From Single Gene to Whole Genome Studies of Human Transcription Regulation

ALVARO RADA-IGLESIAS
Dissertation presented at Uppsala University to be publicly examined in Rudbeck Hall, Rudbeck Laboratory, Dag Hammerskjöld, 20, Uppsala, Friday, March 2, 2007 at 13:15 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English.

Abstract

Transcriptional regulation largely determines which proteins and the protein levels that are found in a cell, and this is crucial in development, differentiation and responses to environmental stimuli. The major effectors of transcriptional regulation are a group of proteins known as transcription factors, which importance is supported by their frequent involvement in mendelian and complex diseases.

In paper I, we attempted to establish the importance of DNA sequence variation in transcriptional control, by analyzing the potential functionality of polymorphic short repetitive elements as cis-regulatory elements. However, the relevance of this study was constrained by the limited number of analyzed sequences and the in vitro nature of the experiments. To overcome these limitations, (paper II) we optimized an in vivo large-scale technology named ChIP-chip, which couples chromatin immunoprecipitation and microarray hybridization. We successfully identified the binding profiles of metabolic-disease associated transcription factors in 1% of the human genome, using a liver cellular model, and inferred the binding sites at base pair resolution.

Another important characteristic of transcriptional regulation is its plasticity, which allows adjusting the cellular transcriptome to cellular and environmental stimuli. In paper III, we investigated such plasticity by treating HepG2 cells with butyrate, a histone deacetylase inhibitor (HDACi) and interrogating the changes in histone H3 and H4 acetylation levels in 1% of the genome. Observation of frequent deacetylation around transcription start sites and hyperacetylation at the nuclear periphery challenges pre-assumed HDACi mechanisms of action.

Finally, in paper IV we extended the DNA binding profiles of the medically relevant transcription factors, USF1 and USF2, and H3 acetylation to the whole non-repetitive fraction of the human genome. Using motif finding tools and chromatin profiling, we uncovered the major determinants of USF-DNA interactions. Furthermore, USFs and H3ac were clearly localized around transcription start sites, frequently in the context of bidirectional promoters.

Keywords: transcription, ChIP-chip, genome-wide, transcription factors, in vivo

Alvaro Rada-Iglesias, Department of Genetics and Pathology, Rudbecklaboratoriet, Uppsala University, SE-75185 Uppsala, Sweden

© Alvaro Rada-Iglesias 2007

ISSN 1651-6206
urn:nbn:se:uu:diva-7463 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-7463)
That one that insists in hitting the moon with a stone will not succeed, but at least will learn how to use the sling.

Arab proverb

In memory of my grandfather, Candi.
List of publications

This work is based on the following publications and manuscript:


List of additional publications

I. The ENCODE project consortium, The ENCODE pilot project: identification and analysis of functional elements in 1% of the human genome. Submitted to Nature.


III. Adam Ameur, Stefan Enroth, Alvaro-Rada Iglesias, Ola Wallerman, Claes Wadelius, Jan Komorowski, Base pair localization of in vivo transcription factor binding from low resolution ChIP-chip data. Submitted to BMC Bioinformatics.

IV. Michael Draminski, Alvaro Rada-Iglesias, Stefan Enroth, Claes Wadelius, Jacek Koronacki, Jan Komorowski, Monte Carlo feature selection for supervised classification. Submitted to Bioinformatics.


Contents

Introduction ................................................................................................................................. 9
RNA Polymerase II machinery ................................................................................................. 10
Transcription Factors .............................................................................................................. 12
Histone Code ............................................................................................................................ 14
Recruitment is the key: cooperativity and synergism ............................................................ 16
Transcription and nuclear organization .................................................................................. 19
DNA sequence and DNA conformation .................................................................................. 20

Methods .................................................................................................................................. 27
Electrophoretic mobility shift assay (EMSA) ........................................................................ 27
Chromatin immunoprecipitation (ChIP) .................................................................................. 28
ChIP-chip (Chromatin immunoprecipitation and microarray hybridization) ...................... 29

Aims ........................................................................................................................................ 32
General Aim ............................................................................................................................. 32
Specific Aims ............................................................................................................................. 32

Paper I .................................................................................................................................... 33
Paper II .................................................................................................................................... 35
Paper III ................................................................................................................................... 37
Paper IV ................................................................................................................................... 39

Acknowledgements .................................................................................................................. 41

References ................................................................................................................................. 44
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACH</td>
<td>active chromatin hub</td>
</tr>
<tr>
<td>BRE</td>
<td>TFIIB recognition element</td>
</tr>
<tr>
<td>ChIP</td>
<td>chromatin immunoprecipitation</td>
</tr>
<tr>
<td>CT</td>
<td>chromosome territory</td>
</tr>
<tr>
<td>CTD</td>
<td>C-terminal domain of RNA polII</td>
</tr>
<tr>
<td>DPE</td>
<td>downstream promoter element</td>
</tr>
<tr>
<td>EMSA</td>
<td>electrophoretic mobility shift assay</td>
</tr>
<tr>
<td>FOXA</td>
<td>forkhead box A</td>
</tr>
<tr>
<td>H3ac</td>
<td>acetylated histone 3</td>
</tr>
<tr>
<td>H4ac</td>
<td>acetylated histone 4</td>
</tr>
<tr>
<td>HAT</td>
<td>histone acetyl transferase</td>
</tr>
<tr>
<td>HDAC</td>
<td>histone deacetylase</td>
</tr>
<tr>
<td>HDACi</td>
<td>histone deacetylase inhibitor</td>
</tr>
<tr>
<td>HMT</td>
<td>histone methyl transferase</td>
</tr>
<tr>
<td>HNF-3β</td>
<td>hepatocyte nuclear factor 3 beta</td>
</tr>
<tr>
<td>HNF-4α</td>
<td>hepatocyte nuclear factor 4 alpha</td>
</tr>
<tr>
<td>IC</td>
<td>interchromatin compartment</td>
</tr>
<tr>
<td>LCR</td>
<td>locus control region</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>OMIM</td>
<td>Online Mendelian Inheritance in Man database</td>
</tr>
<tr>
<td>PIC</td>
<td>preinitiation complex</td>
</tr>
<tr>
<td>PR</td>
<td>perichromatin region</td>
</tr>
<tr>
<td>PSE</td>
<td>proximal sequence element</td>
</tr>
<tr>
<td>PTM</td>
<td>posttranslational modification</td>
</tr>
<tr>
<td>RNA PolII</td>
<td>RNA polymerase 2</td>
</tr>
<tr>
<td>rRNA</td>
<td>ribosomal RNA</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td>TBP</td>
<td>TATA box binding protein</td>
</tr>
<tr>
<td>TF</td>
<td>transcription factor</td>
</tr>
<tr>
<td>TFIs</td>
<td>general transcription factors</td>
</tr>
<tr>
<td>TFO</td>
<td>triplex formation oligonucleotide</td>
</tr>
<tr>
<td>tRNA</td>
<td>transfer RNA</td>
</tr>
<tr>
<td>TSS</td>
<td>transcription start site</td>
</tr>
<tr>
<td>USF1</td>
<td>upstream stimulatory factor 1</td>
</tr>
<tr>
<td>USF2</td>
<td>upstream stimulatory factor 2</td>
</tr>
</tbody>
</table>
Introduction

The completion of the human genome sequence and its subsequent analysis during the last few years, have resulted in numerous surprises and paradoxes. First of all, the number of protein-coding genes, was much lower than expected, with current estimations fluctuating around 20-25 000. This is similar to the number found in mouse, and not much more than the fly and the worm.

Consequently, scientific attention has been focused on deciphering processes and mechanisms that could help explain our complexity in spite of our low number of genes. Several areas of research are trying to reveal what makes us human. Firstly, it should be emphasised that the number of proteins that an organism can encode from a pool of 20-25 000 genes can largely exceed this number. This is possible by combining several mechanisms like alternative promoter usage, mRNA editing, alternative mRNA splicing, post-translational protein modifications, etc. Secondly, the traditional biological dogma based on the three-step process in which DNA is transcribed into mRNA that is then translated into a protein, might need to be re-evaluated. It seems that numerous DNA loci are transcribed into RNAs that do not generate proteins [1-4] and as a total, a large fraction of our genome is actively transcribed, with antisense transcription being a very common event [4-6]. Among these non-coding RNAs, a group of small molecules (snoRNAs, miRNAs, siRNAs, and small dsRNAs) can regulate gene expression at multiple levels [7, 8]. Furthermore, recent hypothesis suggest that RNAs may be a key component in achieving complexity in higher eukaryotes through controlling multiple, large regulatory networks [9].

