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.

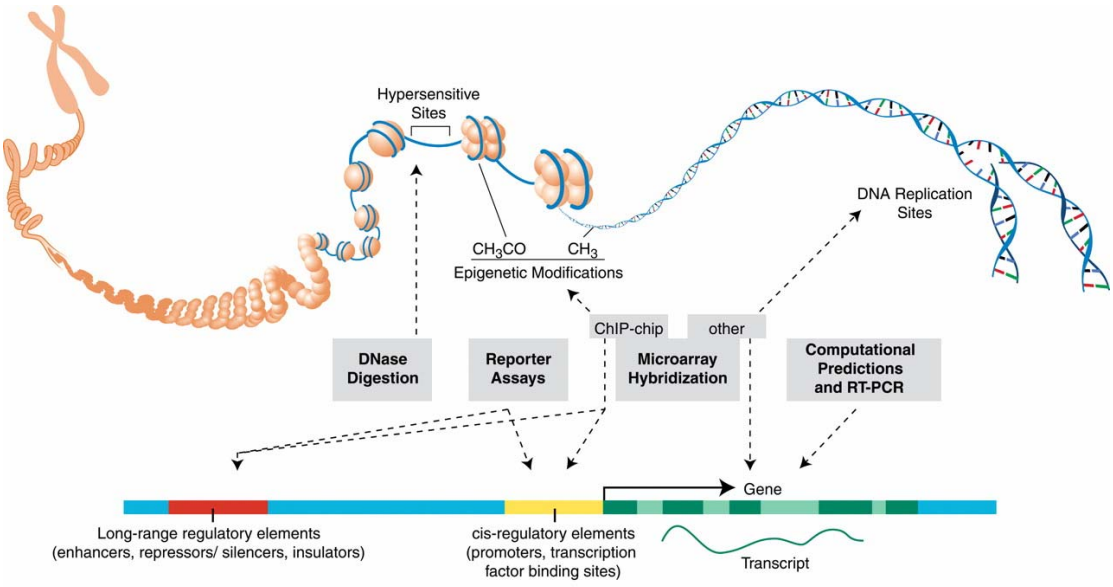

Figure 1.1 Functional genomic elements aimed to be identified by the ENCODE pilot phase. The indicated methods are being employed to this end. This figure is adapted from reference ENCODE, 2004.

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.

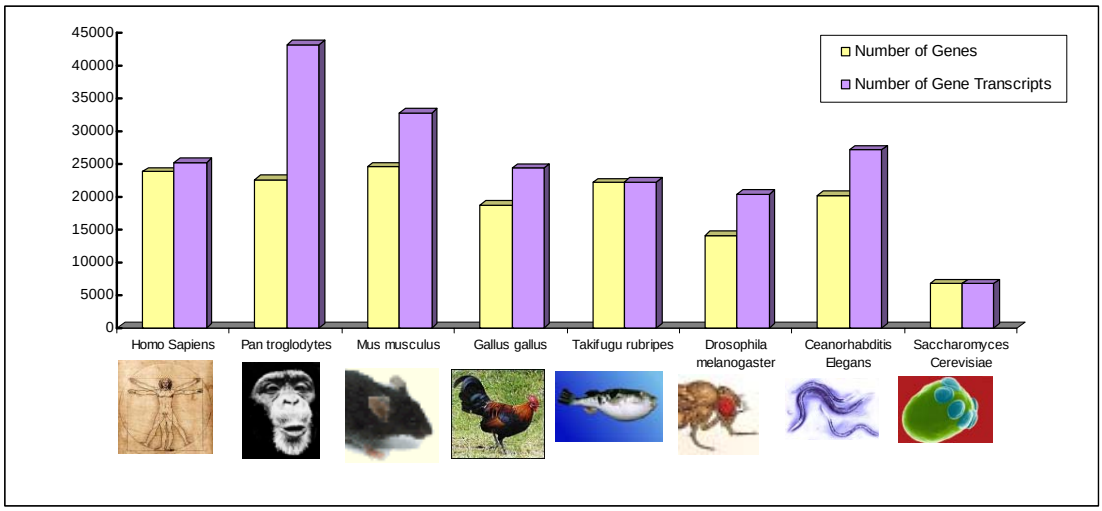

Figure 1.2 The number of genes and transcripts are taken from Ensembl version 39, NCBI build 36.

