Genome-Wide Studies of Transcriptional Regulation in Human Liver Cells by High-throughput Sequencing

MADHUSUDHAN REDDY BYSANI
Dissertation presented at Uppsala University to be publicly examined in Rudbeck hall, The Rudbeck Laboratory, Dag Hammarskjölds väg 20, Uppsala, Monday, June 10, 2013 at 09:15 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English.

Abstract

The human genome contains slightly more than 20 000 genes that are expressed in a tissue specific manner. Transcription factors play a key role in gene regulation. By mapping the transcription factor binding sites genome-wide we can understand their role in different biological processes. In this thesis we have mapped transcription factors and histone marks along with nucleosome positions and RNA levels. In papers I and II, we used ChIP-seq to map five liver specific transcription factors that are crucial for liver development and function. We showed that the mapped transcription factors are involved in metabolism and other cellular processes. We showed that ChIP-seq can also be used to detect protein-protein interactions and functional SNPs. Finally, we showed that the epigenetic histone mark studied in paper I is associated with transcriptional activity at promoters. In paper III, we mapped nucleosome positions before and after treatment with transforming growth factor β (TGFβ) and found that many nucleosomes changed positions when expression changed. After treatment with TGFβ, the transcription factor HNF4α was replaced by a nucleosome in some regions. In paper IV, we mapped USF1 transcription factor and three active chromatin marks in normal liver tissue and in liver tissue of patients diagnosed with alcoholic steatohepatitis. Using gene ontology, we as expected identified many metabolism related genes as active in normal samples whereas genes in cancer pathways were active in steatohepatitis tissue. Cancer is a common complication to the disease and early signs of this were found. We also found many novel and GWAS catalogue SNPs that are candidates to be functional. In conclusion, our results have provided information on location and structure of regulatory elements which will lead to better knowledge on liver function and disease.

Keywords: ChIP-seq, Transcription factors, Alcoholic steatohepatitis, Genome-wide, GWAS, SNPs

Madhusudhan Reddy Bysani, Uppsala University, Department of Immunology, Genetics and Pathology, Rudbecklaboratoriet, SE-751 85 Uppsala, Sweden.

© Madhusudhan Reddy Bysani 2013

ISSN 1651-6206
ISBN 978-91-554-8671-6
urn:nbn:se:uu:diva-198579 (http://urn.kb.se/resolve?urn=nbn:se:uu:diva-198579)
To my parents and well wishers
This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


IV Bysani MSR, Wallerman O, Bornelöv S, Zatloukal K, Komorowski J Wadelius C. ChIP-seq in steatohepatitis and normal liver tissue identifies candidate disease mechanisms related to progression to cancer (Manuscript).

*These authors contributed equally in this work.

Reprints were made with permission from the respective publishers.
Abbreviations

3C  Chromosome Conformation Capture
ASH  Alcoholic Steatohepatitis
Bp  Base pair
ChIA-PET  Chromatin Interaction Analysis by Paired End Tags
ChIP  Chromatin Immunoprecipitation
ChIP-seq  Chromatin Immunoprecipitation with Sequencing
CNV  Copy Number Variation
Co-IP  Co-Immunoprecipitation
CTCF  CCCTC binding factor
DHS  DNase 1 Hyper Sensitivity
DNA  Deoxyribonucleic acid
DNMT  DNA methyl transferase
ENCODE  ENCyclopedia of DNA Elements
FCHL  Familial Combined Hyperlipedemia
GABP  GA Binding Protein
GTF  General Transcription Factor
GWAS  Genome Wide Association Study
H3K27ac  Histone 3 lysine 27 acetylation
H3K4me1  Histone 3 lysine 4 mono-methylation
H3K4me3  Histone 3 lysine 4 tri-methylation
HCC  Hepatocellular carcinoma
HCV  Hepatitis C Virus
HepG2  Cell line from Hepatocellular Carcinoma
HGP  Human Genome Project
HNF4α  Hepatocyte Nuclear Factor 4α
ICR  Imprinted Control Region
LCR  Locus Control Region
lincRNA  long intervening noncoding RNA
lncRNA  long noncoding RNA
miRNA  micro-RNA
MNase1  Micrococcal Nuclease 1
MODY1  Maturity Onset Diabetes of Young 1
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>NASH</td>
<td>Non-alcoholic steatohepatitis</td>
</tr>
<tr>
<td>NGS</td>
<td>Next Generation Sequencing</td>
</tr>
<tr>
<td>NRF</td>
<td>Nuclear Respiratory Factor</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase Chain Reaction</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>rRNA</td>
<td>ribosomal RNA</td>
</tr>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
</tr>
<tr>
<td>TF</td>
<td>Transcription Factor</td>
</tr>
<tr>
<td>TGF</td>
<td>Transforming Growth Factor</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>UTR</td>
<td>Untranslated Region</td>
</tr>
</tbody>
</table>
Introduction

The decoding of a near complete human genome in 2001 is one of the largest achievements in science since the discovery of the structure of deoxyribonucleic acid (DNA) by Watson and Crick in 1953 [1-3]. The human genome project (HGP) revealed that there are approximately three billion nucleotides organized in the 22 pairs of chromosomes and the sex chromosomes X and Y (XX in women and XY in men). Only a small proportion 1.5%, of the genome codes for mRNA. There are slightly more than 20 000 human protein coding genes. In addition, an unknown number of DNA loci are transcribed into non-coding RNAs such as ribosomal RNA (rRNA), transfer RNA (tRNA), micro-RNA (miRNA), long non-coding RNA (lnc-RNA) or transcripts of unknown function [4].

The successful completion of the HGP has provided a road map to many large international and collaboration projects like the HapMap project, the ENCyclopedia of DNA Elements (ENCODE) project and the 1000 genomes project.

Genetic variation

The current estimate is that the human genome contains approximately 38 million single nucleotide polymorphisms (SNPs) and around 10 thousand of common SNPs are associated with common diseases, whereas other variants are rare and possible association to disease is not known. The HapMap project was aimed to identify common patterns of genetic variation in diverse human populations [5]. The purpose of the 1000 genomes project was to characterize the genetic variants with at least 1% frequency in the population using high-throughput sequencing [6]. The second phase of this project used 1092 individuals to provide the current information map of 38 million SNPs, 1.4 million short indels and 14 000 large deletions [7] that have occurred at different times during human evolution. Common variants are old and rare variants are younger. A recent large scale exome-seq study of ~15 000 genes in 6 515 individuals revealed that 86% of the SNPs that were predicted to be harmful had arisen in the last 5 000 years [8]. Selection may not have had time to act on these variants and remove them from the population. There are a large number of copy number variations (CNV) varying in size from ~1 kb to several mega base long in human genome. CNVs have been associated
with many diseases e.g. autism [9]. Genetic variants like SNPs and indels may also affect the binding of transcription factors (TFs) [10]. Also others have later shown that genetic variation changes the heritable chromatin status and TF binding [11, 12].

The ENCODE Project

The ENCODE project was started in 2003 to map the functional elements in the human genome. The pilot phase of the project mapped TF binding sites, histone marks, DNase 1 hypersensitive sites (DHS) etc. in 1% of the human genome [13]. The next phase of the project was focused on continuing the annotations of protein coding, non-protein coding genes and their transcripts and transcription regulatory regions in the whole genome using massively parallel next generation sequencing technology [14]. In September 2012, ENCODE consortia published a series of papers in *Nature*, *Science*, *Genome Research* and *in Genome Biology* to describe the results from 1 640 genome-wide datasets (ChIP-seq, RNA-seq, DNase-seq, FAIRE-seq, 5C etc.) (Figure 1) from 147 cell types. The ENCODE data summary paper [15] claimed that 80.4% of the human genome is biochemically active with approximately 400 000 enhancer-like characters and 70 000 promoter-like characters. Their findings will be further discussed in the coming sections.

