CHAPTER 1
INTRODUCTION

Introduction

The first part of the general introduction will be addressing common features of the regulation of eukaryotic transcription, with the aim of giving a broad outline of factors involved in the transcriptional process. The second part of the introduction will be concerned with erythropoiesis, haemoglobin and the structure and regulation of the mouse and human globin genes.

Transcription

The process of transcriptional regulation is complex and involves many different players. DNA sequence itself, in the form of *cis*-regulatory elements of genes, chromatin and its higher order structures, histone modification and the diverse chromatin modification complexes, all play specific and important roles in eukaryotic gene regulation. In the following paragraphs these features will be addressed.

Chromatin

Eukaryotic chromosomes contain 2m of DNA when stretched and need to be packaged to fit into the nucleus of the cell (with a diameter of 10μm). The packaging of the DNA to form a compact structure is achieved with the aid of specific proteins that are complexed with the DNA to form chromatin (Fig.1). All DNA processes in the nucleus, e.g. replication and transcription, take place in the context of chromatin. The picture that is emerging is that chromatin does *not* present an obstacle to these processes (as was previously thought) but, instead, plays a leading role in their regulation. In this part of my thesis I will consider the role of chromatin in transcriptional regulation. The general structure of chromatin will be addressed first.

Chromatin organisation

The primary proteins involved in chromatin structure are the histones. There are five types of histones that fall into two categories. Firstly, there are the core histones H2A, H2B, H3 and H4 and, secondly, the linker histone H1. The core histones are evolutionary conserved both in size and amino acid sequence (de Lange *et al.*, 1969). They are small proteins, rich in arginine and lysine residues (van Holde, 1989). They contain a globular domain, which is important for histone-histone and histone-DNA interactions, as well as a N-terminal tail domain, with the exception of histone H2A that also has a tail at the C-terminus (Bohm and Crane-Robinson, 1984). The linker histones are less well conserved, are rich in lysine residues and contain a globular domain, which is flanked on both sides by a tail.

The core histones interact with DNA and with each other to form a histone octamer (Fig. 1D). The core histone interacts with the DNA via its globular domain. DNA is wrapped around the histone octamer. The octamer is composed of a tetramer formed by H3 and H4, which, after binding to DNA, is bound by two dimers of H2A-H2B resulting in the final histone octamer (Eickbush and Moudrianakis, 1978; Hayes *et al.*,1990, 1991 and Arents *et al.*,1991). The octamer forms the basis of the nucleosome. The nucleosome is the fundamental repeating unit of chromatin and consists of the histone octamer and approximately 146 bp of DNA wrapped around the octamer in 83 bp superhelical loops (Fig. 1D) (Kornberg, 1974). The DNA helix is thus packaged into nucleosomal cores giving a characteristic “beads on a string” appearance of the chromatin fiber when visualised using electron microscopy (Olins and Olins, 1974). The first level of organisation of DNA into nucleosomal cores forms the 10nm chromatin fiber (Fig. 1B).

The nucleosome model has been analysed in detail by solving the nucleosomal structure at a resolution of 2.8 Angstrom (Luger *et al.*, 1997) and 1.9 Angstrom (Davey *et al.*, 2002), showing a twist of the DNA molecule on the surface of the nucleosome. The “twisting” of the DNA around the nucleosomes results in the formation of minor and major grooves along the backbone of the helix. The minor grooves form the channels through which the N-terminal tails of histone H2B and H3 pass. The tails of histones H2A and H4 pass through the gyres of the DNA superhelix. In these ways the tails not only interact with the DNA but may also interact with neighbouring nucleosomes. The tails are targets for modifications such as acetylation and phosphorylation due to their localisation on the surface of the nucleosome. Such modifications can lead to changes in the binding affinity of nucleosomes for DNA, for example, as a result of the neutralisation of the positively charged tails by acetylation which leads to a decrease in affinity of the nucleosome for DNA, thus resulting in alterations in local chromatin states.

Individual nucleosomes are spaced by approximately 80 bp of linker DNA. Micrococcal nuclease digestions followed by sucrose gradient and mobility shift assays of the smallest (mononucleosomal) fractions, showed that part of the linker DNA is bound by histone H1 (Varshavsky *et al.*, 1976). This histone is required for the maintenance of the stability of the core histones in the histone octamer and the stability of the higher order chromatin structures (Dasso *et al.*, 1994; Wolffe, 1998 and Carruthers and Hansen, 2000). The tails of histone H1 bind to the DNA within the nucleosome and with the linker DNA (Fig. 1D), with the C-terminal tail folding into α-helices that associate with the major groove of the linker DNA (Clark *et al.*, 1988). The backbone of the DNA is neutralised by the histone H1 tails and the 10nm fiber can thus fold into a higher order structure, compacting the DNA into what is known as the 30nm fiber (Fig. 1C) (Clark and Kimura, 1990). Although linker histones are important for the neutralisation of the backbone of DNA, it is the tails of the core histones that are essential for the folding of chromatin into the 30nm fiber. Using defined chromatin model systems with core histones lacking their tails, it was demonstrated that the tail-less chromatin fiber could not fold into 30nm fibers, even with linker histones present (Carruthers and Hansen, 2000). Thus tails of the core histones and the nucleosomes play an important role in internucleosomal contacts within the 30nm fiber and are critical for the self-assembly of the condensed fibers in higher order structures (Fletcher and Hansen, 1996 and Anderson and Widom, 2000).

The 30nm fiber itself is organised into loops estimated to be between 30-100 kb in size. These sizes were estimated using electron microscopy and by distance measurements using fluorescent *in situ* hybridisation (Yokata *et al.*, 1995). These loops have been proposed to play an important role in the regulation of gene expression, e.g. in the interaction between enhancers and promoters over long distances. Two nucleoprotein structures, which have been identified using different nuclear extraction methods have been proposed to play a role in this loop formation: the matrix associated regions (MARs) and the scaffold attachment regions (SARs) (Mirkovitch *et al.*, 1984 and Cockerill and Garrard, 1986). MARs and SARs do not have consensus sequences but contain AT-rich stretches recognised by linker histone H1 and topoisomerase II. This recognition is thought to lead to the anchoring of the 30nm chromatin fiber to the chromosome scaffold or nuclear matrix. Proteins already bound to the MARs and SARs then secure the chromatin fiber into structural loops (Laemmli *et al.*, 1992), thus achieving an even further packaging of the DNA. Other proteins binding to MAR and SAR elements, like SATB1, GATA3 and p300, have been implicated to play a role in gene regulation by introducing chromatin changes through binding to MARs and SARs (Kieffer *et al.*, 2002 and Martens *et al.*, 2002). SATB1 and GATA3 have both been shown to bind MARs in the CD8 gene, and via this binding have been suggested to play a role as epigenetic regulators of CD8 expression (Kieffer *et al.*, 2002).

Euchromatin and heterochromatin

Chromatin has been cytologicaly divided into heterochromatin and euchromatin. (Heitz, 1928). Heterochromatin is densely stained (by carmin acetic acid) throughout the cell cycle indicative of a constitutively condensed chromatin structure. Euchromatin, on the other hand, is more lightly stained and decondenses as the cell progresses from the metaphase after mitosis to the interphase of the cell cycle.

Heterochromatic genomic domains consist predominantly of repetitive DNA, including satellite sequences, and is mostly found in pericentromeric and telomeric regions. Heterochromatin contains generally few genes whereas euchromatin is gene-rich. Other characteristics that distinguish heterochromatin from euchromatin include: (i) higher order chromatin structure, which shows a much more regular nucleosomal organisation for heterochromatin (Wallrath and Elgin, 1995 and Sun *et al.*, 2001); (ii) differences in histone modifications, with heterochromatin being rich in methylated/hypoacytelated histones whereas euchromatin is enriched in acetylated/non-methylated histones (Strahl and Allis, 2000 and Rea *et al.*, 2000); (iii) differences in replication timing, with heterochromatin replicating late in S-phase versus early for euchromatin. Euchromatin is thus generally viewed as the chromatin compartment that supports transcription whereas heterochromatin represses transcription.

Two key observations have coupled formation of heterochromatic structures to silencing of genes that are normally found in euchromatin. The first is X-inactivation in dosage compensation, in which the inactive X-chromosome showed the same cytological staining and molecular characteristics, such as hypoacetylated histones, as heterochromatin, indicative of heterochromatinisation as cause of silencing (Schoenherr and Tilghman, 2000).

The second observation was first described in *Drosophila* in which a chromosomal translocation of the *white* gene, normally residing in euchromatin, placed it close to pericentromeric heterochromatin. This gave rise to a variegated phenotype of white and red patches in the *Drosophila* eye (Muller, 1930). Variegated *white* expression appeared to be due to the variable silencing of the gene in different cells of the eye tissue. Thus in some cells the *white* gene expresses giving rise to red colour, whereas in other cells the gene is inactivated giving rise to *white* colour. The variability in expression of the *white* gene in the eye is thought to reflect differences in the extent of heterochromatinisation at the site of the chromosomal translocation between different eye cells. It is thought that heterochromatin is laid down in a window of opportunity early in development and spreads from so-called nucleation centres along the chromatin fiber in the centromeres, with the extent of spreading varying from one cell to another. However, once heterochromatin has been laid down it becomes fixed thus "freezing" the differences in heterochromatin spreading between cells, which are then clonally inherited in subsequent cell generations. Therefore, in a cell where the *white* gene has been heterochromatinised it will be inactive, whereas in another cell in the same tissue where the gene has not been embedded in heterochromatin, it will be active. Variegated expression patterns also appear to be stable through subsequent cell divisions and are thus clonally inherited. This effect on gene expression is referred to as position effect variegation (PEV) and is defined as the heritable silencing through multiple cell divisions resulting from translocation or integration of a gene in a position close to heterochromatin.

Mutations affecting PEV have been extensively studied and have resulted in the identification of more than 30 genetic modifiers of PEV (Jenuwein and Allis, 2001). These can be divided into two groups, with antagonizing effects. The first group consists of the Su(Var) group of proteins, which when mutated suppress variegation. The second group is the E(Var) group, which when mutated enhance the variegation of expression of the integrated gene. It can therefore be inferred from these phenotypes that the Su(Var) proteins are normally involved in promoting heterochromatinisation, and hence silencing, whereas the E(Var) proteins normally counteract heterochromatinisation and hence promote expression (Wallrath, 1998 and Eissenberg *et al.*, 1990).

One of the best studied proteins of the first group of modifiers is heterochromatin protein 1 (HP1), which is encoded by Su(var) 2-5 (Eissenberg *et al.*, 1990 and Eissenberg *et al.*, 1992). Mutations in HP1 increase expression of the *white* gene, whereas an additional copy of HP1 reduces *white* expression, thus enhancing variegation (Cryderman *et al.*, 1998). Recent studies aimed at understanding the basis for HP1-mediated repression have shown that histone modifications play a role in placing heterochromatic marks. First it was shown that histone deacetylase inhibitors result in disruption of heterochromatin and HP1 binding (Taddei *et al.*, 2001). Secondly, a link between methylation of H3 and heterochromatin formation was identified. The Su(var)39h genes encode histone methyltransferases and have been shown to have high specificity for lys-9 of histone H3 (Rea *et al.*, 2000). A connection between a histone deacetylase and Su(var)39h has recently been shown in *Drosophila* thus providing a mechanism to convert an acetylated histone into a methylated one (Czermin *et al.*, 2001). The link between Su(var)39h and HP1 was found in primary mouse fibroblasts from double null Su(var)39h mice. These studies, using immunostaining to detect HP1 distribution, indicated that the methylation of H3 lys-9 by Su(var)39h is important for HP1 localisation and heterochromatinisation (Bannister *et al.*, 2001 and Lachner *et al.*, 2001). E(Var) proteins on the other hand interact with proteins that remodel chromatin to allow activation of transcription, like the SWI/SNF and Brahma complexes (Tsukinama and Wu, 1997 and Kal *et al.*, 2000).

Besides the Su(var) and E(var) proteins other factors influence the variegation of gene expression. The integration of β-globin and CD2 transgenes containing a full locus control region in pericentromeric regions resulted in normal levels of expression of the transgenes (Festenstein *et al.*, 1996 and Milot *et al.*, 1996). Deletion of part of the LCR, however, led to PEV, which suggest that the loss of hypersensitive sites results in a decrease in accessibility of the regulatory elements of the transgene to the transcriptional machinery. These studies led to the formulation of a probability model for LCR function in which sufficient binding sites for positively acting factors would completely overcome the packaging of DNA into heterochromatin (Festenstein and Kioussis, 2000).

Also transcription factors have an influence on PEV. Studies in which the doses of transcription factors like EKLF and Sp1 in variegating mouse lines was either overexpressed or reduced, showed a higher or lower expression of the variegating transgene, respectively, compared to endogenous genes, which do not show a PEV (Lundgren *et al.*, 2000 and McMorrow *et al.*, 2000). All the results presented above indicate that the decision to silence or activate a gene in a heterochromatic environment is the result of a balance between positive and negative factors.

Nuclear organisation and gene expression

The organisation of euchromatin and heterochromatin in discrete nuclear domains also plays an important role in gene expression. Within the nucleus euchromatin is mainly found in the interior of the nucleus, whereas heterochromatin is found near the nuclear periphery and around the nucleolus. The different chromatin domains delineate different subnuclear compartments, which exert different effects on the transcriptional regulation of genes. For example, a group of proteins, shown to accumulate in heterochromatic domains in the nucleus, are the Sir-proteins, which play a role in gene silencing in yeast (Gasser, 2001).

An example of this is demonstrated when genes become integrated at telomeres and, as a result, become repressed through association with silencer proteins (Sir 2p, 3p and 4p), which bind to histones H3 and H4 and deacetylate the amino-terminal tails of these histones (Grunstein, 1998). The sites of Sir-mediated binding have been localised by double DNA *in situ* hybridisations and immunofluorescence and shown to be clustered in specific foci within the nucleus around the nuclear periphery. Subsequently it has been shown that the nuclear organisation of the telomeres is of critical importance for the repression via Sir proteins. Disruption of the telomeric organisation proved to result in delocalisation of the Sir-proteins and a loss of telomeric silencing (Gotta *et al.*, 1996 and Galy *et al.*, 2000). The HM-locus of *Saccharomyces cerevisiae* has also been shown to be silenced by the interaction of Sir-proteins with silencers present in the HM loci. Studies on HM silencing in *Saccharomyces cerevisiae* show that the localisation of a locus to the nuclear periphery helps to establish transcriptionally silent heterochromatin domains. It was also shown that the HM-locus could be silenced, even with a defective silencer, by artificially anchoring the locus to the nuclear periphery (Andrulis *et al.*, 1998).

Studies in yeast have shown that repressor proteins are present in the nucleus in limiting amounts and are recruited to specific heterochromatic compartments (Gasser, 2001). In addition, the spatial distribution of repetitive DNA forms subcompartments in the nucleus that favour the packaging of chromatin in a repressive state (Gasser, 2001). In higher eukaryotes a nuclear compartmentalisation similar to that observed in yeast has also been shown. Genes that are not transcribed are recruited to heterochromatic compartments and when activated "move" to the periphery of those compartments to be transcribed (Brown *et al.*, 1997 and 1999 and Francastel *et al.*, 1999). Using immunofluorescence *in situ* hybridisations, the Ikaros transcription factor was shown to be localised in constitutively heterochromatic centromeric foci in interphase nuclei. Several genes which are regulated by Ikaros were tested for their nuclear localisation in expressing and non-expressing cells. It was shown that inactive genes associate with Ikaros-heterochromatin foci, whereas the active genes are not associated with the heterochromatin foci (Brown *et al.*, 1997).

Furthermore, not the transcriptional activation of a gene but the hyperacetylation of the histones in the promoter region of the gene that "signals" the re-localisation of the gene within the nucleus. This has been shown using a β-globin gene without a LCR. When integrated in a heterochromatic domain after transfection in MEL cells, the hyperacetylation of the β-promoter resulted in the relocalisation of the gene to the periphery of the heterochromatic compartment that it was previously located in (Schubeler *et al.*, 2000). Also studies with a λ5 transgene inserted in the mouse γ satellite repeat showed using *in situ* hybridisation techniques that while a repressed gene was localised in a heterochromatic compartment, the binding of specific transcription factors relocated the transgene to the periphery of the compartment without activating it. Transcriptional activation of the gene required a strong transactivator, which could presumably overcome the repressing activity of the repressor proteins (Lundgren *et al.*, 2000). These observations indicate a specific sequence of events for gene activation which is related to the gene's subnuclear localisation, histone modification and the action of specific transcription factors.

Chromatin and transcription

Transcription takes place in a chromatin context. A wealth of recent research evidence has revealed the many ways in which the cell not only deals with chromatin in transcription utilises it as a significant regulatory factor (Narlikar, 2002).

Nucleosome positioning

Nucleosomes assembled on DNA generally render the DNA inaccessible to the transcriptional machinery *in vitro* (Svaren and Horz, 1996). However, DNA that is being transcribed or replicated remains nucleosomal, thus indicating that nucleosomes do not form an impossible obstacle for these processes (Studitsky *et al.*, 1995 and Felsenfeld 1996).

The orientation of a nucleosome on DNA can be defined by its translational and rotational positioning. The rotational position defines which of the DNA sequences wrapped around the nucleosome are facing outwards, towards the surrounding solution and may thus be accessible to transcription factors. The translational position of the nucleosome determines the precise site where the DNA and histone contacts begin and end.

The positioning of nucleosomal arrays on promoters and its function in repression or activation of genes has been studied most extensively in yeast, for example, using the PHO5 gene. Promoters of genes of mouse and human origin have also been studied to gain better insight into the role of positioned nucleosomes in the accessibility of transcription factors to their recognition sites regulating the induction of transcription.

The yeast PHO5 gene encodes a secreted acid phosphatase and is activated when phosphate is limiting in the cell. The PHO5 gene otherwise remains silent by the positioning of four nucleosomes over its promoter (Svaren and Horz, 1997). The activation of PHO5 requires two factors, PHO4 and PHO2, and the repositioning of the nucleosomes since at an active PHO5-promoter no nucleosomes are detected (as measured by an increase in deoxyribnuclease I (DNase I) accessibility). PHO4 binds to the sequences UASp1 and UASp2 present in the linker DNA between two precisely positioned nucleosomes. PHO4 binding leads to local chromatin remodelling allowing additional transcription factor binding and PHO5 activation (Haswell and O'Shea, 1999). The precise mechanism of the nucleosome remodelling by PHO2 and PHO4 remains unclear, however, recent evidence using yeast nuclear extracts has shown that PHO4 and PHO2 do not need nucleosome remodelling factors, but do need an ATP-dependent activity and core histone acetylation using acetyl CoA (Terrell *et al.*, 2002 and Haswell and O'Shea 1999).

The mammary tumour virus (MMTV) -promoter has also been studied in understanding how regulatory factors can recognise their cognate sequences when these are embedded in nucleosomal arrays. The MMTV-promoter is induced by glucocorticoids via the glucocorticoid receptor (GR) and NF1. Two GR-bindingsites sites are exposed on the surface of a nucleosome by the rotational positioning of the nucleosomes in the MMTV-promoter. This positioning makes it possible for a GR-dimer to bind to its sites on the nucleosome. Binding of the hormone to a GR-bindingsites site induces a change in the translational positioning of the nucleosomes at the MMTV-promoter, such that the binding sites for NF1 become accessible and, together with the glucocorticoid receptor, NF1 can activate the MMTV-promoter (Eisenfeld *et al.*, 1997 and Belikov *et al.*, 2001).

