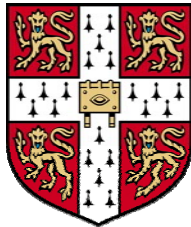


Identification and Characterisation of Regulatory Elements on Human Chromosome 20q12-13.2

Pelin Akan

A thesis submitted in partial fulfilment of the requirements of University of
Cambridge for the degree of
Doctor of Philosophy



Clare Hall, University of Cambridge

September 2006

This dissertation is the result of my own work and includes nothing that is the outcome of work done in collaboration. The dissertation does not exceed the length limit set by the Biology Degree Committee

ABSTRACT

A nearly-finished sequence of the human genome was published in 2004 (IHGSC, 2004). The task ahead is now to complete the structural and functional annotation of the genome. Although much progress is being made in annotating the coding part of the genome, identification of regulatory regions remains a challenge in current genomics. The aim of this thesis is to identify and characterise promoters and other regulatory elements in a 10 Mb region on human chromosome 20q12-13.2. This region was chosen because of (i) its biological importance, as it is associated with a number of medical conditions such as type II diabetes and obesity (ii) the availability of a detailed transcription map of the region, which is particularly valuable for the experimental and computational analyses carried out in this study.

Firstly, I describe the identification and characterization of core promoter elements using computational methods. Here, promoters are studied at the sequence and structural level in an attempt to discover novel signals for promoter identification *in silico*. Candidate promoters are also correlated to genomic features and expression data from two cell lines, HeLa S3 and NTERA-2 clone D1.

In the subsequent chapter, I describe the systematic validation of annotated (candidate) promoters using dual luciferase reporter assays in the two cell lines, HeLa S3 and NTERA-2 clone D1. Analyses include the assessment of promoter activities in synergy with the SV40 enhancer. The differential response of core promoters to the enhancer is then associated with the presence of transcription factor binding motifs predicted by a transcription factor binding motif prediction program (MAPPER).

In the final results chapter, I present my findings in chromatin immunoprecipitation (ChIP) studies carried out on an in-house spotted DNA array. This array was constructed with 2kb overlapping plasmid clones spanning 3.5 Mb of the investigated

region (gene rich segment; 72 protein-coding genes). ChIP analyses (both cell lines, each in triplicate) included 7 antibodies against modified histones, one antibody against RNA polymerase II and one antibody against the transcription factor CTCF to investigate the histone code and transcriptional activity of the region in two cell lines. Additionally, the sequence features of potential distal regulatory elements are studied *in silico* to recognize any common features of such elements.

Acknowledgements

I would like to thank my supervisor Panos Deloukas, who always believed in me and helped me in every way throughout my PhD. I feel truly privileged to work with him.

I would also like to thank Rhian Gwilliam for her help with the construction of the array and guiding me during my years at Sanger, Phillippe Couttet for his invaluable advice, Team 66, Team 67, and Team 62 for allowing me use their hybridisation station. I would also like to show my appreciation for both members of my thesis committee, David Vetrie and Richard Clarkson.

My friends, Paula, Lawrence, Samuel (dear Sammy), Mark (the funniest guy ever, can't live without his jokes) and Joel! You were always there for me, and I wouldn't have finished if it wasn't for you. Paula, I can't find words to describe your loyal friendship, Mark is the luckiest guy ever! Lawrence, our long conversations, complicated feelings and past, tennis, psychology, philosophy... You will always be a part of me. I love you! Sammy, oh my dear Sammy, you are my hero, your integrity, strength, warm personality, there is nothing in the world you cannot have... Mark, you are the most sane among all of us, your acute comments always fine-tuned my thoughts sometimes even my actions! And your *Best of Pink Floyd* album is on its way! Joel, a bird who adores freedom, I miss you very much, I am still waiting for the credits to finish before leaving the movie theatre sometimes! When are we going to cook together again and watch your wonderful pictures around the world?

Friedel, your common-sense, great taste in music and everything, really. When I think about you, I feel like it is all worth it since I have such a friend like you. Please send more pictures of cutest little Karl Kasimir!

And more, Caroline (my gorgeous blonde and smart friend with a beautiful personality, ok, I think I just described a perfect girl-friend), Amina, Richard, Rhian...

I cannot thank you enough...

My sweetheart Nils Jean Nikolaj Martin! First my dear friend then my love. I feel like you are my other half, I don't think anyone can understand me as the way you do. You brighten my life, my heart. I now understood why I had taken such painful decisions in the past, it was all for you. I adore your good heart. Seni çok seviyorum.

And my brother, Ozgun, and dad, without your support and confidence in me, I cannot get through my most difficult times during these years. Ozgun, you always said right things to me, your thoughts always amazed me, my most talented brother, I love you. Dad, you were not only my father but my friend who always listened to me and believed in me and gave me strength...

TABLE OF CONTENTS

ABSTRACT	I
ACKNOWLEDGEMENTS	III
LIST OF TABLES	IX
LIST OF FIGURES	XII
LIST OF FIGURES	XII
LIST OF ABBREVIATIONS.....	XIX
1 INTRODUCTION	1
1.1 Why is it important to find regulatory elements?.....	2
1.2 Regulatory DNA Elements.....	4
1.2.1 Promoters.....	5
1.2.2 Enhancers	8
1.2.3 Silencers	11
1.2.4 Insulators (Boundary Elements)	12
1.2.5 Locus Control Regions	14
1.2.6 Experimental and Computational Efforts for Locating Regulatory Sequences	16
1.3 Transcriptional Machinery.....	21
1.3.1 RNA Polymerase II Transcription Machinery	21
1.3.2 Transcription Factors	26
1.4 Chromatin and Transcription	35
1.4.1 Histone Phosphorylation.....	41
1.4.2 Histone Ubiquitinylation	42
1.4.3 Histone Sumoylation	43
1.4.4 Histone Acetylation	43
1.4.5 Histone Methylation	46
1.5 Chromosome 20	51
1.5.1 Zooming into 20q12-13.2.....	52
2 MATERIALS AND METHODS	55
2.1 Gene Reporter Assays	55
2.1.1 Primer Design.....	55
2.1.2 Sub-cloning.....	57
2.1.3 Cloning inserts into Gene Reporter Vectors	61
2.1.4 Ligations	65
2.1.5 Transfections	66
2.1.6 Dual Luciferase Reporter Assays	67
2.2 Cell Culture	67
2.2.1 HeLa S3 cell line	67
2.2.2 NTERA-2 clone D1 cell line	68
2.2.3 Trypsinizing Cells.....	68

