*Year :* 2016

# Transcriptional programs during mammalian cell prolifération

Rib Leonor

Rib Leonor, 2016, Transcriptional programs during mammalian cell prolifération

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive http://serval.unil.ch
Document URN : urn:nbn:ch:serval-BIB_51F1A1F006D84

UNIL | Université de Lausanne
Faculté de biologie et de médecine

Centre Intégratif de Génomique

# Transcriptional programs during mammalian cell proliferation

**Thèse de doctorat ès sciences de la vie (PhD)**

présentée à la

Faculté de biologie et de médecine de l'Université de Lausanne

par

Leonor Rib

Informaticien diplômée par l'Universitat Politècnica de Catalunya, Espagne

## Jury

Dr. Werner Held, President, Université de Lausanne, Switzerland
Dr. Winship Herr, Thesis supervisor, Université de Lausanne, Switzerland
Dr. Nicolas Guex, Thesis co-supervisor, Swiss Institute of Bioinformatics, Switzerland
Dr. Philipp Bucher, Expert, École polytechnique fédérale de Lausanne, Switzerland
Dr. Liliane Michalik, Expert, Université de Lausanne, Switzerland
Dr. Ueli Schibler, Expert, Université de Genève, Switzerland

Lausanne, 2016

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

**Ecole Doctorale**
**Doctorat ès sciences de la vie**

# Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

| | |
|---|---|
| **Président · e** | Monsieur Prof. Werner **Held** |
| **Directeur · rice de thèse** | Monsieur Prof. Winship **Herr** |
| **Co-directeur · rice** | Monsieur Dr Nicolas **Guex** |
| **Experts · es** | Monsieur Dr Philipp **Bucher** |
| | Madame Dre Liliane **Michalik** |
| | Monsieur Prof. Ueli **Schibler** |

le Conseil de Faculté autorise l'impression de la thèse de

## Madame Maria Leonor Rib Perez

Master Bioinformatique de l' "Universitat Politecnica de Catalunya, Espagne

intitulée

## Transcriptional programs during mammalian cell proliferation

Lausanne, le 6 juillet 2016

pour le Doyen
de la Faculté de biologie et de médecine

Prof. Werner Held

# Abstract

Gene transcription is a precise and complex process that initiates the expression of the genetic code. Transcription of genes can lead to highly coordinated cellular processes such as cell proliferation and eventually to physiological changes such as mouse liver regeneration. RNA polymerases are the major enzymes involved in transcription and their action is regulated by different elements that bind to the chromatin and modify its activity. Host Cell Factor 1 (HCF-1) is one case of a transcriptional co-regulator. The HCF-1 precursor protein is proteolytically cleaved for its maturation into two subunits (HCF-$1_N$ and HCF-$1_C$) that remain bound non-covalently and become active. The mature HCF-1 regulates transcription via chromatin association with transcription factors and chromatin remodelers in gene promoters. Furthermore, HCF-1 is required for proper progression of mammalian cell division, especially for the passage from G1 to S phase and for proper mitosis. The advent of high-throughput sequencing technologies in the past ten years has permitted transcriptional studies at a genome-wide scale. A genome-wide study of HCF-1 showed that it is a common component of active CpG-island promoters and coincides with the occupancy of the transcription factors ZNF143, THAP11, YY1 and GABP.

In this dissertation, I show novel insights about the genome-wide regulation of mammalian transcription in cancerous and differentiated cells. Initially I evaluate the use of paired-end sequencing to study genome-wide binding of transcription regulators to the chromatin. This technology proves to have advantages compared to the traditional single-end sequencing. Subsequently, I report new insights about the chromatin binding of HCF-1 along the cell division of HeLa cells. In the *CDC6* promoter, paired-end sequencing revealed two HCF-1 binding sites with different underlying DNA motifs associated to the transcription factors E2F1 and THAP11/ZNF143. This suggests that HCF-1 could bind to the chromatin through different transcription factors in the same promoter. Interestingly, the individual association with these transcription factors appears to vary during the course of the cell cycle. In this work, I also investigate transcription regulation in the mouse liver. I initially characterize the genome-wide transcriptional responses to partial hepatectomy in the mouse liver, showing that the mouse liver undergoes two different transcriptional cycles: a sham-like cycle and second cycle linked to cell proliferation. Additionally, I describe that the genic accumulation of H3K36me2 in the regenerating mouse liver and in HeLa cells accumulates at the 5' end of transcriptional units whereas H3K36me3 accumulates towards the 3' end. This observation was already reported only in Drosophila which suggests potential similar mechanisms for Pol2 elongation. And lastly, I show that HCF-1 in the mouse liver is a versatile component for the regulation of genes involved in diverse cellular functions in which the two HCF-1 subunits display different chromatin associations.

# Résumé pour le grand public

Nos cellules comportent chacune deux séries d'instructions que nous avons héritées de nos parents. Ces instructions sont compactées de façon remarquable pour tenir dans la très petite taille du noyau cellulaire. Ceci est en parti réalisé en enroulant la séquence d'ADN autour de protéines appelées histones, formant la chromatine. Cependant, les cellules possèdent divers mécanismes de régulation pour la lecture des instructions, appelée transcription. Ceci implique la participation d'ARN polymérases (les lecteurs), ainsi que d'autres facteurs de régulation comme des co-activateurs ou des co-répresseurs, par exemple Host Cell Factor 1 (HCF-1). Ils coexistent tous avec des facteurs de remodelage de la chromatine qui peuvent lire et modifier les marques épigénétiques sur les histones. Au début du 21ème siècle, de nouvelles technologies ont été mises au point pour permettre la visualisation de la position de ces tous petits éléments sur l'ensemble des gènes contenus dans les cellules. Cette révolution dans les sciences du vivant contribue à élucider les mécanismes de régulation de la transcription lors de processus cellulaires tels que la division cellulaire.

Dans ce manuscrit, je présente différents modèles de régulation de la transcription des gènes que j'ai observés en étudiant les positions de régulateurs. J'ai pu étudier ceci dans deux types de cellules : a) des cellules humaines cancéreuses qui se multiplient constamment et b) des cellules saines du foie de souris. Le foie est un organe possédant une capacité de régénération remarquable. Lorsqu'il est lésé, il a la capacité de multiplier les cellules restantes pour restaurer sa masse et sa fonction. Lors de la régénération, j'ai observé que les ARN polymérases peuvent lire les gènes rapidement ou lentement selon le besoin. De plus, la position des marques épigénétiques H3K36me2 et H3K36me3 lors de la régénération donne des pistes sur le mécanisme de lecture.

J'ai également étudié le rôle du cofacteur de transcription HCF-1. A la fois dans les cellules cancéreuses et dans les cellules du foie, HCF-1 est une protéine polyvalente qui peut agir différemment d'un promoteur à l'autre. HCF-1 est de plus impliqué dans la lecture de gènes associés à des fonctions cellulaires très variées. Ceci en fait une protéine très intéressante à étudier, puisqu'elle permet d'obtenir des pistes sur différents modes de régulation.

En conclusion, bien que les deux séries d'instructions reçues de nos parents soient statiques, les éléments qui interagissent avec elles sont extrêmement divers et dynamiques, et agissent de manière précise pour que les cellules réalisent leurs fonctions lorsque nécessaire. Une meilleure compréhension de cette diversité et de cette précision permettra dans le futur d'aider à concevoir de meilleures drogues pour traiter les maladies.

# Acknowledgements

During these five years of doctoral studies I have been fortunate to have the support and teachings of some people that I would like to acknowledge here.

First, I would like to acknowledge Winship Herr, my supervisor, for accepting the challenge to teach a computer scientist on how to do research in molecular biology during these five years, and for even accepting that I conducted my own research in the wet lab. I appreciate very much his teachings on the value of doing good science and the transparency with which he does science as this has also given me the opportunity to develop many complementary soft skills.

I would also like to thank Nicolas Guex, my co-supervisor. I would like to thank him for his availability during these five years, his teachings and his motivation. The collaboration with him has been very interesting and creative.

