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 haematpoietic 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 (Azuara *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 committment (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+ monocytes were chosen for the study of chromatin state during lineage commitment as hESCs are pluripotent while monocytes represent a multipotent lineage committed progentior 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 x 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 x 10^7 cells are obtained from each donor sample which limits the number of ChIP-chip experiments that can be performed using standard protocols.

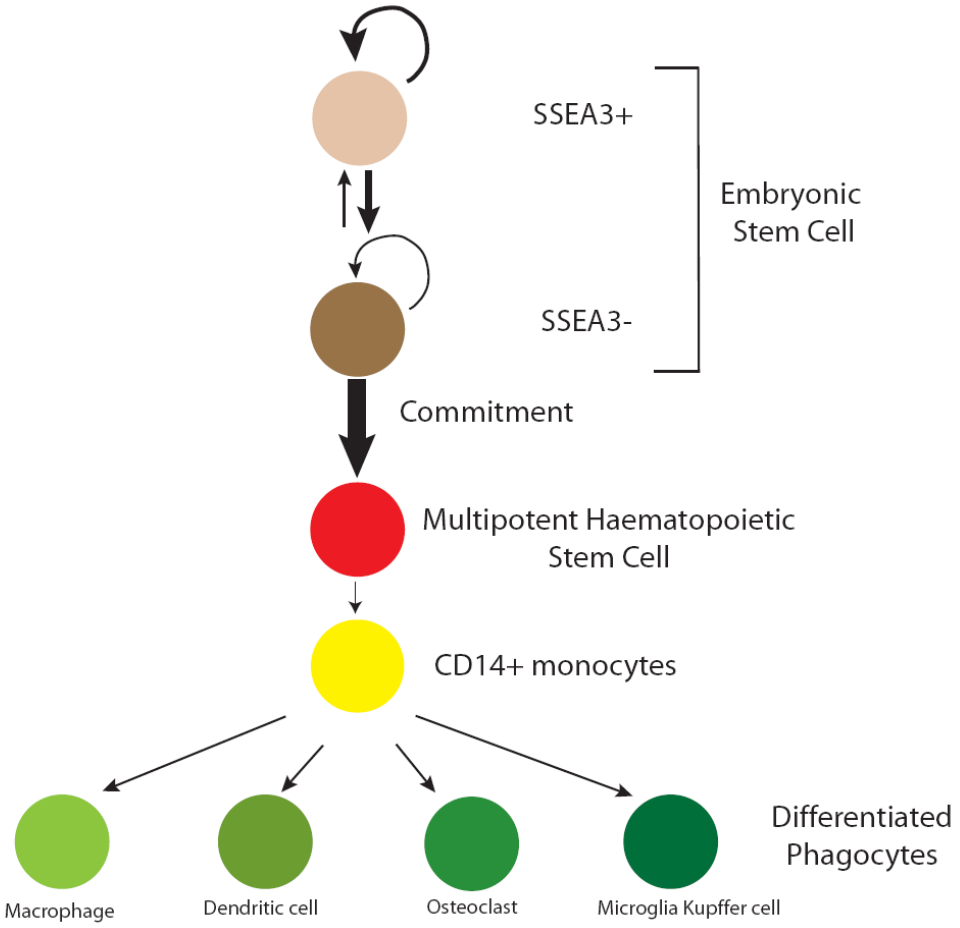


Figure 6.1: The relationship between SSEA3+ and SSEA3- H9 hESCs and CD14+ monocyte cells. SSEA3+ cells represent undifferentiated embryonic stem cells which can self-renew indefinitely in culture. SSEA3+ cells can be induced to spontaneously differentiate in culture and the first cell surface marker 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 HeK27me3 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 in 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.

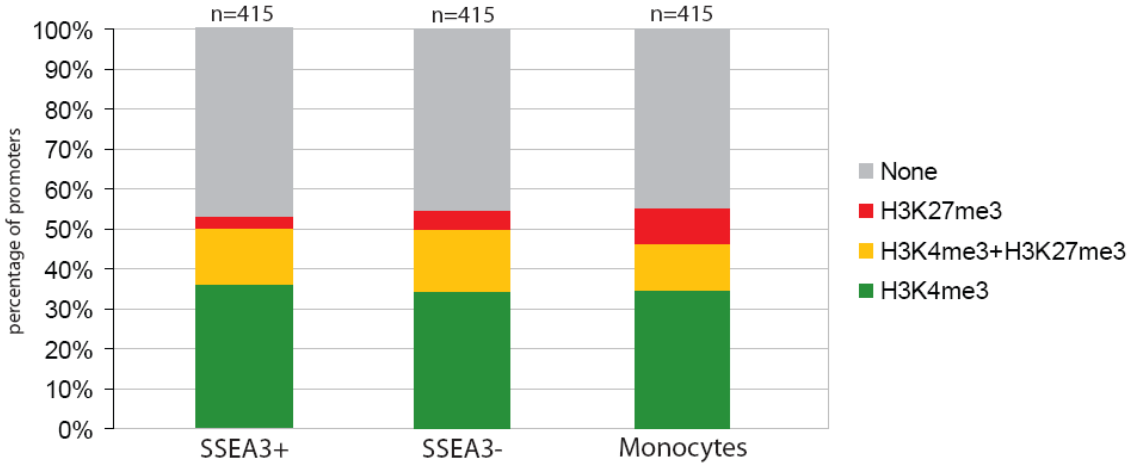


Figure 6.2: Chromatin modification patterns of promoters in human embryonic stem cells and monocyte cells. The promoters of 415 well defined RefSeq genes (Pruitt *et al*., 2007) were examined for H3K4me3 and H3K27me3 modification patterns in SSEA3+ hESCs, SSEA3- hESCs, and CD14+ monocytes. The percentage 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 classsified as HCPs or LCPs based on the system of Saxanov and colleagues (Saxanov *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.

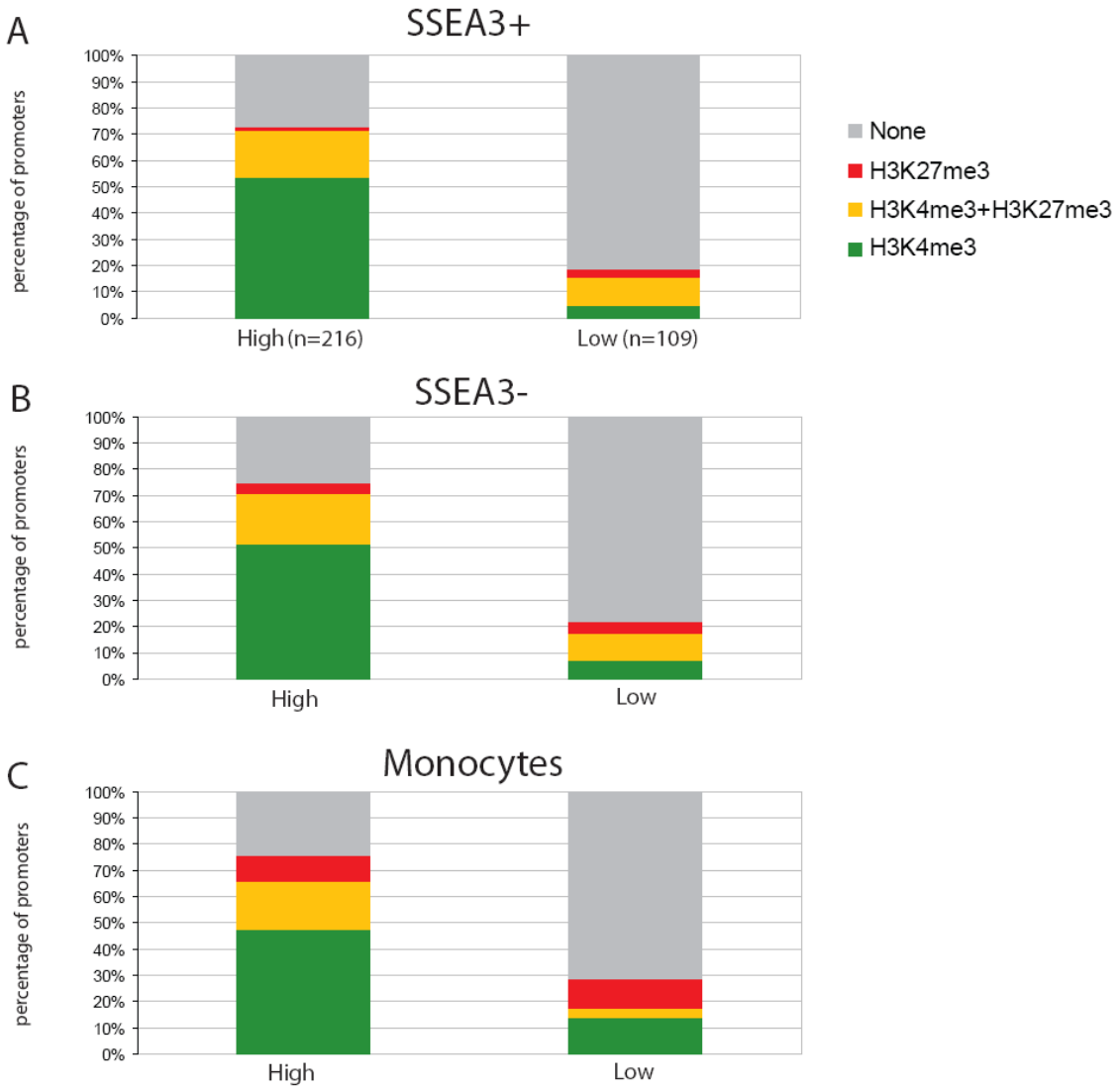


Figure 6.3: Distinct histone modification profiles at high and low CpG content promoters. Panels A - C illustrates the chromatin modification pattern of promoters classified as high CpG (n=216) or low CpG content promoters (n=109). The percentage of promoters which were found to significantly enriched 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).