Finally, even if there is an exact set of genes in common between us and other organisms (e.g. chimpanzee), it would still be possible to create enormous phenotypic differences through defining how much, for how long, when and where every single gene is expressed. The control of these four coordinates affecting the expression of a gene can be summarized under the term transcriptional regulation, which largely depends on three major components: the DNA sequence, the proteins that directly or indirectly bind to the DNA and the epigenetic modifications of DNA (i.e. cytosine methylation.
in CpG dinucleotides) and DNA-bound proteins like the histones. Recent data suggest that a fourth component, non-coding RNAs, could be also considered [10-13], but due to the still few examples known in humans and their poorly characterized mechanisms of action, the following sections will focus mainly on the other aspects of transcriptional control.

RNA Polymerase II machinery

The function and phenotype of a cell is largely the result of the proteins and protein levels that the cell possesses. The cell regulates which and how much of the proteins that are expressed, by keeping a tight control of gene transcription. Transcription is the process in which a DNA sequence is copied into a complementary RNA sequence. This is, in eukaryotic cells, carried out by RNA polymerases: RNA polI, which catalyzes the synthesis of rRNAs, RNA polII that produces the precursors of mRNAs, and RNA polIII, that synthesizes small RNAs (tRNAs, 5s rRNA). Transcription of protein-coding genes is performed by RNA polII, which is seen as a holoenzyme (Fig1) formed by a number of units that constitute its core enzymatic activity and other group of protein complexes, including the general transcription factors TFIID, TFIIB, TFIIF, TFIIE, TFIIH [14-19]. The general factors are major players in (a) the recognition of core-promoter sequences, (b) responses to regulatory TFs (activators and repressors) and (c) conformational changes essential for RNA polII activity. (a) Core-promoter recognition is essential for a proper positioning and assembly of the whole PolII machinery. The core promoter is the minimal DNA sequence needed to specify non-regulated or basal transcription, and it consists of one or more sequence elements such as the TATA box (bound by TBP-a subunit of TFIID), BRE (TFIIB recognition element), Initiator and DPE (downstream promoter element), but none of them is absolutely essential. Recent large-scale characterization of human and mouse core-promoters have demonstrated that none of these elements is very abundant and that in many cases transcription can initiate from several alternative nearby positions, with the exception of TATA-box containing genes which initiate transcription at a well-defined position [6, 20, 21]. (b) Regulatory transcription factors, in this case activators, can affect transcription by stimulating the recruitment of the RNA polII holoenzyme in core promoters, in a state generally termed the pre-initiation complex (PIC). However, they can also act in the post-recruitment step, stimulating the activity of RNA polII after it has been assembled in core-promoters. Furthermore, activators can indirectly affect PolII activity through interaction with and targeting of chromatin-modifying enzymes to core promoters. Most of these effects are transmitted from activators to the
RNA polIII machinery through a protein complex termed Mediator [22, 23]. (c) Transcriptional initiation requires phosphorylation of the C-terminal domain of RNA polII (CTD). This is accompanied by profound remodelling of 11-15 bp of DNA surrounding the TSS. After transcribing around 30 bp, the PolIII loses its contact with the core promoter and the rest of the transcriptional machinery, thereby entering the process termed transcriptional elongation, which is also tightly regulated. It seems that after transcriptional initiation, many of the general TFs remain bound to proximal promoters, forming a scaffold that accelerates subsequent rounds of transcription by facilitating the whole recruitment process [24]. Activators can also modulate the stability of such scaffolds and thus, regulate the level of transcription.

Figure 1. Schematic representation of the RNA polymerase II holoenzyme bound to a core-promoter sequence. In the upper part, several of the general transcription factors (TFIIs) are displayed, interacting with RNA PolIII and different regions of the core promoters. Among such regions there are several well-characterized regulatory elements, presented in the lower part of the picture (Adapted from [19], Reprinted, with permission, from the Annual Review of Biochemistry, Volume 72, 2003 by Annual Reviews www.annualreviews.org).

The above scenario corresponds to a classical view of transcription where the RNA PolIII machinery is recruited to the gene promoter (PIC formation)
and then slides through the coding sequence (elongation) generating a mRNA molecule. However, an alternative model proposes that the RNA PolIII machinery occupies rather fixed positions in the nucleus and the genes move and localize to such nuclear regions in order to be transcribed, where they slide through the RNA PolIII, generating mRNA [25, 26]. Those nuclear foci where the RNA PolIII machinery is found have been named transcription factories and they are rich in RNA PolIII holoenzyme and general transcription factors. Furthermore, they seem to be anchored to a nuclear sub-structure and relatively stable, resembling a functional nuclear subcompartment [27]. This is supported by most actively transcribed genes localizing to one of such factories. The number of active genes (thousands) in a given cell largely exceeds the number of transcription factories (hundreds), suggesting that a same factory can be shared by more than one gene, which in fact seems to occur often and in a non-random manner. This model can in fact accommodate the observation that most “active” genes display short-periodic bursts of transcriptional activity and long periods of inactivity, corresponding with their localization into and out from transcription factories [26, 28-30]. Finally, this model offers alternative explanations for the mechanism of action of distal regulatory elements (enhancers, locus control regions, insulators) and allows to integrate transcriptional regulation and nuclear organization in a more comprehensive manner [27, 31, 32]. In the following sections, although focusing on the classical view of transcription, certain aspects will be also explained under the perspective of the transcription factory model, as we will refer to it from now on.

Transcription Factors

Regulatory TFs control transcriptional rates by affecting the initiation of transcription and/or the level at which a gene is transcribed through direct (i.e. direct interaction with RNA PolIII machinery) and/or indirect mechanisms (i.e. recruitment of coactivators/corepressors). Regulatory TFs can be divided into activators and repressors. They are generally characterized by having DNA binding capacity which varies in its sequence specificity between different factors. The DNA sequences that are recognized by these TFs can be found on various types of regulatory elements which are classified according to their distance to TSS and their effects on transcription as promoters, enhancers, LCR, insulators, etc [33]. These different regulatory elements are typically formed by clusters of TF binding sites bound by numerous TFs, both activators and repressors. TFs are modular proteins and together with a DNA binding domain (Zinc-finger, helix-loop-helix, helix-
turn-helix, leucine zipper, etc), they possess an activator/repressor domain and domain/s for establishing protein-protein interactions [15, 34]. Among the proteins that interact with TFs, there are both co-activators and co-repressors, which interact with TFs rather than with DNA and by that mechanism regulate transcription levels. In fact, the dynamic recruitment of these co-regulators seems to be the key for changing transcriptional responses [35]. In other words, the composition of the complexes bound to promoters and/or enhancers is very important in determining the transcriptional response.

Apart from their final output in transcriptional levels (i.e. activators or repressors), TFs can also be divided based on their expression profiles. There are TFs that are ubiquitously expressed and are found in a large variety of tissues. These TFs are typically bound to proximal promoters and tend to be involved in transcription initiation, conferring basal transcriptional levels. Other TFs are differentially expressed in different cell types and developmental stages, and they preferentially bind regulatory elements located at larger distance from TSS: enhancers. These TFs confer transcriptional regulation in a time and space dependent manner. However, both types of TFs work by facilitating RNA polII recruitment and/or facilitating recruitment of coactivators.