The modENCODE project was started to identify the functional elements in *Caenorhabditis elegans* and in *Drosophila melanogaster*. modENCODE project has generated more than 700 datasets and they have identified protein coding and non-coding genes, regulatory, replication and chromatin elements in *Drosophila melanogaster*. They have also identified new functions of annotated genes, general and tissue-specific regulatory elements. The complex modENCODE data generated today covers 82% of *D. melanogaster* genome which is four times larger than the previously annotated protein coding exons [16]. The modENCODE project of *C. elegans* was aimed at the transcriptome profiling during a developmental time course, identification of TF binding sites and generation of chromatin organization maps. They have also focused on TF binding networks and miRNA interactions [17].
Sequencing technology

The launch of Next Generation Sequencing (NGS) technology, played a key role in the ENCODE project, 1000 genomes project and many other projects. Rapid development in NGS technology reduced the sequencing cost and today the whole human genome can be sequenced in couple of hours.

For this thesis, we have used Illumina and Life technologies SOLiD sequencing platforms to map the regulatory regions, histone marks, nucleosome positioning etc. We have mapped liver specific TFs, using Chromatin Immunoprecipitation in combination with NGS (ChIP-seq).

Transcriptional regulation and RNA polymerases

Transcription is the process where an equivalent copy of RNA is created from a DNA sequence. Transcription is performed by three different multi-subunit RNA polymerases. RNA polymerase I synthesizes 45S pre rRNA which then matures into 28S, 18S and 5.8S rRNA [18]. RNA polymerase II synthesizes pre-mRNAs, miRNAs and snRNAs [19]. RNA polymerase II coordinates with RNA polymerase III to maintain gene expression globally [20]. RNA polymerase II forms a holozyme along with several enzymes and general TFs (GTFs). Transcription begins once the TATA binding protein (TBP), subunit of TFIID, binds to the TATA box which allows the assembly of GTFs and RNA polymerase II. A recent study showed that many promot-
ers lack a classical TATA box, but instead a more degenerate TATA-like sequence has been found in *Saccharomyces* which might also exist in humans [21].

RNA polymerase III is involved in the synthesis of tRNAs, 5S rRNAs and other small RNAs. RNA polymerase III transcription is needed for the generation of adapter tRNA molecules that are used for translation of the genetic information in mRNA into protein. RNA polymerase III binds to only a few hundreds of tRNAs and a study in six mammalian species showed that RNA polymerase III is evolutionarily conserved and that RNA polymerase III binding varies between species in terms of strength and location [22].

Promoters are necessary for transcription and are located upstream of the transcription start site (TSS). Promoters have been divided into core promoters and proximal promoters. Core promoters are located within 50 bp of the TSS and are important for the formation of the pre-initiation complex (PIC) and the binding of GTFs. The TATA-box is located 20-30 bp upstream of the TSS and is found in many promoters. Core promoters are usually silent and need a proximal promoter to initiate transcription. Proximal promoter starts where the core promoters end and are located +/-250 bp of the TSS (Figure 2). The ENCODE project and others have studied GC-rich and AT-rich promoters at the TSS of protein coding transcripts and found that these promoters have distinct histone modification and TF patterns [15, 23]. GC-rich promoters are most common and have the highest activity in all tissues. Most of the housekeeping genes tend to be associated with GC-rich promoters [24].

In addition to promoters, gene activity is also regulated by the binding of sequence specific activator and repressor proteins to DNA elements termed enhancers, operators and silencers. Enhancers/silencers can be located many kilo base pairs from the TSS and are involved in the activation or repression of transcription [25, 26]. P300 is a well-studied enhancer protein whose effect on enhancer activity in different cells was measured by ChIP-seq [27] and with other methods. In the ENCODE project the carbon copy of chromosome conformation capture (5C) method was used to map more than a thousand long range interactions between the promoters and distal elements in each cell type and to correlate their interaction with gene expression [28]. Using one of the recently developed methods, STARR-seq, many functional enhancers have been identified in *Drosophila* [29]. In this method tentative enhancers are cloned in a vector downstream of the TSS so that they are transcribed. Their abundance reflects the enhancer activity and can be measured by RNA-seq.

Silencers are sequence specific DNA binding elements located close to or far from genes and involved in the silencing of a target gene [30]. Locus control regions (LCRs) have similar functions as enhancers/silencers but have the ability to enhance at a greater distance and to regulate many genes in tissue specific manner [31].
RNA polymerase II and GTFs are sufficient for promoter-specific transcription but additional factors such as mediators are important in order to respond to the activators. The mediator complex plays an important role in transcription and acts as a co-activator, co-repressor and sometimes as a GTF during transcription (Figure 2) [32, 33].

Open chromatin sites hypersensitive to the enzyme deoxyribonuclease 1 (DNase 1) are found in all classes of regulatory elements. ENCODE project has identified 2.9 million DHS sites from 129 cell and tissue types by using DNase-seq [15, 34]. Most of the DHS sites they have identified showed cell specific regulation. They also identified many novel TF binding motifs (n=683) using DNase 1 footprints [35]. The majority of the disease associated GWAS SNPs are located at DHS sites and some of these affect TF binding, chromatin states and are involved in the formation of regulatory networks [36].

Figure 2. Schematic illustration of transcriptional regulation. The figure illustrates part of the transcription machinery at core promoter, proximal promoter and TSS. Figure adapted with the permission from Ong C.T & Corces V.G, Nature review genetics, 2011 [37].

The transcriptome
RNA transcripts are involved in many cellular functions either directly or indirectly. Technological advances in RNA sequencing has enabled the prediction of many RNA types, their location and function [38]. Only a small portion of the transcripts code for protein and the remaining are non-coding. There are many types of non-coding RNAs in the genome such as tRNA, rRNA, miRNA and lncRNA. The main function of tRNA is to translate the triplet amino acid code into amino acid sequence of proteins. rRNAs are present in the ribosomes and play an important role in translation.

miRNAs are ~22 nucleotides long small non-coding RNA molecules that control gene expression by degrading mRNA molecules or by inhibiting translation [39-42]. miRNAs binds at 3’un-translated region (3’UTRs) of the
target gene [43]. According to the miRNA data base (miRBASE 19, August 2012; www.mirbase.org), there are about 21,264 miRNAs in 193 species. Among them 2,019 were discovered in humans but the majority are not validated yet. Some of these miRNAs play an important role in organ development, for e.g. miR-127 in lung development and some of them are involved in disease processes, e.g. miR-122 in HCV replication and miR-124 acts as a tumor suppressor [44, 45].

Long non-coding RNAs (lnc-RNA) are at least 200 nucleotides long and show less expression than protein coding genes. lnc-RNAs are expressed in a cell specific manner [46] and can form networks to control gene expression and function. They are also involved in the regulation of the three-dimensional (3D) structure of chromosomes [47]. Most of the disease associated SNPs are found in non-protein coding regions [48] and some of them may affect the function of lnc-RNAs and thereby predispose to disease.

Some lnc-RNAs are called long intervening noncoding RNAs (linc-RNA) and were discovered by Guttman and co-workers [49] by searching outside of non-protein coding genes in regions marked by histone 3 lysine 4 trimethylation (H3K4me3) and Histone 3 lysine 36 tri-methylation (H3K36me3). These linc-RNAs are involved in many biological processes such as dosage compensation and imprinting [50] and many are associated with chromatin remodeling complexes and regulate histone modification patterns and gene expression [51, 52].

Transcription factors

TFs are a group of proteins that regulate the gene expression by binding to DNA in a sequence specific manner. TFs affect transcription alone or in a group by activating or repressing the recruitment of RNA polymerase II. There are ~1,850 TFs in the human genome [30]. Based on their DNA binding motif they have been grouped into helix-loop-helix, leucine-zipper, homeo domain, zinc finger and helix-turn-helix [53]. In a recent study, 830 TF binding motifs were identified by using high throughput SELEX and ChIP-seq [54]. However, the DNA-binding motifs remain to be defined for many TFs.