2.2.4	Freezing the Cells	68
2.2.5	Thawing the Cells	69
2.3	Affymetrix Expression Arrays.....	69
2.3.1	Isolation of Total RNA from cells	69
2.3.2	Checking the quality of the RNA	69
2.3.3	Preparation of RNA for hybridization to Affymetrix Expression Arrays	70
2.3.4	Fragmentation of Biotin Labelled cRNA.....	73
2.3.5	Eukaryotic Target Hybridization	73
2.3.6	Washing and Staining of the probe array.....	74
2.3.7	Array Scanning	74
2.4	Construction of Tilepath Arrays	74
2.5	Chromatin Immunoprecipitation (ChIP)	76
2.5.1	Cell Harvesting	76
2.5.2	Chromatin Fixation.....	76
2.5.3	Cell Lysis and extraction of chromatin.....	78
2.5.4	Sonication	78
2.5.5	Antibodies.....	79
2.5.6	Pre-clearing of the Chromatin.....	79
2.5.7	Immunoprecipitation	80
2.5.8	Addition of Protein G Agarose Beads to Antibody/Chromatin Mixture	80
2.5.9	Washing Antibody/Chromatin/Bead Complexes.....	80
2.5.10	Elution of the Antibody/Chromatin Complexes	80
2.5.11	Reversal of Cross-linking and Digestion of RNA	81
2.5.12	Digestion of Proteins and Recovery of DNA	81
2.6	Real-time PCR	82
2.6.1	Assessing ChIP efficiency	82
2.6.2	Validation of ChIP on chip enrichment	83
2.7	Preparation of ChIP samples for hybridization onto microarrays	83
2.7.1	Labelling of ChIP samples.....	83
2.7.2	Removal of unlabelled nucleotides.....	84
2.7.3	Competitive Hybridization of Labelled Samples onto microarrays.....	84
2.7.4	Array Scanning	85
2.8	Solutions	86
3	PROMOTER AND GENE EXPRESSION PROFILING ON HUMAN CHROMOSOME 20 CYTOBAND Q12-13.2.....	89
3.1	Overview of the region	89
3.2	Feature Predictions.....	90
3.2.1	CpG islands	90
3.2.2	Promoter Predictions	92
3.2.3	Transcription Start Site Prediction - Eponine	93
3.2.4	Overall Summary of Predictions.....	94
3.3	Sequence and Structure Topology of Promoter Sequences.....	95
3.3.1	Sequence Topology of Promoter Sequences.....	95
3.3.2	Structural Topology of Promoter Sequences	103
3.4	Expression Profile of 20q12-13.2 using Affymetrix Arrays	108
3.4.1	Coding Genes and Transcripts.....	109
3.4.2	Expression Profiles and CpG Islands.....	110
3.4.3	Pseudogenes and Processed Transcripts	110

4 IDENTIFICATION AND CHARACTERIZATION OF CORE PROMOTER ELEMENTS BY GENE REPORTER ASSAYS..... 111

4.1	Reporter Genes	112
4.1.1	Chloroamphenicol Acetyltransferase (CAT)	113
4.1.2	Green Fluorescent Protein (GFP)	113
4.1.3	Luciferase	114
4.2	Dual Luciferase Assays	115
4.2.1	Positive and Negative Controls.....	117
4.2.2	Determining Promoter Activities.....	117
4.3	Cloning of candidate promoter fragments	119
4.3.1	Optimisation of Reporter Assays.....	121
4.4	Cell lines	123
4.4.1	HeLa S3	123
4.4.2	NTERA2 clone D1	123
4.5	Transfection Results	124
4.5.1	Promoter Activities in HeLa S3.....	124
4.5.2	Promoter Activities in NTERA-D1	124
4.5.3	Comparison of promoter activities between the cell lines	125
4.6	Promoter Activities in synergy with SV40 Enhancer	131
4.6.1	Promoters in synergy with SV40 Enhancer in HeLa S3 cells.....	131
4.6.2	Promoters in synergy with SV40 Enhancer in NTERA-D1 cells	134
4.6.3	Sp1 binding sites on 71 promoters.....	137
4.6.4	Promoters active only in the presence of SV40 Enhancer	139
4.6.5	Promoters Repressed by SV40 Enhancer	143
4.7	Summary	147

5 ANALYSIS OF 3.5 MB REGION ON 20Q12-13.2 USING CHIP-CHIP . 149

5.1	Unsuccessful Antibodies.....	151
5.2	ChIP.....	152
5.2.1	Optimization of Crosslinking Conditions	155
5.2.2	Immunoprecipitation and subsequent steps in ChIP on chip	157
5.3	Custom-made 3.5 Mb Tilepath Array of human 20q12-13.2.....	159
5.4	Determining Spot Intensities	162
5.4.1	Validation of ChIP-chip enrichments by Real-time PCR	165
5.5	Analysis of RNA polymerase II binding sites by ChIP.....	175
5.5.1	Results in HeLa S3 cells	176
5.5.2	Results in NTERA-D1 cells.....	180
5.6	Histone Modifications on Transcription Start Sites	184
5.6.1	Results in HeLa S3 cells	185
5.6.2	Results in NTERA-D1 cells.....	196
5.7	Histone Modifications marking possible regulatory elements.....	204
5.7.1	HeLa S3 cells.....	206
5.7.2	NTERA-D1 cells	215
5.8	Histone Modification involved in transcriptionally inactive regions	220
5.8.1	H3K27me3	220
5.8.2	H3K9me2	227