Furthermore, nucleosome rearrangements can be induced by transcription factors. This was originally observed for HNF3 and the foetal serum albumin enhancer. It was found that the nucleosomal organisation of the liver-specific enhancer was random in all tissues except liver where it is active. In liver, the chromatin structure of the enhancer consisted of an ordered array of three precisely positioned nucleosomes. Using *in vitro* chromatin assembly assays it was shown that for this nucleosomal organisation the binding of proteins related to HNF3 to the enhancer was required. This led to the suggestion that certain transcription factors can induce nucleosomal rearrangements (McPherson *et al.*, 1993). Other transcription factors like Gal4, TFIIIA, Sp1, TBP, GATA4, Fos and Jun have also been shown to play a role in nucleosomal rearrangements. They bind to their recognition sites on the nucleosomal DNA and do so in competition with histones (Cirillo and Zaret, 1999, Blomquist *et al.*, 1999 and Ng *et al.*, 1997). Their binding changes the nucleosomal structure and subsequent binding of other transcription factors can take place. Studies with the HNF3 and GATA4 transcription factors, that bind to the albumin gene enhancer, provide good examples. To test how these two can bind to their recognition sites in "repressed" chromatin, *in vitro* experiments were done using nucleosomal arrays containing albumin enhancer sequences which were compacted with linker histones. HNF3 and GATA4 were able to bind to their recognition sites and open the nucleosomal arrays without the presence of ATP-dependent remodelling factors, whereas other transcription factors like NF1 and C/EBP could not. The opening of chromatin by HNF3 is mediated by a high affinity DNA binding site and the C-terminal domain of the protein which binds histone H3 and histone H4 and thus facilitates nucleosome rearrangements (Cirillo *et al.*, 2002).

Not all transcriptional activators are able to compete with the histones in the nucleosomal array. This could be because of inaccessibility of the DNA targets site for the transcription factor through nucleosomal positioning or because of very low affinity of the transcriptional activator for nucleosomal DNA. To facilitate the binding of these transcription factors two distinct types of enzymatic activities have been described: enzymatic histone tail modifications and ATP-dependent chromatin remodelling.

Histone tail modifications

The tails of the core histones in nucleosomes are rich in lysine and arginine residues, which can be targets for posttranslational modifications. Unmodified histone tails possess a positive charge, which results in a closer interaction with the negatively charged nucleosomal DNA. Thus, histone tail modifications that change the positive charge are likely to influence interactions with the DNA.

To date, several histone tail modifications have been described (Strahl and Allis, 2000 and Turner, 2000). Regulated acetylation and deacetylation of specific lysine residues of histone H3 and histone H4 are correlated with gene activation and silencing, respectively. Phosphorylation of histone H3 ser-10 has been suggested to be important for transcription activation and chromosome condensation during mitosis (Cheung *et al.*, 2000). The methylation of arginine residues of arg-3, -17, -26 of histone H3 and arg-3 of histone H4 have been shown to play a role in gene activation (Chen *et al.*, 1999 and Wang *et al.*, 2001), whereas methylation of lys-9 of histone 3 plays a role in gene silencing through heterochromatinisation (Rea *et al.*, 2000). Another histone tail modification is the ubiquitination of lys-123 at the C-terminal tail of histone H2B. Mutation of the ubiquitination site of histone H2B causes defects in meiosis and mitosis in yeast (Robzyk *et al.*, 2000). Histone H1 has also been shown to be ubiquitinated by *Drosophila* TAFII250 (Pham and Sauer, 2000). TAFII250 is recruited to promoters, and loss of TAFII250 in the fly embryo has been shown to result in a reduced expression of the Dorsal activator (Pham and Sauer, 2000). This leads to the suggestion that ubiquitination of histone H1 can regulate chromosomal gene activity in a promoter specific manner.

The experimental evidence so far, suggests that histone tail modifications form part of an enzymatic cascade, which leads to specific changes in chromatin structure resulting in the repression or activation of transcription. The end result on transcription or other nuclear functions dictated by particular combinations of histone tail modifications has been called the histone code (review Jenuwein and Allis, 2001).

Histone tail acetylation has been a major focus of investigation. Acetylation of the four core histones, at specific lysine residues, occurs in all animals and plants studied to date (Csordas, 1990). The first evidence that acetylation of histones correlates with transcriptional activity came from studies in yeast (Grunstein *et al.*, 1992) and immunofluorescence studies using specific α-Acetyl-histone antibodies (Jeppesen and Turner, 1993). Transcriptionally active domains have thus been correlated with general histone hyperacetylation (Hebbes *et al.*, 1994 and Lee *et al.*, 1993), whereas inactive domains appear hypoacetylated (Turner *et al.*, 1992). Acetylation occurs on specific lysine residues. For heterochromatin in yeast and *Drosophila* it has been shown that only H4-lys-12 is acetylated, whereas in euchromatin of yeast, *Drosophila* and humans different combinations of acetylated lysines of histone H3 and histone H4 have been found (Turner *et al.*, 1992; Clarke *et al.*, 1992; O'Neil and Turner, 1995; Bannister *et al.*, 2000 and Rojas *et al.*, 1999). Transcriptionally active chromatin has been shown to be acetylated at lys-14 of histone H3 and at lys-8 and -12 of histone H4 (Cheung *et al.*, 2000; Lo *et al.*, 2000 and Wang *et al.*, 2001). Newly synthesised histones appear to be acetylated at different residues than the histones in active chromatin. In newly synthesised histones, histone H4 is acetylated at lys-5 and -12 and at histone H2A lys-5 (Grant and Berger, 1999).

The effect of acetylation of the N-terminal histone tails is the neutralisation of the positive charged on the lysine residues, resulting in a decrease of histone tail affinity for negatively charged nucleosomal DNA. This has an effect on nucleosome conformation and inter-nucleosomal interactions (Hamiche *et al.*, 1999; Langst *et al.*, 1999; Whitehouse *et al.*, 1999; Clapier *et al.*, 2000 and Oliva, *et al.*, 1990), resulting in a "loosening up" of the nuclesosomal structure thus making the DNA more accessible for the transcription machinery (Lee *et al.*, 1993; Vetesse-Dadey *et al.*, 1996 and Sewack *et al.*, 2001).

An important tool in determining the histone tail acetylation status of specific genes of interest and relating it to transcriptional activation, has been the use of specific antibodies against acetylated histones H3 and H4 in the immunoprecipitation of formaldehyde cross-linked chromatin (review Orlando, 2000). Due to the fact that formaldehyde cross-links are reversible, specific DNA sequences can be recovered and tested for enrichment in acetylated histones (Saitoh and Wada, 2000, review by Forsberg and Bresnick, 2001). An example is a study in which the acetylation patterns of histones H3 and H4 of the human β-globin locus were analysed in mouse erythroleukemic (MEL) cells. The acetylation status of the complete locus was compared to that of the locus after a deletion removing HS2-5 of the LCR and another deletion removing HS2-5 as well as an additional 27 kb of upstream sequences. The full locus and the first deletion showed similar basal levels of acetylation throughout the locus, whereas the second deletion line showed a pattern of hypo-acetylation. Furthermore, the full locus, which is transcriptionally active, showed peaks of histone H3 acetylation at the LCR and the active promoters, whereas in both of the deleted loci, which are transcriptionally inactive, no such peaks were observed (Schubeler *et al.*, 2000).

The enzymes responsible for the acetylation of the histone tails are the histone acetyltransferases (HATs). There are two classes of HATs: the nuclear HATs involved in transcriptional regulation and the cytoplasmic HATs involved in the acetylation of newly synthesised histones. The first nuclear HAT was identified in 1996 by Brownwell *et al.* and corresponds to yeast GCN5. GCN5 had already been identified as transcriptional co-activator (Georgakopoulos and Thireos, 1992) and was later shown, using yeast mutants, to be important for gene activation and acetylation of H3-lys-14 and H4-lys-8 and -16 (Kuo *et al.*, 1996 and 1998).

Following identification of GCN5 as a histone acetyltransferase, more proteins were recognised with HAT activity. Many of them are part of multi-protein complexes recruited to promoters, by interaction with DNA-bound activator proteins (Utley *et al.*, 1998) thus playing a direct role in activation of transcription (Larschan and Winston, 2001, Bhaumik and Green, 2001). Examples include PCAF, a human transcriptional co-activator similar to GCN5, p300/CBP, which are global co-activators and TFII250, which is a TBP-associated factor and plays a direct role in transcription initiation (Sterner and Berger 2000).

Besides histone tails, some transcription factors, including p53, EKLF, TFIIEβ, HMG1, GATA1 and ACTR are also substrates for HATs, potentially influencing their roles in the transcriptional process (Gu and Roeder, 1997; Boyes *et al.*, 1998; Chen *et al.*, 1999; Marizo *et al.*, 2000 and Zhang and Bieker, 1998). In the cases of p53, EKLF and GATA1, the acetylation site(s) map next to the DNA-binding domains. Acetylation may thus affect the DNA binding properties of these factors, for example, it may have a stimulatory effect (Gu and Roeder, 1997; Boyes *et al.*, 1998; Zhang and Bieker, 1998 and Martinez-Balbas *et al.*, 2000). On the other hand, acetylation of HMG1 results in disruption of DNA binding, because the acetylated lysines fall directly within the DNA-binding domain (Ugrinova *et al.*, 2001). Therefore, the common view that acetylation is positively affecting transcription does not always seem to hold true when it comes to the acetylation of transcription factors.

Other proteins like α-tubulin (Sterner *et al.*, 1979 and L'Herault and Rosenbaum, 1985) and the importin-α family of nuclear importer factors (Bannister *et al.*, 2000) are also target for acetylases.

These findings suggest that acetyltransferases have a wide range of proteins as substrates: DNA-binding proteins (histones and transcription factors), non-nuclear proteins (α-tubulin) and proteins that shuttle between the nucleus and the cytoplasm. This, in turn, indicates that acetylation has diverse consequences: chromatin and nucleosome remodelling, DNA-binding (see transcription factors), or protein-protein interactions, for example in the generation of a recognition site for bromodomain binding via the acetylation of histones (Dhalluin *et al.*, 1999). Finally, acetylation also seems to influence protein stability, as a correlation has been described between the acetylation of α-tubulin and the stability of microtubules (Takemura *et al.*, 1992).

The effect of acetyltransferases can be antagonised by the histone deacetylases or HDACs, first isolated by Taunton *et al.* (1996). The antagonistic effect of HDACs results in repression of transcription. There are three classes of HDACs that have been described so far (Taunton *et al.*, 1996; Verdel and Khochbin, 1999 and Imai *et al.*, 2000). The first class is related to the yeast protein RPD3, which was the first HDAC to be described (Taunton *et al.*, 1996). The second class of HDAC are related to yeast HDA1 (Rundlett *et al.*, 1996), later human and mouse homologues were also identified (Verdel and Kochbin, 1999; Fischle *et al.*, 1999 and Grozinger *et al.*, 1999). The third class takes its name from the SIR2 protein first identified in yeast, but which also includes recently identified mammalian homologs (Imai *et al.*, 2000). The existence of three different classes indicate that deacetylases, like acetyltransferases, have diverse activities.

Another class of enzymes that has gained a lot of attention in the unravelling of histone modifications and their role in chromatin remodelling and transcriptional regulation, are the histone methyltransferases. The modification of histones by methylation was described for the first time in 1964 by Murray, though it has gained considerable importance only in the last three years. Methylation takes place at lys-4, -9, and -27 of histone H3 and lys-20 of histone H4. The first methyltransferase to be discovered was Suv39 which methylates lys-9 of histone H3 (Rea *et al.*, 2000).

The enzymatic activity of histone methyltransferases resides in their SET domain (Kouzarides, 2002). At the moment, four groups of proteins have been identified possessing a distinct SET domain, the Suv39 group and the SET1, SET2 and RIZ groups. The proteins in the first group specifically methylate lys-9 of histone H3 (O'Carrol *et al.*, 2000; Tachibana *et al.*, 2001 and Yang *et al.*, 2002), whereas proteins belonging to the SET1 group have been shown to methylate lys-4 of histone H3. Proteins belonging to the SET2 group methylate histone H3, although the specific residues have not yet been identified. Finally, for the RIZ group no methylase activity has yet been demonstrated (Kouzarides, 2002).

Methylation of histones can correlate with either a repressed state or an active state of transcription. Methylation of lys-9 on histone H3 leads to the recruitment of HP1 thus resulting in heterochromatinisation and silencing. However, before methylation of lys-9 on histone H3 can occur, this has to be deacetylated. Deacetylation of lys-9 on histone H3 is a result of the recruitment of deacetylases by repressor proteins to the site where histone methylation has to take place. An example of such a repressor protein is the retinoblastoma co-repressor protein (RB). Studies in RB null cells showed that the *cyclin E*-promoter is undermethylated at lys-9 on histone H3 and no association of HP1 with the *cyclin E*-promoter occurs (Nielsen *et al.*, 2001). This indicates that RB activity plays a role in the methylation of lys-9 on histone H3 and the subsequent recruitment of the HP1 repressor (Brehm and Kouzarides, 1999).

On the other hand, methylation of lys-4 of histone H3 is correlated with an active state of transcription. Studies using the mating type loci in fission yeast have shown that inactive chromatin is enriched in lys-9 methylation and devoid of lys-4 methylation of histone H3. The reverse was found in transcriptionally active chromatin (Litt *et al.*, 2001 and Noma *et al.*, 2001).

Besides lysine methylation, the arginine residues arg-3, -17 and -26 of histone H3 and arg-3 of histone H4 are also targets for methyltransferases. To date, there are five arginine methyltransferases known. Only recently, using chromatin immunoprecipitation assays, has it become clear that arginine methylation can be correlated with an active state of transcription, much like histone acetylation (Ma *et al.*, 2001 and Bauer *et al.*, 2002).

An example of the histone code and the interplay between histone modifications is that of the modification status of lys-9 of histone H3. When acetylated by HATs, lys-9 on histone H3 codes for an active status of transcription, however, when it becomes deacetylated by repressor proteins such as

RB, lys-9 on histone H3 becomes the target for methyltransferases. The resulting methylation of lys-9 on histone H3 now codes for an inactive transcription state (Noma *et al.*, 2001).

Another example is that of acetylation of lys-14 of histone H3, which correlates with a transcriptional active state. This acetylation is preceded by and depends on the phosphorylation of ser-10 of histone H3 (Cheung *et al.*, 2000 and Lo *et al.*, 2000). Another combination of histone modifications that marks a transcriptionally active state is the methylation of arg-3 on histone H4, which precedes the acetylation of lys-8 and lys-12 (Wang *et al.*, 2001).

This implies that existing histone modifications can "recruit" new modifications, leading to the recruitment of proteins, or protein complexes, that alter the chromatin structure. This alternating results in either activation or repression of transcription (Strahl and Allis, 2000).

ATP-dependent chromatin remodelling

In order to facilitate transcription, nucleosomal arrays in the chromatin fiber often have to be remodelled or disrupted in such a way that transcription factors can bind to their recognition sites. Besides the histone tail modifications, ATP-dependent chromatin remodelling complexes have also been implicated in this process.

Two genetic screens for altered gene expression in *Saccharomyces cerevisiae*, one showing the importance of Snf genes for SUC2 expression and the second showing the necessity of Swi genes for HO expression, led to the discovery of the first chromatin remodelling complex SWI/SNF (Winston and Carlson, 1992). The complex has a molecular weight of 2 MDa and contains 11 proteins, of which the Swi2/Snf2 protein is responsible for nucleosome disruption. Complexes homologous to SWI/SNF include BRG1 and hBRM in human and the Brahma complex in *Drosophilla* (Khavari *et al.*, 1993 and Muchardt and Yaniv, 1993).

The disruption of the nucleosomal array by SWI/SNF is an ATP-dependent process. The SWI2/SNF2 subunit contains an ATP-ase and a helicase domain, both of which are necessary for chromatin remodelling (Cote *et al.*, 1994 and Kwon *et al.*, 1994).

Genome-wide expression studies in yeast, have shown that the action of SWI/SNF is limited to specific promoters rather than to chromosomal domains. (Sudarsanam *et al.*, 2000). The question remains as to how SWI/SNF binds to target genes since there are no known DNA binding activities in the complex . Two models have been proposed for the binding of SWI/SNF to its targets. The first model suggests that SWI/SNF is recruited to promoters by RNA Pol II (Wilson *et al.*, 1996 and Cho *et al.*, 1998) and the second model argues that SWI/SNF is recruited to promoters by transcriptional activators and thus before RNA Pol II is present (Neely *et al.*, 1999 and Natarajan *et al.*, 1999). Several *in vitro* and *in vivo* studies in yeast provide evidence for the occurrence of both models (Cho *et al.*, 1998; Wilson *et al.*, 1996 and Gregory *et al.*, 1999). Support for the first model is provided by studies showing co-immunoprecipitation of components of the SWI/SNF complex with the RNA Pol II holoenzyme. Using antibodies against SRB, which has been shown to be tightly associated with the C-terminal repeat domain of RNA Pol II in the holoenzyme, and SWI/SNF components, in both cases showed co-immunoprecipitation of SRB with SWI/SNF components. Subsequent purification of the RNA Pol II holoenzyme showed that the SWI/SNF proteins co-elute with known components of the holoenzyme. These data indicate that SWI/SNF, together with SRB, is bound to the RNA-pol II holoenzyme. This binding facilitates the recruitment of SWI/SNF to the gene promoters by RNA-Pol II and the subsequent remodelling of the chromatin template (Wilson *et al.*, 1996).

An example of the second model is provided by studies on the HO gene in yeast. Chromatin immunoprecipitation studies at the HO-promoter have shown that SWI/SNF is recruited to the promoter via an interaction with the swi5 transcription activator (Cosma *et al.*, 1999). These data are consistent with *in vitro* studies, which have shown that SWI/SNF, when purified or in whole cell extracts, directly interacts with swi5 (Neely *et al.*, 1999). After recruitment by swi5 of the SWI/SNF to the HO-promoter, this can remodel the nucleosomal array thus allowing the binding of other factors involved in the transcription of the HO gene.

The two models mentioned also imply different roles for SWI/SNF remodelling at different promoters in transcriptional activation. When a transcriptional activator, like swi5, can bind strongly to DNA, SWI/SNF may play a role in further remodelling the nucleosomal array to allow other proteins to bind such that transcriptional activation can take place. However, when a transcriptional activator has only a low binding activity for DNA, as is the case for activators of the Gal4 gene (Burns and Peterson, 1997), SWI/SNF activity is necessary to remodel the nucleosomal array before the first activator can bind, which will then be followed by the binding of other proteins of the transcriptional complex (Neely *et al.*, 1999; Natarajan *et al.*, 1999 and Wallberg *et al.*, 2000). Interaction between SWI/SNF and activators could thus vary at different promoters or SWI/SNF could function in multiple ways at a single promoter depending on the activator present.

Interactions of human SWI/SNF homologs with tissue specific transcriptional activators have also been reported. The hSWI/SNF-containing E-RC1 complex purified from erythroid cells directly interacts with the zinc finger domain of EKLF and is functionally important for the efficient transcription of the β-globin gene in *in vitro* assays (Armstrong *et al.*, 1998; Kadam *et al.*, 2000 and Lee *et al.*, 1999).