TFs can bind various regulatory regions, i.e. at promoters, enhancers/silencers, LCRs and insulators etc. TFs are modular proteins that contain activator/repressor domains along with the DNA-binding domain [30]. TFs also contain domains for protein-protein interactions. Some proteins called co-activators or co-repressors regulate transcription by binding to TFs instead of DNA [55]. Transcription begins with the binding of one or several TFs to their respected TF binding site and also the recruitment of co-factors. The co-factors help to remove nucleosomes to allow binding of the transcription machinery to the promoters. Co-factors also modify the nucleosomes at promoters and at distal regions.
Regulatory networks

The genomic information is almost the same for all multi-cellular organisms but there is a lot of phenotypic variation. A recent technological advance in the mapping of regulatory regions i.e. DHS sites and TF binding sites has identified regulatory networks in humans and other species. The regulatory networks are involved in the development, metabolism and hormonal stimulation etc. in a tissue specific manner [56]. To know the function of a specific TF, it is important to know which genes it regulates. Mapping of hepatocyte nuclear factors (HNF) in mouse embryonic stem cells showed that correct regulation of the HNF family network is required for differentiation and metabolism. Promoter array analysis on three TFs in liver and in endocrine pancreas showed that there is a transcriptional hierarchy between HNF1 and HNF4α [57] suggesting that regulatory networks play a role in maintaining different biological processes. ENCODE project has identified the regulatory networks between 119 TFs and found that these factors co-bind in a combinatorial fashion. These networks are highly cell specific. They also found that there is a significant difference between the networks for proximal and distal regulatory elements. Strongly connected network elements are more evolutionarily conserved than the weaker ones [15, 58].

Transcription factors in disease

TFs are involved in the control of gene expression and any abnormality of their binding to regulatory regions may change the expression status and thus lead to a disease. This may be caused by different types of mutations in TFs like missense, nonsense, duplications or deletions. To date mutations in more than 400 TFs and other nuclear proteins are known to cause human disease (Wadelius C, personal communication). One example is mutations in HNF4α described further below. Mutations in this protein change the affinity for its motif [59]. Mutations or variants at regulatory regions may affect the binding of TFs which may lead to diseases such as cancer, diabetes, obesity, autoimmune and cardio vascular diseases etc. [60]. Many SNPs have been associated to diseases or common traits and this information is collected in a catalogue at NIH (http://www.genome.gov/26525384). As of 19 April, 2013 the catalogue includes 1 571 publications and 9 906 SNPs. Most of these SNPs are thought to affect gene regulation by changing the interaction with TFs. E.g. there has been studies showing that variants in the TF gene TCF7L2 are associated with the risk of type 2 diabetes [61, 62]. C-MYC is one of the well-studied TF in many tumors and it regulates many genes involved in tumorigenesis.
Transcription factors in reprogramming

TFs also play a key role in the development and reprogramming of cells. Induced pluripotent stems cells (iPS) were generated from adult human dermal fibroblasts using the TFs Oct3/4, Sox2, Klf4 and c-Myc (Figure 3). These factors could reprogram mouse embryonic or adult fibroblasts to iPS cells [63, 64]. For this work, Professor, Yamanaka was awarded the Nobel Prize in 2012. In a study by P. Huang et al., adult fibroblasts were directly converted into iHep cells by transduction of the TFs Gata4, Hnf1α and FoxA3 (Figure 3) [65]. In a parallel study, mouse embryonic and adult fibroblasts were directly converted to cells that are closely resembled to hepatocytes. For this study, they used combinations of Hnf4α plus any one of FoxA proteins (FoxA1/FoxA2/FoxA3) in vitro (Figure 3) [66]. Thus, some defined factors could soon be used to generate hepatocyte-like cells for therapeutic purposes. Further studies on these factors and processes could contribute to a breakthrough in tissue transplantation.

Figure 3. Illustration of the reprogramming of cells from one state to another by selected transcription factors. Figure adapted with the permission from Lee Tong I and Richard A. Young, Cell, 2013 [60].

We have studied the TFs FOXA1, FOXA2, FOXA3, HNF4α, GABP and USF1 for this thesis. These factors are briefly described below.

FOXA: s

FOXA or hepatocyte nuclear factor 3 (HNF3) TFs belong to the forkhead box/winged helix family. They are called pioneer factors since they can bind to DNA that it is packed in nucleosomes and removes the nucleosomes to make place for other TFs [67]. There are three FOXAs, namely FOXA1 (HNF3α), FOXA2 (HNF3β) and FOXA3 (HNF3γ). FOXA proteins play a major role in early development, organogenesis and carbohydrate metabolism [68]. The forkhead box sequences of the three FOXAs are highly conserved and are 95% identical. All FOXA proteins share weaker homology...
outside the fork head box, whereas there is higher degree of homology at the N and C terminal ends. FOXA proteins interact with other TFs like USF1, HNF1\(\alpha\), HNF4\(\alpha\) and HNF6 for the regulation of target genes [69]. The FOXA family gene expression during embryogenesis and combined with liver specific FOXA target genes has been interpreted as evidence that the FOXA TFs regulate hepatogenesis. There are ChIP-chip, ChIP-seq and EM-SE studies on FOXA TFs in mouse, human tissues and in HepG2 but the three factors have not been studied together in HepG2 cells previously.

**HNF4\(\alpha\)**

Hepatocyte nuclear factor 4 alpha (HNF4\(\alpha\)) is a TF that belongs to the zinc finger protein family. HNF4\(\alpha\) plays a key role in the formation of TF networks that regulate the development of vital organs such as the kidney, pancreas and liver [70, 71]. There are three members in the HNF4 family: HNF4\(\alpha\), HNF4\(\beta\) and HNF4\(\gamma\). HNF4\(\alpha\) regulates genes involved in lipid and carbohydrate metabolism [72]. HNF4\(\alpha\) is also directly involved in the regulation of genes in glucose transport and in glycolysis. Any abnormality in HNF4\(\alpha\) regulation may lead to diabetes and cancer [73]. Mutations in HNF4\(\alpha\) lead to Maturity onset diabetes of young 1 (MODY1) which is an autosomal dominant form of diabetes or to hyperinsulinaemic hypoglycemia [74, 75]. A recent study found that mutations in the DNA binding domain of HNF4\(\alpha\) protein as well as in the nearby hinge region cause MODY1 by affecting the binding of HNF4\(\alpha\) to DNA [59]. HNF4\(\alpha\) is also associated to the common form of diabetes [76].

**GABP**

GA binding protein (GABP) is a TF that belongs to the ETS family. It is abundantly expressed in liver and muscle tissue. GABP binds to DNA as a heterotetrameric complex and forms two GABP\(\alpha\) and GABP\(\beta\) subunits. GABP mostly binds to GAA enriched sequences and its preferred motif is CCGGAA. GABP regulates genes involved in the cell cycle, protein synthesis and metabolism. An alternative name for GABP is NRF2 since it regulates nuclear respiratory factors [77].

**USF1**

Upstream stimulatory factor 1 (USF1) belongs to the helix-loop-helix leucine zipper family and binds to the E-box sequence CA[C/T]GTG. USF1 can interact with its targets as a homo-dimer or hetero-dimer. USF1 is along with USF2 is involved in the recruitment of histone modifying enzymes and they are enriched at the sites of H3K4me2 and H3ac [78]. USF1 is ubiquitously expressed and involved in disorders of lipid and carbohydrate metabolism. SNPs in the USF1 gene are strongly associated with the disease familial combined hyperlipidemia (FCHL) which results from defects in the liver and other organs. FCHL is phenotypically associated with type 2 diabetes and the metabolic syndrome [79].
Epigenetics

After several decades of genetics research, some focus is now redirected towards epigenetics. The definition of epigenetics is “the study of changes in the gene function that are mitotically and/or meiotically heritable and that do not entail a change in DNA sequence” [80]. This means that epigenetic modifications are heritable from cell to cell just like genetic modifications. There are mainly two types of epigenetic mechanisms, DNA methylation and modification of histone tails.

Chromatin structure and organization

The eukaryotic nucleus harbors genetic information consisting of several billion nucleotides. To maintain all this genetic information in the nucleus, the DNA is packed tightly to form nucleosomes. A nucleosome consists of approximately 147 bp of DNA wrapped twice around an octamer of eight core histones, namely two of each histone H2A, H2B, H3 and H4. Due to this packing DNA is 10 fold compacted [81, 82]. Nucleosomes are connected to each other by a ~20 bp long linker bound by histone H1, which seems to regulate the distance between the nucleosomes (Figure 4).

Nucleosomes are further subjected to compaction to form higher order chromatin fibers. However this tightly packed chromatin structure is virtually inaccessible for the transcriptional machinery. A prerequisite for the transcription of a certain gene segment is the structural reorganization of the gene locus in question. Change in nucleosome positioning, post-translational modification of histone tails and methylation of CpG dinucleotides constitute various levels of chromatin modifications that affect DNA accessibility [83]. These alterations occur in a highly ordered fashion, where selective modifications allow for tissue-specific expression and heritability at cell division; two hallmarks of epigenetic regulation.

