Chapter 7

Summary and Future work

7.1. Summary of the work presented in this thesis

Understanding how gene expression is regulated is a fundamental issue in molecular biology. It is known that non-coding regulatory elements such as promoters, enhancers, and insulators play a central role in the regulation of gene expression in complex eukaryotic genomes. Therefore a detailed understanding of the location and properties of these elements in the human genome sequence would help further our understanding of gene regulation. However, the identification of these elements in the human genome is a difficult task as enhancers and insulators are often located far from genes in large stretches of non-coding DNA. Therefore, the aim of the work presented in this thesis was to perform large-scale profiling of a range of *in vivo* DNA-protein interactions which could be used to identify the location of regulatory elements in 1% of the human genome. This was achieved by the development of a number of ChIP-chip assays which were used in combination with a highly sensitive and reproducible ENCODE microarray platform. In addition, an improved ChIP-chip method was also developed for the large scale study of DNA-protein interactions in limited cell populations. This method was then applied to investigate the chromatin state of undifferentiated human embryonic stem cells, differentiated human embryonic stem cells, and human CD14+ monocytes. These cell types represented different stages of cellular differentiation and the dynamics of chromatin-mediated gene regulation during differentiation was analysed. The results and interpretations of this work were presented in Chapters 3, 4, 5, and 6 and are summarised below. Future work which follows on from the work presented in this thesis is also outlined in this concluding chapter.

7.1.1. Application of microarray technology for large-scale characterisation of promoter and distal enhancer/repressor elements in the human genome (Chapter 3)

In Chapter 3, a number of ChIP-chip assays were applied to characterise regulatory elements in the human genome on a large scale using the ENCODE microarray. Those assays included detecting genomic regions associated with H3 acetylation, H4 acetylation, H3K4me1, H2K4me2, and H3K4me3. Histone H2B and H3 density, taken to reflect nucleosome occupancy, was also assessed using this array platform. Un-amplified ChIP DNA material derived from 10^7 cultured cells could be used to perform hybridisations with the ENCODE array. The use of un-amplified ChIP DNAs reduced sources of non-biological biases resulting from amplification methods, ensuring that the ENCODE platform performed reproducibly in independently performed assays.

A detailed analysis of H3 acetylation, H4 acetylation, H3K4me1, H3K4me2, and H3K4me3 distribution showed that promoter and distal enhancer/repressor elements could be distinguished in terms of their histone K4 methylation signatures. H3K4me3 was found to be most prominent at promoters of active genes while H3K4me1 was the predominant modification at distal elements. Both promoters of actively transcribed genes and distal regulatory elements associated with histone modifications were shown to be associated with H2B and H3 nucleosome depletion; this feature was not evident at promoters of non-transcribed genes. Taken together these results were the first demonstration that the ENCODE microarray could be used for the large-scale identification and characterisation of regulatory elements in the human genome.

7.1.2. ChIP-chip analysis of the insulator binding protein CTCF and its associated transcription factors (Chapter 4)

Chapter 4 describes the development of ChIP-chip assays for the study of insulator elements in the human genome. A ChIP-chip assay was developed to detect genomic sequences associated with the insulator binding protein CTCF. This assay was developed with the SCL tiling microarray and then applied to study CTCF in the ENCODE regions. The CTCF ChIP-chip assay was validated by comparisons with other experimental defined CTCF binding sites located in the ENCODE regions. Many of the CTCF binding

sites found in this study contained a previously characterised consensus CTCF binding motif. CTCF binding sites were found to be very common in the genome as over 500 binding sites were identified in the ENCODE regions in K562 cells – scaled to the whole genome, this would suggest that there as many as 50,000 CTCF sites genome-wide.

CTCF binding sites were found at promoters, exons, introns and intergenic regions. Over 50% of genes were found to be flanked in their entirety by two CTCF sites and CTCF sites were also found to flank clusters of related or unrelated genes. Multiple CTCF binding sites within genes were also observed. Over half of CTCF binding sites were located far (> 5kb) from TSSs, consistent with a role in insulator function. The remainder was located at TSSs and distal putative enhancer/repressor elements and may be involved in transcriptional activation/repression. Insulator elements may be associated with accessible chromatin domains as CTCF sites were found to co-localise with DNase I hypersensitive sites, FAIRE peaks and regions of histone H2B/H3 depletion. Furthermore, a number of CTCF sites were located at the boundary between active and inactive chromatin domains defined by the presence of H3K27me1 and H3K27me3 respectively, suggestive of a barrier insulator function. CTCF binding sites were found to be highly conserved between K562 and U937 cells indicating that the location of insulators is conserved between cell types.