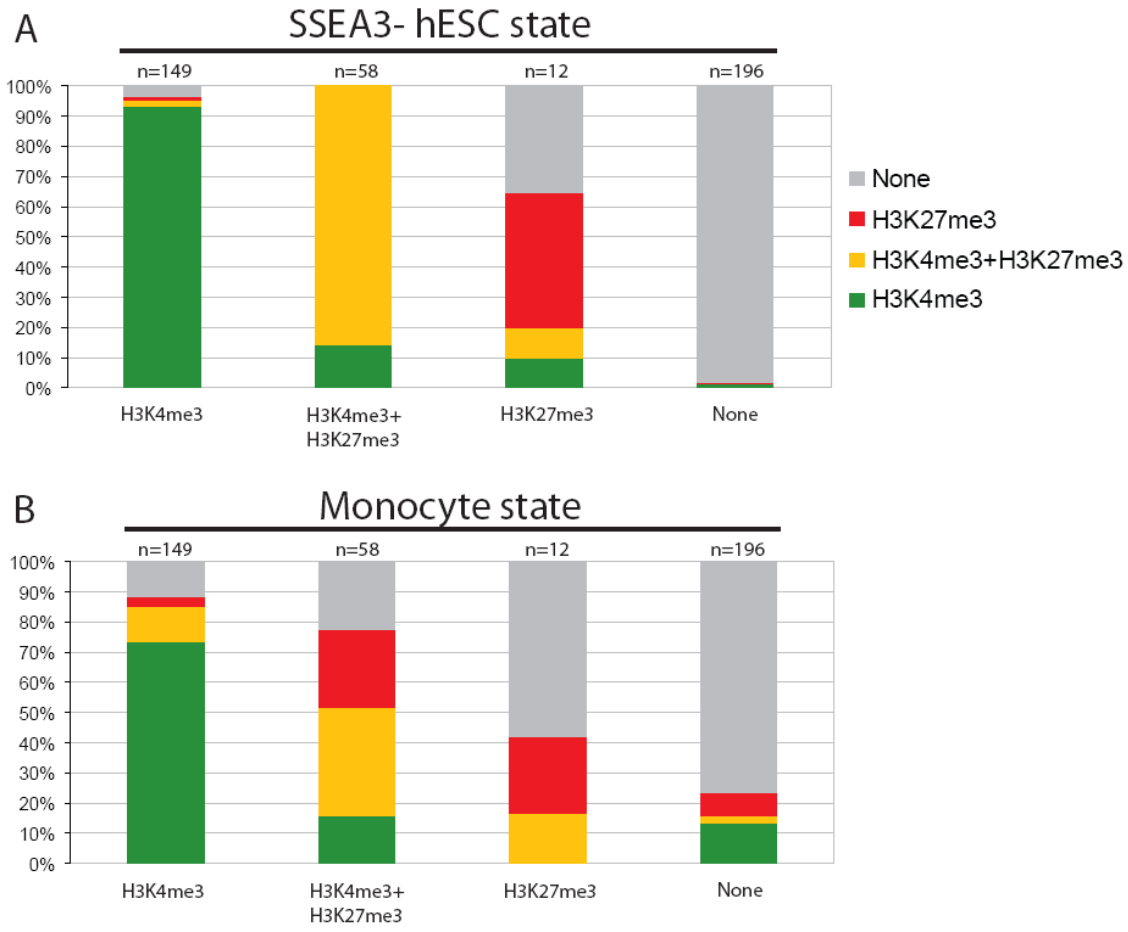


Figure 6.4: Promoter chromatin state and developmental potential. The chromatin state of 415 promoters had previously been determined for SSEA3+ hESCs resulting in the classification of promoters in this cell type into four groups – monovalent H3K4me3, monovalent H3K27me3, bivalent H3K4me3+H3K27me3 and 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 summaried 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.

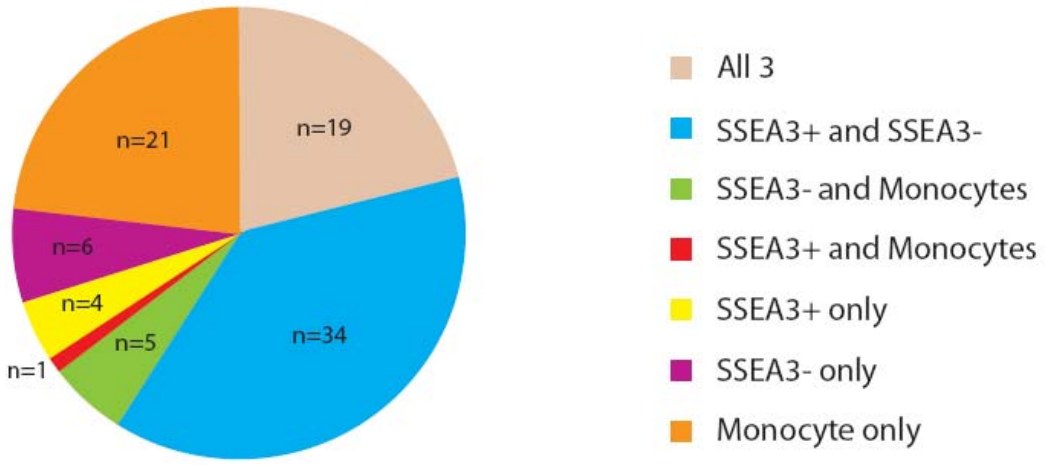


Figure 6.5: Conserved and cell-type specific bivalent promoters. 90 gene promoters were associated with a bivalent chromatin structure in SSEA3+ hESCS, SSEA3- hESCS and monocytes. This pie-chart illustrates that many of these promoters were bivalent in more than one cell type while 19 (21%) of promoters 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.