I would like to thank all the members of the lab for their support, teachings and friendship. I would like to specially thank Dominic Villeneuve, who from the first moment has been very inspiring to me. He answered all the questions I asked him regularly about biology and experiments, and I was very fortunate to have him supervising most of my experimental research making it a very inspiring and exciting scientific experience. I would also like to thank Viviane Praz, because she has always been available to answer my questions. I learned a lot from our discussions that often took longer than expected. I would also like to especially thank Laura Sposito, as I have found it very easy to collaborate with her. I would like to particularly thank her for the long discussions and support while I did my experimental research, and for her friendship in the difficult moments. Further, I would like to thank very much Nathalie Clerc, the administrative assistant in the lab, for her constant administrative support in any issue that could happen. Her work is of great help. I would also like to thank Harmonie Senez for the thorough translation of the abstracts into French.

Furthermore, I would like to thank the CycliX consortium for all the feedback. Especially, I would like to thank Nouria Hernandez, the coordinator of the consortium, who was always very supportive and gave very interesting inputs for my projects. Also, I would like to thank Ueli Schibler for his advice in the project of the mouse liver regeneration. He suggested to do Sham controls, which has been key for my research, and he also shared protocols for the mice work. I would also like to thank Cristian Carmeli because I enjoyed very much collaborating with him and I learnt a lot about the way he analyzed the data. Also, I would like to thank Olivier Martin. He developed the viewer I have used all these years for inspecting the sequencing data and he always

provided me support in any question or did the necessary improvements, in a very efficient way. I would also like to thank him for his kind friendship all over these years.

I would also like to thank Keith Harshman, the director of the IECB doctoral program, for his support, because he believed that computer scientists could learn to do integrated experimental and computational biology, and thanks to this I had this great opportunity to do it. Further, I would like to thank Corinne Dentan, the administrative assistant in the IECB doctoral program. She has always been very efficient and has given me support for any question or issue that could happen.

During these years, I could also be part of the Vital-IT department at the Swiss Institute of Bioinformatics (SIB), as an embedded bioinformatician. I would like to thank Ioannis Xenarios for making me part of Vital-IT and thus allowing me to participate in many scientific activities for computational biologists. Further, I would like to thank the rest of the Vital-IT team, particularly to those maintaining the computing infrastructure at the UNIL campus. They have always been very nice and efficient in their work and I could learn a lot from them on how to work better in such infrastructure.

I collaborated with members in the Genomics Technologies Facility (GTF) to which I am thankful. Particularly, Keith Harshman, the director of the GTF, for his support in all the sequencings we did. I would also like to thank particularly to Corinne Peter and Alexandra Paullison, for their kind support and teachings about sequencing technologies.

I had a very important personal support during these years from my family and friends. I would like to give my greatest thank to Alfonso Buil, my husband, for his support in the good and the difficult moments throughout these years. For being such a great person and always being able to draw a smile on my face. I would also like to thank my parents, Antonia Perez and Juan Rib, and my brothers and their partners, Aris, Esther, Humberto and Carolina. I would like to thank them for their constant love and support even from the distance. Further, I would like to thank Alfonso's family as they always gave me their kind support to hang on and complete my thesis. Thanks also my dear close friends in Barcelona: Almudena, Eli and Oriol, Espe and Albert, Isa, Laura and Santi, Mari, Marta and Martí, Pablo and Sharon, and Sergio, as even from the distance, they have taken care of me. Last but not least, I would like to thank Bill Stone, my dear friend and mentor in Barcelona. Thanks to his advice and support I could take the decision to come to the lab of Winship Herr. I will always be very thankful for opening me the door to this opportunity and for being a good friend all these years.

# Table of contents

# List of figures and tables

## Chapter I: Introduction

## Chapter II: Evaluation of the paired-end sequencing technology for ChIP-seq studies

## Chapter III: HCF-1 chromatin binding in HeLa cells along the cell division cycle



## Chapter IV: Genome-wide transcriptional responses to partial hepatectomy in the mouse liver

## Chapter V: Insights on the accumulation of H3K36me2 and H3K36me3 in the mammalian genome

## Chapter VI: The role of HCF-1 in the mouse liver chromatin

# Thesis scope

The objective of this thesis is to provide insights of the regulation of transcription during mammalian cell proliferation in both differentiated and cancer cell lines.

Chapter I provides a general introduction into the research fields relevant for my PhD thesis, and, from Chapters II to VI, I describe results from my research where I have integrated experimental and computational approaches.

Chapter II describes a comparison between the traditional single-end sequencing technology and the more recent paired-end sequencing technology. I evaluate whether paired-end sequencing is more useful for handling redundancy in the data and I also show interesting insights of the usage of different in-silico size selections to study the chromatin mark H3K4me3.

Chapter III provides insights of the genome-wide binding of HCF-1 throughout the cell division cycle in a human cancer cell line. Single-end sequencing data was analyzed that suggests a preference for HCF-1 to regulate progression of the cell cycle from G1 phase to M phase. Additionally, the later use of paired-end data allows the precise identification of insights of the binding of HCF-1 in the promoter of the *CDC6* gene.

In Chapter IV, I characterize the genome-wide transcriptional responses to partial hepatectomy in the mouse liver. An analysis of the transcriptome shows how partial hepatectomy induces the entry into two cycles: a Sham control-like cycle and a second cycle during the cell proliferation and division time phases. The Pol2 molecules performing transcription and histone marks of active transcription show a coordinated response between transcription activation and its outcome that is particularly fast during the early responses. Furthermore, I identify especial cases that provide examples worth further studying.

In Chapter V, I show insights about the accumulation of H3K36 methylation in the mouse liver genome and its link with splicing regulation. Notably, within transcription units, a transition from H3K36me2 to H3K36me3 is observed that could only be observed previously in Drosophila. This transition tends to occur by the 5' end of annotated first introns of transcribed genes, suggesting its link with splicing. Nevertheless, in I also identify cases of transcription units where H3K36 methylation accumulates but no splicing is observed, suggesting further roles of H3K36 methylation in non-spliced transcription units.

In Chapter VI, I investigate the association of HCF-1 with the mouse liver chromatin. Little is known about the murine HCF-1, such as the similar structure to the human HCF-1 and the proteolytic cleavage of the precursor HCF-1 producing the HCF-$1_N$ and HCF-$1_C$ subunits. Thus, first, I show that in the mouse liver the two subunits of HCF-1 associate, as it occurs in human cells, and that HCF-1 binds to the mouse liver chromatin, especially to promoters, in response to environmental cues. Further, I describe different modes of binding of the two subunits of HCF-1, which suggests that the two subunits can be functional alone and that HCF-1 regulates transcription initiation in different ways.

Chapter VII provides conclusions and reflections of my thesis.

At the end of the document, references of information sources mentioned through-out the thesis are provided.

This thesis document also includes a CD-ROM that includes the scripts used during the thesis. In the main directory of this CD-ROM there is one document describing its contents.

# Chapter I: Introduction

Two complete sets of instructions are contained within the genomes we inherit from our parents. A compact packaging of the instructions is necessary to fit them inside of a cell nuclei. Nevertheless, cells possess mechanisms to make them accessible for reading. Controlled expression patterns of these two sets of instructions guide a single cell – the zygote – through different processes of cell proliferation, differentiation and death. As a result, the single zygotic cell is directed to become an adult human being. In this doctoral thesis I study the regulation of gene expression in two experimental contexts: human cancer cells and the mouse liver.

## I.1 Mammalian cell proliferation

Cell proliferation is a fundamental process whereby cells reproduce themselves by growing and then dividing into two daughter cells. The entry of cells into a proliferative state is the result of external stimuli. In mammals, growth factors transmit this message to cells to grow (Figure I-1). At this point the biosynthetic activities of cells resume at a high rate. After a so-called 'restriction point' (R) cells are committed to enter the cell cycle independently of cellular signaling. Waves of programmed and orchestrated changes occur inside cells that make them grow and multiply the number of organelles, during Interphase. During the synthesis phase the genomic contents of cells are duplicated via DNA replication. This process starts at specific locations in the genome called origins of replications. The DNA is unwind at those positions and synthesis of the two strands results in replication forks through which the replication machinery travels composed by DNA Polymerase and other enzymes such as DNA helicases, topoisomerases, gyrases, etc. Finally, during M phase, two daughter cells are produced equal to the mother cell. The progression of cell division is carefully guarded at checkpoints, where cell proliferation for example can be arrested in response to DNA damage or incomplete replication via the p53 tumor suppressor system, or mitosis by the spindle assembly checkpoint.