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.

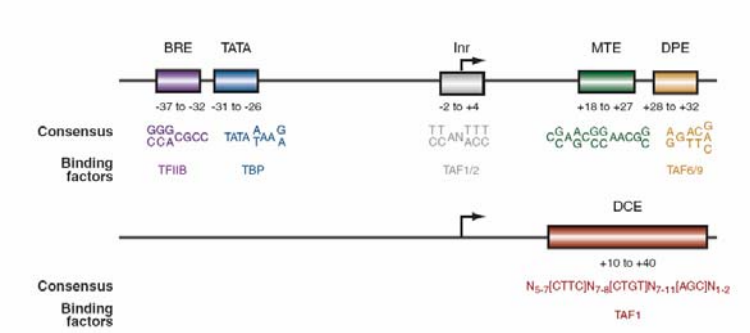

Figure 1.3 Sequence elements that are found on metazoan core promoters; BRE (TFIIB-recognition element), TATA (TATA-box binding protein binding motif), Inr (Initiation element), MTE (motif ten element), DPE (downstream promoter element) and DCE (downstream core element). Transcription initiation site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

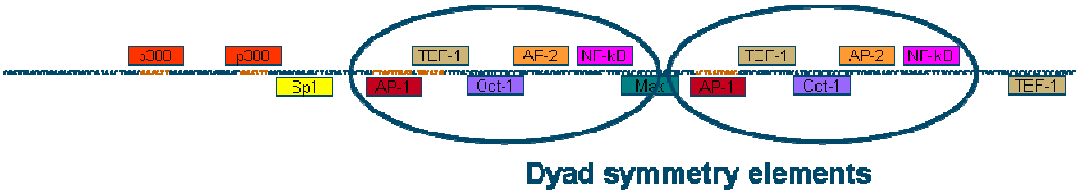

Figure 1.4 Positions of transcription factor binding sites on SV40 enhancer. This enhancer can drive the expression of promoters in a orientation and position independent manner.

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).

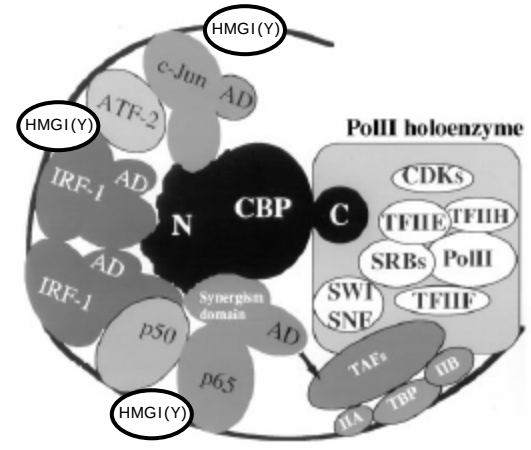

Figure 1.5 Structure of an enhanceosome where an architectural protein such as HMGI(Y), bends DNA and allow cooperative binding of further transcription factors to the enhancer and enables their contacts with the promoter bound complexes. This figure is adapted from reference (Merika et al., 1998).

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).