Besides the "classical" role in gene activation, SWI/SNF also displays repressor activities (Sudarnasam *et al.*, 2000 and Holstege *et al.*, 1998). Possible mechanisms by which this repressors activity could take place have been reported. It has been shown that SWI/SNF can remodel nucleosomes between two states, "inactive" and "active" with equal ability (Schnitzler *et al.*, 1998), which could indicate that the SWI/SNF can also repress genes by creating an inactive nucleosomal state. SWI/SNF could bind repressors, instead of transcriptional activators, and finally the combination of SWI/SNF with histone modifications could also lead to repression (Strahl and Allis, 2000). The Mi-2 complex (now called NuRD for nucleosome remodelling and deacetylation) demonstrates that last point. This complex is composed of the Mi-2 protein, which is a member of the Snf2 superfamily of ATP-ases and the RPD3 protein, which is a HDAC. More complexes containing an ATP-ase and HDAC subunit have been identified and are thought to repress transcription through the combined properties of ATP-dependent remodelling and HDAC activity (Tong *et al.*, 1998 and Zhang *et al.*, 1998).

Additional ATP-dependent chromatin remodelling complexes besides SWI/SNF have also been identified and can be classified into four families depending on the ATPase activities they contain: SWI/SNF, ISWI, Mi-2 and Ino-80 (Cairns 1998; Alfas and Kingston, 2000; Shen *et al.*, 2000; and Vignal *et al.*, 2000).

The ISWI family has also been studied extensively. Family members of ISWI all contain the same ATP-dependent chromatin remodelling subunit ISWI (Corona *et al.*, 1999). Three ISWI-containing members have been isolated from *Drosophila*: NURF, ACF and CHRAC (Ito *et al.*, 1997; Varga-Weisz *et al.*, 1997 and Tsukiyama and Wu, 1995). NURF contains 4 subunits and has a molecular mass of 500 kDa (Tsukiyama *et al.*, 1995 and Xiao *et al.*, 2001). Like SWI/SNF, NURF is involved in allowing the binding of specific activators to DNA, for example, GAGA to the hsp70-promoter (Tsukiyama *et al.*, 1994) or of activator proteins to the Gal4-E4-promoter (Mizuguchi *et al.*, 1997 and 2001). NURF binds stably to nucleosomes in an ATP-dependent manner and facilitates nucleosome repositioning to allow factor binding. Unlike SWI/SNF, however, ISWI ATPase activity is simulated by nucleosomal DNA, whereas SWI/SNF also acts on free DNA. ACF contains 2 subunits and can order an evenly spaced array of randomly assembled nucleosomes on DNA and can further mobilize nucleosomes to facilitate interactions with DNA-binding proteins (Ito *et al.*, 1997 and 1999). Related to ACF is CHRAC, which contains 4 subunits with the acf1 subunit also present in the ACF complex (Ebenharter *et al.*, 2001). CHRAC also appears to function as a nucleosome-spacing factor thus enhancing the accessibility of chromatin (Langst and Becker, 2001).

To facilitate transcription, co-operation between the chromatin remodelling complexes and complexes affecting histone modifications must take place. For the HO gene in yeast, it has been shown that both SWI/SNF and the HAT complex SAGA are recruited one after the other to the promoter. Both complexes are necessary for transcriptional activation (Bhoite *et al.*, 2001).

In conclusion, the process of chromatin remodelling in transcriptional activation encompasses three collaborating events: first the ATP-dependent chromatin remodelling complexes are recruited to the target of activation or repression, where they alter nucleosomal structure, secondly the histone modifying enzymes like HATs, HDACs and methyltransferases are recruited, which further modify the chromatin structure, and enable the third step, the binding of a third group of proteins, which alter the non-histone part of the chromatin thus affecting transcriptional activity.

Polycomb and Trithorax

Two other groups of proteins that regulate transcription at the chromatin level are the Polycomb group (PcG) and the Trithorax group (TrxG). In general the PcG proteins are repressors that maintain the off-state of a gene whereas the TrxG proteins are activators that maintain transcription (Simon and Tamkun, 2002). These complexes appear to be involved in the maintenance and not in the establishment of active or repressive states of chromatin and transcription (Simon, 1995 and Kennison, 1995).

The function of PcG was first suggested by mutations in *Drosophila*. These mutants showed posterior transformations caused by the derepression of homeotic genes (Simon, 1995). Many of the TrxG members were identified in genetic screens as suppressors of the phenotypes of PcG mutations (Kennison and Tamkun, 1988).

It is now clear that PcG and TrxG are not involved in the establishment of the precise patterns of expression of the homeotic loci, which are set up by (short-lived) specific transcription factors earlier in *Drosophila* development (Bienz and Muller, 1995). After the "decay" of these proteins, the repressed or activated states of homeotic genes are maintained by the PcG and TrxG proteins (Pirrotta, 1998).

Mammalian homologs of PcG and TrX-G have also been identified (Jacobs and van Lohuizen, 2002). In addition to their function in the control of homeotic gene expression, PcG and Trx-G complexes are also involved in the control of several other processes including cell proliferation, haematopoiesis, neuronal development, sex determination and cell cycle regulation (Jacobs and van Lohuizen, 2002 and Lund and van Lohuizen, 2002). Oncogenic and tumour suppressive activities have been implicated for PcG and Trx-G proteins. An example of this is the mammalian TrX-G MLL homologue, which has been shown to be involved in myeloid and lymphoid leukemias (Rubnitz *et al.*, 1996 and Corral *et al.*, 1996). The mammalian PcG BMIl gene has also been associated with lymphogenesis and was originally identified in a screen for oncogenes (van Lohuizen *et al.*, 1991 and Haupt *et al.*, 1991). The BMI1 gene has been shown to interact with distinct heterochromatin domains in tumour cell lines. The Ink4 tumour suppressor locus has been shown to be a critical target of the transcriptional repressor BMI1. In BMI1 knockout mice an increase in apoptosis is observed, which can be rescued by deleting the Ink4 locus (Jacobs *et al.*, 1999). Recent experiments, in which lung tumours were compared with healthy lung samples, indeed showed an inverse correlation between the expression of the Ink4 locus and the BMI1 gene. Ink4 expression levels were high in normal tissues and low in the tumour tissues, with BMI1 expression levels showing the opposite patterns. These data support the suggestion for a role of BMI1 in carcinogenesis (Vonlanthen *et al.*, 2001).

There are specific elements through which PcG and TrxG confer their action, the so-called Polycomb and Trithorax response elements (PRE and TRE). These elements can vary in size from 100 bp to a few kb and different PcG or TrxG complexes can bind and confer their action via discrete response elements within the same regulator region of the gene (Tillib *et al.*, 1999).

It appears that TRE and PRE elements are located close to each other, sometimes separated by 30-40 bp (Brock and van Lohuizen, 2001). This suggests an intermingling of these elements in the activation and repression of genes. Evidence for this came from studies in *Drosophila* in which it has been shown that maintenance of repression through the *Fab-7* and *Scr* PREs is affected by both PcG and TrxG mutations (Hagstrom *et al.*, 1997; Gindhart and Kaufman, 1995 and Cavalli and Paro, 1995). It was therefore suggested to rename PREs and TREs as maintenance elements (ME) (Cavalli and Paro, 1995).

How repression or activation through MEs is regulated is not clear. It is not a question of competition because both PcG and TrxG can bind to the same ME regardless of the activity of the target locus. This was shown in studies in which the repression of a gene by PcG proteins was reversed by the co-expression of Trx-G proteins (Cavalli and Paro, 1995 and Zink and Paro, 1995). This shows that although a target locus is suppressed, Trx-G proteins can still bind to their recognition site and reverse the activity of the target locus. It could be a case of other proteins recruiting the PcG or TrxG to the right ME, or rather a histone modification that marks the ME such that only one of the two protein groups can bind (Ekwall *et al.*, 1997).

Insight in the molecular basis of PcG-mediated repression and TrxG-mediated activation has been obtained through the isolation and characterization of protein complexes. Presently there are two complexes that contain PcG proteins: a 1-2M Da complex, PRC1 (Shao *et al.*, 1999 and Saurin *et al.*, 2001), and a 600 kDa complex, ESC/E(Z) (Tie *et al.*, 2001 and Chang *et al.*, 2001). In addition, four complexes containing TrxG proteins have been identified: a 2 MDa BRM complex, a 2 MDa ASH1complex, a 0.5 MDa ASH2 complex, and a 1 MDa TRX complex (TAC1) (Papoulas *et al.*, 1998 and Petruk *et al.*, 2001).

These complexes appear to have different activities and may thus play different roles in gene repression and activation. PRC1 inhibits the remodelling of nucleosomal arrays by SWI/SNF (Breiling *et al.*, 1999 and Francis *et al.*, 2001) and interacts with TBP-associated factors, as shown by co-purifications (Saurin *et al.*, 2001). The functional consequence of this interaction is, however, not clear. The second PcG complex ESC-E(Z) co-immunoprecipitates and co-fractionates with HDAC1 and HDAC2 in extracts from human cells (van der Vlag and Otte, 1999). Purification of the complex from *Drosophila* extracts followed by mass spectrometric analysis and co-immunoprecipitations confirm this finding (Tie *et al.*, 2001). This indicates that HDAC function in PcG repression is linked via the ESC-E(Z) complex. For the trithorax complexes also different activities have been described. The BRM complex has been suggested to play a role in ATP-dependent chromatin remodelling, since *in vitro* studies have shown that the BRM complex catalyses ATP-dependent alterations in nucleosomal organisation (Kal *et al.*, 2000) and TAC1 has been shown to have a role in histone modification (Petruk *et al.*, 2001). The characterisation of the TAC1 complex has shown that TAC1 contains a member of the CBP/p300 HAT family. When this subunit of TAC1 is mutated there is a decrease in transcription of HOX genes and reporter genes (Petruk *et al.*, 2001).

How do these complexes together play a role in the maintenance of “on” and “off” states of chromatin? A model for the multistep action of TrxG and PcG complexes has been proposed (Simon and Tamkun, 2002). In the Trx-G pathway, TAC1 first acetylates histone tails, which may lead to the recruitment of the BRM remodelling complex, thus remodelling the nucleosomal array and rendering the DNA more accessible to the transcriptional machinery. In the PcG pathway, the ESC-E(Z) complex, which acts earlier in development, first deacetylates nucleosomes, thus creating a histone code that attracts subsequent binding of PRC1. When PCR1 is bound, it can maintain the silenced state by counteracting remodelling complexes (Simon and Tamkun, 2002).

DNA methylation

DNA methylation represents another mechanism for global gene repression. When a promoter and/or other regulatory sequences of a gene become methylated, gene expression is repressed. The C-5 position of cytosine in 5’-CpG-3’ dinucleotides is the target for DNA methylases. CpGs from both DNA strands are symmetrically methylated to result in the final silencing of a gene (Antequera and Bird, 1993 and Ohki *et al.*, 2001). 60-90% of all the CpG sequences present in mammalian genomic DNA are methylated.

CpGs that are not methylated are often clustered in so-called CpG-islands, usually found in the regulatory elements (Antequera and Bird, 1993). It has been previously shown that, at least in some cases, DNA methylation can prevent the binding of transcription factors to their DNA binding sites, as is the case for CREB, for example (Iguchi-Ariga and Schaffner, 1989). However, other transcription factors, like Sp1, can still bind to their methylated recognition site, indicating that another mechanism is also playing a role in the inhibition of the binding of transcription factors by methylation. *In vitro* studies have shown that specific proteins (MeCP1 and MeCP2) can repress transcription by binding to methylated DNA, thus preventing the binding of transcription factors (Boyes and Bird, 1991 and 1992). *In vivo* both mechanisms may play a role in transcriptional regulation.

In addition, a relationship between DNA methylation and chromatin structure was hinted at when early on it was shown that artificially methylated DNA can result in a distinctive chromatin structure when integrated into the genome (Keshet *et al.*, 1986). Further studies implicated a connection between histone deacetylation and DNA methylation when it was shown that a specific inhibitor for HDACs, trichostatin A (TSA), can substitute for the DNA-demethylating agent 5-aza-2’-deoxycytidine, in that both could restore transcription from a repressed methylated template (Kass *et al.*, 1997; Chen and Pikaard, 1997 and Eden *et al.*, 1998). Another link between DNA methylation and histone deacetylation was the finding that MeCP2, a protein that specifically binds to methylated DNA via its methyl CpG-binding domain (Lewis *et al.*, 1992; Nan *et al.*, 1993 and Nan *et al.*, 1996), copurifies with a component of the mSin3A/HDAC complex (Nan *et al*., 1998 and Jones *et al.*, 1998). Co-immunoprecipitations showed that mSin3A is the preferred partner of MePC2, suggesting that MeCP2 recruits HDAC to the chromatin template of methylated DNA (Nan *et al*., 1998). This interaction can provide a mechanistic link between DNA-methylation and chromatin structure and transcriptional repression.

Recently, another intriguing connection between DNA-methylation and histone modifications has been made. It was shown in *Neurospora* that the dim2 gene, which encodes a protein with a C-terminal domain homologous to known histone methyltransferases, is responsible for all cytosine methylation (Kouzminova and Selker, 2001). This observation may provide a direct evolutionary link between DNA methylation and histone methylation in transcriptional repression.

There are two classes of DNA methylases: the de novo methylases, Dnmt3a and Dnmt3b, and maintenance methylase, Dnmt1. The de novo methylases add a methyl group to unmethylated CpG-pairs on both strands of DNA without high sequence specificity (Okano *et al.*, 1998 and Lyko *et al.*, 1999). Maintenance methylation by Dnmt1 provides the methylation of hemi-methylated CpG-pairs on the newly replicated DNA strand (Lyko *et al.*, 1999 and Bestor *et al.*, 2000).

Both de novo methylases and maintenance methylases play a role during mammalian embryogenesis. During the embryo implantation stage, genome-wide methylation patterns are lost (with very few exceptions, e.g. imprinted regions). The new methylation patterns after this period of embryogenesis are established through the combined action of de novo methylases and subsequently by maintenance methylases (Okano *et al.*, 1999). Knockout mouse models for both types of methylases have shown that both Dnmt1 and Dnmt3b are necessary for embryonic development and Dnmt3a for postnatal development (Lie *et al.*, 1996 and Okano *et al.*, 1999).

Besides these enzymatic activities, Dnmt1 has been shown to interact with HDAC1 and HDAC2. It was shown that during late S-phase HDAC2 co-localises with Dnmt1 in heterochromatin (Rountree *et al.*, 2000). This interaction could indicate that Dnmt1 directly represses transcription together with HDAC. Dnmt3a and b have been implicated to play a role in differentiation and cell growth, because it methylates unmethylated CpG-pairs creating hemi-methylated pairs which can be methylated by the maintenance methylases. In addition, Dnmt3a and b have been implicated to play a role in altered methylation patterns in tumourigenesis (Nakao, 2001).

DNase sensitivity

Chromatin associated with transcriptionally active states is more sensitive to the action of nucleases (Weintraub and Groudine, 1976 and Wood and Felsenfeld, 1982), for example, DNase I. DNase I nicks double stranded DNA in a non-sequence-specific manner, but with a distinct preference for active chromatin in contrast to inactive chromatin. This distinction has often been used as a molecular tool to study chromatin structure.

Two types of DNase I sensitivity have been recognised: general sensitivity and hypersensitivity. General sensitivity to DNaseI digestion is found in areas of gene domains that are transcriptional active or potentially active. Hypersensitivity to DNase I digestion is found in smaller areas of around 200-600 bp in size within the areas of general DNase I sensitivity. These sites are often located within the regulatory elements of genes like enhancers, promoters and locus control regions and are called hypersensitive sites. The first demonstration of a DNAse I hypersensitive site was shown by Wu *et al.* in 1979 in the *Drosophila* hsp 70 gene. Hypersensitive sites are thought to reflect areas with a less dense nucleosomal packaging which contain multiple sequences for promoter-specific DNA binding proteins (Emerson *et al.*, 1985). However, the exact structural basis for DNase I sensitivity is not clear. Chromatin features such as the absence of linker histone H1 and increased histone acetylation in DNase I sensitive areas have been suggested as an explanation (Smith *et al.*, 1984 and Hebbes *et al.*, 1994). DNase I hypersensitive sites can be classified as constitutive, inducible, tissue-specific and developmental stage specific (review Gross and Garrard, 1988).

An example of a gene locus containing well characterised DNase I hypersensitive sites is the human β-globin locus. Two kinds of hypersensitive sites were mapped within the locus. Minor hypersensitive sites are observed when higher concentrations of DNase I are used and have been found close to the promoters and local enhancers of the β-globin genes (Stalder *et al.*, 1980; Charnay *et al.*, 1984 and Forrester *et al.*, 1986). Major hypersensitive sites were mapped in the locus with lower concentrations of DNase I. Of these, five map within the locus control region upstream of the genes and one maps to the 3' of the locus (Tuan *et al.*, 1985; Grosveld *et al.*, 1987 and Forrester *et al.*, 1987).

Transcription initiation, elongation and termination

After chromatin has been made accessible by remodelling and histone modification, the transcriptional machinery can "land" on the DNA and the process of transcription can take place. Gene transcription is carried out by three classes of polymerases: RNA polymerase I, II and III. The three RNA polymerases have different gene specificities: RNA pol I is responsible for transcribing ribosomal RNA, RNA Pol II synthesises all the heterocellular genes and finally RNA Pol III is responsible for the transcription of tRNA, 5 S RNA and small nuclear RNA (Ogbourne and Antalis, 1998). The initiation of transcription is a highly regulated event and plays a major role in gene regulation. The transcription of the globin genes plays an important role in this thesis, therefore, I will focus on the role of RNA Pol II in the transcriptional process.

RNA Pol II cannot by itself initiate transcription at a specific promoter region and requires a set of transcription factors to do so. These factors have to assemble at the promoter of the gene before transcription can take place. This process can be influenced by regulatory signals within the cell and in this way transcription can be accelerated or repressed. The factors needed for transcriptional activation were identified as basal transcription factors (TF) which assemble into a complex at the promoter of the gene and subsequently recruit RNA Pol II to the promoter (Johnson and McKnight, 1989). Most of the TFs have two distinct domains, one for the binding to the specific regulatory sequences in the DNA and one that interacts with the transcription machinery and accelerates the rate of transcription initiation. The activator domains of TFs can be classified into three classes; acidic, glutamine-rich and proline-rich domains (Alberts *et al.*, 1994).

Some steps in the assembly of the transcriptional complex can be rate-limiting and the TFs play a role in overcoming this limiting step. For example the TFs with acidic domains have been shown to overcome the rate-limiting step of TFIIB entry into the complex (Lin and Green, 1991).

TFs themselves are also regulated. This can be through ligand binding, protein phosphorylation, addition of a secondary subunit or the release from a tight complex with a specific inhibitor (Berk, 1989 and Hunter and Karin, 1992).

For the specific and regulated initiation of transcription a large transcriptional complex of RNA Pol II together with TFs, has to be formed (Buratowski *et al.*, 1989 and Conaway and Conaway, 1993). The assembly of the transcriptional apparatus starts with the binding of TFIID to a TATA-sequence (Fig. 2). This sequence is located around 30 bp upstream of the transcription start site. TFIID itself is composed of several subunits: TBP and TBP-associated factors (TAFs) (Dynlacht *et al.*, 1991; Tanese *et al.*, 1991 and Zhou *et al.*, 1992). TBP is the factor responsible for binding to the TATA-box and the TAFs are needed in order to mediate transcription regulation by upstream activating factors. After TFIID, TFIIA is recruited to the complex. TFIIA is important for the stabilisation of TFIID on the promoter and to counteract inhibitory factors, which cause TFIID to dissociate (Zawel and Reinberg, 1993). The TFIID-TFIIA complex undergoes a conformational change to allow TFIIB to bind (Horikoshi *et al.*, 1988 and Chi and Carey, 1996). TFIIB then facilitates the binding of the non-phosphorylated form of the TFIIF-RNA-pol II sub-complex to the transcription complex (Maxon *et al.*, 1994). The complex thus assembled is thought to melt the DNA around the start site of transcription to form an open complex. To be able to continue into the next step of transcription, that of elongation, promoter clearance has to take place. For this step, TFIIH and TFIIE are thought to be responsible (Goodrich and Tjian, 1994). TFIIE is thought to recruit TFIIH to the initiation complex with TFIIH being responsible for the actual promoter clearance. TFIIH has many subunits of which one is a protein kinase. This kinase phosphorylates the C-terminal Domain (CTD) of RNA Pol II (Maxon *et al.*, 1994), resulting in the release of RNA Pol II from the initiation complex and transcription can proceed.

