Chapter 1

Introduction

1.1 Summary

Many gene loci are regulated by an extensive organization of *cis*-regulatory DNA elements and an intricate system of cooperating *trans*-acting factors. The study of two of those loci, the β-globin loci of man and mouse is described in this thesis. Several *cis*-regulatory DNA elements that are distributed throughout the globin locus and a large number of *trans*-acting factors are essential for its proper transcriptional regulation. In the nucleus, chromatin is largely constrained in its motion, thereby limiting *cis*-regulatory DNA elements to a small volume, while *trans*-acting factors can diffuse "freely". Since the initial (bio)chemical reaction rates are proportional to the concentration of the molecules involved, the different dynamic behavior of *cis*-regulatory DNA elements and *trans*-acting factors raises the question of how they are able to reach a sufficiently high concentration to drive transcription. One possible strategy is to increase a local concentration of *trans*-acting factors in spatial proximity to the *cis*-regulatory DNA elements. Indeed, the nucleus has a number of separate compartments that all enclose transiently accumulated factors. The paradigm of such a nuclear compartment is the nucleolus, dedicated to the transcription of ribosomal RNA (rRNA) genes. It is a substructure that holds constrained chromatin as well as resident proteins that enter and exit this compartment at a high rate. This thesis describes data that provide insight into the regulation of the globin genes, which are transcribed by RNA polymerase (RNAP) II. Much like the nucleolus, the results described in this thesis suggest that there is a nuclear compartment consisting of clusters of *cis*-regulatory DNA elements, which in turn may mediate a transient accumulation of *trans*-acting factors. This nuclear compartment dedicated to RNAP II transcription of globin encoding genes has been named the "active chromatin hub" (ACH).

1.2 Transcription

Eukaryotes contain three different RNAP enzymes (RNAP I, II, and III) each dedicated to the transcription of a distinct set of genes. RNAP II transcribes mainly protein encoding genes and is the topic of this introduction. The transcription of a gene depends on several distinct features of RNAP II and the transcription machinery. First, it must be able to recognize gene-specific sequences in the bulk of DNA in order to initiate transcription properly. Second, the transcription of many genes takes place in a time and tissue-specific manner, while other genes are kept silent. Third, DNA in the nucleus is packaged into chromatin (figure 1.2.1). The ground state of native chromatin in eukaryotes is thought to be restrictive and not transcribed [1], while actively transcribed chromosomal domains have an altered chromatin structure (see box: DNase I) [2, 3]. Hence, there must be mechanisms to alleviate the restrictive forces. Two cooperating components are key in regulating RNAP II transcription. One component acts *in cis* and requires an element that is located on the same DNA molecule as the transcribed gene, the so-called *cis*-regulatory DNA elements. The other component acts *in trans*, i.e. it requires molecules separate from the gene containing DNA molecule, here referred to as *trans*-acting factors, which are a diverse group of proteins (and RNA) that act on the *cis*-regulatory elements and the transcribed gene.

Box: DNase I

Distortions of chromatin structure can often be detected by an increase in sensitivity to nucleases such as the endonuclease DNase I. This enzyme hydrolyzes double- or single-stranded DNA preferentially in a relative random manner at sites adjacent to pyrimidine nucleotides. Susceptibility to DNase I may therefore be used as an assay to measure the general accessibility of DNA regions in chromatin. Two types of DNase I sensitivity can be observed: general sensitivity and hypersensitivity. General sensitivity is often found in transcriptionally active chromosome domains. Hypersensitivity is restricted to small regions, often located at *cis*-regulatory DNA elements, within general sensitive domains. These so-called hypersensitive sites (HSs) are thought to reflect areas with a less dense nucleosomal packaging or nucleosome-free areas.

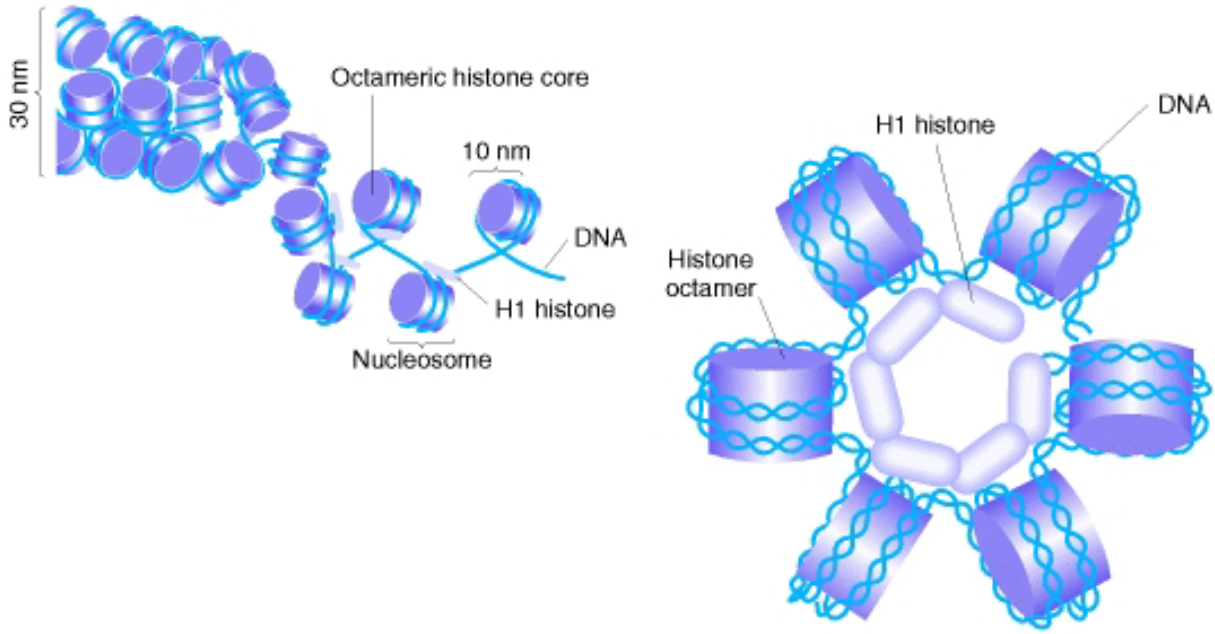


Figure 1.2.1. DNA packaged into chromatin by wrapping around nucleosomes.

The basic building block of chromatin is the nucleosome, which contains ~147 base pairs (bp) of DNA (blue fiber) wrapped 1.7-times around a core histone octamer (purple disk) in a 10 nm structure [288]. The core histone octamer consists of two copies of each histone H2A, H2B, H3, and H4. In the genome, thousands of nucleosomes are organized on a continuous DNA helix in strings separated by 10 to 60 bp of linker DNA. Each nucleosome can associate with histone H1 (grey/white rods) followed by coiling of the chromatin fiber into a solenoid structure with a diameter of 30 nm (on the left a side view and on the right top a view of the 30 nm chromatin fiber) (reviewed in [289]).

1.2.1 Cis-regulatory DNA elements

Promoters

The promoter specifies the site of transcriptional initiation *in vitro* and *in vivo.* The core RNAPII promoter is located approximately at -40 to +40 nucleotides relative to the RNA start site and contains at least one or a mixture of the following sequence elements, the TATA box, the TFIIB recognition element (BRE), the initiator (Inr), or the downstream promoter element (DPE) [4]. These sequences are sufficient to drive RNAP II transcription from a 'naked', i.e. not packaged into chromatin, DNA template *in vitro* [5]. From about -50 to -200 bp relative to the RNA start site there are typically multiple recognition sites for a subgroup of sequence-specific *trans*-acting proteins [4], here referred to as transcription factors. Together with the core promoter, these sequences can drive RNAP II in a reconstituted system [6] if the appropriate *trans*-acting factors are present. In mice, globin transgenes linked to their promoters are not transcribed to levels equal to that of endogenous genes [7]. *In vivo*, the promoters of active endogenous globin genes are DNase I HS [8].

Enhancers

Elements that increase transcription from a linked promoter are referred to as enhancers. Originally, they were defined in transient expression assays in cultured cells [9]. However, the definition does not specify the level of transcriptional enhancement or the assay used to measure it [10]. The enhancer effect is orientation-independent and shows some flexibility with respect to distance. In fact, distantly placed enhancers (> 1 kb) either upstream or downstream of a promoter are typical of higher eukaryotic genes [11]. Enhancers consist of DNA sequences that are recognized by transcription factors. The enhancers of the human β-globin gene can be detected as DNaseI HSs [8].

Enhancers, when linked to a promoter, seem to increase the probability of a gene establishing and maintaining an active transcriptional state in several experiments [11]. However, the presence of an enhancer is usually not sufficient to ensure high level of expression or expression in all cells that should express the gene when introduced into the mouse, or cultured cells, as a transgene. This is observed in various experiments. First, when a transgene is integrated stably into the genome of mammalian cell lines large differences in expression level are found, not related to presence or absence of an enhancer [12]. An increase in expression levels was found in mice carrying transgenes of β-globin constructs including its proximal enhancer elements when compared to β globin genes without the enhancers. However, expression levels did not correlate to the copy number of the integrated transgenes and were much lower than the expression levels of the endogenous genes [13, 14]. Similar results have been found for many genes and show that enhancers increase the proportion of the cells expressing a transgene, but cannot reproduce the normal expression pattern and/or level. In particular, they are often unable to overcome the restrictive effects of the chromatin structure at the site of integration of the transgene known as position effects. Interestingly, multiple copies inserts of a transgene carrying an enhancer can result in position-independent and copy-number dependent expression [15], suggesting that multimers of *cis*-regulatory DNA elements can cooperate to obtain proper levels of RNAP II transcription.

Locus Control Regions

Originally, Locus Control Regions (LCRs) were defined as dominant regulatory sequences that specify tissue specific, position-independent and copy-number dependent expression on a linked transgene in mice [16]. LCRs are capable of activating transcription of a single copy integrated transgene and mRNA levels per transgene copy are equal to mRNA levels of the endogenous gene [10]. Structurally, LCRs are composed of varying numbers of tissue-specific DNase I HSs and each site is composed of an array of sequences that can be bound by *trans*-acting factors [10, 17-19]. LCRs and the genes they regulate are often also separated by long stretches of intervening DNA. Furthermore, the location of individual HSs of a particular LCR can be spread out (figure 1.2.2) [10, 17, 20]. Many combinations of *cis*-regulatory DNA elements have been characterized in various vertebrate species that meet the functional definition of an LCR (reviewed [18]).

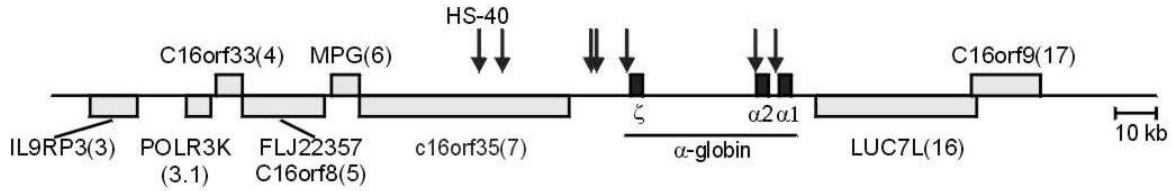
The most prominent property of LCRs is their strong, transcription-enhancing activity. When the LCR is absent, transcription of human β-globin transgenes is absent or very low [7, 21]. Deletion of the LCR from its endogenous position also eliminates or dramatically reduces transcription levels [22-26]. Another property of the LCR is that it can overcome a position effect induced mosaic expression pattern of a transgene, referred to as position effect variegation (PEV). Mice carrying a CD2 or β-globin transgene linked to their LCR show expression in all T cells or erythroid cells, respectively, even when it was integrated into the heterochromatic regions of the centromeres. In contrast, mice that lacked parts of the LCR show PEV [27, 28]. Importantly, the transgene constructs were DNase I insensitive in the nontranscribing portion of the cells [27, 28]. This indicates that the LCR operates by ensuring an open chromatin configuration that allows RNAP II transcription even in highly restrictive chromatin domains, such as (peri)centromeric heterochromatin (see below).

Insulators

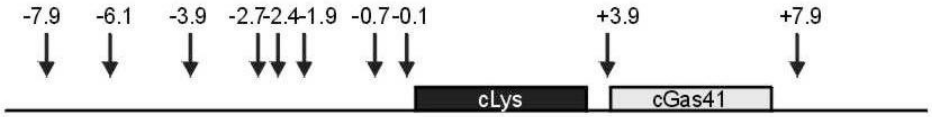
Insulators or boundary elements are DNA sequences defined operationally by two characteristics. They interfere with enhancer-promoter interactions when present between the enhancer and promoter. This can be experimentally defined in an enhancer-blocking assay.They are also thought to buffer transgenes from chromosomal position effects, which is experimentally measured by the ability to protect the gene in *cis* against PEV [29, 30]. The position of the insulator relative to the gene(s) and regulatory elements is essential to their function in that they must be placed between the anticipated source of silencing (e.g. chromatin) or activation (e.g. enhancer) and the affected gene locus (figure 1.2.3) [31]. Like other *cis*-regulatory DNA elements insulators are preferentially digested by nucleases [32, 33]. In addition, they contain consensus sequences for specific DNA binding proteins [34, 35].

A number of insulators have been reported in vertebrates (reviewed in [31, 36]) of which the chicken β-globin 5' HS4 (cHS4) is most relevant to our work [33, 37, 38]. The cHS4 was found to shield a reporter gene from the activating effects of a nearby mouse β-globin LCR in the human erythroleukemic cell line K562 [33]. Conversely it also protects against classical position effects in cell culture as HS4 protects the transcription of a stably transfected reporter gene in the absence of drug selection [38]. Most of the cHS4 insulating activity lies in a 250-bp CpG island (core element), which are often associated with promoters of housekeeping genes [37]. The human 5' HS5 was also found to have insulating properties in primitive cells, but not definitive cells [35]. In addition, a colony assay with 3' HS1 in beween the 5' HS2 enhancer and a reporter gene resulted in a 3-fold reduction of expression compared to a control construct without the 3' HS1. The authors interpreted this as an enhancer-blocking activity of 3' HS1 [39].