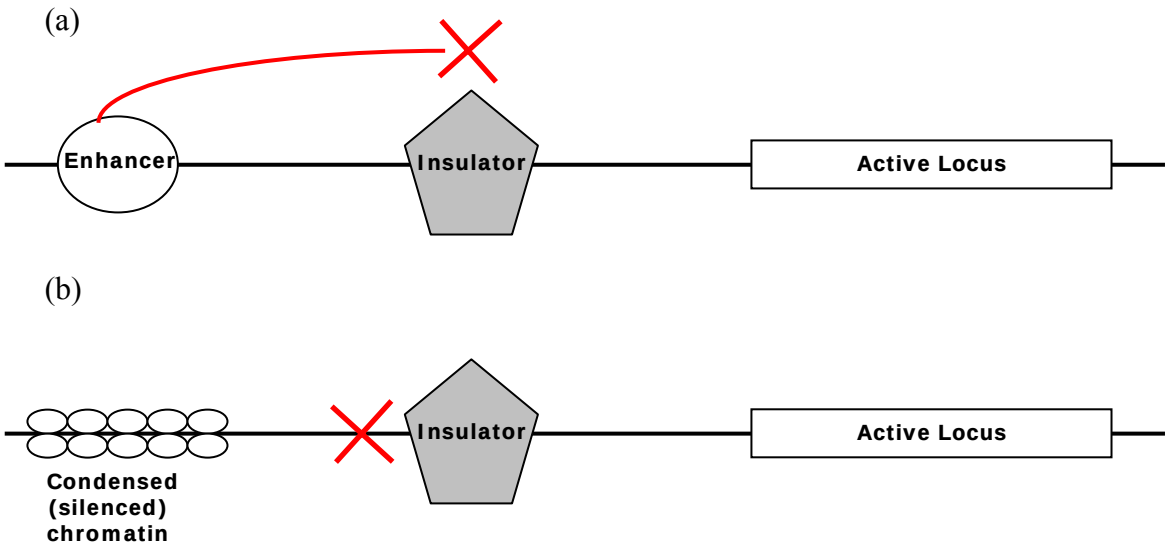

Figure 1.6 Two possible mode of action of an insulator (a) it can block the communication between an enhancer and an active locus or (b) it prevents the spread of condensed chromatin structure onto actively transcribed regions.

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.

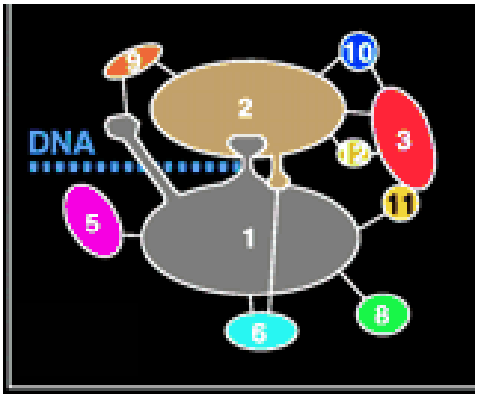

Figure 1.7 Relative positions of the 12 subunits of the RNA polymerase II transcriptional machinery and DNA. The straight lines map interactions between corresponding subunits. Not all subunits can be visualised on this view. This display is reproduced from reference (Cramer et al., 2000).

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 Å resolution (Cramer et al., 2001). The dynamic of the transcription initiation is shown in Figure 1.8.

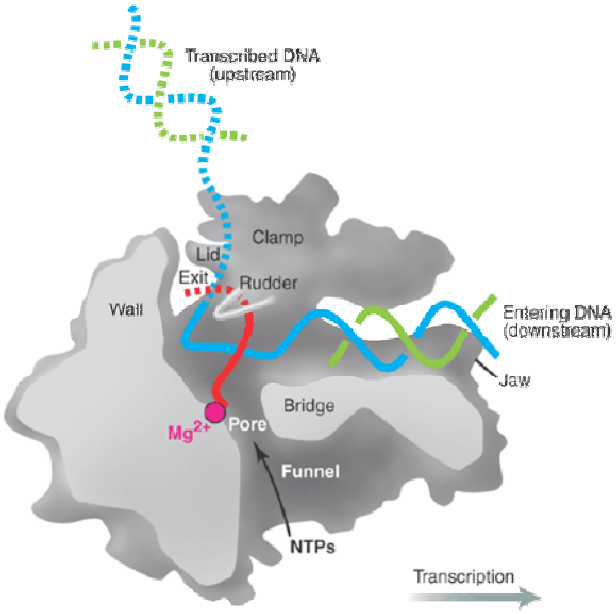

Figure 1.8 Side view of RNA (red) synthesis by RNA polymerase II machinery from a DNA template (template strand blue and coding strand is green). Cut surfaces of the protein, in the front, are lightly shaded and the remainder, at the back, are darkly shaded. By convention, the polymerase is moving on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