Probably, binding to DNA is the most defining characteristic of TFs, which is largely determined by the particular DNA-binding domain that each factor possesses [34]. Most characterized TFs have a reported consensus binding sequence, that represents the DNA sequence to which the TF binds with high affinity. However, such DNA sequences although informative, are largely derived from in vitro experiments (i.e. DNA-affinity chromatography, electrophoretic mobility shift assays, DNA footprinting, etc) [15]. The in vivo interaction between TFs and DNA sequences occurs in a chromatin protein-rich context, which can result in major discrepancies between all potential binding sites and those actually being occupied by certain TF [36-39]. We can conclude that TFs and chromatin mutually influence each other, with certain TFs being able to provoke changes in chromatin organization through recruitment of histone modifying enzymes [40-43] and/or nucleosome remodelling complexes [44-47], while chromatin is one of the critical components in determining the affinity between a TF and a DNA sequence [36, 39, 48]. The importance of studying transcriptional regulation in a chromatin context is further discussed below.
Histone Code

One important, though easily forgotten, variable to be considered in transcriptional processes is that RNA polII does not face a naked DNA waiting to be transcribed, but rather a complex and compacted structure generally known as chromatin, in which DNA is folded around nucleosomes, formed by octamers of a set of proteins called histones [49, 50]. It has been shown that nucleosomes represent a barrier for transcription, but in the last years various studies have shown how this might be overcome by both ATP-mediated remodelling of nucleosomes [18, 51-54] and post-translational modifications of histones [55-60]. The histones are globular proteins with N-terminal tails, which can be post-translationally modified in several ways like phosphorylation, acetylation, methylation, ubiquitination, sumoylation, etc. Depending on the type of modification and aminoacidic residue being modified, the correlation with transcriptional activity and chromatin state (euchromatin or heterochromatin) changes. The different modifications identified so far seem to be part of an important layer of regulatory information that has been largely unknown until recent years, and that has received the name of “Histone Code” [56, 61, 62]. This code seems to be a rather complex language, with combinatorial crosstalk between different modifications. In other words, combinations (sentences) rather than individual modifications (words) are more meaningful. Furthermore, the histone code contains important epigenetic information, that can be transmitted through cell generations and maintained through out the cell cycle [63, 64]. Some components of the code have been deciphered in recent years, elucidating how certain marks are correlated with gene activity (H3K9 acetylation, H3K4 methylation, H3S10 phosphorylation), while others seem to be associated with inactive genes and heterochromatin (H3K9 methylation or H3K27 methylation) [65-70]. However, we are far from a proper understanding of this complicated and intricate language, and for instance not all histone modifications are known as recently demonstrated by the identification of several new modifications located in the globular or core part of the histones, rather than in their tails, which may result in a redefinition of the envisioned code [71].

On the other hand, recent single-nucleosome profiles of several histone modifications in yeast suggest that only a few of all possible combinations of histone modifications exist in vivo, indicating redundancy, with multiple modifications sharing the same role [72]. These and other genome-wide studies have elucidated certain patterns of histone modification clearly associated with gene activity (e.g. gradient of H3K4me3, H3K4me2 and H3K4me from TSS to 3’end of active genes [65], but they also revealed that regions around TSS seem to be depleted of nucleosomes [73, 74], both in active and inactive genes, although upon activation nucleosome occupancy can further decrease and extend to downstream coding sequences [75-77].
Another important aspect in chromatin function in general and its specific role in transcription in particular, is the existence of different histone variants. Among the best characterized histone variants, we could mention H2AX, H2AZ and H3.3 [78-80]. H2AX has been associated with DNA repair mechanisms, which are highly relevant but not in the scope of this thesis. H2AZ seems to play different roles depending on the organism considered. While in yeast it prevents heterochromatic spreading and it is associated with transcriptional activity, in mammalian cells it is preferentially found at heterochromatin foci [58, 78, 81-83]. On the other hand, H3.3, has been associated with transcriptionally active genes in higher eukaryotes, where it is mainly found close to TSS [78, 84, 85]. Interestingly, H3.3 is highly enriched in post-translational modifications (PTMs) (e.g. K9/K14 acetylation, K36/K79 methylation) associated with gene activation, suggesting that the observed enrichments for these PTM around TSS might be confined to this H3 variant [86]. Recently, new hypothesis suggesting a major role of histone variants in demarcating chromosomal territories have been proposed (i.e. H3.3-euchromatin, H3.2-facultative heterochromatin, H3.1-constitutive heterochromatin), together with histone PTMs (i.e. histone code) [84]. The importance of histone variants is further emphasized by the discovery of histone chaperones (i.e. proteins specialized in histone eviction/deposition/exchange) specific for different histone variants (e.g. H3.1 and CAF-1, H3.3 and HIRA) [79].

Histone PTMs and the other major epigenetic modification, DNA methylation, have been shown to be interconnected and different types of combinatorial cross-talks have been proposed [87, 88]. Cytosine methylation in CpG dinucleotides is the only known epigenetic modification directly affecting DNA in eukaryotes, and it has crucial roles in regulation of various cellular processes. DNA methylation is particularly abundant in constitutive heterochromatin, where it mainly occurs in repetitive sequences such as retrotransposons that are therefore kept in a silent state. In such context, recent models proposed that such silencing is firstly achieved by establishment of H3K9 methylation and HP1 recruitment by SUV39H1 that it self can interact and recruit DNA methyltransferases to further stabilize the silent chromatin [89-91]. Furthermore, in some organisms it has been demonstrated that H3K9 methylation is directed to such regions by siRNAs originating from repetitive sequences that act as nucleation centres for heterochromatin formation [92-95]. In the context of facultative heterochromatin, such dependency of DNA methylation on certain histone PTMs has also been observed [96]. The polycomb protein EZH2 represses transcription by methylating H3 at lys 27, and this PTM recruits additional polycomb proteins. Moreover, EZH2 interacts with and recruits DNA methyltransferases to polycomb repressed genes.
and it is required for DNA methylation of EZH2-target promoters [89, 97]. The crosstalk of these processes is also observed at CpG islands, which are typically found in proximity of TSS and are not methylated with some exceptions. DNA methylation generally results in transcription inhibition, either through directly impeding TFs to access DNA or indirectly through methyl-CpG-binding proteins, which recognize the methylated DNA and recruit silencing complexes. In such complexes, HDACs and HMTs are major components, which establish a typical repressed/closed chromatin state, with hypoacetylated and H3K9 methylated nucleosomes [87, 88]. It has been observed that gene silencing can be established in the absence of DNA methylation, just by the combination of certain histone PTMs, but that the stability of such silencing is largely extended by DNA methylation that confers long-term/epigenetic memory [98, 99].

Examples of crosstalk have also been reported between different histone PTMs. For example, the long-standing observation of histone hyperacetylation and H3K4 methylation at promoters of active genes, seems to depend of the recognition of methylated lysine 4 by components of HAT complexes that contain chromodomains and PHD domains [100-102]. In fact, it is becoming a recurring topic that HAT, HDAC and HMT complexes include proteins with different domains (e.g. bromo and chromodomains, PHD domains, etc) that can recognize acetyl and methyl groups [58, 103, 104]. Furthermore, nucleosome remodelling could be also interconnected with histone PTMs. ATP-dependent nucleosome remodelling alters chromatin structure to facilitate PIC formation or elongation of RNA PolII. The complexes responsible of such remodelling were traditionally believed to be recruited to promoters of active genes by sequence specific transcription factors [47], but recent reports suggest that in addition, these complexes can contain protein domains that recognize certain histone PTMs, participating in the recruitment process [105, 106] (e.g. NURF is recruited to chromatin trimethylated at lys 4 of H3 by its PHD finger domain).

Recruitment is the key: cooperativity and synergism

In summary, transcriptional regulation is achieved by the recruitment of large protein complexes that are able to directly or indirectly communicate with each other. These processes need to be specific and flexible at the same time in order to achieve the desired responses. All human cell types are able to achieve a great number of phenotypes and transcriptional programs, and even if the number of proteins devoted to transcriptional control may seem
large (5-10% of protein coding genes), it is usually observed that the same TF can be involved in different transcriptional networks. This indicates that a TF can be reutilized for different purposes, depending on its “partners” (i.e. other TF, coactivators, etc) with which it collaborates in each particular case. In other words, by combinatorial recruitment of a limited number of TF and other regulatory proteins, a great number of different transcriptional outputs can be achieved [35, 42, 53, 107-109]. The different protein-protein and protein-DNA interactions involved in these processes are proposed to follow the two major principles of cooperativity and synergism. Transcriptional control is a highly complex process, well represented by multiple transcriptional networks acting on the different developmental stages and stimuli that a certain cell encounters [109, 110]. In these circumstances, groups rather than individual genes need to be commonly regulated. This complexity might be partially explained by the mentioned principles of cooperativity and synergism, as in the following hypothetical scenario (Fig 2):