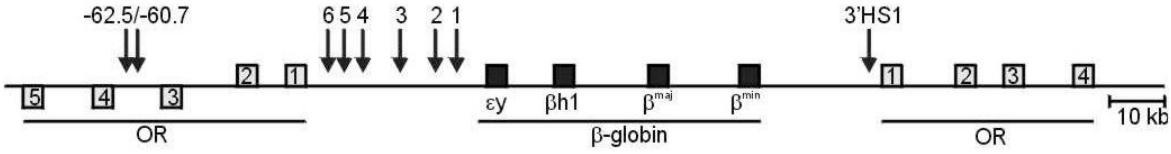
human α-globin cluster



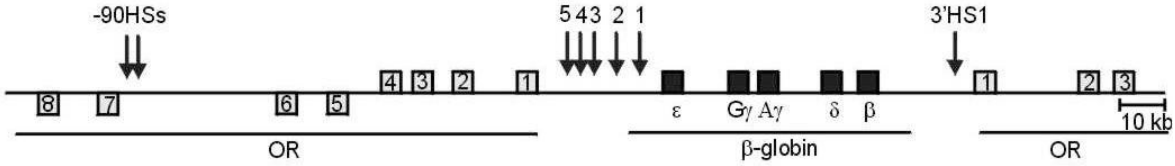
chicken lysozyme locus



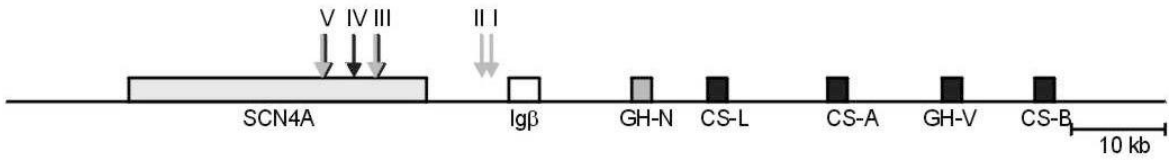
mouse β-globin cluster



human β-globin cluster



human Growth Hormone cluster



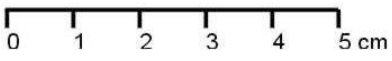


Figure 1.2.2. Genomic organization of several gene loci.

Boxes above the horizontal lines represent genes transcribing from left to right, boxes below represent genes transcribed in the opposite direction. Differentially expressed genes are indicated by different shades of grey and occur within one locus, but not separated by insulator or boundary elements. Arrows indicate known DNase I HSs (with grey/black colors corresponding to their target genes). The HSs are located throughout the depicted gene loci and often separated by a few kb to tens of kb. Note that the scale in kbs is indicated below each genomic fragment, except for the chicken lysozyme locus, where the numbering of HSs corresponds to the distance in kb from the transcriptional start site of the lysozyme gene. This figure was adapted from [20].

Synergy between elements

Despite their differences in functional definitions, there are two common features of *cis*-regulatory DNA elements: they can be detected as DNase I HSs *in vivo* and they are relatively small DNA fragments (200-300bp) that contain multiple binding sites for sequence-specific *trans*-acting factors. Those two properties suggest that *cis*-regulatory DNA elements represent regions that are bound by *trans*-acting factors. Clearly *cis*-regulatory systems somehow need to communicate with each other, in light of the fact that multiple *cis*-regulatory elements appear to be required to drive full transcription at endogenous loci and that combinations can overcome position effects in transgenic experiments.

1.2.2 *Trans*-acting factors

The transcription process starts with the assembly of a multiprotein complex at the promoter, the so-called preinitiation complex (PIC). Approximately 70 *trans*-acting factors are part of this preinitiation complex and required for its assembly [5]. Each of those factors contributes to the complexity of the initiation of RNAP II transcription. Transcriptional initiation culminates in the assembly of the general transcription machinery at the promoter. The key components are TATA-binding protein (TBP) and RNAP II.

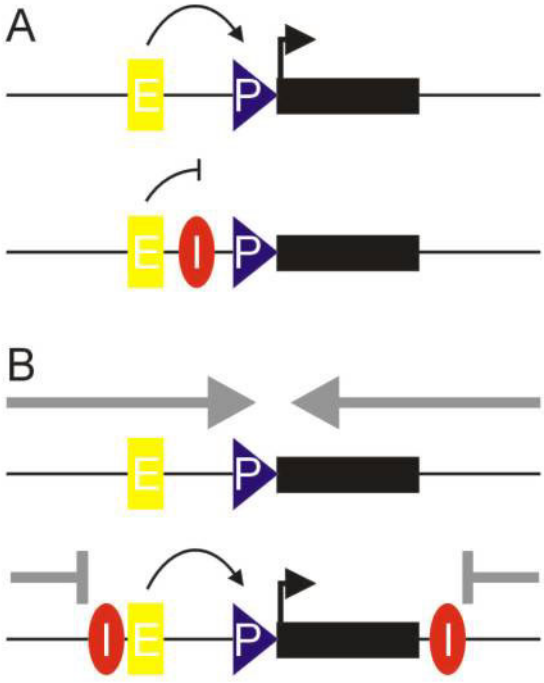


Figure 1.2.3. The position of an insulator relative to a gene is essential for proper functioning.

An enhancer (E, yellow box) can activate transcription (arrow) from a promoter (P, blue triangle) of a gene linked *in cis* (black rectangle) (A, top). In contrast, the enhancer activities are blocked when an insulator element (I, red oval) is located between the enhancer and promoter (A, bottom). Heterochromatin (grey arrows) spreads into the locus inhibiting transcription (B, top), while two flanking insulator elements act as a boundary and stop the heterochromatin spreading, consequently transcription takes place (B, bottom).

RNA polymerase II

RNAP II catalyzes DNA-dependent synthesis of RNA. It is a ~550 kDa protein complex composed of 12 subunits whose sequences are conserved among many diverse species. The largest subunit of RNAP II has a carboxy terminal domain (CTD), which is specific to RNAP II and plays key roles in the regulation of transcription initiation and coordination of co-transcriptional RNA processing events [40]. The CTD consists of a heptapeptide repeat with the consensus sequence Tyr-Ser-Pro-Thr-Ser-Pro-Ser. The CTD may be long, e.g. it consists of 26 repeats in yeast and 52 in humans, and therefore extends far from the main body of RNAP II. During PIC assembly the CTD is hypophosphorylated, however, after transcription initiation the CTD is rapidly phosphorylated [41]. Although RNAP II catalyzes transcription of protein encoding genes, it is unable to initiate promoter-dependent transcription in the absence of other factors [5].

General transcription factors

Direct basal level of *in vitro* RNAP II transcription from a 'naked' DNA template, containing a promoter sequence, is facilitated by a group of proteins called the general transcription factors (GTFs). *In vitro* studies have revealed a stepwise recruitment of GTFs and RNAP II at the promoter to complete PIC assembly. This process commences with a single polypeptide that binds the TATA box of the promoter; this is the TATA-binding protein (TBP). Subsequent to formation of the TBP-DNA complex transcription factor II B (TFIIB) enters the PIC and its C-terminal domain binds the promoter through the BRE downstream of the TATA box. The binding by TFIIB is likely to account for the direction of transcription. Next, RNAP II associated to TFIIF enters the DNA-TBP-TFIIB ternary complex, which has a stabilizing effect. Binding of the RNAP II-TFIIF complex at the promoter allows entry of the next two GTF complexes, TFIIE and TFIIH. First, TFIIE is recruited, which affects late events in PIC assembly, including TFIIH recruitment and regulation. TFIIH is a large and complex GTF, consisting of 9 subunits and it is the only GTF with defined enzymatic activities. Two of its subunits are ATP-dependent DNA helicases of opposite polarity (XPB and XPD), which also have a role in nucleotide excision repair (NER) [290-292]. Another subunit is a cyclin-dependent protein kinase (cdk7-cyclin H), which is essential for transcription intitiation while it is dispensable for NER. Both TFIIE and TFIIH are required for ATP-dependent formation of the open promoter complex prior to initiation of transcription (reviewed in [5, 40]). However, this stepwise assembly only holds true for assembly of the transcription machinery *in vitro*. *In vivo* several reports show that the order of events may differ per gene locus rather than a 'general' order of recruitment [42-48]. Furthermore, in reconstituted systems and *in vivo* chromatin, additional factors are required to initiate transcription, these include several types of co-factors and sequence-specific transcription factors.

Co-factors

Two groups of co-factors have been described (although some can be placed in both categories): protein components of, or interacting with the transcription machinery, and modifiers of chromatin structure. First, TBP-associated factors (TAFs) form a complex with TBP, called TFIID, which was originally defined as a general transcription factor. The hallmark of TFIID is its ability to support activated transcription in reconstituted systems. TAFs can serve as core promoter recognition factors by binding to the Inr and DPE sequences, which are located downstream of the TATA box. This stabilizes the TFIID on the core promoter, enhancing or replacing TBP-TATA box interactions. Furthermore, certain TAFs are also co-factors that are capable of binding to activation domains of sequence-specific transcription factors and exhibit enzymatic activities that are capable of modifying

proteins and potentially modulating their activities (reviewed in [5]). Another co-factor complex that has been described is Mediator, a modular complex that 'mediates' interactions between the activation and repression domains of TFs and the CTD of RNAP II. Thus, it transduces both positive and negative regulatory information from TFs to the basal transcription machinery. Mediator is composed of approximately 25 proteins and is unable to bind specific DNA sequences. To date, the mechanism of Mediator action is unknown (reviewed in [6, 40]).

The chromatin modifying co-factors can be divided into two distinct sets based on their enzymatic activities. First, there are co-factors that covalently modify proteins by either adding chemical compounds to or removing them from other proteins. Second, there are co-factors that use the energy of ATP hydrolysis in order to perturb or reorganize chromatin structure. These are collectively referred to as chromatin remodeling enzymes.

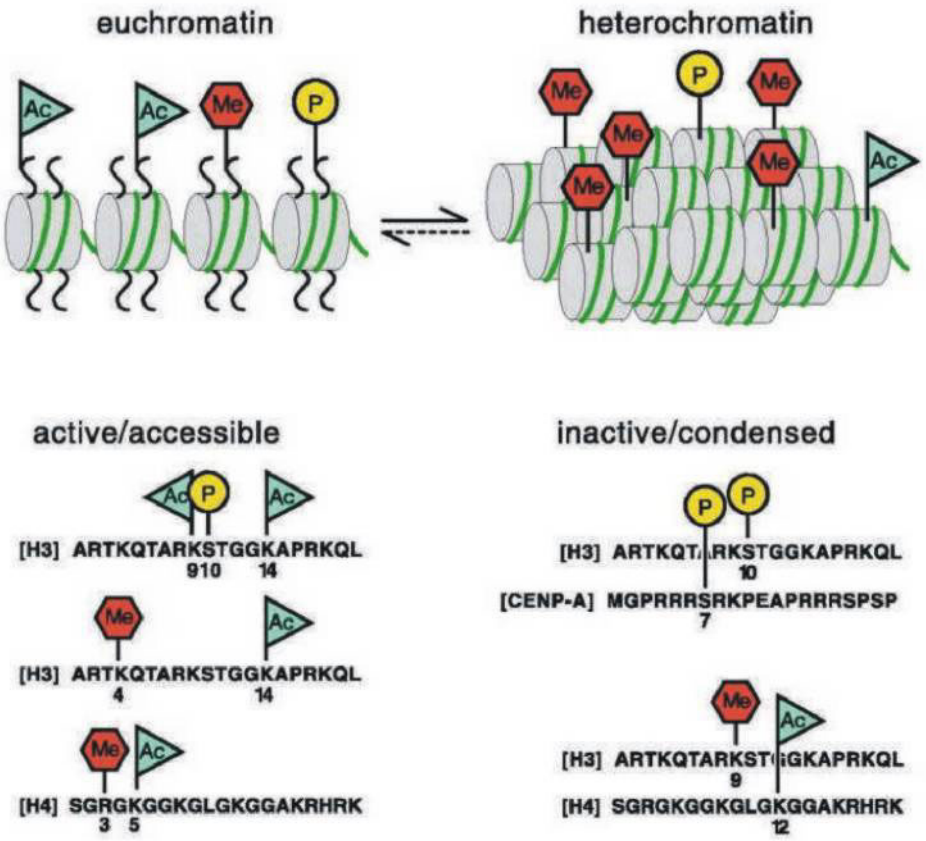


Figure 1.2.4. The 'histone code': models for euchromatic or heterochromatic histone tail modifications.

Top panel shows a schematic representation of euchromatin and heterochromatin as accessible or condensed nucleosome fibers containing acetylated (Ac, green triangles), phosphorylated (P, yellow circles), and methylated (Me, red hexagonals) histone NH_2-termini. Bottom panel illustrates examples of combinatorial modifications in histone NH_2-termini that are likely to represent "imprints" for active or inactive chromatin. Single-letter abbreviations for amino acid residues: A, Ala; E, Glu; G, Gly; H, His; K, Lys; L, Leu; M, Met; P, Pro; Q, Gln; R, Arg; S, Ser; and T, Thr. This figure was adapted from [53].

Co-factors that covalently modify proteins can alter the histones by a collection of post-translational modifications [49]. These modifications take place on the 'tail' domains of histones, which extend out of their globular domain, and some are known to correlate with nuclear processes (figure 1.2.4). For instance, histone acetyl transferases (HATs) function enzymetically by transferring an acetyl group from acetyl-coenzyme A to the ε-amino group of certain lysine residues within the histone tail [50]. This can result in histone hyperacetylation, which correlates with increased transcription [51]. The counterparts of HATs are histone deacetylases (HDACs) that remove the acetyl groups that leads to hypoacetylation, which correlates with transcriptional repression [51]. Histone methyl transferases (HMTases) target either lysine or arginine residues with different outcomes on transcriptional state. Methylation of lysine 9 and lysine 27 of histone H3 (H3-meK9 and H3-meK27, respectively) will lead to silencing and subsequent heterochromatinization. In contrast, H3-meK4 and arginine methylation at several residues of H3 and H4 correlates with an active state of transcription [52]. In the past, two hypothetical mechanisms have been proposed that explain these correlations between modification and nuclear processes. First, histone tails bind the DNA through charge interactions and modifications may alter chromatin structure by influencing histone-DNA and histone-histone contacts. This may lead to positive and negative changes in chromatin accessibility for other *trans*-acting factors. Second, a histone 'language' may be encoded on the tail domains. This 'language' is named the 'histone code' and may be read by other factors that are the actual agents of activity [49]. In fact, there are two protein domains found that associate with specific histone modifications. The bromodomain can selectively interact with acetylated lysines and chromodomains appear to be targeting modules for methylation marks [53]. Obviously, these two mechanisms may not be mutually exclusive but could very well co-exist *in vivo*.