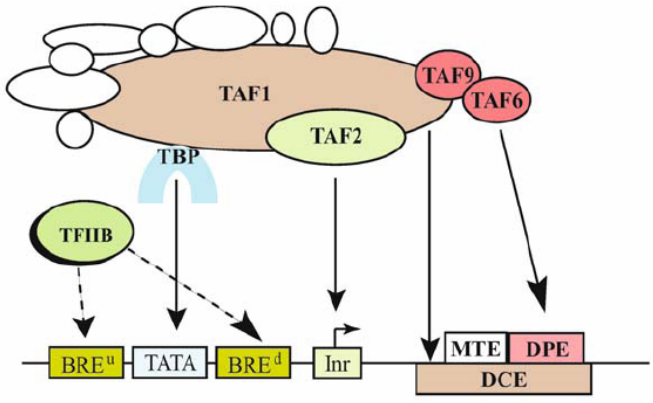

Figure 1.9 Known contacts between TATA box binding protein associated factors (TAFs) and core promoter elements as explained in Figure 1.3. This figure is adapted from reference (Thomas and Chiang, 2006).

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.

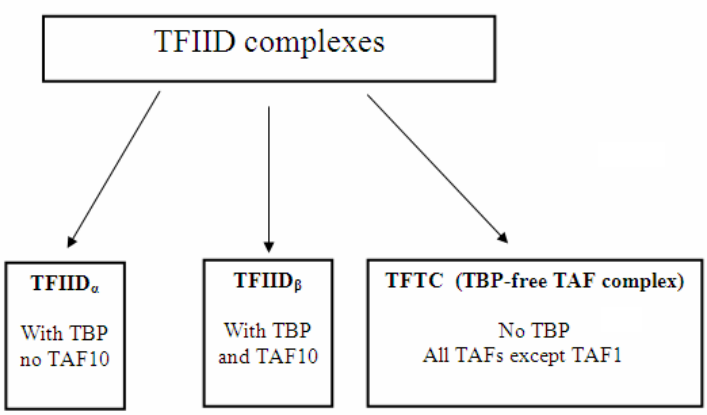

Figure 1.10 Three different TFIID complexes depending on the inclusion of TBP and TAF10.

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).

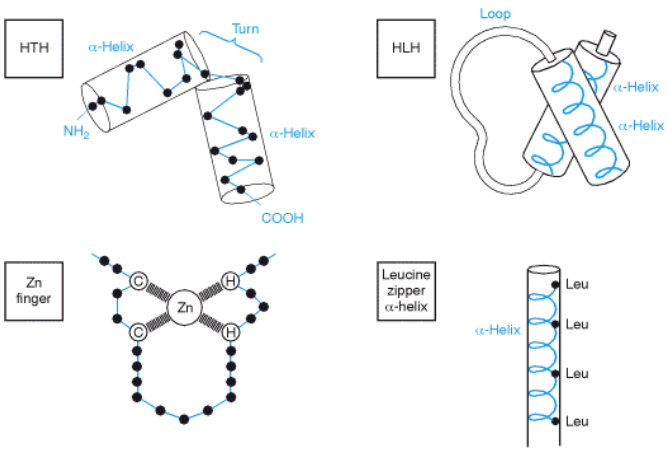

Figure 1.11 Schematic representation of four common DNA binding domains in transcription factors. Abbreviations: HTH, helix-turn-helix; HLH, helix-loop-helix, Zn, zinc; Leu, leucine. This figure is reproduced from (Strachan and Read, 2003).

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).

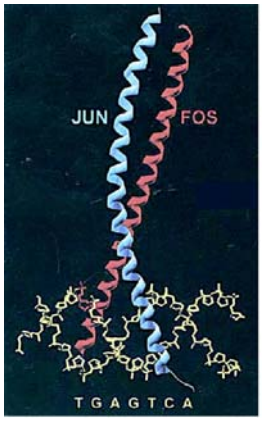

Figure 1.12. Jun and fos proteins both have leucine zipper motif to form a heterodimer (AP-1) to bind to their palindromic recognition sequences on DNA. This figure is reproduced from reference (Hess et al., 2004).

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).