5.9	CTCF	228
5.10	Summary	238
6	DISCUSSION AND CLOSING REMARKS.....	255
	REFERENCES	312

LIST OF TABLES

Table 1.1. A number of diseases associated with mutations on regulatory DNA elements and the genes affected. This table is reproduced from reference (Maston et al., 2006).	5
Table 1.2 General Transcription Factors (GTFs), their protein composition and possible functions in the transcription initiation process. TAF corresponds to TATA-box binding protein associated factor. This figure is adapted from reference (Thomas and Chiang, 2006).	24
Table 1.3 A list of coactivators engaging different enzymatic processes to activate transcription. This table is adapted from reference (Lonard and O'Malley, 2005).	35
Table 1.4 Chemical properties of histone proteins	36
Table 1.5. Lysine residues that are acetylated on amino terminal tails of histones	43
Table 1.6 Histone Acetyltransferases that are active in humans and their known cellular functions.	45
Table 1.7 Histones and their particular residues that can accept methyl groups and the associated enzymatic machinery. * denotes HMT that does not contain a SET domain.	47
Table 1.8. Gene Distribution of human chromosome 20q12-13.2. Human Genome Organization (HUGO) nomenclature are used for all genes.....	54
Table 2.1 Recipe for the PCR for amplification of candidate promoters.....	57
Table 2.2 Recipe for Ligation reaction for TA cloning	58
Table 2.3 Recipe for colony PCR	59
Table 2.4 Recipe for Restriction Enzyme Digestion	60
Table 2.5 Restriction Enzyme Pair for the digestion of vectors	63
Table 2.6 Restriction enzyme digestions of cloning vectors	64
Table 2.7 Reaction set up for vector dephosphorylation	65
Table 2.8 Total RNA yields and quality from both cell lines	69
Table 2.9 Critical Specifications of Human Genome U133A 2.0 Expression Analysis Array.....	70
Table 2.10 Recipe for the hybridization cocktail.....	73
Table 2.11. PCR recipe for amplification of DNA fragments to be spotted on the array.	75
Table 2.12 Chromatin Fixation Conditions	77
Table 2.13 Sonication Conditions	78
Table 2.14 Antibodies used in this study.....	79
Table 2.15 Reaction set up for real time PCR with Eurogentec real-time PCR kit	82
Table 2.16. Slide washing and drying protocol on Tecan HS4800 Pro hybridization station.	85
Table 3.1 Number of transcripts associated with CpG islands	92
Table 3.2. Bendability scores of all possible trinucleotides. Higher values (less negative) translate to higher bendability towards major groove.	104
Table 3.3. Expression Profiles of 96 representative transcripts associated or not associated with CpG islands	110
Table 4.1. Details of promoter fragments cloned to pGFP-1 reporter vector. TSS is denoted as +1....	122
Table 4.2 Promoter activities and transcript types.....	130
Table 4.3 Expression Profile and Promoter Activity of 50 genes in HeLa S3 and NTERA-D1 cell lines.	130
Table 4.4 Summary of the results obtained with SV40 Enhancer in HeLa S3 and NTERA-D1 cell lines	137
Table 4.5. The left column lists transcription factors that have binding sites on SV40 enhancer and right column lists its interaction partners.	141

Table 4.6 Transcription Binding Sites of three constructs carrying ZPF161 binding site downstream of TSS. Binding site coordinates are given relative to TSS and “R” denotes for the binding motif on the opposite strand. Note that ELMO2-003, which was repressed under the effect of SV40 enhancer, does not contain activators such as YY1 or NFKB1.	145
Table 4.7 Transcription Binding Site profile of the constructs that are repressed in HeLa S3 cells. Binding site coordinates are given relative to TSS and “R” denotes for the binding motif on the opposite strand.	145
Table 4.8 Transcription factor binding site profile of promoters whose activities are greatly enhanced in synergy with SV40 enhancer. The response levels to enhancer is normalized to map onto 0 to 1 range. Binding sites were given relative to the TSS at +1 and “R” denotes the opposite strand. .	146
Table 5.1. Working antibodies used in this study.	151
Table 5.2 The concentrations of crosslinking agents and corresponding incubation times to crosslink transcription factors to DNA.	156
Table 5.3 Expression profile of the genes whose TSSs are not contained by any working spot on the array in HeLa S3 and NTERA-D1 cells.	161
Table 5.4. Enrichment levels of seven antibodies used in this study for 23 transcripts representing 22 genes in 3.5 Mb region in HeLa S3 cells. “u” marks a signal coming from ~2kb upstream of the annotated start site of the region and “d” marks a signal coming from ~2 kb downstream of the annotated start site of the region. The “expression” column displays the expression status of the corresponding gene; “A” (Absent) stands for no expression while “P” (Present) means the gene is expressed. * The polII enrichment on this gene is reported although it is below the selected threshold. H3K27me3 and H3K9me2 antibody columns are omitted since none of the spots showed any enrichment with these antibodies.	188
Table 5.5. Enrichment levels of 30 H3K4me3 enriched TSSs with other antibodies used in this study in NTERA-D1 cells. In the expression column “A” stands for no expression and P denotes that the gene is expressed in NTERA-D1 cells. “u” and “d” denote that the signal is detected 2 kb upstream or downstream of the TSS respectively. * This signals is placed around 4 kb upstream of the KCNS1 TSS.	198
Table 5.6 Number of occurrences of different histone combinations at one site in HeLa S3 and NTERA-D1 cells.	205
Table 5.7 The enrichment profiles of 28 H3K4me2 enriched spots. Empty cells means that there was no significant enrichment with that of specific antibody. First two columns list the genomic start and end coordinates (NCBI, version 36) of the enriched spots, while the spot information column gives a short description of the sequence of the spot.	207
Table 5.8 The enrichment profiles of 56 enriched spots not in close proximity of any annotated start sites in NTERA-D1 cells. Empty cells means that there was no significant enrichment with that specific antibody. The first two columns list the genomic start and end coordinates (NCBI, version 36) of the enriched spots, while the spot information column gives a short description of the sequence of the spot.	219
Table 5.9 The enrichment profiles of 15 H3K27me3 enriched spots in HeLa S3. Empty cells means that there was no significant enrichment with that of specific antibody. First two columns list the genomic start and end coordinates (NCBI, version 36) of the enriched spots, while the spot information column gives a short description of the sequence of the spot.	222
Table 5.10 The enrichment profiles of 49 H3K27me3 enriched spots in NTERA-D1 cells. Empty cells means that there was no significant enrichment with that specific antibody. The first two columns list the genomic start and end coordinates (NCBI, version 36) of the enriched spots, while the spot information column gives a short description of the sequence of the spot.	226
Table 5.11 The summary of CTCF enriched spots that are close to or contain an annotated start site in both cell lines. H3K4me3 enrichments of the spots are also given in both cell lines.	229
Table 5.12. CTCF enriched spots that lie within intergenic regions. The adjacent CTCF-enriched spots are treated as one region.	232
Table 5.13 CTCF enriched regions which fall within introns.	236
Table 5.14. Expression profile of the genes whose start sites are enriched with CTCF in HeLa S3 and NTERA-D1 cells.	239