Chromatin organization in the cell nucleus is not random. It is organized in a spatial manner and plays an important role in gene regulation. Chromosome conformation capture (3C) based methods have been widely used to study spatial chromatin organization and their interactions [84]. The new Hi-C method can be used to view the 3D interactions of the chromatin genome-wide [85].
Figure 4. Compaction of DNA into chromosomes: A) DNA anti-parallel strands base pair each other and form a double helical structure. The DNA double helix is tightly wrapped around the histone core octamer to form a nucleosome. Nucleosomes are compacted and folded into 30 nm fibers for higher order of compaction and for the chromosome formation. B) Nucleosome core particle: 3D image of nucleosome core particle showing that DNA is wrapped around the histone core octamer. Figure adapted with the permission from Felsenfeld & Groudine, Nature, 2003 [86].
Nucleosome positioning

Rapid advances in sequencing allowed mapping of nucleosome positions in different species and in different cell types. At least in some genes nucleosomes are positioned in order at promoters while the positioning is more random in the gene body (Figure 5) [87]. We have seen well-positioned nucleosomes downstream and upstream of a TSS and nucleosome free region in between them. However, this average nucleosome positioning signal was rarely observed at the TSS of individual genes. Some CTCF binding sites contain well positioned nucleosomes but nucleosomes are not necessarily well positioned on both sides of a TF bound region. This has been observed in our data at JUND binding sites (Paper III). In addition, some nucleosome positioning studies have shown that exons contain positioned nucleosomes marked with specific histone modifications reflecting the expression level [88, 89]. We have studied the nucleosome positioning pattern in HepG2 cells after the treatment with transforming growth factor β (TGFβ) (Paper III).

![Figure 5 Nucleosome positioning pattern. Figure illustrates the positioning pattern at the TSS and at nucleosome free regions. Figure adapted with the permission from Jiang & Pugh, Nature review genetics, 2009 [87].](image)

Post-translational histone modifications

Histone tails undergo a wide range of post-translational modifications at their N-terminal ends. Modified histones are involved in biological processes such as transcription, replication and maintain the chromatin states. There are at least eight known histone modifications. They are acetylation and methylation of lysines (K) and arginines (R), ubiquitinylation of lysines (K) and phosphorylation of serines (S) and threonines (T) (Figure 6) [82]. The pattern of modifications on histone tails are complex and to further increase the variation, many of the modifications can occur at different residues of the histone tail. However, some histone modifications are frequently observed in either active (euchromatin) or silent chromatin (heterochromatin) [90]. The
complex language of histone modifications is the subject of continuous research and great progress has been made in recent years.

Figure 6. Histone octamer and modification of their tails at N-terminal ends. Figure adapted with the permission from Paul A. Marks et al, Nature reviews Cancer 2001 [91].

Acetylation and methylation of histones are the most commonly studied chromatin marks. Acetylation is usually localized at the active regions of chromatin; on the other hand methylation plays a dual role and is associated with both transcriptional activation and silencing. As shown in Table 1, histone 3 lysine 27 tri-methylation (H3K27me3) is associated with transcriptional silencing. By contrast previous studies have shown that nucleosomes with H3K4me3 are linked to transcriptional activity. High levels of H3K4me3 is associated with 5’ regions of all active genes. There is a strong positive correlation between this modification and RNA polymerase II activity, transcription rates, histone acetylation [92, 93].

Table 1. Different histone marks and their functional association in the genome. Table obtained and modified from Mehdi Motallebipour PhD thesis (2008).

<table>
<thead>
<tr>
<th>Histone</th>
<th>Residue</th>
<th>Modification</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>H2A</td>
<td>Ser1</td>
<td>Phosphorylation</td>
<td>Mitosis</td>
</tr>
<tr>
<td>H2A</td>
<td>Lys5</td>
<td>Acetylation</td>
<td>Activation</td>
</tr>
<tr>
<td>H2A</td>
<td>Lys119</td>
<td>Ubiquitylation</td>
<td>Spermatogenesis</td>
</tr>
<tr>
<td>H2A</td>
<td>Thr120</td>
<td>Phosphorylation</td>
<td>Mitosis</td>
</tr>
<tr>
<td>Histone</td>
<td>Position</td>
<td>Modification</td>
<td>Activity</td>
</tr>
<tr>
<td>---------</td>
<td>----------</td>
<td>--------------</td>
<td>----------</td>
</tr>
<tr>
<td>H2A Ser139</td>
<td>Phosphorylation</td>
<td>Repair</td>
<td></td>
</tr>
<tr>
<td>H2A Tyr142</td>
<td>Phosphorylation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H2B Lys5,12,15, 20</td>
<td>Acetylation</td>
<td>Activation</td>
<td></td>
</tr>
<tr>
<td>H2B Ser14</td>
<td>Phosphorylation</td>
<td>Apoptosis</td>
<td></td>
</tr>
<tr>
<td>H2B Lys120</td>
<td>Ubiquitylation</td>
<td>Elongation?</td>
<td></td>
</tr>
<tr>
<td>H3 Arg2, 8</td>
<td>Methylation</td>
<td>Activation</td>
<td></td>
</tr>
<tr>
<td>H3 Thr3</td>
<td>Phosphorylation</td>
<td>Mitosis</td>
<td></td>
</tr>
<tr>
<td>H3 Lys4</td>
<td>Methylations 1-2-3</td>
<td>Activation</td>
<td></td>
</tr>
<tr>
<td>H3 Thr6</td>
<td>Phosphorylation</td>
<td>Activation</td>
<td></td>
</tr>
<tr>
<td>H3 Arg8</td>
<td>Methylation</td>
<td>Activation</td>
<td></td>
</tr>
<tr>
<td>H3 Lys4, 9,14,18, 27, 36</td>
<td>Acetylation</td>
<td>Activation</td>
<td></td>
</tr>
<tr>
<td>H3 Lys9</td>
<td>Methylation</td>
<td>Silencing</td>
<td></td>
</tr>
<tr>
<td>H3 Ser10, 28</td>
<td>Phosphorylation</td>
<td>Mitosis</td>
<td></td>
</tr>
<tr>
<td>H3 Thr11</td>
<td>Phosphorylation</td>
<td>Mitosis</td>
<td></td>
</tr>
<tr>
<td>H3 Lys27</td>
<td>Methylations 1</td>
<td>Activation</td>
<td></td>
</tr>
<tr>
<td>H3 Lys27</td>
<td>Methylations 2</td>
<td>More active than repression</td>
<td></td>
</tr>
<tr>
<td>H3 Lys36</td>
<td>Methylations 1, 3</td>
<td>Activation</td>
<td></td>
</tr>
<tr>
<td>H3 Thr45</td>
<td>Phosphorylation</td>
<td>Replication</td>
<td></td>
</tr>
<tr>
<td>H3 Lys56</td>
<td>Acetylation</td>
<td>Repair</td>
<td></td>
</tr>
<tr>
<td>H3 Lys79</td>
<td>Methylation</td>
<td>Activation</td>
<td></td>
</tr>
<tr>
<td>H4 Ser1</td>
<td>Phosphorylation</td>
<td>Activation</td>
<td></td>
</tr>
<tr>
<td>H4 Arg3</td>
<td>Methylation</td>
<td>Activation</td>
<td></td>
</tr>
<tr>
<td>H4 Lys5, 8, 12, 16</td>
<td>Acetylation</td>
<td>Activation</td>
<td></td>
</tr>
<tr>
<td>H4 Lys20</td>
<td>Methylation</td>
<td>Silencing</td>
<td></td>
</tr>
<tr>
<td>H4 Lys91</td>
<td>Acetylation</td>
<td>Repair</td>
<td></td>
</tr>
</tbody>
</table>

ChIP-seq studies suggest that histone modifications can predict the activity state of chromatin. In two large scale ChIP-seq studies of 19 methylated histone marks and 18 acetylated histone marks the localization of these marks was determined in CD4+T cells and correlated to promoters and other gene features (Table 1) [94, 95]. One can use the histone modification to predict gene expression [96]. It is also possible to detect enhancer elements using histone modifications. For example, by using acetylated histone 3 lysine 27 (H3K27ac) and mono-methyl histone 3 lysine 4 (H3K4me1), active enhancers can be distinguished from inactive/poised enhancers [97].
DNA methylation

DNA methylation is a biological process in which a methyl group is added to the 5’ positions of cytosine residues within CpG dinucleotide by DNA methyl transferases (DNMTs). CpGs occur in lower frequencies than expected in the human genome and 70% of the CpGs are present at promoters. DNA methylation is one of the important and widely studied epigenetic modifications that is involved in controlling cellular processes such as X-chromosome inactivation, transcription, embryonic development, cell differentiation, chromosome stability and genomic imprinting [98]. DNA methylation is mainly associated with gene silencing and heterochromatin formation (Figure 7). DNA methylation along with histone modification plays a role in gene activation, repression and nuclear structure. Two of the studies headed by Dirk Schubeler showed that local DNA sequences such as TF binding sequences (motifs) and respective TFs are the primary determinants of target specification and regulation of DNA methylation in mammals [99, 100].