We can imagine a certain cell during the early phases of embryogenesis, when that cell starts receiving signals that will initiate its differentiation into a particular cell type. These signals can result in increased levels of a certain TF, which will be able to bind directly to a specific sequence of nucleosomal DNA (“pioneer factors”, FOXA, GATA, etc) [111-113]. By doing so, it will create marks or “holes” in the chromatin, resulting in regulatory elements being accessible across the genome. Later in the differentiation process, new signals might increase the levels of other TFs, that may require chromatin free of nucleosomes in order to recognize their target sequences. The targets of this second TF, that might be in proximity to the binding sites of the first TF, would be then accessible. The second TF, different in its structure to the first one, might posses multiple protein-protein interaction domains (“scaffold” TF), which allows it to establish interactions and recruit different proteins like coactivators, other TFs, members of the RNA polII machinery, etc. This results in the accumulation of a number of proteins in the initial chromatin marks. Some of these proteins might for instance possess HAT or HMT activity, which could acetylate/methylate the histones in the nucleosomes nearby, resulting in the recruitment of even more proteins and the establishment of a heritable epigenetic mark. This type of large protein complexes is usually known as the enhanceosome [107, 114]. The two initial TFs could not have initiated enhanceosome assembly without each other, i.e. they work cooperatively. The final effect is not just a linear or additive response depending on the number of proteins recruited; the absence of some of the proteins will completely disrupt transcription, while their incorporation will dramatically increase the response: this more than additive effect is called synergy.
This scenario could also be applied to transcriptional changes in response to internal or external stimuli (e.g. heat-shock, hormonal changes, etc) and has been represented according to the classical view of transcription. However, under the previously presented model of transcription factories, some of the mechanistic assumptions could be slightly different. The initial binding of TFs to distal regulatory elements (i.e. enhancers) could result in such DNA regions moving from a silenced territory in the nucleus to a transcription factory, due to interactions between TFs and components of the factory [25, 31, 115]. Upon full assembly of the enhanceosome, recruitment of the promoter to the transcription factory can be achieved by loops similar to those proposed on Fig 2, involving protein-protein interactions, that finally can result in the gene being transcribed. Under this model, TFs could play a major role in determining chromatin dynamics and moving genes into and out from transcription factories, through interactions with DNA regulatory ele-
ments and other proteins (i.e. coactivators, general transcription factors, etc) (Fig 3).

Figure 3. Enhancer-promoter communication in the context of transcription factories. From top left corner and clockwise are depicted four different stages of the presented model. The green square represents an enhancer or distal regulatory element, while the red (inactive) or green (active) arrowheads denote a transcription start site. Purple and red circles correspond to TFs that can interact with regulatory elements located at the enhancer or promoter regions, respectively.

Transcription and nuclear organization

Chromosome organization is not restricted to mitosis/meiosis, since in the interphase nuclei DNA seems to be organized as chromosome territories (CT), separated by DNA-free regions known as interchromatin compartments (IC), where protein-rich nuclear speckles/bodies can be observed
On the other hand, there are two major types of chromatin organization: an open transcription permissive state (euchromatin) and a closed transcriptionally silenced state (facultative and constitutive heterochromatin). In a nuclear architecture context, heterochromatin could be considered the major component of chromosome territories, while euchromatin could correspond to decondensed chromatin around CT named perichromatin region (PR), probably consisting of chromatin loops protruding from CT into IC, also known as active chromatin hubs (ACH). This model of nuclear organization is supported by observations of position effect silencing or the preferential nuclear localization of pericentromeric repeats in the nuclear periphery [25, 27, 31, 117-119]. Furthermore, it should be emphasized that chromatin is a highly dynamic structure, where DNA-DNA and DNA-protein interactions could be constantly generated and lost, resulting in localization changes [49, 116].

In terms of transcriptional regulation, the described nuclear architecture accommodates recent observations of long-range DNA interactions and the proposed transcription factory model. Recent reports have demonstrated proximity between distal DNA regions located in the same or different chromosomes [120-125]. Such proximity has been interpreted as the result of protein-protein interactions, mediating enhancer-promoter communications in the context of gene activation [126, 127] or communication between differentially methylated regions (DMRs) at imprinted regions [128-130]. Furthermore, such inter or intrachromosomal connections imply formation of chromatin loops that in the case of a gene being activated could correspond to formation of ACH, resulting in the gene moving into a transcription factory and being transcribed. However, currently the techniques used to investigate these processes (e.g RNA TRAP, 3C), can not differentiate between direct physical interaction (in cis or trans) or just physical proximity due to, for instance, colocalization into the same transcription factory.

DNA sequence and DNA conformation
Coding DNA sequences contain the encrypted information that is translated into the aminoacids that will form a protein. However, most of the genome is formed by DNA sequences that do not code for proteins and that in some cases contain other types of information: splicing signals, transcription regulatory elements, non-coding RNAs like tRNAs, rRNAs, miRNAs, etc. Many of these informative sequences serve as binding sites for proteins involved in transcriptional regulation: transcription factors (general transcription factors, activators, repressors), RNA polII holoenzyme, etc. As previously men-
tioned, regulatory sequences can be classified depending on their location and effect on transcription[19, 131]:

**Core promoters:** minimal sequence that is needed to drive basal or non-regulated transcription. This sequence includes one or more regulatory elements (TATA-box, initiator element, downstream promoter elements (DPE), TFIIB recognition elements (BRE), proximal sequence element (PSE), but none of them is absolutely essential.

**Enhancers/Silencers:** regulatory sequences that can be at great distance from transcription start sites (TSS), upstream or downstream, and that are responsible for the regulated transcription of a gene. Traditionally, part of the definition of an enhancer is its independence of sequence direction, since when cloned in one or the opposite direction from a reporter gene, its transcriptional effect does not change. Enhancers/silencers are typically a few hundred base pairs long, containing numerous binding sites for different activators/inhibitors, that through establishing multiple protein-protein interaction eventually will result in the assembly of large protein complexes called the enhanceosome or silenceosome.

**Locus control regions:** sequences similar to enhancers, but that can reside at long distances from transcription units, even on different chromosomes. These sequences typically regulate genes with highly complex expression patterns, usually those that are key components of developmental programs.

**Insulators/boundary elements:** sequences that prevent the transmission of the chromatin-structural features associated with the boundaries between repressive or active domains. When a certain enhancer is located in the intergenic region between two genes with opposite directions, the enhancer may exert its regulatory function in only one of the genes. This might be caused by an insulator located between the enhancer and the gene that does not respond to the enhancer.

All these different regulatory sequences have some aspects in common. Firstly, they are characterized by including in a limited sequence stretch, high densities of protein binding sites. Secondly, these regulatory elements are difficult to identify. This can even apply to core-promoters, since it is becoming clear that many genes can utilize multiple/alternative core-promoters, introducing another level of complexity in transcriptional control [3, 20, 132]. There are different strategies that have been used to identify these regulatory sequences. One approach is sequence analysis of regions generally located upstream from TSS, where a number of binding sites for different proteins can be identified, by comparing to previously established TF consensus binding sequences. There are programs and databases that store consensus sequences for multiple TFs that can be used for this purpose [133]. One drawback is that most of these consensus sequences are derived from experimental data obtained in vitro. Furthermore, the bioinformatically
identified sequences are usually validated also by similar *in vitro* techniques (i.e. EMSA). Another approach is derived from the observation that sequences bound by proteins and usually located near active genes, are more susceptible to degradation by DNAse I. These DNAse I hypersensitive sites usually coincide with regulatory sequences, where the chromatin is in a more relaxed/opened state [134]. A third possibility is to clone larger DNA sequences located upstream and/or downstream from TSS into a reporter gene and perform serial deletions. By comparing the expression level of the reporter gene in the different constructs harbouring deletions of different sizes, it is possible to map those sequence elements that affect the gene expression the most. Other strategies have arisen since the completion of several eukaryotic genomes, which allows for multiple species sequence comparisons. This has resulted in the identification of highly conserved non-coding elements, which in some cases are even more conserved than coding sequences and have been named as ultraconserved elements [135, 136]. Most of these sequences are expected to be functional, but this has not been verified on a large scale and it is not clear yet if they might have a common function. Preliminary studies have evaluated this by cloning some of these sequences upstream of reporter genes and some of them seem to act as enhancers [137]. Finally, large scale *in vivo* approaches have arisen during the past few years. Chromatin immunoprecipitation is performed in order to enrich the *in vivo* genomic targets of a certain protein/TF, and later they are identified either by extensive sequencing [138-140] or by microarray hybridization (ChIP-chip, genome-wide location analysis) [21, 110, 141-145]. These techniques allow the interrogation of large genomic regions or even whole genomes for identification of TF binding sites *in vivo* and the inference of consensus binding sequences based on *in vivo* data. Furthermore, with these methodologies one does not need, in principle, any prior knowledge about the targets of a TF. However, the appropriate antibody and appropriate biological material (cell line, tissue) expressing the protein of interest are required.

