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

LIST OF TABLES

LIST OF FIGURES

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
D. melanogaster	Drosophila melanogaster
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

1 INTRODUCTION

The International Human Genome Sequencing Consortium reported a nearly finished human genome sequence in 2004 (IHGSC, 2004). This gold standard sequence has an error rate of only 1 per 100,000 bases, contains 2.85 billion nucleotides interrupted by only 341 gaps and covers ~99% of the euchromatic genome. Having a nearly complete genome sequence in our hands, the task ahead is its structural annotation. The current version of the human gene catalogue (Ensembl, NCBI build 36) contains 23,341 (21,206 known and 2,135 novel) protein-coding genes. The annotation of non-coding transcribed elements (currently 719 pseudogenes and 1,430 RNA genes) is constantly improving as more experimental data and computational tools become available. All these functional genomic elements encode the information to generate the molecular machinery that carries out biological processes in our bodies, yet they need to be orchestrated in both time and space by regulatory regions in the genome. However, the identification and annotation of these regulatory sequences is quite challenging as we do not possess enough information on their sequence and structure characteristics. In 2003, the ENCODE (Encyclopaedia of DNA Elements) project was launched which aims to identify and annotate all functional elements in the human genome (ENCODE, 2004). The project is to be implemented in three phases; pilot, technology development and production phase. The pilot phase started by selecting representative regions of 0.5-2 Mb totalling 30 Mb (1% of the human genome) to apply and assess a battery of experimental approaches available as well as to develop novel approaches. Half the ENCODE regions were selected manually in order to include well-characterized genes and/or other functional elements (such as α and ß-globin gene clusters), and the regions where a number of multi-species sequence data are available, such as the locus containing *CFTR* (cystic fibrosis transmembrane conductance regulator). Remaining targets were chosen at random by means of an algorithm that ensured that the complete set of targets represented the range of gene content and level of non-exonic conservation (relative to mouse) found in the human genome. The ENCODE project is currently applying technologies for large-scale identification of functional elements in the target regions, specifically genes, promoters, enhancers, repressors/silencers, exons, origins of replication, sites of replication termination, transcription factor binding sites, methylation sites, deoxyribonuclease I (DNase I) hypersensitive sites, chromatin modifications, and multi-species conserved sequences of yet unknown function (Figure 1.1). Such projects will provide a systematic understanding of functional genomic elements beyond coding regions.

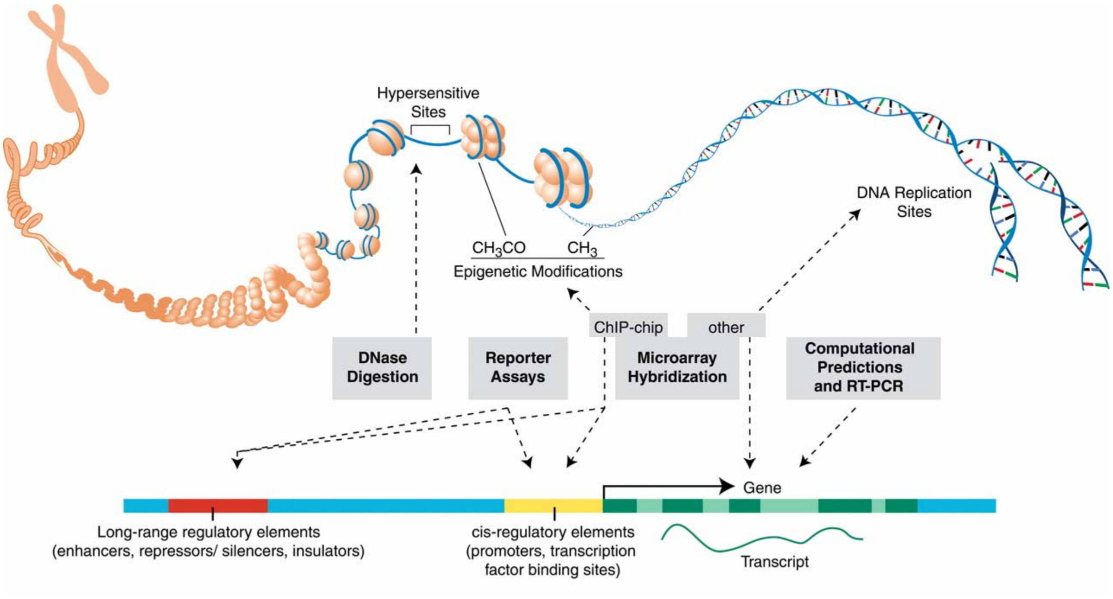

Hypersensitive Sites
CH_3CO
CH_3
Epigenetic Modifications
DNA Replication Sites
ChIP-chip
other
DNase Digestion
Reporter Assays
Microarray Hybridization
Computational Predictions and RT-PCR
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript

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.

1.1 Why is it important to find regulatory elements?

The C value paradox states that the genome size of an organism is not correlated with its biological complexity (Cavalier-Smith, 1978). As more genome sequences of diverse species become available, another paradox (N value paradox) emerged, gene number does not reveal biological complexity either (see Figure 1.2). This leaves us with the notion that there should be another scale on which biological complexity correlates with genomic data.

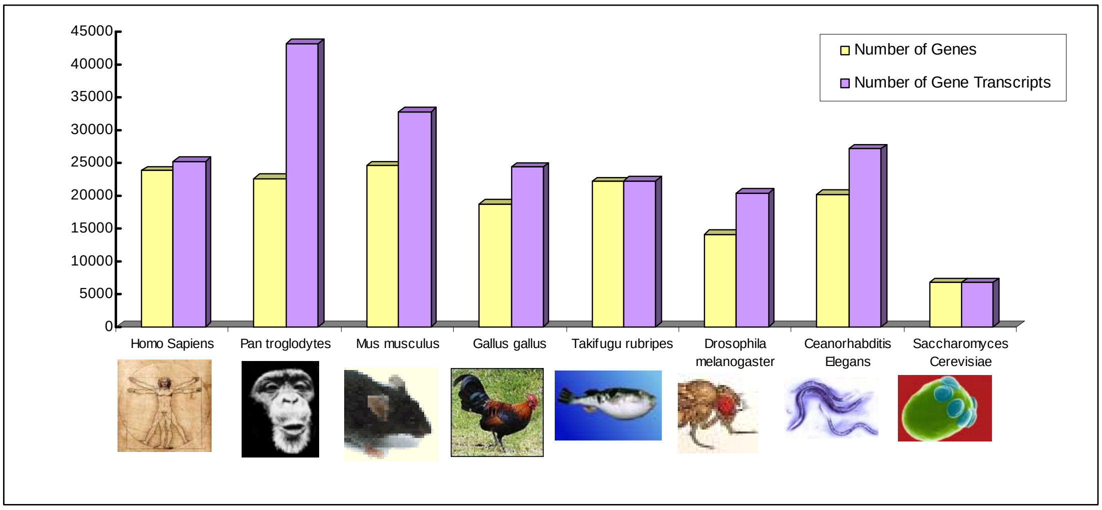


Figure 1.2 The number of genes and transcripts are taken from Ensembl version 39, NCBI build 36.

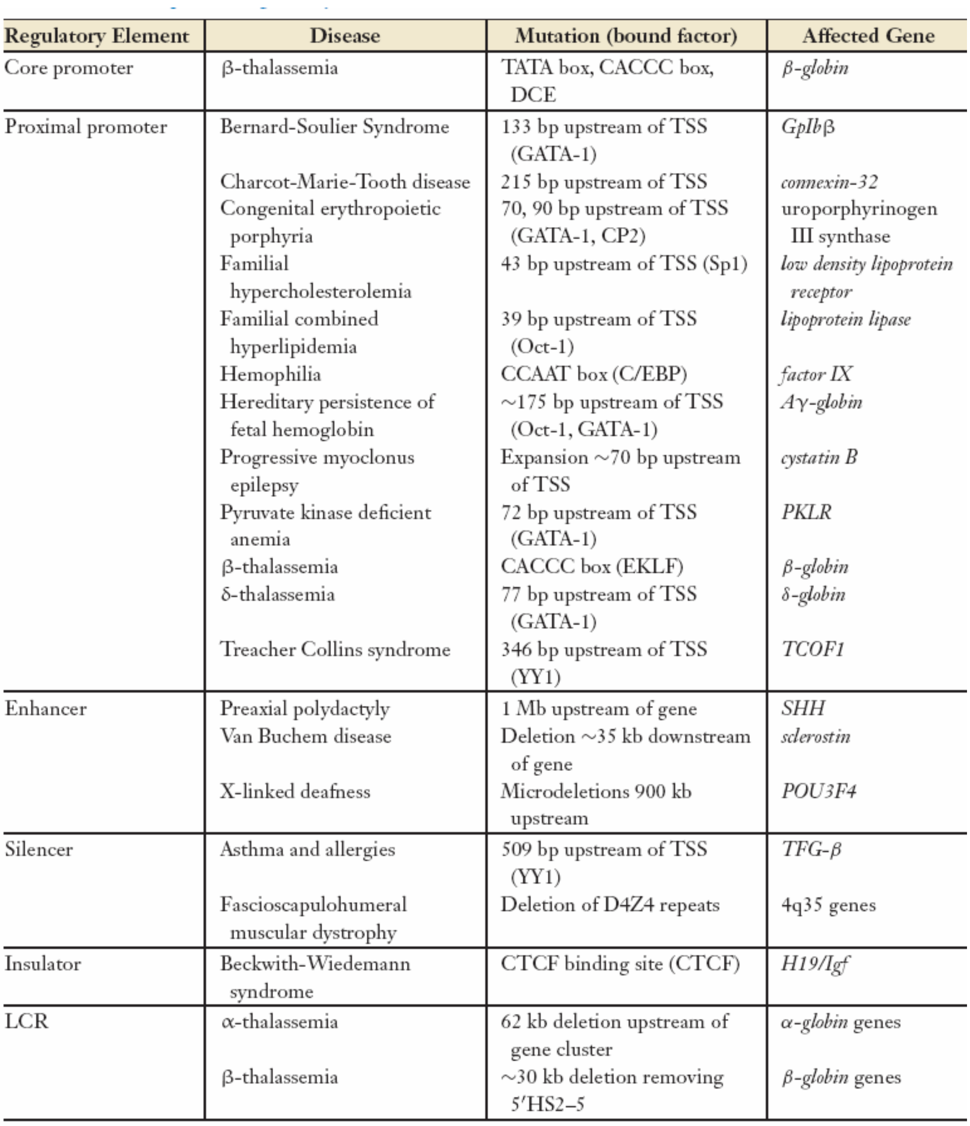
Regulatory networks may be the answer to the complexity paradox as their sophistication differs greatly between organisms (Kauffman, 1995). A better understanding of regulatory networks is also essential for the analysis of interactions between different cellular processes and/or genes.

Gene regulation lies at the heart of regulatory networks. It is a complex process, requiring a large number of proteins acting in a strictly co-operative manner on specific regions of the DNA, called regulatory regions. The best-known regulatory elements are promoter regions found at the 5’ UTR end of genes, whose transcriptional activity they control. There are a number of other regulatory elements such as enhancers, silencers or insulators - mostly located distant from the promoters – that they affect. These elements contain binding sites to recruit specific protein assemblies at the correct place in the genome. Proteins involved in transcriptional processes are called transcription factors. These proteins have the ability to bind at specific DNA sites, and activate or repress transcription processes through their interactions with the DNA and other factors. DNA is wrapped with packaging proteins called histones and the amino-terminal modifications of these histone proteins regulate the accessibility of the DNA to other proteins, hence playing a vital role in the gene regulation processes (Turner, 2001).

Numerous diseases have been associated with mutations in transcriptional regulatory elements, some of which are listed in Table 1.1. Mutations in regulatory DNA elements can result in reduced binding of a functional transcription factor on the site, such as in familiar hypercholesterolemia, where mutations in the proximal promoter of lipoprotein receptor gene cause reduced binding of the Sp1 transcription factor leading to incorrect regulation of the gene (Koivisto et al., 1994). Insertion or deletion mutations that change the spatial distribution of the regulatory elements in the genome are critical disease-causing factors, presumably by preventing the synergistic operation of the factors on those elements (Lalioti et al., 1999). Although there are relatively few diseases known to be caused by defects in regulatory elements, this could simply be due to our incomplete understanding of gene regulatory systems.

1.2 Regulatory DNA Elements

Every gene has a specific expression pattern; regulatory elements ensure that this pattern is achieved at the correct time and tissue. While many genes involved in basic cell functions are expressed constitutively, others are induced for example, only during cell differentiation or in response to a stimulus (Locker, 2001). Both constitutive and inducible gene expression is controlled by trans-acting proteins that are recruited on cis-acting regulatory DNA sequences.



Regulatory Element	Disease	Mutation (bound factor)	Affected Gene
Core promoter	β-thalassemia	TATA box, CACCC box, DCE	*β-globin*
Proximal promoter	Bernard-Soulier Syndrome	133 bp upstream of TSS (GATA-1)	*GpIbβ*
	Charcot-Marie-Tooth disease	215 bp upstream of TSS	*connexin-32*
	Congenital erythropoietic porphyria	70, 90 bp upstream of TSS (GATA-1, CP2)	uroporphyrinogen III synthase
	Familial hypercholesterolemia	43 bp upstream of TSS (Sp1)	*low density lipoprotein receptor*
	Familial combined hyperlipidemia	39 bp upstream of TSS (Oct-1)	*lipoprotein lipase*
	Hemophilia	CCAAT box (C/EBP)	*factor IX*
	Hereditary persistence of fetal hemoglobin	~175 bp upstream of TSS (Oct-1, GATA-1)	*Aγ-globin*
	Progressive myoclonus epilepsy	Expansion ~70 bp upstream of TSS	*cystatin B*
	Pyruvate kinase deficient anemia	72 bp upstream of TSS (GATA-1)	*PKLR*
	β-thalassemia	CACCC box (EKLF)	*β-globin*
	δ-thalassemia	77 bp upstream of TSS (GATA-1)	*δ-globin*
	Treacher Collins syndrome	346 bp upstream of TSS (YY1)	*TCOF1*
Enhancer	Preaxial polydactyly	1 Mb upstream of gene	*SHH*
	Van Buchem disease	Deletion ~35 kb downstream of gene	*sclerostin*
	X-linked deafness	Microdeletions 900 kb upstream	*POU3F4*
Silencer	Asthma and allergies	509 bp upstream of TSS (YY1)	*TFG-β*
	Fascioscapulohumeral muscular dystrophy	Deletion of D4Z4 repeats	4q35 genes
Insulator	Beckwith-Wiedemann syndrome	CTCF binding site (CTCF)	*H19/Igf*
LCR	α-thalassemia	62 kb deletion upstream of gene cluster	*α-globin* genes
	β-thalassemia	~30 kb deletion removing 5′HS2–5	*β-globin* genes

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).

1.2.1 Promoters

An essential part of the regulatory machinery of a gene is its promoter region found immediately upstream of the point where transcription starts. A promoter can be grouped into two regions; the core and the proximal promoter region. Typically, the core promoter is found within -40 to +40 nucleotides relative to transcription start site (TSS) where the basal transcription machinery is recruited (reviewed in Smale and Kadonaga, 2003). Core promoters contain several sequence motifs such as TATA-box, BRE (TFIIB-recognition element), DPE (downstream promoter element), DCE (downstream core element), Initiator element (Inr) and MTE (motif ten element) (Lim et al., 2004). The positional preferences along a typical core promoter is shown in Figure 1.3.

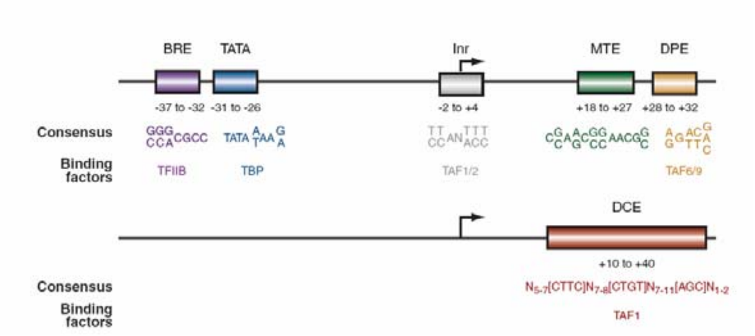


Figure 1.3 Sequence elements that are found on metazoan core promoters; BRE (TFIIB-recognition element), TATA (TATA-box binding protein binding motif), Inr (Initiation element), MTE (motif ten element), DPE (downstream promoter element) and DCE (downstream core element). Transcription initiation site 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).

Statistical analysis on circa 10,000 predicted core promoters has shown that these sequence elements are not as common as previously thought (Gershenzon and Ioshikhes, 2005). The initiator element is the most common one, found in nearly half of the promoters, whereas DPE and BRE are found in a quarter of them. Strikingly, the TATA box is only found in one in eight of the predicted core promoters. Other recent studies suggest that more general sequence features such as ATG deserts mark promoter regions (Lee, Howcroft et al., 2005). It was also shown that mammalian and plant core promoter sequences can be differentiated on the basis of their DNA structure (Florquin et al., 2005).

The assembly of the transcription initiation machinery on core promoters is partly regulated by the promoter-proximal region which is located upstream of the core promoter. This control region is a few hundred base pairs long and typically contains multiple recognition sites for transcription factors (TFs) that regulate the stability of the transcription machinery on the core promoter.

Human promoters can broadly be classified into two groups depending on the presence or not of CpG islands on their proximal promoter regions. CpG islands are long stretches of DNA sequences (between 500 bp to 2 kb in length) that have a high G+C nucleotide (GC) content and a high frequency of the CpG dinucleotide (a C nucleotide immediately followed by a G nucleotide). The current definition of a CpG island is that (i) it should be at least 500 bp long (ii) the GC content should be higher than 55% and (iii) the ratio of the observed number CpG dinucleotides to the expected number of CpG dinucleotides should be higher than 0.65 (Takai and Jones, 2002). This definition sets more stringent criteria than the original one by Frommer et al (Gardiner-Garden and Frommer, 1987) in order to exclude Alu repetitive elements which are short interspersed sequences with high GC content and frequency of CpG dinucleotides. CpG dinucleotides found in CpG islands on promoter sequences are normally not methylated, whereas elsewhere in the genome are typically methylated at the fifth carbon position of the cytosine base (Larsen et al., 1992). Other regions in mammalian genomes contain relatively fewer CpG dinucleotides since methylated cytosines are mutational hotspots (Coulondre et al., 1978) and replaced by TpG dinucleotides. Moreover, methylation of CpG dinucleotides on promoters is associated with gene silencing (Bird, 2002), genomic imprinting (Feil and Khosla, 1999), X-chromosome inactivation (Panning and Jaenisch, 1998), silencing of intra-genomic parasites (Yoder et al., 1997) and carcinogenesis (Baylin et al., 1998; Jones and Laird, 1999). Methylated CpG may block the binding of activating transcription factors to their recognition sequences due to steric hindrance of methyl groups (reviewed in Maston et al., 2006). Additionally, repressor proteins such as MeCp2 (methyl CpG binding protein 2) specifically binds to methylated CpG dinucleotides and recruit protein complexes that achieve a repressive chromatin environment (Jones et al., 1998).

CpG island are found in circa ~60% of human promoters and they are often used to locate promoters (Larsen et al., 1992). Most housekeeping genes as well as many tissue-specific genes are associated with CpG islands around their start sites (Gardiner-Garden and Frommer, 1987; Larsen et al., 1992). In a recent study, it was found that BREs are more common in promoters associated with CpG islands, whereas TATA boxes are more common in promoters that do not have a CpG island (Gershenzon and Ioshikhes, 2005; Maston et al., 2006).

1.2.2 Enhancers

The promoters' ability to drive expression may also be dependant on regulatory sequences called enhancers which can be located as far as hundreds kilo bases upstream or downstream of the promoter itself. Enhancers can also act across chromosomes (transvection), for example a recent study reported a single enhancer acting on several promoters independent of their chromosomal locations in mouse olfactory cells (Lomvardas et al., 2006). Enhancers contain multiple binding sites for activatory proteins and in this respect, they are rather similar to proximal promoter elements except their distant localization from the promoter. These elements are usually modular, such that a single promoter can be affected by a specific set of enhancer elements at different times or in different tissues or in response to a different stimuli (reviewed in Maston et al., 2006).

The first enhancer was identified in the simian virus 40 (SV40) genome by showing that it could markedly increase the transcription of a heterologous promoter (Banerji et al., 1981); named SV40 enhancer. It contains binding sites for common as well as tissue-specific activator proteins and can activate the transcription of a promoter under its control (Parsons et al., 2004; Song, 2004; Song, 2005), although it had a silencing effect in some cases depending on the tissue it operates in (Yamaguchi et al., 1989).

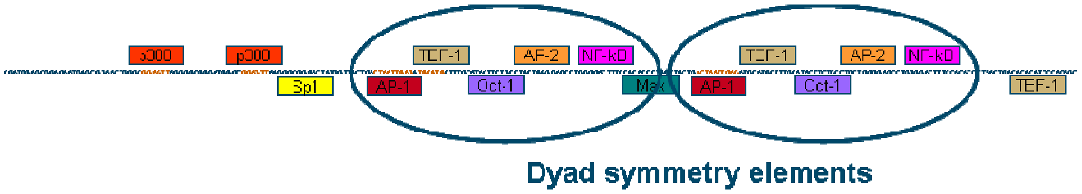


Figure 1.4 Positions of transcription factor binding sites on SV40 enhancer. This enhancer can drive the expression of promoters in a orientation and position independent manner.

The SV40 enhancer is 186 bp long and contains two 72 bp long repeats called dyad symmetry elements (Figure 1.4). Dyad symmetry elements contain binding sites for several activators such as activatory protein-1 (AP-1) and activatory protein-2 (AP-2), octamer binding factor-1 (Oct-1 or POU2F1), NF-kB (NFKB1, nuclear factor of kappa light polypeptide gene enhancer in B-cells) and transcription enhancer factor-1 (TEF-1). There are also binding sites for p300, a common enhancer binding factor and transcriptional activator with a histone acetyltransferease activity, and specificity protein-1 (Sp1), a common transcription factor, outside the dyad symmetry elements. No mutation on these binding sites totally abolishes the enhancer effect, which suggests a redundancy of functional information (Atchison, 1988).

How enhancers perform their actions over long distances in the genome is not well understood. DNA scanning is one of the proposed mechanisms where enhancer-promoter contact is achieved via enhancer-bound factors that move continuously along the DNA until they encounter their cognate promoter. However, this mechanism cannot explain the transvection, where promoter and enhancer reside in different chromosomes, and activation from a tailed hairpin that extends outwardly from a double-stranded circle (Plon and Wang, 1986). Another possible mechanism is called 'facilitated tracking' where an enhancer-bound complex tracks via small steps (perhaps scanning) along the chromatin until it encounters the cognate promoter, at which point a stable looped structure is formed (Blackwood and Kadonaga, 1998). However, recent studies are in favour of a third mechanism called "DNA looping" where the enhancer and the core promoter are brought into close proximity by looping out the intervening DNA (Tolhuis et al., 2002). DNA looping is consistent with long distance and orientation-independent transcriptional activation, the action of boundary elements and transvection. Also, a recent study, where it was shown that an enhancer bound factor is able to induce DNA looping, lends additional support to the above mechanism (Petrascheck et al., 2005).

Enhanceosomes

An enhanceosome is essentially a higher order three dimensional protein complex formed by architectural proteins that are able to bend DNA to allow specific contacts between enhancer and promoter bound proteins and other associated factors (Bazett-Jones et al., 1994) (Figure 1.5). Enhanceosomes typically contain an architectural protein that is not an activator by itself but it facilitates enhanceosome assembly by binding to several sites on the enhancer (Thanos and Maniatis, 1995) (Grosschedl, 1995). They differ from other regulatory complexes formed on regulatory elements since experimental studies have shown that the precise arrangement of the associated factors and their cooperative binding strictly determine the level of transcriptional activation (Carey, 1998).

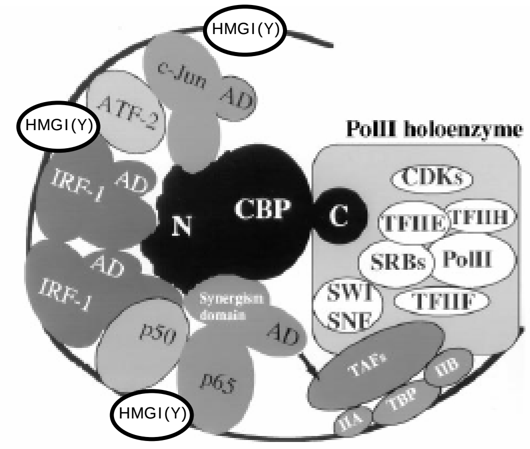


Figure 1.5 Structure of an enhanceosome where an architectural protein such as HMGI(Y), bends DNA and allow cooperative binding of further transcription factors to the enhancer and enables their contacts with the promoter bound complexes. This figure is adapted from reference (Merika et al., 1998).

1.2.3 Silencers

Silencers are regulatory sequences that reduce promoter activity, i.e. they have the opposite effect of an enhancer. Classic silencers operate in a position and orientation independent manner and are usually located within introns, intergenic regions or proximal promoter elements. They contain binding sites for repressor proteins, which in turn recruit co-repressors to inhibit transcription (Burke and Baniahmad, 2000). These co-repressors mediate silencing either by directly inhibiting the transcription machinery (Ayer et al., 1995; Hurlin et al., 1997) or recruiting chromatin modifying protein complexes to establish a repressive chromatin environment (Srinivasan and Atchison, 2004). Like enhancers, silencers act over long distances and thus the proposed mechanisms for enhancer action are also valid for silencers, for example a distant silencer is brought into close proximity of a promoter via DNA looping or another yet unknown mechanism requiring more complex tertiary DNA structures.

Some positional-dependent silencers have also been identified (Ogbourne and Antalis, 1998). They are usually found in proximal promoters, as well as introns, exons and various flanking sequences. They exert their effect by (i) physically inhibiting the interaction of transcription factors with their specific DNA binding sites, (ii) interfering with specific signals that control transcriptional events such as splicing and polyadenylation or (iii) affecting transcriptional elongation. The position-dependent silencer found in c-fos promoter is an example of such elements, where position-dependent silencer is bound by the nuclear factor ying yang-1 (YY1), which induces a specific bend in the promoter that blocks the interaction of an activator with the promoter (Natesan and Gilman, 1993). Silencers located in intron are especially interesting since they can physically repress the transcription by either presenting a binding site for a repressor that will halt the transcriptional elongation (Yuan, 2000), or preventing the recognition of intronic splice sites (Carstens et al., 2000).

There are also some orientation-dependent silencers known where they can only exert their effect when placed in a certain orientation, but their function has not yet been fully understood (Ogbourne and Antalis, 1998).

1.2.4 Insulators (Boundary Elements)

Insulators are DNA sequence elements that can protect genes from inappropriate signals emanating from their surrounding environment (West et al., 2002). Insulators mediate this protective function in two ways. The first way is to block the effect of a distal enhancer on a promoter. In this case, the insulator should be positioned between the promoter and the enhancer to exert its blocking function (Figure 1.6a). The second way is when insulators act as “walls” to prevent the spread of silenced chromatin into actively transcribed regions (Figure 1.6b). These two actions can be separable at least for some insulators which means that there are insulators which functions in only one of the ways described above (Recillas-Targa et al., 2002).

Typically, insulators are 0.5-3 kb in length, and they function in an orientation independent manner (reviewed in Maston et al., 2006).

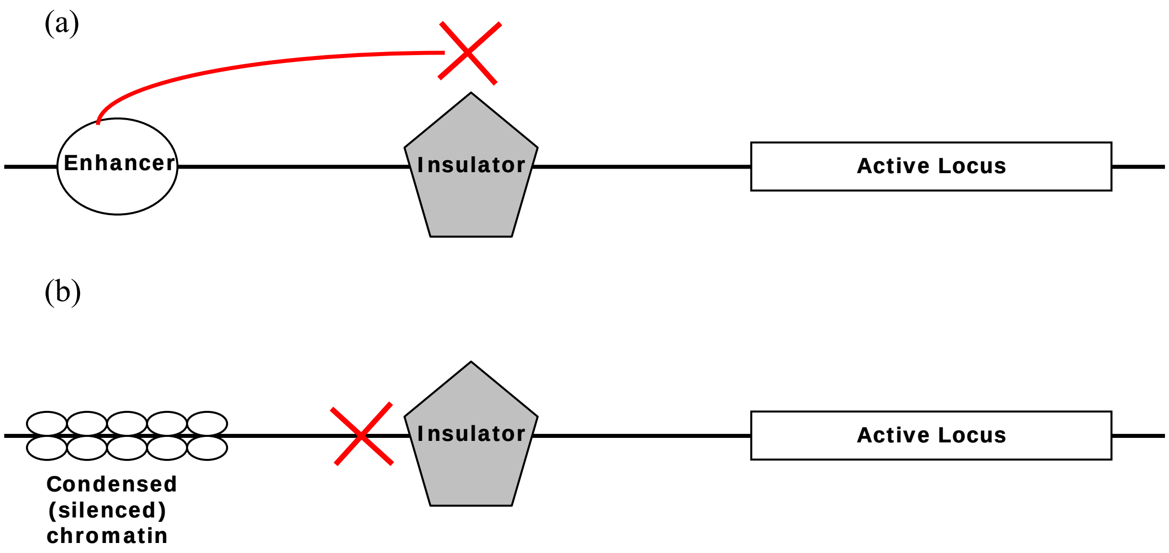


Figure 1.6 Two possible mode of action of an insulator (a) it can block the communication between an enhancer and an active locus or (b) it prevents the spread of condensed chromatin structure onto actively transcribed regions.

The first vertebrate insulator was found near the 5’ end of the chicken β-globin locus (Chung et al., 1993); a homologous insulator element resides in the human β-globin locus (Li et al., 1999). Another well-known insulator element is located upstream of the human H19 gene and controls the allele-specific expression of the H19 and Igf2 genes (Kaffer et al., 2000). This insulator is inactive as a result of hypermethylation in the paternally inherited locus, and hence not able to block the expression of the paternal Igf2 gene, whereas in the maternal locus, this insulator is not methylated and thus able to block the expression of the maternal Igf2 gene.

Insulators are dynamic elements in that they can alter their blocking functions depending on the proteins bound onto them (Bell et al., 2001). Currently, the CCCTC-binding protein (CTCF) is the only known factor mediating insulator activities in vertebrates. CTCF is an evolutionary conserved zinc finger protein that binds to its ~50 bp long target sites through combinatorial use of its 11 zinc finger motifs (Ohlsson et al., 2001). It functions in gene activation (Vostrov and Quitschke, 1997), and repression (Kuzmin et al., 2005) as well as chromatin insulation. CTCF is sufficient for the insulator activity in vertebrates (Bell et al., 1999) but its mechanism is not yet clear. Recent studies showed that CTCF co-localizes with the nuclear matrix binding proteins on insulator elements (Dunn et al., 2003; Yusufzai and Felsenfeld, 2004). This tethering of CTCF with nuclear matrix proteins might aid to generate separate chromatin loops which separate the promoter and enhancer chromatin environments hence blocking their communication (Dunn et al., 2003; Yusufzai and Felsenfeld, 2004).

An interesting property of insulators is that if two insulator elements are positioned between a promoter and an enhancer, the insulator activity is abolished (Cai and Shen, 2001; Muravyova et al., 2001). This type of neutralization favours the chromatin loop model, however the type of structure that could mediate such neutralisation is not obvious. This may suggest that insulator action involves rather more complex steps than those in the simple loop domain models (Bell et al., 2001).

1.2.5 Locus Control Regions

Locus control regions (LCRs) are complex enhancer assemblies controlling a set of physically linked genes in an orientation and position independent manner. Their main difference of classic distal regulatory elements came from transgenic mice experiments, where LCRs were shown to exert an enhancing effect on their target promoters independently of the insertion point within the host genome (Grosveld et al., 1987). LCRs contain regions that are preferentially digested by DNase I, called DNase I hypersensitive (HS) regions (Weintraub and Groudine, 1976). These HS regions probe for active chromatin domains and often exhibit a tissue specific pattern. The β-globin gene loci in chicken, mouse and human are controlled by LCRs located in the upstream and downstream sequence. The question that remains unanswered is whether LCRs have the ability to open chromatin. Experiments designed to answer this question offered different answers in different organisms. In mouse, when β-globin LCRs are deleted, there is no major change in DNase I hypersensitivity and chromatin architecture, but gene expression is reduced to 1-4% of that of the wild-type, suggesting that LCRs do not have a major effect on chromatin environment but have a major enhancer activity (Bender et al., 2000). However, when the human β-globin LCRs were deleted, transcription was halted and the chromatin environment was altered, suggesting that LCRs are important in recruiting protein assemblies to create an open chromatin environment (Schubeler et al., 2000; Dean, 2006). These differences may just reflect that mouse and human loci are regulated in different fashions.

One interesting structural property of LCRs is that they are able to form chromatin loops, presumably to bring target promoters and the appropriate enhancers into close proximity. This looping is verified using two different experimental approaches. The first one uses fluorescence in situ hybridization (FISH) and real-time PCR to locate regions in close proximity to each other *in vivo*, and application of this approach to the murine β-globin locus showed that HS sites are physically interacting with each other *in vivo* (Carter et al., 2002). The second one uses an elegant method called chromatin conformation capture (3C). In 3C, crosslinked DNA-protein material from intact cells is digested by restriction enzymes and then ligated at very low DNA concentrations, which favours intramolecular ligations rather than random intermolecular ones (Dekker et al., 2002). This way, DNA fragments located far away in the genome but which are spatially close to each other *in vivo* will be crosslinked together and can then be detected via quantitative PCR with locus specific primers. Application of this method to the murine locus again showed that HS are in contact with each other (Tolhuis et al., 2002). These experiments could also support the "DNA looping" model for promoter and distal regulatory elements interactions (see 1.2.2).

1.2.6 Experimental and Computational Efforts for Locating Regulatory Sequences

Locating promoter elements is a relatively uncomplicated task in experimental terms as they are located at the 5' end of genes. A powerful approach to locate promoters utilizes an oligo-capping approach to select only for full-length cDNAs that are then aligned to the human genome to locate TSSs which ultimately mark promoter regions (Suzuki and Sugano, 2001). TSSs found using this approach are collected in a database called DBTSS (Database of Human Transcription Start Sites) and this database currently contains TSS information for 8,308 human genes and 4,276 mouse genes. Gene reporter assay is another popular method for both verifying and assessing promoter activity in different tissues *in vivo* (see Chapter 3). The activity of 921 predicted promoter sequences (by aligning full-length cDNAs onto human genome) in the ENCODE regions were assessed using gene reporter assays in 16 cell lines (Cooper et al., 2006). In this study, 42% of the promoters showed activity in at least one cell line. Such studies verify promoter sequences in the human genome but defining a promoter region requires additional experiments to exclude regions not affecting transcriptional activity. Even then, in order to gain a complete picture of the promoter, such experiments should be performed in each tissue.

The combinatorial usage of regulatory elements in the genome to achieve a timely and spatially correct regulation poses the main difficulty in understanding regulatory networks. Genome-wide studies, generating promoter activity data across different tissues, aim at classifying promoters in broad groups with the help of chromatin structure and gene expression information (Kim et al., 2005). Yet, each promoter tells a different story when combined with its distal regulatory elements and trans-acting factors, that is not captured by reporter assays. Therefore, although locating promoter elements is straightforward once we have a complete list of gene structures, it is still a daunting task to functionally characterise a promoter. This is where computational approaches could be of use, since they are able to find information-rich segments (e.g. binding site for a transcription factor) within a promoter. Unfortunately, this is currently not achievable as the sequence information used by proteins to bind to the promoter elements is extremely fuzzy to us. Therefore, such computational efforts predict too many potential sites whereas only a small fraction is biologically relevant.

Since we neither hold a complete catalogue of protein coding genes and non-coding transcripts operating in different tissues nor have their complete structures, we cannot locate all the promoters in our genome. Promoter prediction *in silico* is therefore one way to locate promoter regions in a given sequence. Computational programs use sequence features known to be enriched in promoters such as CpG islands (Ioshikhes and Zhang, 2000; Davuluri et al., 2001; Hannenhalli and Levy, 2001; Ponger and Mouchiroud, 2002; Bajic and Seah, 2003), short promoter sequence motifs (see 1.2.1) or increased frequency of putative transcription factor binding sites (Fickett and Hatzigeorgiou, 1997; Prestridge, 2000; Down and Hubbard, 2002). There are also programs that utilize current gene annotation (Ohler et al., 2000; Down and Hubbard, 2002; Solovyev and Shahmuradov, 2003), statistical analysis of nucleotide distribution in promoters (Knudsen, 1999; Davuluri et al., 2001; Reese, 2001; Down and Hubbard, 2002; Bajic and Seah, 2003; Bajic and Seah, 2003; Bajic et al., 2003) and homology with orthologous promoters (Solovyev and Shahmuradov, 2003). A recent study by Bajic et al. employed several of these computational promoter prediction programs in genome-wide scale comparison (Bajic et al., 2004). Most programs failed in predicting promoters on such scale, while programs such as FirstEF, Eponine and CpGproD performed relatively better. Promoters associated with CpG islands were fairly easy to predict; most programs could predict ~40% of those promoters by making one false prediction for every two true predictions (Down and Hubbard, 2002; Ponger and Mouchiroud, 2002). Essentially no program was able to predict promoters not associated with CpG islands satisfactorily. For instance, a commonly used promoter prediction program FirstEF (Davuluri et al., 2001) predicted only 4% of such promoters while making 16 false predictions for every true prediction. These results demonstrate that there is an urgent need for other criteria that will enable us to identify promoter sequences in a more efficient way. One recent study utilized secondary structure of promoter elements and obtained promising results (Florquin et al., 2005). In this study, a dataset that contained 25% promoter sequences and 75% non-promoter sequences was assessed with a number of structural models. Structural methods such as DNA bendability, nucleosome position preference, DNA denaturation, DNA deformation by protein binding etc. discriminate 75-82% of the promoter sequences from non-promoters. Such findings stress the point that we need to investigate promoter sequences in higher dimensions rather than on primary sequence level where only the adjacency of bases is accounted for.

When it comes to locating distal regulatory elements, not only is there a handful of well-characterized such elements but the space to search is vast. There are three main difficulties in locating and characterizing such elements. Firstly, we do not possess enough information about the genomic environment of these elements or their protein interaction profile. Recently, it was found that LCRs are open chromatin regions (Elefant et al., 2000; Schubeler et al., 2001; Fu et al., 2002), but unfortunately this finding cannot be extended to other distal regulatory elements. Chromatin immunoprecipitation (see chapter 5) is an approach with great potential in characterising distal elements as long as we know the proteins that are bound to these elements or their histone codes. Moreover, once we know enough proteins or modified histones related to these elements, chromatin immunoprecipitation experiments could aid us to classify distal enhancers. Secondly, distal elements are scattered around the genome and there is no positional information like the proximity of promoters to the 5' end of genes. Most likely, the location of these elements is not random and are positioned within a higher order chromatin and/or nuclear scaffold structure (Dunn et al., 2003; Yusufzai and Felsenfeld, 2004). Lastly, a distal element does not contain direct information for its target promoter, i. e. there is virtually no way to know solely from its sequence which promoter(s) a particular distal element affects. Experimental techniques such as chromatin conformation capture (Dekker et al., 2002) is very promising in finding distal elements and their target regions since the positional information of distal elements in the genome is conserved during experiment (see section 1.2.5).

The availability of genome sequences from multiple species enables us to identify regions that are under selective pressure in evolution. Coding regions are well conserved across many organisms, together with many regulatory elements although the latter often have lower conservation scores. Human mouse comparison reveals that nearly ~5% of the mammalian genome is under selection (Waterston et al., 2002). Since only ~1.5% of the human genome is coding (Lander et al., 2001; Venter et al., 2001), there should be additional functional elements (e.g. regulatory regions, non-protein coding genes, chromosomal structural elements) under selection. Comparison studies using a number of genomes detected many conserved blocks of non-coding regions across the human genome (Margulies et al., 2003). These non-coding conserved sequences are called conserved non-genic sequences (CNGs) and are defined as a region of at least 100 bp long and 70% identical in an ungapped alignment (Duret et al., 1993). CNGs do not share any sequence similarity between each other and they do not align to other regions in the genome either (Dermitzakis et al., 2003). CNGs generally populate non-genic regions and are uniformly distributed across intergenic regions (Dermitzakis et al., 2005). A subset of CNGs are called ultra conserved elements (UCEs), as they are conserved also in chicken and dog genomes with up to 95 to 99% identity (Bejerano et al., 2004). Some UCEs are even conserved in fish and there is evidence that these sequences may play a vital role in development (Woolfe et al., 2005). Many studies reported the regulatory nature of CNGs. A comparative analysis of 209 kb mouse BAC clone containing the *GDF6* and flanking regions, to homologous sequence data from 14 species revealed a 404 bp enhancer important in limb joints development (Portnoy et al., 2005). Another study showed four different single nucleotide substitutions with full penetrance on an intronic enhancer, which is also a ~750 bp long CNG, residing within *LMBR1* on the autosomal-dominant Preaxial Polydactyly disease locus on chromosome 7q36 (Lettice et al., 2003). There are several instances where a genomic rearrangement is responsible for a disease phenotype by presumably replacing or deleting the regulatory elements (de Kok et al., 1995; Wirth et al., 1996; Bishop et al., 2000). However, the functional nature of how such rearrangements may affect nearby CNGs has not yet been established. Nevertheless, two ultra-conserved elements with several kilobases of ultra-conserved sequences are located within the global control locus of *HoxD* cluster, of which its strictly controlled expression is required in limb development (Spitz et al., 2003).

These conserved sequences although non-coding are there for a reason of which we do not yet have a clear understanding. CNGs may be target sequences for matrix attachment proteins and play a vital role in keeping the chromosomal architecture in place (Spitz et al., 2003). However, this idea needs more support due to the fact that nucleosomal architecture is mostly tissue-specific (Nielsen et al., 2002). Another possibility is that CNGs play a role in developmental processes. Then again, a deletion study which removed two large non-coding segments, containing several

CNGs, from the mouse genome generated mice homozygous for these deletions that were all alive and did not show any aberrant developmental features or other disease phenotypes (Nobrega et al., 2004). However, it is important to note that none of these CNGs were conserved in fish and the ones that are conserved even in fish are those implicated in development.

1.3 Transcriptional Machinery

1.3.1 RNA Polymerase II Transcription Machinery

The core promoter region of a gene is the assembly point for the protein complexes to initiate transcription. The central dogma in biology dictates that, information flows from DNA to RNA to protein (Crick, 1958). Transcription is the step of copying sequence information in the DNA to RNA molecules. Transcription is achieved by a specific polymerase called DNA directed RNA polymerase, which is essentially a nucleotidyl transferase that polymerizes ribonucleotides at the 3' end of an RNA transcript from a DNA template (Weiss and Gladstone, 1959). In eukaryotes, there are three different RNA polymerases; RNA polymerase I is responsible for transcribing the genes for the 18S and 28S ribosomal RNA, RNA polymerase III is transcribing the transfer RNA genes and 5S ribosomal RNA, and RNA polymerase II (polII) is responsible for transcribing all other genes including small nuclear RNAs and micro RNA genes (Lee et al., 2004). PolII generates messenger RNAs (mRNAs). These three polymerases are highly conserved proteins, two large subunits of these holoenzymes are homologous in structure and function with the two subunits of the prokaryotic RNA polymerase (Tsonis, 2003). The polII machinery (basal transcription machinery) and its associated proteins are central to this study. PolII is composed of 12 subunits, each being encoded by a separate gene. However, additional subunits are required for polII to function correctly. The schematic representation of the 12 subunits of polII machinery is shown in Figure 1.7.

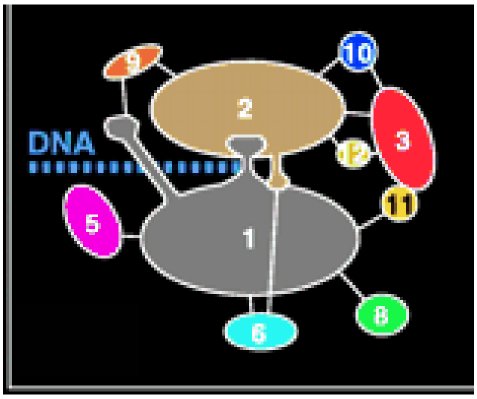


Figure 1.7 Relative positions of the 12 subunits of the RNA polymerase II transcriptional machinery and DNA. The straight lines map interactions between corresponding subunits. Not all subunits can be visualised on this view. This display is reproduced from reference (Cramer et al., 2000).

The largest subunit of polII (labelled as 1 in Figure 1.7) contains a carboxyl terminal domain composed of heptapeptide repeats that are essential for polymerase activity. These repeats contain serine and threonine residues that are phosphorylated in actively transcribing polII complexes (Komarnitsky et al., 2000). In addition, this subunit in combination with several other polymerase subunits, forms the DNA binding domain of the polymerase, a groove in which the DNA template is transcribed into RNA (Davis et al., 2002). The second largest subunit of the polII machinery ('2' in Figure 1.7) is responsible, in combination with at least two other polymerase subunits, to form a structure that maintains the DNA template and the newly synthesized RNA in contact with the active site (Cramer et al., 2000). The fifth largest subunit of the polII machinery ('5' in Figure 1.7) was shown to interact with a hepatitis virus transactivating protein suggesting that interactions between transcriptional activators and the polymerase could be mediated by this subunit (Cheong et al., 1995). A tertiary structure of the polII machinery is available at 2.8 Ǻ resolution (Cramer et al., 2001). The dynamic of the transcription initiation is shown in Figure 1.8.

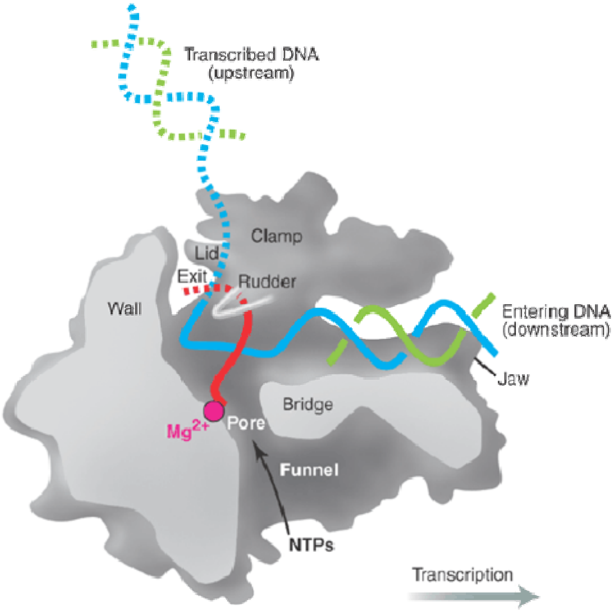
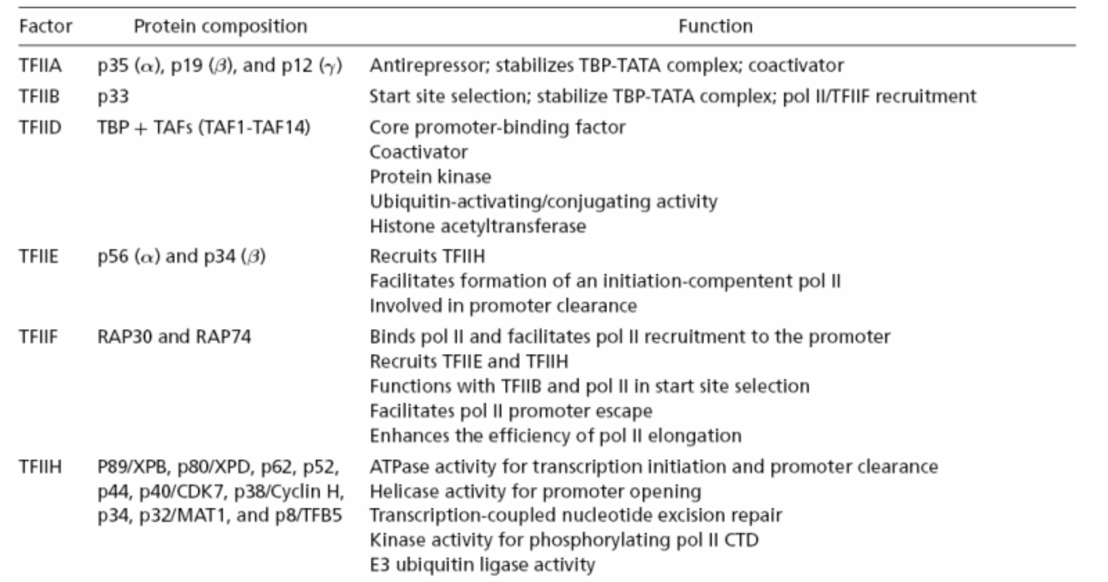


Figure 1.8 Side view of RNA (red) synthesis by RNA polymerase II machinery from a DNA template (template strand blue and coding strand is green). Cut surfaces of the protein, in the front, are lightly shaded and the remainder, at the back, are darkly shaded. By convention, the polymerase is moving on 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).

Although the polII machinery is able to initiate and synthesize RNA from any DNA template, it requires accessory proteins, called general transcription factors (GTFs), to recognize the start sites of genes, i.e. promoters, for correct initiation. There are six different general transcription factors listed in Table 1.2. Each GTF name is composed of the ‘TF’ prefix for Transcription Factor, ‘II’ for RNA polymerase II machinery, and a ‘letter’ corresponding to which chromatographic fraction the specific GTF was isolated from.



Factor	Protein composition	Function
TFIIA	p35 (α), p19 (β), and p12 (γ)	Antirepressor; stabilizes TBP-TATA complex; coactivator
TFIIB	p33	Start site selection; stabilize TBP-TATA complex; pol II/TFIIF recruitment
TFIID	TBP + TAFs (TAF1-TAF14)	Core promoter-binding factor Coactivator Protein kinase Ubiquitin-activating/conjugating activity Histone acetyltransferase
TFIIE	p56 (α) and p34 (β)	Recruits TFIIH Facilitates formation of an initiation-compentent pol II Involved in promoter clearance
TFIIF	RAP30 and RAP74	Binds pol II and facilitates pol II recruitment to the promoter Recruits TFIIE and TFIIH Functions with TFIIB and pol II in start site selection Facilitates pol II promoter escape Enhances the efficiency of pol II elongation
TFIIH	P89/XPB, p80/XPD, p62, p52, p44, p40/CDK7, p38/Cyclin H, p34, p32/MAT1, and p8/TFB5	ATPase activity for transcription initiation and promoter clearance Helicase activity for promoter opening Transcription-coupled nucleotide excision repair Kinase activity for phosphorylating pol II CTD E3 ubiquitin ligase activity

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).

When the cell decides to transcribe a gene, the first step towards forming a functional initiation machinery is the recruitment of TFIID that recognizes a core promoter element TATA box. Then TFIIA joins and further stabilizes the complex. The third GTF entering the complex is TFIIB, which mainly functions in recognizing the correct site for initiation and induces recruitment of polII machinery. Once the D-A-B complex forms on DNA, the polII machinery together with TFIIF binds to it. TFIIF was shown to reduce non-specific DNA contacts of polII *in vitro*, therefore it might play a role in the recruitment of polII to correctly positioned D-A-B complexes (Finkelstein et al., 1992). Once polII is assembled on the correct initiation site, then TFIIE binds directly to polII at the position of the jaw (see Figure 1.8). TFIIE may be involved in regulating jaw opening and closing (Leuther et al., 1996) (Svejstrup, 2004). Then, TFIIH joins the complex, a step which completes the preinitiation complex. TFIIH has several vital duties; (i) it has an helicase activity to open promoter sequences for transcription (Schaeffer et al., 1993), (ii) it phosphorylates the CTD of the largest subunit of polII by its kinase activity and (iii) it aids the polII complex to escape the ties between the promoter and become engaged in mRNA production by its ATPase activity (promoter clearance) (Svejstrup, 2004).

As mentioned earlier, TFIID, which is composed of TATA binding protein (TBP) and its associated factors, (TAFs) makes the first contact with the promoter to initiate transcription. TAFs are required for making direct contact with core promoter elements hence recognizing the correct initiation site. The contact points of different TAFs with core promoter elements are shown in Figure 1.9. Also, TFIID mediates an open chromatin conformation through its TAF1 subunit (Mizzen et al., 1996). However, whole-genome microarray analyses showed that individual TAFs are required for the expression of only a subset of genes (Chen and Hampsey, 2002) (Lee et al., 2000) with several promoters requiring different subsets of TAFs for activation (Lee et al., 2000). Nevertheless, TAFs are crucial for transcription but their role is not well understood.

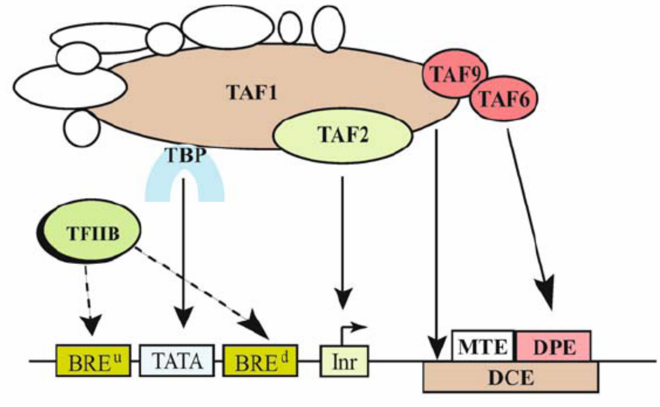


Figure 1.9 Known contacts between TATA box binding protein associated factors (TAFs) and core promoter elements as explained in Figure 1.3. This figure is adapted from reference (Thomas and Chiang, 2006).

Until recently, TFIID was thought to be universal for all polymerase initiation complexes (PICs). However, three different PICs have been found, the composition of which is given in Figure 1.10 (Wieczorek et al., 1998). The TFIID$_{\alpha}$ and TFIID$_{\beta}$ have TBP but only TFIID$_{\beta}$ has TAF10. TFTC, TBP-free TAF complex, has all TAFs except TAF1.

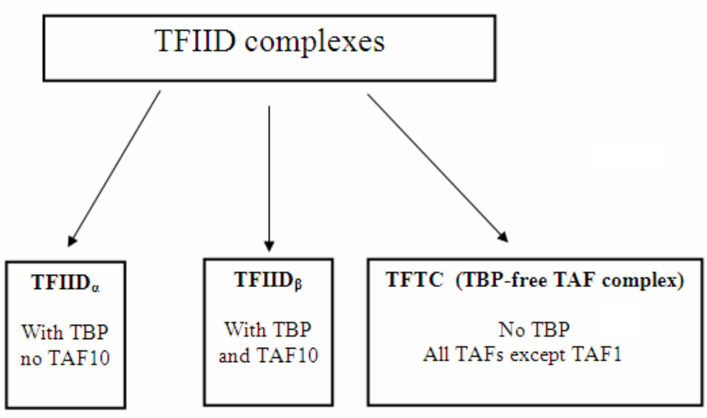


Figure 1.10 Three different TFIID complexes depending on the inclusion of TBP and TAF10.

TAF10 was found to be essential for early mouse development since there were no viable mice without TAF10 (Mohan et al., 2003). A recent study showed that TAF10 is only required in foetal skin development but not in adult stages, meaning that different PIC assemblies can be dynamic depending on the cellular environment and developmental stage of the cell (Indra et al., 2005).

1.3.2 Transcription Factors

Transcription initiation and elongation are achieved via polII machinery (section 1.3.1), which can drive transcription in cell-free systems to significant levels. However, measurable transcription obtained *in vivo* requires the action of regulatory proteins called transcription factors (TFs) . TFs interact directly or indirectly with the regulatory DNA sequences and modulate the assembly or disassembly of the basal transcription machinery. They can be grouped into two broad categories, the sequence specific regulators and the coregulators (Locker, 2001). Sequence specific regulators can interact with DNA sequences and they are modular in nature meaning that they use different domains to recognize and bind to DNA and exert their regulatory effects. They are subdivided into two groups as activators and repressors according to their positive or negative action on the transcription process respectively. However, the regulatory action of these sequence specific factors can be dependent on the cellular context in which they function and/or the DNA binding site. Coregulators do not bind directly to DNA, but they function via protein-protein interactions with sequence specific regulators and other coregulators. They are also subdivided into categories as coactivators or coreppessors depending on the nature of their action on the transcription.

1.3.2.1 Sequence Specific Transcription Factors

Sequence specific transcription factors interact directly with the regulatory DNA sequences via their DNA binding domain. Although these domains vary a great deal in nature, the majority of them has a design to position mainly an α-helix in the major groove within the binding site to make sufficient and specific molecular interactions with the charged phosphate backbone of the DNA as well as with the bases (Locker, 2001). The most common DNA binding domains are described below.

The Leucine Zipper DNA binding motif

The leucine zipper DNA binding motif is formed by a helical stretch of amino acids with leucine occurring once every seven residues, i.e. once every two turns of the helix (Figure 1.11).

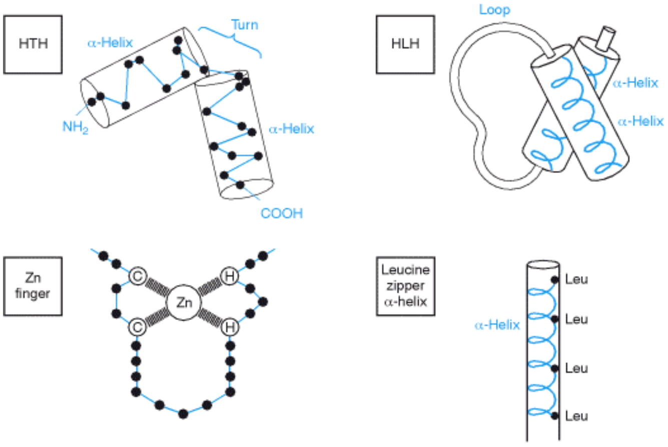


Figure 1.11 Schematic representation of four common DNA binding domains in transcription factors. Abbreviations: HTH, helix-turn-helix; HLH, helix-loop-helix, Zn, zinc; Leu, leucine. This figure is reproduced from (Strachan and Read, 2003).

The helix is formed in such a way that polar amino acids face one side of the helix and the hydrophobic ones face the other side. This unit needs to dimerize through another molecule with a leucine zipper domain to form a 'Y' shaped structure. Then the dimer is able to bind to DNA by gripping the double helix much like a clothes peg grips a clothes line (Strachan and Read, 2003). Representatives of this family are the well known FOS (c-fos, FosB,Fra-1 and Fra-2) and JUN (c-jun, JUNB and JUND) gene families which can hetero-dimerize and form the AP-1 transcription factor as shown in Figure 1.12 (Hess et al., 2004). Although dimerization is achieved through the leucine zipper motif, these proteins also form together a basic domain to bind to their palindromic recognition sequences on DNA. Each heterodimer has a different activation potential, such that some heterodimers can even have repressor effect by forming inactive dimers to compete for binding sites (Hess et al., 2004). Members of the JUN family can be phosphorylated by mitogen-activated protein kinases (MAPK), which could then increase their gene activation potential. The latter plays a role in cell proliferation and apoptosis which are triggered by extracellular stimuli (Behrens et al., 1999). The AP-1 family is also shown to play an important role in differentiation (Angel and Karin, 1991) (Eferl et al., 1999) (Behrens et al., 2001).

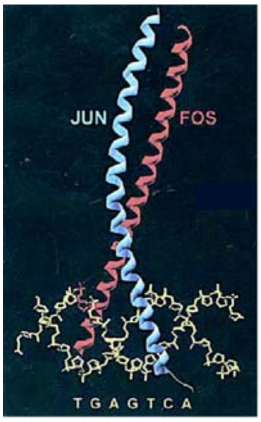


Figure 1.12. Jun and fos proteins both have leucine zipper motif to form a heterodimer (AP-1) to bind to their palindromic recognition sequences on DNA. This figure is reproduced from reference (Hess et al., 2004).

CREB1 (cAMP responsive element binding protein 1) is another protein carrying a leucine zipper binding motif. This protein homodimerizes and binds to an octamer palindromic DNA sequence called cyclic-AMP responsive element (CRE) (Deutsch et al., 1988). Phosphorylation of this protein induces transcription of several genes in response to hormonal stimulation of the cyclic AMP (cAMP) pathway (Cardinaux et al., 2000).

The helix-loop-helix DNA binding motif

The helix-loop-helix (HLH) motif is related to the leucine zipper in terms of the structural conformation for DNA binding (Figure 1.13). It consists of two α-helices, one short and one long, connected by a flexible loop and DNA binding is achieved through dimerization with another (HLH) domain containing protein (Figure 1.11 and Figure 1.13). HLH mediates its homo or heterodimerization through its hydrophobic residues (Murre et al., 1994).

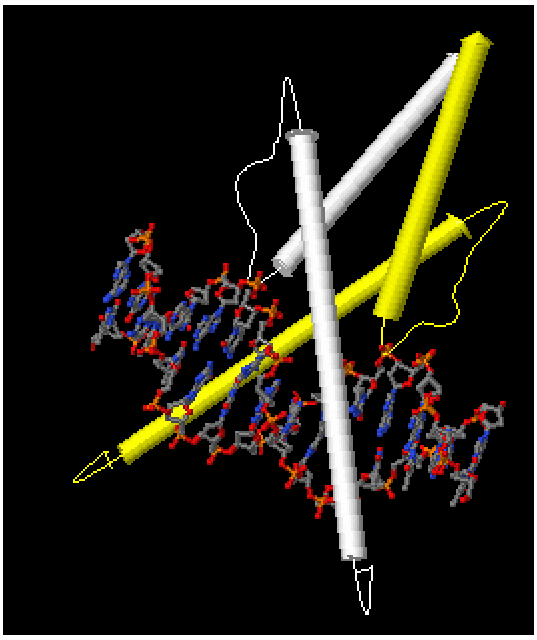


Figure 1.13 The helix-loop-helix motif carrying dimer protein bound to DNA. Different subunits are shown in white and yellow. This illustration is adapted from URL site reference URL1, 2004.

HLH proteins are well conserved in evolution, present from yeast to human and play an important role in developmental processes (Atchley and Fitch, 1997) such as heart, pancreatic, muscle, B and T cell development, neurogenesis and hematopoiesis (Massari and Murre, 2000). An oncogene, MAX, also contains the HLH motif and is able to form dimers with other HLH proteins such as MYC, MAD or Mxi1. These dimers are implicated in cell proliferation, differentiation and apoptosis (Grandori et al., 2000).

The helix-turn-helix proteins

The helix-turn-helix (HTH) motif consists of two short α-helices connected with a short loop which introduces a turn so that the two helices do no lie in the same plane (Figure 1.11). In this motif, one of the helices, also called the recognition helix, fits into the major groove of DNA and participates in sequence-specific recognition of DNA, whereas the N-terminal helix functions primarily as a structural component that aids to position the recognition helix (Alberts et al., 2001). Many bacterial repressors utilize HTH motif to bind to DNA, although crucial activator factors within the basal initiation machinery also contain HTH, indicating the functional diversity of this motif (Ohlendorf et al., 1983; Gribskov and Burgess, 1986). This versatility in function of the proteins carrying this motif comes from the fact that their structure outside the helix differs a lot from protein to protein, which enables each protein to employ its HTH motif to bind to a variety of DNA sequences with the help of the rest of their structure (Alberts et al., 2001). For instance, chromatin proteins like histone H1 protein and basal transcription factors TFIIB and TFIIH utilizes this motif to bind to their cognate DNA sites (Aravind et al., 2005). A specialized form of HTH is called homeodomain, (Figure 1.14), a vital domain used in developmental processes in every eukaryotic organism studied up to date. Homeodomain is also used by the POU and

MYB family transcription factors, which play a vital role in B-cell development and cell cycle progression respectively.

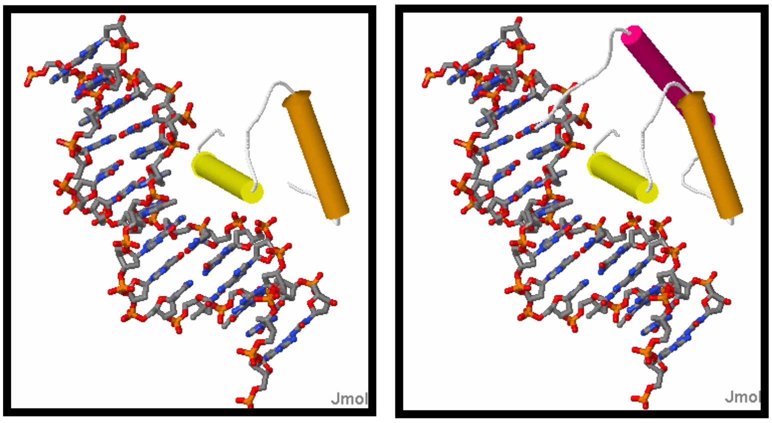


Figure 1.14 Helix-turn-helix (HTH) motif bound to its DNA on the left and a specialized form of HTH motif called homeodomain bound to its DNA. Homeodomain contains an extra α-helix presumably for further stabilization of DNA binding. These figures are adapted from URL site reference URL2, 2004.

Zinc finger proteins

The zinc finger motif is composed of a loop of polypeptide chain held in a hairpin bend bound to a zinc atom (Figure 1.11). Usually, the zinc atom is held by two conserved cysteine and histidine residues although a number of different forms exist (Strachan and Read, 2003). This motif can be tandemly repeated within a protein which enables it to bind long and diverse DNA sequences. CTCF (CCCTC-binding protein) presents a unique example since it contains 11 structurally adjacent zinc-finger motifs (Ohlsson et al., 2001), which enables CTCF to bind to around 50 bp long sequences.

Sp1 also carries three classic zinc-finger motifs for DNA binding. This protein is a common transcriptional activator, recognizing GC-rich sequences within the proximal promoter regions called GC-boxes (Letovsky and Dynan, 1989). It is also required to prevent the methylation of CpG islands (Brandeis et al., 1994) which is crucial for CpG islands found in promoters (see section 1.2.1).

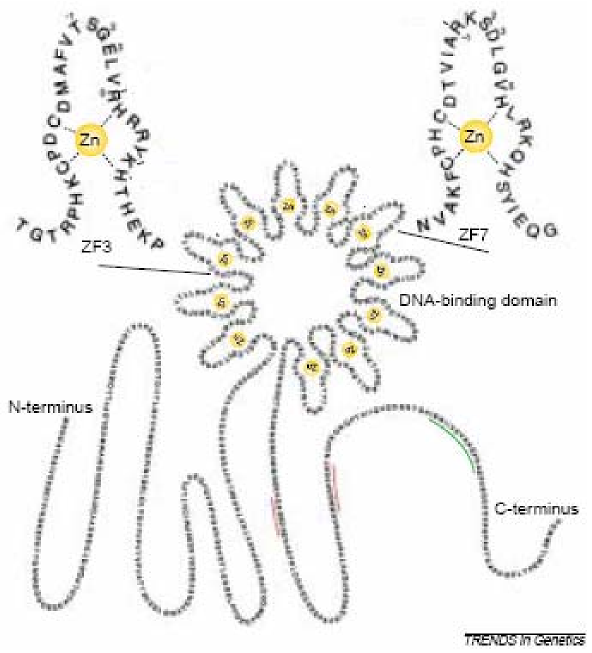


Figure 1.15. Structure of DNA binding domain of CTCF composed of 11 adjacent zinc finger domain. This figure is adapted from (Ohlsson et al., 2001).

YY1 protein is another factor containing four zinc fingers and it can either act as an activator or as a repressor depending on the cellular context in which it functions (Hahn, 1992) (Shi et al., 1997). It also has the ability to induce bends in the DNA structure which mediates its repression function by preventing protein contacts for gene activation (Natesan and Gilman, 1993; Kim and Shapiro, 1996).

As mentioned earlier, transcription factors are modular molecules which use different domains for DNA binding and for activation or repression. A TF may contain more than one activation domain and these domains can be of different type, which will increase the number of proteins they can interact with (Triezenberg, 1995).

A common activation domain is the acidic domain which is composed of one amphipathic α-helix in which all the negatively charged amino acids are displayed on one side of the helix creating a net negative charge (Latchman, 1998). It has been shown that mutations in the activation domain that increase the net negative charge, increase this domain’s ability to activate transcription, whereas mutations that reduce the negative charge result in the opposite effect (Gill and Ptashne, 1987).

Another activation domain is the glutamine-rich domains, which has 25% glutamine and contains very few negatively charged residues (Gill and Ptashne, 1987). Sp1 mediates its activatory functions through a glutamine-rich domain and interestingly, substitution of its activation domain with another glutamine-rich domain with no apparent sequence homology did not affect its activatory function (Latchman, 1998).

A third type is composed of 25% proline residues. Oncogenes such as JUN, AP2 or GTFIIIA (general transcription factor IIIA) and POU2F2 (Oct-2) all contain this domain and use it to activate the transcription of their target genes (Latchman, 1998).

Different subunits of the transcriptional machinery interact with specific activation domains; for example, TFIIH interacts with factors carrying an acidic domain (Xiao et al., 1994), whereas factors with proline-rich activation domains bind to TFIIB for its recruitment to initiation complex (Lonard and O'Malley, 2005)

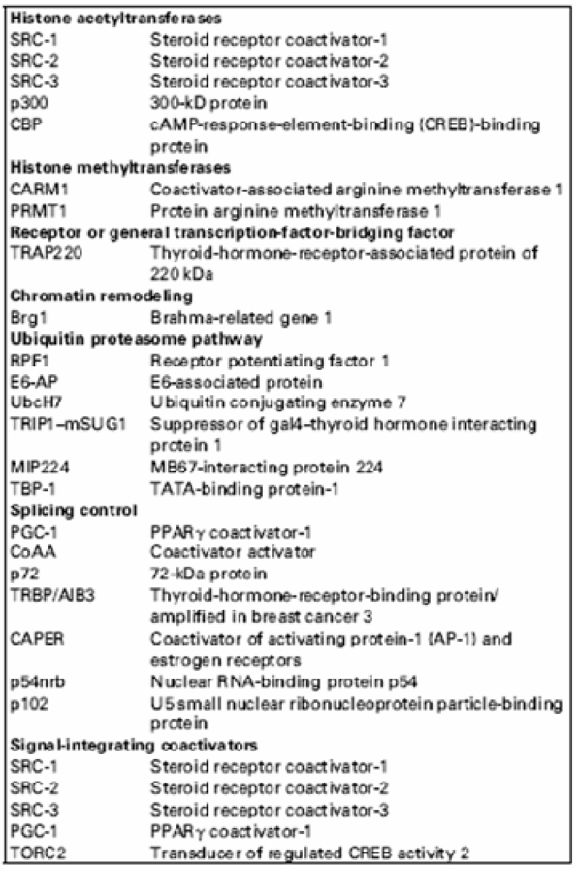
TFs can also have repressive effects on transcription. Although many studies characterising repressive domains have been reported (Leichter and Thiel, 1999; Peng, Begg, Harper et al., 2000; Peng, Begg, Schultz et al., 2000), little information is available about their structural nature. One well-known transcriptional repressor is MECP2 (methyl CpG binding protein 2) that can bind to methylated CpG dinucleotides on promoter sequences and repress transcription by recruiting chromatin modifying protein complexes to create a repressive chromatin environment. Another repressor protein REST (RE1-silencing transcription factor) binds to a repressor DNA motif called RE1 (repressor element 1) and reduces the transcription of RE1 containing (mainly neuronal) genes two to ten fold in non-neuronal cells (Schoenherr and Anderson, 1995).

Also, many TFs exert their negative effects via their DNA binding domains by either competing with activatory factors or producing such DNA structures which disables interactions between activatory proteins and basal transcription machinery.

1.3.2.2 Coregulators

Coregulators represent a diverse set of proteins that have the ability to activate or repress transcriptional processes via their protein-protein contacts. More importantly, they provide the functional link between the cellular and extra-cellular signals and basal transcription machinery. They mainly interact with nuclear receptors (NRs), which are ligand-regulated factors that transduce hormonal signals from steroid hormones and other lipophillic ligands (Lonard and O'Malley, 2005). These coregulators affect the transcriptional machinery via their enzymatic functions by changing the chromatin environment of the regulatory region, or modifying other transcription factors by phosphorylation, ubiquitinylation or sumoylation. Coactivators interact with the basal transcription machinery or its associated factors in a positive way to start or enhance the transcription levels of a gene. Table 1.3 lists a number of coactivators with different enzymatic capabilities to affect transcription.

Corepressors, in contrast, have a negative effect on gene transcription. They can be recruited to a variety of DNA regulatory elements and they exert their silencing effects mainly by creating a repressive chromatin environment. CoREST is a co-repressor, which interacts with REST and mediate the repression of neuronal genes in non-neuronal cells to maintain the cell's identity (Andres et al., 1999). CoREST is mainly associated with HDAC1/2 and AOF2 (a histone demethylase, see section 1.4.5) suggesting that coREST-mediated repression involves in changes in histone code of the locus it operates on (Lee, Wynder et al., 2005).



Histone acetyltransferases	
SRC-1	Steroid receptor coactivator-1
SRC-2	Steroid receptor coactivator-2
SRC-3	Steroid receptor coactivator-3
p300	300-kD protein
CBP	cAMP-response-element-binding (CREB)-binding protein
Histone methyltransferases	
CARM1	Coactivator-associated arginine methyltransferase 1
PRMT1	Protein arginine methyltransferase 1
Receptor or general transcription-factor-bridging factor	
TRAP220	Thyroid-hormone-receptor-associated protein of 220 kDa
Chromatin remodeling	
Brg1	Brahma-related gene 1
Ubiquitin proteasome pathway	
RPF1	Receptor potentiating factor 1
E6-AP	E6-associated protein
UbcH7	Ubiquitin conjugating enzyme 7
TRIP1–mSUG1	Suppressor of gal4-thyroid hormone interacting protein 1
MIP224	MB67-interacting protein 224
TBP-1	TATA-binding protein-1
Splicing control	
PGC-1	PPARγ coactivator-1
CoAA	Coactivator activator
p72	72-kDa protein
TRBP/AIB3	Thyroid-hormone-receptor-binding protein/ amplified in breast cancer 3
CAPER	Coactivator of activating protein-1 (AP-1) and estrogen receptors
p54nrb	Nuclear RNA-binding protein p54
p102	U5small nuclear ribonucleoprotein particle-binding protein
Signal-integrating coactivators	
SRC-1	Steroid receptor coactivator-1
SRC-2	Steroid receptor coactivator-2
SRC-3	Steroid receptor coactivator-3
PGC-1	PPARγ coactivator-1
TORC2	Transducer of regulated CREB activity 2

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).

Corepressors can also interact with transcriptional activators and block their activating function by masking their activation domains. One example is the regulation of the activity of E2F, a cell cycle transcription factor, by the RB tumour suppressor protein (Weintraub et al., 1995). When E2F is not bound to RB, it acts as a transcriptional activator. However, when RB binds to E2F it blocks E2F activation domain and actively silences transcription by recruiting other silencers (Brehm et al., 1998).

1.4 Chromatin and Transcription

Transcriptional activity and chromatin structure are tightly coupled. Human genomic DNA is tightly wrapped around small basic proteins called histones, restricting its accessibility to factors involved in DNA replication and transcription. The genomic DNA together with its associated proteins is called chromatin. All histones are small proteins, highly basic and rich in lysine (K) and arginine (R), amino acids that are positively charged at cellular pH (Turner, 2001). There are five types of histone proteins, their molecular properties are given in Table 1.4 and structural organisation is shown in Figure 1.16.

Histones		**Number of Residues**	**Residues/mol (%)**		
			Lysine	**Arginine**	**Net Charge**
Core	H2A	129	14 (10.9)	12 (9.3)	+15
	H2B	125	20 (16.0)	8 (6.4)	+19
	H3	135	13 (9.6)	18 (13.3)	+20
	H4	102	11 (10.8)	14 (13.7)	+16
Linker	H1	224	66 (29.5)	3 (1.3)	+58

Table 1.4 Chemical properties of histone proteins

Two molecules of H2A, H2B, H3 and H4 make up the core nucleosome particle that wraps 146 bp long DNA sequence. Although core histones do no share significant sequence homology, they all share a structural domain called histone fold (Mersfelder and Parthun, 2006).

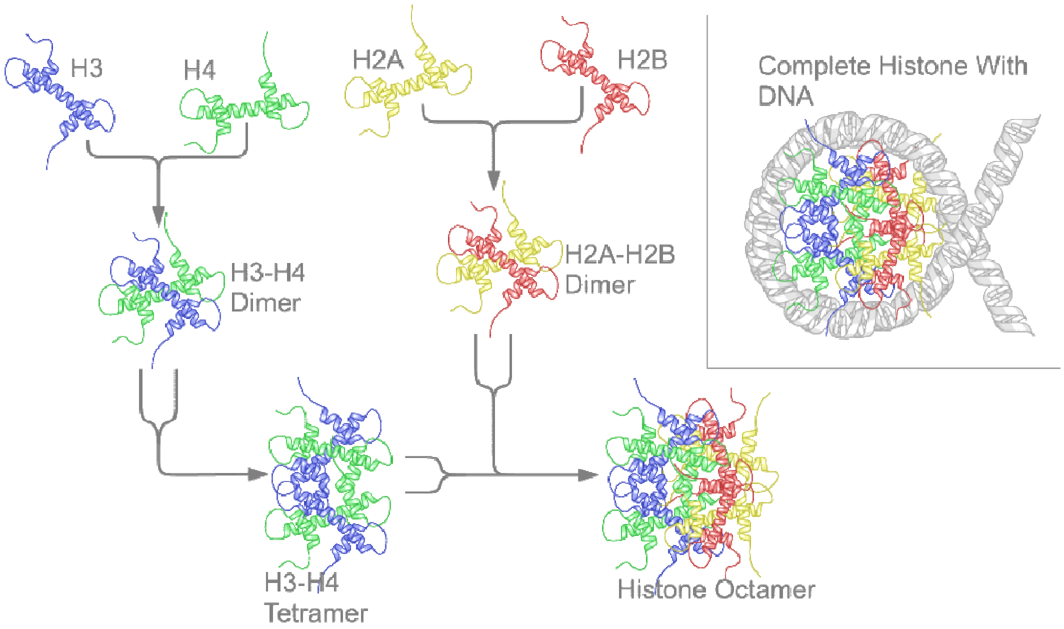


Figure 1.16. Structural organisation of core histones in the nucleosome core particle. This figure is adapted from reference (Alberts et al., 2001).

The linker histone H1 binds to approximately 20 bp of DNA and positions close to where the DNA strand enters and exits the nucleosome. The nucleosome core particle together with the histone H1 is called chromatosome and each chromatosome is linked with around 40 bp of linker DNA. A proposed model for further genomic DNA packing is a helical array consisting of eight chromatosomes, which represents 10 nm chromatin fiber structures (Figure 1.17B). Chromatin is proposed to be further packed in higher order structures (30 nm fiber) called solenoids (Figure 1.17C).

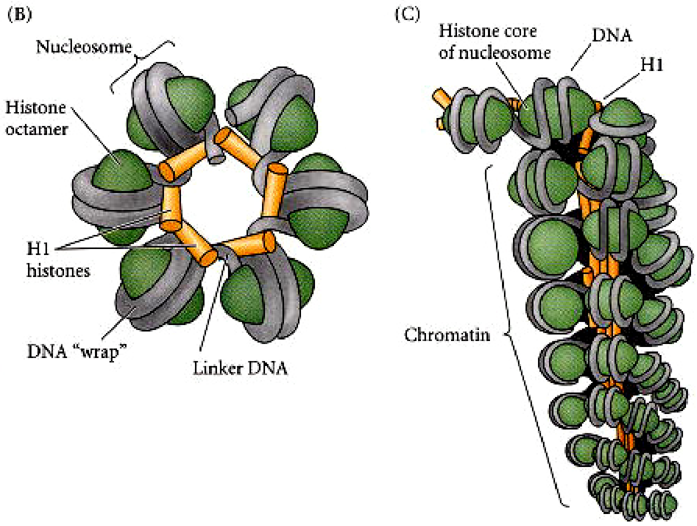


Figure 1.17. (B) About 160 base pairs of DNA encircle each histone core particle, nucleosome, and about 40 base pairs of DNA link the nucleosomes together. (C) Model for the arrangement of nucleosomes in the highly compacted solenoidal chromatin structure. This figure is adapted from reference (Turner,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

These solenoid structures are thought to be attached to the nuclear matrix by matrix (or scaffold) attachment sites (MARs or SARs) and form DNA loops (see Figure 1.18). Then these scaffold attached loops of DNA, so called chromatin fibers, are further packed and form chromosomes (see Figure 1.18).

Two endonucleases that digest DNA in a non-sequence-specific manner within internal sites (as opposed to digesting from the ends) are particularly useful for mapping nucleosome organization. Micrococcal nuclease, which has to surround DNA in order to cut it, only cuts the linker DNA and leaves intact the DNA wrapped twice around the nucleosomal core. This enzyme is particularly important for understanding nucleosomal positioning in different parts of the genome. It was shown that nucleosomes are translationally positioned on the active allele and rotationally positioned on the inactive allele of the human *HPRT* promoter using differential cutting patterns by micrococcal nuclease (Chen and Yang, 2001).

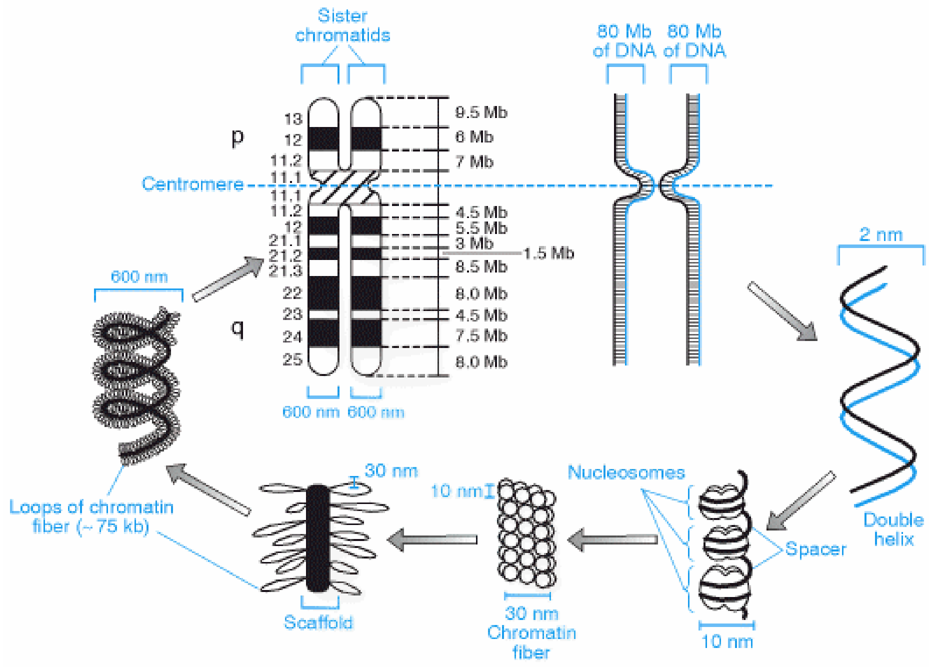


Figure 1.18. Higher order of DNA packaging in nucleus. This figure is reproduced from reference (Strachan and Read, 2003).

Deoxyribonuclease I (DNase I) operates rather differently introducing nicks in one or the other strand of DNA. So, this enzyme cuts the DNA that is wrapped around the nucleosomes but the sequences that are furthest away from the nucleosomal surface can be nicked better (Turner, 2001). Genomic sequences that are free or depleted from nucleosomes can be easily digested by DNase I in a uniform cutting manner, therefore it is possible to map such regions. These sites are called DNase I hypersensitive sites (HSs) and they are mostly free or depleted of nucleosomes. These sites are thought to be more accessible to DNA binding proteins and enriched in locus control regions, regulatory elements and replication origins. DNase I hypersensitivity depends on the developmental stage or cell type. A novel method was developed to map such sites in the human genome using genomic microarrays (Sabo et al., 2006).. The above study verified many well-known DNase I hypersensitive sites. In the ENCODE project, promoters, 3' end of genes and intronic regions were found enriched in DNase I HSs, whereas distal intergenic sites were depleted in such sites. The finding that the 3' UTRs also contains DNase I HSs may suggest their possible role in the process of transcription termination or regulation of antisense transcripts (Sabo et al., 2006). Figure 1.19 displays four different types of genomic regions in terms of their DNase I hypersensitivity.

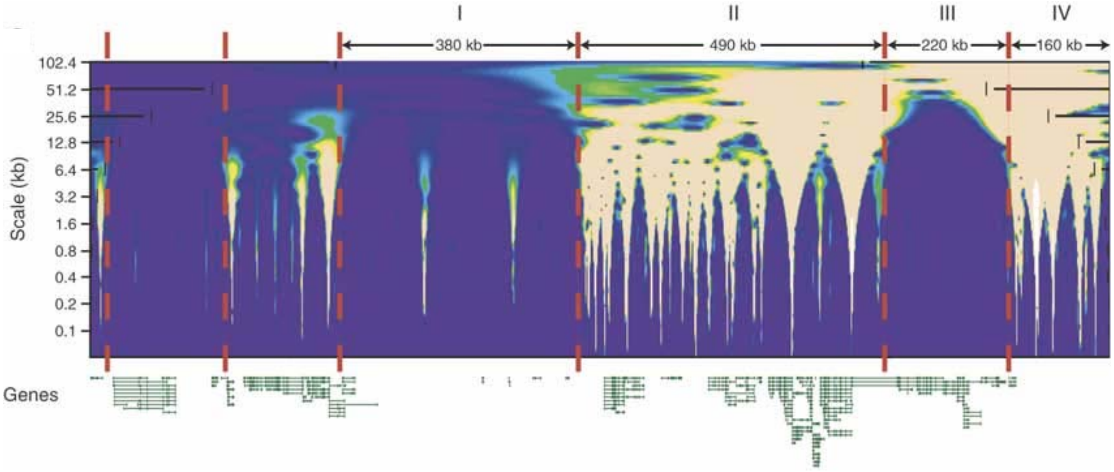


Figure 1.19. Continuous wavelet transform heat map of chromatin accessibility across a 1.7-Mb segment of chromosome 21 containing the Down Syndrome critical region29 (x axis, genomic position; y axis, wavelet scale), genes are shown below heat map. Four broad classes of chromatin domains are thus distinguished 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).

Interestingly, gene-poor regions that neighbour a gene-rich region are also densely populated with DNase I HSs, verifying that distal regulatory elements are also enriched with DNase I HSs.

There are two types of chromatin structures in the nuclei of many higher eukaryotic cells: a highly condensed form, called heterochromatin, and a less condensed form, called euchromatin. Euchromatic regions decondense during the interphase of the cell cycle and contain most of the genes coding for cellular proteins, while heterochromatic regions stay condensed even in the interphase and are generally silent (Grunstein, 1997). Although heterochromatic and euchromatic regions have nearly the same histone composition, heterochromatin includes additional proteins for further packing of the chromatin and different post-translational modifications are present in each form (Craig, 2005).

Heterochromatin can be classified into two major subtypes, namely facultative and constitutive. Facultative heterochromatin is similar to euchromatin in terms of gene density and sequence characteristics, but it is highly packed with accessory proteins, like heterochromatin protein-1 (HP1), has silencing histone modifications, and is transcriptionally inactive (Turner, 2001). Conversely, constitutive heterochromatin occurs in large blocks near centromeres and telomeres and contains mostly repetitive elements (Dimitri et al., 2005). Constitutive chromatin is always silent, whereas facultative chromatin is formed when it is necessary to permanently silence genes or regions. Facultative heterochromatin formation follows a complex pattern, it is developmentally regulated and genomic regions can selectively become heterochromatic to establish and maintain cell identity (Craig, 2005). It is known that polycomb proteins form protein complexes and function in the formation of heterochromatic regions for the transcriptional repression of the genes involved in embryogenesis, cell cycle and tumorigenesis (Orlando, 2003).

Structural changes in the chromatin play a vital role in the control of gene expression and are governed by complexes that remodel chromatin and enzymes that post-translationally modify histones. Chromatin-remodelling complexes (SWI/SNF family) mobilize nucleosomes, causing the histone octamers to move short distances along the DNA to make sequences accessible to regulatory proteins (Becker and Horz, 2002). Each remodelling complex has ATPase activity to provide the necessary energy.

Nucleosomes can be covalently modified and regulate expression. Different combinations of histone modifications in a locus will result in different expression patterns. These different covalent modifications are collectively referred to as the histone code since it actually encodes the regulatory pattern of the genes by changing the chromatin environment (Strahl and Allis, 2000).

Histones can be post-translationally modified either in their amino terminal tails which includes the first 25-40 amino acids, or on their core domains. Histone amino tails pass between the gyres of the DNA helix and the structure of the first 20-25 amino acids on the amino terminal cannot be determined; (Figure 1.16) they come across as random coils (Luger et al., 1997). Modifications include phosphorylation, ubiquitinylation, sumoylation, acetylation or methylation. They provide the means of communication between chromatin and non-histone proteins such as transcription factors. Modifications within the histone core are classified into three groups in terms of their effects (Mersfelder and Parthun, 2006):

- Solute accessible face; are able to change higher order chromatin structure and chromatin protein interactions
- Histone lateral surface; mainly affect the interactions between histone and DNA
- Histone histone interface; affect nucleosome stability.

This study is mainly focused on modifications occurring in the amino terminal tails of histone H3 and H4 and which therefore are described in more detail below

1.4.1 Histone Phosphorylation

Histone H3 can be phosphorylated on serine 10, a modification associated with chromosome condensation in mitosis (Hans and Dimitrov, 2001) in several organisms although the molecular mechanism is not yet known. In humans, the same modification is shown to aid HP1 disassociate from chromatin in mitosis, which may help repositioning structural chromosomal proteins (Hirota et al., 2005). Also, histone H3 is rapidly phosphorylated in response to growth factors and protein synthesis inhibitors, presumably to help activation of related genes (Mahadevan et al., 1991). A more recent study showed that histone phosphorylation occurs synergistically with histone acetylation in response to growth factors on target promoter regions (Cheung et al., 2000). These findings are important since they provide a link between the extracellular environment and the histone code and need to be investigated further.

Linker histone H1 can also be phosphorylated and this modification is associated with chromosome condensation in various organisms (Hans and Dimitrov, 2001), although its role in human has not yet been established. Histone H2B also has been shown to be phosphorylated universally in apoptotic cells and associated with apoptosis-specific nucleosomal DNA fragmentation (Ajiro, 2000).

1.4.2 Histone Ubiquitinylation

Ubiquitin is a ubiquitously expressed 76 amino acid protein that can be covalently attached to target proteins. Target proteins can be mono- or poly-ubiquitinylated, and these modifications control protein degradation, stress response, endocytic trafficking, chromatin structure and DNA repair (Di Fiore et al., 2003) (Zhang, 2003). Histones H1, H2A, H2B, H3 and H4 can all be ubiquitinylated at their lysine residues and affect transcription in a positive or negative manner. Recently, ubiquitinylated histone H2A at K119 was associated with X chromosome inactivation and polycomb mediated gene silencing (de Napoles et al., 2004; Wang, Wang et al., 2004). It is known that this modification is somehow related to histone H3 methylation at K27, although the molecular role of this relation is not yet clear (Shilatifard, 2006). Histone 2B can also be ubiquitinylated and this modification sets a regulatory mark in signalling for histone methylation (Wood et al., 2005). This modification is shown to be associated with transcriptional elongation (Xiao et al., 2005).

A protein complex that can ubiquitinylate histones H3 and H4 proteins was purified and used to show *in vivo* and *in vitro* evidences that these modifications are associated with cellular response to DNA damage (Wang et al., 2006). TAF1 can ubiquitinylate linker histone H1 *in vitro*, although the functional identification and characterization of this modification *in vivo* has not yet been done (Belz et al., 2002).

1.4.3 Histone Sumoylation

Small ubiquitin-related modifier (SUMO) is a ubiquitin-like molecule that can also be covalently linked to target proteins controlling their interactions with other proteins, cellular localization or their degradation (Melchior, 2000). Many transcriptional factors and chromatin modifying enzymes have been shown to be sumoylated reversibly. Sumoylation is nearly always associated with transcriptional silencing (Gill, 2005). Histone H4 can be sumoylated mediating gene silencing through recruitment of histone deacetylases and HP1 (Shiio and Eisenman, 2003). This is currently the only histone sumoylation known in humans.

1.4.4 Histone Acetylation

Acetylation of histones is controlled by the histone acetyl transferase (HAT) and histone deacetylase (HDAC) enzymes. Table 1.5 shows the known acetylation sites of different histones at their lysine residues.

Histone	Lysines that can be acetylated
H2A	5, 9 (minor)
H2B	5, 12, 14, 20
H3	9, 14, 18, 23
H4	5, 8, 12, 16

Table 1.5. Lysine residues that are acetylated on amino terminal tails of histones

HATs transfer an acetyl group from acetyl coenzyme A to the ε-amino group of lysine residues of core histones (Table 1.5 and Figure 1.20).

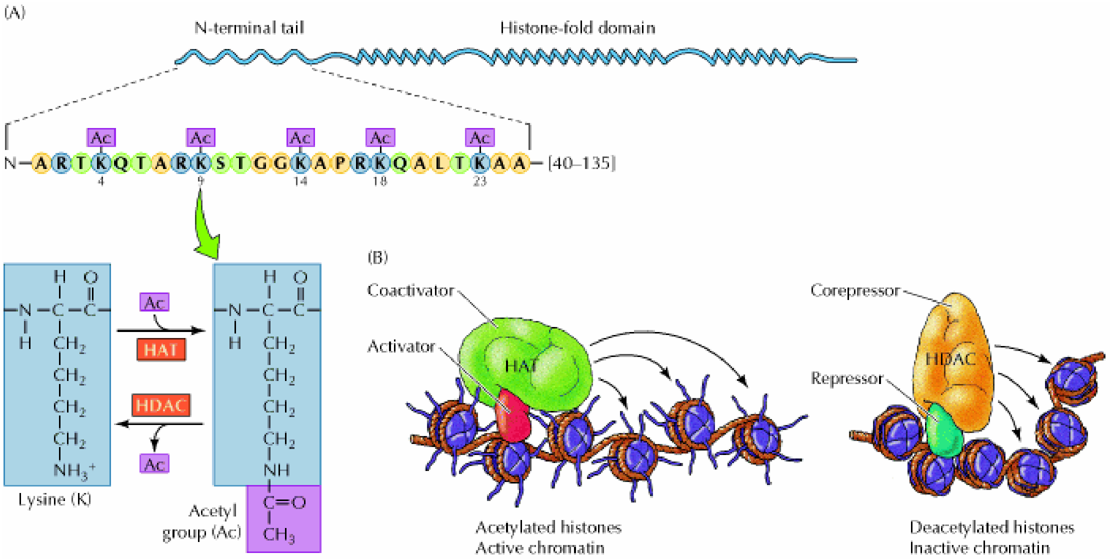


Figure 1.20. (A) The N-terminal tails of the core histones (e.g., H3) are modified by the addition of acetyl groups (Ac) to the side chains of specific lysine residues. (B) Transcriptional activators and repressors are associated with coactivators and corepressors, which have histone acetyltransferase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

Histone acetylation has been associated with transcriptional activation and active chromatin regions in humans (Allfrey et al., 1964) (Strahl and Allis, 2000). Lysine residues in histones carry a positive charge which increases the histone tail's affinity for DNA. Addition of acetyl groups reduces this affinity and helps decondensing the chromatin structure, although there is no direct evidence for this mechanism. Many transcriptional activators, including TAF1, contain a domain, namely bromodomain, that binds to acetylated histones (Jacobson et al., 2000), suggesting that histone acetylation may help recruiting the basal transcription machinery or co-activators such as PCAF (p300/CREB binding protein) (see Figure 1.20B) (Zeng and Zhou, 2002). Chromatin-remodelling enzymes also contain bromodomains, which could initiate the remodelling of chromatin structure for gene activation (Zeng and Zhou, 2002). In contrast, HDACs remove acetyl groups from lysine residues of histones and this action is associated with transcriptional repression, mediated by repressor proteins interacting with HDAC complexes (see Figure 1.20B). For instance, histone deacetylase 1 (HDAC1) interacts with MeCP2 which mediates repression of several genes (Suzuki et al., 2003).

There are several HATs in humans, which are involved in different cellular processes and have different affinities for different histones (Table 1.6).

HAT	Function	Preferred Histone
GCN5/PCAF family		
GCN5L2	Co-activator	H3
PCAF	Co-activator	H3
MYST family		
MYST1	Co-activator	H4
MYST2	Initiation of DNA replication (Iizuka et al., 2006)	H3, H4
MYST3	Leukeomogenesis (Deguchi et al., 2003)	
MYST4	Co-activator	
Nuclear Hormone Receptor Family		
NCOA1	Hormone signalled transcription (Onate et al., 1995)	H3, H4
NCOA3	Hormone signalled transcription (Chen et al., 1997)	H3, H4
Other HATs		
TAF1	TFIID subunit	H3
HAT1	Replication-dependent chromatin assembly (Makowski et al., 2001)	H4

Table 1.6 Histone Acetyltransferases that are active in humans and their known cellular functions.

As mentioned earlier, histone acetylation is involved in mainly transcriptional activation or active chromatin conformation. Although little is known about the function of histones H2A and H2B acetylation in humans, histone H3 and H4 acetylation have been associated with a number of cellular processes. Acetylated chromatin has increased sensitivity to DNase I, hence more accessible to interacting proteins *in vivo* (Krajewski and Becker, 1998). Also, histone acetylation is shown to be depleted in heterochromatic regions (Eberharter and Becker, 2002). Histones H2A, H3 and H4 are underacetylated in inactive female X chromosome, and histone H2A has a different acetylation pattern than H3 and H4 along the chromosome (Chen et al., 1997). It is important to note that acetylated histone H3 is not marking all open chromatin regions (such as intra-genic transcribed regions,) it is mostly concentrated around TSSs (Liang et al., 2004). Yet, histone H3 acetylation at K9 correlate with nucleosome depletion on promoter regions even if the gene is not actively expressed (Nishida et al., 2006).

The study by Shogren-Knaak et al. in which histone H4 acetylation at K16 prevents heterochromatin formation offers a direct evidence *in vivo* that histone acetylation indeed modify chromatin state (Shogren-Knaak et al., 2006). Moreover, loss of histone H4 acetylation only at K16 has been associated with several cancers in humans (Fraga et al., 2005; Fraga and Esteller, 2005). The link between this modification and cancer might be due to repression of tumour suppression genes, those have to be kept active otherwise for proper cell cycle regulation.

1.4.5 Histone Methylation

Histones H3 and H4 can be methylated on their ε-amino groups of lysine and arginine residues. Figure 1.21 shows that up to three methyl groups can be added on the same residue but this modification does not change the charge of the histones.

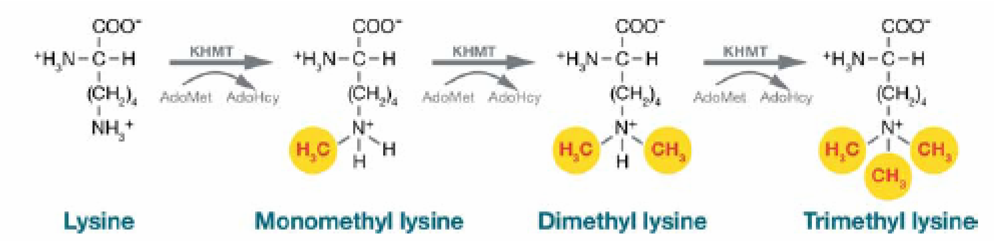


Figure 1.21 The chemistry of methylation on lysine residues of histones. Adomet (S-adenosyl-L-methionine or SAM) is a cofactor which carries the methyl group to be transferred. This figure is reproduced from reference (Shilatifard, 2006).

Enzymes that transfer methyl groups from the donor S-adenosyl-L-methionine (SAM) are called histone methyltransferases (HMTs) and differ greatly depending on the particular residue to be methylated on the histone. Table 1.7 lists known HMTs in humans and their preferred substrate on the tails of histones.

HMTs can be classified into three groups, those responsible for (i) methylating lysine 4, 9, 27 and 36 of histone H3 and K20 of histone H4 that carry a SET domain (ii) methylating K79 of histone H3 that have no SET and (iii) methylating arginine 2, 17 and 26 of histone H3 and arginine 3 of histone H4 (Shilatifard, 2006). HMT that methylates K59 of histone H3 in humans or any other organism has not been found yet.

Histone	Residue	Position	Index	Enzymatic Machinery
Histone H3	Lysine	Amino Tails	4	MLL
			9	SUV39H1
		Solute accessible core domain	27	EED-EZH2
			36	SETD2
		Histone-histone interface	79	DOT1L*
	Arginine	Tail	2	CARM1
		Tail	17	
		Tail	26	
Histone H4	Lysine	Tail	20	SET7/SET8
		Histone-histone interface	59	?
	Arginine	Tail	3	PRMT1

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.

The SET domain takes its name from the Drosophila proteins Su(var)3-9, Enhancer of Zeste [E(z)], and Trithorax(SET) (Jones and Gelbart, 1993). This domain is around 130 to 140 amino acid long and has a unique narrow channel structure that connects the methyl group donor (SAM) on one surface with the substrate binding site on the opposite surface of the domain (Xiao, Wilson et al., 2003). Also the geometry and shape of the side of channel facing SAM molecule seems to be determining the number of methyl groups to be added to substrate lysine residues (Xiao, Jing et al., 2003). One interesting property of proteins carrying this domain is their exceptional substrate specificity, which is mediated by a module within the SET domain that varies in length and has no significant sequence conservation between family members (Xiao, Wilson et al., 2003).

Histone methylation affects transcriptional accessibility of chromatin in both positive and negative manner. It is now well-established that promoter regions of most (if not all) of the actively transcribed genes are associated with tri-methylated histone H3 at K4 (H3K4me3) (Santos-Rosa et al., 2002; Kim et al., 2005). A recent study that examines the histone H3 methylation patterns at the promoters of key adipogenic genes during adipocyte differentiation showed that promoters of adipogenic genes are enriched in H3K4me2 (di-methylated histone H3 at K4) where none of these genes are yet expressed (Musri et al., 2006). They then showed that H3K4me2 is restricted to the promoter regions of adipogenic genes in undifferentiated cells and associated with RNA polymerase II loading. At the later stages of adipogenesis, the activation of these genes coincided with promoter histone H3 hyperacetylation and tri-methylation at K4. These results suggest that H3K4me2 serves as a preparatory signal for the start of transcription whereas H3K4me3 is the signal that actually marks transcription. Mono-methylation of histone H3 (H3K4me) is not restricted to TSSs and its functional importance in humans has not been clearly established yet.

Histone methylation at K4 requires MLL (myeloid/lymphoid or mixed-lineage leukemia), the human homologue of Drosophila trithorax protein (Milne et al., 2002). This protein was cloned over 15 years ago and associated with the pathogenesis of several different forms of haematological malignancies, including acute myeloid leukemia (AML) (Shilatifard, 2006). However, MLL and MLL chimeras found in translocations associated with leukemia that activates gene expression, appear to have no effect on histone methylation (Quentmeier et al., 2004).

HP1 can specifically recognize histone H3 methylated at K9 (Lachner et al., 2001) and this recognition is partly required for the establishment and maintenance of heterochromatin (Shilatifard, 2006). Although methylated histone H3 at K9 is mainly associated with silent regions of the genome, the number of methyl groups on K9 determines distinct chromatin structures (Rice et al., 2003). While mono and di-methylated K9 of histone H3 marks silenced euchromatic regions, tri-methylated

H3K9 is present in heterochromatin regions (Rice et al., 2003). Methylation on histone molecules is known as a more stable mark than phosphorylation and acetylation (Waterborg, 1993). However, it is very recently reported that during mitosis, there are significant changes in H3K9me3 patterns on specific chromosomal regions (McManus et al., 2006). This study identified a mitosis-specific tri-methylation of H3K9 in pericentromeric heterochromatin that functions in the faithful segregation of chromosomes (McManus et al., 2006).