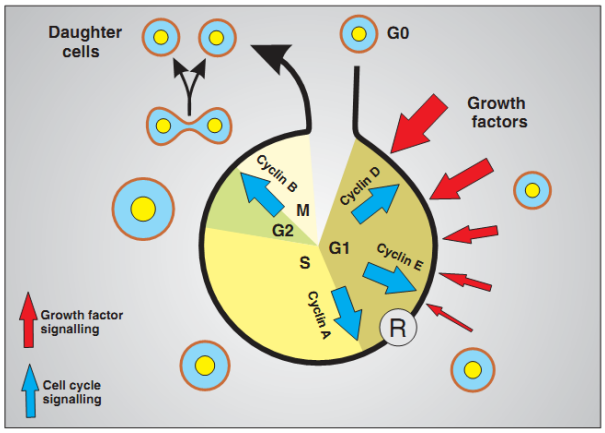


**Figure I-1**. Description of the cell cycle phases, the increasing growth of cells at each stage and the times when growth factors (red arrows) and cyclins (blue arrows) are relevant for the regulation of the mammalian cell proliferation. Growth factor signaling are key during G1 to drive cells towards the restriction (R) point, when cells pass the checkpoint controlling their conditions to start dividing. After (R) the control is taken by the internal cell cycle signaling system (blue arrows). The figure has been taken from the book [Berridge et al., 2006].

The eukaryotic cell cycle is controlled by a regulatory network, the general features of which are conserved from yeast to humans [Morgan et al., 2007]. Some of the elements that regulate the major events of the cell cycle are cyclins and cyclin dependent kinases (CDKs). Cyclins were discovered in 1982 by R. Timothy Hunt [Evans et al., 1983]. They were originally named because their concentration varied in a cyclical fashion during cell proliferation (Figure I-2) as a result of their expression and degradation at specific times of the cell cycle (Figure I-1). In contrast, Cyclin Dependent Kinases are more constantly expressed during cell proliferation. CDKs become active after binding to Cyclins and being phosphorylated by a Cyclin-dependent kinase-Activating Kinase (CAK). These cyclin/CDK complexes are responsible for specific events during cell division.

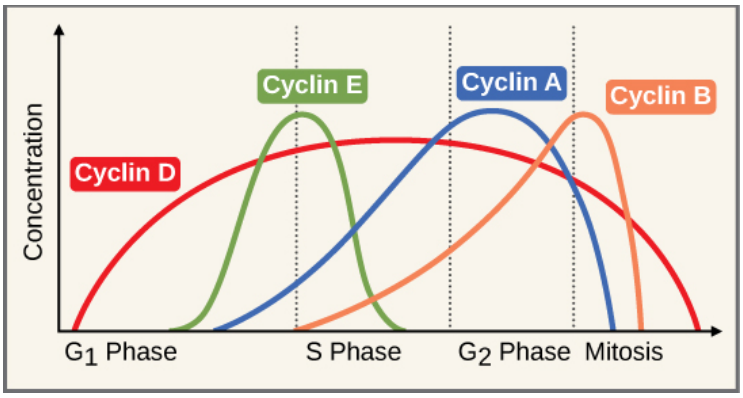



**Figure I-2**. Cyclin expression cycle. Source: website www.boundless.com/. This figure describes approximated protein concentration levels of cyclins D, E, A and B throughout the cell cycle phases (G1, S, G2 and M).

In backstage, orchestrated waves of gene expression control the progression of cell proliferation. Cascades of transcription factor activity regulate transcription of genes including some cyclins and CDKs. In most eukaryotes, cell cycle-regulated transcription can be grouped into three main waves that coincide with the transition points G1-to-S, G2-to-M and M-to-G1 [Bahler et al., 2005]. Some key transcription factors belong

to the family of E2Fs. They are present in higher eukaryotes and regulate the transition between G1 to S phase by targeting promoters of cyclins, CDKs, checkpoint regulators, DNA repair and replication genes.

## I.2 Chromatin

In the nucleus of mammalian cells, the chromosomal DNA is packed with different levels of compaction through association with proteins, forming the chromatin (Figure I-3). The nucleosome is the basic unit of compaction. Within a nucleosome roughly 145 bp of DNA are wrapped around histone octamers formed by a pair of each of the histone proteins H2A, H2B, H3 and H4 (Figure I-4) [Luger et al., 1997]. This unit is repeated throughout the genome and is further condensed through association of additional proteins, for example the linker histone H1.

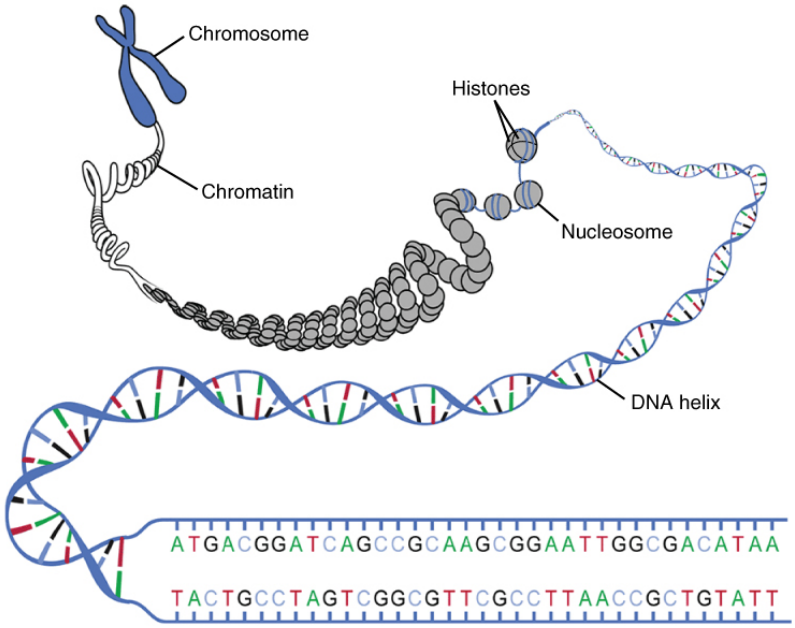



**Figure I-3**. DNA compaction levels. The figure illustrates different compaction levels of the DNA double-stranded sequence, starting by the double helix, following by the basic compaction unit that is the nucleosome and followed by further structures that ultimately form the compact chromosomes observed during mitosis.

The chromatin is highly dynamic. Nucleosomes can be displaced, histones can be post-translationally modified, evicted by chromatin remodelers or replaced by histone variants [Li et al., 2007]. All these changes affect the

accessibility of proteins to the DNA and, therefore, the progression of molecular events, such as gene transcription required during cellular processes such has cell proliferation.

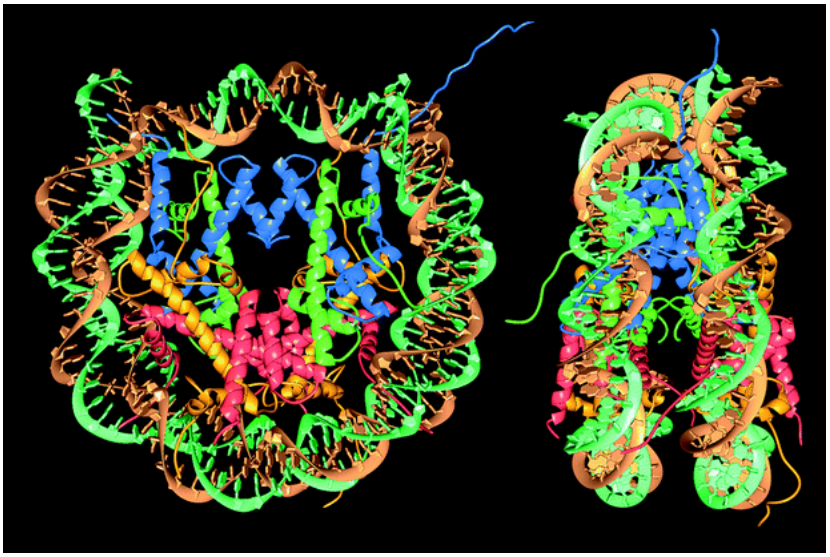


**Figure I-4**. X-ray crystal structure of the nucleosome core particle. Source: [Luger et al., 1997]. On the left the nucleosome core particle is viewed down the superhelical axis. On the right the particle is rotated 90º on the Y-axis. The different histone proteins have been colored: H2A (yellow), H2B (red), H3 (blue) and H4 (green). The two strands of DNA are shown in light green and orange.