Recent *in vivo* genome-wide characterization of TF binding profiles using the aforementioned techniques has demonstrated that *in silico* predictions based on matches to consensus binding sequences results in high rates of both false positives and false negatives [37, 38, 146, 147]. This suggests that only a fraction of all consensus-matching sequences are occupied by a TF in a given condition or cell type *in vivo*, and that, in at least some cases, sequences differing from the consensus can be bound. Part of these discrepancies can be explained by DNA-TF interactions occurring in a chromatin context, with most DNA sequences being occupied by nucleosomes, therefore restricting their accessibility to TFs [36, 39, 48]. It has been proposed that such accessibility can be modified by establishing certain histone PTMs (histone code hypothesis) and/or ATP-dependent nucleosome remodelling.
Recent high-resolution in vivo mapping of nucleosome occupancy combined with computational modelling in yeast, have demonstrated that the genome sequence can explain up to 50% of the nucleosome positions [73, 74]. This is predicted to have an important impact in transcription processes, and for instance, it was observed that a region of 200 bp upstream of transcription start sites of protein-coding genes is typically nucleosome-free, while surrounded by several well-positioned nucleosomes. Furthermore, functional TF binding sites were more frequently in nucleosome-free contexts than non-functional sites or than expected by chance. In conclusion, the genome sequences contain non-coding information that plays important transcriptional regulatory functions, typically by serving as TF binding sites [58, 73, 74]. However, of the large number of potential binding sites for a TF, only a small fraction will be occupied in vivo. This occupancy can be controlled partially by proteins that can modify histones or remodel-nucleosomes, but it can also be affected by the DNA-sequence context which can have an impact on the nucleosome positions and consequently on the availability of the binding sites.

All the mentioned sequence elements and most likely the majority of the genomic sequence, is in the archetypical B-DNA conformation, consisting of a DNA complementary, anti-parallel double helix, with Watson-Crick base pairing. However, since the proposed model for this structure back in 1953 by Crick and Watson, several alternative DNA conformations have been described (Fig 4): triplexes, Z-DNA, bent DNA, cruciforms, G4-tetrad (tetrplexes), sticky DNA, etc [148]. These alternative DNA conformations are preferentially adopted by simple DNA repeat sequences, which, though typically found in the orthodox B-DNA structure, eventually may adopt some of these other conformations, usually involving non Watson-Crick base pairing (i.e. Hogsteen). This type of repetitive sequences and the conformations they are able to adopt, seem to have some kind of biological significance. The main evidences regarding their functionality are related to maintenance of telomeres, chromosomal rearrangements/genomic instability and transcriptional regulation.[148-151]
Polypyrimidine/polypurine sequences are the preferred sequences for triplex conformation. Although several models have been proposed, the so-called H-DNA conformation is the best characterized among triplexes, involving either pyr*pur-pyr or pur*pur-pyr, with the third strand being parallel or antiparallel respectively. Both inter- and intramolecular triplexes have been reported, and typically the triplex is accompanied by partial single-stranded DNA loops [152-154]. The biological importance of these sequences comes from several lines of evidence, and in most cases the functionality is ascribed to their capacity to adopt triplexes:
• Effects on transcription levels have been shown in several cases [155-158].
• Alternative conformations (B-DNA and triplex) and specific protein binding of the different states have been documented. Several proteins able to bind the single-stranded DNA accompanying the triplex formation have been reported [153, 154, 159-162].
• Importance in chromosome and/or chromatin organization have also been reported, with these sequences not favouring nucleosome positioning [163, 164].
• Evidence for their in vivo occurrence have also been generated [165, 166].

In the previous paragraphs, the functional relevance, in terms of transcriptional regulation, of DNA sequences and DNA conformation has been addressed separately, while most likely they are interdependent elements affecting each other. Sequence-specific TF interactions to DNA generally depends on DNA sequence, but in some cases, the DNA conformation seems to be more crucial in the affinity of such interactions than the underlying sequence [167]. At the same time, those preferred conformations are preferentially adopted by DNA sequences with certain characteristics (e.g. polypyrimidine tracts, AT-rich sequences, etc). This is well represented by the matrix-associated regions (MARs) DNA binding protein SAT1B [168]. This protein preferentially binds to AT-rich sequences with only C in one strand and G on the other. These ATC sequences have strong potential for extensive unpairing and unwinding, which are characteristics of MARs, and such structural features are the crucial factors in determining SAT1B binding. Recent data indicates that through binding to these DNA sequences/structures, SAT1B can organize chromatin into multiple small chromatin loops, all anchored to SAT1B at their base, and that such chromatin structure has important impact on the transcriptional activity of the implicated loci [169].

During the last decades, several genes involved in mendelian diseases have been identified. In many cases mutations in coding sequences that result in a deleterious effect (aminoacid change, truncated protein, etc) have been reported (OMIM, Online Mendelian Inheritance in Man database). However, very few genes involved in more common complex diseases, such as type 2 diabetes, schizophrenia or cancer, have been reported. The same applies to quantitative phenotypes, such as height or blood cholesterol levels, where even though the genetic component is clear, the genes involved in them are poorly known. In all these cases, groups rather than individual genes seem to be acting together, with each of them being partially responsible of the considered phenotype. Recently, some single nucleotide polymorphisms (SNPs) located in non-coding sequences of genes have been reported
to be associated with complex diseases. In most cases, disruption or creation of a TF binding site has been suggested as functional mechanism[170-173]. This has contributed to the great interest in the identification of SNPs at a large scale and resolution, and databases with large numbers of SNPs are being generated (dbSNP database, http://www.ncbi.nlm.nih.gov/projects/SNP/).

However, there are other types of polymorphic sequences that can be key players behind the generation of quantitative genetic variation, and among them, simple sequence repeats like microsatellites, are excellent candidates due to some of their properties [174]. They are widely dispersed in eukaryotic genomes, but in some instances their distributions are not random, with some type of repetitive sequences being more frequent nearby transcription start sites [150, 175]. They are involved in transcriptional control, sometimes acting as regulatory elements, and can be also found in coding regions affecting the function of proteins for example in some triplet repeats disorders [176, 177]. They present high mutation rates and thereby they are highly polymorphic, which can affect their functionality in a quantitative manner. These quantitative effects can lead to quantitative phenotypes [178-180]. The biological functions of microsatellites are tightly associated with their capacity to adopt non-B DNA conformations, being more important in some cases than in others, notably pyr/pur repeats and C/G rich repeats. Finally, the usage of microsatellite polymorphisms as a source of functional diversity seems to be evolutionary conserved, since it is also exploited by certain prokaryotic organisms [181]. Therefore, the identification of polymorphic microsatellites in the appropriate locations and with certain sequence characteristics is of interest, and could result in a better understanding of how quantitative genetic variation is achieved.
Methods

A brief comparison of the key techniques used in this work will be presented: EMSA vs ChIP and ChIP-chip. This comparison does not aim to rank them, but rather indicate their complementarities and suitability for different purposes.

Electrophoretic mobility shift assay (EMSA)

This technique is generally used for investigating if individual short (20-50 bp) DNA sequences are bound by a given protein/s, or if a certain protein binds a defined DNA sequence. In the first option, nuclear protein extracts are generally used, and if a particular protein present in the extract is binding the DNA sequence, it can be identified by using an antibody against it, in what is called as a supershift assay. In the second case, a recombinant or purified protein is typically studied, analyzing its binding capacity to a cohort of DNA sequences.