As mentioned earlier, recent findings associate methylation of H3K27 with gene silencing and X-chromosome inactivation (Shilatifard, 2006). A specific case for silencing is the repression of certain developmental genes by polycomb proteins with the help of this epigenetic modification (Cao et al., 2002), and this repression mechanism is implicated in preserving the pluripotency of stem cells. Although the silencing mechanism is not yet clear, H3K27me3 facilitates the binding of polycomb protein, part of the silencing complex, to histone H3 (Min et al., 2003) in Drosophila. Another study showed that decreasing levels of H3K27me3 on the promoter regions of certain silenced genes is associated with their activation which may contribute to oncogenesis depending on the functional nature of erroneously activated silenced genes (Cha et al., 2005). X-chromosome inactivation in Drosophila females has also been linked to tri-methylation of K27 of the histone H3 (Plath et al., 2003). The inactive X chromosome is enriched with H3K27me3 along with other known markers such as DNA methylation or histone hypoacetylation.

All these findings suggests a possible silencing role for this modification in various organisms, yet determining its function in humans still demands further experimental studies.

Histone methylation was proposed as a stable epigenetic marker, since it was thought to be irreversible (Waterborg, 1993). The discovery of various histone demethylases

(HDMs) though and findings supporting a dynamic reversible methylation pattern in specific phases of the cell cycle have cast doubt (McManus et al., 2006). Still histone methylation has a better potential than the other modification such as phosphorylation or acetylation to be a stable epigenetic marker, since its turnover rate in mammalian cells is much lower (Trojer and Reinberg, 2006). Especially the fact that histone methylation plays a role in heterochromatin formation and/or maintenance lend further support to this claim. However, we still do not know how histone methylation patterns will be passed onto progenitor cells.

AOF2 (amine oxidase (flaving containing) domain 2, LSD1) demethylases mono- and di-methylated K4 of histone H3 (Shi et al., 2004). AOF2 is found within repressor complexes, which include HDAC1/2, coREST and its associated factors. Also it has been shown that hyperacetylated histone H3 is less susceptible to AOF2-mediated demethylation, suggesting that hypoacetylated histone H3 are the potential physiological substrates (Shi et al., 2005).

The members of protein family JMJD2 remove a methyl group from methylated H3K9 and H3K36 in humans, they also function as histone deacetylases and transcriptional co-repressors (Klose et al., 2006; Whetstine et al., 2006). JMJD2 family contains a domain called jumonji, which is responsible for demetyhlation activity. One member of this family, namely JMJD2C, removes a methyl group from di or tri-methylated K9 of histone H3 (Cloos et al., 2006). At present, there are no other known demethylases in human, although most of enzymes that can remove methyl group from mono-, di- and tri-methylated histones are known in higher eukaryotes (Trojer and Reinberg, 2006).

There is no enzyme known to remove a methyl group from a H3K4me3, which marks actively transcribed genes. Since most genes cannot be on all the time, there should exist either, an enzymatic machinery to remove methyl groups from tri-methylated histones that we are not yet aware of, or more epigenetic and/or cellular signals to turn a gene on or off.

1.5 Chromosome 20

Chromosome 20 (HCHR20) is a metacentric chromosome and represents ~1.82% of the human genome. It contains 553 protein coding (476 known and 77 novel) genes, 251 (19 known, 76 putative and 156 novel) processed transcripts and 181 (164 processed and 12 unprocessed) pseudogenes according to Vega Genome Browser version 19 (see also (Deloukas et al., 2001)). It has a gene density of 8.86 per Mb, which is an intermediate between the lowest gene density of chromosome 18 (4.4 genes per Mb) (Nusbaum et al., 2005) and highest of chromosome 19 (26.9 genes per Mb) (Grimwood et al., 2004). HCHR20 is reported to be linked to 236 disorders according to the Online Mendelian Inheritance in Man Database (OMIM, (McKusick, 1998) including Creutzfeldt-Jakob disease (mutations in the *PRNP*) and severe combined immunodeficiency (*ADA*), the Alagille (mutations in the *JAG1*) but also multi-factorial diseases such as type 2 diabetes (T2D), obesity, cataract and asthma (reviewed in Deloukas et al., 2001). T2D and obesity have been linked to 20q12-13.2 region. This region also harbors a commonly deleted region (CDR) found in patients with myeloproliferative disorders and myelodisplastic syndromes (Bench et al., 2000). Additionally, a recent study reported a strong association between increased copy number of 20q13.2 and advanced tumour stage (Dimova et al., 2005) in ovarian cancers.

Figure 1.22 shows the gene density along the HCHR20, which peaks at 20q13.1 cytoband. Due to its medical interest, 20q12-13.2 has been selected as a test region by the Deloukas lab at the Sanger Institute to establish and optimise molecular tools for structural and functional annotation of genomic sequences. A detailed transcription map of this region was produced by both experimental and *in silico* approaches including systematic human:mouse comparative analysis (Stavrides, 2002). This thesis describes work in the same region, 20q12-13.2, aimed at identifying and characterising promoter and other regulatory elements as well as assessing existing experimental and computational approaches.

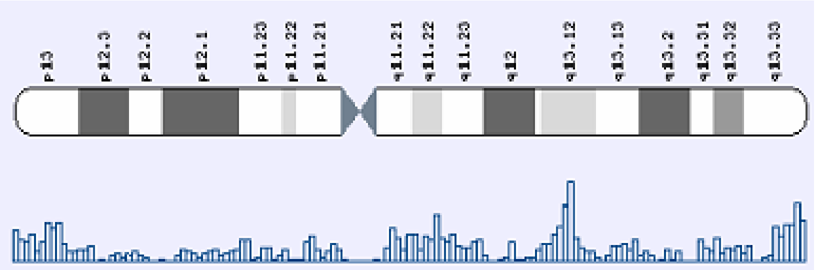


Figure 1.22. Cytoband view of chromosome 20 on the right and the gene number histogram along the chromosome on the left.

1.5.1 Zooming into 20q12-13.2

The 20q12-q13.2 region spans 10,099,058 bp; it starts at 38,106,381 bp and ends at 48,205,439 bp of HCHR20 according to NCBI 36 assembly. All genomic coordinates given within this study are NCBI 36. In bacterial clone sequence coordinates of the region, it starts at the first base of AL009050 and ends at the 113,589th base of AL034423. The region codes for 103 protein coding genes where the gene names and summary of their function can be found in Table A1 (Appendix A). The gene distribution along the region is not homogenous (Table 1.8).

ADA and *HNF4A,* which are associated with SCID (several combined immunodeficiency disease) and monogenic autosomal dominant non-insulin-dependent type I diabetes, respectively, map to the region. There are 14 members of the whey-acidic protein (WAP) domain family where each member contains eight characteristically spaced cysteine residues forming four di-sulphide bonds to perform their possible protease inhibition functions. The region also contains five members of solute carrier transporters involved in transportation of different ion and metabolites across membranes and seven genes encoding transcription factors with a zinc-finger motif. The protein products of 31 genes in the region has not yet been assigned for a function yet.

The promoter regions of *ADA*, *PTGIS*, *PPGB*, *MYBL2*, *TOP1*, *PI3* and *HNF4A* have been investigated by several groups (Berkvens et al., 1987; Yokoyama et al., 1996; Xie and Bikle, 1997; Sala et al., 1999; Keller et al., 2002; Bagwell et al., 2004; Chowdhury et al., 2006), including distal regulatory region of *ADA* (Aronow et al., 1989). However, no systematic large-scale studies aiming for regulatory regions have been undertaken so far.

The region also harbours 63 processed transcripts and 37 (36 processed and one unprocessed) pseudogenes. Processed transcripts are those that are identical or homologous to cDNAs or splice ESTs from the same species or proteins from all species but no unambiguous open reading frame can be assigned to them (Ashurst and Collins, 2003). Pseudogenes are homologous to known genes and proteins but with a disrupted ORF (Ashurst and Collins, 2003). Processed pseudogenes that lack introns and are thought to arise from reverse transcription of mRNA followed by reinsertion of DNA into the genome whereas unprocessed pseudogenes that can contain introns as they are produced by gene duplication (Ashurst et al., 2005).

MAFB TOP1	PLCG1 ZHX3 LPIN3 EMILIN3 CHD6	PTPRT	SFRS6 L3MBTL SGK2 IFT52 MYBL2 FAM112A C20ORF100	JPH2 C20ORF111 GDAP1L1 C20ORF142 R3HDML HNF4A C20ORF62 C20ORF121 SERINC3 PKIG ADA WISP2 KCNK15 RIMS4 YWHAB C20ORF119 TOMM34 STK4	KCNS1 WFDC5 WFDC12 PI3 SEMG1 SEMG2 SLPI MATN4 RBPSUHL SDC4 DBNDD2 C20ORF10 PIGT WFDC2 SPINT3 WFDC6 SPINLW1 WFDC8 WFDC10A WFDC9 WFDC11 WFDC13 WFDC10B SPINT4 C20ORF168 WFDC3 DNTTIP1 UBE2C TNNC2 SNX21 ACOT8 ZSWIM3 ZSWIM1 C20ORF165 PPGB NEURL2 PLTP C20ORF67 ZNF335 MMP9 SLC12A5	NCOA5 CD40 CDH22 SLC35C2 ELMO2 C20ORF157 ZNF334 C20ORF123 SLC13A3 TP53RK SLC2A10 EYA2	PRKCBP1 NCOA3 SULF2	PREX1 ARFGEF2 CSE1L	STAU1 DDX27 ZNFX1 KCNB1 PTGIS B4GALT5 SLC9A8 SPATA2 ZNF313 SNAI1 UBE2V1
38.1 MB	**39.1 MB**	**40.1 MB**	**41.1 MB**	**42.1 MB**	**43.1 MB**	**44.1 MB**	**45.1 MB**	**46.1 MB**	**47.1 MB**

Table 1.8. Gene Distribution of human chromosome 20q12-13.2. Human Genome Organization (HUGO) nomenclature are used for all genes.

2 MATERIALS AND METHODS

In this section, all experiments employed in this study were explained in detail. Recipes for the solutions used in this study can be found in section 2.8. Primers for all the candidate promoters can be found in Appendix B.

2.1 Gene Reporter Assays

General scheme of this experiment is as summarized in Figure 2.1.

2.1.1 Primer Design

Primers to amplify candidate promoter regions were designed using the Primer3 program (Rozen and Skaletsky, 2000). Jilur Ghori at the Wellcome Trust Sanger Institute (WTSI) kindly wrote a program which adds different restriction enzyme sites at the end of each primer for directional cloning of the amplicons. Restriction enzyme pairs used for the inserts are;

- Sac I and BamH I
- BamH I and Nhe I
- Sac I and Hind III
- Hind III and Nhe I

Primers were synthesised in house or ordered via Sigma

(http://www.genxy.com/index.html). Each primer is provided with its optical density measurement at 260 nm (OD_{260}). The average molecular weight of each nucleotide is assumed as 330 ng. For a single stranded primer with length N, working dilutions were calculated according to the formula given below;

Volume of each primer (µl) = (Total amount of primer required (ng))/(330xNxOD_{260})

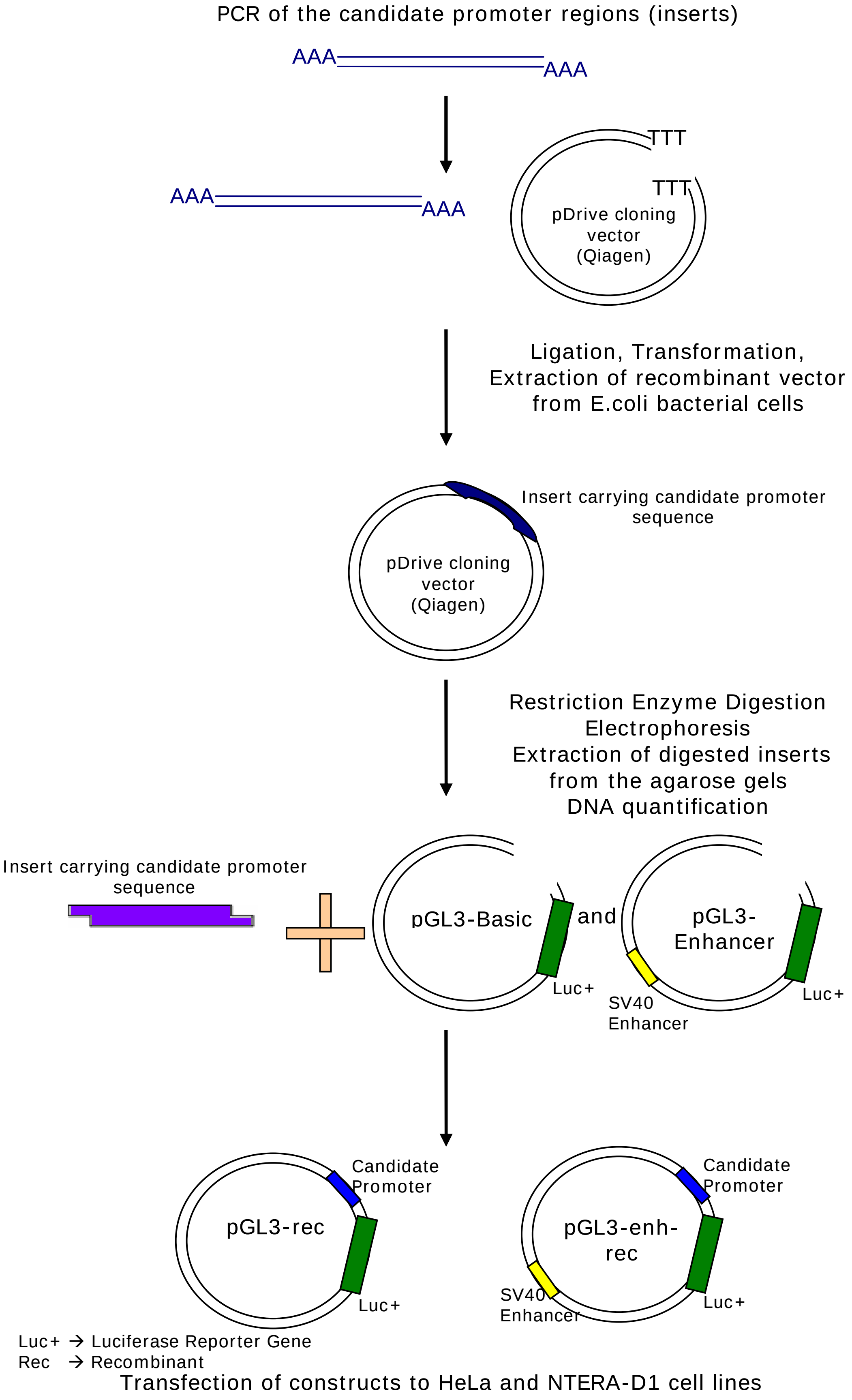

PCR of the candidate promoter regions (inserts)
AAA
AAA
TTT
TTT
AAA
AAA
pDrive cloning vector (Qiagen)
Ligation, Transformation, Extraction of recombinant vector from E.coli bacterial cells
Insert carrying candidate promoter sequence
pDrive cloning vector (Qiagen)
Restriction Enzyme Digestion
Electrophoresis
Extraction of digested inserts from the agarose gels
DNA quantification
Insert carrying candidate promoter sequence
pGL3-Basic
Luc+
and
pGL3-Enhancer
SV40 Enhancer
Luc+
pGL3-rec
Candidate Promoter
Luc+
pGL3-enh-rec
Candidate Promoter
SV40 Enhancer
Luc+
Luc+ → Luciferase Reporter Gene
Rec → Recombinant
Transfection of constructs to HeLa and NTERA-D1 cell lines

Figure 2.1 General Scheme of Cloning Procedure

Primers were stored at 4 ºC for short-term (less than one month) or –20°C for long term and working dilutions were prepared at 100 ng/µl for each primer with $T_{0.1}E$.

2.1.2 Sub-cloning

2.1.2.1 Polymerase Chain Reaction (PCR)

BD Titanium Taq Polymerase (Clontech, #639208) was used for all PCRs. The recipe for the reaction is given in Table 2.1;

Reagents	Amount (µl)
DD water	0
Sucrose Creosol Red	5
PCR buffer (10X)	1.5
10 mM dNTPs	0.6
Primer Mix (100 ng/µl each)	1.5
Human Genomic Template (30 ng/µl)	2.38
5 M Betaine (Sigma, #B0300)	3.9
Titanium Taq Polymerase (5 U/µl)	0.12
Total Volume	15

Table 2.1 Recipe for the PCR for amplification of candidate promoters.

CEPH sample NA12149A was used as the human genomic template to amplify sequences of interest.

The following PCR conditions were used for all reactions

1. 95 ºC for 2 min (Enzyme Activation)
2. 95 ºC for 30 sec (Melting)
3. 60 ºC for 30 sec (Annealing)
4. 72 ºC for 50 sec (Extension)

Repeat step from 2 to 4 for 44 times

5. 72 ºC for 10 min (for extra addition of A-tail to products).

2.1.2.2 Electrophoresis

Electrophoresis was carried out in 1% agarose gel made in 1XTBE buffer with ethidium bromide (250 ng/µl). PCR reactions were loaded directly on the gels and run on 200 volts for 50 min. 100 bp DNA ladder (Invitrogen, #15628) was also loaded onto the gel as a marker. DNA was visualised using a UV transilluminator and photographed digitally using LabWorks Image Acquisition and Analysis Software (UVP Bioimaging Systems).

2.1.2.3 TA cloning

pDrive cloning vector (Qiagen, #231122) was used for sub-cloning purposes. The map of this vector is shown in Figure 2.2.

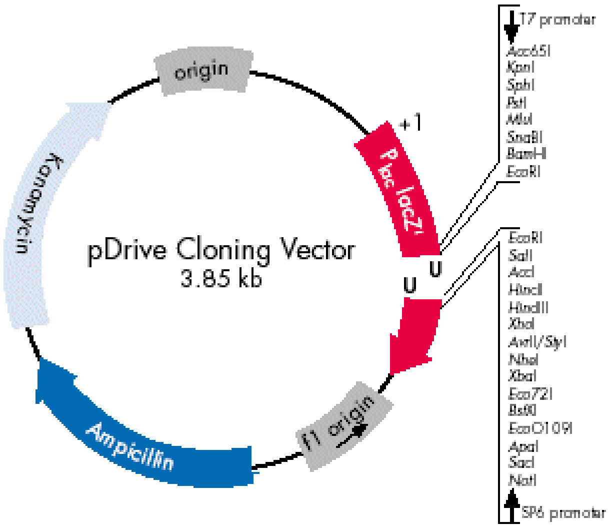


Figure 2.2 Map of pDrive TA cloning vector

T4 DNA Ligase kit (Roche, #481220) was used for ligations. The recipe for ligation reaction is as follows;

Reagents	Amount (µl)
PCR reaction containing the insert	1
pDrive vector (5ng/µl)	1
T4 Ligation Buffer (10X)	1
T4 Ligase (1 U/µl)	0.2
DD water	6.8

Table 2.2 Recipe for Ligation reaction for TA cloning

All ligation reactions were incubated overnight at 4 ºC.

2.1.2.4 Transformation

25 µl of JM109 E.coli competent cells (Promega, #L2001) were incubated with 2 µl of ligation reaction on ice for 10 min. The cells were then incubated at 42 ºC for 50

sec and put on ice immediately for 2 min. 475 µl of SOC medium were added to the cells and incubated for 2 h at 37 ºC on a shaker at 300 rpm. Cells were spun at 4000 rpm (1600xg) for 5 min and resuspended in 100 µl of SOC medium. All of the transformed cells were spread onto LB-agar plates containing 50 µg/ml ampicillin and incubated for 14-16 h at 37 ºC. Then, 3 colonies from each plate were picked to check for the presence of the recombinant vector by colony PCR.

2.1.2.5 Colony PCR

The colony of interest was picked in 100 µl of DD water and diluted in 1:3 with DD water. It is heated to 95 ºC for 5 min for colony lysis. 1 µl of the colony pick solution was used for PCR. AmpliTaq Gold (Applied Biosciences, #N808055) DNA polymerase was used for the amplification whose recipe is given in Table 2.3.

Reagents	Amount (µl)
DD water	0
Sucrose Creosol Red	5
NEB PCR buffer (10X)	1.5
10 mM dNTPs	0.6
Primer Mix (100 ng/µl each)	1.5
Colony Lysate	2.38
5 M Betaine (Sigma, B0300)	2
AmpliTaq Gold Polymerase (1 U/µl)	0.12
Total Volume	15

Table 2.3 Recipe for colony PCR

PCR amplification steps were as follows;

1. 95 ºC for 1 min
2. 95 ºC for 20 sec
3. 60 ºC for 30 sec
4. 72 ºC for 50 sec

Repeat step from 2 to 4 for 34 times

PCR reactions were then loaded onto 2% agarose gel and run for 50 min at 100 V to check the insert amplification. Colonies which carries the corresponding recombinant vector were picked and inoculated in 10 ml of LB growth medium and incubated at 37 ºC on a shaker for about 16 h. 1 ml of bacterial culture is mixed with 500 µl of 50 % glycerol and stored at -70 ºC. 3 ml of bacterial culture was used for the extraction of the recombinant vector according to the mini-prep protocol described in section 2.1.2.6.

2.1.2.6 Mini-prep protocol using Qiaprep Spin Miniprep Kit

The mini-preps were prepared using QIAprep Spin Miniprep Kit (QIAGEN, #27104) and all centrifugations were performed using a standard table-top micro-centrifuge.

2.1.2.7 Restriction Enzyme Digestion

pDrive recombinant vectors were digested with appropriate restriction enzymes to cut the insert (section 2.1.1). All enzymes are ordered from NEB Life Sciences. The recipe for the digestion reaction is shown in Table 2.4. Promega Buffer E was used for all reactions.

Reagents	Amount (µl)	Final amounts
Restriction Enzyme I	0.5-2	10 Units
Restriction Enzyme II	0.5-2	10 Units
Promega Buffer E (10X)	4	1X
BSA	0.4	0.1 µg/µl
pDrive recombinant vector	30	~ 5 µg
DD water	up to 40	n/a
Total Volume	40	n/a

Table 2.4 Recipe for Restriction Enzyme Digestion

The reactions were incubated at 37 ºC for 3 h. Then, 8 µl of Sucrose/Creosol Red was added to the digests. They were then loaded to 1 % agarose gel and run for 3 h at 100 V. The inserts were cut from the gel and extracted from the gel using QIAquick Gel Extraction Kit (Qiagen, #28704) according to protocol described below.

2.1.2.8 Extraction of DNA from Agarose Gel

All centrifugation steps were performed at 13,000 rpm on a Eppendorf desktop microfuge and all buffers were provided with the kit. The extraction procedure is given in step by step fashion below;

- Add 450 µl of Buffer QG
- Incubate @ 50 ºC for 10 min or until the gel slice has completely dissolved
- Check that the colour of the mixture is still yellow (If it is orange or violet, add 10 µl 3 M sodium acetate at pH 5.2 and mix)
- Apply the sample to the QIAquick column and centrifuge for 1 min. Discard flow-through
- Add 500 µl of Buffer QG and centrifuge for 1 min. Discard flow-through.
- Pipette out the remaining drops of yellow liquid from the top of the column.
- To wash, add 750 µl Buffer PE and centrifuge for 1 min. Discard flow-through
- Add 750 µl Buffer PE, leave for 10 min and centrifuge for 1 min
- Discard flow-through and centrifuge for 1 min again
- Place the column into a clean 1.5 ml microfuge tube
- To elute DNA, add 30 µl Buffer EB to the centre of the column’s membrane, let it stand for 15-30 min. Centrifuge for 1 min.

2.1.3 Cloning inserts into Gene Reporter Vectors

2.1.3.1 Gene Reporter Vectors

pGL3 vectors from Promega were used as reporter vectors. Candidate promoters were cloned into pGL3-basic (#E1751) and pGL3-enhancer (#E1771) vectors. pGL3-promoter (#E1761) and pGL3-control vector (#E1741) were used as positive controls. pRL-SV40 vector (#E2231) was used as the internal control vector. The maps of these vectors can be found in the figures from 2.3 to 2.7.

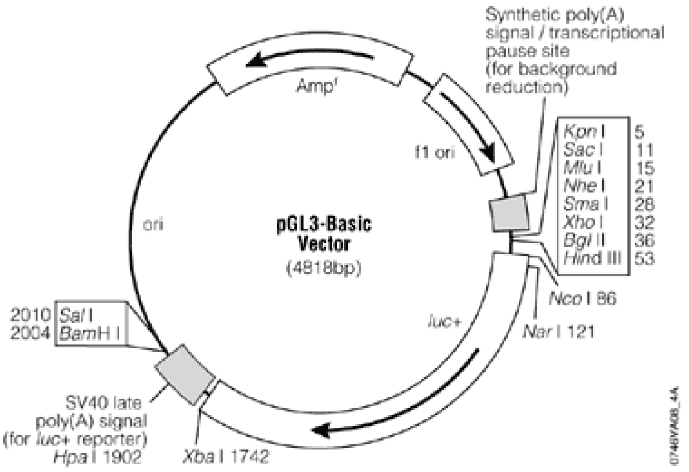


Figure 2.3 pGL3-basic vector map

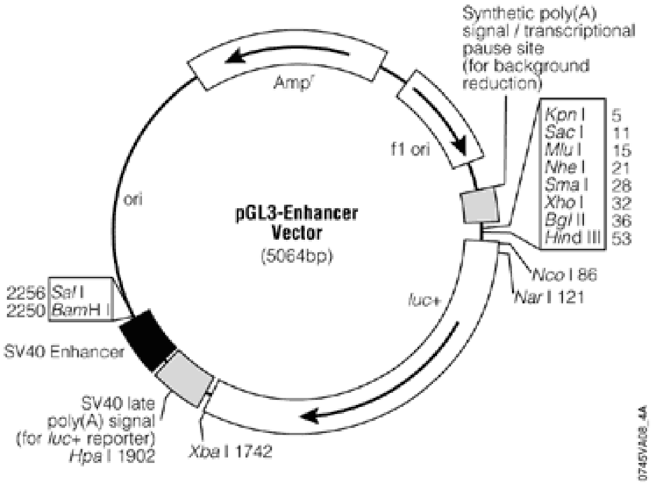


Figure 2.4 pGL3-enhancer vector map

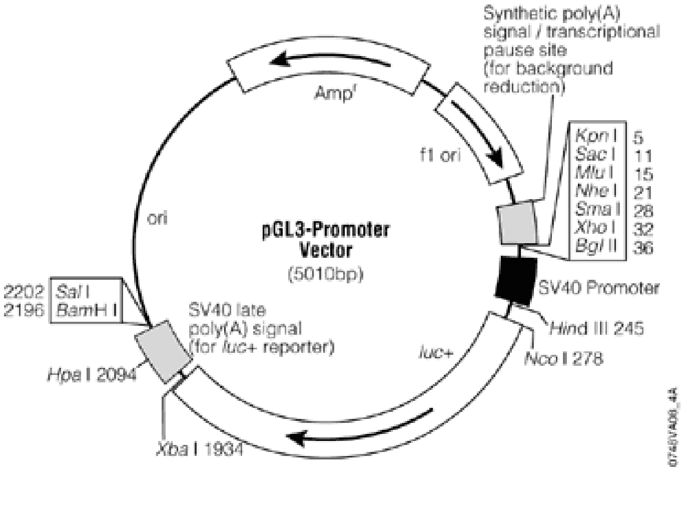


Figure 2.5 pGL3-promoter vector map

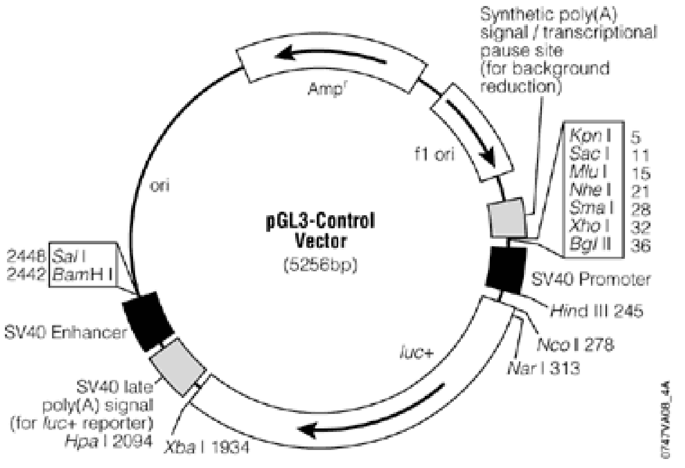


Figure 2.6 pGL3-control vector map

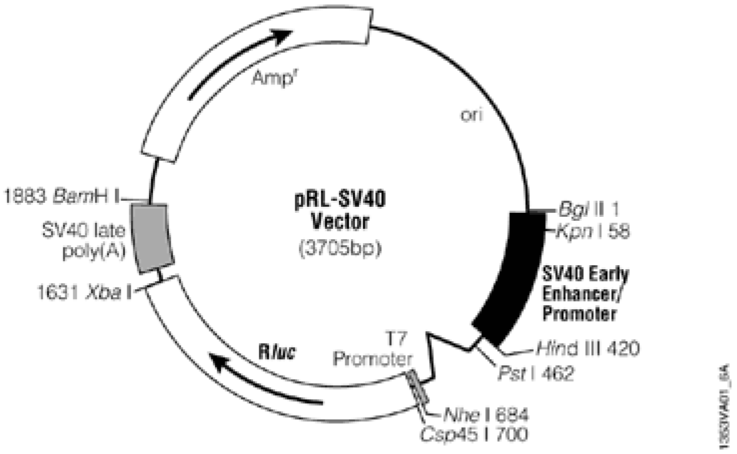


Figure 2.7 PRL-SV40 vector map

2.1.3.2 Preparation of the cloning vectors

pGL3-basic and pGL3-enhancer vectors were digested with the appropriate enzymes (Table 2.5) sequentially.

Insert Digestion	Corresponding vector digestion
Bam HI and Nhe I	Bgl II and Nhe I
Sac I and Bam HI	Sac I and BamH I
Sac I and Hind III	Sac I and Hind III
Hind III and Nhe I	Hind III and Nhe I

Table 2.5 Restriction Enzyme Pair for the digestion of vectors

The digestion reaction set up is shown in Table 2.6. The reaction was incubated at 37 ºC for 3 h.

Reagents	Amount (μl)	Final amounts
Restriction Enzyme	0.5-2	10 Units
NEB Buffer (10X)	10	1X
BSA	1	0.1 μg/μl
Vector	variable	10 μg
DD water	up to 100	n/a
Total Volume	100	n/a

Table 2.6 Restriction enzyme digestions of cloning vectors

The digest is ethanol-precipitated using Pellet Paint (Novagen, #69049) as follows; 2 μl of Pellet Paint, 0.1 volume of 3 M sodium acetate at pH 5.2 and 2 volumes of absolute ethanol were added to the sample, mixed and incubated for 2 min at room temperature. Samples were spun at 14,000 rpm on a desktop microfuge for 5 min and washed with 500 μl of 70% ethanol by spinning 5 min at 13,000 rpm. Pellets were air-dried for 2 min and resuspended in 30 μl $T_{0.1}E$.

Samples were then digested with the second enzyme as described above, ethanol precipitated using Pellet Paint, and resuspended in 30 μl $T_{0.1}E$. 10 μl of Sucrose/Creosol Red was added to samples which were loaded to 1 % agarose gel and run for 6 h at 50 V to eliminate undigested vector band. The cut vector band was extracted from the gel using QIAquick gel extraction kit as described in 2.1.2.8. The digested vectors were quantified using NanoDrop spectrophotometer (NanoDrop Technologies) and dephosphorylated as described below.

Shrimp Alkaline Phosphatase (USB, #70092Y) was used for dephosphorylation of double digested vectors. For every 0.62 μg of pGL3-basic and 0.59 μg of pGL3-enhancer vectors, 1 unit of SAP was used to dephosphorylate 1 pmol of DNA termini. It is assumed that 1.0 μg of 3 kb plasmid contains 1.0 pmole of DNA termini. Reaction set up was listed in Table 2.7.

Reagents	Amount (µl)
Double digested vector	30
10X Reaction Buffer	1
SAP	variable (1 U per 1 picomole of DNA termini)
DD water	up to 50

Table 2.7 Reaction set up for vector dephosphorylation

Reactions were incubated at 37 ºC for 1 hour. The enzyme was heat-inactivated at 65 ºC for 15 min. The dephosphorylated vector was purified using phenol/chloroform three times as described in section 2.1.3.3. Then, 0.1 volumes of 3 M Sodium Acetate at pH 5.2 and 2 volumes of absolute ethanol were added to the purified sample and incubated for 1 hour at -70 ºC. The sample was spun at 13,000 rpm at 4 ºC for 15 min to precipitate DNA. Then, it was washed with 500 µl of 70 % Ethanol by centrifuging for 5 min at 13,000 rpm. The pellet is air-dried for 10 min and resuspended in 30 µl of $T_{0.1}E$.

2.1.3.3 Phenol – chloroform purification of DNA

Equal volume of phenol:chloroform:isoamyl alcohol (25:4:1) (Invitrogen, #15593031) was added to the sample, mixed well and spun at 13,000 rpm on a microfuge for 5 min. The aqueous layer (bottom layer) was then transferred to a new tube.

2.1.4 Ligations

Roche Rapid DNA Ligation Kit (Roche Applied Sciences, #11635379001) was used to generate recombinants vectors carrying candidate promoters. 10 ng of vector (pGL3-basic or pGL3-enhancer) was used for each ligation reaction. Amount of insert required to set up ligation reaction was calculated according to equation below;

$$\textbf{Amount of insert} = \frac{\textbf{Size of insert (bp)}}{\textbf{Size of vector (bp)}} \times (\textbf{Amount of vector}) \times \textbf{3}$$

Insert and vector were mixed and the volume is completed up to 10 µl by 1X DNA dilution buffer. 10 µl of 2X DNA ligation buffer and 1 µl of (1 U) of T4 DNA ligase were added and the reaction was incubated for 30 min at room temperature. 2 µl of the ligation reaction was used to transform E. coli cells according to protocol described in section 2.1.2.4. On the next day, 3 colonies from each plate were checked by colony-PCR using the protocol described in section 2.1.2.5. Colonies which carries the recombinant vector were picked and inoculated in 10 ml of LB growth medium and incubated at 37 ºC on a shaker for about 16 h. For long-term storage, 1 ml of bacterial culture is mixed with 500 µl of 50 % glycerol and stored at -70 ºC. For plasmid extraction, 2 ml of bacterial culture was used and extraction was performed according to the mini-prep protocol described in section 2.1.2.6. The mini-preps were quantified using NanoDrop spectrophotometer.

2.1.5 Transfections

Human adherent cell lines (HeLa S3 and NTERA-D1) were transfected using QIAGEN Effectene Transfection Reagent (QIAGEN, #301425) with recombinant reporter vectors. The transfection was performed in 48-well plate format, with the reagents provided with the kit, according to the protocol described below;

- The day before transfection, seed $2x10^4$ cells per well in a 48-well plate (Corning, #3548) in 200 µl appropriate growth medium containing serum and antibiotics.
- Incubate the cells under their normal growth conditions (at 37 °C and 5% CO_2). The wells should be 40–80% confluent on the day of transfection.
- The day of transfection, dilute 150 ng DNA and 50 ng of internal control vector DNA (pRLSV-40) with the DNA-condensation buffer, Buffer EC, to a total volume of 50 µl. Add 1.2 µl Enhancer and incubate at room temperature for 2-5 min.
- Dilute 4 µl of Effectene Reagent in 50 µl of Buffer EC and add to the DNA-Enhancer mixture. Mix by pipetting up and down 5 times
- Incubate the samples for 5–10 min at room temperature to allow transfection-complex formation.
- While complex formation takes place, gently aspirate the growth medium from the plate, and wash cells once with 200 µl PBS. Add 100 µl fresh growth medium to the cells.
- Add 150 µl growth medium to the wells containing the transfection complexes. Mix by pipetting up and down twice, and immediately add the

transfection complexes drop-wise onto the cells in the 48-well dishes. Gently swirl the dish to ensure uniform distribution of the transfection complexes.

- Incubate the cells with the transfection complexes under their normal growth conditions for 48 h for expression of the transfected gene. Assay the cells for expression of the transfected gene using a luminometer.

2.1.6 Dual Luciferase Reporter Assays

A dual luciferase assay (Promega, #E1910) was used to measure the luciferase and renilla activity within the transfected cells. Two days after transfection, growth media was removed and cells were washed with 100 µl of 1XPBS. Then, cells were incubated in 100 µl of 1X Passive Cell Lysis Buffer for 20 min at room temperature for cell lysis. 20 µl of the lysate were transferred to luminometer plates (Greiner Bio-One Inc, #655083) and they are ready for assay. Luciferase assays were performed with MicroLumat Plus LB 96V luminometer (Berthold Technologies). First, 50 µl of Luciferase Reagent II was injected to the lysate and luciferase activity was measured for 10 sec. Then, 50 µl of *Stop and Glo*® reagent was injected to the lysate to stop the luciferase activity and catalyse the renilla reaction, incubated for 1.6 sec and then renilla activity was measured for 10 sec.

2.2 Cell Culture

2.2.1 HeLa S3 cell line

HeLa S3, human cervical carcinoma cell line, was kindly provided by Team 76 (Cancer Genome Project) at the Sanger Institute. This adherent cell line was grown in high glucose (5 g/l) DMEM (Gibco, #41966029) supplemented by 10 % FBS (Gibco, #10108165) and 1X antibiotic/antimycotic solution (Gibco, #10378016) in an humidified incubator at 37 ºC and 5 % CO_2. Cells were spun at 1200 rpm (250 g) for 5 min and resuspended in appropriate amount of complete growth media. Cell were trypsinized every 3-4 days (section 2.2.3) and $4x10^4$ cells were seeded per cm^2 of flask area. Doubling time of this cell line is ~ 1.5 days.

2.2.2 NTERA-2 clone D1 cell line

NTERA-2 clone D1 (NTERA-D1), human Caucasian pluripotent embryonal carcinoma cell line was purchased from European Collection of Cell Cultures (ECAC) (ECAC, #93021013). This adherent cell line was grown in high glucose (5 g/l) DMEM (Gibco, #41966029) supplemented by 10 % FBS and 1X antibiotic/antimycotic solution in an humidified incubator at 37 ºC and 5 % CO_2. Cells were spun at 1200 rpm (250 g) for 5 min and resuspended in appropriate amount of complete growth media. Cells were trypsinized every 3-4 days (section 2.2.3)and $8x10^4$ cells seeded per cm^2 of flask area and Doubling time of this cell line is about 3 days.

2.2.3 Trypsinizing Cells

When the cells were 90-100% confluent, they were taken from the incubator, the growth media was discarded and appropriate amount of 1XPBS (3 ml per 25 cm^2 flask area) was added onto cells to remove the residual of FBS. Cells were washed twice by 1XPBS by rocking the flasks for 5 sec and the solution was discarded. Then, appropriate amount of Trypsin/EDTA solution (Gibco, #15240096) (2 ml per 25 cm^2 flask area) was added and the cells were incubated for 5-6 min in the 37 ºC incubator for detachment. Then 1 ml of complete growth media was added to trypsinized cells to inhibit trypsin. Cell were counted and seeded at the required density to the cell flasks or plates.

2.2.4 Freezing the Cells

The cells which are in their early passage (at most 3 passages) were trypsinized and counted. Then, 2 million cells was resuspended in 1 ml of 95 % FBS/5% DMSO solution and put in 1 ml cryogenic vials and stored at -70 ºC for one day. Vials were transferred to the liquid nitrogen tank (-180 ºC) the next day for long-term storage.

2.2.5 Thawing the Cells

Cells were taken from the liquid nitrogen tank and rapidly thaw in a water-bath at 37 ºC. Cells were transferred to a 15 ml eppendorf tube and 7 ml of complete growth media was added onto cells. Then, cells were spun at 1200 rpm (250 g) for 5 min, the media was removed and cells were resuspended in 5 ml of 1XPBS completely and spun again at 1200 rpm (250 g) for 5 min. 1XPBS was discarded and cells were resuspended in 10 ml of complete growth media and seeded onto 25 cm^2 flasks.

2.3 Affymetrix Expression Arrays

2.3.1 Isolation of Total RNA from cells

RNeasy^{TM} Mini Kit (QIAGEN, #74104) was used to isolate total RNA from HeLa S3 and NTERA-2 clone-D1 cells according to the manufacturer's instructions. The total RNA yields and quality from both cell lines are given in Table 2.8 below.

Cell Type	Number of cells	Amount of total RNA (µg)	A_{260}/A_{280}	A_{260}/A_{230}
HeLa S3	$4x10^6$	32.2	2.12	2.10
NTERA-2 clone D1	$4x10^6$	61.5	2.13	2.17

Table 2.8 Total RNA yields and quality from both cell lines

2.3.2 Checking the quality of the RNA

The quality of RNA samples were checked by running on 1 % agarose gel in 1X MOPS (3-(N-morpholino) propanesulfonic acid) buffer. Agarose gel was prepared as described here; 0.4 g of agarose was weighed, 8 ml of 5XMOPS buffer and DEPC-treated water up to 40 ml was added. The mixture was boiled in the microwave oven until melted. The solution was cooled for 2-3 min before adding 1 ml of 38% FA in a fume hood. A small gel tank was cleaned with RNaseZap® (Ambion, #9780) and the gel solution was poured to the tank and let to solidify for 30 min.

2 µl of RNA loading buffer was added to the samples and they are incubated at 65 ºC for 10 min to denature any RNA secondary structure. The samples were put immediately placed on ice for 2 min, short-spun to be ready for loading to the gel. RNA ladder (Promega, #G3191) was used to assess to quality of RNA extraction.

2.3.3 Preparation of RNA for hybridization to Affymetrix Expression Arrays

GeneChip$^{\text{TM}}$ Human Genome U133A 2.0 Expression Analysis Array (Affymetrix, #900466) was used to determine the gene expression profile of both cell lines. One Cycle Target Labelling and Control Reagents Kit (Affymetrix, #900493) which includes all the reagents and materials was used for the preparation of the total RNA to the arrays. The preparation steps of total RNA content of the cells for the hybridization to arrays are schematically represented in Figure 2.8. The critical specifications of GeneChip$^{\text{TM}}$ Human Genome U133A 2.0 Expression Analysis Array are given in Table 2.9.

Critical Specifications of Human Genome U133A 2.0 Expression Analysis Array	
Feature Size	11 µm
Probe Pairs/Sequence	11
Hybridization Controls Poly-A Controls Normalization Controls Housekeeping/Control Genes	bioB, bioC, bioD, cre dap, lys, phe, thr 100 probe sets GAPDH, beta-Actin, ISGF-3 (STAT1)
Array Format	49
Fluidics Protocol	EukGE-WS2v5
Hybridization Volume	200 µl
Library Files	HG-U133 Plus 2.0

Table 2.9 Critical Specifications of Human Genome U133A 2.0 Expression Analysis Array

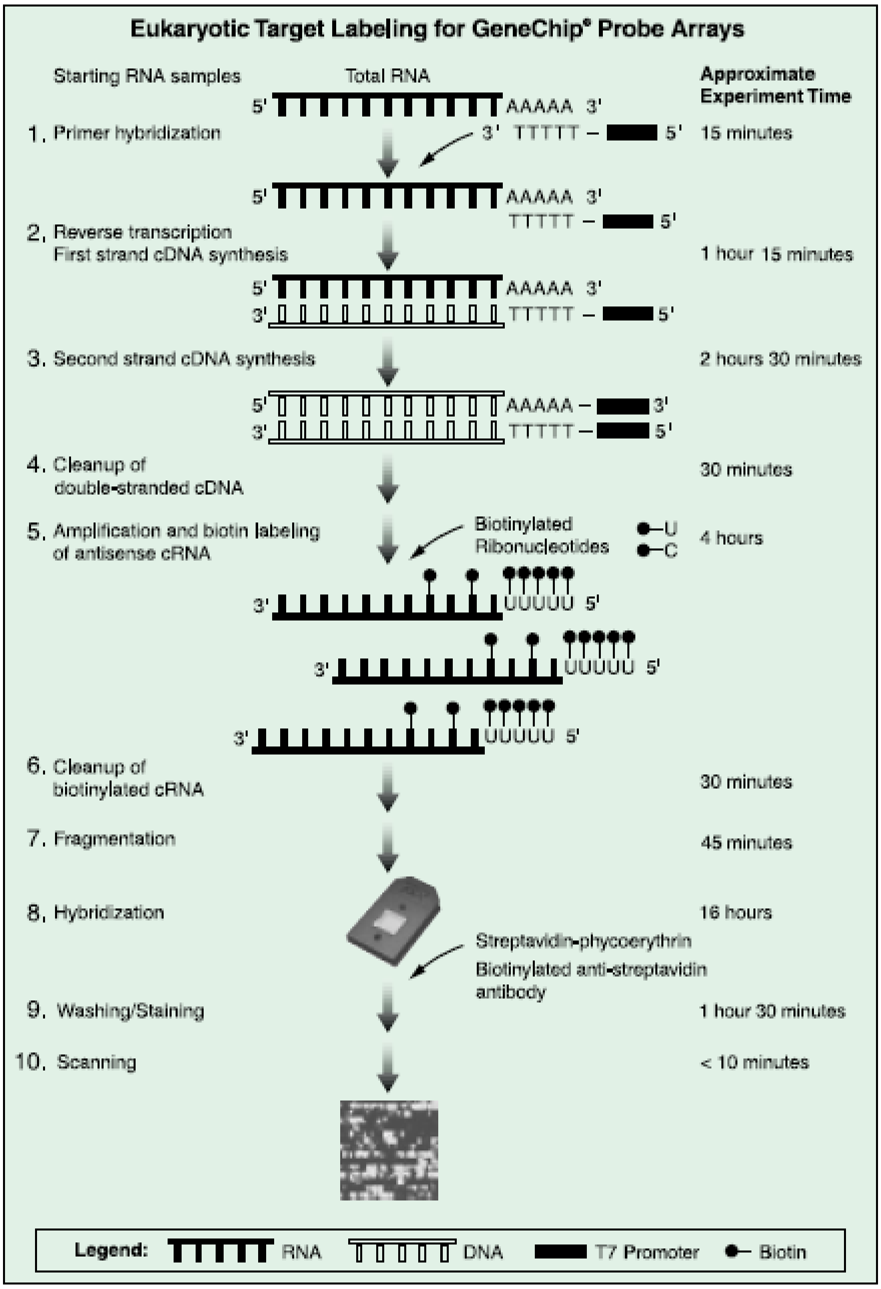

Eukaryotic Target Labeling for GeneChip® Probe Arrays
Starting RNA samples
Total RNA
Approximate Experiment Time
5' AAAAA 3'
3' TTTTT 5'
1. Primer hybridization
15 minutes
2. Reverse transcription First strand cDNA synthesis
1 hour 15 minutes
3. Second strand cDNA synthesis
2 hours 30 minutes
4. Cleanup of double-stranded cDNA
30 minutes
5. Amplification and biotin labeling of antisense cRNA
Biotinylated Ribonucleotides
U
C
4 hours
3' UUUUU 5'
6. Cleanup of biotinylated cRNA
30 minutes
7. Fragmentation
45 minutes
8. Hybridization
16 hours
Streptavidin-phycoerythrin
Biotinylated anti-streptavidin antibody
9. Washing/Staining
1 hour 30 minutes
10. Scanning
< 10 minutes
Legend:
RNA
DNA
T7 Promoter
Biotin

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

2.3.3.1 First Strand cDNA synthesis

Total RNA is first reverse transcribed using SuperScript II reverse transcriptase and T7-Oligo(dT) Promoter Primer in the first strand cDNA synthesis reaction according to the manufacturer’s instructions

(http://www.affymetrix.com/support/technical/manual/expression_manual.affx). For reverse transcription reaction, 5 µg of total RNA was used. In order to monitor the labelling efficiency independent from the quality of the starting total RNA, four *B. subtilis* polyadenylated transcripts (lys, phe, thr, dap) were added to the reaction at known concentrations (set by the manufacturer). Each GeneChiP^{TM} array contains probes for these transcripts and their signal will be accepted as reference to evaluate the labelling efficiency of the sample.

2.3.3.2 Second Strand Synthesis

After the first strand synthesis, RNase H is added to the reaction to digest the RNA from RNA:cDNA hybrid. Then, E. coli DNA ligase and E.coli DNA Polymerase I were added for second strand cDNA synthesis. All reactions were set up according to the manufacturer's instructions

(http://www.affymetrix.com/support/technical/manual/expression_manual.affx).

2.3.3.3 Cleanup of Double Stranded cDNA

Double Stranded cDNA was cleaned up using the cDNA Cleanup Spin Columns according to the manufacturer's instructions

(http://www.affymetrix.com/support/technical/manual/expression_manual.affx).

2.3.3.4 Synthesis of Biotin Labelled cRNA

Double stranded cDNA containing T7-promoter sequence is the template for the synthesis of antisense cRNA in the presence of four natural ribonucleotides and one biotin-conjugated nucleotide analogue for labelling. The reaction was set up according to the manufacturer's instruction

(http://www.affymetrix.com/support/technical/manual/expression_manual.affx). The labelling reaction was incubated for 16 h at 37 ºC.

2.3.3.5 Cleanup of Biotin Labelled cRNA

Biotin labelled cRNA was cleaned up using the cRNA Cleanup Spin Columns according to the manufacturer's instructions

(http://www.affymetrix.com/support/technical/manual/expression_manual.affx).

2.3.4 Fragmentation of Biotin Labelled cRNA

Biotin Labelled cRNA was quantified using NanoDrop spectrophotometer and 20 µg cRNA was break down to 35 to 200 base fragments by metal-induced hydrolysis. For fragmentation, 8 µl of 5X Fragmentation Buffer was added to 20 µg of biotin labelled cRNA and the volume is completed up to 40 µl with RNase-free water. The reaction was incubated for 94 ºC for 35 min and then immediately placed on ice. The fragmentation efficiency was checked by running 2 µl of the reaction on an RNA gel as described in 2.3.2.

2.3.5 Eukaryotic Target Hybridization

The following hybridization cocktail was prepared for 49 Format Array as described in Table 2.10.

Fragmented cRNA	15 µg
Control Oligonucleotide B2 (3 nM)	5 µl
20X Eukaryotic Hybridization Controls (bioB, bioC, bioD,cre)	15 µl
Herring Sperm DNA (10 mg/ml) (Promega D1811)	3 µl
BSA (5 mg/ml) (Invitrogen, #15561-020)	3 µl
2X Hybridization Buffer	150 µl
DMSO (Sigma, #D5879)	30 µl
DD water	to a final volume of 300 µl

Table 2.10 Recipe for the hybridization cocktail

The probe array was equilibrated to room temperature before use. Hybridization cocktail was heated to 99 ºC for 5 min and incubated at 45 ºC for 5 min. Meanwhile, the probe array was wet by filling it with 200 µl of 1X Hybridization Buffer and incubated at 45 ºC for 10 min with rotation. The hybridization cocktail was spun at maximum speed on a table-top centrifuge for 5 min to remove any insoluble material and transferred to a new tube avoiding any insoluble matter at the bottom of the tube. After 10 min of incubation of the probe array at 45 ºC with rotation, the buffer was removed from the array and the probe array was filled with 200 µl of hybridization cocktail and placed in hybridization oven at 45 ºC and incubated for 16 h while rotating 60 rpm.

2.3.6 Washing and Staining of the probe array

After 16 h of hybridization, the hybridization cocktail was removed from the probe array and the probe array is filled with 200 µl of Non-stringent wash buffer A. The probe array was washed and stained using Affymetrix Fluidics Station 450 according to the manufacturer's instructions

(http://www.affymetrix.com/support/downloads/manuals/expression_analysis_technical_manual.pdf). The program named EUkGE-WS2v4_450 was used for washing and staining of the array.

2.3.7 Array Scanning

After the probe array washed and stained, it was scanned using Affymetrix Scanner controlled by Affymetrix Microarray Suite (GCOS). A quality control report was generated for each array scanned to assess the efficiency of the experiment.

2.4 Construction of Tilepath Arrays

In order to generate a high density single nucleotide polymorphism (SNP) map of human chromosome 20, chromosome sequence was sheared into small fragments of size around 2 kb (Spencer et al., 2006). These DNA fragments were then cloned into pUC18 vector, amplified and re-sequenced for finding SNPs, and they were stored in

$T_{0.1}E$ -70 ˚C for further use. For our study, we picked around 1800 of those fragments that will cover 3.5 Mb region on human chromosome 20q12-13.2. These fragments were amplified with pUC18 primers where the forward primer contained an amino group linked to sixth carbon atom at the 5' terminal to enable them for printing onto a microarray. The amino-linked primer were ordered from Operon (www.operon.com).

The PCR amplification protocol was given in Table 2.11. Fragment that are picked were diluted 1:10 with TE buffer for use in PCR.

Reagent	**Amount (µl)**
DNA Template (1:10 diluted)	1
10X CHIP PCR Buffer	6
10 mM dNTPs	3
5' amino linked forward primer (100 ng/µl)	0.75
Reverse primer (100 ng/µl)	0.75
AmpliTaq Gold Polymerase	0.3
DD water	48.2

Table 2.11. PCR recipe for amplification of DNA fragments to be spotted on the array.

PCR amplification steps are as follows;

1. 95 ˚C for 5 min
2. 95 ˚C for 1 min
3. 60 ˚C for 1 min
4. 72 ˚C for 4 min

Repeat steps 2-4 for 30 times

5. 72 ˚C for 5 min

Quality and size of the PCR reactions were checked by running 2 µl of the reactions on 1 % agarose gels with ethidium bromide at 100 volts for 2 h.

Then, PCR reactions (~55 µl) were transferred to filter plates (pore size of 0.65 µm) (Millipore, #MSDV6550), 15 µl of 4X Spotting Buffer were added and centrifuged at 2000 rpm for 10 min into 96-well plates. Centrifuged samples were now ready for spotting onto microarrays or alternatively they can be kept at -20 ˚C until further use. Samples were sent to Microarray Facility at WTSI for printing. Each fragment had three replicates on each array. In total, 162 arrays were printed in 4 batches and less than 5% of the spots were unusable on each batch.

2.5 Chromatin Immunoprecipitation (ChIP)

The experimental protocol is kindly provided by Vetrie Lab at Sanger Institute and some modifications have been introduced. ChIP experiments were performed on HeLa and NTERA-D1 cell lines.

2.5.1 Cell Harvesting

HeLa or NTERA-D1 cell lines were grown in 500 cm^2 cell culture dishes (Corning, #431110) under their normal growth conditions until they reach 100 % confluency. Growth media was removed and cells were washed with 25 ml of 1XPBS to be ready for chromatin fixation.

2.5.2 Chromatin Fixation

Fixation conditions vary depending on the antibody used in further steps of the experiment. Table 2.12 lists the fixation conditions depending on the antibody. Dimethyl Adipimidate.2HCl (DMA) (Pierce, #20660), Disuccinimidyl Glutarate (DSG) (Pierce, cat. no.20593) and Disuccinimidyl Suberate (DSS) (Fluka, #80424) which cross-links the proteins in the cells were used to increase cross-linking efficiency. Formaldehyde (FA) solutions were prepared from 38% stock solution (BDH Analar, #101135B) in serum-free growth media and supplemented by 1 mM $MgCl_2$ to keep cells adherent during fixation.

Antibodies recognising	First Fixation Solution	Duration (min)	Second Fixation Solution	Duration (min)
Modified histones	0.37 % FA	15	-	-
CTCF	1% FA	15		
RNA polymerase II	1 % FA	60	-	-
	10 mM DMA	45	1% FA	15
	10 mM DMA	60	1% FA	15
	2 mM DSS	52.5	1 % FA	15
	2 mM DSG	52.5	1 % FA	15

Table 2.12 Chromatin Fixation Conditions

For fixation, 50 ml of fixation solution (see Table 2.12) was added onto the cells in 500 cm^2 culture dishes and incubated at room temperature on orbital shaker. For the sequential fixation, cells were first incubated with 50 ml of first fixation solution (see Table 2.12) on an orbital shaker. The solution was discarded and 50 ml of 1 % formaldehyde solution was added onto the cells for further 15 min incubation on orbital shaker. Then, 3.3 ml of 2 M Glycine solution was added (0.125 mM of final concentration) to the plates and incubated for 5 min on orbital shaker to stop the fixation reaction. The solution was discarded and cells were washed with 20 ml of 1X PBS by rocking the plate for 5 sec.

One protease tablet (Roche Applied Sciences, # 11697498001) was dissolved in 20 ml of ice-cold 1XPBS. 2 ml of ice-cold 1XPBS supplemented by proteases were added onto plates and cells were scraped from the plates using a universal plate lid. One plate of HeLa cells or two plates of NTERA-D1 cells (~150 million cells in total) were collected to one 15 ml eppendorf tube.

Cells were spun at 2000 rpm (825 g) for 6 min ºC, the supernatant discarded. The cell pellet was fully resuspended in 1.5 ml of ice-cold 1XPBS supplemented by proteases and spun again at 2000 rpm (825 g) for 5 min at 4 ºC. Now, the cells are ready for lysis and chromatin extraction.

2.5.3 Cell Lysis and extraction of chromatin

Cells were fully resuspended in 3 ml of Cell Lysis Buffer by gently pipetting up and down and incubated for 10 min on ice for complete lysis. Then, the lysate was centrifuged down at 3200 rpm for 5 min at 4 ºC to pellet nuclei. The supernatant was carefully discarded and the pellet were fully resuspended in 1.2 ml of Nuclei Lysis Buffer by gently pipetting up and down and incubated for 10 min on ice. After incubation, 720 µl Dilution Buffer, the sample was mixed gently and transferred to a 5 ml falcon tube. Now, the samples are ready for sonication.

2.5.4 Sonication

An MSE Sanyo Soniprep 150 sonicator was used to sonicate the chromatin samples. Sonication was performed to fragment the chromatin between 300 – 600 bp fragments. Higher amplitude settings were used for samples fixed with either 1% formaldehyde or other protein cross-linking chemicals. The sonication conditions were listed in Table 2.13.

Cell Type	**HeLa S3**		**NTERA-D1**	
Fixation Conditions	**0.37% FA**	**1% FA**	**0.37% FA**	**1% FA**
Amplitude (microns)	14	15.2 (maximum)	14	15.2 (maximum)
Number of Bursts	8	8	6	6
Length of bursts (sec)	30	30	30	30

Table 2.13 Sonication Conditions

The sonication probe (Exponential microprobe, 3 mm in diameter) was placed approximately 1-2 cm below the surface of the sample. Sample was kept on ice-water bath at all times during sonication. Also, one min break was allowed between each burst to prevent over-heating of the sample.

After sonication, the lysate were transferred to a 2 ml eppendorf tube and centrifuged for 10 min at 13,000 rpm at 4 ºC to remove insoluble cellular debris. Then, the lysate was transferred to a 15 ml falcon tube and 4.1 ml Dilution Buffer was added to bring the ratio of Nuclei Lysis Buffer to Dilution Buffer to 4:1. At this point, chromatin is ready for immunoprecipitation. For long term storage, the chromatin was flash-frozen by keeping the samples for 5 min in dry ice and stored in -70 ºC until further use.

2.5.5 Antibodies

Antibodies used in this study are listed in Table 2.14.

Company	Cat. No.	Raised for	Modification	Residue	Raised in	Abbreviation
Santa Cruz Biotechnology	sc-15914 (C-20)	CTCF	-	-	Goat	CTCF
Abcam	ab7766	Histone 3	Di-methylated	Lysine 4	Rabbit	H3K4me2
Abcam	ab8580	Histone 3	Tri-methylated	Lysine 4	Rabbit	H3K4me3
Abcam	ab8895	Histone 3	Mono-methylated	Lysine 4	Rabbit	H3K4me
Abcam	ab9045	Histone 3	Di-methylated	Lysine 9	Rabbit	H3K9me2
Abcam	ab7312	Histone 3	Tri-methylated	Lysine 27	Rabbit	H3K27me3
Upstate	06-599	Histone 3	Acetylated	Lysine 9, 14	Rabbit	H3Ac
Upstate	06-866	Histone 4	Acetylated	Lysine 5, 8,12 and 16	Rabbit	H4Ac
Abcam	ab5131	PolII CTD	Phosphorylated	Serine 5	Rabbit	polII

Table 2.14 Antibodies used in this study

2.5.6 Pre-clearing of the Chromatin

To reduce the non-specific binding, chromatin was treated with the isotype of the antibody which will be later used for immunoprecipitation. 100 µg of either Rabbit IgG (Upstate, #12-370) or Goat IgG (Santa Cruz Biotechnology, #sc-3850) was added to the chromatin according to the antibody used for the immunoprecipitation and incubated for 1 hour in the cold room on a rotator. Then, 200 µl of homogenous Protein G Agarose Beads (Roche Applied Sciences, #1719-416) were added to the chromatin and incubated 3 h in the cold room on a rotator. After incubation, the sample was centrifuged at 3000 rpm for 2 min at 4 ºC to pellet the beads. The supernatant is transferred to a new tube and it is ready for immunoprecipitation. 270 µl of pre-cleared chromatin was taken as the input chromatin (complete DNA content of the cells) and stored in -20 ºC until further use.

2.5.7 Immunoprecipitation

For immunoprecipitation, 10 µg of antibody was incubated with 1080 (modified histones and CTCF) or 1350 µl (RNA polymerase II) of chromatin. The negative control antibody (the isotype of the corresponding antibody) was always included in the experiment to determine the background level. All antibody incubations were performed overnight (14-16 h) in the cold room on a rotator.

2.5.8 Addition of Protein G Agarose Beads to Antibody/Chromatin Mixture

Chromatin/Antibody complex was centrifuged at 13,000 rpm for 5 min at 4 ºC on a table-top centrifuge and supernatant was transferred to a new tube. 100 µl of fully resuspended Protein G Agarose Beads were added to each mixture and incubated 3 h in the cold room on a rotator. Then, the samples were centrifuged at 13,000 rpm for 20 sec at 4 ºC to pellet the beads, the supernatant was discarded and the beads are ready for washing.

2.5.9 Washing Antibody/Chromatin/Bead Complexes

This step is necessary to remove non-specific complexes and unbound protein and DNA molecules. The beads were washed with 750 µl of Wash Buffer I twice. For each wash, the sample was vortexed for 3 sec and centrifuged at 7500 rpm for 2 min at 4 ºC. The tubes were left undisturbed for one min before removing the supernatant at each wash. Then, the beads were washed once with 750 µl of Wash Buffer II and twice 750 µl TE at pH 8.0 as described above. Now, the antibody/chromatin complex is ready to be separated from the beads.

2.5.10 Elution of the Antibody/Chromatin Complexes

225 µl of Elution Buffer was added to the washed beads, vortexed for 3 sec and incubated for a min at room temperature. Then, the samples were vortexed again and

centrifuged at 7500 rpm for 2 min at room temperature. The supernatant was carefully transferred to a new tube. This step was repeated and elutions (containing antibody/chromatin complexes) were collected in the same tube.

2.5.11 Reversal of Cross-linking and Digestion of RNA

1 µl of RNase (ICN Biochemicals, #101076) from 2 mg/ml stock solution and 27 µl of 5 M NaCl (final concentration of 0.3 M) were added to the samples. Also, 1 µl of RNase and 16.2 µl of 5 M NaCl were added to input chromatin samples. Then, the samples were incubated at 65 ºC for 6 h to reverse the cross-linking.

2.5.12 Digestion of Proteins and Recovery of DNA

After reversal of cross-linking, 9 µl of Proteinase K (Gibco-BRL #25530-031) from 10 mg/ml stock solution were added to the samples and the samples were incubated at 45 ºC overnight to digest proteins.

To recover DNA, 2 µl of tRNA (Gibco-BRL 15401-029) from 5 mg/ml stock solution was added immediately before adding 500 µl of water saturated phenol (Rathburn Chemicals, cat. no RP3024). The sample was mixed well and centrifuged at 13,000 rpm on a table-top centrifuge for 5 min. The aqueous layer was transferred to a new tube and 500 µl of chloroform (Rathburn Chemicals, RH1009) was added. The sample mixed well and centrifuged at 13,000 rpm on a table-top centrifuge for 5 min and the aqueous layer was transferred to a new tube. 1 µl of tRNA, 5 µl glycogen (Roche Applied Sciences, #901 393), 50 µl 3 M Sodium Acetate and 1200 µl absolute ethanol were added to the sample and the sample is put at -70 ºC for 30 min for DNA precipitation. Then, the sample was centrifuged at 13,200 rpm for 20 min at 4 ºC to precipitate DNA. The DNA pellet was washed with 500 µl of 70 % ethanol by centrifugating at 13,000 rpm for 5 min at room temperature. The supernatant was removed and DNA pellet was air-dried for 10 min. 100 µl or 50 µl of dd water were added to input chromatin or ChIP samples respectively.

5 µl of Sucrose/Creosol Red was added to 5 µl of each ChIP sample or 1 µl of input chromatin and the mixes were loaded on 1 % agarose gel and run for 2 h at 100 V to check the recovery of DNA.

2.6 Real-time PCR

2.6.1 Assessing ChIP efficiency

A Real-time PCR kit from Eurogentec (#RT-SN2X-03+WOUN) was used to assess the crosslinking efficiency or antibody performance. This kit provides a 2X reaction buffer which contains the PCR buffer, dNTPs, HotGoldStar DNA polymerase, MgCl2, SYBR© Green I, stabilizers and passive reference. The real-time PCR reactions are set up according to Table 2.15.

Reagents	Volume (µl)
2X Reaction Buffer	6.25
1.5 µM Primer Mix	2.5
DNA template	variable
DD water	up to 13
Total Volume	13.00

Table 2.15 Reaction set up for real time PCR with Eurogentec real-time PCR kit

There are 11 primer pairs designed to span promoter region of C20orf121 (-1946 bp to+393 bp; TSS is at +1) used for real-time PCR (see section Appendix B for primer list). ChIP sample and input chromatin were diluted in 1:10 and 1:20 respectively, and 2 µl of the diluted materials were used as templates for their corresponding reaction. Human genomic DNA (50 ng) was used as the positive control. The real-time PCRs are performed on ABI PRISM® 7700 Sequence Detection System in a 96-well plate format according to manufacturer’s instructions. The real time PCR steps are as follows; incubation for 2 min at 50 ºC followed by 10 min incubation at 95 ºC, then

40 cycles of 15 sec incubation at 95 ºC and 1 min incubation at 60 ºC. Ct values and amplification curves were generated automatically by the ABI PRISM® 7700 software.

2.6.2 Validation of ChIP on chip enrichment

The real-time PCR kit mentioned in section 2.6.1 is used to validate the enrichments obtained from ChIP on chip experiments. Input chromatin is diluted in 1:40 for amplification. Then, the reactions were set up according to Table 2.15, 0.1 µl DNA template (ChIP material or 1:40 diluted input chromatin) was used. The real-time PCR steps and analysis are described in section 2.6.1. Primers that are used to validate ChIP on chip results were listed in Appendix B.

2.7 Preparation of ChIP samples for hybridization onto microarrays

2.7.1 Labelling of ChIP samples

ChIP'ed DNA and input chromatin were randomly labelled using Bioprime Labelling Kit (Invitrogen, cat.no. 18094-011). 40 µl of ChIP sample were mixed with 60 µl 2.5X random primers solution and make up to 130.5 µl with dd water. For each ChIP sample, 2 µl of input chromatin was mixed with 60 µl 2.5X random primers solution and make up to 130.5 µl with dd water. The samples were denatured at 100 ºC for 15 min and immediately cool on ice for 5 min.

15 µl of deoxynucleotides mix (2 mM dATP, 2mM dGTP, 2 mM dTTP and 1 mM dCTP) was added to the reaction. 1.5 µl of cyanine-3 labelled dCTP analogue (NEN Life Sciences, #NEL576) was added to the ChIP'ed DNA and 1.5 µl of cyanine-5 labelled dCTP analogue (NEN Life Science, #NEL577) was added to input chromatin reaction. Then, 3 µl of Klenow Fragment was added to both reactions and they were

incubated at 37 ºC overnight in the dark. After the incubation, 15 µl of Stop Buffer was added to the reactions.

2.7.2 Removal of unlabelled nucleotides

Micro-spin G50 columns from Pharmacia Amersham (#275330-01) was used to remove unlabelled nucleotides. Since the capacity of each column is 50 µl, 3 columns were used for each sample. The resin in the columns was resuspended by gentle vortexing and the bottom closure of the columns was snapped off. The caps of the columns were loosened by one-quarter turn and the columns were placed in 1.5 ml eppendorf tubes. The columns were centrifuged at exactly 735xg for 1 min and water was discarded. 50 µl of HPLC purified water was applied to each column and centrifuged at 735xg for 1 min. Then, the columns were placed onto 2 ml eppendorf tubes and 50 µl of the labelling reactions was applied to the centre of the angled surface of the resin of the column and the columns were centrifuged for 2 min at 735xg. The columns were discarded and the flow-through samples were combined in the same tube. 5 µl of the samples were run on a 1 % agarose gel for 2 h at 100 V to check the labelling efficiency.

2.7.3 Competitive Hybridization of Labelled Samples onto microarrays

2.7.3.1 Preparation of the samples for the hybridization

The following reactions were prepared for each sample

Cy3 Labelled DNA	~180 ul
Cy5 Labelled DNA	~180 ul
Human Cot1 DNA	135 ul
3 M sodium acetate at pH 5.2	55 ul
Absolute Ethanol	1200 ul

The reactions were mixed gently, and put at -70 ºC for 1 hour to precipitate the DNA. Meanwhile, the hybridization buffer was heated in a 70 ºC heat block. The samples were spun at 13,000 rpm at 4 ºC for 15 min. The supernatant was removed and 500 µl

80 % ethanol was added to the samples and spun at 13,000 rpm for 5 min. The supernatant was removed and the samples re-spun at 13,000 rpm for 1 more min and residual ethanol was take off with a small tip. The pellets were dried for 10 min in the dark. 125 µl of hybridization buffer and 3 ul yeast tRNA were added to the pellet and the samples were left for 2-3 min in a 70 ºC heat block before resuspending the pellet.

The samples were denatured for 15 min at 100 ºC and immediately cool on ice for 5 min. They were pulse-spun and incubated 1 hour at 37 ºC before the hybridization onto arrays.

2.7.3.2 Hybridization

Tecan HS4800 Pro® hybridization station was used for hybridizations of the samples onto microarrays. The small TECAN chambers with dimensions 50.8 mm by 18 mm were used for hybridizations. 100 µl of the sample was injected onto chambers and hybridized for 45 h at 37 ºC in the dark. Then, the slides were washed and dried according to the protocol given in Table 2.16.

Type	Number of Runs	Solution	Wash Time (sec)	Soak Time (sec)	Temperature (ºC)
Wash	10	1XPBS/0.05% Tween20	60	30	37
Wash	5	0.1X SSC	60	120	52
Wash	10	1XPBS/0.05% Tween20	60	30	23
Wash	2	DD water (HPLC)	30	0	23
Drying	8	CO_2	30	-	23

Table 2.16. Slide washing and drying protocol on Tecan HS4800 Pro hybridization station.

2.7.4 Array Scanning

Slides were scanned using ScanArray Express HT microarray scanner (Perkin Elmer). Two laser of 633 and 543 nm wavelength were used to detect signal from Cy5 and Cy3 labelled molecules respectively. PMT gains used for Cy5 and Cy3 lasers were 65 and 67% respectively.

2.8 Solutions

$T_{0.1}E$

10 mM Tris-HCl (pH8.0)
0.1 mM EDTA

10X TBE

Add the following to 800 ml distilled water

- 108g Tris base (Sigma-Aldrich, #T3253)
- 55g Boric acid (Fisher, B/3800/53)
- 9.3g EDTA (NBS Bio, #0105)

Adjust volume to 1 L with additional distilled water

Dilute 1:10 to obtain 1XTBE agarose gel running buffer

SOC medium

2.0 g	Bacto®-tryptone
0.5 g	Bacto®-yeast extract
1 ml	1M NaCl
0.25 ml	1M KCl
1 ml	Mg2+ stock (1M MgCl2 • 6H2O, 1M MgSO4 •7H2O), filter-sterilized
1 ml	2M glucose, filter-sterilized

- Add Bacto®-tryptone, Bacto®-yeast extract, NaCl and KCl to 97ml distilled water
- Stir to dissolve
- Autoclave and cool to room temperature
- Add 2M Mg2+ stock and 2M glucose stock
- Filter the complete medium through a 0.2µm filter unit.
- Adjust pH to 7.0

LB Media

10 mg/ml Bacto®-tryptone
5 mg/ml Bacto®-yeast extract
10 mg/ml NaCl
Adjust pH to 7.4

LB-agar

(for LB-agar plates)
1 L of LB media
15 g Agar

Autoclave before use

LB growth medium

92.4 ml LB broth
7.5 ml 100% glycerol
0.1 ml 25 mg/ml chloroamphenicol

10X NEB PCR Buffer

For 100 ml;
8 g Tris
2.2 g $(NH_4)_2SO_4$
6.7 ml 1 M $MgCl_2$

Adjust pH to 8.8 and add DD water up to 100 ml. Filter sterilize before use.

Cresol red solution

0.1 g/l cresol red in $T_{0.1}E$

Sucrose/cresol red (1 litre of 40%)

400 g sucrose
0.1 g cresol red
Made up to 1000 ml with DD water

5X MOPS Buffer

0.2 M MOPS
50 mM Sodium Acetate
5 mM EDTA

Treated with DEPC by adding 1 ml of DEPC per litre of solution and incubate overnight, then autoclave.

2X RNA Loading Buffer

50% deionized formamide
1X MOPS Buffer
6.5% formaldehyde
5.4% saturated bromophenol dye
5.4% glycerol
15 ug/ml Ethidium Bromide

Cell Lysis Buffer (ChIP)

10 mM Tris-HCl pH 8.0
10 mM NaCl
0.2% NP40
10mM sodium butyrate

1 Roche Protease Tablet was added per 10 ml of solution

Nuclei Lysis Buffer (ChIP)

50 mM Tris-HCl pH 8.1
10 mM EDTA
1% SDS
10mM sodium butyrate

1 Roche Protease Tablet was added per 10 ml of solution

IP Dilution Buffer (ChIP)

20mM Tris-HCl pH 8.1
150 mM NaCl
2mM EDTA
1% Triton X-100
0.01% SDS
10mM sodium butyrate

1 Roche Protease Tablet was added per 10 ml of solution

IP Wash Buffer I (ChIP)

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

IP Wash Buffer II (ChIP)

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

IP Elution Buffer (ChIP)

100 mM $NaHC0_3$
1% SDS

Hybridization Buffer (ChIP)

2x SSC
50% deionised formamide
10mM Tris-HCl pH7.4
5% dextran sulphate
0.1% Tween 20

Promega Buffer E (10X)

6 mM Tris-HCl
6 mM $MgCl_2$
100 mM NaCl
1 mM DTT
Adjust pH to 7.5 at 37 ºC.

10X PCR CHIP Buffer

500 mM KCl
50 mM Tris-HCl, pH 8.5
25 mM $MgCl_2$

4X Spotting Buffer

1 M NaH_2PO_4
Adjust pH to 8.5

Add 0.001% Sarkosyl

1X PBS at pH 7.4

(0.2 M phosphate + 1.5 M NaCl)

2.28 g NaH_2PO_4 (mw=120)
11.5 g Na_2HPO_4 (mw=141.96)
43.84 g NaCl
Bring final volume to 1 L with dd water
Adjust pH to 7.4

5X MOPS Buffer

0.2 M MOPS at pH 7.0
50 mM sodium acetate
5 mM EDTA

Autoclave before use

DEPC-treatment of DD water

1 L DD water
1 ml DEPC

Incubate overnight and autoclave before use.

2X RNA Loading Buffer

50% deionized formamide
5.4% saturated bromophenol/xylene
cyanol dyes
5.4% glycerol
15 ug/ml Ethidium Bromide
6.5% formaldehyde
1x MOPS

3 PROMOTER AND GENE EXPRESSION PROFILING ON HUMAN CHROMOSOME 20 CYTOBAND q12-13.2

This chapter describes computational efforts on identifying promoter elements in a 10 Mb region of 20q12-13.2 and assesses the success rate of the current prediction programs. Besides, I also attempted to uncover novel features to further improve the efficiency of such programs. Finally, I present the expression profile of the genes in this region obtained with the Affymetrix Gene Expression Arrays in two different cell lines and correlate these results with the profile of the promoters obtained *in silico*.

3.1 Overview of the region

As described in section 1.5.1, the selected chromosomal region at 20q12-13.2 spans 10,099,058 bp and has, in brief, 103 protein-coding genes, 36 pseudogenes and 43 processed transcripts annotated in the Vega genome browser (version 19). Based on the current annotation, 54% (56/103) of the protein coding genes have more than one transcript, and many of these annotated alternative transcripts are based on alignments with expressed sequences (cDNAs and ESTs) and many either be incomplete or represent aberrant transcripts. In total, there are 402 different transcripts of which 35% (142/402) do not encode a protein. Out of the remaining 260 coding transcripts, 177 are controlled by different promoters assuming that two transcripts utilize the same promoter if their TSSs are less than 150 bp apart from each other. Analysis focused on this set. In all instances of multiple transcripts of the same gene using the same promoter, the transcript that has more biological evidence (cDNAs, EST, promoter or transcription site predictions) was included in selected dataset. The schematic representation of the selection procedure for the 177 transcripts is given in Figure 3.1.

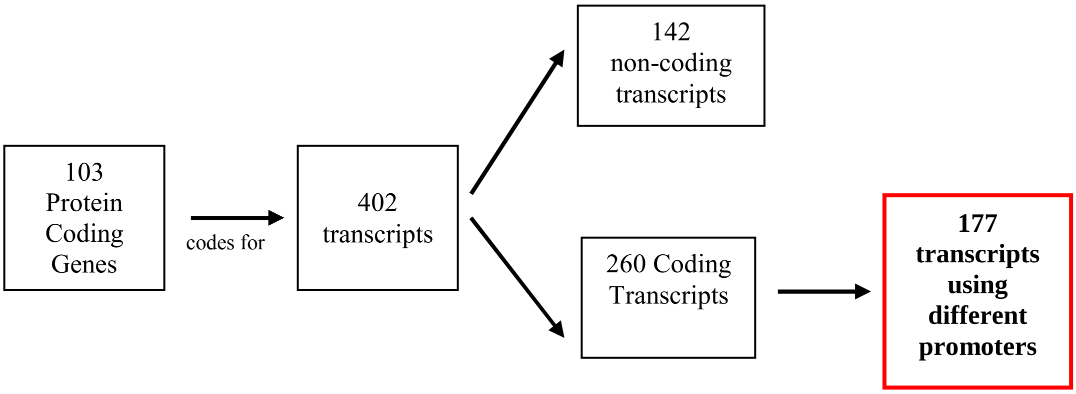


Figure 3.1 Schematic representation of the method for choosing 177 transcripts using different promoters.

The 177 selected coding transcripts are classified as representative transcript (RT) when they correspond to the longest 5' transcript, and alternative transcript (AT). It cannot be excluded that the annotation of both representative and alternative transcripts is incomplete.

3.2 Feature Predictions

The 177 candidate promoter regions (500 bp upstream and downstream of annotated TSS) were subjected to *in silico* analysis using CpG island, promoter and TSS prediction programs.

3.2.1 CpG islands

Promoter regions were searched for CpG islands using a program called "CpG Island Searcher" (Takai and Jones, 2002). Originally, CpG islands are defined as regions of greater than 200 bp with a %GC greater than 50% and an observed to expected CG dinucleotide (CpG) rate greater than 0.6 (Gardiner-Garden and Frommer, 1987). However, this software applies a set of more stringent conditions in order to exclude Alu repeats (see section 1.2.1). According to this program, 46% of the 177 promoters are associated with a CpG island. While 64% (66/103) of the RTs are associated with a CpG island, this percentage drops to 20% (15/74) for ATs.

CpG islands are also annotated in genome browsers such as Vega, Ensembl and UCSC and they all use the original definition of CpG islands. Vega and Ensembl use a program called "newcpgreport" (Micklem, 1999) to predict CpG islands. Newcpgreport finds CpG-rich regions of any length to find smaller CpG islands but it does not take into account the %GC of the region. Vega and Ensembl accept CpG islands found by this program only if their size and %GC are greater than 1000 bp and 50% respectively and the ratio of observed to expected number of CpG is equal or greater than 0.6. There are 97 CpG islands annotated in Vega Browser for the 20q12-13.2 and 75% (72/97) of them are associated with a TSS. Out of the selected 177 candidate promoters, 41% are associated with a CpG island according to Vega (Table 3.1). Three CpG islands annotated in Vega (associated with transcripts of GDAP1L1-001, GDAP1L1-006 and WFDC2-002) were not detected by "CpG island searcher" and both genes have restricted expression according to Unigene expression profiles (Wheeler et al., 2003). "CpG island searcher" detected 12 new CpG islands (10 out of 12 are associated with representative transcripts) and 75% of those transcripts are widely expressed according to Unigene expression profiles (Wheeler et al., 2003). That is in agreement with housekeeping genes being associated with CpG islands (Larsen et al., 1992). None of the newly added CpG islands contains ALU repeats and they are probably rejected by "newcpgreport" (used by VEGA and Ensembl genome browsers) since they are all smaller than 1000 bp.

The UCSC genome browser has more annotated CpG islands (118) in the region since it has a lower threshold for size (>300 bp) of a CpG island.

Type of Transcript	CpG island containing promoters		Total number of transcripts
	VEGA Genome Browser	CpG Island Searcher	
Representative Transcripts	58 (56%)	66 (64%)	103
Alternative Transcripts	14 (19%)	15 (20%)	74
All Transcripts	72 (41%)	81 (46%)	177

Table 3.1 Number of transcripts associated with CpG islands

Table A2 (Appendix A) lists CpG and GC content of a portion of the promoter regions (250 bp upstream and 50 bp downstream of TSS). A number of promoters (CDH22-001, C20orf142-001, ZSWIM3-001 and DNTTIP1-005) are associated with a CpG island despite their low CpG content since their CpG islands span downstream of their first exon.

3.2.2 Promoter Predictions

Promoter prediction in complex genomes is one of the most challenging tasks in computational biology for there are no known clear sequence signatures indicating its presence. PromoterInspector was the first prediction program with an acceptable sensitivity rate (Scherf et al., 2000). In a previous study, PromoterInspector was used to predict promoters in 20q12-13.2, where the program predicted 47.4% of the promoters correctly, producing 1.3 false predictions for every correct prediction, i. e. with a 43% sensitivity rate (Stavrides, 2002). Nowadays, several algorithms are available to predict promoters, nevertheless they all suffer from high false positive rates (Bajic et al., 2004). The program First Exon Finder (FirstEF) succeeds to predict a diverse set of promoters with a relatively low false positive rate (Bajic et al., 2004) (see section 1.2.6). Therefore, I chose FirstEF to predict promoter sites in the region.

FirstEF produced 198 predictions in the region and 68 (34%) of them are associated with the 5' of a coding transcript. The remaining predictions are not associated with 5' end of transcripts of protein coding genes. Interestingly, although 54% of the promoter regions of RTs are associated with a FirstEF prediction, this percentage is only 19% for the promoter regions of ATs. In terms of its accuracy and sensitivity, FirstEF predicted 66% of the promoters correctly while producing around 2.4 false predictions for every correct prediction, which translates to 30% sensitivity.

FirstEF utilizes the compositional features of promoters such as presence of a CpG island, therefore it is relevant to explore the GC content of its predictions to see if there is any bias towards GC-rich promoters. Within the selected promoter set, 88% (60/68) of the FirstEF predictions were associated with a CpG island. The remaining eight FirstEF predictions overlapping with promoters not associated with a CpG island have a significantly higher GC content (p value <0.0005). This result shows the success bias of the predictor toward CpG-island associated promoters. Additionally, 42% of the false predictions are overlapping with internal splicing sites of the genes; this is also expected since first slice-donor sites are also used for predicting promoters by the program.

3.2.3 Transcription Start Site Prediction - Eponine

While promoter prediction programs aim to find “promoter regions” in the genome, they are unable to locate the exact start position of transcription. Eponine is developed to locate the TSSs in mammalian genomes by exploiting (i) CpG enrichments downstream of TSSs (ii) the presence of a TATA-box binding motif centred around 30 bp upstream of a TSS (Down and Hubbard, 2002). The model constraints are schematically described in Figure 3.2.

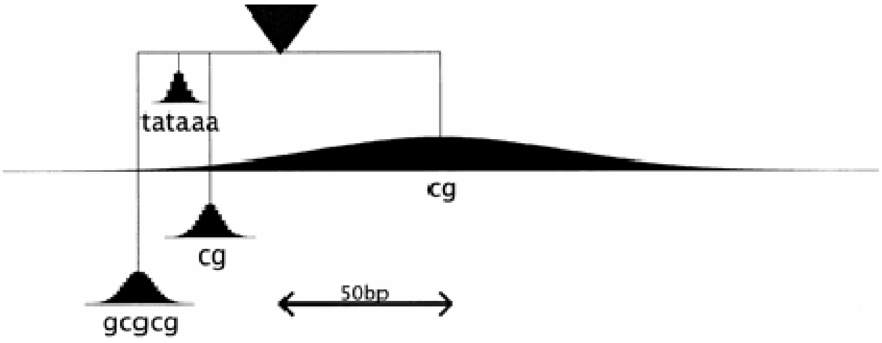


Figure 3.2. Sequence signals utilized by Eponine. Arrow head on the top marks the true TSS (taken from Eukaryotic Promoter Database (EDP) (Cavin Perier et al., 1998)). This figure is reproduced from reference Down and Hubbard, 2002.

Eponine predicted 267 TSSs in the region and 42 of them are located within 25 bp upstream or downstream of an annotated TSS. Out of these 42 predictions, 27 are associated with a coding transcript. As expected, all coding transcripts associated with an Eponine prediction contain a CpG island. Interestingly, only 4 of these predictions contain a TATA box binding motif (JASPAR model number M00980 (Sandelin et al., 2004)).

3.2.4 Overall Summary of Predictions

Out of the 177 candidate promoters, 23 are associated with all three prediction programs and 88 are not associated with any. Out of the 23 promoters with three predictions, 78% (18/23) correspond to a RT, while out of 88 promoters with no predictions, only 35% (57/88) corresponds to a RT. The prediction distribution over transcripts is shown in Figure 3.3.

Table A3 and Table A4 (Appendix A) list all the RTs and ATs in the region with the predictions associated.

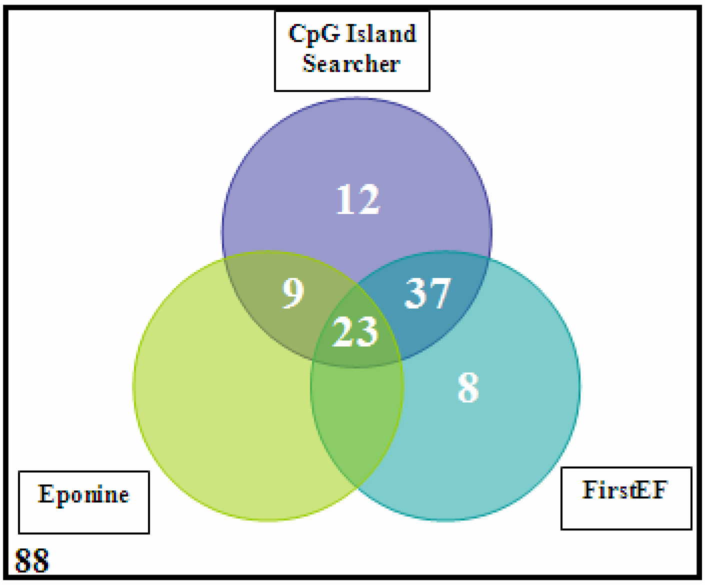


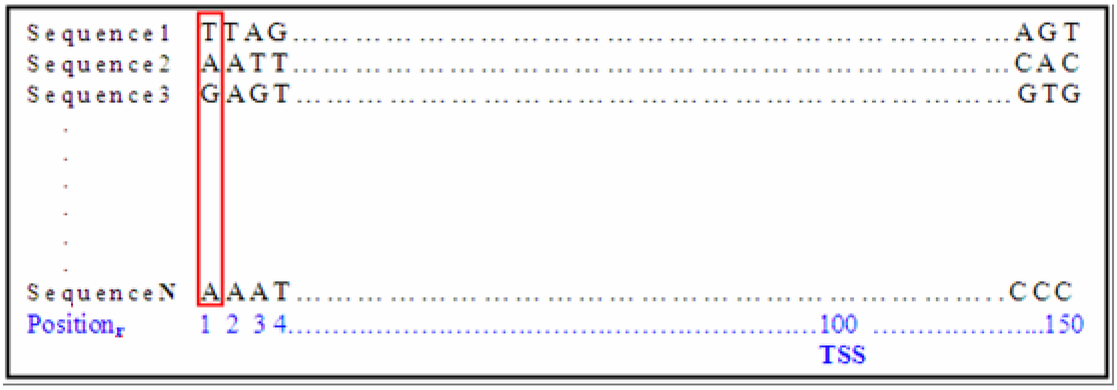
Figure 3.3. Number of promoters associated with predictions. There are 88 promoters not associated with any prediction.

In this study, promoters of 72 protein coding genes were detected either with the help of a CpG island or a promoter or TSS prediction program, but only 6 out of these 72 promoters were not associated with a CpG island. Promoters of 31 protein coding genes were not detected by any prediction programs; none of them is associated with a CpG island. These 31 genes have tissue specific expression according to Unigene Expression Profiles (Wheeler et al., 2003). These results clearly show the bias of the current promoter prediction programs towards CpG island associated promoters. Certainly, these prediction programs require additional signals to be able to detect promoters not associated with CpG islands.

3.3 Sequence and Structure Topology of Promoter Sequences

3.3.1 Sequence Topology of Promoter Sequences

Promoter sequences contain binding motifs which direct the recruitment of the trancription machinery onto it. The presence of these motifs may well set specific constraints on the sequence of promoters. Therefore, I investigated the frequency of nucleotides at each position of the promoter sequence relative to the TSS. To this end, sequences of extended core promoters (100 bp upstream and 50 bp downstream of TSS) are aligned relative to the TSS. Then, the number of each nucleotide, represented by B (B → [A, C, G, T]), in each position ($f(B)_r$, Equation 3.1) is counted in all sequences as it is illustrated in Figure 3.4.



```
Sequence1   TTAG.............................................AGT
Sequence2   AATT.............................................CAC
Sequence3   GAGT.............................................GTG
   .
   .
   .
   .
   .
   .
SequenceN   AAAT.............................................CCC
Position_r  1 2 3 4..............................100 ...........150
                                                TSS
```

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.

$$f(B)_r = \sum_{i=1}^{N} g(B)_r \ \textbf{where} \begin{Bmatrix} \textbf{if} \quad (\mathbf{B}_\mathbf{r})_N = B, & \mathbf{g(B)_r} = 1 \\ \textbf{else} \qquad\qquad , & \mathbf{0} \end{Bmatrix} \qquad \text{Equation 3.1}$$

$f(B)_r$ → number of nucleotide B on position r

B → [A, C, G, T]

r → position in the sequence [1..R] where R is the sequence length and

N → number of sequences

Nucleotide frequencies are normalized by subtracting the observed frequencies of each nucleotide at any random site in the human genome (denoted by O(B), Equation 3.2). These observed frequencies of nucleotides A, C, G and T at any random site in human genome are 30%, 20%, 20% and 30% respectively (Wu et al., 2005).

$$F(B)_r = \left(\frac{f(B)_r}{R} \times 100 \right) - O(B) \qquad \text{Equation 3.2}$$

$F(B)_r$ → normalized frequency of nucleotide B on position r

O(B) → observed frequency of nucleotide B in the genome

Also, nucleotide frequencies in ten adjacent positions are averaged to smooth out the local fluctuations.

Figure 3.5 shows the percentage deviation of each nucleotide from its randomly observed frequency at each position in the sequences of the 177 putative promoters under investigation.

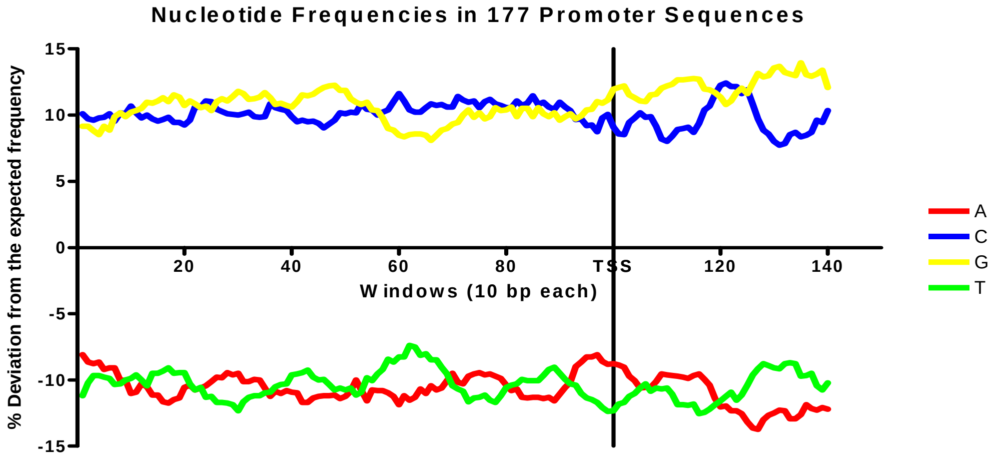


Figure 3.5. Frequency plots of each nucleotide relative to their distance to TSS in 177 promoters. Frequencies are averaged out in windows of ten nucleotides to eliminate local fluctuations. Solid black line at x=100 marks the TSS.

Figure 3.5 shows that nucleotides C and G are observed around 10% more than their expected frequencies whereas there is a 10% drop in observed frequencies of A and T nucleotides in promoter sequences.

Promoters can be categorized into two broad classes; TATA-box enriched promoters, which have relatively well-defined initiation sites, and CpG rich promoters (Carninci et al., 2006). The 177 putative promoter sequences were divided into two groups according to presence or absence of a CpG island. For illustration purposes, I summed the normalized frequencies of C and G nucleotides and then plotted for each class the summed nucleotide frequencies (Figure 3.6).

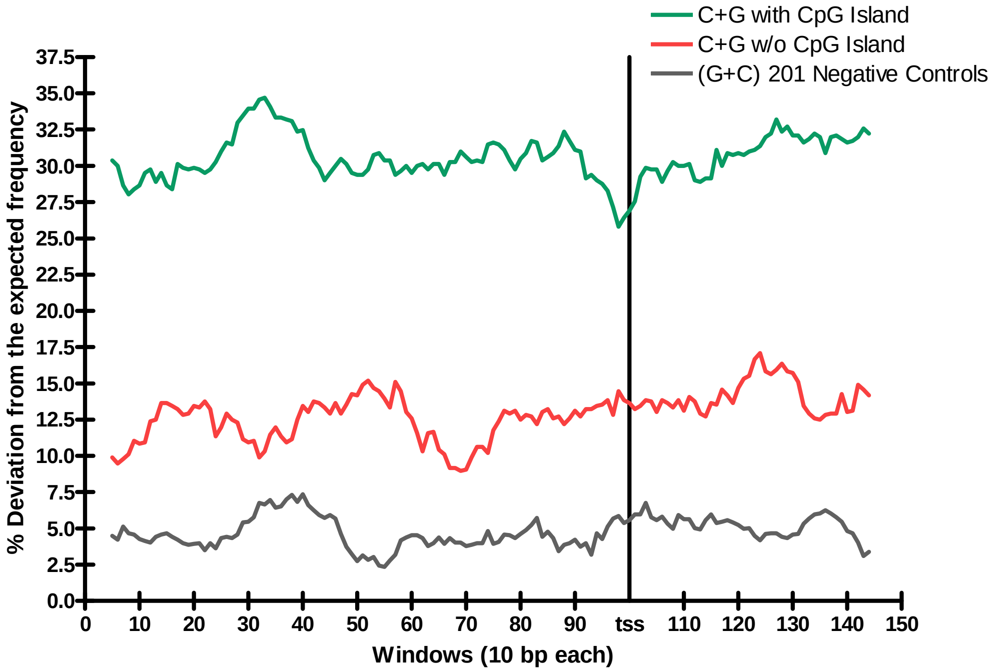


Figure 3.6. Summed nucleotide frequencies in promoters with, without CpG islands and 201 negative controls. Solid black line at x=100 marks the TSS. Frequencies are averaged out in windows of ten nucleotides to eliminate local fluctuations.

As expected, CpG island containing promoters have much higher GC content (30.6% ± 1.6%) as opposed to promoters without CpG islands (12.9% ± 1.7%). Both groups have however higher GC content compared to randomly selected negative controls (201) (4.8% ± 1.0%) as shown in Figure 3.6.

I also profiled 75 bp upstream and downstream of 3' UTR end of the 177 selected transcripts to investigate GC enrichment of such sequences (Figure 3.7).

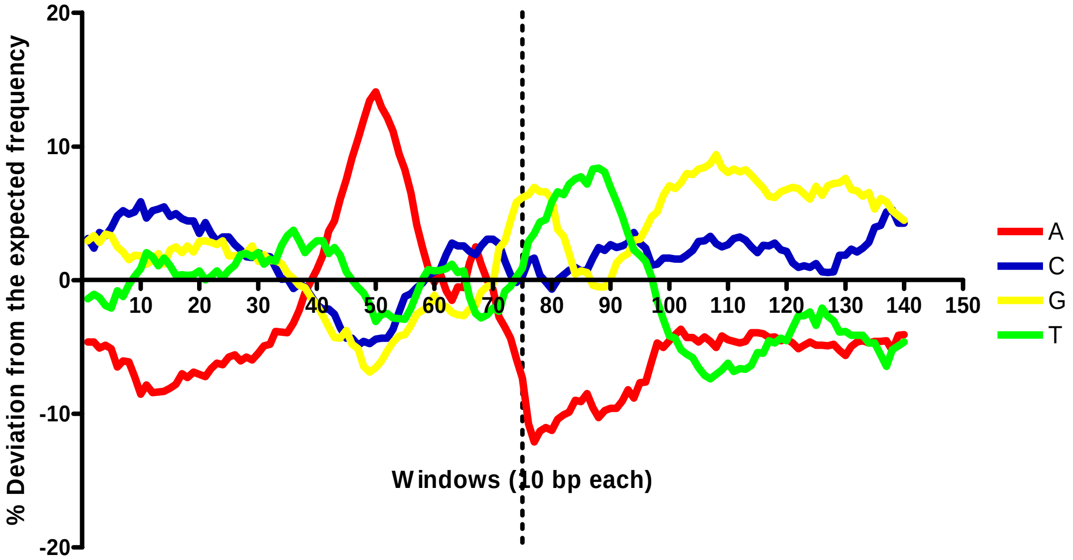


Figure 3.7 Frequency plots of each nucleotide relative to their distance to the 3’ end of the 177 promoters. Frequencies are averaged out in windows of ten nucleotides to eliminate local fluctuations. Solid black line at x=75 3’ end of the transcript.

As shown in Figure 3.7, the sequence profile of the 3’ ends of the transcripts are different than that of promoter sequences. There is no C:G enrichment across the sequences and different peaks are observed; a strong A peak at around 50 bp upstream of the 3’ end of the transcript and G peak at around 30 bp downstream of the 3’ end of the transcript.

3.3.1.1 Sequence Topology and Prediction Programs

Promoter and TSS prediction programs seek for specific sequence motifs such as CpG islands, TATA-box or initiator sequence. In the investigated promoter dataset, 44% (77/177) of the promoters are associated with either a promoter (FirstEF) or a TSS (Eponine) prediction. A+T and C+G nucleotide frequency plots of the two subgroups i.e. promoters with or without associated predictions are shown in Figure 3.8.

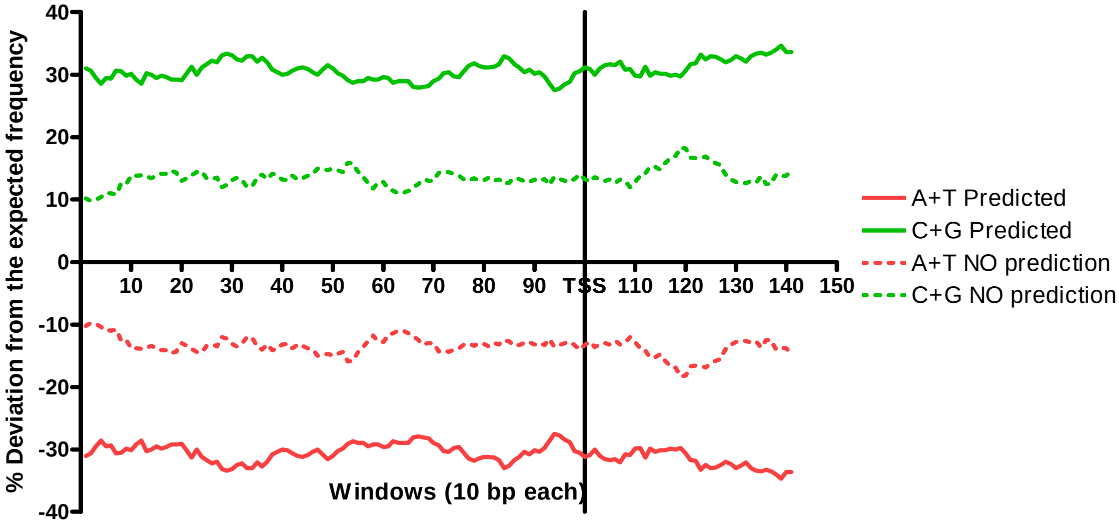


Figure 3.8 Summed nucleotide frequencies of 77 promoters associated with at least one prediction (dotted lines) and 100 promoters not associated with any prediction. Frequencies are averaged out in windows of ten nucleotides to eliminate local fluctuations.

Promoters predicted computationally are very GC-rich compared to the other subgroup. This is expected since 90% of the promoters with a prediction are associated with a CpG island. The similarity of the nucleotide frequency profiles between the two subgroups provides supports that the sequences without predictions are most likely false negatives.

A promoter with a CpG island has a higher chance to be predicted *in silico* since it has a distinct sequence feature. To see whether there are any sequence features in promoters that are missed by the prediction programs, I examined the nucleotide frequencies of promoters not predicted by any program. Additionally, I subdivided promoters that are not associated with any prediction according to their transcription class since 70% (72/103) of the RTs are associated with at least one prediction and/or a CpG island while only 23% (17/74) of the ATs are associated with at least one prediction and/or a CpG island

Figure 3.9 displays the summed nucleotide frequencies of RTs and ATs not associated with any prediction or a CpG island.

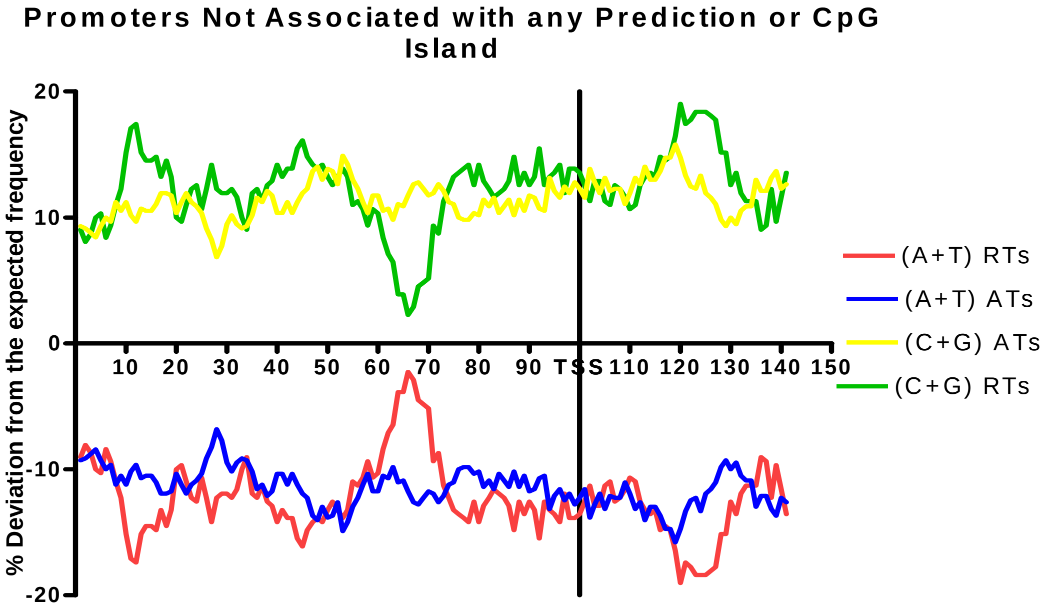


Figure 3.9. Summed nucleotide frequencies of RTs not associated with any prediction or a CpG island and ATs not associated with any prediction or a CpG island. Frequencies are averaged out in windows of ten nucleotides to eliminate local fluctuations.