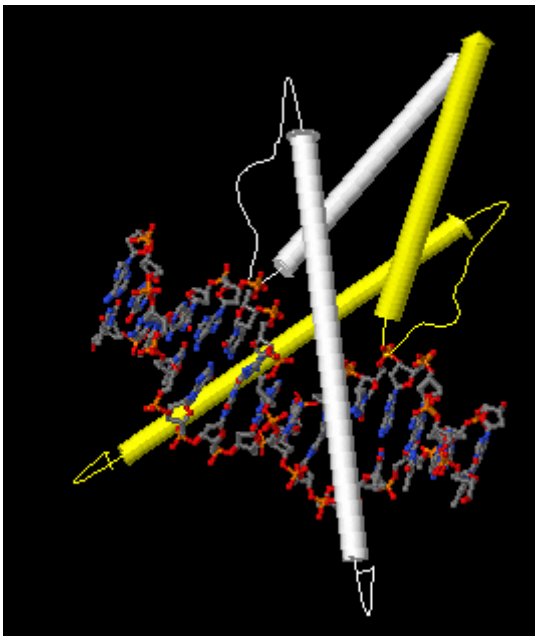

Figure 1.13 The helix-loop-helix motif carrying dimer protein bound to DNA. Different subunits are shown in white and yellow. This illustration is adapted from URL site reference URL1, 2004.

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.

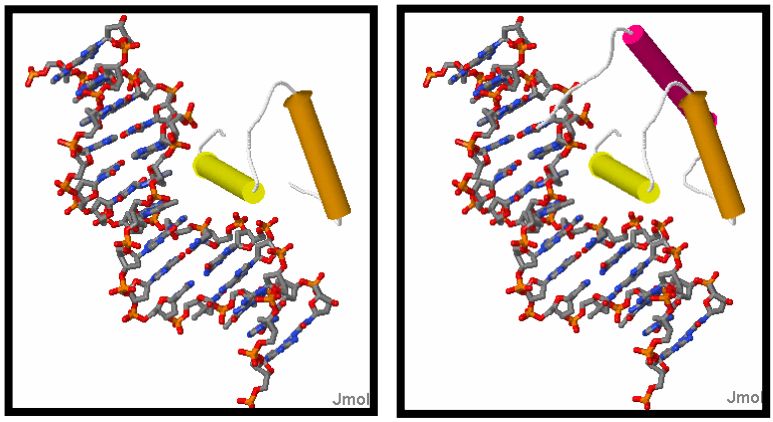

Figure 1.14 Helix-turn-helix (HTH) motif bound to its DNA on the left and a specialized form of HTH motif called homeodomain bound to its DNA. Homeodomain contains an extra α-helix presumably for further stabilization of DNA binding. These figures are adapted from URL site reference URL2, 2004.

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).

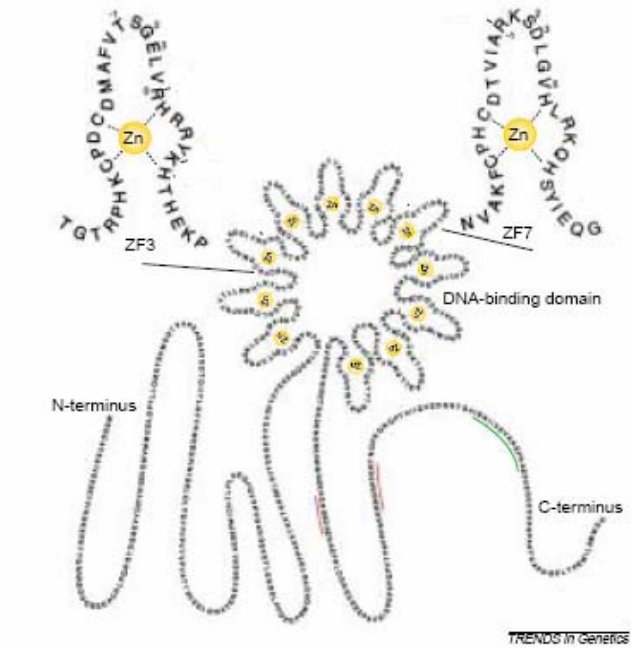

Figure 1.15. Structure of DNA binding domain of CTCF composed of 11 adjacent zinc finger domain. This figure is adapted from (Ohlsson et al., 2001).

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	U5 small 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).