Table 5.15. ChIP Signals of 83 coding transcripts obtained from nine antibodies used in HeLa S3 cells. Since H3K27me3 did not show any enrichment on any start site, it is omitted. (*) Signals coming from <i>NEURL2</i> are attributed to <i>PPGB</i> since the latter is expressed while the former is not. (#) This signal is omitted since the corresponding spot has a high sequence similarity to a ubiquitously expressed gene elsewhere in the genome.	250
Table 5.16 ChIP Signals of 83 coding transcripts obtained from nine antibodies used in HeLa S3 cells. Since H3K9me2 did not show any enrichment on any start site, it is omitted. Signals coming from <i>NEURL2</i> are attributed to <i>PPGB</i> since the latter is expressed while the former is not. (#) This signal is omitted since the corresponding spot has a high sequence similarity to a ubiquitously expressed gene elsewhere in the genome.	253
Table 5.17 Enrichment profiles of nine genes whose start sites were enriched with H3K27me3 in NTERA-D1 cells in both cell lines.	254

LIST OF FIGURES

Figure 1.1 Functional genomic elements aimed to be identified by the ENCODE pilot phase. The indicated methods are being employed to this end. This figure is adapted from reference ENCODE, 2004.....	2
Figure 1.2 The number of genes and transcripts are taken from Ensembl version 39, NCBI build 36.....	3
Figure 1.3 Sequence elements that are found on metazoan core promoters; BRE (TFIIB-recognition element), TATA (TATA-box binding protein binding motif), Inr (Initiation element), MTE (motif ten element), DPE (downstream promoter element) and DCE (downstream core element). Transcription initiation site is shown by the black arrow at +1 bp position. DCE is shown on a different construct for illustration purposes only, although this element can occur together with BRE, TATA and Inr elements, it presumably does not occur together with MTE and DPE. The figure is reproduced from reference (Maston et al., 2006).	6
Figure 1.4 Positions of transcription factor binding sites on SV40 enhancer. This enhancer can drive the expression of promoters in a orientation and position independent manner.....	9
Figure 1.5 Structure of an enhanceosome where an architectural protein such as HMGI(Y), bends DNA and allow cooperative binding of further transcription factors to the enhancer and enables their contacts with the promoter bound complexes. This figure is adapted from reference (Merika et al., 1998).....	11
Figure 1.6 Two possible mode of action of an insulator (a) it can block the communication between an enhancer and an active locus or (b) it prevents the spread of condensed chromatin structure onto actively transcribed regions.	13
Figure 1.7 Relative positions of the 12 subunits of the RNA polymerase II transcriptional machinery and DNA. The straight lines map interactions between corresponding subunits. Not all subunits can be visualised on this view. This display is reproduced from reference (Cramer et al., 2000)..	22
Figure 1.8 Side view of RNA (red) synthesis by RNA polymerase II machinery from a DNA template (template strand blue and coding strand is green). Cut surfaces of the protein, in the front, are lightly shaded and the remainder, at the back, are darkly shaded. By convention, the polymerase is moving on the DNA from left to right. The double stranded DNA is gripped by protein “jaws” where the upper jaw cannot be seen on this side view. The subunits named as “wall” in this figure blocks the straight passage of nucleic acids through the enzyme, therefore DNA:RNA hybrid makes almost a right angle with the axis of entering DNA. Importantly, this bend exposes the end of DNA:RNA hybrid for the addition of substrate nucleoside triphosphates (NTPs). NTPs could enter through funnel shaped opening at the bottom. Only nine base pair long DNA:RNA hybrid is allowed within the polymerase; a loop of proteins (rudder) mediates this by separating DNA from RNA. This figure is adapted from reference (Cramer et al., 2000).	23
Figure 1.9 Known contacts between TATA box binding protein associated factors (TAFs) and core promoter elements as explained in Figure 1.3. This figure is adapted from reference (Thomas and Chiang, 2006).	25
Figure 1.10 Three different TFIID complexes depending on the inclusion of TBP and TAF10.	26
Figure 1.11 Schematic representation of four common DNA binding domains in transcription factors. Abbreviations: HTH, helix-turn-helix; HLH, helix-loop-helix, Zn, zinc; Leu, leucine. This figure is reproduced from (Strachan and Read, 2003).	27
Figure 1.12. Jun and fos proteins both have leucine zipper motif to form a heterodimer (AP-1) to bind to their palindromic recognition sequences on DNA. This figure is reproduced from reference (Hess et al., 2004).	28
Figure 1.13 The helix-loop-helix motif carrying dimer protein bound to DNA. Different subunits are shown in white and yellow. This illustration is adapted from URL site reference URL1, 2004. ...	29
Figure 1.14 Helix-turn-helix (HTH) motif bound to its DNA on the left and a specialized form of HTH motif called homeodomain bound to its DNA. Homeodomain contains an extra α -helix presumably for further stabilization of DNA binding. These figures are adapted from URL site reference URL2, 2004.	31
Figure 1.15. Structure of DNA binding domain of CTCF composed of 11 adjacent zinc finger domain. This figure is adapted from (Ohlsson et al., 2001).	32

Figure 1.16. Structural organisation of core histones in the nucleosome core particle. This figure is adapted from reference (Alberts et al., 2001). 36