The two transcript types have rather similar C+G and A+T content but points of clear difference are also visible. The A+T peak in the RTs plot around 35 bp upstream of the TSS indicates the presence of a TATA-box. There is no significant change in the A+T and C+G distribution between the two classes at 25 bp upstream and downstream of TSS. However, there is an increase in C+G content in the RTs 30 bp downstream of the TSS, which might be a sign for the downstream promoter element (DPE). For further analysis of the differences between the nucleotide frequency distributions between RTs and ATs with no prediction or CpG island, I fitted curves to each C+G nucleotide frequency distribution plot and subtracted fitted C+G curves of different transcripts respectively and plotted difference curve in Figure 3.10.

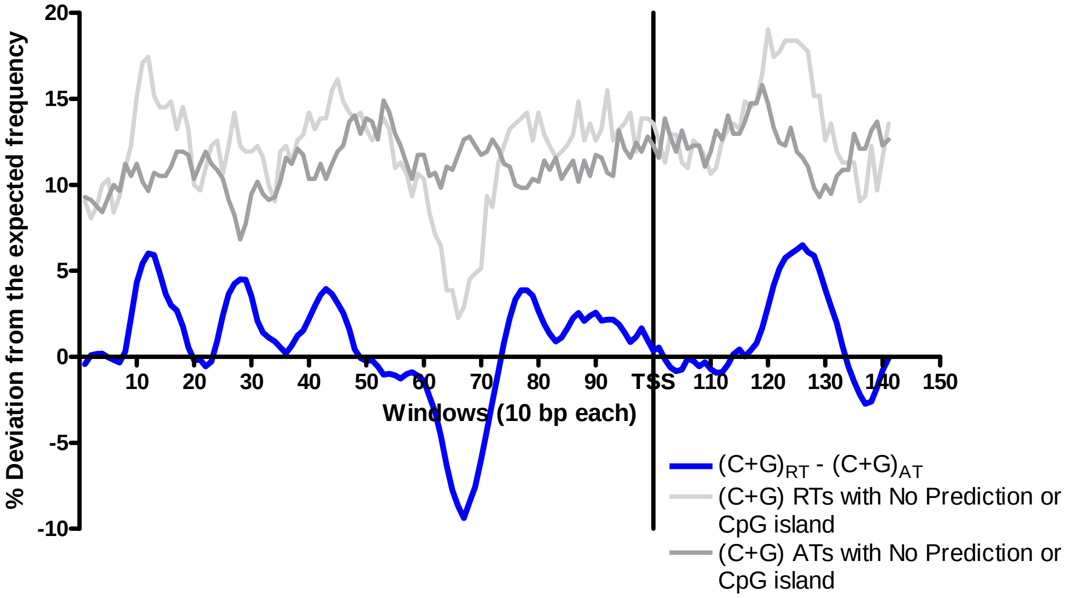


Figure 3.10 (C+G) plot of RTs and ATs not associated with any prediction are curve-fitted (grey curves) and subtracted from each other (blue curve).

The (C+G) peaks between 100 to 50 bp upstream of the TSS in Figure 3.11 may corresponds GC-boxes (consensus sequence, GGGCGGG) which serve as binding sites for the transcription factor Sp1 (Fukue et al., 2005).

In summary, promoters display a unique sequence pattern where (C+G) content is on average 20% higher from its expected frequency. Interestingly, (C+G) and (A+T) contents do not fluctuate greatly, which means that changes in the frequency of nucleotide C is mostly compensated by nucleotide G and changes in the frequency of nucleotide A is compensated by nucleotide T.

Computational efforts for predicting promoters or TSS seems quite biased towards (C+G) content of sequence as promoters with low (C+G) content cannot be detected by these programs (Figure 3.8). These programs certainly have room for improvement to detect more promoters on the sequence level, as Figure 3.10 clearly shows that representative transcripts which are not associated with any prediction or CpG island have distinct sequence features such as Sp1 and TBP binding sites as well as DPE. It is intriguing that although alternative transcripts not associated with any prediction or a CpG island, do not have such distinct sequence motifs, they do retain the high GC content that may indicate promoter sequences.One can speculate that these promoters may be very tissue-specific and/or tightly-controlled, as they do not have clear binding sites for common activation factors.

3.3.2 Structural Topology of Promoter Sequences

Initiation of transcription is a complex process requiring a number of proteins acting co-operatively on the promoter region. DNA bending is a vital factor in protein binding and nucleosome positioning; a straight and rigid double helix cannot accommodate deformed structures which allows functional protein binding (Travers, 1989). It was therefore interesting to examine the bendability profile of the 177 putative promoter sequences in this study together with non-promoter sequences and finally correlate the bendability profiles of different types of promoters to structural features.

3.3.2.1 DNA Bending

DNA interactions with DNA I nuclease were used to understand the DNA bending as the binding of the nuclease is largely determined by the flexibility of the sequence towards the major groove (Brukner et al., 1990). It is possible to extract a measure of bending ability of a DNA sequence using experimental data produced by employing variety of DNA sequences and their corresponding DNase I cutting frequencies (Brukner et al., 1995). To this end, bendability figures for all possible trinucleotide sequences were produced as listed in Table 3.2; the measure is in arbitrary scale and higher values (less negative) means higher bendability towards the major groove (Brukner et al., 1995).

Sequence	Reverse Complemented	Bendability Score
AAA	TTT	-0.274
AAC	GTT	-0.205
AAG	CTT	-0.081
AAT	ATT	-0.28
ACA	TGT	-0.006
ACC	GGT	-0.32
ACG	CGT	-0.033
ACT	AGT	-0.183
AGA	TCT	0.027
AGC	GCT	0.017
AGG	CCT	-0.057
ATA	TAT	0.182
ATC	GAT	-0.11
ATG	CAT	0.134
CAA	TTG	0.015
CAC	GTG	0.04
CAG	CTG	0.175
CCA	TGG	-0.246
CCC	GGG	-0.012
CCG	CGG	-0.136
CGA	TCG	-0.003
CGC	GCG	-0.077
CTA	TAG	0.09
CTC	GAG	0.031
GAA	TTC	-0.037
GAC	GTC	-0.013
GCA	TGC	0.076
GCC	GGC	0.107
GGA	TCC	0.013
GTA	TAC	0.025
TAA	TTA	0.068
TCA	TGA	0.194

Table 3.2. Bendability scores of all possible trinucleotides. Higher values (less negative) translate to higher bendability towards major groove.

Bendability scores were produced for the 177 putative promoters using 100 bp upstream and 50 bp downstream sequence of the TSS. A bendability score was assigned to every trinucleotide along the sequence according to Table 3.2 and the scores of every 12 nucleotides (10 trinucleotides) were averaged out to smooth the local fluctuations as it is illustrated in Figure 3.11.

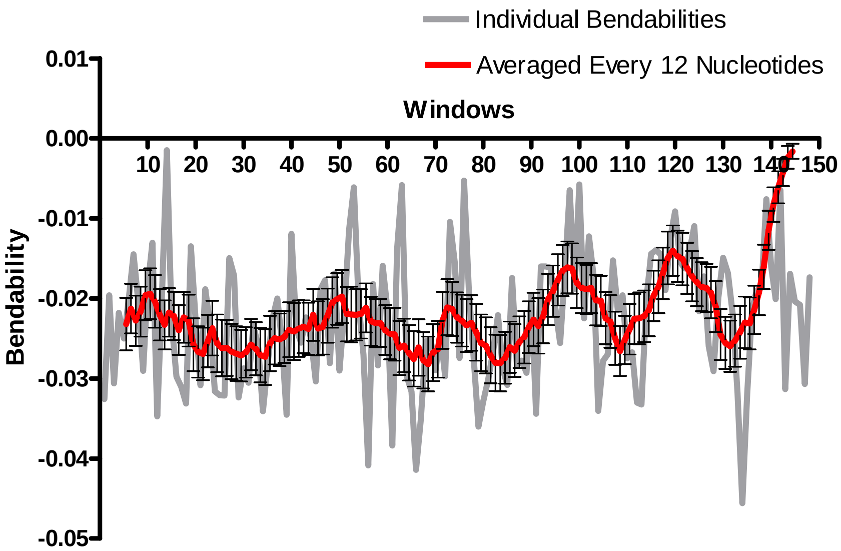


Figure 3.11. Average bendability score of every trinucleotide along the sequence of 177 promoter sequences (grey line) and bendability scores averaged at every 12 nucleotide (10 trinucleotides) shown with error bars (red line).

Figure 3.12 shows the bendability profile of 201 randomly selected intergenic and intra-genic human sequences (negative controls) alongside those of the 177 putative promoters.

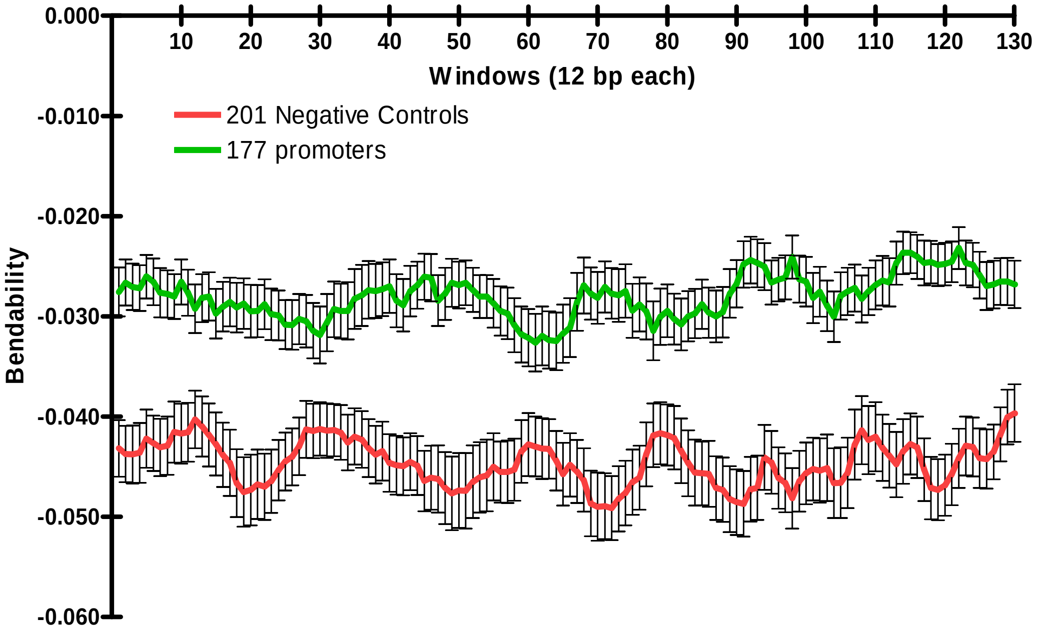


Figure 3.12. Bendability scores of 201 negative control sequences (red line with error bars) versus 177 promoter sequences (green line with error bars) averaged out in every 12 nucleotides.

Promoter sequences have a higher mean bendability (0.02704 ± 0.004546) compared to the negative control set (0.04302 ± 0.00687). The high bendability of promoter sequences might offer a more flexible structure to DNA helix to accommodate protein binding events.

Open-chromatin and gene rich regions of the genome are enriched with GC-rich sequences, which can have higher bendabilities (Vinogradov, 2003). The bendability profiles of putative promoters associated or not associated with a CpG island are shown in Figure 3.13.

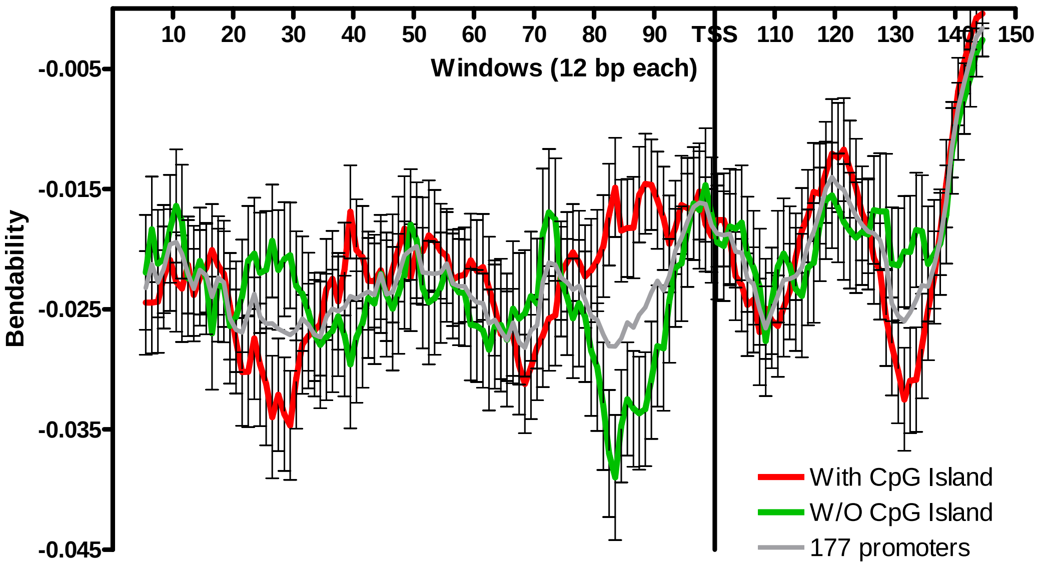


Figure 3.13. Bendability scores of 81 promoters associated with a CpG island (red line with error bars) and 86 promoters not associated with a CpG island (green line with error bars). Grey line indicates the bending profile of all the dataset.

The bendability profiles of the two sets overall do not differ greatly. However, there is a difference in bendability scores around -20 bp, which coincides with the position of TATA-box motif. However, analysis of a genome-wide set will be required to assess the significance of this observation.

Such different DNA conformations might help recruiting different sets of protein complexes or even different type of initiation complexes (Wieczorek et al., 1998).

Since there is a significant difference in nucleotide distribution of representative and alternative transcripts not associated with any prediction or CpG island (see Figure

3.10), I assessed the bending profiles of these two types of transcripts. The plots of bendability scores of the two transcript groups are presented in Figure 3.14.

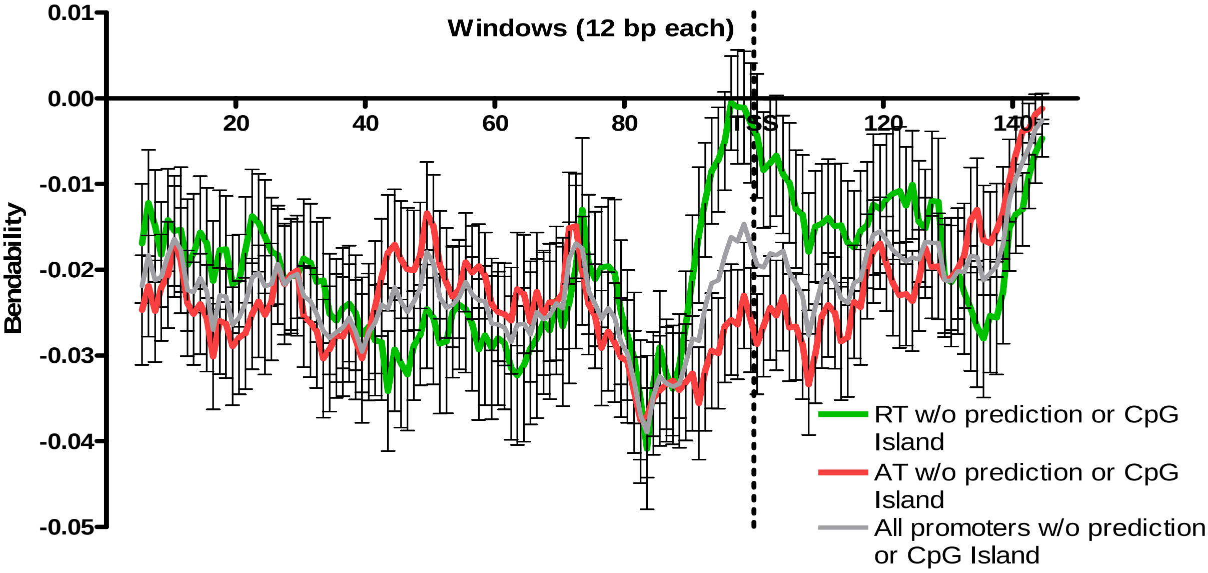


Figure 3.14. Bendability scores of RTs (green line with error bars) and ATs (red line with error bars) not associated with any promoter or TSS prediction or a CpG island and grey line denotes the bendability of all promoters sequences not associated with any prediction or a CpG island.

The two groups show significant differences in their bending profiles most notably around the TSS where RTs not associated with a CpG island or prediction have much higher bendability towards the major groove. This high bending might be the result of (A+T)-rich sequence just before the TSS found only in representative transcripts (see Figure 3.9).

From these analyses, we can conclude that there is a specific bendability profile in promoter sequences and this profile is affected by the presence of a CpG island. As expected, there is a correlation between the sequence content and the bending capacity of the sequence as higher bendability is observed in the presence of (A+T)-rich sequence profile in case of the promoters of representative transcripts not associated with a CpG island or prediction.

3.4 Expression Profile of 20q12-13.2 using Affymetrix Arrays

Affymetrix Human Expression Array U133 plus 2.0 was used to determine the expression status of genes in the 20q12-13.2 region on HeLa S3 and NTERA-D1 cell lines. These two cell lines were used to characterise experimentally the 177 putative promoter sequences by gene reporter assays (see Chapter 4) and chromatin immunoprecipitation assays (see Chapter 5). The arrays were hybridized with the total RNA extract of each cell line as described in section 2.3. A script was written to extract all the probes on the corresponding array perfectly aligning to the sequence of the investigated region. There were 280 probes aligning to 20q12-13.2 and 205 of them aligned to a coding transcript. There were 6 and 13 probes representing non-coding transcripts and pseudogenes respectively. The remaining 56 probes were either

- representing partial cDNAs
- ambiguous
- matching on the reverse strand of an annotation
- matching a partial gene, or exon prediction
- not matching with any annotated feature or experimental evidence

Affymetrix probes are classified into 4 groups according to the uniqueness of the probe (Ivanova et al., 2006). Probes that have a;

- "**_at**" suffix represent probes designed to detect a unique sequence of a single gene
- "**_a_at**" suffix represent probes designed to detect multiple alternative transcripts of a single gene
- "**s_at**" suffix represent probes designed to detect multiple transcripts of different genes
- "**x_at**" suffix represent probes that can cross-hybridize to unrelated sequences; these probes should be treated with caution.

I removed all the probes with an "x_at" suffix since some of the signal of this type of probes might be produced by an unrelated sequence. This left 188 unique probes.

3.4.1 Coding Genes and Transcripts

Out of 103 genes, 99 (96%) genes are represented by at least one probe on Affymetrix U133 Plus 2 expression array respectively. At the gene level, 47 (47.5%) genes were expressed in both cell lines and 29 (30%) genes are not expressed in any of the two cell lines used in this study. As shown in Figure 3.16, 17 genes are only expressed in NTERA-D1 and 6 genes are only expressed in HeLa S3 cell line. Therefore, NTERA-D1, a testis cell line, has a broader expression pattern as expected, since testis has already shown as one of the tissue source of one of the most complex and diversified transcriptome (Jongeneel et al., 2005). The expression status of all the genes is listed in Table A5 (Appendix A).

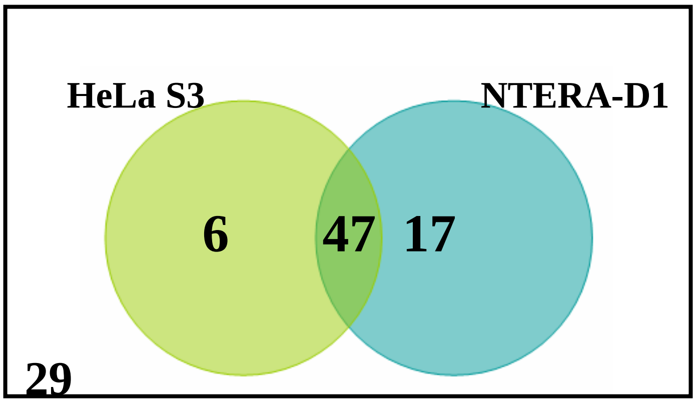


Figure 3.15. Expression profile of 96 genes represented by probes on Affymetrix U133 Plus 2 expression array, where 29 genes are not expressed by neither of the cell lines.

Table A6 (Appendix A) lists the transcripts that are represented by each probe on the expression array, unfortunately none of the probes was able to differentiate the expression of the alternative transcripts of the same gene.

3.4.2 Expression Profiles and CpG Islands

Out of the 99 genes represented by at least one probe on Expression Array, 57 contain a CpG island (see Table 3.3). Of those 57, 36 (63.2%) were expressed in both cell lines. Out of the 42 genes not associated with a CpG island, 11 (23.8%) were expressed in both cell lines while 21 (50%) were not expressed in either. In summary, 86% of the genes associated with a CpG island are expressed in at least one cell line whereas this figure is 50% for the genes not associated with a CpG island.

	Representative Transcripts Associated with a CpG Island	**Representative Transcripts NOT Associated with a CpG Island**
Expressed in both cell lines	36 (**63.2**%)	10 (**23.8**%)
Expressed only in HeLa S3	1 (1.8%)	5 (11.9%)
Expressed only in NTERA-D1	12 (21.1%)	6 (14.3%)
No expression	8 (**14.0**%)	21 (**50.0**%)
Total number of transcripts	56	42

Table 3.3. Expression Profiles of 96 representative transcripts associated or not associated with CpG islands

3.4.3 Pseudogenes and Processed Transcripts

There are seven probes aligning to processed pseudogenes and four of them gave high positive signals, but all these pseudogenes show sequence homologies with the ubiquitously expressed ribosomal protein genes. There are six probes designed to detect to expression of processed transcripts. One of these probes (“226835_s_at” aligning to RP4-686N3.3 processed transcript) showed a very high positive signal. However, this probe cross hybridizes the mRNA of E3 ubiquitin-protein ligase gene (NEDD4). So it is not possible to differentiate the signal since this gene is widely expressed.

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

Promoters carry the central regulatory information of genes, therefore their correct annotation and characterization is vital to understand gene function. The idea that gene expression might be controlled by specific regions in the genome came from Scaife and Beckwith in 1966 (Scaife and Beckwith, 1966); they identified several genomic mutations decreasing or abolishing the activity of the lac operon genes in *E. coli*. They also confirmed that these mutations act at the sequence level as the insertion of a second lac region into the genome did not relieve the mutational effect. Two years later, Ippen and co-workers used bacterial strains with mutations at the start of the lac operon and located a control region that they call promoter where the transcription initiates (Ippen et al., 1968). An elegant study from Block et al. in 1971 successfully confirmed that, indeed, transcription initiates on lac promoter, using *in vitro* transcription (Eron and Block, 1971).

In vitro transcription was the first method of choice for promoter identification in mammalian cells (Weil et al., 1979) (Manley et al., 1980). Although this method is powerful for confirmation of promoters, it is not suitable for studying its specific regulation. Another approach to assess promoter activity is to measure the RNA levels of its gene. However, RNA quantitation can be a tedious task; template amplification may not be always successful and it is often difficult to have accurate and reproducible measurements for genes with low expression.

In late 1970's, an alternative method emerged to study promoters *in vivo* where the promoter of interest is joined with a "reporter" gene whose product can be assayed to monitor the activity of the promoter controlling the reporter gene (Ota et al., 1979). This method has been successfully adapted to mammalian cells using suitable reporter gene such as chloroamphenicol acetyltransferease (Gorman et al., 1982), luciferase (de Wet et al., 1987), green fluorescent protein (GFP) (Kain et al., 1995). In such assays, the reporter gene is encoded by a plasmid of between 4-6 kb in size with a multiple cloning site placed immediately upstream of the reporter gene and an antibiotic resistance gene for selection of recombinant clones (see section 2.1.3.1). The plasmid carrying a candidate promoter fragment in front of the reporter is transiently transfected into mammalian cells to detect promoter activity (see section 2.1.5). Typically, transfected cells are incubated for 24 – 72 h depending on the cell type, doubling time of the cells or regulatory nature of promoter of interest. Detection of the reporter protein in the cells and its expression level is then correlated with promoter activity. Alternatively, cells can be stably transfected with plasmid which integrates into the genome of the host cell and is retained beyond division (Pellicer et al., 1980). Reporter assays allow us to do large-scale promoter screening due to its rapid and scalable nature (Trinklein et al., 2003; Cooper et al., 2006). In this study, I have screened 74 candidate promoter fragments using dual luciferase reporter assay in two different human cell lines in HeLa S3 and NTERA2 clone-D1 (NTERA-D1).

4.1 Reporter Genes

The main considerations in choosing a reporter gene are (i) its activity ideally should be absent in the host cell; (ii) other enzymatic activities of the host cell should not interfere with the reporter activity; (iii) its detection assay should be rapid, reproducible and sensitive without using any hazardous chemical to the host cells.

Reporter genes are employed for assessing promoter activity (Gorman et al., 1982; de Wet et al., 1987; Kain et al., 1995), subcellular protein localization (Kain et al., 1995), to assess gene delivery methods into cells (Ewert et al., 2004) and *in vivo* detection of protein-protein interactions (Mitra et al., 1996). There are several alternative reporter genes to monitor promoter activity *in vivo* and each has advantages and disadvantages depending on its purpose of use. Here, three common reporter genes for promoter assays will be described in brief and compared.

4.1.1 Chloroamphenicol Acetyltransferease (CAT)

CAT is encoded by a bacterial drug-resistance gene and inactivates the antibiotic chloroamphenicol by acetylating at one or both of its hydroxyl groups (Shaw, 1975). First, the cells are lysed and incubated with radioactively labelled chloroamphenicol in the presence of the cofactor n-Butyryl Coenzyme A. CAT activity can then be assayed by autoradiography of the lysate subjected to thin layer chromatography (TLC). Acetylated and non-acetylated chloroamphenicol are separated by TLC, and the presence of the acetylated form correlates to the expression of CAT. In order to avoid using radioactivity, an enzyme-linked immunosorbent assay (ELISA) has been adapted to detect CAT activity (Gao et al., 2002).

4.1.2 Green Fluorescent Protein (GFP)

GFP is a 27-kDa monomeric protein from the jellyfish *Aequorea* widely used as a reporter protein in fixed and live tissues and no substrate is required for its visualization. Wild type GFP when excited by violet (395 nm) light emits green (509 nm) light which can be detected by a fluorometer. For protein localization and interactions, GFP is the reporter of choice since it is possible to detect its signal in intact cells. GFP signal cannot be enhanced enzymatically like the luciferase signal (see section 4.3.1), which leads to lower sensitivity. Yet, it has been shown that GFP is as sensitive as luciferase when the GFP expression is quantified by flow cytometry where the emitted fluorescence intensity can be measured at the single-cell level without any processing steps (Ducrest et al., 2002).

4.1.3 Luciferase

Luciferase is a 62 kDa beetle enzyme that catalyses the reaction of luciferin, to oxyluciferin in the presence of ATP and O_2 and Mg^{2+} and yellow light (560 nm) is produced as a result. It is a chemiluminescent assay since it requires a chemical modification to give the luminescence that can be detected by a luminometer. It can detect very low levels of gene expression (Promega, 2006). But, the delivery of luciferin to different cell types is difficult and unlike GFP, it is not possible to detect directly.

Nowadays, luciferase reporter assays are commonly used in monitoring promoter activity in mammalian cells and are also the method of choice in this study. Figure 4.1 shows that the bioluminescent reactions of firefly and renilla luciferase enzymes which catalyse the two luciferin molecules used in this study. Firefly luciferase, using ATP, catalyses the two-step oxidation of luciferin to oxyluciferin, which yields light at 560 nm. Renilla luciferase catalyses the oxidation of coelenterazine to coelenteramide, which yields light at 480 nm.

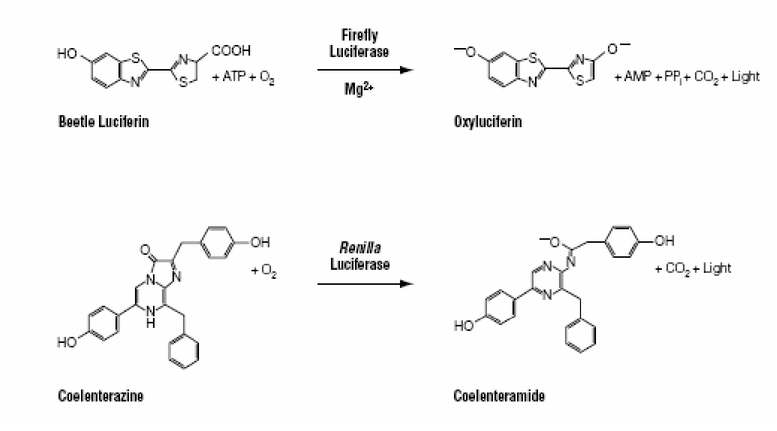


Figure 4.1. Bioluminescence reactions catalysed by firefly and renilla luciferase. This figure is reproduced from Promega® Product Technical Manual No 0.40.

Luciferase assays are faster than any other available reporter assay requiring cell lysis and it has been shown that they are up to 100 times more sensitive than CAT assays (Shaw-Jackson and Michiels, 1999). Figure 4.2 shows the linear ranges of firefly and renilla luciferases that can detect concentrations as low as 10^{-20} and 3×10^{-19} moles respectively.

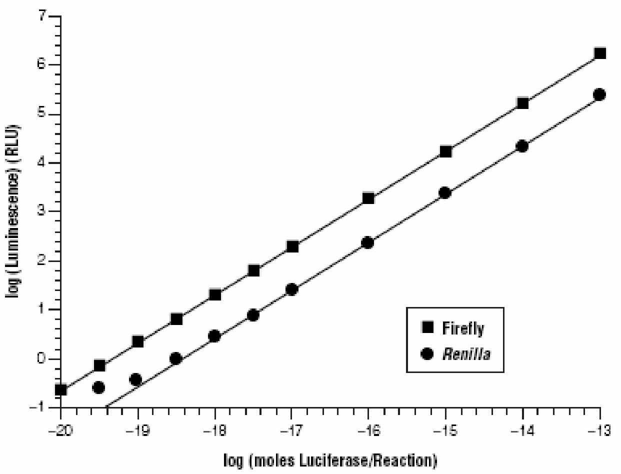


Figure 4.2. Linear ranges of firefly and renilla luciferases. The linear range of the firefly luciferase assay is seven orders of magnitude, providing detection sensitivity of 1 femtogram (approximately 10^{-20} mole) of experimental reporter enzyme. The renilla luciferase assay has a linear range of 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.

Another advantage of luciferase is its short half-life (~3 h) in mammalian cells compared to CAT (~50 h) (Promega, 2006). Due to its short half-life, it can reflect better changes in its promoter activity.

4.2 Dual Luciferase Assays

Reproducibility in reporter assays mainly depends on the efficiency of plasmid delivery into host cells assuming a constant amount of transfected plasmid. There might also be factors interfering with transfection or gene expression that are inherent to the experimental system. In such cases, an internal control plasmid can be used to normalize experimental variations. This control plasmid carries a different reporter gene under the control of a constitutively active promoter and it is co-transfected with the main reporter plasmid to the host cell. Activities of these reporter genes can then be detected by separate means.

In this study, dual luciferase reporter assays have been employed where the internal control plasmid pRL-SV40 (see map in section 2.1.3.1) carries the coelenterazine reporter gene under the control of the SV40 promoter. The internal control plasmid was co-transfected with the pGL3-basic reporter plasmid (see map in section 2.1.3.1) carrying the beetle luciferin under the control of a putative promoter fragment. First, the signal obtained by the oxidation of beetle luciferin catalysed by firefly luciferase was detected, a reagent was then added to quench the firefly luciferase action and catalyse the renilla luciferase reaction which oxidizes coelenterazine. The signal from the renilla luciferase was then detected (see section 2.1.6). The normalized signal was calculated by dividing the firefly to the renilla signal. Figure 4.3 shows the signals generated by both luciferase enzymes used in dual reporter assays and the residual activity of luciferin molecule (middle column) where the firefly luciferase activity is quenched by greater than five orders of magnitude.

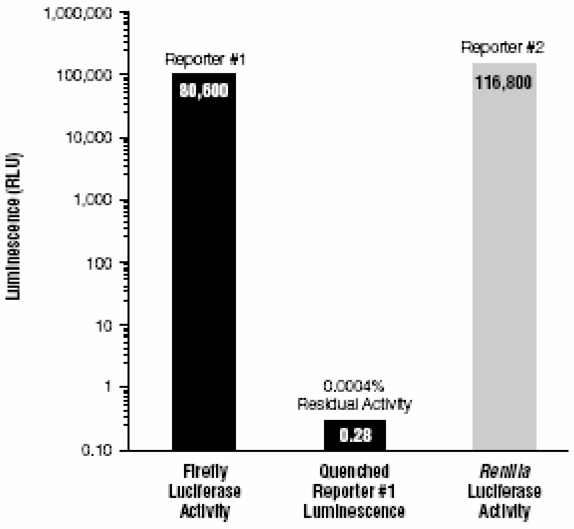


Figure 4.3. Measurement of luciferase activities before and after the addition of Stop& Glo® Reagent which quench the activity of beetle luciferase and initiate the renilla luciferase reaction. Beetle luciferase luminescence was quenched by greater than 5 orders of magnitude. This figure is reproduced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

All transfections were carried out in 48-well plate format. Each plasmid carrying a putative promoter fragment (construct) was transfected in triplicate in each experiment and transfection experiment was performed in duplicate. The methodology is described in section 2.1.5 in detail.

4.2.1 Positive and Negative Controls

A plasmid carrying the beetle luciferin reporter gene under the control of the strong SV40 promoter is used as positive control (pGL3-promoter plasmid, see Figure 2.5 for the map) to confirm transfection success and assess its efficiency.

A promoter is stripped of its genomic context in reporter assay studies and this may lead to non-specific promoter activity. That is because the complete regulatory information of a promoter is encoded in its histone code and that of distal genomic regulatory elements. To estimate the degree of non-specific promoter activity, five randomly selected inter- or intra-genic fragments with no known or predicted promoter activity were also included in the study as negative controls. Their observed activation levels were used to determine the degree of noise in the applied reporter assay methodology.

4.2.2 Determining Promoter Activities

Transfection of each promoter construct (with or without SV40 enhancer) was performed in triplicate per experiment and two experiments were performed per construct. Thus, in total six replicates were generated per construct. The signal from each well is calculated by dividing the luciferin (Luc) to coelenterazine (renilla, Ren) signal (Equation 4.1). The mean raw signal of ith promoter in jth experiment (Mean_Raw_Signal$_{ij}$) is calculated as in Equation 4.2. (k is the replicate index within one specific experiment).

$$Raw\_Signal_{ijk} = \frac{Luc_{ijk}}{\mathrm{Re}\, n_{ijk}} \qquad \text{Equation 4.1}$$

$$Mean\_Raw\_Signal_{ij} = \left(\sum_{k}^{3} Raw\_Signal_{ijk} \right) \times \frac{1}{3} \qquad \text{Equation 4.2}$$

Then, the mean signal of each negative control is calculated according to Equation 4.2 and the negative control construct that gives the highest signal is taken as the background signal (Background$_{ij}$). The background signal was subtracted from the signal of each promoter to remove the background noise (Equation 4.3).

$$Signal_{ij} = Mean\_Raw\_Signal_{ij} - Background_{ij} \qquad \text{Equation 4.3}$$

The signals taken from each experiment (Signal$_{i1}$, Signal$_{i2}$) are averaged out to calculate the mean signal of the ith promoter (Equation 4.4).

$$Mean\_Signal_{i} = \frac{\left(\sum_{j=1}^{2} Signal_{ij} \right)}{2} \qquad \text{Equation 4.4}$$

The standard deviation of the mean raw signal of the ith promoter at the jth experimental replicate is calculated according to (Equation 4.5).

$$\sigma_{ij} = \sqrt{\sum_{k=1}^{3} \frac{(Raw\_Signal_{ijk} - Mean\_Raw\_Signal_{ij})^2}{2}} \qquad \text{Equation 4.5}$$

Note that this expression employs the unbiased estimate method where the square root of squared deviations of each measurement from the mean is divided by *n-1* where n is the number of measurements (n=3 in this case) (Kenney, 1962)

To determine confidence intervals for whether a fragment is a promoter or not, we assume that the measured firefly/renilla luciferase signal from a promoter inherently contains some background activity. This background activity is set as the activity of the negative control fragment that gives the highest firefly/renilla luciferase signal. Since we assume that we can measure the background independently, the errors in their measurement will be independent as well. This means that the standard deviation (σ) of a real promoter activity will be the sum of the standard deviation of the measured signal and the standard deviation of the background (Equation 4.6) (Abramowitz, 1972).

$$\bar{\sigma}_{ij} = \sqrt{\left(\sigma_{Signal_{ij}}\right)^2 + \left(\sigma_{Background_{ij}}\right)^2} \qquad \text{Equation 4.6}$$

Equation 4.6 will give the standard deviation of the mean signal for each promoter in one experiment. The standard deviation of the replicated measurement will be simply the sum of the standard deviations of each measurement in each experiment (Equation 4.7).

$$\sigma_i = \frac{\sqrt{\sum_{j=1}^{2} (\bar{\sigma}_{ij})^2}}{2} \qquad \text{Equation 4.7}$$

If we assume that all the errors in measurements follow a Gaussian distribution, then our confidence intervals to decide whether a fragment is a promoter or not, i.e. showing an activity higher than the background are as follows (Abramowitz, 1972);

- $\text{Mean\_Signal}_i \pm \sigma_i$ → 68.3 % confidence that it is a promoter
- $\text{Mean\_Signal}_i \pm 2\sigma_i$ → 95.4 % confidence that it is a promoter
- $\text{Mean\_Signal}_i \pm 3\sigma_i$ → 99.7 % confidence that it is a promoter.

4.3 Cloning of candidate promoter fragments

Each candidate promoter fragment cloned to pGL3-basic plasmid was approximately 300 bp in size including 250 bp upstream and 50 bp downstream of the annotated TSS. Here, only the activity of extended core promoters was investigated since proximal promoter elements (especially between 500 bp to 1000 bp upstream of TSS) often have potential silencer elements (Cooper et al., 2006). Although core promoter region is typically accepted as 100 bp upstream to 50 bp downstream of TSS, here a longer fragment size was used to include possible activatory binding sites to increase the chance of obtaining a promoter activity from the selected fragments.

The 300 bp putative promoter fragments do not contain the complete proximal promoter region. Therefore, fragments that require enhancer elements located in proximal regions for activation cannot be detected using gene reporter assays. In an attempt to recover the activity of such fragments, 300 bp putative promoter fragments were also cloned to pGL3-enhancer plasmid which carries SV40 enhancer (see map in section 2.1.2.9). SV40 enhancer is a ubiquitously active regulatory element which has binding sites for several common activator transcription factors (see section 1.2.2). SV40 enhancer may recover the activity of candidate core promoter fragments that do not contain activatory proximal promoter elements. Also the differential response of candidate promoters to SV40 enhancer was investigated for further characterization of binding motifs in these putative fragments.

To investigate the activation recovery ability of SV40 enhancer, I cloned 17 putative promoter fragments that were 600 bp long (550 bp upstream and 50 bp downstream of the annotated TSS) to pGL3-basic plasmid and compared their activity to those of 300 bp with or without SV40 enhancer (Figure 4.4) in HeLa S3 cells. Four 300 bp constructs did not give activity, but showed activity when they were cloned in front of SV40 enhancer or when the proximal promoter was included (600 bp construct). Two constructs did not show any activity in any configuration and the remaining 11 constructs showed activity with all types of constructs. This suggested that SV40 enhancer can successfully replace the proximal promoter elements of core promoters.

As mentioned earlier, there are 103 protein coding genes and 177 coding transcripts utilizing different promoters in the 20q12-13.2 region (section 3.1). Primers were designed to amplify extended core promoter regions of 132 coding transcripts for cloning (see section 2.1.1). Of these, 109 were successfully amplified by PCR (see section 2.1.2.1). In total, 74 and 76 fragments were successfully cloned to pGL3-basic and pGL3-enhancer plasmids respectively and 71 candidate promoter fragments were cloned to both pGL3-basic and pGL3-enhancer plasmids.

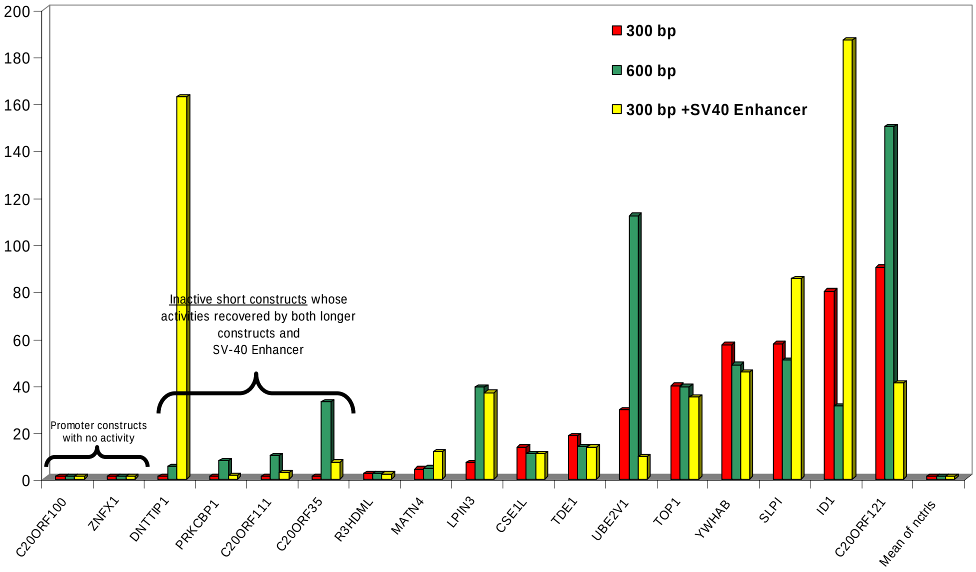


Figure 4.4. Comparison of activities between candidate promoter fragments of 300 and 600 bp in size, and 300 bp fragments cloned together with SV40 enhancer.

4.3.1 Optimisation of Reporter Assays

Selection of Reporter Gene

The use of GFP was considered as a reporter gene first since it can be detected in intact cells using a fluorescent microscope and its signal can be detected using a fluorometer without lysing cells (see section 4.1). Six promoter constructs, whose details are given in Table 4.1, were cloned to pEGFP-1 (Clontech, #6086-1) plasmids and pEGFP-N1 (Clontech, #6085-1), which carries GFP under the control of the strong CMV promoter and was used as positive control.

Four fragments showed activity higher than the background (nctrl2), but the activities between these promoters did not differ significantly, which suggested that the GFP reporter assay can only tell promoter status but it is not sensitive enough to determine the fold difference between promoters of different strength. Therefore GFP reporter assays were abandoned and dual luciferase assays were used instead.

Gene Name	Promoter prediction	CpG island prediction	BAC Clone Name (Ensembl)	Clone coordinates of the cloned region for promoter activity	Coordinates of the cloned region relative to TSS of the gene
ZHX3	yes	yes	DJ796I11	65473..67029	-940..+616
LPIN3	no	no	DJ450M14	11949..12990	-398..+643
ZNFX1	yes	yes	DJ66I20	78596.. 80221	-743..+882
C20orf130	yes	yes	DJ620E11	8596.. 9636	-546..+236
C20orf10	no	no	DJ453C12	89414..90400	-737..+249
nctrl2	no	no	DJ148H17	53971..54733	-

Table 4.1. Details of promoter fragments cloned to pGFP-1 reporter vector. TSS is denoted as +1.

Transfections were performed according to protocol in section 2.1.5 and results are shown in Figure 4.5.

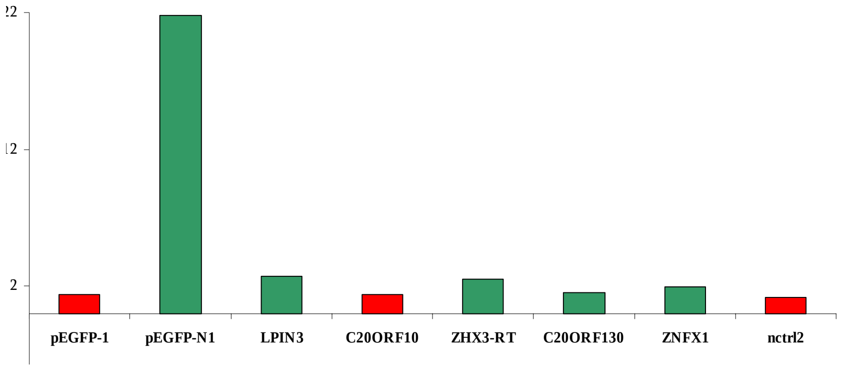


Figure 4.5 Results of transfection with GFP reporter gene. PEGFP-N1 is the positive control plasmid which carries strong CMV promoter. Fragments that did not give significant activity over background (nctrl2) were shown in red.

Optimisation of Dual Luciferase Assays

A key step in reporter assays is the transfection. A series of exprements were carried out to optimise transfection conditions. These included tests for optimising:

- the amount of internal control plasmid
- plate format and
- the number of cells to be transfected per well.

In brief, optimised conditions include the transfection of 50 µg of pRL-SV40 plasmid (internal control plasmid) with 100 µg of pGL3-basic or enhancer plasmids carrying a putative promoter fragments. Transfections were performed in 48-well plate format instead of 96-well format since the former gave better reproducibility between replicates with 2x10^4 cells per well (HeLa S3 or NTERA-D1). The method is detailed in section 2.1.5.

4.4 Cell lines

4.4.1 HeLa S3

HeLa S3 is a sub-clone of parent HeLa cell line which is a cervical carcinoma cell line. HeLa S3 is an epithelial cell line and it grows as monolayers. This cell line is commonly used and is chosen to be used for ENCODE consortium as well. HeLa S3 is also relatively easy to transfect with foreign DNA. Therefore this cell line has been selected for this study.

4.4.2 NTERA2 clone D1

NTERA-2 clone D1 (NTERA-D1) is a pluripotent embryonal carcinoma cell line. It is a sub-clone of NTERA-2 cells which were originally isolated from a lung metastasis of 22 year old patient with primary embryonal carcinoma of the testis. Treatment of this cell line with retinoic acid (RA) results in differentiation to neuronal and other cell types (Mavilio et al., 1988; Pleasure and Lee, 1993; Segars et al., 1993). A testis cell line was desired as the second cell line in this study since testis cDNA libraries contained the highest number of transcribed sequences in the region (Stavrides, 2002). NTERA-D1 is one of the two commercially available testis cell line. The other line (Hs1-tes, ATCC) is a normal human testis cell line. Hs1-tes cells were the first choice but they have a doubling time of ~3 days and they also have finite life span. Therefore this cell line was abandoned due to its incompatibly with large-scale experimental approaches, and NTERA-D1 cell was used instead.

4.5 Transfection Results

4.5.1 Promoter Activities in HeLa S3

As mentioned 74 candidate promoter fragments were successfully cloned to pGL3-basic plasmid. Their promoter activities were assessed using dual luciferase assays in HeLa S3 cells. Out of the 74 constructs, 30 (40.5%) showed an activity with 99.7% confidence.

Figure 4.6 shows background subtracted activities of the 74 (candidate) promoters as calculated in section 4.2.2. In Figure 4.7, confidence levels are shown for each promoter; each bar represents the background subtracted mean signal in units of its standard deviation. In this study, a candidate fragment was accepted as a promoter if its activity is higher than three times of its standard deviation (corresponding to 99.7% confidence level), which means that fragments with an acceptance rate (background subtracted mean signal/its standard deviation) of greater than 3 in these units are accepted as promoters.

The confidence level of a fragment does not depend only on its actual signal, it also takes the uncertainty of the actual signal (its standard deviation) into consideration. This means that the more consistent the activity of a fragment among replicates, the higher a confidence level it will have.

4.5.2 Promoter Activities in NTERA-D1

As in 4.5.1, candidate promoters were also assessed in NTERA-D1 cells where 45 of the 74 fragments tested (60.8%) showed a significant activity. This figure is higher than HeLa S3 cells where only 40.5% of the constructs showed activity which suggests that the testis cells indeed exhibit a diversified transcriptome (Jongeneel et al., 2005). Background subtracted activities of all fragments is shown in Figure 4.8. Also, Figure 4.9 displays confidence intervals for each candidate promoter.

4.5.3 Comparison of promoter activities between the cell lines

Out of 74 candidate promoters, 30 (40.5%) showed activity in both cell lines. There are 16 inactive promoters in HeLa S3 which showed activity in NTERA-D1 cell line. All promoters that showed activity in HeLa S3 were also active in NTERA-D1.

4.5.3.1 Transcript Type and Promoter Activity

Of the 74 candidate promoters, 50 and 24 correspond to representative and alternative transcripts respectively. Table 4.2 summarizes the activity status of both transcript types in the two cell lines. While the promoters of 14 representative transcripts inactive in HeLa S3 were active in NTERA-D1, this was the case for only promoters of two alternative transcripts. This observation is supportive of alternative transcripts being tightly-regulated with restricted expression although the explanation may simply be incomplete annotation.

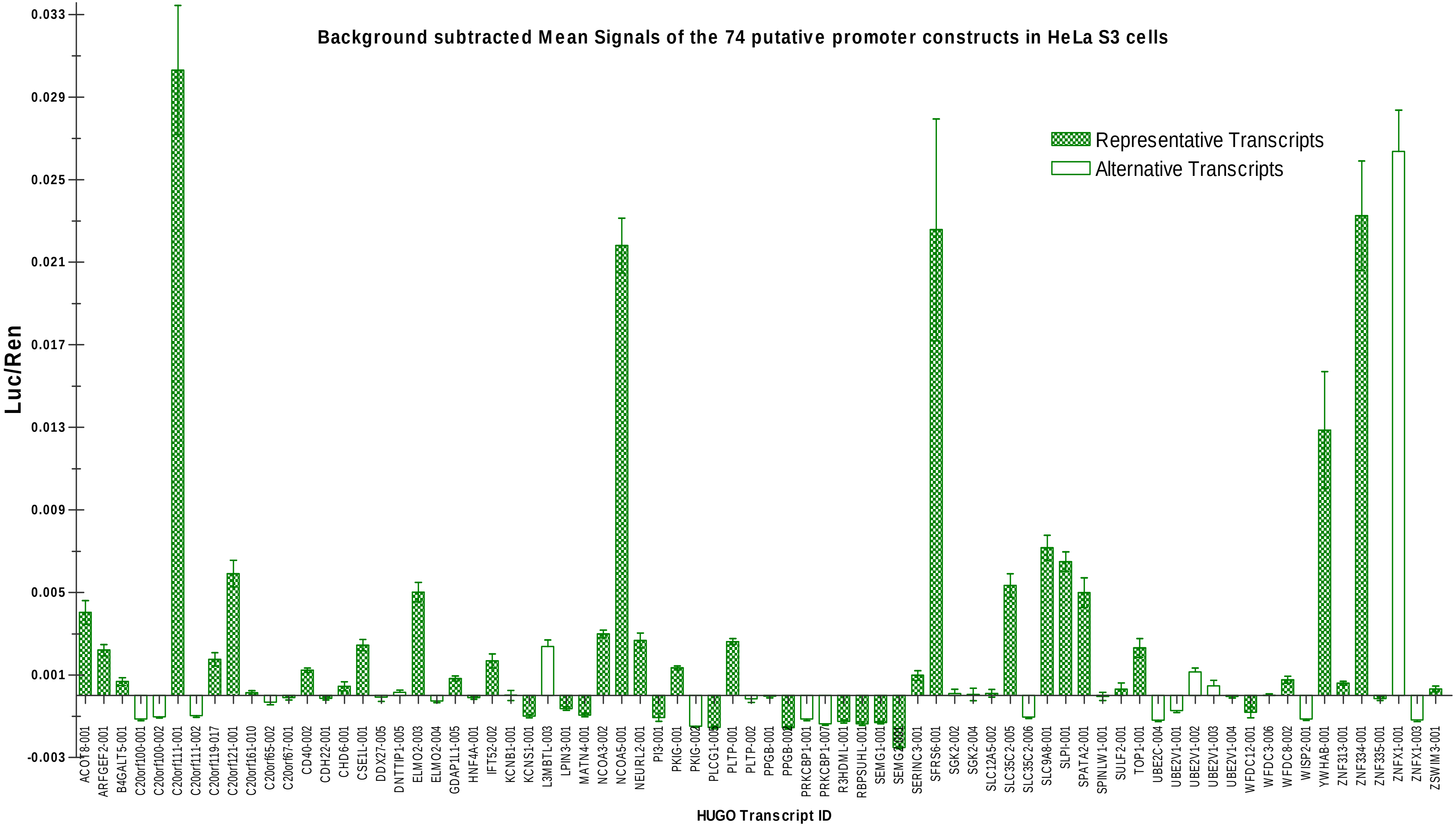


Figure 4.6 Background subtracted Luciferase/Renilla signals of 74 candidate promoters in HeLa S3 cell line. Representative transcripts (RTs) are shown with columns with crossed pattern and alternative transcripts (ATs) are shown with unfilled columns.

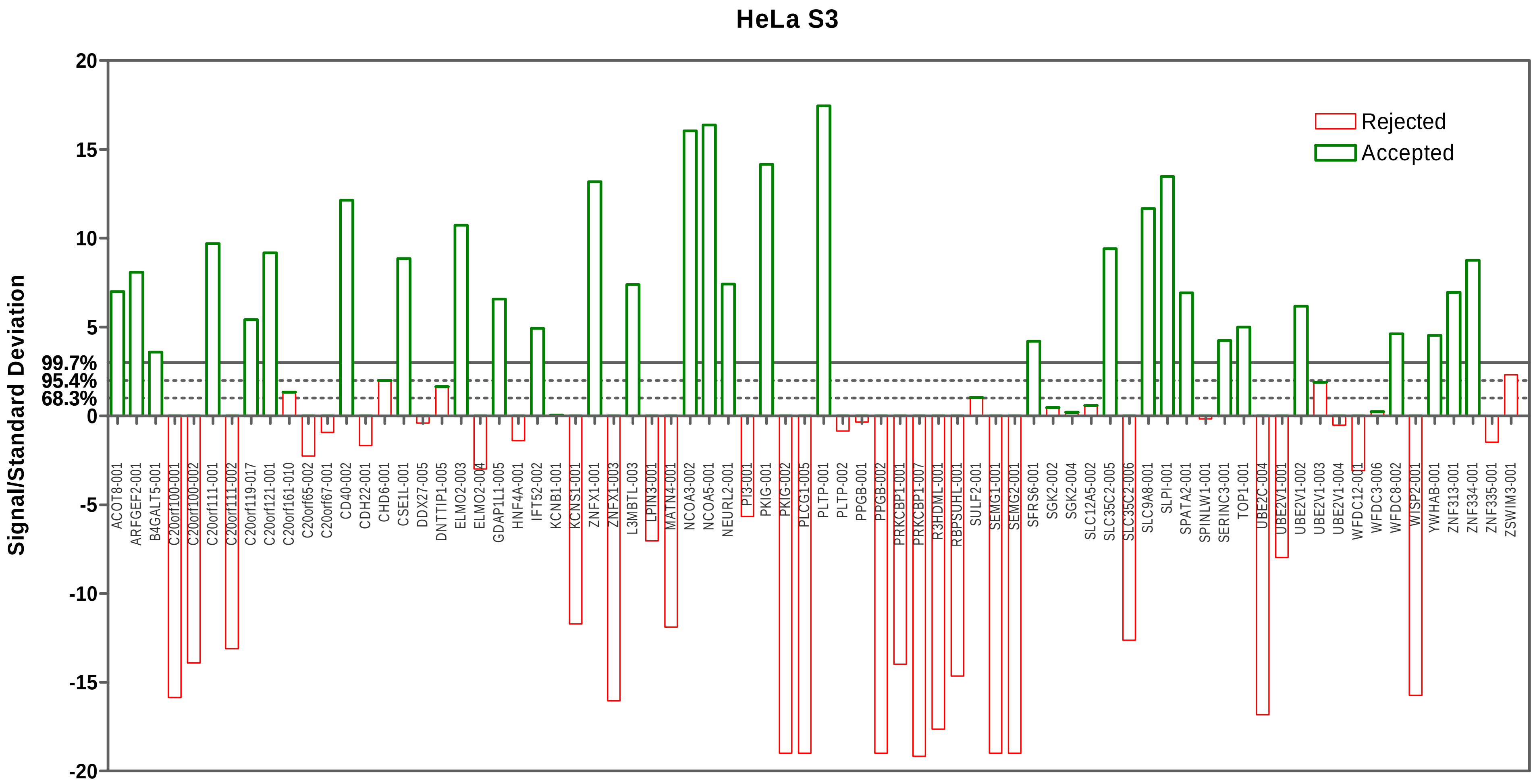


Figure 4.7 Confidence levels to accept (green bars) or reject (red bars) a fragment as promoter in HeLa S3. Here, fragments which shows 3*σ higher than the background are accepted as a promoter and the confidence level for this decision is 99.7% assuming the error distribution follows Gaussian. Dotted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

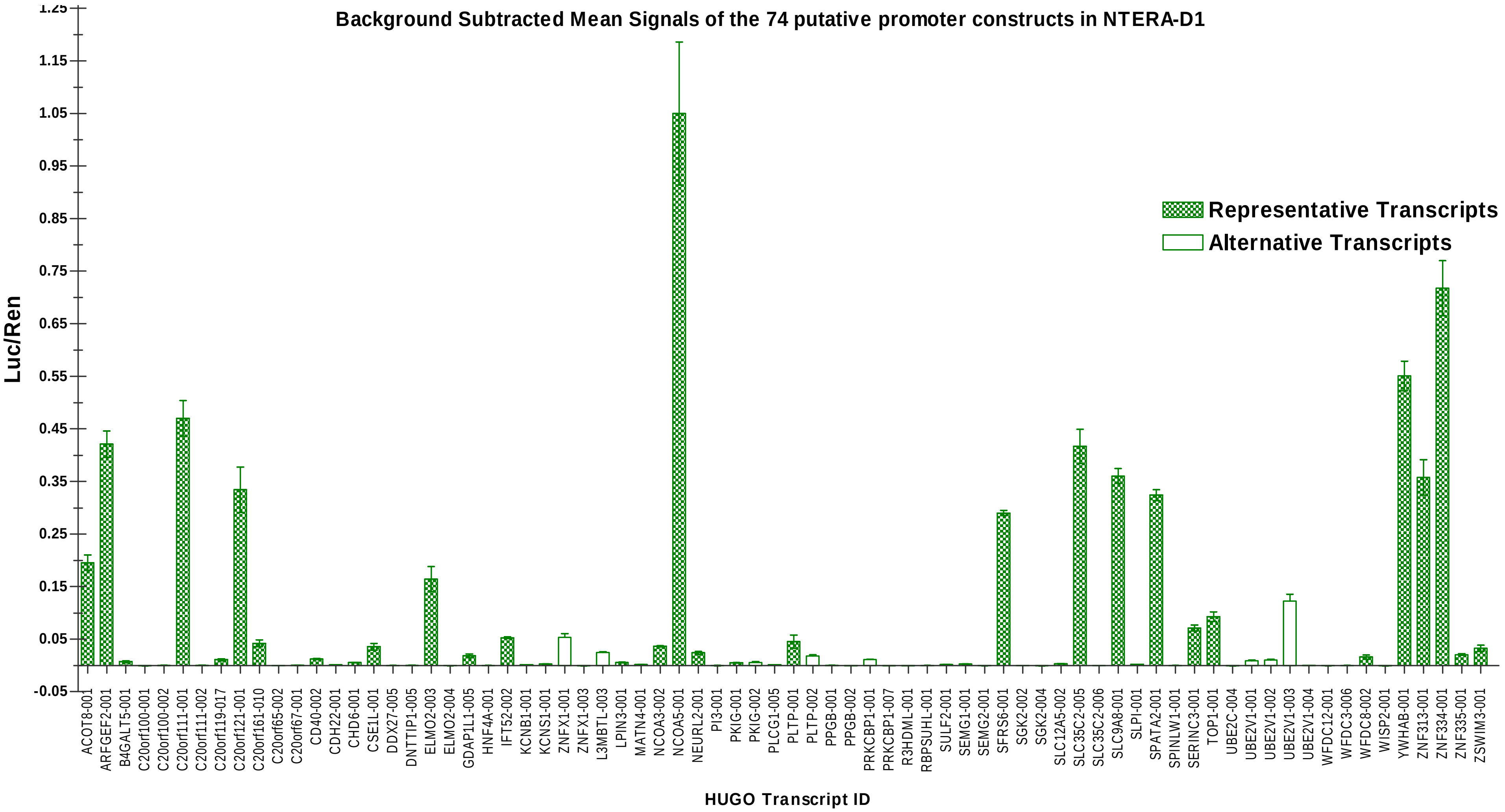


Figure 4.8 Background subtracted Luciferase/Renilla signals of the 74 candidate promoters in NTERA-D1 cell line. Representative transcripts (RTs) are shown with columns with crossed pattern and alternative transcripts (ATs) are shown with unfilled columns.

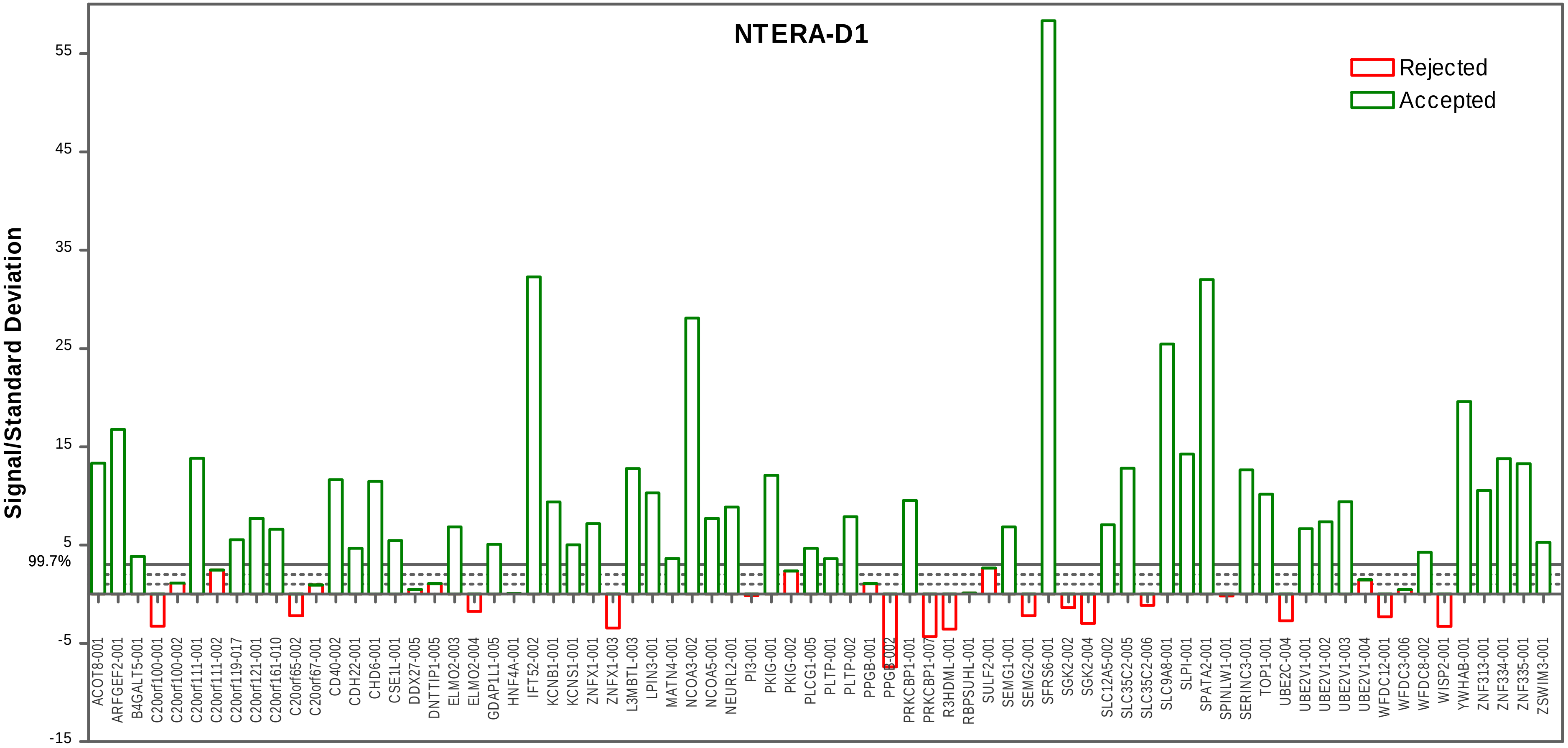


Figure 4.9 Confidence levels to accept (green bars) or reject (red bars) a fragment as promoter in NTERA-D1 cell line. Here, fragments which shows 3*σ higher than the background are accepted as a promoter and the confidence level for this decision is 99.7% assuming the error distribution follows Gaussian. Dotted lines show lower confidence levels.

	HeLa S3			NTERA-D1		
Transcript Type/Promoter Status	Active	No activity	Total	Active	No activity	Total
Representative Transcripts	27 (54%)	25	52	42 (81%)	10	52
Alternative Transcripts	3 (13%)	19	22	5 (29%)	17	22
All Transcripts	30 (42%)	44	74	47 (64%)	27	74

Table 4.2 Promoter activities and transcript types

4.5.3.2 Expression Profile and Promoter Activity

As discussed in section 3.4, both cell lines were profiled with Affymetrix Expression Arrays. Expression profiles of the 50 genes (corresponding to 74 promoter fragments) were correlated with their promoter activity and the results are summarized in Table 4.3. While 70% of the expressed genes in HeLa S3 showed an activity in reporter assays, this number is 89% for NTERA-D1 cells.

	HeLa S3			NTERA-D1		
Promoter Activity	Active	No activity	Total	Active	No activity	Total
Expressed Genes	21	9	30	31	4	35
Not Expressed Genes	6	14	20	8	7	15

Table 4.3 Expression Profile and Promoter Activity of 50 genes in HeLa S3 and NTERA-D1 cell lines.

A higher proportion of expressed genes compared to non expressed genes showed promoter activity in reporter assays. Genes that are expressed but with no promoter activity reported here may correspond to promoters with an epigenetic or distal activation mechanism; their promoter cannot be activated when stripped from its genomic environment. Also, 30% and 53.3% of promoters of non-expressed genes showed a significant activity in HeLa S3 and NTERA-D1 cell lines respectively. This result is expected since reporter assays cannot fully mimic the endogenous status of a promoter since the promoter lacks its histone code and cis-acting distal regulatory signals that might have a repressive effect.

4.6 Promoter Activities in synergy with SV40 Enhancer

In section 4.5, I described the activity of 74 candidate promoter fragments in pGL3-basic plasmids. Out of these, 46 showed promoter activity in at least one of the two cell lines. As described in section 4.3, 76 candidate promoter fragments (includes 71 of the above) were cloned in to pGL3-enhancer plasmid which contains the SV40 enhancer, in order to examine changes in promoter activity due to the enhancer.

4.6.1 Promoters in synergy with SV40 Enhancer in HeLa S3 cells

First, the effect of SV40 enhancer was assessed in HeLa S3 cells. Out of 76, 46 (60.5%) fragments drove the expression of the luciferase reporter gene under the effect of enhancer. The background subtracted mean activities for these constructs are shown in Figure 4.10. As expected, the magnitude of signals obtained from promoter-enhancer constructs is much stronger (approximately six fold) than of those obtained without enhancer. Figure 4.11 displays the confidence levels applied to accept whether a given construct gave or not activity under the effect of the enhancer. Constructs accepted as responsive to the enhancer if they showed activity 3*σ higher than the background signal (representing 99.7% confidence level assuming the errors follows Gaussian distribution).

As mentioned, 71 promoters were tested both with and without enhancer. Out of these 71, 24 (corresponding to 21 representative and 3 alternative transcripts) showed activity with and without enhancer, whereas 20 (16 representative and 4 alternative transcripts) showed activity only in the presence of the enhancer. Also, 4 active promoters did not show any activity in synergy with the enhancer, and 23 promoters (9 representative and 14 alternative transcripts) did not show any activity with or without the enhancer.

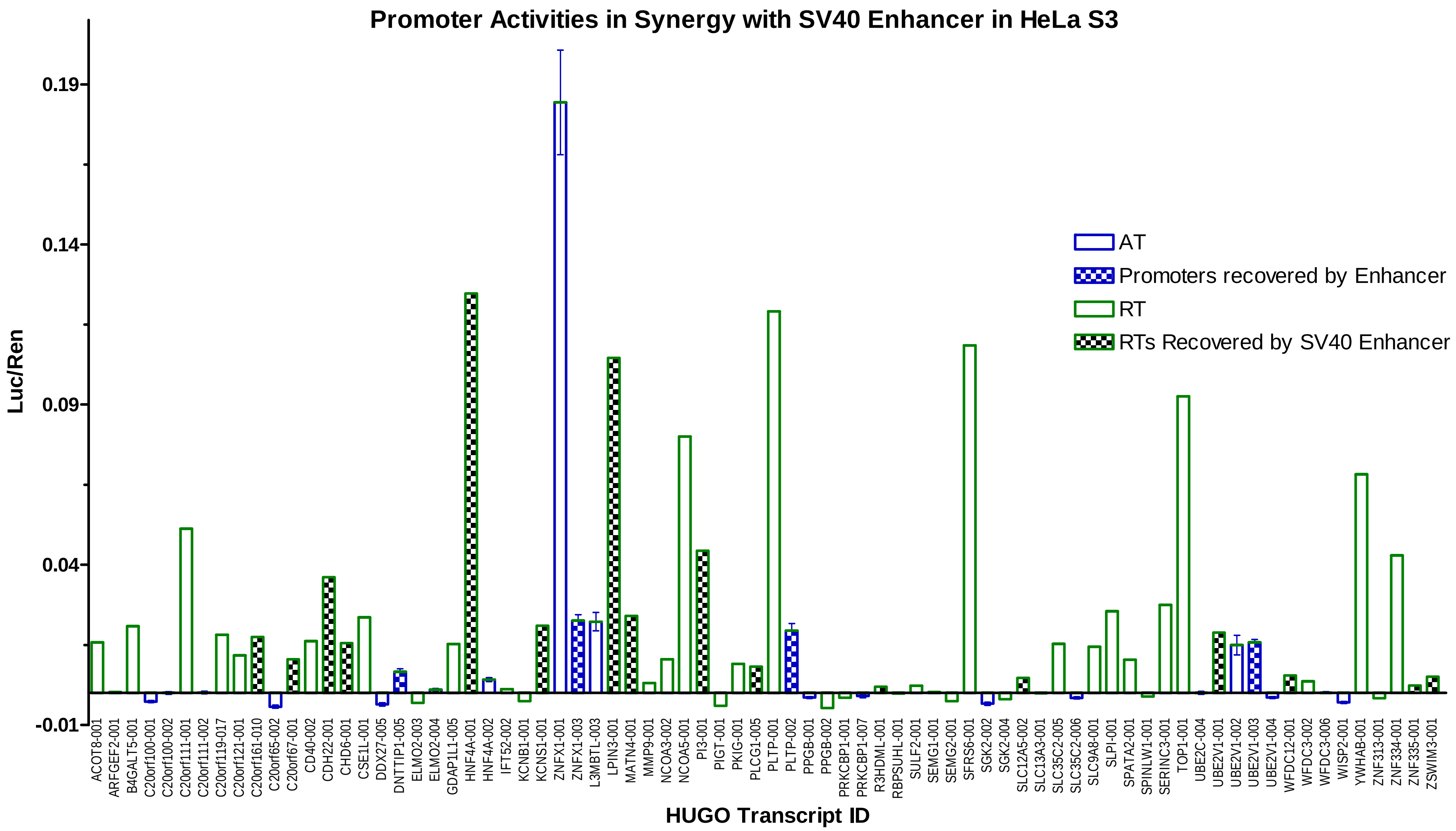


Figure 4.10 Background subtracted activity of 76 candidate promoter fragments in synergy with SV40 Enhancer in HeLa S3 cells. The blue bars represents the promoter activities of alternative transcripts and the bars with check pattern are the fragments which showed activity only in synergic with the enhancer.

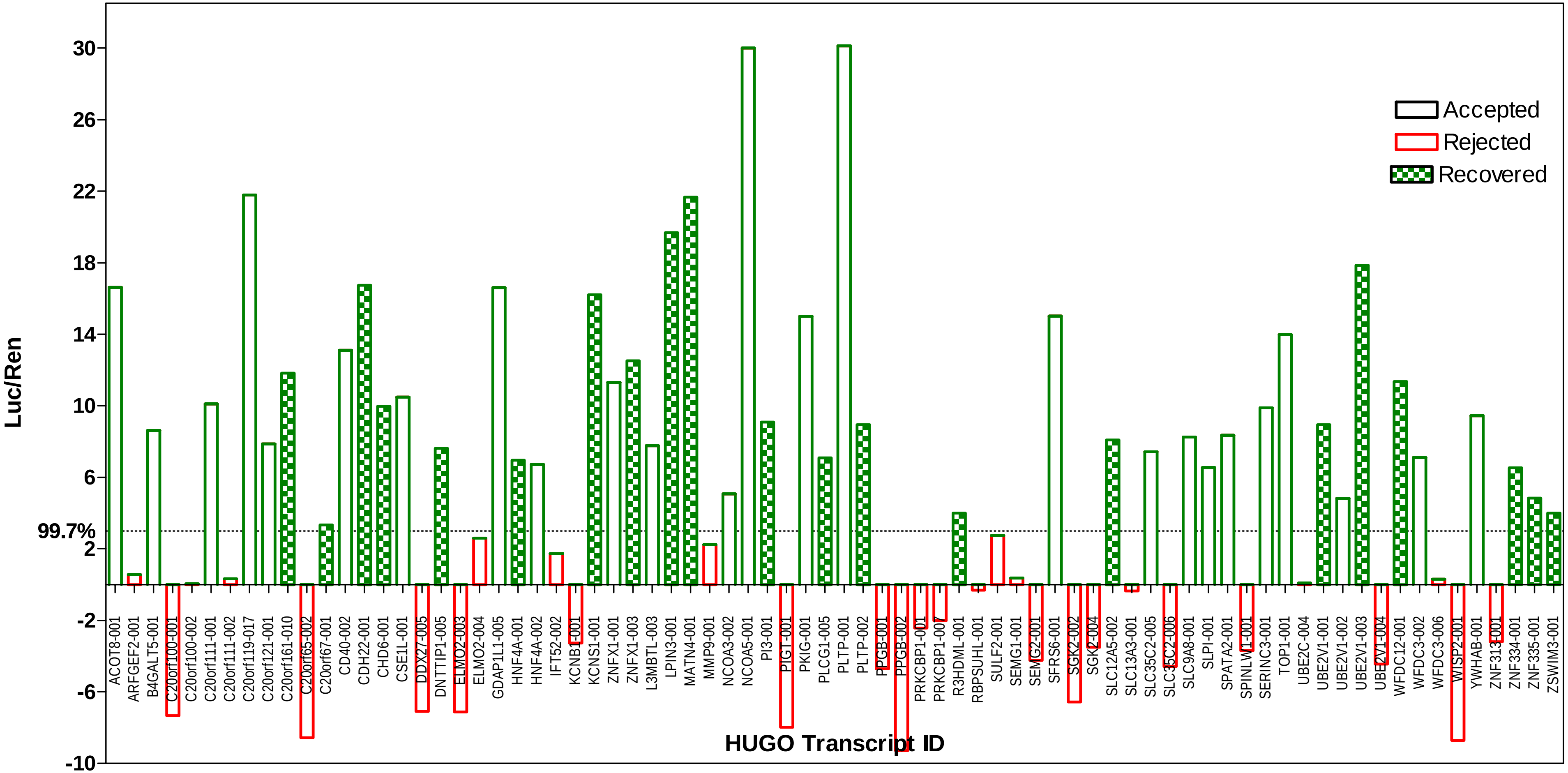


Figure 4.11 Confidence levels to accept (green bars) or reject (red bars) the activation response in HeLa S3 cell line. Here, fragments which shows 3*σ higher than the background are accepted as activated by the enhancer and the confidence level for this decision is 99.7% assuming the error distribution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

4.6.2 Promoters in synergy with SV40 Enhancer in NTERA-D1 cells

The promoter activities of the 76 constructs were also assessed under the effect of the SV40 enhancer in NTERA-D1 cells and 61 (80%) of them showed activity. The background subtracted mean activity of these candidate promoters is shown in Figure 4.12, and Figure 4.13 displays confidence levels of the constructs to the responsiveness to the enhancer.

Of the 76 constructs, 71 were also assessed without enhancer. Of these, 42 showed activity with the enhancer. There are 13 putative promoters (representing 6 representative and 6 alternative transcripts) that showed activity only in synergy with the enhancer and 14 promoters (representing 5 representative and 9 alternative transcripts) that showed no activity with or without the enhancer. Also, 2 constructs (ELMO2-003, WFDC3-006) that showed activity without the enhancer were not active in the presence of the enhancer.

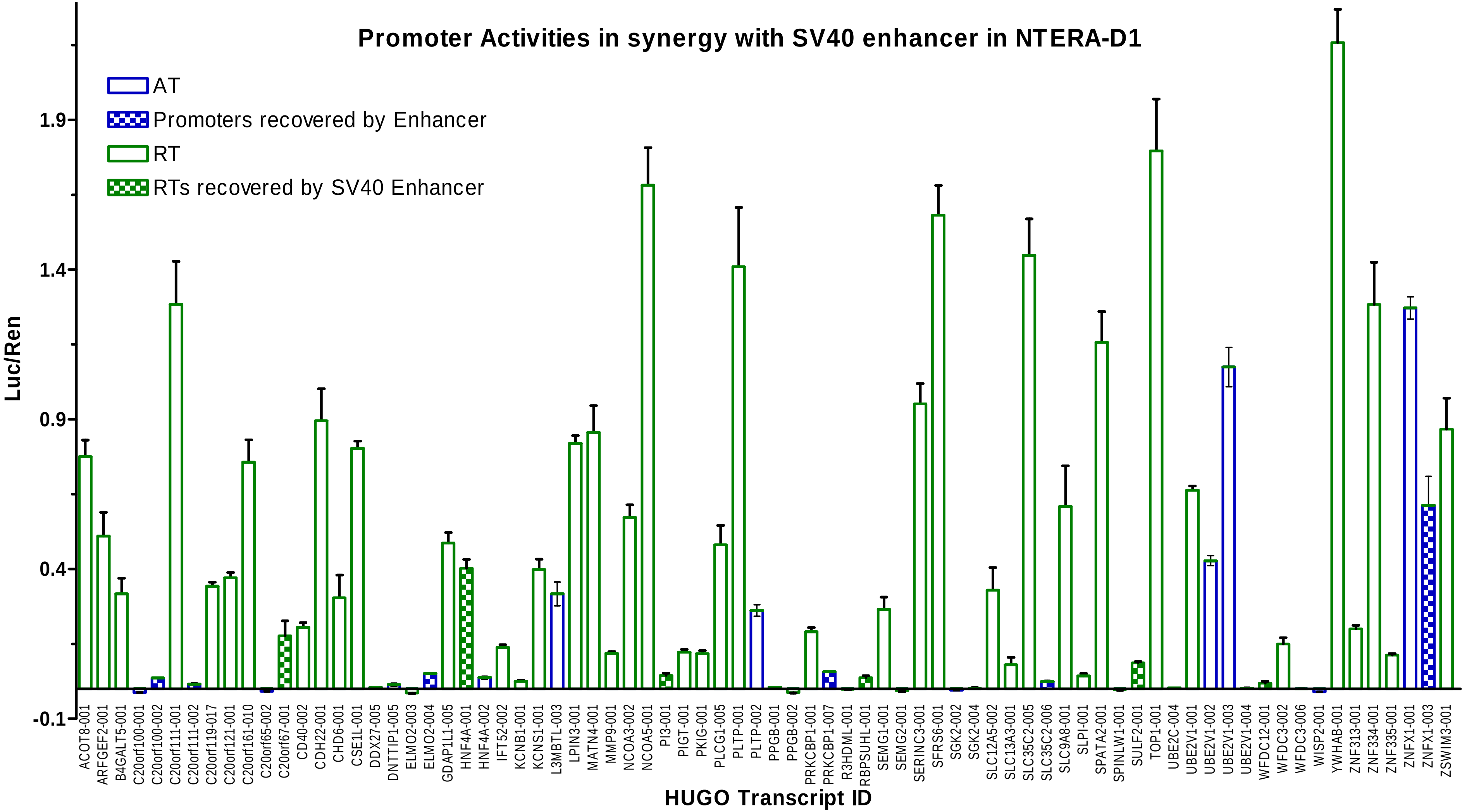


Figure 4.12 Background subtracted activity of 76 candidate promoters in synergy with SV40 enhancer in NTERA-D1 cells. The blue bars represents the promoter activities of alternative transcripts and the bars with checked pattern are the fragments which showed activity only in synergy with the enhancer.

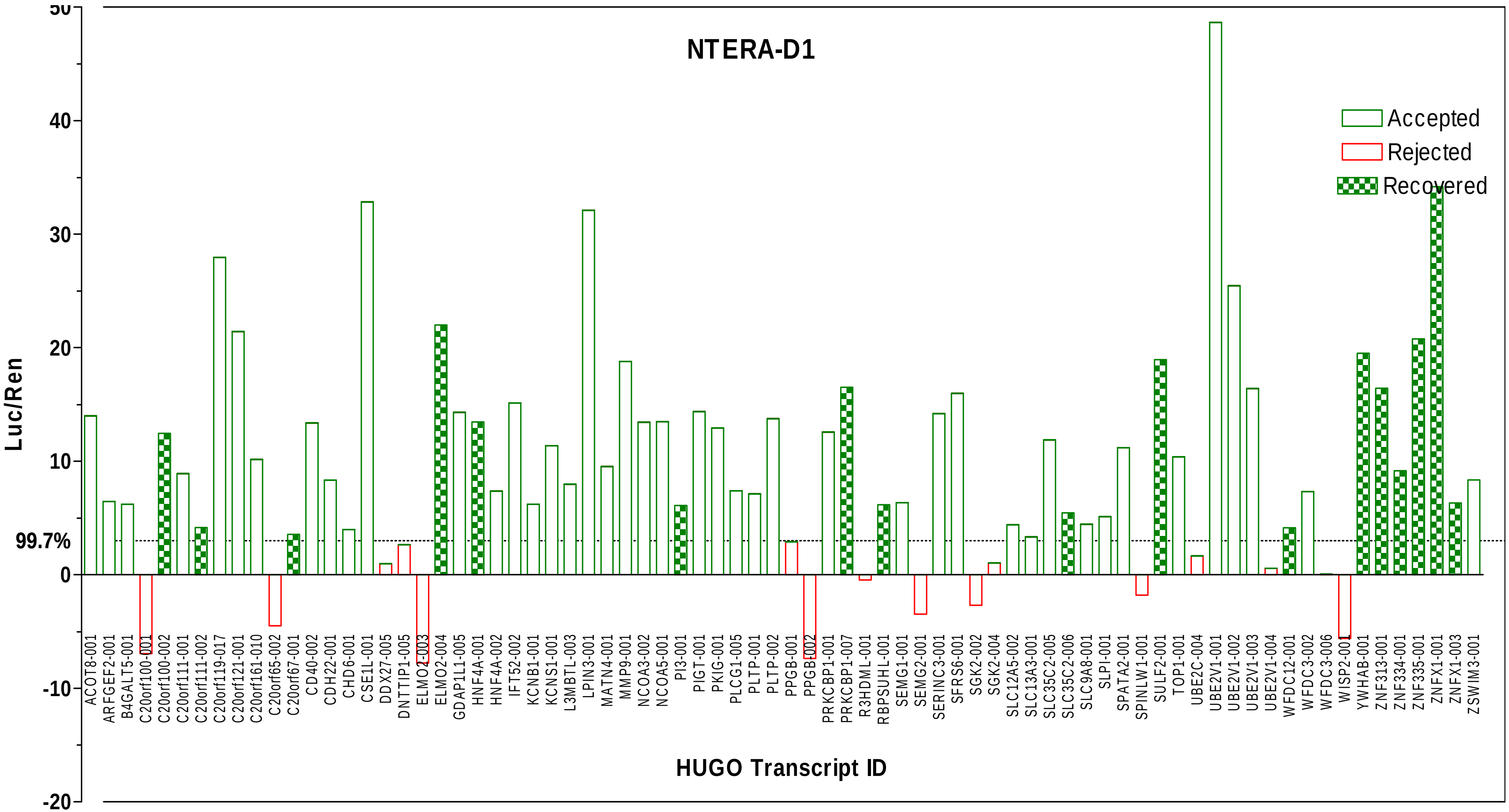


Figure 4.13 Confidence intervals to accept (green bars) or reject (red bars) the activation response in NTERA-D1 cells. Here, fragments which shows 3*σ higher than the background are accepted as activated by the enhancer and the confidence level for this decision is 99.7% assuming the error distribution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

The response status of the candidate promoters to SV40 enhancer is summarized in Table 4.4 where 44 constructs (37 of them belong to representative transcripts) showed in both cell lines activity in synergy with the enhancer and 14 (5 of them belong to representative transcripts) did not. Also, 75% of constructs activated by the enhancer in both cell lines are associated with a promoter (FirstExon) and/or TSS (Eponine) prediction whereas only 14% of the inactive constructs are associated with a promoter and/or TSS prediction. This may mean that prediction programs perform better on sequences carrying higher number of transcription factor binding motifs since the activation by the enhancer mainly depends on this feature.

Construct Summary	Number of constructs activated by enhancer in NTERA-D1	Number of constructs that did not show activity with enhancer in NTERA-D1	Total Number of Constructs
Number of constructs activated by enhancer HeLa S3	**44**	**2**	**46**
Number of constructs that did not show activity with enhancer in HeLa S3	**16**	**14**	**30**
Total Number of Constructs	**60**	**16**	**76**

Table 4.4 Summary of the results obtained with SV40 Enhancer in HeLa S3 and NTERA-D1 cell lines

4.6.3 Sp1 binding sites on 71 promoters

There are 71 candidate promoter fragments, whose activities were assessed both in the absence and presence of SV40 enhancer in each cell line. These constructs are grouped as non-responsive, the ones that did not show any activity in synergy with the enhancer, and responsive. SV40 enhancer contains binding sites for Sp1, POU2F1 and NFKB1 transcription factors that interact with Sp1 transcription factor to mediate promoter activation. So, I searched for Sp1 binding sites in these 71 sequences using the program MAPPER (Marinescu et al., 2005) and attempted to correlate the Sp1 binding site profiles to their activation levels. Table A7 (Appendix A) lists the Sp1 binding site coordinates in the 71 constructs and each construct's response to SV40 enhancer. Out of 71, 13 constructs did not have any Sp1 binding sites.

Since Sp1 acts as an activator in a cooperative manner, a higher number of Sp1 bound to a promoter means a better chance for activation (Anderson and Freytag, 1991). Therefore, I generated the number of putative Sp1 binding sites on each of the 71 candidate promoter fragments activated or not by SV40 in both cell lines to generate frequency distribution plots for the number of putative binding sites and fitted curves for the frequency distribution plots. Figure 4.14 shows that promoters which are not activated in synergy with the enhancer, have a lower number of Sp1 binding sites.

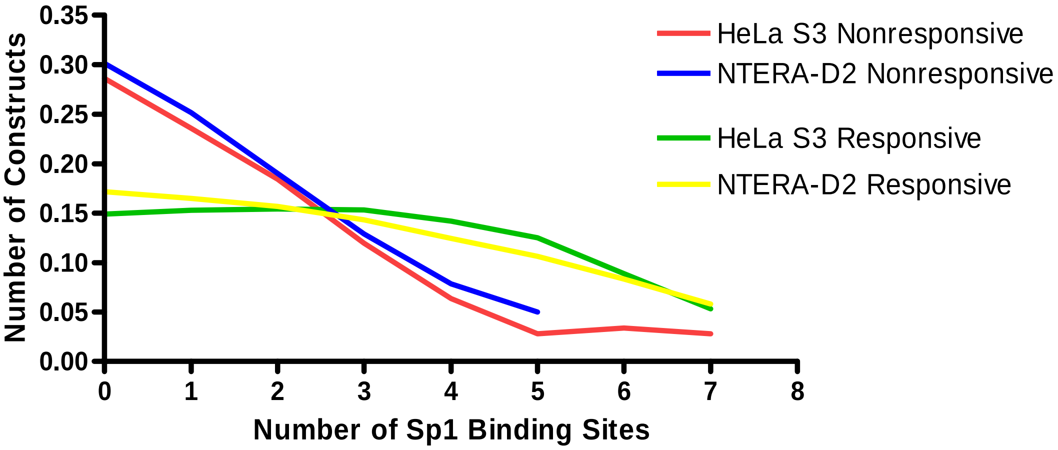


Figure 4.14. Frequency distribution of putative Sp1 binding sites on promoters responding or not responding to SV40 Enhancer.

This is not to say that higher number of putative Sp1 binding sites means activation since the promoters which were activated by the SV40 enhancer have variable number of putative Sp1 binding sites. Interestingly, only a few promoters with a high number of putative Sp1 binding sites (>4) are not activated by the enhancer, which suggests that collective Sp1 binding may indeed increase a promoter's chance for activation by the SV40 enhancer. This is in agreement with Sp1 playing a pivotal role in recruiting several activator proteins due to its diverse interactions (see Table 4.5).

4.6.4 Promoters active only in the presence of SV40 Enhancer

Out of 71 candidate promoters cloned to both vectors (with and without SV40 enhancer), 20 (16 representative and 4 alternative transcripts) showed activity only in the presence of the enhancer in HeLa S3 cells. Of these 20, half belong to genes that were expressed in HeLa S3 according to the Affymetrix Expression Array. In NTERA-D1 cells, 12 candidate promoters (6 representative and 6 alternative transcripts) were active only in the presence of the enhancer of which 9 were expressed in this cell line.

The activities of the promoters recovered by the action of the enhancer are shown in Figure 4.15. in a scale between -100 to 100 in order to compare different data sets. The candidate promoters that did not show activity above background were divided by the lowest activity within that dataset and multiplied by 100 (so that the lowest activity will be -100) and the candidate promoters that showed activity above background were divided by the highest activity within that dataset and multiplied by 100 (so that the highest activity will be 100). Five candidate promoters (denoted by a star in Figure 4.15) whose activities recovered in both cell lines.

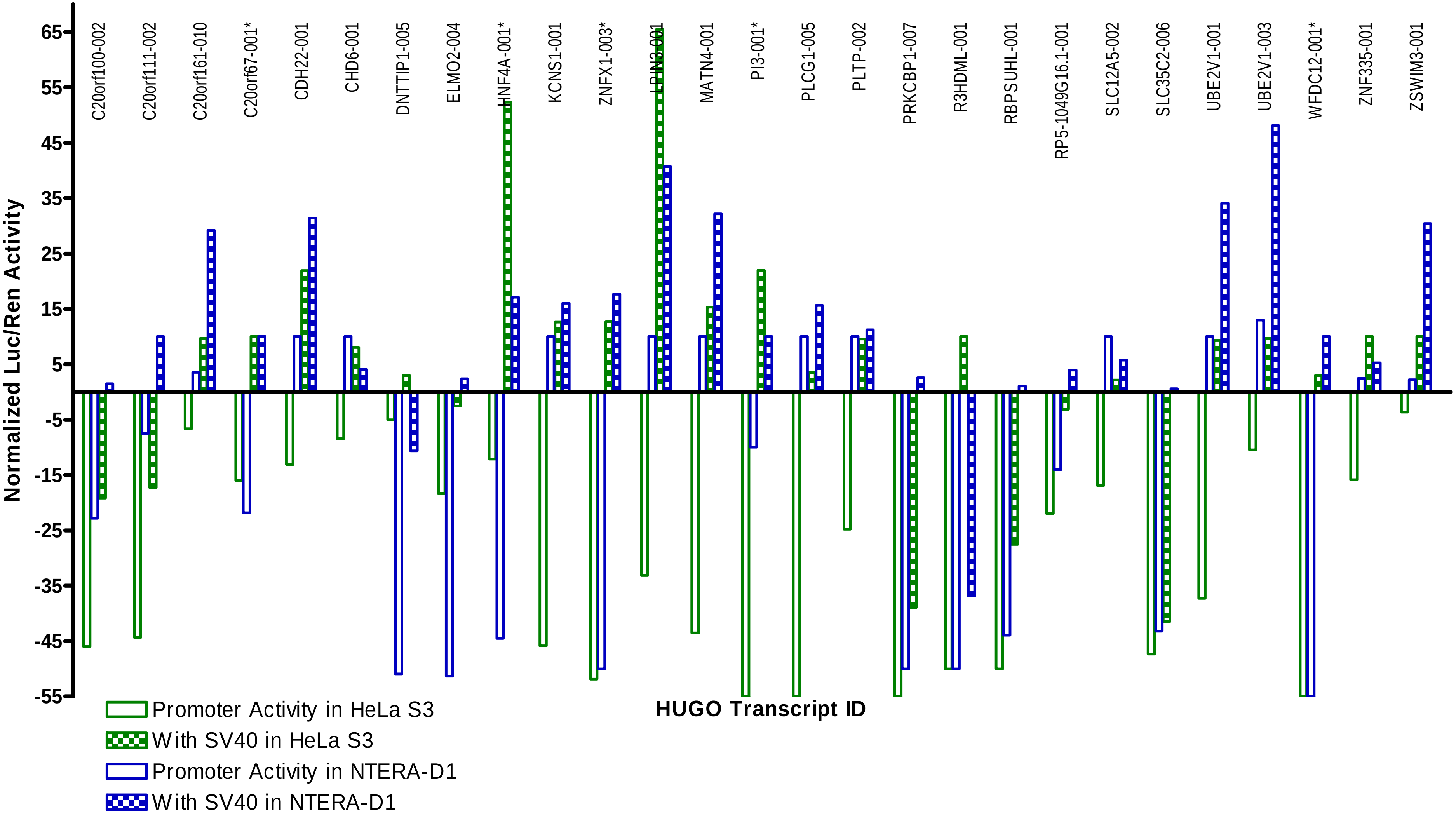


Figure 4.15 Scaled activities of promoters active only in synergy with SV40 enhancer in HeLa S3 and/or NTERA-D1 cells. Scaling was performed by dividing the activity of each promoter to the highest activity within all constructs. Constructs with (*) are recovered by the enhancer in both cell lines.

For a promoter to be responsive to an enhancer, it should contain binding site(s) for the proteins interacting with factors that bind directly or indirectly to the enhancer. The transcription factors that bind to SV40 enhancer are listed in Table 4.5. The 27 candidate promoters active only in the presence of SV40 enhancer were searched with MAPPER for binding sites of CREB, YY1, Sp1,TBP, NFKB1, AP1 and MYC transcription factors for an attempt to detect any correlation between the presence of the binding sites of these factors and the recovery ability of the promoter by the enhancer.

SV40 Enhancer Binding Factors	Interacting Factors
p300	CREB, YY1
Sp1	Sp1, YY1, TBP, NFKB1
AP-1	AP-1, AP-2
TEF-1	TFIID
POU2F1 (oct-1)	Sp1, GR-alpha, TBP
AP-2	AP-2, AP-4
NFKB1	Sp1
MAX	AP-1

Table 4.5. The left column lists transcription factors that have binding sites on SV40 enhancer and right column lists its interaction partners.

Also, the binding site profiles of the promoters not activated by the enhancer were generated and compared. Only putative sites with an 85% confidence level were included in the analysis to avoid false positive hits.

The transcription binding site profiles of the candidate promoters which gave activity only in the presence of the SV40 enhancer were given in Table A8 (HeLa S3) and Table A9 (NTERA-D1) (Appendix A). Also, Table A10 (Appendix A) lists the transcription binding site profiles of the candidate promoters that did not give activity in any cell line. No clear correlation between the binding sites and the ability of the enhancer to recover activity can be derived. Note that presence of a putative binding site in a promoter does not mean that the promoter is under the effect of this protein since the position of the binding site relative to TSS is also crucial. Therefore, rather than looking only for the presence of a putative binding site, I also looked at its position relative to the TSS. Accordingly, out of 27 candidate promoters recovered by the enhancer in at least one cell line, 7 carry a putative CREB protein binding site around 250 bp upstream of the TSS. None of the 12 inactive putative promoter fragments in both cell lines had such binding site in that position. The promoter of C20orf100-001 transcript, which did not give any activity with or without enhancer in any cell line, also contains a putative CREB binding site around 70 bp upstream of the TSS.

YY1 (Ying Yang 1) is a zinc finger protein that can activate, repress or initiate transcription in different gene contexts (Shrivastava and Calame, 1994). It initiates transcription when present around the TSS (Shi et al., 1997). In other positions, it can behave as an activator or a repressor according to the promoter context. Putative binding sites on 71 candidate promoters were plotted with respect to their position relative to TSS in Figure 4.16. There are two peaks at 250 downstream of TSS, and around TSS where YY1 may play a role in transcription initiation.

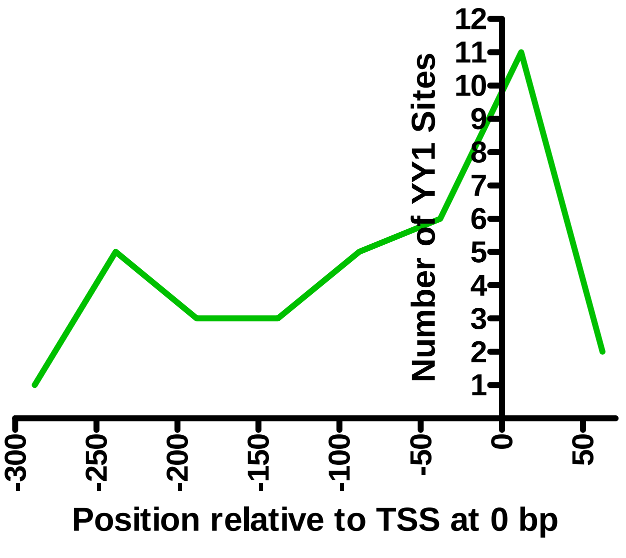


Figure 4.16 Number of putative YY1 binding sites plotted against their position relative to the TSS at 0 bp.

YY1 interacts with Sp1 and p300 that have binding sites on SV40 enhancer (Seto et al., 1993) (Austen et al., 1997). Six promoters recovered by the SV40 enhancer which contain putative YY1 binding sites between 90 to 250 upstream of the TSS, whereas none of the inactive promoters contain YY1 putative binding sites in this region. Since YY1 has an ability to produce sharp bends on the sequence it binds, such positional constraints for it to exert its activation or repression function on the promoter is expected (Kim and Shapiro, 1996).

4.6.5 Promoters Repressed by SV40 Enhancer

Four candidate promoter fragments were repressed by SV40 enhancer in HeLa S3 cells. The promoter of ELMO2-003 was also repressed in NTERA-D1 cells. The sequence of ELMO2-003 was searched with MAPPER (Marinescu et al., 2005) for putative transcription binding sites that may be responsible for the observed silencing effect. Interestingly, there is a ZFP161 (Zinc finger protein 161) binding site located at around 36 bp upstream of the TSS. ZFP161 contains a POZ (Poxvirus and zinc finger) domain found at the amino termini of several zinc finger transcription factors (Bardwell and Treisman, 1994). This domain has been shown to mediate protein oligomerization which subsequently prevents high-affinity DNA binding (Numoto et al., 1999). It can form dimers but also interacts with non-POZ domain containing proteins (Kaplan and Calame, 1997). It acts as a repressor on MYC and β-actin promoters, but also activates the HIV-1 long terminal repeats (LTR). ELMO2-003 has 6 binding sites for Sp1 (one overlapping with ZPF161 binding site) but no binding sites for the activators interacting with SV40 enhancer binding proteins. It is shown that Sp1 and ZPF161 can act together to repress transcription (Kaplan and Calame, 1997). There are 2 other transcripts (SLC12A5-002, SNX21-010) containing ZPF161 binding sites downstream of TSSs (46 and 50 bp respectively), their transcription factor binding motif profile for Sp1, YY1 and NFKB1 activator proteins are given in

Table 4.6. Unlike ELMO2-003, promoters of above two transcripts contain YY1 and NFKB1 binding sites that can act as activators by interacting with SV40 binding factors p300 and Sp1, overcoming the repression effect of ZPF161.

Interestingly a putative binding site cyclic AMP element responsive binding (CREB) protein binding site was found around 140 bp upstream of the TSS site of the three candidate promoters repressed in HeLa cells (active in NTERA-D1 cells) by SV40 enhancer. CREB protein interacts with p300 that has a binding site on SV40 enhancer. It is known that CREB protein mediate repression by forming homodimers, or heterodimers with potential activators to prevent their activation function (Costa and Medcalf, 1996) (Van et al., 2001). In this case, CREB protein binding stabilized by SV40 enhancer on these candidate promoters might mediate recruitment of certain repressor proteins depending on the transcription factor binding site motif profiles of these promoters. Note that there are other candidate promoters carrying a CREB protein binding site around the same location which means that this repression effect is most likely dependant on the nature of the promoter sequences.

Table 4.8 lists some of the promoters whose activities are greatly enhanced by SV40 enhancer. As can be seen from the table, except HNF4A-001, remaining promoters have putative binding sites for at least one activator protein interacting with factors that have binding site(s) on SV40 enhancer.

	HeLa S3			**NTERA-D1**			**Binding Site Coordinates relative to TSS**			
HUGO Transcript ID	**Promoter Activity**	**Enhancer Response Activity**	**Response to Enhancer**	**Promoter Activity**	**Enhancer Response Activity**	**Response to Enhancer**	**Sp1**	**NFKB1**	**YY1**	**ZPF161**
ELMO2-003	0.172	-0.708	repressed	0.143	-1.000	repressed	-226;-182R;-84;-63;35;	0	0	36
SLC12A5-002	-0.169	0.022	recovered	0.003	0.058	active	-202R;-93;-44;33;45;	0	-241	46
SNX21-010	-0.066	0.097	recovered	0.036	0.292	active	-223;-75;-63;29;49;58R	-160	0	50

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.

	HeLa S3		**NTERA-D1**		**Transcription Factor Binding Site Coordinates relative to TSS**				
HUGO Transcript ID	Promoter Activity	Response to Enhancer	Promoter Activity	Response to Enhancer	**Sp1**	**CREB**	**YY1**	**MYC**	**ZF161**
ARFGEF2-001	6.595	-18.510	53.784	14.938	-240R;-116;-38R;-22R;	-167R;	-156;-57R;	-226R;-225;-181;-174R;	
ZNF313-001	1.641	-51.291	39.882	8.968	-107;-87R;-70R;-56R;47;	-125;	0;	80R;81;	
IFT52-002	3.143	-13.918	7.422	6.079	-123R;	-115R;34;			
ELMO2-003	0.172	-0.708	0.143	-1.000	-226;-182R;-84;-63;35;				36

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.

HUGO Transcript ID	Promoter Activity	Response to Enhancer	Promoter Activity	Response to Enhancer	Transcription Factors and their binding sites
LPIN3-001	inactive	0.65	inactive	0.60	1 AP1 binding site at +51 bp 6 Sp1 binding sites NFKB1 binding site at -125 bp
HNFA-001	inactive	0.53	inactive	0.17	No sites found
SFRS6-001	0.30	0.64	0.43	0.70	YY1 binding sites at -146R bp 66R bp MYC binding site at +5 bp
TOP1-001	0.05	0.54	0.10	0.70	CREB binding site at -93R bp
ZNFX1-001	0.97	0.99	0.04	0.64	6 Sp1 binding sites CREB binding site at -188 bp
PLTP-001	0.10	0.79	0.01	0.45	TATA-box at -41 bp NFKB1 binding site at -196R bp MYC binding site at -162 bp 3 Sp1 binding sites
MATN4-001	inactive	0.15	0.01	0.32	YY1 binding site at -91 bp

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.