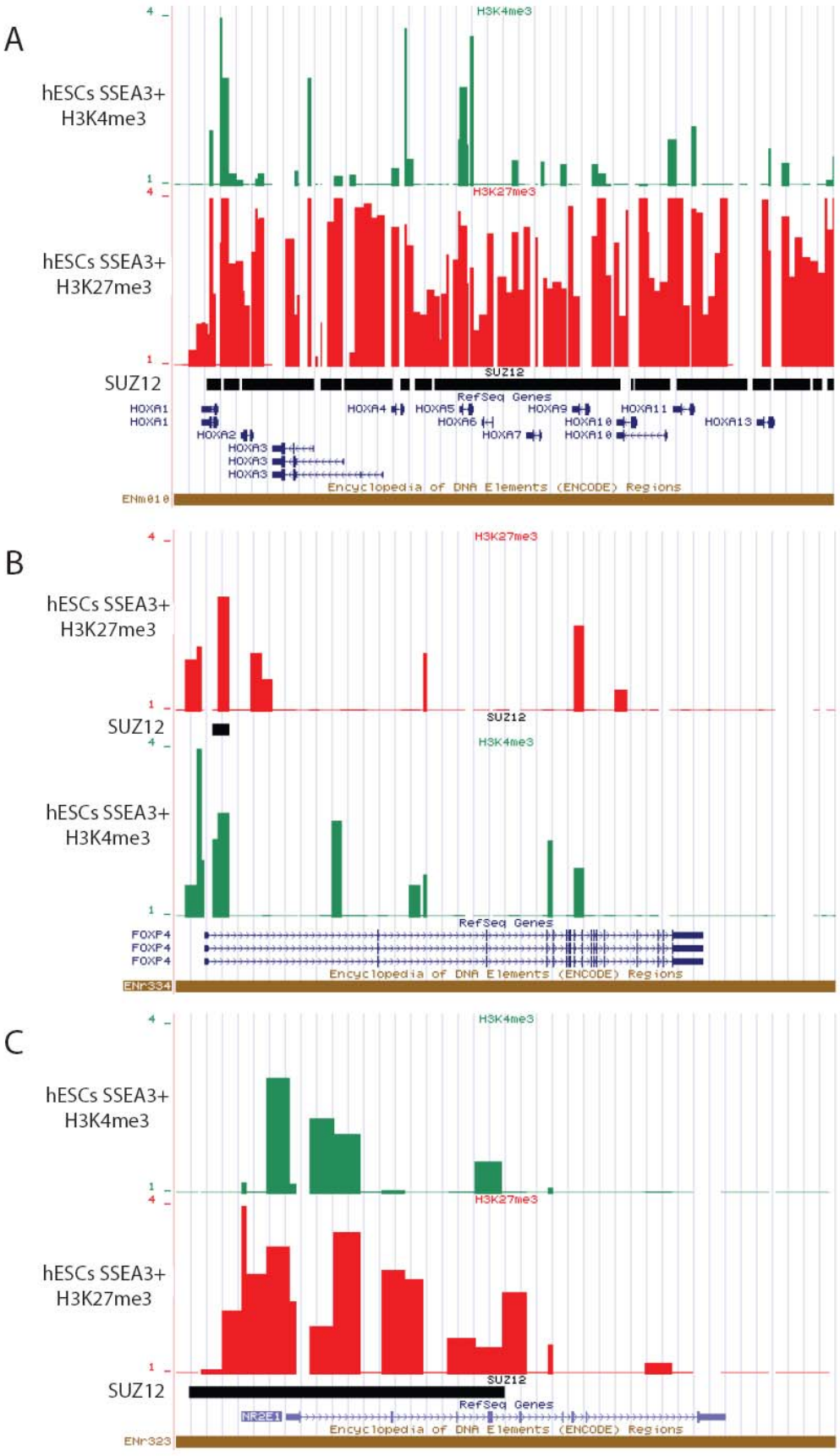

A
hESCs SSEA3+
H3K4me3
H3K4me3
hESCs SSEA3+
H3K27me3
H3K27me3
SUZ12
SUZ12
RefSeq Genes
HOXA1
HOXA1
HOXA2
HOXA3
HOXA3
HOXA3
HOXA4
HOXA5
HOXA6
HOXA7
HOXA9
HOXA10
HOXA10
HOXA11
HOXA13
Encyclopedia of DNA Elements (ENCODE) Regions
ENm010
B
H3K27me3
hESCs SSEA3+
H3K27me3
SUZ12
SUZ12
H3K4me3
hESCs SSEA3+
H3K4me3
RefSeq Genes
FOXP4
FOXP4
FOXP4
Encyclopedia of DNA Elements (ENCODE) Regions
ENr334
C
H3K4me3
hESCs SSEA3+
H3K4me3
H3K27me3
hESCs SSEA3+
H3K27me3
SUZ12
SUZ12
RefSeq Genes
NR2E1
Encyclopedia of DNA Elements (ENCODE) Regions
ENr323

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_2 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).

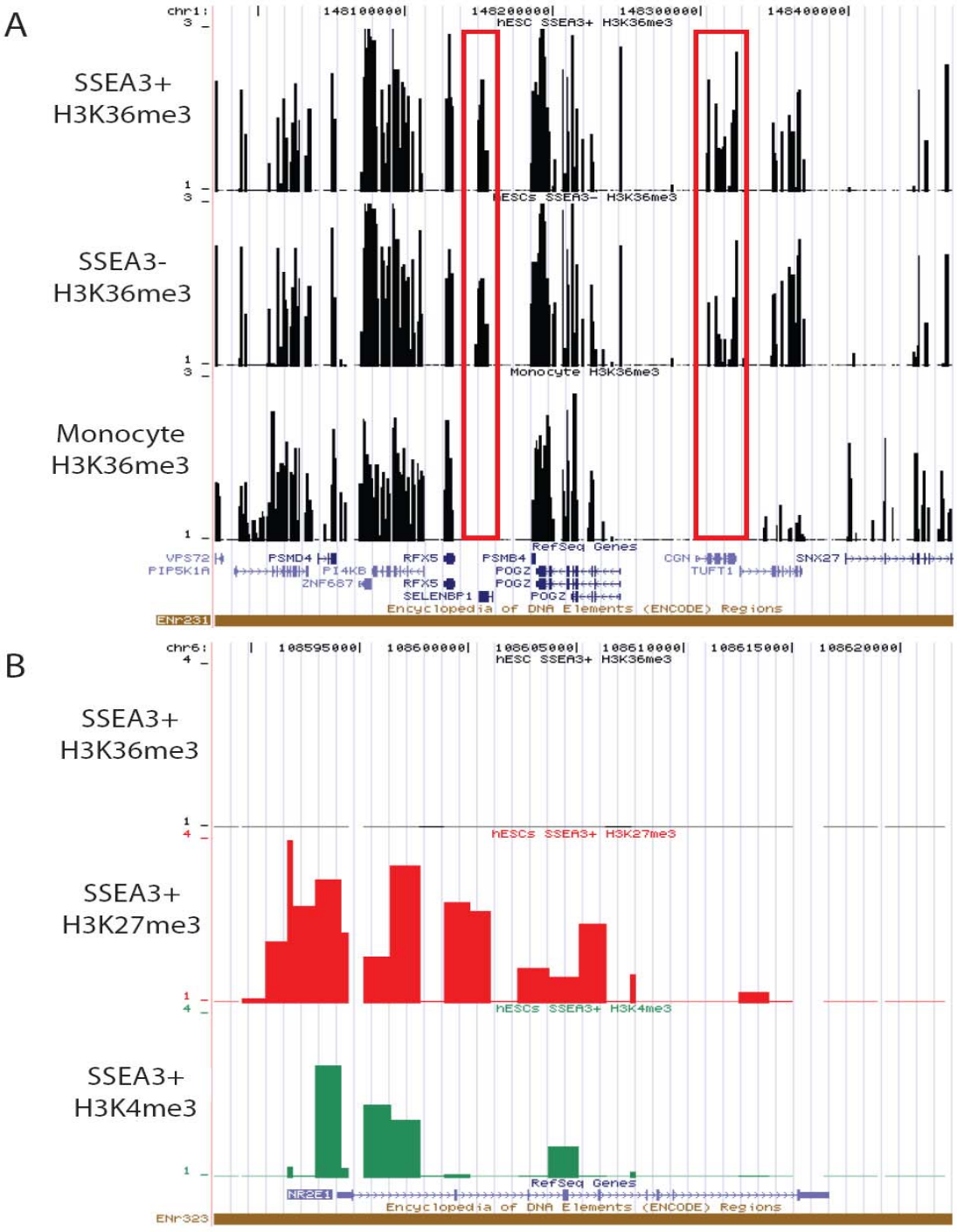


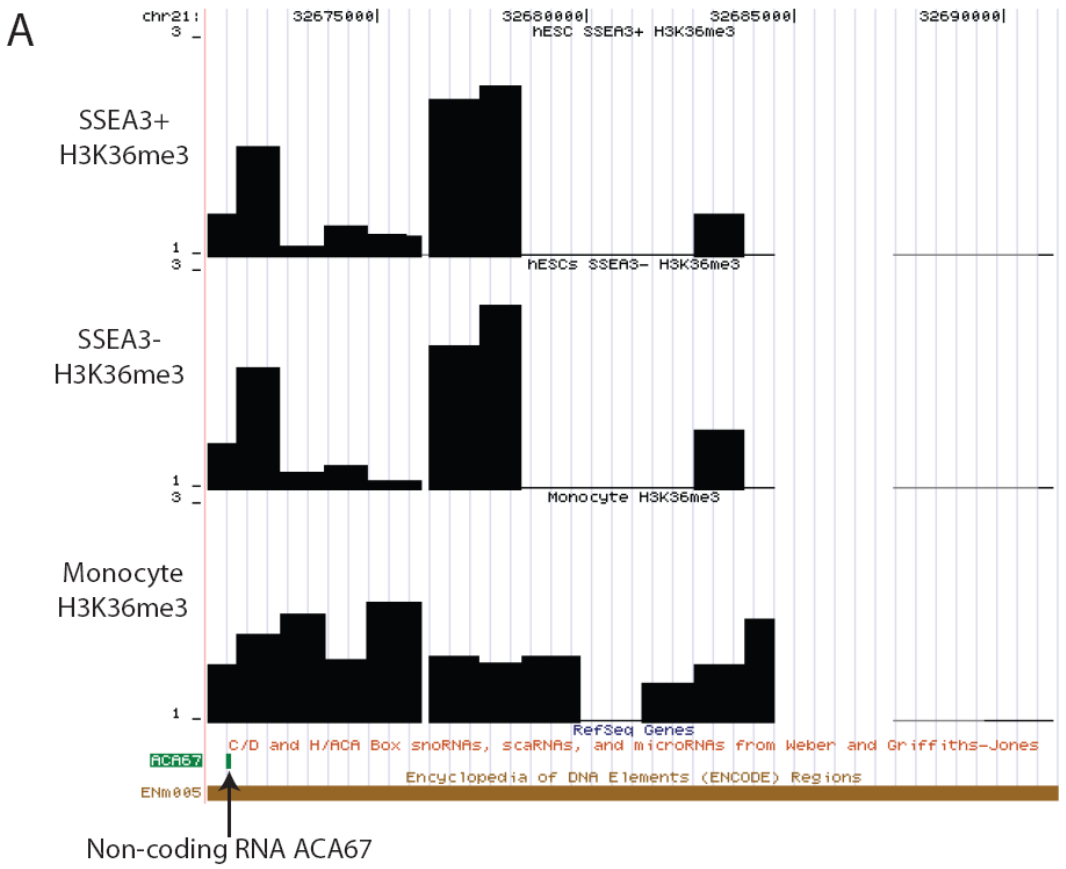
Figure 6.7: H3K36me3 annotates transcribed genes in hESCs and monocytes. Two screenshots from the UCSC genome browser (Kuhn *et al*., 2007) showing the relationship between H3K36me3 and transcriptional activity. Panel A: ENCODE region Enr231 is a gene dense region in which the majority of genes are associated 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 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.