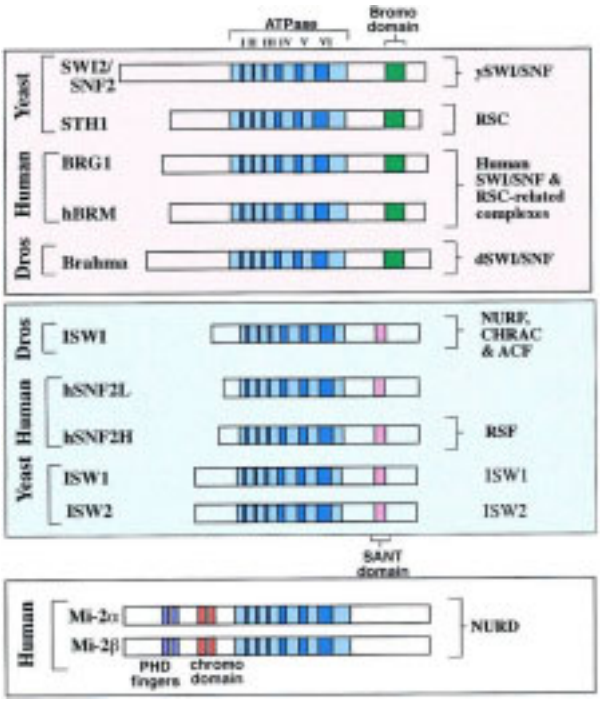


Figure 1.2.5 . ATPase subunits of chromatin remodeling complexes.

Conserved domains are labeled. Protein names are shown at left; names of complexes at right. SWI/SNF family complexes are depicted in the pink box (top), ISWI-family complexes in the blue box (middle) and Mi-2 complexes in the white box (bottom). This figure was adapted from [61].

Some HATs have been shown to acetylate nonhistone substrates as well as histones In fact, several erythroid-specific transcription factors are substrates for their HAT activities. GATA-1 can be acetylated *in vitro* and *in vivo* by p300/CBP and increases GATA-1-dependent transcription [54]. *In vivo* an acetylated form of erythroid kruppel like factor (EKLF) exists and interacts with PCAF and p300/CBP. However, *in vitro* it is only an acetylation substrate for p300/CBP [55]. The p300/CBP interacts with NF-E2, resulting in increased p300/CBP nucleosomal HAT activity and acetylation of NF-E2 [56]. These interactions suggest that posttranslational modifications can influence transcription factor activity and recruitment of HATs to specific *cis*-regulatory DNA elements.

Chromatin remodeling co-factors can perturb or reorganize chromatin structure in an ATP-dependent manner and this activity might relieve the restrictive and transcriptionally silent ground state. 'Chromatin remodeling' refers to numerous *in vitro* ATP-dependent changes in a chromatin substrate [57]. The *in vivo* activities of chromatin remodeling complexes are poorly understood. Only a few examples of *in vivo* chromatin remodeling activities have been reported [46, 58, 59]. Remodeling complexes are classified into three groups (figure 1.2.5) based on their biochemical properties and the overall sequence similarity of their ATPase units. First, the SWI/SNF-family of complexes mostly play a role in the activation of transcription. Second, ISWI members appear dedicated to transcriptional repression pathways and nucleosome assembly. Third, Mi-2/CHD complexes are also involved in transcriptional repression and they usually form complexes with HDACs [60, 61].

Some sequence-specific transcription factors interact with chromatin remodeling factors [62, 63]. For instance, two subunits of a SWI/SNF complex are necessary and sufficient for targeted chromatin remodeling and transcriptional activation by EKLF *in vitro*. They interact directly to generate a DNase I HS within the chromatin-assembled β-globin promoter. In contrast, these two subunits do not interact or function on the same promoter with two unrelated transcription factors [64]. In addition, SWI/SNF complexes contain a bromodomain, which has been shown to selectively interact with acetylated lysines [53, 61], suggesting an alternative targeting pathway mediated indirectly by HATs. Furthermore, human SWI/SNF complexes are phosphorylated during mitosis, which correlates with removal of the complexes from condensing chromosomes. Recent studies also suggest the existence of 'shielding' factors, which block ATP-dependent remodeling of chromatin fibers [65, 66]. This suggests that chromatin remodeling factors might be tightly regulated and may interact with other *trans*-acting factors, which would result in a proper targeting and timing of their activation or repressive activities.

Sequence-specific transcription factors

Transcription factors can be defined as proteins that recognize and bind to specific sequences that are usually located in *cis*-regulatory DNA elements and regulate transcription of specific genes through interactions with other factors. They contain a DNA-binding domain and one or more regulatory domains [50]. Transcription factors comprise a huge number of highly diverse proteins that can be expressed ubiquitously, tissue-specifically, or in a cell cycle-dependent manner. Describing all the members of this group is beyond the scope of this introduction. Here some general properties are recapitulated. First, several types of post-translational modifications mediated by other proteins can modulate the regulatory properties of certain transcription factors. For instance, the erythroid cell specific factors NF-E2, GATA-1, and EKLF can be phosphorylated [67-69], and acetylated [54-56]. Second, transcription factors may be involved in many protein-protein interactions, including those with the basal transcription machinery, other transcription factors and chromatin modifying co-factors. In this context, NF-E2 may modulate transcription through direct interaction with the basal transcription apparatus component $TAF_{II}130$ [70] and EKLF needs a specific chromatin remodeling complex, E-RC1, to regulate transcription [63, 64]. Recently, GATA-1 was found to interact with several proteins *in*

vivo using a single step purification method ([71] and Patrick Rodriguez, personal comm.), some of these are thought to be activating and others repressing transcription. Third, some factors can bind chromatin independently, while others require separate co-factors to facilitate binding in reconstituted systems. NF-E2 binding results in an ATP-dependent disruption of nucleosome structure, suggesting that chromatin remodeling factors present in the *Drosophila* embryonic assembly extract may be involved [72, 73]. EKLF requires the E-RC1 complex as separate co-factor to bind chromatin [63]. Recently, it has been shown that transcription factors HNF3 and GATA-4 can bind their specific binding-sequence without the activities of chromatin remodeling factors. [74]. The authors speculate that the ability of GATA-4 to bind and open chromatin may relate to the ability of the erythroid cell specific factor GATA-1 to disrupt a subset of histone-DNA contacts on a mononucleosome [54]. The highly conserved zinc finger motif of this family of factors may generally possess a chromatin disrupting function. The sequence-specific chromatin binding activity and the fact that transcription factors are involved in many protein-protein interactions provides a mechanistic framework for recruitment of other *trans*-acting factors at any given *cis*-regulatory DNA element.

How do *cis*-regulatory elements and *trans*-acting factors collaborate efficiently on the restrictive chromatin template to assemble the preinitiation complex at the promoter, which is followed by RNAP II transcription? Each *cis*-regalutory DNA element by itself appears to be a target of certain sequence-specific transcription factors and their interacting partners. However, many gene loci have widely separated regulatory elements that all appear to contribute to proper transcriptional regulation. One important aspect of (bio)chemistry is that the initial rate of chemical reactions is proportional to the concentration of molecules involved. An economical way to bring the different components together might be through compartmentalization [75].

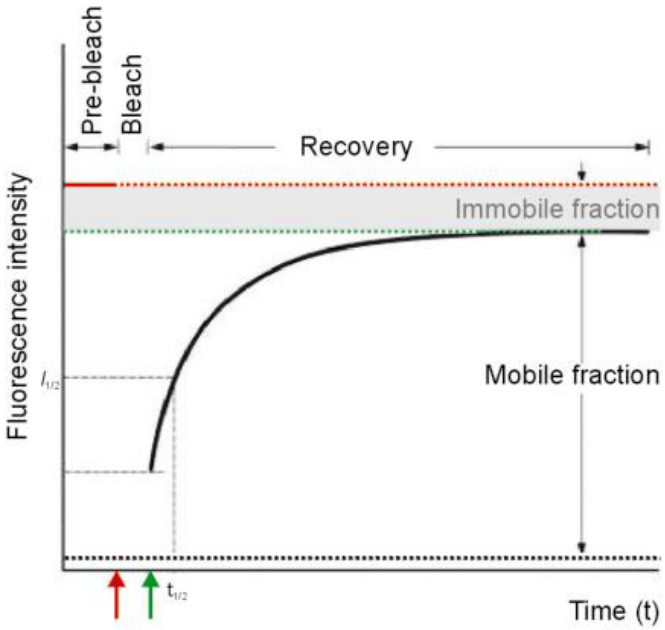


Figure 1.3.1. Quantitative fluorescence recovery after photobleaching (FRAP).

An idealized plot of fluorescence intensity as a function of time shows the parameters of a quantitative FRAP experiment. The bleach region is monitored during a pre-bleach period to determine the initial intensity (red solid/dotted line). This region is bleached using high-intensity illumination (red arrow), and recovery is monitored starting at t_0 (green arrow) until the fluorescence intensity reaches a final value, when no further increase can be detected (green dotted line). Some methods calculate the effective diffusion coefficient, D_{eff}, directly from the time ($t_{1/2}$) to reach half final intensity ($I_{1/2}$). To calculate D_{eff} as accurately as possible, the fluorescence intensity must be corrected for the background intensity (black dotted line) and the amount of total fluorescence removed by the bleach. The mobile fraction is the proportion of fluorescence that is regained, indicated by the difference between the initial and final fluorescence (grey area). This figure was adapted from [76].

1.3 The nucleus

Eukaryotic cells contain a specialized cell organelle that holds the genetic material (DNA), referred to as the nucleus. Based on microscopic studies, it has been suggested that the nucleus is organized into many distinct compartments. These so-called nuclear compartments were originally defined by the following parameters: they are morphologically identifiable structures using either light or electron microscopy, and contain a particular subset of resident proteins. In addition, some compartments can be biochemically isolated in an enriched form, although they lack any delineating membranes. In addition, kinetic studies (see box: FRAP and FLIP) demonstrate that the nucleus can also be subdivided into highly dynamic components and sections with constrained mobility.

Box: FRAP and FLIP

Quantitative fluorescence recovery after photobleaching (FRAP) generates information about the relative mobility of the fluorescently (e.g. GFP) tagged protein: the effective diffusion coefficient (D_{eff}). In FRAP a region is bleached irreversibly and then recovery of fluorescence in the bleached zone is monitored. To derive D_{eff}, the recovery is plotted as a function of time (figure 1.3.1). By comparing pre-bleach fluorescence intensity and the recovered intensity the protein can be separated in mobile and immobile fractions. Proteins are considered to have a high mobility, if they have both a high D_{eff} (i.e. fast diffusion rate) and a high mobile fraction. Studies based on fluorescence loss in photobleaching (FLIP) show continuity or transport between different populations of proteins, e.g. when fluorescently tagged proteins occupy distinct nuclear compartments. In FLIP, a region is repeatedly bleached, and the loss of fluorescence from outside the bleached zone is monitored. The rate of loss of fluorescence signal is an indicator of the mobility or exchange rate of the protein. The fluorescence signal from freely mobile molecules decreases rapidly, the fluorescence signal from slowly mobile molecules decreases slowly. Note that in this method the movement of unbleached molecules is monitored excluding the possibility that damage by the bleach pulse affects the mobility of the monitored proteins.
(For details see [76, 77])

1.3.1 The nuclear envelope

The nucleoplasm (i.e. the interior of the nucleus) is physically separated from the cytoplasm by a set of membranes and an associated protein network, collectively labelled the nuclear envelope (NE). Structurally the membrane components are an outer nuclear membrane (ONM) that faces the cytoplasm, and an inner nuclear membrane (INM) that faces the nucleoplasm. The INM has associated to it a unique set of integral membrane proteins, e.g. Lamin B receptor (LBR), lamina associated polypeptides (LAPs), and emerin. Most of those proteins interact with a network of polymers formed by lamins (i.e. nuclear intermediate filament proteins) called the nuclear lamina that is closely associated with the INM. Due to its polymeric nature, the nuclear lamina provides a scaffolding structure of the NE (reviewed in [78]).

Enclosure of nuclear content in eukaryotes implies that cells require sites in the NE that mediate exchange of material between nucleus and cytoplasm, given that RNA synthesis occurs in the nucleoplasm and protein synthesis occurs in the cytoplasm. The nuclear pore complexes (NPCs), 125 MDa macromolecular complexes, are responsible for this transport. [79, 80]. In vertebrates, NPCs consitst of 50-100 proteins, termed nucleoporins. Each NPC can be divided into three structural elements: cytoplasmic fibrils, the central core, and the nuclear basket. Transport through NPCs can be rapid (i.e. several hundred molecules/pore/second) by either passive diffusion of relatively small

molecules or active transport of complexes of up to several megadaltons. Each pore facilitates both import and export, using import and export receptors and small GTPases that travel with their cargoes through the pores. These cargoes include proteins and various types of RNA and RNP molecules [79, 80].

Dynamics of the nuclear envelope

Each cell division requires the complete breakdown and re-assembly of the NE, implying that the NE must have a high level of plasticity. In the early stages of mitosis, nuclear proteins are distributed throughout the cytoplasm [81, 82]. Shortly after the onset of anaphase the nuclear membranes assemble around chromosomes. Fusion proteins of green fluorescent protein (GFP) with the INM proteins LBR and emerin and the lamina protein lamin B1 demonstrated that these proteins localize to the daughter chromosomes in this very early telophase-stage [81, 82]. Despite its tremendous size and complicated molecular architecture, the NPC also breaks down and reassembles at cell division in vertebrates within a short period of time [79]. The timing of recruitment and spatial location of several nucleoporins coincides with recruitment of emerin and LBR to the daughter chromosomes. Nuclear import is recovered slightly later and chromosome decondensation occurs only after this recovery [82].

In interphase nuclei, kinetic studies of Lamin C and B1 fused to GFP show a very slow exchange rate when integrated into the NE, whereas nucleoplasmic GFP-lamin C diffuses rapidly. This has been interpreted to be due to the formation of higher-order polymers by lamin proteins and supports the view that the lamina is required for maintaining nuclear size and shape [81, 83]. Although incorporation and exchange of lamins is slow, the lamina itself shows considerable dynamic movements through folding and indentations. In addition, central core components of NPCs were found stably attached to a flexible lamina framework, but peripheral components can exchange rapidly with an intranuclear pool [84, 85].