4.7 Summary

This study investigated 74 putative promoter fragments in HeLa S3 and NTERA-D1 cell lines. Of these, 30 and 47 showed activity in HeLa S3 and NTERA-D1 cells respectively. The activity of promoter fragments is shown in Figure 4.17.

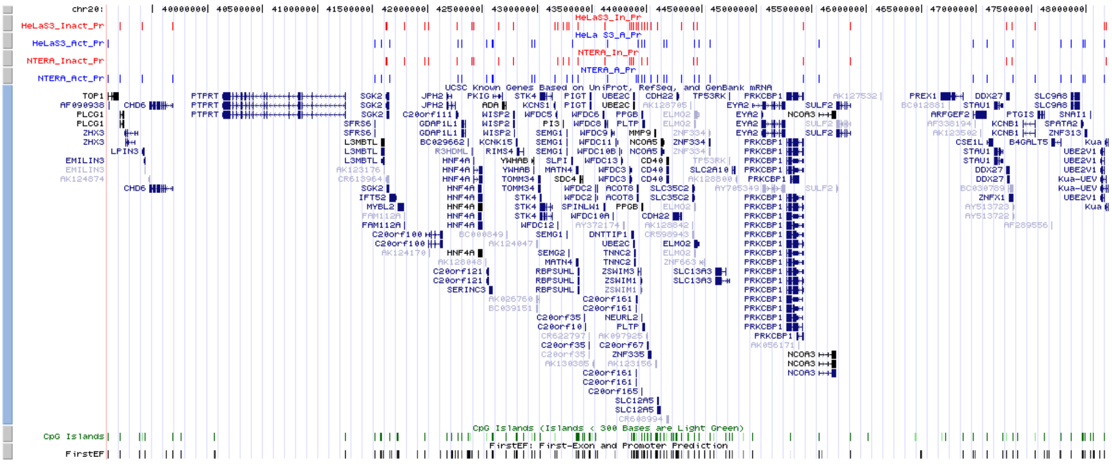


Figure 4.17 Promoter activities of 74 putative promoter fragments in HeLa S3 and NTERA-D1 cell lines. Inactive promoters are shown in red and active promoters were shown in blue. The annotation is taken from UCSC Genome Browser.

Then, 71 of these 74 putative fragments were investigated under the effect of SV40 enhancer in both cell lines. Of the 71 fragments, 46 and 60 fragments showed activity in synergy with the enhancer, and the activities of 16 and 13 putative fragments were recovered using SV40 enhancer in HeLa S3 and NTERA-D1 cells respectively. The activity of putative promoter fragments in synergy with the enhancer is shown in Figure 4.18.

Of the 103 representative transcripts, core promoter activities of 50 were investigated in two cell lines and 35 (70%) showed activity in at least one cell line. When the 50 core promoters were analysed in synergy with SV40 enhancer, 46 (92%) showed activity in at least one cell line. Therefore, the promoter activity of 92% of the investigated putative fragments were verified using two cell lines. In another study, gene reporter assays were employed to verify the activities of 921 putative promoter fragments in Ensembl region using 16 different cell lines (Cooper et al., 2006).

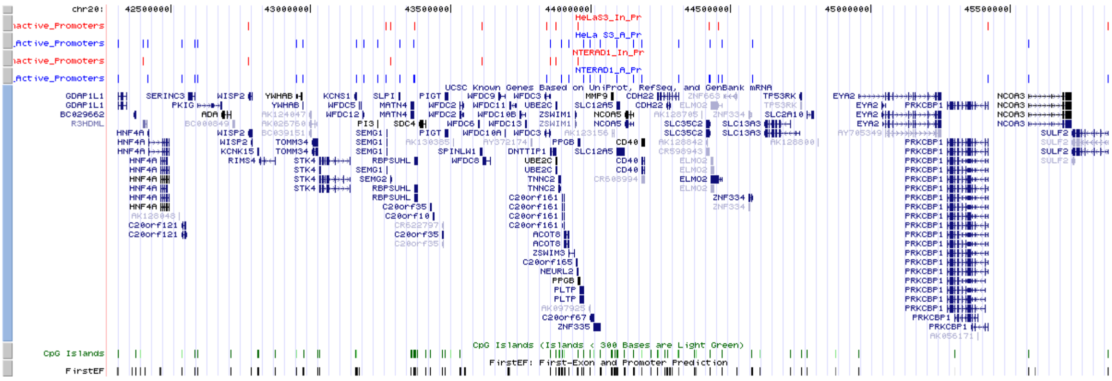


Figure 4.18. Promoter activities of 71 putative promoter fragments in synergy with SV40 enhancer in HeLa S3 and NTERA-D1 cell lines. Inactive promoters are shown in red and active promoters were shown in blue. The annotation is taken from UCSC Genome Browser.

Of these, 387 (42%) fragments showed activity in at least one cell line. Based on the high promoter activity recovery rate obtained in this study with the help of SV40 enhancer in two cell lines, such an approach appears to be more cost-effective to investigate the promoter activity of a putative fragment. If a given genomic fragment has a tissue-specific promoter activity, it may still give activity in a cell line in which it is inactive with the help of a strong enhancer since most enhancers helps recruiting ubiquitously expressed activators onto core promoter regions. Promoters that did not show activity in any configuration (with or without SV40 enhancer) can be further tested by using a fragment longer than 300 bp in a reporter assay. If such fragments are indeed promoters, they might still need their specific enhancer elements located in proximal promoter regions. However, in this case, it is possible that the promoter has a tissue specific expression, therefore more cell lines should be employed as well.

In this study, candidate promoter fragments were selected based on the previous annotation of the region and since 92% showed activity in at least one configuration, it conforms the accurate annotation status of transcripts in this region.

5 ANALYSIS OF 3.5 MB REGION ON 20q12-13.2 USING ChIP-CHIP

Gene regulation is a complex process, requiring a large number of proteins acting cooperatively on specific DNA sequences, which are wrapped with histone proteins. Although instructions for the recruitment of the correct factors onto transcription initiation sites are stored in the DNA sequence, it should be decorated with the correct epigenetic markers to permit functional interactions between the trans acting factors and the nucleosome-bound DNA. Until recently, gene regulation was studied *in vitro* with purified proteins and naked DNA sequences (Kim and Ren, 2006), but such methods cannot mimic *in vivo* the genomic environment of regulatory sequences. Gene reporter assays offer the possibility to investigate regulatory sequences within cells where trans-acting factors are present. However, this approach also lacks the ability to investigate a regulatory sequence in the context of its chromatin environment as well as the effect of possible distal cis-regulatory elements acting on it.

A method developed in 1985 by Solomon et al covalently crosslinks DNA to the proteins bound to it in intact cells using formaldehyde, HCHO, (FA) without damaging the DNA (Solomon and Varshavsky, 1985). FA generates DNA-protein (Varshavsky et al., 1979), RNA-protein (Moller et al., 1977) and protein-protein (Jackson, 1978) crosslinks. In *in vivo* crosslinking, unbound proteins (i.e. not bound to DNA) are not crosslinked to the DNA (Solomon and Varshavsky, 1985). This method is very powerful in understanding the cellular nature of DNA-protein interactions, since it enables to take a snapshot of any given region in the genome together with its associated factors. Solomon et al. used *in vivo* FA crosslinking to provide the first evidence that actively transcribed genes are not free of histone proteins in *D.*

melanogaster (Solomon et al., 1988). Since then, this method rapidly became the method of choice to map proteins onto their genomic targets (Orlando and Paro, 1993) (Gohring and Fackelmayer, 1997) (de Belle et al., 1998) (de Belle et al., 2000) (Wells and Farnham, 2002). Chromatin Immunoprecipitation (ChIP) involves two fundamental steps (i) *in vivo* crosslinking of intact cells, followed by (ii) selective immunoprecipitation of the cross-linked DNA:protein complexes using an antibody directed against the DNA-binding protein of interest (Kuo and Allis, 1999).

In 2000, Ren et al combined *in vivo* FA-crosslinking with DNA microarray technology (ChIP on chip) which emerged as an extremely powerful system to perform large-scale DNA-protein interaction studies (Ren et al., 2000). A number of studies applied this methodology to map common transcription factor binding sites on either a collection of selected sites (Li et al., 2003; Rada-Iglesias et al., 2005) or on a particular genomic region or chromosome (Mao et al., 2003) (Martone et al., 2003) (Horak et al., 2002) or on the entire genome (Kim et al., 2005).

In parallel, *in vivo* FA-crosslinking has been extensively utilized to understand the chromatin context of functional genomic sites. It is well established that histone tails can be covalently modified by acetylation, phosphorylation, sumoylation, ubiquitination and methylation (section 1.4). In 2001, it was shown that in humans 5' ends of transcriptionally active genes are enriched in tri-methylated histone H3 at K4 (H3K4me3) (Litt et al., 2001). Transcriptionally repressed chromatin regions are enriched in tri-methylated histone H3 at K9 (H3K9me3), tri-methylated histone 3 at lysine 27 (H3K27me3) and acetylated histone H4 at lysine 20 (H4K20ac) (Shilatifard, 2006) (Fraga et al., 2005). Moreover, phosphorylated histone H3 at serine 10 was linked to both transcriptional activation and chromosome condensation during mitosis(Cheung et al., 2000), whereas ubiquitinylated histone 2A at lysine 119 and tri-methylated histone 3 at K27 are required for both gene silencing and X-chromosome inactivation (Cao et al., 2002; Plath et al., 2003). The above are examples common histone modifications involved in gene regulation though there are several histone modifications implicated in other cellular processes (Strahl and Allis, 2000).

In this study, I explored the transcriptional activity on the region of interest using ChIP with an antibody against RNA polymerase II. I then investigated the histone code of a 3.5 Mb sub-region of 20q12-13.2 using *in vivo* FA-crosslinking or chromatin immunoprecipitation (ChIP) with antibodies against seven posttranslationally modified histones. Additionally, I employed ChIP with an antibody against CTCF (CCCTC-binding factor) to locate possible insulators within the region. The antibodies used are listed in Table 5.1 in detail. These antibodies were employed successfully in ChIP experiments in this study and will be discussed in detail.

Antibody	Abbreviation	Expected Genomic Regions	Cross Ref.
CTD of RNA polymerase II (phosphorylated at Serine 5)	polII	Actively transcribed regions	1.3
mono-methylated Histone H3 at lysine 4	H3K4me	?	1.4.5
di-methylated Histone H3 at lysine 4	H3K4me2	active promoters	1.4.5
tri-methylated Histone H3 at lysine 4	H3K4me3	active promoters	1.4.5
di-methylated Histone H3 at lysine 9	H3K9me2	Heterochromatin?	1.4.5
tri-methylated Histone H3 at lysine 27	H3K27me3	Silenced genes and heterochromatin?	1.4.5
acetylated Histone H3 at lysine 9 and 14	H3Ac	active promoters	1.4.4
acetylated Histone H4 at Lysines 5,8,12 and 16	H4Ac	?	1.4.4
CTCF	CTCF	mainly insulator elements	1.3.2.1

Table 5.1. Working antibodies used in this study.

5.1 Unsuccessful Antibodies

My strategy was to employ a battery of antibodies to characterise promoter elements in depth. However, many of the antibodies proved difficult to work with and lead to unsuccessful results. Note that there is little or no evidence in the literature for most of these antibodies to have been successfully used in ChIP studies in humans. ChIP experiments were performed with antibodies recognizing TBP (Upstate, #06-241; Santa Cruz Biotechnology, #sc-273; Abcam, #ab818), TAF1 (Santa Cruz Biotechnology, #sc-735; Abcam, #ab14211) and TAF5 (Santa Cruz Biotechnology, #sc-743) proteins. Their use could have helped to differentiate different transcription initiation complexes assembled on promoter sequences (section 1.3.1) and correlate this with the activity profile and sequence characteristics of promoters. Likewise, an antibody recognising YY1 protein (Santa Cruz Biotechnology, (H-10), #sc-7341) was tested. YY1 can bind to initiator elements on TATA-less promoters (Usheva and Shenk, 1996) and have the ability to produce sharp DNA bends like TBP (Kim and Shapiro, 1996), which makes it a good candidate for a TBP substitute in TBP-free initiation complexes (TFTC). I also used an antibody recognizing Sp1 transcription factor (Upstate, #07-645) (section 1.3.2.1). The rationale was to examine sequence features of promoters (GC% or presence of a CpG island) activated by Sp1.

Antibodies recognizing H3K9me and H3K9me3 were employed in ChIP experiments to detect heterochromatic regions in 20q12-13.2. These experimental failures could be due to badly designed antibodies not being able to bind to epitopes. Alternatively, the region of interest might lack true heterochromatic segments, which may be the case since 20q12-13.2 has the highest gene density along chromosome 20 (see Figure 1.22). Furthermore, the fact that the region does not seem to contain any annotated developmental genes which are often in regions needing silencing also supports this alternative.

5.2 ChIP

Crosslinking is the process of chemically joining two or more molecules by a covalent bond. Crosslinking reagents contain reactive ends to form covalent bonds with specific functional groups (primary amines, sulfhydryls, etc.) on proteins or other molecules. Crosslinking has many useful applications such as solid-phase immobilization, preparing antibody-enzyme conjugates, immunotoxins or other labelled protein reagents. It is also used for modification of nucleic acids, drugs and solid surfaces.

FA (formaldehyde) is a high resolution crosslinking agent since it can bridge distances of 2 Ǻ (Jackson, 1978). It is a reactive dipolar compound where its carbon atom is the nucleophilic centre (see Figure 5.1).

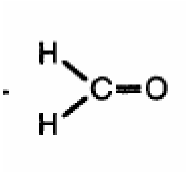


Figure 5.1 Chemical structure of formaldehyde.

Amino and imino groups of proteins (e.g. the side chains of lysine and arginine) and of nucleic acids (e.g. cytosine) react with FA forming a Schiff base (reaction I in Figure 5.2). This intermediate can then react with a second amino group (reaction II in Figure 5.2) and condenses (Orlando et al., 1997). A simple heat treatment is sufficient to reverse the reaction equilibrium toward de-crosslinking (Solomon and Varshavsky, 1985).

FA-crosslinking was the method of choice in this study although there are other crosslinking agents available such as ultra-violet light or laser which have been applied successfully in other studies (Lejnine et al., 1999). FA-crosslinking is commonly used since its crosslinking effect can be fully reversible as well as the chromatin structure being faithfully preserved during and after its application (Orlando, 2000). FA is also proven to be particularly effective in crosslinking lysine and arginine rich histone proteins to DNA (Kuo and Allis, 1999).

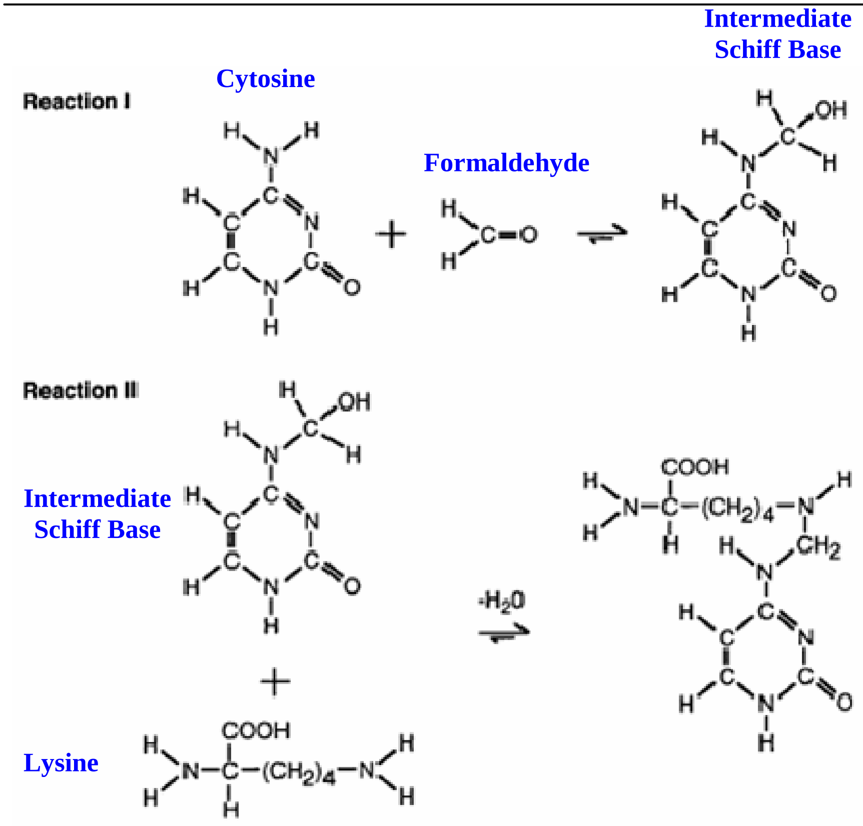


Figure 5.2. Crosslinking of Cytosine to a Lysine by formaldehyde. This figure is reproduced from reference (Orlando et al., 1997).

In this study, HeLa S3 and NTERA-D1 cells were incubated with various concentrations of FA and time courses depending on the nature of the protein targeted to crosslink to the DNA. Although the methodology for crosslinking histone proteins to DNA is well established in the literature (Schubeler et al., 2000) (Bernstein et al., 2005) (Nguyen et al., 2001), crosslinking other DNA binding proteins such as transcription factors requires a great deal of optimization (see section 5.2.1). This may be due to the transient and/or weaker interactions transcription factors have with DNA. Their relative positions within the DNA-protein interaction complex *in vivo* is an additional factor. Finally, the design of antibody epitopes against the protein of interest is also crucial. If the region of the protein where the antibody is supposed to bind (epitope) is occupied by another protein *in vivo*, one cannot precipitate any desired DNA-protein complex with such an antibody irrespective of the efficiency of crosslinking.

A schematic description of ChIP methodology is given in Figure 5.3.

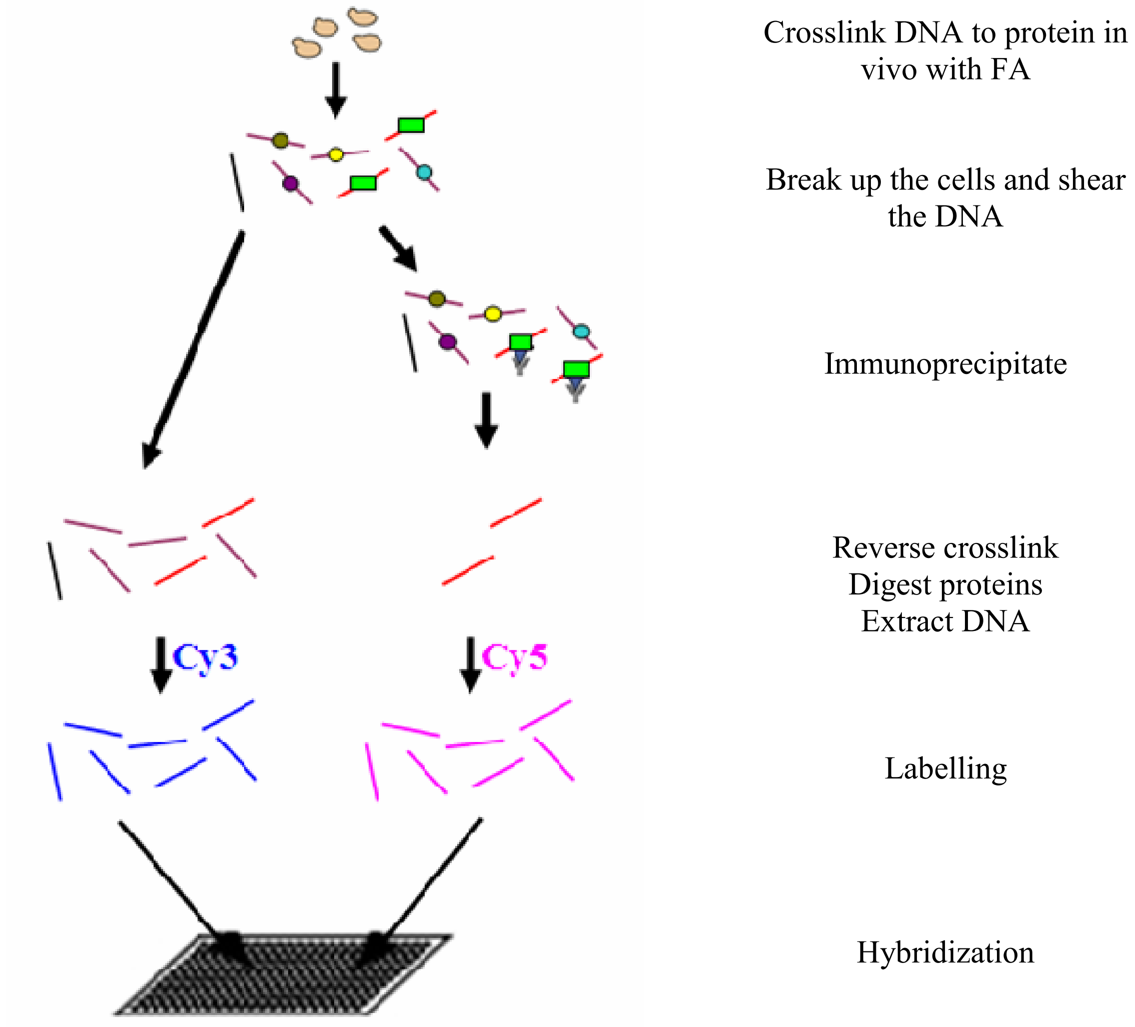


Figure 5.3. Schematic description of Chromatin Immunoprecipitation coupled with DNA microarrays

5.2.1 Optimization of Crosslinking Conditions

An already optimized crosslinking protocol, kindly provided by David Vetrie Lab (WTSI), was applied to crosslink histone proteins to DNA using ChIP, where cells were incubated with 0.37% FA solution for 15 min to crosslink histone proteins to DNA.

Condition Index	Crosslinking Agent	Incubation Time (min)	Crosslinking Agent	Incubation Time (min)
1	1% FA	15	n/a	
2		30		
3		45		
4		60		
5	DMA	30	1% FA	15
6		45		
7		60		

Table 5.2 The concentrations of crosslinking agents and corresponding incubation times to crosslink transcription factors to DNA.

To crosslink RNA polymerase II (polII) to DNA, I incubated cells with different concentrations of FA using different timings. I also used additional crosslinking agents such as Dimethyl Adipimidate.2HCl (DMA) (see section 2.5.2). The crosslinking conditions and timings used in the optimization process are listed in Table 5.2. The crosslinking efficiency was checked using real-time PCR with the primers aligning to the sequence of an active promoter (C20orf121) in 20q12-13.2 (see section 2.6). ChIP experiments were performed as described in section 2.5, then the ChIP sample, which should contain the DNA fragments where the protein of interest is bound *in vivo* (determined by the antibody used to co-immunoprecipitate crosslinked DNA-protein complexes), was used as a template for real-time PCR. The crosslinking conditions that gave the highest amplification with the real-time PCR were chosen for further experiments. Cells that were crosslinked with 1% FA for 60 min (condition 4) showed higher amplification levels than all other timings used. However, the amplification rates obtained from ChIP experiments with polII antibody were ~ten fold lower than those obtained with histone antibodies. For an attempt to improve crosslinking efficiency, I tried an additional crosslinking agent, DMA, which is a strong protein-protein crosslinker, bridging distances of 7.4 Ǻ, and has been successfully applied to crosslink a number of transcription factors with the help of FA (Kurdistani et al., 2002; McCabe and Innis, 2005). DMA is particularly useful if the protein of interest is not directly bound to DNA, most likely because it specifically crosslinks proteins using their primary amine groups (Green et al., 2001). The polII antibody used in this study recognizes the carboxyl terminal domain of largest subunit of polII, but it is known that polII initiation complex *in vivo* achieves its direct contact with the promoter DNA via its TFIID subunit (see section 1.3.1). Therefore, I used DMA to first crosslink protein complexes and then FA to crosslink proteins to DNA to increase crosslinking efficiency. Crosslinking with DMA for 60 min followed by 15 min crosslinking with 1% FA (condition 7; Table 5.2) showed the highest amplification, but similar to those of condition 4. Crosslinking agents such DSG and DSS, which can bridge distances of 6.2 and 8.8 Ǻ respectively, were also tested instead of DMA, but both of them failed to give any amplification. Therefore, condition 4 was used as the simplest most efficient option for crosslinking polII to DNA for both cell lines.

The crosslinking condition for CTCF was kindly provided by David Vetrie Lab, where cells were incubated with 1% FA solution for 15 min for crosslinking.

5.2.2 Immunoprecipitation and subsequent steps in ChIP on chip

After crosslinking, the cell lysate was sonicated to obtain fragments of around 300-500 bp in size. At this point, a small amount of lysate (input chromatin) was taken and stored at -20 ˚C. This sample contains the full genomic content of a cell and serves as reference to estimate levels of enrichment of DNA sequences in later stages of the experiment. Following sonication, the lysate was incubated with the appropriate antibody for each protein to be investigated. Following overnight incubation of the cell lysate with an antibody, DNA-protein-antibody complexes were immunoprecipitated using protein G-coupled agarose beads, where protein G recognizes the immunoglobulin type of the antibody (Rabbit IgG or Goat IgG) used in this study. DNA-protein-antibody complexes were then eluted and crosslinking was reversed by incubation at 65 ˚C. From this point on, input chromatin was also included. For a given antibody, the immunoprecipitated samples and the corresponding input chromatin were first treated overnight with Proteinase K to remove proteins and then DNA was extracted using phenol-chloroform; DNA should include most, if not all, of the sites where the protein of interest was already bound during the crosslinking treatment. After this step, there are several ways to analyse ChIP samples. Real time PCR or quantitative PCR are commonly used methodologies where the end product of ChIP and input chromatin are used as template to amplify targeted regions of interest (Johnson and Bresnick, 2002) (Weinmann et al., 2001) (Wang, Derynck et al., 2004). Although PCR methods are shown to be both useful and sensitive, they are too laborious to scale up for whole genome and limited by the amplification efficiency of the regions of interest (Shieh and Li, 2004). An alternative methodology is the utilization of DNA microarrays to analyse ChIP material. DNA arrays can have either spotted overlapping DNA fragments or long oligonucleotides to cover the stretch of DNA under investigation. DNA arrays overcome the limitations of the PCR-based methodologies, as with current array spotting densities several megabases of DNA can be accommodated on a single chip. Thus, this approach was chosen for analysing 3.5 Mb of the 20q12-13.2 region. The construction of the custom-made DNA arrays is described in section 2.4. Input chromatin was labelled with Cyanine 3 coupled cytosine nucleotide (Cy3-dCTP) analogue, and ChIP material was labelled with Cyanine 5 (Cy5-dCTP) coupled cytosine nucleotide analogue. Approximately 500 ng of input chromatin and 4/5 of a ChIP sample (containing 800 to up to 4000 ng of DNA depending on the antibody used) was used for labelling. After removing unlabelled nucleotides from the samples, the labelled ChIP sample and input chromatin were combined and ethanol precipitated together with Human

Cot-1 DNA. Human Cot-1 DNA is a fraction of DNA consisting largely of highly repetitive sequences obtained from total genomic DNA by selecting for rapidly re-associating DNA sequences during renaturation of DNA (Strachan, 2003). The DNA mixture (containing ChIP material, input chromatin and Human Cot-1 DNA) was resuspended and then denatured. Upon denaturation, unlabelled human Cot-1 DNA will readily associate with complementary strands of the repetitive sequences within the labelled probe, thereby effectively blocking the repetitive parts of the labelled sequences. This step is crucial as highly-represented labelled repeat sequences can cause unreliable data (Mantripragada et al., 2004). This repeat masked material containing input chromatin and ChIP material was then hybridized for 48 h to the custom-made DNA microarray described in section 5.3. Hybridized arrays were then scanned using appropriate lasers for Cy3 and Cy5 dyes and light intensity on each spot was digitalized using a software called ProScanArray® Express (Perkin Elmer LAS Inc.). Spots that have a higher signal from Cy5 (ChIP material) channel are the potential binding sites for the protein of interest.

5.3 Custom-made 3.5 Mb Tilepath Array of human 20q12-13.2

As part of the chromosome 20 SNP discovery project at the WTSI (Spencer et al., 2006) resource of the 2 kb plasmid clones across the entire chromosome was available. These clones were generated from flow-sorted chromosome 20 DNA from four different lymphoblastoid cell lines. Each library was sequenced to a two fold depth. The paired reads (forward and reverse) were used to generate a chromosome 20 tilepath (Zemin Ning at the WTSI) of the clones.

The need to pick clone templates manually from the libraries and PCR amplify all selected clones was the main reason for selecting a sub-region of 3.5 Mb rather than the whole 10 Mb region at 20q12-13.2. This sub-region is from 42,274,163 to

45,850,636 bp and was selected to encompass the most gene rich section. In total, 1875 clones were amplified by PCR and 1795 were successful. Further, four primer pairs were designed to close gaps that contained an annotated TSS, and of those, three were successfully amplified. In total, 1798 amplicons were successful and therefore spotted on the array.

The Microarray Facility at the WTSI spotted the above PCR products to generate a custom-made array (see section 2.4). This array contains 1798 spots and has a resolution of approximately 1.8 kb (2067 ± 483 bp). There are 511 gaps of various sizes ranging from 2 bp to 4182 bp. The distribution of gap sizes is shown in Figure 5.4, where 75% of gaps are less than 800 bp long.

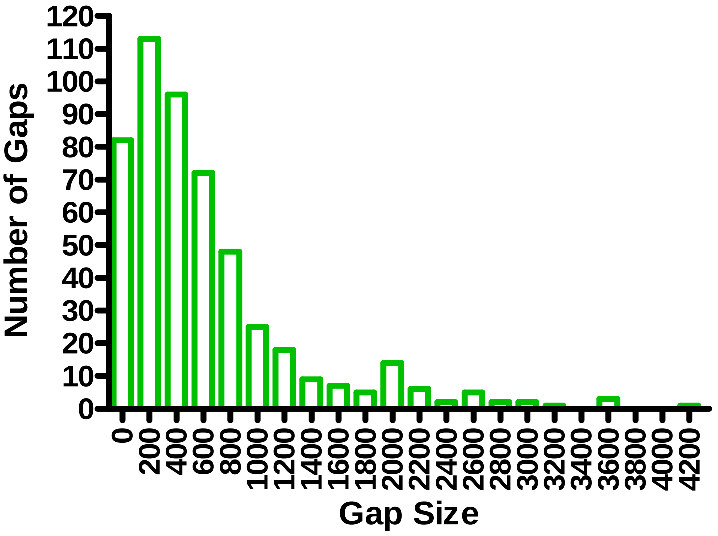


Figure 5.4 Distribution of the gaps in 3.5 Mb tilepath array. The 25% and 75% percentile of the gaps is 157 and 799 bp respectively with a median of 420 bp.

On the spotted array, there are 1723 working spots since 75 spots did not show any hybridization signal. Of these 75 non-working spots, five contain the annotated start site of five different genes (see Table 5.3).

The array contains 72 genes, covering 70% of all annotated genes at 20q12-13.2. Out of the 72 genes, 66 are represented on the array by at least one working spot containing the TSS of one of its coding transcripts. The TSS of 86% (150/175) of the coding transcripts was mapped to at least one working spot on the array. Out of the 91 non-coding transcripts in the 3.5 Mb sub-region, the 5'ends of 79 (87%) were mapped on at least one working spot. There are 26 processed transcripts (transcripts with ambiguous ORF) with the 5' end of 24 processed transcripts being mapped to at least one working spot. Also, the 5'ends of 16 out of the 18 pseudogenes were also mapped to at least one working spot.

As mentioned earlier, the two cell lines HeLa S3 and NTERA-D1 were analysed. Based on the gene expression analysis described in section 3.4, 33 and 39 genes were expressed in HeLa S3 and NTERA-D1 cell respectively. Table 5.3 lists the genes whose TSS is not represented on the array. Therefore, one expects 28 and 35 spots showing positive signals with the antibodies that detect active TSSs in HeLa S3 and NTERA-D1 cells respectively.

Gene Name	Expression in HeLa S3	Expression in NTERA-D1
ADA	P	P
KCNK15	P	A
SLPI	P	A
DNTTIP1	P	P
SNX21	P	P
CDH22	A	P

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.

Gene predictions were also taken into consideration while assessing the gene feature content of the spots on the array (see below). There are 103 gene predictions annotated in VEGA Genome Browser (version 19) and the 5' end of 62 gene predictions were included within the boundaries of at least one spot on the array.

Figure A1 (Appendix A) displays three scanned arrays carrying signals obtained with ChIP experiments performed with antibodies recognizing H3K4me3, H3K4me2 and H3K4me.

5.4 Determining Spot Intensities

For each spot on the array, there are two signals; one comes from the cyanine 3 channel produced by the Cyanine 3 labelled DNA sequences (ChIP material) hybridizing on the spot and the other signal is produced by cyanine 5 channel by the Cyanine 5 labelled DNA sequences (input chromatin) hybridizing on the same spot. If a spot carries target sites for the protein of interest, then the cyanine 3 signal intensity should be higher than cyanine 5 since ChIP material should be enriched in fragments with binding sites for this protein. The input chromatin contains the same number of each genomic fragment. Therefore, in order to find out whether a spot is "enriched", the normalized cyanine 3 signal is divided by the cyanine 5 signal, and a high signal indicates the presence of possible binding sites of the protein of interest. This ratio determines the signal of a spot.

There are three replicates of each spot on each array (technical replicates). Each ChIP experiment for a given antibody was done in triplicate (biological replicates). For each replicate the negative control antibody (Rabbit IgG or Goat IgG) carrying the isotype of the main antibody was included. Let x_{ij} be the signal from ith technical replicate of the jth biological replicate, and bg_{ij} be the signal of the negative control antibody (background signal) from ith technical replicate of the jth biological replicate. The mean signal ($\bar{x}_j$) and background signal ($\overline{bg_j}$) obtained from ith technical replicate of jth biological replicate were both calculated as below (Equation 5.1).

$$\overline{a_j} = \left(\sum_{i=1}^{3} a_{ij}\right) \times \frac{1}{3} \ \textbf{where} \ \mathbf{a} \in \{\mathbf{x}, \mathbf{bg}\} \quad \text{Equation 5.1}$$

Also, the standard error of the mean signal (σ_{X_j}) at each technical replicate is given as;

$$\sigma_{a_j} = \sqrt{\frac{1}{2}\times\sum_{i=1}^{3}(\overline{a_j}-a_{ij})^2\times\frac{1}{3}}\ \textbf{where a}\in\{\mathbf{x,bg}\} \qquad \text{Equation 5.2}$$

The standard error of the mean background signal (σ_{bg_j}) was calculated according to Equation 5.2

It is assumed that the signal coming from the negative control antibody is also included in the actual measured signal since the antibody and the negative control antibody share the same isotype. Therefore, the actual signal (S_j) is calculated by subtracting the background signal ($\overline{bg_j}$) from the mean signal ($\overline{x}_j$) at jth technical replicate. This subtraction aims to eliminate spots that are enriched in a non-specific manner.

To assess the distribution of the Cy3 signals, the raw Cy5 signals (horizontal axis) are plotted against the raw Cy3 signals obtained from ChIP-chip experiments performed with rabbit IgG (negative control antibody) and tri-methylated K4 of histone H3 (H3K4me3) in NTERA-D1 cells (Figure 5.5). The spots that are on the upper left space in panel B (red circle) correspond to those that have a high Cy3 to Cy5 signal and present sites that are potentially enriched specific to the protein of interest.

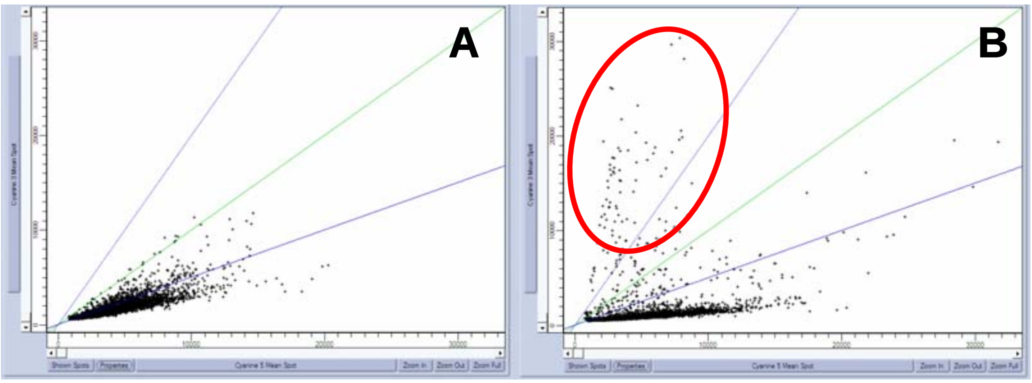


Figure 5.5 The graphs above show raw Cy5 (input chromatin, horizontal axis) relative to the raw Cy3 (antibody, vertical axis) signals for Rabbit IgG (A) and tri-methylated K4 of Histone H3 (B) in NTERA-D1 cells. In graph A, there are no spots on the array, which produce high Cy3 to Cy5 signals, whereas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

The standard error of the actual signal will be the square root of the sum of the squared standard errors of the measured signal (σ_{X_j}) and the background signal (σ_{bg_j}), since the background (noise) signals were measured independently (Equation 5.3) (Abramowitz, 1972).

$$\sigma_{S_j} = \sqrt{\left(\sigma_{x_j}\right)^2 + \left(\sigma_{bg_j}\right)^2} \qquad \text{Equation 5.3}$$

Then, the average of actual signals obtained from each biological replicate (S_j) was taken to calculate the actual mean signal (S) (Equation 5.4)

$$S = \left(\sum_{j=1}^{3} S_j\right) \times \frac{1}{3} \qquad \text{Equation 5.4}$$

The standard error of the actual signal (σ) propagated over the biological replicates can be calculated according to Equation 5.5 (Abramowitz, 1972).

$$\sigma = \frac{1}{3}\sqrt{\sum_{j=1}^{3}\left(\sigma_{S_j}\right)^2} \qquad \text{Equation 5.5}$$

The above allows associating every spot on the array with a signal and standard error for a given antibody. The signal should give the enrichment level specific to the antibody used since the spot signal coming from the negative control antibody (Rabbit or Goat IgG) is subtracted. Due to low resolution of the array, high signals come from one or at most two adjacent spots. For example, all spots that carry a TSS represent the centre of a high signal whereas neighbouring spots show no such high signal (<2) if they have no sequence overlap with the fragment carrying the TSS. Therefore, no peak finding algorithms was necessary to find "enriched" spots (Kim et al., 2005).

Figure 5.6 shows the background subtracted mean signals of the spots for H3K4me3 in NTERA-D1 cells.

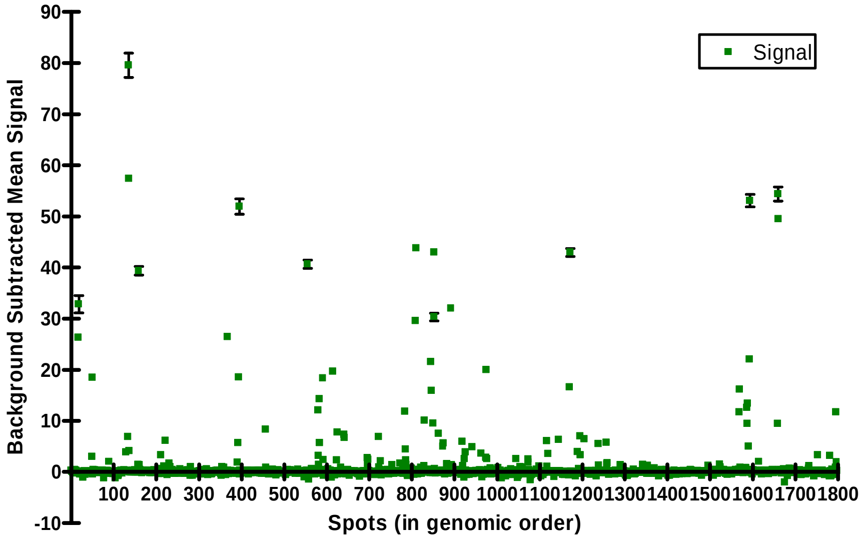


Figure 5.6 The background subtracted mean signals and standard errors obtained from ChIP-chip using antibody recognising H3K4me3 in NTERA-D1 cells. The horizontal axis corresponds to the spots, and are ordered according to the genomic coordinates of the sequences they carry.

The signal does not follow a normal distribution. Therefore statistical analysis valid for datasets with normal distributions could be misleading. The ‘outliers’ most likely represent the biological genomic targets of the protein of interest. However, since the total number of true positives and true negatives is not known, it is difficult to determine the signals that are the true biological targets. Here, I employ an empirical data analysis method to set a threshold to estimate the number of true positives with a false positive ratio below 5%. Signals obtained from spots that are most likely true positives (for example TSSs for polII and H3K4me3) and true negatives (inter-genic regions with no enrichment with any antibody) were validated using real-time PCR (method described in section 2.6.2). The amplification level in real-time PCR was correlated with the enrichment levels in ChIP on chip.

5.4.1 Validation of ChIP-chip enrichments by Real-time PCR

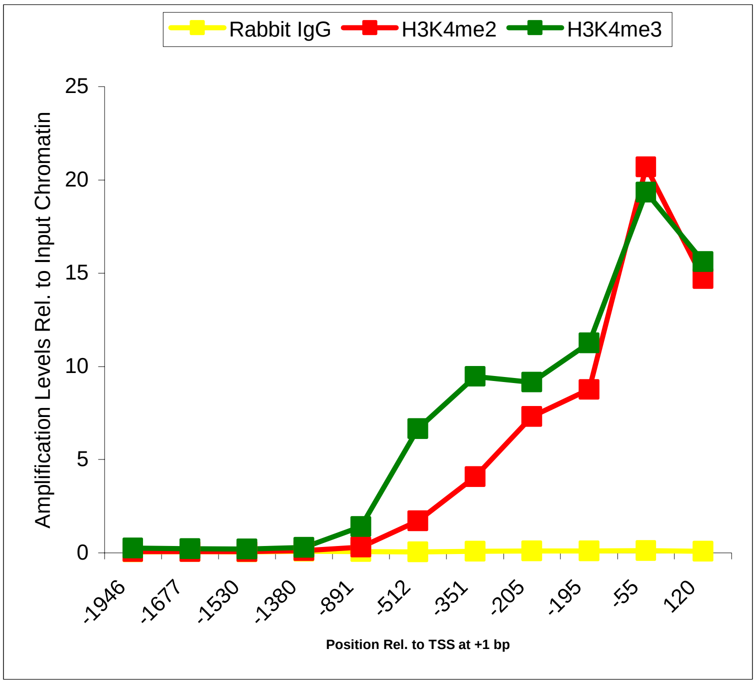
Promoter regions of three genes (*C20orf121*, *ADA*, *ZNF335*) were selected and primers were designed to amplify around 1000 bp upstream to 300 bp downstream of their annotated TSS. Also, primers were designed to amplify the TSSs of *C20orf142*, *TOMM34*, *YWHAB*, *SLPI* and *PIGT*. For H3K27me3 and CTCF antibodies, four regions with a high signal were selected for real-time PCR validations. Finally, 25 inter- and intra-genic assays were designed to amplify regions where no signal was obtained with any antibody used in this study (in either of the cell lines). These will be referred as 'negative real-time PCR controls' All primer sequences are listed in Appendix B. ChIP and input chromatin material obtained from several experiments were quantified using NanoDrop®, and it is estimated that 1:40 diluted input chromatin material contains approximately the same amount of DNA with the ChIP material. Therefore, 0.1 µl of ChIP material or 0.1 µl of diluted input chromatin (1:40) was used for each amplification. Ct values are determined and the amplification level (E) for each antibody is calculated according to the equation given below;

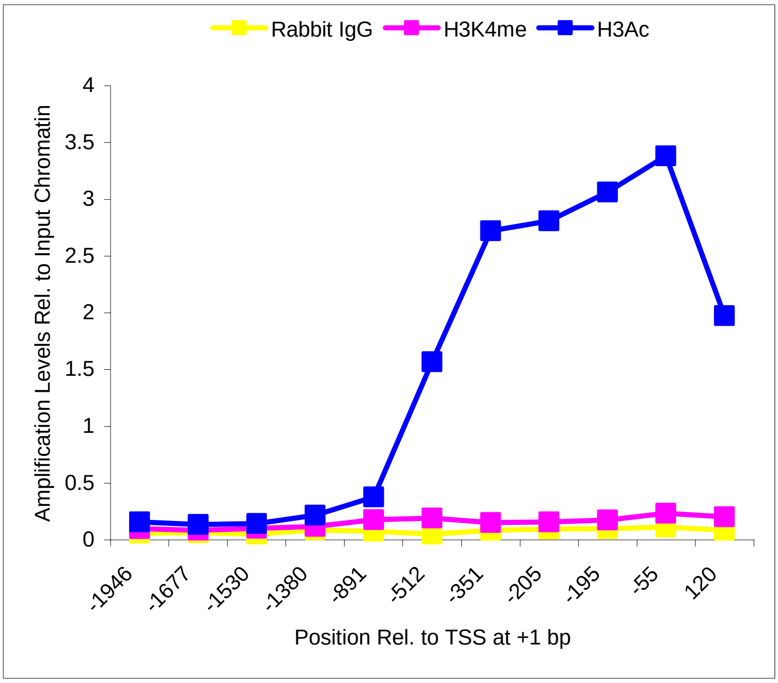
$$E = 2^{(-(Ct(ChIP)-Ct(InputChromatin))}$$

Real-time PCRs were performed with all designed primer pairs (Appendix B) for each ChIP material and the corresponding input chromatin. Promoter region of *C20ORF121* is given as an example for all but H3K27me3 and CTCF antibodies.

Figure 5.7 shows the amplification levels relative to input chromatin using ChIP materials obtained by five antibodies recognizing Rabbit IgG, H3K4me, H3K4me2, H3K4me3 and H3Ac in NTERA-D1 cells respectively. Amplification levels by real-time PCR fully mirror the enrichment level obtained with ChIP on chip, although they are not on the same scale. The spot neighbouring the TSS-carrying spot did not give any signal above background, and no amplification can be seen by real-time PCR.

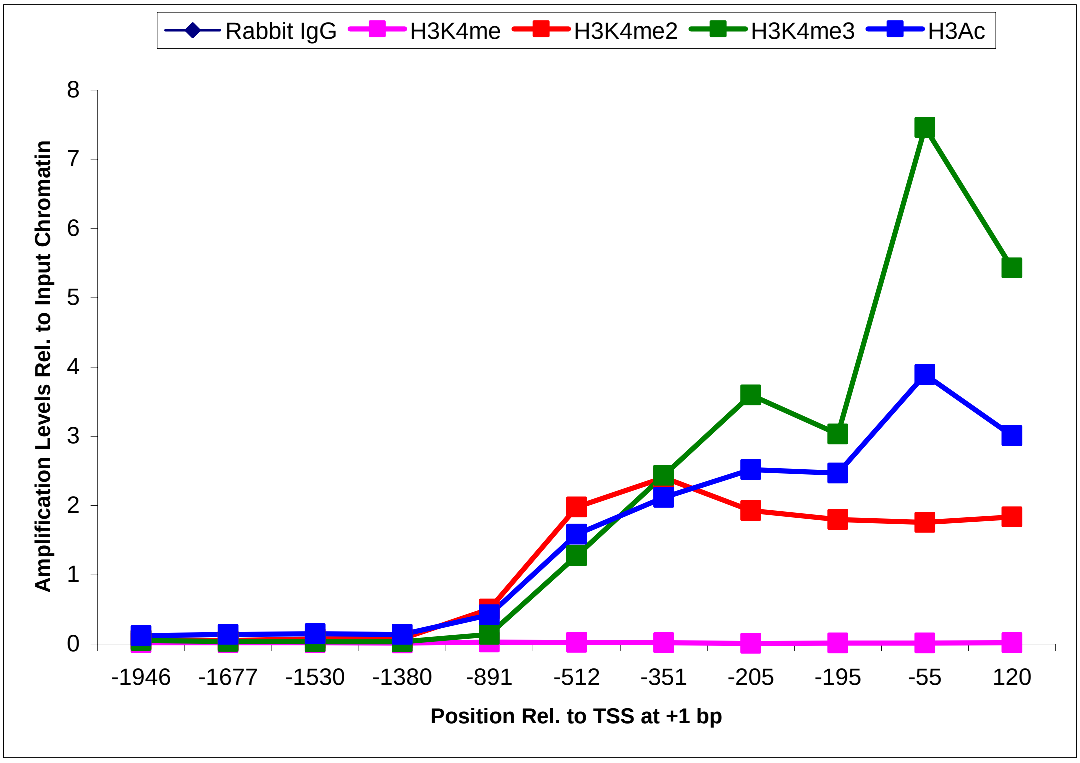
The same analysis was performed with ChIP material obtained with these antibodies in HeLa S3 cells and the amplification levels are shown in Figure 5.8.





Start (bp)	End (bp)	Spot Information	H3K4me	H3K4me2	H3K4me3	H3Ac
-2830	-957		-0.10	0.03	0.21	0.07
-308	435	carries TSS of C20orf121	-0.02	25.77	79.56	7.10

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.

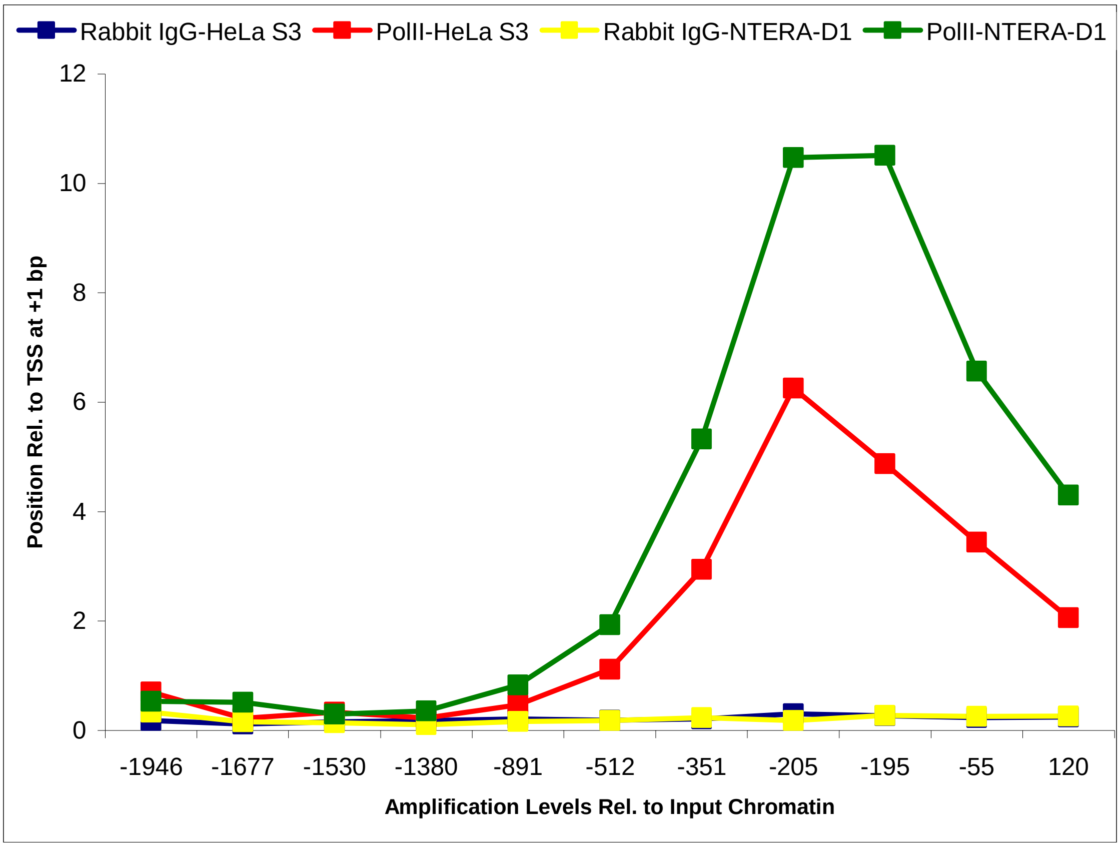


Start (bp)	End (bp)	Spot Information	H3K4me	H3K4me2	H3K4me3	H3Ac
-2830	-957		1.01	0.36	0.40	0.46
-308	435	carries TSS of C20orf121	1.29	17.97	36.72	14.14

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.

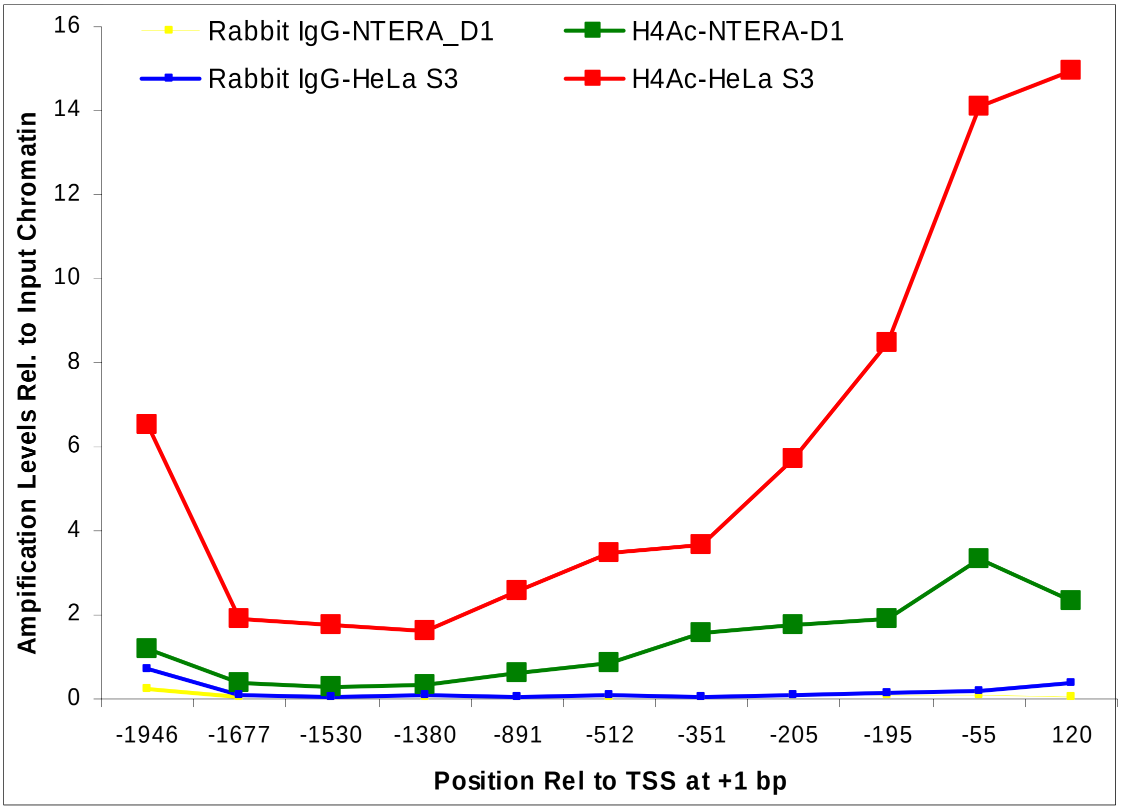
Amplification levels of the ChIP material obtained from HeLa S3 with above listed antibodies relative to input chromatin validated the enrichment levels by ChIP on chip. Note that real-time PCR and ChIP on chip signal intensities follow the same trend in the two cell lines, i.e. higher in NTERA-D1 than in HeLa S3.

The enrichment levels obtained by the antibody recognising RNA polymerase II (polII), H3K4Ac and H3K9me2 were validated with real-time PCR in both HeLa S3 and NTERA-D1 cells. Figure 5.9 – 5.11 show the amplification levels of the *C20orf121* promoter region.



Start (bp)	End (bp)	Spot Information	polII –HeLa S3	PolII – NTERA-D1
-2830	-957		0.10	0.26
-308	435	carries TSS of C20orf121	5.15	7.44

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.

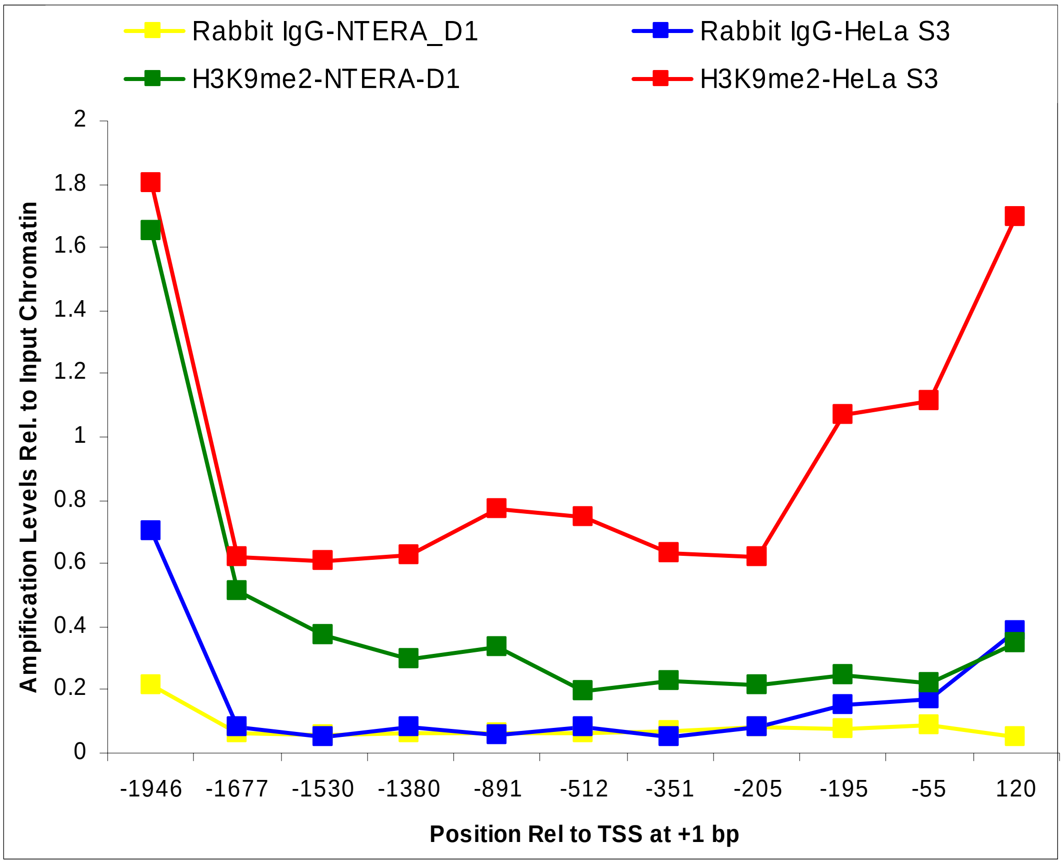


Start (bp)	End (bp)	Spot Information	H4Ac –HeLa S3	H4Ac – NTERA-D1
-2830	-957		0.74	-0.05
-308	435	carries TSS of C20orf121	6.51	0.16

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.

Amplification levels are in concordance with the enrichment levels obtained by ChIP on chip with antibodies above in both cell lines.

In Figure 5.10, ChIP material obtained by H4Ac antibody in NTERA-D1 cells showed ~2 fold amplification relative to the input chromatin whereas no significant enrichment was observed in ChIP on chip experiment. Real-time PCR appears to be more sensitive than ChIP on chip since a low resolution array was used to detect enrichments.



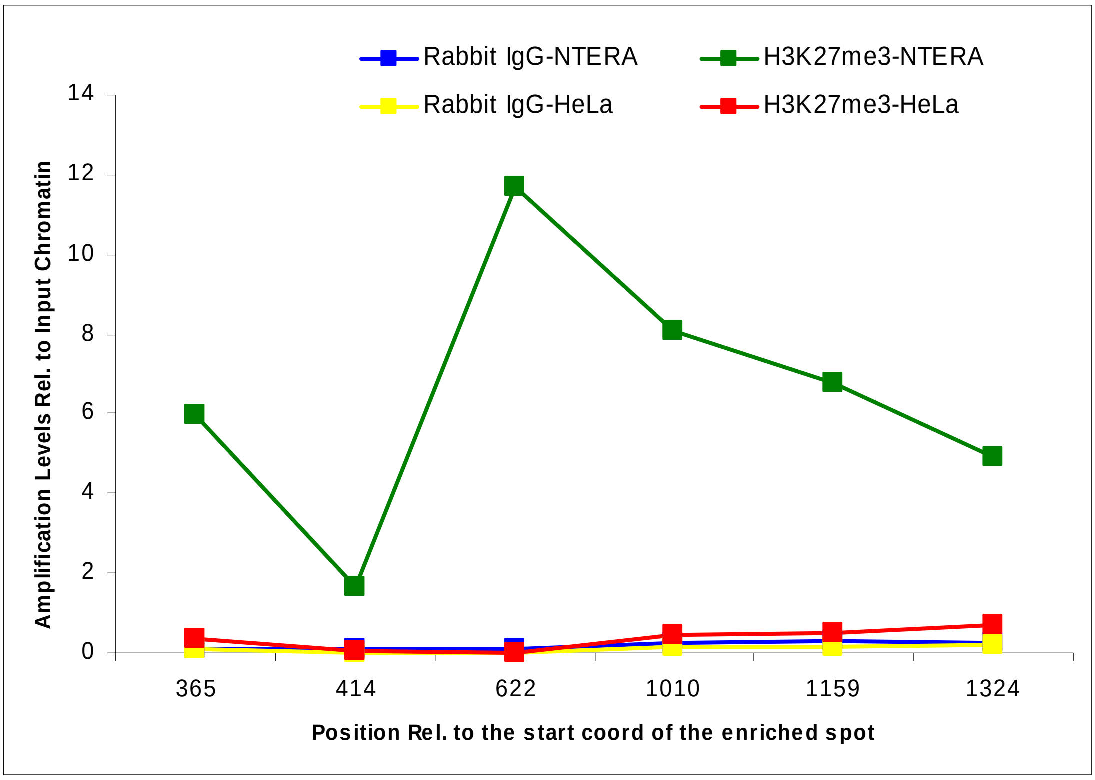
Start (bp)	End (bp)	Spot Information	H3K9me2 – HeLa S3	H3K9me2 – NTERA-D1
-2830	-957		-0.02	0.02
-308	435	carries TSS of C20orf121	-0.52	-0.89

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.

Amplification levels using ChIP material obtained by H3K9me2 antibody in HeLa S3 were slightly higher than those of Rabbit IgG, however it is very low compared to the enrichment levels obtained by other antibodies. Real-time PCR reactions performed for regions elsewhere did not show high amplification levels either (<2.5). This antibody did not show high enrichment levels on ChIP on chip experiments in both cell lines.

For H3K27me3 validation, four regions with high signal were used in real-time PCR. Figure 5.12 shows the amplification levels in both cell lines (H3K27me3_Pr21 to

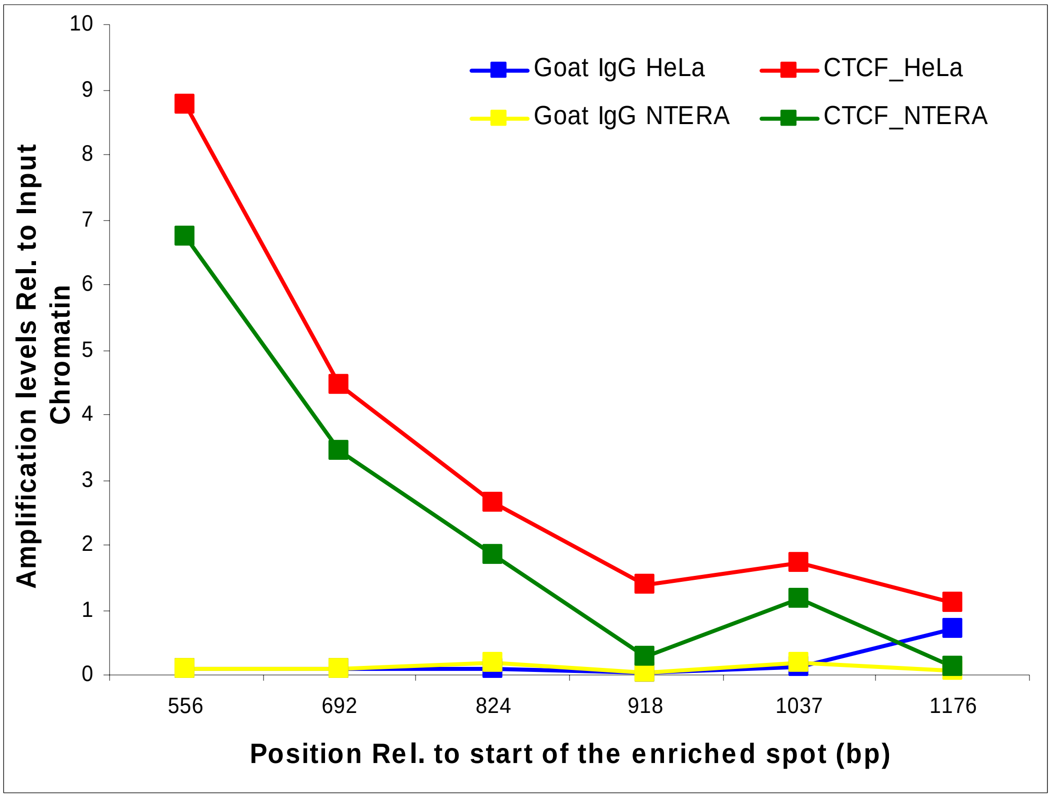
H3K27me3_Pr26, Appendix B) only enriched by H3K27me3 in NTERA-D1 cells. The spots that were enriched in HeLa S3 cells were also validated with real-time PCR (data not shown).



Start (bp)	End (bp)	Spot Information	H3K27me3 –HeLa S3	H327me3 – NTERA-D1
44,094,336	44,095,961		0.65	7.20

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.

Lastly, enrichment levels obtained by the antibody recognising CTCF were validated by designing real-time PCR primers to four regions enriched in CTCF in one or both cell lines. Figure 5.13 shows the amplification levels on a region enriched by CTCF in both cell lines.



Start (bp)	End (bp)	Spot Information	CTCF –HeLa S3	CTCF – NTERA-D1
-1421	+1171		8.67	5.69
+1	+1393		7.49	5.30

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.

Results for ADA, ZNF335 and the TSS-containing amplicons gave comparable results for the corresponding antibodies (data not shown). Finally, none of the 25 negative real-time PCR control regions showed significant amplification (<2-fold) above input chromatin (all antibodies in both cell lines).

After validation using real-time PCR, the following strategy was employed for H3K4me3 and polII to estimate the number of true enriched regions. The background-subtracted mean signals produced by H3K4me3 antibody were considered to select a threshold. First, the spots that gave enrichment more than one-fold over the background were taken. Within this set, the signals coming from the spots carrying a

TSS were considered and the minimum signal among those spots was determined. Then, the spots that gave a more than one-fold enrichment with more than one antibody were taken and the minimum signal among those spots was also determined since such spots are assumed to be more likely to be true positives. A threshold was selected to include the maximum number of "enriched" spots that are either carrying a TSS or are enriched with two or more antibodies while keeping the number of spots that are "enriched" with only one antibody to no more than 5% of the total number of "enriched" spots. I used this as an estimate of the false positive ratio of the experiment.

As an example, Figure 5.14 shows the case of the H3K4me3 antibody in NTERA-D1 cells, in which the signals are plotted in relation with the type of the spot and the signals obtained with other antibodies. For ease of visualisation, only the spots that gave a signal between 1 and 5 were shown. All spots showing signal values higher than 5 either carry or neighbour a TSS, or showed enrichment with more than one antibody. Here, the threshold for including the maximum number of possible true positive spots while keeping false positive ratio at most 5% was 1.5.

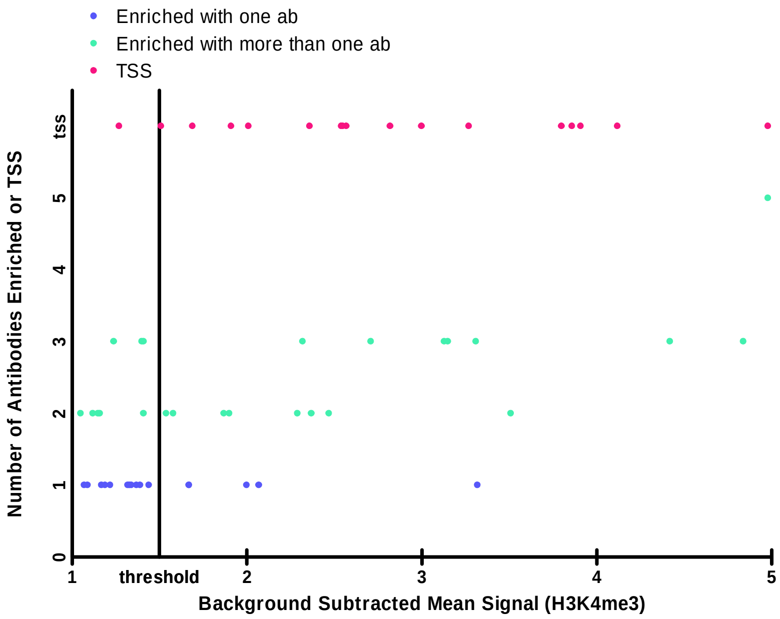


Figure 5.14 The H3K4me3 signals in NTERA-D1 cells. Spots that showed signals between 1 and 5 were shown here for the ease of visualisation. Vertical axis denotes the number of antibodies that a corresponding spot showed “enriched” signal. “tss” denotes that these spots either carry or neighbours a TSS.

The same threshold was determined for H3K4me3 antibody in HeLa S3 cells. The threshold for RNA polymerase II antibody was set to 1.75 in both cell lines to achieve a false positive ratio below 5%.

The same strategy cannot be directly applied to other histone antibodies since we have a limited knowledge of known true targets. Therefore, the threshold 1.5 was also selected for other histone antibodies based on the real-time PCR data and the assumption that enrichment patterns between different histone antibodies will be similar. As only real-time PCR data were available for CTCF antibody, conservative threshold 1.75, was selected.

5.5 Analysis of RNA polymerase II binding sites by ChIP

ChIP experiments were performed to screen for RNA polymerase II (polII) binding sites in HeLa S3 and NTERA-D1 cells. The polII antibody (Abcam) #ab5131 was raised in rabbit and it recognizes the phosphorylated serine found in the amino acid 5 position (serine-5) of the carboxyl terminal domain (CTD) repeat YSPTSPS of the protein. The unphosphorylated form of CTD is required for the initiation of the transcription and this form also interacts with a wide range of general transcription factors (Proudfoot et al., 2002). However, once the transcription is initiated, serine-5 at CTD is phosphorylated by the basal transcription factor TFIIH at the promoter, and the mRNA capping enzyme is brought to the initiation complex via binding to this modified form of CTD (Kim et al., 2004) (see section 1.3.1). During elongation, serine-5 phosphorylation drops and the capping enzyme dissociates (Komarnitsky et al., 2000). Therefore, this antibody should be able to detect DNA regions where RNA polymerase II assembles itself to initiate transcription.

For each transcript, the spot carrying a TSS plus the spots carrying the right and left flank were searched for an enrichment in signal. This roughly corresponds to 2 kb upstream and downstream of the TSS. Heat maps were generated for these sites. A heat map is a false-coloured graph representation of the signal intensities plotted against the positions of the feature in the genome. Higher signal values are represented as red while lower signal values are represented as white.

In a second round, spots that gave a high signal but not overlapping within ±2 kb of TSS of a coding transcripts were discussed. Such signals were analysed in the context of annotated features such as processed transcripts, pseudogenes or gene predictions including multi-species conservation of the DNA sequence of these spots.

5.5.1 Results in HeLa S3 cells

There were 62 spots which gave a signal intensity higher than the selected threshold (1.75) in HeLa S3 cells using the RNA polymerase II antibody (Abcam, #5131). Of these, 36 enriched spots were located ±2 kb of the TSS of an annotated feature. There were 25 spots within the genes whose TSSs showed polII enrichment. The remaining three spots were not in close proximity of any annotated TSS.

With respect to coding transcripts, the 35 enriched spots near a TSS represent 16 coding transcripts (15 genes). Per signal the intensity of the enriched spot and that of the three adjacent spots on either side were plotted against their relative position to the start site of the coding transcript they contain or neighbour. Figure 5.15 shows the signals of the 16 coding transcripts in the colour coded manner known as a heat map, red representing the highest and white the lowest signal.

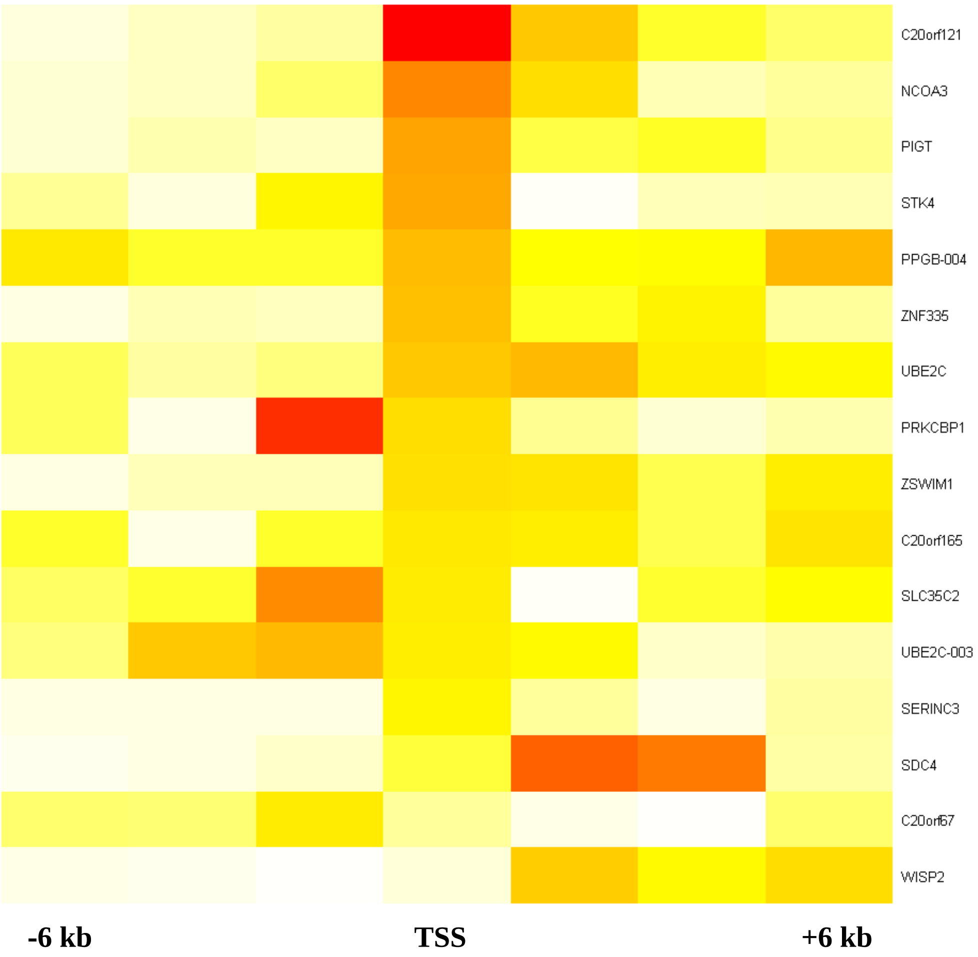


Figure 5.15 Heat map which displays spot intensities within ±6 kb distance of 16 coding transcripts in HeLa S3 cells. Highest signal is shown as red while the lowest signal is denoted as white. Representative transcripts are labelled with gene name only.

The heat map that displays the spot intensities of all 79 transcripts (corresponding to 66 genes) represented on the array can be found in Figure A2 (Appendix A).

Figure 5.15 shows *PPGB* among the positives. Note that the 5' ends of *NEURL2* and *PPGB* were contained in the same spot. Since *NEURL2* is not expressed but *PPGB* has a very high expression according to the Affymetrix Expression Arrays in HeLa S3 (Table A5, Appendix A), the observed enrichment is attributed to *PPGB* only.

Figure 5.15 shows that the phosphorylated form of RNA polymerase II recruits itself on the annotated TSS of ten transcripts, although there are exceptions at which a higher signal is observed on either upstream (PRKCBP1-007, SLC35C2-006 and UBE2C-003) or downstream (PPGB-001, SDC4-001 and WISP2-002) of the TSS. The higher signals upstream of the UBE2C-003 transcript is the signal of the upstream representative transcript of the gene (UBE2C in Figure 5.15). In the case of PRKCBP1-007 (its TSS is located at the far end of the spot) and SLC35C2-006, the higher signals were produced mostly by the adjacent spots which span 2 kb of the upstream region of these genes. However, in the case of SDC4-001 and WISP2-002, the neighbouring spots which carry the first intron of these genes showed higher signal than the spots carrying the TSS of these genes. This variation might be due to the annotated TSS being imprecise. This possibility will be discussed in later sections by comparing the enrichment levels by different modified histones on these spots.

The annotated TSS of *KCNK15* is not present on the array. However, the spot carrying the neighbouring 2.2 kb fragment which spans the first intron (its genomic coordinates are 42,809,326..42,811,520) gave a 3.3 fold enrichment. Also, the *KCNK15* is expressed in HeLa S3 cells according to the Affymetrix Expression Arrays which supports the fact that *KCNK15* is active in HeLa S3 cells.

There is one polII enriched spot (42,754,161 to 42,755,685 bp), which does not contain any potential TSS in its close proximity. The region does not contain any known micro RNA nor any transcriptionally active regions (TARs) reported by high density oligonucleotide tiling arrays (Bertone et al., 2004). On the other hand, this region was found to be enriched with mono and di-methylated histone H3 at K4 (~4-5 fold). Therefore, it will be discussed in detail in section 5.7.1 for their possible regulatory functions.

The other polII enriched spot (located between 42,813,344 and 42,816,090) contains 3' UTR end of *RIMS4*, which is not expressed and its TSS is not enriched with polII HeLa S3 cells. The same spot also showed an enrichment with mono-methylated K4 of histone H3 (~1.75 fold) and will be discussed in section 5.7.1. The polII spot located between 42,469,627 and 42,471,927 bp lies within the *HNF4A* and will be discussed in 5.10.

In summary, gene initiation activities of 15 genes (16 transcripts) were detected by ChIP experiment using the polII antibody and 13 of them are expressed in HeLa S3 cells according to Affymetrix Expression Arrays. The *ZNF335* and *C20ORF165*, which gave an enriched signal but are not expressed may correspond to transcripts with post-transcriptional regulation. There are 15 more genes that are expressed in HeLa S3 cells but no polII enrichment was detected. This could be simply due to suboptimal working experimental conditions. Alternatively, polII initiation complexes on the TSSs of these genes may be short-lived which inevitably results in lower or no signal from these sites by ChIP. Another possibility would be that the epitope of polII recognized by the antibody is competed *in vivo* by other proteins hindering the detection of initiation complexes with this antibody.

It is important to note that all polII enriched start sites, except that of *SLC35C2*, were also enriched with tri-methylated histone H3 (H3K4me3) at K4 and acetylated histone H3 (H3Ac) as well. Note that these are the epigenetic markers found mainly on the start sites of actively transcribed genes (Bernstein et al., 2005) (see sections 1.4.4 and 1.4.5).

5.5.2 Results in NTERA-D1 cells

ChIP experiments with the polII antibody in NTERA-D1 cells resulted in 89 enriched spots. Of these, 42 were located within ±2kb of the TSS of an annotated coding transcript. While seven of the remaining 47 enriched spots were within intergenic regions, the rest of the enriched spots were contained in the intra-genic regions of active genes.

In NTERA-D1, annotated TSS of 17 coding transcripts (representing 14 genes) gave an enrichment and the heat map of these transcripts is shown in Figure 5.16. Note that all these polII enriched start sites showed an enrichment with H3K4me3 and H3Ac as well. The spot containing *C20orf67* did not produce any signal on a specific batch of the arrays used (marked as white in Figure 5.6), however, the adjacent spot (end of this spot is 400 bp upstream of the TSS of *C20orf67*) gave around 3 fold enrichment which attributed to *C20orf67* promoter activity.

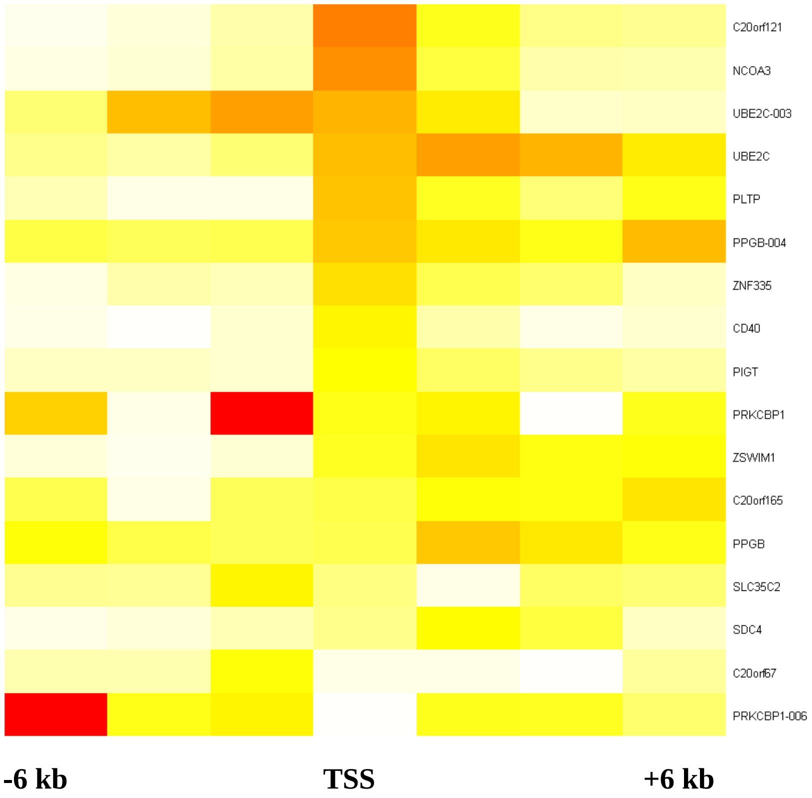


Figure 5.16 Heat map which displays spot intensities within ±6 kb distance of 17 coding transcripts in NTERA-D1 cells. Highest signal is shown as red while the

lowest signal is shown as white. Note that representative transcripts were denoted by gene name only.

The heat map that displays the spot intensities of all 79 transcripts (corresponding to 66 genes) represented on the array is given Figure A2 (Appendix A).

Figure 5.17 shows the enrichment levels of *PRKCBP1* together with its annotation (Ensembl Genome Browser; version 39).

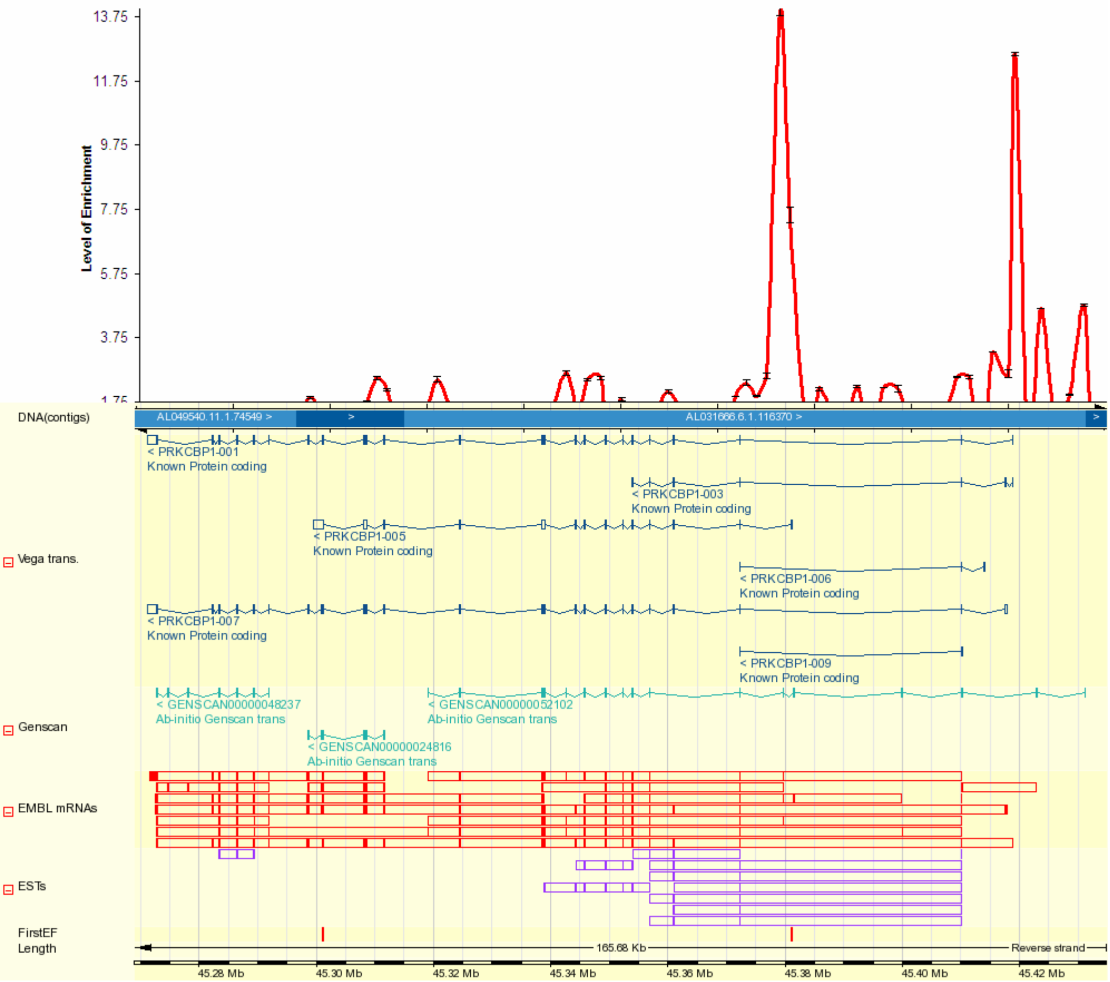


Figure 5.17. Enrichment levels (only signals above threshold were shown) on PRKCBP1 gene shown together with the annotation tracks reproduced from Ensembl Genome Browser. Track information is displayed on the left hand side of the annotation window.

The region downstream of the annotated TSS of the PRKCBP1-005 transcript showed 13.5 fold polII enrichment, while the proximal promoter region of the *PRKCBP1* (representative transcript) showed ~12 fold enrichment. Interestingly, spots carrying the annotated TSSs of these transcripts showed only ~2 fold polII enrichment. Also, there are two polII-enriched region (4.2 fold) at around 2 kb and 10 kb upstream of the annotated start site of *PRKCBP1* representative transcript. The +10 kb region is also enriched with H3K4me3 (53 fold) and H3Ac (3.8 fold). This region contains the 5' end of a gene prediction (GENSCAN00000035740). Further, there is a human spliced EST (CN335160, UCSC Genome Browser) from embryonic stem cells. The above strongly supports an alternative *PKRCBP1* transcript whose 5' end is 10 kb further upstream of the currently annotated start site. None of these upstream peaks were observed in HeLa S3 cells which suggests that the putative alternative transcript described above is tissue-specific.

All 14 genes except *C20orf121*, *NCOA3* and *CD40* showed polII enrichment across their gene sequence while these three genes specifically showed polII enrichment on their annotated start site.

As mentioned earlier, there are seven polII enriched spots that do not contain by any gene or lie within 2 kb of a start site of a genic feature. Two of them are located upstream of *PKRCBP1* and explained above. Another one is located between 43,354,154 and 43,355,325 bp, and showed a 2.5 fold polII enrichment. This region also showed a 40 fold enrichment with H3K4me3, and a 4 fold enrichment with H3Ac. It is adjacent to the 3'end of the MATN-4 gene, which did not give any polII enrichment on its start site and is not expressed in NTERA-D1 according to Affymetrix Expression Arrays. Within this enriched spot, there is a FirstExon promoter prediction (on reverse strand) and a CpG island (692 bp) immediate upstream of the promoter prediction. Additionally, it also contains the candidate first exon of a human spliced EST (DB444499, UCSC Genome Browser) found in testis.

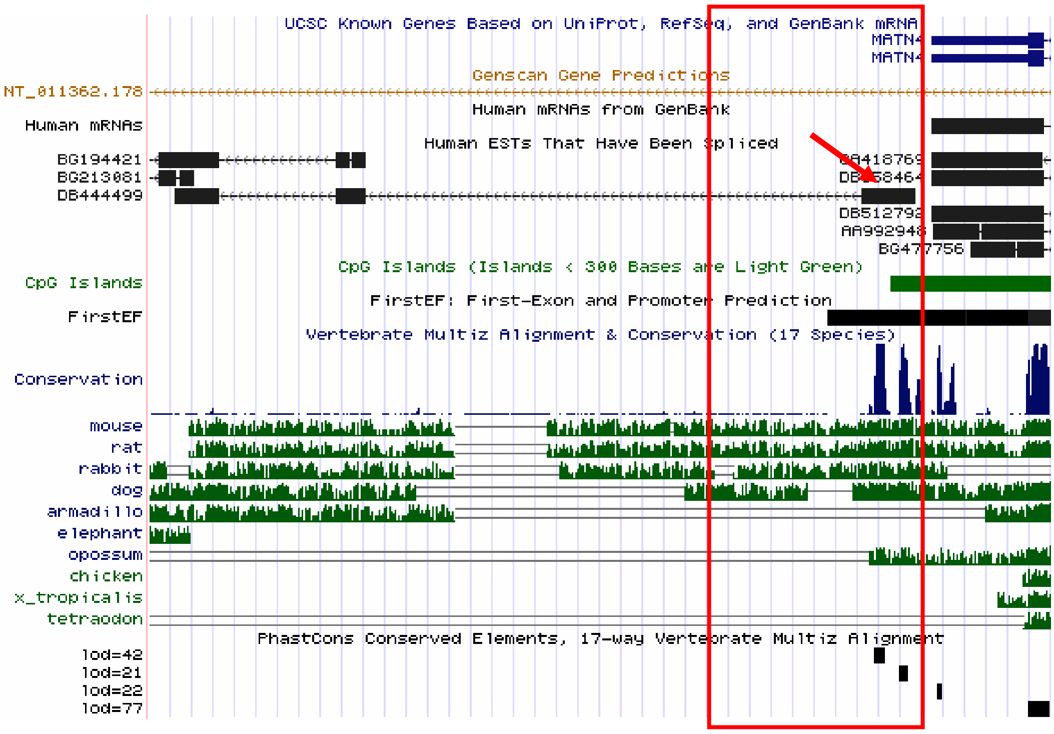

UCSC Known Genes Based on UniProt, RefSeq, and GenBank mRNA
MATN4
MATN4
Genscan Gene Predictions
NT_011362.178
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
BG194421
BG213081
DB444499
BG418769
DB458464
DB512792
AA992948
BG477756
CpG Islands (Islands < 300 Bases are Light Green)
CpG Islands
FirstEF: First-Exon and Promoter Prediction
FirstEF
Vertebrate Multiz Alignment & Conservation (17 Species)
Conservation
mouse
rat
rabbit
dog
armadillo
elephant
opossum
chicken
x_tropicalis
tetraodon
PhastCons Conserved Elements, 17-way Vertebrate Multiz Alignment
lod=42
lod=21
lod=22
lod=77

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.

Figure 5.18 shows feature annotations (UCSC Genome Browser) mapped onto this polII enriched spot (in red square). Combined, the presence of a promoter prediction, a CpG island and a spliced EST, strongly supports the existence of a new gene, which is tissue-specific (not present in HeLa S3).

There are four adjacent spots (located between 43,821,287 and 43,829,560 bp) which gave polII (~5 fold) as well as H3K4me3 (~11 fold) enrichment, and they do not coincide with any genic feature. There is no evidence such as gene or promoter predictions or high sequence conservation across species, yet this polII enrichment may not be an artefact. Further experimental testing is required but none of these regions contained any known micro RNA, or any TARs.

In summary, out of 35 genes expressed in NTERA-D1, 5’ ends of 14 were enriched with RNA polymerase II. TSSs of two non-expressed genes (*C20orf165* and *WFDC10A*) were enriched with polII (2- and 1.9 fold respectively).

The genes, *PLTP*, *WISP2* and *ZNF334* were enriched by polII only in NTERA-D1 but not in HeLa S3 cells, and these three genes were only expressed in NTERA-D1 cells. Therefore, it was possible to detect expression profile differences of the genes in ChIP experiments in contrast to the gene reporter assays where there was less correlation between the promoter activity of the gene and its expression status (section 4.5.3.2). This result is expected as ChIP experiments can successfully capture the genomic environment of the genes while gene reporter assays only mimic the trans regulatory content of the gene.

SERINC3 and *CD40* are expressed in both cell lines, but *SERINC3* gave polII enrichment only in HeLa S3 cells, and *CD40* only in NTERA-D1 cells. This may indicate a different mode of gene regulation depending on the cell type. Where there is no ChIP signal, polII complexes might be hindered by other proteins and its epitope not accessible for the antibody in one cell line, although in the other cell line, there might be another set of proteins used to activate the gene that might not have the same hindrance effect. However, the fact that in each cell line circa 50% of the expressed genes did not give a polII enrichment favours the explanation of sensitivity.

It is well established that actively transcribed genes are decorated by a set of specifically modified histones (Allfrey et al., 1964) (Litt et al., 2001). As mentioned, H3Ac and H3K4me3 are the most common markers found on actively transcribed genes. Below, I will discuss the epigenetic markers involved in transcriptionally active genes and correlate the activating histone code with the observed polymerase II activity described so far.

5.6 Histone Modifications on Transcription Start Sites

ChIP experiments were performed with antibodies recognising histone H3 tri-methylated at K4 position (H3K4me3) and histone H3 acetylated on lysine 9 and 14

(H3Ac) in both HeLa S3 and NTERA-D1 cells. The ChIP samples were characterized using the custom 3.5 Mb tile-path array.

5.6.1 Results in HeLa S3 cells

5.6.1.1 H3K4me3

An antibody recognizing H3K4me3 was used to locate transcriptionally active promoter regions. The H3K4me3 polyclonal antibody (Abcam, #ab8580) was raised in rabbit and it recognizes tri-methylated K4 residues on histone H3 proteins. This antibody has a weak reactivity to di-methylated K4 of histone H3 but it has no cross reactivity with any modified form of K9 of histone H3.

A lower threshold value (1.50) was used to decide whether a spot enriched or not by H3K4me3 since the relative distribution of the background and actual signal with this antibody differs from that obtained with the polII antibody. On the array 50 spots gave enrichment with H3K4me3 antibody in HeLa S3 cells. Of those, 42 were within 2 kb of annotated TSSs of alternative transcripts representing 22 genes. Out of eight enriched spots not associated with any TSS, two were not considered as true positives due to possible cross-hybridization, since they had over 90% sequence identity with ubiquitously expressed genes elsewhere in the genome. One of the spots (located between 44,159,993 and 44,162,444 bp) contains a processed pseudogene (*RPL13P2*), which has a 90% sequence identity with ubiquitously expressed *RPL13* ribosomal protein gene. The other spot (located between 45,817,478 and 45,819,561 bp) does not contain any genic feature but it has 85% sequence identity with *GTFIIIC* (general transcription factor IIIC polypeptide I), a ubiquitously expressed gene required for RNA polymerase III transcription. The remaining six enriched spot will be discussed in further sections since they gave higher signals with other antibodies used in this study.

The heat map of H3K4me3 signals within ~6 kb upstream and downstream of the annotated start sites of the 23 positive transcripts (representing 22 genes) is given in Figure 5.19.

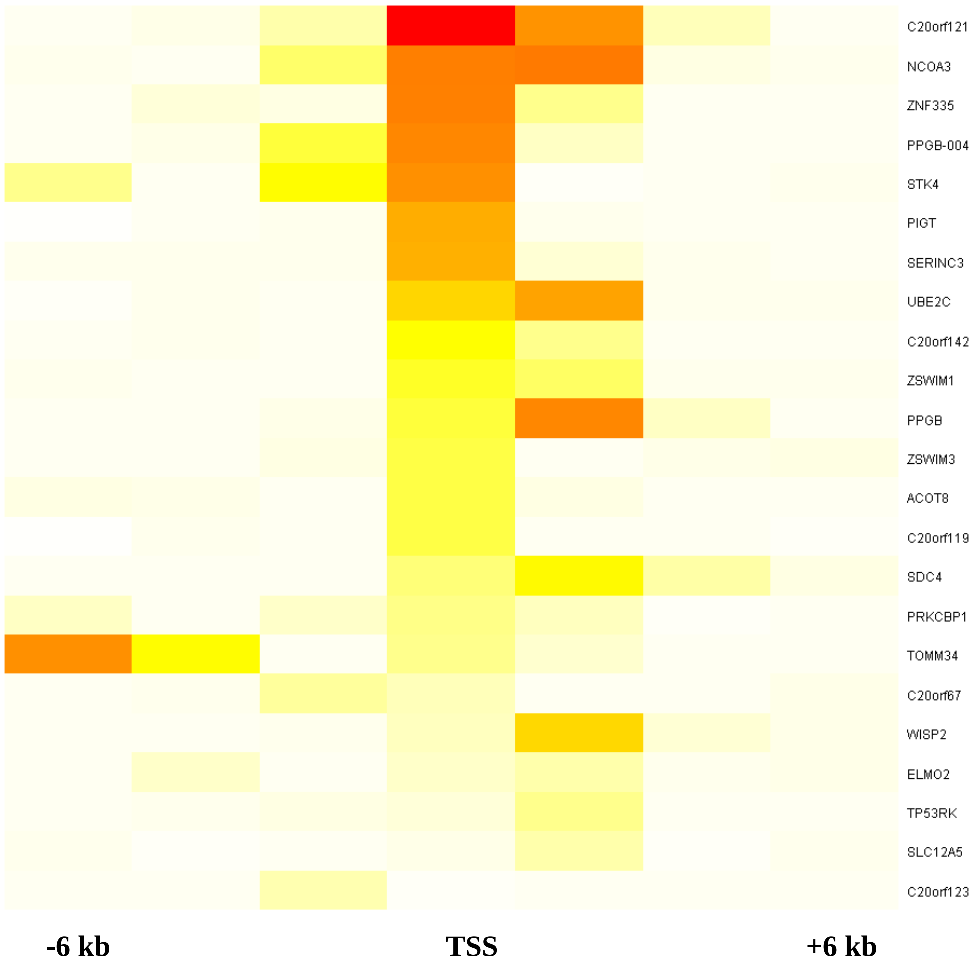


Figure 5.19 The heat map displaying ~6 kb upstream and downstream of the annotated start sites of 22 genes that showed an H3K4me3 enrichment around their TSS.

Also, Table 5.4 shows the enrichment levels of H3K4me3 enriched start sites of the 23 transcripts (representing 22 genes) by other 8 antibodies used in this study. Out of 28 genes expressed in HeLa S3 according to Affymetrix Expression Arrays, 20 were enriched with H3K4me3 around their TSS, and out of these 20 H3K4me3 enriched genes, 13 were enriched with RNA polymerase II as well.

C20orf62 gave an enrichment on the spot carrying its TSS, which also contains a pseudogene (*RPL37AP1*) that has 96% sequence identity (across 700 bp) to a ubiquitously expressed ribosomal protein (*RPL37A*). Since *C20orf62* gene is not expressed in HeLa S3 cells, this enrichment was attributed to a cross hybridization, therefore it is not included within the active gene set.

The start site of *SLC12A5*, which is not expressed in HeLa S3 cells was enriched with H3K4me3. Interestingly, the H3K4me3 enrichment was observed downstream rather than on the actual TSS. The downstream spot is also enriched with H3Ac (Table 5.4). *SLC12A5* is not enriched anywhere across its sequence with polII. Also, its core promoter region did not show any activity in the gene reporter assays (section 4.5.1). In NTERA-D1 cells, its start site as well as its downstream region are enriched with H3K4me3 and H3Ac, but also with H3K27me3, an epigenetic marker associated with silenced genes (Cao et al., 2002). Hence, the enrichment profile across this gene will be discussed in detail in section 5.10.

There are five genes expressed in HeLa S3 (*TOMM34*, *ACOT8*, *C20orf142*, *ELMO2* and *TP53RK*) that gave an H3K4me3 enrichment but no polII enrichment on their TSS. Reporter assays for *ACOT8*, *ELMO2* and *TP53RK* genes gave strong core promoter activity (section 4.5.1). This indicates that the lack of enrichment by polII on the start site of these genes is most likely due to experimental limitations as explained in section 5.5.1, since detecting a promoter activity by reporter assay confirms the lack of any dominant trans-acting silencing on the promoter.

Histone H3 K9 di-methylation (H3K9me2) and K27 tri-methylation (H3K27me3) are associated with heterochromatic regions and silencing (Shilatifard, 2006) and will be discussed in section 5.8. However, it is worth mentioning here that it is not surprising to see no H3K4me3 enriched region being enriched with either of these two epigenetic markers in HeLa S3 cells.

Index	HUGO Gene ID	polII	H3K4me	H3K4me2	H3K4me3	H3Ac	H4Ac	CTCF	Expression
1	ACOT8		2.0d	3.5	6.3	2.2			P
2	C20orf119			4.6	6.2	2.9			P
3	C20orf121	7.4	3.3d	18	36.7	14.4	6.5		P
4	C20orf123				2.4u				No probe
5	C20orf142			4.6	9	3.3			P
6	C20orf67	2.2u		2.2u	2	1.6			P
7	ELMO2		2.9d	2.1d	2.6d	1.6d	1.6u		P
8	NCOA3	4.4	0	6.7	22.8	11.4	3.4		P
9	PIGT	3.8	0	7.3	17.9	7.9	3.7		P
10	PPGB (RT)	1.39*	2.70	4.59	6.62	2.84	2.55		P
11	PPGB (AT)	3.2	2.9	14.3	21.8	9.7	2.5		P
12	PRKCBP1	2.5	3.1	7.2	3.9	2.6	3.3		P
13	SDC4	5.3d	2.5d	2.9	4.5	2.1			P
14	SERINC3	2		7.3	17.5	5.7	1.9		P
15	SLC12A5				2.6d	2.2d			A
16	STK4	3.7	1.7	9.2	21	12.1	2.5		P
17	TOMM34		2.7d	4.2d	3.7	2.1			P
18	TP53RK		2.1d	4.4d	3.6d	3.7d			P
19	UBE2C	3	2.0d	4.2	13.4	2.9			P
20	WISP2	2.9d	4.5	2.7	1.9	2.5	2		P
21	ZNF335	3.2	3.4d	9.1	22.7	12.9	6.7	4.7d	A
22	ZSWIM1	2.4	1.9d	3	7.4	3.3	1.7		P
23	ZSWIM3		2.0u	3.5	6.3	2.2			P

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.

Note that none of the polII enriched spots, which have no association to any gene annotations, showed enrichment with H3K4me3 (see section 5.5.1). However, they were enriched with other modified histones, H3K4me and H3K4me2 and will be discussed in section 5.7.1.

As described earlier, the TSS of *KCNK15* is absent but the neighbouring spot spanning 2.2 kb of the first intron (42,809,326..42,811,520 bp) showed 3.3 fold enrichment for polII. Likewise, there is 3.7 and 3.6 fold enrichment with H3K4me3 and H3Ac antibodies respectively, which also supports the fact that the gene is actively transcribed in HeLa cells.

There is another H3K4me3 enriched spot (~1.8 fold) carrying the second 5' exon *SULF2*. The annotated start site of this gene did not give any enrichment above threshold with any antibody. In NTERA-D1 cells, there is a strong H3K4me3 enrichment on the actual TSS as well as the second 5' exon. There is also a promoter prediction (FirstExon) around the second 5' exon. These observations combined might indicate a tissue specific alternative TSS for this gene.