After these steps, the next action of transcription, elongation, can take place. There are two phases in transcription elongation. The early phase, which is characterised by a hypophosphorylated CTD of RNA Pol II (by DSIF and Factor 2) and a late elongation phase, which is characterised by a hyperphosphorylated CTD. The hypophosphorylated RNA Pol II has to be re-phosphorylated to proceed with elongation. Several factors have been found to be involved and include DmsII, TFIIF, ELL and Elongin (for review see Shilatifard *et al.*, 1997). If re-phosphorylation of the CTD does not occur, elongation is blocked 500 bp downstream of the start site (Chodosh *et al.*, 1989). If CTD phophorylation does take place, the late elongation phase proceeds and nascent RNA (nRNA) is synthesised until it is stopped by a termination signal.

Termination of transcription by RNA polymerases, at least for RNA pol I and RNA Pol III, appears to be mediated by specific DNA sequences. RNA pol I recognises a specific18 bp recognition sequence downstream of the mature 3' end of the newly synthesised RNA (Proudfoot, 1991). RNA pPol III terminates at uridine 2, 3 or 4 in a stretch of four uridines which are surrounded by GC pairs at the 3'end of the gene (Platt, 1986). The specific termination signal for RNA Pol II, however, has still to be determined.

After termination of transcription the nascent RNA (nRNA) matures into mRNA. During maturation, both the 5'end and the 3'end of the nRNA are covalently modified. The 5'end is capped by the addition of a methylated G nucleotide. This cap plays an important role in the initiation of protein synthesis. The 3'end of the nRNA is modified by the addition of a poly-A tail. This tail is important for the export of mRNA from the nucleus and for the stability of the mRNA in the cytoplasm. It also serves as a recognition signal for the ribosomes. Furthermore the intron sequences are removed from the nRNA during the process called splicing, leaving just the mRNA (Alberts *et al.*, 1994).

Transcriptional regulation

The regulation of transcriptional activation is achieved through the interactions of distinct elements such as enhancers, silencers and insulators. These elements are discussed below.

Enhancers and silencers

Enhancers were originally identified in transient transfection assays as sequences that were able to activate transcription in an orientation and distance independent manner with respect to the promoter to which they are linked (Banerji *et al.*, 1981 and Moreau *et al.*, 1981).

Enhancers can be found located from within the promoter of a gene up to several kilobases downstream and/or upstream of a gene. The first enhancer to be identified using transient transfection assays was a 200 bp element from the SV40 virus (Banerji *et al.*, 1981). The sizes of enhancer elements vary between 50 bp to 1.5 kb and contain a collection of protein binding sites to which both tissue specific and ubiquitous transcription factors can bind. By multimerising transcription factor binding sites a functional enhancer can be created *in vitro* (Zenke *et al.*, 1986). Interactions with different proteins give rise to cell-type and developmental stage specificity in the action of enhancers (Dynan, 1989), the first tissue-specific enhancer transcribed is the B-cell specific enhancer located at the 3'-end of the rat IgH locus (Pettersson *et al.*, 1990).

Enhancers have to interact directly with the promoter of the gene, as they themselves cannot initiate transcription. Three models have been proposed for this interaction. The prevailing favoured model is that of DNA looping. The proteins bound to the enhancer interact with the proteins bound to the promoter and intervening DNA sequences are looped out (Ptashne, 1988 and Ptashne and Gann, 1997). Experiments in which it was shown that the transcription of a β-globin gene could be stimulated using a biotin-streptavidin bridge between the β-promoter and enhancer sequences which resided on different molecules (Muller *et al.*, 1989) and transvection studies in *Drosophila*, in which an enhancer is shown to activate a gene in trans (Henikoff,1997), have provided support for the looping model.

Two other models have been proposed. First, a tracking mechanism, in which the transcription factors use the enhancer site as a landing platform for the assembly of a nucleoprotein complex which then travels along the chromatin fiber to activate the first gene it encounters (Heredeen *et al.*, 1992). Second, the accessibility model, in which the enhancer provides a favourable chromatin environment for transcription of a gene, by antagonising repressive chromatin structures (Weintraub, 1988; Walters *et al.*, 1995 and 1996). The tracking and accessibility models are not able to explain the results of the trans-activation experiments described above and are thus less likely to explain the physical interaction between an enhancer and promoter. The interaction models described also apply with some variations to the function of the LCR of the β-globin locus. These models are described in the section concerning the locus control region.

The presence of an enhancer identified in transfection assays as part of a transgene, is in most cases unable to direct normal expression levels of the transgene in mice or in stably transfected cells. Instead, expression levels of the transgene are often low and the tissue and developmental regulation are disturbed. This is the result of random integration of the transgene into the host chromosomes, such that integration of the same transgene at different positions often shows different expression patterns. Classical enhancers are not able to shield a transgene from these effects, resulting in the absence of normal expression levels. This led to the identification and formulation of a new functional element: the locus control region. Such an element was first identified in the human β-globin locus (Grosveld *et al.*, 1987)(for more information about the human β-globin locus control region see section on the locus control region). More LCR elements have been identified since then, like the LCR from the mouse β-globin locus and the LCR of the CD2 gene (Greaves *et al.*, 1989 and Festenstein and Kioussis, 2000).

A LCR is functionally defined as an element that can confer position independent, copy number dependent and tissue specific activation of a gene in transgenic mouse assays (Grosveld *et al.*, 1987). A LCR resembles an enhancer in that it contains a high concentration of binding sites for specific transcription factors and it can activate genes over long distances (Talbot *et al.*, 1989; Dillon and Grosveld, 1993 and Kioussis and Festenstein, 1997). At the same time, a LCR is distinguishable from a classical enhancer by its property to confer chromosomal position independent transcription in transgenic assays (Grosveld *et al.*, 1987 and Blom van Assendelft *et al.*, 1989). Whereas a LCR can counteract the negative influences of chromatin, the classical enhancer is affected by the structure of surrounding chromatin at the site of transgene integration (Wilson *et al.*, 1990). LCRs and enhancers are therefore defined by different functional assays. Transient assays in which DNA is not integrated in chromatin can be used to identify enhancers, whereas, LCR activity will not necessarily be observed in these assays (Hug *et al.*, 1992; Tuan *et al.*, 1989; Pruzina *et al.*, 1991; Philipsen *et al.*, 1990 and Fraser *et al.*, 1990). Some elements will merely show activity in one of the assays, such as HS3 and HS4 of the human β-globin LCR, which are only active in a chromatin context (Pruzina *et al.*, 1991; Philipsen *et al.*, 1990 and Fraser *et al.*, 1990), whereas other elements, like HS2 of the LCR, show activity in both assays (Tuan *et al.*, 1989; Sorrentino *et al.*, 1990 and Talbot and Grosveld, 1991). This indicates that there is functional overlap between a classical enhancer and the LCR.

Silencer elements resemble enhancers, in that they contain multiple factor binding elements and they act on a promoter in an orientation and position independent manner (Brand *et al.*, 1985). However, instead of activating a gene they suppress it. The first silencing elements were identified in the yeast mating type loci (Brand *et al.*, 1985), and subsequently in many more gene systems, like the human ε-globin gene (Cao *et al.*, 1989) and the human thyrotropin-β gene (Kim *et al.*, 1996). Often, enhancers and silencers work together to ensure that gene expression is tissue and developmental stage specific (Huang *et al.*, 1993 and Trepicchio *et al.*, 1993). To date, a number of silencers have been described, which all confer different actions on genes and their promoters (Ogbourne and Antalis, 1998).

Silencers are divided into two functional groups: classical silencers, as described by Brand, and negative regulatory elements (NREs). Classical silencers are thought to confer their action through the binding of repressor proteins and subsequent interaction with promoters, and/or by changing chromatin structure or via physical interference with the transcription initiation process by blocking TFIIH activity at the transcription start site (Kim *et al.*, 1996 and Liu *et al.*, 1996). However, no definite mechanism has yet been described.

NREs are position-dependent silencers and have been identified in promoters, introns and exons (Clark and Docherty, 1993). They passively repress transcription by binding repressor proteins and thereby physically inhibit the binding of transcription factors or other factors that play a role in transcriptional activation (Dong and Lim, 1996; Gumucio *et al.*, 1993 and Peters *et al.*, 1993). An example of NRE function can be seen in the deletion of an element of 70 bp in the third exon of the human α1-chimaerin gene. This deletion resulted in a 5-6 fold increase of promoter activity (Dong and Lim, 1996). Addition of the NRE element to a heterologous TK element showed an orientation independent but position dependent repression of promoter activity (Dong and Lim, 1996).

Insulators

The presence of elements like insulators was proposed when the idea of different functional domains in the genome was put forward. Inherent to this was the suggestion that elements must exist that would prevent enhancers from acting on the wrong domain (Eisenberg and Elgin, 1991).

Two predictions were made for the function of such elements. The first one predicts that when such an element is placed on both sides of a transgene it should protect, or insulate, this from chromosomal position effects. Secondly, when placed between an enhancer and a promoter, activation of the gene should be blocked by this element (Eisenberg and Elgin, 1991). Using these assays, Kellum and Schedl showed that the scs and scs' elements from the 87A7 *Drosophila* hsp70 heat shock locus could block the expression of a gene when placed between the enhancer and promoter, while other elements from the same locus could not. In the same studies they also showed that the scs and scs' elements themselves do not contain enhancer activities (Kellum and Schedl, 1992). In addition, using the mini *white* gene of *Drosophila* as a reporter, they showed by scoring for eye colour in transgenic flies that the scs and scs' elements could insulate a gene from position effects (Kellum and Schedl, 1991).

The assays led to the further description of additional insulators of which the best studied, in addition to the scs and scs' sequences (Kellum and Schedl, 1991), are the *gypsy* transposon (Corces and Geyer, 1993) and HS4 of the chicken β-globin locus (Chung *et al.*, 1993).

Enhancer blocking assays showed that the binding sites for *Drosophila* suppressor of Hairy-wing [su(hw)] protein in the *gypsy* retrotransposon also confer insulator properties (Corces and Geyer, 1993). Recent results suggest, that the gypsy insulator affects the enhancer-promoter interactions by affecting chromatin structure. It was shown that in the presence of the Su(Hw) protein and a second component, the modifier of mdg4 protein, accessibility of DNA for nucleases was increased in the promoter-proximal but not in the promoter-distal region (Chen and Corces, 2001).

The first vertebrate insulator described was the HS4 of the chicken β-globin locus (Chung *et al.*, 1993). This element was tested using the enhancer blocking assays in human erythroid cell-lines. It was shown that cHS4 can insulate a reporter gene containing the β-globin promoter from the effects of the LCR (Chung *et al.*, 1993). The minimal core required for insulation contains binding sites for the CTCF transcription factor. CTCF binding sites are necessary and sufficient for the insulating effects of HS4 in this assay (Bell *et al.*, 1999 and Bell *et al.*, 2001). The core of HS4 was also tested in the position effect insulation assay. The positioning of two cHS4 elements around a transgene showed that the cHS4 can protect a transgene from position effects, this has been shown both in *Drosophila* and transgenic mice (Chung *et al.*, 1993 and Pikaart *et al.*, 1998). However, for this insulation the CTCF sites were not necessary as was shown by mutation analysis. These observations suggest that there are two overlapping insulator activities within the HS4 of the β-globin locus.

Exactly how insulators work is still a matter of debate. Two models have been proposed to account for the basis of insulator function: the local interaction and the structural model (Zhan *et al.*, 2001 review). The first model suggests a local interaction between proteins bound to the insulator element and proteins bound to the enhancer. This blocks the interaction of proteins in the enhancer with the promoter and the gene is thus repressed. Variations on the principle put forward by this model are the decoy looping model (Fig.3A), in which the looping of the enhancer to the promoter is blocked by the formation of an additional loop (Geyer, 1997 and Gerasimova and Gorces, 1998), and the derailment of tracking model (Fig. 3B) (Dorsett, 1993), in which the spreading of activation signals from the enhancer are blocked. The two models can be distinguished by the directionality of the insulator and the bidirectionallity of the enhancer.

The structural model proposes that insulating function is coupled to a structural role in higher order nuclear organisation. This could be established by the formation of loops by insulator proteins and attachment to the nuclear scaffold at MARs and SARs (Cai and Chen, 2001 and Murayova *et al.*, 2001). This would constrain the chromatin thus hindering transcriptional activation.

The HS5 of the human β-globin locus has also been reported to contain insulator properties like cHS4 (Li and Stamatoyannopoulos, 1994). Studies in which the LCR is reversed, show a downregulation of the expression of all the globin genes, indicating that HS5 might shield the genes form the other hypersensitive sites (Tanimoto *et al.*, 1999). However, other studies show contrary results. For example, the deletion of HS5 of the mouse β-globin locus, which is highly homologous to the human HS5, has only a minimal effect on transcription, indicating that HS5 is not necessary to protect the β-globin genes against surrounding chromatin (Bender *et al.*, 1998 and Farell *et al.*, 2000). Furthermore HS5 cannot shield the globin genes from the effect of the LCR in transgenis mice (Zafarana, thesis 2001). These observations indicate that HS5 by itself does not have (strong) insulator capacities. With the different studies contradicting each other it is not clear whether HS5 does or does not act as an insulator.

Regulation of gene expression

The dogma on gene expression is, that transcription activation and gene expression are following a rate or analog model, also called deterministic or gradient model. This model implies that gene expression levels are regulated by the rate of transcription as response upon a stimulus in the environment of the cell. An other model for gene activation is a probability or stochastic model. This model states that the probability and frequency of gene activation determines the level of gene expression rather then the increase of gene expression per cell upon a stimulus.

Several *in vitro* studies using inducible reporter gene assays show evidence for a probability or stochastic gene expression model. They find bimodal expression patterns, either a cell does or does not express and upon induction just more cells start expressing instead of higher expression levels per cell (Femino *et al.*, 1998; Ross *et al.*, 1994 and Fiering *et al.*, 1990). An other feature of a probability or stochastic model, a normal distribution of expressing cells, is also noted in these studies.

If stochastic models are indeed true for general gene transcription, how are then the high and constant levels of gene expression established and/or maintained? The probability that a gene will express is based upon different probabilities, the intrinsic probability of the gene itself and the transcription probabilities of the gene's regulatory elements which are occupied by diverse transcription factors. These probabilities are multiplied during the process of transcriptional activation, leading to, if required, a high probability for gene expression. Constant expression levels of a gene are reached if the half-live of the mRNA or protein of a gene is longer then the time required for a second round of transcription. If the half-live is shorter then this period, the transcription of the gene will be observed in pulses (Hume, 2000).

Data on enhancer function too indicate a probability or stochastic nature of gene expression. Also on enhancer function the dogma is a rate model: the presence of an enhancer extends the rate of transcription and thereby increases levels of expression of a gene. However, this idea is the result of the early studies on enhancer function, which were done using bulk assays and not on single-cell level. This way no distinction could be made between a rate and a probability model (Fiering *et al.*, 2000). Single cell studies, using the β-gal gene with the SV-40 or β-globin HS2 enhancer, show now that it is not a higher level of gene expression per cell, but the number of cells expressing that is influenced by the enhancer (Walter *et al.*, 1995). Enhancers thus affect rather the on/off status of a cell then the rate of transcription of a gene. This indicates that enhancers act in a stochastic fashion to increase the probability that a gene will be transcribed (Fiering *et al.*, 2000).

If indeed a probability or stochastic model determines gene activation and expression it also indicates that any (inducible) gene will show a percentage of mono-allelic expressing cells, which has nothing to do with specific regulation of expression levels. Several studies describe and mention indeed mono-allelic expression or stochastic gene expression, however, these are all studies on genes involved in cellular responses, dosage compensation or lineage commitment (Goto and Monk, 1998; Nemazee, 2000; Chess *et al*., 1994; Held *et al*., 1999; Hollander *et al*., 1998; Bix and Locksley, 1998 and Riviere *et al*., 1998). Recently several studies have been published, which indeed show allelic expression patterns indicative of a stochastic nature for general gene activation and expression (Chapter 2; Elowitz *et al*., 2002 and Obzudak *et al*., 2002).

If all this leads to the acceptance of a probabilistic or stochastic nature for general gene expression, also the description of gene transcription has to be changed. "The production of mRNA occurs in *pulses*. The *mean frequency* of pulses is the major determinant of mRNA production and is determined by the *probability* of formation of a preinitiation complex" (Hume, 2000).

Erythropoiesis

Erythropoiesis is a multi-step process in which early progenitor cells differentiate to become erythrocytes (Fig. 4). This process starts when a multipotent haematopoietic stem cell (HSC) undergoes multilineage commitment, followed by proliferation and maturation into the erythroid committed progenitor cell (Ogawa, 1993). This maturation takes place via a series of intermediate precursor cells, like burst and colony forming units (BFU-e and CFU-e). Besides the commitment to erythroid progenitors, cells originating from HSC, can also form all of the other cell lineages, such as lymphocytes, mast cells, megakaryocytes, macrophages and neutrophils (Metcalf, 1998).

The first haematopoietic cell, which can be identified, arises from the blood islands of the yolk sac, the primitive erythrocytes. These erythrocytes, however, cannot reconstitute the multiple haematopoietic lineages as definitive HCS can (Medvinsky *et al.*, 1993 and Rich, 1995). HCS are identified within the mouse embryo in the aortic/gonad/mesonephros (AGM) region, and are the key elements responsible for the maintenance of blood cell formation throughout life (Medvinsky and Dzierzak, 1996 and Sanchez *et al.*, 1996). A second origin of haematopoiesis in the embryo has been identified in the AGM. Cells from this region have been shown to reconstitute the multiple haematopoietic lineages in an irradiated adult recipient, and are definitive haematopoietic stem cells (Medvinsky and Dzierzak, 1996; Sanchez *et al.*, 1996 and Mukouyama *et al.*, 1998). It appears that there are two sites from which the haematopoietic cells in the embryo originate. First from the yolk sac producing large numbers of primitive erythrocytes, followed by the production of definitive haematopoietic stem cells in the AGM. After the embryonic phase the site of erythropoiesis shifts to the foetal liver and finally to the bone marrow, the sites of definitive erythropoiesis (Moore and Metcalf, 1970).

Primitive and definitive erythrocytes also differ morphologically, with the primitive erythrocyte nucleated and the definitive erythrocyte enucleated. Primitive and definitive erythrocytes also differ in the complement of globin genes that are activated, resulting in the formation of embryonic-, foetal- and adult-stage haemoglobin in man (Stamatoyannopoulos and Grosveld, 2001).

Haemoglobin

Haemoglobin is synthesised in erythrocytes as a heterotetrameric protein and is responsible for the transport of oxygen. The heterotetramer consists of two α-like and two β-like globin chains, these polypeptides bind one haem group each. The α-like globin chains and β-like globin chains synthesised by each erythrocyte can differ depending on whether the cell is part of the primitive or definitive lineage and also depending on the species. This results in man in embryonic haemoglobin, type Gower I $\zeta_2\varepsilon_2$, type Gower II $\alpha_2\varepsilon_2$, type Portland I $\zeta_2\gamma_2$ type Portland II $\zeta_2\gamma_2$, foetal haemoglobin HbF, $\alpha_2\gamma_2$, and adult haemoglobin HbA and HbA_2, $\alpha_2\beta_2$ and $\alpha_2\delta_2$ (Bunn and Forget, 1986).