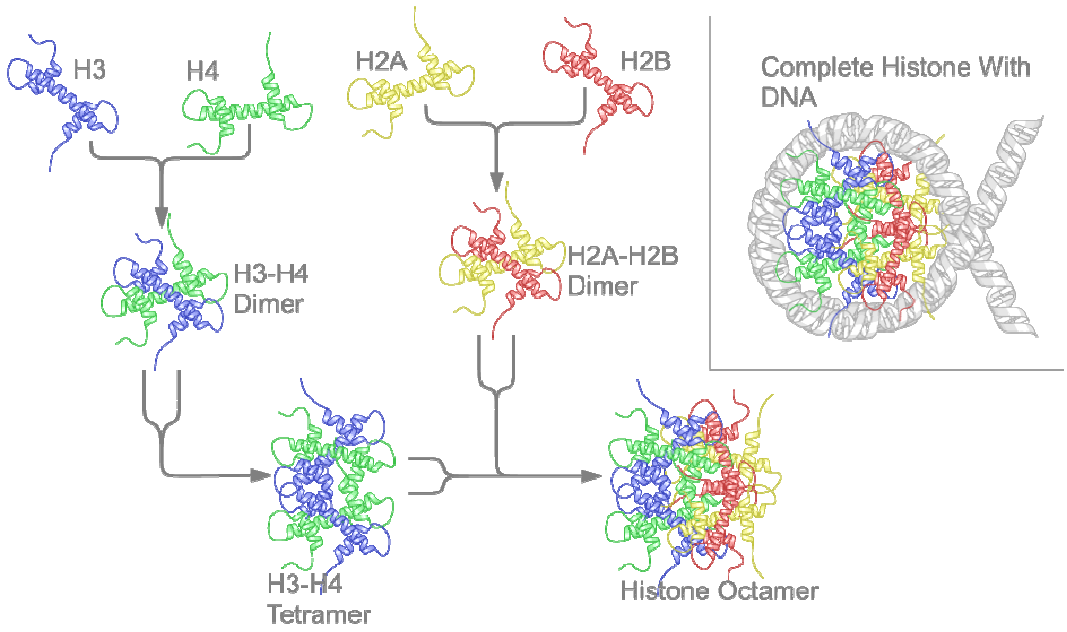

Figure 1.16. Structural organisation of core histones in the nucleosome core particle. This figure is adapted from reference (Alberts et al., 2001).

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).

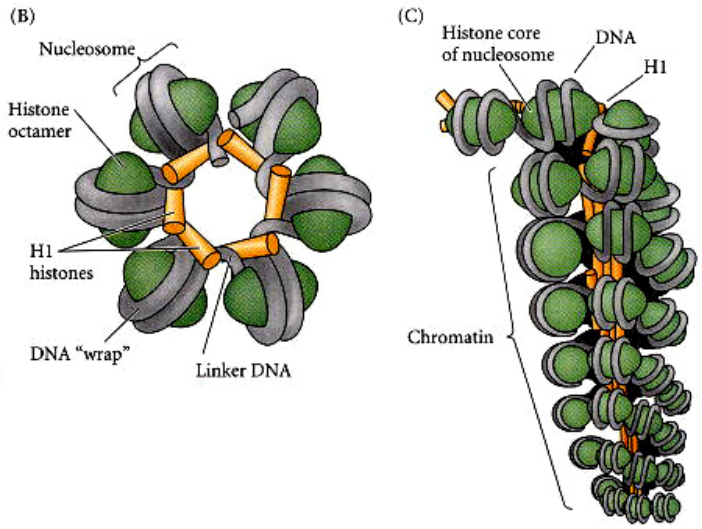

Figure 1.17. (B) About 160 base pairs of DNA encircle each histone core particle, nucleosome, and about 40 base pairs of DNA link the nucleosomes together. (C) Model for the arrangement of nucleosomes in the highly compacted solenoidal chromatin structure. This figure is adapted from reference (Turner,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