Figure 1.17. (B) About 160 base pairs of DNA encircle each histone core particle, nucleosome, and about 40 base pairs of DNA link the nucleosomes together. (C) Model for the arrangement of nucleosomes in the highly compacted solenoidal chromatin structure. This figure is adapted from reference (Turner, 2001). 37

Figure 1.18. Higher order of DNA packaging in nucleus. This figure is reproduced from reference (Strachan and Read, 2003). 38

Figure 1.19. Continuous wavelet transform heat map of chromatin accessibility across a 1.7-Mb segment of chromosome 21 containing the Down Syndrome critical region²⁹ (x axis, genomic position; y axis, wavelet scale), genes are shown below heat map. Four broad classes of chromatin domains are thus distinguished based on TSS density and chromatin activity: I, TSS-poor, inactive chromatin; II, TSS-rich, DNase I hypersensitive site-rich active chromatin; III, TSS-rich, inactive chromatin; IV, TSS-poor, DNase I hypersensitive site-rich active chromatin. This figure is adapted from reference (Sabo et al., 2006). 39

Figure 1.20. (A) The N-terminal tails of the core histones (e.g., H3) are modified by the addition of acetyl groups (Ac) to the side chains of specific lysine residues. (B) Transcriptional activators and repressors are associated with coactivators and corepressors, which have histone acetyltransferase (HAT) and histone deacetylase (HDAC) activities, respectively. This figure is adapted from reference (Cooper, 2002). 44

Figure 1.21 The chemistry of methylation on lysine residues of histones. Adomet (S-adenosyl-L-methionine or SAM) is a cofactor which carries the methyl group to be transferred. This figure is reproduced from reference (Shilatifard, 2006). 46

Figure 1.22. Cytoband view of chromosome 20 on the right and the gene number histogram along the chromosome on the left. 52

Figure 2.1 General Scheme of Cloning Procedure 56

Figure 2.2 Map of pDrive TA cloning vector 58

Figure 2.3 pGL3-basic vector map 62

Figure 2.4 pGL3-enhancer vector map 62

Figure 2.5 pGL3-promoter vector map 62

Figure 2.6 pGL3-control vector map 63

Figure 2.7 PRL-SV40 vector map 63

Figure 2.8 Schematic Representation of eukaryotic RNA labelling assay for expression profiling using GeneChip™ expression arrays 71

Figure 3.1 Schematic representation of the method for choosing 177 transcripts using different promoters. 90

Figure 3.2. Sequence signals utilized by Eponine. Arrow head on the top marks the true TSS (taken from Eukaryotic Promoter Database (EDP) (Cavin Perier et al., 1998)). This figure is reproduced from reference Down and Hubbard, 2002. 94

Figure 3.3. Number of promoters associated with predictions. There are 88 promoters not associated with any prediction. 95

Figure 3.4 Schematic representation of calculating frequencies of each nucleotide at a given position (r) in N number of promoter sequences. The number of each nucleotide is counted at a given position (r) along the sequences (denoted by red box) to obtain the frequency of the nucleotide at that position. 96

Figure 3.5. Frequency plots of each nucleotide relative to their distance to TSS in 177 promoters. Frequencies are averaged out in windows of ten nucleotides to eliminate local fluctuations. Solid black line at x=100 marks the TSS. 97

Figure 3.6. Summed nucleotide frequencies in promoters with, without CpG islands and 201 negative controls. Solid black line at x=100 marks the TSS. Frequencies are averaged out in windows of ten nucleotides to eliminate local fluctuations. 98

Figure 3.7 Frequency plots of each nucleotide relative to their distance to the 3' end of the 177 promoters. Frequencies are averaged out in windows of ten nucleotides to eliminate local fluctuations. Solid black line at x=75 3' end of the transcript. 99

Figure 3.8 Summed nucleotide frequencies of 77 promoters associated with at least one prediction (dotted lines) and 100 promoters not associated with any prediction. Frequencies are averaged out in windows of ten nucleotides to eliminate local fluctuations. 100

Figure 3.9. Summed nucleotide frequencies of RTs not associated with any prediction or a CpG island and ATs not associated with any prediction or a CpG island. Frequencies are averaged out in windows of ten nucleotides to eliminate local fluctuations. 101

Figure 3.10 (C+G) plot of RTs and ATs not associated with any prediction are curve-fitted (grey curves) and subtracted from each other (blue curve). 102

Figure 3.11. Average bendability score of every trinucleotide along the sequence of 177 promoter sequences (grey line) and bendability scores averaged at every 12 nucleotide (10 trinucleotides) shown with error bars (red line). 105

Figure 3.12. Bendability scores of 201 negative control sequences (red line with error bars) versus 177 promoter sequences (green line with error bars) averaged out in every 12 nucleotides. 105

Figure 3.13. Bendability scores of 81 promoters associated with a CpG island (red line with error bars) and 86 promoters not associated with a CpG island (green line with error bars). Grey line indicates the bending profile of all the dataset. 106

Figure 3.14. Bendability scores of RTs (green line with error bars) and ATs (red line with error bars) not associated with any promoter or TSS prediction or a CpG island and grey line denotes the bendability of all promoters sequences not associated with any prediction or a CpG island. 107

Figure 3.15. Expression profile of 96 genes represented by probes on Affymetrix U133 Plus 2 expression array, where 29 genes are not expressed by neither of the cell lines. 109

Figure 4.1. Bioluminescence reactions catalysed by firefly and renilla luciferase. This figure is reproduced from Promega® Product Technical Manual No 0.40. 114

Figure 4.2. Linear ranges of firefly and renilla luciferases. The linear range of the firefly luciferase assay is seven orders of magnitude, providing detection sensitivity of 1 femtogram (approximately 10^{-20} mole) of experimental reporter enzyme. The renilla luciferase assay has a linear range of greater than five orders of magnitude and allows for the detection of approximately 30 femtograms (approximately 3×10^{-19} moles) of control reporter enzyme. This figure is reproduced from Promega® Product Technical Manual No 0.40. 115