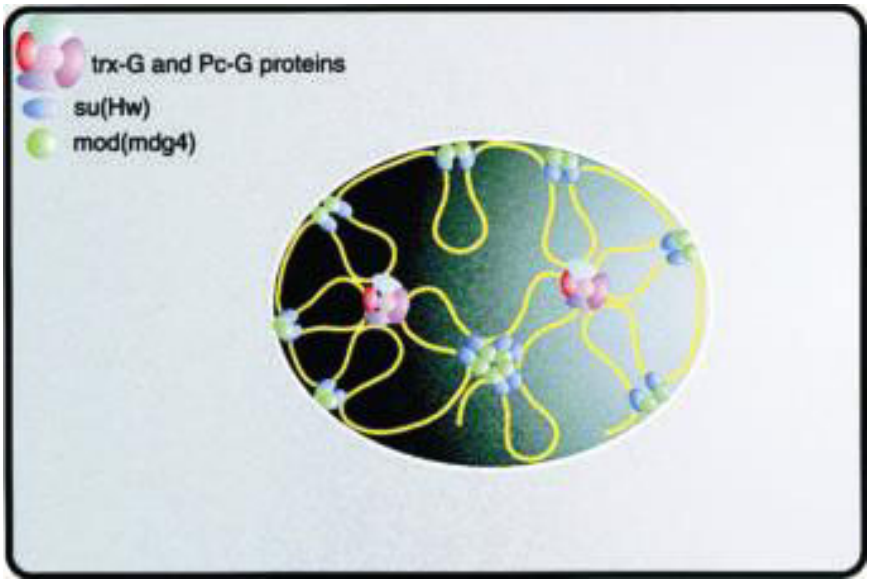


Figure 1.3.2. Schematic model that illustrates the role of *trx*G proteins in the function of the *gypsy* insulator higher-order chromatin domains.

The diagram represents a cell with a nucleus (oval). The chromatin fiber is represented as a yellow line, and proteins are represented as ovals: Mod(mdg4) (green), SuHw (dark blue) and various TrxG and/or PcG proteins (all other colored ovals). This figure was adapted from [92].

The nuclear envelope and transcription

Lamins bind to DNA, chromosomes and histones *in vitro* and the INM proteins LBR and LAP2 also bind to chromatin. Changes in lamin expression often correlate with transitions in development or differentiation, when gene expression patterns change. Thus, it has been proposed that the NE might influence the capacity for gene expression [78]. However, the actual effects of the NE on transcriptional regulation are largely unknown, while conflicting data have been published. Several indications suggest that the NE represents a repressive chromatin environment. First, heterochromatin is preferentially located near the envelope [78], while in the late stages of S phase, this heterochromatic DNA is released from the NE, moves to replication centers inside the nucleus and then returns to the envelope [86]. Second, it was shown that the yeast *HMR* locus, with a defective silencer element, exhibited improper silencing unless it was tethered to the nuclear periphery [87]. Third, native yeast telomeres, which have a heterochromatic structure, are highly dynamic but remain within a restricted volume adjacent to the NE during interphase [88-90]. Fourth, a nuclear periphery protein (Esc1) is a component of a pathway that functions to localize silencing complexes to the nuclear periphery [91].

In *Drosophila* a link between the NE and transcriptional activation, rather than silencing was observed. Here, a *cis*-regulatory DNA element, the *gypsy* transposon insulator, localizes to the NE in conjunction with *trans*-acting factors Su(hw) and mod(mdg4) to form so-called insulator bodies. A mutation in Su(hw) that destroys its insulator capacity causes a detachment form the NE. In addition, there was a correlation found between the organization imposed by the *gypsy* insulator and the ability of genes to be transcribed. Heat-shock response (i.e. heat-shock genes are turned on, the rest of the genes is turned off) causes redistribution of Su(hw) and mod(mdg4) in a diffuse pattern and reshuffling of insulator-linked DNA into the interior of the nucleus due to alterations in the chromatin organization imposed by the *gypsy* insulator [92]. Thus, *gypsy* may be required for establishing higher-order chromatin domains (figure 1.3.2), association with the NE and regulating gene expression.

In mammalian cells, the results are conflicting. In general, actively expressed genes are not found associated with the NE (reviewed in [93]). For instance, immunoglobulin (Ig) loci are selectively activated for transcription during B lymphocyte development. Ig heavy (H) and Igκ loci are preferentially positioned at the nuclear periphery in progenitors and pro T-cells but not associated with heterochromatin. In transcriptionally active pro-B cell nuclei they are centrally configured in the nucleus [94]. In addition, gene poor chromosomes are more likely to be at the nuclear periphery, while gene-rich chromosomes are localized to the nuclear centre [95, 96]. On the other hand, chromatin at the nuclear periphery is more DNase-sensitive than 'bulk' chromatin, probably reflecting transcriptionally active chromatin domains. The nuclear periphery has been proposed to function as a compartment for the spatial coupling of transcription and nucleocytoplasmic transport. Highly expressed genes might associate with NPCs to facilitate export of their mRNA [97-99]. Furthermore, it has been shown that a nuclear lamin mutant has defects in transcript localization (reviewed in [93]). Lamins are distributed throughout the nucleoplasm and disruption of the nuclear network formed by lamins is accompanied by a dramatic reduction in transcription by RNAPII that involves TBP, but does not detectably inhibit RNA polymerases I and III [100]. The authors suggest that lamins may act as a scaffold upon which the RNAP II basal transcription machinery is organized. Finally, a growing number of transcription factors either co-localize with the nuclear lamina or interact with proteins anchored to the lamina [101]. It should be noted in this context that expressed globin genes do not localize to the NE (Mariëtte van de Corput, personal comm.).

Thus, the NE may be involved in the establishment or maintenance of higher-order chromatin domains. On the other hand, the NE may be involved in a compartment dedicated to efficient transcription and subsequent nucleocytoplasmic transport, putatively located near NPCs. It is also

possible that the higher-order chromatin and nuclear compartment functions of the NE co-exist in a single nucleus.

1.3.2 Nuclear bodies

Nuclear bodies were the first substructures that were designated compartments. Their discovery depended mostly on studies involving light and electron microscopy. When light microscopy advanced into fluorescent and confocal microscopy, the number and types of nuclear bodies rapidly increased. Studies based on fluorescent fusion proteins, *in situ* hybridization of RNA and DNA, and immunolabeling of proteins, showed foci with an accumulated set of proteins (figure 1.3.3) and, in some cases, particular chromatin domains. Here, the best-characterized nuclear bodies are discussed.

The nucleolus

The nucleolus is assembled around clusters of tandemly repeated ribosomal RNA genes (rRNA genes) which are transcribed by RNAP I. The human rRNA genes are located in approximately 180 copies of tandemly arranged rDNA repeats on five separate chromosomes. These tandem arrays form the nucleolar-organizing regions and are the basis of the structural organization of the nucleolus, responsible for targeting of all processing and assembly components required for the nucleolus' major function, ribosome biosynthesis (reviewed in [102]). Morphologically it can be separated into three distinct substructures: fibrilar centres (FCs), dense fibrilar components (DFCs), and granular components. A typical human nucleolus is composed of ~30 FCs, each accommodating about four rRNA genes. They are surrounded by DFCs, while the granular components radiate out of the DFCs (figure 1.3.4). High-resolution *in situ* hybridization and Br-UTP-incorporation studies revealed that active rRNA genes are restricted to the periphery of the FCs, excluding their interior. The primary transcript (pre-rRNA) is approximately 13 000 nucleotides, known as 45S rRNA, and it is cleaved into 18S, 5.8S, and 28S rRNA [103]. Pre-rRNA processing is initiated in DFCs and finished in the granualar components [102, 104].

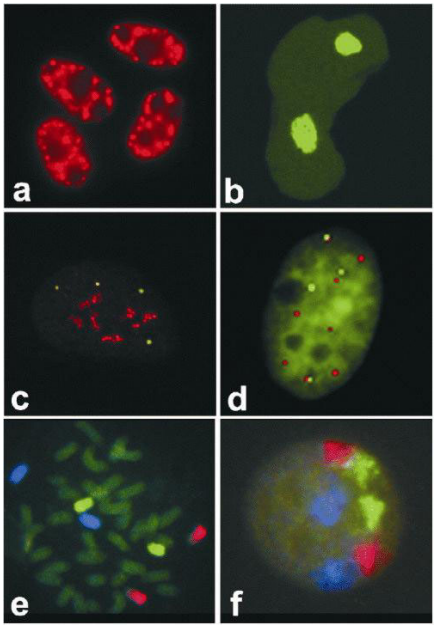


Figure 1.3.3. Nuclear compartments visualized by fluorescence microscopy.

Shown are splincing factor compartments (A, red), the nucleolus (B, green), the Cajal bodies and the nucleolus (C, green and red, respectively), PML bodies and Cajal bodies (D, red and green, repectively), and several mouse chromosomes (E) during metaphase. In interphase nuclei, chromosomes appear as finite and mutually exclusive entities, referred to as chromosome territories (F). Here, chromosomes 13 (blue), 14 (red) and 15 (green) are shown at both cell cycle phases (E and F). This figure was adapted from [102].

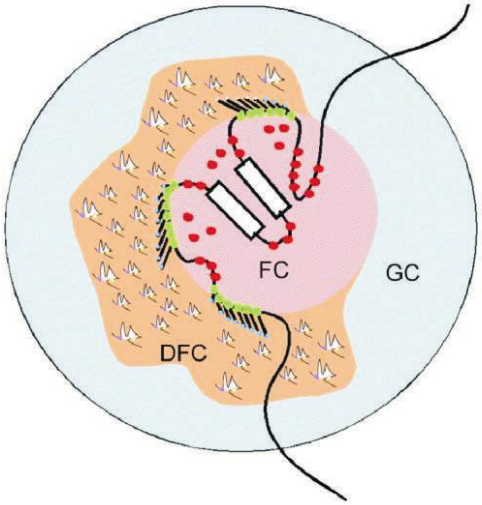


Figure 1.3.4. Organization of the nucleolar substructures.

The nucleolus is organized from chromatin-containing rRNA genes, which are anchored in the fibrillar centre (FC). Transcriptionally active rRNA genes are located at the periphery of FCs. Nascent pre-rRNA transcripts are entering DFCs, where they undergo a series of processing steps. Later processing and assembly steps occur in the granular component (GC). This figure was adapted from [102].

The rRNA genes are highly expressed, each active gene is associated with 100-120 engaged RNAP I molecules and transcripts are synthesized at a rate of ~2.5 kb per minute. Given that each transcript is about 13.3 kb, it takes approximately 5 minutes to transcribe an entire rRNA gene [105, 106]. Transcription of rRNA genes is essential for establishment and maintenance of the nucleolus. Upon inhibition of transcription the nucleolus disassembles (reviewed in [102]). Under these conditions only 11 different resident proteins were found by proteomics, whereas in transcribing nucleoli 271 different resident proteins were identified [107]. This demonstrates that transcription is the key to normal organization of the nucleolus. Furthermore, nucleolar components preferentially interact with each other and introduction of extrachromosomal rRNA genes triggers the spontaneous formation of novel nucleoli [102, 108]. Thus, interactions between rRNA genes and *trans*-acting factors appear to result in a transcriptionally active nucleolus.

Like the NE, mammalian nucleoli disassemble at the onset of mitosis and reassemble during telophase, indicating high structural plasticity. Partially processed pre-rRNA is preserved in association with processing components in the perichromosomal regions in particles called nucleolus-derived foci during mitosis. During telophase the nucleolus-derived foci disappear with a parallel appearance of material in the reforming nuclei. Prenucleolar bodies appear in nuclei in early telophase and gradually disappear as nucleoli are formed, strongly suggesting the transfer of prenucleolar body components to newly formed nucleoli. In interphase, nucleolar resident proteins are steadily and rapidly exchanged between the nucleoplasm and the nucleolus [77, 109]. The short residence times (order of seconds) and high exchange rate (thousands of molecules per second) demonstrate a large flux of proteins through the apparently stable nucleolus. RNAPI analysis shows that only 7 to 10% of the enzymes are actually engaged in transcription, leaving the vast majority moving freely at any given time [109]. The authors suggest a kinetic framework for RNAP I transcription *in vivo*. RNAP I machineries might be reassembled at each round of transcription in a sequential manner via meta-stable intermediates, each with increasing stability.

Transcription factories

Some research has focused on whether compartmentalization is involved in RNAP II and RNAP III transcription, analogous to the nucleolus. Several observations illustrate parallels between the different polymerases. First, in HeLa cells the number of nascent RNA transcripts per nucleus was determined to be ~90 000, with ~15 000, ~65 000, and ~10 000 being made by RNAPs I, II, and III,

respectively [105, 110]. Fluorescent detection of nascent transcripts in HeLa nuclei revealed a much lower number of foci of ongoing transcription, so-called transcription sites. It was estimated that each nucleus contains 5000 to 8000 RNAP II and about 2000 RNAP III transcription sites. Furthermore, many genes, but not all (e.g. globin genes [111]) were found associated with only one elongating complex at any given time [105, 110, 112]. The authors concluded from this that RNAP II and RNAP III transcription sites might represent a clustering of transcribed genes, the so-called 'transcription factories', that resembles rRNA gene clustering. Second, centers of RNAP II gene expression appear to be organized such that the chromatin template, transcription machinery and nascent product each occupy distinct zones (figure 1.3.5) [113] much like the structural organization of rRNA genes. Third, the kinetics of RNAP II enzymes are similar to that of RNAP I, the majority of nuclear RNAP II enzymes move freely throughout the nucleus and only 25% is engaged in transcription [114]. Fourth, several sequence-specific transcription factors fused to GFP were found to co-localize with transcription sites. They accumulated in bright foci that remained stationary in the nuclear space. However, kinetic studies showed a fast diffusion rate and only a fraction of the proteins was immobile [115]. Similar to the resident proteins of nucleoli, these transcription factors transiently accumulate at transcription sites, which results in visually stable structures. However, the analysis of co-localisation between transcription factors and transcription sites does not demonstrate whether transcription actually occurs once factors are recruited to a target gene [113].

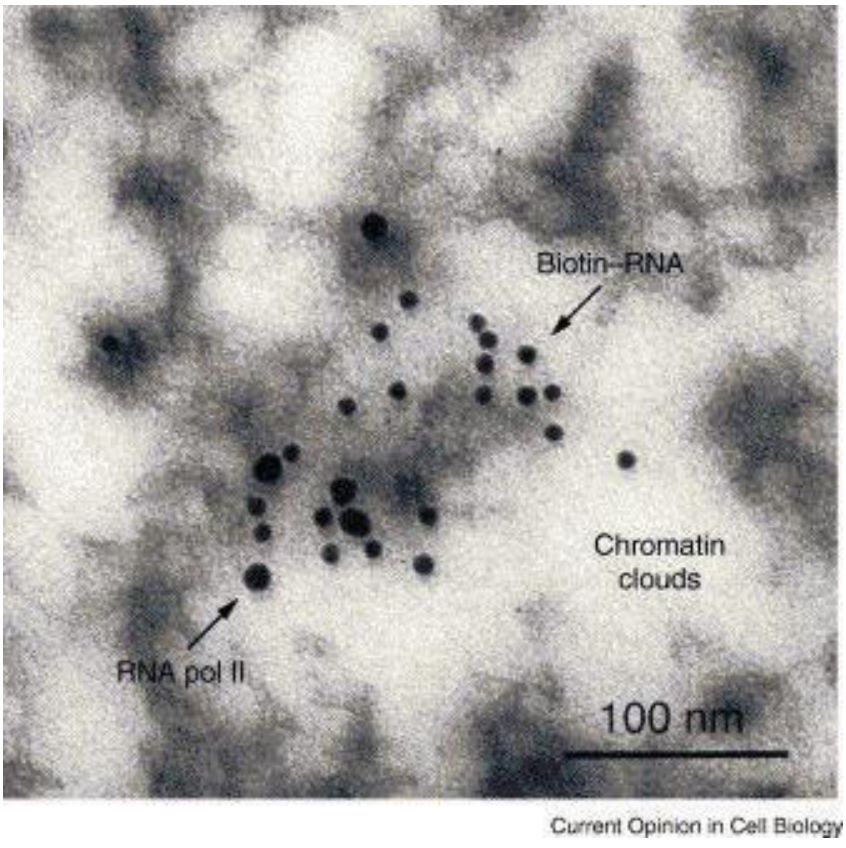


Figure 1.3.5. The anatomy of a transcription site.

