicition 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.

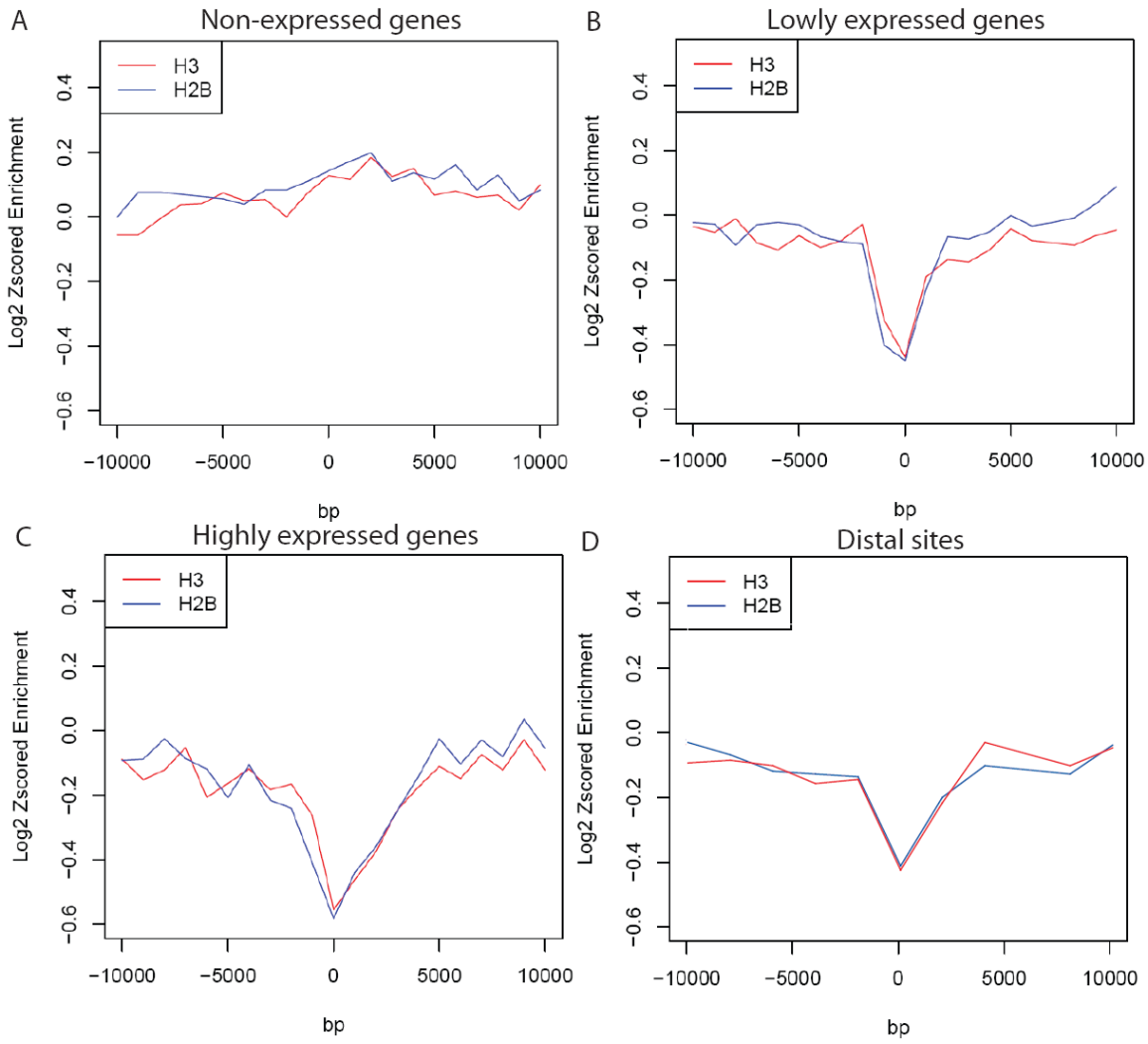

Figure 3.17: Nucleosome occupancy at transcription start sites and distal elements in K562 cells. The average z-scored $\log_2$ histone H2B and H3 values are presented at TSSs and surrounding regions (+/- 10 kb) of non-expressed (panel A), lowly expressed (panel B) and highly expressed (panel C) genes 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.

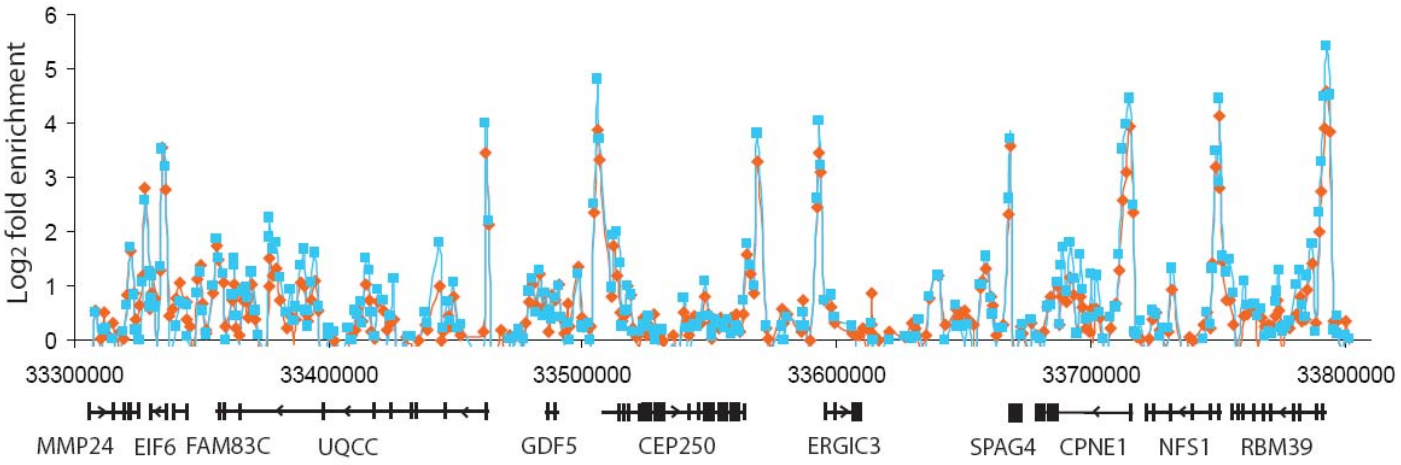

Figure 3.18: Histone modification levels relative to nucleosome occupancy in K562 cells. An example of the effect of normalizing histone modification data with nucleosome data is presented for ENCODE region Enr333. Histone H3 acetylation levels are presented before (orange profile) and after (blue profile) 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 where 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 insulators 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 confirmation 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 continous 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 immunoprecipate 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 efficiences 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.