The principle in EMSA (Fig 5) is based on the fact that the mobility in a polyacrylamide gel of a labelled DNA probe is reduced when such DNA is bound by a protein/s. The mobility can be further retarded if the protein-DNA complex is bound by an antibody that recognizes the protein. The specificity of the investigated interactions can be established by performing competition experiments with excess of unlabelled DNA sequences, both with the same sequences as the labelled one or with unrelated ones. In the first case, the binding represented by the shifted complex should be abrogated/competed, while in the second case no effect should be observed (Fig 5).
Chromatin immunoprecipitation (ChIP)

Chromatin immunoprecipitation is the state-of-the-art technique employed for investigating \textit{in vivo} protein-DNA interactions. Typically, crosslinking of DNA-protein interactions is required, and this should be optimized in such a way that those interactions are preserved without introducing unspecificity. After fragmentation of chromatin into a desired size (by sonication, MNaseI, restriction enzyme digestion), the DNA regions bound by the protein of interest are immunoprecipitated and separated from the unbound...
DNA sequences. This is achieved by using an antibody that recognizes the DNA-bound protein, and that then can be precipitated e.g. through its interaction with protein-G/protein-A coupled to agarose or sepharose beads. The primary antibody can be also directly coupled to some substrate that might allow its precipitation (agarose beads, magnetic beads). Once the protein-bound DNA has been purified, the interaction of the investigated protein with different genomic loci can be established by semi-quantitative or preferably quantitative PCR. Such analysis can estimate the enrichment levels of the ChIP DNA over a fraction of the total genomic DNA originally employed in the ChIP and/or over the DNA obtained in a mock ChIP (i.e. using IgG or no antibody).

**ChIP-chip (Chromatin immunoprecipitation and microarray hybridization)**

Although large-scale analysis of ChIP DNA has been accomplished by intensive cloning and sequencing methodologies, it is the combination of ChIP and microarray technologies (ChIP-chip, ChIP-on-chip, genome-wide location analysis) that has been most widely used [141, 182]. Compared to EMSA or ChIP analysis by PCR, where one single locus can be investigate at a time, ChIP-chip technology allows the parallel interrogation of large genomic regions in a single experiment. The extent of such interrogation largely depends on the microarray characteristics, ranging from arrays covering selected genomic regions (e.g. CpG islands, promoters) or arrays completely covering (i.e. tilling arrays) a certain genomic fraction (e.g. ENCODE regions, individual chromosomes) or a whole genome. Furthermore, the resolution of the analysis will be positively and negatively correlated with the density and size of the microarray probes respectively. The resolution is also affected by the size range of the fragmented DNA obtained by sonication or other methods.

In a typical ChIP-chip experiment (Fig 6), the ChIP DNA and a fraction of the total genomic DNA from the analyzed cells, are differentially labelled and hybridized onto the array. Subsequently, the dye ratios are calculated which allows establishing the enrichment levels throughout the genomic regions printed in the array, which reflect the binding of the investigated protein in such DNA locations.
The described methods have some advantages and disadvantages. The main drawback in EMSA is that it is an in vitro technique, so the results always have to be interpreted with caution, since extrapolation to the in vivo situation could be misleading in many instances. Another problematic aspect is that the technique is difficult to scale up, and a limited number of DNA se-

Figure 6. Flowchart of a typical ChIP-chip experiment. Depending on the initial number of cells employed, amplification of the immunoprecipitated DNA might need to be amplified prior to hybridization (adapted from [141]). Reprinted from: ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **83**, 349-60 (2004), with permission from Elsevier).
quences and/or proteins can be studied in parallel, which results in time-consuming and rather tedious experiments. However, it is definitely an option when a first glimpse of the regulatory regions of a single gene or the binding characteristics of a single protein are of interest. EMSA has also the advantage that it can be employed to study the functional effect of polymorphic regulatory elements, since several DNA sequences differing in, for instance, a bp (SNP) or the number of repetitive units (microsatellite), can be compared for their protein binding properties. Finally, it is a rather inexpensive and simple technique that can be performed in standard equipped laboratories.

Most of the advantages and disadvantages just mentioned are reversed when it comes to ChIP-chip. This novel technique is rather sophisticated and expensive, requiring non-standard equipment which limits its applicability. Moreover, it does not allow investigating, in a straightforward and simple manner, the effects of variation in regulatory sequences. Nevertheless, its benefits are remarkable, and can be summarized by the fact that the generated data represent the in vivo binding profile of a particular protein in large genomic regions. This allows the analysis, at a genomic scale, of various aspects of transcriptional control, such as transcriptional initiation, establishment of transcriptional programs, cooperativity between transcription factors, the histone code, etc. The results gain importance in parallel with the genomic coverage of the arrays employed and with their resolution. Sufficient resolution could result in the identification of consensus binding sequences for the investigated protein based on in vivo data, and the inference of the binding sites at base pair resolution. Another important aspect is that, in theory, there is no need for previous knowledge about a TF in order to decipher its binding profile by ChIP-chip. The limiting factors are the availability of an antibody suitable for ChIP and the proper biological system/cell type expressing such protein.

The complementarity of both methods can be exemplified by a study of a certain TF by ChIP-chip, where a number of binding locations are generated. A particular regulatory element suspected to control a gene of interest can then be further studied, resulting in the prediction of the bases bound by the protein at that location. It might be the case that polymorphic variation is known to occur in this sequence, and thus, could affect the binding of the protein. This matter can easily be studied by EMSA, which could result in the identification of a functional polymorphism with perhaps quantitative phenotypic effects or even deleterious ones causing disease. For these and other methods employed in our four papers, more detailed protocols can be found in the actual publications included in this work.
Aims

General Aim
- Achieve a better understanding of transcriptional regulation in human cells and how its alteration can lead to disease.

Specific Aims
- Evaluate the functional potential of microsatellites as cis-regulatory elements, which due to their abundance and polymorphism could represent an important source of quantitative genetic variation (paper I).
- Establish and develop technologies allowing large-scale in vivo analysis of transcriptional regulation processes (papers II-IV).
- Decipher the binding profiles of transcription factors implicated in metabolic diseases (e.g. type 2 diabetes, familial combined hyperlipidaemia, metabolic syndrome), in order to uncover the transcriptional networks being disturbed in such disorders (papers II and IV).
- Improve the understanding of the trans-activating mechanisms employed by transcription factors and the crucial components controlling their binding properties (paper IV).
- Investigate the suitability of ChIP-chip technology for detection of dynamic changes in protein-DNA interactions (paper III).
- Characterize the genome-wide effects of histone deacetylase inhibitor treatment on histone acetylation levels in an unbiased manner (paper III).
Paper I

All kinds of repeats, with the exceptions of those having structural roles in centromeres and telomeres, are usually considered as “junk” DNA, even though accumulating evidence indicate their potential roles in genomic stability/instability or transcriptional regulation.

The aim of our first manuscript was to evaluate the functional potential of microsatellites (short tandem repeats) as polymorphic regulatory elements. In order to do that, a database with all short tandem repeats, from 2-10 bases as repetitive unit, located in the first 10 kilobases 5’ of TSS of all annotated human genes was created. Several microsatellites were selected based on two criteria: relative position to TSS and number of repetitive units, with those located close to TSS and having larger number of repeats being preferred, since functionality and polymorphism are more likely under those conditions. Selected microsatellites were genotyped in DNA samples from 48 unrelated individuals, which allowed the identification of common polymorphisms, and we identified 51 polymorphic short tandem repeats and selected a subset of 12 for further characterization. The in vitro functional evaluation of the different repeats was mainly focused on two aspects: their capacity to act as traditional transcription factor binding sequences and/or adopt non-B DNA conformations. The most thorough assessment was performed for CTC/GAG (polypyrimidine/polypurine) repeats, of which two polymorphic representatives were found in proximal promoters of two different genes. This type of repeats was found to bind, as double-stranded DNA, with high affinity and specificity to the ubiquitous transcription factor Sp1. Moreover, it was also shown that these sequences can form triplexes (H-DNA), where single-stranded GAG interacts with the CTC/GAG duplex, while the remaining CTC strand loops out as single-stranded DNA. Due to this observation, we investigated the protein binding properties of ssCTC repeats and we demonstrated that they were able to interact with polypyrimidine tract-binding protein 1 (PTBP1) and potentially with other hnRNP proteins as hnRNP K. All these observations, together with the skewed distribution of CTC repeats towards TSS, their occurrence in TATA-less, growth-related gene promoters, led to the proposal of a model for the involvement of these repeats on transcription initiation, were two alternative states with different DNA conformations are proposed. Importantly, one of
the polymorphic CTC repeats, located upstream of the \textit{PAX7} gene, was previously reported to result in differences in gene expression depending on the number of repeats. We have shown in our work that longer repeats seem to bind Sp1 and form triplexes with higher efficiency, which could explain the reported differences in expression.