A
chr21:
32675000
32680000
32685000
32690000
hESC SSEA3+ H3K36me3
SSEA3+
H3K36me3
hESCs SSEA3- H3K36me3
SSEA3-
H3K36me3
Monocyte H3K36me3
Monocyte
H3K36me3
RefSeq Genes
C/D and H/ACA Box snoRNAs, scaRNAs, and microRNAs from Weber and Griffiths-Jones
ACA67
Encyclopedia of DNA Elements (ENCODE) Regions
ENm005
Non-coding RNA ACA67

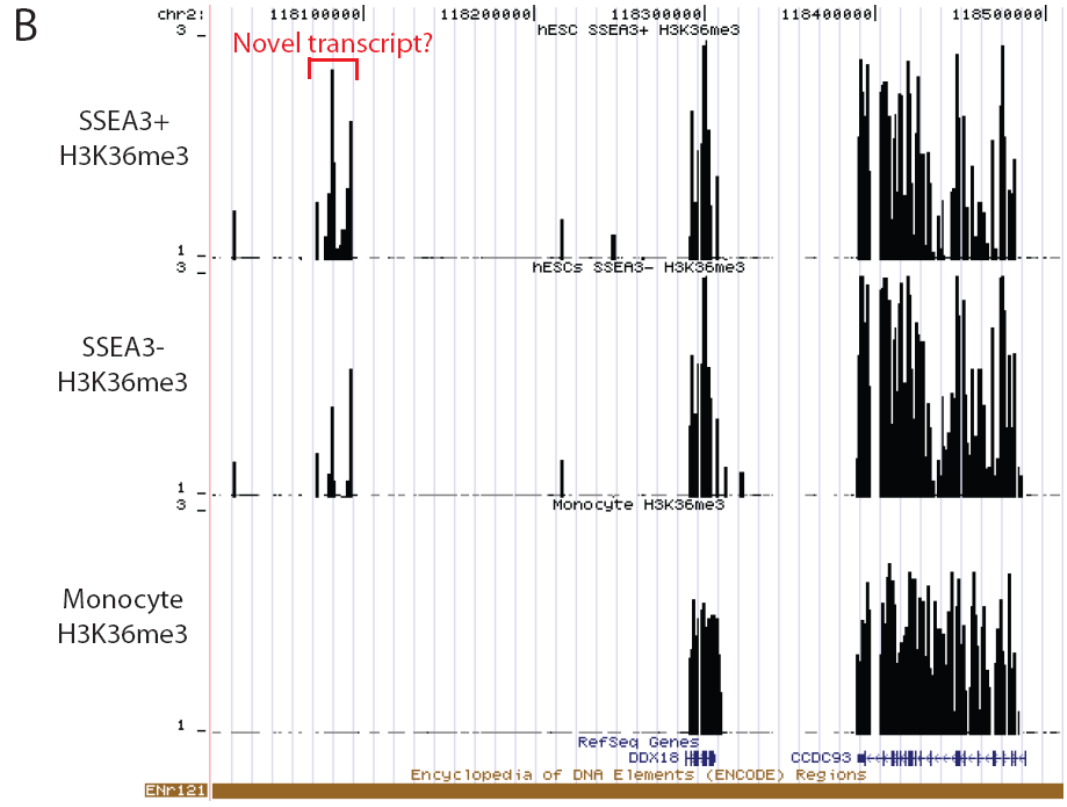
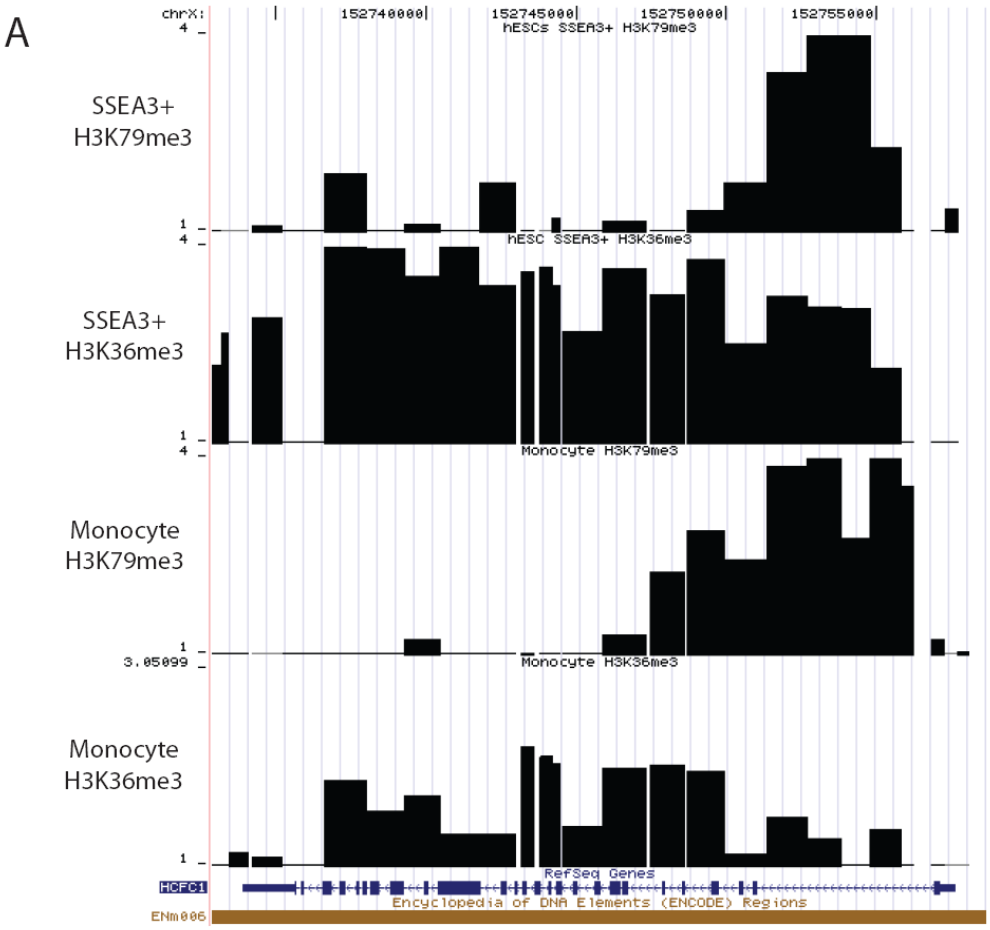

B
chr2:
118100000
118200000
118300000
118400000
118500000
Novel transcript?
hESC SSEA3+ H3K36me3
SSEA3+
H3K36me3
hESCs SSEA3- H3K36me3
SSEA3-
H3K36me3
Monocyte H3K36me3
Monocyte
H3K36me3
RefSeq Genes
DDX18
CCDC93
Encyclopedia of DNA Elements (ENCODE) Regions
ENr121

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 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.


A
chrX:
152740000
152745000
152750000
152755000
hESCs SSEA3+ H3K79me3
SSEA3+
H3K79me3
hESC SSEA3+ H3K36me3
SSEA3+
H3K36me3
Monocyte H3K79me3
Monocyte
H3K79me3
Monocyte H3K36me3
Monocyte
H3K36me3
RefSeq Genes
HCFC1
Encyclopedia of DNA Elements (ENCODE) Regions
ENm006