A study in colon cancer has identified methylation alterations at lower CpG density sites called CpG island shores. CpG island shores are located neither at promoters nor at CpG islands but are in the proximity of 2 kb from CpG islands. Just like at CpG islands, methylation at CpG island shores is strongly correlated with gene expression [101, 102]. Hypermethylation at CpG island shores has been observed during normal tissue differentiation [101, 103]. This is in line with data which shows that methylation with the help of DNMTs plays a role in normal mammalian development [104].

DNA methylation has been associated with many diseases in addition to cancer, e.g. systemic lupus erythematosis (SLE), mental retardation, Prader Willi syndrome [98] and many other disorders. Prader Willi syndrome is due to effect in imprinting discussed further below. DNA methylation suppresses the expression of many genes and changes of methylation status i.e. hypo or hypermethylation has been observed in many cancers. Mutations in DNMT3A that reduce methylation have been observed in acute myeloid leukemia (AML). Differential methylation patterns have been observed in the pancreatic islet cells of type 2 diabetic individuals suggesting that methylation also plays a role in type 2 diabetes [105]. A human methylation study in rheumatoid arthritis patients revealed differential methylation in the major histocompatibility complex region and it was suggested to mediate the genetic risk of rheumatoid arthritis [106].
Figure 7. Illustrates different methylation mechanisms in the genome. Figure adapted with the permission from Smith Z.D & A. Meissner, Nature review genetics 2013 [104].
Imprinting and X-chromosome inactivation

Imprinting is an epigenetic phenomenon in which genes are expressed on the basis of their parental origin. There are about 150 imprinted genes observed in the mouse. Imprinted genes can form clusters of 20-3 700 kb of DNA. The expression of these genes is regulated by an imprinted control region (ICR). ICR is usually associated with parent origin specific methylation and histone modifications. Notable imprinted clusters are H19/Igf2 and Gtl2/Dlk1 in which the paternal alleles are methylated, and the non-coding RNAs Kcnqot1 and Snprn, where the ICRs of the maternal allele are methylated [107]. One well-known locus in tumorigenesis is IGF2-H19 in which loss of imprinting is associated with cancer [108] and in this region CTCF regulates the methylation pattern (Figure 8).

In mammals, females have two X- chromosomes and males have one X and one Y-chromosome. One of the X-chromosomes is inactivated in females by a 170 000 bp long non-coding RNA, X-inactivation specific transcript (XIST) [109, 110]. It plays an important role since XIST mediates the epigenetic silencing of the genes on the X-chromosome. XIST is also involved in cancer and in reprogramming [107].

![Diagram showing normal and cancer states of H19/IGF2 loci](image)

*Figure 8. The mechanism of methylation and imprinting at H19/IGF2 loci in the maternal and paternal allele. CTCF binds to the unmethylated maternal allele of the ICR and prevents activation of IGF2. Modified figure adapted with the permission from, Plass. C and P.D Soloway, European Journal of Human Genetics, 2002 [108].*
TGFβ signaling

TGFβ is involved in many cellular processes such as differentiation, proliferation, apoptosis and development. There are three highly homologous isoforms in the TGFβ family which are expressed in a tissue specific manner [111]. TGFβ is involved in immune cell maturation and regulation of the gene expression.

TGFβ signaling is complex and is mediated e.g. by the SMAD TFs. Abnormality in TGFβ signaling may contribute to cancer, diabetes, heart failure etc. TGFβ can act as both tumor suppressor and tumor promoter and is associated to many cancers [112]. TGFβ is one of the main regulators of liver disease and liver cancer and participates in liver inflammation, fibrosis and cirrhosis in hepatocellular carcinoma (HCC) (Figure 9). Due to the effect of the TGFβ signaling pathway in many diseases, it has been identified as a promising drug target [113].

![Diagram of TGFβ effects](image)

*Figure 9. Roles of TGFβ in different cells and its positive and negative effects Figure adapted with the permission from Dooley. S & P. ten Dijke, 2012, Cell and tissue research [114].*

Alcoholic steatohepatitis

Excess alcohol consumption may cause fatty liver disease or alcoholic steatohepatitis (ASH) which may lead to ballooned hepatocytes, mallory bodies, inflammation, fibrosis, cirrhosis and malignant transformation to HCC. There are similarities between ASH and the disease non-alcoholic steatohepatitis (NASH). One of the main sign of ASH is jaundice with fever and prox-
imal muscle loss [115]. Both genetic and environmental factors contribute to this disease. Aldehyde dehydrogenase (ALDH) genes, alcohol dehydrogenase (ADH) genes and catalase genes participate in alcohol metabolism and change in their function may contribute to the disease (Paper IV) and [116].

In hepatocytes alcohol is converted to acetaldehyde through oxidation and then to acetate. Downstream in the pathway ethanol affects the lipid metabolism through the TFs PPAR-α, AMPK and SREBP1 leading to an increase in fatty acid (FA) and steatosis. Lipopolysaccharides (LPS) in gram negative bacteria may also contribute to inflammation in the liver (Figure 10).

ASH can be treated with proper diet control, using lactulose and gut cleaning antibiotics. Corticosteroids can be used to inhibit the TFs AP-1 and NF-Kb. Other drugs such as Pentoxifylline and Infliximab antibodies that inhibit the TFs that contribute to this disease can be used to treat ASH. In the worst scenario, liver transplantation may be a final option.

**Figure 10.** An illustration of alcoholic steatohepatitis. Adapted with the permission from Lucey. M. R et al, New England Journal of Medicine, 2009 [115].
Materials and Methods

Cells and Tissues
In this study we used the HepG2 cell line which is derived from a hepatocellular carcinoma in a Caucasian male. HepG2 cells were cultured in RPMI1640 medium containing 10% fetal bovine serum (FBS), 1% L-glutamine at 37°C with 5% CO2. Human liver tissue from people diagnosed with alcoholic steatohepatitis and normal liver tissue were obtained from the Graz biobank, Medical University of Graz, Austria.

Methods
ChIP-seq is the main method used in this thesis. Apart from this, we performed Nucleosome-seq and RNA-seq for paper III. For the follow-up studies, we used semi-quantitative PCR, qPCR, ChIP-reChIP, co-Immunoprecipitation (Co-IP) and Western blotting.

Chromatin Immunoprecipitation (ChIP)
Cells or finely chopped tissue was crosslinked with formaldehyde to keep the protein-DNA interactions stable. Crosslinking was stopped by adding glycine and the chromatin was sonicated to 100-300 bp fragments by using Bioruptor. Sonication can be replaced by an enzyme called micrococcal nuclease 1 (MNase 1). MNase 1 cleaves the nucleosome free regions and leaves the nucleosomes. The main advantage of using MNase 1 for shearing is that chromatin marks can be mapped at the highest possible resolution and assigned to individual nucleosomes. MNase 1 was used to map the histone mark H3K4me3 in paper I and for mapping of the nucleosome positioning in paper III.