Protein binding was also found for AAACA and AAAAT repeats, and even though the identity of the proteins was not discovered, we were able to verify differences in binding patterns depending on the number of repetitive units in both cases. In spite of most of our results being generated through \textit{in vitro} experiments, our data suggest that certain microsatellites, when located in proximity of TSS and having certain sequence composition, can act as cis-regulatory elements, thereby affecting gene expression through transcription factor binding and/or switching DNA conformation. Furthermore, due to their high polymorphism and abundance, they might be an important source of quantitative genetic variation.
Transcriptional regulation is a highly complex and regulated process, where principles as synergism, cooperativity and combinatorics are used by DNA-regulatory elements and transcription factors in order to achieve both the required specificity and flexibility. Despite this complexity, most of our knowledge of transcriptional control comes from analysis of single genes, with a clear bias towards sequences located in proximity and upstream of TSS. As a result, limited numbers of regulatory elements have been identified, generally by employing in vitro methodologies as in our first paper, and consequently the conclusions drawn from such studies should be taken with caution. Therefore, there is a clear need for in vivo large-scale approaches, where both more genes and larger genomic-regions can be interrogated for their role in transcriptional regulation. In order to achieve these goals, a large-scale in vivo characterization of hepatocyte transcriptional regulation in HepG2 cells is presented in our second study. By combining chromatin immunoprecipitation (ChIP) and microarray technologies (ChIP-chip), we analyzed the binding patterns of metabolic disease related transcription factors (HNF-4α, HNF-3β and USF1) and acetylated histone H3 in 1% of the human genome as selected in the ENCODE project. The ENCODE project was launched by the NIH to evaluate strategies for the identification of all functional elements in the human genome, and as a starting point, 1% of the genome was covered, including both manually and randomly selected regions. In collaboration with the Welcome Trust Sanger Institute, PCR fragment based genomic tilling path arrays covering the ENCODE regions at an average resolution of 1,1 Kb were used in this study.

Considerable effort was made in performing various and exhaustive quality controls in order to establish the specificity and reproducibility of ChIP-chip experiments. To this end we made comparison to negative control arrays, verification of enrichments by traditional ChIP experiments, analysis of antibody specificity by western-blotting and immunohistochemistry, validation of enrichments with two different antibodies against a common protein (i.e. USF1) and comparison to already published data. More interestingly, our data suggest that HNF-4α and HNF-3β, which were commonly bound to distal regulatory elements, may cooperate in the regulation of a large fraction of the liver transcriptome, since they share many common targets. Further-
more, both HNF-4α and USF1 may promote histone H3 acetylation to many of their targets, especially those located in proximal promoters, where, most of the H3ac and USF1 bindings were located. Importantly, bioinformatic analysis of the sequences bound by each transcription factor using motif finding programs, resulted in the identification of consensus binding sequences highly similar to the previously established ones, and allowed the inference of the exact location of transcription factor binding sites at base pair resolution, that we referred as tentative binding sites (TBS). Some of these TBS were shown to be in fact bound by the correspondent TF by EMSA using HepG2 nuclear extracts, while other TBS coincided with previously established binding sites, mainly located in the well-characterized apolipoprotein A4/C3/A1 cluster. Intriguingly, when the HNF-4α targets were separated based on their distance to TSS and their sequences analyzed by BioProspector, the resulting consensus was very similar to the previously established one in the case of targets located at some distance, either upstream or downstream, from TSS, while it was very different when using targets located within 5Kb of TSS. This suggests that a large fraction of HNF-4α bindings in proximal promoters are probably due to indirect interactions, and this might be a recurrent theme for certain transcription factors.

Finally, our study has clear implications for future studies aiming at the profiling of TF binding sites in the whole genome, since it indicates that a resolution of 1Kb could be enough to map with high precision most of the regulatory elements. Moreover, by establishing the quality of an antibody against a particular TF and using the appropriate biological material (cell type, tissue, etc), it should be possible to extend these type of studies to uncharacterized/predicted TFs, in order to gain broad biological insight into their functions.
Cancer, one of the major causes of death in the Western-World, is a complex disease where both environmental and genetic factors play important roles. Early research efforts demonstrated the importance of genetic and cytogenetic alterations in carcinogenesis. However, recent data suggest that cancer is a genetic as much as an epigenetic disease where both DNA methylation and histone modifications are abnormally altered in tumour cells.

The development of anti-tumorogenic drugs is of great interest and in particular those trying to correct cancer-associated epigenetic alterations are receiving great attention. Among the last ones, HDACi are natural or synthetic drugs which inhibit the action of class I and class II histone deacetylases and as a result increase the global levels of histone acetylation. This type of histone modifications is one of the best characterized histone PTM, is typically found close to TSS of active genes and display high turnover rates. Therefore, it has been hypothesized and demonstrated in a few cases (e.g. WAF1/P21), that HDACi can increase histone acetylation at promotors of epigenetically silenced tumor suppressor genes. These genes can subsequently be activated, leading to growth arrest and/or apoptosis of tumour cells.

However, transcriptome analysis of cells treated with HDACi typically identifies equal numbers of genes being up or downregulated. Furthermore, most of the efforts in elucidating the mechanisms of action of these drugs have been concentrated on a few genes, typically those being upregulated, while ignoring the common gene repression events. Therefore, we decided to investigate in an unbiased manner, the effects of butyrate, which is a naturally occurring HDACi, on the acetylation patterns of DNA-bound histones H3 and H4. Such analysis was performed in HepG2 cells treated with Na-butyrate for 12 hours and we interrogated histone acetylation in 1% of the human genome as defined by the ENCODE consortium.

After confirming that butyrate increases the global levels of H3ac and H4ac in HepG2 cells, we performed ChIP-chip analysis before and after 12h treatment with the drug. Our results demonstrated loss of H3ac and H4ac at
many promoter regions, which was confirmed by standard PCR analysis of ChIP DNAs and was found to be reversible upon withdrawal of the drug. Moreover, similar promoter deacetylations were observed when cells were treated with another HDACi, Trichostatin A, or when another cell line, HT29, was treated with butyrate. We found only few instances of increased histone acetylation and they were not located in promoter regions. More detailed time-course analysis of histone acetylation after butyrate treatment indicates that hyperacetylation might occur early and transiently after treatment.

The observed histone deacetylations were not the result of altered nucleosome occupancy, since total H3 content or H3K4me3 were not affected by the drug at investigated promoters. Furthermore, several genes where we observed loss of histone acetylation displayed decreased expression and reduced RNA PolII initiation/elongation. Finally, the global increase in histone acetylation was mainly localized at the nuclear periphery, probably associated to heterochromatic regions and/or free histones. In summary, our study offers a mechanistic explanation for the commonly observed gene repression under HDACi treatment and suggests that previously proposed mechanisms of action of these drugs might need to be revised.
In Western-societies, the incidence of complex metabolic disorders such as obesity, type2 diabetes or the metabolic syndrome is reaching epidemic proportions. Several aetiological characteristics associated with these disorders are shared with familial combined hyperlipidaemia (FCHL), which is the most frequent form of dyslipidaemia, involving elevated serum triglycerides and/or cholesterol levels. Recently, several reports have revealed the association of FCHL with non-coding variants of the USF1 gene. Consequently, USF1 has received great attention and its functional role as a central metabolic regulator has been further emphasized by associations with the metabolic syndrome or the risk of cardiovascular disease.

Transcription factors are optimal candidates to be involved in complex diseases, since both coding and non-coding mutations can induce subtle changes in the level or activity of the TF and as a result, the expression of many genes can be altered. Therefore, identification of the genes and pathways under the control of USF1 could be of relevance for the understanding of metabolic transcriptional regulation and its alteration in metabolic disorders. As a crucial step towards that goal and extending the work presented in paper II, we have investigated by ChIP-chip, the binding profiles of USF1, USF2 and H3ac in the whole non-repetitive fraction of the human genome with an average resolution of 35 bp, using HepG2 cells as a liver cellular model. Similarly to paper II, great effort was devoted to establish the quality of our ChIP-chip experiments, by a combination of qPCR, siRNA and comparison to previous data sets, resulting in very limited false positive rates.