5.6.1.2 H3Ac

ChIP experiments in HeLa S3 cell were performed with the antibody (Upstate #06-599), which recognizes histone H3 acetylated at lysine 9 and 14. This polyclonal antibody is raised in rabbit and has no cross-reactivity with other modified histones. There are 60 spots on the array enriched with H3Ac. Out of these, 52 were within 2 kb of a TSS of 17 transcripts (17 genes) in the region. Figure 5.20 shows the heat map displaying the signals within 6 kb upstream and downstream of these transcripts.

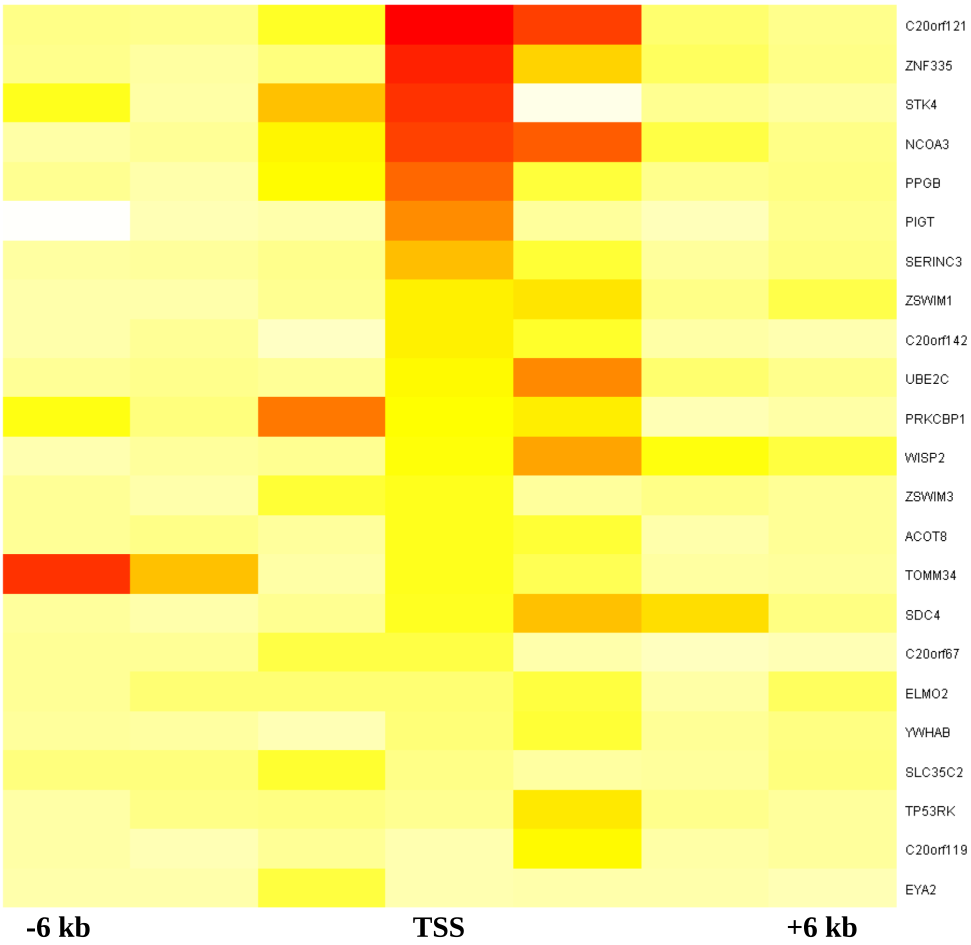


Figure 5.20 Heat map displaying the signals around 6 kb upstream and downstream of the annotated start sites of 17 H3Ac enriched transcripts in HeLa S3 cells. Higher signal is indicated with colours towards red while lower signals would be towards white.

Unlike H3K4me3 enrichment signals where the highest signal is mainly located on the TSS spot (see Figure 5.19), histone acetylation seems to be more widespread with a trend toward downstream sequences. This may mean that the factors which play a role in accurate positioning of the initiation machinery require H3K4me3, while histone 3 acetylation is needed for recruiting the initiation machinery around the start site.

Out of 24 H3K4me3 enriched start sites, only one was not enriched with acetylated histone H3. This finding is in agreement with previous studies (Kim et al., 2005) (Bernstein et al., 2005) where the acetylated histone H3 is found in 90-99% of H3K4me3 enriched genes. Also, it is shown that MLL, which is responsible for histone H3 methylation at K4 residues is stimulated by acetylated histone H3 peptides (Milne et al., 2002) (Nakamura et al., 2002). Therefore, histone H3 acetylation might play an important role in initiation of histone H3 tri-methylation at K4.

C20orf123, which showed H3K4me3 but no H3Ac enrichment on upstream of its start site did not give any enrichment with other antibodies either.

The remaining eight sites that gave enrichments with H3Ac gave higher enrichments with other antibodies, therefore the features of these sequences will be discussed in 5.7.1.

5.6.1.3 H3K4me and H3K4me2

In 5.6.1.1, I showed that tri-methylation of histone H3 at K4 on the TSS is a feature of transcriptionally active genes. To find out whether mono and di-methylation of histone H3 at K4 shows a similar pattern, ChIP experiments were carried out as before with polyclonal antibodies Abcam #ab8895 and Abcam #ab7766, which recognize the H3K4me and H3K4me2 form respectively. Both antibodies are raised in rabbit and do not have cross reactivity to histone H3 methylated at K9. The mono- and di-methylation profiles of the 6 kb distances of the start sites of genes enriched with H3K4me3 are displayed in Figure 5.21 which shows that tri-methylation is more centred on the actual TSS whereas there is virtually no tri-methylation in the flanking sequences. As mentioned earlier, this may indicate that this modification plays a role in fine positioning of the polII transcription initiation complex on the actual start site (see section 5.6.1.2).

High H3K4me2 enrichment was observed on H3K4me3 enriched start sites (upper half of the H3K4me2 heat map in Figure 5.21). TSSs that are not strongly tri-methylated are relatively more enriched with H3K4me2. This may be due to the fact that H3K4me2 is an intermediate state to H3K4me3. The H3K4me2 type of enrichment pattern was also observed with H3K4me.

As shown in Table 5.4, only two H3K4me3 enriched start sites (of *C20orf123* and *SLC12A5*) were not enriched with either H3K4me2 or polII. Yet, these genes are not expressed in HeLa S3 cells. This observation supports the idea that di-methylation of histone H3 at K4 on TSSs can be also seen as a mark for active genes. Again H3K4me2 is an intermediate to H3K4me3. However, as will be discussed in later sections (5.6.1.3 and 5.7.1), H3K4me2 can exist as an independent epigenetic functional marker from H3K4me3 in the genome. Histone H3 mono-methylation (H3K4me) was observed in 60% of the H3K4me3 enriched start sites. H3K4me is the first step in the methylation process of histone H3, so the presence of H3K4me on start sites is to be expected. On the other hand, HK4me has recently been found in distant regulatory sequence elements (Ren, B., unpublished data).

In summary the above results suggest that promoter regions are clearly marked with tri-methylated or di-methylated histone H3 at K4, whereas mono-methylated histone H3 at K4 does not have the same discriminatory power.

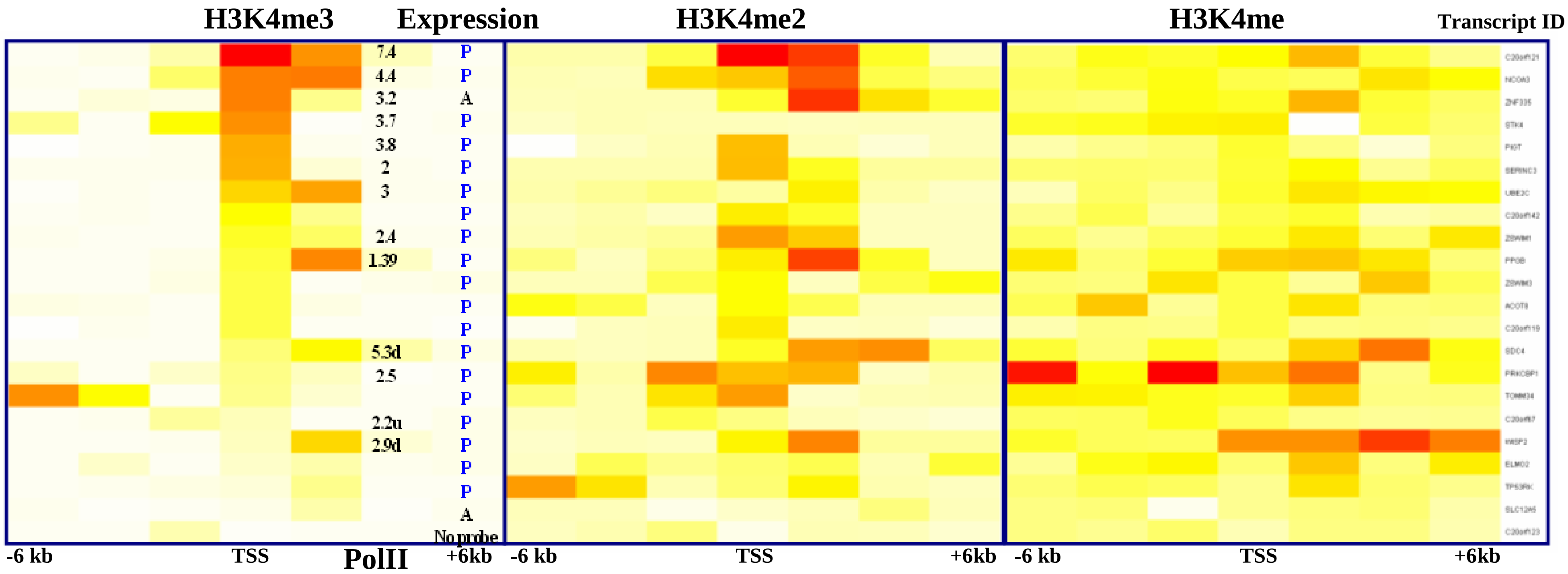


Figure 5.21 H3K4me3, H3K4me2 and H3K4me profiles of 22 genes that showed H3K4me3 enrichments on their start sites in HeLa S3 cells. Each profile is presented as a heat map which displays the signals obtained within ~6 kb distance of TSSs. The first column on the left lists the polII enrichments and the 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.

5.6.1.4 H4Ac

An antibody recognizing the acetylated form of histone H4 at lysine 5,8,12 and 16 was employed to determine genomic sites enriched with this modified histone type. The antibody, Upstate #06866 is raised in rabbit and has a weak cross-reactivity with acetylated histone H3. There were 60 spots on the array that showed enrichment with this modified histone but only 19 were also enriched with H3K4me3. Only half (53%) of the H3K4me3 enriched sites were also enriched with H4Ac. This is expected as histone H4 acetylation plays a major role in chromatin structure changes and protein interactions (Shogren-Knaak et al., 2006).

A nice illustration of how the above ChIP results can be combined to improve the current annotation comes from the following example. A spot containing the TSSs of four non-coding transcripts of *DBNDD2* (*C20ORF35*) (DBNDD2-007, -009, -011 and -013) gave an 8.8 fold H3K4me3 enrichment. Interestingly, this spot also carries the start sites of three coding transcripts of *C20orf169*. However, the gene record for *C20orf169* has been removed from the Entrez Gene database

(http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene) since, except its last exon, all other exons are shared with the *DBNDD2*, and the mRNA evidence supporting *C20orf169* gene is used in the annotation of *DBNDD2*. The annotation of the region is shown in Figure 5.22.

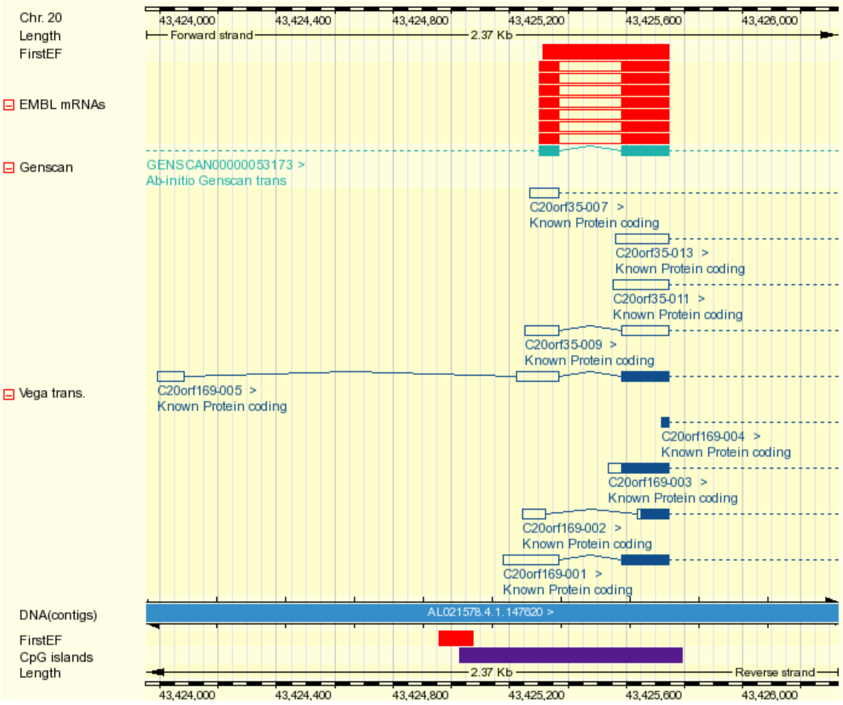


Figure 5.22. Ensembl annotation of the spot (its genomic coordinates 43,423,952..43,426,324 bp) that showed a 8.8 fold H3K4me3 and 5.3 fold H3K4me2 enrichments. It carries the start site of four non-coding transcripts of *DBNDD2* (*C20orf35*). It also carries the first exon of 5 coding transcripts of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

The fact that this region gave a significant 8.8 fold enrichment for H3K4me3 as well as H3K4me, H3K4me2 and H4Ac enrichments, there are two possible explanations: either, the annotated non-coding transcripts of *DBNDD2* are transcriptionally active or *C20orf169* actually is a real gene. *DBNDD2* is "expressed" in HeLa S3 cells but none of the start sites of its coding transcripts showed enrichment with the antibodies used in this study. Given that the spot in question contains a promoter prediction (FirstExon) and a CpG island, I believe that it is necessary to perform further experiments to reassess annotation of the region and re-instate the *C20orf169*, which is most likely a true gene.

5.6.2 Results in NTERA-D1 cells

5.6.2.1 H3K4me3

There were 83 H3K4me3 enriched spots in the ChIP experiments carried out in NTERA-D1 cells. Of these, 73 carried sequences within 2 kb distance of a TSS. Three spots carrying pseudogenes were eliminated due to possible cross hybridization with ubiquitously expressed genes elsewhere in the genome. Other enriched spots that are not close to any TSS were already discussed in section 5.5.2.

Out of the 66 genes whose TSS was represented by at least one spot on the array, 30 showed H3K4me3 enrichment. Of the 35 genes represented on the array and expressed in NTERA-D1 cells, 25 showed enrichment. The enrichment levels within 6 kb distance of the start site of the 30 genes are shown in Figure 5.23.

Table 5.5 lists the signals from the H3K4me3 enriched start sites together with the signals obtained with the other antibodies used in this study. Of the 30 genes enriched with H3K4me3, 13 were enriched with RNA polymerase II as well (Table 5.5). The four genes (*KCNS1*, *SLC12A5*, *WFDC10A* and *C20ORF165*) that are not expressed but whose start sites are enriched with H3K4me3, showed enrichments with either CTCF or H3K27me3, which play a role in silencing genes.

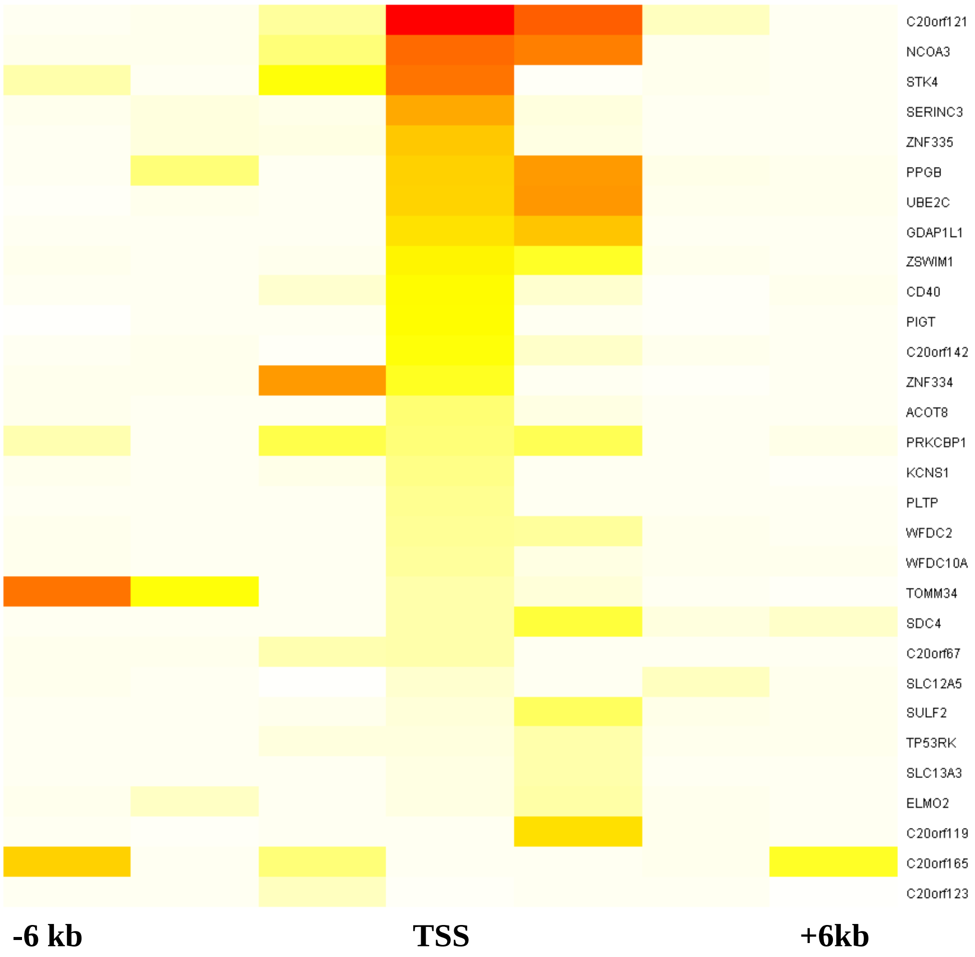


Figure 5.23 Heat map displaying the H3K4me3 enrichment levels of spots spanning 6 kb upstream and downstream sequences of 30 enriched TSSs of representative transcripts in NTERA-D1 cells.

Index	HUGO Transcript ID	polII	H3K4me	H3K4me2	H3K4me3	H3Ac	H4Ac	H3K27me3	CTCF	Expression
1	ACOT8			4.02	10.05					P
2	C20orf119			5.26	26.42	2.76				P
3	C20orf121	5.15	1.8d	25.77	79.56	7.1				P
4	C20orf123			1.6u	3.91u					No probe
5	C20orf142			7.68	18.47	2.66				P
6	C20orf165	1.9d		5.8u	9.5u				2.2u	A
7	C20orf67	1.9u		3.93	5.61	1.55				P
8	CD40	2.18		12.59	19.97	1.74				P
9	ELMO2			1.62	6.02d					P
10	GDAP1L1			14.25	26.25	3.4			3.3d	P
11	KCNS1			7.24	8.33			0	3.20u*	A
12	NCOA3	4.75		16.28	54.36	5.18				P
13	PIGT	1.98		7.57	19.6	2.2				P
14	PLTP	3.49			7.49	1.79				P
15	PPGB	3.34		21.91	42.93	4.79				P
16	PRKCBP1	1.75	2.5	7.75	9.39	2.37	3.46			P
17	SDC4	2.1d		4.21	5.65	1.84				P
18	SERINC3			12.24	39.36	3.89				P
19	SLC12A5			8.55	2.54			9.52		A
20	SLC13A3			7.6	5.475					P
21	STK4			18.35	51.9	4.65				P
22	SULF2			9.4d	1.91	1.91d				P
23	TOMM34			5.5d	5.7					P
24	TP53RK			1.67	1.51					P
25	UBE2C	3.62		6.47	29.56	1.81				P
26	WFDC10A			7.99	6.89				4.5	A
27	WFDC2			7.4	7.29	1.51				P
28	ZNF334			9.6	42.93	4.21				P
29	ZNF335	2.74		11.5	31.97	3.43			3.9d	P
30	ZSWIM1	2.6d		4.53	21.51	1.66d				A

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.

There are three other genes (*KCNS1*, *WFDC10A* and *C20orf165*) that are not expressed in NTERA-D1 cells, yet they showed enrichments with H3K4me3 (but not with H3Ac) around their start sites. Interestingly, the TSSs of *WFDC10A* and *C20orf165* also showed enrichment with CTCF protein, which can act as a silencer, as well as the region 4 kb upstream of *KCNS1* (see section 5.9). CTCF can silence promoter regions of its target genes, and in the context of these three genes, it may function as a repressor since they are not expressed. Note that there is also a CTCF enrichment around 4 kb upstream of the *KCNS1* in HeLa S3 cells, while, unlike in NTERA-D1, there is no other H3K4me3 or polII enrichment on the gene in HeLa S3 cells and the gene is not expressed in HeLa S3 cells either. This indicates that if CTCF does function as a silencer on *KCNS1*, its recruitment is probably not dependent of the H3K4me3 status of the region since its binding is detected in two different chromatin environments.

The promoter region of *WFDC10A* is also enriched with CTCF protein in both cell lines. However, there is an H3K4me3 enrichment only in NTERA-D1 cells and this gene is not expressed in either cell line. This observation represents another case where CTCF binding is not dependant on at least the H3K4me3 status of the start site.

SDC4 showed a similar H3K4me3 and H3Ac enrichment pattern in both cell lines such that the higher enrichments were located around 2 kb downstream of the TSS (see Figure 5.23 and Figure 5.25). In NTERA-D1 cells, there is a second polII (~3 fold) and H3K4me3 peak (~12 fold) approximately 5 kb downstream of the start site (Figure 5.24). There is an mRNA (G43350, located between 43,407,194 and 43,407,343 bp), located 1 kb upstream of the second peak, which may represent an as yet un-annotated first exon. Except the two multi-species conserved blocks located within the spot giving the second peak, there is no other mRNA or EST evidence for an alternative exon. Based on these observations, the 5' annotation of this gene needs to be inspected for a mis-annotated first exon or a tissue-specific second exon.

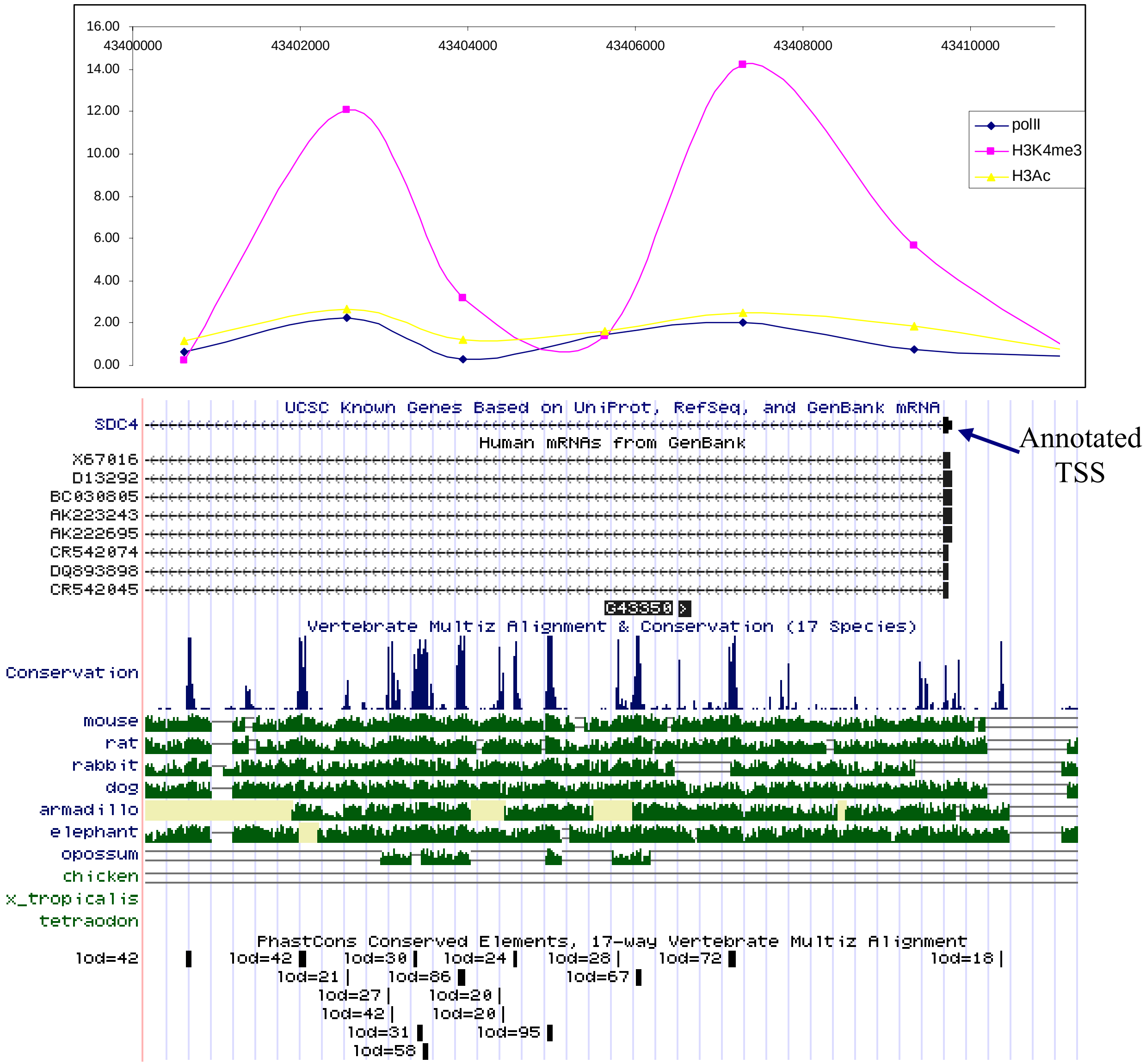


Figure 5.24 H3K4me3, polII and H3Ac peaks within the first 10 kb of SDC4 gene. The annotation is reproduced from UCSD Genome Browser.

As expected, except one case (*SLC12A5*), none of the H3K4me3 enriched sites gave enrichments with H3K27me3 or H3K9me2.

5.6.2.2 H3Ac

There are 50 spots that were enriched with acetylated histone H3 and 40 of them carry sequences that are within 2 kb distance of a TSS. Of the remaining 10, the profiles of seven were discussed in the previous section since they coincided with H3K4me3 enrichments. Finally, the remaining three spots will be discussed in 5.7.2.

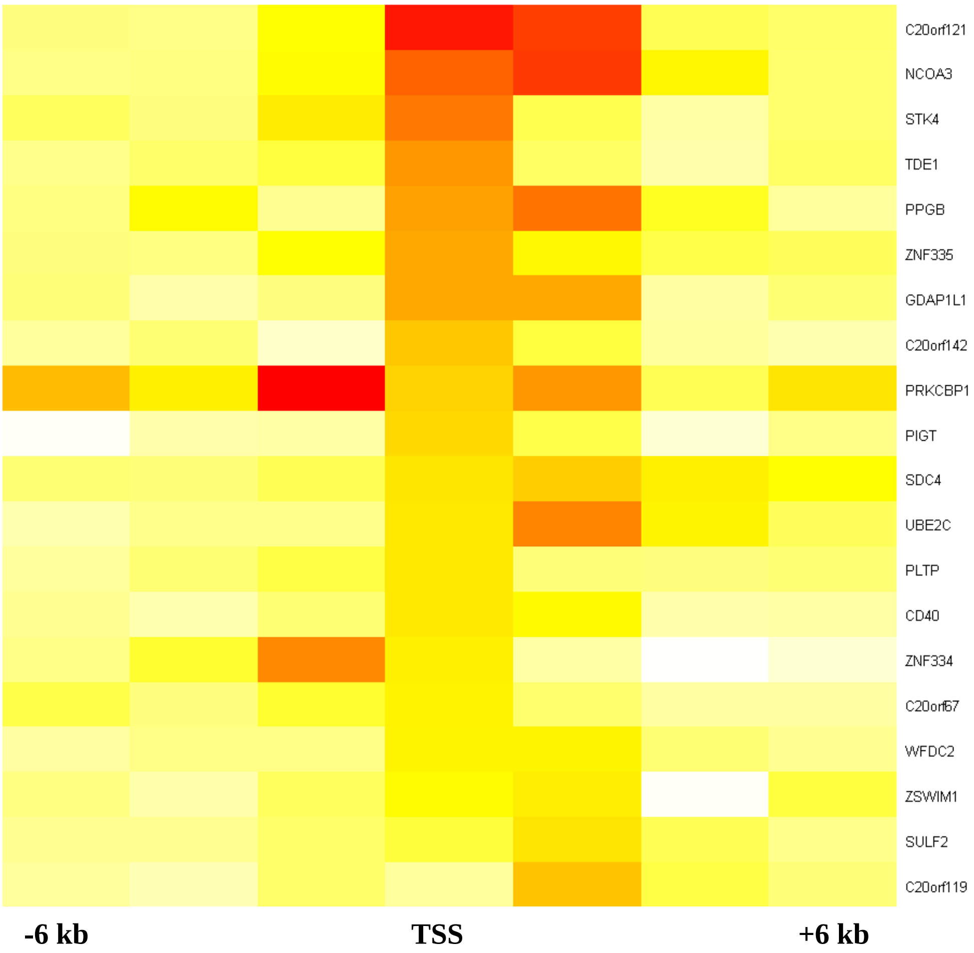


Figure 5.25 Heat map displaying the signals around 6 kb upstream and downstream of the annotated start sites of 21 H3Ac enriched transcripts in NTERA-D1 cells. Higher signal is indicated with colours towards red while lower signals would be towards white.

Start sites of 21 transcripts that showed H3Ac enrichment (40 spots) and the signals within 6 kb distance of these start sites are shown in Figure 5.25. Like in HeLa S3 cells, histone acetylation appears to be more widespread around the TSS whereas tri-methylation of histone H3 at K4 is mostly centred on the actual TSS (see section 5.6.1.2 and 5.6.2.1).

5.6.2.3 H3K4me and H3K4me2

The H3K4me and H3K4me2 enrichment profiles of the TSSs that are enriched with H3K4me3 in NTERA-D1 cells are shown in Figure 5.26. As observed in HeLa S3 cells (see section 5.6.1.3), H3K4me3 is very specific to the TSS, H3K4me2 is observed both on and immediately downstream of the start site. On the other hand, H3K4me does not appear to have such a location preference; the enrichment levels do not change greatly around the TSS. However, genes exhibit different H3K4me patterns such as the *PRKCB1*, which is enriched up to 6 kb either way of the start site, and *NCOA3*, *ZNF335* and *SDC4*, which are enriched with H3K4me across the downstream sequences (Figure 5.26).

5.6.2.4 H4Ac

Interestingly, in NTERA-D1 cells, the antibody recognizing acetylated histone 4 at lysines 5,8,12 and 16 (H4Ac) showed enrichments in only three spots. All enriched spots coincide with the start site of *PRKCBP1*. The poor performance of this antibody is intriguing since it worked well in HeLa S3 cells. Also, all antibodies except H4Ac gave higher signals in NTERA-D1 than HeLa S3 cells.

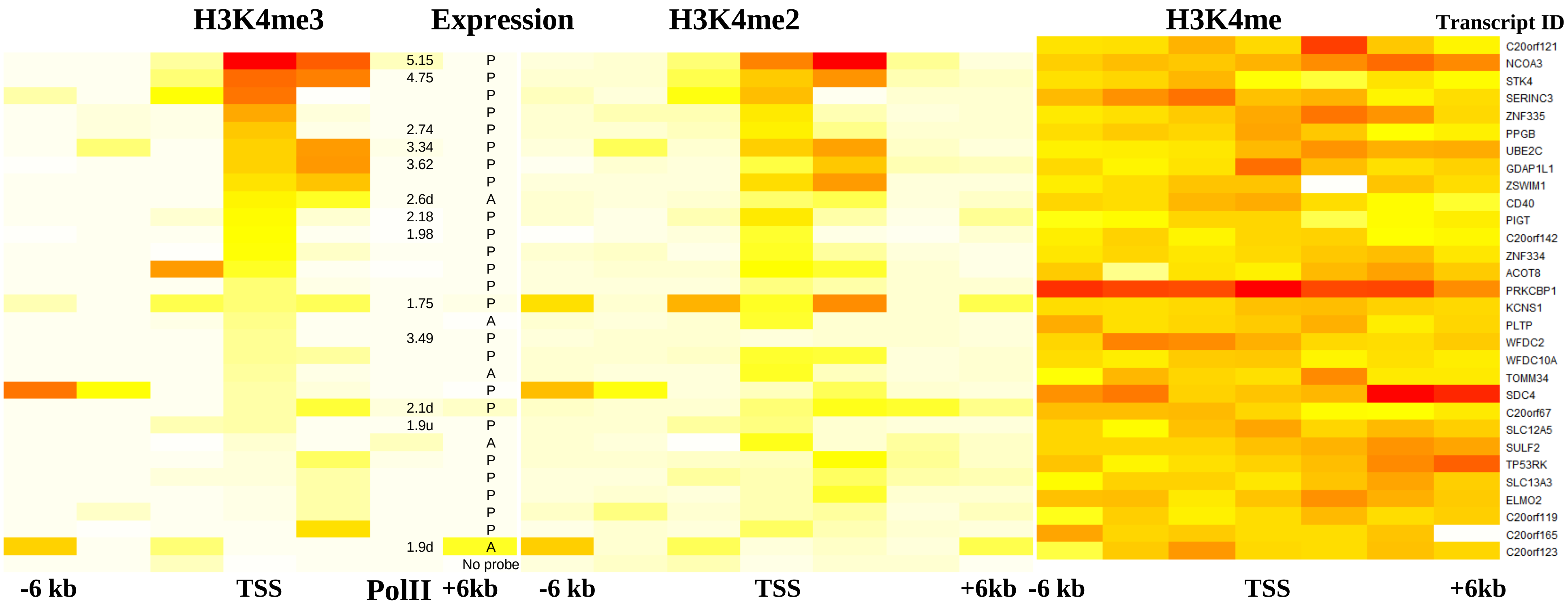


Figure 5.26 H3K4me3, H3K4me2 and H3K4me profiles of 23 genes that showed H3K4me3 enrichments on their start sites in NTERA-D1 cells. Each profile is presented as a heat map which displays the signals obtained within ~6 kb distance of TSSs. The first column on the left lists the polII enrichments and 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.

5.7 Histone Modifications marking possible regulatory elements

In this study, antibodies recognizing seven different modified histones were employed to derive a partial histone map of a 3.5 Mb region at 20q12-13.2. So far, I have described the results obtained with five antibodies namely, mono-, di- and tri-methylated histone H3 at K4 (H3K4me, H3K4me2 and H3K4me3), acetylated histone H3 at lysine 9 and 14 (H3Ac) and acetylated histone H4 at lysines 5,8,12 and 16 (H4Ac), all of which are commonly found in euchromatic regions. In order to look at the number of occurrences of observing each possible combination of modified histones at a given site, all sites carrying different combinations of modified histones were listed (Table 5.6) and then plotted as shown in Figure 5.27. However, since there are only three sites enriched with H4Ac in NTERA-D1 cells, as opposed to 64 sites in HeLa S3 cells, H4Ac combinations were excluded from the plots (Figure 5.27).

In both cells, there are no spots that are enriched only with H3K4me + H3K4me3 or H3K4me + H3K4me3 + H3Ac as expected, since methylation of histone H3 at K4 is performed in a stepwise manner by the same enzyme (MLL) and H3K4me2 is the intermediate molecule between H3K4me and H3K4me3.

In HeLa S3 cells, there are three times the number of sites enriched with only H3K4me than in NTERA-D1 cells. Among the sites that are only enriched with H3K4me, 80% and 85% are within inter- or intra-genic regions in HeLa S3 and NTERA-D1 cells respectively, meaning that they do not possess the potential for further methylation. This observation again supports the opinion that H3K4me is not associated with TSSs and probably exists as a marker for other regulatory regions, most likely in combination with other modified histones.

H3K4me	H3K4me2	H3K4me3	H3Ac	H4Ac	# of occurrences in HeLa S3	# of occurrences in NTERA-D1
H3K4me	H3K4me2	H3K4me3	H3Ac	H4Ac	14	2
H3K4me	H3K4me2	H3K4me3	H3Ac		9	2
H3K4me	H3K4me2	H3K4me3			2	3
H3K4me	H3K4me2		H3Ac	H4Ac	6	
H3K4me	H3K4me2		H3Ac		11	3
H3K4me	H3K4me2			H4Ac	11	
H3K4me	H3K4me2				17	10
H3K4me		H3K4me3	H3Ac	H4Ac		
H3K4me		H3K4me3				
H3K4me			H3Ac	H4Ac	15	
H3K4me			H3Ac			1
H3K4me				H4Ac	9	
H3K4me					71	25
	H3K4me2	H3K4me3	H3Ac	H4Ac	5	
	H3K4me2	H3K4me3	H3Ac		12	35
	H3K4me2	H3K4me3			3	38
	H3K4me2		H3Ac	H4Ac		
	H3K4me2		H3Ac		1	3
	H3K4me2			H4Ac		
	H3K4me2				10	52
		H3K4me3	H3Ac	H4Ac		
		H3K4me3	H3Ac		3	1
		H3K4me3			4	
			H3Ac	H4Ac		
			H3Ac		2	2
				H4Ac	3	1

Table 5.6 Number of occurrences of different histone combinations at one site in HeLa S3 and NTERA-D1 cells.

Next, 10 out of 11 (90%) and 40 out of 54 (73%) sites which were only enriched with H3K4me2, were within inter- or intra-genic regions in HeLa S3 and NTERA-D1 cells respectively. Sites that were enriched only with H3K4me and H3K4me2 were not also close to annotated start sites. However, nearly 90% of the sites that were enriched with H3K4me + H3K4me2 + H3Ac were close to TSSs. This observation can lead to two conclusions; (i) H3 acetylation may indeed have a stimulatory effect for a site to be enriched in H3K4me3 (Milne et al., 2002; Nakamura et al., 2002) and (ii) H3Ac may not be a common epigenetic marker in regulatory elements other than promoters.

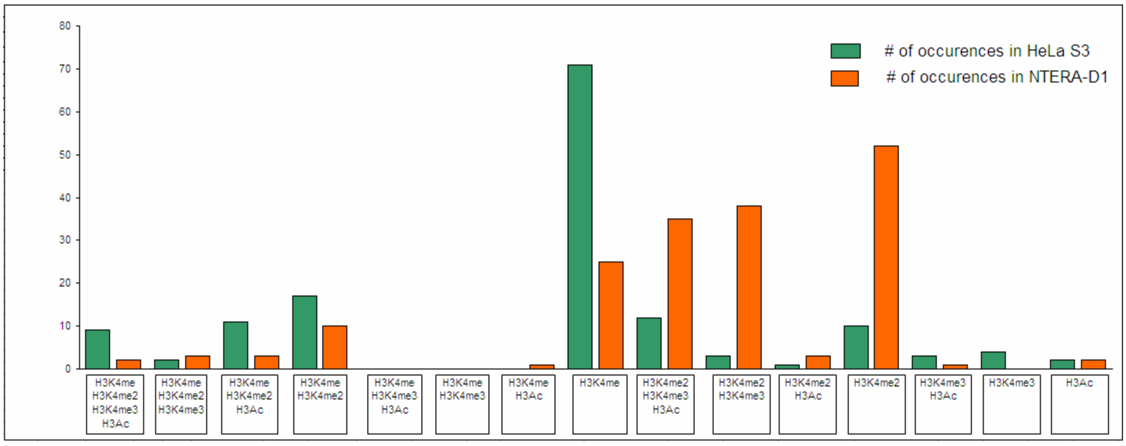


Figure 5.27 Plots of number of occurrences of all possible modified histone combinations in HeLa S3 and NTERA-D1 cells

5.7.1 HeLa S3 cells

It is interesting to look at the spots that gave enrichment with a number of modified histones but are not located in close proximity of any TSS. In HeLa S3 cells, there are 99 sites that showed enrichments with H3K4me2; 45 of which showed enrichment also with H3K4me3 and are associated with TSSs. Out of the remaining 54 H3K4me2 enriched sites, 80% showed enrichment also with H3K4me. The spots that showed low enrichment with only one antibody will not be discussed here due to insufficient experimental evidence for their potential functions. Table 5.7 lists 28 selected spots that gave H3K4me2 enrichment and their short feature descriptions.

Index	Start Coordinates	End Coordinates	Spot Information	polII	H3K4me	H3K4me2	H3Ac	H4Ac	CTCF
SP1	42469627	42471927	within HNF4A gene; contains one highly conserved region	6.71		3.23			
SP2	42670305	42672476	low conservation		1.74	1.84		2.3	6.59
SP3	42672013	42674371	within PKIG gene; no features		2.71	2.15		2.18	7.55
SP4	42704657	42706871	no conservation		2.63	1.77			
SP5	42754161	42755685	contains highly conserved regions	5.98	3.47	2.92		2.18	
SP6	42755951	42756933	low conservation	1.58	9.61	2.21		3.56	
SP7	42756907	42757833	low conservation	1.25	5.91	5.71	2.15	2.84	
SP8	42757760	42758740	no conservation	1.5	5.29	4.76	1.6	2.99	
SP9	42760403	42761222	contains one ultra conserved region		9.73	2.24	2.04	4.98	
SP10	43235675	43237871	contains 5' end of inactive PI3 gene		3.23	2.27			
SP11	43381309	43383563	Contains one highly conserved region		2.25	1.6			2.78
SP12	43839764	43842639	contains 5' end of inactive WFDC3 gene		2.99	1.64			
SP13	43883251	43885299	contains 3' end of TNNC2 gene		2.73	1.52			
SP14	43896779	43899401	contains 5' end of non-coding transcript of SNX21 gene		3.79	1.51			
SP15	43922315	43924699	within ZSWIM3 gene; contains one highly conserved region		2.88	2.25			
SP16	44147270	44149246	within NCOA5 gene; contains one highly conserved region	1.83	2.73	2.91			
SP17	44275725	44277729	within CDH22 gene; contains one highly conserved region		5.49	2.06			
SP18	44380475	44382544	Contains one highly conserved region		2.77	3.83			
SP19	44461256	44463475	within ELMO2 gene; contains highly conserved regions		1.77	2.5			
SP20	44622588	44624763	no conservation			1.7			4.74
SP21	44624669	44627010	contains an exon of SLC13A3 gene			2.85			5.29
SP22	45395404	45396897	within PRKCBP1 gene; contains a weak conserved region		2.52	3.39	1.62		
SP23	45453630	45455876	no conservation	2.65	2.43	1.51		4.31	
SP24	45516032	45518287	contains highly conserved regions		4.95	1.83		2.63	
SP25	45553177	45556238	contains highly conserved regions		1.87	1.9			
SP26	45636634	45639142	within NCOA3 gene; contains highly conserved regions		4.8	5.26			
SP27	45638797	45641114	within NCOA3 gene; contains highly conserved regions		3.91	3.55			
SP28	45648944	45651138	within NCOA3 gene; low conservation		3.05	2.99		1.53	

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.

One region encompassing the SP5 to SP9 (Table 5.7) showed a particularly interesting enrichment pattern. This region does not contain any annotated coding features although it contains a novel transcript with no open reading frame (RP11-445H22.4). An Affymetrix expression probe (241759_at) designed mRNAs, AK090842 and CR597563, (13 spliced ESTs), gave an expression signal only in HeLa S3 cells. The annotation of the region is given in Figure 5.28, together with the enrichment levels of the different proteins. The region is overall enriched with H3K4me (~5-10 fold) and H4ac (~2-4 fold). Also, the spot containing the mRNA and spliced ESTs were enriched with H3K4me2 (6 fold) and H3Ac (2.2 fold), the epigenetic markers which are often found on the TSSs of active genes.

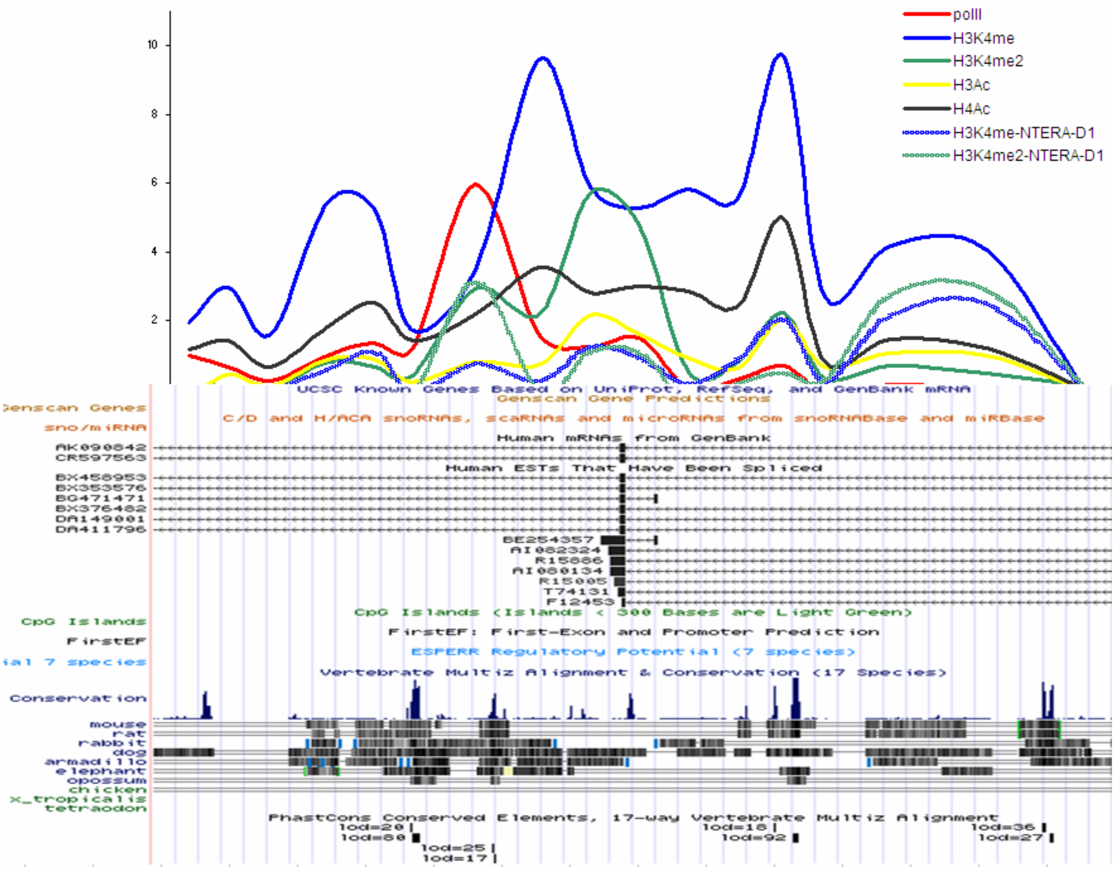
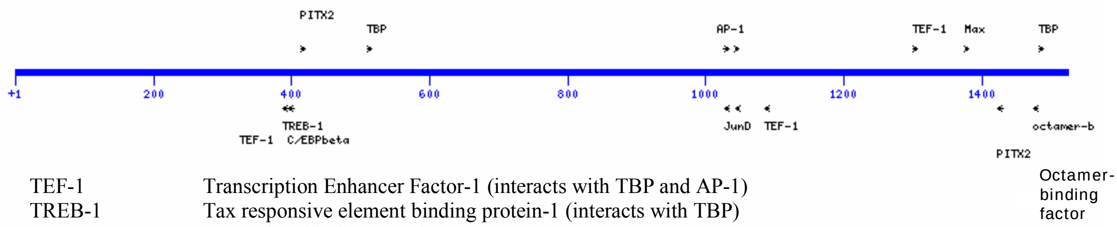


Figure 5.28 Annotation taken from UCSD Genome Browser for the region between 42,749,148 and 42,766,245 bp together with enrichment levels with proteins RNA polymerase II (polII), H3K4me, H3K4me2, H3Ac and H4Ac in HeLa S3. Also the thick blue and green lines displays enrichment levels with H3K4me and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

There is no H3K4me3 enrichment but SP5 shows a 5.9 fold RNA polymerase II peak. Despite the fact that the spot which showed polII enrichment is not close to mRNA or EST evidence in the region (circa 3 kb upstream), still the Affymetrix probe showed an expression for this mRNA in HeLa S3 cells. To explore whether SP5 has any promoter characteristics, I searched the sequence for putative transcription factor binding sites using the program MAPPER (Marinescu et al., 2005).

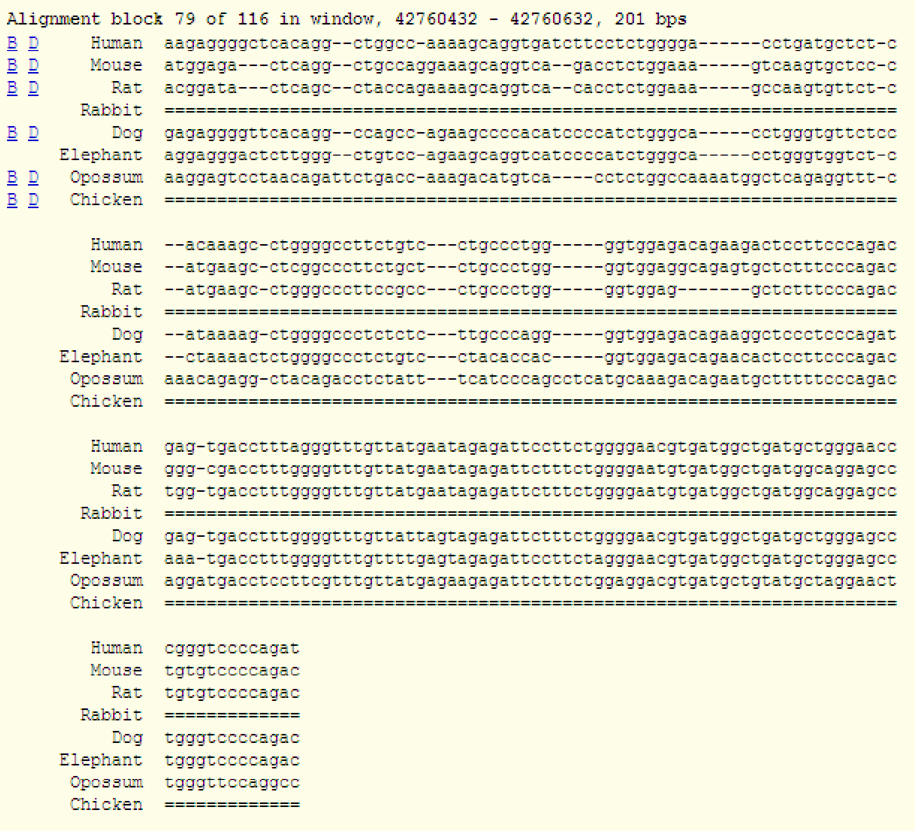


TEF-1	Transcription Enhancer Factor-1 (interacts with TBP and AP-1)
TREB-1	Tax responsive element binding protein-1 (interacts with TBP)
TBP	TATA-box binding protein
AP-1	Activator protein-1
JunD	Jun D proto oncogene (functional part of AP-1 activation complex)
Max	Myc binding protein X (interacts with TEF-1)
PITX2	paired-like homeodomain transcription factor 2 (expressed only in HeLa S3 cells
C/EBPbeta	CCAAT/enhancer binding protein beta subunit

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.

Figure 5.29 shows a graphical display of the putative binding sites on the polII enriched region for the human transcription factors. Only factors expressed in HeLa S3 cells were included in this analysis. Interestingly, this region contains two perfect TATA boxes and binding sites for TEF-1 protein which is an activator that binds to enhancers and also interacts with TATA-box binding protein (TBP) (Gruda et al., 1993). This region also has binding sites for common transcription factors such as AP-1, JunD and Max that act as activators together on a number of human promoters. More importantly, it also contains a binding site for enhancer binding factor C/EBPβ, which interacts with RNA polymerase II initiation complex (Mo et al., 2004). However the lack of H3K4me3 and H3Ac enrichment raises questions whether this region has potential to be a real TSS. In addition, no common promoter elements such as GC-boxes (Sp1 binding sites), initiator element or CAAT-box were found.

Within SP9 (Table 5.7), there is a 201 bp long region which is highly conserved across five species. This spot showed H3K4me (~10 fold), H3K4me2 (~2.2 fold), H3Ac (~2 fold) and H4Ac (~5 fold) enrichments. The alignment of this conserved region across five species is shown in Figure 5.30.



```
Alignment block 79 of 116 in window, 42760432 - 42760632, 201 bps
B D     Human  aagaggggctcacagg--ctggcc-aaaagcaggtgatcttcctctgggga------cctgatgctct-c
B D     Mouse  atggaga---ctcagg--ctgccaggaaagcaggtca--gacctctggaaa-----gtcaagtgctcc-c
B D       Rat  acggata---ctcagc--ctaccagaaaagcaggtca--cacctctggaaa-----gccaagtgttct-c
       Rabbit  ======================================================================
B D       Dog  gagaggggttcacagg--ccagcc-agaagccccacatccccatctgggca-----cctgggtgttctcc
     Elephant  aggagggactcttggg--ctgtcc-agaagcaggtcatccccatctgggca-----cctgggtggtct-c
B D   Opossum  aaggagtcctaacagattctgacc-aaagacatgtca----cctctggccaaaatggctcagaggttt-c
B D   Chicken  ======================================================================

        Human  --acaaagc-ctggggccttctgtc---ctgccctgg-----ggtggagacagaagactccttcccagac
        Mouse  --atgaagc-ctcggcccttctgct---ctgccctgg-----ggtggaggcagagtgctctttcccagac
          Rat  --atgaagc-ctgggcccttccgcc---ctgccctgg-----ggtggag-------gctctttcccagac
       Rabbit  ======================================================================
          Dog  --ataaaag-ctggggccctctctc---ttgcccagg-----ggtggagacagaaggctccctcccagat
     Elephant  --ctaaaactctggggccctctgtc---ctacaccac-----ggtggagacagaacactccttcccagac
      Opossum  aaacagagg-ctacagacctctatt---tcatcccagcctcatgcaaagacagaatgcttttcccagac
      Chicken  ======================================================================

        Human  gag-tgacctttagggtttgttatgaatagagattccttctggggaacgtgatggctgatgctgggaacc
        Mouse  ggg-cgacctttggggtttgttatgaatagagattctttctggggaatgtgatggctgatggcaggagcc
          Rat  tgg-tgacctttggggtttgttatgaatagagattctttctggggaatgtgatggctgatggcaggagcc
       Rabbit  ======================================================================
          Dog  gag-tgacctttggggtttgttattagtagagattctttctggggaacgtgatggctgatgctgggagcc
     Elephant  aaa-tgacctttggggtttgttttgagtagagattccttctagggaacgtgatggctgatgctgggagcc
      Opossum  aggatgacctccttcgtttgttatgagaagagattctttctggaggacgtgatgctgtatgctaggaact
      Chicken  ======================================================================

        Human  cgggtccccagat
        Mouse  tgtgtccccagac
          Rat  tgtgtccccagac
       Rabbit  =============
          Dog  tgggtccccagac
     Elephant  tgggtccccagac
      Opossum  tgggttccaggcc
      Chicken  =============
```

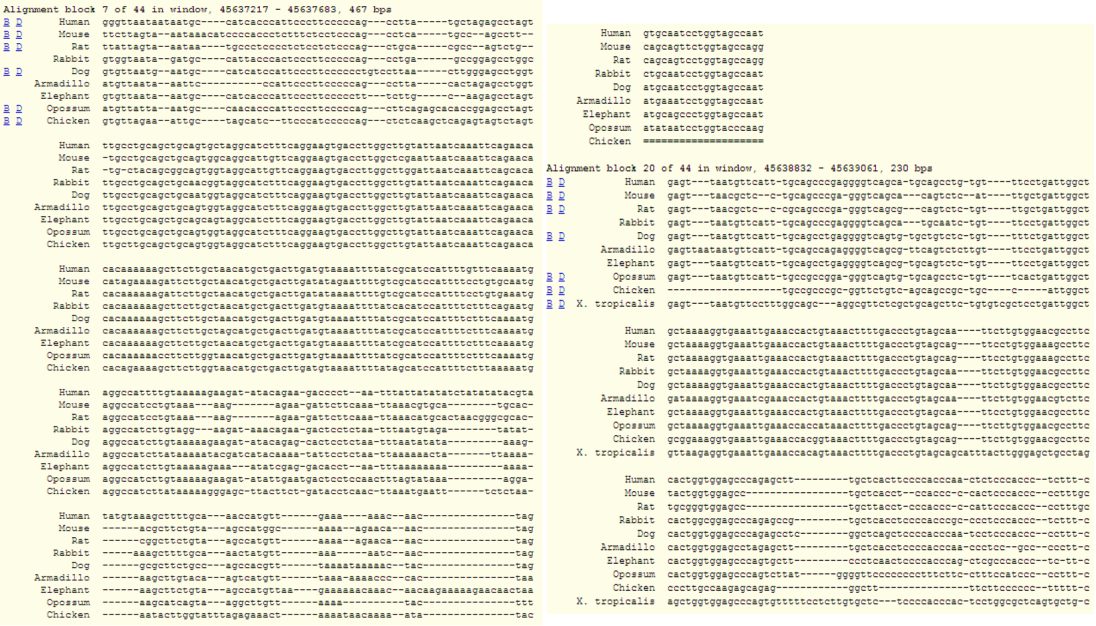
Figure 5.30 The alignment of the human DNA sequence spanning between 42,760,432 and 42,760,632 bp coordinates to mouse, rat, dog, elephant and opossum sequences. This alignment is reproduced from UCSD Genome Browser. This conserved region is within the spot that gave enrichments with H3K4me, H3K4me2,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

This highly conserved region was searched for putative binding sites of transcription factors by the program MAPPER and it was found to contain binding sites for AP-2, Max-1, SPI-B, CD28RC and PU-1 transcription factors.

In light of these findings, this region (SP5 to SP9) may well be a tissue specific distal regulatory element whose function seems to be regulated by different histone modifications especially by mono-methylated histone H3 at K4 and acetylated histone H4. This region may also have strong interactions with the initiation complex on the promoter of its target gene located elsewhere on the genome and as a result, it gets crosslinked and co-immunoprecipitated as if polII bridged these two sequences. This hypothesis can be tested by assessing the enhancer activity of this region on a promoter (or a number of promoters) via gene reporter assays. The promoter activity of the region with a polII enrichment can also be tested using gene reporter assays. Note that this region (SP5 to SP9) is enriched with H3K4me and H3K4me2 also in NTERA-D1 cells (thick blue and green curves in Figure 5.28). Interestingly, this may indicate that the above combination of histone modifications can mark distal elements irrespective of their activity status in different cell types. Additionally, SP5, which shows a polII enrichment in HeLa S3 cells, did not show any polII but CTCF (3.2 fold) enrichment in NTERA-D1 cells. This region might be repressed in NTERA-D1 cells due to CTCF binding which can function as a repressor (see section 1.3.2.1).

SP4 is located within the first intron of the *ADA* (42,704,657 to 42,706,871 bp) and showed 2.6 and 1.8 fold enrichment with H3K4me and H3K4me2. In NTERA-D1 cells, SP4 including its left and right spots (42,702,249 to 42,708,995) also showed H3K4me and H3Kme2 enrichments (see Table 5.8). An intronic enhancer of the ADA promoter has already been located between 42,706,026 and 42,709,030 region that contains three DNase hypersensitive sites (Aronow et al., 1989). This intronic enhancer is contained within the spots that showed enrichment with H3K4me and H3K4me2 in both cell lines. This observation supports the hypothesis that mono-methylated histone H3 can indeed be a marker for distal enhancer elements (Ren, B., unpublished data).

SP26 is enriched with H3K4me (3.9 fold) and H3K4me2 (3.6 fold). This region lies within the intron of NCOA3 gene and it contains two ultra conserved regions of 176 and 179 bp long; the genomic alignments of these regions are given in Figure 5.31. Analysis of the first region with the program MAPPER found a perfect-match binding site for the six members of ETS family transcription factors. These proteins are transcriptional activators; mainly signalling pathways such as MAP kinases or Ca^{+2} specific signals activated by growth factors or cellular responses converge on the ETS family proteins, controlling their activity, interactions and specification of their downstream targets (Yordy and Muise-Helmericks, 2000). In addition, there is a binding site for a nuclear receptor factor (NR5A2), an enhancer binding factor (Li et al., 1998), which interestingly is known to interact with NCOA3 protein (Ortlund et al., 2005). To estimate the probability of finding a putative binding site for NR5A2 by chance, a 30 kb intergenic sequence was searched for its binding site, one NR5A2 binding site was found per 2.63 kb. I also searched a 600 kb sequence containing several genes, and the program found one NR5A2 binding site approximately per 2.94 kb. Thus, there is a high probability that the putative NR5A2 binding site found within this 176 bp highly conserved site is real.



```
Alignment block 7 of 44 in window, 45637217 - 45637683, 467 bps
B D      Human  gggttaataataatgc----catcacccattcccttcccccag---cctta-----tgctagagcctagt
B D      Mouse  ttcttagta--aataaacatccccaccctctttctcctcccag---cctca-----tgcc--agcctt--
B D        Rat  ttattagta--aataa----tgccctcccctctcctctcccag---ctgca-----cgcc--agtctg--
        Rabbit  gtggtaata--gatgc----cattacccactcccttcccccag---cctga------gccggagcctggc
B D        Dog  gtgttaatg--aatgc----catcatccattcccttccccctgtccttaa-----cttgggagcctggt
     Armadillo  atgttaata--aattc---------ccattcccttcccccag---cctta-----cactagagcctggt
      Elephant  gtgttaata--aatgc----catcacccattcccttccccctt---tcttg-----c--aagagcctagt
B D    Opossum  atgttatta--aatgc----caacacccattcccttcccccag---cttcagagcacaccggagcctagt
B D    Chicken  gtgttagaa--attgc----tagcatc--ttcccatcccccag---ctctcaagctcagagtagtctagt

         Human  ttgcctgcagctgcagtgctaggcatctttcaggaagtgaccttggcttgtattaatcaaattcagaaca
         Mouse  -tgcctgcagctgcagtggcaggcattgttcaggaagtgaccttggctcgaattaatcaaattcagaaca
           Rat  -tg-ctacagcggcagtggtaggcattgttcaggaagtgaccttggcttggattaatcaaattcagcaca
        Rabbit  ttgcctgcagctgcaacggtaggcatctttcaggaagtgaccttggcttgtattaatcaaattcagaaca
           Dog  ttgcctgcagctgcaatggtaggcatctttcaggaagtgaccttggcttgtattaatcaaattcagaaca
     Armadillo  ttgcctgcagctgcagtggtaggcatctttcaggaagtgaccttggcttgtattaatcaaattcagaaca
      Elephant  ttgcctgcagctgcagcagtaggcatctttcaggaagtgaccttggcttgtattaatcaaattcagaaca
       Opossum  ttgcctgcagctgcagtggtaggcatctttcaggaagtgaccttggcttgtattaatcaaattcagaaca
       Chicken  ttgcttgcagctgcagtggtaggcatctttcaggaagtgaccttggcttgtattaatcaaattcagaaca

         Human  cacaaaaaagcttcttgctaacatgctgacttgatgtaaaattttatcgcatccatttttgtttcaaaatg
         Mouse  catagaaaagattcttgctaacatgctgacttgatatagaattttgtcgcatccattttcctgtgcaatg
           Rat  cacaaaaaagattcttgctaacatgctgacttgatataaaattttgtcgcatccattttcctgtgaaatg
        Rabbit  cacaaaaaagcttcttgctaacatgctgacttgatgtaaaattttatcacatccatttttctttcagaatg
           Dog  cacaaaaaagcttcttgctaacatgctgacttgatgtaaaattttatcgcatccatttttctttcaaaatg
     Armadillo  cacaaaaaagcttcttgctagcatgctgacttgatgtaaaattttatcgcatccatttttctttcaaaatg
      Elephant  cacaaaaaagcttcttgctaacatgctgacttgatgtaaaattttatcgcatccatttttctttcaaaatg
       Opossum  cacaaaaaaccttcttggtaacatgctgacttgatgtaaaattttatcgcatccatttttctttcaaaatg
       Chicken  cacagaaaagcttcttggtaacatgctgacttgatgtaaaattttatagcatccatttttctttaaaaatg

         Human  aggccattttgtaaaaagaagat-atacagaa-gacccct--aa-tttattatatatctatatatacgta
         Mouse  aggccatcctgtaaa---aag-------agaa-gattcttcaaa-ttaaacgtgca--------tgcac-
           Rat  aggccatcctgtaaa---aag-------agaa-gattcttcaaa-ttaaacatgcactaacgggcgcac-
        Rabbit  aggccatcttgtagg---aagat-aaacagaa-gactcctctaa-tttaatgtaga--------tatat-
           Dog  aggccatcttgtaaaaagaagat-atacagag-cactcctctaa-tttaatatata---------aaag-
     Armadillo  aggccatcttataaaaatacgatcatacaaaa-tattcctctaa-ttaaaaaacta-------ttaaaa-
      Elephant  aggccatcttgtaaaaagaaa---atatcgag-gacacct--aa-tttaaaaaaaa---------aaaa-
       Opossum  aggccatcttgtaaaaagaagat-atattgaatgactcctccaactttagtataaa---------agga-
       Chicken  aggccatcttataaaaagggagc-ttacttct-gatacctcaac-ttaaatgaatt------tctctaa-

         Human  tatgtaaagcttttgca---aaccatgtt------gaaa---aaac--aac--------------tag
         Mouse  ------acgcttctgta---agccatggc------aaaa--agaaca--aac--------------tag
           Rat  ------cggcttctgta---agccatgtt------aaaa--agaaca--aac--------------tag
        Rabbit  -----aaagcttttgca---aactatgtt------aaa----aatc--aac--------------tag
           Dog  ------gcgcttctgcc---agccacgtt------taaaataaaaac--tac--------------tag
     Armadillo  ------aagcttgtaca---agtcatgtt------taaa-aaaaccc--cac--------------taa
      Elephant  ------aagcttctgta---agccatgttaa----gaaaaaacaaac--aacaagaaaaagaacaactaa
       Opossum  ------aagcatcagta---aggcttgtt------aaaa----------tac--------------ttt
       Chicken  ------aatacttggtatttagagaaact------aaaataacaaaa--ata--------------tac
```

```
         Human  gtgcaatcctggtagccaat
         Mouse  cagcagttctggtagccagg
           Rat  cagcagtcctggtagccagg
        Rabbit  ctgcaatcctggtagccaat
           Dog  atgcaatcctggtagccaat
     Armadillo  atgaaatcctggtagccaat
      Elephant  atgcagccctggtagccaat
       Opossum  atataatcctggtacccaag
       Chicken  ====================

Alignment block 20 of 44 in window, 45638832 - 45639061, 230 bps
B D          Human  gagt---taatgttcatt-tgcagcccgaggggtcagca-tgcagcctg-tgt----ttcctgattggct
B D          Mouse  gagt---taacgctc--c-tgcagcccga-gggtcagca---cagtctc--at----ttgctgattggct
B D            Rat  gagt---taacgctc--c-cgcagcccga-gggtcagcg---cagtctc-tgt----ttgctgattggct
            Rabbit  gagt---taatgttcatt-tgcagcccgaggggtcagca---tgcaatc-tgt----ttcctgattggct
B D            Dog  gagt---taatgttcatt-tgcagcctgaggggtcagtg-tgctgtctc-tgt----tttctgattggct
         Armadillo  gagttaataatgttcatt-tgcagccagaggggtcagcg-ttcagtctcttgt----ttcctgattggct
          Elephant  gagt---taatgttcatt-tgcagcctgaggggtcagcg-tgcagtctc-tgt----ttcctgattggct
B D        Opossum  gagt---taatgttcatt-tgccgccgga-gggtcagtg-tgcagcctc-tgt----tcactgattggct
B D        Chicken  ------------------tgccgcccgc-ggttctgtc-agcagccgc-tgc----c-----attggct
B D  X. tropicalis  gagt---taatgttcctttggcagc---aggcgttctcgctgcagcttc-tgtgtcgctcctgattggct

             Human  gctaaaaggtgaaattgaaaccactgtaaacttttgaccctgtagcaa----ttcttgtggaacgccttc
             Mouse  gctaaaaggtgaaattgaaaccactgtaaacttttgaccctgtagcag----ttcctgtggaaagccttc
               Rat  gctaaaaggtgaaattgaaaccactgtaaacttttgaccctgtagcag----ttcctgtggaaagccttc
            Rabbit  gctaaaaggtgaaattgaaaccactgtaaacttttgaccctgtagcaa----ttcttgtggaacgccttc
               Dog  gctaaaaggtgaaattgaaaccactgtaaacttttgaccctgtagcaa----ttcttgtggaacgccttc
         Armadillo  gataaaaggtgaaatcgaaaccactgtaaacttttgaccctgtagcaa----ttcttgtggaacgtcttc
          Elephant  gctaaaaggtgaaattgaaaccactgtaaacttttgaccctgtagcaa----ttcctgtggaacgccttc
           Opossum  gctaaaaggtgaaattgaaaccaccataaacttttgaccctgtagcag----ttcttgtggaacgccttc
           Chicken  gcggaaaggtgaaattgaaaccacggtaaacttttgaccctgtagcag----ttcttgtggaacgccttc
     X. tropicalis  gttaagaggtgaaattgaaaccacagtaaacttttgaccctgtagcagcatttacttgggagctgcctag

             Human  cactggtggagcccagagctt---------tgctcacttccccacccaa-ctctcccaccc--tcttt-c
             Mouse  tactggtggagcc----------------tgctcacct--ccaccc-c-cactcccaccc--cctttgc
               Rat  tgcgggtggagcc----------------tgcttacct-cccaccc-c-cattcccaccc--cctttgc
            Rabbit  cactggcggagcccagagccg---------tgctcacctccccacccgc-ccctcccaccc--tcttt-c
               Dog  cactggtggagcccagagcctc--------ggctcagctccccacccaa-tcctcccaccc--ccttt-c
         Armadillo  cactggtggagcctagagctt---------tgctcacctccccacccaa-ccctcc--gcc--ccctt-c
          Elephant  cactggtggagcccagtgctt---------ccctcaactccccacccag-ctcgcccaccc--tc-tt-c
           Opossum  cactggtggagcccagtcttat------ggggttccccccctttcttc-ctttccatccc--ccttt-c
           Chicken  cccttgccaagagcagag------------ggctt---------------ttcttccccc--ttttt-c
     X. tropicalis  agctggtggagcccagtgtttttcctcttgtgctc---tccccacccac-tcctggcgctcagtgctg-c
```

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.

NCOA3 directly controls the expression of genes important for initiation of DNA replication and it has been linked to multiple types of human cancer due to its frequent over-expression (Louie et al., 2006). Moreover, NCOA3 has been shown to bind to its own promoter to direct its auto-regulation in a positive manner (Louie et al., 2006). Therefore, SP26 may actually play a role in auto-regulation of the *NCOA3*. This possibility surely requires further experimental verification, such as assessing the activity of SP26 on the *NCOA3* promoter by using for example, gene reporter assays. A ChIP experiment using an antibody recognizing the NCOA3 protein could also offer important insights on the regulatory elements of this gene. The second conserved site had putative binding sites for interferon regulatory factors, but no other relevant binding site could be found.

SP28 is also located in an intron of the *NCOA3*. It was enriched in H3K4me (3.1 fold), H3K4me2 (3 fold) and H4Ac (1.5 fold). This region contains a single short highly conserved site and it has moderate conservation with dog and armadillo (UCSC Genome Browser). Again, SP28 was searched by MAPPER and several putative binding sites for three relevant factors were found: five binding for the enhancer binding factor TEF-1, POU2F1 (Oct-1) and NR5A2 protein. The graphical display of the binding sites is shown in Figure 5.32. All these factors are known enhancer binding factors, and the arrangement of binding sites of TEF-1 and POU2F1 is quite interesting due to its similarity to that in the SV40 enhancer, on which they bind in symmetry to exert their activation functions (see Figure 1.4). For these reasons, SP28 is also a good candidate to be tested for enhancer activity.

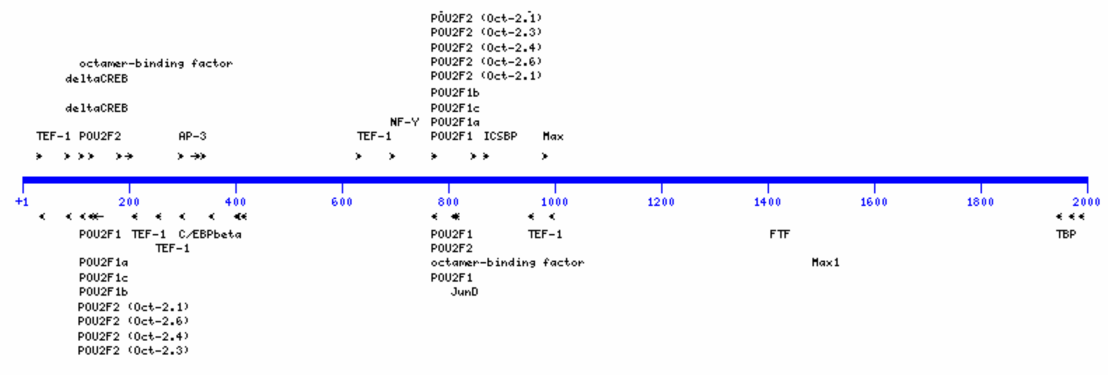


Figure 5.32 The putative binding sites on the sequence spanning from 45,648,944 to 46,651,138 bp. SP28 was enriched with H3K4me, H3K4me2 and H3Ac proteins.

SP24 showed enrichment with H3K4me (5 fold), H3K4me2 (1.8 fold) and H4Ac (2.6 fold) and contains multi-species conserved regions. No relevant binding sites were found by MAPPER on the conserved sites. SP24 contains six binding sites for Topors (topoisomerase I binding, arginine/serine rich) protein. It is unusual to find six Topors binding sites within 2255 bp since its consensus binding site sequence is 22 bp long. Topors have been shown to regulate the activity of p53 by ubiquitination and sumoylation, acting as tumour suppressors (Rajendra et al., 2004) (Weger et al., 2005). It also has trans-activating activities by binding to promoters and other regulatory elements of its target genes (Chu et al., 2001). Taking these observations together, this region seems to have a potential to be a distal regulatory element that requires further testing.

Within the *NCOA5*, SP16 is enriched in polII (1.8 fold), H3K4me (2.7 fold) and H3K4me2 (2.9 fold) and it has one highly conserved block of 226 bp. This region contains multiple putative binding sites for TEF-1, AP-1, RUNX1 (an enhancer binding protein) and SOX9 (enhancer binding protein). It is known that TEF-1 and AP-1 act in synergy on SV40 enhancer and exert their activation functions. In gene reporter assays, *NCOA5* showed one of the highest responses to SV40 enhancer which also carries binding sites for AP-1 and TEF-1 factors. Therefore this possible regulatory element can also have some activatory potential on the promoter of *NCOA5* in which this element resides.

It is important to note that for *NCOA5*, the TSS was not enriched in any protein studied, although *NCOA5* was highly expressed in both cell lines. The core promoter of this gene also showed very high activity (~900 fold increase compared to background constructs in HeLa cells) in reporter assays in both cell lines. No apparent problem was detected with the spot carrying the TSS. If the lack of any sort of enrichment on the TSS is not an experimental error, this gene may have a unique regulatory pattern where its promoter is not marked with common markers as H3K4me3 and H3Ac. Just to add here that there are 10 genes in HeLa S3 and 7 genes in NTERA-D1 cells that are expressed but their start sites are not enriched with polII or H3K4me3. Also, in a genome-wide promoter study by Ren et al, it was found that the start sites of 15% of expressed genes were not enriched with either polII, H3K4me3 or H3Ac (Kim et al., 2005).

SP1 lies within the *HNF4A* and showed polII (6.7 fold) and H3K4me2 (3.2 fold) enrichments. It contains a highly conserved region of 291 bp in which MAPPER found putative binding sites for interacting factors SMACA3 and Sp3. SMACA3 is a member of SWI/SNF chromatin modelling factors and Sp3 is a transcription factor acting as an activator or a repressor on several human promoters. No other binding sites for relevant proteins were found in SP1. The enrichment profile of SP1 its flanking region will be discussed in detail in section 5.10.

5.7.2 NTERA-D1 cells

This cell line exhibited a different histone enrichment profile than HeLa S3. In NTERA-D1, only 25 spots were enriched with H3K4me whereas this number was 71 in HeLa S3 cells. Also, in NTERA-D1 cells, 70% of the H3K4me enriched spots were in close proximity of annotated start sites while only 45% of those spots were in close proximity of annotated start sites in HeLa S3 cells. There were 52 spots enriched only with H3K4me2 in NTERA-D1 but only 10 spots in HeLa S3 cells. In NTERA-D1, 35% of these H3K4me2 enriched spots were in close proximity of annotated start sites.

Also, the H3Ac enrichment profiles were different between these two cell lines; the percentage occurrences of H3Ac enriched spots in combination with other modified histones is shown in Figure 5.33. In NTERA-D1, H3Ac enrichment was closely associated with H3K4me3 enrichment, whereas in HeLa S3, H3Ac was equally present together with H3K4me and H3K4me3 enrichments.

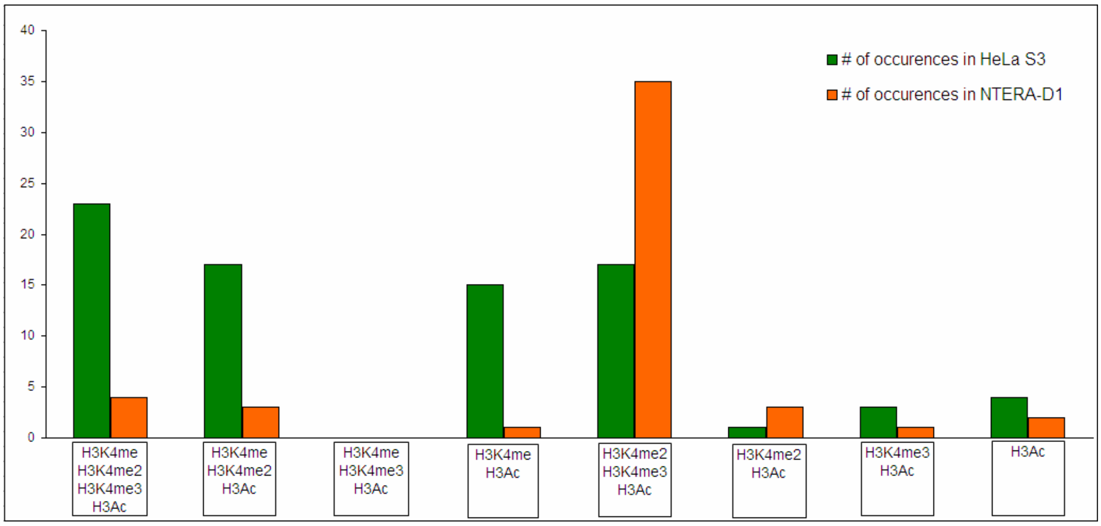


Figure 5.33 Percentages of the spots enriched with possible combinations of modified histones that includes H3Ac.

The enrichment profile of NTERA-D1 with H4Ac was very poor, only three spots showed enrichment higher than the selected threshold and all of them were in close proximity of the TSS of *PRKCBP1*. There were several enriched spots with H4Ac but they did not exceed the threshold level. Maybe this antibody did not work optimally in this cell type, although this is highly improbable since it worked very well in HeLa S3 and NTERA-D1 cells showed overall higher enrichment levels with most antibodies (except H4Ac and H3K9me2) compared to HeLa S3 cells.

It cannot be excluded that the observed differences are due to the use of tissue-specific combinations of modified histones in order to recruit different sets of activators on distal regulatory sites. A set of ChIP experiments using antibodies recognizing different sets of modified histones can be performed in NTERA-D1 cells to see their potential for marking distal regulatory elements.

These differences can also be due to harvesting the two cell lines at different time points in the cell cycle. This possibility can be tested by performing ChIP experiments with synchronized cell lines.

Index	Start Coord.	End Coord.	Spot Information	H3K4me	H3K4me2	H3Ac	H3K27me3	H3K9me2	CTCF
SP1	42385488	42387258	contains moderately conserved regions		2.038				
SP2	42586728	42588766	contains two short highly conserved region		1.718				
SP3	42670305	42672476	within PKIG gene; low conservation		1.583				4.1
SP4	42702249	42704639	within ADA gene; no conservation		3.845				
SP5	42704657	42706871	within ADA gene; no conservation	2.91	5.957	1.84			
SP6	42706827	42708995	within ADA gene; no conservation	2.289	2.717				
SP7	42754161	42755685	no conservation		3.17				
SP8	42754237	42755121	no conservation		2.933				
SP9	42755040	42757910	Low conservation		1.692				
SP10	42760403	42761222	contains one highly conserved region	2.046					
SP11	42761596	42763328	no conservation	2.167	2.758				
SP12	42762937	42764899	no conservation	2.541	2.917				
SP13	42789250	42791223	contains 3' end of WISP2 gene	1.952	1.78				
SP14	42790775	42793081	no conservation		2.987				
SP15	42825732	42827861	no conservation	1.69					
SP16	42845959	42847575	within RIMS4 gene; no conservation		1.73				
SP17	42866586	42869054	within RIMS4 gene; contains two highly conserved region	1.67					
SP18	43146874	43149361	no conservation		2.402				
SP19	43270902	43273021	contains 3' end of SEMG1 gene		2.948				
SP20	43319244	43320783	no conservation		2.095				
SP21	43398547	43400836	within SDC4 gene; contains one highly conserved region	2.346	2.43				
SP22	43400608	43402868	within SDC4 gene; contains two highly conserved regions	2.079	2.028				
SP23	43827441	43829560	no conservation		4.257				
SP24	43884167	43885901	contains 3' end of TNNC2 gene	2.03					
SP25	44063015	44065393	contains moderately conserved regions with mouse and rat	1.739					
SP26	44083316	44085609	contains 5' end of RP11-465L10.10 processed transcript; contains one promoter prediction		2.132				
SP27	44117171	44119404	within SLC12A5 gene; contains highly conserved regions		1.897		7.787	1.267	
SP28	44198835	44200465	no conservation		5.552				
SP29	44256074	44258486	within CDH22 gene; contains one ultra conserved region and a spliced EST	1.681					
SP30	44263883	44266386	no conservation	2.154					
SP31	44265868	44267498	no conservation	1.928	1.967				
SP32	44277449	44279433	within CDH22 gene; contains a highly conserved region		1.57				

SP33	44298319	44300488	within CDH22 gene; contains a highly conserved region		2.667				
SP34	44388316	44389991	contains one ultra conserved region	2.136					
SP35	44404114	44405724	no significant conservation			1.506			
SP36	44410407	44412877	contains the 3' end of SLC35C2 gene	2.47	3.077	1.546			3.9
SP37	44682236	44684551	within SLC123A3 gene; no conservation		1.807				
SP38	44736072	44738637	contains moderately conserved regions		2.018				
SP39	44746060	44748961	contains 3' end of TP53RK gene		1.937				
SP40	44764688	44766247	contains moderately conserved regions		1.88				
SP41	44775942	44778380	no conservation		1.947				
SP42	44939262	44941435	no conservation		5.843				
SP43	45007120	45009126	within EYA-2 gene; contains one ultra conserved region		3.875				
SP44	45053697	45056047	within EYA2 gene; no conservation		1.785				
SP45	45085601	45088163	within EYA2 gene; no conservation		1.715				
SP46	45179954	45182499	within EYA2 gene; contains two highly conserved region		2.675				
SP47	45433592	45436002	contains one moderately conserved region		2.185				
SP48	45581712	45582950	within NCOA3 gene; no conservation		4.692				
SP49	45617423	45620210	within NCOA3 gene; no conservation		4.438				
SP50	45620195	45622416	no conservation		2.775				
SP51	45636634	45639142	within NCOA3 gene; contains ultra conserved regions		1.828				
SP52	45648944	45651138	within NCOA3 gene; moderately conserved		2.557				
SP53	45748303	45750946	within SULF2 gene; low conservation		2.962				
SP54	45819538	45820937	within SULF2 gene;contains one highly conserved region		2.137				
SP55	45824411	45826796	one short highly conserved stretch			1.708			
SP56	45828163	45830521	within SULF2 gene;no conservation		4.222				

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.

Table 5.8 lists all enriched spots and a short sequence feature description in NTERA-D1 that are not in close proximity (more than 2 kb apart from) with a TSS. Several of these sites contain highly conserved sequences, while some are more than 95% conserved over 100 bp long stretches. The intronic sites were searched for putative binding sites for enhancer-binding proteins such as SMAD-3 and SMAD-4, Octamer binding proteins, TEF-1, AP-1 and AP-3. Since most of these spots were only enriched with one histone modifications, more evidence is needed for them to be treated as potential distal regulatory sites.

5.8 Histone Modification involved in transcriptionally inactive regions

ChIP experiments were performed using antibodies recognizing tri-methylated histone H3 at K27 (H3K27me3) and di-methylated histone H3 at K9 (H3K9me2) in both HeLa S3 and NTERA-D1 cells. These modified histones are often associated with X-chromosome inactivation (Plath et al., 2003), gene silencing (Plath et al., 2003) and heterochromatin formation and maintenance (Richards and Elgin, 2002) (Grewal and Moazed, 2003).

5.8.1 H3K27me3

5.8.1.1 HeLa S3

None of the spots containing TSS were found to be enriched with H3K427 higher than the threshold. A heat map was generated that displays the H327me3 signals within 6 kb distance of all the genes in comparison with H3K4me3 signals (Figure 5.34).

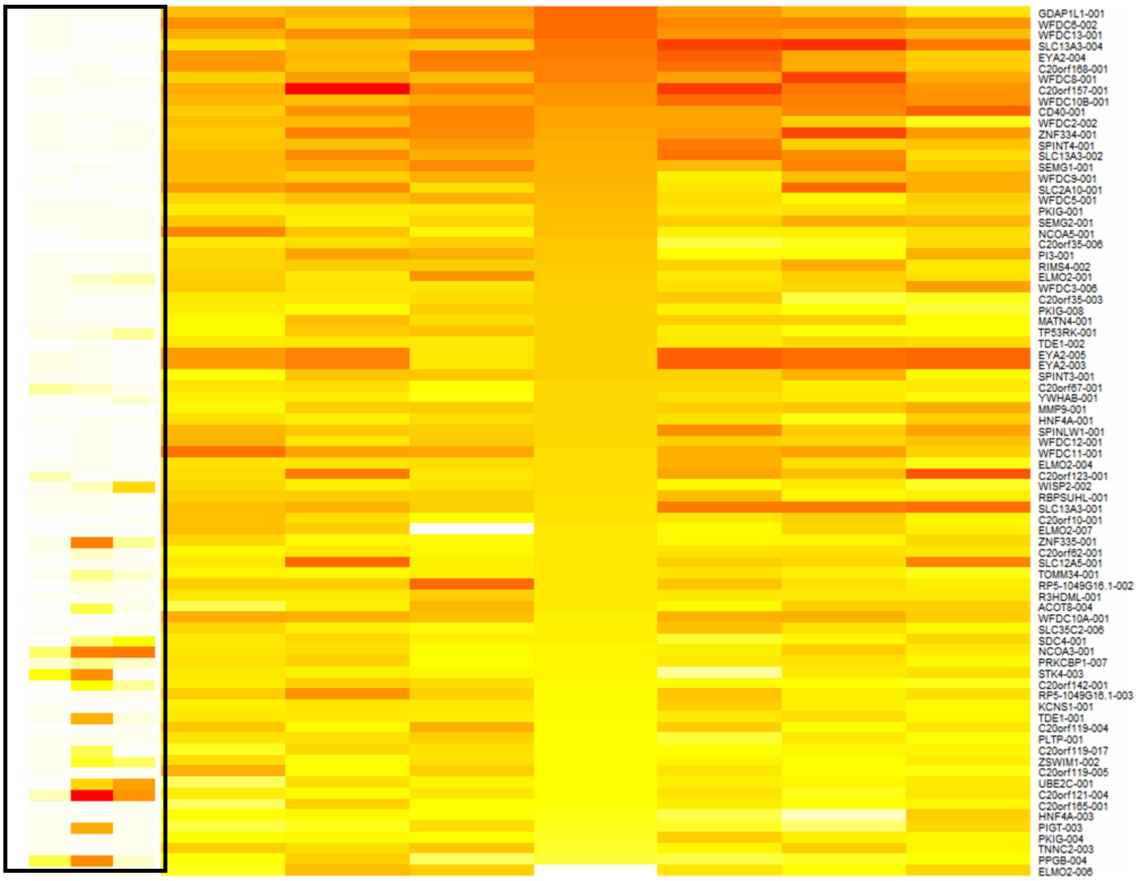


Figure 5.34 Heat map displaying H3K27me3 signals of 6 kb distance of 79 TSSs in HeLa S3 cells. The small heat map on the left displays the H3K4me3 signals of 2 kb distances of the corresponding start sites.

Figure 5.34 shows that H3K27me3 signals are negatively correlated with H3K4me3 signals at the start sites of genes in the region. This confirms that genes indeed carry different histone codes depending on their transcriptional activity status. Table 5.9 lists the 15 spots that gave signals above the selected threshold (1.50) with H3K27me3 antibody together with a short feature description of their sequences. One interesting point is that all the spots that are H3K27me3 enriched are located after a gene rich segment of the region. This may be a sign of a start of a different chromatin domain within the region.

Index	Start Coord.	End Coord.	Spot Information	H3K27me3	CTCF
SP1	44163647	44166500	no conservation	1.547	
SP2	44190355	44192401	contains 3' end of CD40 gene	2.502	4.844
SP3	44198835	44200465	no conservation	1.537	
SP4	44202007	44204336	contains short conserved regions	2.188	
SP5	44359438	44361613	5' end of a GENSCAN gene prediction; no conservation	1.62	
SP6	44360913	44362559	5' end of a GENSCAN gene prediction; contains one short highly conserved region	1.567	
SP7	44382683	44384608	no conservation	1.885	
SP8	44399167	44401245	contains one highly conserved region	1.765	
SP9	44401188	44402830	contains one short highly conserved region	1.505	
SP10	44540333	44542416	no conservation	1.771	
SP11	44637193	44639415	within SLC13A3 gene; contains one short moderately conserved region	3.175	
SP12	44705037	44707354	within SLC13A3 gene; no conservation	1.823	
SP13	45179954	45182499	within EYA2 gene; contains two highly conserved regions	1.663	
SP14	45807042	45808720	within SULF2 gene; no conservation	2.093	
SP15	45808066	45810484	within SULF2 gene; no conservation	2.311	

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.

Since there are several sites that are H3K27me3 enriched, this modified histone can be a functional marker of distal regulatory elements as well. It is known that histone H3 K27 tri-methylation facilitates binding of polycomb protein (a member of poycomb complex) to histone H3 and regulates the silencing of polycomb group genes (Cao et al., 2002). This constitutes an example of how different histone codes on a genomic site direct the recruitment of different sets of factors on a site. Therefore, I postulate that such regions have the potential to be functional regulatory elements and needs to be pursued further by ChIP experiments performed with antibodies recognizing a variety of transcription factors.

5.8.1.2 NTERA-D1

In NTERA-D1 cells, 49 spots were enriched with H3K27me3 and the histone enrichment profile is different than HeLa S3 cells. A heat map was generated that shows the enrichment levels within 6 kb of the TSSs in the region (Figure 5.35).

Nine genes showed H3K27me3 enrichment within 2 kb distance of their TSSs, and four of them also showed enrichment with H3K4me3. This is quite different from HeLa S3 cells in which no H3K4me3 enriched genes showed enrichment around the start site. On the other hand, none of the H3K27me3 enriched spots showed enrichment with polII or acetylated histone H3 or H4 in both cell lines. Table 5.10 lists the spots that are enriched with H3K27me3 along with the enrichment levels of the spots with other proteins used in this study.

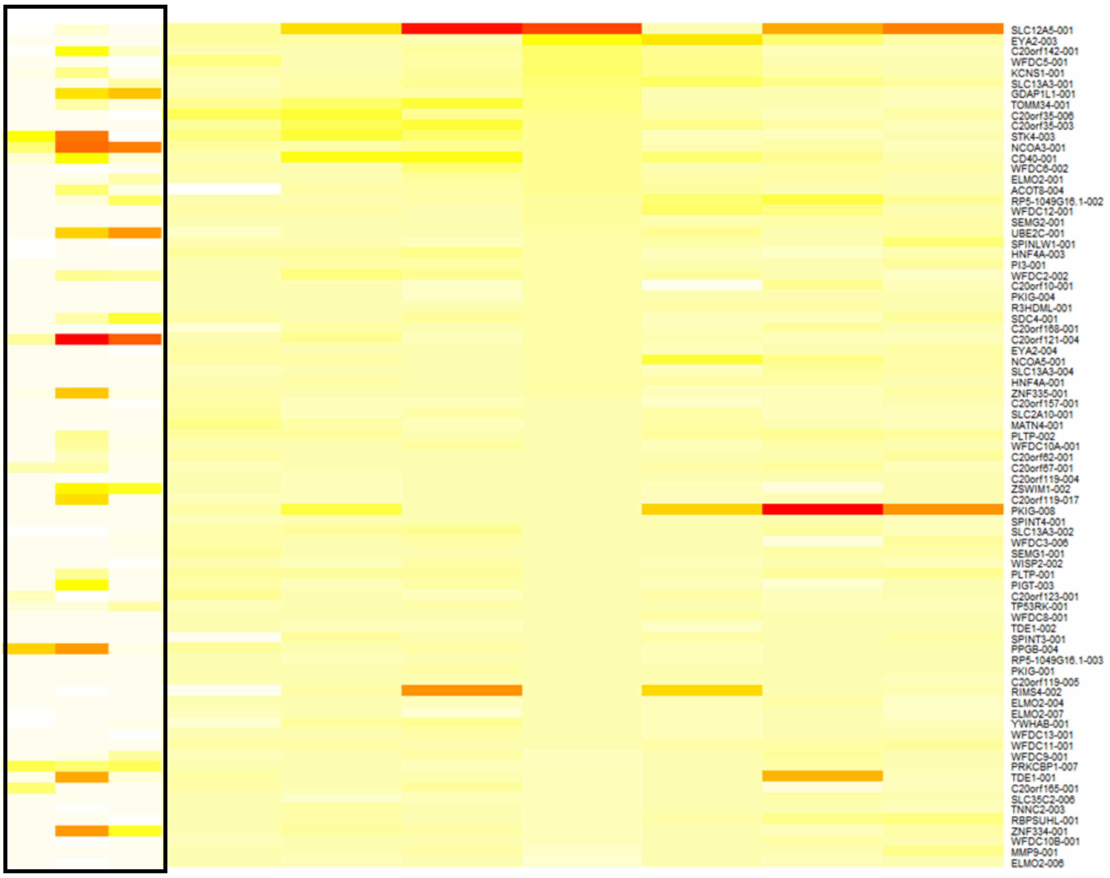


Figure 5.35 Heat map displaying H3K27me3 signals of 6 kb distance of 79 TSSs in NTERA-D1 cells. The small heat map on the left displays the H3K4me3 signals of 2 kb distances of the corresponding start sites.

Index	Start Coordinates	End Coordinates	Spot Information	H3K4me	H3K4me2	H3K4me3	H3K27me3	H3K9me2	CTCF
SP1	42387722	42389987	contains a promoter prediction				3.26		
SP2	42451724	42453980	within HNF4A gene; low conservation				1.58		
SP3	42580025	42581664	within SERINC3 gene; no conservation				5.1		
SP4	42677109	42679491	close to 5' end of PKIG gene	1.984			4.07		
SP5	42681139	42682610	contains 3' end of ADA gene				12.23		
SP6	42683191	42685461	within ADA				6.48		
SP7	42685423	42687531	contains 3' end of PKIG gene				1.69		
SP8	42689279	42691036	close to 5' end of non-coding transcript of ADA gene	1.657	10.82	3.273	21		
SP9	42868663	42871244	close to 5' end of RIMS4 gene				3.86		
SP10	42873659	42875941	close to 5' end of RIMS4 gene				6.44		
SP11	43024757	43026900	no conservation				1.61		
SP12	43465098	43467459	close to 5' end of DBNDD2 gene				1.6		
SP13	44086840	44088464	contains two highly conserved regions				3.72		
SP14	44088445	44090983	close to 5' end of SLC12A5 gene				11.46		
SP15	44090749	44092135	Contains 5' end of SLC12A5 gene		8.547	2.54	9.52		
SP16	44092559	44094553	close to 5' end of SLC12A5 gene		2.453	3.863	5.68		
SP17	44094336	44095961	within SLC12A5 gene				7.2		
SP18	44095668	44097726	within SLC12A5 gene				6.13		
SP19	44097020	44099292	within SLC12A5 gene				9.07		
SP20	44099144	44100963	within SLC12A5 gene				2.18		
SP21	44101265	44103436	contains 3' end of SLC12A5				1.72		
SP22	44103491	44106129	within SLC12A5 gene				3.87	1.082	
SP23	44104967	44107170	within SLC12A5 gene				2.16		
SP24	44106646	44108979	within SLC12A5 gene				1.51		
SP25	44113102	44115230	within SLC12A5 gene				1.59		
SP26	44114762	44117270	within SLC12A5 gene				8.4		
SP27	44117171	44119404	within SLC12A5 gene		1.897		7.79	1.267	
SP28	44120237	44121938	close to end of SLC12A5 gene (see section 5.5.2.1)		13.56	4.843	1.78		
SP29	44149238	44150590	close to 5' end of NCOA5 gene				1.55		
SP30	44176229	44177856	contains one ultra conserved region				1.93		
SP31	44177841	44179824	close to 5' end of CD40 gene		1.632	2.821	1.93		
SP32	44231425	44234066	contains short moderately conserved regions				2.83		
SP33	44233981	44235883	contains short moderately conserved regions				3.05		
SP34	44307554	44309441	within CDH22 gene; contains 3 ultra conserved regions				2.66		

SP35	44310685	44312911	close to 5' end of CDH22 gene		3.587	2.568	3.69		
SP36	44311936	44314006	contains 5' end of CD22 gene				1.69		
SP37	44313930	44316183	close to 5' end of CDH22 gene				2.54		
SP38	44364653	44367357	contains one moderately conserved region				2.08		
SP39	44366974	44369194	contains moderately conserved regions		2.217	1.873	2.69		2.884
SP40	44371086	44372681	contains 5' end of a gene prediction				3.45		
SP41	44375439	44377976	contains one highly conserved region				1.93		
SP42	44956175	44958701	contains 5' end of EYA2 gene				2.04		
SP43	44958627	44959383	close to 5' end of EYA2 gene				3.27		
SP44	45093985	45096221	within EYA2 gene; no conservation				1.51		
SP45	45096065	45098568	within EYA2 gene; contains short highly conserved regions				2.22		2.803
SP46	45533578	45536011	no conservation				14.04		
SP47	45535724	45537859	no conservation				22.34		
SP48	45588657	45591804	within NCOA3 gene; low conservation				2.02		
SP49	45591796	45593160	within NCOA3 gene; contains one highly conserved region				2.45		

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.

There is one particularly interesting region located between 45,533,578 to 45,537,859 bp which showed the highest enrichment (~20 fold) with H3K27me3 in NTERA-D1 cells. This spot is not within or close to any gene and it has no significant sequence conservation either. A TF binding site search of the region with MAPPER gave 13 different binding sites for PITX2 protein, a homeodomain containing transcription factor, and five sites for AREB6, another homeodomain zinc-finger protein. As mentioned, tri-methylation of histone H3 at K27 serves as a signal for recruiting polycomb gene silencing complex (the proteins necessary to maintain the silence state of *HOX* cluster) (Cao et al., 2002) on promoter regions of target genes for silencing. Since this region does not lie in close proximity of any promoter region, we cannot speculate a similar mechanism. However, it is known that polycomb proteins can exert their silencing effects via long range interactions where their distal DNA binding sites is looped onto the target promoters (Min et al., 2003) (Paro and Hogness, 1991) (Simon and Tamkun, 2002). Based on these putative binding sites for PITX2 and high levels of H3K27me3, these regions may harbour a cis-acting regulatory element of an as yet unknown target promoter. Sequences of some other H3K27me3 enriched spots were searched for putative binding regions, where no such significant sites can be found.

5.8.2 H3K9me2

ChIP experiments performed with this antibody did not show enrichment above the selected threshold of 1.5 in both cell lines. However, in HeLa S3 cells, the TSSs of three genes, *GDAP1L1*, *RIMS4* and *SLC13A3* showed ~1.5 fold enrichment. None of these genes showed any enrichment with polII or other modified histones and they were not expressed in HeLa S3 cells. Although the enrichment levels are relatively low, these results are in agreement with the findings that histone H3 methylation at K9 is associated with inactive chromatin regions (Shilatifard, 2006).

5.9 CTCF

An antibody recognizing the CCCTC-binding protein, CTCF, was employed in ChIP experiments in both HeLa S3 and NTERA-D1 cells in order to locate binding sites within the region of interest. CTCF is a versatile protein functioning as an activator or repressor on promoters or silencer sequences, or a chromatin insulator protein (Ohlsson et al., 2001). It is also located as part of multi-protein complexes regulating histone acetylation and deacetylation (Ohlsson et al., 2001). Here, I am reporting several potential CTCF binding sites and elaborate on their functional roles.

In HeLa S3 cells, there were 55 spots enriched with CTCF above the selected threshold level (1.75). Of these, 43 were also enriched with CTCF in NTERA-D1 cells. Similarly, out of 48 spots that showed enrichment with CTCF in NTERA-D1 cells, 45 were also enriched in HeLa S3 cells. Among CTCF enriched spots, 11 were within 2 kb distance of an annotated TSS, while 21 of them resided within the introns of annotated genes. Also, 12 intergenic regions were enriched with CTCF. A study was performed to locate CTCF binding sites on human chromosome 22 using ChIP assays, where they found around 200 binding sites within genes, promoters or intergenic regions (Mukhopadhyay et al., 2004). This is in agreement with the distribution of CTCF binding sites observed in this study. The summary of CTCF enriched spots that are close to or contain an annotated start site is given in Table 5.11.

Index	Start Coordinates	End Coordinates	Spot Information	CTCF peak HeLa S3 cells	Expression in HeLa S3 cells	CTCF peak NTERA-D1	Expression in NTERA-D1 cells
SP1	43950364	43952334	~1 kb upstream of C20orf165 gene		A	2.21	A
SP2	42309966	42311607	~1 kb upstream of GDAP1L1 gene	1.92	A	3.3	P
SP3	43173696	43176384	~2 kb downstream of WFDC5 gene; no conservation	2.26	A		A
SP4	45266520	45269030	~2 kb upstream of PKRCBP1 gene		P	1.78	P
SP5	42809326	42811520	2 kb downstream of 5' end of KCNK15 gene	6.62	P	4.61	A
SP6	43366633	43368867	contains 5' end of MATN4 gene	6.49	A	4.11	P
SP7	43368246	43370427	contains 5' end of RBPSUHL gene	8.67	A	5.46	A
SP8	43423952	43426324	contains 5' end of rejected C20orf169 gene	2.3	No probe		No probe
SP9	43691558	43692227	contains 5' end of WFDC10A gene	4.57	A	4.5	A
SP10	43692227	43694607	contains 5' end of WFDC9 gene	2.14	A		A
SP11	44031431	44032950	~ 2 kb upstream of ZNF335 gene	4.70	A	3.86	P
SP12	45724594	45727077	Contains 5’ end of SULF2 gene	3.79	A	5.15	P

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.