Figure 4.3. Measurement of luciferase activities before and after the addition of Stop& Glo® Reagent which quench the activity of beetle luciferase and initiate the renilla luciferase reaction. Beetle luciferase luminescence was quenched by greater than 5 orders of magnitude. This figure is reproduced from Promega® Product Technical Manual No 0.40. 116

Figure 4.4. Comparison of activities between candidate promoter fragments of 300 and 600 bp in size, and 300 bp fragments cloned together with SV40 enhancer. 121

Figure 4.5 Results of transfection with GFP reporter gene. PEGFP-N1 is the positive control plasmid which carries strong CMV promoter. Fragments that did not give significant activity over background (nctrl2) were shown in red. 122

Figure 4.6 Background subtracted Luciferase/Renilla signals of 74 candidate promoters in HeLa S3 cell line. Representative transcripts (RTs) are shown with columns with crossed pattern and alternative transcripts (ATs) are shown with unfilled columns. 126

Figure 4.7 Confidence levels to accept (green bars) or reject (red bars) a fragment as promoter in HeLa S3. Here, fragments which shows $3 \cdot \sigma$ higher than the background are accepted as a promoter and the confidence level for this decision is 99.7% assuming the error distribution follows Gaussian. Dotted lines show lower confidence levels. Note that rejection rates smaller than -10 is not shown on the plot. 127

Figure 4.8 Background subtracted Luciferase/Renilla signals of the 74 candidate promoters in NTERA-D1 cell line. Representative transcripts (RTs) are shown with columns with crossed pattern and alternative transcripts (ATs) are shown with unfilled columns. 128

Figure 4.9 Confidence levels to accept (green bars) or reject (red bars) a fragment as promoter in NTERA-D1 cell line. Here, fragments which shows $3 \cdot \sigma$ higher than the background are accepted

as a promoter and the confidence level for this decision is 99.7% assuming the error distribution follows Gaussian. Dotted lines show lower confidence levels. 129

Figure 4.10 Background subtracted activity of 76 candidate promoter fragments in synergy with SV40 Enhancer in HeLa S3 cells. The blue bars represents the promoter activities of alternative transcripts and the bars with check pattern are the fragments which showed activity only in synergic with the enhancer. 132

Figure 4.11 Confidence levels to accept (green bars) or reject (red bars) the activation response in HeLa S3 cell line. Here, fragments which shows $3*\sigma$ higher than the background are accepted as activated by the enhancer and the confidence level for this decision is 99.7% assuming the error distribution follows Gaussian. Bars with squared patterns denote the promoters that gave activity only in the presence of the enhancer..... 133

Figure 4.12 Background subtracted activity of 76 candidate promoters in synergy with SV40 enhancer in NTERA-D1 cells. The blue bars represents the promoter activities of alternative transcripts and the bars with checked pattern are the fragments which showed activity only in synergy with the enhancer..... 135

Figure 4.13 Confidence intervals to accept (green bars) or reject (red bars) the activation response in NTERA-D1 cells. Here, fragments which shows $3*\sigma$ higher than the background are accepted as activated by the enhancer and the confidence level for this decision is 99.7% assuming the error distribution follows Gaussian. Bars with squared patterns denote the promoters that gave activity only in the presence of the enhancer..... 136

Figure 4.14. Frequency distribution of putative Sp1 binding sites on promoters responding or not responding to SV40 Enhancer. 138

Figure 4.15 Scaled activities of promoters active only in synergy with SV40 enhancer in HeLa S3 and/or NTERA-D1 cells. Scaling was performed by dividing the activity of each promoter to the highest activity within all constructs. Constructs with (*) are recovered by the enhancer in both cell lines..... 140

Figure 4.16 Number of putative YY1 binding sites plotted against their position relative to the TSS at 0 bp..... 142

Figure 4.17 Promoter activities of 74 putative promoter fragments in HeLa S3 and NTERA-D1 cell lines. Inactive promoters are shown in red and active promoters were shown in blue. The annotation is taken from UCSC Genome Browser. 147

Figure 4.18. Promoter activities of 71 putative promoter fragments in synergy with SV40 enhancer in HeLa S3 and NTERA-D1 cell lines. Inactive promoters are shown in red and active promoters were shown in blue. The annotation is taken from UCSC Genome Browser..... 148

Figure 5.1 Chemical structure of formaldehyde. 153

Figure 5.2. Crosslinking of Cytosine to a Lysine by formaldehyde. This figure is reproduced from reference (Orlando et al., 1997). 154

Figure 5.3. Schematic description of Chromatin Immunoprecipitation coupled with DNA microarrays 155

Figure 5.4 Distribution of the gaps in 3.5 Mb tilepath array. The 25% and 75% percentile of the gaps is 157 and 799 bp respectively with a median of 420 bp. 160

Figure 5.5 The graphs above show raw Cy5 (input chromatin, horizontal axis) relative to the raw Cy3 (antibody, vertical axis) signals for Rabbit IgG (A) and tri-methylated K4 of Histone H3 (B) in NTERA-D1 cells. In graph A, there are no spots on the array, which produce high Cy3 to Cy5 signals, whereas in the graph B, a number of spots (red circle) have high Cy3 to Cy5 signals and are potential biological targets of the protein of interest. 163

Figure 5.6 The background subtracted mean signals and standard errors obtained from ChIP-chip using antibody recognising H3K4me3 in NTERA-D1 cells. The horizontal axis corresponds to the spots, and are ordered according to the genomic coordinates of the sequences they carry..... 165

Figure 5.7 Amplification levels of the promoter region of *C20orf121* by real-time PCR using ChIP material obtained by using antibodies listed above and the input chromatin in NTERA-D1 cells. The table below shows the enrichment level of the spot carrying the TSS of the gene and the upstream neighbouring spot in NTERA-D1 cells. The spot coordinates are given according to the TSS. 167