## I.3 Gene transcription

Transcription is the first step in the process of gene expression and is performed by RNA polymerases. RNA polymerases are complexes of proteins that are able to read a template DNA sequence and carry out the synthesis of different kinds of RNA. In eukaryotes there are different types of RNA polymerases: Pol1, Pol2 and Pol3. Pol1 and Pol3 transcribe a limited number of genes encoding ribosomal RNAs, transfer RNAs and non-coding RNAs. In contrast, Pol2 is responsible for transcription of the thousands of protein-encoding genes and most snoRNAs, miRNAs and long non-coding RNAs (lncRNAs). After further maturation steps, the synthesized pre-mRNA is translated into a protein or kept as a mature mRNA.

In proliferating cells, transcription-replication interference can occur. Nevertheless, like bacteria, eukaryotic cells have evolved ways to deal with these collisions, such as by polar replication barriers and blocking

replication forks that are progressing opposite to the direction of transcription [Kobayashi et al., 1996; Pasero et al., 2002]. Furthermore, it is believed that DNA replication and transcription are segregated events in the nucleus which helps ensure that the replication and transcription machineries do not encounter one another. [Nakayasu and Berezney, 1989; Jackson et al, 1993; Spector et al, 1991; Wei et al, 1998].

## I.3.1 RNA polymerase II

The existing crystals of RNA polymerase II (Pol2) show that yeast Pol2 is a complex of approximately 550 kD with 10 to 12 subunits [Cramer, Bushnell and Gnatt et al et al., 2000] (Figure I-5). RPB1 is the largest subunit followed by RPB2, which in combination with several other polymerase subunits forms the DNA binding domain of the polymerase, a groove in which the DNA template is transcribed into RNA. Furthermore, RPB1 has a unique C-terminal domain (CTD) consisting of heptapeptide (YSPTSPS) repeats that is key for transcription.

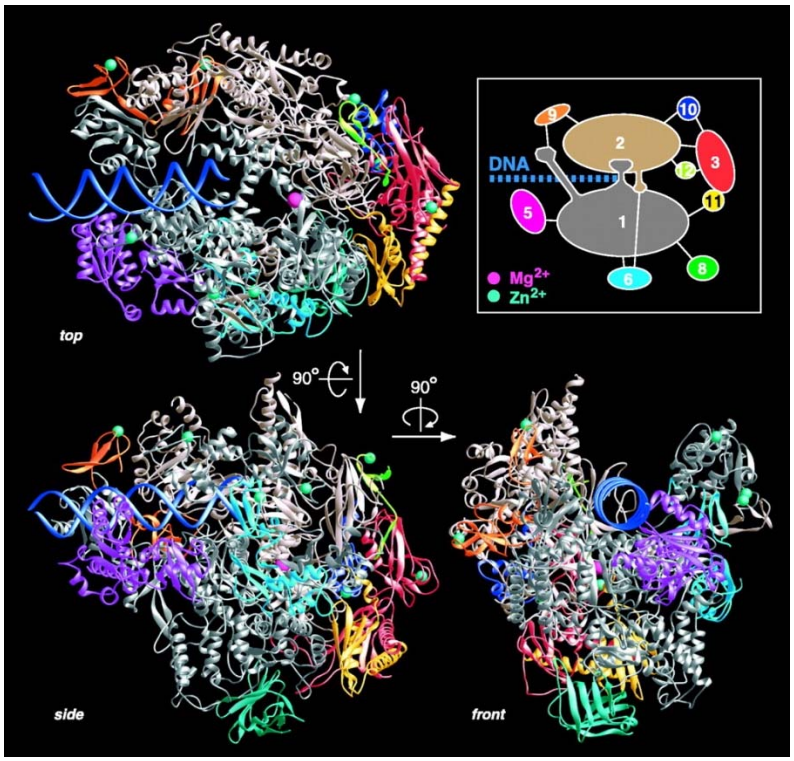



**Figure I-5**. Architecture of yeast Pol2. Source: [Cramer, Bushnell and Gnatt et al et al., 2000]. Three views of the yeast Pol2 are displayed from different angles by 90° rotations. Ribbon diagrams depict the backbones for the 10 Pol2 subunits. The downstream DNA sequence has been artificially added in the location identified by crystallographic studies [Poglitsch et al.,

1999]. Eight zinc atoms (green spheres) and the active site magnesium (pink sphere) are shown. The box in the upper right area contains a key to the subunit color code and an interaction diagram.

Pol2 by itself is capable of unwinding an open DNA double helix, polymerizing RNA, and repairing the 3' end of the nascent transcript. Nevertheless, Pol2 stands at the center of a complex machinery called the Pol2 holoenzyme that regulates this process and whose composition changes in the course of transcription. Initially, the preinitiation complex (PIC) forms with Pol2 (Figure I-6) with at least six general transcription factors (TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH). Because Pol2 alone is not able to recognize promoter sequences this complex helps to position Pol2 in promoters with sequence specificity. The assembly process begins when the general transcription factor TFIID binds to a short double-helical DNA sequence primarily composed of T and A nucleotides. For this reason, this sequence is known as the TATA box, and the subunit of TFIID that recognizes it is called TATA-Binding Protein (TBP). The preinitiation complex is also responsible for melting the double stranded DNA allowing for initiation of transcription of a single strand from 5' to 3' [Roeder, 1996; Conaway and Conaway, 1997; Kornberg, 1999; Lee and Young, 2000]. Once Pol2 sits in a promoter sequence, typically a Mediator complex of about 1.2 MD attaches and transmits regulatory information from transcription activators and repressors [Myers and Kornberg, 2000]. For productive elongation of the transcript new proteins are recruited. In some promoters Pol2 is paused in a promoter proximal position by negative elongation factors and later released by positive elongation factors such as P-TEFb [Conaway and Conaway, 1999; Shilatifard, 1998].

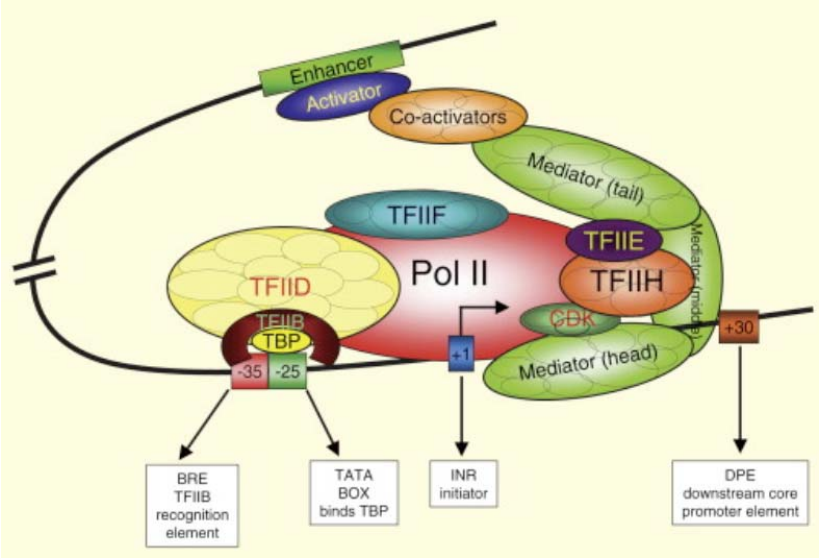


**Figure I-6.** Scheme of the key elements involved in the formation of the Pol2 preinitiation complex (PIC) when a TATA box is present in the proximal promoter. Source: [Krishnamurthy and Hampsey, 2009]. The figure illustrates the core promoter elements at the positions -35, -25, +1 (Transcription Start Site or TSS) and +30 bp. The TBP subunit of the TFIID complex binds to the TATA box, followed by binding of the general transcription factors TFIIB, Pol2/TFIIF, TFIIE and TFIIH. The assembly is promoted by the binding of an activator to an enhancer sequence that tethers coactivator complexes that typically include chromatin modifiers, and a mediator complex that interacts directly with Pol2 and the general transcription factors.

Regulatory proteins interact with Pol2 co-transcriptionally that will maturate the nascent mRNA. This occurs during the processes of capping of the 5' end of the pre-mRNA and for splicing the exons by excising the introns (Figure I-7). Finally, during termination of transcription, additional complexes complete the maturation of the mRNA by means of cleavage and polyadenylation of the 3' end [Hirose and Manley, 2000; Proudfoot, 2000] (Figure I-7).