In addition, known binding partners of CTCF - USF1, USF2, and mSin3a – were studied in K562 cells. Analysis of USF1 and USF2 binding sites revealed a widespread distribution with no bias towards promoter, exonic, intronic and intergenic locations. However, mSin3a interactions were found to be heavily biased towards promoter locations and binding was found to be a good predictor of active gene expression. One third of CTCF sites overlapped with interactions for one or more of the three factors, over half of which were located at TSSs, while 22% and 23% of overlapping interactions were located at distal enhancer/repressor elements and other locations (thought to be insulator elements) respectively. Therefore, the majority of mSin3a, USF1, and USF2 overlapping interactions with CTCF may be involved in regulating gene expression by interacting with promoter and distal elements. However, only a minority of putative insulators were associated with CTCF and either USF1 or USF2, suggesting that their co-localisation may not be a general paradigm for insulator barrier function in the human genome.

7.1.3. Development of a modified ChIP-chip method for use with fewer cells (Chapter 5)

Chapters 3 and 4 demonstrated that ChIP-chip assays could be used to identify various types of regulatory elements using immortalised human cell lines. However, when working with many types of primary human cells and other scarce samples, cell numbers limit the application of ChIP-chip technology. Therefore ChIP assays were developed which could be used in combination with microarray hybridization to detect histone modifications from as few as 10^4 K562 cells with similar efficiencies as assays performed with 10^7 cells. This protocol was developed using the detection of regions enriched for histone acetylation and H3K4me3 at the SCL locus as a model system. The antibody to chromatin ratio was found to be critical for the reliable detection of known regulatory elements when cell numbers were reduced. Other factors such as the level of protein-G agarose and data normalization procedures to account for non-specific interactions were also assessed as means of improving the detection of *bona fide* regulatory elements. The sensitivity of the method was confirmed by quantitative PCR and further validated with ChIP-chip assays for a number of other histone modifications including H3K27me3, H3K36me3 and H3K79me3.

7.1.4. Application of the modified ChIP-chip method to study chromatin regulation during differentiation (Chapter 6)

In this Chapter the modified ChIP-chip approach was applied to study the patterns of histone modifications at various stages of cellular differentiation. The patterns of H3K4me3, H3K27me3, H3K36me3, and H3K79me3 were examined in SSEA3+ hESCs, SSEA3- hESCs, and CD14+ monocytes. Analysis of H3K4me3 and H3K27me3 at promoters revealed three classes of promoters: H3K4me3 monovalent, H3K4me3/H3K27me3 bivalent promoters, and H3K27me3 monovalent promoters. Bivalent promoters were found to be a feature of all three cell types, many of which were associated with genes involved in the transcriptional regulation of developmental. A number of bivalent promoters were common to the three cell types but a large proportion of monocyte bivalent promoters were specific to this cell type. Analysis of promoter CpG

content showed that the majority of high CpG content promoters (often associated with housekeeping genes) were enriched with H3K4me3(alone or bivalent) in the three cell types while few low CpG content promoters (oftent associated with tissue-specific promoters) were enriched for H3K4me3. Comparisons of chromatin state during differentiation showed that promoters in SSEA3+ and SSEA3- hESCs displayed a similar histone modification profile for many of the same genes whilst monocytes had histone modifications associated primarily with different sets of genes. H3K36me3 was found to be associated with the entire length of primary transcripts in hESCs and monocytes. A number of non-coding RNAs were also associated with H3K36me3 in hESCs. H3K79me3 was located at the 5' end of actively transcribed genes in the three cell types.

7.1.5. Identification of a consensus histone code in human monocytes (Chapter 6)

The modified ChIP-chip method allowed for the detailed study of 19 histone modifications in monocytes. Chromatin modification maps were created for four histone acetylation modifications (H3K9ac, H3K18ac, H3K27ac, and H4K16ac) and the mono-, di-, and tri-methylation states of H3K4, K27, K36, and K79. This resulted in the identification of a detailed consensus histone code for promoters of transcribed genes, transcribed coding regions and distal enhancer/repressor elements. A histone code for inactive genes was not readily identifiable based on the modifications studied in this thesis. Promoters of transcribed genes were associated with H3K9ac, K18ac, and K27ac as well as all three H3K4 methylaytion states, in particular H3K4me2. Actively transcribed coding regions were associated with high levels of H3K36me3 and all three states of H3K79 methylation. Distal enhancer/repressors were associated with the three histone H3 acetylation events as well as the three H3K4 methylation states. H3K4me1 and H3K4me2 were most prominent at distal elements. H3K9, K27 and K36 mono-methylation states were also associated with distal elements. Taken together these results suggest that the modified ChIP-chip method can be used to provide a detailed characterization of the histone code of regulatory and coding elements in the human genome.