Components involved in RNA synthesis are shown in a high-resolution electron microscopy (EM) image. In this example, the dense chromatin regions have a white/grey `cloud-like' appearance. Nascent transcripts are visualised with 5 nm gold particles. The transcription machinery is visualised with 10 nm gold particles. This figure was adapted from [113].

Splicing-factor compartments, Cajal bodies and PML bodies

In addition to the nucleolus, other morphological structures have been identified (reviewed in [102, 108, 116]). Three of these are described briefly: splicing-factor compartments (SFCs), Cajal bodies (CBs) and promyelocytic leukaemia oncoprotein (PML) bodies. SFCs or nuclear speckles are composed of dense granules (diameter 20-25 nm) connected by fibers. Its resident proteins are mainly pre-mRNA splicing factors. CBs are spherical structures (diameter 0.1-1.0 μm) that contain a large number of components, including spliceosomal snRNPs, and GTFs. Their function is unknown but transport/maturation of snRNPs and assembly site for the basal transcription machinery have been proposed. PML bodies are small spherical domains (diameter 0.2-1.0 μm) and functions have been implicated in terminal differentiation, transcription regulation, nuclear storage, growth control, and apoptosis. In addition to the PML protein, these bodies contain many other proteins.

In relation to transcription, CBs associate preferentially with specific genomic loci (snRNA genes and histone gene clusters). SFCs are not primary sites of pre-mRNA splicing, rather they provide storage and assembly of the splicing machinery. However, most active genes are found at the periphery of SFCs, which suggests that all RNA-processing machines are linked physically to the transcription process. Similar to what is observed in nucleoli and 'transcription factories', SFCs, CBs and PML bodies behave in a dynamic way. SCFs undergo continuous changes in shape, suggesting high internal dynamics [117]. CBs and PML bodies move in an ATP-dependent manner [118, 119]. A rapid and continuous flux of proteins between nuclear bodies and the nucleoplasm has also been demonstrated [118, 120].

1.3.3 Chromatin compartments

The interior of the nucleus is clearly not homogeneous, when examined with either light or electron microscopy. Two types of chromatin can be distinguished: one that remains condensed throughout the cell cycle and another chromatin fraction, which decondenses as the cell progresses from metaphase to interphase, referred to as heterochromatin and euchromatin, respectively (reviewed in [121]). These two types of chromatin can be considered nuclear compartments, each with its distinct properties (table 1.3.1) and resident proteins.

Euchromatin

The differences between euchromatin and heterochromatin are detectable at the DNA sequence level. Characteristically, euchromatin is largely composed of unique sequences that contain a relatively high to variable density of genes. In addition, it is relatively GC-rich mainly because CpG islands are found at the 5' end of many mammalian genes (approx 60%), including most house keeping genes and 40% of genes that have a tissue-restricted pattern of expression [122]. In CpG islands the dinucleotide CpG occurs with a 10-fold higher frequency than in the rest of the genome. In addition, the cytosine residues of CpG islands are unmethylated, while they can be methylated outside CpG islands ([123] and references therein). DNA methylation is associated with heterochromatin and transcriptional repression. Furthermore, CpG islands contain very little histone H1 (a histone involved in chromatin condensation) [124] and they are initiation sites for both transcription and replication of DNA [125]. Euchromatic domains have irregular nucleosomal arrays due to the presence of DNase I HSs. Euchromatic domains usually contain hyperacetylated histones and methylation of histones can be found in the form of H3-meK4 (see also figure 1.2.4) [124, 126]. In conclusion, euchromatin forms a decondensed or open chromatin compartment with many biochemical 'marks' associated with active transcription and high gene density.

Table 1.3.1 Distinctions between euchromatic and heterochromatic compartments.

Feature	Euchromatin	Heterochromatin
Staining/packaging in interphase	Dispersed/decondensed	Appears condensed, heteropycnotic
DNA sequence	Predominantly unique	Predominantly repetitive (satellites, derivatives of viruses, transposons, etc.)
Presence of genes	High/variable density	Low density
Meiotic (reciprocal) recombination	Normal frequency	Low frequency
Replication timing	Throughout S phase	Late S phase
Chromatin structure	DNase I HSs, irregular nucleosomes, accessible to nucleases	Loss of DNase I HSs, regular nucleosomal array, less accessible to nucleases
Activity state		
Euchromatic genes	Genes inducible	Genes silenced (variegated)
Heterochromatic genes	Genes silenced (variegated)	Genes inducible
Characteristic modifications	Histone hyperacetylation Histone H3-meK4 present Cytosine (CpG) hypomethylation	Histone hypomethylation Histone H3-meK9 present Cytosine (CpG) hypermethylation

Adapted from [126].

Heterochromatin

Heterochromatin is a more condensed form of chromatin and can be classified into two populations: constitutive and facultative heterochromatin. The former remains visibly condensed during interphase and is mainly located at pericentromeric chromosomal domains in mammals. Facultative heterochromatin is conceptually defined as developmentally regulated heterochromatin capable of undergoing a transition between the heterochromatic and euchromatic states [121]. This transition can occur during development and in differentiation pathways when a particular subset of the active genes is silenced. It can also occur in PEV when normally active chromatin can be silenced when placed next to heterochromatin. It is presently not clear whether facultative heterochromatin shares all of the structural features of constitutive heterochromatin.

Constitutive heterochromatin contains many repetitive sequences, particularly at the centromeres with long stretches of satellite repeats. Presence of long tandem arrays of a sequence motif will have the effect of generating relatively uniform sequence content over a large region. Recent data obtained in yeast indicate these may form heterochromatin as a result of an RNAi-type repression [127-129]. Three well-characterized covalent modifications that ‘mark’ heterochromatin have been described: hypoacetylation of histone lysine residues, histone methylation in the form H3-meK9 (see also figure 1.2.4), and DNA methylation at the cytosine residues. These modifications are associated with transcriptional repression and may be involved in packaging heterochromatin and stable silencing of euchromatic genes [53, 126]. Large-scale heterochromatinization of the genome is observed in terminally differentiated cells, such as plasma cells, and glial cells. In addition, the inactive X chromosome in mammals forms a densely stained structure (the Barr body) in which the core histones are hypoacetylated and H3 K9 is methylated [121]. In addition to transcriptional silencing, heterochromatin may also be required for chromosome stability during cell divisions. In *S. pombe* the heterochromatic protein Swi6 is required for efficient cohesion between sister centromeres [130]. Mice with a null mutation of *Suv39h*, encoding an H3-meK9 specific HMTase, no longer bind heterochromatin protein 1 (HP1) at the centromere and show increased chromosome instability with a higher frequency of non-homologous pairing in male meiosis [131].

Chromatin dynamics

In situ data demonstrated that repositioning of genes, proteins and *cis*-regulatory elements relative to a particular chromatin compartment occurs in response to alterations of transcriptional activity and mitogenic activity. The insertion of heterochromatin into the euchromatic *brown* gene (*Drosophila*) results in PEV and aberrant association of the gene and its homologous copy with the heterochromatin compartment in interphase nuclei [132]. In the nucleus of B lymphocytes Ikaros proteins (required for normal T, B, and NK cell development) localize to discrete centromeric heterochromatin containing foci and associate with transcriptionally inactive but not transcriptionally active genes. In addition, in developing T cells two gene loci were shown to relocate to centromeric domains following heritable gene silencing. These findings suggest that repressed genes are selectively recruited into centromeric domains [133, 134]. However, several transcriptionally repressed genes were not associated with centromeric heterochromatin in mature resting B lymphocytes. Mitogenic activation of these cells resulted in association with centromeric heterochromatin. This indicates that the spatial organization of genes in cycling and non-cycling lymphocytes may be different and suggests that locus repositioning to the heterochromatin compartment may be a feature of heritable gene silencing [134].

The dynamics of the chromatin template were monitored in living cells after the development of a Lac^O/LacR tagging tool. This tool originally consisted of 256 direct repeats of the bacterial Lac^O sequence that were introduced into the genome of cell lines as large tandem arrays due to gene amplification using dihydrofolate reductase (DHFR) as a selectable marker. Next, the Lac^O array was tagged with a fusion between GFP and LacR, which is the cognate DNA binding protein of Lac^O. This allowed a live analysis of these cells using fluorescent microscopy [135]. In yeast, this revealed movements of ~0.5 μm in ~10-second intervals for the internal chromosomal loci in G_1 -phase nuclei. Chromatin movement seemed to be undirected, sensitive to ATP depletion and to changes in metabolic status in the cell. Therefore, the authors proposed that the movements reflect the action of large ATP-dependent enzymes involved in transcription or chromatin remodeling [90, 136]. The same fast short-range movement has been reported for mammalian and *Drosophila* spermatocyte nuclei [137, 138]. However, it should be noted that this type of movement in mammalian nuclei restricts a locus to only ~$1/1000^{th}$ of the nuclear volume [137], while in yeast half the radius of the nucleus can be crossed with the same type of movement (figure 1.3.6) [90, 136]. In addition, loci at nucleoli or the nuclear periphery are significantly less mobile than other, more nucleoplasmic loci. Although the difference in linear distance is merely 1.5 to 2-fold, this corresponds to a further restriction of 4 to 8-fold in the nuclear space [137]. Thus, chromatin motion appears to be confined to a limited volume of the nucleus and the rapid random walk movements are thought to reflect constrained diffusion. In *Drosophila*, an additional slower component was detected that is limited to a much larger chromosome-sized domain [138]. This is in agreement with fluorescence *in situ* hybridization analyses that have demonstrated that chromosomes occupy a finite and mutually exclusive fraction of the nuclear volume, the so-called chromosome territories (reviewed in [99]).

The effects of transcriptional activation on chromatin were also examined in living cells with the Lac^O/LacR tagging tool. A 90 Mbp array of Lac^O sequences stained with LacR-GFP as condensed foci of approximately 1 μm in diameter [135], whereas a fusion of LacR-GFP to the potent VP16 transcriptional activation domain resulted in a dramatic decondensation of these fluorescent foci [139]. A similar dramatic decondensation upon transcriptional activation was reported elsewere [140]. Decondensation was accompanied by an increase in acetylation of all four core histones, recruitment of several HATs, SWI/SNF components, and activation of transcription. Interestingly, inhibition of transcription with α-amanitin does not block the decondensation [139, 141].

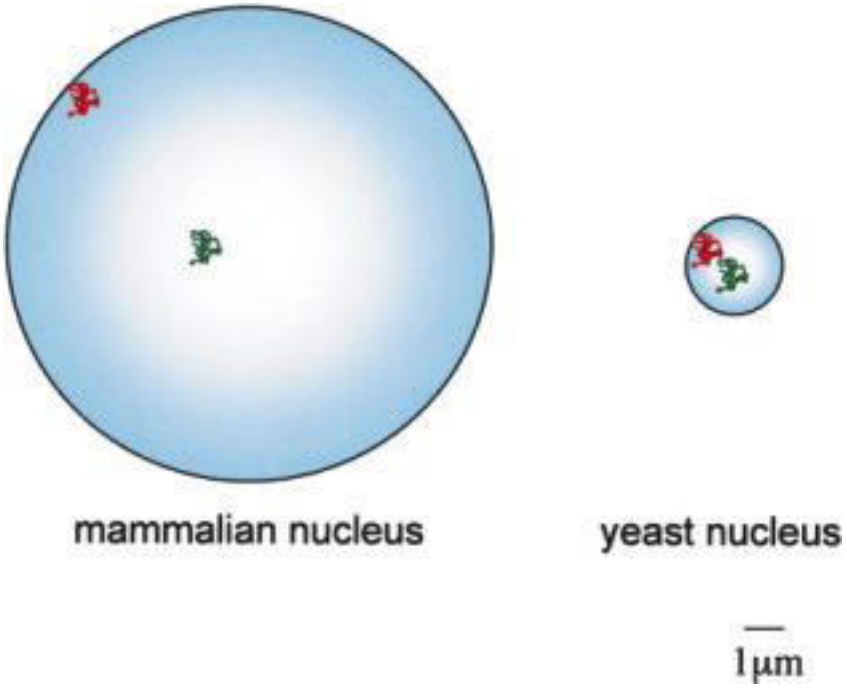


Figure 1.3.6. Constrained chromatin motion in mammalian and yeast nuclei.

The random-walk, but constrained (0.5 μm average range) chromatin motion of a locus in both mammalian and budding yeast nuclei allows it to sample only a portion of the nuclear volume. A locus positioned close to the nuclear periphery (red) may encounter a local environment that contains different *trans*-acting factors than a locus distant from the periphery (green). For instance, the nuclear periphery is associated with gene silencing and presumably a local accumulation of silencing protein (see paragraph 1.3.1). A centrally located locus is less likely to encounter these 'silencing' regions with its limited mobility. However, it should be noted that data regarding the transcriptional activity of the mammalian nuclear periphery are conflicting (paragraph 1.3.1). This figure was adapted from [273].

Dynamics of chromatin proteins and interacting proteins

The most prominent proteins of chromatin are the core histones (H2A, H2B, H3, and H4 see figure 1.2.1) that form an octamer with the DNA wrapped around this core. They are largely responsible for the repressive ground state of chromatin [1] and their dynamic behavior has recently become of great interest. Kinetic experiments revealed that the inner core of the nucleosome (containing H3 and H4) is very stable, whereas H2B on the surface of active nucleosomes exchanges both rapidly and slowly. The rapidly exchanging fraction was proposed to represent H2B in transcriptionally active chromatin [142].