- Figure 5.8 Amplification levels of the promoter region of *C20orf121* by real-time PCR using ChIP material obtained by using antibodies listed above and the input chromatin in HeLa S3 cells. The table below shows the enrichment level of the spot carrying the TSS of the gene and the upstream neighbouring spot in HeLa S3 cells. The spot coordinates are given according to the TSS..... 168
- Figure 5.9 Amplification levels of the promoter region of *C20orf121* by real-time PCR using ChIP material obtained by using antibodies recognising Rabbit IgG and PolII, and the input chromatin in HeLa S3 and NTERA-D1 cells. The table below shows the enrichment level of the spot carrying the TSS of the gene and the upstream neighbouring spot. The spot coordinates are given according to the TSS..... 169
- Figure 5.10 Amplification levels of the promoter region of *C20orf121* by real-time PCR using ChIP material obtained by using antibodies recognising Rabbit IgG and H4Ac and the input chromatin in HeLa S3 and NTERA-D1 cells. The table below shows the enrichment level of the spot carrying the TSS of the gene and the upstream neighbouring spot. The spot coordinates are given according to the TSS..... 170
- Figure 5.11 Amplification levels of the promoter region of *C20orf121* by real-time PCR using ChIP material obtained by using antibodies recognising Rabbit IgG and H3K9me2 and the input chromatin in HeLa S3 and NTERA-D1 cells. The table below shows the enrichment level of the spot carrying the TSS of the gene and the upstream neighbouring spot. The spot coordinates are given according to the TSS..... 171
- Figure 5.12 Amplification levels of an H3K27me3 enriched spot in NTERA-D1 cells by real-time PCR using ChIP material obtained by using antibody recognising Rabbit IgG and H3K27me3 and the input chromatin in HeLa S3 and NTERA-D1 cells. The table below shows the enrichment level of the spot in both cell lines. 172
- Figure 5.13 Amplification levels of an CTCF enriched spot in HeLa S3 and NTERA-D1 cells by real-time PCR using ChIP material obtained by using antibody recognising Goat IgG and CTCF and the input chromatin in HeLa S3 and NTERA-D1 cells. The table below shows the enrichment level of the spot and its upstream neighbouring spot in both cell lines. 173
- Figure 5.14 The H3K4me3 signals in NTERA-D1 cells. Spots that showed signals between 1 and 5 were shown here for the ease of visualisation. Vertical axis denotes the number of antibodies that a corresponding spot showed “enriched” signal. “tss” denotes that these spots either carry or neighbours a TSS..... 175
- Figure 5.15 Heat map which displays spot intensities within ± 6 kb distance of 16 coding transcripts in HeLa S3 cells. Highest signal is shown as red while the lowest signal is denoted as white. Representative transcripts are labelled with gene name only. 177
- Figure 5.16 Heat map which displays spot intensities within ± 6 kb distance of 17 coding transcripts in NTERA-D1 cells. Highest signal is shown as red while the lowest signal is shown as white. Note that representative transcripts were denoted by gene name only..... 180
- Figure 5.17. Enrichment levels (only signals above threshold were shown) on PRKCBP1 gene shown together with the annotation tracks reproduced from Ensembl Genome Browser. Track information is displayed on the left hand side of the annotation window. 181
- Figure 5.18 The annotation of the an polII enriched region (43,354,134..43,355,325 bp on chromosome 20) reproduced from UCSC Genome Browser. The red squared denotes the boundaries of the enriched spot (3.8 fold polII and 40 fold H3K4me3 enrichments). Track information is displayed on the left hand side of the annotation display. 183
- Figure 5.19 The heat map displaying ~ 6 kb upstream and downstream of the annotated start sites of 22 genes that showed an H3K4me3 enrichment around their TSS..... 186
- Figure 5.20 Heat map displaying the signals around 6 kb upstream and downstream of the annotated start sites of 17 H3Ac enriched transcripts in HeLa S3 cells. Higher signal is indicated with colours towards red while lower signals would be towards white..... 190
- Figure 5.21 H3K4me3, H3K4me2 and H3K4me profiles of 22 genes that showed H3K4me3 enrichments on their start sites in HeLa S3 cells. Each profile is presented as a heat map which displays the signals obtained within ~ 6 kb distance of TSSs. The first column on the left lists the polII enrichments and the sec column lists the expression profiles (“P” stands for the gene is expressed and “A” denotes no expression) for the corresponding genes. “d” and “u” means that the signal is detected at ~ 2 kb downstream or upstream of TSS respectively. 193

Figure 5.22. Ensembl annotation of the spot (its genomic coordinates 43,423,952..43,426,324 bp) that showed a 8.8 fold H3K4me3 and 5.3 fold H3K4me2 enrichments. It carries the start site of four non-coding transcripts of *DBNDD2 (C20orf35)*. It also carries the first exon of 5 coding transcripts of the *C20orf169* whose record was removed from the Entrez Gene Database. 195

Figure 5.23 Heat map displaying the H3K4me3 enrichment levels of spots spanning 6 kb upstream and downstream sequences of 30 enriched TSSs of representative transcripts in NTERA-D1 cells. . 197

Figure 5.24 H3K4me3, polII and H3Ac peaks within the first 10 kb of SDC4 gene. The annotation is reproduced from UCSD Genome Browser. 200

Figure 5.25 Heat map displaying the signals around 6 kb upstream and downstream of the annotated start sites of 21 H3Ac enriched transcripts in NTERA-D1 cells. Higher signal is indicated with colours towards red while lower signals would be towards white..... 201

Figure 5.26 H3K4me3, H3K4me2 and H3K4me profiles of 23 genes that showed H3K4me3 enrichments on their start sites in NTERA-D1 cells. Each profile is presented as a heat map which displays the signals obtained within ~6 kb distance of TSSs. The first column on the left lists the polII enrichments and the sec column lists the expression profiles (“P” stands for the gene is expressed and “A” denotes no expression) for the corresponding genes. “d” means that the signal is detected at ~2 kb downstream of TSS..... 203

Figure 5.27 Plots of number of occurrences of all possible modified histone combinations in HeLa S3 and NTERA-D1 cells..... 206

Figure 5.28 Annotation taken from UCSD Genome Browser for the region between 42,749,148 and 42,766,245 bp together with enrichment levels with proteins RNA polymerase II (polII), H3K4me, H3K4me2, H3Ac and H4Ac in HeLa S3. Also the thick blue and green lines displays enrichment levels with H3K4me and H3K4me2 in NTERA-D1 cells. 208