7.2. Future Work

The work described in this thesis has shown how *in vivo* DNA-protein interactions can be assayed across large regions of the genome in numerous cell types using a microarray-based method. Here, the Sanger ENCODE microarray resource was used to detect other DNA-protein interactions associated with regulatory elements in 1% of the human genome, but the data could be extrapolated to suggest mechanisms and functional relationships which represent found across the entire human genome. For example, there are at least 34 amino acid residues in the core histone proteins that are chemically modified in human cells (Garcia *et al*., 2007) and the distribution and function of many of these modifications is not well understood. In addition, approximately 2600 transcription factors have been identified in the human genome (Babu *et al*., 2004), and how these factors act individually and collectively to regulate transcription programs is a major challenge in the post-genome era. However, for genomic annotation and identification of all regulatory sequences genome-wide, there is a need to completely interrogation of the entire genome using high-thoughput methods such as CHIP-chip. Other techniques which could be used for the high-throughput identification of genome features which influence gene expression are also outlined below.

7.2.1. Further development of ChIP-chip assays for characterising regulatory elements and the role of histone modifications in gene regulation

The work presented in this thesis involved mapping the location of a number of histone modifications to analyse their distribution in 1% of the human genome and explore the relationships between modifications. However, while this represented a detailed study of histone modifications, it was by no means exhaustive for the number of modifications that could be investigated. For example, methylation of H4K20 was not examined in this study and H4K20me3 has been shown to be associated with repressive chromatin while the mono-methylation form has been found at the promoter and coding regions of active genes (Schotta *et al*., 2004, Talasz *et al*., 2005, Vakoc *et al*., 2006). In addition, histone argine methylation was not examined in this study; histone H3 arginine methylation has been shown to regulate pluripotency in the early mouse embryo (Torres-Padilla *et al*., 2007) while asymmetrically dimethylated H3R2 has been shown to regulate the

deposition of H3K4me3 in the yeast genome (Kirmizis *et al*., 2007). The function of arginine methylation is still very much unexplored in the human genome. Phosphorylation of histone residues has been implicated in both transcriptional activation and repression (Nowak and Corces, 2004). Ubiquitylation and sumoylation are two large histone modifications whose roles in regulating gene expression are least understood. Ubiquitylation can be repressive or activating depending on the site modified (Zhu *et al*., 2005; Wang *et al*., 2004). Sumoylation is the only repressive modification described in yeast (Shiio and Eisenman, 2003) but its role in mammalian cells is not known.

Furthermore, an understanding of how the various histone modifications interact with other proteins and how this is translated into specific biological outputs is not well understood. Two models have been proposed - the direct and effector mediated models (Jenuwein and Allis, 2001; Strahl and Allis, 2000). In the direct model, histone modifications such as acetylation and phosphorylation directly affect interactions between basic histone proteins and negatively charged DNA (Shogren-Knaak *et al*., 2006; Ahn *et al*., 2005). The effector model proposes that histone modifications are 'read' by cognate non-histone binding proteins, known as effectors. Effector proteins can compact chromatin by cross-linking nucleosomes (Francis *et al*., 2004), enhance the binding of the RNA polymerase complex (Vermeulen *et al*., 2007) or recruit chromatin remodeling factors (Jenuwein and Allis, 2001). To date there are 11 known protein domain families which 'read' the various histone acetylation, methylation and phosphorylation modifications (Taverna *et al*., 2007). Effector proteins, such as those which contain bromodomains, are known to interact with lysine acetylation marks (Dhalluin *et al*., 1999) and are often found in histone acetyltransferases and chromatin remodeling complexes (Zeng and Zhou, 2002). Chromodomains of HP1 and Polycomb are known to recognize di- and tri-methyl H3K9 and H3K27 respectively (Lachner *et al*., 2001; Min *et al*., 2003). Double chromodomain, double tudor domain and PHD finger containing proteins have been shown to recognize H3K4me3 (Flanagan *et al*., 2005; Huang *et al*., 2006; Pena *et al*., 2006; Li *et al*., 2006). The implications of multiple domains which recognize the same modification is not known.