Immunoprecipitation was performed using the antibodies against the TFs or the histone marks of interest. The immunoprecipitated complex was attached to the solid phase of Protein-G agarose/sepharose beads or magnetic beads. After the extensive washing to remove all unbound protein-DNA complexes the protein-DNA complexes were eluted. Crosslinks were reversed the protein-bound DNA was treated with RNase A and proteinase K.
and extracted with phenol, chloroform followed by ethanol precipitation. A fraction of chromatin called input was not subject to immunoprecipitation and was saved and purified after reverse-crosslinking the DNA. Input DNA was later used to measure the background of the ChIP experiments. ChIP enriched DNA can be used to verify known regions by PCR or can be used in large scale sequencing.

**Figure 11.** Chromatin Immunoprecipitation. Chopped tissues or cells are crosslinked then cells are lysed and chromatin is sheared into small fragments and immunoprecipitated and DNA is extracted. This DNA can be used in PCR, qPCR, Microarray and in sequencing. Figure adapted with the permission of publishers, Philippe Collas, 2008 [117].
Verification of ChIP DNA by PCR

Enrichment in immunoprecipitated DNA can be verified by semi-quantitative PCR or preferably real time quantitative PCR (qPCR) in known target regions of the specific TF or histone modification. ChIP enrichment by semi-quantitative PCR is usually evaluated by separating the PCR products on an agarose gel and visualizing under UV light. qPCR with SYBR green is a more sensitive and accurate way of measuring ChIP enrichment as it enables measurements to be done in the exponential phase of amplification. Serial dilutions of input DNA can be used to obtain a standard curve and from this it is possible to quantify the ChIP enrichment. Enrichment can be compared with that in a negative control ChIP using IgG as antiserum or with the enrichment obtained from unbound regions of the TF or histone mark.

Library preparation for ChIP-sequencing

ChIP combined with high-throughput sequencing is a powerful method to map the regulatory elements across the genome. For ChIP-seq, ChIP-DNA is end-polished with end-polishing enzymes T4 DNA polymerase, Klenow DNA polymerase and T4 polynucleotide kinase to create blunt ends. The end-repaired DNA fragments are A-tailed for the Illumina pipeline and ligated with adapters. The ligated fragments are amplified using adapter specific primers and the number of cycles should be low to reduce duplicates. Since the chemistry of the sequencing instruments only allows reading of smaller fragments, the amplified PCR product is separated on an agarose gel and the fragments are size selected and purified. It is also possible to use AMPure XP beads to size select the fragments. The purified DNA fragment sizes should be checked on a BioAnalyzer and can be used for large scale sequencing. The above protocol is an outline for different instruments but, each NGS platforms have their own adapters and library protocols. There may be changes between the steps for example; SOLiD adapters need to be nick-translated before the amplification.

Next generation Sequencing

Sanger sequencing has been the main source for any type of sequencing until the launch of the 454 sequencing instrument in 2005. The 454 technique is based on pyro-sequencing and uses sequencing by synthesis [118]. The initial read length for 454 was 100-150 bp but, now it is up to 1000 bp. In 2006, Illumina-Solexa released their Genome Analyzer which is also based on sequencing by synthesis. After that Life technology’s SOLiD (Sequencing by Oligo Ligation Detection) was launched into the market. Today there are
many other NGS instruments on the market, for example PGM and Proton (Life tech/Ion torrent), RS (Pacific Biosciences) and ARGUS (Opgen) (Table 2). For all platforms the sequence reads come from the ends of the ChIP-enriched fragments. The reads are aligned to the genome and peaks are called using resources like MACS available on the internet or using in-house technology.

Table 2. Different NGS platforms, their sequencing type, chemistry.

<table>
<thead>
<tr>
<th>Platform</th>
<th>Sequencing Type</th>
<th>Amplification</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina</td>
<td>Synthesis</td>
<td>Bridge PCR</td>
<td>Short reads</td>
</tr>
<tr>
<td>SOLiD</td>
<td>Ligation</td>
<td>Emulsion PCR</td>
<td>High accuracy</td>
</tr>
<tr>
<td>Ion Torrernt</td>
<td>Synthesis, H+ detection</td>
<td>Emulsion PCR</td>
<td>short Time</td>
</tr>
<tr>
<td>454</td>
<td>Synthesis</td>
<td>Emulsion PCR</td>
<td>Long reads</td>
</tr>
<tr>
<td>Pac Bio</td>
<td>Synthesis</td>
<td></td>
<td>Single molecule sequencing</td>
</tr>
<tr>
<td>Ion Proton</td>
<td>Synthesis</td>
<td>Emulsion PCR</td>
<td>short Time</td>
</tr>
</tbody>
</table>

For this thesis, we have used Illumina and SOLiD platforms and I will briefly discuss their chemistry and current status.

Illumina sequencing system is one of the leaders in NGS. It is based on sequencing by synthesis technology. In this system, denatured single stranded libraries with adapters are hybridized to the flow cell. Formamide is used for denaturation and Bst polymerase is used for the extension and amplification to create clusters of individual molecules. The double stranded library is made single stranded. Each base is labeled with a unique fluorophore and one base at a time is incorporated and detected by a CCD camera. The HiSeq 2000 machine can generate up to 200 GB of data per run with the read length 2x100 bp.

SOLiD sequencing is based on ligation. Quantified SOLiD libraries are added to paramagnetic beads and emulsion PCR is performed. Beads with template are enriched and the selected beads undergo 3’ modification to attach covalently to the slide or flow cell. During sequencing, a universal sequencing primer hybridizes to the P1 adapter on the template beads and an eight base labeled probe ligates to the sequencing primer. The first base of this probe contains ligation site, a cleavage site at the 5\textsuperscript{th} base and the last base contains one of four fluorescent dyes. The images from the fluorescent signal are recorded, the probe is cut at the cleavage site and the next probe is ligated to generate a complex data in which each base is sequenced twice to achieve an accurate sequence. SOLiD has recently launched the 5500 W instrument in which a flow chip is used in place of emulsion PCR for the amplification. Using this system we can get the data with 99.99% accuracy.
ChIP-reChIP

ChIP-reChIP is a method to detect two different proteins or histone marks present at the same location in the genome. In this method, ChIP is performed with an antibody against the first protein. After the elution of the protein-antibody-DNA complex from the solid phase; the complex is again immunoprecipitated with a second antibody against the second protein. This complex is eluted from the solid phase and DNA is extracted as described in the ChIP method. Enrichment of ChIP-reChIP DNA can be verified by PCR with known genomic loci to confirm that the studied proteins are located at the same region in the genome. ChIP-reChIP enriched DNA could also be used in sequencing to map all common target regions.

Co-Immunoprecipitation

Co-IP is a method used to find protein-protein interactions *in vivo* using Western blotting. In this method, the extracted cell lysate is incubated with an antibody on a solid phase and washed vigorously to remove the unbound proteins. Later, the purified protein lysate is boiled and separated on a poly-acrylamide gel and the separated proteins are transferred on to a PVDF membrane. The membrane is probed with the antibody against the interacting protein of interest.

TGFβ treatment, nucleosome and RNA preparation

HepG2 cells were cultured as described above and serum starved overnight and then treated with 2.5ng/mL of TGFβ for one hour. The cells were used either for nucleosome preparation or RNA extraction. Nuclei were prepared by homogenization with a Dounce homogenizer and the nuclear suspension was slowly layered with the buffer containing 30% of sucrose. The nuclei were collected and treated with 300U of MNase 1 at 37°C for 5 minutes. Mono-nucleosome size DNA was prepared and DNA was gel purified before the library preparation (Paper III).

Total RNA was extracted by using TRIzol-chloroform method according to the manufacturer’s protocol. The quality of RNA was measured by using a BioAnalyzer. The SOLiD transcriptome whole library preparation kit (Ambion) was used to prepare the strand-specific library.
Present Investigations

Aims of the present studies

The main aim of this thesis was to map regulatory elements in liver cells and disease tissue by high-throughput sequencing. The specific aims are as follows:

**Paper I and II**
To investigate the TF binding sites for FOXA1, FOXA2, FOXA3, HNF4α, GABP and their interaction in the human genome. The TFs were mapped genome wide in the liver cell line HepG2 using ChIP-seq.

**Paper III**
To evaluate the effect of TGFβ on nucleosome positioning and RNA expression by nucleosome and RNA-sequencing.