The interaction of Pol2 with enzymes also results in the phosphorylation of some serine residues in the CTD tail. The TFIIH and mediator complex phosphorylate serine 5 (Ser5) of the CTD heptapeptide repeat sequence during elongation, whereas positive elongation factor P-TEFb phosphorylates serine 2 (Ser2) [Hirose and Ohkuma, 2007]. These serine phosphorylations are not only essential for transcription but also a platform for RNA processing and chromatin regulation, as different factors involved in those processes recognize and bind to the modified serines, including splicing factors and histone methyltransferases (HMT).

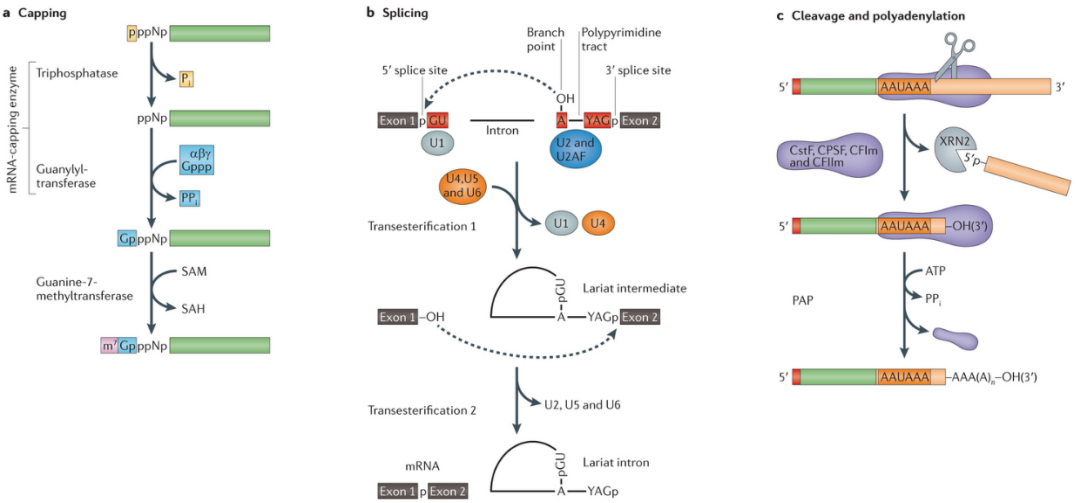


**Figure I-7**. Scheme of the co-transcriptional processes of 5' end capping, splicing and 3' cleavage and polyadenylation. Source: [Bentley, 2014]. a) 5' end capping. The 5' end of the nascent transcript (green) has a cap added after three steps by two different enzymes. The mRNA-capping enzyme in metazoans functions as a triphosphatase and guanylyl-transferase that first removes the γ-phosphate (yellow) of the transcript and then transfers GMP from the GTP donor (blue). The methyl donor S-adenosyl-l-methionine (SAM) is converted to S-adenosyl-l-homocysteine (SAH), becoming a 7-methylguanosine cap (pink). b) Splicing mechanism. The intron is cut out and the flanking exons are ligated by means of two transesterification reactions. In red the conserved intronic splicing elements are highlighted. A simplification of the spliceosomal proteins is depicted. Only spliceosomal U1, U2, U4, U5 and U6 snRNPs are shown together with the U2 auxiliary factor (U2AF). c) Formation of the 3' end of the transcript by cleavage and polyadenylation. A consensus sequence (AAUAAA) signals the cleavage site located approximately 25 bp downstream. The cleavage is performed by a complex (purple) formed by the cleavage stimulation factor (CstF), the cleavage and polyadenylation specificity factor (CPSF) that carries an endonuclease, and cleavage factors I and II (CFIm and CFIIm). Later, the enzyme Poly(A) polymerase (PAP) performs the addition of the Poly(A) tail to the 3' end. Finally, the RNA exonuclease (XRN2) degrades RNA downstream of the cleavage site which facilitates termination of transcription.

## I.3.2 Transcription factors, co-transcription factors and chromatin modifiers

Further combinations of proteins can bind the chromatin providing a unique regulation of gene transcription. Transcription factors bind the DNA of enhancer or promoter proximal sites. They bind the DNA in a sequence-dependent manner through their DNA-binding domains. Approximately 3,000 of the about 30,000 human genes encode for proteins that have DNA-binding domains, and most of them are presumed to be transcription factors [Babu et al., 2004]. Some factors, called pioneering factors, can still bind their DNA-recognition-sites even if the sites sit on compacted nucleosomal DNA. But for most of the transcription factors the accessibility of the DNA sequence is required. This can be acquired beforehand by the action of chromatin remodelers and the help of thermal fluctuations of the chromatin. Some transcription factors have the ability to activate the initiation of transcription and others to repress it by either stabilizing or blocking the binding of the preinitiation complex and RNA polymerase. On top of that, transcription factors can recruit further co-transcriptional factors that can't recognize DNA sequences on their own and co-activate or co-repress transcription. These complexes can also tether chromatin modifiers capable to affect the chromatin accessibility, hence, facilitating or hindering transcription [Li et al., 2007].

### I.3.3 Histone post-translational modifications and their link with transcription

The four core histones share a similar structure, with a globular hydrophobic internal region and a flexible N-terminal tail region that extends out of the nucleosome [Luger et al., 1997]. A major component in the regulation of cellular processes by chromatin structure is the post-translational modification of these N-terminal tails. Histone are subject to modifications such as methylation, citrullination, acetylation, phosphorylation, SUMOylation, ubiquitination and ADP-ribosylation. The addition of modifications can have different effects that can regulate the progression of transcription (Table I-1). On the one hand, some modifications such as acetylation can alter the non-covalent bonds between DNA and histones provoking a transient release of their interaction and the accessibility for other proteins to regulate transcription. Further, other modifications such as methylation can provide recognition sites for reader enzymes involved in chromatin remodeling and/or transcription regulation. In genes, the addition of histone modifications can occur at different positions. Although their role is still not completely understood, some are typically associated to specific transcription outcomes (Figure I-8).

| **Modification** | **Histone** | **Position** | **Role in transcription** |
|---|---|---|---|
| Methylation | H3 | K2 | Activation |
| | | K4 | Activation |
| | | K9 | Activation and repression |
| | | R17 | Activation |
| | | R26 | Activation |
| | | K27 | Repression |
| | | K36 | Recruiting the Rpd3S to repress internal initiation |
| | | K79 | Activation |
| | H4 | R3 | Activation |
| | | K20 | Repression |
| Phosphorylation | H3 | S10 | Activation |
| Ubiquitination | H2A | K119 | Repression |
| | H2B | K120/123 | Activation |
| Acetylation | H3 | K9 | Activation and repression |
| | H3 | K56 | Activation |
| | H4 | K16 | Activation |

| | | | |
|---|---|---|---|
| | H2A.Z | K14 | Activation |

**Table I-1**. Histone post-translational modifications that regulate transcriptional activity. Source: [Li et al., 2007]. This table provides a summary of the most studied histone modifications known to regulate transcription. The column 'Histone' indicates the different Histones described and their specific positions and residues altered by post-translational modifications are indicated in the column 'Position' (K: Lysine; S: Serine; R: Arginine).

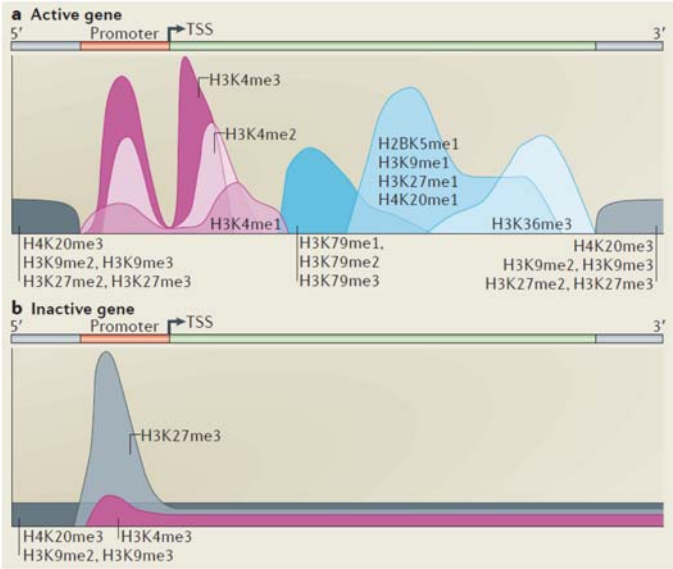



**Figure I-8**. Schematic of the distribution of specific methylations in the histone tails and their associated transcription role. Source: [Kooistra and Helin, 2012]. In both panels a gene (green box) is illustrated on the top that is transcribed from left to right. A promoter sequence is depicted upstream of the Transcriptional Start Site (TSS) indicated with an arrow. Curves represent patterns of methylation determined by genome-wide studies. a) Distribution of histone methylations associated with actively transcribed genes. b) Distribution of histone methylations associated to non-transcribed genes.