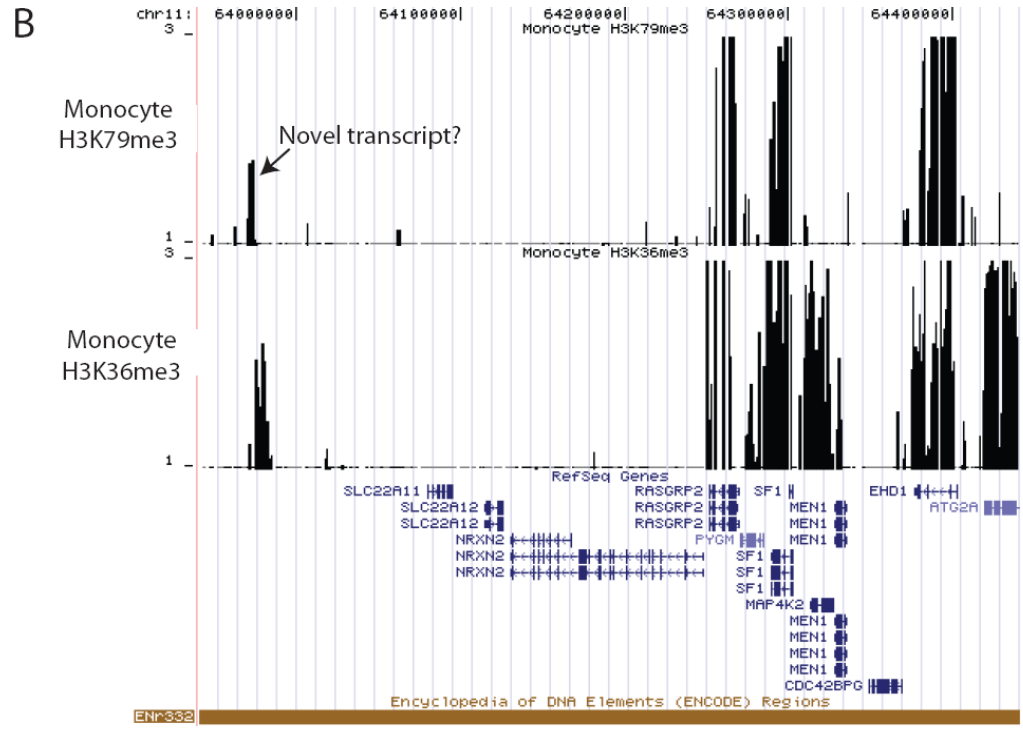

B
chr11:
64000000
64100000
64200000
64300000
64400000
Monocyte H3K79me3
Monocyte
H3K79me3
Novel transcript?
Monocyte H3K36me3
Monocyte
H3K36me3
RefSeq Genes
SLC22A11
SLC22A12
SLC22A12
NRXN2
NRXN2
NRXN2
RASGRP2
RASGRP2
RASGRP2
PYGM
SF1
SF1
SF1
SF1
MAP4K2
MEN1
MEN1
MEN1
MEN1
MEN1
MEN1
MEN1
CDC42BPG
EHD1
ATG2A
Encyclopedia of DNA Elements (ENCODE) Regions
ENm332

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 tri-methylation 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 liked 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 H14K16 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.

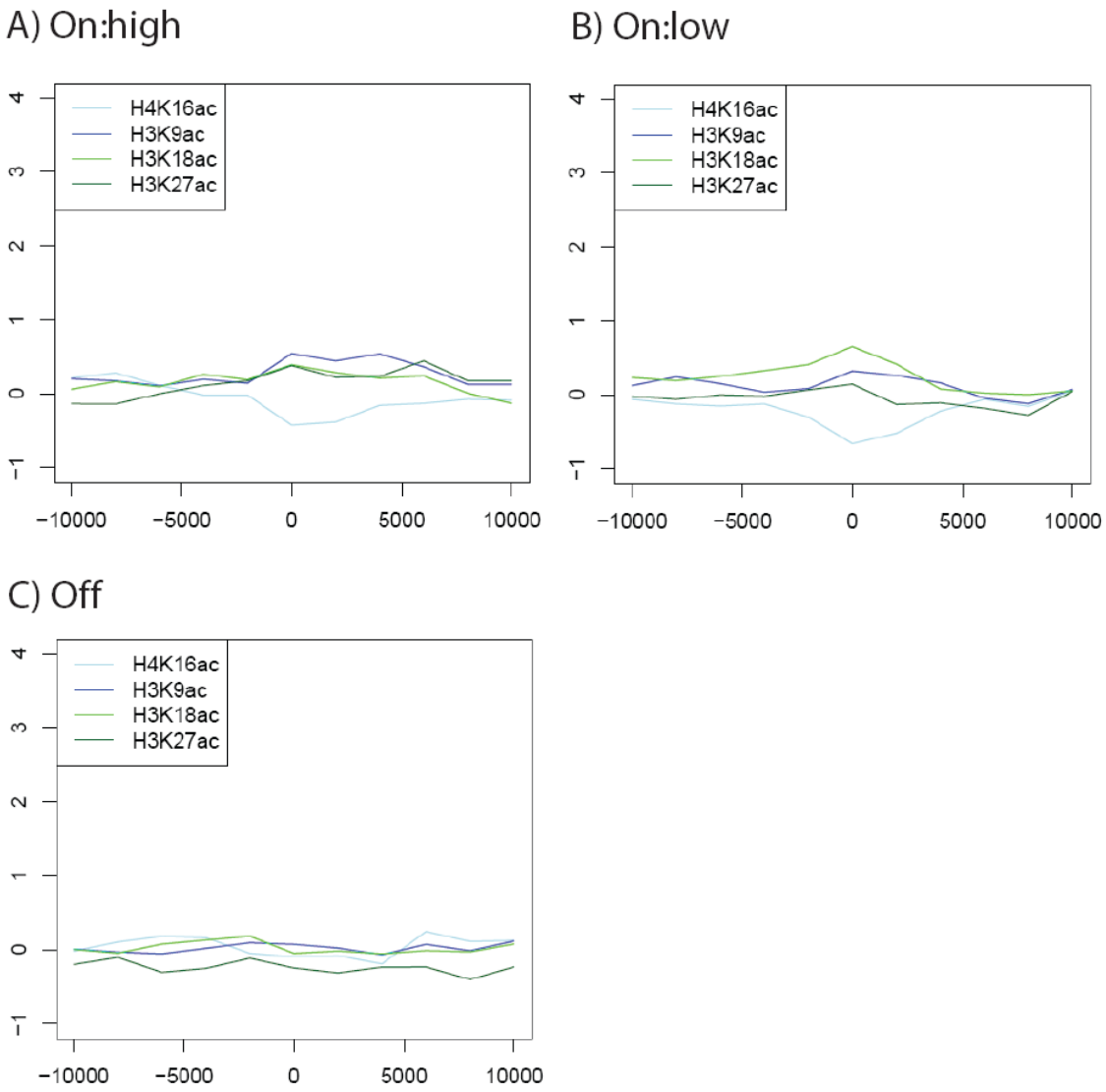


Figure 6.10: Histone acetylation patterns at active and inactive genes. H3K9ac, H3K18ac, H3K27ac, and H4K16ac average modification patterns are presented 10 kb upstream and downstream of TSSs associated with highly expressed (panel A), lowly expressed (panel B) and inactive (panel C) genes. The scale on the x-axis of each plot represents distances (bp’s) upstream (negative values) and downstream (positive values) of TSSs (represented by 0). Log_2 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.