In addition to the core histones, histone H1, HMG proteins and HP1 are also chromatin proteins, which are responsible for higher-order chromatin structure. Almost the entire population of histone H1 is bound to chromatin at any one time, however, it is exchanged continuously between chromatin regions. The behavior of histone H1 can be summarized in a 'stop-and-go' model. A histone H1 molecule binds chromatin for about 1 to 2 minutes, falls off, and then diffuses freely through the nucleoplasm for a short period of time until it encounters a free binding site [143]. HMG proteins also behave in a 'stop-and-go' fashion although their exchange rate is faster (in the order of seconds) and they have a significantly larger unbound fraction that moves rapidly throughout the entire nucleus [77]. Recent studies on the binding of HP1 protein in heterochromatin indicate that the individual

components within these domains are not bound stably, but in constant flux. The residence time of HP1 in euchromatin was 5 to 90 seconds and in heterochromatin between 60 and 200 seconds [144, 145]. The residence times of histone H1 and HP1 are further reduced by alterations in the histone hyperacetylation and in cells lacking the histone H3K9 methylase Suv39 [143, 144].

Transcription factors directly interact with chromatin, since they bind specific DNA sequences. Hence, their binding kinetics to specific DNA sequences may provide insight in how they activate or repress transcription. With the use of a tandem array of reporter elements and a glucocorticoid receptor labeled with GFP, targeting of the receptor to response elements in live mouse cells was observed. Photobleaching experiments provide direct evidence that the glucocorticoid receptor-GFP as well as its co-factor (GRIP-1) fused to GFP undergo rapid exchange between chromatin and the nucleoplasmic compartment [146, 147]. Similar dynamics have been observed for the intranuclear estrogen receptor (ER) and the steroid receptor co-activator 1 (SRC-1) [148]. Similarly the RUNX proteins and the AhR/ARNT transcription factor complex were shown to behave in a dynamic manner similar to that observed for the glucocorticoid and estrogen receptors [115, 149]. Thus, transcription factors appear to move very rapidly throughout the entire nucleus and their interaction with target sites in chromatin is transient. Considering the high DNA content and the large amounts of RNAs and proteins, one might intuitively expect the nucleus to be a viscous environment, however photobleaching experiments have now shown that many proteins are highly mobile within the nucleus [150]. The dynamic properties of nuclear proteins also suggest that a stochastic mechanism provides the basis of gene expression and nuclear architecture [150]. In such a dynamic environment, the fate of a chromatic domain is likely determined by a competition for binding between activators and repressors. Binding is dependent on sequence-specific elements recruiting these proteins, the residence time of each factor at its cognate binding site in chromatin, and the relative local concentrations of both activator and repressor [151].

The nuclear organization depends on dynamic interactions between components with constrained and high mobility. The rapid exchange of proteins binding to or dissociating from macromolecules and moving in or out of bodies generates a dynamic but stable nuclear organization of the larger structures. A striking feature of many proteins found in nuclear compartments is the presence of self-interacting domains (reviewed in [152]). These characteristics are supportive of a concept of self-organization in the formation of nuclear structures. In a self-organizing system, the interactions of its molecular parts determine its architectural and functional features. This may provide an elegant mechanism not only to effectively concentrate factors where they are needed to drive biochemical reactions, but also to segregate factors away from sites where they are not required [152]. This brings up the question whether *cis*-regulatory DNA elements contribute to the nuclear organization because these elements bind a multitude of factors. In the nucleolus, active rRNA genes and presumably their *cis*-regulatory DNA elements are restricted to the periphery of the FCs and CBs associate preferentially with specific genomic loci. The AhR/ARNT complex binds to specific *cis*-regulatory DNA elements of numerous target genes and accumulates in multiple bright foci that co-localize with active transcription sites. This suggests that gene sequences might provide nucleation sites for the formation of nuclear structures [152].

1.4 Globins

The globin genes have a long history as a model system to study gene regulation. One compelling reason is that mutations in or deletions of globin genes are responsible for the most widespread genetic disorders, thalassemia and sickle cell anemia [153]. The products of the globin genes, globin mRNA and hemoglobin protein (see box: hemoglobin) occur in large amounts in erythroid cells and can be isolated easily. This contributed to the fact that globin genes and disease were among the first to be studied by biochemical and molecular biological methods. In the context of transcriptional regulation, the globin genes are an important system to study tissue-specific and developmentally-regulated transcription. Over the past, the globin genes have played a major role in identifying *cis*-regulatory DNA elements that contribute to the proper regulation of transcription.

Box: hemoglobin
Hemoglobin is a heterotetramer that contains two α-globin peptide chains and two β-globin peptide chains. Each globin chain binds one heme-group that binds oxygen in a reversible manner. Hemoglobin is abundant in erythrocytes that circulate in the blood stream and functions as a carrier of oxygen from the lungs to respiring tissues. Conversely, it can transport carbondioxide from the respiring tissues to the lungs.

1.4.1 The human and mouse β-globin loci

The globin genes encode peptide chains that form a heterotetramer called hemoglobin. In vertebrates, two types of globin genes exists due to gene duplications [154], the α-like globin genes and the β-like globin genes. This introduction is focussed on the β-like globin genes, since the research data in this thesis have been obtained from the human and mouse β-globin loci (see chapters 2, 3, and 4).

The β-like globin genes encode relatively small proteins of 146 amino acids. All genes have three coding regions (exons) separated by two intervening sequences (introns) and are immediately flanked by regulatory elements *in cis*. These elements are sufficient and necessary for tissue-specific expression and developmental timing of the individual genes. They include promoters, enhancers and silencers; the latter are counterparts of enhancers and enable the silencing of genes linked *in cis* in a distance and orientation independent manner. Most of the data investigating there functional activities are based on the human globin genes often inserted as transgenes integrated in the mouse genome.

The globin promoters can be detected as DNase I HSs in the chromatin of erythroid cells. The predominant sites in definitive erythroid cells are located at the 5' sides of the globin genes within 200 bp of the respective transcriptional start site. In human fetal livers, the HSs of the fetal and adult genes can be identified, while in bone marrow only the sites near the adult genes are present [8]. Thus, the DNase I hypersensitive state of globin promoters is controlled developmentally. All globin promoters contain similar but not identical TATA, CAAT and CAC box sequences (figure 1.4.1). The LCR strongly stimulates expression from a promoter possessing only a TATA box, however, the combination of the LCR and the CAC/CAAT elements is necessary for erythroid-specific transcription [155]. Like the β-globin promoters, the enhancers are also detectable as DNase I HSs [8]. They are located in close proximity or even within the β-globin like genes, 3' of the $^{A}\gamma$ and β-globin genes as well as in the second intron of the β-gene [13, 14, 156]. Deletion of the 3' β-globin enhancer showed decreased transcription of the β-globin gene [157]. Silencer elements are present at various positions in the β-globin locus, 5' of ε and the γ genes [158, 159].

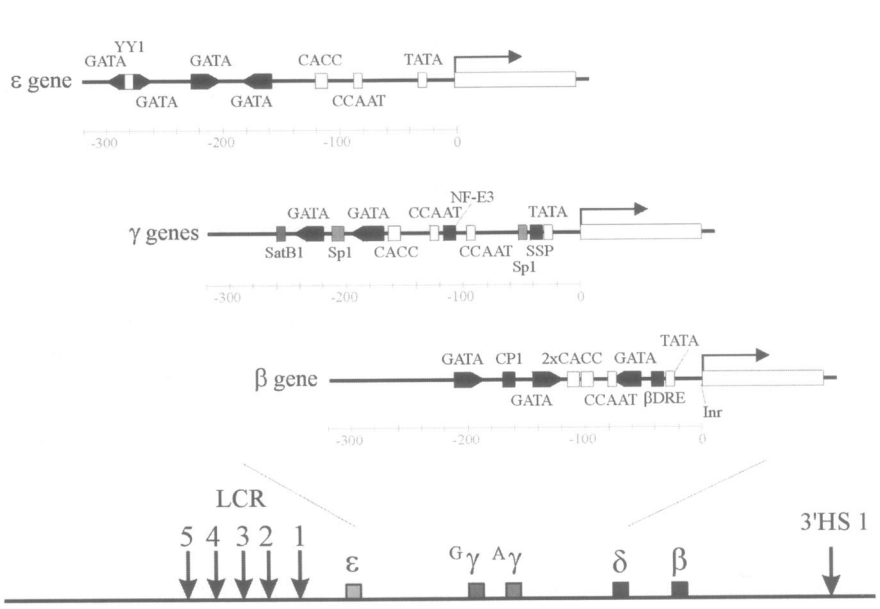


Figure 1.4.1. The promoter regions of the human ε-, γ-, and β-globin genes.

Shown are the locations of conserved *trans*-acting factor binding sites. The type of *trans*-acting factor that associates with a particular type of binding site can differ per promoter.

Genomic organization

Humans and mice have a similar, but not identical genomic organization of their β-globin loci (figure 1.4.2). The *cis*-regulatory DNA elements and genes are highly conserved between the two loci [160-162]. Both have approximately a 40% GC-content and almost an identical percentage of repetitive DNA (~30%) [162]. The human locus contains , they are arranged starting at the 5' side with ε, ^{G}γ, ^{A}γ, δ, and β to the 3' side. In the mouse, a similar arrangement can be found of its four genes, 5'- εy-βh1-β^{maj}-β^{min}-3'. Regions of synteny containing olfactory receptor (OR) genes are flanking the locus [161, 162]. Both loci have erythroid-specific distally located *cis*-regulatory DNA elements, including an LCR, 3'HS1 [163], and upstream located HSs. In mouse, these sites are located 85/84 kb (5' HS-84/-85) and 62/60 kb (5' HS-60/-62) upstream of εy [39, 164]. In humans, a homolog of the mouse 5' HS-62 was found at 110kb and an additional site at 107 kb upstream of ε (5' HS-107/-110) [39].

The first indications of the existence of a distally located *cis*-regulatory DNA element came from a Dutch family carrying a thalassemia deletion that did not express any β-like globin genes from the affected chromosome. In this deletion, the γ and δ genes were absent but the β-globin gene was still present toghether with its proximal *cis*-regulatory DNA elements (figure 1.4.2). The unaffected chromosome did express globin genes, hence it was concluded that a deletion far from the β-globin gene resulted in the suppression of its activity rather than a defect in *trans*-acting environment [22, 165]. It later transpired that the deletion included a series of DNase I HSs located upstream of the globin genes [163, 166]. In transgenic mice, linkage of these sites to a β-globin gene resulted in tissue-specific, position-independent and copy-number dependent expression [16], while this was not the case for globin transgenes without an LCR [7, 13, 14]. The mouse LCR was identified on basis of sequence homology between the human and mouse loci [167]. An analysis of the human LCR chromatin in transfected mouse erythroleukemia (MEL) cells suggested that the formation of the HSs

precedes β-globin transcription [168]. However, the LCR needs to be linked to an active promoter to stay hypersensitive [169-171]. HS formation requires binding sites for multiple transcription factors, including GATA-1, NF-E2, EKLF, and Sp1 [172-174]. Each HS holds an array of binding sites for transcription factors with similarities among the individual sites (figure 1.4.3) [172, 175, 176] although their functional properties are not equivalent.

In humans, 5' HS1 directly to a transgene does not result in expression [177], but contributes in context of the complete locus [28]. 5' HS2 shows classical enhancer properties [178] and contributes equally to overall expression of the locus throughout development [177]. The enhancer property is mediated by a tandem NF-E2 binding site that is not found as a tandem array in any of the other HSs [179]. Only multi-copy integration of 5' HS2 linked transgenes is sufficient to drive position-independent and copy-number dependent expression [15]. HS3 is the most active site during the embryonic period, and the only site capable of high-level expression of the gamma genes during fetal hematopoiesis [177]. It can control transcription of single copy integrations and establish/maintain DNase I HSs, which suggests a chromatin opening or remodeling activity [180]. 5' HS3 activity largely depends on EKLF [173, 181]. The highest activity of 5' HS4 is observed during the adult stage of development [177]. Insertion of the 5' HS4 element in place of 5' HS3 results in significant expression changes at every developmental stage [182]. 5' HS5 was originally identified as a constitutive HS with boundary/insulator activities [163, 183]. 5' HS5 also contains binding sequences for an enhancer-blocking transcription factor, called CTCF [184]. However, recently it was shown that 5' HS5 is erythroid cell-specific with enhancer-blocking activities only at the embryonic stages of development but not at later stages [35]. Thus, the LCR is composed of multiple HSs with separable functions. A possible *in vivo* mechanism for LCR activity was proposed based on the chromatin opening activity of 5' HS3. In this model, chromatin at the LCR is first opened by *trans*-acting factor activities at 5' HS3 and subsequently spreads throughout the LCR. Protein-protein interactions between factors bound at the individual sites were postulated to establish a 'holocomplex' [180]. This model is supported by the observation that an incomplete LCR often shows reduced levels of expression in transgenic mice and loses its properties of position-independent and copy-number dependent expression when integrated into heterochromatin [28]. In addition, 5' HS3 activity largely depends on EKLF, which interacts *in vitro* with the chromatin remodeling factor E-RC1 [63, 173].

In mice, the different properties of the individual HSs are less clear. Targeted deletions of 5' HS1, 5' HS2 and 5' HS4 reduce expression levels of all the genes similar to what was observed for the human sites [185, 186]. A targeted deletion of 5' HS3 results in a minimal reduction of embryonic genes and 30% reduction of adult genes [187], a deletion of 5' HS5 has minimal effects on expression [188]. In all cases, the deletion did not affect the formation of the remaining sites [188, 189]. This suggests that in the endogenous mouse locus no dominant or initiating site whose formation must precede the other HSs exists. The discrepancies found between human and mouse may reflect differences in the functional assays used, transgenic mice versus targeted deletions or the fact that the sites show some difference in numbers and affinities of the individual factor binding sites. Alternatively, the endogenous mouse locus may contain redundant elements that are missing from the transgenic constructs. Recently, two more 5' HSs of the LCR have been identified, 5' HS6 and 5' HS7 [161]. Targeted deletion of 5' HS6 from the endogenous mouse locus has minimal effects on expression [188] with again no effect on formation of other sites [189]. The functional activities of equivalent human elements have not yet been determined. Lastly, the functional significance of the distally located HSs (see figure 1.4.2) was until recently unknown. Novel data imply an active role of some of those elements in β-globin gene transcription (chapter 3 and 4) [190, 191].