**Paper IV**
To identify disease mechanisms of alcoholic steatohepatitis by mapping the active chromatin marks H3K4me1, H3K4me3 and H3K27ac and the TF USF1 in diseased and normal liver tissue.
We performed ChIP-seq for whole genome analysis of the TFs FOXA1, FOXA3 and the histone mark, H3K4me3 on the SOLiD platform. We obtained FOXA2 data from paper II. From ChIP-seq data, we identified 8 175 FOXA1 peaks, 4 598 FOXA3 peaks and 41 480 H3K4me3 regions that correspond to 160 000 nucleosomes across the genome. Only a fraction of FOXA1 (5.2%) and FOXA3 (12.2%) peaks were located within 1kb of TSS and the majority of them were located more than 1 kb away from a TSS. H3K4me3 signals were mainly localized downstream of the TSS. Of the H3K4me3 regions 42% were located within 1 kb of a TSS. To identify promoter patterns, we compared H3K4me3 data with CAGE (Cap-analysis of gene expression) tag data of HepG2 and found that 28% of H3K4me3 regions were not associated with any known transcript.

To study the relationship between FOXA proteins, we compared the ChIP-seq data of the TFs FOXA1, FOXA2 and FOXA3 and found that these three datasets had 2 304 regions in common. This suggested that the TFs not only interact with DNA but also with each other. When we tested these interactions by Co-IP and ChIP-reChIP, we found that FOXA1 and FOXA3 bind to FOXA2, but there is no evidence of binding between FOXA1 and FOXA3. Based on these results, we classified the binding sites of these factors pair-wise and found that 51% of FOXA2 and FOXA3 binding occurs within 5kb of TSS whereas only 10% of FOXA2-FOXA1 bind within the same distance.

We clustered H3K4me3 signals into seven clusters by K-means clustering and each of these clusters has individual H3K4me3 signatures. All seven clusters showed different levels of signals and expression. We found that the genes with the highest expression had high levels of H3K4me3 enrichment in HepG2 cells. Clusters I-III had the highest H3K4me3 enrichment upstream of the TSS and we found that it could be due to bidirectional promoters. When we looked at the CpG density at bidirectional promoters, we found that there was high CpG density in all clusters except in cluster V. Cluster V contains genes with TATA and CAAT boxes. Comparisons of the clusters with CAGE tag data at TSSs of a gene revealed that 30% of genes in each of these clusters had bidirectional promoters. Confirming of this bi-directionality, we also found an H3K4me3 double peak at FOXA1-2-3 binding sites. Another finding was that 11% of the genes in cluster-I had a FOXA3 binding site within 1kb of TSS.

In summary, we showed that FOXA1 and FOXA3 TFs had differential binding patterns in HepG2 cells and no verified interaction in vivo. Most of the FOXA1, FOXA2 and FOXA3 bindings coincide with H3K4me3 and the data suggested a directionality of this histone mark even though there is no
known TSS in close vicinity. We also showed that ChIP-seq can be used to find mono-allelic binding of TFs or histone marks at SNPs. SNPs present inside the regulatory regions have the capacity to differentially control the activity of many genes and some of these SNPs could be associated with disease. This tells us that ChIP-seq can also be used to identify disease-associated SNPs.

Paper II

Molecular interactions between HNF4a, FOXA2 and GABP identified at regulatory DNA elements through ChIP-sequencing

In paper II we performed ChIP-seq of three TFs FOXA2, HNF4α and GABP (NRF2) in HepG2 cells using the Solexa/Illumina platform. Before starting this work, a paper was published by the group on USF1 and USF2 which found that most of the USF2 binding was at distal positions and its binding coincided with motifs for FOXA2 and HNF4α. They also found that there are GABP binding motifs in close vicinity of TSSs bound by USF.

We found 18,693 HNA4α binding sites, 7,253 FOXA2 binding sites and 3,060 sites for GABP in HepG2 cells. Among these, 10% of HNF4α, 6.3% of FOXA2 and 85.2% GABP binding sites were located within 500 bp of a TSS so most of the HNF4α and FOXA2 binding sites are far from TSSs. We also found that there was a large number of overlap between HNF4α and GABP peaks at promoters. We have identified many variants of expected motifs for FOXA2 and HNF4α and these motifs were highly similar to the motifs already present in Transfac database. For GABP, we were able to identify two motifs that were enriched equally for both GABP/NRF2 and for NRF1. Although many GABP peaks contain motifs for both NRF1 and NRF2, NRF1 motifs have a tight distribution in the peaks and some of the GABP peaks contain an NRF1 motif only. Correlation of HepG2 expression data with ChIP-seq data showed that the genes bound by HNF4α and FOXA2 had higher expression than average expression in HepG2 cells but, GABP did not show such an association in HepG2.

We also applied Co-IP studies as in paper I to see the possible interaction between HNF4α, FOXA2 and GABP and identified that these factors were interacting in vivo but, we did not see any interaction with USF2 except the ChIP-seq overlap between the data sets.

In conclusion, we showed that ChIP-seq is a powerful method to map the regulatory regions in the genome. Apart from this, ChIP-seq can be used to identify SNPs associated with common diseases in TF binding regions.
In this paper, we performed high-throughput sequencing of nucleosomal DNA and total RNA using the SOLiD platform before and after 1 hour of TGFβ treatment. When we looked at the nucleosome positioning pattern around the TSS, we found that there was a well-positioned nucleosome downstream of TSS, another well-positioned nucleosome upstream of the TSS and a nucleosome free region between them. However, we did not see any flanking positioned nucleosomes at 33% of TSSs.

We also observed a well-positioned nucleosome at exons as seen before in CD4+ T cells. There was a higher density of nucleosomes in exons than in introns and intergenic regions. When we separated the nucleosomes based on their fuzzyness, exonic regions contained more fuzzy nucleosomes than phased nucleosomes. This is compatible with the fact that there are AT-rich sequences flanking exons which disfavor positioned nucleosomes.

When we looked the regulatory potential of TGFβ1 in HepG2 cells, 81% of the TGFβ+ nucleosome positions matched with TGFβ- and a surprising 11% did not match with each other. After the strict filtering of the data, 24318 nucleosome depleted loci were identified in TGFβ+ cells. When we looked at the expression of genes with nucleosomes depleted in the vicinity, after TGFβ1 treatment, 20% of those genes changed their expression.

To understand the role of TGFβ in regulation, we searched for TF motifs using the JASPAR database at nucleosome depleted regions and found that these regions were highly enriched with TF motifs. Some of these TFs, for example SPI1, SMADs, FEV, ELF5, and HNF4α etc., are regulated by TGFβ1. By coupling these results with the RNA-seq expression change after TGFβ1 treatment it was found that many nucleosomes with TF motifs that changed positions were related to the gene expression change. Of nucleosomes in introns that had TF motifs and that were depleted after treatment, 61% were associated with exon expression change. Out of 2469 intronic distal loci, 1173 were associated with expression change in 306 genes.

When we compared the expression change after TGFβ1 treatment, 591 genes were up-regulated and 196 were down-regulated. Genes with neuronal function, ion channels and ion transporters, neuro-transmitters, G-protein coupled receptors were up-regulated and the TFs that were known to be regulated by TGFβ1 were among the most up-regulated genes.

Using HNF4α ChIP-seq data from paper II, we identified 37 candidate HNF4α binding sites at nucleosome depleted regions in TGFβ- cells compared to TGFβ+ cells and the ChIP-qPCR validations supported our findings.

In summary, our data showed that 1 hour of TGFβ1 treatment depleted many nucleosomes and some of these depletions were associated with
change in the expression of genes. In addition, chromatin is dynamic and plays an important role in transcriptional regulation in general. TGFβ1 therefore has an important impact on chromatin organization and function.

Paper IV

*ChIP-seq in steatohepatitis and normal liver tissue identifies candidate disease mechanism related to progression to cancer.*

To learn more of disease mechanisms we mapped the active chromatin marks H3K4me1, H3K4me3 and H3K27ac and the TF USF1 in ASH patients and in controls using ChIP-seq. We identified 2 054 USF1 peaks in control and 1 766 for ASH and out of these 44% of peaks in control and 54% in ASH contained a USF1 motif. Peak calling for chromatin marks gave 10 256 H3K4me1, 14 668 H3K4me3 and 11 656 H3K27ac peaks in the control sample and 10 674 for H3K4me1, 18 358 H3K4me3 and 12 385 H3K27ac in ASH sample. For each histone modification, the top 1 000 differentially enriched peaks were identified.