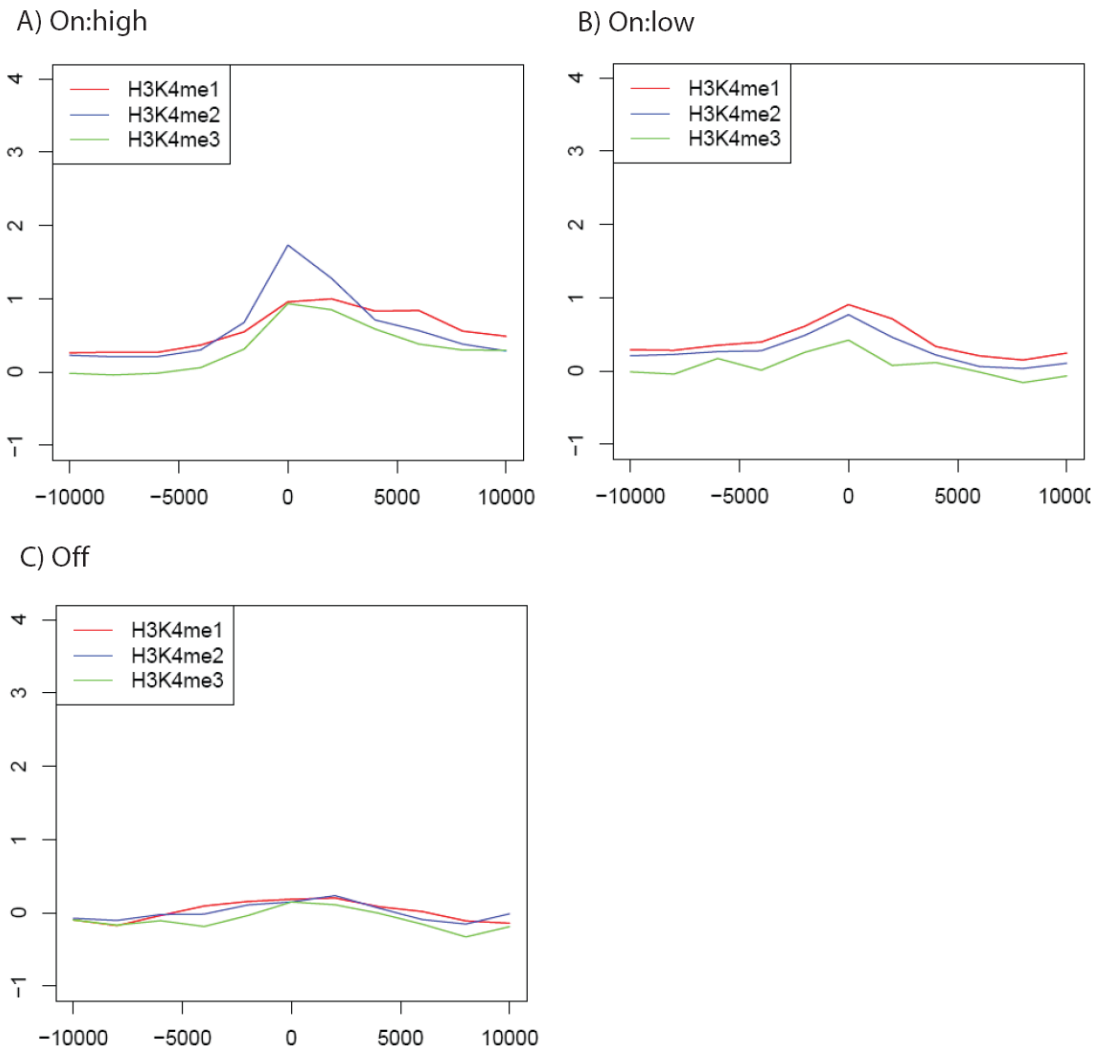


Figure 6.11: Histone H3K4 methylation patterns at active and inactive genes. H3K4me1, H3K4me2, and H3K4me3 average modification patterns are presented 10 kb upstream and downstream of promoters associated with highly expressed (panel A), lowly expressed (panel B) and inactive (panel C) genes. The scale on the x-axis of each plot represents distances (bp's) upstream (negative values) and downstream (positive values) of known TSSs and z-scored $\log_2$ 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.

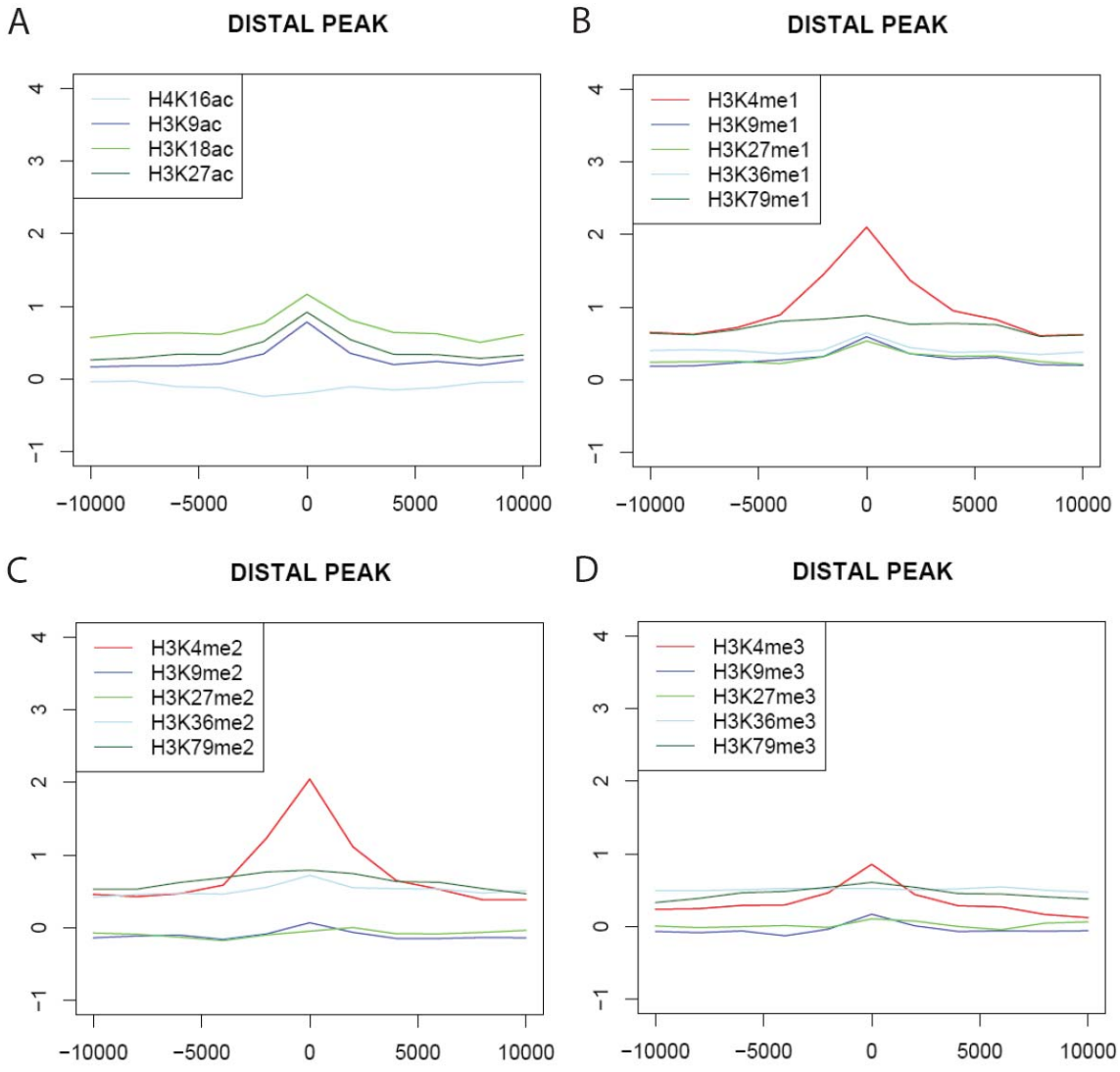


Figure 6.12: Histone acetylation and methylation patterns at distal enhancer/repressor elements. Histone acetylation and methylation patterns are presented 10 kb upstream and downstream of distal elements. The average pattern of H3K9, H3K18, H3K27 and H4K16 acetylation at distal elements is presented 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 examine at the promoter regions of these genes (Figure 6.13).

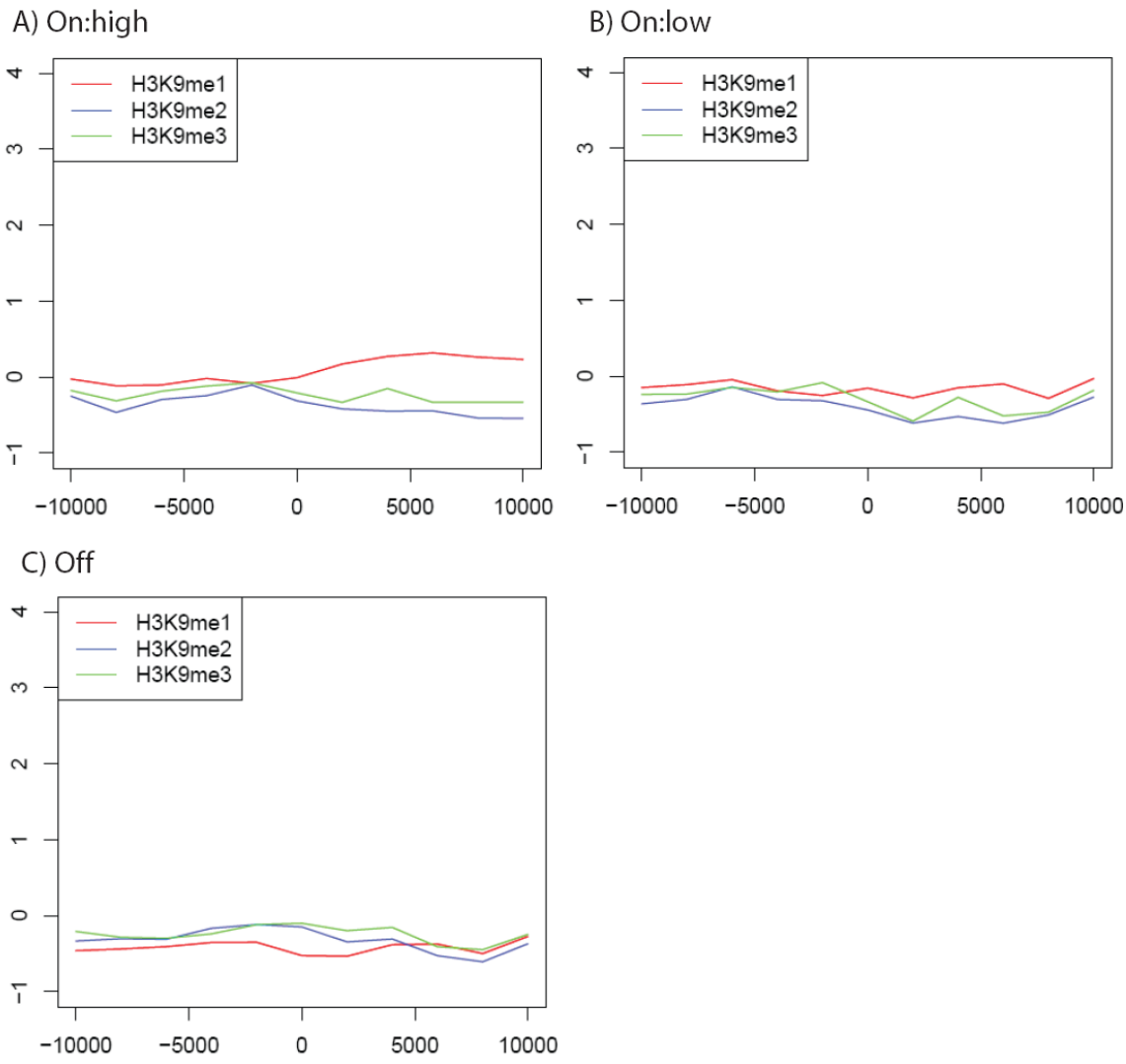


Figure 6.13: Histone H3K9 methylation patterns at active and inactive genes. H3K9me1, H3K9me2, and H3K9me3 average modification patterns are presented 10 kb upstream and downstream of TSSs associated with highly expressed (panel A), lowly expressed (panel B) and inactive (panel C) genes. The scale on the x-axis of each plot represents distances (bp's) upstream (negative values) and downstream (positive values) of known TSSs and z-scored $\log_2$ 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.