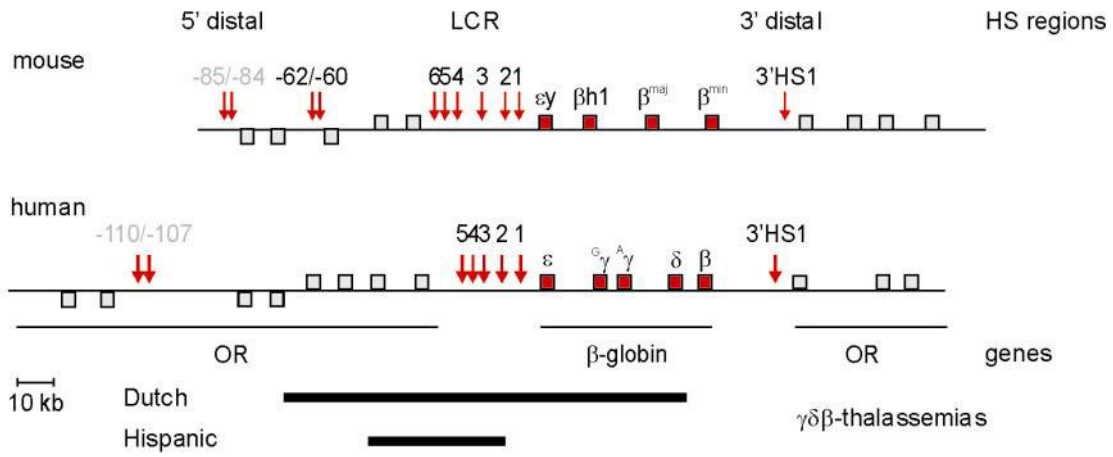


Figure 1.4.2. The mouse and human β-globin loci.

The β-globin loci of man and mouse are shown schematically. Boxes above the horizontal lines represent genes transcribing from left to right, boxes below represent genes transcribed in the opposite direction. The globin genes are indicated (red) and corresponding erythroid cell-specific DNase I HSs are represented by red arrows. The *cis*-regulatory DNA elements of the loci are positioned in distinct regions. Each region contains one or more HSs and names of each site are specified. HSs that are involved in globin gene expression are labeled in black and HSs with currently unknown function in grey. Olfactory receptor genes (OR, grey) flank and partially overlap both loci. Below the human locus the size of two deletions (Dutch and Hispanic) that are known to cause γδβ-thalassemia are indicated (see text for details).

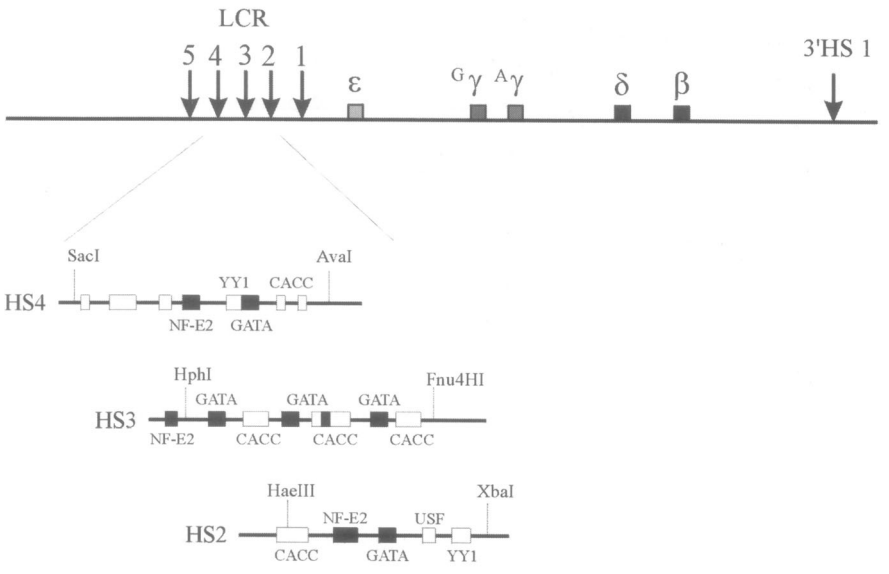


Figure 1.4.3. The core regions of 5' HS2 , 3, and 4 of the human LCR.

The binding motifs of various sequence-specific transcription factors are depicted.

Chromatin state

Recently, it was reported that a domain of DNase I sensitivity in the endogenous mouse locus extends from approximately 10 kb upstream of 5' HS-60/-62 to a few kb downstream of 3' HS1 [39]. The human locus is also more sensitive to DNaseI I than 'bulk' DNA in erythroid cells [2]. In a Hispanic thalassemia patient, a deletion was found that extended from 9.5 kb to 39 kb upstream of the ε-globin gene, leaving the distally located HSs (i.e. 5' HS-107/-110 and 3' HS1), 5' HS1 of the LCR and all the genes intact (figure 1.4.2). However, the genes were found to be inactive [23]. In addition, the remaining parts of the locus had become DNase I-resistant and HSs were not formed [192]. These findings support the transgenic data that the LCR is required for chromatin opening of the whole locus. In contrast, deleting the human endogenous LCR in an erythroid background did not show any changes in chromatin structure of the human locus [25]. The authors interpreted this as the LCR being necessary to open, but not maintain an open chromatin formation. Interestingly, a targeted deletion of the mouse LCR shows that the locus is still DNase I sensitive and that the remaining HSs were present. This was even detected after germline passage of the deletion. This suggests that elements elsewhere in the mouse locus are sufficient to establish and maintain an open chromatin conformation [26]. The transgenic experiments have clearly demonstrated that the human LCR mediates the open chromatin structure of the β-globin locus and is supported by the data on patients. However, these are in apparent contradiction with the data on the endogenous mouse locus and the deletion of the human LCR after introduction in erythroid cells. It is presently not clear why, although recent data on erythroid precursor cells show that there are intrinsic differences in chromatin modifications between the two loci. Remarkably, this difference is maintained in the human locus in transgenic mice suggesting that the primary DNA sequence rather than the organism determines the difference in epigenetic code [193].

DNA methylation at the cytosine residues is different in nonexpressing versus expressing erythroid cells. In expressing cells, the regions surrounding and including the genes showed little modification, which was not the case in neighboring DNA and the entire DNA locus in nonexpressing cells [194]. The mouse β-globin locus showed erythroid cell-specific hyperacetylation and histone H3 MeK4 at its *cis*-regulatory DNA elements. In definitve erythroid cells, four regions within the locus exhibited hyperacetylation: 5' HS-62/-20, the LCR and transcriptionally active β^{maj} and β^{min}-globin genes [39, 195]. Other regions, including the inactive genes and 3' HS1, were modestly enriched in acetylation and methylation, while the region encompassing the OR genes in between LCR and 5' HS-62/-60 showed similar levels as inactive control genes [39]. In embryonic yolk sac, which expresses εy and βH1, the LCR and both active and inactive promoters were hyperacetylated [195]. This is analogous to the developmental changes in DNase I hypersensitivity found at the promoters of the human β-globin genes [8]. Chromatin analyses of erythroid precursor cells, which express the globin genes at basal levels [196, 197], showed that the β^{maj} promoter have low to moderate acetylation levels and the mouse 5' HS3 and β^{maj} promoter are already accessible to DNase I. In fully differentiated cells, which express the genes at high levels, acetylation and accessibility were further increased [193]. In the human locus, acetylation levels were tested in MEL cell hybrids. Human chromosomes carrying either a wild-type locus or a locus with a deleted LCR showed a similar acetylation levels throughout the locus. Conversely, a chromosome carrying the Hispanic deletion was hypoacetylated [198]. The chromatin of the LCR in precursor cells is acetylated, with the exception of 5' HS3, and accessible to DNase I. Accessibility and acetylation, including 5' HS3, further increased in mature erythroid cells. Chromatin at the human γ- and β-promoters revealed a curiously high level of histone H3 acetylation in precursors and this decreased during differentiation. Conversely, the β promoter is inaccessible to DNase I in precursors, while upon differentiation accessibility increased [193]. The authors proposed that this transient chromatin acetylation may be involved in gene-specific

potentiation in precursor cells (i.e. before extensive chromatin remodeling and transcription take place in erythroid cells).

1.4.2 Hemoglobin gene switching

The β-globin genes have a well-characterized developmental transcription pattern. In all species that contain β-globin genes, the switch in globin gene expression coincides with changes in morphology of the erythroid cell, the site of erythropoiesis, and hemoglobin composition [19]. Hemoglobin gene switching happens differently between humans and mice (figure 1.4.4). In humans, a switch from embryonic (ε) to fetal (γ) globin gene expression and a second one from fetal (γ) to adult (δ/β) globin expression can be observed, whereas in mice expression only switches from embryonic (εy/βh1) to adult (βmajor/βminor). The order of globin genes reflects their expression pattern during ontogeny (figure 1.4.2) [19, 199]. The expression of the embryonic genes occurs in primitive cells, which are derived from the embryonic yolk sac. In contrast the fetal/adult genes are expressed in definitive cells, which are in origin derived from stem cells born in the AGM region of the developing embryo [200].

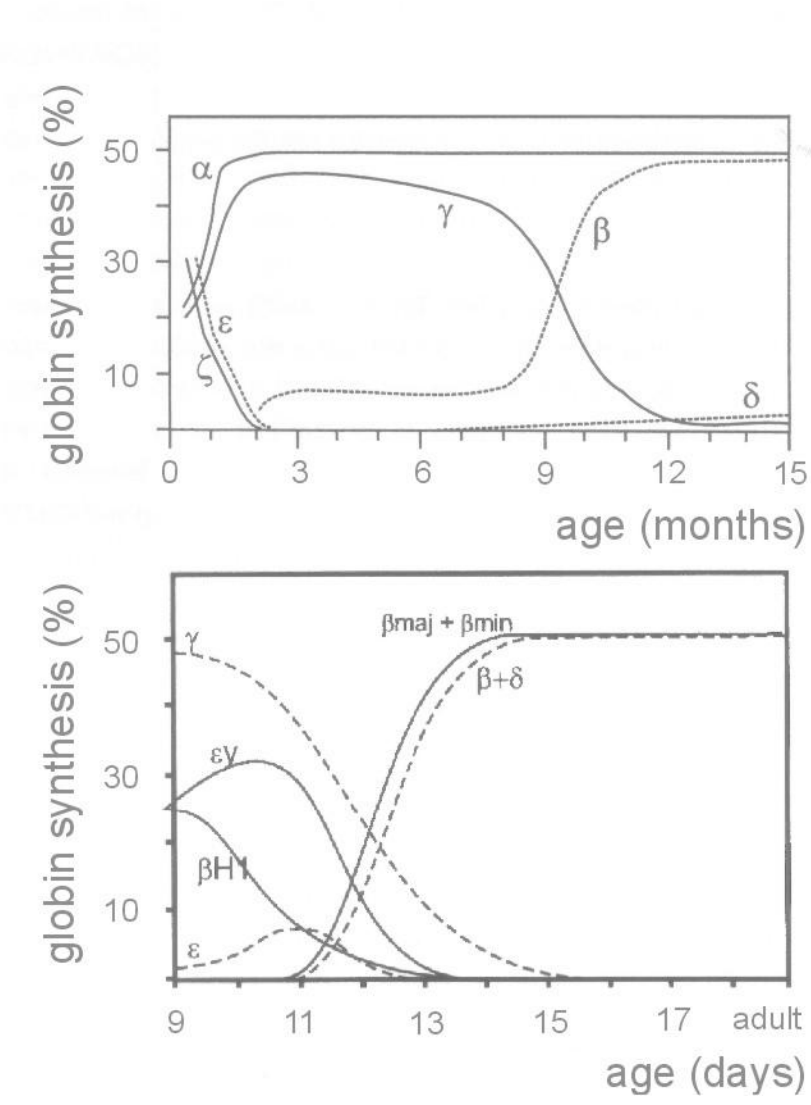


Figure 1.4.4. Developmental regulation of globin synthesis.

Top panel shows the expression levels of the individual human globin genes during development. The α- and ζ-globin genes are located in the α-globin locus. Bottom panel shows the expression levels of the individual mouse globin genes (solid lines) and human transgenes (dotted lines).

Molecular mechanism of switching

Several studies have led to the proposal of a dual molecular mechanism of hemoglobin gene switching: autonomous gene control and gene competition for direct interaction with the LCR [201, 202]. The concept of autonomous gene control is largely based on human globin genes in transgenic mice. Early experiments showed that individual human transgenes with just proximal *cis*-regulatory DNA elements (i.e. promoters, enhancers and silencers) express tissue-specifically and with the correct developmental timing, albeit in a position dependent manner [7, 14, 21, 203]. Transgenes of the ε and γ-globin genes, including their proximal *cis*-regulatory DNA elements, linked to an LCR are properly silenced at later stages of development [204, 205], silencing elements reside in their promoter regions. The 5' ε-element reduces expression of CAT-reporter construct [158], deletion or mutations lead to improper developmental timing of the ε-globin gene [206, 207]. In addition, this region is necessary for ε and γ activation at the embryonic stage [157]. Suppression of ε transcription requires a direct repeat element in its promoter that can be bound by a novel protein complex *in vitro* [208, 209]. In humans, point mutations in the promoters of the γ genes give rise to non-deletion hereditary persistence of fetal hemoglobin (HPFH), which is a genetically inherited condition that results in continued expression of the fetal γ-globin genes in adult life [19]. A targeted deletion of the endogenous mouse LCR gave reduced expression levels, but tissue-specificity and developmental timing is not affected [26]. Altogether these data demonstrate that the human ε- and γ-globin genes are accurately silenced independent of LCR functioning. The proximal regions of the embryonic/fetal genes are sufficient and necessary for autonomous silencing.

The first indications for gene competition emerged from HPFH patients that showed mutations in the γ-globin genes, which caused increased γ-gene expression in adult life. Importantly, this leads to a decrease of β-globin gene expression from the same allele while expression of the non-mutated allele is unaffected [210]. In a 'classical' enhancer assay, a small 244 bp fragment of polyoma virus DNA was found to enhance transcription of a linked rabbit β1-gene, while a complete copy of polyoma virus DNA interfered with the enhancement [211]. Expression of a β-globin transgene was found at all developmental stages when it was directly linked to the LCR [16, 212]. This is further supported by the DNase I hypersensitivity of the promoter regions of all globin genes at early stages of development although transcription of the adult genes is absent [8]. Correct developmental timing of β-globin gene expression was achieved by introducing a γ- or α-globin gene in-between the LCR and the β-gene. This correct timing was not achieved when the β-gene was the globin gene closest to the LCR [213]. This suggests competition of the genes for LCR function based on gene order or relative distance. This competition model was confirmed in the context of the complete locus. Introducing a second 'marked' β-gene (βm), located proximal to the LCR with respect to the β-gene, demonstrated a competitive advantage of the proximal βm-gene in terms of expression levels. When the relative distance was reduced between the genes and the LCR the competitive advantage of proximal βm-gene was decreased [214]. In addition, inverting the gene order of the locus, thereby altering relative distances, activated the β-gene at early stages and abolishes ε -gene expression [215]. At the single cell level, *in situ* hybridization of pre-mRNA transcripts supported the competition mechanism by demonstrating that the β-globin genes are alternately transcribed, i.e. only one gene is transcribed at any given moment [111, 216]. A similar alternating transcription was observed for the mouse β-globin genes [217]. Furthermore, the actual switching between γ and β-gene transcription is a continuous dynamic process (a flip-flop mechanism) in which both genes can be alternately transcribed until the γ-genes are permanently silenced [111]. Recently, a structural basis for this competitive interaction was revealed by showing *in vivo* that actively transcribed mouse adult globin genes physically contact the LCR while the intervening DNA bearing inactive embryonic genes is looped out (see Chapter 3) [190,

218]. Moreover, the LCR most likely operates as a single unit, or holocomplex, in these LCR-gene interactions [190].