The three-dimensional protein structure of the globin protein family members has been resolved and consists of a four of α-helices, forming the haem-binding pocket, which is characteristic for all family members (Dickerson and Geis, 1983).

Although haemoglobin was first characterised in vertebrates it is highly conserved throughout evolution and present in invertebrates, plants, and also in several species of eubacteria, *Saccharomyces cerevisiae*, protist and protozoa (Hardison, 1996). This shows the importance of globin throughout evolution.

Haemoglobinopathies

Several hereditary blood disorders, the haemoglobinopathies, have been described as a result of mutations or deletions in the α- and/or β-globin gene loci. The analysis of the molecular basis for these disorders has been very important in the understanding of globin gene regulation.

Sickle cell disease

Sickle cell disease results from a single T→A base change causing the substitution of one amino acid from valine to glutaminic acid in the amino terminus of the β-globin chain, resulting in Sickle haemoglobin HbS. This change causes the β-globin chain to polymerise when it is deoxygenated and to form aggregates in the cell (for review see Stamatoyanopoulos and Grosveld, 2001). As a result of this, the shape of red blood cells changes from a round shape to that of a sickle. In acute cases, sickling causes vaso-occlusion and severe anaemia. Occurrence of sickle cell anaemia is correlated with areas where malaria is endemic. People heterozygous for HbS are relative resistant to malaria infection caused by *Plasmodium falciparum* (Allison, 1957). In these areas carriers have a selective advantage in survival explaining the high frequency within the population.

Thalassemias

Thalassemias are the most common single gene disorders in the world. The disease is characterised by abnormal globin gene expression, which results in the reduction of the α- or β-globin chains, giving rise to either α- or β-thalassemias. The reduced production of one of the globin chains results in the accumulation of the intact globin chains in the erythroid cells. The intact chains precipitate in the erythroid precursor cells and form inclusion bodies. The inclusion bodies cause membrane damage and premature destruction of the erythroid cells, both mature and precursor erythroid cells, resulting in ineffective erythropoiesis and anaemia (Orkin, 1986 and Thein, 1993).

Thalassemias are defined both clinically and genetically. Using the clinical definition, thalassemias are divided into major forms, which are severe and transfusion dependent and minor forms, which are asymptomatic and often resemble the carrier state or trait.

The genetic classification describes the globin chains affected and the amount of chain production. The most common thalassemias genetically defined are the α-thalassemias, with the absence (α^0-thalassemia) or reduction (α^+-thalassemias) of the α-globin chains. β-thalassemias are similarly classified as β^0-thalassemia and β^+-thalassemias. The δβ-thalassemias are characterised by the production of HbF in adult life and are genetically classified by the amount of γ chains that are produced. $(\delta\beta)^0$ and $({}^A\gamma\delta\beta)^0$ -thalassemias, are similar to HPFH condition which is discussed below. Most of the α-thalassemias are caused by deletions of the locus, while the majority of the β-thalassemias are caused by non-deletion mutations.

Although uncommon, deletion types of the β-thalassemias have played an important role in gaining insight on the expression mechanisms of the globin genes and the discovery of important regulatory elements (Collins and Weissman, 1984 and Stamatoyannopoulos and Nienhuis, 1996). Two deletion types, the Dutch and the Hispanic thalassemias have played an important role. The Dutch deletion has a size of 100 kb and removes the whole β-globin locus leaving just the β gene and its promoter intact (van der Ploeg *et al.*, 1980; Kiousis *et al.*, 1983 and Taramelli *et al.*, 1986). The Hispanic deletion removes 30 kb upstream of the β-globin locus leaving the genes and their proximal regulatory elements intact (Driscoll *et al.*, 1989). In both cases there is no globin expression, although the genes are intact. It is now clear that the region upstream of the genes, deleted in both the Dutch and the Hispanic thalassemias, contains an important regulatory element, identified as the locus control region, which is necessary for the normal expression of the globin genes (Grosveld *et al.*, 1987).

Non-deletion β-thalassemias are caused by a variety of mutations (Huisman and Carver, 1998). The majority of the non-deletion β-thalassemias are caused by either insertion or deletion of a nucleotide or a nonsense mutation, all leading to a frame shift and/or a preliminary stop codon. Furthermore, mutations which affect the start codon or the correct splicing of the β-globin RNA have also been described. Also the non-deletion β-thalassemias have played a role in the characterisation of regulatory elements. For example mutations in the CACCC box of the β-globin promoter result in down regulation of the β-globin gene expression (Kulozik *et al.*, 1991). This element is the binding site for the erythroid-specific transcription factor EKLF (Wijgerde *et al.*, 1996).

HPFH

Hereditary Persistence of Foetal Haemoglobin (HPFH) is not considered a disease but rather a condition, because individuals with HPFH are clinically normal. HPFH disorders are heterogeneous and show an elevated level of HbF in adult life. Molecular analysis of HPFH conditions has shown that in some cases the β-globin cluster is intact. These are called the non-deletion HPFH conditions. In some cases, the 3'end of the locus, encompassing the δ and the β gene is deleted, thus giving rise to deletion HPFH. This last group resembles δβ-thalassemias which are also associated with elevated HbF. Distinction between the two disorders is made on a number of features: δβ-thalassemia patients show between 5-15% HbF in adult life and hypochromic and microcytic red cells, whereas HPFH heterozygotes show between 15-30% HbF and normal red cells.

Several hypotheses have been put forward to explain the differences in the HbF levels observed between HPFH and δβ-thalassemias. They fall into three categories: deletion of regulatory sequences within the genecluster; powerful enhancers downstream of the locus brought closer to the γ genes due to the large deletions; and competition between the γ- and β-genes for the activation by the LCR.

The first hypothesis was based on the comparison of deletions causing either δβ-thalalassemia or HPFH. Since the report of these initial studies, many more deletions causing either βδ-thalassemia or HPFH have been described. These include deletions causing δβ-thalassemia which have a 5' breakpoint upstream of all previously reported HPFH 5' breakpoints, thus deleting the same or even more of the possible regulatory sequences within the region (Wood, 1993). These data make the first hypothesis less likely. Transgenic studies, in which several sequences of the region have been deleted (Peterson *et al.*, 1995; Zhang *et al.*, 1997 and Calzolari *et al.*, 1999), also suggest that there are no essential silencer regions elements outside the γ-genes. Still it could be that some intergenic sequences act as positive or negative regulators and that the deletion of these could determine the observed phenotypes.

The second hypothesis is based on a HPFH deletion in which the β 3'enhancer is brought 25 kb closer to the γ genes (Feingold and Forget, 1989) and also on the HPFH1 and HPFH2 deletions in which a region normally located 120 kb from the γ genes and containing enhancer properties is brought within 10 kb of the γ genes (Feingold and Forget, 1989). This latter element has been shown in transgenic mouse models to give rise to elevated HbF (Arcasoy *et al.*, 1997).

The last hypothesis is based on the observation that elevated HbF is observed only when both the δ and the β genes are deleted. Furthermore in non-deletion HPFH there is evidence for competition between the γ and the β genes in the adult. However, this competition for the LCR cannot explain the differences observed between δβ-thalassemias and HPFH in HbF levels, it can only explain why higher levels of γ expression are observed in the adult. Another argument against this hypothesis is that in transgenic mice containing a LCR coupled to the γ genes, no γ expression is observed in the adult transgenic mice (Dillon and Grosveld, 1991).

In conclusion neither of the three hypotheses can explain all of the conditions and the differences observed between HPFH and the δβ-thalassemias in increased HbF. These hypotheses are not mutually exclusive and it is therefore likely that a combination of the mechanisms and different contributions of each mechanism between the various deletions account for the observed HbF increases and differences between δβ-thalassemias and HPFH deletions.

Non-deletion HPFH result from point mutations in the promoters of the γ genes (mutations in the $^{A}\gamma$ -promoter give rise to $^{A}\gamma$ HPFH and mutations in the $^{G}\gamma$-promoter give rise to $^{G}\gamma$ HPFH). These HPFHs are characterised by the δ-and β-globin chain synthesis in *cis* with the HPFH determinant. Transgenic mouse studies in which the $^{A}\gamma$-promoter is mutated, at -117 causing Greek non-deletion HPFH, indeed showed elevated $^{A}\gamma$ expression in the adult mice, in contrast to control mice which carried the same locus but without the mutation (Peterson *et al.*, 1995). An other study, in which the same mutation was introduced in mice, also showed persistence of γ-globin expression in the adult and a concomitant decrease in β-globin expression (Berry *et al.*, 1992).

The HPFH phenotypes are of clinical importance since they modulate the effect of both the β-thalassemia and HbS mutations. Patients heterozygous for both HPFH and either β-thalassemia or HbS show a much milder condition (Fessas and Stamatoyannopoulos, 1964 and Stamatoyannopoulos *et al.*, 1975). It is thus of great interest to elucidate the mechanisms for γ gene silencing and how these are apparently reversed in HPFH phenotypes, in order to develop therapeutic alternatives for the treatment of haemoglobinopathies.

Structure of globin gene loci

It is thought that the α-and β-globin geneclusters originated from one ancestral globin gene and have been duplicated and diverged into two geneclusters during evolution. In the beginning these geneclusters were linked, as suggested by the fact that in primitive vertebrates, such as the zebrafish *Danio rerio* (Chan *et al.*, 1997) and the frog *Xenopus* (Hosbach *et al.*, 1983), the α-and β-globin genes are still physically linked. In other vertebrates the split of the α- and β-globin genes onto separate clusters is thought to have taken place about 300 million years ago, before mammals and birds diverged during evolution. This has been deduced from data showing that in both mammals and birds the globin genes are situated on separate chromosomes (Deisseroth *et al.*, 1976 and Hughes *et al,*. 1979).

In mammals, the α- and β-globin geneclusters are located on different chromosomes and are independently regulated. In humans, the β-globin genecluster resides on chromosome 11p (Deisseroth *et al.*, 1978) and the α-globin genecluster on chromosome 16p (Deisseroth *et al.*, 1977). In mouse the β-globin genecluster is located on chromosome 7 (Jahn *et al.*, 1980), and the α-globin genecluster on chromosome 11 (Fig. 5).

The mammalian α- and β-globin geneclusters differ in a number of respects: they are embedded in different chromatin environments, with the α-cluster residing in a constitutively 'open', CG-rich chromatin domain, containing mostly housekeeping genes (Craddoch *et al.*, 1995). The β-cluster is found in a 'closed', AT-rich chromatin domain and is flanked by olfactory receptor genes. The different chromatin environments are also reflected on the replication timing of the two loci, such that the α-cluster is replicated early in S-phase in most cell types (Epner *et al.*, 1981 and Furst *et al.*, 1981), whereas the β-cluster is late replicating except in eryhroid cells (Kitsberg *et al.*, 1993 and Aladjem *et al.*, 1998). Furthermore the α-cluster does show differential methylation (Bird *et al.*, 1987), whereas the β-cluster shows methylation in non-erythroid cells (van der Ploeg and Flavell, 1980). Other differences between the clusters are found in specific DNA sequences, for example, the α-cluster contains Alu-repeats, whereas the β-cluster contains LINE-elements and Alu-repeats; the α-contains CpG-islands (Pondel *et al.*, 1995), which are not found in the β-genecluster (Collins and Weisman, 1984). Finally the α-cluster does not have MARs, in contrast to the β-cluster (Fischel-Ghodsian *et al.*,1987 and Jarman and Higgs, 1988).

The human β-globin cluster is composed of five developmentally regulated genes, arranged in the order in which they are expressed during development: 5'ε-$^{A}\gamma$-$^{G}\gamma$-δ-β 3'. The locus is regulated by an element located upstream of the genes called the locus control region (LCR) (Grosveld *et al.*, 1987). This element comprises five tissue-specific DNase I hypersensitive sites. The LCR will be discussed in more detail below.

The ε gene is expressed in the blood islands of the yolk sac and is detectable between the third and the tenth week of gestation in man. At about five weeks of gestation the site of haematopoiesis changes to the foetal liver. At the same time the switch from ε to γ expression starts to take place and is complete by around ten weeks of gestation (Huehns *et al.*, 1964). Until the twentieth week of gestation the foetal liver remains the main site of erythropoiesis and it gradually switches to the spleen and bone marrow, until at around birth the bone marrow becomes the main site of erythropoiesis (Bloom and Bartelmez, 1940; Knoll and Pingel, 1949 and Wintrobe and Shumacker, 1935). During the same period γ expression decreases and around birth switches to expression of the adult β and δ genes, with δ-globin making only a minor (3%) contribution to adult globin chain synthesis (Fig. 6).

The mouse β-globin locus contains two embryonic globins, the εy and βH1 genes, which are expressed in the embryonic yolk sac. The two adult globin genes, β_{min} and β_{maj}, are expressed during the foetal liver and adult bone marrow stages with the β_{maj} being the dominant gene. In mice, therefore, there are no distinct foetal genes and the switch from embryonic to adult globin gene expression takes place around day 11.5 of gestation (Farace *et al.*, 1984 and Chada *et al.*, 1986). The mouse β-globin locus is also regulated by a LCR, which contains six DNase I hypersensitive sites (Moon and Ley, 1990).

The major α-globin locus regulatory element comprises only one DNase I hypersensitive site, located 40 kb, in the case of the human locus, and 26 kb, in the case of the mouse locus, upstream of the ζ gene (Higgs *et al.*, 1990). The adult α gene is expressed during all developmental stages and is only in the embryonic stage accompanied by expression of the embryonic-specific ζ-globin gene (Rohrbaugh and Hardison, 1983; Leder *et al.*, 1985). The ζ-gene follows the expression of the ε gene. Whereas the α gene is expressed throughout development, it is from week six of gestation that the adult expression levels are reached. Both human and mouse α-globin geneclusters express the same genes and have very similar expression profiles.

Pseudo genes with structural homology to the globin genes are also present in the globin gene loci and appear to be the result of gene duplication events. The duplication is followed by mutation and inactivation of the duplicated gene, which is followed by subsequent mutations due to loss of selective pressure. In the β-globin cluster there is one pseudo gene, ψβ while three pseudo genes are present in the α-globin genecluster: ψζ and $\psi\alpha_1$ $\psi\alpha_2$ (Forget, 2001) .

Structure and regulation of the globin genes

Human globin genes are relatively small and contain three exons and two introns. The exons code for 141 and 146 amino acid peptides for the α- and the β-globin chains, respectively. Intron 2 (IVS-2) appears to be important for polyadenylation and for the release of the transcript from the template, such that the transport from the nucleus to the cytoplasm can take place (Collis *et al.*, 1990 and Antoniou *et al.*, 1998).

Each gene is flanked by promoters, enhancers and silencers, important for the correct tissue-specific and developmental stage-specific expression. These elements are thought to interact with the LCR, in the case of the β-globin genes and with the αMRE, in the case of the α-globin genes, in order to activate transcription. The fact that the hypersensitive sites of the LCR and αMRE as well as the proximal regulatory elements of the individual genes contain some of the same transcription factor binding sequences, suggest an interaction between the different regulatory elements. The globin gene promoters, while sharing several characteristics, also have unique, distinguishing features. The α- and β-globin gene promoters contain a TATA box and a CCAAT box (Efstratiadis *et al.*, 1980), but differ otherwise (Fig. 7A). Because the main part of this thesis will concentrate on the human β-globin genecluster I will only briefly discuss the promoters and regulation of the individual α-globin genes and concentrate more in detail on the human β-globin genes.

All β-globin gene promoters show with minor variations the same recognition sequences for transcription factors. The sequences identified are ATAA, CCAAT and CACCC located respectively at -30, -70-80 and -80-140 nucleotides from the cap-site (Meyers *et al.*, 1986). A fourth sequence which is also found in all promoters is the (A/T) GATA (A/G) motif (Martin *et al.*, 1989 and Tsai *et al.*, 1989).

Looking at the genes individually, differences in both regulatory sequences and regulation of the genes can be observed. The individual genes and their regulatory elements are discussed below.

The ε-globin gene

The region containing the ε gene and its regulatory sequences spans approximately 3.7 kb of DNA, with 2 kb of upstream sequences containing regulatory elements. In this region the TATA, CACCC and CCAAT boxes and GATA1 binding sites have been identified. These sites play a role both in the activation and the developmental stage-specific silencing of the ε gene, as has been shown in a number of transfection experiments using cultured cells (Gong and Dean, 1993; Gong *et al.*, 1991 and Walters and Martin, 1992).

Transcription factors Sp1 (Yu *et al.*, 1991), FKLF1 (Asano and Stamatoyannopoulos, 1999) and FKLF2 (Asano and Stamatoyannopoulos, 2000) have been shown to bind to the CACCC box in the ε-promoter, although Sp1 binding does not appear to have a significant influence on ε-globin expression

(Yu *et al.*, 1991). Of the two FKLFs, FKLF1 is the predominant factor activating ε expression, whereas FKLF 2 activates ε expression to a lesser degree in stable and transient transfection assays. However, the *in vivo* role for both FKLF1 and 2 remains to be determined (Asano and Stamatoyannopoulos, 1999 and Asano and Stamatoyannopoulos, 2000).

Using transgenic mice, several sequences have been indicated to play a role in the developmental stage-specific silencing of the ε gene. A silencer element from –304 to –179 5' of the ε gene was identified containing binding sites for GATA and YY1 factors (Cao *et al.*, 1989). Binding of GATA1 at –208 together with binding of YY1 at position –269 appear to be involved in ε repression (Raich *et al.*, 1995). Deletion of this silencer element in the context of a YAC containing the complete human β-globin locus, did not lead to high levels of ε expression in the foetal stage as would be expected, but instead to a decline of ε expression in the yolk sac (Liu *et al.*, 1997). Together with the data from Raich *et al.*, this suggests that the 5'silencer of ε has a dual role both in suppressing ε in the foetal stage and in maintaining ε expression in the embryonic stage. Furthermore, is has been shown that the binding of NF-E3, which is immunologically related to COUP-TF, to the DR1 element located near the CCAAT box, results in the repression of ε expression (Filipe *et al.*, 1999). The replacement of the DR1 element by a high-affinity CACC-binding site for EKLF and a 4 bp substitution in the ε-globin CAAT sequence, also disrupting a DR element, led to the identification of a protein complex which mediates the suppression of ε-globin transcription during definitive erythropoeisis, called DRED (direct repeat erythroid-definitive) (Tanimoto *et al.*, 2000). Recently, two core components of DRED, nuclear orphan receptors TR2 and TR4 have been shown to bind to the DR1 element present in the ε-globin promoter and to repress ε-globin expression in definitive erythroid cell (Tanabe *et al.*, 2002).

The ε gene also needs the LCR for expression since an ε-globin gene alone, or a human β-globin locus with deletions in the LCR, result in undetectable ε expression in transgenic mice (Raich *et al.*, 1990; Shih *et al.*, 1990; Navas *et al.*, 1998 and Chapter 5 of this thesis). Additionally, ε-globin is said to be autonomously regulated during development, in that it does not require the presence of other globin genes in order to be silenced in the foetal and adult stages in transgenic mice (Shih *et al.*, 1990).

The γ-globin genes

The regulation of the γ-globin genes has been studied intensively because even slightly elevated HbF in the adult, as seen in HPFH conditions, can alleviate the effects of β-thalassemia and sickle cell disease. Some mutations resulting in HPFH map within transcription factor binding sites in γ regulatory sequences, giving rise to the creation of new protein binding sites or the destruction of existing ones.