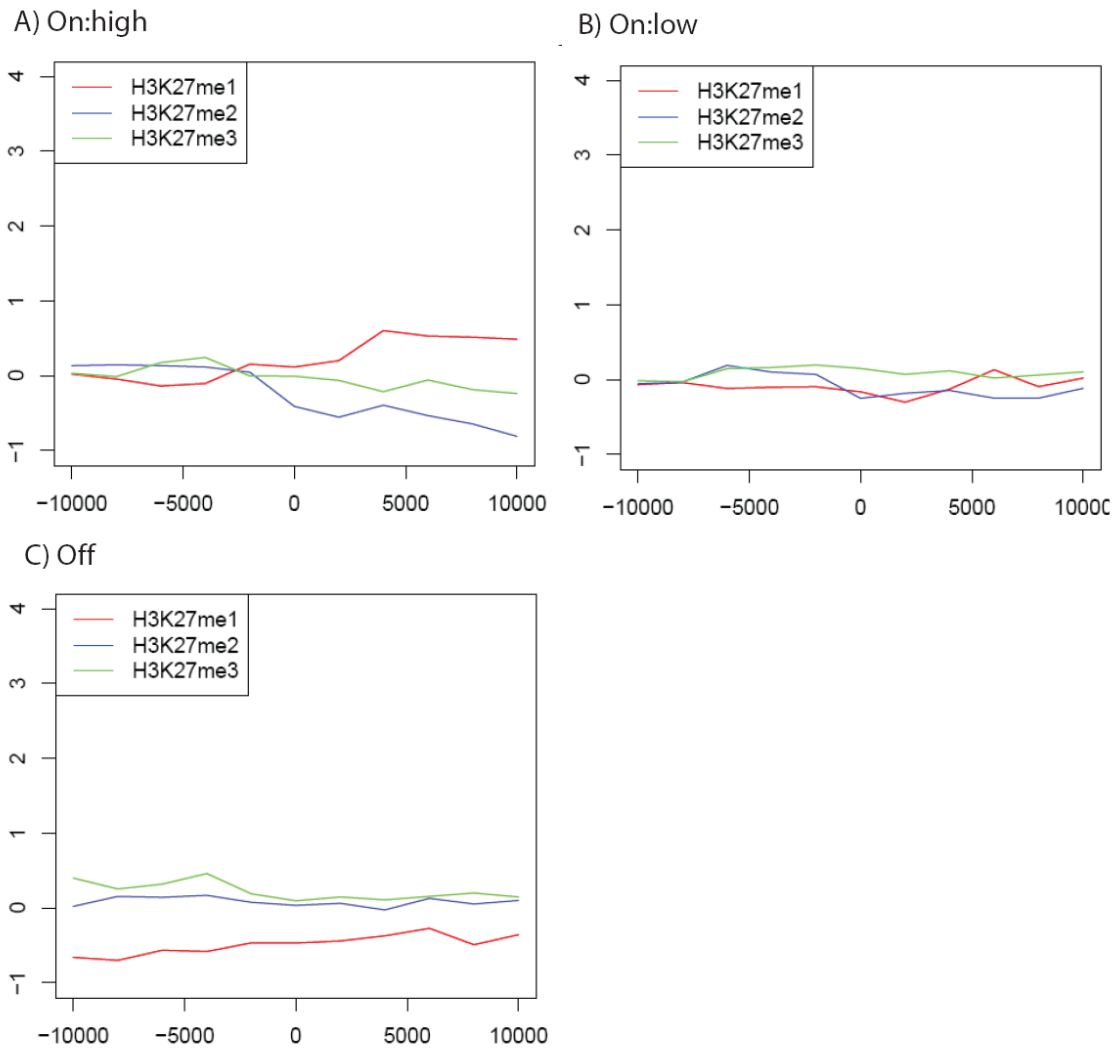


Figure 6.14: Histone H3K27 methylation states at active and inactive genes. Average H3K27me1, H3K27me2, and H3K27me3 modification patterns are presented 10 kb upstream and downstream of TSSs associated with highly expressed (panel A), lowly expressed (panel B) and inactive (panel C) genes. The scale on the x-axis of each plot represents distances (bp's) upstream (negative values) and downstream (positive values) of known TSSs and z-scored $\log_2$ 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.

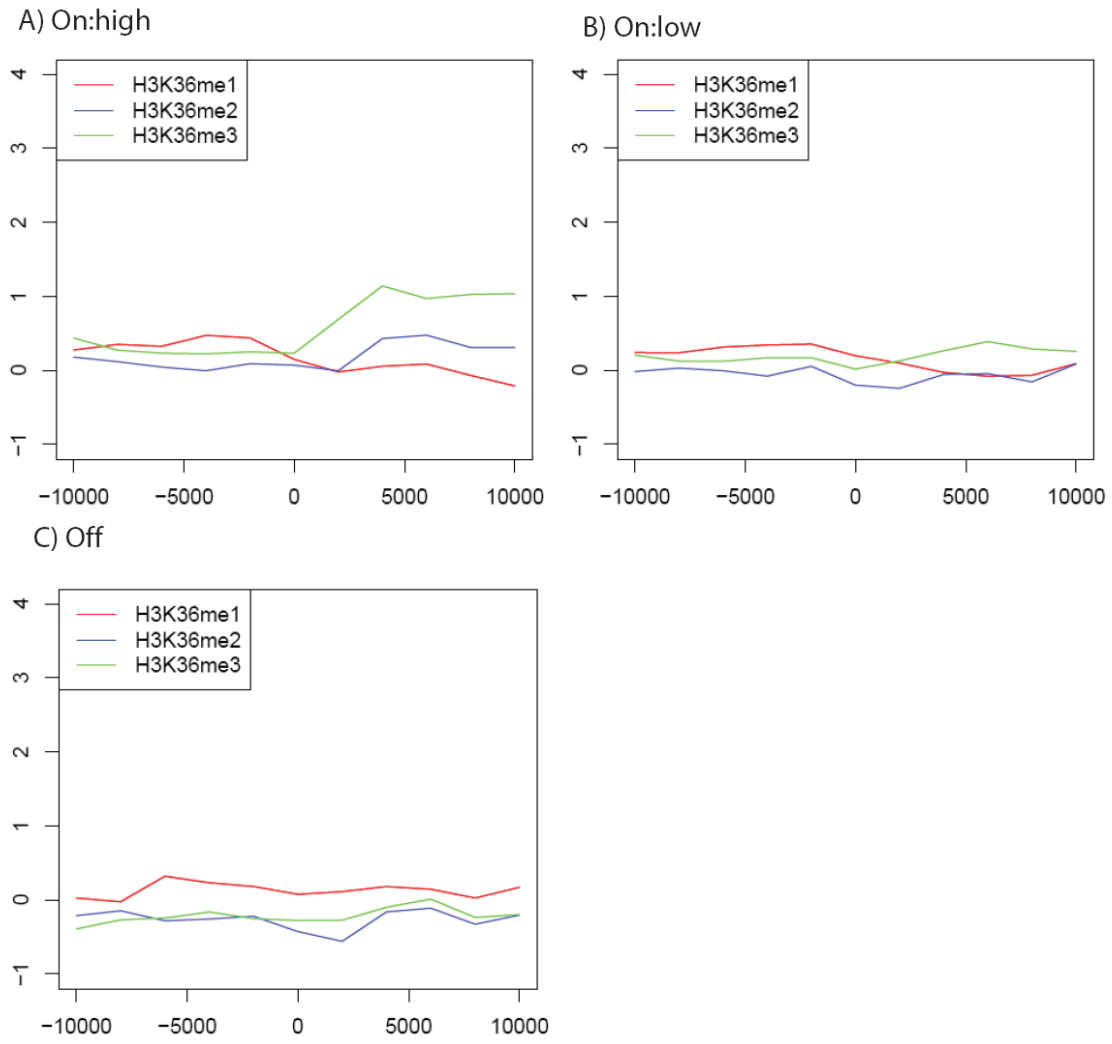


Figure 6.15: Histone H3K36 methylation patterns at active and inactive genes. H3K36me1, H3K36me2, and H3K36me3 average modification patterns are presented 10 kb upstream and downstream of TSSs associated with highly expressed (panel A), lowly expressed (panel B) and inactive (panel C) genes. The scale on the x-axis of each plot represents distances (bp's) upstream (negative values) and downstream (positive values) of known TSSs and z-scored $\log_2$ 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).