The role of histone modifications in regulating gene expression and other nuclear processes is further complicated by the realization that histone modifications rarely occur

in isolation - mass spectrometry studies are beginning to show that multiple modifications occur on the same histone tail (Cosgrove, 2007). The promiscuity of histone modifications with respect to recruiting effector binding proteins (there are more than 10 effector proteins known to bind H3K4me3) suggests that one histone modification is not sufficient to stably recruit a complex (Ruthenburg *et al*., 2007). The co-existence of multiple modifications within a given tail or chromatin domain may serve to dictate the recruitment of complexes (Ruthenburg *et al*., 2007). For example, the PHD finger and bromodomain of BPTF could be used to simultaneously bind H3K4me3 and an acetylated lysine residue. It will be important to gain an increased understanding of how histone modification combinations and effector proteins operate and dictate functional outcomes.

Future ChIP-based approaches will be important in determining the co-localisation of modifications and their effector proteins. However, the ability to perform ChIP assays to detect histone modifications and the binding of effector proteins is dependent on the production of antibodies which perform in ChIP assays and recent years has seen a large increase in the number of 'ChIP grade' antibodies becoming available from commercial suppliers. The continued availability of antibodies to novel histone modifications and histone associated proteins will ensure that future ChIP studies can be performed to gain a more complete understanding of the relationship between histone modifications and the proteins which recognize these marks.

7.2.2. Applications of microarrays for characterising other features of the human genome

While nucleosomes represent the fundamental repeating unit of chromatin and play a crucial role in a number of genome functions, higher order chromatin structures also contribute to the regulation of the genome. It has been suggested that chromatin forms 50–200 kb chromosomal loops which are attached to the nuclear matrix (Bode *et al*., 2000; Heng *et al*., 2004). The anchor points of the DNA to the nuclear matrix have been termed matrix attachment regions (MARs), which have been implicated in the control DNA replication and gene expression (Jenke et al., 2004; Amati and Gasser, 1990). The DNA and proteins associated with the nuclear matrix can be isolated by extraction of

histones with high salt or mild detergent followed by restriction enzyme or DNase I treatment to remove all DNA except matrix attached DNA. Sumer and colleagues isolated MAR DNA and hybridized it to a BAC/PAC array to define a 2.5 Mb region of MAR enriched DNA at a human neocentromere (Sumer *et al*., 2003). Ioudinkova and colleagues have also used arrays to map MARs at the chicken α-globin domain (Ioudinkova *et al*., 2005). However high-resolution mapping of MARs across a large region of the human genome has yet to be carried out.

Physical interactions between regulatory elements play an important role in gene regulation. Current evidence suggests that chromatin loops allow these elements to physically interact with their target genes to repress or silence transcription (Spilianakis *et al*., 2005; Splinter *et al*., 2006). Two genomic regions can be tested for interaction using the chromosome conformation capture (3C) technique (Dekker *et al*., 2002), in which protein-DNA interactions are cross-linked using formaldehyde, followed by digestion with a restriction enzyme. Interacting fragments are ligated together and can be quantified individually by qPCR. The 3C method has recently been modified for the high-throughput detection of sequences which interact with a particular DNA fragment in a method known as 5C (3C carbon copy) (Dostie *et al*., 2006). In this method T7 and T3 primers are ligated to interacting fragments and used as primers for PCR amplification. Custom microarrays can then be used to analyse the junction fragments or alternatively high-throughput sequencing can be used to detect interacting fragments. A circular chromosome conformation capture (4C) method has also been developed (Zhao *et al*., 2006 b), which involves the circularization of 3C products. Inverse PCR from the known test fragment was used to amplify interacting fragments which were then cloned and sequenced. These sequences were then used to create a customized microarray. Another 4C method (3C-on-chip) was developed by Simonis and colleagues (2006) in which circular inverse PCR was used to amplify interacting sequences which were then identified directly using a custom designed array. These methods could be used to identify which genes are controlled by particular regulatory elements, which is often difficult to decipher based purely on their proximity to the gene which they are thought to control.