The promoters of the two γ genes are identical and contain the conserved CCAAT and CACCC boxes, a TATA box, DRE 1 sites and an OCT1 binding site (ATTTGCAT) flanked by two GATA1 binding sites. Between the CCAAT and TATA boxes a so-called stage selector element (SSE) has been reported (Jane *et al.*, 1992 and Jane *et al.*, 1993).

Different transcription factors bind to these sequences and play a role in the regulation of the γ genes. Transcription factors suggested to act as activators for γ expression are SSP, binding to the SSE (Jane *et al.*, 1992), CP1 binding to the CCAAT box (Skalnik *et al.*, 1991) and FKLF2 binding to the CACCC box (Asano *et al.*, 2000). The binding of SSP to SSE has been proposed to provide the γ gene with a competitive advantage over the β gene in the foetal stage. Transgenic studies in which the SSE has been mutated show a down-regulation of γ expression only when the gene is in competition with the β gene (Jane *et al.*, 1992). CP1 is ubiquitously expressed and interacts with both CCAAT boxes, however, there is no *in vivo* evidence to support CP1's role as a positive regulator of γ gene expression. Mutations in the CACCC box of the γ-promoter have been shown in transgenic mice to result in severe decrease of expression, indicating that the binding of transcription factors to this site plays an important role in γ-globin gene regulation (Stamatoyannopoulos *et al.*, 1993 and Duan *et al.*, 2001). Of these factors, Sp1, Sp3 do not appear to play a role *in vivo* (Marin *et al.*, 1997) and for BKLF/TEF2 there is no evidence at all for an effect on γ-expression (Crossley *et al.*, 1996). FKLF1 and FKLF2 bind to the γ-CACCC box *in vitro* and have been implicated as transcriptional activators, however, their role *in vivo* remains to be determined (Asano *et al.*, 1999 and Asano *et al.*, 2000).

Other transcription factors binding to the γ-promoter, such as CDP1 (Skalnik *et al.*, 1991), NF-E3, DRED and GATA1 (Gumucio *et al.*, 1988; Mantovani *et al.*, 1988 and Berry *et al.*, 1992), have been suggested to act as transcriptional repressors, although no *in vivo* experimental evidence, e.g. from gene knockouts, exists to support these assertions. CDP, at least *in vitro*, acts as a transcriptional repressor and competes with CP1 for binding at the CAAT-boxes (Skalnik *et al.*, 1991). A G →A mutation in the distal CAAT-box showed decreased binding of NF-E3 and GATA1 and led to the association of these factors with γ-globin gene repression. However, other mutations which decrease NF-E3 and GATA1 binding to the CAAT-box did not result in an increase of γ expression in transgenic mice (Ronchi *et al.*, 1996). Thus, the exact role of these factors in the γ-globin gene repression is still under debate. Mutations in the DR 1 binding site of DRED resulted in an elevation of γ expression in the adult, at least *in vitro*, indicative of a suppressor role of DRED for γ expression in the adult (Tanabe *et al.*, 2002).

Other binding sites for transcription factors have been found in the region upstream of the γ-promoter, having been associated with mutations that result in HPFH. Two of these regions which have been associated with elevated γ expression, are the –175 and –200 regions (Surrey *et al.*, 1988; Stoming *et al.*, 1989 and Jane *et al.*, 1995).

Regulatory sequences besides the promoter include an enhancer element reported at 750 bp downstream of the γ genes (Bodine and Ley, 1987) and sequences –382 to –370 5' to the ^{A}γ-promoter, which have been shown to contain an adult specific silencer in transgenic studies using constructs containing the γ region with different truncations of the 5'region of the ^{A}γ-promoter coupled to a μLCR (Stamatoyannopoulos *et al.*, 1993). The 3' γ gene enhancer however does not significantly affect γ gene transcription when deleted in transgenic mice carrying a YAC construct containing the β-globin locus (Puruker *et al.*, 1990 and Liu *et al.*, 1998). Instead this element has been proposed to play a role in the protection against chromosomal position effects and in the stabilisation of the interaction of the LCR with the γ-promoters (Li and Stamatoyannopoulos, 1994 and Stamatoynnopoulos *et al.*, 1997). Like the ε gene, the γ gene was also shown to be autonomously regulated. When a γ gene is linked to a LCR, it is silenced or expressed at very low levels in adult transgenic mice, indicative of an autonomous control of the γ gene (Dillon and Grosveld, 1991).

The δ-globin gene

The next genes to be developmentally activated are the adult δ- and β-globin genes. The δ gene is very similar to the β gene in its 5' region but is distinct in its 3' region (Spritz *et al.*, 1980 and Martin *et al.*, 1983). The similarity of the 5' region between the β and the δ genes would suggest a similar mode of regulation. This, however, is not the case and is due to the deletion of the CACCC box and mutations found in the CCAAT box. The former is the major reason for the low δ expression levels, because this deletion results in the loss of an EKLF binding site (Donze *et al.*, 1996 and Tang *et al.*, 1997). Addition of the β-globin CACCC box to the δ gene, results in a 10-fold upregulation of δ expression in transfection assays (Donze *et al.*, 1996).

The β-globin gene

β-globin is the dominantly expressed gene in the adult stage. CCAAT and CACCC boxes have also been identified in the β-promoter. In contrast to the γ genes, however, there are two CACC boxes and one CCAAT box in the β-promoter. Both boxes bind proteins involved in the activation of the β gene. CP1, GATA1 and NF-E6 bind to the CCAAT box (Antoniou and Grosveld, 1990; Antoniou *et al.*, 1988; de Boer *et al.*, 1988; Berry *et al.*, 1992 and Wall *et al.*, 1996). Of these, CP1 is thought to be a positive regulator, at least *in vitro*. GATA1 binds weakly and may not be of functional importance (Li *et al.*, 1998). NF-E6 seems to have a role *in vivo* since overexpression in transgenic mice of a dominant negative NF-E6 mutant leads to a shift in the ratio of γ to β expression, resulting in a higher expression of γ (Zafarana *et al.*, 2000).

The CACCC boxes in the β-promoter bind several factors *in vitro* (Hartzog and Myers, 1993), however, *in vivo* studies have shown that EKLF is the functional protein binding at this site (Miller and Bieker, 1993 and Feng *et al.*, 1994). EKLF has been shown to be the major regulator of β gene expression. In transgenic mice heterozygous for the EKLF gene knockout, lower β-globin levels were observed (Nuez *et al.*, 1995). The complete gene knockout of EKLF proved to be lethal due to severe anaemia immediately after the γ- to β-globin expression switch (Perkins *et al.*, 1995 and Nuez *et al.*, 1995).

There are a number of reasons for the profound effect of EKLF on β gene expression but not on ε and γ gene expression. The β-globin CACC box has a much higher affinity for binding EKLF than those of ε and γ. In addition, the CACC box in the γ-promoter is flanked by a CCTTG repeat which has been shown to be a repressor for the recruitment of EKLF (Donze *et al.*, 1995 and Lee *et al.*, 2000).

Besides the promoter, two other regulatory elements have been described. An enhancer element located downstream of the poly-A signal containing four GATA1 sites has been shown to stimulate the activity of a linked promoter in transfection studies (Antoniou *et al.*, 1988). Furthermore, the deletion of this enhancer resulted in a decrease of β expression in transgenic mice (Liu *et al.*, 1997). Another enhancer element described in cell transfection and transgenic mouse studies is located near the junction of intron 2 and exon 3 (Antoniou *et al.*, 1988 and Behringer *et al.*, 1987). Its *in vivo* role in the context of the whole locus has not been determined yet.

In contrast to the other genes silencer elements have not been identified in the neighbourhood of the β gene.

The ζ-globin gene

The ζ-globin promoter contains a TATA-box, a CAAT-box, a CACC-box and DR-repeats. Transcription factors involved in the regulation of the ζ gene include CP2, which binds to the CAAT-box (Lim *et al.*, 1992), Sp1-like proteins, which bind in the CACC-box region (Watt *et al.*, 1990 and Yu *et al.*, 1990) and GATA1, which has a strong binding site overlapping the Sp1-binding site, the two proteins binding in a competitive manner. There is also a GATA1 site in the upstream part of the promoter. Both GATA1 sites are necessary for interactions with αMRE, as has been shown in transient transfection studies (Zhang *et al.*, 1993).

Besides the promoter, another positive regulatory element containing a GATA1 site flanked at the 3' by a CACC-box, has also been described (Sabath *et al.*, 1995).

Like the ε gene the developmental regulation of the ζ gene is autonomous: all elements required for the silencing of the ζ gene can be found in the sequences flanking it (Sabath *et al.*, 1993; Albitar *et al.*, 1991 and Pondel *et al.*, 1992).

The α-globin gene

The promoter of the α-globin gene differs from those of all other globin genes. There is no CACC-box, but there is a GC-rich area which forms part of a methylation-free island extending into the gene (Flint *et al.*, 1997 and Shewchunk and Hardison, 1997). It was shown in transfection studies that the α gene could be expressed in non-erythroid cells without additional enhancer elements, probably due to the presence of the methylation-free island (Humphries *et al.*, 1982). For the expression of α-globin in transgenic mice, the α gene requires the presence of the αMRE or the β-globin LCR (Hanscombe *et al.*, 1991 and Higgs *et al.*, 1990).

Transcription factors that bind to the α-promoter overlap those that bind to the β-globin-like promoters, like GATA1 and CP1, but there are also proteins that only bind to the α-promoter, like the inverted repeat protein (Lim *et al.*, 1992; Kim *et al.*, 1990; Lim *et al.*, 1993 and Swendeman *et al.*, 1994).

The Locus Control Region

The existence of important regulatory sequences upstream of the β-globin genes became clear from the molecular analysis of large deletions in the β-globin locus that give rise to thalassemias (van der Ploeg *et al.*, 1980; Driscoll *et al.*, 1989 and Kulozik *et al.*, 1991). In particular, the Dutch and Hispanic thalassemias are characterised by the deletion of 100 kb and 40 kb of sequence, respectively, upstream of the gene locus. These deletions left part, or all, of the β-globin gene locus intact but the genes were transcriptionally inactive. Although the intact genes are still able to express, as was shown by cloning and expressing them in transfection studies (Kioussis *et al.*, 1983; Ryan *et al.*, 1989). In the deleted locus the genes are silent and embedded in an inactive chromatin structure (Kioussis *et al.*, 1983; Forrester *et al.*, 1990 and Schubeler *et al.*, 2000).

DNase I hypersensitivity studies in the region usptream of the β-globin gene locus revealed the presence of five tissue-specific hypersensitive sites located between 6 kb and 25 kb 5' of the ε gene (Forrester *et al.*, 1986, Tuan *et al.*, 1985). These sites have been termed 5'hypersensitive site 1-5 (5'HS1- 5'HS5). The importance of these hypersensitive sites for globin expression was demonstrated in transgenic mouse studies, where this region was coupled to a β-globin gene. Expression levels proportional to transgene copy number and comparable (per copy) to endogenous mouse globin levels were observed (Grosveld *et al.*, 1987).

Transgenic mouse studies in which the LCR was coupled to a β-globin gene showed that the LCR can drive tissue-specific expression of the transgene independently of its (random) site of chromosomal integration, thus conferring copy number dependent levels of expression (Grosveld *et al.*, 1987 and Blom van Assendelft *et al.*, 1989). These properties form the defining characteristics of LCRs and suggest that one of the fundamental aspects of LCR function is the organisation of a chromatin domain that will support transcriptional activation (Festenstein and Kioussis, 1997 and 2000; Fraser and Grosveld, 1998; Grosveld, 1999).

Although the LCR was suggested by both transgenic mouse assays and the analysis of deletions in thalassemias as having chromatin activating capacities, recent studies in mice in which the LCR was deleted from the endogenous mouse β-globin locus have led to a discussion about this function of the LCR (Reik *et al.*, 1998; Epner *et al.*, 1998 and Bender *et al.*, 2000). In mice carrying a deletion of the endogenous mouse β-globin LCR, it was shown that the locus maintained DNase I hypersensitivity, however gene expression levels had dropped to just a small fraction of the wild type expression levels (Epner *et al.*, 1998; Reik *et al.*, 1998 and Bender *et al.*, 2000). From this, it was suggested that the LCR is not involved in the chromatin opening of the gene domain. One possible reason for these observations could be that there are differences between the mouse and the human LCRs (Higgs, 1998 and Grosveld, 1999). However, a satisfying conclusion reconciling the differences between the human genetic data on human β-globin LCR function and the LCR deletions in mice, has not yet been suggested.

The hypersensitive sites

The five hypersensitive sites of the LCR can be sub-divided into the erythroid-specific, HS1-4 (Forrester *et al.*, 1986 and Tuan *et al.*, 1985), and the constitutive HS5 (Tuan *et al.*, 1985 and Dhar *et al.*, 1990) although additional studies indicate that HS5 is present in most haematopoietic cells (Li *et al.*, 1999 and Zafarana *et al.*, 1995). Construction of a micro-LCR (μLCR) containing small regions holding each hypersensitive site showed that these retain the functional activity of the LCR (Talbot *et al.*, 1989). Mapping of each individual hypersensitive site revealed core sequences of around 250-500 bp (Philipsen *et al.*, 1990; Talbot and Grosveld, 1991; Pruzina *et al.*, 1991 and Lowrey *et al.*, 1992).

All the hypersensitive sites contain binding sites for the erythroid specific factors NF-E2 and GATA1, as well as GT-sequences to which factors like EKLF and Sp1 can bind (Talbot *et al.*, 1990; Strauss and Orkin, 1992 and Ikuta *et al.*, 1996). Hypersensitive sites 2, 3 and 4, contain these binding sites, however, in different combinations (Fig. 7B) (Talbot *et al.*, 1990 and 1991; Pruzina *et al.*, 1991; Stamatoyannopoulos *et al.*, 1995; Walters *et al.*, 1991; Strauss and Orkin, 1992 and Ikuta and Kan, 1991).

HS2 contains two NF-E2, two GATA1, one Sp1 and two Tal1/USF binding sites (Ney *et al.*, 1990; Talbot and Grosveld, 1991 and Lui *et al.*, 1992). Mutagenesis of the GATA1 sites in a synthetic HS2 fragment shows reduced activity of the HS2 when coupled to a β-globin gene and transfected into MEL cells (Ellis *et al.*, 1993). The significance of the NF-E2 sites was shown both in transient transfection assays using MEL cells and in transgenic mouse studies demonstrating that they are necessary for full LCR and HS2 activity (Caterina *et al.*, 1991; Moi and Kan, 1990 and Liu *et al.*, 1992).

HS3 contains one NF-E2 site, a triple tandem repeat of GATA1 sites and GT-sequences (Philipsen *et al.*, 1990 and Strauss and Orkin, 1992). In transgenic mice it has been demonstrated that the GATA1 sites are required for LCR activity (Philipsen *et al.*, 1993). The GT sequence in HS3 is bound by EKLF *in vivo* (Gillemans *et al.*, 1998). The lack of binding of EKLF at the GT-sequence of HS3 results in changes in chromatin structure, as detected by DNase I sensitivity studies (Gillemans *et al.*, 1998 and Wijgerde et al., 1996).

Finally, HS4 has an AP1/NF-E2 site followed by a Sp1 site and two GATA1 binding sites (Pruzina *et al.*, 1991 and Lowrey *et al.*, 1992). The GATA1 sites in HS4 are inverted and are required for hypersensitive site formation (Lowrey *et al.*, 1992 and Stamatoyannopoulos *et al.*, 1995). Other factors that have been reported to interact with the LCR include USF and YY1. The functional relevance of their binding is not clear.

The role of the hypersensitive sites in the LCR

The role of each hypersensitive site within the LCR has been investigated in transgenic mouse studies and cell transfection assays. Transgenic mice in which a single hypersensitive site was coupled to a β-globin gene, showed different levels of expression. HS3 showed approximately 70% of full LCR activity, HS2 and HS4 30% and HS1 less then 10%, while HS1 did not show any activity (Fraser *et al.*, 1990). Similar results were obtained using stable transfections in MEL cells (Collis *et al.*, 1990).

In transient transfection assays, only HS2 showed classical enhancer activity (Ney *et al.*, 1990 and Tuan *et al.*, 1989). This enhancer activity is dependent on the tandem repeat of NF-E2 sites, which are only found in HS2 (Ney *et al.*, 1990 and Pruzina *et al.*, 1991). The role of HS3 became apparent in single copy transgenic studies, since only HS3 was able to drive β-globin gene expression. The other hypersensitive sites tested needed to integrate as multiple copies to activate globin expression (Ellis *et al.*, 1996 and Ellis *et al.*, 1993). From these studies it was suggested that HS3 contains chromatin opening activities.

The role of the hypersensitive sites has also been studied in transgenic mice carrying the complete human globin locus bearing a deletion of each one of the hypersensitive sites. The results of these studies showed a loss of chromatin opening activity by the LCR, resulting in position effects and lower expression levels of the transgenes (Milot *et al.*, 1996; Peterson *et al.*, 1996; Bungert *et al.*, 1995; Bungert *et al.*, 1999 and Chapter 5 of this thesis). Integration of the transgene in pericentromeric regions resulted in two types of position effects. Some lines showed a classical position effect, PEV, in which a sub-population of the cells does not express the transgene. Other lines showed a new kind of position effect, in which the transgene is expressed in all cells but for a shorter period of time during the cell cycle. This type of position effect was called cell timing position effect. A detailed description and discussion of the various HS-deletion studies in transgenic mice, is given in Chapter 5.

The observations on the hypersensitive sites led to the suggestion that the LCR functions as one unit called a holocomplex, which interacts with the β-globin genes in a developmental order and with only one gene at the time and that there is a developmental stage specificity in the interaction of the hypersensitive sites with the globin genes (Fraser *et al.*, 1993 and Dillon *et al.*, 1996).

αMRE, the locus control element of the α-globin locus

In contrast to the β-globin locus, the α-globin genecluster does not have an equivalent of a locus control region but one hypersensitive site located 40 kb in the human locus and 26 kb in the mouse locus upstream of the ζ gene (Higgs *et al.*, 1990). The critical region in αMRE has been localised in a 350 bp core element and contains binding sites for NF-E2, GATA1 and Sp1 (Jarman *et al.*, 1991). Deletion of this site in a MEL cell line containing a human chromosome 16, resulted in the down-regulation of the α genes (Bernet *et al.*, 1995). When the element was coupled to the α-globin genes in transgenic mice correct tissue specific and developmentally regulated expression was observed, although the α gene was silenced in the adult (Higgs *et al.*, 1990). In the absence of αMRE, α-globin transgenes do not normally express in transgenic mice. These results indicate, that this HS site is very important for the expression of the α-globin genes.

Haemoglobin switching

The developmental expression patterns of globin genes is characterised by the switching of expression of one globin gene to another. Developmental switching occurs twice in the human β-globin locus, from expression of ε-globin to γ and from γ to β. In the mouse there is only developmental switch in the expression of the β-globin genes from the embryonic εy/βH1 to the adult β_{min} and β_{maj} genes.

Transgenic mouse studies have been used to elucidate the mechanism of switching. Integration of the human γ or β genes without the LCR in transgenic mice, resulted in very low levels of globin expression, but expression was tissue- and developmental-stage specific (Kollias *et al.*, 1986; Trudel *et al.*, 1987; Trudel and Costantini, 1987 and Townes *et al.*, 1985). These observations indicate that the elements responsible for developmental-stage specificity lie within the regions flanking the genes.