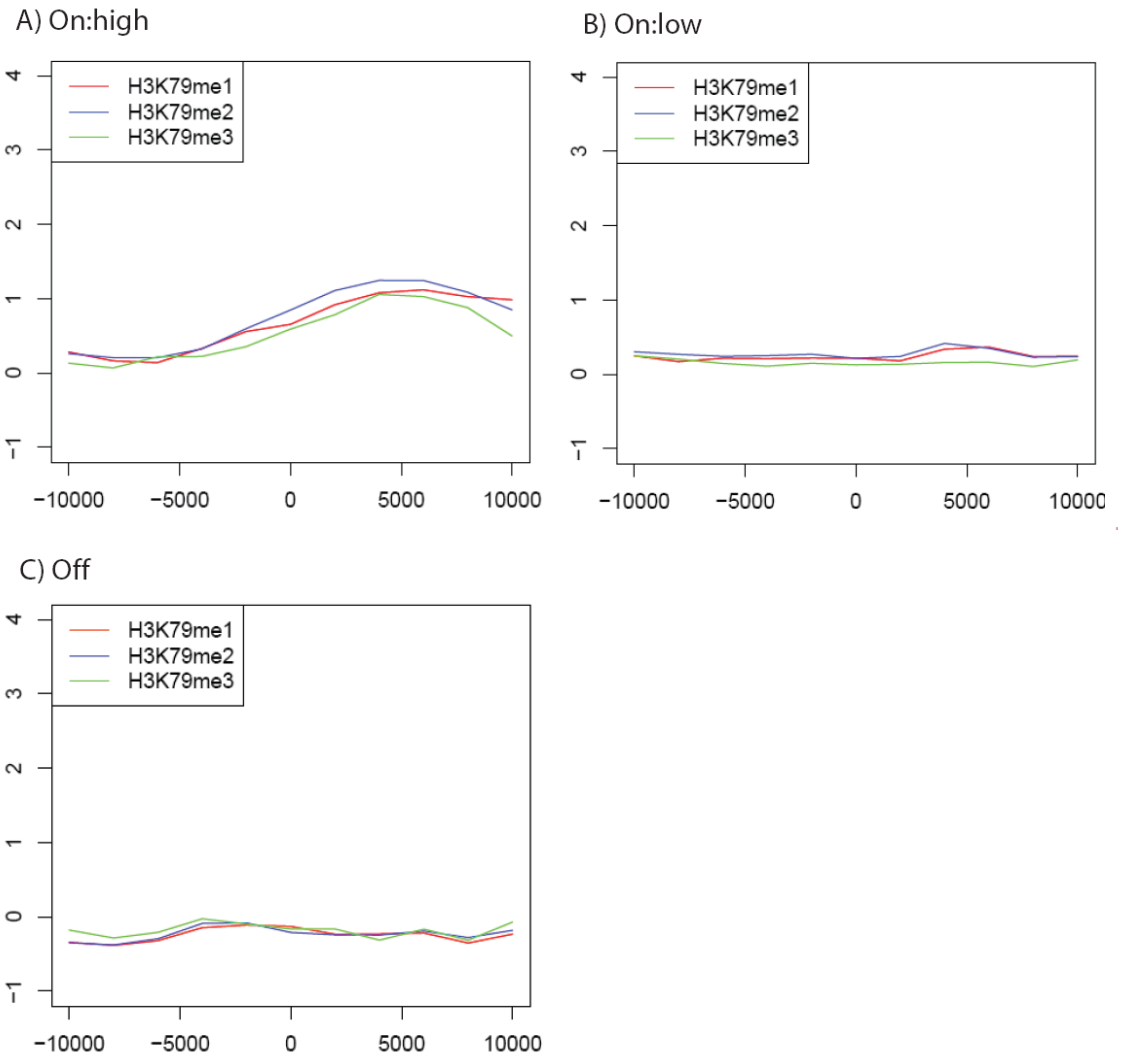


Figure 6.16: Histone H3K79 methylation patterns at active and inactive genes. H3K79me1, H3K79me2, and H3K79me3 average modification patterns are presented 10 kb upstream and downstream of TSSs associated with highly expressed (panel A), lowly expressed (panel B) and inactive (panel C) genes. The scale 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 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 of 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+ and SSEA3- cells which may reflect the similar developmental state of these two cell types. In addition, SSEA3- cells are derived from SSEA3+ cells while the monocytes used in this study were not derived from SSEA3+ cells and are therefore less likely to have similar characteristics.

While SSEA3+ only or SSEA3- 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+ hESCs, SSEA3- 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 compexes, 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+ 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.

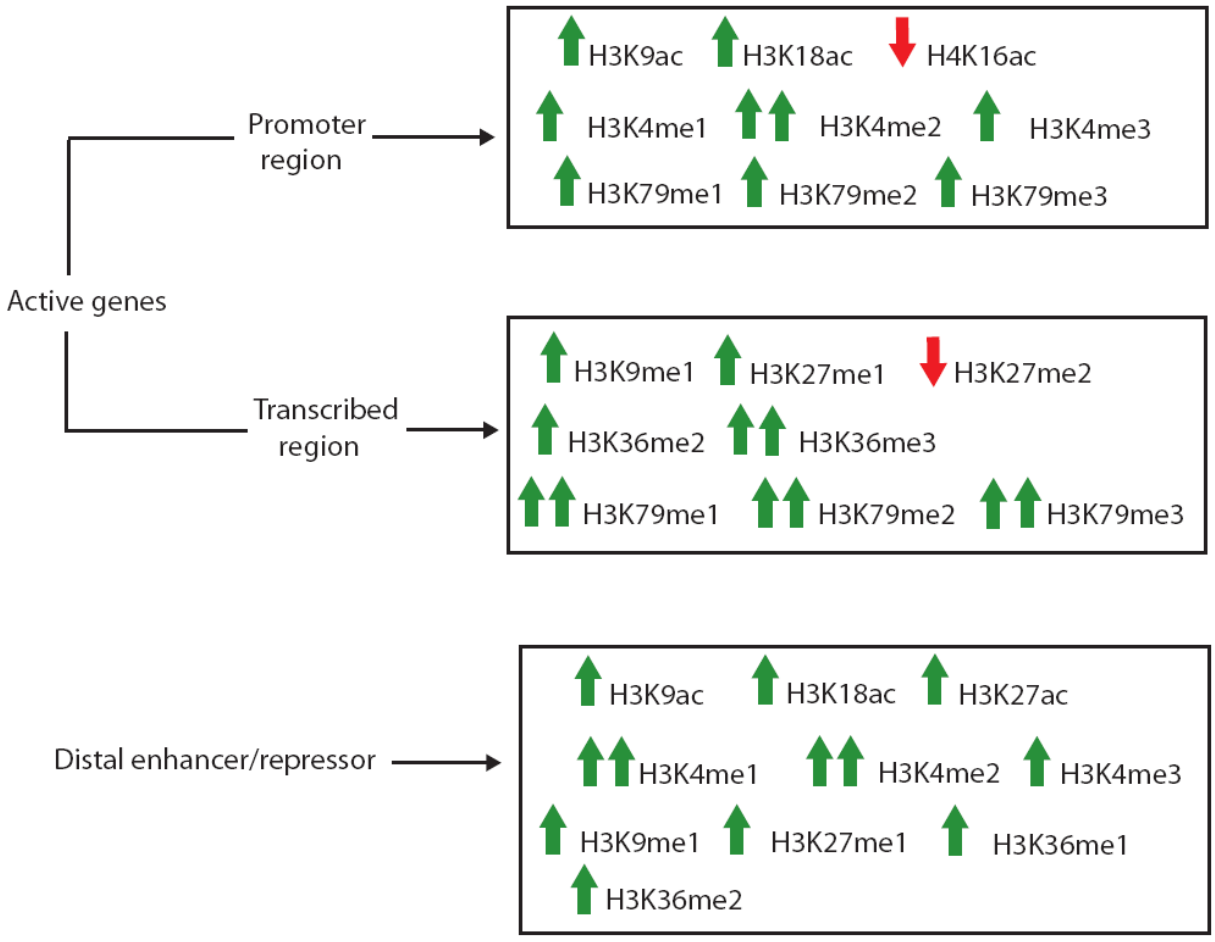


Figure 6.17: A consensus histone code for regulatory and transcribed regions in human monocytes. This code was proposed based on observations on the presence/absence of 19 histone acetylation and methylation modifications in monocyte cells. The modifications are shown associated with active promoters, 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 (Heintzamn *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+ 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.