Out of eight genes whose start sites were enriched with CTCF, six of them were not expressed in HeLa S3 cells. The spot (SP8 in Table 5.11) containing the TSS of rejected *C20orf169* also showed enrichment with H3K4me3. Since C20orf169 has an ambiguous annotation (see Figure 5.22), the role of CTCF on the putative promoter of this gene is not clear. On the other hand, it is possible that CTCF acts as a repressor on the promoter of genes that are not expressed. It is known that CTCF can repress promoters by assisting histone deacetylase complexes assemble on promoter regions (Lutz et al., 2000). The fact that the putative *C20orf169* promoter was not enriched with acetylated histone H3 may be indicative of such a repression mechanism.

The *ZNF335* showed a very interesting enrichment pattern. It is enriched with CTCF, polII as well as modified histones marking for active genes (H3K4me3 and H3Ac) although no significant mRNA expression was detected by Affymetrix Expression Arrays. This gene encodes for a zinc finger protein which indirectly activates ligand-bound nuclear hormone receptors (Mahajan et al., 2002), and it also has a potential phosphorylation site (Beausoleil et al., 2004). The core promoter region of this gene does not contain any Sp1 binding site but a putative AP1 binding site on 14 bp downstream of its annotated start site, and this core promoter did not show any activity but its activity was restored when it was cloned with SV40 enhancer (see section 4.6.1). Since no mRNA expression can be detected despite the fact that the promoter region seems to be actively transcribed, there should be a mechanism which can either halt the initiation complex or decrease the mRNA stability. It has been shown that CTCT can act as a silencer by binding on promoter sites of the genes to be silenced (Klochkov et al., 2006) (Rakha et al., 2004). In the case of *ZNF335*, CTCF can be acting as a repressor by having a negative effect on the elongation process of the transcription.

In NTERA-D1 cells, there are four non-expressed genes, which were enriched with CTCF around their start sites. CTCF binding may be the cause of no transcription. On the other hand, the promoter region of *GDAP1L1* is enriched with CTCF and this gene is actively transcribed in NTERA-D1 cells. As discussed CTCF is pluripotent and in this instance may be part of an activation complex. Furthermore, a study by Norton et al showed that CTCF interacts with the large subunit of RNA polymerase II *in vitro* (Klenova et al., 2002), which can lend support to this hypothesis.

I attempted to localize CTCF binding sites within the promoter regions listed in Table 5.11 by MAPPER. Unfortunately, the program could not locate CTCF binding sites on most sequences within the threshold set for the search (90% confidence level). This was somewhat expected since CTCF uses its 11 zinc-finger motifs to bind sequences up to 50 bp long (section 1.3.2.1) which leads to many possible binding sequences that cannot be detected using one single consensus sequence.

There are 12 intergenic regions that are enriched with CTCF protein which are given in Table 5.12 and Figure 5.36 displays their genomic positions relative to gene features and other potential regulatory elements.

Since CTCF is mainly found acting on insulator elements, firstly I explored the possibility of these 12 regions being insulators. So many insulator elements within an area of less than 2.1 Mb may seem unrealistic, although it is known that insulator elements are mainly found in regions with high density of coding or regulatory sequences, which is the case for this specific region (Fourel et al., 2004). Also, detecting potential insulators only concentrated within the region of highest gene density supports these findings (see Figure 5.36).

Regions	Start Coordinates	End Coordinates	Spot Information	CTCF peak HeLa S3 cells	CTCF peak NTERA-D1 cells
R1	42744108	42746320	intergenic; no conservation	4	2.78
	42745943	42748103	intergenic; no conservation	4.35	3.29
R2	43199283	43201824	intergenic; no conservation	5.02	6.23
	43201123	43203622	intergenic; no conservation	4.87	4.18
R3	43288822	43291796	intergenic; no conservation	4.22	2.88
	43291402	43293747	intergenic; no conservation	2.28	
R4	43297185	43299177	intergenic; no conservation	3.21	3.65
R5	43352110	43353759	intergenic; no conservation	2.86	2
R6	43381309	43383563	intergenic; no conservation	2.78	2
R7	43507579	43509757	intergenic; no conservation	4.16	
R8	43518653	43521025	intergenic; no conservation		2.46
R9	43832093	43834748	intergenic; low conservation	2.07	
R10	44366974	44369194	intergenic; contains two short highly conserved regions	5.26	2.88
R11	44408689	44411452	intergenic; contains short highly conserved regions	3.11	3.76
R12	44820067	44822119	Intergenic; contains short moderately conserved regions	6.63	5.39

Table 5.12.CTCF enriched spots that lie within intergenic regions. The adjacent CTCF-enriched spots are treated as one region.

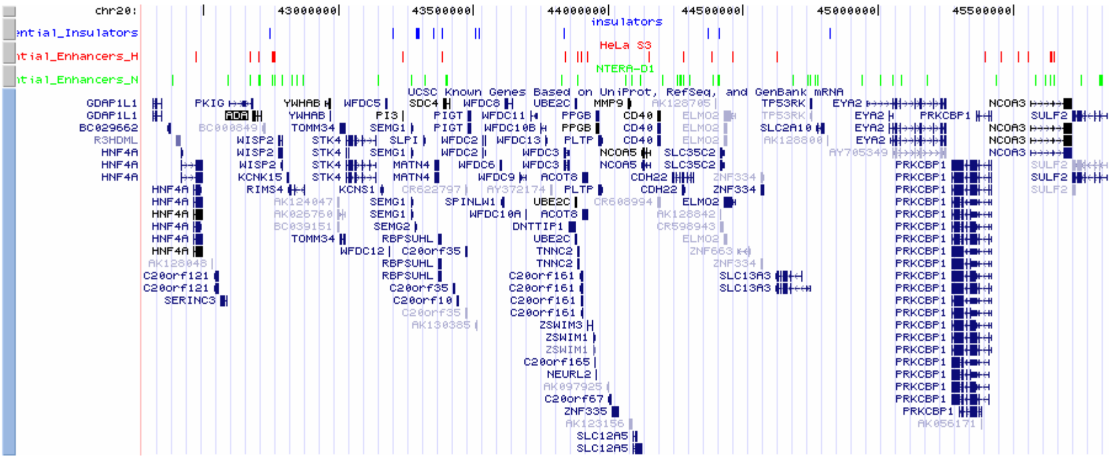


Figure 5.36 The genomic positions of 12 CTCF enriched regions (blue track) listed in Table 5.12 together with the H3K4me and H3K4me2 enriched regions in HeLa S3 (red track) and NTERA-D1 (green track) cells which are seen as potential distal regulatory elements. The region shown here covers the ~3.5 Mb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

Region 1 (R1;Table 5.12) is enriched with CTCF in both cell lines and it is a potential insulator since it lies between the intronic enhancer of the house-keeping *ADA* (section 5.7.1) and the promoter of the tissue-specific *WISP2*. The schematic representation of the region is shown in Figure 5.37.

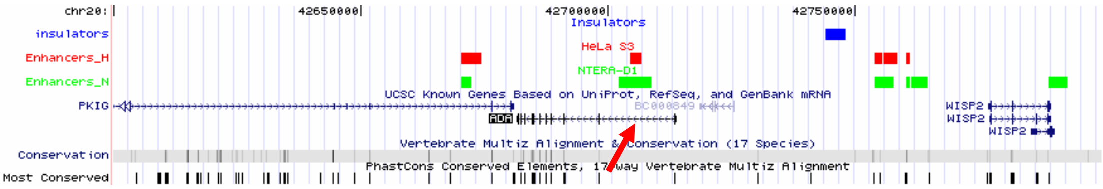

chr20:
42650000|
42700000|
42750000|
insulators
Insulators
Enhancers_H
HeLa S3
Enhancers_N
NTERA-D1
UCSC Known Genes Based on UniProt, RefSeq, and GenBank mRNA
PKIG
BC008849
ADA
WISP2
Vertebrate Multiz Alignment & Conservation (17 Species)
Conservation
PhastCons Conserved Elements, 17-way Vertebrate Multiz Alignment
Most Conserved

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.

Figure 5.37 also displays the regions that are enriched with H3K4me and H3K4me2 in HeLa and NTERA-D1 cells. The intronic *ADA* enhancer falls onto one of those enriched regions in both cell lines. This enhancer is a very strong regulatory element that can increase the *ADA* activity in a tissue-independent manner in *in vivo* studies (Aronow et al., 1989). Enhancer elements exert their effects over long distances in an orientation-independent manner. So this enhancer can be a problem if it was to act on a promoter such as that of the *WISP2* which has tissue specific expression and is probably functioning in bone turnover. An insulator element could solve this problem by blocking the communication between the enhancer and the promoter located on the other side of the insulator. The R1 CTCF-enriched site can be such an element.

There are five candidate insulator elements between 43,100,000 and 43,500,000 bp (Table 5.12). Figure 5.39 displays these putative insulator elements together with the regions that are seen as the possible cis-acting elements in section 5.7.1 and 5.7.2 due to their enrichments with H3K4me and H3K4me2. Each putative insulator lies between two candidate enhancers thereby blocking the communication between these elements to ensure proper regulation of the neighbouring genes. It is important to note that all the genes in this window exhibit a tissue-specific expression pattern.

A novel experimental design was developed to test potential insulators (Mukhopadhyay et al., 2004) and can be used to verify the putative insulator elements described here. The schematic diagram of the reporter construct used in this method is given in Figure 5.29.

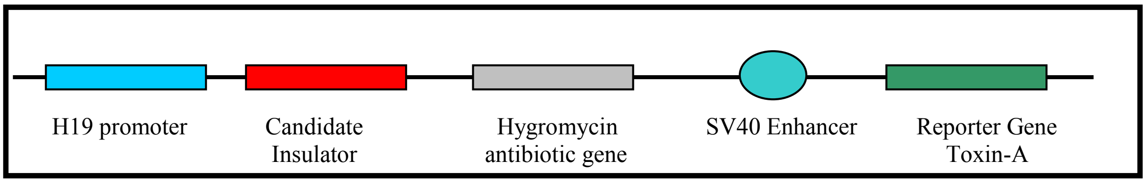


Figure 5.38 An insulator trap reporter vector construct where the candidate insulator is placed between H19 promoter and SV40 enhancer and promoter-enhancer activity is monitored by toxin-A reporter gene. In order to discriminate a possible silencing activity of the candidate fragment, hygromycin gene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

In this insulator trapping technique, a potential insulator is placed between a promoter (H19 in this example) and SV40 enhancer and the promoter-enhancer activity is monitored by toxin-A reporter gene. If the fragment is indeed an insulator, it will block the promoter-enhancer communication and toxin-A gene will not be expressed, hence the cells will survive. The number of cells survived will be determined by a cell-counting assay. An antibiotic hygromycin gene is also placed in downstream of the candidate insulator to discriminate its possible silencing activity since cells were also treated by hygromycin. So, if the fragment is a silencer, then it will silence the hygromycin gene where cells will die.

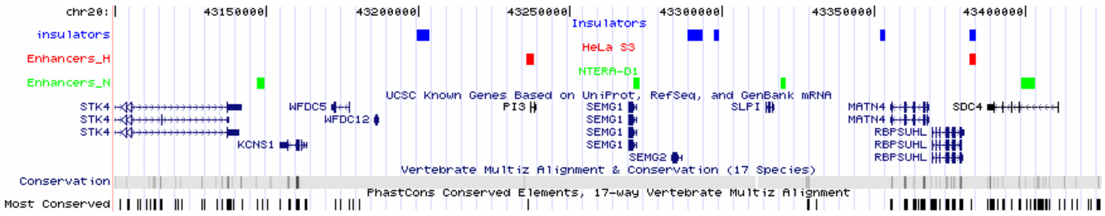


Figure 5.39 The region spanning from 43,100,000 to 43,425,000 bp where there are five candidate insulators shown as blue boxes on the insulator track. There are two more tracks, displaying H3K4me and H3K4me2 enriched regions in HeLa S3 (Enhancers_H track) and NTERA-D1 (Enhancers_N track) as possible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

There are 17 more CTCF enriched regions falling within the introns of genes in the regions. These regions are listed in Table 5.13 and schematically represented on the genome in Figure 5.40.

Regions	Start Coordinates	End Coordinates	Spot Information	CTCF peak in HeLa S3	CTCF peak in NTERA-D1 cells
R1	42449867	42451865	within HNF4A gene; one moderately conserved short region	3.90	2.85
R2	42509500	42511603	within HNF4A gene; one highly conserved short region	2.38	2.62
R3	42670305	42672476	within PKIG gene; no conservation	6.59	4.10
	42672013	42674371	within PKIG gene; no conservation	7.55	6.01
R4	42861319	42863581	within RIMS4 gene; contains one highly conserved short region	2.00	2.09
R5	43071862	43073714	within STK4 gene; contains moderately conserved short regions	5.79	3.95
R6	43105913	43106599	within STK4 gene; no conservation	7.31	3.47
R7	43158018	43159290	within KCNS1 gene; no conservation	3.20	3.20
R8	43860206	43861861	within DNTTIP1 gene; low conservation	2.27	
R9	44029429	44031576	within ZNF335 gene; no conservation	2.07	
	44031431	44032950	within ZNF335 gene; no conservation	4.70	3.86
R10	44073900	44075605	within MMP9 gene; low conservation	3.97	2.28
R11	44622588	44624763	within SLC13A3 gene; no conservation	4.74	2.77
	44624669	44627010	within SLC13A3 gene; no conservation	5.29	3.78
R12	44742326	44744698	within SLC13A3 gene; no conservation	2.88	
R13	45034995	45037048	within EYA2 gene; contains short highly conserved regions	3.28	3.50
R14	45096065	45098568	within EYA2 gene; contains short highly conserved regions	4.20	2.80
R15	45212776	45215255	within EYA2 gene; no conservation	3.15	2.68
R16	45724594	45727077	within SULF2 gene; no conservation	5.15	3.79
R17	45738270	45740862	within SULF2 gene; no conservation	8.67	5.69
	45739691	45741084	within SULF2 gene; no conservation	7.49	5.30

Table 5.13 CTCF enriched regions which fall within introns.

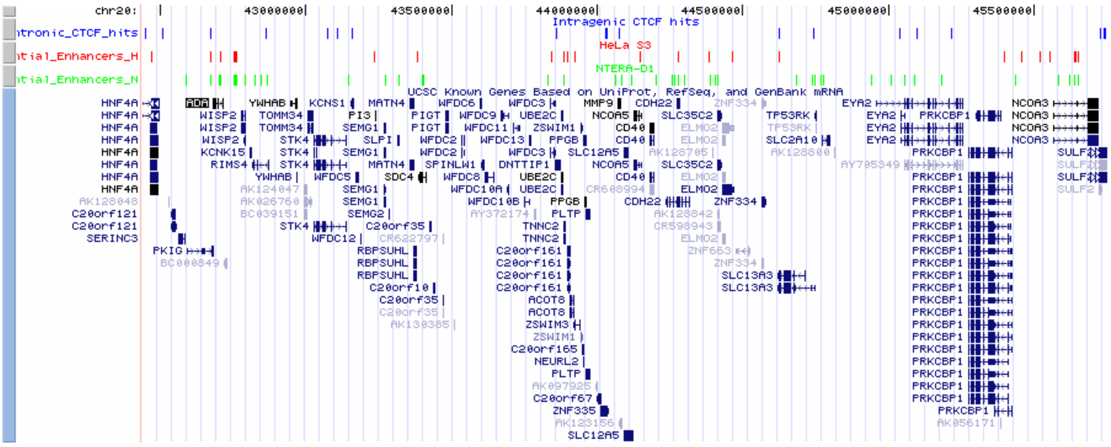


Figure 5.40 CTCF regions that falls within intronic regions shown in the blue track. Red and green tracks display the H3K4me and H3K4me2 enriched regions in HeLa S3 and NTERA-D1 cells respectively.

Interestingly, except *STK*-4, none of the remaining genes containing an intronic CTCF-enriched region showed strong H3K4me3 or polII enrichments on their start sites. While *HNF4A*, *RIMS4*, *KCNS1* and *EYA2* are not expressed in any, *PKIG*, *STK4* and *DNTTIP* genes are expressed in both cell lines according to Affymetrix expression arrays. Surely these regions may also function as insulators, however it would be unfeasible for the cell to use an intra-genic site as a silencer, since insulators need certain dynamic structural requirements such as looping over long distances, and this can be a problem when the gene carrying the insulator needs to be transcribed. On the other hand, some of these regions can be silencers and it is well established that CTCF plays a major role as a repressor in many silencer complexes (Lutz et al., 2000; Ohlsson et al., 2001; Klochkov et al., 2006). Long-range acting silencers are usually placed within introns or 3' ends of their target genes and they possess binding sites for repressor and co-repressor proteins. These regions need to be further investigated for their candidate silencing functions experimentally. This can be tested by replacing these regions together with a constitutively active strong promoter such as CMV or SV40 promoter and investigate their possible silencing effect using gene reporter assays.

5.10 Summary

Table 5.15 and Table 5.16 lists the signals obtained from eight of the antibodies used in this study in HeLa S3 and NTERA-D1 cells respectively. In both cell lines, more than half of the expressed genes (57%) showed binding of either polII, H3K4me3 or H3Ac. The remaining expressed genes did not give an enrichment with any of the above proteins. This may simply reflect suboptimal experimental conditions. However, ChIP experiments performed with H3K4me3 antibody showed very high enrichments especially in NTERA-D1 cells (~70-100 fold). It may also be that the remaining expressed genes are associated with a different set of modified histones, but this possibility is not favoured since in similar studies, most of the expressed genes are associated with H3K4me3 or H3Ac (Kim et al., 2005). Note that, the resolution of the custom-made array (~2 kb) is much lower than arrays used in similar studies (up to ~50 bp). This may also be a reason for not detecting all expressed genes via ChIP on chip. As the DNA size on the spot increases, the signal is diluted. Hence, it will be difficult to detect genes that are not highly enriched with these modifications. This will also be the case for genes that are activated in a transient fashion. Note that no correlation was observed between the expression signal of the gene obtained from Affymetrix Expression Arrays and ChIP enrichment on the start site.

Another issue is the reliability of the gene expression data. All expression arrays use probes which will detect either sense or antisense transcript of genes. Since antisense transcripts do not correspond to an actively expressed gene, a probe detecting such transcripts will lead to a false positive signal. It is not currently possible to determine the fraction of genes that produce an antisense transcript, although a new microarray platform is being developed to detect sense and antisense transcripts of a gene separately (Clevebring et al., 2006). Such a microarray platform will certainly improve the correlation between gene expression platforms and activity information obtained by ChIP experiments. They can also assist to resolve ChIP enrichment like those obtained from genes such as *MATN4* and *SLC12A5*, in this study where their 3'end was also enriched with H3K4me3 (see sections 5.4.1 and 5.5.2).

There are certain cases where annotated start sites produced lower enrichments than the flanking regions: for example, *PRKCBP1* and *SDC4* (see sections 5.4.1 and 5.4.2). The TSSs of these genes need further experimental examination although these variations may simply be due to a mis-assembled tilepath or a mixed well during PCR to construct the array.

A small subset of genes that are not expressed also showed enrichments with either polII, H3K4me3 or H3Ac (see Table 5.15 and Table 5.16). In HeLa S3, there are four such cases; *ZNF335*, *SLC12A5*, *EYA2* and *C20ORF165*. *ZNF335* also has a CTCF enrichment on its start site (see section 5.5.1). In NTERA-D1 cells, there are four such cases; *KCNS1*, *SLC12A5*, *WFDC10A* and *C20ORF165*, and all but *KCNS1* showed CTCF enrichment on the TSS. CTCF may be part of a silencing mechanism of the above genes. CTCF enrichment and the gene expression status of the CTCF enriched genes is given in Table 5.14

CTCF-enriched genes in	Expressed	Not expressed
HeLa S3	1	8
NTERA-D1	5	4

Table 5.14. Expression profile of the genes whose start sites are enriched with CTCF in HeLa S3 and NTERA-D1 cells.

CTCF obviously operates differently in the two cell lines (Table 5.14). In HeLa S3, it most likely acts as a repressor, whereas in NTERA-D1 cells, it is also seen on the TSSs of expressed genes. Three expressed genes (*MATN*4, *RBPSUHL* and *SULF2*) that showed CTCF enrichment contain no other signal on the start site in NTERA-D1, which may suggest that these genes have a transient expression pattern whose control involves CTCF. Yet, the start of these genes was enriched with CTCF also in HeLa S3 where they are not expressed. This may be an example of how the activity of a transcription factor, in this case CTCF, depends on the context of the cell it operates in.

The antibody recognizing H4Ac showed 60 enriched spots in HeLa S3 whereas only three in NTERA-D1 cells. Figure 5.41 shows the difference between the enrichment obtained by Rabbit IgG (negative control antibody) and H4Ac in NTERA-D1 cells on a small section of the array.

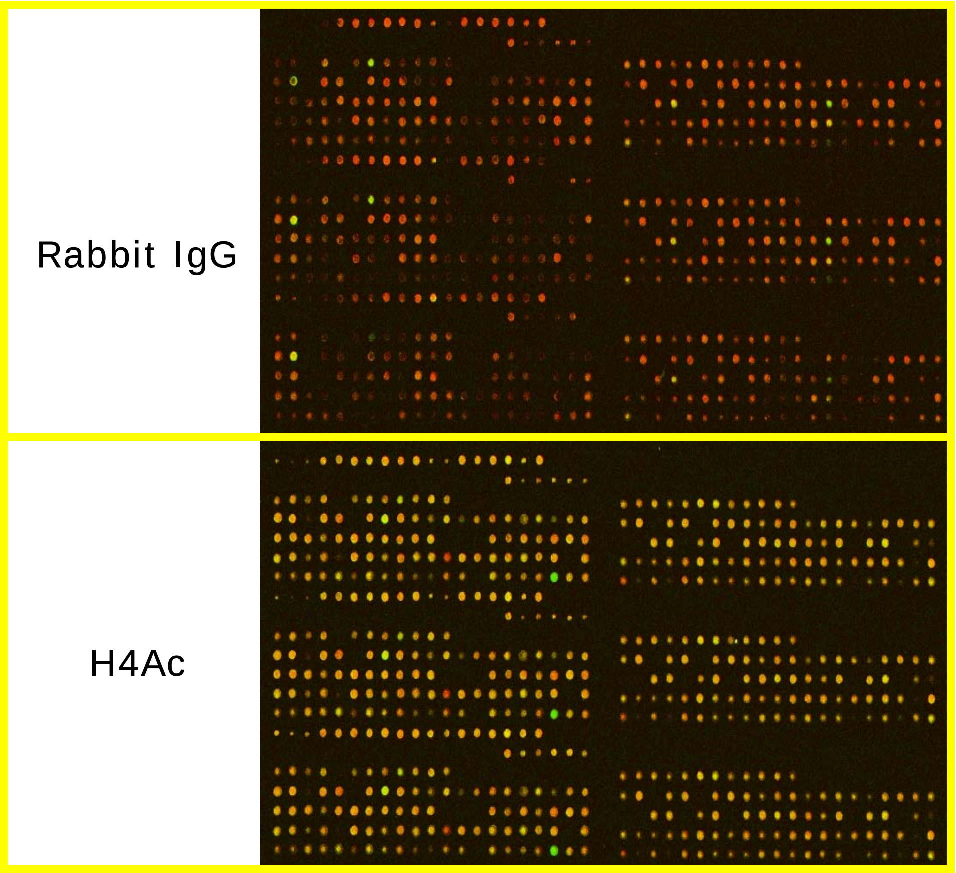


Figure 5.41 The enrichment difference between Rabbit IgG (negative control) (top) and H4Ac (bottom) antibodies on a small section of the custom-made array in NTERA-D1 cells. The green enriched spot on the bottom array carries upstream sequence of *PKRCBP1*.

The bottom array in Figure 5.41 displays a pattern as if all sites were enriched with H4Ac compared to the top array which shows Rabbit IgG enrichments. This difference may be an experimental artefact due to different working efficiencies of the antibody. However, Figure 5.42 shows the signals on the same section of the array obtained with a non-working antibody, Sp1 (see section 5.2.1).

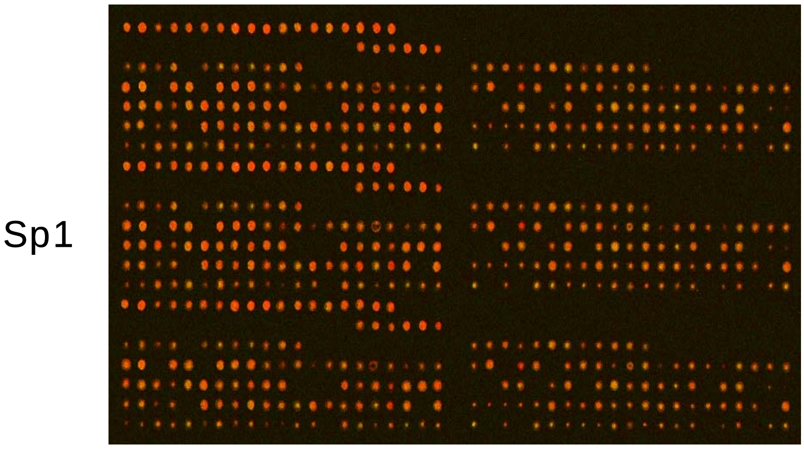


Figure 5.42. Enrichment profile of Sp1 on a subsection (the same section as in Figure 5.32) in NTERA-D1 cells.

The enrichment profile of H4Ac and Sp1 are different, that of Sp1 (non-working antibody) is similar to Rabbit IgG. This may suggest that H4Ac may be working in this cell line but the histone H4 acetylation pattern is widespread across the region. The fact that most of the other working antibodies performed relatively better in NTERA-D1 than in HeLa S3 further supports this claim. As mentioned in section 4.4.2, NTERA-D1 is established from a malignant germ cell tumour and it differentiates to neurons in response to RA (Mavilio et al., 1988; Pleasure and Lee, 1993; Segars et al., 1993). HeLa S3 cells are not from germ lines but epithelial tumour. Therefore, it is expected to see major differences in the histone code of these two cell lines.

H3K27me3 enrichment showed also very different patterns between the two cell lines. In HeLa S3, none of the gene start sites showed enrichment with H3K27me3 and the H3K27me3 enrichment was negatively correlated with that of H3K4me3 (see section 5.8.1.1). However, nine annotated start sites showed H327me3 enrichments in NTERA-D1 cells and four were also enriched with H3K4me3. Also, in HeLa S3 there were 15 spots enriched with H3K27me3 in total whereas this number was 50 in

NTERA-D1 cells. H3K27me3 is an assisting factor in polycomb-mediated gene silencing and it is shown that HOX cluster is enriched with this modification, a gene cluster which is developmentally regulated. The difference in H3K27me3 enrichment profile between the two cell lines might again be due to their different origin. The nine genes whose start sites were enriched with H3K27me3 have the same expression pattern in both cell lines, therefore H3K27me3 does not seem to play a role in the expression of these genes, but the histone enrichment profile of these genes is quite different between the two cell lines (Table 5.17).

One interesting case is *SLC12A5* where there is strong H3K4me3 on downstream of the TSS together with an H3K27me3enrichment. Also, this gene is enriched with H3K27me3 across its sequence as shown in Figure 5.43. Yet, there is a moderate H3K4me3 enrichment together with a strong H3K4me2 enrichment on the spot carrying the 3’ UTR. This spot also carries a CpG island (length=1856 bp, GC%=64.3, Obs(CpG)/Exp(CpG)=0.757), a promoter prediction (FirstExon) and TSS predictions (Eponine).

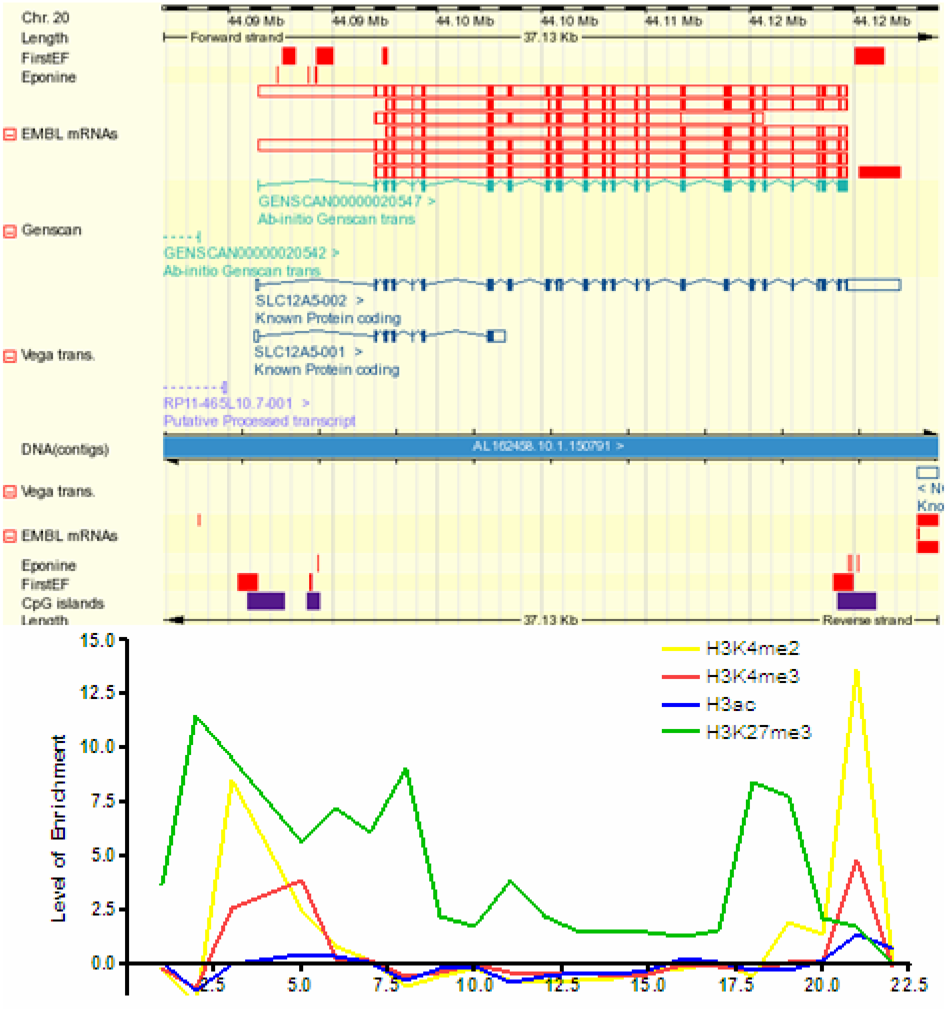


Figure 5.43 Enrichment levels on SLC12A5 gene with antibodies recognizing H3K4me2, H3K4me3, H3Ac and H3K27me3 together with the annotation of the gene taken from Ensembl Genome Browser.

SLC12A5 encodes for a membrane protein responsible for co-transporting potassium and chloride ions across the cell membrane. It has a tissue specific expression pattern, restricted to neurons in the central nervous system and retina, and it plays an important role in neuronal development (Hebert et al., 2004). There is no expression of this gene is in either HeLa S3 or NTERA-D1 cells. In HeLa S3, there is a ~2.5 fold H3K4me3 and H3Ac enrichment on the TSS, but there is no other significant enrichment with any other proteins used overall of the gene sequence. It is intriguing that NTERA-D1 requires such epigenetic marking across the entire gene, while in HeLa S3 cells such marking seems unnecessary to keep the gene silent. Such differences in epigenetic signatures in the same gene show that the histone code is dependent on the cellular context. It might be that H3K27me3 modification interferes with an activatory complex that exists in only in NTERA-D1 cells to prevent the activation of the gene. It is still not clear why the gene has such H3K4me2 and H3K4me3 peaks on its 3' end. A possible explanation for these enrichments is the presence of a silencing mechanism by antisense RNA where the H3K4me peaks may regulate the transcription of an antisense RNA to ensure that the gene is not expressed. These results suggest that H3K27me3 has implications in regulatory mechanisms other than gene silencing.

Histone modifications such as H3K4me and H3K4me2 were found within intergenic and intra-genic regions some of which contained ultra conserved CNGs. The regulatory potential of such regions needs to be verified experimentally by either using gene reporter assays or ChIP analysis performed with antibodies recognizing common or specific enhancer binding factors such as p300 or TEF-1. The differential histone code on these elements supports the fact that different sets of histone codes are employed in different cell lines most probably depending on the activity status of the element (see section 5.6.1 and 5.6.2).

A subset of polII enriched regions in HeLa S3 cells were within inter-genic regions and the cis-acting regulatory potential was discussed in section 5.6.1. A polII enriched spot located between 42,469,627 and 42,471,927 lies within the *HNF4A* which encodes a transcription factor regulating the expression of hundreds of genes in liver and pancreas cells (Odom et al., 2004). In the beta cells of pancreas, HNF4A regulated the insulin secretion in response to glucose level. Mutations in *HNF4A* are associated with monogenic autosomal dominant non-insulin-dependent diabetes mellitus type I, also four SNPs found in the 10.7 kb region encompassing P2 promoter is shown to be associated with T2D in some populations (Bagwell et al., 2005). *HNF4A* showed an interesting enrichment pattern across the gene (Figure 5.44).

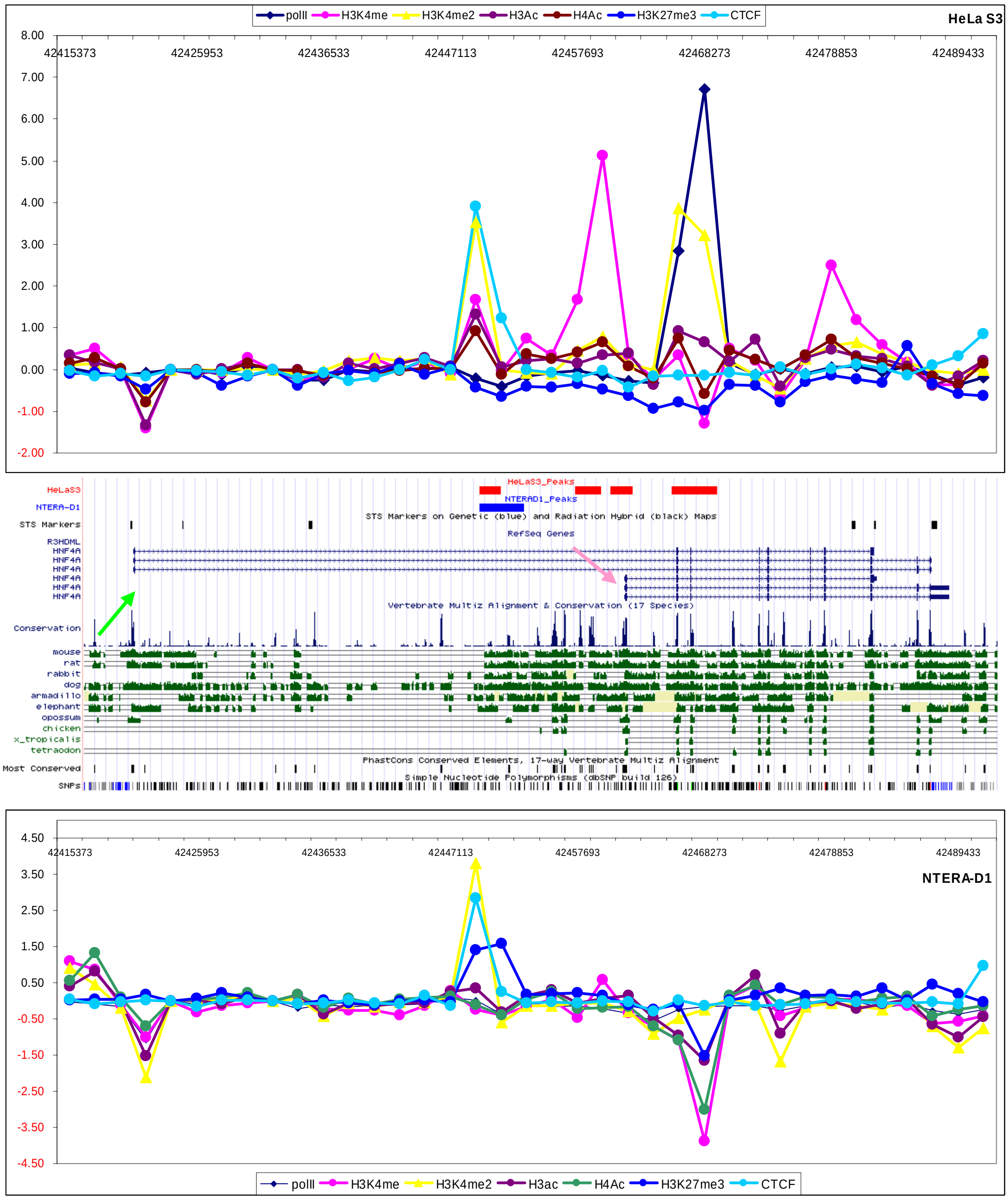


Figure 5.44. Enrichment profile across *HNF4A* in HeLa S3 and NTERA-D1 cells. The annotation is reproduced from UCSC genome browser. Green arrow denotes P2 promoter and pink arrow denotes the P1 promoter. The peaks are also displayed as custom annotation tracks (red and blue boxes).

None of the known enhancer elements of *HNF4A* showed enrichment in the two cell lines used (Mitchell et al., 2002). P2 promoter (green arrow in Figure 5.44) showed depletion for several antibodies in both cell lines. P1 promoter (pink arrow in Figure 5.44) showed H3K4me enrichment (~5.8 fold) in HeLa S3 but none in NTERA-D1 cells. The region around 32 kb upstream of the P2 promoter (first red and blue boxes in Figure 5.44) is enriched with CTCF in both cell lines. There is another region located around 7.5 kb downstream of the P1 promoter (fourth red box in Figure 5.44) which is enriched with polII as well as H3K4me2. Interestingly, this is where most of the antibodies show a depleted signal in NTERA-D1. The polII peak in this region might again be a sign of a distal regulatory element (see section 5.6.1). Note that all these regions contain CNGs.

HNF4A is not expressed in any of the two cell lines investigated. As expected, the P2 promoter did not show any activity but it did show strong activity in synergy with the SV40 enhancer in both cell lines (see section 4.6). The cloning of the P1 promoter to pGL3-basic (enhancer-less) plasmid was unsuccessful but it was cloned to pGL3-enhancer plasmid (carrying SV40 enhancer) and P1 promoter-SV40 enhancer construct showed strong activity in both cell lines. These results suggest that *HNF4A* is under the effect of an epigenetic silencing (probably involved with intra-genic regions mentioned above) rather than a dominant trans-acting silencing mechanism. These two regions need to be investigated further in other cell lines where *HNFA* is active to see any alteration of the histone code on these elements.

In a recent study, the transcriptional map of approximately 30% of the human genome was produced at 5-nucleotide resolution using DNA microarrays (Cheng et al., 2005). Repeat-masked sequences of ten human chromosomes, including chromosome 20, were presented on high density arrays using 25-mer oligonucleotides spaced every 5 bp on average (i.e., 20 bp overlap), and the sites of transcription for poly A+ cytosolic RNA derived from eight cell lines were mapped. Approximately 9% of 74,180,611 total probe pairs detected per transcription per cell line and per chromosome, and the average number of transfrags (transcribed fragments) per cell line and per chromosome was found to be 16,864. Also, a considerable proportion of the detected transcription is cell-line specific. Strikingly, 31.8% of the detected cytosolic poly A+ sequences do not overlap with any well-characterised exon, mRNA or EST annotation

(UCSC Genome Browser, b34). Inter- and intra-genic regions that showed enrichment with several antibodies including polII and H3K4me3 were investigated whether they contain or lie within any of the transfrags obtained from Cheng et al. study. To this end, transfrags obtained from eight cell lines that are longer than 300 bp were taken (9,439) and the enrichment patterns of these fragments were investigated by all antibodies used. In total, 17 non-TSS containing enriched spots overlapped with a transfrag, and are listed in Table A.11a (HeLa S3) and Table A.11b (NTERA-D1) in Appendix A. Four and three polII enriched spots in HeLa S3 and NTERA-D1 overlapped with a transfrag. Also, seven and five CTCF-enriched spots in HeLa S3 and NTERA-D1 respectively overlapped with transfrags. It still remains unknown whether transfrags serve a function or are simply products of randomly assembled initiation machineries on sequences other than promoters. The fact that some enriched regions overlap with transfrags brings the possibility that certain protein assemblies or histone signatures might favour the assembly of polymerase II initiation machinery irrespective of the DNA sequence of the region, which will then produce such transcribed sequences. However this scenario cannot shed a light on functional role of transfrags.

This study produced a number of potential distal regulatory sequences, which need to be tested further in order to derive a regulatory map of the region. Inclusion of more cell lines to such studies will improve the coverage of such analysis since most probably many regions will produce signals only in the cells in which they function.

Transcript Type	HUGO Transcript ID	polII	H3K4me	H3K4me2	H3K4me3	H3Ac	H4Ac	H3K9me2	CTCF	Expression
RT	C20orf121-004	**7.44**	3.33	17.97	**36.72**	**14.44**	6.51			P
RT	NCOA3-001	**4.40**		6.66	**22.81**	**11.39**	3.39			P
RT	PPGB-004	**3.24**	2.94	14.29	**21.83**	**9.69**	2.53			P
RT	STK4-003	**3.68**	1.70	9.19	**20.95**	**12.13**	2.47			P
AT	UBE2C-003	**2.12**	1.97	10.55	**19.00**	**8.11**	0.00			P
RT	PIGT-003	**3.75**		7.28	**17.91**	**7.93**	3.70			P
RT	SERINC3-001	**1.95**		7.33	**17.51**	**5.66**	1.91			P
RT	UBE2C-001	**2.97**	1.97	4.21	**13.35**	**2.91**	0.00			P
RT	C20orf142-001	**0.00**		4.59	**8.96**	**3.30**	0.00			P
RT	ZSWIM1-002	**2.41**	1.89	3.04	**7.42**	**3.31**	1.74			P
AT	PPGB-001	**3.24**	2.70	4.59	**6.62**	**2.84**	2.55			P
RT	ACOT8-004	**0.00**	2.01	3.50	**6.28**	**2.19**	0.00			P
RT	ZSWIM3-001	**0.00**	2.01	3.50	**6.28**	**2.19**	0.00			P
RT	C20orf119-017	**0.00**		4.61	**6.22**	**2.90**	0.00			P
RT	SDC4-001	**5.28**	2.51	2.86	**4.46**	**2.07**	0.00			P
RT	PRKCBP1-007	**2.48**	3.10	7.18	**3.93**	**2.61**	3.28			P
RT	TOMM34-001	**0.00**	2.68	4.16	**3.73**	**2.14**	0.00			P
RT	TP53RK-001	**0.00**	2.05	4.37	**3.63**	**3.67**	0.00			P
RT	ELMO2-001	**0.00**	2.92	2.05	**2.63**	**1.59**	1.57			P
RT	C20orf67-001	**2.17**	0.00	2.16	**1.97**	**1.57**	0.00			P
RT	WISP2-002	**2.85**	4.45	2.67	**1.87**	**2.51**	2.05			P
AT	PRKCBP1-006		5.36	7.86	**1.82**	**3.41**	2.24			P
RT	YWHAB-001		2.07	2.05		**1.80**				P
RT	SLC35C2-006	**2.17**	1.81	1.66		**1.85**				P
RT	C20orf10-001									P
AT	C20orf119-004									P
AT	C20orf119-005									P
RT	DBNDD2-003									P
AT	DBNDD2-006									P
RT	CD40-001									P
AT	ELMO2-004									P
AT	ELMO2-006									P
AT	ELMO2-007									P
RT	NCOA5-001									P

RT	PI3-001		3.23	2.27						P
RT	PKIG-001									P
RT	PLTP-001		3.47				2.20			P
AT	PRKCBP1-009									P
RT	SPINLW1-001									P
RT	SPINLW1-002									P
AT	SERINC3-002									P
RT	WFDC2-002									P
RT	ZNF335-001	**3.15**	3.44	9.15	**22.69**	**12.91**	6.68		**4.70**	A
RT	NEURL2-001*	**3.24**	2.94	14.29	**21.83**	**9.69**	2.53			A
RT	SLC12A5-001				**2.61**	**2.17**				A
RT	C20orf62-001#				**1.73**					A
RT	EYA2-003		9.82	5.11		**1.59**	2.51			A
RT	C20orf165-001	**2.25**	1.91							A
AT	EYA2-004									A
RT	GDAP1L1-001							**1.13**	**1.92**	A
RT	HNF4A-001									A
AT	HNF4A-003		5.13							A
RT	KCNS1-001									A
RT	MATN4-001								**6.49**	A
RT	MMP9-001									A
RT	R3HDML-001									A
RT	RBPSUHL-001								**8.67**	A
RT	RIMS4-002							**1.15**		A
RT	SULF2-002									A
AT	SULF2-003								**5.15**	A
RT	SEMG1-001									A
RT	SEMG2-001		1.51							A
RT	SLC13A3-001									A
AT	SLC13A3-002									A
AT	SLC13A3-004							**1.03**		A
RT	SLC2A10-001									A
RT	SPINT3-001									A
RT	TNNC2-003		1.57							A
RT	WFDC10A-001								**4.57**	A

RT	WFDC10B-001									A
RT	WFDC11-001									A
RT	WFDC12-001									A
RT	WFDC13-001									A
RT	WFDC3-006		2.99	1.64						A
RT	WFDC5-001								**2.26**	A
RT	WFDC6-002									A
RT	WFDC8-001									A
RT	WFDC9-001								**2.14**	A
RT	ZNF334-001									A
RT	C20orf123-001				**2.40**					No probe
RT	C20orf157-001									No probe
RT	C20orf168-001									No probe
RT	SPINT4-001									No probe

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 *NEURL2* are attributed to *PPGB* 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.

Transcript Type	HUGO Transcript ID	polII	H3K4me	H3K4me2	H3K4me3	H3Ac	H4Ac	H3K27me3	CTCF	expression
RT	C20orf121-004	**5.15**	1.75	25.77	**79.56**	**7.10**				P
RT	NCOA3-001	**4.75**		16.28	**54.36**	**5.18**				P
RT	STK4-003			18.35	**51.90**	**4.65**				P
AT	UBE2C-003	**3.81**		1.70	**43.81**	**4.32**				P
RT	PPGB-004	**3.34**		21.91	**42.93**	**4.79**				P
RT	SERINC3-001			12.24	**39.36**	**3.89**				P
RT	ZNF335-001	**2.74**		11.50	**31.97**	**3.43**			**3.86**	P
AT	PPGB-001	**3.34**		15.98	**30.33**	**3.63**				P
RT	UBE2C-001	**3.62**		6.47	**29.56**	**1.81**				P
RT	C20orf119-017			5.26	**26.42**	**2.76**				P
RT	GDAP1L1-001			14.25	**26.25**	**3.40**			**3.30**	P
RT	ZSWIM1-002	**2.62**		4.53	**21.51**	**1.66**				P
RT	CD40-001	**2.18**		12.59	**19.97**	**1.74**		1.93		P
RT	PIGT-003	**1.98**		7.57	**19.60**	**2.20**				P
RT	C20orf142-001			7.68	**18.47**	**2.66**				P
RT	ZNF334-001			7.49	**16.58**	**1.58**				P
AT	PRKCBP1-006	**2.21**	1.67	24.62	**12.53**	**3.84**				P
RT	ACOT8-004			4.02	**10.05**					P
RT	ZSWIM3-001			4.02	**10.05**					P
RT	PRKCBP1-007	**1.75**	2.50	7.75	**9.39**	**2.37**	3.46			P
RT	PLTP-001	**3.49**			**7.49**	**1.79**				P
RT	WFDC2-002			7.40	**7.29**	**1.51**				P
RT	ELMO2-001			1.62	**6.02**					P
RT	TOMM34-001			5.47	**5.70**			1.61		P
RT	SDC4-001	**2.03**		4.21	**5.65**	**1.84**				P
RT	C20orf67-001	**1.90**		3.93	**5.61**	**1.55**				P
RT	SLC13A3-001			7.60	**5.48**					P
RT	SULF2-002			9.40	**1.91**	**1.91**				P
RT	TP53RK-001			1.67	**1.51**					P
RT	DBNDD2-003							1.60		P
RT	NCOA5-001							1.55		P
AT	PRKCBP1-009		2.72	3.50		**1.87**				P
RT	SEMG1-001			2.95						P

RT	YWHAB-001			2.82						P
RT	MMP9-001			1.59						P
RT	SLC35C2-006	**2.19**								P
RT	C20orf10-001									P
AT	C20orf119-004									P
AT	C20orf119-005									P
AT	DBNDD2-006									P
AT	ELMO2-004									P
AT	ELMO2-006									P
AT	ELMO2-007									P
RT	MATN4-001								**4.11**	P
RT	PKIG-001									P
RT	RBPSUHL-001								**5.46**	P
AT	SULF2-003								**3.79**	P
AT	SLC13A3-002									P
AT	SLC13A3-004									P
RT	SLC2A10-001									P
AT	SERINC3-002									P
RT	NEURL2-001*	**3.34**		21.91	**42.93**	**4.79**				A
RT	C20orf165-001	**1.89**		5.84	**9.51**				**2.21**	A
RT	KCNS1-001			7.24	**8.33**					A
RT	WFDC10A-001			7.99	**6.89**				**4.50**	A
RT	WFDC9-001									A
RT	C20orf62-001$^{\#}$				**3.80**					A
RT	SLC12A5-001			8.55	**2.54**			9.52		A
RT	RIMS4-002							3.86		A
RT	EYA2-003							2.04		A
AT	EYA2-004			1.61						A
RT	TNNC2-003		2.40							A
RT	HNF4A-001									A
AT	HNF4A-003									A
RT	PI3-001									A
RT	R3HDML-001									A
RT	SEMG2-001									A
RT	SPINLW1-001									A

RT	SPINLW1-002									A
RT	SPINT3-001									A
RT	WFDC10B-001									A
RT	WFDC11-001									A
RT	WFDC12-001									A
RT	WFDC13-001									A
RT	WFDC3-006									A
RT	WFDC5-001									A
RT	WFDC6-002									A
RT	WFDC8-001									A
RT	WISP2-002									A
RT	C20orf123-001			1.56	**3.91**					No probe
RT	C20orf157-001									No probe
RT	C20orf168-001									No probe
RT	SPINT4-001									No probe

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 *NEURL2* are attributed to *PPGB* 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.

Cell Type	HUGO Gene ID	polII	H3K4me	H3K4me2	H3K4me3	H3Ac	H4Ac	H3K9me2	H3K27me2	expression
HeLa S3	SLC12A5				**2.61**	**2.17**				A
	RIMS4							**1.15**		A
	EYA2		9.82	5.11		**1.59**	2.51			A
	CD40									P
	TOMM34		2.68	4.16	**3.73**	**2.14**				P
	DBNDD2									P
	NCOA5									P
Cell Type	**HUGO Gene ID**	**polII**	**H3K4me**	**H3K4me2**	**H3K4me3**	**H3Ac**	**H4Ac**	**H3K9me2**	**H3K27me2**	**expression**
NTERA-D1	SLC12A5			8.55	**2.54**				9.52	A
	RIMS4								3.86	A
	EYA2								2.04	A
	CD40	**2.18**		12.59	**19.97**	**1.74**			1.93	P
	TOMM34			5.47	**5.70**				1.61	P
	DBNDD2								1.60	P
	NCOA5								1.55	P

Table 5.17 Enrichment profiles of nine genes whose start sites were enriched with H3K27me3 in NTERA-D1 cells in both cell lines.

6 Discussion and Future Work

6.1 Discussion

While a significant number of SNPs in putative promoters are already available as a matter of course from the genome project and SNP ascertainment projects (Sachidanandam et al. 2001; Consortium 2005b; Hinds et al. 2005), there have been almost no efforts of any scale to specifically mine promoter sequences for polymorphisms. Buckland et al were the first group to re-sequence promoters across many genes, but their panel was small, ethnically heterogeneous and gave limited information about allele frequencies, as well as suffering from significant ascertainment bias as reported by the authors themselves (Buckland et al. 2005). This project has carried out the deepest available re-sequencing of promoters currently available, with considerably more power to detect rare polymorphisms than the Buckland project despite there still being some ascertainment bias away from rare SNPs. In a surprising result, essentially no difference was found between overall mutation rates in promoters and in chromosome 22 overall apart from those explainable by elevated GC content. This is despite the naïve assumption that the promoters would have suppressed C-T mutation rates compared to the rest of the chromosome. Some reasons why this might be the case have been outlined in section 4.3. However, an interesting avenue for further investigation would be to look at the history of C-T mutations in order to see whether the rate in the genome as a whole has slowed over time. This could be done by using a measure such as extended haplotype heterozygosity to estimate an age profile for C/T SNPs versus other SNPs, to see whether C/T SNPs are generally older (although this would depend on whether such a slowdown had happened within human evolutionary timescales).

Rockman and Wray have previously estimated a rate of 0.94 functional SNPs per kb in the 850 base pair sequences upstream of TSSs (Rockman and Wray 2002). This was likely to be an underestimate, as the majority of functional variants in the promoters studied have probably not been identified. The chromosome 22 project identified between 0.73 and 0.98 functional SNPs per kb, depending on the number of unconfirmed SNPs that are taken as being real. This is from an average of 630 base pairs upstream. These numbers are in remarkable agreement considering the very different methods used to obtain them, and suggest that the significantly greater degree of functional variation observed here compared to the Buckland set should not be considered surprising.

What is still unclear is how much of this promoter variation that is detectable by isolating the promoter remains significant when all the other regulatory inputs found in a native genome are added? This work has not been carried out on a consistent set of promoter polymorphisms such as that produced here. However, literature surveys suggest that a significant proportion of SNPs with functional effects in reporter assays also have further evidence of function either on a biochemical or disease level phenotype. Indeed, for a set of 107 genes with published functional promoter polymorphism, 59% and 71% respectively also had published evidence of such phenotypes (Rockman and Wray 2002). These figures may be affected by publication bias as a result of underreporting of negative results, and this is probably not possible to quantify, but nevertheless the link between reporter assays and an *in vivo* function does exist and can be amply demonstrated with current methods, many of which are now being developed to a high-throughput capability (Knight et al. 2003; Linnell et al. 2004). There is also considerable evidence of extensive allele-specific variation in gene expression (Yan et al. 2002b; Pastinen et al. 2005) as well as association between *cis*-acting loci and gene expression levels (Monks et al. 2004; Cheung et al. 2005; Stranger et al. 2005) that suggest the presence of a lot of *cis*-regulatory variation in the genome. Essentially all these studies have been carried out on subsets of the same CEPH families from which the panel for this project was drawn. Even though this does not say anything about the *in vivo* functionality of the particular functional SNPs discovered, it does demonstrate that there is ample potential for them to have phenotypic consequences at least on expression phenotypes in the 48-person CEPH panel, if not at the level of disease and/or organismal phenotype. While no evidence was found for an association of any of the *in vitro* functional SNPs with expression phenotypes in the HapMap individuals in the panel, this may well have been due to the low power afforded from an overlap of only 31 individuals. The lack of power would be exacerbated by the failure to obtain genotypes from the re-sequencing for a subset of individuals in each SNP. This would lead to an even smaller number of informative individuals for whom functional data was available, and was not an uncommon occurrence. The net result was to make it relatively unlikely for any association to survive the correction for multiple testing.

A crucial result of this project was the lack of enrichment of functional SNPs in putative regulatory elements including TFBS and ultraconserved regions. This is surprising given that the traditional model for the action of functional promoter SNPs has been the perturbation of TFBS. Buckland et al reported that only 35% of the functional SNPs were in a TFBS (Buckland et al. 2005). However, the absolute numbers of putative TFBS present in a promoter, as determined by any of a number of possible tools and databases is largely a function of the parameters used for the search and the quality of the position weight matrices in the database. It may therefore be more meaningful to compare the rates of functional and non-functional SNPs in TFBS using consistent parameters and express this as an enrichment factor. To my knowledge, this is the first project to explore the enrichment of putative TFBS for functional SNPs, although others have used TFBS as a criterion to predict functional SNPs (Mottagui-Tabar et al. 2005). The lack of enrichment suggests that current models of TFBS are inadequate and not useful for predicting whether promoter SNPs are likely to be functional. This is despite ample evidence that some regulatory SNPs do function by altering the affinity of a TFBS, as evidenced by EMSA experiments using allelic probes and transient transfection assays in parallel (Rockman and Wray 2002). However, it is often the case in the literature that one set of experiments is done without the other, making it difficult to assess how much known functional variation can be accounted for in this manner. Limited evidence from a small number of experiments has suggested that between 70 and 80% of SNPs in TFBS within conserved regions can alter the binding of a TF *in vitro* according to EMSA experiments (Belanger et al. 2005; Mottagui-Tabar et al. 2005). Even if these results were representative, it is still the case that not all SNPs in binding sites cause functional differences, and indeed it may be that only a minority of sites do so (Rockman and Wray 2002). The lack of functionality of SNPs in some binding sites (even ones experimentally verified by EMSA), as well as the number of functional promoter SNPs apparently not within any known binding sites points to one or both of two possibilities; that there is a significant number of binding sites still to be discovered or that these SNPs are exerting their effects by a mechanism other than direct perturbation of a binding site.

Several analyses of human promoters using various methods, often heavily reliant on evolutionary conservation, have found conserved motifs that are enriched at promoters (Xie et al. 2005; Robertson et al. 2006). This enrichment, and the fact that many known motifs have been re-discovered with these methods, suggests that they may indeed be functional, although the resulting elements have yet to be functionally tested (for example by deletion analysis in reporter constructs). It is therefore not unlikely that our knowledge of the number of regulatory elements is far from complete, although it has been proposed that many of the remaining motifs may be rare and/or only functional in restricted biological conditions (Buckland 2006).

There is also evidence that non-binding site-dependent mechanisms may be important in explaining promoter SNP effects. These SNPs may function by altering the conformational properties of the DNA upstream of the TSS, and thus altering the dynamics of TF interactions with each other and the promoter without necessarily being in a binding site (Buckland 2006). The inherent curvature of DNA is often higher at promoters, and this has been shown to be an important factor in the activation of at least some eukaryotic genes (Nishikawa et al. 2003). Manipulations of cloned promoters in reporter vectors have shown that promoters with higher inherent curvature can promote transcription markedly more efficiently than the same promoter carrying mutations that reduced this curvature (Kim, Klooster, and Shapiro 1995). The addition of intercalators that abrogated this curvature greatly reduced this activity difference (Kim, Klooster, and Shapiro 1995). While structural studies show that some TFs, including TBP and p53 (Nagaich, Appella, and Harrington 1997; de Souza and Ornstein 1998), alter the conformation of DNA on binding, it is also the case that DNA which is already in a favourable conformation pre-binding can drastically increase binding affinity (Parvin et al. 1995). Alteration of TF binding efficiency by the introduction of artificial substitutions outside the TFBS that alter conformation has been demonstrated in yeast (Acton, Zhong, and Vershon 1997), although the presence or extent of natural SNPs that function in this way is unknown. A distinct but related property of the DNA itself that can be important in TF binding is the flexibility, or the ability of DNA conformation to be altered by the binding of proteins. This can be important in allowing multiple protein-DNA interactions in close proximity by relieving steric hindrances (either by one factor binding multiple sites or by multiple factors) or by allowing the DNA to loop and bring distant bound factors into contact (Mastrangelo et al. 1991; Suzuki and Yagi 1995; Nagaich, Appella, and Harrington 1997).

The results produced in this project and other evidence presented above have important implications for efforts to predict functional polymorphisms by using models of TFBSs. While several such attempts have been made, usually claiming at least moderate success, they are often tested using an inadequately small number of actual functional experiments (Belanger et al. 2005; Mottagui-Tabar et al. 2005). This makes their success hard to quantify, although the fact that even small scale predictions were not confirmed more than 50% of the time suggests there is still some way to go before such predictive methods become reliable. There is some evidence that even using position weight matrices rather than simple consensus sequences may not enable the true deduction of the effect of a base change on a binding site, and that more complete experimental characterization of TFBS may be necessary for this (Bulyk, Johnson, and Church 2002). The presence of an unbiased potential training set of functional polymorphisms may be very important in developing new *in silico* methods for regulatory polymorphism discovery. *In silico* analysis of the effect of the functional SNPs discovered here and by Buckland et al on the DNA conformation may shed more light on the putative importance of this mechanism. Collaboration with other groups to analyse the performance of some of the novel motifs discovered by comparative genomics (Xie et al. 2005) may also shed more light on the utility of conservation for predicting functional variation.

This project has also explored the qualitative relationship between promoter activity and *in vivo* expression. This has confirmed that promoter sequences contain many of the elements that determine whether a gene is expressed or not, and therefore that the promoter really does integrate the majority of signals in the transcription initiation pathway. Other work has found a more quantitative relationship between promoter activity and gene expression (Cooper et al. 2006), but this was not reproduced here. As suggested before, this may be due to the relative quantitative potential of RT-PCR (as used by Cooper et al) and Affymetrix arrays. Another factor may be the difference in the controls used for the luciferase assays, where a single promoterless plasmid was used in this project versus the average of 102 cloned non-functional DNA elements by Cooper et al. This latter control may form a more consistent baseline as any non-specific activation of transcription due to stochastic biological variation in different cell growths would perturb the baseline by relatively low levels. Indeed, Promega have recently released the pGL4 luciferase plasmid series, where a large number of cryptic TFBS were removed from the vector backbone relative to the pGL3 plasmids. These may have been a source of variation in background levels.

The finding that upregulatory mutations are skewed towards higher derived allele frequencies relative to downregulatory mutations may have implications for the evolutionary mechanisms of gene regulation. The expansion of derived alleles of functional SNPs has been observed previously, with 7 out of 21 known functional SNPs having derived major alleles, and 11 of the remainder having either allele as the major allele in different populations (Rockman and Wray 2002). However, greater tendency for upregulatory changes in promoters to expand relative to downregulatory changes is a novel finding, and suggests that upregulatory promoter changes may be more amenable to positive selection than downregulatory ones, and may therefore be more likely to have positive fitness consequences. If this were the case, it may be important to understanding the mechanistic basis of transcriptome evolution. The known phylogeny between primates is recapitulated by expression variation between species (Gilad et al. 2006), and levels of selective constraint on gene expression levels and coding sequence coincide (Khaitovich et al. 2005). Interestingly, despite more constraint on interspecific gene expression variation in the brain in primates (Khaitovich et al. 2005), there has been an acceleration in gene expression changes in the human lineage (Enard et al. 2002), and this difference is made up largely of upregulations rather than downregulations (Caceres et al. 2003). Upregulations in gene expression in the human lineage have also occurred in human versus chimpanzee TF genes (Gilad et al. 2006) and in fibroblasts (Karaman et al. 2003), although the bias in favour of upregulations is much less clear in the latter case.

The bias towards expansion of upregulatory changes seems at odds with some theoretical models of transcriptome evolution, which propose that downregulatory changes should be more common that upregulatory ones (Khaitovich, Paabo, and Weiss 2005). It also does not agree with recent findings by the Dermitzakis lab at the Sanger Institute that SNPs found by whole genome association to expression phenotypes agree with this model (Stranger et al unpublished). However, it is important to note that while Stranger et al were measuring mRNA levels, this project was measuring *in vitro* promoter activity, with the latter being a component of the former. A possible explanation for the discrepancy is that these association studies may be finding regulatory SNPs in distant enhancer or silencer elements rather than the promoter, and that such functional SNPs may have more powerful effects than those at promoters. This is suggested by the fact that the majority of SNPs identified by Stranger et al are more than 10 kb away from the TSS of the genes they influence (data not shown). The effects of these elements on transcription may be sufficiently powerful that where they contain functional variation, this dominates over promoter sequence variation, and precludes it from identification in association studies. This may also explain discrepancies in the difference between human and chimpanzee promoter activities and the corresponding difference in transcript levels (Heissig et al. 2005). Heissig and colleagues found seven genes that showed significant differences between chimpanzees and humans both in luciferase reporter assays and measures of transcript abundance. However, in 4/7 genes these differences were in the opposite direction to each other (Heissig et al. 2005). It may therefore be proposed that globally, variation in proximal promoters and in distal regulatory elements are influenced differently by selection.

6.2 Future work

Following on from the generation of a set of functional promoter polymorphisms *in vitro*, the next natural step is to investigate the effects of these SNPs *in vivo* in order to determine whether they are still functional in their native genomic contexts. There are several experimental methods for doing this, all of which give subtly different levels of information on the SNPs under investigation.

The most obvious method would be to look for differences in the mRNA transcripts produced by variant promoters. The most well-established method for doing this is probably quantitative RT-PCR from cell lines or mRNA from heterozygotes (Yan et al. 2002b; Bray et al. 2003; Pastinen and Hudson 2004). This would require the identification of individuals who were heterozygous both for the promoter haplotypes of interest and for a transcribed marker SNP that could be used to distinguish the two transcripts. It would also necessitate knowledge of the phase of the promoter haplotypes and the maker SNP in order to be able to say which promoter haplotype is driving the expression of which transcript, enabling the assignment of direction to functional changes. With the HapMap project now having completed phase 1, there is a ready source of cell lines from a range of individuals that can be used for this kind of work (Consortium 2005b). The genotype information would also enable the inference of phase between transcribed markers and the promoter SNPs (Stephens, Smith, and Donnelly 2001).

The advent of chromatin immunoprecipitation combined with a quantitative genotyping method also allows direct assay of differential RNA polymerase II loading on polymorphic promoters in a heterozygote, a technique dubbed the haploChIP method (Knight et al. 2003). This would involve chromatin IP with an antibody to RNA Pol II phosphorylated at serine 5, which is enriched at the 5' end of transcripts. This would be followed by quantitative assessment of fragments from the two promoter alleles by primer-extension and mass spectrometry analysis (Knight et al. 2003). This method has the advantage of not requiring a transcribed marker SNP,as well as the ability to yield information on multiple heterozygous promoters in a single chromatin immunoprecipitation sample, hence making it suitable for high throughput applications.

If a complete set of *in vitro* and *in vivo* data for a set of promoter SNPs could be produced, it would then be desirable to explain the mechanistic basis of any functional differences, either in terms of TF binding or mechanisms related to structural conformation of DNA. Again, an established method for this already exists; electrophoretic mobility shift assays (EMSA). To apply it to SNPs, radioactively labelled oligonucleotide probes would be synthesised containing the putative binding site, with one probe per allele per polymorphism. The allelic probes would then be allowed to bind proteins from cellular extracts and run down an agarose gel to look for a band shift indicating binding. Relative binding abilities would be assessed by using a non-specific competitor oligonucleotide. This currently remains a low throughput process, and would probably be a bottleneck in any large scale pipeline. One advantage however is that it introduces the possibility of identifying unknown TFs binding to SNPs that are not in known sites by mass fingerprinting. A more high-throughput possibility would be to use a haploChIP-style method, but assessing binding of TFs rather than Pol II. However, this will be limited only to the relatively few TFs for which antibodies are available.

Another area of interest would be to study the population history of functional polymorphisms and examine the relative importance of regulatory variation and coding variation. It is now relatively easy to design genotying assays for a known polymorphism, and the facilities available at the Sanger Institute would enable rapid and thorough genotyping of several hundred putative regulatory SNPs across the entire HapMap population panel. This would enable studies both within and across continental populations, and could make possible the use of robust statistical methods for inferring selection. Importantly, full genotyping in the HapMap individuals of a large panel of functional SNPs would make it easy to repeat the association studies with the whole genome expression data and obtain far more robust associations (and where an association couldn't be shown, this would again be a more convincing negative result).

Finally, current knowledge of functional promoter polymorphisms can be used to build a database of polymorphisms for which function is known *a priori*, and use this for meta-analysis to examine the properties of functional SNPs more thoroughly. This database can then be put through the above battery of methods in order to complete the knowledge required for each of the polymorphisms. Two sets of promoter polymorphisms tested *in vitro* under homogeneous experimental conditions are already available; the data presented in the thesis and that produced by Buckland et al. Together, these consist of 79 isolated and confirmed promoter polymorphisms. There have also been two efforts to curate information from the wider literature, which contains data on many more promoter variants distributed among a large number of papers. Rockman and Wray produced a survey of 140 functional SNPs tested in reporter assays in the literature, and in many cases were able to find supporting published evidence in the form of EMSA experiments or associations with expression or disease phenotypes. In addition, the ORegAnno database of regulatory elements (Montgomery et al. 2006) contains a set of 172 promoter polymorphisms that have been partly manually curated and partly submitted by external contributors. In both these curated datasets, the SNPs are not always clearly-mapped to the genome and the evidence supporting each SNP is very heterogeneous (in some cases, for example, there is differential binding data from EMSA but no luciferase assay). These would then be put through the methods proposed to complete the evidence for them, with the expectation that there would be a high rate of functional confirmation. Indeed, I was able to construct a preliminary database that would hold the integrated results of such a meta-analysis of published functional promoter SNPs, and was able to populate it with data from both the Buckland set and from individual papers. This work was not presented in this thesis, as more work is needed to establish an ontology for populating it with a dataset that can be consistently analysed.

Eventually, these methods would lead to a set of promoter polymorphisms where data was available for every potential step in the process of explaining their mechanistic basis; *in vitro* function in isolation from confounding regulatory inputs, effect on TF binding, carry-over of the *in vitro* effect *in vivo* and population data to study the selection history of the polymorphism. Such a dataset has never been accumulated before, and could be the turning point for efforts to understand the mechanisms of promoter variation effects. It would be an excellent training set for computational methods that could then be used to predict the effect of promoter SNPs. If these methods could be perfected on the strength of such a training dataset, it would have potential implications for human health, allowing better assessments for non-coding pathogenic variants whose function could not be predicted in the same way as deleterious coding functions.

REFERENCES

Abramowitz, M. a. S., I. A. (Eds.) (1972). Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables. New York, Dover.

Ajiro, K. (2000). "Histone H2B phosphorylation in mammalian apoptotic cells. An association with DNA fragmentation." J Biol Chem **275**(1): 439-43.

Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter (2001). Molecular Biology of the Cell, Taylor & Francis.

Allfrey, V. G., R. Faulkner and A. E. Mirsky (1964). "Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis." Proc Natl Acad Sci U S A **51**: 786-94.

Anderson, G. M. and S. O. Freytag (1991). "Synergistic activation of a human promoter in vivo by transcription factor Sp1." Mol Cell Biol **11**(4): 1935-43.

Andres, M. E., C. Burger, M. J. Peral-Rubio, E. Battaglioli, M. E. Anderson, J. Grimes, J. Dallman, N. Ballas and G. Mandel (1999). "CoREST: a functional corepressor required for regulation of neural-specific gene expression." Proc Natl Acad Sci U S A **96**(17): 9873-8.

Angel, P. and M. Karin (1991). "The role of Jun, Fos and the AP-1 complex in cell-proliferation and transformation." Biochim Biophys Acta **1072**(2-3): 129-57.

Aravind, L., V. Anantharaman, S. Balaji, M. M. Babu and L. M. Iyer (2005). "The many faces of the helix-turn-helix domain: transcription regulation and beyond." FEMS Microbiol Rev **29**(2): 231-62.

Aronow, B., D. Lattier, R. Silbiger, M. Dusing, J. Hutton, G. Jones, J. Stock, J. McNeish, S. Potter, D. Witte and et al. (1989). "Evidence for a complex regulatory array in the first intron of the human adenosine deaminase gene." Genes Dev **3**(9): 1384-400.

Ashurst, J. L., C. K. Chen, J. G. Gilbert, K. Jekosch, S. Keenan, P. Meidl, S. M. Searle, J. Stalker, R. Storey, S. Trevanion, L. Wilming and T. Hubbard (2005). "The Vertebrate Genome Annotation (Vega) database." Nucleic Acids Res **33**(Database issue): D459-65.

Ashurst, J. L. and J. E. Collins (2003). "Gene annotation: prediction and testing." Annu Rev Genomics Hum Genet **4**: 69-88.

Atchison, M. L. (1988). "Enhancers: mechanisms of action and cell specificity." Annu Rev Cell Biol **4**: 127-53.

Atchley, W. R. and W. M. Fitch (1997). "A natural classification of the basic helix-loop-helix class of transcription factors." Proc Natl Acad Sci U S A **94**(10): 5172-6.

Austen, M., B. Luscher and J. M. Luscher-Firzlaff (1997). "Characterization of the transcriptional regulator YY1. The bipartite transactivation domain is independent of interaction with the TATA box-binding protein, transcription factor IIB, TAFII55, or cAMP-responsive element-binding protein (CPB)-binding protein." J Biol Chem **272**(3): 1709-17.

Ayer, D. E., Q. A. Lawrence and R. N. Eisenman (1995). "Mad-Max transcriptional repression is mediated by ternary complex formation with mammalian homologs of yeast repressor Sin3." Cell **80**(5): 767-76.

Bagwell, A. M., A. Bailly, J. C. Mychaleckyj, B. I. Freedman and D. W. Bowden (2004). "Comparative genomic analysis of the HNF-4alpha transcription factor gene." Mol Genet Metab **81**(2): 112-21.

Bagwell, A. M., J. L. Bento, J. C. Mychaleckyj, B. I. Freedman, C. D. Langefeld and D. W. Bowden (2005). "Genetic analysis of HNF4A polymorphisms in Caucasian-American type 2 diabetes." Diabetes **54**(4): 1185-90.

Bajic, V. B. and S. H. Seah (2003). "Dragon Gene Start Finder identifies approximate locations of the 5' ends of genes." Nucleic Acids Res **31**(13): 3560-3.

Bajic, V. B. and S. H. Seah (2003). "Dragon gene start finder: an advanced system for finding approximate locations of the start of gene transcriptional units." Genome Res **13**(8): 1923-9.

Bajic, V. B., S. H. Seah, A. Chong, S. P. Krishnan, J. L. Koh and V. Brusic (2003). "Computer model for recognition of functional transcription start sites in RNA polymerase II promoters of vertebrates." J Mol Graph Model **21**(5): 323-32.

Bajic, V. B., S. L. Tan, Y. Suzuki and S. Sugano (2004). "Promoter prediction analysis on the whole human genome." Nat Biotech **22**(11): 1467-1473.

Banerji, J., S. Rusconi and W. Schaffner (1981). "Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences." Cell **27**(2 Pt 1): 299-308.

Bardwell, V. J. and R. Treisman (1994). "The POZ domain: a conserved protein-protein interaction motif." Genes Dev **8**(14): 1664-77.

Baylin, S. B., J. G. Herman, J. R. Graff, P. M. Vertino and J. P. Issa (1998). "Alterations in DNA methylation: a fundamental aspect of neoplasia." Adv Cancer Res **72**: 141-96.

Bazett-Jones, D. P., B. Leblanc, M. Herfort and T. Moss (1994). "Short-range DNA looping by the Xenopus HMG-box transcription factor, xUBF." Science **264**(5162): 1134-7.

Beausoleil, S. A., M. Jedrychowski, D. Schwartz, J. E. Elias, J. Villen, J. Li, M. A. Cohn, L. C. Cantley and S. P. Gygi (2004). "Large-scale characterization of HeLa cell nuclear phosphoproteins." Proc Natl Acad Sci U S A **101**(33): 12130-5.

Becker, P. B. and W. Horz (2002). "ATP-dependent nucleosome remodeling." Annual Review of Biochemistry **71**: 247-73.

Behrens, A., K. Sabapathy, I. Graef, M. Cleary, G. R. Crabtree and E. F. Wagner (2001). "Jun N-terminal kinase 2 modulates thymocyte apoptosis and T cell activation through c-Jun and nuclear factor of activated T cell (NF-AT)." Proc Natl Acad Sci U S A **98**(4): 1769-74.

Behrens, A., M. Sibilia and E. F. Wagner (1999). "Amino-terminal phosphorylation of c-Jun regulates stress-induced apoptosis and cellular proliferation." Nat Genet **21**(3): 326-9.

Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick and D. Haussler (2004). "Ultraconserved elements in the human genome." Science **304**(5675): 1321-5.

Bell, A. C., A. G. West and G. Felsenfeld (1999). "The protein CTCF is required for the enhancer blocking activity of vertebrate insulators." Cell **98**(3): 387-96.

Bell, A. C., A. G. West and G. Felsenfeld (2001). "Insulators and boundaries: versatile regulatory elements in the eukaryotic." Science **291**(5503): 447-50.

Belz, T., A. D. Pham, C. Beisel, N. Anders, J. Bogin, S. Kwozynski and F. Sauer (2002). "In vitro assays to study protein ubiquitination in transcription." Methods **26**(3): 233-44.

Bench, A. J., E. P. Nacheva, T. L. Hood, J. L. Holden, L. French, S. Swanton, K. M. Champion, J. Li, P. Whittaker, G. Stavrides, A. R. Hunt, B. J. Huntly, L. J. Campbell, D. R. Bentley, P. Deloukas and A. R. Green (2000). "Chromosome 20 deletions in myeloid malignancies: reduction of the common deleted region, generation of a PAC/BAC contig and identification of candidate genes. UK Cancer Cytogenetics Group (UKCCG)." Oncogene **19**(34): 3902-13.

Bender, M. A., M. Bulger, J. Close and M. Groudine (2000). "Beta-globin gene switching and DNase I sensitivity of the endogenous beta-globin locus in mice do not require the locus control region." Mol Cell **5**(2): 387-93.

Berkvens, T. M., E. J. Gerritsen, M. Oldenburg, C. Breukel, J. T. Wijnen, H. van Ormondt, J. M. Vossen, A. J. van der Eb and P. Meera Khan (1987). "Severe combined immune deficiency due to a homozygous 3.2-kb deletion spanning the promoter and first exon of the adenosine deaminase gene." Nucleic Acids Res **15**(22): 9365-78.

Bernstein, B. E., M. Kamal, K. Lindblad-Toh, S. Bekiranov, D. K. Bailey, D. J. Huebert, S. McMahon, E. K. Karlsson, E. J. Kulbokas, 3rd, T. R. Gingeras, S. L. Schreiber and E. S. Lander (2005). "Genomic maps and comparative analysis of histone modifications in human and mouse." Cell **120**(2): 169-81.

Bertone, P., V. Stolc, T. E. Royce, J. S. Rozowsky, A. E. Urban, X. Zhu, J. L. Rinn, W. Tongprasit, M. Samanta, S. Weissman, M. Gerstein and M. Snyder (2004). "Global identification of human transcribed sequences with genome tiling arrays." Science **306**(5705): 2242-6.

Bird, A. (2002). "DNA methylation patterns and epigenetic memory." Genes Dev **16**(1): 6-21.

Bishop, C. E., D. J. Whitworth, Y. Qin, A. I. Agoulnik, I. U. Agoulnik, W. R. Harrison, R. R. Behringer and P. A. Overbeek (2000). "A transgenic insertion upstream of sox9 is associated with dominant XX sex reversal in the mouse." Nat Genet **26**(4): 490-4.

Blackwood, E. M. and J. T. Kadonaga (1998). "Going the distance: a current view of enhancer action." Science **281**(5373): 61-3.

Brandeis, M., D. Frank, I. Keshet, Z. Siegfried, M. Mendelsohn, A. Nemes, V. Temper, A. Razin and H. Cedar (1994). "Sp1 elements protect a CpG island from de novo methylation." Nature **371**(6496): 435-8.

Brehm, A., E. A. Miska, D. J. McCance, J. L. Reid, A. J. Bannister and T. Kouzarides (1998). "Retinoblastoma protein recruits histone deacetylase to repress transcription." Nature **391**(6667): 597-601.

Brukner, I., V. Jurukovski and A. Savic (1990). "Sequence-dependent structural variations of DNA revealed by DNase I." Nucleic Acids Res **18**(4): 891-4.

Brukner, I., R. Sanchez, D. Suck and S. Pongor (1995). "Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides." Embo J **14**(8): 1812-8.

Burke, L. J. and A. Baniahmad (2000). "Co-repressors 2000." Faseb J **14**(13): 1876-88.

Cai, H. N. and P. Shen (2001). "Effects of cis arrangement of chromatin insulators on enhancer-blocking activity." Science **291**(5503): 493-5.

Cao, R., L. Wang, H. Wang, L. Xia, H. Erdjument-Bromage, P. Tempst, R. S. Jones and Y. Zhang (2002). "Role of histone H3 lysine 27 methylation in Polycomb-group silencing." Science **298**(5595): 1039-43.

Cardinaux, J. R., J. C. Notis, Q. Zhang, N. Vo, J. C. Craig, D. M. Fass, R. G. Brennan and R. H. Goodman (2000). "Recruitment of CREB binding protein is sufficient for CREB-mediated gene activation." Mol Cell Biol **20**(5): 1546-52.

Carey, M. (1998). "The enhanceosome and transcriptional synergy." Cell **92**(1): 5-8.

Carninci, P., A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. Semple, M. S. Taylor, P. G. Engstrom, M. C. Frith, A. R. Forrest, W. B. Alkema, S. L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawaji, C. Kai, M. Nakamura, H. Konno, K. Nakano, S. Mottagui-Tabar, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S. M. Grimmond, C. A. Wells, V. Orlando, C. Wahlestedt, E. T. Liu, M. Harbers, J. Kawai, V. B. Bajic, D. A. Hume and Y. Hayashizaki (2006). "Genome-wide analysis of mammalian promoter architecture and evolution." Nat Genet **38**(6): 626-635.

Carstens, R. P., E. J. Wagner and M. A. Garcia-Blanco (2000). "An intronic splicing silencer causes skipping of the IIIb exon of fibroblast growth factor receptor 2 through involvement of polypyrimidine tract binding protein." Mol Cell Biol **20**(19): 7388-400.

Carter, D., L. Chakalova, C. S. Osborne, Y. F. Dai and P. Fraser (2002). "Long-range chromatin regulatory interactions in vivo." Nat Genet **32**(4): 623-6.

Cavalier-Smith, T. (1978). "Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox." J Cell Sci **34**: 247-78.

Cavin Perier, R., T. Junier and P. Bucher (1998). "The Eukaryotic Promoter Database EPD." Nucleic Acids Res **26**(1): 353-7.

Cha, T. L., B. P. Zhou, W. Xia, Y. Wu, C. C. Yang, C. T. Chen, B. Ping, A. P. Otte and M. C. Hung (2005). "Akt-mediated phosphorylation of EZH2 suppresses methylation of lysine 27 in histone H3." Science **310**(5746): 306-10.

Chen, B. S. and M. Hampsey (2002). "Transcription activation: unveiling the essential nature of TFIID." Curr Biol **12**(18): R620-2.

Chen, C. and T. P. Yang (2001). "Nucleosomes are translationally positioned on the active allele and rotationally positioned on the inactive allele of the HPRT promoter." Mol Cell Biol **21**(22): 7682-95.

Chen, H., R. J. Lin, R. L. Schiltz, D. Chakravarti, A. Nash, L. Nagy, M. L. Privalsky, Y. Nakatani and R. M. Evans (1997). "Nuclear receptor coactivator ACTR is a novel histone acetyltransferase and forms a multimeric activation complex with P/CAF and CBP/p300." Cell **90**(3): 569-80.

Cheng, J., P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, D. Stern, H. Tammana, G. Helt, V. Sementchenko, A. Piccolboni, S. Bekiranov, D. K. Bailey, M. Ganesh, S. Ghosh, I. Bell, D. S. Gerhard and T. R. Gingeras (2005). "Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution." Science **308**(5725): 1149-54.

Cheong, J. H., M. Yi, Y. Lin and S. Murakami (1995). "Human RPB5, a subunit shared by eukaryotic nuclear RNA polymerases, binds human hepatitis B virus X protein and may play a role in X transactivation." Embo J **14**(1): 143-50.

Cheung, P., K. G. Tanner, W. L. Cheung, P. Sassone-Corsi, J. M. Denu and C. D. Allis (2000). "Synergistic coupling of histone H3 phosphorylation and acetylation in response to epidermal growth factor stimulation." Mol Cell **5**(6): 905-15.

Chowdhury, M. A., H. Kuivaniemi, R. Romero, S. Edwin, T. Chaiworapongsa and G. Tromp (2006). "Identification of novel functional sequence variants in the gene for peptidase inhibitor 3." BMC Med Genet **7**: 49.

Chu, D., N. Kakazu, M. J. Gorrin-Rivas, H. P. Lu, M. Kawata, T. Abe, K. Ueda and Y. Adachi (2001). "Cloning and characterization of LUN, a novel ring finger protein that is highly expressed in lung and specifically binds to a palindromic sequence." J Biol Chem **276**(17): 14004-13.

Chung, J. H., M. Whiteley and G. Felsenfeld (1993). "A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila." Cell **74**(3): 505-14.

Clevebring, D., V. Wirta, A. Skollermo, A. Persson, M. Uhlen, P. Nilson and J. Lundeberg (2006). A new microarray-based method to study sense and antisense transcripts. Human Genome Meeting 2006. Helsinki, Finland.

Cloos, P. A., J. Christensen, K. Agger, A. Maiolica, J. Rappsilber, T. Antal, K. H. Hansen and K. Helin (2006). "The putative oncogene GASC1 demethylates tri- and dimethylated lysine 9 on histone H3." Nature **442**(7100): 307-11.

Cooper, G. M. (2002). The Cell: A Molecular Approach, Sinauer Associates Inc,.

Cooper, S. J., N. D. Trinklein, E. D. Anton, L. Nguyen and R. M. Myers (2006). "Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome." Genome Res **16**(1): 1-10.

Costa, M. and R. L. Medcalf (1996). "Differential binding of cAMP-responsive-element (CRE)-binding protein-1 and activating transcription factor-2 to a CRE-like element in the human tissue-type plasminogen activator (t-PA) gene promoter correlates with opposite regulation of t-PA by phorbol ester in HT-1080 and HeLa cells." Eur J Biochem **237**(3): 532-8.

Coulondre, C., J. H. Miller, P. J. Farabaugh and W. Gilbert (1978). "Molecular basis of base substitution hotspots in Escherichia coli." Nature **274**(5673): 775-80.

Craig, J. M. (2005). "Heterochromatin--many flavours, common themes." Bioessays **27**(1): 17-28.

Cramer, P., D. A. Bushnell, J. Fu, A. L. Gnatt, B. Maier-Davis, N. E. Thompson, R. R. Burgess, A. M. Edwards, P. R. David and R. D. Kornberg (2000). "Architecture of RNA polymerase II and implications for the transcription mechanism." Science **288**(5466): 640-9.

Cramer, P., D. A. Bushnell and R. D. Kornberg (2001). "Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution." Science **292**(5523): 1863-76.

Crick, F. H. (1958). "On protein synthesis." Symp Soc Exp Biol **12**: 138-63.

Davis, J. A., Y. Takagi, R. D. Kornberg and F. A. Asturias (2002). "Structure of the yeast RNA polymerase II holoenzyme: Mediator conformation and polymerase interaction." Mol Cell **10**(2): 409-15.

Davuluri, R. V., I. Grosse and M. Q. Zhang (2001). "Computational identification of promoters and first exons in the human genome." Nat Genet **29**(4): 412-7.

de Belle, I., S. Cai and T. Kohwi-Shigematsu (1998). "The genomic sequences bound to special AT-rich sequence-binding protein 1 (SATB1) in vivo in Jurkat T cells are tightly associated with the nuclear matrix at the bases of the chromatin loops." J Cell Biol **141**(2): 335-48.

de Belle, I., D. Mercola and E. D. Adamson (2000). "Method for cloning in vivo targets of the Egr-1 transcription factor." Biotechniques **29**(1): 162-9.

de Kok, Y. J., G. F. Merkx, S. M. van der Maarel, I. Huber, S. Malcolm, H. H. Ropers and F. P. Cremers (1995). "A duplication/paracentric inversion associated with familial X-linked deafness (DFN3) suggests the presence of a regulatory element more than 400 kb upstream of the POU3F4 gene." Hum Mol Genet **4**(11): 2145-50.

de Napoles, M., J. E. Mermoud, R. Wakao, Y. A. Tang, M. Endoh, R. Appanah, T. B. Nesterova, J. Silva, A. P. Otte, M. Vidal, H. Koseki and N. Brockdorff (2004). "Polycomb group proteins Ring1A/B link ubiquitylation of histone H2A to heritable gene silencing and X inactivation." Dev Cell **7**(5): 663-76.

de Wet, J. R., K. V. Wood, M. DeLuca, D. R. Helinski and S. Subramani (1987). "Firefly luciferase gene: structure and expression in mammalian cells." Mol Cell Biol **7**(2): 725-37.

Dean, A. (2006). "On a chromosome far, far away: LCRs and gene expression." Trends Genet **22**(1): 38-45.

Deguchi, K., P. M. Ayton, M. Carapeti, J. L. Kutok, C. S. Snyder, I. R. Williams, N. C. Cross, C. K. Glass, M. L. Cleary and D. G. Gilliland (2003). "MOZ-TIF2-induced acute myeloid leukemia requires the MOZ nucleosome binding motif and TIF2-mediated recruitment of CBP." Cancer Cell **3**(3): 259-71.

Dekker, J., K. Rippe, M. Dekker and N. Kleckner (2002). "Capturing chromosome conformation." Science **295**(5558): 1306-11.

Deloukas, P., L. H. Matthews, J. Ashurst, J. Burton, J. G. Gilbert, M. Jones, G. Stavrides, J. P. Almeida, A. K. Babbage, C. L. Bagguley, J. Bailey, K. F. Barlow, K. N. Bates, L. M. Beard, D. M. Beare, O. P. Beasley, C. P. Bird, S. E. Blakey, A. M. Bridgeman, A. J. Brown, D. Buck, W. Burrill, A. P. Butler, C. Carder, N. P. Carter, J. C. Chapman, M. Clamp, G. Clark, L. N. Clark, S. Y. Clark, C. M. Clee, S. Clegg, V. E. Cobley, R. E. Collier, R. Connor, N. R. Corby, A. Coulson, G. J. Coville, R. Deadman, P. Dhami, M. Dunn, A. G. Ellington, J. A. Frankland, A. Fraser, L. French, P. Garner, D. V. Grafham, C. Griffiths, M. N. Griffiths, R. Gwilliam, R. E. Hall, S. Hammond, J. L. Harley, P. D. Heath, S. Ho, J. L. Holden, P. J. Howden, E. Huckle, A. R. Hunt, S. E. Hunt, K. Jekosch, C. M. Johnson, D. Johnson, M. P. Kay, A. M. Kimberley, A. King, A. Knights, G. K. Laird, S. Lawlor, M. H. Lehvaslaiho, M. Leversha, C. Lloyd, D. M. Lloyd, J. D. Lovell, V. L. Marsh, S. L. Martin, L. J. McConnachie, K. McLay, A. A. McMurray, S. Milne, D. Mistry, M. J. Moore, J. C. Mullikin, T. Nickerson, K. Oliver, A. Parker, R. Patel, T. A. Pearce, A. I. Peck, B. J. Phillimore, S. R. Prathalingam, R. W. Plumb, H. Ramsay, C. M. Rice, M. T. Ross, C. E. Scott, H. K. Sehra, R. Shownkeen, S. Sims, C. D. Skuce, M. L. Smith, C. Soderlund, C. A. Steward, J. E. Sulston, M. Swann, N. Sycamore, R. Taylor, L. Tee, D. W. Thomas, A. Thorpe, A. Tracey, A. C. Tromans, M. Vaudin, M. Wall, J. M. Wallis, S. L. Whitehead, P. Whittaker, D. L. Willey, L. Williams, S. A. Williams, L. Wilming, P. W. Wray, T. Hubbard, R. M. Durbin, D. R. Bentley, S. Beck and J. Rogers (2001). "The DNA sequence and comparative analysis of human chromosome 20." Nature **414**(6866): 865-71.

Dermitzakis, E. T., A. Reymond and S. E. Antonarakis (2005). "Conserved non-genic sequences - an unexpected feature of mammalian genomes." Nat Rev Genet **6**(2): 151-7.

Dermitzakis, E. T., A. Reymond, N. Scamuffa, C. Ucla, E. Kirkness, C. Rossier and S. E. Antonarakis (2003). "Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs)." Science **302**(5647): 1033-5.

Deutsch, P. J., J. P. Hoeffler, J. L. Jameson, J. C. Lin and J. F. Habener (1988). "Structural determinants for transcriptional activation by cAMP-responsive DNA elements." J Biol Chem **263**(34): 18466-72.

Di Fiore, P. P., S. Polo and K. Hofmann (2003). "When ubiquitin meets ubiquitin receptors: a signalling connection." Nat Rev Mol Cell Biol **4**(6): 491-7.

Dimitri, P., N. Corradini, F. Rossi and F. Verni (2005). "The paradox of functional heterochromatin." Bioessays **27**(1): 29-41.

Dimova, I., A. Yosifova, B. Zaharieva, S. Raitcheva, N. Doganov and D. Toncheva (2005). "Association of 20q13.2 copy number changes with the advanced stage of ovarian cancer-tissue microarray analysis." Eur J Obstet Gynecol Reprod Biol **118**(1): 81-5.

Down, T. A. and T. J. Hubbard (2002). "Computational detection and location of transcription start sites in mammalian genomic DNA." Genome Res **12**(3): 458-61.

Ducrest, A. L., M. Amacker, J. Lingner and M. Nabholz (2002). "Detection of promoter activity by flow cytometric analysis of GFP reporter expression." Nucleic Acids Research (Online) **30**(14): e65.

Dunn, K. L., H. Zhao and J. R. Davie (2003). "The insulator binding protein CTCF associates with the nuclear matrix." Exp Cell Res **288**(1): 218-23.

Duret, L., F. Dorkeld and C. Gautier (1993). "Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression." Nucleic Acids Res **21**(10): 2315-22.

Eberharter, A. and P. B. Becker (2002). "Histone acetylation: a switch between repressive and permissive chromatin. Second in review series on chromatin dynamics." EMBO Rep **3**(3): 224-9.

Eferl, R., M. Sibilia, F. Hilberg, A. Fuchsbichler, I. Kufferath, B. Guertl, R. Zenz, E. F. Wagner and K. Zatloukal (1999). "Functions of c-Jun in liver and heart development." J Cell Biol **145**(5): 1049-61.

Elefant, F., N. E. Cooke and S. A. Liebhaber (2000). "Targeted recruitment of histone acetyltransferase activity to a locus control region." J Biol Chem **275**(18): 13827-34.

ENCODE (2004). "The ENCODE (ENCyclopedia Of DNA Elements) Project." Science **306**(5696): 636-40.

Eron, L. and R. Block (1971). "Mechanism of initiation and repression of in vitro transcription of the lac operon of Escherichia coli." Proc Natl Acad Sci U S A **68**(8): 1828-32.

Ewert, K., N. L. Slack, A. Ahmad, H. M. Evans, A. J. Lin, C. E. Samuel and C. R. Safinya (2004). "Cationic lipid-DNA complexes for gene therapy: understanding the relationship between complex structure and gene delivery pathways at the molecular level." Curr Med Chem **11**(2): 133-49.

Feil, R. and S. Khosla (1999). "Genomic imprinting in mammals: an interplay between chromatin and DNA methylation?" Trends Genet **15**(11): 431-5.

Fickett, J. W. and A. G. Hatzigeorgiou (1997). "Eukaryotic promoter recognition." Genome Res **7**(9): 861-78.

Finkelstein, A., C. F. Kostrub, J. Li, D. P. Chavez, B. Q. Wang, S. M. Fang, J. Greenblatt and Z. F. Burton (1992). "A cDNA encoding RAP74, a general initiation factor for transcription by RNA polymerase II." Nature **355**(6359): 464-7.

Florquin, K., Y. Saeys, S. Degroeve, P. Rouze and Y. Van de Peer (2005). "Large-scale structural analysis of the core promoter in mammalian and plant genomes." Nucleic Acids Res **33**(13): 4255-64.

Fourel, G., F. Magdinier and E. Gilson (2004). "Insulator dynamics and the setting of chromatin domains." Bioessays **26**(5): 523-32.

Fraga, M. F., E. Ballestar, A. Villar-Garea, M. Boix-Chornet, J. Espada, G. Schotta, T. Bonaldi, C. Haydon, S. Ropero, K. Petrie, N. G. Iyer, A. Perez-Rosado, E. Calvo, J. A. Lopez, A. Cano, M. J. Calasanz, D. Colomer, M. A. Piris, N. Ahn, A. Imhof, C. Caldas, T. Jenuwein and M. Esteller (2005). "Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer." Nat Genet **37**(4): 391-400.

Fraga, M. F. and M. Esteller (2005). "Towards the human cancer epigenome: a first draft of histone modifications." Cell Cycle **4**(10): 1377-81.

Fu, X. H., D. P. Liu and C. C. Liang (2002). "Chromatin structure and transcriptional regulation of the beta-globin locus." Exp Cell Res **278**(1): 1-11.

Fukue, Y., N. Sumida, J. Tanase and T. Ohyama (2005). "A highly distinctive mechanical property found in the majority of human promoters and its transcriptional relevance." Nucleic Acids Res **33**(12): 3821-7.

Gao, C., X. Hou, F. Zhang, W. Zhou, Y. Yuan and X. Dong (2002). "[Establishment of a sandwich ELISA method for detection of reporter chloramphenicol acetyltransferase gene]." Zhonghua Shi Yan He Lin Chuang Bing Du Xue Za Zhi **16**(1): 69-73.

Gardiner-Garden, M. and M. Frommer (1987). "CpG islands in vertebrate genomes." J Mol Biol **196**(2): 261-82.

Gershenzon, N. I. and I. P. Ioshikhes (2005). "Synergy of human Pol II core promoter elements revealed by statistical sequence analysis." Bioinformatics **21**(8): 1295-300.

Gill, G. (2005). "Something about SUMO inhibits transcription." Curr Opin Genet Dev **15**(5): 536-41.

Gill, G. and M. Ptashne (1987). "Mutants of GAL4 protein altered in an activation function." Cell **51**(1): 121-6.

Gohring, F. and F. O. Fackelmayer (1997). "The scaffold/matrix attachment region binding protein hnRNP-U (SAF-A) is directly bound to chromosomal DNA in vivo: a chemical cross-linking study." Biochemistry **36**(27): 8276-83.

Gorman, C. M., L. F. Moffat and B. H. Howard (1982). "Recombinant genomes which express chloramphenicol acetyltransferase in mammalian cells." Mol Cell Biol **2**(9): 1044-51.

Grandori, C., S. M. Cowley, L. P. James and R. N. Eisenman (2000). "The Myc/Max/Mad network and the transcriptional control of cell behavior." Annu Rev Cell Dev Biol **16**: 653-99.

Green, N. S., E. Reisler and K. N. Houk (2001). "Quantitative evaluation of the lengths of homobifunctional protein cross-linking reagents used as molecular rulers." Protein Sci **10**(7): 1293-304.

Grewal, S. I. and D. Moazed (2003). "Heterochromatin and epigenetic control of gene expression." Science **301**(5634): 798-802.

Gribskov, M. and R. R. Burgess (1986). "Sigma factors from E. coli, B. subtilis, phage SP01, and phage T4 are homologous proteins." Nucleic Acids Res **14**(16): 6745-63.

Grimwood, J., L. A. Gordon, A. Olsen, A. Terry, J. Schmutz, J. Lamerdin, U. Hellsten, D. Goodstein, O. Couronne, M. Tran-Gyamfi, A. Aerts, M. Altherr, L. Ashworth, E. Bajorek, S. Black, E. Branscomb, S. Caenepeel, A. Carrano, C. Caoile, Y. M. Chan, M. Christensen, C. A. Cleland, A. Copeland, E. Dalin, P. Dehal, M. Denys, J. C. Detter, J. Escobar, D. Flowers, D. Fotopulos, C. Garcia, A. M. Georgescu, T. Glavina, M. Gomez, E. Gonzales, M. Groza, N. Hammon, T. Hawkins, L. Haydu, I. Ho, W. Huang, S. Israni, J. Jett, K. Kadner, H. Kimball, A. Kobayashi, V. Larionov, S. H. Leem, F. Lopez, Y. Lou, S. Lowry, S. Malfatti, D. Martinez, P. McCready, C. Medina, J. Morgan, K. Nelson, M. Nolan, I. Ovcharenko, S. Pitluck, M. Pollard, A. P. Popkie, P. Predki, G. Quan, L. Ramirez, S. Rash, J. Retterer, A. Rodriguez, S. Rogers, A. Salamov, A. Salazar, X. She, D. Smith, T. Slezak, V. Solovyev, N. Thayer, H. Tice, M. Tsai, A. Ustaszewska, N. Vo, M. Wagner, J. Wheeler, K. Wu, G. Xie, J. Yang, I. Dubchak, T. S. Furey, P. DeJong, M. Dickson, D. Gordon, E. E. Eichler, L. A. Pennacchio, P. Richardson, L. Stubbs, D. S. Rokhsar, R. M. Myers, E. M. Rubin and S. M. Lucas (2004). "The DNA sequence and biology of human chromosome 19." Nature **428**(6982): 529-35.

Grosschedl, R. (1995). "Higher-order nucleoprotein complexes in transcription: analogies with site-specific recombination." Curr Opin Cell Biol **7**(3): 362-70.

Grosveld, F., G. B. van Assendelft, D. R. Greaves and G. Kollias (1987). "Position-independent, high-level expression of the human beta-globin gene in transgenic mice." Cell **51**(6): 975-85.

Gruda, M. C., J. M. Zabolotny, J. H. Xiao, I. Davidson and J. C. Alwine (1993). "Transcriptional activation by simian virus 40 large T antigen: interactions with multiple components of the transcription complex." Mol Cell Biol **13**(2): 961-9.

Grunstein, M. (1997). "Histone acetylation in chromatin structure and transcription." Nature **389**(6649): 349-52.

Hahn, S. (1992). "The Yin and the Yang of mammalian transcription." Curr Biol **2**(3): 152-4.

Hannenhalli, S. and S. Levy (2001). "Promoter prediction in the human genome." Bioinformatics **17 Suppl 1**: S90-6.

Hans, F. and S. Dimitrov (2001). "Histone H3 phosphorylation and cell division." Oncogene **20**(24): 3021-7.

Hebert, S. C., D. B. Mount and G. Gamba (2004). "Molecular physiology of cation-coupled Cl- cotransport: the SLC12 family." Pflugers Arch **447**(5): 580-93.

Hess, J., P. Angel and M. Schorpp-Kistner (2004). "AP-1 subunits: quarrel and harmony among siblings." J Cell Sci **117**(Pt 25): 5965-73.

Hirota, T., J. J. Lipp, B. H. Toh and J. M. Peters (2005). "Histone H3 serine 10 phosphorylation by Aurora B causes HP1 dissociation from heterochromatin." Nature **438**(7071): 1176-80.

Horak, C. E., M. C. Mahajan, N. M. Luscombe, M. Gerstein, S. M. Weissman and M. Snyder (2002). "GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIp-chip analysis." Proc Natl Acad Sci U S A **99**(5): 2924-9.

Hurlin, P. J., C. Queva and R. N. Eisenman (1997). "Mnt, a novel Max-interacting protein is coexpressed with Myc in proliferating cells and mediates repression at Myc binding sites." Genes Dev **11**(1): 44-58.

IHGSC (2004). "Finishing the euchromatic sequence of the human genome." Nature **431**(7011): 931-45.

Iizuka, M., T. Matsui, H. Takisawa and M. M. Smith (2006). "Regulation of replication licensing by acetyltransferase Hbo1." Mol Cell Biol **26**(3): 1098-108.

Indra, A. K., W. S. Mohan, 2nd, M. Frontini, E. Scheer, N. Messaddeq, D. Metzger and L. Tora (2005). "TAF10 is required for the establishment of skin barrier function in foetal, but not in adult mouse epidermis." Dev Biol **285**(1): 28-37.

Ioshikhes, I. P. and M. Q. Zhang (2000). "Large-scale human promoter mapping using CpG islands." Nat Genet **26**(1): 61-3.

Ippen, K., J. H. Miller, J. Scaife and J. Beckwith (1968). "New controlling element in the Lac operon of E. coli." Nature **217**(131): 825-7.

Ivanova, N., R. Dobrin, R. Lu, I. Kotenko, J. Levorse, C. Decoste, X. Schafer, Y. Lun and I. R. Lemischka (2006). "Dissecting self-renewal in stem cells with RNA interference." Nature.

Jackson, V. (1978). "Studies on histone organization in the nucleosome using formaldehyde as a reversible cross-linking agent." Cell **15**(3): 945-54.