## I.3.3.i Methylation

During the past decade, much research has focused on the study of histone methylations that can happen in either lysine or arginine residues. Histone methylation doesn't affect the electrostatic charge of the residues. Hence it is unlikely that methylation of histones impacts nucleosome structures. Instead, these modifications

have been shown to act as signals for downstream effectors [Jenuwein and Allis, 2001]. In genes, different methylation levels can be observed in histones that have been linked to different functional outcomes (Figure I-8).

Histone methyltransferase (HMTase) enzymes use S-adenosylmethionine to add methyl groups to specific histone lysine or arginine residues. This addition was generally believed to be irreversible and that it could thus only be removed by histone eviction or by dilution during DNA replication. However, this notion changed after isolation of two families of enzymes that can demethylate histones: the Lysine-Specific histone Demethylase family (LSD) and the Jumonji C family (JMJC). JMJC demethylases are highly conserved as they can be found from bacteria to humans. JMJC demethylases are able to reverse the three lysine methylation states by a Fe(II)- and α-ketoglutarate-dependent mechanism [Tsukada et al., 2006]. In contrast, LSD demethylases can only demethylate mono- and di-methylated residues and for that they use a flavin adenine dinucleotide (FAD)-dependent amine oxidation reaction to catalyze the demethylation of their substrate [Shi et al., 2004]. The biochemical activities of these histone demethylases towards specific histone residues, and in some cases non-histone substrates, have highlighted their importance in developmental control, cell-fate decisions and disease [Kooistra and Helin, 2012].

### I.3.3.ii H3K4 methylation signatures and their link with gene expression

The Set1 protein from S. cerevisiae was the first histone H3 lysine 4 (H3K4) methyltransferase identified and is the only one able to perform this function in yeast [Briggs et al., 2001; Roguev et al., 2001]. In mammals, in contrast, there is much more redundancy (Figure I-9). There are ten different H3K4 methyltransferases identified. Six of them belong to the MLL family that arose through genome duplication during vertebrate evolution and contain a catalytic domain similar to the yeast Set1 or the Drosophila MLL-related proteins. On the other hand, there are both LSD and JMJC families of H3K4 demethylases (Figure I-10).

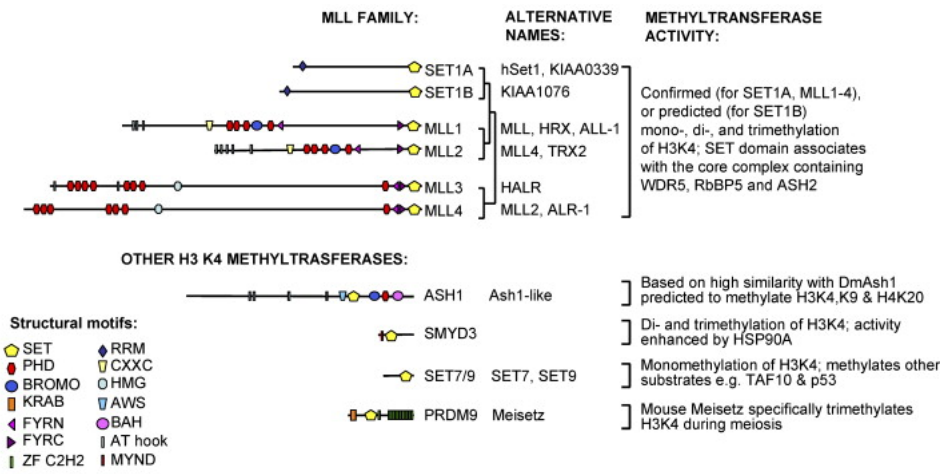



**Figure I-9**. Scheme of the known human H3K4 methyltransferases, their protein domains and the sequence similarity among them. Source: [Ruthenburg, Allis and Wysocka, 2007]. Two main groups are described. On the top the MLL family of proteins that contain SET domains similar to the SET domain of the yeast Set1 and the Drosophila MLL family. Below, the non-related proteins that are still capable to methylate H3K4. The alternative names for each protein are provided and the methyltransferase activity reported in the source of the figure are described.

| Name | Synonyms | Protein structure* | Histone substrates‡ | Non-histone substrates‡ |
|---|---|---|---|---|
| *LSD demethylases* | | | | |
| LSD1 | AOF2, BHC110, KDM1A | SWIRM, Amine oxidase, Spacer region | H3K4me1, H3K4me2 | p53, E2F1, DNMT1 |
| LSD2 | AOF1, KDM1B | CW | H3K4me1, H3K4me2 | |
| *JMJC demethylases* | | | | |
| FBXL10 | JHDM1B, KDM2B | JMJC, CXXC, PHD, FBOX, LRR | H3K4me3 | |
| JARID1B | PLU1, KDM5B | ARID, C5HC2 | H3K4me2, H3K4me3 | |
| JARID1C | SMCX, KDM5C | | H3K4me2, H3K4me3 | |
| JARID1D | SMCY, KDM5D | | H3K4me2, H3K4me3 | |
| JARID1A | RBP2, KDM5A | | H3K4me2, H3K4me3 | |
| NO66 | | | H3K4me2, H3K4me3 | |

**Figure I-10**. Description of structure and histone substrates of the known H3K4 demethylases. Source: Modified from [Kooistra and Helin, 2012]. Proteins from two families of demethylases are known to remove methyl groups from H3K4: LSD, Lysine-Specific histone Demethylase, and JMJC, the Jumonji C, families of demethylases. (*) Protein

structures are grouped according to levels of homology. (‡) Substrates are indicated for all proteins with demonstrated demethylase activity.

Interestingly, the patterns of H3K4 dimethylation (H3K4me2) accumulation in yeast differ from the patterns in vertebrates, suggesting distinct cellular functions of methylases and demethylases between these organisms. In S. cerevisiae H3K4me2 spreads all along genes, peaking by the middle of the coding region and colocalizes with H3K4 monomethylation (H3K4me1) that peaks by the 3' end [Santos-Rosa et al., 2002; Ng et al., 2003; Pokholok et al., 2005]. In contrast, the majority of vertebrate H3K4me2 colocalizes with H3K4 trimethylation (H3K4me3) in a region of 5 to 20 nucleosomes proximal to highly transcribed genes [Schneider et al., 2004; Bernstein et al., 2005].

H3K4me3 is observed in 90% of the promoters occupied by Pol2 and it has been shown that the loss of H3K4me3 is associated with a lowered transcriptional activity. Moreover, the loss also correlates with a decreased binding of TFIID in some promoters missing the canonical TATA box. This has led to the hypothesis that in those situations H3K4me3 may define the core promoter by either anchoring TFIID to activated promoters or by recruiting TFIID during promoter activation [Vermeulen et al., 2007]. Additionally, it has been proposed that H3K4me3 facilitates the efficiency of post-initiation transcriptional processes [Sims et al., 2007].

### I.3.3.iii H3K36 methylation signatures and their link with gene expression

In yeast, H3K36 is solely methylated by the SET domain-containing 2 (Set2) protein. It has the ability to perform the three methylation events of this residue [Lee and Shilatifard, 2007]. In contrast, in metazoan, these modifications are performed by multiple enzymes that are responsible for specific methyl additions, being Set2 the only responsible for the trimethylation H3K36 [Reviewed in Wagner and Carpenter, 2012]. On the other hand, demethylation of H3K36 is performed by a different family of enzymes, the JMJC enzymes (Figure I-12).