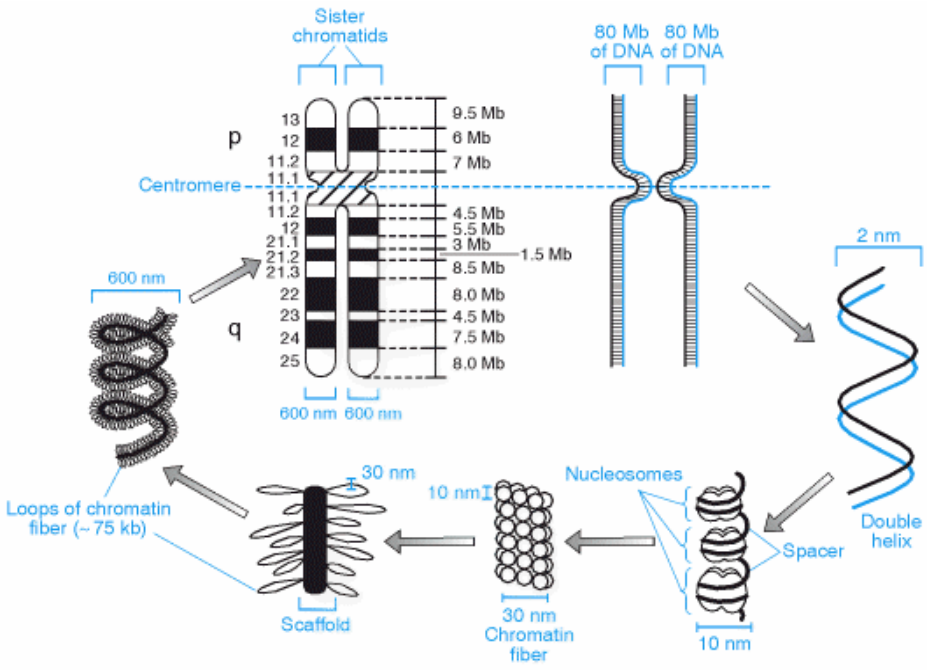

Figure 1.18. Higher order of DNA packaging in nucleus. This figure is reproduced from reference (Strachan and Read, 2003).

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.

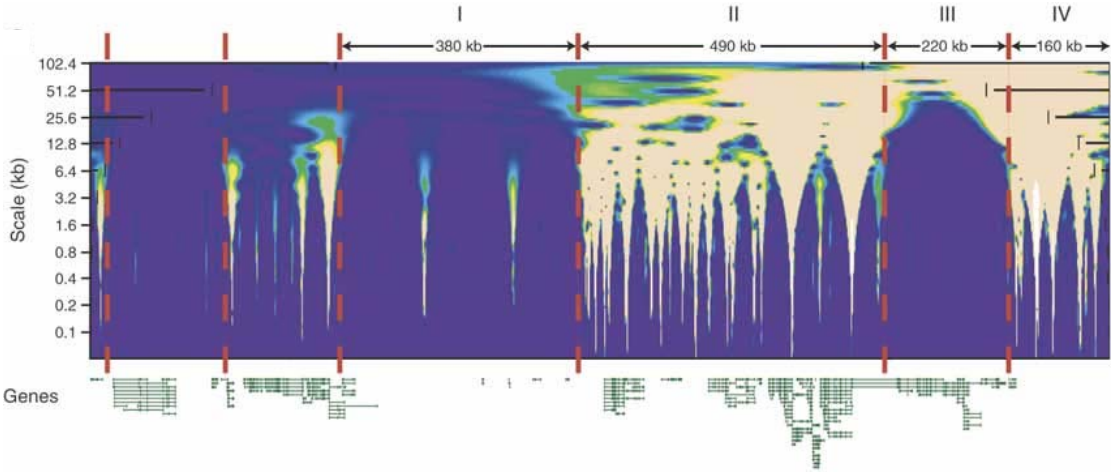

Figure 1.19. Continuous wavelet transform heat map of chromatin accessibility across a 1.7-Mb segment of chromosome 21 containing the Down Syndrome critical region29 (x axis, genomic position; y axis, wavelet scale), genes are shown below heat map. Four broad classes of chromatin domains are thus distinguished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