Jacobson, R. H., A. G. Ladurner, D. S. King and R. Tjian (2000). "Structure and function of a human TAFII250 double bromodomain module." Science **288**(5470): 1422-5.

Johnson, K. D. and E. H. Bresnick (2002). "Dissecting long-range transcriptional mechanisms by chromatin immunoprecipitation." Methods **26**(1): 27-36.

Jones, P. A. and P. W. Laird (1999). "Cancer epigenetics comes of age." Nat Genet **21**(2): 163-7.

Jones, P. L., G. J. Veenstra, P. A. Wade, D. Vermaak, S. U. Kass, N. Landsberger, J. Strouboulis and A. P. Wolffe (1998). "Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription." Nat Genet **19**(2): 187-91.

Jones, R. S. and W. M. Gelbart (1993). "The Drosophila Polycomb-group gene Enhancer of zeste contains a region with sequence similarity to trithorax." Mol Cell Biol **13**(10): 6357-66.

Jongeneel, C. V., M. Delorenzi, C. Iseli, D. Zhou, C. D. Haudenschild, I. Khrebtukova, D. Kuznetsov, B. J. Stevenson, R. L. Strausberg, A. J. Simpson and T. J. Vasicek (2005). "An atlas of human gene expression from massively parallel signature sequencing (MPSS)." Genome Res **15**(7): 1007-14.

Kaffer, C. R., M. Srivastava, K. Y. Park, E. Ives, S. Hsieh, J. Batlle, A. Grinberg, S. P. Huang and K. Pfeifer (2000). "A transcriptional insulator at the imprinted H19/Igf2 locus." Genes Dev **14**(15): 1908-19.

Kain, S. R., M. Adams, A. Kondepudi, T. T. Yang, W. W. Ward and P. Kitts (1995). "Green fluorescent protein as a reporter of gene expression and protein localization." Biotechniques **19**(4): 650-5.

Kaplan, J. and K. Calame (1997). "The ZiN/POZ domain of ZF5 is required for both transcriptional activation and repression." Nucleic Acids Res **25**(6): 1108-16.

Kauffman, S. (1995). At Home in the Universe: The Search for the Laws of Self-Organization and Complexity New York, Oxford University Press.

Keller, C., E. M. Ladenburger, M. Kremer and R. Knippers (2002). "The origin recognition complex marks a replication origin in the human TOP1 gene promoter." J Biol Chem **277**(35): 31430-40.

Kenney, J. F. K., E. S. (1962). The Standard Deviation" and "Calculation of the Standard Deviation. Mathematics of Statistics Princeton, NJ, Van Nostrand**:** 77-80.

Kim, J. and D. J. Shapiro (1996). "In simple synthetic promoters YY1-induced DNA bending is important in transcription activation and repression." Nucleic Acids Res **24**(21): 4341-8.

Kim, M., S. H. Ahn, N. J. Krogan, J. F. Greenblatt and S. Buratowski (2004). "Transitions in RNA polymerase II elongation complexes at the 3' ends of genes." Embo J **23**(2): 354-64.

Kim, T. H., L. O. Barrera, M. Zheng, C. Qu, M. A. Singer, T. A. Richmond, Y. Wu, R. D. Green and B. Ren (2005). "A high-resolution map of active promoters in the human genome." Nature **436**(7052): 876-80.

Kim, T. H. and B. Ren (2006). "Genome-Wide Analysis of Protein-DNA Interactions." Annu Rev Genomics Hum Genet.

Klenova, E., I. Chernukhin, T. Inoue, S. Shamsuddin and J. Norton (2002). "Immunoprecipitation techniques for the analysis of transcription factor complexes." Methods **26**(3): 254-9.

Klochkov, D., H. Rincon-Arano, E. S. Ioudinkova, V. Valadez-Graham, A. Gavrilov, F. Recillas-Targa and S. V. Razin (2006). "A CTCF-dependent silencer located in the differentially methylated area may regulate expression of a housekeeping gene overlapping a tissue-specific gene domain." Mol Cell Biol **26**(5): 1589-97.

Klose, R. J., K. Yamane, Y. Bae, D. Zhang, H. Erdjument-Bromage, P. Tempst, J. Wong and Y. Zhang (2006). "The transcriptional repressor JHDM3A demethylates trimethyl histone H3 lysine 9 and lysine 36." Nature **442**(7100): 312-6.

Knudsen, S. (1999). "Promoter2.0: for the recognition of PolII promoter sequences." Bioinformatics **15**(5): 356-61.

Koivisto, U. M., J. J. Palvimo, O. A. Janne and K. Kontula (1994). "A single-base substitution in the proximal Sp1 site of the human low density lipoprotein receptor promoter as a cause of heterozygous familial hypercholesterolemia." Proc Natl Acad Sci U S A **91**(22): 10526-30.

Komarnitsky, P., E. J. Cho and S. Buratowski (2000). "Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription." Genes Dev **14**(19): 2452-60.

Krajewski, W. A. and P. B. Becker (1998). "Reconstitution of hyperacetylated, DNase I-sensitive chromatin characterized by high conformational flexibility of nucleosomal DNA." Proc Natl Acad Sci U S A **95**(4): 1540-5.

Kuo, M. H. and C. D. Allis (1999). "In vivo cross-linking and immunoprecipitation for studying dynamic Protein:DNA associations in a chromatin environment." Methods **19**(3): 425-33.

Kurdistani, S. K., D. Robyr, S. Tavazoie and M. Grunstein (2002). "Genome-wide binding map of the histone deacetylase Rpd3 in yeast." Nat Genet **31**(3): 248-54.

Kuzmin, I., L. Geil, L. Gibson, T. Cavinato, D. Loukinov, V. Lobanenkov and M. I. Lerman (2005). "Transcriptional regulator CTCF controls human interleukin 1 receptor-associated kinase 2 promoter." J Mol Biol **346**(2): 411-22.

Lachner, M., D. O'Carroll, S. Rea, K. Mechtler and T. Jenuwein (2001). "Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins." Nature **410**(6824): 116-20.

Lalioti, M. D., H. S. Scott and S. E. Antonarakis (1999). "Altered spacing of promoter elements due to the dodecamer repeat expansion contributes to reduced expression of the cystatin B gene in EPM1." Hum Mol Genet **8**(9): 1791-8.

Larsen, F., G. Gundersen, R. Lopez and H. Prydz (1992). "CpG islands as gene markers in the human genome." Genomics **13**(4): 1095-107.

Latchman, D. (1998). Eukaryotic Transcription Factors, Academic Press.

Lee, M. G., C. Wynder, N. Cooch and R. Shiekhattar (2005). "An essential role for CoREST in nucleosomal histone 3 lysine 4 demethylation." Nature **437**(7057): 432-5.

Lee, M. P., K. Howcroft, A. Kotekar, H. H. Yang, K. H. Buetow and D. S. Singer (2005). "ATG deserts define a novel core promoter subclass." Genome Res **15**(9): 1189-97.

Lee, T. I., H. C. Causton, F. C. Holstege, W. C. Shen, N. Hannett, E. G. Jennings, F. Winston, M. R. Green and R. A. Young (2000). "Redundant roles for the TFIID and SAGA complexes in global transcription." Nature **405**(6787): 701-4.

Lee, Y., M. Kim, J. Han, K. H. Yeom, S. Lee, S. H. Baek and V. N. Kim (2004). "MicroRNA genes are transcribed by RNA polymerase II." Embo J **23**(20): 4051-60.

Leichter, M. and G. Thiel (1999). "Transcriptional repression by the zinc finger protein REST is mediated by titratable nuclear factors." Eur J Neurosci **11**(6): 1937-46.

Lejnine, S., G. Durfee, M. Murnane, H. C. Kapteyn, V. L. Makarov and J. P. Langmore (1999). "Crosslinking of proteins to DNA in human nuclei using a 60 femtosecond 266 nm laser." Nucleic Acids Res **27**(18): 3676-84.

Letovsky, J. and W. S. Dynan (1989). "Measurement of the binding of transcription factor Sp1 to a single GC box recognition sequence." Nucleic Acids Res **17**(7): 2639-53.

Lettice, L. A., S. J. Heaney, L. A. Purdie, L. Li, P. de Beer, B. A. Oostra, D. Goode, G. Elgar, R. E. Hill and E. de Graaff (2003). "A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly." Hum Mol Genet **12**(14): 1725-35.

Leuther, K. K., D. A. Bushnell and R. D. Kornberg (1996). "Two-dimensional crystallography of TFIIB- and IIE-RNA polymerase II complexes: implications for start site selection and initiation complex formation." Cell **85**(5): 773-9.

Li, M., Y. H. Xie, Y. Y. Kong, X. Wu, L. Zhu and Y. Wang (1998). "Cloning and characterization of a novel human hepatocyte transcription factor, hB1F, which binds and activates enhancer II of hepatitis B virus." J Biol Chem **273**(44): 29022-31.

Li, Q., M. Zhang, Z. Duan and G. Stamatoyannopoulos (1999). "Structural analysis and mapping of DNase I hypersensitivity of HS5 of the beta-globin locus control region." Genomics **61**(2): 183-93.

Li, Z., S. Van Calcar, C. Qu, W. K. Cavenee, M. Q. Zhang and B. Ren (2003). "A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells." Proc Natl Acad Sci U S A **100**(14): 8164-9.

Liang, G., J. C. Lin, V. Wei, C. Yoo, J. C. Cheng, C. T. Nguyen, D. J. Weisenberger, G. Egger, D. Takai, F. A. Gonzales and P. A. Jones (2004). "Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome." Proc Natl Acad Sci U S A **101**(19): 7357-62.

Lim, C. Y., B. Santoso, T. Boulay, E. Dong, U. Ohler and J. T. Kadonaga (2004). "The MTE, a new core promoter element for transcription by RNA polymerase II." Genes Dev **18**(13): 1606-17.

Litt, M. D., M. Simpson, M. Gaszner, C. D. Allis and G. Felsenfeld (2001). "Correlation between histone lysine methylation and developmental changes at the chicken beta-globin locus." Science **293**(5539): 2453-5.

Locker, J. (2001). Transcription Factors. Oxford, BIOS Scientific Publishers Ltd.

Locker, J. (2001). Transcripton Factors. Oxford, BIOS Scientific Publishers Ltd.

Lomvardas, S., G. Barnea, D. J. Pisapia, M. Mendelsohn, J. Kirkland and R. Axel (2006). "Interchromosomal interactions and olfactory receptor choice." Cell **126**(2): 403-13.

Lonard, D. M. and B. W. O'Malley (2005). "Expanding functional diversity of the coactivators." Trends Biochem Sci **30**(3): 126-32.

Louie, M. C., A. S. Revenko, J. X. Zou, J. Yao and H. W. Chen (2006). "Direct control of cell cycle gene expression by proto-oncogene product ACTR, and its autoregulation underlies its transforming activity." Mol Cell Biol **26**(10): 3810-23.

Luger, K., A. W. Mader, R. K. Richmond, D. F. Sargent and T. J. Richmond (1997). "Crystal structure of the nucleosome core particle at 2.8 A resolution." Nature **389**(6648): 251-60.

Lutz, M., L. J. Burke, G. Barreto, F. Goeman, H. Greb, R. Arnold, H. Schultheiss, A. Brehm, T. Kouzarides, V. Lobanenkov and R. Renkawitz (2000). "Transcriptional repression by the insulator protein CTCF involves histone deacetylases." Nucleic Acids Res **28**(8): 1707-13.

Mahadevan, L. C., A. C. Willis and M. J. Barratt (1991). "Rapid histone H3 phosphorylation in response to growth factors, phorbol esters, okadaic acid, and protein synthesis inhibitors." Cell **65**(5): 775-83.

Mahajan, M. A., A. Murray and H. H. Samuels (2002). "NRC-interacting factor 1 is a novel cotransducer that interacts with and regulates the activity of the nuclear hormone receptor coactivator NRC." Mol Cell Biol **22**(19): 6883-94.

Makowski, A. M., R. N. Dutnall and A. T. Annunziato (2001). "Effects of acetylation of histone H4 at lysines 8 and 16 on activity of the Hat1 histone acetyltransferase." J Biol Chem **276**(47): 43499-502.

Manley, J. L., A. Fire, A. Cano, P. A. Sharp and M. L. Gefter (1980). "DNA-dependent transcription of adenovirus genes in a soluble whole-cell extract." Proc Natl Acad Sci U S A **77**(7): 3855-9.

Mantripragada, K. K., P. G. Buckley, T. D. de Stahl and J. P. Dumanski (2004). "Genomic microarrays in the spotlight." Trends Genet **20**(2): 87-94.

Mao, D. Y., J. D. Watson, P. S. Yan, D. Barsyte-Lovejoy, F. Khosravi, W. W. Wong, P. J. Farnham, T. H. Huang and L. Z. Penn (2003). "Analysis of Myc bound loci identified by CpG island arrays shows that Max is essential for Myc-dependent repression." Curr Biol **13**(10): 882-6.

Margulies, E. H., M. Blanchette, D. Haussler and E. D. Green (2003). "Identification and characterization of multi-species conserved sequences." Genome Res **13**(12): 2507-18.

Marinescu, V. D., I. S. Kohane and A. Riva (2005). "MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes." BMC Bioinformatics **6**: 79.

Martone, R., G. Euskirchen, P. Bertone, S. Hartman, T. E. Royce, N. M. Luscombe, J. L. Rinn, F. K. Nelson, P. Miller, M. Gerstein, S. Weissman and M. Snyder (2003). "Distribution of NF-kappaB-binding sites across human chromosome 22." Proc Natl Acad Sci U S A **100**(21): 12247-52.

Massari, M. E. and C. Murre (2000). "Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms." Mol Cell Biol **20**(2): 429-40.

Maston, G. A., S. K. Evans and M. R. Green (2006). "Transcriptional Regulatory Elements in the Human Genome." Annu Rev Genomics Hum Genet.

Mavilio, F., A. Simeone, E. Boncinelli and P. W. Andrews (1988). "Activation of four homeobox gene clusters in human embryonal carcinoma cells induced to differentiate by retinoic acid." Differentiation **37**(1): 73-9.

McCabe, C. D. and J. W. Innis (2005). "A genomic approach to the identification and characterization of HOXA13 functional binding elements." Nucleic Acids Res **33**(21): 6782-94.

McKusick, V. A. (1998). Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders., Baltimore: Johns Hopkins University Press.

McManus, K. J., V. L. Biron, R. Heit, D. A. Underhill and M. J. Hendzel (2006). "Dynamic changes in histone H3 lysine 9 methylations: identification of a mitosis-specific function for dynamic methylation in chromosome congression and segregation." J Biol Chem **281**(13): 8888-97.

Melchior, F. (2000). "SUMO--nonclassical ubiquitin." Annu Rev Cell Dev Biol **16**: 591-626.

Merika, M., A. J. Williams, G. Chen, T. Collins and D. Thanos (1998). "Recruitment of CBP/p300 by the IFN beta enhanceosome is required for synergistic activation of transcription." Mol Cell **1**(2): 277-87.

Mersfelder, E. L. and M. R. Parthun (2006). "The tale beyond the tail: histone core domain modifications and the regulation of chromatin structure." Nucleic Acids Res **34**(9): 2653-62.

Micklem, G. (1999). cpgreport, Unpublished.

Milne, T. A., S. D. Briggs, H. W. Brock, M. E. Martin, D. Gibbs, C. D. Allis and J. L. Hess (2002). "MLL targets SET domain methyltransferase activity to Hox gene promoters." Mol Cell **10**(5): 1107-17.

Min, J., Y. Zhang and R. M. Xu (2003). "Structural basis for specific binding of Polycomb chromodomain to histone H3 methylated at Lys 27." Genes Dev **17**(15): 1823-8.

Mitchell, S. M., A. L. Gloyn, K. R. Owen, A. T. Hattersley and T. M. Frayling (2002). "The role of the HNF4alpha enhancer in type 2 diabetes." Mol Genet Metab **76**(2): 148-51.

Mitra, R. D., C. M. Silva and D. C. Youvan (1996). "Fluorescence resonance energy transfer between blue-emitting and red-shifted excitation derivatives of the green fluorescent protein." Gene **173**(1 Spec No): 13-7.

Mizzen, C. A., X. J. Yang, T. Kokubo, J. E. Brownell, A. J. Bannister, T. Owen-Hughes, J. Workman, L. Wang, S. L. Berger, T. Kouzarides, Y. Nakatani and C.

D. Allis (1996). "The TAF(II)250 subunit of TFIID has histone acetyltransferase activity." Cell **87**(7): 1261-70.

Mo, X., E. Kowenz-Leutz, H. Xu and A. Leutz (2004). "Ras induces mediator complex exchange on C/EBP beta." Mol Cell **13**(2): 241-50.

Mohan, W. S., Jr., E. Scheer, O. Wendling, D. Metzger and L. Tora (2003). "TAF10 (TAF(II)30) is necessary for TFIID stability and early embryogenesis in mice." Mol Cell Biol **23**(12): 4307-18.

Moller, K., J. Rinke, A. Ross, G. Buddle and R. Brimacombe (1977). "The use of formaldehyde in RNA-protein cross-linking studies with ribosomal subunits from Escherichia coli." Eur J Biochem **76**(1): 175-87.

Mukhopadhyay, R., W. Yu, J. Whitehead, J. Xu, M. Lezcano, S. Pack, C. Kanduri, M. Kanduri, V. Ginjala, A. Vostrov, W. Quitschke, I. Chernukhin, E. Klenova, V. Lobanenkov and R. Ohlsson (2004). "The binding sites for the chromatin insulator protein CTCF map to DNA methylation-free domains genome-wide." Genome Res **14**(8): 1594-602.

Muravyova, E., A. Golovnin, E. Gracheva, A. Parshikov, T. Belenkaya, V. Pirrotta and P. Georgiev (2001). "Loss of insulator activity by paired Su(Hw) chromatin insulators." Science **291**(5503): 495-8.

Murre, C., G. Bain, M. A. van Dijk, I. Engel, B. A. Furnari, M. E. Massari, J. R. Matthews, M. W. Quong, R. R. Rivera and M. H. Stuiver (1994). "Structure and function of helix-loop-helix proteins." Biochim Biophys Acta **1218**(2): 129-35.

Musri, M. M., H. Corominola, R. Casamitjana, R. Gomis and M. Parrizas (2006). "Histone H3 lysine 4 dimethylation signals the transcriptional competence of the adiponectin promoter in preadipocytes." J Biol Chem **281**(25): 17180-8.

Nakamura, T., T. Mori, S. Tada, W. Krajewski, T. Rozovskaia, R. Wassell, G. Dubois, A. Mazo, C. M. Croce and E. Canaani (2002). "ALL-1 is a histone methyltransferase that assembles a supercomplex of proteins involved in transcriptional regulation." Mol Cell **10**(5): 1119-28.

Natesan, S. and M. Z. Gilman (1993). "DNA bending and orientation-dependent function of YY1 in the c-fos promoter." Genes Dev **7**(12B): 2497-509.

Nguyen, C. T., F. A. Gonzales and P. A. Jones (2001). "Altered chromatin structure associated with methylation-induced gene silencing in cancer cells: correlation of accessibility, methylation, MeCP2 binding and acetylation." Nucleic Acids Res **29**(22): 4598-606.

Nielsen, J. A., L. D. Hudson and R. C. Armstrong (2002). "Nuclear organization in differentiating oligodendrocytes." J Cell Sci **115**(Pt 21): 4071-9.

Nishida, H., T. Suzuki, S. Kondo, H. Miura, Y. Fujimura and Y. Hayashizaki (2006). "Histone H3 acetylated at lysine 9 in promoter is associated with low nucleosome density in the vicinity of transcription start site in human cell." Chromosome Res **14**(2): 203-11.

Nobrega, M. A., Y. Zhu, I. Plajzer-Frick, V. Afzal and E. M. Rubin (2004). "Megabase deletions of gene deserts result in viable mice." Nature **431**(7011): 988-93.

Numoto, M., K. Yokoro and J. Koshi (1999). "ZF5, which is a Kruppel-type transcriptional repressor, requires the zinc finger domain for self-association." Biochem Biophys Res Commun **256**(3): 573-8.

Nusbaum, C., M. C. Zody, M. L. Borowsky, M. Kamal, C. D. Kodira, T. D. Taylor, C. A. Whittaker, J. L. Chang, C. A. Cuomo, K. Dewar, M. G. FitzGerald, X. Yang, A. Abouelleil, N. R. Allen, S. Anderson, T. Bloom, B. Bugalter, J. Butler, A. Cook, D. DeCaprio, R. Engels, M. Garber, A. Gnirke, N. Hafez, J. L. Hall, C. H. Norman, T. Itoh, D. B. Jaffe, Y. Kuroki, J. Lehoczky, A. Lui, P. Macdonald, E. Mauceli, T. S. Mikkelsen, J. W. Naylor, R. Nicol, C. Nguyen, H. Noguchi, S.

B. O'Leary, K. O'Neill, B. Piqani, C. L. Smith, J. A. Talamas, K. Topham, Y. Totoki, A. Toyoda, H. M. Wain, S. K. Young, Q. Zeng, A. R. Zimmer, A. Fujiyama, M. Hattori, B. W. Birren, Y. Sakaki and E. S. Lander (2005). "DNA sequence and analysis of human chromosome 18." Nature **437**(7058): 551-5.

Odom, D. T., N. Zizlsperger, D. B. Gordon, G. W. Bell, N. J. Rinaldi, H. L. Murray, T. L. Volkert, J. Schreiber, P. A. Rolfe, D. K. Gifford, E. Fraenkel, G. I. Bell and R. A. Young (2004). "Control of pancreas and liver gene expression by HNF transcription factors." Science **303**(5662): 1378-81.

Ogbourne, S. and T. M. Antalis (1998). "Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes." Biochem J **331 (Pt 1)**: 1-14.

Ohlendorf, D. H., W. F. Anderson and B. W. Matthews (1983). "Many gene-regulatory proteins appear to have a similar alpha-helical fold that binds DNA and evolved from a common precursor." J Mol Evol **19**(2): 109-14.

Ohler, U., G. Stemmer, S. Harbeck and H. Niemann (2000). "Stochastic segment models of eukaryotic promoter regions." Pac Symp Biocomput: 380-91.

Ohlsson, R., R. Renkawitz and V. Lobanenkov (2001). "CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease." Trends Genet **17**(9): 520-7.

Onate, S. A., S. Y. Tsai, M. J. Tsai and B. W. O'Malley (1995). "Sequence and characterization of a coactivator for the steroid hormone receptor superfamily." Science **270**(5240): 1354-7.

Orlando, V. (2000). "Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation." Trends Biochem Sci **25**(3): 99-104.

Orlando, V. (2003). "Polycomb, epigenomes, and control of cell identity." Cell **112**(5): 599-606.

Orlando, V. and R. Paro (1993). "Mapping Polycomb-repressed domains in the bithorax complex using in vivo formaldehyde cross-linked chromatin." Cell **75**(6): 1187-98.

Orlando, V., H. Strutt and R. Paro (1997). "Analysis of chromatin structure by in vivo formaldehyde cross-linking." Methods **11**(2): 205-14.

Ortlund, E. A., Y. Lee, I. H. Solomon, J. M. Hager, R. Safi, Y. Choi, Z. Guan, A. Tripathy, C. R. Raetz, D. P. McDonnell, D. D. Moore and M. R. Redinbo (2005). "Modulation of human nuclear receptor LRH-1 activity by phospholipids and SHP." Nat Struct Mol Biol **12**(4): 357-63.

Ota, Y., A. Kikuchi and M. Cashel (1979). "Gene expression of an Escherichia coli ribosomal RNA promoter fused to structural genes of the galactose operon." Proc Natl Acad Sci U S A **76**(11): 5799-803.

Panning, B. and R. Jaenisch (1998). "RNA and the epigenetic regulation of X chromosome inactivation." Cell **93**(3): 305-8.

Paro, R. and D. S. Hogness (1991). "The Polycomb protein shares a homologous domain with a heterochromatin-associated protein of Drosophila." Proc Natl Acad Sci U S A **88**(1): 263-7.

Parsons, M. J., U. M. D'Souza, M. J. Arranz, R. W. Kerwin and A. J. Makoff (2004). "The -1438A/G polymorphism in the 5-hydroxytryptamine type 2A receptor gene affects promoter activity." Biol Psychiatry **56**(6): 406-10.

Pellicer, A., D. Robins, B. Wold, R. Sweet, J. Jackson, I. Lowy, J. M. Roberts, G. K. Sim, S. Silverstein and R. Axel (1980). "Altering genotype and phenotype by DNA-mediated gene transfer." Science **209**(4463): 1414-22.

Peng, H., G. E. Begg, S. L. Harper, J. R. Friedman, D. W. Speicher and F. J. Rauscher, 3rd (2000). "Biochemical analysis of the Kruppel-associated box (KRAB) transcriptional repression domain." J Biol Chem **275**(24): 18000-10.

Peng, H., G. E. Begg, D. C. Schultz, J. R. Friedman, D. E. Jensen, D. W. Speicher and F. J. Rauscher, 3rd (2000). "Reconstitution of the KRAB-KAP-1 repressor complex: a model system for defining the molecular anatomy of RING-B box-coiled-coil domain-mediated protein-protein interactions." J Mol Biol **295**(5): 1139-62.

Petrascheck, M., D. Escher, T. Mahmoudi, C. P. Verrijzer, W. Schaffner and A. Barberis (2005). "DNA looping induced by a transcriptional enhancer in vivo." Nucleic Acids Res **33**(12): 3743-50.

Plath, K., J. Fang, S. K. Mlynarczyk-Evans, R. Cao, K. A. Worringer, H. Wang, C. C. de la Cruz, A. P. Otte, B. Panning and Y. Zhang (2003). "Role of histone H3 lysine 27 methylation in X inactivation." Science **300**(5616): 131-5.

Pleasure, S. J. and V. M. Lee (1993). "NTera 2 cells: a human cell line which displays characteristics expected of a human committed neuronal progenitor cell." J Neurosci Res **35**(6): 585-602.

Plon, S. E. and J. C. Wang (1986). "Transcription of the human beta-globin gene is stimulated by an SV40 enhancer to which it is physically linked but topologically uncoupled." Cell **45**(4): 575-80.

Ponger, L. and D. Mouchiroud (2002). "CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences." Bioinformatics **18**(4): 631-3.

Portnoy, M. E., K. J. McDermott, A. Antonellis, E. H. Margulies, A. B. Prasad, D. M. Kingsley, E. D. Green and D. P. Mortlock (2005). "Detection of potential GDF6 regulatory elements by multispecies sequence comparisons and identification of a skeletal joint enhancer." Genomics **86**(3): 295-305.

Prestridge, D. S. (2000). "Computer software for eukaryotic promoter analysis." Methods Mol Biol **130**: 265-95.

Promega. (2006). "Technical Resources; FAQ." from http://www.promega.com/faq/lucifer.html.

Proudfoot, N. J., A. Furger and M. J. Dye (2002). "Integrating mRNA processing with transcription." Cell **108**(4): 501-12.

Quentmeier, H., W. G. Dirks, R. A. Macleod, J. Reinhardt, M. Zaborski and H. G. Drexler (2004). "Expression of HOX genes in acute leukemia cell lines with and without MLL translocations." Leuk Lymphoma **45**(3): 567-74.

Rada-Iglesias, A., O. Wallerman, C. Koch, A. Ameur, S. Enroth, G. Clelland, K. Wester, S. Wilcox, O. M. Dovey, P. D. Ellis, V. L. Wraight, K. James, R. Andrews, C. Langford, P. Dhami, N. Carter, D. Vetrie, F. Ponten, J. Komorowski, I. Dunham and C. Wadelius (2005). "Binding sites for metabolic disease related transcription factors inferred at base pair resolution by chromatin immunoprecipitation and genomic microarrays." Hum Mol Genet **14**(22): 3435-47.

Rajendra, R., D. Malegaonkar, P. Pungaliya, H. Marshall, Z. Rasheed, J. Brownell, L. F. Liu, S. Lutzker, A. Saleem and E. H. Rubin (2004). "Topors functions as an E3 ubiquitin ligase with specific E2 enzymes and ubiquitinates p53." J Biol Chem **279**(35): 36440-4.

Rakha, E. A., S. E. Pinder, C. E. Paish and I. O. Ellis (2004). "Expression of the transcription factor CTCF in invasive breast cancer: a candidate gene located at 16q22.1." Br J Cancer **91**(8): 1591-6.

Recillas-Targa, F., M. J. Pikaart, B. Burgess-Beusse, A. C. Bell, M. D. Litt, A. G. West, M. Gaszner and G. Felsenfeld (2002). "Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities." Proc Natl Acad Sci U S A **99**(10): 6883-8.

Reese, M. G. (2001). "Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome." Comput Chem **26**(1): 51-6.

Ren, B., F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell and R. A. Young (2000). "Genome-wide location and function of DNA binding proteins." Science **290**(5500): 2306-9.

Rice, J. C., S. D. Briggs, B. Ueberheide, C. M. Barber, J. Shabanowitz, D. F. Hunt, Y. Shinkai and C. D. Allis (2003). "Histone methyltransferases direct different degrees of methylation to define distinct chromatin domains." Mol Cell **12**(6): 1591-8.

Richards, E. J. and S. C. Elgin (2002). "Epigenetic codes for heterochromatin formation and silencing: rounding up the usual suspects." Cell **108**(4): 489-500.

Sabo, P. J., M. S. Kuehn, R. Thurman, B. E. Johnson, E. M. Johnson, H. Cao, M. Yu, E. Rosenzweig, J. Goldy, A. Haydock, M. Weaver, A. Shafer, K. Lee, F. Neri, R. Humbert, M. A. Singer, T. A. Richmond, M. O. Dorschner, M. McArthur, M. Hawrylycz, R. D. Green, P. A. Navas, W. S. Noble and J. A. Stamatoyannopoulos (2006). "Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays." Nat Methods **3**(7): 511-8.

Sala, A., B. Saitta, P. De Luca, M. N. Cervellera, I. Casella, R. E. Lewis, R. Watson and C. Peschle (1999). "B-MYB transactivates its own promoter through SP1-binding sites." Oncogene **18**(6): 1333-9.

Sandelin, A., W. Alkema, P. Engstrom, W. W. Wasserman and B. Lenhard (2004). "JASPAR: an open-access database for eukaryotic transcription factor binding profiles." Nucleic Acids Res **32**(Database issue): D91-4.

Santos-Rosa, H., R. Schneider, A. J. Bannister, J. Sherriff, B. E. Bernstein, N. C. Emre, S. L. Schreiber, J. Mellor and T. Kouzarides (2002). "Active genes are tri-methylated at K4 of histone H3." Nature **419**(6905): 407-11.

Scaife, J. and J. R. Beckwith (1966). "Mutational alteration of the maximal level of Lac operon expression." Cold Spring Harb Symp Quant Biol **31**: 403-8.

Schaeffer, L., R. Roy, S. Humbert, V. Moncollin, W. Vermeulen, J. H. Hoeijmakers, P. Chambon and J. M. Egly (1993). "DNA repair helicase: a component of BTF2 (TFIIH) basic transcription factor." Science **260**(5104): 58-63.

Scherf, M., A. Klingenhoff and T. Werner (2000). "Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach." Journal of Molecular Biology **297**(3): 599-606.

Schoenherr, C. J. and D. J. Anderson (1995). "The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes." Science **267**(5202): 1360-3.

Schubeler, D., C. Francastel, D. M. Cimbora, A. Reik, D. I. Martin and M. Groudine (2000). "Nuclear localization and histone acetylation: a pathway for chromatin opening and transcriptional activation of the human beta-globin locus." Genes Dev **14**(8): 940-50.

Schubeler, D., M. Groudine and M. A. Bender (2001). "The murine beta-globin locus control region regulates the rate of transcription but not the hyperacetylation of histones at the active genes." Proc Natl Acad Sci U S A **98**(20): 11432-7.

Segars, J. H., T. Nagata, V. Bours, J. A. Medin, G. Franzoso, J. C. Blanco, P. D. Drew, K. G. Becker, J. An, T. Tang and et al. (1993). "Retinoic acid induction of major histocompatibility complex class I genes in NTera-2 embryonal carcinoma cells involves induction of NF-kappa B (p50-p65) and retinoic acid receptor beta-retinoid X receptor beta heterodimers." Mol Cell Biol **13**(10): 6157-69.

Seto, E., B. Lewis and T. Shenk (1993). "Interaction between transcription factors Sp1 and YY1." Nature **365**(6445): 462-4.

Shaw-Jackson, C. and T. Michiels (1999). "Absence of internal ribosome entry site-mediated tissue specificity in the translation of a bicistronic transgene." J Virol **73**(4): 2729-38.

Shaw, W. V. (1975). "Chloramphenicol acetyltransferase from chloramphenicol-resistant bacteria." Methods Enzymol **43**: 737-55.

Shi, Y., F. Lan, C. Matson, P. Mulligan, J. R. Whetstine, P. A. Cole, R. A. Casero and Y. Shi (2004). "Histone demethylation mediated by the nuclear amine oxidase homolog LSD1." Cell **119**(7): 941-53.

Shi, Y., J. S. Lee and K. M. Galvin (1997). "Everything you have ever wanted to know about Yin Yang 1." Biochim Biophys Acta **1332**(2): F49-66.

Shi, Y. J., C. Matson, F. Lan, S. Iwase, T. Baba and Y. Shi (2005). "Regulation of LSD1 histone demethylase activity by its associated factors." Mol Cell **19**(6): 857-64.

Shieh, B. and C. Li (2004). "Multi-faceted, multi-versatile microarray: simultaneous detection of many viruses and their expression profiles." Retrovirology **1**(1): 11.

Shiio, Y. and R. N. Eisenman (2003). "Histone sumoylation is associated with transcriptional repression." Proc Natl Acad Sci U S A **100**(23): 13225-30.

Shilatifard, A. (2006). "CHROMATIN MODIFICATIONS BY METHYLATION AND UBIQUITINATION: Implications in the Regulation of Gene Expression." Annu Rev Biochem **75**: 243-69.

Shogren-Knaak, M., H. Ishii, J. M. Sun, M. J. Pazin, J. R. Davie and C. L. Peterson (2006). "Histone H4-K16 acetylation controls chromatin structure and protein interactions." Science **311**(5762): 844-7.

Shrivastava, A. and K. Calame (1994). "An analysis of genes regulated by the multi-functional transcriptional regulator Yin Yang-1." Nucleic Acids Res **22**(24): 5151-5.

Simon, J. A. and J. W. Tamkun (2002). "Programming off and on states in chromatin: mechanisms of Polycomb and trithorax group complexes." Curr Opin Genet Dev **12**(2): 210-8.

Solomon, M. J., P. L. Larsen and A. Varshavsky (1988). "Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene." Cell **53**(6): 937-47.

Solomon, M. J. and A. Varshavsky (1985). "Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures." Proc Natl Acad Sci U S A **82**(19): 6470-4.

Solovyev, V. V. and I. A. Shahmuradov (2003). "PromH: Promoters identification using orthologous genomic sequences." Nucleic Acids Res **31**(13): 3540-5.

Song, J. S. (2004). "Activity of the human telomerase catalytic subunit (hTERT) gene promoter could be increased by the SV40 enhancer." Biosci Biotechnol Biochem **68**(8): 1634-9.

Song, J. S. (2005). "Enhanced expression of apoptin by the Myc-Max binding motif and SV40 enhancer for SCLC gene therapy." Biosci Biotechnol Biochem **69**(1): 51-5.

Spencer, C. C. A., P. Deloukas, S. Hunt, J. Mullikin, S. R. Myers, B. Silverman, P. Donnelly, D. Bentley and G. McVean (2006). "The influence of recombination on human genetic diversity." PLoS Genetics **preprint**(2006): e148.eor.

Spitz, F., F. Gonzalez and D. Duboule (2003). "A global control region defines a chromosomal regulatory landscape containing the HoxD cluster." Cell **113**(3): 405-17.

Srinivasan, L. and M. L. Atchison (2004). "YY1 DNA binding and PcG recruitment requires CtBP." Genes Dev **18**(21): 2596-601.

Stavrides, G. S. (2002). Human chromosome 20q12-13.2: Structural, comparative and sequence variation studies. Molecular Biology at
The Wellcome Trust Sanger Institute
Research degrees. Cambridge, UK, University of Cambridge. **PhD**.

Strachan, T. (2003). Nucleic Acid Hybridization Assays. Human Molecular Genetics, John Wiley & Sons.

Strachan, T. and A. Read (2003). Human Molecular Genetics, Taylor & Francis Group.

Strahl, B. D. and C. D. Allis (2000). "The language of covalent histone modifications." Nature **403**(6765): 41-5.

Suzuki, M., T. Yamada, F. Kihara-Negishi, T. Sakurai and T. Oikawa (2003). "Direct association between PU.1 and MeCP2 that recruits mSin3A-HDAC complex for PU.1-mediated transcriptional repression." Oncogene **22**(54): 8688-98.

Suzuki, Y. and S. Sugano (2001). "Construction of full-length-enriched cDNA libraries. The oligo-capping method." Methods Mol Biol **175**: 143-53.

Svejstrup, J. Q. (2004). "The RNA polymerase II transcription cycle: cycling through chromatin." Biochim Biophys Acta **1677**(1-3): 64-73.

Takai, D. and P. A. Jones (2002). "Comprehensive analysis of CpG islands in human chromosomes 21 and 22." Proc Natl Acad Sci U S A **99**(6): 3740-5.

Thanos, D. and T. Maniatis (1995). "Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome." Cell **83**(7): 1091-100.

Thomas, M. C. and C. M. Chiang (2006). "The general transcription machinery and general cofactors." Crit Rev Biochem Mol Biol **41**(3): 105-78.

Tolhuis, B., R. J. Palstra, E. Splinter, F. Grosveld and W. de Laat (2002). "Looping and interaction between hypersensitive sites in the active beta-globin locus." Mol Cell **10**(6): 1453-65.

Travers, A. A. (1989). "DNA conformation and protein binding." Annu Rev Biochem **58**: 427-52.

Triezenberg, S. J. (1995). "Structure and function of transcriptional activation domains." Curr Opin Genet Dev **5**(2): 190-6.

Trinklein, N. D., S. J. Aldred, A. J. Saldanha and R. M. Myers (2003). "Identification and functional analysis of human transcriptional promoters." Genome Res **13**(2): 308-12.

Trojer, P. and D. Reinberg (2006). "Histone lysine demethylases and their impact on epigenetics." Cell **125**(2): 213-7.

Tsonis, P. A. (2003). Anatomy of Gene Regulation: a three dimensional structural analysis. Cambridge, Cambridge University Press.

Turner, B. M. (2001). Chromatin and Gene Regulation. Iowa, Iowa State University Press.

Usheva, A. and T. Shenk (1996). "YY1 transcriptional initiator: protein interactions and association with a DNA site containing unpaired strands." Proc Natl Acad Sci U S A **93**(24): 13571-6.

Van, P. L., K. W. Yim, D. Y. Jin, G. Dapolito, A. Kurimasa and K. T. Jeang (2001). "Genetic evidence of a role for ATM in functional interaction between human T-cell leukemia virus type 1 Tax and p53." J Virol **75**(1): 396-407.

Varshavsky, A. J., O. Sundin and M. Bohn (1979). "A stretch of "late" SV40 viral DNA about 400 bp long which includes the origin of replication is specifically exposed in SV40 minichromosomes." Cell **16**(2): 453-66.

Vinogradov, A. E. (2003). "DNA helix: the importance of being GC-rich." Nucleic Acids Res **31**(7): 1838-44.

Vostrov, A. A. and W. W. Quitschke (1997). "The zinc finger protein CTCF binds to the APBbeta domain of the amyloid beta-protein precursor promoter. Evidence for a role in transcriptional activation." J Biol Chem **272**(52): 33353-9.

Wang, H., L. Wang, H. Erdjument-Bromage, M. Vidal, P. Tempst, R. S. Jones and Y. Zhang (2004). "Role of histone H2A ubiquitination in Polycomb silencing." Nature **431**(7010): 873-8.

Wang, H., L. Zhai, J. Xu, H. Y. Joo, S. Jackson, H. Erdjument-Bromage, P. Tempst, Y. Xiong and Y. Zhang (2006). "Histone H3 and H4 ubiquitylation by the CUL4-DDB-ROC1 ubiquitin ligase facilitates cellular response to DNA damage." Mol Cell **22**(3): 383-94.

Wang, J. C., M. K. Derynck, D. F. Nonaka, D. B. Khodabakhsh, C. Haqq and K. R. Yamamoto (2004). "Chromatin immunoprecipitation (ChIP) scanning identifies primary glucocorticoid receptor target genes." Proc Natl Acad Sci U S A **101**(44): 15603-8.

Waterborg, J. H. (1993). "Dynamic methylation of alfalfa histone H3." J Biol Chem **268**(7): 4918-21.

Waterston, R. H., K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S. E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M. R. Brent, D. G. Brown, S. D. Brown, C. Bult, J. Burton, J. Butler, R. D. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A. T. Chinwalla, D. M. Church, M. Clamp, C. Clee, F. S. Collins, L. L. Cook, R. R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. Daly, R. David, J. Davies, K. D. Delehaunty, J. Deri, E. T. Dermitzakis, C. Dewey, N. J. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D. M. Dunn, S. R. Eddy, L. Elnitski, R. D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G. A. Fewell, P. Flicek, K. Foley, W. N. Frankel, L. A. Fulton, R. S. Fulton, T. S. Furey, D. Gage, R. A. Gibbs, G. Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T. A. Graves, E. D. Green, S. Gregory, R. Guigo, M. Guyer, R. C. Hardison, D. Haussler, Y. Hayashizaki, L. W. Hillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. Jackson, D. B. Jaffe, L. S. Johnson, M. Jones, T. A. Jones, A. Joy, M. Kamal, E. K. Karlsson, D. Karolchik, A. Kasprzyk, J. Kawai, E. Keibler, C. Kells, W. J. Kent, A. Kirby, D. L. Kolbe, I. Korf, R. S. Kucherlapati, E. J. Kulbokas, D. Kulp, T. Landers, J. P. Leger, S. Leonard, I. Letunic, R. Levine, J. Li, M. Li, C. Lloyd, S. Lucas, B. Ma, D. R. Maglott, E. R. Mardis, L. Matthews, E. Mauceli, J. H. Mayer, M. McCarthy, W. R. McCombie, S. McLaren, K. McLay, J. D. McPherson, J. Meldrim, B. Meredith, J. P. Mesirov, W. Miller, T. L. Miner, E. Mongin, K. T. Montgomery, M. Morgan, R. Mott, J. C. Mullikin, D. M. Muzny, W. E. Nash, J. O. Nelson, M. N. Nhan, R. Nicol, Z. Ning, C. Nusbaum, M. J. O'Connor, Y. Okazaki, K. Oliver, E. Overton-Larty, L. Pachter, G. Parra, K. H. Pepin, J. Peterson, P. Pevzner, R. Plumb, C. S. Pohl, A. Poliakov, T. C. Ponce, C. P. Ponting, S. Potter, M. Quail, A. Reymond, B. A. Roe, K. M. Roskin, E. M. Rubin, A. G. Rust, R. Santos, V. Sapojnikov, B. Schultz, J. Schultz, M. S. Schwartz, S. Schwartz, C. Scott, S. Seaman, S. Searle, T. Sharpe, A. Sheridan, R. Shownkeen, S. Sims, J. B. Singer, G. Slater, A. Smit, D. R. Smith, B. Spencer, A. Stabenau, N. Stange-Thomann, C. Sugnet, M. Suyama, G. Tesler, J.

Thompson, D. Torrents, E. Trevaskis, J. Tromp, C. Ucla, A. Ureta-Vidal, J. P. Vinson, A. C. Von Niederhausern, C. M. Wade, M. Wall, R. J. Weber, R. B. Weiss, M. C. Wendl, A. P. West, K. Wetterstrand, R. Wheeler, S. Whelan, J. Wierzbowski, D. Willey, S. Williams, R. K. Wilson, E. Winter, K. C. Worley, D. Wyman, S. Yang, S. P. Yang, E. M. Zdobnov, M. C. Zody and E. S. Lander (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature **420**(6915): 520-62.

Weger, S., E. Hammer and R. Heilbronn (2005). "Topors acts as a SUMO-1 E3 ligase for p53 in vitro and in vivo." FEBS Lett **579**(22): 5007-12.

Weil, P. A., D. S. Luse, J. Segall and R. G. Roeder (1979). "Selective and accurate initiation of transcription at the Ad2 major late promotor in a soluble system dependent on purified RNA polymerase II and DNA." Cell **18**(2): 469-84.

Weinmann, A. S., S. M. Bartley, T. Zhang, M. Q. Zhang and P. J. Farnham (2001). "Use of chromatin immunoprecipitation to clone novel E2F target promoters." Mol Cell Biol **21**(20): 6820-32.

Weintraub, H. and M. Groudine (1976). "Chromosomal subunits in active genes have an altered conformation." Science **193**(4256): 848-56.

Weintraub, S. J., K. N. Chow, R. X. Luo, S. H. Zhang, S. He and D. C. Dean (1995). "Mechanism of active transcriptional repression by the retinoblastoma protein." Nature **375**(6534): 812-5.

Weiss, S. and L. A. Gladstone (1959). "A mammalian system for the incorporation of cytidine triphosphate into ribonucleic acid." J. Am. Chem. Soc. **81**: 4118-4119.

Wells, J. and P. J. Farnham (2002). "Characterizing transcription factor binding sites using formaldehyde crosslinking and immunoprecipitation." Methods **26**(1): 48-56.

West, A. G., M. Gaszner and G. Felsenfeld (2002). "Insulators: many functions, many mechanisms." Genes Dev **16**(3): 271-88.

Wheeler, D. L., D. M. Church, S. Federhen, A. E. Lash, T. L. Madden, J. U. Pontius, G. D. Schuler, L. M. Schriml, E. Sequeira, T. A. Tatusova and L. Wagner (2003). "Database resources of the National Center for Biotechnology." Nucleic Acids Res **31**(1): 28-33.

Whetstine, J. R., A. Nottke, F. Lan, M. Huarte, S. Smolikov, Z. Chen, E. Spooner, E. Li, G. Zhang, M. Colaiacovo and Y. Shi (2006). "Reversal of histone lysine trimethylation by the JMJD2 family of histone demethylases." Cell **125**(3): 467-81.

Wieczorek, E., M. Brand, X. Jacq and L. Tora (1998). "Function of TAF(II)-containing complex without TBP in transcription by RNA polymerase II." Nature **393**(6681): 187-91.

Wirth, J., T. Wagner, J. Meyer, R. A. Pfeiffer, H. U. Tietze, W. Schempp and G. Scherer (1996). "Translocation breakpoints in three patients with campomelic dysplasia and autosomal sex reversal map more than 130 kb from SOX9." Hum Genet **97**(2): 186-93.

Wood, A., J. Schneider, J. Dover, M. Johnston and A. Shilatifard (2005). "The Bur1/Bur2 complex is required for histone H2B monoubiquitination by Rad6/Bre1 and histone methylation by COMPASS." Mol Cell **20**(4): 589-99.

Woolfe, A., M. Goodson, D. K. Goode, P. Snell, G. K. McEwen, T. Vavouri, S. F. Smith, P. North, H. Callaway, K. Kelly, K. Walter, I. Abnizova, W. Gilks, Y. J. Edwards, J. E. Cooke and G. Elgar (2005). "Highly conserved non-coding sequences are associated with vertebrate development." PLoS Biol **3**(1): e7.

Wu, X., Y. Li, B. Crise, S. M. Burgess and D. J. Munroe (2005). "Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses." J Virol **79**(8): 5211-4.

Xiao, B., C. Jing, J. R. Wilson, P. A. Walker, N. Vasisht, G. Kelly, S. Howell, I. A. Taylor, G. M. Blackburn and S. J. Gamblin (2003). "Structure and catalytic mechanism of the human histone methyltransferase SET7/9." Nature **421**(6923): 652-6.

Xiao, B., J. R. Wilson and S. J. Gamblin (2003). "SET domains and histone methylation." Curr Opin Struct Biol **13**(6): 699-705.

Xiao, H., A. Pearson, B. Coulombe, R. Truant, S. Zhang, J. L. Regier, S. J. Triezenberg, D. Reinberg, O. Flores, C. J. Ingles and et al. (1994). "Binding of basal transcription factor TFIIH to the acidic activation domains of VP16 and p53." Mol Cell Biol **14**(10): 7013-24.

Xiao, T., C. F. Kao, N. J. Krogan, Z. W. Sun, J. F. Greenblatt, M. A. Osley and B. D. Strahl (2005). "Histone H2B ubiquitylation is associated with elongating RNA polymerase II." Mol Cell Biol **25**(2): 637-51.

Xie, Z. and D. D. Bikle (1997). "Cloning of the human phospholipase C-gamma1 promoter and identification of a DR6-type vitamin D-responsive element." J Biol Chem **272**(10): 6573-7.

Yamaguchi, M., Y. Obata and A. Matsukage (1989). "Paradoxical effect of Simian virus 40 enhancer on the function of mouse DNA polymerase beta gene promoter." Nucleic Acids Res **17**(10): 3725-34.

Yoder, J. A., C. P. Walsh and T. H. Bestor (1997). "Cytosine methylation and the ecology of intragenomic parasites." Trends Genet **13**(8): 335-40.

Yokoyama, C., T. Yabuki, H. Inoue, Y. Tone, S. Hara, T. Hatae, M. Nagata, E. I. Takahashi and T. Tanabe (1996). "Human gene encoding prostacyclin synthase (PTGIS): genomic organization, chromosomal localization, and promoter activity." Genomics **36**(2): 296-304.

Yordy, J. S. and R. C. Muise-Helmericks (2000). "Signal transduction and the Ets family of transcription factors." Oncogene **19**(55): 6503-13.

Yuan, W. (2000). "Intron 1 rather than 5' flanking sequence mediates cell type-specific expression of c-myb at level of transcription elongation." Biochim Biophys Acta **1490**(1-2): 74-86.

Yusufzai, T. M. and G. Felsenfeld (2004). "The 5'-HS4 chicken beta-globin insulator is a CTCF-dependent nuclear matrix-associated element." Proc Natl Acad Sci U S A **101**(23): 8620-4.

Zeng, L. and M. M. Zhou (2002). "Bromodomain: an acetyl-lysine binding domain." FEBS Lett **513**(1): 124-8.

Zhang, Y. (2003). "Transcriptional regulation by histone ubiquitination and deubiquitination." Genes Dev **17**(22): 2733-40.

URL1 (2004) http://arapaho.nsuok.edu/~biology/Tutorials/HelixLH.htm
URL2 (2004) http://arapaho.nsuok.edu/~biology/Tutorials/HelixTH.htm

APPENDIX A

Vega Gene ID	HUGO Gene ID	Gene Description	Summary
OTTHUMG00000033045	ACOT8	acyl-CoA thioesterase 8	The protein encoded by this gene is a peroxisomal thioesterase that appears to be involved more in the oxidation of fatty acids rather than in their formation. The encoded protein can bind to the human immunodeficiency virus-1 protein Nef, and mediate Nef-induced down-regulation of CD4 in T-cells. Multiple transcript variants encoding several different isoforms have been found for this gene.
OTTHUMG00000033081	ADA	adenosine deaminase	Adenosine deaminase catalyzes the hydrolysis of adenosine to inosine. ADA deficiency causes one form of severe combined immunodeficiency disease (SCID), in which there is dysfunction of both B and T lymphocytes with impaired cellular immunity and decreased production of immunoglobulins.
OTTHUMG00000032687	ARFGEF2	ADP-ribosylation factor guanine nucleotide-exchange factor 2 (brefeldin A-inhibited)	ADP-ribosylation factors (ARFs) play an important role in intracellular vesicular trafficking. The protein encoded by this gene is involved in the activation of ARFs by accelerating replacement of bound GDP with GTP and is involved in Golgi transport. It contains a Sec7 domain, which may be responsible for its guanine-nucleotide exchange activity and also brefeldin A inhibition.
OTTHUMG00000033086	B4GALT5	UDP-Gal:betaGlcNAc beta 1,4-galactosyltransferase, polypeptide 5	This gene is one of seven beta-1,4-galactosyltransferase (beta4GalT) genes. They encode type II membrane-bound glycoproteins that appear to have exclusive specificity for the donor substrate UDP-galactose; all transfer galactose in a beta1,4 linkage to similar acceptor sugars: GlcNAc, Glc, and Xyl. Each beta4GalT has a distinct function in the biosynthesis of different glycoconjugates and saccharide structures. As type II membrane proteins, they have an N-terminal hydrophobic signal sequence that directs the protein to the Golgi apparatus and which then remains uncleaved to function as a transmembrane anchor. By sequence similarity, the beta4GalTs form four groups: beta4GalT1 and beta4GalT2, beta4GalT3 and beta4GalT4, beta4GalT5 and beta4GalT6, and beta4GalT7. The function of the enzyme encoded by this gene is not clear. This gene was previously designated as B4GALT4 but was renamed to B4GALT5. In the literature it is also referred to as beta4GalT2.
OTTHUMG00000032582	C20orf10	chromosome 20 open reading frame 10	
OTTHUMG00000032517	C20orf100	chromosome 20 open reading frame 100	
OTTHUMG00000032518	C20orf111	chromosome 20 open reading frame 111	
OTTHUMG00000032553	C20orf119	chromsome 20 open reading frame 119	
OTTHUMG00000032536	C20orf121	chromosome 20 open reading frame 121	
OTTHUMG00000032652	C20orf123	chromosome 20 open reading frame 123	
OTTHUMG00000032522	C20orf142	chromosome 20 open reading frame 142	
OTTHUMG00000032647	C20orf157	chromosome 20 open reading frame 157	

OTTHUMG00000032624	SNX21 (C20ORF161)	sorting nexin family member 21	This gene encodes a member of the sorting nexin family. Members of this family contain a phox (PX) domain, which is a phosphoinositide binding domain, and are involved in intracellular trafficking. This protein does not contain a coiled coil region, like some family members. The specific function of this protein has not been determined. Multiple transcript variants encoding distinct isoforms have been identified for this gene.
OTTHUMG00000032628	C20orf165	chromosome 20 open reading frame 165	
OTTHUMG00000032611	C20orf168	chromosome 20 open reading frame 168	
OTTHUMG00000032576	DBNDD2 (C20ORF35)	dysbindin (dystrobrevin binding protein 1) domain containing 2	
OTTHUMG00000032533	C20orf62	chromosome 20 open reading frame 62	
OTTHUMG00000032510	FAM112A (RP5-10496-16.1)	family with sequence similarity 112, member A	
OTTHUMG00000032635	C20orf67	Chromosome 20 open reading frame 67	
OTTHUMG00000033053	CD40	CD40 molecule, TNF receptor superfamily member 5	The protein encoded by this gene is a member of the TNF-receptor superfamily. This receptor has been found to be essential in mediating a broad variety of immune and inflammatory responses including T cell-dependent immunoglobulin class switching, memory B cell development, and germinal center formation. AT-hook transcription factor AKNA is reported to coordinately regulate the expression of this receptor and its ligand, which may be important for homotypic cell interactions. Adaptor protein TNFR2 interacts with this receptor and serves as a mediator of the signal transduction. The interaction of this receptor and its ligand is found to be necessary for amyloid-beta-induced microglial activation, and thus is thought to be an early event in Alzheimer disease pathogenesis. Two alternatively spliced transcript variants of this gene encoding distinct isoforms have been reported.
OTTHUMG00000033073	CDH22	cadherin-like 22	This gene is a member of the cadherin superfamily. The gene product is composed of five cadherin repeat domains and a cytoplasmic tail similar to the highly conserved cytoplasmic region of classical cadherins. Expressed predominantly in the brain, this putative calcium-dependent cell adhesion protein may play an important role in morphogenesis and tissue formation in neural and non-neural cells during development and maintenance of the brain and neuroendocrine organs.
OTTHUMG00000032487	CHD6	chromodomain helicase DNA binding protein 6	Chromosomal DNA of eukaryotic cells is compacted by nuclear proteins to form chromatin, an organized nucleoprotein structure that can inhibit gene expression. Several multisubunit protein complexes exist to remodel the chromatin to allow patterns of cell type-specific gene expression. The protein encoded by this gene is thought to be a core member of one or more of these complexes. The encoded protein, which is a member of the SNF2/RAD54 helicase family, contains two chromodomains, a helicase domain, and an ATPase domain.

OTTHUMG00000033046	CSE1L	CSE1 chromosome segregation 1-like (yeast)	Proteins that carry a nuclear localization signal (NLS) are transported into the nucleus by the importin-alpha/beta heterodimer. Importin-alpha binds the NLS, while importin-beta mediates translocation through the nuclear pore complex. After translocation, RanGTP binds importin-beta and displaces importin-alpha. Importin-alpha must then be returned to the cytoplasm, leaving the NLS protein behind. The protein encoded by this gene binds strongly to NLS-free importin-alpha, and this binding is released in the cytoplasm by the combined action of RANBP1 and RANGAP1. In addition, the encoded protein may play a role both in apoptosis and in cell proliferation.
OTTHUMG00000033072	DDX27	DEAD (Asp-Glu-Ala-Asp) box polypeptide 27	DEAD box proteins, characterized by the conserved motif Asp-Glu-Ala-Asp (DEAD), are putative RNA helicases. They are implicated in a number of cellular processes involving alteration of RNA secary structure such as translation initiation, nuclear and mitochondrial splicing, and ribosome and spliceosome assembly. Based on their distribution patterns, some members of this family are believed to be involved in embryogenesis, spermatogenesis, and cellular growth and division. This gene encodes a DEAD box protein, the function of which has not been determined.
OTTHUMG00000032610	DNTTIP1	deoxynucleotidyltransferase, terminal, interacting protein 1	
OTTHUMG00000033070	ELMO2	engulfment and cell motility 2	The protein encoded by this gene interacts with the dedicator of cyto-kinesis 1 protein. Similarity to a C. elegans protein suggests that this protein may function in phagocytosis of apoptotic cells and in cell migration. Alternative splicing results in multiple transcript variants encoding the same protein.
OTTHUMG00000046304	EMILIN3	elastin microfibril interfacer 3	
OTTHUMG00000033041	EYA2	eyes absent homolog 2 (Drosophila)	This gene encodes a member of the eyes absent (EYA) family of proteins. The encoded protein may be post-translationally modified and may play a role in eye development. A similar protein in mice can act as a transcriptional activator. Five transcript variants encoding three distinct isoforms have been identified for this gene.
OTTHUMG00000032530	GDAP1L1	ganglioside-induced differentiation-associated protein 1-like 1	The ganglioside GD3 synthase causes cell differentiation with neurite sprouting when transfected into the mouse neuroblastoma cell line Neuro2a. After differentiation, the expression of several genes is upregulated, including one that encodes a protein termed ganglioside-induced differentiation-associated protein 1 (Gdap1). A similar gene was found in humans, and mutations in the human gene are associated with Charcot-Marie-Tooth type 4A disease. The protein encoded by this gene is similar in sequence to the human GDAP1 protein.
OTTHUMG00000032531	HNF4A	hepatocyte nuclear factor 4, alpha	The protein encoded by this gene is a nuclear transcription factor which binds DNA as a homodimer. The encoded protein controls the expression of several genes, including hepatocyte nuclear factor 1 alpha, a transcription factor which regulates the expression of several hepatic genes. This gene may play a role in development of the liver, kidney, and intestines. Mutations in this gene have been associated with monogenic autosomal dominant non-insulin-dependent diabetes mellitus type I. Alternative splicing of this

			gene results in multiple transcript variants.
OTTHUMG00000032513	IFT52	intraflagellar transport 52 homolog (Chlamydomonas)	
OTTHUMG00000033037	JPH2	junctophilin 2	Junctional complexes between the plasma membrane and endoplasmic/sarcoplasmic reticulum are a common feature of all excitable cell types and mediate cross talk between cell surface and intracellular ion channels. The protein encoded by this gene is a component of junctional complexes and is composed of a C-terminal hydrophobic segment spanning the endoplasmic/sarcoplasmic reticulum membrane and a remaining cytoplasmic domain that shows specific affinity for the plasma membrane. This gene is a member of the junctophilin gene family. Alternative splicing has been observed at this locus and two variants encoding distinct isoforms are described.
OTTHUMG00000033051	KCNB1	potassium voltage-gated channel, Shab-related subfamily, member 1	Voltage-gated potassium (Kv) channels represent the most complex class of voltage-gated ion channels from both functional and structural standpoints. Their diverse functions include regulating neurotransmitter release, heart rate, insulin secretion, neuronal excitability, epithelial electrolyte transport, smooth muscle contraction, and cell volume. Four sequence-related potassium channel genes - shaker, shaw, shab, and shal - have been identified in Drosophila, and each has been shown to have human homolog(s). This gene encodes a member of the potassium channel, voltage-gated, shab-related subfamily. This member is a delayed rectifier potassium channel and its activity is modulated by some other family members.
OTTHUMG00000032544	KCNK15	potassium channel, subfamily K, member 15	This gene encodes one of the members of the superfamily of potassium channel proteins containing two pore-forming P domains. The product of this gene has not been shown to be a functional channel, however, it may require other non-pore-forming proteins for activity.
OTTHUMG00000033079	KCNS1	potassium voltage-gated channel, delayed-rectifier, subfamily S, member 1	Voltage-gated potassium channels form the largest and most diversified class of ion channels and are present in both excitable and nonexcitable cells. Their main functions are associated with the regulation of the resting membrane potential and the control of the shape and frequency of action potentials. The alpha subunits are of 2 types: those that are functional by themselves and those that are electrically silent but capable of modulating the activity of specific functional alpha subunits. The protein encoded by this gene is not functional by itself but can form heteromultimers with member 1 and with member 2 (and possibly other members) of the Shab-related subfamily of potassium voltage-gated channel proteins. This gene belongs to the S subfamily of the potassium channel family.
OTTHUMG00000032696	ZNFX1 (KIAA1404)	zinc finger, NFX1-type containing 1	
OTTHUMG00000032503	L3MBTL	l(3)mbt-like (Drosophila)	This gene encodes the homolog of a protein identified in Drosophila as a suppressor of malignant transformation of neuroblasts and ganglion-mother cells in the optic centers of the brain. This gene product is localized to condensed chromosomes in mitotic cells.

			Overexpression of this gene in a glioma cell line results in improper nuclear segregation and cytokinesis producing multinucleated cells. Two transcripts have been identified for this gene.
OTTHUMG00000033056	LPIN3	lipin 3	Humans lipodystrophy is characterized by loss of body fat, fatty liver, hypertriglyceridemia, and insulin resistance. Mice carrying mutations in the fatty liver dystrophy (fld) gene have similar phenotypes. Through positional cloning, the mouse gene responsible for fatty liver dystrophy was isolated and designated Lpin1. The nuclear protein encoded by Lpin1 was named lipin. Lpin1 mRNA was expressed at high levels in adipose tissue and was induced during differentiation of preadipocytes. These results indicated that lipin is required for normal adipose tissue development and provided a candidate gene for human lipodystrophy. Through database searches, mouse and human EST and genomic sequences with similarities to Lpin1 were identified. These included two related mouse genes (Lpin2 and Lpin3) and three human homologs (LPIN1, LPIN2, and LPIN3). Human LPIN1 gene has been mapped to 2p25.; linkages of fat mass and serum leptin levels to this same region have been noted. Human LPIN2 and LPIN3 mapped to chromosomes 1
OTTHUMG00000033052	MAFB	v-maf musculoaponeurotic fibrosarcoma oncogene homolog B (avian)	The protein encoded by this gene is a basic leucine zipper (bZIP) transcription factor that plays an important role in the regulation of lineage-specific hematopoiesis. The encoded nuclear protein represses ETS1-mediated transcription of erythroid-specific genes in myeloid cells. This gene contains no introns.
OTTHUMG00000033043	MATN4	matrilin 4	This gene encodes a member of von Willebrand factor A domain containing protein family. The proteins of this family are thought to be involved in the formation of filamentous networks in the extracellular matrices of various tissues. The specific function of this gene product has not yet been determined. Three alternatively spliced variants have been described.
OTTHUMG00000033044	MMP9	matrix metalloproteinase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV collagenase)	Proteins of the matrix metalloproteinase (MMP) family are involved in the breakdown of extracellular matrix in normal physiological processes, such as embryonic development, reproduction, and tissue remodeling, as well as in disease processes, such as arthritis and metastasis. Most MMP's are secreted as inactive proproteins which are activated when cleaved by extracellular proteinases. The enzyme encoded by this gene degrades type IV and V collagens. Studies in rhesus monkeys suggest that the enzyme is involved in IL-8-induced mobilization of hematopoietic progenitor cells from bone marrow, and murine studies suggest a role in tumor-associated tissue remodeling.
OTTHUMG00000033062	MYBL2	v-myb myeloblastosis viral oncogene homolog (avian)-like 2	The protein encoded by this gene, a member of the MYB family of transcription factor genes, is a nuclear protein involved in cell cycle progression. The encoded protein is phosphorylated by cyclin A/cyclin-dependent kinase 2 during the S-phase of the cell cycle and possesses both activator and repressor activities. It has been shown to activate the cell division cycle 2, cyclin D1, and insulin-like growth factor-binding protein 5 genes. Transcript variants may exist for this gene, but their full-length natures have not

			been determined.
OTTHUMG00000033061	NCOA3	nuclear receptor coactivator 3	The protein encoded by this gene is a nuclear receptor coactivator that interacts with nuclear hormone receptors to enhance their transcriptional activator functions. The encoded protein has histone acetyltransferase activity and recruits p300/CBP-associated factor and CREB binding protein as part of a multisubunit coactivation complex. This protein is initially found in the cytoplasm but is translocated into the nucleus upon phosphorylation. Two transcript variants encoding different isoforms have been found for this gene. In addition, a polymorphic repeat region is found in the C-terminus of the encoded protein.
OTTHUMG00000032639	NCOA5	nuclear receptor coactivator 5	This gene encodes a coregulator for the alpha and beta estrogen receptors and the orphan nuclear receptor NR1D2. The protein localizes to the nucleus, and is thought to have both coactivator and corepressor functions. Its interaction with nuclear receptors is independent of the AF2 domain on the receptors, which is known to regulate interaction with other coreceptors. Two alternatively spliced transcript variants for this gene have been described. However, the full length nature of one of the variants has not been determined.
OTTHUMG00000032626	NEURL2	neuralized-like 2 (Drosophila)	
OTTHUMG00000032567	PI3	protease inhibitor 3, skin-derived (SKALP)	This gene encodes an elastase-specific inhibitor, which contains a WAP-type four-disulfide core (WFDC) domain, and is thus a member of the WFDC domain family. Most WFDC gene members are localized to chromosome 20q12-q13 in two clusters: centromeric and telomeric. This gene belongs to the centromeric cluster.
OTTHUMG00000032574	PIGT	phosphatidylinositol glycan anchor biosynthesis, class T	This gene encodes a protein that is involved in glycosylphosphatidylinositol (GPI)-anchor biosynthesis. The GPI-anchor is a glycolipid found on many blood cells and serves to anchor proteins to the cell surface. This protein is an essential component of the multisubunit enzyme, GPI transamidase. GPI transamidase mediates GPI anchoring in the endoplasmic reticulum, by catalyzing the transfer of fully assembled GPI units to proteins.
OTTHUMG00000033065	PKIG	protein kinase (cAMP-dependent, catalytic) inhibitor gamma	The protein encoded by this gene is a member of the cAMP-dependent protein kinase (PKA) inhibitor family. Studies of a similar protein in mice suggest that this protein acts as a potent competitive PKA inhibitor, and is a predominant form of PKA inhibitors in various tissues. Three alternatively spliced transcript variants encoding the same protein have been reported.
OTTHUMG00000033082	PLCG1	phospholipase C, gamma 1	The protein encoded by this gene catalyzes the formation of inositol 1,4,5-trisphosphate and diacylglycerol from phosphatidylinositol 4,5-bisphosphate. This reaction uses calcium as a cofactor and plays an important role in the intracellular transduction of receptor-mediated tyrosine kinase activators. For example, when activated by SRC, the encoded protein causes the Ras guanine nucleotide exchange factor RasGRP1 to translocate to the Golgi, where it activates Ras. Also, this protein has been shown to be a major substrate for heparin-binding growth factor 1 (acidic fibroblast growth factor)-

			activated tyrosine kinase. Two transcript variants encoding different isoforms have been found for this gene.
OTTHUMG00000033047	PLTP	phospholipid transfer protein	The protein encoded by this gene is one of at least two lipid transfer proteins found in human plasma. The encoded protein transfers phospholipids from triglyceride-rich lipoproteins to high density lipoprotein (HDL). In addition to regulating the size of HDL particles, this protein may be involved in cholesterol metabolism. At least two transcript variants encoding different isoforms have been found for this gene.
OTTHUMG00000033078	PPGB	protective protein for beta-galactosidase (galactosialidosis)	This gene encodes a glycoprotein which associates with lysosomal enzymes beta-galactosidase and neuraminidase to form a complex of high molecular weight multimers. The formation of this complex provides a protective role for stability and activity. Deficiencies in this gene are linked to multiple forms of galactosialidosis.
OTTHUMG00000032667	PRKCBP1	protein kinase C binding protein 1	The protein encoded by this gene is a receptor for activated C-kinase (RACK) protein. The encoded protein has been shown to bind in vitro to activated protein kinase C beta I. In addition, this protein is a cutaneous T-cell lymphoma-associated antigen. Finally, the protein contains a bromodomain and two zinc fingers, and is thought to be a transcriptional regulator. Multiple transcript variants encoding several different isoforms have been found for this gene.
OTTHUMG00000033077	PTGIS	prostaglandin I2 (prostacyclin) synthase	This gene encodes a member of the cytochrome P450 superfamily of enzymes. The cytochrome P450 proteins are monooxygenases which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids. However, this protein is considered a member of the cytochrome P450 superfamily on the basis of sequence similarity rather than functional similarity. This endoplasmic reticulum membrane protein catalyzes the conversion of prostglandin H2 to prostacyclin (prostaglandin I2), a potent vasodilator and inhibitor of platelet aggregation. An imbalance of prostacyclin and its physiological antagonist thromboxane A2 contribute to the development of myocardial infarction, stroke, and atherosclerosis.
OTTHUMG00000033040	PTPRT	protein tyrosine phosphatase, receptor type, T	The protein encoded by this gene is a member of the protein tyrosine phosphatase (PTP) family. PTPs are known to be signaling molecules that regulate a variety of cellular processes including cell growth, differentiation, mitotic cycle, and oncogenic transformation. This PTP possesses an extracellular region, a single transmembrane region, and two tandem intracellular catalytic domains, and thus represents a receptor-type PTP. The extracellular region contains a meprin-A5 antigen-PTP (MAM) domain, Ig-like and fibronectin type III-like repeats. The protein domain structure and the expression pattern of the mouse counterpart of this PTP suggest its roles in both signal transduction and cellular adhesion in the central nervous system. Two alternatively spliced transcript variants of this gene, which encode distinct proteins, have been reported.
OTTHUMG00000032524	R3HDML	R3H domain (binds single-stranded nucleic acids) containing-like	

OTTHUMG00000033055	RBPSUHL	recombining binding protein suppressor of hairless (Drosophila)-like	In mouse, recombining binding protein L (RBP-L) is a transcription factor that binds to DNA sequences almost identical to that bound by the Notch receptor signalling pathway transcription factor RBP-J. However, unlike RBP-J, RBP-L does not interact with Notch receptors. RBP-L has been shown to activate transcription in concert with Epstein-Barr virus nuclear antigen-2 (EBNA2). The protein encoded by this gene is similar in sequence to the mouse RPB-L protein and Drosophila suppressor of hairless protein.
OTTHUMG00000032546	RIMS4	regulating synaptic membrane exocytosis 4	
OTTHUMG00000032685	PREX1 (RP11-269H4.1)	phosphatidylinositol 3,4,5-trisphosphate-dependent RAC exchanger 1	The protein encoded by this gene acts as a guanine nucleotide exchange factor for the RHO family of small GTP-binding proteins (RACs). It has been shown to bind to and activate RAC1 by exchanging bound GDP for free GTP. The encoded protein, which is found mainly in the cytoplasm, is activated by phosphatidylinositol-3,4,5-trisphosphate and the beta-gamma subunits of heterotrimeric G proteins.
OTTHUMG00000032675	SULF2	sulfatase 2	Heparan sulfate proteoglycans (HSPGs) act as coreceptors for numerous heparin-binding growth factors and cytokines and are involved in cell signaling. Heparan sulfate 6-O-endosulfatases, such as SULF2, selectively remove 6-O-sulfate groups from heparan sulfate. This activity modulates the effects of heparan sulfate by altering binding sites for signaling molecules (Dai et al., 2005).[supplied by OMIM]
OTTHUMG00000033083	SDC4	syndecan 4 (amphiglycan, ryudocan)	The protein encoded by this gene is a transmembrane (type I) heparan sulfate proteoglycan that functions as a receptor in intracellular signaling. The encoded protein is found as a homodimer and is a member of the syndecan proteoglycan family. This gene is found on chromosome 20, while a pseudogene has been found on chromosome 22.
OTTHUMG00000032565	SEMG1	semenogelin I	The protein encoded by this gene is the predominant protein in semen. The encoded secreted protein is involved in the formation of a gel matrix that encases ejaculated spermatozoa. The prostate-specific antigen (PSA) protease processes this protein into smaller peptides, with each possibly having a separate function. The proteolysis process breaks down the gel matrix and allows the spermatozoa to move more freely. Two transcript variants encoding different isoforms have been found for this gene.
OTTHUMG00000032566	SEMG2	semenogelin II	The secreted protein encoded by this gene is involved in the formation of a gel matrix that encases ejaculated spermatozoa. Proteolysis by the prostate-specific antigen (PSA) breaks down the gel matrix and allows the spermatozoa to move more freely. The encoded protein is found in lesser abundance than a similar semenogelin protein. The genes encoding these two semenogelin proteins are found in a cluster on chromosome 20.
OTTHUMG00000032502	SFRS6	splicing factor, arginine/serine-rich 6	The protein encoded by this gene is involved in mRNA splicing and may play a role in the determination of alternative splicing. The encoded nuclear protein belongs to the splicing factor SR family and has been shown to bind with and modulate another

			member of the family, SFRS12.
OTTHUMG00000033054	SGK2	serum/glucocorticoid regulated kinase 2	This gene encodes a serine/threonine protein kinase. Although this gene product is similar to serum- and glucocorticoid-induced protein kinase (SGK), this gene is not induced by serum or glucocorticoids. This gene is induced in response to signals that activate phosphatidylinositol 3-kinase, which is also true for SGK. Two alternate transcripts encoding two different isoforms have been described.
OTTHUMG00000032638	SLC12A5	solute carrier family 12, (potassium-chloride transporter) member 5	K-Cl cotransporters are proteins that lower intracellular chloride concentrations below the electrochemical equilibrium potential. The protein encoded by this gene is an integral membrane K-Cl cotransporter that can function in either a net efflux or influx pathway, depending on the chemical concentration gradients of potassium and chloride. The encoded protein can act as a homomultimer, or as a heteromultimer with other K-Cl cotransporters, to maintain chloride homeostasis in neurons.
OTTHUMG00000033042	SLC13A3	solute carrier family 13 (sodium-dependent dicarboxylate transporter), member 3	Mammalian sodium-dicarboxylate cotransporters transport succinate and other Krebs cycle intermediates. They fall into 2 categories based on their substrate affinity: low affinity and high affinity. Both the low- and high-affinity transporters play an important role in the handling of citrate by the kidneys. The protein encoded by this gene represents the high-affinity form. Alternatively spliced transcript variants encoding different isoforms have been found for this gene, although the full-length nature of some of them have not been characterized yet.
OTTHUMG00000032657	SLC2A10	solute carrier family 2 (facilitated glucose transporter), member 10	SLC2A10 is a member of the facilitative glucose transporter family, which plays a significant role in maintaining glucose homeostasis.[supplied by OMIM]
OTTHUMG00000033050	SLC35C2	solute carrier family 35, member C2	Oxygenation levels play an important role in the regulation of cellular invasiveness which occurs during early implantation when the trophoblast cells invade the uterus as well as during tumour progression and metastasis. This gene, which is regulated by oxygen tension, is induced in hypoxic trophoblast cells and is overexpressed in ovarian cancer. Two protein isoforms are encoded by transcript variants of this gene.
OTTHUMG00000032710	SLC9A8	solute carrier family 9 (sodium\/hydrogen exchanger), isoform 8	
OTTHUMG00000033075	SLPI	secretory leukocyte peptidase inhibitor	This gene encodes a secreted inhibitor which protects epithelial tissues from serine proteases. It is found in various secretions including seminal plasma, cervical mucus, and bronchial secretions, and has affinity for trypsin, leukocyte elastase, and cathepsin G. Its inhibitory effect contributes to the immune response by protecting epithelial surfaces from attack by endogenous proteolytic enzymes; the protein is also thought to have broad-spectrum anti-biotic activity.
OTTHUMG00000033048	SNAI1	snail homolog 1 (Drosophila)	The Drosophila embryonic protein snail is a zinc finger transcriptional repressor which downregulates the expression of ectodermal genes within the mesoderm. The nuclear protein encoded by this gene is structurally similar to the Drosophila snail protein, and is also thought to be critical for mesoderm formation in the developing embryo. At least two variants of a similar processed pseudogene have been found on chromosome 2.

OTTHUMG00000032704	SPATA2	spermatogenesis associated 2	
OTTHUMG00000032588	SPINLW1	serine protease inhibitor-like, with Kunitz and WAP domains 1 (eppin)	This gene encodes an epididymal protease inhibitor, which contains both kunitz-type and WAP-type four-disulfide core (WFDC) protease inhibitor consensus sequences. Most WFDC genes are localized to chromosome 20q12-q13 in two clusters: centromeric and telomeric. This gene is a member of the WFDC gene family and belongs to the telomeric cluster. Alternatively spliced transcript variants encoding distinct isoforms have been found for this gene.
OTTHUMG00000032585	SPINT3	serine protease inhibitor, Kunitz type, 3	
OTTHUMG00000032604	SPINT4	chromosome 20 open reading frame 137	
OTTHUMG00000032691	STAU1	staufen, RNA binding protein, homolog 1 (Drosophila)	Staufen is a member of the family of double-stranded RNA (dsRNA)-binding proteins involved in the transport and/or localization of mRNAs to different subcellular compartments and/or organelles. These proteins are characterized by the presence of multiple dsRNA-binding domains which are required to bind RNAs having double-stranded secary structures. The human homologue of staufen encoded by STAU, in addition contains a microtubule- binding domain similar to that of microtubule-associated protein 1B, and binds tubulin. The STAU gene product has been shown to be present in the cytoplasm in association with the rough endoplasmic reticulum (RER), implicating this protein in the transport of mRNA via the microtubule network to the RER, the site of translation. Five transcript variants resulting from alternative splicing of STAU gene and encoding three isoforms have been described. Three of these variants encode the same isoform, however, differ in their 5'UTR.
OTTHUMG00000033059	STK4	serine/threonine kinase 4	The protein encoded by this gene is a cytoplasmic kinase that is structurally similar to the yeast Ste20p kinase, which acts upstream of the stress-induced mitogen-activated protein kinase cascade. The encoded protein can phosphorylate myelin basic protein and undergoes autophosphorylation. A caspase-cleaved fragment of the encoded protein has been shown to be capable of phosphorylating histone H2B. The particular phosphorylation catalyzed by this protein has been correlated with apoptosis, and it's possible that this protein induces the chromatin condensation observed in this process.
OTTHUMG00000033087	SERINC3 (TDE1)	serine incorporator 3	
OTTHUMG00000032623	TNNC2	troponin C type 2 (fast)	Troponin (Tn), a key protein complex in the regulation of striated muscle contraction, is composed of 3 subunits. The Tn-I subunit inhibits actomyosin ATPase, the Tn-T subunit binds tropomyosin and Tn-C, while the Tn-C subunit binds calcium and overcomes the inhibitory action of the troponin complex on actin filaments. The protein encoded by this gene is the Tn-C subunit.
OTTHUMG00000032552	TOMM34	translocase of outer mitochondrial membrane 34	The protein encoded by this gene is involved in the import of precursor proteins into mitochondria. The encoded protein has a chaperone-like activity, binding the mature portion of unfolded proteins and aiding their import into mitochondria. This protein, which is found in the cytoplasm and sometimes associated with the outer mitochondrial

			membrane, has a weak ATPase activity and contains 6 TPR repeats.
OTTHUMG00000033057	TOP1	topoisomerase (DNA) I	This gene encodes a DNA topoisomerase, an enzyme that controls and alters the topologic states of DNA during transcription. This enzyme catalyzes the transient breaking and rejoining of a single strand of DNA which allows the strands to pass through one another, thus altering the topology of DNA. This gene is localized to chromosome 20 and has pseudogenes which reside on chromosomes 1 and 22.
OTTHUMG00000085887	TP53RK	TP53 regulating kinase	
OTTHUMG00000033038	UBE2C	ubiquitin-conjugating enzyme E2C	The modification of proteins with ubiquitin is an important cellular mechanism for targeting abnormal or short-lived proteins for degradation. Ubiquitination involves at least three classes of enzymes: ubiquitin-activating enzymes, or E1s, ubiquitin-conjugating enzymes, or E2s, and ubiquitin-protein ligases, or E3s. This gene encodes a member of the E2 ubiquitin-conjugating enzyme family. This enzyme is required for the destruction of mitotic cyclins and for cell cycle progression. Multiple alternatively spliced transcript variants have been found for this gene, but the full-length nature of some variants has not been defined.
OTTHUMG00000033085	UBE2V1	ubiquitin-conjugating enzyme E2 variant 1	Ubiquitin-conjugating E2 enzyme variant proteins constitute a distinct subfamily within the E2 protein family. They have sequence similarity to other ubiquitin-conjugating enzymes but lack the conserved cysteine residue that is critical for the catalytic activity of E2s. The protein encoded by this gene is located in the nucleus and can cause transcriptional activation of the human FOS proto-oncogene. It is thought to be involved in the control of differentiation by altering cell cycle behavior. Multiple alternatively spliced transcripts encoding different isoforms have been described for this gene. A pseudogene has been identified which is also located on chromosome 20. Co-transcription of this gene and the neighboring upstream gene generates a rare transcript (Kua-UEV), which encodes a fusion protein comprised of sequence sharing identity with each individual gene product.
OTTHUMG00000046331	WFDC10A	WAP four-disulfide core domain 10A	This gene encodes a member of the WAP-type four-disulfide core (WFDC) domain family. The WFDC domain, or WAP signature motif, contains eight cysteines forming four disulfide bonds at the core of the protein, and functions as a protease inhibitor. Most WFDC gene members are localized to chromosome 20q12-q13 in two clusters: centromeric and telomeric. This gene belongs to the telomeric cluster.
OTTHUMG00000046334	WFDC10B	WAP four-disulfide core domain 10B	This gene encodes a member of the WAP-type four-disulfide core (WFDC) domain family. The WFDC domain, or WAP signature motif, contains eight cysteines forming four disulfide bonds at the core of the protein, and functions as a protease inhibitor. Most WFDC gene members are localized to chromosome 20q12-q13 in two clusters: centromeric and telomeric. This gene belongs to the telomeric cluster. Two alternatively spliced transcript variants have been found for this gene, and they encode distinct isoforms.
OTTHUMG00000046330	WFDC11	WAP four-disulfide core domain 11	This gene encodes a member of the WAP-type four-disulfide core (WFDC) domain

			family. The WFDC domain, or WAP signature motif, contains eight cysteines forming four disulfide bonds at the core of the protein, and functions as a protease inhibitor. Most WFDC gene members are localized to chromosome 20q12-q13 in two clusters: centromeric and telomeric. This gene belongs to the telomeric cluster.
OTTHUMG00000046412	WFDC12	WAP four-disulfide core domain 12	This gene encodes a member of the WAP-type four-disulfide core (WFDC) domain family. The WFDC domain, or WAP signature motif, contains eight cysteines forming four disulfide bonds at the core of the protein, and functions as a protease inhibitor. Most WFDC gene members are localized to chromosome 20q12-q13 in two clusters: centromeric and telomeric. This gene belongs to the centromeric cluster.
OTTHUMG00000046333	WFDC13	WAP four-disulfide core domain 13	This gene encodes a member of the WAP-type four-disulfide core (WFDC) domain family. The WFDC domain, or WAP signature motif, contains eight cysteines forming four disulfide bonds at the core of the protein, and functions as a protease inhibitor. Most WFDC gene members are localized to chromosome 20q12-q13 in two clusters: centromeric and telomeric. This gene belongs to the telomeric cluster.
OTTHUMG00000032594	WFDC2	WAP four-disulfide core domain 2	This gene generates multiple alternatively spliced transcript variants, which encode different protein isoforms. These isoforms contain one or two WAP-type four-disulfide core (WFDC) domains, and are thus members of the WFDC domain family. The WFDC domain, or WAP Signature motif, contains eight cysteines forming four disulfide bonds at the core of the protein, and functions as a protease inhibitor in many family members. This gene is expressed in pulmonary epithelial cells, and was also found to be expressed in some ovarian cancers. The encoded isoforms are small secretory proteins, which may be involved in sperm maturation.
OTTHUMG00000032614	WFDC3	WAP four-disulfide core domain 3	This gene encodes a member of the WAP-type four-disulfide core (WFDC) domain family. The WFDC domain, or WAP signature motif, contains eight cysteines forming four disulfide bonds at the core of the protein, and functions as a protease inhibitor. The encoded protein contains four WFDC domains. Most WFDC genes are localized to chromosome 20q12-q13 in two clusters: centromeric and telomeric. This gene belongs to the telomeric cluster. Several alternatively spliced transcript variants have been found for this gene, but the full-length nature of some variants has not been determined.
OTTHUMG00000046411	WFDC5	WAP four-disulfide core domain 5	This gene encodes a member of the WAP-type four-disulfide core (WFDC) domain family. Most WFDC proteins contain only one WFDC domain, and this encoded protein contains two WFDC domains. The WFDC domain, or WAP signature motif, contains eight cysteines forming four disulfide bonds at the core of the protein, and functions as a protease inhibitor. Most WFDC gene members are localized to chromosome 20q12-q13 in two clusters: centromeric and telomeric. This gene belongs to the centromeric cluster.
OTTHUMG00000046354	WFDC6	WAP four-disulfide core domain 6	This gene encodes a member of the WAP-type four-disulfide core (WFDC) domain family. The WFDC domain, or WAP signature motif, contains eight cysteines forming four disulfide bonds at the core of the protein, and functions as a protease inhibitor.

			Most WFDC gene members are localized to chromosome 20q12-q13 in two clusters: centromeric and telomeric. This gene belongs to the telomeric cluster.
OTTHUMG00000046342	WFDC8	WAP four-disulfide core domain 8	This gene encodes a member of the WAP-type four-disulfide core (WFDC) domain family. The WFDC domain, or WAP signature motif, contains eight cysteines forming four disulfide bonds at the core of the protein, and functions as a protease inhibitor. The encoded protein contains a Kunitz-inhibitor domain, in addition to three WFDC domains. Most WFDC genes are localized to chromosome 20q12-q13 in two clusters: centromeric and telomeric. This gene belongs to the telomeric cluster. Two alternatively spliced transcript variants have been found for this gene, and they encode the same protein.
OTTHUMG00000046332	WFDC9	WAP four-disulfide core domain 9	The WAP-type four-disulfide core (WFDC) domain, or WAP signature motif, contains eight cysteines forming four disulfide bonds at the core of the protein, and functions as a protease inhibitor in many members of the WFDC domain family. This gene encodes a protein which contains a WFDC domain, and is thus a member of the WFDC domain family. This gene and several other gene family members are clustered at 20q13.12.
OTTHUMG00000033071	WISP2	WNT1 inducible signaling pathway protein 2	This gene encodes a member of the WNT1 inducible signaling pathway (WISP) protein subfamily, which belongs to the connective tissue growth factor (CTGF) family. WNT1 is a member of a family of cysteine-rich, glycosylated signaling proteins that mediate diverse developmental processes. The CTGF family members are characterized by four conserved cysteine-rich domains: insulin-like growth factor-binding domain, von Willebrand factor type C module, thrombospondin domain and C-terminal cystine knot-like (CT) domain. The encoded protein lacks the CT domain which is implicated in dimerization and heparin binding. It is 72% identical to the mouse protein at the amino acid level. This gene may be downstream in the WNT1 signaling pathway that is relevant to malignant transformation. Its expression in colon tumors is reduced while the other two WISP members are overexpressed in colon tumors. It is expressed at high levels in bone tissue, and may play an important role in modulating bone turnover.
OTTHUMG00000032549	YWHAB	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, beta polypeptide	This gene encodes a protein belonging to the 14-3-3 family of proteins, members of which mediate signal transduction by binding to phosphoserine-containing proteins. This highly conserved protein family is found in both plants and mammals. The encoded protein has been shown to interact with RAF1 and CDC25 phosphatases, suggesting that it may play a role in linking mitogenic signaling and the cell cycle machinery. Two transcript variants, which encode the same protein, have been identified for this gene.
OTTHUMG00000032481	ZHX3	zinc fingers and homeoboxes 3	This gene encodes a member of the zinc fingers and homeoboxes (ZHX) gene family. The encoded protein contains two C2H2-type zinc fingers and five homeodomains and forms a dimer with itself or with zinc fingers and homeoboxes family member 1. In the nucleus, the dimerized protein interacts with the A subunit of the ubiquitous transcription factor nuclear factor-Y and may function as a transcriptional repressor.

OTTHUMG00000032709	ZNF313	zinc finger protein 313	
OTTHUMG00000032654	ZNF334	zinc finger protein 334	
OTTHUMG00000032637	ZNF335	zinc finger protein 335	The protein encoded by this gene enhances transcriptional activation by ligand-bound nuclear hormone receptors. However, it does this not by direct interaction with the receptor, but by direct interaction with the nuclear hormone receptor transcriptional coactivator NRC. The encoded protein may function by altering local chromatin structure.
OTTHUMG00000074023	ZSWIM1	zinc finger, SWIM-type containing 1	
OTTHUMG00000032627	ZSWIM3	zinc finger, SWIM-type containing 3	

Table A. 1 The list of 103 protein coding genes residing on 20q12-13.2 region. First column is the identification tag of the gene in VEGA Genome Browser. Sec column is the gene name set by HUGO gene nomenclature committee, third column is the long description of the gene name and the last column gives a summary about the function of the product of the gene.