When we looked at the genomic localization of differentially enriched histone modification peaks at +/-2 kb of a TSS, a majority of the H3K4me3 peaks were located at TSS as expected, whereas the peaks for H3K4me1 and H3K27ac were distributed both at TSS and non-TSS regions. Surprisingly there was an increased number of peaks in ASH patient compared to control. There were 1 600 genes that contained histone modification peaks only in control and 3 000 genes contain peaks only in patient. Comparison of differentially enriched USF1 peaks with histone modification peaks gave 7 overlaps which suggests that USF1 is just one among many factors that may contribute to differential gene regulation.

Gene ontology of genes associated with differentially enriched histone modification peaks revealed that the genes from control sample were enriched with different metabolic process whereas ASH patient sample were enriched for cancer associated genes. This suggests that the disease in this patient is progressing towards cancer. We observed different genes associated with both ASH and NASH which indicates that common pathophysiological mechanisms exist. We looked at the pattern of histone modifications at the genes involved in alcohol metabolism and disease (*CD14, TLR4, TNF4α and TGFβ*) and found evidence that they contribute to disease in this case.

Using SNP calling in differentially enriched regions, we identified 178 potentially functional SNPs that are present in the GWAS catalogue and additionally 1 237 GWAS SNPs with high LD were identified in the patient and 456 in the control. The majority of these SNPs were located at the binding motifs of different TFs and DHS sites which suggest a regulatory potential of these SNPs. Some of these SNPs are within the binding motif of the
TFs associated with ASH and NASH. We also identified 383 novel SNPs from ASH sample that might be involved in the etiology of the disease.

RNA expression with qPCR at differentially enriched USF1 sites did not follow USF1 ChIP-seq pattern. We compared the histone modification ChIP-seq data with RNA expression by qPCR and found that RNA expression at majority of the sites followed the ChIP-seq pattern. In summary, our study identified many genes associated with liver disorders and we also found many genes associated with cancer pathways. This illustrates that analysis of chromatin marks in disease tissue may provide useful insights into disease mechanisms.
Concluding remarks and future perspectives

For this thesis, we used ChIP with next generation sequencing to map TFs, histone marks and nucleosome positions and to determine gene expression.

Rapid advances in NGS technology have allowed the mapping of regions in the genome in a short time span. Due to high competition, the price for sequencing and reagents has dropped rapidly. At the time of our ChIP-seq paper I and II [10, 119], very few genome-wide ChIP-seq studies had been performed. Today there are thousands of ChIP-seq datasets, mainly from the ENCODE project, and many additional have been published.

We have mapped FOXA TFs, HNF4α, GABP, USF1 and a few histone marks by ChIP-seq. In paper I and II, we showed that ChIP-seq not only can be used to map the TF binding sites, but also to identify protein-protein interactions genome-wide. We also developed a method to identify allele-specific binding of TF or histone modification and identified many functional SNPs. During the time of this thesis work, it was found that the TFs studied in paper I and II are sufficient to reprogram fibroblasts to liver cells which illustrates the importance of these factors. In paper III, we mapped the nucleosome positions across the HepG2 genome before and after TGFβ1 treatment and measured the nucleosome dynamics in response to TGFβ1. We have seen a change in RNA expression of genes at nucleosome depleted regions after treatment. Nucleosome positioning data is valuable for delineating gene regulatory regions. After the extensive studies in the liver cell line HepG2, we focused on liver tissue from ASH patients and controls. We identified differences in histone marks which provide insights into ASH disease mechanisms and in this case its progression towards cancer. There was a correlation between ChIP-seq of histone marks and RNA expression which suggests that chromatin marks could be used to get further insights into disease processes.

Advances in NGS technology has facilitated introduction of new ChIP based methods. ChIA-PET was developed to map binding sites for nuclear proteins and long way chromatin interactions globally and has been applied to TFs and RNA Pol II [120, 121]. ChIP-exo was developed to map TFs at single base-pair resolution [122] and will be useful to identify TF binding sites more precisely.

There are ~1 850 TFs and it is not an easy task to map all TFs in one cell and even more demanding to map them in all cell types and tissues. Consequently there is a need to develop high-throughput methods. Recently an
HT-ChIP method was developed for the preparation of sequencing library samples using a pipetting robot [123]. Our group also developed a new HT-ChIP-library preparation method with which we can prepare many ChIP-libraries in a short time span and perhaps more efficiently than any other method (Manuscript in preparation).

A lot of things have happened in genetics and genomics since the discovery of the structure of DNA exactly 60 years ago, but we still need answer to many questions. Each and every regulatory element in the genome probably has an important function and identification of their location and the gene(s) they act on will provide useful information. We still need to understand the epigenetic behavior in cancer cells and normal cells and high-throughput mapping of several TFs in tumors by ChIP-seq may be a good strategy to understand disease mechanisms. Mapping of regulatory regions, non-coding RNAs, mRNA in dynamic conditions i.e. after treatment with different biological agents may provide clues for many diseases. NGS technology will soon play a great role in disease diagnostics in patients since one can sequence the entire genome in one or a few days using novel bench-top instruments. Today interpretation of genomic sequences is limited to the protein coding regions, but soon we will also be able to deduce the functional consequences of variation in gene regulatory regions.
Acknowledgements

This thesis work would not have been possible without following people. I would like to express my deepest gratitude towards the people who helped me directly or indirectly.

First of all, I would like to thank heartfully to my Supervisor, Professor Claes Wadelius, who gave me the opportunity to pursue my PhD in his group. I am very much thankful for your immense support, valuable discussions and suggestions.

I also would like to thank my co-supervisor Professor Jan Komorowski for great collaboration and group meetings.

I am very much thankful to the former members of the group. Mehdi Motallebi, for all your help during my early days in the lab. Ola Wallerman, for taking care of large datasets and for all the help whenever needed. The present members in the lab, Gang Pan, for all the discussions we have during lunch and work. Marco Cavalli and Helena Nord, for all your help and fun in lab.

Thanks to the present and former members of Komorowski’s lab especially, Susanne, Stefan and Robin for all the data analysis, group meetings we have together.

Professor Marta Alarcon, who gave me the opportunity to work in her lab. Sergey Kozyrev, for all your suggestions and help. Special thanks to Hong Yin, who introduced me to Claes.

Lena Åslund, for all the help during teaching period and for the discussions in the journal club. Marcela Ferella, for being a good friend in the office.

I am grateful to the members of Uppsala Genome Center (Ulf, Inger, Katherina and Cecelia) and Seq & SNP platform for the sequencing service. Thanks to the people at Life technologies, California for the great collaboration.
Thanks to all other Rudbeck friends, Angelica, Ammar, Nikki, Hamid, Sara E, Tanmoy, Jitendra and Kali.


Thanks to all my friends in India, Vidya, Renuka, Aruna, C.V, Raja.B, Raja.O and Ramana. Thank you so much Siva for always taking care of my Tirumala trip.

Thanks to my sisters and brother-in-laws, Aruna-Nageswar Reddy and Prameela-Prathap for all your help. Eswar Reddy mama for always being there for anything. Peddnanna, Jayarami Reddy for all your support and Se-sha Reddy anna for being a good friend than a brother. My wife’s parents, Balarami Reddy mama and Ratna atha, for all your help and support. My Padmavathi pinni, for all your care during my childhood and school days.

Thanks to my parents, Nagi Reddy and Lakshmidevi for your unconditional support even my interests were beyond your limits. Simple thanks may not enough.

Many thanks to my wife, Vijayalakshmi for all the support and love. Finally, my little daughter Akshara, for all the joy you brought into our lives.
References

24. Gagniuc P Fau - Ionescu-Tirgoviste, C. and C. Ionescu-Tirgoviste, Eukaryotic genomes may exhibit up to 10 generic classes of gene promoters. (1471-2164 (Electronic)).


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 904

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.

Distribution: publications.uu.se
urn:nbn:se:uu:diva-198579