USF1 and USF2 proteins displayed highly similar binding profiles, in agreement with their preferential binding capacity as heterodimers. Furthermore, a major fraction of USF bound regions was close to TSS of protein coding genes, frequently in the context of H3 hyperacetylated chromatin. This was in concordance with a positive correlation between USF binding affinity to promoter regions and the expression levels of the corresponding genes. Interestingly, many of the USF and/or H3ac bound promoters were found in a bidirectional conformation, involving pairs of sense and antisense transcripts.
The abundance of USF binding in proximity of TSS allowed us to investigate pathways that might be under the control of these TFs. Such analysis uncovered a previously unknown role of USF proteins in regulating mitochondrial activity and identified several USF-bound genes with important functions in lipid metabolism.

The extent of our binding profiles has also allowed us to investigate the major determinants of USFs binding. Using motif finding tools and chromatin profiling, we found that USF binding mainly occurs in the context of euchromatic regions while its binding affinity largely depends on the presence of E-box or E-box like sequences (i.e. CACGTG, CACGTGAC). These two characteristics are most frequently fulfilled in regions close to transcription start sites, where USFs can interact with RNA PolII machinery, as indicated by our data and previous reports.

In conclusion, our whole genome-characterization of the USF regulome represents a valuable resource to expand our understanding of metabolic transcriptional regulation and identifies relevant genes and pathways for future study in several metabolic disorders.
Acknowledgements

I would like to express my gratitude to a number of people that made this work possible:

Ulf Pettersson and Kenneth Nilsson, former and present directors of the Rudbeck Laboratory, for creating such a pleasant place to do science.

My supervisor, Claes Wadelius, for his never-ending optimism and for his support and confidence in me and my ideas. Actual and former members of Wadelius’ group: Mattias Jansson, who introduced me in this crazy world of research, Benjamin Bakall, Fredrik Aspgren. The ones remaining, Ola Wallermann (thanks a lot for all the help with computers and bioinformatic analysis of various kinds) and Mehdi Motallebipour (those late evenings have been much more pleasant with you in the room. Thanks for tolerating my chaos around you and you can bet I will do that trip to Iran. I really hope you can find that dream place for your future and surely successful career). I will also like to thank to all my students, for “adapting” to my special learning techniques. I hope you all good luck.

Other people at Rudbeck Lab, especially those from Dahl’s group, both actual and former members. I apologize for my characteristic chaos that might have affected some of you. I especially enjoyed the football discussions with Ed and Larry, while pipetting stuff. I am also grateful to the people that shared or shares same office as me: Hanna Göransson, Maja Nyström and Christina Edéby….the smallest and most crowded room, but for sure I will not change it. Other people that are or were at the lab (Marina, Cecilia, Patrick, Sarah, Chris, Ludmila Elfineh, etc), that made those incubation times more enjoyable. Many lunches were also tastier thanks to Marta Alarcon, that let me practice my spanish while eating. Finally, Lucia Cavelier, who I believe is a model for me, not only as a scientist, but mostly as a person. For always offering help without expecting anything in return.

Special mention to the people at Linnaeus Centre for Bioinformatics: Adam Ameur (I will always be in debt to you, not only for your scientific help, but mostly for your patience and tolerance with me), Stefan Enroth (I
am still amazed with how easy you make comprehensible things that for most mortals are just mathematical nightmares...I have learnt a lot from your way of thinking), and their supervisor Jan Komorowski (I will always remember our collaboration mostly from its personal side, and secondly from the scientific one. I really appreciate your way of dealing with both scientific and non-scientific issues). It has been great fun collaborating with you and I hope the future will bring even better things for all of us.

Lots of thanks to Lena Åslund for so many reasons, scientific and non-scientific (specially that fantastic Indian food and Julbord). I hope the future will let you more time to exploit your incredible scientific knowledge that surely will benefit everyone, as it has done with me, first as your student and then as lab colleague.

I will also like to mention some people that contributed in my decision to stay in Sweden, and that made my first years easier and funnier: Emma, Jenny, Leila, David, the “famiglia” of Italians, my rugby team….and all those that I would mention, but I will probably forget some, so thanks to all.

The following will be in Spanish, since most of those people I would like to express my gratitude to do not understand English:

A mi madre, porque siempre quiso, quiere y querrá lo mejor para mí, no importa lo que sea; porque se mantuvo feliz en mis malos momentos y fue aún mas feliz en los buenos; porque me habla cuando necesito escuchar y me escucha cuando lo que quiero es hablar, y sobre todo, porque la verdad es que madre solo hay una, y la mía es la mejor.

A mi padre, que siempre ha tenido tanta confianza y expectaciones puestas en mí, que siempre ha hecho todo lo posible e imposible para ayudarme y hacerme las cosas mas fáciles. Sólo espero que un día te pueda hacer tan orgulloso como mereces, aunque probablemente eso sea imposible

A mi hermano, porque es mi ídolo e imagen en tantas cosas que no imagina: la felicidad es posible, pero exige tomar riesgos. Además, ojalá algún día consiga ser tan buen amigo de mis amigos como tú lo eres de los tuyos. Te echo de menos y espero que el futuro nos ponga otra vez juntos geográficamente, porque mentalmente nunca estuviste lejos de mí.

A la mejor mujer del mundo y parte del universo, que por cuestiones inexplicables tal vez relacionadas con la suerte o el azar, resultó ser mi abuela. No sé muy bien que es lo que he hecho para merecer su amor, alegría,
apoyo, y por supuesto, comida. Hay muchas razones por las que quiero volver pronto a España, pero Mamina, tú eres la principal.

A mis tíos y primos, donde prefiero no dar nombres, pero aquellos a los que me refiero no necesitan que se les nombre porque ellos ya saben quienes son, por nunca olvidarse de mí, y porque aunque les vea poco, siempre me muestran su cariño.

A mis amigos en España, especialmente David, el paramés, Zuma y Jorge, el tasuguillo. Por ser tan buenos amigos, por perdonarme el no responder emails, por las jornadas de pajareo, por los kanel bullar, por Santoña y por Leon (cada uno sabe lo que le toca de esas cosas).

A la panda del moco, o sea, el grupo de amigos, que gracias a tanto asado y fiesta casera, ha hecho y hará más llevadero sobre todo los inviernos suecos. A todos los que están aquí (“el negro” Marcos, “el gordo gris” Sebas, “la amperti” Ruse, “el chasqui” Alvaro, “paulotudo” Paulo, Raulito, “el chorizo” Marcelo, Jeremy, Alex, Rakel, Pepe) y a los que para su bien ya nos dejaron (“Maximus”, “malaguita” Antonio, Eduardo---cuantos trucos de laboratorio te debo!!!) os quiero dar las gracias, deciros que nunca os olvidaré y como dicen los Celtas “Nos vemos en los bares”. No cambiéis nunca.

Todo este tiempo en Suecia, todo el tiempo pasado en la Universidad, este trabajo, y me imagino que todo lo que he conseguido en mi vida adulta habría sido imposible, o al menos poco probable, sin una criatura muy bella y especial: Patricia. Espero poder seguir mirando a esos grandes ojos marrones durante muchos años y encontrar en ellos toda la paz, armonía, amor y comprensión que me has dado hasta ahora. No me puedo imaginar mi vida sin ti, y solo deseo poder quererte la mitad que alguien como tú merece. Como dice la frase, si no existieras, tendría que inventarte. Te quiero, morenucha.

Finalmente me gustaría acabar diciendo que todo este trabajo y el trabajo venidero son mi sencillo y sentido homenaje a mi abuelo Candi. Creo sincera y profundamente que a día de hoy soy biólogo, para lo bueno y lo malo, por él. Solo deseo algún día poder albergar una pequeña fracción de la sabiduría que él atesoraba. Sólo hay una razón por la que a veces dudo de la no existencia de Dios, porque así, quizás, Candi podría estar viéndome, y porque, quizás, algún día yo le pueda volver a ver a él.
References


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 222

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine”.)