A challenge in the past few years has been to decipher how the H3K36 methylation states participate in gene regulation. H3K36 methylation is commonly associated with the transcription of active euchromatin. Nevertheless, it has also been implicated in diverse processes, including alternative splicing, dosage compensation and transcriptional repression, as well as DNA repair and recombination [reviewed in Wagner and Carpenter, 2012]. Disrupted localization of methylated H3K36 can lead to a range of human developmental

defects and diseases [Nimura et al., 2009; Hu et al., 2010; Tomita et al., 2002], highlighting the importance of this modification.

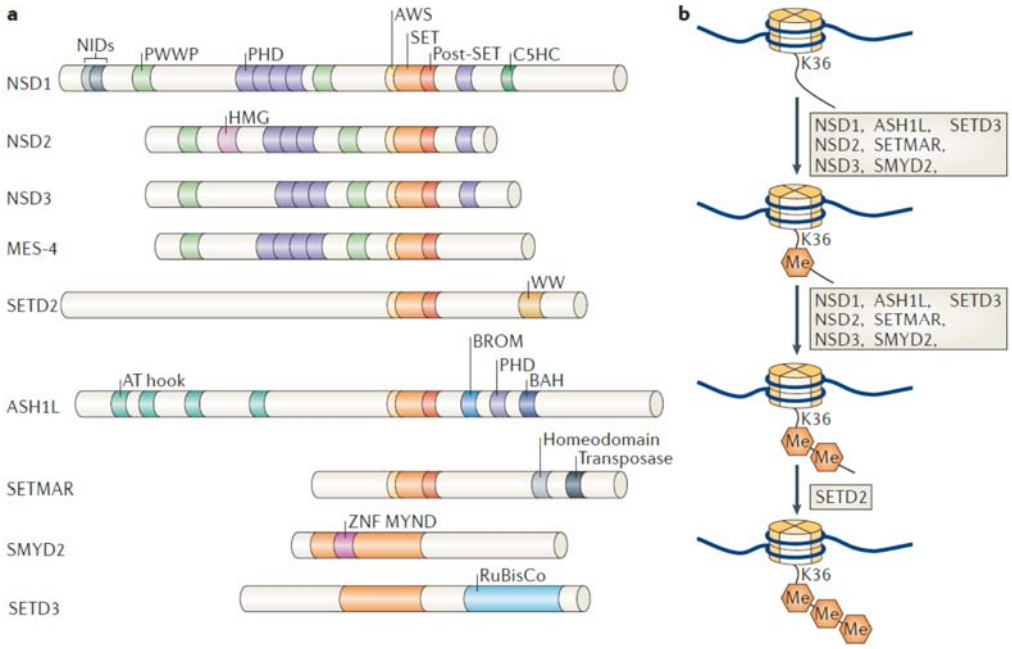



**Figure I-11**. Description of the structure and histone substrates of the known H3K36 methyltransferases. Source: [Wagner and Carpenter, 2012]. a) Schematic of structure of the enzymes. All depict human enzymes, except for the Drosophila MES-4. b) Stepwise transitions between the different levels of H3K36 methylation and the enzymes known to be involved in the addition of the specific mono-, di- and trimethyl groups.

| Name | Synonyms | Protein structure* | Histone substrates‡ | Non-histone substrates‡ |
|---|---|---|---|---|
| JMJD5 | KDM8 | JMJC | H3K36me2 | |
| FBXL10 | JHDM1B, KDM2B | CXXC PHD FBOX LRR | H3K36me1, H3K36me2 | |
| FBXL11 | JHDM1A, KDM2A | | H3K36me1, H3K36me2 | p65, NF-κB |
| JMJD2A | JHDM3A, KDM4A | JMJN TUDOR | H3K36me2, H3K36me3 | |
| JMJD2C | JHDM3C, GASC1, KDM4C | | H3K36me2, H3K36me3 | |
| JMJD2B | JHDM3B, KDM4B | | H3K36me2, H3K36me3 | |
| JMJD2D | JHDM3D, KDM4D | | H3K36me2, H3K36me3 | |
| NO66 | | | H3K36me2, H3K36me3 | |

**Figure I-12**. Description of structure and histone substrates of the known H3K36 demethylases. Source: Modified from [Kooistra and Helin, 2012]. Only proteins from the JMJC family of demethylases are known to remove methyl groups from H3K36. (*) Protein structures are grouped according to levels of homology. (‡) Substrates are indicated for all proteins with demonstrated demethylase activity.

Some studies have given insights into the mechanisms by which H3K36 methylases and demethylases affect eukaryotic gene expression. In yeast, it has been observed that Set2 can bind and methylate H3K36 in the promoters of genes, which represses Pol2 elongation. And antagonistic demethylation by Rpdh1 (the yeast JMJD2C demethylase) promotes elongation [Kim and Buratowski, 2007]. In yeast it has been observed that during Pol2 elongation the Set2 protein stays attached to the phosphorylated CTD of Pol2, which facilitates the addition of methyl groups along the body of genes and polyadenylation sites [Li et al., 2003; Krogan et al., 2003; Xiao et al., 2003; Kizer et al., 2005; Schaft et al., 2003]. Consistent with this, ChIP-seq studies have described the accumulation of H3K36me3 towards the 3' end of genes [Bannister et al., 2005; Pokholok et al., 2005; Rao et al., 2005]. Dimethylation of H3K36 (H3K36me2) has also been observed in genes, although its localization changes across species. In yeast and chicken, it has been observed to coincide with the trimethylation [Bannister et al., 2005; Pokholok et al., 2005; Rao et al., 2005]. In contrast, in drosophila, H3K36me2 is observed closer to the 5' end of genes [Bell et al., 2007]. This suggests distinct regulations of the addition of methyl groups across species. As all these H3K36 methylation marks can be binding sites for further effector proteins possessing domains able to recognize and bind them [Martin and Zhang, 2005; Wysocka et al., 2006; Li et al., 2006; Pray-Grant et al., 2005] these different localizations can lead to different effects in the chromatin and gene expression.

It has been hypothesized that the accumulation of H3K36me3 is linked to the regulation of splicing events. On this regard, it has been observed that the H3K36me3 mark begins to accumulate in genes by the 3 ' end of the 1st intron, the first splicing event occurring within a gene [Huff et al., 2010]. Also it is known that a significant fraction of splicing occurs cotranscriptionally [Reviewed in Pandya-Jones, 2011], and the cotranscriptional splicing apparatus influences the establishment of H3K36me3, as shown after splice site deletions in a tet-inducible CMV β-globin human gene [Kim et al., 2011].

## I.4 The Host-Cell Factor 1 (HCF-1) co-transcriptional regulator

The Host-Cell Factor 1 (HCF-1) is an abundant and highly conserved co-transcriptional regulator protein. It was first discovered as a human host-cell factor used by herpes simplex virus (HSV) to initiate viral transcription [reviewed in Wysocka and Herr, 2003]. Upon HSV infection of permissive cells, HCF-1 stabilizes a multiprotein-DNA transcriptional complex – the VP16-induced complex (VIC) – formed by the viral protein VP16 and a second cellular protein called Oct-1. In animals (e.g., vertebrates, sponges, sea anemones, worms and insects), there are homologous proteins that conserve this function [Kristie et al., 1989; Wilson et al., 1993; Gonzalez et al., unpublished].

In humans, HCF-1 is synthesized as a precursor protein of 2,035 amino acids; the closely related mouse HCF-1 protein, which extends to 2,045 amino acids, differs only by 2% of its sequence and this difference is concentrated around the middle of the molecule (Figure I-13a). Different domains have been described having specific functions and showing distinct interactions with other proteins, such as the N-terminal Kelch domain. The Kelch domain has a beta-propeller structure formed by the HCF-$1_{KEL}$ repeats that can sit on DNA-binding proteins such as VP16 (Figure I-13b) [Wysocka et al., 2001].

HCF-1 undergoes a maturation process by which it is proteolytically cleaved at the HCF-$1_{PRO}$ repeats (Figure I-13a) producing two subunits: the amino- (HCF-$1_N$) and the carboxy- (HCF-$1_C$) terminal subunits. Mature HCF-1 is a heterodimeric complex of an amino- and a carboxy-terminal subunits stably but non-covalently associated by the SAS1N and SAS1C segments (Figure I-13a and b) [Wilson et al., 1993a; Wilson et al., 1995b; Kristie et al., 1995].