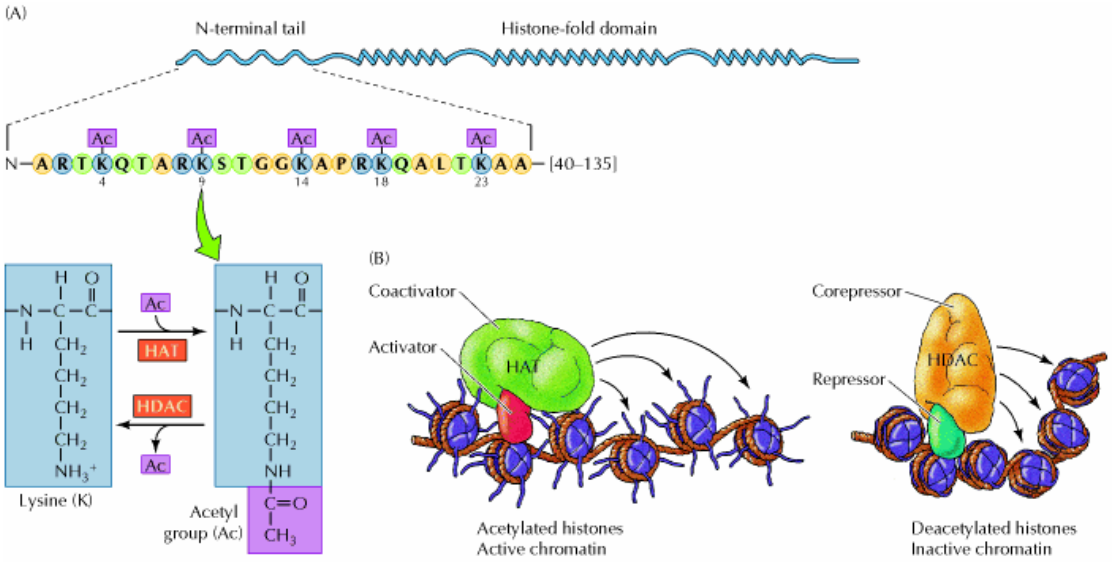

Figure 1.20. (A) The N-terminal tails of the core histones (e.g., H3) are modified by the addition of acetyl groups (Ac) to the side chains of specific lysine residues. (B) Transcriptional activators and repressors are associated with coactivators and corepressors, which have histone acetyltransferase (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.

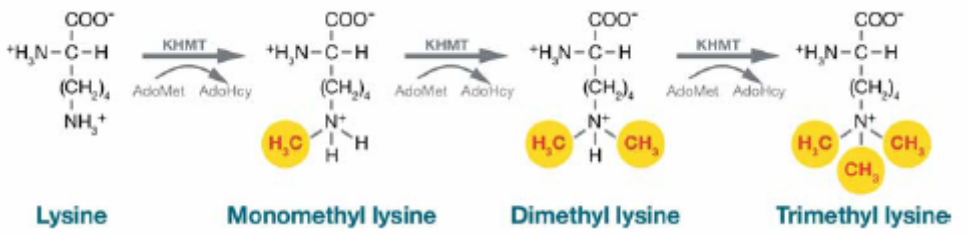

Figure 1.21 The chemistry of methylation on lysine residues of histones. Adomet (S-adenosyl-L-methionine or SAM) is a cofactor which carries the methyl group to be transferred. This figure is reproduced from reference (Shilatifard, 2006).

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.

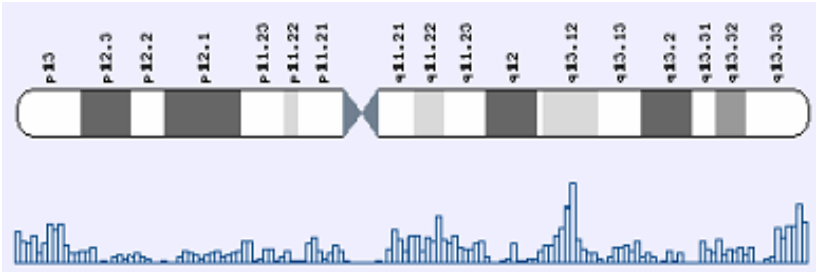

Figure 1.22. Cytoband view of chromosome 20 on the right and the gene number histogram along the chromosome on the left.

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).

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
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

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