Of course, *trans*-acting factors play a crucial role in hemoglobin gene switching. As such, the *trans*-acting environment can favor activation of specific genes regardless of gene order [202]. For instance, complete transgenic loci express the γ-genes much higher than the ε-globin gene in mice [199], suggesting that the *trans*-acting environment in embryonic erythroid cells favors LCR-γ-gene interactions. A protein complex has been purified, called direct repeat erythroid-definitive (DRED), that binds a ε-promoter element *in vitro*, which is required for gene silencing [208]. This is a 540 kDa complex and the core consists of two nuclear orphan receptors (TR2 and TR4) that can bind to direct repeat sites in the ε and γ-promoters. DRED was suggested to act as a stage-specific negative regulator by repressing embryonic and fetal transcription in definitive erythroid cells [209]. In addition to the possible role of DRED, EKLF plays a key role in β and γ-gene competition during hemoglobin gene switching. The fetal livers of mice with a reduced EKLF concentration display a decrease in the number of transcriptionally active β-genes with a reciprocal increase in the number of transcriptionally active γ-genes, although the timing of γ globin gene silencing remains the same. When EKLF is absent, there is a further increase in the number of transcriptionally active γ-genes, while β-gene transcription is abolished [181].

1.4.3 Communication between *cis*-regulatory DNA elements in the β-globin loci

As already mentioned earlier gene loci comprise an extensive *cis*-regulatory system and all elements appear to contribute to proper transcriptional regulation. Although many enhancer-promoter combinations are possible, correct enhancer-promoter specificity is observed [4, 10, 20]. This is also the case in the β-globin loci, where the LCR does not influence the expression of the adjacent OR genes [39, 161]. Thus, *cis*-regulatory DNA elements must have mechanisms to communicate with each other in order to drive expression of the correct set of genes, in the right tissue, at the proper time in development.

Hypothetical mechanisms

In the past, several hypothetical mechanisms have been proposed that may explain the communication between the β-globin genes and the LCR (figure 1.4.5). The first model is based on direct interactions between enhancer and promoter sequences. Proteins bind at widely separated enhancers and promoters in order to regulate gene expression and the intervening DNA loops or bends to allow protein-protein interactions [219]. This mechanism is generally known as the looping model. To ensure activation of the β-globin loci by looping a mass-action mechanism has been proposed (reviewed in [220]). In such a system, a transcription factor occasionally binds its cognate binding-site in inactive chromatin when such a binding site is (temporarily) available. Subsequently, two events may occur either the factor dissociates or, within small window-of-opportunity, a second binding site becomes available as the result of chromatin alterations and a second factor binds at such a neighboring site. This process would repeat itself and be the start of inducing a chain-reaction resulting in activation. The significance of the LCR in this mechanism is that it would contain a sufficiently large number of binding sites for factors to ensure activation would usually take place [220, 221]. This simple model incorporates the properties of nucleosome sliding/remodelling [222] and shows similarities with the stop-and-go properties of *trans*-acting factors exchanging between chromatin and the nucleoplasm [150]. Looping to other sites in a locus would lead to chromatin modifications in a larger domain.

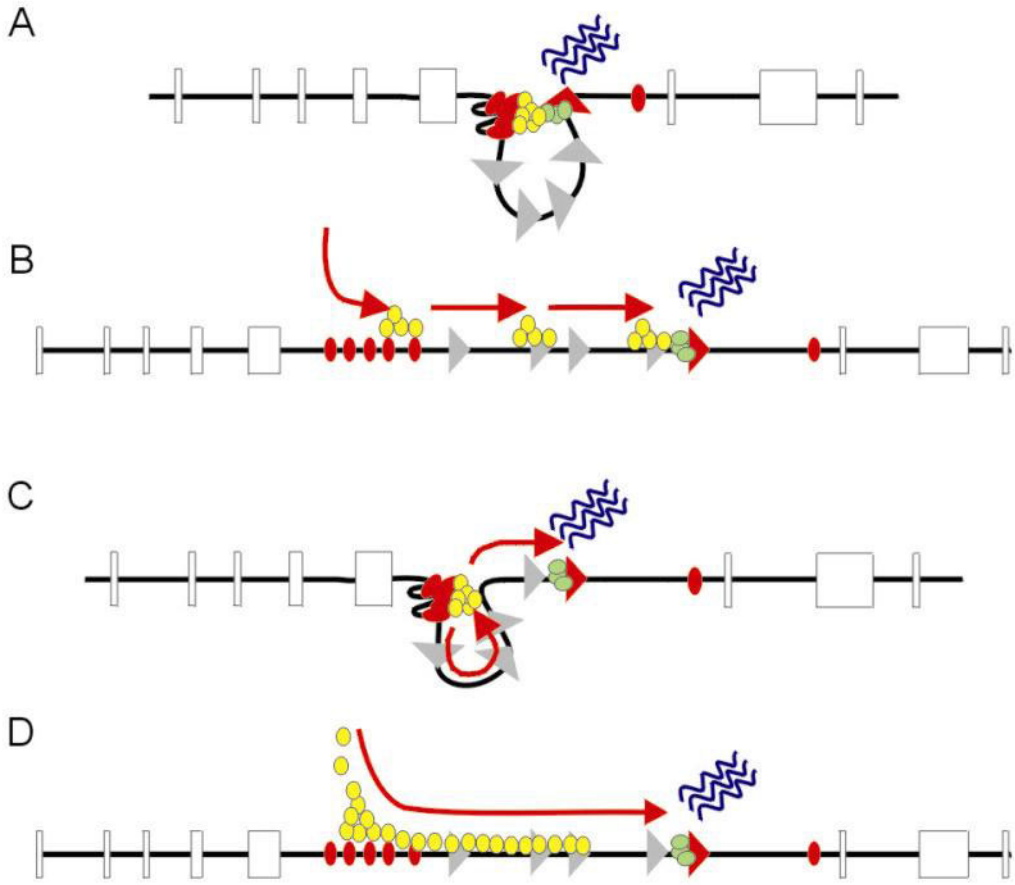


Figure 1.4.5. Hypothetical mechanisms of LCR-globin gene promoter communication.

Globin genes are depicted as triangles with the transcribed gene in red and the nontranscribed genes in grey. DNase I HSs are indicated as red ovals. LCR bound *trans*-acting factors are shown as yellow ovals and promoter bound factors (e.g. general transcription machinery) as green ovals. White boxes are olfactory receptor (OR) genes. The looping model states that DNA-bound protein-protein interactions result in direct contacts between a LCR 'holocomplex' and the promoter of a gene (A). The contacts are established by random-collision and are essential for transcription. As a consequence of the LCR-gene interactions the intervening chromatin template is looped out. The tracking model assumes that factors nucleate at the LCR, which is followed by tracking of those factors to the promoter of the gene (B). The red arrows indicate the movement of the complexes, which depends on the intervening chromatin template. Once a tracking complex reaches the promoter of a potentially active gene then transcription takes place. The facilitated tracking model unites looping and tracking (C). Complexes nucleate and stay bound at the LCR. This LCR-protein complex tracks through the intervening chromatin template towards a potentially active gene (red arrows indicate movement). Upon direct contact between the LCR-protein complex and the promoter transcription occurs. The linking model presumes that LCR-promoter communication is facilitated by transmission of a nucleoprotein structure along the intervening chromatin template (D). This structure originates from the LCR and gradually polymerizes in the direction of the promoter of a gene (red arrow). Transcription happens when the nucleoprotein structure reaches a promoter of a potentially active gene. Note that the looping model, unlike the other models, does not depend on the presence of the intervening chromatin template for transcription, rather transcription depends on random-collision between LCR-bound complexes and promoter-bound complexes.

A second model assumes that enhancer-promoter communication involves proteins that move or track from the enhancer to the promoter using the intervening DNA. This so-called tracking model was supported by the observation that intergenic transcripts were found originating from the 5' HS2 enhancer of the LCR [223]. Similar intergenic transcripts were observed in bacteriophage T4 late genes [224]. This was explained as the basal transcription machinery tracking from enhancer to promoter and consequently long strands of RNA through the intervening DNA are produced.

A third model combines the features of looping and tracking and is therefore referred to as facilitated tracking. As such, an enhancer bound complex 'tracks' through small steps (and perhaps tracking) along the intervening chromatin until it encounters the cognate promoter, at which a stable loop is formed [4]. This model was based on the assumption that the likelihood of forming small loops seems to be higher than the chance of forming large loops, which would occur in the looping model (see box: chromatin contacts). Recent data provided supportive evidence for this model [48]. Chromatin immunoprecipitation (ChIP) analyses at the HNF-4α locus during CaCo-2 cell differentiation demonstrated an independent initial assembly of factors at the enhancer and promoter elements. Subsequently, enhancer-bound factors were present at the intervening DNA, but not upstream of the enhancer, whereas promoter-bound factors stayed fixed at the promoter. Transcription initiation was observed only after the enhancer-bound factors were also detected at the promoter. Detection of the enhancer-bound complex coincided with a unidirectional spreading of histone acetylation [48].

Finally, a fourth model supposes that enhancer-promoter communication needs to be facilitated by transmission of specific nucleoprotein structures from enhancer to promoter; this is referred to as the linking model [225, 226]. Linking was proposed to overcome the prohibitive effects of increasing distance. In addition, spreading of nucleoprotein alters (opens) chromatin structure of intervening DNA and boundary elements determine amount of spreading, resulting in distinct chromosome domains. This model was largely based on the Chip protein in *Drosophila*, which has been suggested to act as a general facilitator of enhancer-promoter communication [225]. Initially it was proposed that enhancer and promoter communicate without directly contacting each other, however, recent discoveries by others necessitated a revision that combined linking and looping in one model [227].

Box: chromatin contacts

The efficiency of bringing chromatin bound-protein factors together by looping depends on the flexibility and conformation of the intervening chromatin template. The stiffness of a chromatin fiber is thought to be determined by interactions between nucleosomes. Thus, if the attractive forces between nucleosomes are relatively weak, the fiber will be soft and bend easily. Based on fluorescence *in situ* hybridization (FISH) data and computational analysis the separation distance of any two given sites that is optimal for interaction can be calculated. For the observed average chromatin fiber optimal site separation is 40 kb with a local concentration of the sites of $8x10^{-9}$ M. However, if the flexibility is increased ~5-fold, presumably due to an open chromatin configuration of the fiber, than the optimal site separation decreases to 10 kb, while the local concentration increases ~100-fold. Local concentrations drop for site separation distances larger or smaller than the optimal site separation, thus the likelihood of interaction will go down. In metaphase chromosomes interactions between distant sites become rather unlikely except for sites that are brought into proximity by the condensation process itself. This effect might also occur within interphase chromatin, for instance as a result of higher-order chromatin folding.
(For details see: [228])

Direct interactions between *cis*-regulatory DNA elements and loop formation

Recently, two independent novel assays demonstrated that the *cis*-regulatory DNA elements of the mouse and human β-globin loci are spatially close and may communicate by interacting with each other in the nucleus of erythroid cells, but not in non-erythroid cells [190, 191, 218]. One of those novel techniques involved RNA TRAP in which HRP-labeled probes are targeted to nascent RNA transcripts and HRP-catalyzed biotin deposition on nearby chromatin is quantified. Clearly, a peak of biotin deposition at 5' HS2 in addition to a peak at the transcribed gene was observed [218]. This finding indicates that the classical enhancer element 5' HS2 of the LCR is in close physical proximity to an actively transcribed β-globin gene *in vivo* [218]. RNA TRAP is dependent on nascent transcripts coming from actively transcribed genes, which can be both an advantage and a disadvantage. The advantage is that nonexpressing cells do not contribute to the measured values, meaning that interactions within the transcriptionally active locus can be picked up even in mixed cell populations. However, to determine whether interactions are specific for the transcriptionally active status of a locus one needs to know its silent conformation, which is an important control that cannot be checked by RNA TRAP. Moreover, it may be difficult to obtain satisfying results by RNA TRAP for genes with moderate transcription rates. The globin genes, which were used to develop RNA TRAP, are transcribed very efficiently, and these genes are exceptionally good targets for visualizing nascent transcripts. The other approach was initially developed in yeast to detect the frequency of interaction between any two genomic sites, which can be used to analyze the overall spatial organization of chromosomes. This methodology is referred to as Chromosome Conformation Capture (3C) [229]. In brief, this method uses formaldehyde to induce protein-protein and protein-DNA cross-links, cleavage of the cross-linked chromatin by a restriction enzyme, intramolecular ligation, and quantitation of ligation products (see chapter 3 and 4) [190, 191, 229]. In potential, 3C can measure cross-linking frequencies at many positions in the locus (e.g. active gene, LCR, intergenic sequence) with the position of restriction sites as the only limitation. Additionally, it can measure cross-linking frequencies in any tissue either expressing a particular gene or repressing it. However, interactions between distal DNA elements are thought to be dynamic, while the measurements represent steady-state average levels. Thus, short-lived but important interactions may give much lower cross-linking frequencies than more long-lived interactions. Furthermore, interpretation of 3C data in a strictly quantitative manner may also be influenced by several other observations. In the first place, we find that additional parameters, e.g. the fragment size, notably affect the cross-linking efficiency. Secondly, changes in fixation conditions differentially affect cross-linking frequencies. Thirdly, it is impossible to say if chromatin is folded as a 10nm fibre, a 30nm fibre or that it adopts yet another conformation. Moreover, chromatin folding is certainly not uniform along the entire β-globin locus, given the presence of DNase I HSs [230]. Therefore, we currently prefer to describe data obtained by 3C technology in a qualitative manner, rather than interpreting the relationship between cross-linking frequency and genomic site separation in terms of real distances. Measured interactions become particularly meaningful if they can be correlated to a phenotype, e.g. if they occur only in a transcriptionally active locus.