Transgenic studies with constructs containing the whole β-globin locus, based on both ligated cosmid and YAC constructs (Stouboulis *et al.*, 1992; Gaensler *et al.*, 1993 and Peterson *et al.*, 1993), show levels of expression of the human globin genes similar to those of the endogenous mouse globin genes, as well as correct developmental switching, indicating that transgenic mice can be used to study the basis of human β-globin switching in the context of the full human β-globin locus (Fig. 8).

There is, however, one difference between the switching in humans and the switching of human β-globin genes in transgenic mice that has to be taken into account. The switch from γ to β takes place around birth in humans, whereas in transgenic mice this is accelerated so that γ to β switching takes place around day 12.5/14.5dpc in the foetal liver.

Switching is not a progressive process in which γ expression is followed by β expression. It is more of a dynamic process in which the γ- and β-globin genes are alternately transcribed during development. This was shown in transgenic mouse studies in which the transcription sites of the γ- and β-globin genes were visualised *in vivo* using primary transcript *in situ* hybridisation techniques. These studies showed that some cells had both γ and β transcripts on the same allele. There were also cells which showed β mRNA in the cytoplasm and a γ signal in the nucleus (Wijgerde *et al.*, 1995 and Gribnau *et al.*, 1998). These results indicate that the LCR flip-flops between the γ and β genes during a period of overlap in the expression of these genes and that the switching from γ to β takes place gradually.

Studies on the basis of haemoglobin switching led to the proposal of a dual mechanism by which switching is regulated: autonomous gene silencing and gene competition.

Autonomous globin gene silencing

Regulation of different human globin genes has been studied in several transgenic mouse models. Transgenic mice containing just the ε gene with 5' and 3' flanking sequences do not express at any developmental stage. Addition of the LCR to the ε gene, however, resulted in detectable expression only in the embryonic stage (Raich *et al.*, 1990 and Shih *et al.*, 1990). This observation suggests that the ε gene is autonomously silenced during development and that it does not require the presence of additional globin genes as would have been predicted in a competitive model. Deletion in transgenic mice of a putative silencer element located upstream of the ε gene, resulted in the continued expression of the ε gene into definitive erythroid cells albeit at low levels (Raich *et al.*, 1992 and Cao *et al.*, 1989).

The γ gene also appears to be regulated by autonomous silencing. Transgenic mice with a γ gene coupled to the LCR show expression of the gene in the foetal stage but no expression in the adult stage (Dillon and Grosveld, 1991; Enver *et al.*, 1989; Enver *et al.*, 1990 and Behringer *et al.*, 1990). Mutations in the promoter sequences of the γ genes, which give rise to a HPFH phenotype, suggest that the promoter of the γ genes play a role in the process of autonomous silencing (Berry *et al.*, 1992 and Ronchi *et al.*, 1996).

Studies with the β gene coupled to the LCR, show immediate activation of the transgene at the embryonic stage which persists all the way into the adult stage (Enver *et al.*, 1990 and Behringer *et al.*, 1990). This suggests that the β-globin gene when linked by itself to the LCR is not appropriately regulated during development in transgenic mice.

Gene competition

The result that the β-globin gene on its own is not developmentally regulated, led to the idea that the correct β expression is regulated through gene competition. Support for the competition model in human β-globin gene regulation came from experiments in which transgenic mice with the LCR coupled to the β gene (Enver *et al.*, 1990) and transgenic mice with a γ gene preceding the β gene were compared (Hanscombe *et al.*, 1991). These studies showed that the presence of the γ gene restored the normal developmental expression patterns of the β-globin gene thus restricting its expression in the foetal liver and adult blood stages.

The functional role of this competition between the genes is thought to be related to keeping balanced β-like globin chain production. That competition is indeed important for balanced haemoglobin production is shown by studies with non-deletion HPFH subjects. In these subjects, γ expression levels are increased with β expression levels decreased to an extent equivalent to the increase in γ expression, indicative of regulation by competition (Giglioni *et al.*, 1984).

Transgenic mouse studies have indicated that several factors play a role in gene competition: gene order (Hanscombe *et al.*, 1991 and Tanimoto *et al.*, 1999), distance between a gene and the LCR (Dillon *et al.*, 1997) and dosage of transcription factors present (Wijgerde *et al.*, 1996).

In transgenic mouse studies where the positions of the β- and the γ-globin genes were changed with respect to the LCR, correct timing of β gene expression depended on a place of the β gene further away from the LCR than the competing γ genes. Placing the γ genes away from the LCR resulted in premature silencing due to competition from the more LCR-proximal β gene (Hanscombe *et al.*, 1991). These results suggest that the difference in relative distance from the LCR plays an important role in gene competition for the LCR. The importance of gene order has also been shown in transgenic mice in which either the gene order or the LCR had been reversed (Tanimoto *et al.*, 1999). Reversal of the order of genes rendered ε-globin the 3'-most and β the 5'-most genes with respect to the LCR. The expression profile of the human globin genes in these transgenic mice was completely changed, with the ε gene no longer expressing, whereas expression of the β gene was found throughout development. One conclusion from these studies was that it is necessary for the embryonic genes to be proximal to the LCR for their transcriptional activation.

The importance of proximity to the LCR was demonstrated in a study in which an additional "marked" β-globin gene was inserted at two different positions in the β-globin locus. The first position is in place of the ε-globin gene proximal to the LCR and the second position is just upstream of the δ gene, i.e. distally to the LCR. The effects of placing the β^{marked} gene proximally or distally to the LCR, on the expression of the β-globin gene in its native position and on the developmental regulation of all globin genes in the locus were assessed in transgenic mice (Dillon *et al.*, 1997). The expression levels of the β^{marked} at the position just in front of δ compared to the β gene, showed that β^{marked} is expressed at 75% and β at 25% of the total expression level of human β compared to mouse β. Both transcriptional interference or competition for the LCR could be causing this result. If the former were the case, placing the β^{marked} at the position of the ε gene should decrease the effect of β^{marked} on the β gene. The analysis of the expression levels of β^{marked} at the ε position versus β, showed that β gene expression is severely reduced to approximately 10%, whereas the β^{marked} was expressed at 90% of the total level of human β expression. In addition, expression of the γ-globin genes at the embryonic stage was similar to the levels of β^{marked} expression. In the early foetal liver stage, however, there is only expression of β^{marked} with no detectable expression for γ-globin. This indicated that normally β is repressed during embryonic and foetal stages owing to its distal position to the LCR. This indicates that indeed the distance from the LCR plays an important role in the correct developmental expression of the human globin genes (Dillon *et al.*, 1997).

Finally the role of trans-acting factors in gene competition has been shown in transgenic studies using compound EKLF knockout/human β-locus transgenic mice (Wijgerde *et al.*, 1996). Mice heterozygous for the knockout allele of EKLF (hum$\beta^{+/+}$/EKLF$^{+/-}$) show a decrease in β transcription and a reciprocal increase of γ transcription. These data indicate that EKLF plays a role in the γ versus β competition and suggest that EKLF is potentially important in stabilising the interaction of the LCR with the β-promoter, thus giving it a competitive advantage (Wijgerde *et al.*, 1996).

In conclusion, autonomous gene silencing and gene competition account for the developmental regulation of globin gene switching, with silencing appearing to be more important for the embryonic to foetal switch and gene competition for the foetal to adult switch.

Models of gene regulation by the LCR

Three models haven been proposed to explain the basis of activation by the LCR: the accessibility model (Martin *et al.*, 1996), the scanning/tracking model (Tuan *et al.*, 1992 and Kong *et al.*, 1997) and the looping model (Fig. 9) (Stamatoyannopoulos *et al.*, 1991; Epner *et al.*, 1992; Dillon *et al.*, 1993; Grosveld *et al.*, 1993 and Hanscombe *et al.*, 1991).

The first model envisions that the LCR's function is to open up the chromatin structure over the entire globin gene domain thus rendering it accessible to transcription factor binding at the regulatory elements of the individual genes. The developmental expression of the genes is then the result of stage-specific binding of transcription factors and transcriptional interference. According to this model, the genes behave independently to each other and there is no competition between them for interaction with the LCR. Since several studies have clearly indicated the existence of gene competition between the globin genes (Enver *et al.*, 1990; Hanscombe *et al.*, 1989 and Giglioni *et al.*, 1984), this model is unlikely to account for the basis of LCR function.

The second model of scanning/tracking suggests that the LCR binds a transcriptional activator complex, which starts scanning along the DNA fiber of the locus activating the first promoter poised for transcription that it encounters. This model can account for the results obtained in the studies on gene order and distance of genes from the LCR (Dillon *et al.*, 1997 and Tanimoto *et al.*, 1999), however, it is difficult to explain the alternating expression of the γ and β genes in the same cell observed in foetal liver cells (Wijgerde *et al.*, 1995) and the order/distance parameter which plays a role in competition.

The third model is that of looping. This model envisions direct interactions of the LCR as a holocomplex and genes in the locus via the "looping out" of intervening DNA sequences. The presence of transcription factors at the gene promoter will secure the binding of the LCR and activation can take place. The LCR could activate one gene and then loop directly to the next gene, thus explaining γ and β transcription at the same time. The presence of the appropriate transcription factors and the strength of binding of the LCR to the promoter will determine the duration of time that the LCR will be present at a promoter and thus the expression levels of a gene. The looping model can also account for the results obtained in the gene order and distance experiments. In this model a gene closer to the LCR would interact more frequently with the LCR than a gene that is more distal, resulting in the higher expression of the more proximal gene. When the proximal gene is placed closer to the distal gene, then the frequency of interaction with the LCR would become less and the advantage of the proximal gene is reduced, resulting in smaller differences in expression between the proximal and distal genes. This is what was observed in the β^{marked} experiments. This model can explain all the results thus far obtained for the expression patterns of the genes and by the time that I will defend my thesis there will be direct proof for this model.

Transcription factors and globin expression

Transcription factors play important roles in the regulation of globin genes. In the following paragraphs the most important factors will be described and their actions on the globin genes summarized.

GATA1

GATA1 was the first member to be identified of a family consisting of six proteins, all recognising the consensus GATA motif (Orkin, 1992). The GATA motif is found in almost all regulatory elements of the globin genes. At first it was thought that GATA1 was erythroid specific, however, it is also present in other haematopoietic lineages such as mast cells, megakaryocytes and eosinophils (Martin *et al.*, 1990 and Crotta *et al.*, 1990) and in the Sertoli cells of testes (Ito *et al.*, 1993).

GATA proteins are characterised by two zinc-fingers which interact with the major groove of the DNA helix (Omichinski *et al.*, 1993). The C-terminal finger is required for the binding of the GATA-motif, whereas the N-terminal finger is important for the stabilisation of this binding (Tsai *et al.*, 1989 and Trainor *et al.*, 1990) and for the interaction with other factors like, for example, friend of GATA (FOG) (Tsang *et al.*, 1997 and Tsang *et al.*, 1998).

GATA1 is thought to carry out several functions. GATA1 overexpression can dominantly affect lineage selection in cell lines (Kulessa *et al.*, 1995 and Visvader *et al.*, 1992). For instance the introduction of GATA1 in a myeloid cell line resulted in the induction of megakaryocytic differentiation (Visvader *et al.*, 1992). Studies indicate that GATA1 plays a role in the regulation of a cascade of downstream pathways in cellular differentiation. GATA1 has also been shown to play an important role in the balance between erythroid cell proliferation and survival (Weiss *et al.*, 1994). Furthermore, it has been shown that upon induction of GATA1 overexpressing MEL cells, cyclin A-dependent kinase activity was decreased much less in the GATA1 overexpressing than in control cells. In the same study it was also shown that GATA1 binds to the retinoblastoma protein. The data together led to the conclusion GATA1 regulates differentiation by affecting the cell-cycle apparatus (Whyatt *et al.*, 1997).

GATA1 knockout embryos die at embryonic day 10 or 11 from severe anaemia. This is caused by the production of erythroid precursors arrested at the pre-erythroblast stage, which then undergo apoptosis (Pevny *et al.*, 1991; Fujiwara *et al.*, 1996 and Weiss *et al.*, 1994). Overexpression of GATA1 showed the opposite effect of stimulation of proliferation of pro-erythroblast cells resulting in inhibition of differentiation (Whyatt *et al.*, 1997). Finally, GATA1 plays a role as transcriptional activator, which correlates with the presence of GATA1 binding sites in the promoters of the globin genes and the core regions of the hypersensitive sites of the LCR. In conventional reporter assays in heterologous cells, GATA1 has been shown to act as a transcriptional activator (Martin and Orkin, 1990). Furthermore GATA1 has been reported to have effects on the expression of the globin genes (see the paragraph on gene regulation). Studies in which GATA1 knockout cells were tested for rescue of differentiation by different forms of GATA1, showed that its transcriptional activation function can be dissociated from its survival and differentiation function (Weiss *et al.*, 1997).

These studies also led to the suggestion that GATA1 probably needs a transcriptional co-activator. In agreement with this suggestion GATA1 has been shown to interact with several other transcription factors via its zinc finger domain. Examples of these factors are EKLF, Sp1 (Merika and Orkin, 1995), p300/CBP (Blobel *et al.*, 1998) and FOG (Tsang *et al.*, 1997 and Tsang *et al.*, 1998). The precise function of these interactions remains to be elucidated. The interactions of GATA1 with different transcription factors suggest that it has a primary role in the formation of a haematopoietic transcription factor complex at specific sites in the globin locus, potentially controlling expression at different developmental time-points (Orkin, 2000). Furthermore, the interaction with p300/CBP, might be a way via which histone acetyltransferases are brought to specific DNA sites (Blobel *et al.*, 1998), resulting in the acetylation of histones and enhancement of transcription of the globin genes (Boyes *et al.*, 1998).

NF-E2

NF-E2 was the second erythroid-specific factor to be identified (Mignotte *et al.*, 1989). It was initially found to bind AP-1 sites in the promoter of the human porphobillinogen deaminase (PBGD) gene. Subsequent studies showed that the AP-1 sites present in HS2 enhanced expression of reporter constructs in transfected cells (Ney *et al.*, 1990). The same activation by HS2 was also observed in transgenic mice (Talbot *et al.*, 1990 and Caterina *et al.*, 1994). NF-E2 binds as a heterodimer and consists of a haematopoietic subunit called p45 NF-E2 and a more widely expressed subunit called p18 NF-E2 or MafK (Andrews *et al.*, 1993a and 1993b). p45 NF-E2 contains the transcriptional activation domain. Both subunits are family members of the basic leucine zipper family. *In vitro* studies using MEL cells support the idea that NF-E2 plays a role in globin expression. MEL cells not expressing p45 NF-E2, cannot sustain high levels of globin expression. Reintroduction of the p45 NF-

E2 subunit restored expression of globin genes (Lu *et al.*, 1994). Furthermore, these studies also indicated that multiple p45 NF-E2 subunits are required for NF-E2-mediated activation (Bean and Ney, 1997). Using chromatin immunoprecipitations a recent study showed that the NF-E2 complex is recruited to both the LCR and the active globin promoters upon induction of MEL cells. This recruitment has been shown to correlate with a 100-fold increase in β_{maj} globin expression. From these results it has been speculated that the recruitment of the NF-E2 complex to both the LCR and the active globin promoters may be a rate-limiting step in the globin gene expression (Sawado *et al.*, 2001).

Although shown to be important for globin expression *in vitro*, the *in vivo* role of NF-E2 is not clear. Knockout mice for p45 NF-E2 only show a subtle reduction in globin expression, however, they do suffer from the loss of production of circulating platelets (Shivadasani and Orkin, 1995 and Shivadasani *et al.*, 1995).

EKLF

CACC motifs are present in many gene promoters, including those of the β-globin genes, and are bound by a number of proteins, like Sp1 and Kruppel related proteins. EKLF has been identified by cDNA subtraction assays between lymphoid and erythroid transcripts (Miller and Bieker, 1993) and is highly erythroid specific (Southwood *et al.*, 1996).

EKLF contains three zinc fingers that bind specifically to the CCACACCCT sequence found in the β-promoter and HS3 of the LCR (Feng *et al.*, 1994 and Gillemans *et al.*, 1998). The transcriptional activity of EKLF seems to be downstream to that of GATA1, as indicated by the presence of GATA1 binding sites in the EKLF-promoter (Crossley *et al.*, 1994). Although present throughout development, with binding sites present in all globin gene promoters (except for the δ-promoter), EKLF only acts on the β gene. This has been shown in knockout mice for EKLF. These mice die from anaemia at the foetal stage due to a deficiency in β-globin synthesis. No effect on expression of the other globin genes is observed in these mice (Perkins *et al.*, 1995 and Nuez *et al.*, 1995; for effects of EKLF on the globin genes also see the paragraph on globin gene regulation). Overexpression of EKLF in transgenic mice, showed a reduction in platelets which suggests that EKLF could also play a role in the balance between megakaryocytic and erythroid lineages (Tewari *et al.*, 1998).

Furthermore, EKLF has been suggested to have an effect on the γ to β switch. In human β globin locus transgenic mice heterozygous for the knockout allele of EKLF (humβ$^{+/+}$/EKLF$^{+/}$), γ expression is increased with a concomitant reduction in β expression during the period of gene competition between γ and β genes (Wijgerde *et al.*, 1996 and Perkins *et al.*, 1996). Finally, EKLF has been shown to play a direct role in LCR function (Gillemans *et al.*, 1998). HS3, thought to play a role in the chromatin opening function of the LCR (Ellis *et al.*, 1996) contains binding sites for EKLF. DNase I hypersensitivity of HS3 is markedly reduced in EKLF knockout mice. In addition, EKLF has been shown to interact *in vitro* with the chromatin remodelling complex E-RC1 (Armstrong *et al.*, 1998). Combination of these data led to the suggestion that EKLF could play a role in LCR activation by binding to HS3 and recruiting E-RC1.

Aim of the PhD project

A number of different elements play a role in the developmentally regulated expression of the human and mouse β- and α-globin geneclusters. The elements, which are involved in this process, have been discussed in the introduction above. During my PhD project I investigated the activation of the mouse α-and β-globin gene loci in their endogenous context. I was involved in the development of novel methodology that allowed the manipulation of the human β-globin locus in the context of a 185 kb PAC by homologous recombination in *E. coli*. I further applied this methodology in deleting separately HS2 and HS3 from the human β-globin LCR and assaying the effects these deletions had on the regulation of the locus in transgenic mice. Finally, I was involved in a project examining the role in the regulation of γ gene expression of putative regulatory sequences located downstream of the $^{A}\gamma$-globin gene in the context of the human β-globin locus in transgenic mice.

The aim of the first project was to investigate the basic mechanisms of transcriptional activation of the mouse globin gene loci. We made use of *in situ* hybridisation techniques to detect nuclear and cytoplasmic patterns of globin gene expression in 14.5 dpc mouse foetal liver cells. We were able to provide strong evidence that globin gene activation takes place in a stochastic, all-or-nothing manner which, once established, is clonally inherited in subsequent cell generations. The results of this project are described in chapter 2 of this thesis.

Chapter 3 describes work which extends our ability to modify the human β-globin in the context of a 185 kb PAC insert by homologous recombination. This method allows us to efficiently manipulate the locus while avoiding the limitations of the cosmid ligation approach.