External Gene ID	CpG Island Topology	# of (CG)s	%GC	CpG island (Vega)	CpG island Searcher	Transcript Type
C20orf168-001	--	0	40.7	no	no	RT
WFDC11-001	--	0	40.7	no	no	RT
CHD6-002	--	0	41.3	no	no	AT
PRKCBP1-007	--	0	39.3	no	no	AT
PRKCBP1-009	--	0	41.3	no	no	AT
ZHX3-002	--	0	34.3	no	no	AT
C20orf10-001	---|--	1	49.0	no	no	RT
C20orf62-001	--|	1	45.7	no	no	RT
PI3-001	--|---------------------------	1	45.3	no	no	RT
SEMG1-001	--|-------	1	41.7	no	no	RT
SPINT4-001	---|-----------------------------------	1	46.0	no	no	RT
WFDC10A-001	--|----	1	25.0	no	no	RT
WFDC10B-001	--|------------------------------------	1	43.3	no	no	RT
SLC13A3-004	---|-------------------------------	1	41.7	no	no	AT
SLC35C2-003	--|----	1	40.0	no	no	AT
UBE2C-003	--|--	1	44.7	no	no	AT
WFDC3-001	---------------------------------|---	1	38.3	no	no	AT
ZHX3-004	--|----------------------	1	41.0	no	no	AT
C20orf123-001	---------------|---------------|---	2	56.0	no	no	RT
C20orf157-001	---|-----------------------|------	2	36.3	no	no	RT
JPH2-001	------------------------------------|--|-------------------	2	55.3	no	no	RT
MATN4-002	---|---------|-----	2	41.0	no	no	RT
R3HDML-001	------------|------|---	2	59.3	no	no	RT
SPINLW1-001	-------|--|-----------------	2	53.3	no	no	RT
WFDC6-002	---|--------------------|---------------------	2	55.7	no	no	RT

<table>
<tr><td>WFDC8-001</td><td>------|---|---</td><td>2</td><td>60.7</td><td>no</td><td>no</td><td>RT</td></tr>
<tr><td>C20orf119-007</td><td>----------------|---------------|---</td><td>2</td><td>47.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>DNTTIP1-003</td><td>---|-------|----------</td><td>2</td><td>43.0</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>GDAP1L1-002</td><td>----------------------------------|--|--</td><td>2</td><td>55.0</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>L3MBTL-004</td><td>---|-|--------------------------</td><td>2</td><td>55.7</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>PKIG-008</td><td>----------------------------------|--------------------------------|--------------------------------</td><td>2</td><td>46.7</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>ZHX3-007</td><td>-------------|--|---</td><td>2</td><td>43.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>SEMG2-001</td><td>-------------|--------|--------|--</td><td>3</td><td>41.7</td><td>no</td><td>no</td><td>RT</td></tr>
<tr><td>SLPI-001</td><td>---|-----------|--------------------------------|-----</td><td>3</td><td>49.0</td><td>no</td><td>no</td><td>RT</td></tr>
<tr><td>SPINT3-001</td><td>--|----------|----|---</td><td>3</td><td>55.0</td><td>no</td><td>no</td><td>RT</td></tr>
<tr><td>WFDC13-001</td><td>--|----------------|-----|--------------------</td><td>3</td><td>40.7</td><td>no</td><td>no</td><td>RT</td></tr>
<tr><td>WFDC9-001</td><td>--|---|---------------------|-----------------</td><td>3</td><td>53.7</td><td>no</td><td>no</td><td>RT</td></tr>
<tr><td>C20orf119-011</td><td>-----------------|-------------------------------------|-------------|------------------------------</td><td>3</td><td>52.0</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>DBNDD2-012</td><td>--------------|------|--|-------------------------</td><td>3</td><td>54.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>ELMO2-005</td><td>---|-------------------------|---|------------</td><td>3</td><td>51.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>PKIG-004</td><td>---------------------------|--|-----------------|-</td><td>3</td><td>49.0</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>SGK2-002</td><td>-------------------------------|--------|-------------|---</td><td>3</td><td>56.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>SPINLW1-002</td><td>----------|-------------------------------|----------------------------------|----------------------</td><td>3</td><td>34.0</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>UBE2V1-007</td><td>----------|---|--|------------------</td><td>3</td><td>53.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>C20orf165-001</td><td>-----------------------|----------------------------------|-------------------------|--------|------</td><td>4</td><td>48.7</td><td>no</td><td>no</td><td>RT</td></tr>
<tr><td>MMP9-001</td><td>----|--|------------|---|---------------------------</td><td>4</td><td>61.7</td><td>no</td><td>no</td><td>RT</td></tr>
<tr><td>WFDC5-002</td><td>----------------------------|---------------|---------------|------------|--------------------------</td><td>4</td><td>55.7</td><td>no</td><td>no</td><td>RT</td></tr>
<tr><td>C20orf119-004</td><td>--------------------|-------------------------------|-|-------------------------------|-------------</td><td>4</td><td>47.7</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>DBNDD2-010</td><td>---|-|----------||----------------------------------</td><td>4</td><td>53.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>FAM112A-002</td><td>----|-----------------------------------|-----------------------------|-----------------------------</td><td>4</td><td>47.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>CHD6-008</td><td>----------|------|--|------------------|----------</td><td>4</td><td>44.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>ELMO2-004</td><td>---|-----|-------|---|------------</td><td>4</td><td>41.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>L3MBTL-011</td><td>-------------|---------------------------------|----------------------------|--------------------|--</td><td>4</td><td>54.3</td><td>no</td><td>no</td><td>AT</td></tr>
</table>

<table>
<tr><td>LPIN3-003</td><td>------------------------|---------------|----------------------|---------|-----------------------</td><td>4</td><td>57.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>WFDC3-006</td><td>----|---------|---|---------------------|-----------------</td><td>4</td><td>50.7</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>CDH22-001</td><td>-------|--------|--|--|--|</td><td>5</td><td>54.3</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>SGK2-005</td><td>----------|--------|---|--------------|-----|-</td><td>5</td><td>55.0</td><td>no</td><td>no</td><td>RT</td></tr>
<tr><td>WFDC12-001</td><td>--|-----|-----|---|-|--</td><td>5</td><td>55.7</td><td>no</td><td>no</td><td>RT</td></tr>
<tr><td>C20orf119-005</td><td>------------------------|---||--|-----------|----</td><td>5</td><td>43.0</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>ELMO2-006</td><td>---------|------------------|-------|---|--|----------------</td><td>5</td><td>51.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>ZNFX1-003</td><td>-|------------|------------|---|--------|-----------</td><td>5</td><td>48.7</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>L3MBTL-010</td><td>---------------||-----|----------------|----------------|---</td><td>5</td><td>59.0</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>PLCG1-004</td><td>------|------|----|---|-----|--------------------------</td><td>5</td><td>56.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>PRKCBP1-006</td><td>-----|-----|--|---|----------------------|--------</td><td>5</td><td>48.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>STAU1-008</td><td>|------||----------------|--------------|---</td><td>5</td><td>44.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>SERINC3-002</td><td>-------------|--------------|-----------------|---------------------------------------||----------</td><td>5</td><td>46.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>TNNC2-003</td><td>------||-----------------|-------------------------------|---------------------------|-------------</td><td>5</td><td>60.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>ZHX3-008</td><td>---------|---------------------------------|-|------------------------------------|---------------|</td><td>5</td><td>59.0</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>C20orf142-001</td><td>------------------|--|--------------------|---------------|-------------------------------|--------</td><td>6</td><td>43.0</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>FAM112A-001</td><td>-----|---------|---------------|---|----------|--|--------</td><td>6</td><td>46.3</td><td>no</td><td>no</td><td>RT</td></tr>
<tr><td>C20orf119-009</td><td>--|-----||-|--------------||------</td><td>6</td><td>56.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>PREX1-002</td><td>-------------------|---------------------------------|----|-----------------------|-|-------------|</td><td>6</td><td>58.7</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>EYA2-004</td><td>----|-------|----------|----------||------------------------|---------------|----------------------</td><td>7</td><td>53.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>SLC35C2-007</td><td>-----------------------------------||-------|----------|--------|----------|-------------|--------</td><td>7</td><td>56.0</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>SLC13A3-003</td><td>---------------------|--------------|----|---------------------------|-|-------------|--------|----</td><td>8</td><td>61.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>ZSWIM3-001</td><td>-------|------|-----|---------------|------------------|-------|---|--|-----------------------|----</td><td>9</td><td>52.3</td><td>no</td><td>yes</td><td>RT</td></tr>
<tr><td>DNTTIP1-005</td><td>--------------------------------|--|-|---------|--------|-----|--------------------|----|----|----</td><td>9</td><td>52.3</td><td>no</td><td>yes</td><td>AT</td></tr>
<tr><td>L3MBTL-005</td><td>--------------|-|--------------|----------------|---------------|-----|---------|-----------|----|-</td><td>9</td><td>66.3</td><td>no</td><td>no</td><td>RT</td></tr>
<tr><td>ELMO2-007</td><td>--------------------------|--|---|----||------|---------------------------------|----|------|------</td><td>9</td><td>51.7</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>SULF2-003</td><td>----|-------------------------|--------------------------|--------------|---|----|----|----|-|-----</td><td>9</td><td>59.7</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>PRKCBP1-003</td><td>--------------------|-----|--------||----|-|----|------|----|---------------------------------|----</td><td>10</td><td>48.3</td><td>no</td><td>no</td><td>RT</td></tr>
</table>

<table>
<tr><td>DBNDD2-006</td><td>-------|-|-------|-----------|--|------------------||----|-----|--------------------------------|---</td><td>10</td><td>61.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>HNF4A-003</td><td>-----------|------------|--------------|----||--|--------|--------------------|---------------|--|-</td><td>10</td><td>60.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>SULF2-005</td><td>|--|----|-----------------------|---|-------|-------|--------|--------------------------|---|----</td><td>10</td><td>60.7</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>SGK2-007</td><td>-----------------|-|-------------------|--|------||-|----------|-----------------------|-----------</td><td>10</td><td>63.0</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>SLC35C2-006</td><td>----|---------------------------|-----------|---------|------------|-----|---|-|-----|-------|--</td><td>10</td><td>60.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>UBE2V1-004</td><td>---|---------------------|--------|---------------------------------|--|------|---|-|--|--|------</td><td>10</td><td>59.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>HNF4A-001</td><td>---|--------------------------|-------|-----------|----------|-------------|-|---|--|----|--------</td><td>11</td><td>60.0</td><td>no</td><td>no</td><td>RT</td></tr>
<tr><td>C20orf100-001</td><td>-------------------------|---------|-----------|-----------|--||----------|---|--|----|-|---------</td><td>11</td><td>55.0</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>C20orf100-004</td><td>-----------------------||-------------|------|----------------|---|------|--------|----------|----</td><td>11</td><td>53.7</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>JPH2-002</td><td>---|----------------------------|----------|--||--|--|---------|--|----||---|---------------------</td><td>12</td><td>64.0</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>SLC2A10-001</td><td>|---|------------||-----------------------|--------------------|---|--|----|----|---------||-|---</td><td>13</td><td>66.7</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>ACOT8-001</td><td>----|||-||--------|------|------|--------------|------------------|--------|--|--|-----------------</td><td>13</td><td>54.2</td><td>no</td><td>yes</td><td>AT</td></tr>
<tr><td>WISP2-002</td><td>---||-------------|----|--------|-|--|-----|---------------|---------------------|---||------|-----</td><td>13</td><td>63.3</td><td>no</td><td>no</td><td>RT</td></tr>
<tr><td>TP53RK-001</td><td>-------------------|--------|-----------------------------|--------||||--|--||--|--|---------|-|-</td><td>14</td><td>49.7</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>WISP2-001</td><td>--------------------------------||-------|-------------------------|--------------------|||||||--|</td><td>14</td><td>57.7</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>SERINC3-001</td><td>----------------|--|---|-||||-|--||-|-||----|-|----|---</td><td>16</td><td>54.0</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>C20orf111-002</td><td>|------|-----||----|------|--|------|-|---------|-------|------------|---||----------------|-|-----</td><td>16</td><td>59.0</td><td>yes</td><td>yes</td><td>AT</td></tr>
<tr><td>CD40-001</td><td>------|-|---------------------|----|----|-------------|------|--|---|------|----|------------|||---</td><td>16</td><td>64.0</td><td>no</td><td>no</td><td>RT</td></tr>
<tr><td>CHD6-006</td><td>||--||----||-----|----|---|------------|--|---------||---||---|-|----------------------------------</td><td>17</td><td>56.0</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>EYA2-005</td><td>----------------------------|---------|-||----------------|----|--|------------||--||--|-----||||-|</td><td>17</td><td>60.7</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>PLTP-002</td><td>---------------|--|||-|--||--|-|-----|---------|--|-------|--|----------|-----|------------------|-</td><td>17</td><td>61.0</td><td>yes</td><td>yes</td><td>AT</td></tr>
<tr><td>UBE2C-004</td><td>-|---||-------------------||----|--|--------|----------|-|--|-----||---|---------|-----|-------|---</td><td>17</td><td>65.0</td><td>YES</td><td>yes</td><td>AT</td></tr>
<tr><td>GDAP1L1-001</td><td>--|--------------------|----|-------------|---||---------|---|------|-----|-|--------||------|---||</td><td>17</td><td>66.3</td><td>yes</td><td>no</td><td>RT</td></tr>
<tr><td>DDX27-001</td><td>----------------------------------|--|-----------|---|-|---|----|-|----||--|-----|--|---|--||---|-</td><td>17</td><td>49.3</td><td>no</td><td>yes</td><td>RT</td></tr>
<tr><td>KCNS1-001</td><td>-----------------------------|---|-------|-|-|-----------------|-----||-||-|-----|-|-||---||------</td><td>17</td><td>59.0</td><td>no</td><td>yes</td><td>RT</td></tr>
<tr><td>ZSWIM1-002</td><td>-|------|--------------------------|-||||-|-|--|------|------|--|--|-----------|--------||---------</td><td>17</td><td>57.7</td><td>no</td><td>no</td><td>RT</td></tr>
<tr><td>PKIG-002</td><td>-----------------------||----------------|------------|---|-----|---|----|---|-----||-|--||-|----||</td><td>18</td><td>60.3</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>KCNK15-001</td><td>------------------|---|----|-------|------|----------------|--|---|---|--|--|-|------|-|-|-||----|-</td><td>19</td><td>65.3</td><td>yes</td><td>yes</td><td>RT</td></tr>
</table>

<table>
<tr><td>WFDC2-002</td><td>---------------------|---|-|---|----------|------------------|-|--|--------||-----|---|||---|----|-</td><td>19</td><td>66.3</td><td>yes</td><td>no</td><td>RT</td></tr>
<tr><td>LPIN3-001</td><td>----|------|----|------|-|-|---|--|-|---|----------------|--|-----|--------|-|-----|---------|---||</td><td>19</td><td>69.0</td><td>no</td><td>no</td><td>RT</td></tr>
<tr><td>CSE1L-001</td><td>-------------||---------|-|-----------------|----|-|---|----------|---|-||--|-----|-|---|-----|-||-</td><td>20</td><td>61.7</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>PLTP-001</td><td>-------|------------------||---|---------|--|-||------------------|--|---|------|--|||-----|---|-||</td><td>20</td><td>69.3</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>SULF2-001</td><td>-||--------|------------|----|--|---|----|---|--|------------|--|-------|-------------|---|---|||--</td><td>20</td><td>59.3</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>TOMM34-001</td><td>----------------------|------------------|------|--||-----------|-|--------|-||--||-|----|--|--||-|</td><td>20</td><td>58.7</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>UBE2V1-009</td><td>--------|---------------|-|---|-||-------|--|-------|--|-|||--------|-|--------|--|-----|--|-|-----</td><td>20</td><td>54.3</td><td>no</td><td>no</td><td>AT</td></tr>
<tr><td>C20orf119-017</td><td>------------------------|---|---------------|---------|---||-|-|--|--||-|----|-||-----|-|-||-----|-</td><td>21</td><td>69.0</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>IFT52-001</td><td>-------|-------|------|--|-||--|-|--|-||-|--||---|--|||||--</td><td>21</td><td>54.3</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>PTGIS-001</td><td>--|------------------------------||-|--|-|----------|--|----|---|||--|---------|--|--|--|--||---||-</td><td>21</td><td>75.3</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>ACOT8-004</td><td>-----|-----|------|--------||-||-|----------|-|----------||||-----|-|--|-|---||--|-----|-----------</td><td>22</td><td>63.3</td><td>no</td><td>yes</td><td>RT</td></tr>
<tr><td>ZNF334-001</td><td>|-|------|-------------------------|-|||------|-||--|---|--|-|---|---------|--|-|---|------|----||-</td><td>22</td><td>60.0</td><td>no</td><td>yes</td><td>RT</td></tr>
<tr><td>SLC13A3-001</td><td>-------------|---|----------|-|----|---|-------------|---|---|-------|||--|---|-||-|-|-|--||---|---</td><td>23</td><td>69.0</td><td>no</td><td>yes</td><td>RT</td></tr>
<tr><td>SLC12A5-001</td><td>----|----|-------|-----------|---|---------|-----|----|---|-----||----|||--|--|-|---|||-|------|-|-</td><td>24</td><td>68.0</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>ZNF335-001</td><td>-|-----|--||---------------|||---------|-----|-|--|-------|--|----|----|---|-|-----|-||-|----|--|--</td><td>24</td><td>56.7</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>PIGT-003</td><td>-----|----|------|-----|---------------||-|-||---|--|-|-|-|-----|-----|----|-||--|---|--||--|----|-</td><td>25</td><td>61.7</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>YWHAB-002</td><td>|------|---------|-|------|--------|--|--|------|-------|-|--|-|---|-|------|--|-|---||--|-||----|-</td><td>25</td><td>69.7</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>GDAP1L1-006</td><td>--||------|---|||---------|----|---------------||--|-|---|--||-|-|--|------------------|----||--|-|</td><td>25</td><td>70.0</td><td>yes</td><td>no</td><td>AT</td></tr>
<tr><td>C20orf121-002</td><td>--------------------------|---|------------|-|--|--------|||-||--||--|--|--|--|--|||-|--|----||--||</td><td>25</td><td>65.3</td><td>no</td><td>yes</td><td>RT</td></tr>
<tr><td>ELMO2-003</td><td>---------|-|-||----|---|-----------------|||-|----|-|--|-|----|-||----||-|-|--------|------|---|||-</td><td>26</td><td>68.7</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>UBE2C-001</td><td>--||-|---------|----|--|---||--------|---|-||--------|-----||-------||||-|------|--|------||-|-----</td><td>26</td><td>59.0</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>TP53RK-002</td><td>---|-||-|---|-|--|---------------|----|----|---|--|-|-|------|------||-|--|-------|----||----||----</td><td>26</td><td>59.7</td><td>yes</td><td>yes</td><td>AT</td></tr>
<tr><td>C20orf67-001</td><td>|-|----||-|-||--|--|--|-----|-|----------|----------|------|-------|-----|---|------||||||----|--|-</td><td>27</td><td>67.0</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>PPGB-002</td><td>----|---------|----||------|--||---|-----|||----||------||--|-|-------|-||||---|---------------|-|-</td><td>27</td><td>66.7</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>RBPSUHL-003</td><td>|--|-|------|-|-||--|--|-----||--|-|----|---------|--|----|-|------------|--|---------|-----|--||--</td><td>27</td><td>70.0</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>SNX21-006</td><td>-|----|||----|-|----|-|||---|------|---|----------|-------|--|--------|----|-|-|-----||------|-|--|</td><td>28</td><td>73.0</td><td>yes</td><td>yes</td><td>AT</td></tr>
<tr><td>PPGB-005</td><td>---|---|---||-----|----|------|||-|-|-|--|---|---|-|---|------|----|-||--|-|-------||------|-------</td><td>28</td><td>66.8</td><td>yes</td><td>yes</td><td>AT</td></tr>
<tr><td>UBE2V1-014</td><td>----------|-|----------|-|----------------|-|---||-|-|||-|-|||-|--|--|--------|----||------|-|-|-||</td><td>28</td><td>66.7</td><td>yes</td><td>yes</td><td>AT</td></tr>
</table>

<table>
<tr><td>DBNDD2-005</td><td>----||----|---|--|-|----|---|-|----|------|||--||-|-------||-|||--|--------||------------||-------|</td><td>29</td><td>72.7</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>KCNB1-001</td><td>----------|-|-|--------||---||---|--|-----|--|-----|---------|-------|---|||-|---|-||------||-||||-</td><td>29</td><td>72.7</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>PPGB-001</td><td>----|--|||---||-|---||--||----||---|--|--|-----|||------|-|-|------|-------||----||-----|-|--------</td><td>29</td><td>69.7</td><td>yes</td><td>yes</td><td>AT</td></tr>
<tr><td>NCOA5-002</td><td>---------|---|-|-------|-------|---|-|---|-------||-----|--||--|--||-|-|||-------|-||---|-||-|--|--</td><td>30</td><td>66.7</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>NEURL2-001</td><td>--|-|----|----||-|---|--|----||---||-|-|-|-----|--------|------||-------|-|--||-|----|------|---|-|</td><td>30</td><td>72.0</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>SNX21-003</td><td>--||--|-----|-------|-----------|--||-||-||---|-||----|||--|---||----|-|----|----|---|||----|--|--</td><td>31</td><td>75.7</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>SLC35C2-005</td><td>----|--------|----|-----------|----||-|---||--|-|-|------||--|-|||-|||-|-|--|---|------|-----||--||</td><td>31</td><td>69.0</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>SLC9A8-001</td><td>-|-|------|---|--------|--||--|---------|-----|---|----|-||||--|-||-|||--|-|---|----||-|-|-----|---</td><td>31</td><td>69.3</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>SPATA2-001</td><td>----|-----||--|-----|-||-|-|-|--|---|-----||-|--||-|--||-|-----|-|------|-|--|-|---|---------||----</td><td>31</td><td>66.7</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>NCOA3-001</td><td>----------|-|--|----|--|-|-||--|||--||------|----|||-----|---|--|-------|-||-------|--||-|-|--|-|-</td><td>32</td><td>69.0</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>C20orf111-001</td><td>-|---|------|------|-|-|----|-|-|---|-||-|-|-||-||-|--------------|-|-|-|--|||-|------|--|--|----|</td><td>33</td><td>68.3</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>ZNFX1-004</td><td>----|-|-||--|------------|--|---||---|-|||||------|----||------||--|-----||--|---|-|--||--|--|||--</td><td>33</td><td>63.3</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>SDC4-001</td><td>------------|---------||-|--|-------|||-|--|--|-----|-|-|||-|-|-||-|--|-----||||---||--|----||----</td><td>33</td><td>74.0</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>TNNC2-001</td><td>|------|---|||-|----|-|----|||----|---|-----||-----||--|--||-|----|||---|-|||-||--|------------|---</td><td>33</td><td>76.0</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>ZNF313-001</td><td>--|--|-----|--|--|-----------|-----|---|-|----|-|--|--|---|-|-|--|-|----|-|||||-|----||--|-|---|--|</td><td>33</td><td>69.0</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>STAU1-002</td><td>----||--------|--|---|-|-|--|--|--|---||--||--||---|----||-----|-|---|-|-|-||----|----|-||--|---|-|</td><td>34</td><td>76.7</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>C20orf100-002</td><td>--|---|-|-|----|-------------||--|||------------||-|-|-||-||----||-|--|-|---||---|---|-|--|-|-|--</td><td>34</td><td>68.3</td><td>yes</td><td>yes</td><td>AT</td></tr>
<tr><td>TOP1-001</td><td>-----||---|------|-|-||-|--|-----|-|--|-|-|-------|-|-||||----|||-|----|------------|||-|---||||----</td><td>35</td><td>66.3</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>ZNFX1-002</td><td>---||-|||-----|--|-|-|||-----|-------|-------|-||-|---|--|-|-|--|||-|--|---|----|||----|||-|---|-----</td><td>35</td><td>78.0</td><td>yes</td><td>yes</td><td>AT</td></tr>
<tr><td>STK4-003</td><td>----------|-|---|---|-|-|---|-------|-|---|----|-|-||-||----||-|---|--|-|-||||-|--|-|------||----|</td><td>35</td><td>69.7</td><td>no</td><td>yes</td><td>RT</td></tr>
<tr><td>DBNDD2-008</td><td>|-||-|-|----|||---||||-|------|--|------|----|------|||----|-|||-|-|--|-|-----|-|---|-|---||--|----</td><td>36</td><td>73.3</td><td>yes</td><td>yes</td><td>AT</td></tr>
<tr><td>DNTTIP1-006</td><td>--|-|--|-----||-|-----|--|-|---||-|||-----|-|----|-----||--||-|---|--|||-|||---|-----||||-|--------</td><td>36</td><td>72.3</td><td>no</td><td>yes</td><td>RT</td></tr>
<tr><td>WFDC3-002</td><td>---------|-||-|-----|---||-|-||-||---|||---||----|----|-|-----|||--||--|-|--|------||-|-----|--|-|-</td><td>36</td><td>72.1</td><td>no</td><td>yes</td><td>RT</td></tr>
<tr><td>SFRS6-002</td><td>-------|-------------|--|-||--|-|--------||||-||-||--|---|--||---|---||||------||-----||-|--|-|-|-||</td><td>37</td><td>66.0</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>SNAI1-001</td><td>|-|-|-|-|---||---|||||-|--||---|---||-|--|-||-----|----|---------|------||-----||||--|----|---|-||--</td><td>37</td><td>72.3</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>ZHX3-006</td><td>-|--|---------|-||----|-|||---|-||--|-|-|--|--|------|--|-|--|---|--||-|-|-|--||-|-|----|-||---||</td><td>38</td><td>77.7</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>MAFB-001</td><td>-----|--|--||---||-|-||-|---||-|-----|-|-||--||-|----||-||-|-|--------|---||||------||-|-|-||--|----</td><td>39</td><td>76.3</td><td>yes</td><td>yes</td><td>RT</td></tr>
<tr><td>EMILIN3-001</td><td>|-||-|-|------|-----|||-|------------|-|--|||-|--|---||||||------|---|-|---||--|--||--|||-|||-|-||</td><td>40</td><td>77.3</td><td>yes</td><td>yes</td><td>RT</td></tr>
</table>

MYBL2-001	---|-|-||----|--------|-|--||-|---|-------||-||---|||||-|--------||||----|-|-||----|-|-||||--||-|--	41	77.7	yes	yes	RT
EYA2-003	-||--|-----||||-|-----|--|-||---|||--|||||--||-||||--||-----|-------|--||---||----|--|-|----|-||---|	42	79.7	yes	yes	AT
ADA-001	|-----|-------|---|--||--|||----||-||-|--|-|||-||-------||--|||-|--||--|-|-||||-|-|--||---|-|-|--|-	44	78.0	yes	yes	RT
ARFGEF2-001	---|-|---|-|||--||-|-|--|-|-||---|||--|-|---|||||-|-|-||-||-|---|-|-||-|||-|--||-|-----------|-|---	45	76.7	yes	yes	RT
SULF2-002	-||-|||---|--||-|-|-|-||-------|-|-|--|---||-|-|-|----||||---|-|-|||--|---------|||----|-||||||-|-|	45	82.7	yes	yes	AT
C20orf100-003	-----||--|||--|-----|-||--|--|---||-|---||||-|-|-|-|----|-||--|-|-|-|-----|||----||-||||||||--||---	46	84.0	yes	yes	RT
PLCG1-005	|-||-||-|-|-|----|||-|||-|||----|-|||---|||-|-------||||---||-|----|||-||--|--------||||---||----||	46	81.7	yes	yes	RT
B4GALT5-001	||-------|---|-|------|--|---|----|||--||||-||---||-||---|--|||-|-|-||-|||-|---|-|---|-|||||--||-||	48	77.7	yes	yes	RT
PTPRT-004	||||---|--|-|--|-|-|---|--|||||---|-||-|--||---||---||||||---|||---|-|-||-----|||-|-|-||||---------	48	81.7	yes	yes	RT
PREX1-001	-----|--||-|---|---|-|---|-||--||--|-|--||----|||||-|-||--|-|||-||---|-|---|-|||---||||||----|-|-|-	50	84.7	yes	yes	RT
UBE2V1-005	||||------|---|------|-|-|----|---|--|---|-||---|-||-|--||-|||-|---|||||--|-|||||||--|--||--|||-|--	51	79.3	yes	yes	RT
RIMS4-002	--|-|-||--||---|-||-|||-|||||||----|-|-||||----|||-||||--|-|-|-|-|-||-|||||-||---|-----|-|||||||-||	62	86.7	yes	yes	RT

Table A. 2 CpG content of 177 promoter regions. 250 bp upstream and 50 bp downstream of the TSS is taken as the promoter sequence. Each "CG" dinucleotide is shown with "|", other dinucleotides are shown with "-". If there are no 3 consecutive "CG" dinucleotides in the sequence, it is represented as "-" symbol, all other combinations are represented "|" for better display purposes. The grey boxes represent Vega CpG islands rejected by a more stringent definition (see 3.2.1). Yellow boxes represent new CpG islands meeting the conditions of the new definition but not detected by Vega. Green boxes represent CpG islands extending to the first exon of the gene.

External Gene ID	Transcript Type	CpG island Searcher	FirstEF	Eponine
ZHX3-006	RT	yes	yes	yes
SFRS6-002	RT	yes	yes	yes
C20orf100-003	RT	yes	yes	yes
DNTTIP1-006	RT	yes	yes	yes
WFDC3-002	RT	yes	yes	yes
PREX1-001	RT	yes	yes	yes
ARFGEF2-001	RT	yes	yes	yes
PTPRT-004	RT	yes	yes	yes
PLTP-001	RT	yes	yes	yes
SNAI1-001	RT	yes	yes	yes
MAFB-001	RT	yes	yes	yes
ELMO2-003	RT	yes	yes	yes
KCNS1-001	RT	yes	yes	yes
PLCG1-005	RT	yes	yes	yes
SDC4-001	RT	yes	yes	yes
UBE2V1-005	RT	yes	yes	yes
RIMS4-002	RT	yes	yes	yes
EMILIN3-001	RT	yes	yes	yes
IFT52-001	RT	yes	yes	no
C20orf111-001	RT	yes	yes	no
YWHAB-002	RT	yes	yes	no
C20orf119-017	RT	yes	yes	no
PIGT-003	RT	yes	yes	no
DBNDD2-005	RT	yes	yes	no
TNNC2-001	RT	yes	yes	no
SNX21-003	RT	yes	yes	no
NEURL2-001	RT	yes	yes	no
ZSWIM3-001	RT	yes	yes	no
C20orf67-001	RT	yes	yes	no
ZNF335-001	RT	yes	yes	no
SLC12A5-001	RT	yes	yes	no
ZNF334-001	RT	yes	yes	no
SULF2-001	RT	yes	yes	no
SLC13A3-001	RT	yes	yes	no

External Gene ID	Tanscript Type	CpG island Searcher	FirstEF	Eponine
DBNDD2-008	AT	yes	yes	yes
SULF2-002	AT	yes	yes	yes
ZNFX1-002	AT	yes	yes	yes
UBE2C-004	AT	yes	yes	yes
EYA2-003	AT	yes	yes	yes
C20orf100-002	AT	yes	yes	no
DNTTIP1-005	AT	yes	yes	no
SNX21-006	AT	yes	yes	no
PLTP-002	AT	yes	yes	no
PPGB-005	AT	yes	yes	no
TP53RK-002	AT	yes	yes	no
PPGB-001	AT	yes	no	yes
UBE2V1-014	AT	yes	no	yes
C20orf111-002	AT	yes	no	no
ACOT8-004	AT	yes	no	no
GDAP1L1-006	AT	no	yes	no
HNF4A-003	AT	no	yes	no
ZHX3-002	AT	no	no	no
ZHX3-004	AT	no	no	no
ZHX3-007	AT	no	no	no
ZHX3-008	AT	no	no	no
CHD6-002	AT	no	no	no
CHD6-008	AT	no	no	no
L3MBTL-004	AT	no	no	no
L3MBTL-010	AT	no	no	no
L3MBTL-011	AT	no	no	no
FAM112A-002	AT	no	no	no
C20orf100-001	AT	no	no	no
C20orf100-004	AT	no	no	no
GDAP1L1-002	AT	no	no	no
C20orf119-005	AT	no	no	no
C20orf119-007	AT	no	no	no
C20orf119-009	AT	no	no	no

ACOT8-001	RT	yes	yes	no
CSE1L-001	RT	yes	yes	no
SLC35C2-005	RT	yes	yes	no
KCNB1-001	RT	yes	yes	no
RBPSUHL-003	RT	yes	yes	no
STK4-003	RT	yes	yes	no
NCOA3-001	RT	yes	yes	no
PKIG-002	RT	yes	yes	no
DDX27-001	RT	yes	yes	no
CDH22-001	RT	yes	yes	no
PTGIS-001	RT	yes	yes	no
SERINC3-001	RT	yes	yes	no
ZNFX1-004	RT	yes	yes	no
SLC9A8-001	RT	yes	yes	no
TP53RK-001	RT	yes	yes	no
TOMM34-001	RT	yes	no	yes
NCOA5-002	RT	yes	no	yes
TOP1-001	RT	yes	no	yes
MYBL2-001	RT	yes	no	yes
ADA-001	RT	yes	no	yes
B4GALT5-001	RT	yes	no	yes
C20orf121-002	RT	yes	no	yes
CHD6-006	RT	yes	no	no
C20orf142-001	RT	yes	no	no
KCNK15-001	RT	yes	no	no
SLC2A10-001	RT	yes	no	no
STAU1-002	RT	yes	no	no
SPATA2-001	RT	yes	no	no
ZNF313-001	RT	yes	no	no
UBE2C-001	RT	yes	no	no
EYA2-005	RT	yes	no	no
PPGB-002	RT	yes	no	no
L3MBTL-005	RT	no	yes	no
HNF4A-001	RT	no	yes	no
WFDC2-002	RT	no	yes	no
C20orf119-011	AT	no	no	no
DBNDD2-006	AT	no	no	no
DBNDD2-010	AT	no	no	no
DBNDD2-012	AT	no	no	no
SPINLW1-002	AT	no	no	no
DNTTIP1-003	AT	no	no	no
WFDC3-001	AT	no	no	no
WFDC3-006	AT	no	no	no
TNNC2-003	AT	no	no	no
PRKCBP1-006	AT	no	no	no
PRKCBP1-007	AT	no	no	no
PRKCBP1-009	AT	no	no	no
SULF2-003	AT	no	no	no
SULF2-005	AT	no	no	no
PREX1-002	AT	no	no	no
STAU1-008	AT	no	no	no
ZNFX1-003	AT	no	no	no
JPH2-002	AT	no	no	no
UBE2C-003	AT	no	no	no
EYA2-004	AT	no	no	no
SLC13A3-003	AT	no	no	no
SLC13A3-004	AT	no	no	no
SLC35C2-003	AT	no	no	no
SLC35C2-006	AT	no	no	no
SLC35C2-007	AT	no	no	no
SGK2-002	AT	no	no	no
SGK2-007	AT	no	no	no
LPIN3-003	AT	no	no	no
ELMO2-004	AT	no	no	no
ELMO2-005	AT	no	no	no
ELMO2-006	AT	no	no	no
ELMO2-007	AT	no	no	no
WISP2-001	AT	no	no	no
PLCG1-004	AT	no	no	no
UBE2V1-004	AT	no	no	no
UBE2V1-007	AT	no	no	no
JPH2-001	RT	no	yes	no

CD40-001	RT	no	yes	no
LPIN3-001	RT	no	yes	no
FAM112A-001	RT	no	no	no
R3HDML-001	RT	no	no	no
GDAP1L1-001	RT	no	no	no
C20orf62-001	RT	no	no	no
SEMG1-001	RT	no	no	no
SEMG2-001	RT	no	no	no
PI3-001	RT	no	no	no
C20orf10-001	RT	no	no	no
SPINT3-001	RT	no	no	no
SPINLW1-001	RT	no	no	no
SPINT4-001	RT	no	no	no
C20orf168-001	RT	no	no	no
C20orf165-001	RT	no	no	no
C20orf157-001	RT	no	no	no
C20orf123-001	RT	no	no	no
PRKCBP1-003	RT	no	no	no
MATN4-002	RT	no	no	no
MMP9-001	RT	no	no	no
SGK2-005	RT	no	no	no
SLPI-001	RT	no	no	no
WFDC11-001	RT	no	no	no
WFDC10A-001	RT	no	no	no
WFDC9-001	RT	no	no	no
WFDC13-001	RT	no	no	no
WFDC10B-001	RT	no	no	no
WFDC8-001	RT	no	no	no
WFDC6-002	RT	no	no	no
WFDC5-002	RT	no	no	no
WFDC12-001	RT	no	no	no
WISP2-002	RT	no	no	no
ZSWIM1-002	RT	no	no	no

Table A. 3 The prediction profiles of 103 representative transcripts (RT)

UBE2V1-009	AT	no	no	no
SERINC3-002	AT	no	no	no
PKIG-004	AT	no	no	no
PKIG-008	AT	no	no	no
C20orf119-004	AT	no	no	no

Table A. 4 Prediction profiles of alternative transcripts (AT)

Expression in HeLa S3	Expression in NTERA-D1	Vega Gene ID	HUGO Gene Name	Transcript count
P	P	OTTHUMG00000033045	ACOT8	3
P	P	OTTHUMG00000033081	ADA	3
P	P	OTTHUMG00000032687	ARFGEF2	1
P	P	OTTHUMG00000033086	B4GALT5	1
P	P	OTTHUMG00000032582	C20orf10	1
A	P	OTTHUMG00000032517	C20orf100	4
P	P	OTTHUMG00000032518	C20orf111	2
P	P	OTTHUMG00000032553	C20orf119	8
P	P	OTTHUMG00000032536	C20orf121	3
NO PROBE	NO PROBE	OTTHUMG00000032652	C20orf123	1
P	P	OTTHUMG00000032522	C20orf142	1
NO PROBE	NO PROBE	OTTHUMG00000032647	C20orf157	1
A	A	OTTHUMG00000032628	C20orf165	1
NO PROBE	NO PROBE	OTTHUMG00000032611	C20orf168	1
A	A	OTTHUMG00000032533	C20orf62	2
P	P	OTTHUMG00000032635	C20orf67	1
P	P	OTTHUMG00000033053	CD40	3
A	P	OTTHUMG00000033073	CDH22	1
P	P	OTTHUMG00000032487	CHD6	4
P	P	OTTHUMG00000033046	CSE1L	1
P	P	OTTHUMG00000032576	DBNDD2	13
P	P	OTTHUMG00000033072	DDX27	1
P	P	OTTHUMG00000032610	DNTTIP1	5
P	P	OTTHUMG00000033070	ELMO2	8
P	P	OTTHUMG00000046304	EMILIN3	1
A	A	OTTHUMG00000033041	EYA2	3
A	A	OTTHUMG00000032510	FAM112A	2
A	P	OTTHUMG00000032530	GDAP1L1	5
A	A	OTTHUMG00000032531	HNF4A	4
P	P	OTTHUMG00000032513	IFT52	2
A	A	OTTHUMG00000033037	JPH2	2
A	A	OTTHUMG00000033051	KCNB1	1

P	A	OTTHUMG00000032544	KCNK15	1
A	A	OTTHUMG00000033079	KCNS1	1
P	P	OTTHUMG00000032503	L3MBTL	13
A	A	OTTHUMG00000033056	LPIN3	2
A	P	OTTHUMG00000033052	MAFB	1
A	P	OTTHUMG00000033043	MATN4	3
A	P	OTTHUMG00000033044	MMP9	1
P	P	OTTHUMG00000033062	MYBL2	1
P	P	OTTHUMG00000033061	NCOA3	3
P	P	OTTHUMG00000032639	NCOA5	2
A	A	OTTHUMG00000032626	NEURL2	1
P	A	OTTHUMG00000032567	PI3	1
P	P	OTTHUMG00000032574	PIGT	3
P	P	OTTHUMG00000033065	PKIG	7
P	P	OTTHUMG00000033082	PLCG1	2
A	P	OTTHUMG00000033047	PLTP	2
P	P	OTTHUMG00000033078	PPGB	5
A	P	OTTHUMG00000032685	PREX1	2
P	P	OTTHUMG00000032667	PRKCBP1	5
A	P	OTTHUMG00000033077	PTGIS	1
A	A	OTTHUMG00000033040	PTPRT	7
A	A	OTTHUMG00000032524	R3HDML	1
A	P	OTTHUMG00000033055	RBPSUHL	3
A	A	OTTHUMG00000032546	RIMS4	1
P	P	OTTHUMG00000033083	SDC4	1
A	P	OTTHUMG00000032565	SEMG1	1
A	A	OTTHUMG00000032566	SEMG2	1
P	P	OTTHUMG00000033087	SERINC3	2
P	P	OTTHUMG00000032502	SFRS6	2
A	A	OTTHUMG00000033054	SGK2	6
A	A	OTTHUMG00000032638	SLC12A5	2
A	P	OTTHUMG00000033042	SLC13A3	4
A	P	OTTHUMG00000032657	SLC2A10	1
P	P	OTTHUMG00000033050	SLC35C2	7

P	P	OTTHUMG00000032710	SLC9A8	1
P	A	OTTHUMG00000033075	SLPI	1
A	P	OTTHUMG00000033048	SNAI1	1
P	P	OTTHUMG00000032624	SNX21	12
P	P	OTTHUMG00000032704	SPATA2	1
P	A	OTTHUMG00000032588	SPINLW1	5
A	A	OTTHUMG00000032585	SPINT3	1
NO PROBE	NO PROBE	OTTHUMG00000032604	SPINT4	1
P	P	OTTHUMG00000032691	STAU1	8
P	P	OTTHUMG00000033059	STK4	2
A	P	OTTHUMG00000032675	SULF2	4
A	A	OTTHUMG00000032623	TNNC2	2
P	P	OTTHUMG00000032552	TOMM34	1
P	P	OTTHUMG00000033057	TOP1	1
P	P	OTTHUMG00000085887	TP53RK	2
P	P	OTTHUMG00000033038	UBE2C	3
P	P	OTTHUMG00000033085	UBE2V1	9
A	A	OTTHUMG00000046331	WFDC10A	1
A	A	OTTHUMG00000046334	WFDC10B	1
A	A	OTTHUMG00000046330	WFDC11	2
A	A	OTTHUMG00000046412	WFDC12	1
A	A	OTTHUMG00000046333	WFDC13	1
P	P	OTTHUMG00000032594	WFDC2	1
A	A	OTTHUMG00000032614	WFDC3	5
A	A	OTTHUMG00000046411	WFDC5	2
A	A	OTTHUMG00000046354	WFDC6	2
A	A	OTTHUMG00000046342	WFDC8	2
A	A	OTTHUMG00000046332	WFDC9	1
P	A	OTTHUMG00000033071	WISP2	3
P	P	OTTHUMG00000032549	YWHAB	4
A	A	OTTHUMG00000032481	ZHX3	8
P	P	OTTHUMG00000032709	ZNF313	1
A	P	OTTHUMG00000032654	ZNF334	1
A	P	OTTHUMG00000032637	ZNF335	1

P	P	OTTHUMG00000032696	ZNFX1	4
P	A	OTTHUMG00000074023	ZSWIM1	1
P	P	OTTHUMG00000032627	ZSWIM3	1

Table A. 5 Expression profile of 103 protein-coding genes in human chromosome 20q12-13.2 region. The first and sec columns represent the expression status of the genes in HeLa S3 and NTERA-D1 cell lines respectively. A stands for “Absent” which means it is not expressed and “P” stands for “Present” meaning that it is expressed. M stands for “marginal” where no decision can be made for the expression status of the gene

Probe ID	HeLa-S3 STATUS	HeLa S3 Signal	NTERA-D1 STATUS	NTERA-D1 Signal	Matching Transcript 1	Matching Transcript 2	Matching Transcript 3	Matching Transcript 4	Matching Transcript 5	Matching Transcript 6	Matching Transcript 7
236514_at	P	49.3	P	75.8	ACOT8-001	ACOT8-003	ACOT8-004				
204639_at	P	118.6	P	124.3	ADA-001						
218098_at	P	364.8	P	351.6	ARFGEF2-001						
221485_at	P	359.2	P	857.2	B4GALT5-001						
238470_at	P	11.4	P	36.1	C20orf10-001						
209020_at	P	242.4	P	407.4	C20orf111-001	C20orf111-002					
231838_at	P	20.9	P	14	C20orf119-007	C20orf119-009					
228031_at	P	60.3	P	241.7	C20orf121-001						
226805_at	P	157.3	P	69.3	C20orf142-001						
1553960_at	P	20.4	P	103.1	SNX21-011						
218094_s_at	P	93.3	P	170.3	DBNDD2-001	DBNDD2-002	DBNDD2-003	DBNDD2-004	DBNDD2-005	DBNDD2-008	
222044_at	P	137.9	P	77.2	C20orf67-001						
35150_at	P	45.5	P	93	CD40-001	CD40-002	CD40-006				
225031_at	P	41.3	P	134.5	CHD6-001						
201111_at	P	788.9	P	1793.8	CSE1L-001						
221780_s_at	P	203	P	325.6	DDX27-001						
224825_at	P	240.6	P	490.1	DNTTIP1-002	DNTTIP1-004					
55692_at	P	72.4	P	221	ELMO2-001	ELMO2-003	ELMO2-006	ELMO2-007	ELMO2-018		
228307_at	P	140.1	P	119.4	EMILIN3-001						
218709_s_at	P	118	P	403.7	IFT52-001	IFT52-002					
225076_s_at	P	96.5	P	141.2	ZNFX1-001	ZNFX1-004					
213837_at	P	16.6	P	18.8	L3MBTL-003	L3MBTL-005	L3MBTL-010	L3MBTL-011			
201710_at	P	356.7	P	912.2	MYBL2-001						
207700_s_at	P	115.7	P	140.6	NCOA3-001	NCOA3-002					
225145_at	P	123.7	P	323.7	NCOA5-001	NCOA5-002					
217770_at	P	265.7	P	191.2	PIGT-001						
202732_at	P	72.7	P	560.5	PKIG-001	PKIG-002	PKIG-003	PKIG-004	PKIG-006	PKIG-007	PKIG-008
202789_at	P	81.3	P	463.7	PLCG1-005						
200661_at	P	667.7	P	416.6	PPGB-001	PPGB-002	PPGB-004				
209048_s_at	P	172.7	P	794.8	PRKCBP1-001	PRKCBP1-007	PRKCBP1-005	PRKCBP1-006	PRKCBP1-009	PRKCBP1-003	
202071_at	P	402	P	875.7	SDC4-001						

208804_s_at	P	456.8	P	997.9	SFRS6-001	SFRS6-002					
225037_at	P	121.3	P	232.7	SLC35C2-001	SLC35C2-002	SLC35C2-003	SLC35C2-005	SLC35C2-013		
212947_at	P	40	P	100.4	SLC9A8-001						
204433_s_at	P	67.8	P	168.7	SPATA2-001						
211505_s_at	P	273	P	606.2	STAU1-001	STAU1-002	STAU1-003	STAU1-004	STAU1-006	STAU1-007	
211085_s_at	P	31.8	P	52.9	STK4-003	STK4-006					
221471_at	P	564.1	P	624.8	SERINC3-001						
201870_at	P	271.9	P	438.5	TOMM34-001						
208901_s_at	P	825.5	P	1751.5	TOP1-001						
225402_at	P	132.2	P	530.6	TP53RK-001	TP53RK-002					
202954_at	P	1838.5	P	4331.4	UBE2C-001	UBE2C-004					
223186_at	P	268.4	P	279.2	UBE2V1-001	UBE2V1-005					
201002_s_at	P	445.7	P	1772.6	UBE2V1-004						
203892_at	P	14	P	652.8	WFDC2-002						
208743_s_at	P	1550.8	P	1713.6	YWHAB-001	YWHAB-002					
200867_at	P	179.2	P	514.5	ZNF313-001						
228223_at	P	46.5	P	68.9	ZSWIM3-001						
220540_at	P	82.1	A	0.4	KCNK15-001						
41469_at	P	34.6	A	12.9	PI3-001						
203021_at	P	2732.7	A	22.1	SLPI-001						
206318_at	P	12	A	1	SPINLW1-001						
205792_at	P	1709.3	A	14.2	WISP2-001	WISP2-002	WISP2-003				
224909_s_at	M	21.4	P	75.2	PREX1-001						
228737_at	A	2.3	P	26.9	C20orf100-001	C20orf100-002	C20orf100-004				
1569679_at	A	14.4	P	48.1	CDH22-001						
219668_at	A	12.1	P	92.6	GDAP1L1-001	GDAP1L1-004					
218559_s_at	A	1.6	P	543.9	MAFB-001						
203936_s_at	A	15.2	P	69.4	MMP9-001						
202075_s_at	A	6.5	P	293.4	PLTP-001						
210702_s_at	A	1	P	15.8	PTGIS-001						
221377_s_at	A	14.3	P	8.8	RBPSUHL-001	RBPSUHL-002	RBPSUHL-003				
233555_s_at	A	10.5	P	43.9	SULF2-001	SULF2-003	SULF2-005				
206442_at	A	0.8	P	66.1	SEMG1-001						
205244_s_at	A	1.4	P	87.6	SLC13A3-001	SLC13A3-003	SLC13A3-004				
219480_at	A	23.4	P	50.5	SNAI1-001						

220022_at	A	0.6	P	73.8	ZNF334-001						
78330_at	A	8.8	P	15.5	ZNF335-001						
226670_s_at	A	39.8	M	101.2	C20orf119-001						
1553752_at	A	1.2	A	0.7	C20orf165-001						
1570327_at	A	10.3	A	8.1	C20orf62-001	C20orf62-002					
232820_s_at	A	1.9	A	3.6	FAM112A-001	FAM112A-002					
209692_at	A	1	A	22.6	EYA2-003	EYA2-005					
243652_at	A	0.8	A	1.8	EYA2-003	EYA2-004	EYA2-005				
216889_s_at	A	4.7	A	4.4	HNF4A-002	HNF4A-003					
220385_at	A	1.9	A	0.9	JPH2-001	JPH2-002					
211006_s_at	A	9.3	A	18.3	KCNB1-001						
207366_at	A	4.1	A	46.2	KCNS1-001						
232966_at	A	2.3	A	4.1	LPIN3-001	LPIN3-003					
230283_at	A	10	A	40.4	NEURL2-001						
205948_at	A	1.9	A	3.6	PTPRT-001	PTPRT-002	PTPRT-003	PTPRT-004	PTPRT-005	PTPRT-006	PTPRT-007
234774_at	A	9	A	3.9	R3HDML-001						
233299_at	A	17	A	22.8	RIMS4-002						
216030_s_at	A	6.8	A	6.4	SEMG2-001						
220357_s_at	A	15.8	A	16.8	SGK2-001	SGK2-004	SGK2-005	SGK2-006			
210040_at	A	5.1	A	20.8	SLC12A5-002						
215503_at	A	1.1	A	1	SPINT3-001						
205388_at	A	5.7	A	2.3	TNNC2-001	TNNC2-003					
233913_at	A	2.5	A	2.5	WFDC10A-001						
1552999_a_at	A	0.9	A	4.2	WFDC10B-001						
1552608_at	A	1.1	A	9.3	WFDC11-001	WFDC11-002					
1553081_at	A	1.7	A	17.8	WFDC12-001						
1553052_at	A	7.4	A	1.5	WFDC13-001						
232602_at	A	41.7	A	2	WFDC3-001	WFDC3-002	WFDC3-003				
242204_at	A	3.1	A	12.1	WFDC5-001	WFDC5-002					
1552396_at	A	2.1	A	1.4	WFDC6-001	WFDC6-002					
1554156_a_at	A	3	A	0.7	WFDC8-001	WFDC8-002					
1552415_a_at	A	20.4	A	12	WFDC9-001						
217367_s_at	A	18.3	A	37.1	ZHX3-007	ZHX3-008					
223607_x_at*	P	52.2	A	50.5	ZSWIM1-001						
221024_s_at	A	0.3	P	181.6	SLC2A10-001						

207123_s_at	A	10.2	P	31.8	MATN4-001	MATN4-002	MATN4-003				

Table A. 6 Expression Profile of 198 transcripts in human 20q12-13.2 using Affymetrix U133 Plus 2 Expression Arrays on HeLa S3 and NTERA-D1 cell line total RNA. First column denotes the probe name, sec and fourth columns represents whether the gene is expressed or not in HeLa S3 or NTERA-D1 respectively. “A” stands for “absent and “P” stands for present. Third and fifth columns represent the signal intensity of the probe. Transcripts that are represented by each probe are listed in the remaining columns. ”*” The probe for SWIM1 gene is an x_at type probe although it has no cross-hybridization anywhere in the human genome.

HUGO Transcript ID	Response to Enhancer in HeLa S3	Response to Enhancer in NTERA-D1	Transcript Type	Number of Sp1 Binding Sites	Putative Sp1 Binding Site Coordinates Relative to TSS at +1												
WISP2-001	NR	NR	AT	1	-235												
FAM112A-002	NR	NR	AT	1	-99												
DDX27-005	NR	NR	AT	1	47												
SEMG2-001	NR	NR	RT	1	12R												
WFDC3-006	NR	NR	AT	1	-45												
PPGB-002	NR	NR	RT	2	-149	-67											
SGK2-004	NR	NR	RT	2	-73	76											
SPINLW1-001	NR	NR	RT	2	-130R	3											
UBE2V1-004	NR	NR	AT	3	-148	-113	-73										
PPGB-001	NR	NR	AT	3	-121	-40	-9										
C20orf100-001	NR	NR	AT	4	-118R	-93	-46	-40R									
SGK2-002	NR	NR	AT	4	-224R	-183	-3	35R									
UBE2C-004	NR	NR	AT	5	-182R	-116	-84	-63R	22								
ELMO2-003	NR	NR	RT	6	-226	-182R	-84	-63	-53	35							
R3HDML-001	R	NR	RT	3	-225R	-138	-20										
DNTTIP1-005	R	NR	AT	7	-275R	-239R	-152	-141R	-63	6R	14						
ELMO2-004	NR	R	AT	0													
PRKCBP1-007	NR	R	AT	1	-189												
C20orf100-002	NR	R	AT	1	58												
SEMG1-001	NR	R	RT	2	-157	-21R											
PRKCBP1-001	NR	R	RT	2	-123	-97											
SULF2-001	NR	R	RT	3	-170R	-144R	-79R										
SLC35C2-006	NR	R	AT	3	-183R	-32R	18										
C20orf111-002	NR	R	AT	4	-202R	-141	-89R	-79R									
IFT52-002	NR	R	RT	5	-146R	-123R	-41R	-21R	28								
RBPSUHL-001	NR	R	RT	6	-180R	-133	-114R	-97R	-79	-10							
ZNF313-001	NR	R	RT	6	-107	-100R	-87R	-70R	-56R	47							
ARFGEF2-001	NR	R	RT	7	-240R	-220	-140	-116	-38R	-29	-22R						
KCNB1-001	NR	R	RT	10	-318	-267	-221R	-211	-203R	-187	-162	-139R	-112	-71			
SLPI-001	R	R	RT	0													

ZNF335-001	R	R	RT	0													
MATN4-001	R	R	RT	0													
ZNF334-001	R	R	RT	1	-150												
ACOT8-001	R	R	RT	1	-189R												
ZSWIM3-001	R	R	RT	1	-51												
ZNFX1-003	R	R	AT	1	-213												
CSE1L-001	R	R	RT	2	-76	-51R											
SERINC3-001	R	R	RT	2	-142R	-87											
UBE2V1-003	R	R	AT	2	-108	-5R											
WFDC12-001	R	R	RT	2	-135R	-63R											
CDH22-001	R	R	RT	3	-124R	-50R	43R										
PI3-001	R	R	RT	3	-99R	-87	-33R										
NCOA3-002	R	R	RT	4	-278	-186R	-164	-104									
TOP1-001	R	R	RT	4	-243	-179	-131	6R									
PKIG-001	R	R	RT	4	-133R	10R	18	53									
SLC35C2-005	R	R	RT	4	-143R	-127R	-98R	-38R									
SPATA2-001	R	R	RT	4	-155R	-84R	-67R	64R									
SLC9A8-001	R	R	RT	4	-175R	-140R	-118R	-103R									
UBE2V1-002	R	R	AT	4	-220R	-72	-15	12									
KCNS1-001	R	R	RT	4	-149R	-71R	-39R	-11R									
CD40-002	R	R	RT	5	-111R	-69R	-56	-39	60								
HNF4A-001	R	R	RT	5	-241R	-220	-59R	71R	91								
L3MBTL-003	R	R	AT	6	-272	-222	-174	-124	-73	5							
NCOA5-001	R	R	RT	6	-228	-146R	-102R	-27R	-8R	17R							
SNX21-010	R	R	RT	6	-223	-75	-63	29	49	58R							
SFRS6-001	R	R	RT	7	-224	-81R	-69R	-26	-13R	3	25R						
C20orf111-001	R	R	RT	8	-169	-158	-147	-136R	-88	-66	-20	-13R					
C20orf121-001	R	R	RT	8	-136R	-71	-35R	-26R	3	17	64R	86R					
YWHAB-001	R	R	RT	8	-227R	-207R	-186R	-172R	-161R	-112R	-98	-50					
B4GALT5-001	R	R	RT	8	-80R	-60R	-48R	-29	-14R	37	77	119					
PLTP-002	R	R	AT	8	-208	-188	-168	-114	-71	-37	-19	38					
CHD6-001	R	R	RT	8	-292R	-252R	-228R	-218R	-186	-157	-136	-54R					
C20orf67-001	R	R	RT	8	-243R	-235R	-209R	-194R	-143	-32	12R	57					
GDAP1L1-005	R	R	RT	9	-270	-242R	-215	-197	-185	-147R	-106	-88	-20R				

PLTP-001	R	R	RT	9	-244	-235	-115R	-99R	-86	-63	-52	-17R	22R				
LPIN3-001	R	R	RT	9	-243	-218	-204	-182R	-150	-132	-103	-93	38				
C20orf119-017	R	R	RT	10	-177R	-167	-107	-83	-66R	-37	-21	-13R	9R	26			
ZNFX1-001	R	R	AT	10	-247R	-230R	-183R	-174	-131R	-118R	-88R	-78R	-33	64R			
UBE2V1-001	R	R	RT	10	-230	-206	-192	-176R	-161R	-135R	-68R	-29R	0	11			
SLC12A5-002	R	R	RT	10	-202R	-176R	-140R	-93	-44	33	45	56	68R	82			
PLCG1-005	R	R	RT	13	-244R	-226R	-194	-183	-171	-161R	-116	-95R	-74R	-26R	-14R	48R	69

Table A. 7 Sp1 binding site coordinates of 71 candidate promoter sequences together with their activation status in synergy with SV40 enhancer in HeLa S3 and NTERA-D1 cells. Coordinates are given relative to TSS assigned to +1 and R denotes that the motif is located on the opposite strand of the direction of the transcription.

	HeLa S3		NTERA-D1		Transcription Factor Binding Site Coordinates relative to TSS						
HUGO Transcript ID	Promoter Activity	Response to Enhancer	Promoter Activity	Response to Enhancer	Transcript Type	AP1	Sp1	NFKB1	CREB	YY1	MYC
SNX21-010	-6.603	9.658	3.573	29.175	RT		-223;-75;-63;49;58R;	-160;-160R;-105;	-233R;		
C20orf67-001	-15.971	0.786	-21.828	1.533	RT		-235R;-209R;-143;-32;			24;	
CDH22-001	-13.116	21.891	0.060	31.387	RT		-124R;-50R;43R;	15R;	-226R;-134R;		
CHD6-001	-8.434	8.059	0.633	4.102	RT		-292R;-252R;-228R;-218R;-186;-157;-136;			-33;	-338;-215R;-214;
DNTTIP1-005	-5.036	2.977	-50.950	-10.642	AT	-230	-152;6R;14;	-42R;	-257R;	-250R;23;	-258;
HNF4A-001	-12.118	52.349	-44.500	17.122	RT		-220;91;				
KCNS1-001	-45.841	12.657	0.145	16.039	RT			-158R;			
ZNFX1-003	-51.849	12.698	-66.822	17.642	AT			-208;		1;	
LPIN3-001	-33.133	65.502	0.627	40.684	RT	51	-243;-218;-204;-182R;-150;-132;	-125;			
MATN4-001	-43.471	15.320	0.037	32.157	RT					-91R;	
PI3-001	-60.178	21.980	-99.632	1.246	RT		-99R;-87;-33R;	-114R;			
PLCG1-005	-64.626	3.486	0.061	15.622	RT		-171;-161R;-116;-95R;-74R;-14R;48R;				-71;
PLTP-002	-24.783	9.568	1.754	11.203	AT	67	-208;-188;-168;-114;-37;				
R3HDML-001	-53.963	0.363	-71.446	-36.871	RT		-225R;-138;-20;			-100R;	
SLC12A5-002	-16.852	2.160	0.280	5.753	RT		-202R;-93;-44;33;45;			-241R;	
UBE2V1-001	-37.261	9.265	0.758	34.093	RT		-230;-206;-192;-176R;-29R;		-51R;		-187R;
UBE2V1-003	-10.475	9.693	12.976	48.084	AT		-108;-5R;	-12;	-122R;		
WFDC12-001	-59.061	2.959	-59.858	0.293	RT			-149R;		-224R;-121R;13;	-232R;-231;
ZNF335-001	-15.852	0.642	2.499	5.279	RT	14			-250R;-248;		
ZSWIM3-001	-3.640	0.937	2.195	30.418	RT		-51;			-28;	

Table A. 8 Transcription Binding Site Profile of candidate promoters that showed activity only in synergy with SV40 enhancer in HeLa S3 cells. Binding site coordinates are given in relative to TSS at +1. Transcripts coloured yellow denotes the ones that gave activity only in synergy to enhancer in both cell lines. All promoter activities are scaled from 0 to 100.

	HeLa S3		NTERA-D1			Transcription Factor Binding Site Coordinates relative to TSS					
HUGO Transcript ID	**Promoter Activity**	**Response to Enhancer**	**Promoter Activity**	**Response to Enhancer**	**Transcript Type**	**AP1**	**Sp1**	**NFKB1**	**CREB**	**YY1**	**c-myc**
C20orf100-002	-45.973	-19.161	-22.819	1.521	AT			-133;-7;-7R;	-219;		16;
C20orf111-002	-44.335	-17.244	-7.465	0.245	AT		-141;-79R;	-180;-56;			
C20orf67-001	-15.971	0.786	-21.828	1.533	RT		-235R;-209R; -143;-32;			24;	
ELMO2-004	-18.330	-2.533	-51.339	2.406	AT						38R;
HNF4A-001	-12.118	52.349	-44.500	17.122	RT		-220;91;				
ZNFX1-003	-51.849	12.698	-66.822	17.642	AT			-208;		1;	
PI3-001	-60.178	21.980	-99.632	1.246	RT		-99R;-87;-33R;	-114R;			
PRKCBP1-007	-58.050	-38.907	-76.783	2.558	AT		-189;				
RBPSUHL-001	-60.823	-27.486	-43.882	1.053	RT		-133; -97R;-79;	-126R;			
SULF2-001	-21.963	-3.153	-14.019	3.982	RT	-89	-79R;		-163;	-31R;	
SLC35C2-006	-47.335	-41.458	-43.198	0.584	AT	-21	-183R; -32R;18;				
WFDC12-001	-59.061	2.959	-59.858	0.293	RT			-149R;		-224R; -121R;13;	-232R;-231;

Table A. 9 Transcription Binding Site Profile of candidate promoters that showed activity only in synergy with SV40 enhancer in NTERA-D1 cells. Binding site coordinates are given in relative to TSS at +1. Transcripts coloured yellow denotes the ones that gave activity only in synergy to enhancer in both cell lines. All promoter activities are scaled from 0 to 100.

	HeLa S3		NTERA-D1			Transcription Factor Binding Site Coordinates relative to TSS					
HUGO Transcript ID	**Promoter Activity**	**Response to Enhancer**	**Promoter Activity**	**Response to Enhancer**	**Transcript Type**	**AP1**	**Sp1**	**NFKB1**	**CREB**	**YY1**	**MYC**
C20orf100-001	-50.180	-60.975	-65.563	-92.177	AT		-118R;-93;		-70R;		
FAM112A-002	-26.372	-93.324	-35.584	-73.521	AT						
DDX27-005	-24.872	-80.994	-29.494	-45.016	AT						
PPGB-001	-9.867	-36.907	-14.991	-0.713	AT		-121;-40;	33;		4;	-162R;
PPGB-002	-64.579	-100.0	-66.191	-90.674	RT		-67;				
SEMG2-001	-99.896	-69.880	-52.575	-65.664	RT						
SGK2-002	-19.588	-77.879	-27.874	-53.351	AT		-224R;	-86R;-46;-46R;		-9;43R;	
SGK2-004	-30.894	-58.041	-55.173	-20.228	RT			-103R;			
SPINLW1-001	-23.233	-34.263	-37.073	-42.911	RT		-130R;			19R;	
UBE2C-004	-51.751	-20.342	-61.014	-14.463	AT		-84;-63R;22;	-109;-69;-69R;		-35;	
UBE2V1-004	-10.536	-38.391	-23.355	-42.128	AT		-148;				-197R;
WISP2-001	-49.812	-64.179	-65.818	-80.927	AT		-235;			-81;	

Table A. 10 Transcription Binding Site Profile of candidate promoters that did not show in any cell line with or without enhancer. Binding site coordinates are given in relative to TSS at +1. All promoter activities are scaled from 0 to 100.

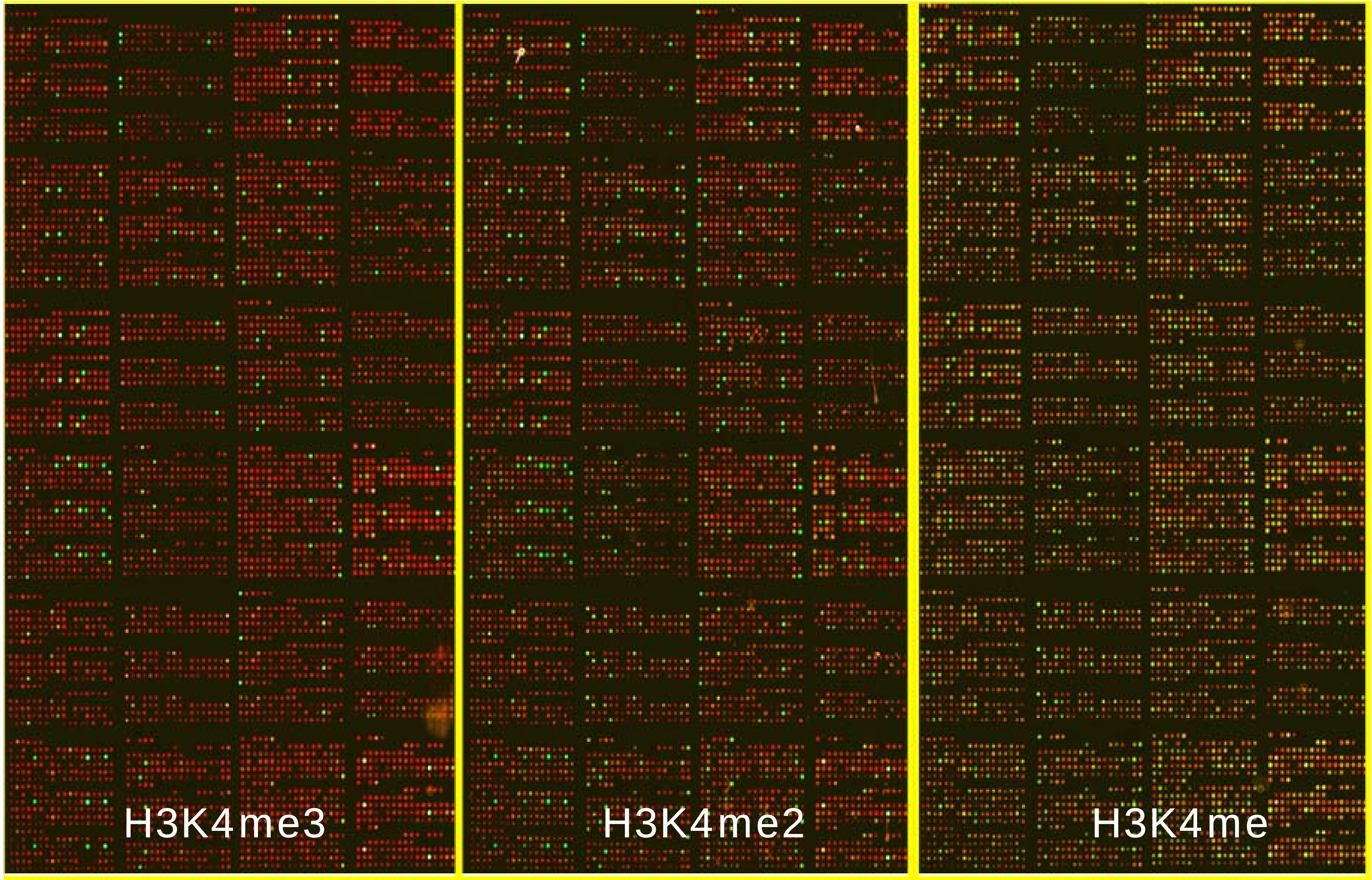


Figure A1. The scanned arrays of H3K4me3, H3K4me2 and H3K4me hybridizations. Spots that are enriched with the antibody bound DNA fragments are green.

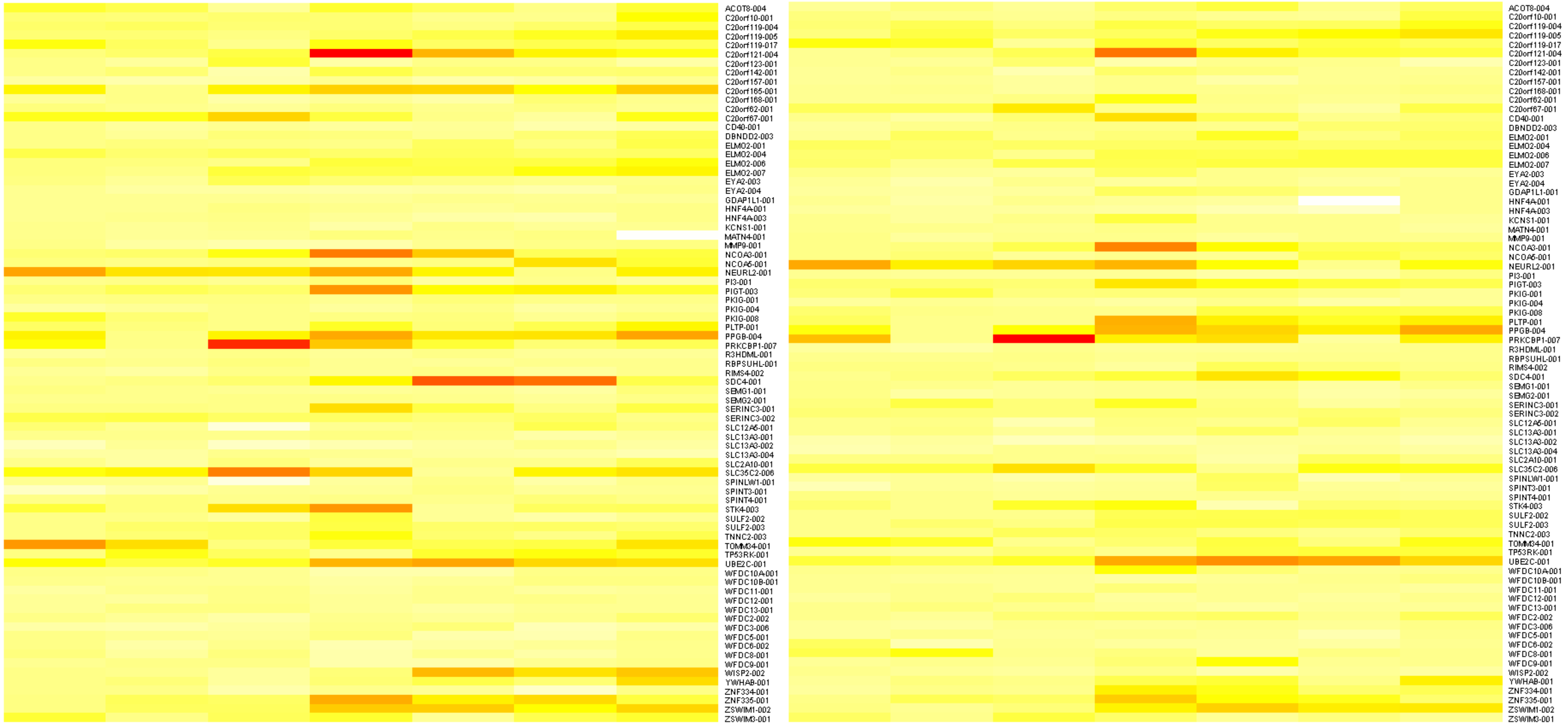


Figure A2. The heat maps that displays the spot intensities of all 79 transcripts represented on the array (see section 5.3), corresponding to 66 genes in HeLa S3 cell (left) and NTERA-D1 cells (right)

Cell Type	Start	End	Length	Spot St.	Spot End	Spot Information	polII	H3K4me	H3K4me2	H3K4me3	H3ac	H4Ac	CTCF
Jurkat	42671737	42672051	314	42670305	42672476		0.64	1.74	1.84	0.31	1.30	2.30	6.59
NCCIT	42671747	42672056	309										
FHs738Lu	43131582	43131910	328	43130169	43132851		0.42	1.79	0.26	0.04	0.11	0.83	0.08
FHs738Lu	43132024	43132330	306										
FHs738Lu	43158352	43159058	706	43158018	43159290		-0.12	1.00	0.45	-0.02	1.26	0.61	3.20
HepG2_PolyA-	43360758	43361072	314	43360657	43363032		-1.00	-0.56	-1.58	-0.32	-2.08	-0.43	1.73
HepG2_PolyA+	43401492	43402875	1383	43400608	43402868		2.08	4.18	0.86	0.10	0.92	0.62	-0.04
HepG2_PolyA-	43401512	43402875	1363	43402560	43405196		3.01	5.42	7.82	1.61	4.42	1.30	-0.06
HepG2_PolyA+	43403188	43404327	1139	43403941	43405207		0.50	1.04	1.79	0.47	0.61	0.42	0.16
HepG2_PolyA-	43403188	43404327	1139										
HepG2_PolyA+	43404378	43404818	440	43403941	43405207		0.50	1.04	1.79	0.47	0.61	0.42	0.16
HepG2_PolyA-	43404378	43404795	417										
NCCIT	43902419	43902983	564	43901741	43903780	3' end of ACOT8 gene	1.28*	0.02	-0.51	-0.17	0.04	0.00	0.01
A375	43902484	43902983	499										
Jurkat	43902484	43902983	499										
SK-N_AS	43902484	43902983	499										
HepG2_PolyA+	43902489	43902983	494										
PC3	43902489	43902983	494										
HepG2_PolyA+	44047015	44047509	494	44046483	44048188		-0.23	0.04	-0.16	-0.05	0.55	0.19	1.74*
A375	44047020	44047509	489										
NCCIT	44047020	44047509	489										
Jurkat	44047025	44047509	484										
SK-N_AS	44047025	44047509	484										
FHs738Lu	44047040	44047509	469										
U87	44047040	44047509	469										
HepG2_PolyA-	44047045	44047509	464										
PC3	44047045	44047509	464										
A375	44047581	44048239	658										
HepG2_PolyA+	44047581	44048239	658										
HepG2_PolyA-	44047581	44048239	658										
Jurkat	44047581	44048239	658										
NCCIT	44047581	44048239	658										
PC3	44047581	44048239	658										

SK-N_AS	44047581	44048239	658										
U87	44047581	44048239	658										
U87	44147844	44148155	311	44147270	44149246		1.83	2.73	2.91	0.30	0.93	0.20	0.93
PC3	44147850	44148180	330										
HepG2_PolyA-	44270148	44270483	335	44269828	44270823		2.11	3.67	8.15	2.50	4.79	3.57	-0.09
A375	44298702	44299421	719	44298319	44300488		-0.10	-0.00	-0.55	-0.01	-0.27	-0.03	-0.07
FHs738Lu	44298702	44299426	724										
HepG2_PolyA+	44298702	44299416	714										
HepG2_PolyA-	44298702	44299426	724										
PC3	44298702	44299421	719										
U87	44298702	44299426	724										
HepG2_PolyA+	44382036	44382579	543	44380475	44382544	5' end of GENSCAN37458	0.73	2.77	3.83	1.25	1.02	1.00	-0.05
Jurkat	44382036	44382579	543										
PC3	44382036	44382586	550										
A375	44382147	44382569	422										
U87	44382147	44382558	411										
A375	44820422	44820843	421	44820067	44822119		-0.20	-0.17	0.08	-0.03	0.08	-0.19	6.63
U87	44820422	44820843	421										
SK-N_AS	45034269	45034740	471	45034995	45037048		-0.13	-0.08	0.22	-0.06	-0.50	-0.18	3.28
HepG2_PolyA+	45035321	45035682	361	45034995	45037048		-0.13	-0.08	0.22	-0.06	-0.50	-0.18	3.28
NCCIT	45035321	45035677	356										
PC3	45035326	45035677	351										
A375	45035336	45035667	331										
Jurkat	45035336	45035682	346										
SK-N_AS	45035336	45035677	341										
FHs738Lu	45035351	45035677	326										
U87	45035356	45035662	306										

Table A 11a. The coordinates and cell specificity of the transfrags obtained in the study by Cheng J et al, that showed enrichments with antibodies used in this study in HeLa S3 cells. * denotes a signal which is reported here although it is below the threshold used in this study.

Cell Type	Start	End	Length	Spot St	Spot End	Spot Information	polII	H3K4me	H3K4me2	H3K4me3	H3ac	H4Ac	CTCF
Jurkat	42671737	42672051	314	42670305	42672476		0.18	0.69	1.58	0.26	0.68	0.17	4.10
NCCIT	42671747	42672056	309										
FHs738Lu	43158352	43159058	706	43158018	43159290		0.01	0.06	0.10	-0.11	0.52	0.02	3.20
HepG2_PolyA+	43401492	43402875	1383	43400608	43402868		0.66	2.08	2.03	0.23	1.18	0.60	0.03
HepG2_PolyA-	43401512	43402875	1363	43402560	43405196		2.24	1.00	13.85	12.06	2.65	0.13	0.05
HepG2_PolyA+	43403188	43404327	1139	43403941	43405207		0.27	2.03	3.18	3.15	1.22	0.06	0.59
HepG2_PolyA-	43403188	43404327	1139										
HepG2_PolyA+	43404378	43404818	440	43403941	43405207		0.27	2.03	3.18	3.15	1.22	0.06	0.59
HepG2_PolyA-	43404378	43404795	417										
NCCIT	43902419	43902983	564	43901741	43903780	3' end of ACOT8 gene	1.90	-0.16	-0.24	-0.32	0.41	0.21	0.10
A375	43902484	43902983	499										
Jurkat	43902484	43902983	499										
SK-N_AS	43902484	43902983	499										
HepG2_PolyA+	43902489	43902983	494										
PC3	43902489	43902983	494										
A375	43903806	43904471	665	43903837	43906350	3' end of C20orf161-001 gene	2.05	-0.26	-0.08	-0.05	0.33	0.27	0.13
FHs738Lu	43903806	43904116	310										
HepG2_PolyA+	43903806	43904471	665										
Jurkat	43903806	43904471	665										
NCCIT	43903806	43904471	665										
PC3	43903806	43904471	665										
SK-N_AS	43903806	43904471	665										
A375	44298702	44299421	719	44298319	44300488		-0.01	0.44	2.67	0.35	0.67	0.04	-0.04
FHs738Lu	44298702	44299426	724										
HepG2_PolyA+	44298702	44299416	714										
HepG2_PolyA-	44298702	44299426	724										
PC3	44298702	44299421	719										
U87	44298702	44299426	724										
NCCIT	44737903	44738536	633	44736072	44738637		0.61	1.22	2.02	0.20	0.84	0.33	-0.03
PC3	44764446	44764757	311	44764688	44766247		0.09	0.83	1.88	0.16	0.51	0.03	-0.01
A375	44820422	44820843	421	44820067	44822119		-0.16	0.26	0.34	0.03	-0.07	-0.05	5.39
U87	44820422	44820843	421										
SK-N_AS	45034269	45034740	471	45034995	45037048		-0.04	-0.06	-0.08	-0.09	-0.22	0.12	3.50
HepG2_PolyA+	45035321	45035682	361	45034995	45037048		-0.04	-0.06	-0.08	-0.09	-0.22	0.12	3.50

NCCIT	45035321	45035677	356										
PC3	45035326	45035677	351										
A375	45035336	45035667	331										
Jurkat	45035336	45035682	346										
SK-N_AS	45035336	45035677	341										
FHs738Lu	45035351	45035677	326										
U87	45035356	45035662	306										
A375	45321779	45322470	691	45319908	45322097		1.64	-0.04	-0.52	-0.53	-0.03	0.37	-0.12
PC3	45321779	45322470	691										
SK-N_AS	45321779	45322465	686										
HepG2_PolyA+	45321784	45322465	681										
U87	45321799	45322465	666										
NCCIT	45321804	45322455	651										
FHs738Lu	45321824	45322285	461										
Jurkat	45321824	45322465	641										
HepG2_PolyA+	45343564	45343975	411	45341963	45344242		1.78	0.52	-0.42	-0.48	0.18	0.21	-0.08
HepG2_PolyA+	45346308	45346638	330	45345737	45348010		1.63	0.01	-0.03	-0.02	0.21	0.32	-0.07

Table A 11b. The coordinates and cell specificity of the transfrags obtained in the study by Cheng J et al, that showed enrichments with antibodies used in this study in NTERA-D1 cells.

APPENDIX B

B1. Primers for Amplification of Putative Promoter Fragments (See Chapter 4)

Primer Name (20ace database)	Primer1	Primer2
ADA_80509	AAGGATCCCCGGCCACGCTCTTCTTAAC	AAGCTAGCGCCGGGGGGAAGGCAGCGCC
C20ORF100_79329	AAGCTAGCCATGCACTGAGTACAGGTAT	AAGGATCCCCCCATCCCATTCGTATAAG
C20ORF111_79334	AAGGATCCGCTGCAGCCAAAGCACAAAG	AAGCTAGCTCGGCCCCACTCCTTTCTTC
C20ORF121_79370	AAGCTAGCTTCCTTCCTTCTCTGAGCCT	AAGGATCCTCCCTCTCCTCAGCTCCCAG
C20ORF169_79455	AAGCTAGCTCCTGTCTGTCAGCGCTGTT	AAGGATCCGTCTGCATGAGGACGATCTG
C20ORF35_79441	AAGCTAGCCCGGCGTCGGGGGGCGCGGC	AAGGATCCTTGATGCTCTGCGCCTTTTG
ELMO2_80468	AAGGATCCTCTCTTCCCTGCGTCTTTCG	AAGCTAGCGTTTCTTTAAGAGGCCTCCA
KCNS1_80507	AAGGATCCCTCTCCTTCCTTCTTCCTTG	AAGCTAGCGTACCCTGGCACATACTAAA
KIAA1404_79649	AAGGATCCATGCCCAGTACATGGAGAAC	AAGCTAGCAACCGCAGACCACCATGGTC
LPIN3_80393	AAGCTAGCGGAGGTGGGGCAGGCGGTGG	AAGGATCCAGAGGGTCATCTCTTCTCCC
MATN4_80334	AAGGATCCAGCAAAGAAGGCCTCTCATG	AAGCTAGCCAGAGCAAGACTCTATCTAA
OVCOV1_80365	AAGGATCCGAATCTCCCCACCGGAAGTG	AAGCTAGCCAGACAATTACACGGCTGTG
OVCOV1_80366	AAGGATCCCACCGCCTTCCACAAAAAGG	AAGCTAGCGGGACCTGTGAGCGCTGCAT
PRKCBP1_79596	AAGGATCCATGGAATGCTGTCATGAGCC	AAGCTAGCATGTTTCTGAAAGTGTTTTC
TOMM34_79390	AAGGATCCCGAGTTGGGAGCTCCTTCCT	AAGCTAGCAAATGAGAGTCTTCGTTTGC
UBE2C_80312	AAGCTAGCTGATGCGGCCTACCACTCGC	AAGGATCCTTGAGCGTCTTCATCTAGGC
UBE2V1_80529	AAGGATCCAGATCACGCCGCCCAATGAC	AAGCTAGCCGACGCGCCGCGAGCTCCCA
UBE2V1_80531	AAGGATCCTACTCGCATGGATTCCTTGG	AAGCTAGCAACCATTTTAATCTTCCCCA
ZNF335_79553	AAGGATCCTAACTCTACCCGAAGCTCAC	AAGCTAGCATGACGTCACAACCACTTCC
ARFGEF2_79627_S	AAGAGCTCGCCTTAGGCTCCGCCCCGCT	AAGGATCCTCGGCTAGGATCTTCTCCAG
B4GALT5_80543_S	AAGGATCCTAGAGAGCCAGGCCGGGCCT	AAGAGCTCGATTACTCCATGCGGGCTGC
BA269H4.1_79623_S	AAGGATCCCTGGGCGCCTCCATTCTAG	AAGAGCTCGTGGCACTTGGCAGACGGTC
C20ORF100_79330_S	AAGAGCTCGGCTTGCGTGCCTGGAACCC	AAGGATCCACGAAAAGCCTGCGAGGTTG
C20ORF122_79408_S	AAGGATCCTCACAAGAACGAGAGACACC	AAGAGCTCCTGGTGGCGCCTAATGGTCA
C20ORF130_80415_S	AAGGATCCTGCCGGTATTCAGCCTTTAC	AAGAGCTCCACGGAGGCTGGAATCAGAG
C20ORF161_79534_S	AAGAGCTCTGGGACAGATTGGAGATGAC	AAGGATCCTGCGTCCCACGGTGCATTCA
C20ORF164_79540_S	AAGAGCTCCATACACAGAACCTGACTCT	AAGGATCCTGGCCTATTCTGGAACTCAG

C20ORF169_79457_S	AAGAGCTCGAACTCCTGACCTTGTGATC	AAGGATCCTATTCCTACTCCAGCAGACG
C20ORF170_79482_S	AAGGATCCTTCCCTATGGAGACAGCTTC	AAGAGCTCCTCCGTGTGGGCATCAGAGC
C20ORF1901_01027_S	AAGGATCCCCATTCTTGACGCGGAGTAC	AAGAGCTCCCAAGCCCGCCTCGCGCTGC
C20ORF35_79439_S	AAGAGCTCGCATCCGAGCGAGCGGAGAC	AAGGATCCTCTGACTCAGTCCGTCCGCT
CDH22_80491_S	AAGGATCCTTCCTGCGCAGAAAGGATGC	AAGAGCTCTAGGAGATTGGGAGTACAGG
CHD6_79270_S	AAGGATCCCAAGGACACCCTAACCCATC	AAGAGCTCCCGCCCCCTCGGCCCGCGGG
ELMO2_80469_S	AAGGATCCTGGTCGATTTCAAGGAGCTG	AAGAGCTCTGTCCCATTTCTTCTTGCTG
EYA2_80324_S	AAGAGCTCGTGAGCGCGCTCTGCTCGGC	AAGGATCCTACTCACCGCGCGCAGTGGAG
EYA2_80326_S	AAGAGCTCTGTTCTAGCTGAGTCTCTCA	AAGGATCCTTCCTGAGCCTGCAGCCGAG
HNF4A_79362_S	AAGAGCTCCTTCTGCTCCGGCCCTGTCC	AAGGATCCTTCACGCTGACCATGGCCAA
KCNB1_80374_S	AAGGATCCTGCGAGGCAGAAAGCGAAAG	AAGAGCTCGGCTCCAGGGGCATCTCTTG
KCNK15_79378_S	AAGAGCTCGGGAGGAAGGGAGGGGAGGG	AAGGATCCTCCCAACCTGCTCCGTGTG
KIAA1404_79647_S	AAGGATCCGCGAGGCAAGGAAGAATCAG	AAGAGCTCGGTCCGCTCCCCTCAGCCGC
L3MBTL_79296_S	AAGAGCTCCTCTGGGACTGGGCCTGTGG	AAGGATCCTACTCGCTTATCCGGAACTG
C20ORF142_79342_S2	AAGGATCCATTATTCCTAAGGAAACTTC	AAGCTAGCCTCCTCTCCCCGAGCCTC
R3HDML_79344_S	AAGAGCTCAATATGACTTGGGGTAAGGT	AAGGATCCCTGAGCGTTCTCCAACAGTG
TDE1_80544_S	AAGGATCCTCCTGAGGCTGCTTTCTAAC	AAGAGCTCGCCTCTCTGTACACTTCTGT
TOP1_80397_S	AAGAGCTCCTGCTTGGGGTGGGACGCCG	AAGGATCCAGACTCCAGAAACGGCTGAG
YWHAB_79386_S	AAGAGCTCCCCACCTCCTTTTCCCGCTC	AAGGATCCTGACCTCTTCCTTCCTAAGC
WISP2_80484_S2	AAGAGCTCTTCGGTGTGCCTCTCATGTC	AAGGATCCAAGCTGTCACAGGCTCTTGG
CSE1L_80345_S	AAGAGCTCTGCCAAGCCCTGACATCAAA	AAGGATCCTTCAAACCCCGGCAAAATGG
ADA_80509_S2	AAGGATCCAAAGTTGCTGGAGGAGCCGG	AAGAGCTCTCGGTGGCCGCTCGGCTTTC
B4GALT5_80543_S2	AAGGATCCGATTACTCCATGCGGGCTGC	AAGAGCTCTCCCGCAGCTCCCCGTCC
BA269H4.1_79623_S2	AAGGATCCTGGGCGCCTCCATTCTAGC	AAGAGCTCTGCACGCGCAGAGCGCCTCT
C20ORF10_79460_S2	AAGGATCCTGAGCATGGTTTGGGAATCC	AAGAGCTCACCTGCAGCTGTGACATCAC
C20ORF111_79333_S	AAAAGCTTGAGCGTCTCTCCCAAATCTC	AAGAGCTCGGTCGGGAGCAAATCCACGC
C20ORF119_79391_S2	AAGCTAGCTCCCAGTCCTCGGCTCTTTC	AAAAGCTTAAGCAGGAGCCGGGTCACC
C20ORF138_79493_S2	AAGAGCTCAAAAAATTCAATCCTCTGTG	AAGGATCCTTTGACCAGGCACTGCTGTG
C20ORF146_79490_S2	AAGAGCTCATGACTCTGATCAGATCGTG	AAGGATCCCAGTTAATTCTAAGCAACTC
C20ORF162_79537_S2	AAGCTAGCAGTAACGTGTCCTTCCAGTC	AAGGATCCACTTTCACCCCTTCAGGAAG
C20ORF163_79539_S	AAAAGCTTGACCTTGACCCGGCATTGAC	AAGAGCTCCTCCTCAGCGCTTCCGTGCA
C20ORF165_79541_S2	AAGGATCCTTCCATCTCTGGTCATCTGC	AAGAGCTCCGGTTTGTGAGGGAATAGTG
C20ORF167_79502_S2	AAGAGCTCACAACTCCGCTGGACTCTGTC	AAGGATCCCCCCAGCCTACAGAAAACTG
C20ORF167_79503_S2	AAGAGCTCAGATCCGTCACCATCTCTCC	AAAAGCTTTTGGCGTTTTCGCAGCCCGA
C20ORF169_79455_S2	AAGCTAGCGTCTGTCAGCGCTGTTTTGG	AAGGATCCTAGCTGCGGAACTGACCCG

C20ORF169_79457_S	AAGAGCTCGAACTCCTGACCTTGTGATC	AAGGATCCTATTCCTACTCCAGCAGACG
C20ORF170_79482_S	AAGGATCCTTCCCTATGGAGACAGCTTC	AAGAGCTCCTCCGTGTGGGCATCAGAGC
C20ORF190_101027_S	AAGGATCCCCATTCTTGACGCGGAGTAC	AAGAGCTCCCAAGCCCGCCTCGCGCTGC
C20ORF35_79439_S2	AAGAGCTCCATCAAGTGTGCGTGGGCAG	AAGGATCCTGGGGCTGCGCCTCCAGGG
C20ORF64_79581_S2	AAGGATCCGCCAGCTCTGAGTCTCAATT	AAGAGCTCAAATCTTCGGTGACTCTCGC
C20ORF65_79314_S2	AAGGATCCAATCTGCTGAGTGGGATTCG	AAGAGCTCTGCTATAGTTCTGTCTCGGG
C20ORF67_79550_S2	AAGAGCTCTGGTACCGCCTCTGAGGGAC	AAGGATCCTTCTAAGACAGCCCCCACGC
C20ORF9_79318_S2	AAGAGCTCTTTACCTGAGGCCCAGCGAC	AAGGATCCTTTGGGCACTTGGGTCCTAC
CHD6_79270_S2	AAGGATCCTAGCCAGCTACCTGAGAGAC	AAGAGCTCTTCGCCGGTCCAAAACACAG
DDX27_80489_S2	AAGAGCTCGCAGAAATGTGTGCTAAAGC	AAGGATCCAGCTGCATACTGTCTTGCAG
DJ1049616.1_79606_S2	AAGGATCCAATCGCAGGGCAAACCGAG	AAGAGCTCTCCCCTTAGCTAGCAACTCG
dJ688G8.2_79485_S2	AAGGATCCTATCTCGGGTTCTAACTGCC	AAGCTAGCACCTCTTTTGCTGCCACTAC
dJ688G8.5_79486_S2	AAGGATCCCAGCCAAATTCTACCAGAC	AAGCTAGCTCATTTTTTTGCGGGGCAGG
EYA2_80324_S2	AAGAGCTCTTGCAGGGAGGATGTGCTGC	AAGGATCCGTTGCTGTCTCTGCCGTTGC
GDAP1L1_79360_S2	AAGCTAGCATGGGCCACCAGCTGCAGTT	AAAAGCTTTGGCCCAATACACTGAGCTG
HNF4A_79363_S2	AAGAGCTCACCTGCCCTACCCTGGGCG	AAGGATCCCGTGAGTCATGATGCCTGCC
JPH2_80307_S2	AAGGATCCCTTCTTGCACCAAGTTCTCC	AAGCTAGCAGAACACCCAGTCCTGGAAC
KCNB1_80374_S2	AAGGATCCCGTCTTCTCACCTCCATCCC	AAGAGCTCGAGTTTCAGCACTCTAAGGG
L3MBTL_79296_S2	AAGAGCTCTTCGCCTCATGCCAGCTCAC	AAGGATCCAGTTGAAGCACTCCTAGGCC
MMP9_80337_S2	AAGAGCTCTGCCAGAGGCTCATGGTGAG	AAGGATCCCCCACAAGCTCTGCAGTTTG
MYBL2_80408_S2	AAGAGCTCGTCAGGACCCGGGCTGCTC	AAGGATCCCACAACTCCGAAGTAGCGGC
NCOA3_80404_S2	AAGCTAGCTTAAATCGGAAACTCGCCGC	AAAAGCTTAAATTAAGGGCAGGGCTAGG
NCOA5_79559_S	AAGGATCCTAGGACAAAGGCGCCACCAAC	AAGAGCTCTGAAAACAGAAAAAAAGTAC
OVCOV1_80363_S	AAGGATCCATTCTTACCTTACACAATGG	AAGAGCTCGGCATCTTAAGTTTTAGATA
OVCOV1_80365_S2	AAGGATCCGCACACCCACTCACCTCGG	AAGCTAGCGGTTCCAAAGCAAGCCCCTG
PI3_79418_S	AAGAGCTCTCTACTCTGTGAGAAAGTAG	AAGGATCCACACCACCACGATCAAGAAG
PIGT_79434_S	AAGAGCTCCATGAACTAGGTGCGGCCTC	AAGGATCCATGACAAGTTCCTCCCGCAG
PKIG_80411_S2	AAGAGCTCACTGTCCCCCTTTCCGTATC	AAAAGCTTTCCTGCTCACCTGCGGTCTC
PKIG_80412_S	AAGAGCTCAAGGAAGGATATTAGGCAAG	AAGGATCCTCCTCAAAATCCAGTGGTCC
PLCG1_80514_S	AAGAGCTCGCGCTCCCGCCGCCATCGCG	AAGGATCCAGCCGTTGGCGCAAGGGGAC
PLTP_80348_S2	AAAAGCTTTCACGTGGGATGGCGGGCA	AAGAGCTCGTCCCAGCAAAGTGGGATTG
PLTP_80349_S	AAGGATCCTCCTTGACTCCACCTTTCTG	AAGAGCTCGGCAAAGAAGGCCACTTCTA
PPGB_80497_S2	AAGAGCTCAGCTCTTTCTCCTCGATCTC	AAGGATCCGGGAAATGCCTTACAAGGTC
PPGB_80498_S	AAGAGCTCGAGCCAGGAGGGGTCGCTGC	AAGGATCCTCCTCGTGGAACCATATCTG
PPGB_80500_S	AAGAGCTCACTCCCTTCCCCGAGCCTCT	AAGGATCCTGCCAGCTCACCTCTGCTC

PRG5_79407_S	AAGGATCCATTAACACCTGCAGCCTCAG	AAGAGCTCCCTCAACAACCTTCACCCTG
PRKCBP1_79590_S	AAGGATCCAAGTCGAGCTTACCTCTGTG	AAGAGCTCGGGCCCCTTTCCCAAGTTTT
PTE1_80338_S	AAGGATCCTCTAGTTCAATGCTGCAGGC	AAGAGCTCGGTGTATTTGCTATCTGACA
PTGIS_80496_S	AAGGATCCAGCAACAGTGCGGCCAGGAG	AAGAGCTCATCCCTTCCGCCCCCTCCCC
PTPRT_80315_S	AAGGATCCAGCTGCAGCCTCAGGAGCAG	AAGAGCTCAACGGCGGCGGCGTTAGGAC
RBPSUHL_80389_S	AAGAGCTCGCGTGGTGGCGTGGCAGCGA	AAGGATCCTCTGGACGAGTCCAGTGCTG
SDC4_80515_S2	AAGGATCCCTGCCTGGCAGTGGGTCAAA	AAGAGCTCAGCAGCGCGAACAGACGGG
SEMG1_79416_S	AAGAGCTCTGTAAAATTAAAGTGATCTG	AAGGATCCCTCCACTCACCTTTTTGTCC
SEMG2_79417_S	AAGAGCTCAATTACTTTTGTAACCTGAA	AAGGATCCATGTTCCTTGAGTGGGTGTG
SFRS6_79292_S2	AAGAGCTCTCTGCGGCTGGATTAGAGAC	AAGGATCCAAGCCAGTAACCGCGGTGC
SGK2_80383_S	AAGAGCTCGCCAGTCTTGGGTCTCTCTG	AAGGATCCAGCTCCAGAAGTCCAAGATG
SGK2_80385_S	AAGAGCTCAGGCTCAGGGAGCTGAAGTC	AAGGATCCCACCTGAAAACTCCAGGTTC
SLC12A5_79558_S	AAGAGCTCACGCAATCCCCCAGTTTTTG	AAGGATCCCAGGTTGTTTAGCATGGTGG
SLC13A3_80329_S2	AAGGATCCAGTACGCGCCTTAATCCTCG	AAGAGCTCACACCTTCTTGGCCGCTGCT
SLC2A10_79578_S	AAGAGCTCTGGGCGAGGGAGGGGGTCCT	AAGGATCCTGTCTACACCCTGGGAAGGG
SLC9A81_06483_S	AAGAGCTCCCCTGTGATGGGGAAAAGCC	AAGGATCCAAAAAGCCCACTCACTCCTC
SLPI_80494_S2	AAGGATCCGACTTCATGGTGAAGGCAGG	AAGAGCTCTAATGGCCTGGGATCTTGTG
SNAI1_80350_S	AAGAGCTCGCGCTGCGCCAGCGAACCCC	AAGGATCCGTAGTTAGGCTTCCGATTGG
SPATA2_79658_S	AAAAGCTTCGAGGGAAGCAAGCGAGAG	AAGAGCTCGCCTGTTTACTTCCGGGTCC
SPINLW1_79467_S2	AAGGATCCGAGGCTCAAAAGTCCAGAAG	AAGAGCTCTCTGAAGGTAGCCTGGAAAG
SPINLW1_79468_S	AAGGATCCTAGTATCACTTACCCCGGAG	AAGAGCTCAGTTACTTAAATTTCATGGG
STAU_79633_S	AAGGATCCAAACGCTGAAGAGCCGCTCA	AAGAGCTCAGTCCAGCAGGCAGCGCGAA
TIX-1_79257_S2	AAGGATCCTCTGAAGGCTGGAATTCTGG	AAGAGCTCAAGGCAGAAGCTCATCTTTC
TIX-1_79258_S	AAGGATCCCAGCTGCACCAACCGACTC	AAGAGCTCACTCGAGGAGGCTGACTGATG
TIX-1_79259_S	AAGGATCCATGCTGGCATCTTGCAACAC	AAGAGCTCTCATATGCGGGAAGGCTATG
TNFRSF5_80377_S	AAGAGCTCCCTGGGGGCAAAGAAGAAGA	AAGGATCCGGGCAAAAACAACTCACAGC
TNNC2_79522_S	AAAAGCTTACTGGGATGACTCTGCGGCA	AAGCTAGCCGGAGGGGCTCAGGACCCTC
TNNC2_79524_S	AAGGATCCAAGTCCCCTCTTGTCCTTAC	AAGAGCTCACTCCAAGGCAGTGGAACAG
UBE2V1_80530_S	AAGGATCCTCTTGCGTCGCTCTTGCTTG	AAGAGCTCTCACAATCGCGCCTTCCCAA
UBE2V1_80532_S2	AAGGATCCATGTTTAGCAGCGGCAGCAG	AAGAGCTCGGGTCATTTCGGCTCATCTC
WFDC2_79476_S	AAGAGCTCGTCAGGAGGGAGGCGTGGAC	AAGGATCCACTCACCTGAGACTAGGGTG
WFDC3_79509_S	AAAAGCTTTAAGTGTAACTGTCCTGGGC	AAGAGCTCCCTCACCGGGGGCGCCCTCA
WFDC3_79513_S	AAAAGCTTTACCTCTACCCTTTTCCAGG	AAGAGCTCAAGCCACCATTGCGGCACTG
ZNF313_79663_S	AAGAGCTCAGTGGTGCCGAACTAACGAT	AAGGATCCTGTACCGGCTTCTCGTACAC
ZNF334_79575_S	AAGGATCCCCTATAGGTTGTGGGTGCAC	AAGAGCTCCGGTGAAACGGATATGAAAC

B2. Primers for Amplification of 600 bp constructs (see section 4.3.1)

Primer Name (20ace database)	Primer1	Primer2
C20ORF100_79329	AAGCTAGCATATCTCCTCCCTAGTG	AAGGATCCCCCCATCCCATTCGTATAAG
C20ORF111_79334	AAGGATCCGCTGCAGCCAAAGCACAAAG	AAGCTAGCAGGTAAGTGCCCTGACGTGAC
C20ORF121_79370	AAGCTAGCGCTCCAGAGCAAGAATCAGG	AAGGATCCTCCCTCTCCTCAGCTCCCAG
C20ORF35_79441	AAGCTAGCAGTATCTCAGAGAGATCCCC	AAGGATCCTTGATGCTCTGCGCCTTTTG
KIAA1404_79649	AAGGATCCATGCCCAGTACATGGAGAAC	AAGCTAGCGCACGTTCATTGTCGTTCTG
LPIN3_80393	AAGCTAGCTTACAACTGTGGGCTATGCG	AAGGATCCAGAGGGTCATCTCTTCTCCC
MATN4_80334	AAGGATCCAGCAAAGAAGGCCTCTCATG	AAGCTAGCAGAGCATTTAACACTGTGCC
PRKCBP1_79596	AAGGATCCATGGAATGCTGTCATGAGCC	AAGCTAGCTGGAAGGTAAGCAAACAGGC
SLPI_80494	AAGGATCCAGTGACTCTGATGGCCAATG	AAGCTAGCGGAGCTCTTCTTCAGCTTTC
UBE2V1_80531	AAGGATCCTACTCGCATGGATTCCTTGG	AAGCTAGCGTAAAAAGGCAAACCTGCCC
CSE1L_80345	AAGAGCTCAGAGGAACAGGAAGAAGGTG	AAGGATCCTTCAAACCCCGGCAAAATGG
R3HDML_79344	AAGAGCTCTGGTCTCTCGTGCTTCTCTG	AAGGATCCCTGAGCGTTCTCCAACAGTG
TDE1_80544	AAGGATCCTCCTGAGGCTGCTTTCTAAC	AAGAGCTCGAAATGGGACGTTCTCACTC
TOP1_80397	AAGAGCTCGTGGGCGTGAAATAATCCAG	AAGGATCCAGACTCCAGAAACGGCTGAG
YWHAB_79386	AAGAGCTCTTTCTGTTCTTCCCTGGCTC	AAGGATCCTGACCTCTTCCTTCCTAAGC
C20ORF167_79503	AAGAGCTCTGACAGAGTCCAGCGGAGTTG	AAAAGCTTAACCTGGACATCCCTGCTTC
ID1_long	CCGAGCTCAGGAGCTGCAAATTCAAG	AAGGATCCAGCCCGAAGCAGATAC
ID1_short	AAGAGCTCTTCCAGAGGAGCCCAG	AAGGATCCAGCCCGAAGCAGATAC

B3. Primers for Negative Control Fragments (300 bp) (see Chapter 4)

Primer Name (20ace database)	Primer1	Primer2
stSG116340	GCTATAGCACCATGCTGCAG	GGAGCAGAAGGGAAGGATCT

stSG116341	CCCTACCAAAAACAGAAAACA	GACAATGTCAGCTCCTGCTG
stSG116342	AATCCTCTGTTTCCCCGTTT	TGTTTGGTGAGTCTCTGGGA
stSG116343	AAACTGAAAATGCATTATTGGT	ACAATTAAAGTGAGGCAAGGG
stSG116344	TCCCATTTTCCCTTACTCCC	TATTTGTGGTGGCCAGGAAT

B4. Primers to check success of cloning to pDrive subcloning vector (see section 2.1.2 and 2.1.3)

Primer Name (20ace database)	Primer1	Primer2
pDriveinsertcheck	GTAAAACGACGGCCAGT	AACAGCTATGACCATG
stSG1163639	TACTAACATACGCTCTCCATC	TTCCATCTTCCAGCGGATAG

Primers for Real Time PCR (see section 5.2)

Primer Name (20ace database)	Primer1	Primer2
C20orf121_RT_01	GCTGGTCTCGAACTCCTGAACT	AAAACCCCCCAAAACAGCC
C20orf121_RT_02	CCACCACGCCCAACTAATTTT	GGAGTATGATGAGACCACCCCA
C20orf121_RT_03	CCGGCCTTACCTTACGTTTTAA	ACTGACTGATCCTCCTGCTTCC
C20orf121_RT_04	AGAGGTGACAACAGAAGCCAGA	GTTTGTAGAGGCCACCTGCAT
C20orf121_RT_05	GTCAAAACTCATGATTCCCAGG	AAGAGGAGTCCCAAGAAGATGG
C20orf121_RT_06	AGGTAAGGAAACGGAAGCACAA	CCCACTATGGCTGCTATCAGGA
C20orf121_RT_07	GGTTTGAAGCCCTTTACTGCCT	CTGCCATTTTCCAAGAACATGC
C20orf121_RT_08	ACCTCCTTCCTTCTCTGAGCCT	AGCTGCGGACCTGGAGTG
C20orf121_RT_09	TTCTCTGAGCCTCAGTTTCCCC	GAGCTCCGAGCTGCGGAC
C20orf121_RT_10	AGATGGCTTCCTGACCTCTCCT	TGGTAACGGTGAAGCCAAAGA
C20orf121_RT_11	CTTCACCGTTACCAATGGTGTC	TGGAGCTTCATGTTGACTCTCC
ADA_RT_01	GGCCTGATCATACAGCTGAGCT	TCTGCTTCTGCCCCTGACTTGA
ADA_RT_08	GTCTCTGCCGGCTCGGTG	CCCGGCCCGTTAAGAAGA
ADA_RT_07	GGCCACGCTCTTCTTAACGG	CAGGAAATGCGCGATCCA
ADA_RT_09	TGGAATTCTGGACCCGGCGT	AAGGCAGCGCCCAGCGAG
ADA_RT_02	CACACCAGCGCTATTCCGAATA	TGTCTGGCTGAAGTTCATTCCG
ADA_RT_03	CGGAATGAACTTCAGCCAGACA	CCATTGTATGCAGTTTTCCGCA

ADA_RT_04	TGTTTAGACACATGCATCGGTG	ACCCGCGTGTTATTTCCCT
ADA_RT_05	GTGCCCTGCTAAGTTTTGGATT	TCATGTGAGGTCAGGTGTTCG
ADA_RT_06	GCTTCCCAAGGCGTGATTA	AGCTACCATCCACCCCAGTACT
ZNF335_RT_01	GCCTAACTCTACCCGAAGCTCA	GAGAGGAACGTGGCTACGAAA
ZNF335_RT_02	AGGAACCCATCGGCCTATTGT	TCCGGAACACTGAAAATGTGTC
ZNF335_RT_03	TTTGGCACTGAGCCCACTATG	CGTAACACTCCGCCGAGATAAA
ZNF335_RT_04	TGCAACCCAATCGAGCTCT	CCCATACCCAGACCTCAGTTTC
ZNF335_RT_05	GCTGGTCCAAAATCACATCCA	GCCATCCGTACTTAGCTCCTGA
ZNF335_RT_06	CTTGATCCCAGGAGTTCCAAAC	CACCATGCCTGGCTAATTTTTC
ZNF335_RT_07	AAAATTAGCCAGGCATGGTGG	AGCCTCAACCTCCCAGACTCAA
c20orf142_RT	CCCTGGAGAATTGGAGCAGTAG	TCCGAGGCTCCAAGAAAGATC
YWHAB_RT	AACCCTGCCTTTCGCTCTT	CCCAAACCGTGTTCTCATTG
TOMM34_RT	CGAATATCTCCGCCCAACA	GCCTAAGTTCCAAACCGAGCA
SLPI_RT	GCCAATGCCAAACCTCACTATT	GGATTTCTGGTCTCTGGCTACG
PIGT_RT	AGCGGAAGTCACTTCTCACACG	CCATGAAAAGCCGGTATCCAAT
H3K27me3_Pr11	ACCACGCCCGGCTAATTTT	GGCGGATCATGAGGTCAGG
H3K27me3_Pr12	CTGACCTCATGATCCGCCC	GCCAGGTACAGTGGCTCACG
H3K27me3_Pr13	TGCAGTCTCCATGTGGCAAG	CCTGCCAGCCTTAACTCCATC
H3K27me3_Pr14	GATGGAGTTAAGGCTGGCAGG	CCACATCCAGGCATCCCTC
H3K27me3_Pr15	CTGCTGTGTGACCCCTGGTTA	CCGCTCTGCTCCAAGCACT
H3K27me3_Pr16	GAAGAAAGTGCCCTGCCCA	TACAGGCGTGAGCCACTGC
H3K27me3_Pr17	GTCAGGAATTCGAGACCAGCC	CACCCACCATCACACCCAA
H3K27me3_Pr21	CTGAAAGTGACGACACAGCCA	CCTGAGGTGAAAGGATGTTTCC
H3K27me3_Pr22	CATAACTCCAGCCCCCTTTTTT	CCTGTTCTCAGCACATCATGGT
H3K27me3_Pr23	CCAAAGACTTTTCCTGAGCACC	GGTGCAATCGTCAGCTCTTTCT
H3K27me3_Pr24	CACTTGGAATTTGATCAGCGG	GGCCCCATAGGTACATTTCCTT
H3K27me3_Pr25	CCCTGGTGCCCAAGATACTTTC	AGCAAAGAGGCTAGGCAGGAGA
H3K27me3_Pr26	ACTTGAGTGCCCTACAGTGCCT	GAATGGTTGGATGATGGTGCTA
H3K27me3_Pr27	ATCCTAGGTTCTTCCTCTCCCC	CCTGACCAGAGCAATGTAGAGC
CTCF_Pr11	CCCTCACGGTTCCTATTCTGC	TGACTCACCACGCCCCATA
CTCF_Pr12	CGACACGTCACTCAATCGCTT	CAGTAACCCAGCGAGGTGGA
CTCF_Pr13	CTCACACACAAGCGCCGTG	GGCCAAGGTGACCAGTTGG
CTCF_Pr14	CAACTGGTCACCTTGGCCA	GCCATGCCAACCA CTTACCT
CTCF_Pr15	ACCCTTGGAGACCCCACGT	GGTGAATCGGTGAGGACGG
CTCF_Pr16	GATTTCCCGTCCATATCCGC	GACTGAGTCCCCGATCTCCTG

CTCF_Pr21	ATCGGCAAGAACACAAGACTGG	CCTGGCCTTGTCGTTCGTC
CTCF_Pr22	GACGAACGACAAGGCCAGG	CCTCTACTGGCCAAGTCGGAG
CTCF_Pr23	GGCCAGTAGAGGCAGTGAGGA	TGAGCCCTATGTCTGGCCAG
CTCF_Pr24	ATCAGAGTCATGGCACCCCA	CAGCCCATCACCTACTCTGCA
ChIP_nctrl01	TCAGGCTTTTTCACTGCCTT	GTGTATGGGGAATGGAGGTG
ChIP_nctrl02	TTGAGCATCTGCTATGTGCC	GCTGTACTTCAGCCTGGGAG
ChIP_nctrl03	TCACTTTGATGCTTGGCTTG	ATCCCCAGAACCTGTGAGTG
ChIP_nctrl04	GCGGTCTTTGTAAAACCCAA	AGTTCTGCCTGGACTCCTGA
ChIP_nctrl05	TAACGTGCCACCTACTTCCC	GTGCACCAGGCCTTTAACAT
ChIP_nctrl06	TCACCTGAGGTCAGGAGCTT	CTTGCTCAAGGTTTCAAGGC
ChIP_nctrl07	ATAGTCGGTGGCCAACAAAG	TGATCTCTCCAGCCTCAGGT
ChIP_nctrl08	CTTGAGCCAGTCCCTTTCTG	TCTAGTGCAAGGCCACCTCT
ChIP_nctrl09	GACTGTGGTGATGGTTGCAC	CAGGTGGTTTCAGGTGGTCT
ChIP_nctrl10	GTGATTCTTCTGCCTCAGCC	ACTGGAAACACCATTGGCTC
ChIP_nctrl11	AGAGGGGCATTGTGGTGTAG	TGAGGCGATTCAACTCTGTG
ChIP_nctrl12	CTACCGCTTGATGGCTTCTC	TTCCCTCTGTGTGCTGAGTG
ChIP_nctrl13	ACCCTGAACCTGCCATACTG	CATCACAGCAAGCCTTTGAA
ChIP_nctrl14	AAATCCAGACACCTGGCAAC	GTCAGGAGTTCGAGACCAGC
ChIP_nctrl15	TGTGGTAGTGTGCCCCTGTA	TTCAACTCTTTGCCAGCCTT
ChIP_nctrl16	ACCACTTGAGTCTTGTGGGG	TCCTGTGATTGTTCAGCAGC
ChIP_nctrl17	CTGAAACCAAGCAGAAAGCC	GTCAGGCTGGTCTCGAACTC
ChIP_nctrl18	GAGCTGGGATTCGTGGATAA	AAGTGCCCTGGTGACATAGG
ChIP_nctrl19	CTCGTAGCCTCAAGCAATCC	CACCATGCCTGGCTAATTTT
ChIP_nctrl20	ATTACAGGCATGCAACACCA	CTTGCCTCCAAGGAACTCAG
chiparray_nctrl21	CATCCATTCAGTCATGTCGC	CACACTGTTTCCCCTCCACT
chiparray_nctrl22	TGCAGTAGGAACCCAGCTCT	ATGAGCGTGCCCATAAAAAC
chiparray_nctrl23	GCTGCATTGTAAACCACCCT	GGTAGACTGCTTGAGCCCAG
chiparray_nctrl24	CTTCCAGGGAGAAAGCACAG	ATTGTTCTGGCCAAAATTGC
chiparray_nctrl25	AACAAAATGTACCTTGCGGC	CTGACTCTGCACCAAGTGGA

B5. Primers to amplify fragments printed onto the 3.5 Mb custom-made array (see section 5.2)

	aminolinked sense	antisense
pUC18	CCAGTCACGACGTTGTAA	CGGATAACAATTTCACACA

B6. Primers Designed to fill gaps containing TSSs in 3.5 Mb custom made array (see section 5.3)

Primer Name (20ace database)	Primer1	Primer2
C20ORF146	CCAGTCACGACGTTGTAACATAAAACATTTTGTAATTAG	CCTTTCACTATGGGATGTGCCATTTC
NCOA5	CCAGTCACGACGTTGTAACATTTGGTAAATAGCAGTAC	TTAAATGAGATGATATGTGTAAAGGATTC
UBE2C	CCAGTCACGACGTTGTAACCTGAACCAATGTTGCCTTT	CAGTCACAGGAGGAGTGTCTTGCTCC
C20ORF170	CCAGTCACGACGTTGTAACTTTACCTTGCTCAGGCTCT	GACTCCATTGATATAGAACAGAAGC
PPGB	CCAGTCACGACGTTGTAAGAAGCAAACACGTCCACCAC	GAAATGCCTTACAAGGTCGCC
NCOA3	CCAGTCACGACGTTGTAAGTCATAAATACGCTGTGGGA	GATCTGAAGCCGCTGGCTCTCGCAG
OVCOV1	CCAGTCACGACGTTGTAATCGAAGGCTTAGTCTCTATT	CTTCTGGATCGCCCAAGTTTTAAG
TOMM34	CCAGTCACGACGTTGTAATGCCGGAGTCGGAAGGGGCT	CTCAGCATGTTGGCTTTTCATACTTG