While chromatin modifying enzymes are known to co-localize to DNA replication sites (McNairn and Gilbert, 2003), the relationship between replication timing, chromatin and the regulation of gene expression is not fully understood. Microarrays can be used to assess replication timing in the human genome (Woodfine *et al*., 2004; Jeon *et al*., 2005) and could also be correlated with information on the location of DNA replication origins and chromatin modifications. The replication of DNA requires the formation of origin recognition complexes (ORCs) (Gilbert, 2007). ChIP-chip analysis of ORC proteins has been used to identify replication origins in the yeast genome (Wyrick *et al*., 2001; Xu *et al*., 2006) but has approach has not been applied to identify replication origins in the human genome. ORC complexes may be only formed during the S-phase of the cell cycle, meaning that human cells may need to be synchronized to identify ORC-DNA interactions.

7.2.3. Testing the proposed histone code

The analysis of 19 histone modifications in CD14+ monocytes led to model for a histone code, in which a number of key histone modifications were associated with various regulatory elements and gene activity. This histone code is consistent for the genes contained on the ENCODE array in human monocytes, however it remains to be seen if this code is consistent for all the genes in the human genome. This would be an extremely large undertaking considering the number of histone modifications analysed in this study. Furthermore, a complete understanding of the histone code will be much more complex to decipher than the genetic code given the large number of known histone modifications. Multiple histone modifications can also be located on the same histone tail and different histone modifying enzymes are often responsible for catalysing the same modification. Different effector proteins may be recruited by various histone modification combinations, adding further complexity to understanding the code. Methods such as sequential ChIP and mass spectrometry will allow for histone modification and modification/effector combinations to be determined.

Finally, elucidating the role of these epigenetic events in normal development and disease is the ultimate goal. Having developed a modified ChIP method for the study of histone modifications in limited cell populations, there is scope for employing this method to

study the histone code in rare cell populations at different stages of development. This could be used to gain a greater understanding of the chromatin basis of cellular differentiation and pluripotency. The development of the new generation sequencing technologies, which may be more cost effective and experimentally-efficient than whole genome microarray studies, will also help to realize the goal of annotating regulatory elements in the human genome during normal development and disease. Multiple layers of ChIP information will be overlaid onto the genetic code – an undertaking much more complex project than even the sequencing of the the human genome. It will also be important to understand how chromatin state is altered during abnormal development and disease (De Gobbi *et al*., 2006). Such chromatin-based studies will be logical extensions of the sequencing of human genomes associated with pathological states. Once this has been done, the relationships between specific genomic sequence elements, their sequence variants, chromatin function and disease will be routinely determined.

However, when complete, this information will give us an unprecedented understanding of our genome. There is also the possibility of extending our understanding of the histone code to other mammalian genome in order to determine the conservation of the code amongst species. A more complex histone code may provide a distinction between higher and lower eukaryotes and there is already evidence for more elaborate patterns of histone modifications in metazoans (Garcia *et al*., 2007).

7.2.4. Further characterisation of putative CTCF insulator elements

CTCF binding sites identified in this study could be tested for enhancer-blocking activity or barrier activity using functional plasmid-based assays developed by Felsenfeld and colleagues (Chung *et al*., 1993; Pikaart *et al*., 1998). Enhancer-blocking activity is tested by cloning a CTCF binding site between an erythroid-specific enhancer and promoter driving expression of G418 resistance. The construct is then stably transfected into K562 cells and a CTCF binding site which displays enhancer-blocking activity gives a reduced number of colonies relative to the control plasmid when grown on agar. A CTCF binding site can also be tested for barrier activity by flanking the interleukin 2 receptor (IL-2R) expression cassette with CTCF binding sites. The construct is then stably transfected into the human genome and expression of IL-2R is monitored by flow cytometry. Expression

of any transgene stably integrated into the genome is normally silenced after a prolonged period in culture due to the spread of inactivating histone modifications. However, when an IL-2R transgene is flanked by CTCF sites associated with barrier activity constant expression is maintained. However, these assays are not suitable for high-throughput screening of insulator function. Mukhopadhyay and colleagues developed a technique for the high-throughput screening of insulator function (Mukhopadhyay *et al*., 2004), in which sequences from a mouse CTCF target site library were inserted into a plasmid to interfere with SV40 enhancement of toxin-A reporter gene activity. The number of cell clones increased dramatically when the toxin-A gene was insulated from the SV40 enhancer. Total DNA and DNA from the emerging clones could be differentially labeled and hybridized to a microarray. An increase in microarray signal would be observed for functionally selected sequences.

7.3. Final thoughts

The findings presented in this thesis illustrate how microarray technology is being applied to study global gene regulatory events. Given the recent advances in microarray and sequencing technology, the post-genome goal of annotating the complete repertoire of functional elements in the human genome sequence will surely be realized. The findings of this thesis will make an initial contribution towards this goal.