The manipulation of the human β-globin locus by homologous recombination in *E. coli*, was applied in deleting two putative regulatory elements located downstream of the $^{A}\gamma$-globin gene and assaying the effects of this deletion on human globin gene regulation in transgenic mice. The two elements Enh and F have been associated with naturally occurring deletions that give rise to elevated γ-globin gene expression in the adult stage. The 5' breakpoints of these deletions map within the $^{A}\gamma$- and δ-globin intergenic region and it has been postulated that this region harbours *cis*-regulatory elements important for γ gene silencing in the adult stage. Consistent with this hypothesis, the Enh and F elements had previously shown to exhibit silence activity in transient transfection assays (Kosteas *et al.*, 1993and 1994). We tested whether this is indeed the case by deleting the two elements together in the context of the 185 kb human β-globin locus PAC. As described in Chapter 4, analysis of this deletion showed that Enh and F indeed act as locus-wide embryonic stage-specific transcriptional repressors, but are not involved in the regulation of γ switching in the foetal liver and adult stages.

In Chapter 5 we applied the method of homologous recombination in separately deleting HS2 and HS3 from the human β-globin LCR. This project extends on earlier work on the deletion of HS2 and HS3 in the context of a smaller (70 kb) human β-globin locus construct (Milot *et al.*, 1996). This work showed different chromosomal position effects in transgenic mice resulting from the deletion of HS2 and HS3. For example, HS2 deletion led to a classical effect of PEV, whereas deletion of HS3 resulted in a novel type of position effect called cell timing position effect (Milot *et al.*, 1996). These observations raised the prospect that the deletion of specific HS sites gave rise to different chromosomal position effects. These studies, however, were done in multiple copy mice with only few transgenic lines available. We extended this work by obtaining additional HS2- and HS3-deleted lines. Only single copy lines were analysed. The results of this study revealed no correlation between deletion of a specific hypersensitive site and a specific type of position effect, in contrast to the effects observed in the studies by Milot *et al.*, 1996. The results of the study and discussion on the differences between this study and earlier studies on transgenic mice carrying the human or mouse β-globin locus with deletions in the LCR are described in Chapter 5.

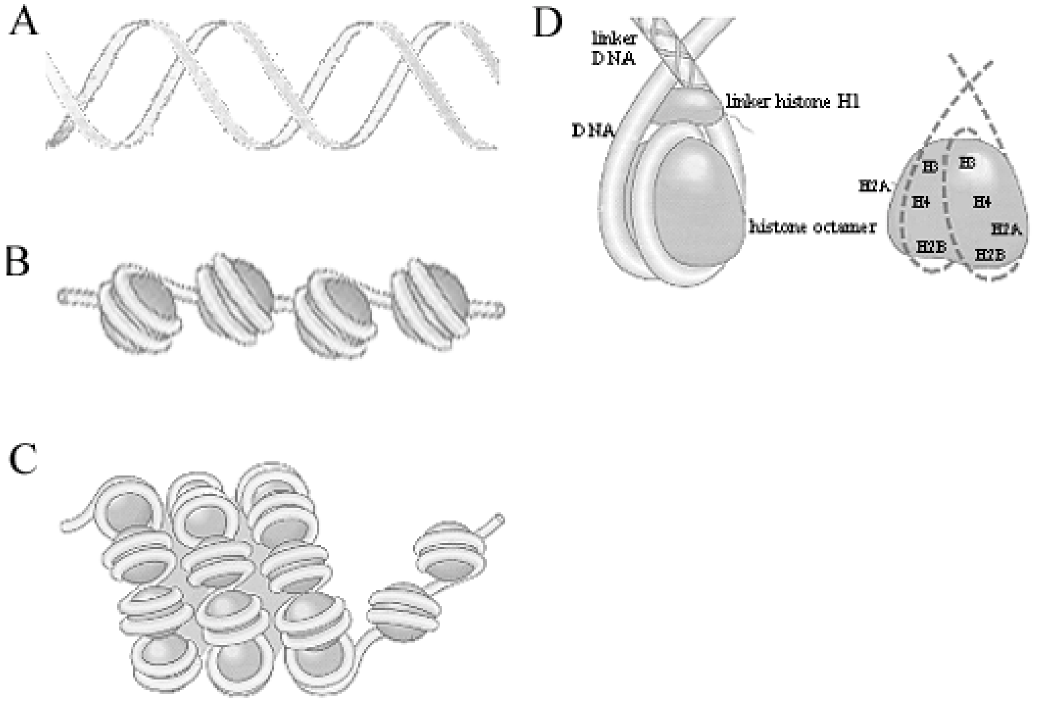

Figure 1: *Chromatin structure*—A: DNA helix, B: "beads on a string", DNA wrapped around histone octamers resulting in a 10nm fiber, C: compaction of the chromatin fiber into the 30nm fiber, D: left, a schematic representation of the organisation of a nucleosome; right, the organisation of the core histones in the histone octamer. (adapted from Wolfe; Molecular and cellular biology, 1993).

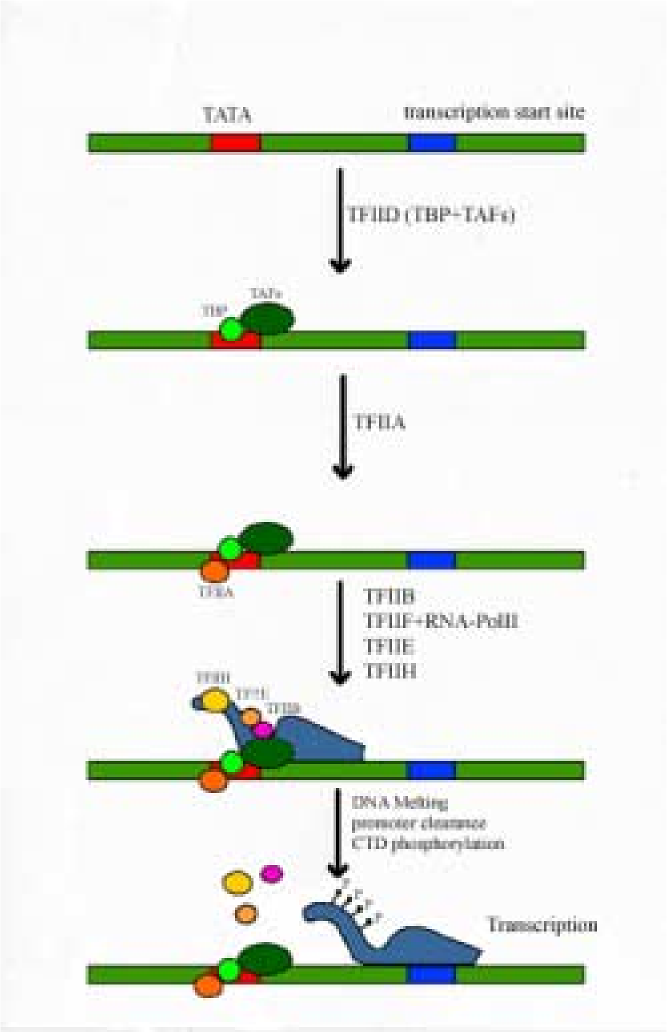

Figure 2*: Transcription initiation*—The start of transcription initiation, binding of the TFIID complex, composed of TATA box binding protein (TBP) and TBP-associated factors, to the TATA box, followed by the recruitment of TFIIA, which stabilises the TFIID-DNA interaction. This complex undergoes a conformational change allowing TFIIB to bind and the recruitment of the TFIIF-RNA-Pol II complex, which is followed by the binding of the two last components of the complex, TFIIE and TFIIH. The final step of initiation includes, DNA melting; promoter clearance, for which the TFIIH and TFIIIE are responsible; and the CTD phosphorylation of the RNA-Pol II, done by TFIIH. The phosphorylated RNA-Pol II is released from the complex and transcription can take place. (adapted from Ogbourne and Antalis, 1998).

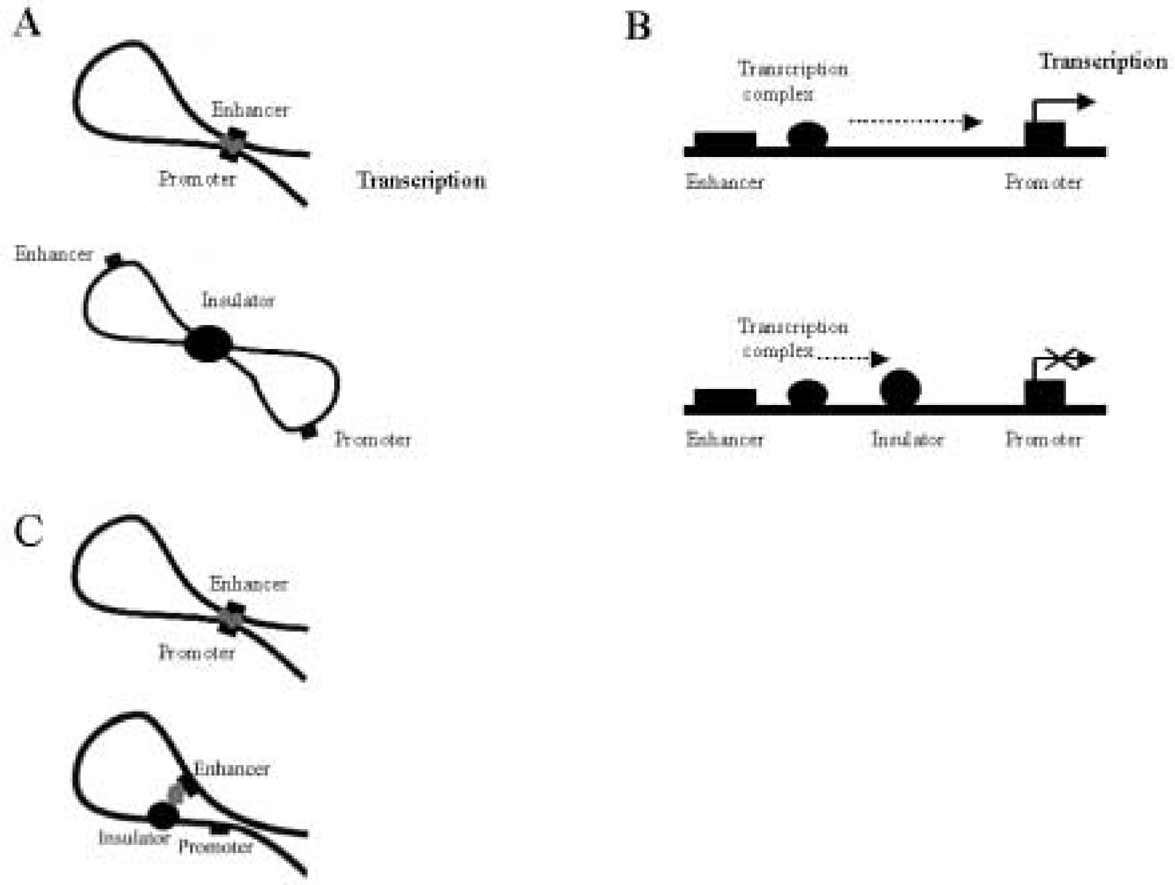

Figure 3: *Local interaction models for insulator function*—A: the decoy looping model, the interaction between an enhancer and a promoter, as in the top panel, is prevented by the formation of an additional loop by insulator binding, placing enhancer and promoter in two separate loops, bottom panel. B: the derailment or tracking model, in which the spreading of activation signals, as in top panel, is blocked by the binding of an insulator, bottom panel. C: straight competition.

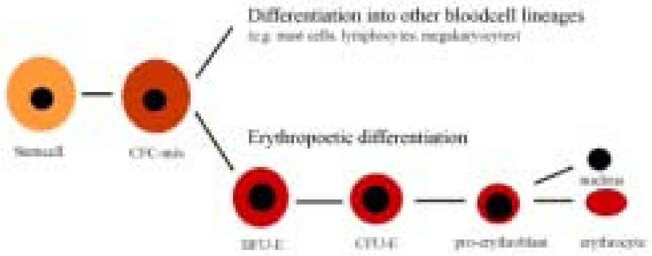
Differentiation into other bloodcell lineages
(e.g. mast cells, lymphocytes, megakaryocytes)
Stemcell
CFC-mix
Erythropoetic differentiation
nucleus
BFU-E
CFU-E
pro-erythroblast
erythrocyte

Figure 4: *Erythroid differentiation pathway*—The erythrocytes arise from a pluripotent stem cell which divides and produces more pluripotent stem cells or commited progenitor cells (CFC-mix). Cells from the CFC-mix start to differentiate into the different bloodcell lineages under the influence of specific growth factors. During erythropoetic differentiation the nucleus becomes smaller because of chromatin condensation and in the final step of differentiation enucleation takes place, resulting in the erythrocyte. Accumulation of haemoglobin in the red blood cells takes place during the terminal phase of differentiation.
BFU-E: burst forming unit erythroid; CFU-E: colony forming unit erythroid; CFC-mix: colony forming cells.

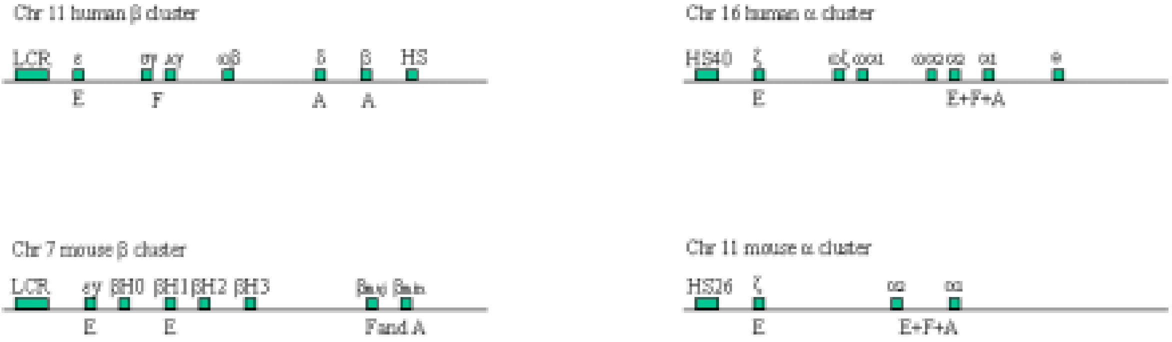

Figure 5: *Schematic representation of the α- and the β-globin loci of human and mouse*—E: gene expressed during embryonic phase; F: gene expressed during foetal phase; A: gene expressed in the adult.

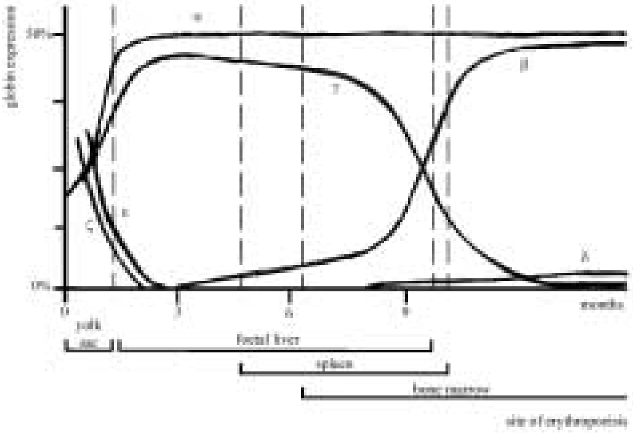

Figure 6: *Expression patterns of the human α-and β-globin genes*—The site of erythropoeisis during development is depicted below the expression profiles of the genes.

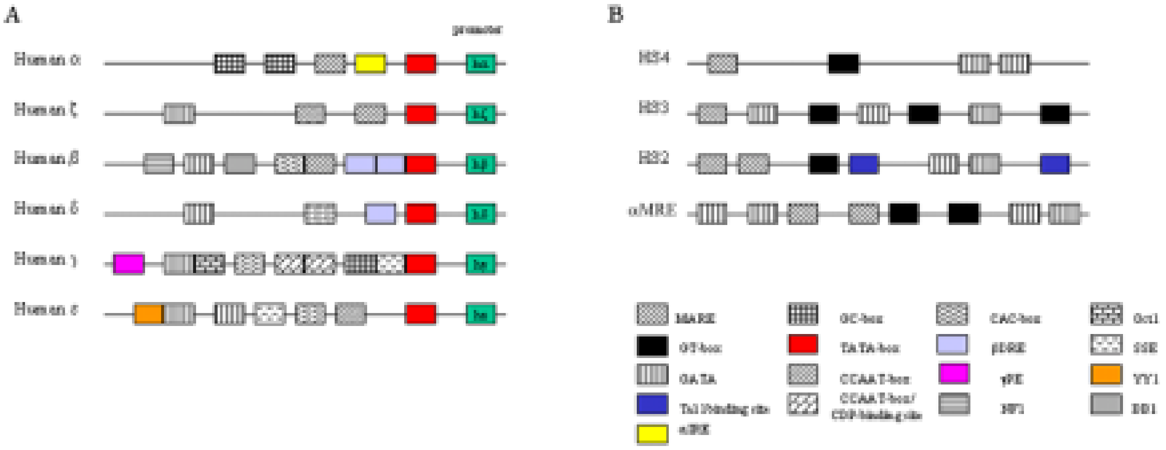

Figure 7: *Schematic representation of the regulatory elements of the human α- and β-globin loci*—A: representation of the transcription factor binding sites in the regions upstream of the individual globin gene promoters. B: representation of the transcription factor binding sites present in hypersensitive sites 2, 3 and 4 of the human β-globin LCR and the αMRE of the human α-globin locus.

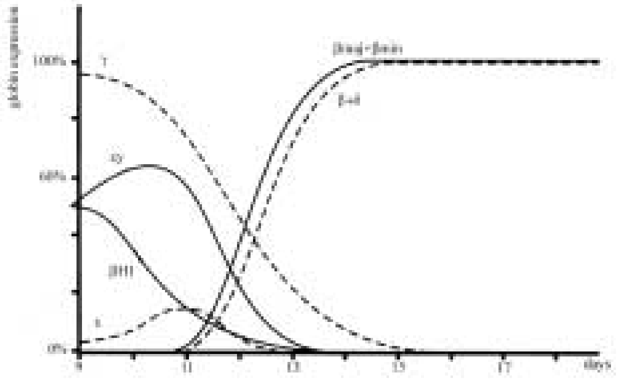

***Figure 8:** Comparison of the expression patterns of the endogenous mouse β-globin genes and the expression patterns of the human β-globin genes in transgenic mice*—Dotted lines represent the human genes and the black lines the mouse genes.

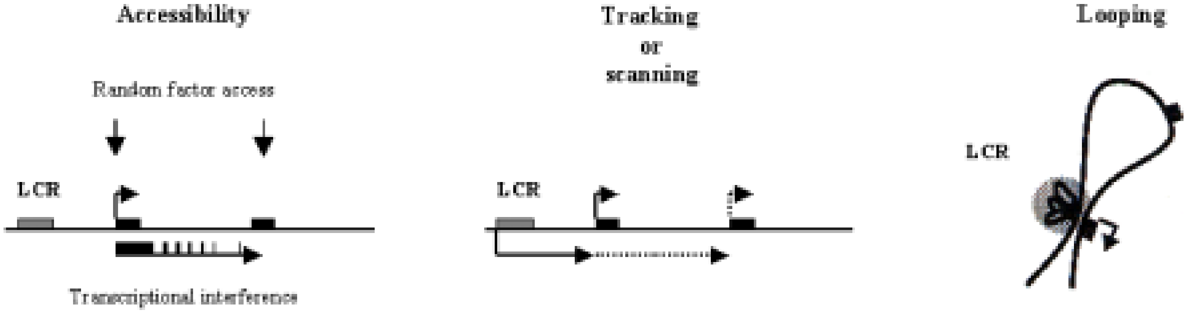

***Figure 9:** Long-range activation models proposed for the activation of the human β-globin genes by the LCR*—The arrows on the black boxes indicate transcriptional active genes. The arrow beneath the genes in the accessibility model indicates, that the transcriptional interference caused by an upstream gene decreases with distance.