Figure 5.29 Transcription factor binding profile of the region spanning from 42,754,161 to 42,755,685 bp that gave a 5.5 fold enrichment with RNA polymerase II. This figure is the reproduced from the graphical output of program MAPPER used to search putative binding sites. 209

Figure 5.30 The alignment of the human DNA sequence spanning between 42,760,432 and 42,760,632 bp coordinates to mouse, rat, dog, elephant and opossum sequences. This alignment is reproduced from UCSD Genome Browser. This conserved region is within the spot that gave enrichments with H3K4me, H3K4me2, H3Ac and H4Ac modified histones..... 210

Figure 5.31 The alignments across multi-species of the regions spanning from 45,637,267 to 45,637,440 bp on the left and 45,638,816 to 45,638,995 bp on the right. The alignments were taken from UCSC Genome Browser. 212

Figure 5.32 The putative binding sites on the sequence spanning from 45,648,944 to 46,651,138 bp. SP28 was enriched with H3K4me, H3K4me2 and H3Ac proteins..... 214

Figure 5.33 Percentages of the spots enriched with possible combinations of modified histones that includes H3Ac. 216

Figure 5.34 Heat map displaying H3K27me3 signals of 6 kb distance of 79 TSSs in HeLa S3 cells. The small heat map on the left displays the H3K4me3 signals of 2 kb distances of the corresponding start sites. 221

Figure 5.35 Heat map displaying H3K27me3 signals of 6 kb distance of 79 TSSs in NTERA-D1 cells. The small heat map on the left displays the H3K4me3 signals of 2 kb distances of the corresponding start sites. 224

Figure 5.36 The genomic positions of 12 CTCF enriched regions (blue track) listed in Table 5.12 together with the H3K4me and H3K4me2 enriched regions in HeLa S3 (red track) and NTERA-D1 (green track) cells which are seen as potential distal regulatory elements. The region shown here covers the ~3.5 Mb region spanning from 42,274,163 to 45,850,636 bp. The gene annotations are taken from UCSC Genome Browser..... 233

Figure 5.37 A region that contains the intronic enhancer of house-keeping ADA gene (shown with the red arrow) and a possible insulator element (blue box in “Insulators” track) that can block the activity of this intronic enhancer on the tissue-specific promoter of WISP2 gene. ”Enhancers_H” and “Enhancers_N” track displays the regions that are enriched with H3K4me and H3K4me2 in HeLa S3 and NTERA-D1 cells as possible enhancer elements..... 233

Figure 5.38 An insulator trap reporter vector construct where the candidate insulator is placed between H19 promoter and SV40 enhancer and promoter-enhancer activity is monitored by toxin-A reporter gene. In order to discriminate a possible silencing activity of the candidate fragment, hygromycin gene is placed its downstream and the cells transfected with this construct is screened with hygromycin antibiotic. 234

Figure 5.39 The region spanning from 43,100,000 to 43,425,000 bp where there are five candidate insulators shown as blue boxes on the insulator track. There are two more tracks, displaying H3K4me and H3K4me2 enriched regions in HeLa S3 (Enhancers_H track) and NTERA-D1 (Enhancers_N track) as possible cis-acting regulatory elements. 235

Figure 5.40 CTCF regions that falls within intronic regions shown in the blue track. Red and green tracks display the H3K4me and H3K4me2 enriched regions in HeLa S3 and NTERA-D1 cells respectively. 237

Figure 5.41 The enrichment difference between Rabbit IgG (negative control) (top) and H4Ac (bottom) antibodies on a small section of the custom-made array in NTERA-D1 cells. The green enriched spot on the bottom array carries upstream sequence of *PKRCBP1*. 240

Figure 5.42. Enrichment profile of Sp1 on a subsection (the same section as in Figure 5.32) in NTERA-D1 cells. 241

Figure 5.43 Enrichment levels on SLC12A5 gene with antibodies recognizing H3K4me2, H3K4me3, H3Ac and H3K27me3 together with the annotation of the gene taken from Ensembl Genome Browser. 243

Figure 5.44. Enrichment profile across *HNF4A* in HeLa S3 and NTERA-D1 cells. The annotation is reproduced from UCSC genome browser. Green arrow denotes P2 promoter and pink arrow denotes the P1 promoter. The peaks are also displayed as custom annotation tracks (red and blue boxes). 245

LIST OF ABBREVIATIONS

PCR	Polymerase Chain Reaction
E. coli	Escherichia Coli
DD	Double distilled
EDTA	Ethylene diamine tetraacetic acid
TE	Tris – EDTA
LB	Luria Bertani
CTD	C-terminal domain
dATP	2'-deoxyadenosine 5'-triphosphate
dCTP	2'-deoxycytidine 5'-triphosphate
dGTP	2'-deoxyguanosine 5'-triphosphate
dTTP	2'-deoxythymidine 5'-triphosphate
DMEM	Dulbecco's Modified Eagle Medium
FBS	Foetal Bovine Serum
IVT	In Vitro Translation
CMV	Cytomegalovirus
SV40	Simian Virus 40
TSS	Transcription Start Site
CpG	A cytosine nucleotide immediately followed by a guanine nucleotide on DNA sequence
FirstEF	First Exon Finder

cDNA	complementary DNA
EST	Expressed Sequence Tag
ORF	Open Reading Frame
<i>D. melanogaster</i>	<i>Drosophila melanogaster</i>
OMIM	Online Mendelian Inheritance in Man Database
RT	Representative Transcript
AT	Alternative Transcript
K	lysine
Ac	Acetylated
Me	Methylated
H3K4me	mono-methylated lysine 4 of histone H3
H3K4me2	di-methylated lysine 4 of histone H3
H3K4me3	tri-methylated lysine 4 of histone H3
H3K9me2	di-methylated lysine 9 of histone H3
H3K27me3	tri-methylated lysine 27 of histone H3
H3Ac	Acetylated lysines 9 and 14 of histone H3
H4Ac	Acetylated lysines 5,8,12 and 16 of histone H4
PolII	RNA polymerase II
CCTF	CCCTC-binding protein
WTSI	Wellcome Trust Sanger Institute
FA	Formaldehyde