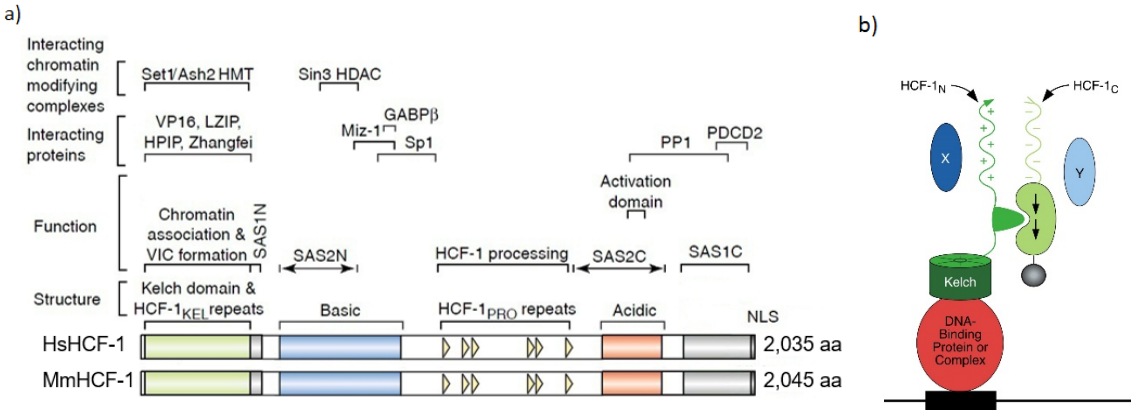



**Figure I-13**. Schematic representation of human and mouse full length HCF-1 (HsHCF-1 and MmHCF-1). Only 2% of the protein sequence is different between these two species. This

difference is present by the central HCF-$1_{PRO}$ processing repeats. a) Structural or sequence elements along with functional regions are described. b) Schematic of the interactions of a processed HCF-1. HCF-1 is represented in green colors, being the dark green region the N-terminal subunit and the light green the C-terminal subunit.

The glycosyltransferase O-GlcNAc transferase OGT is the enzyme responsible for the cleavage of HCF-1 and it does so in an unusual way; OGT glycosylates and proteolytically cleaves HCF-1 at the six central HCF-$1_{PRO}$ repeats [Capotosti et al., 2011]. The proteolytic cleavage of the precursor protein can produce different polypeptides depending on the proteolytic repeat used for cleavage. [Wilson et al, 1993a] showed the diversity of polypeptides obtained from nuclear HCF-1 purification of HeLa cells (Figure I-14). Polypeptides were observed with atomic masses of 300 (full length), 150, 127, 125, 116 and 110 kD. And smaller degradation products of size 66 kD were observed as well. Sequencing of the observed bands showed that both amino- and carboxy-terminal subunits of HCF-1 were contained [Wilson et al., 1993b].

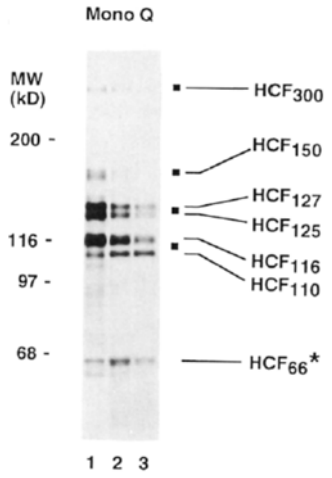



**Figure I-14**. The HCF-1 family of polypeptides. Source: [Wilson et al., 1993a]. Several purification steps were done to retrieve nuclear HCF-1 from HeLa cells, the last step being Mono-Q chromatography. Three Mono-Q peak fractions were electrophoresed on a 7% denaturing SDS-polyacrylamide gel and visualized by silver staining. Multiple HCF-1 polypeptides were observed with sizes 110, 116, 125, 127, 150 and 300 kD, being the last one the full length HCF-1. Also degradation products were identified with size 66 kD. Sequencing of the bands showed that each of them contained both amino- and carboxy-terminal HCF-1 subunits.

## I.4.1 HCF-1 is a regulator of cell proliferation

Because viruses often target key regulators of cellular function to promote infection, HCF-1 was thought to be involved in key cellular functions. Research shows that HCF-1 is required for multiple steps of cell proliferation. Interestingly, each subunit is required for different steps of the cell cycle [Julien and Herr, 2003; Goto et al., 1997] (Figure I-15). In the absence of HCF-$1_N$ subunit function, mammalian cells enter a stable G1-phase arrest. HCF-1 regulates the G1/S passage by interacting at least in part with proteins from the E2F family, known to regulate the cell cycle. And, in the absence of the HCF-$1_C$ subunit cells proliferate but display multiple M-phase defects including defective regulation of histone H4 lysine 20 (H4K20) methylation, chromosome segregation, and cytokinesis, the latter resulting in multinucleated cells [Julien and Herr, 2004].

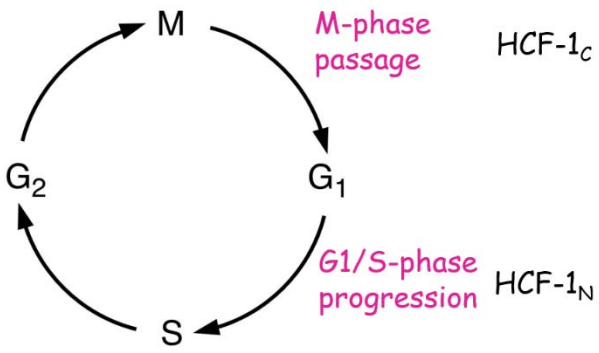



**Figure I-15**. Summary of the roles of the HCF-1 subunits in the progression of the cell cycle. Source: Figure from Winship Herr. As shown in [Julien and Herr, 2003; Goto et al., 1997], HCF-$1_N$ is required for G1 to S transition and HCF-$1_C$ is required for proper M-phase passage.

## I.4.2 The spectrum of HCF-1 protein associations

HCF-1 is a versatile adaptor protein that mediates the interaction of different combinations of DNA-binding transcription factors and chromatin remodelers (Figure I-16). As a first step, HCF-1 is able to associate with DNA-binding proteins through the Kelch domain beta propeller structure via a 4-amino-acid sequence motif called the 'HCF-Binding Motif' (HBM) found in the target proteins [Freiman and Herr, 1997; Lu et al., 1998]. The consensus HBM sequence is defined as B/ZHxY (B denotes aspartate or asparagine, Z denotes glutamate or glutamine, H denotes Histidine, x means any amino acid and Y denotes Tyrosine). Interestingly, not all the transcription factors known to associate with HCF-1 contain an HBM, suggesting that alternative binding mechanisms occur.

The proteins that associate with HCF-1 fall into four different categories (Figure I-16): 1) promoter-specific transcription factors, 2) chromatin modifiers, 3) adaptor proteins like HCF-1 and 4) proteins with other types of functions. The first category includes the DNA-binding transcription factors LZIP/Luman [Freiman and Herr 1997; Lu et al. 1997], Zhangfei [Lu and Misra, 2000], GABPβ [Vogel and Kristie, 2000; Vercauteren et al., 2008], Sp1 [Gunther et al., 2000], Miz1 [Piluso et al., 2002], Krox20 [Luciano and Wilson, 2003], E2F1, E2F3, and E2F4 [Knez et al. 2006; Tyagi et al. 2007], THAP1 and THAP3 [Mazars et al., 2010], THAP11/Ronin [Dejosez et al. 2010], YY1 [Yu et al., 2010], ZNF143/Staf [Michaud et al., 2013], Myc [Thomas et al., 2015] and Mad1 [Ding and Herr, ms. in prep. ]. The second category includes the Set1A/B and mixed-lineage leukemia MLL1 and MLL2 histone H3 lysine 4 (H3K4) methyl transferases [Wysocka et al., 2003; Yokoyama et al., 2004; Shilatifard, 2012], the histone demethylases LSD1 [Liang et al. 2009] and PHF8 [Liu et al. 2010], MOF histone acetyl transferase [Smith et al., 2005; Cai et al., 2010], Sin3 histone deacetylase (HDAc) [Wysocka et al., 2003], OGT [Wysocka et al., 2003; Cai et al. 2010; Yu et al., 2010; Daou et al., 2011], the phosphatase PP1 [Ajuh et al. 2000], and the ubiquitin hydrolase BAP1 [Machida et al., 2009; Misaghi et al., 2009; Yu et al., 2010]. The third and fourth categories include the adaptor proteins PGC-1α and PGC-1β [Lin et al., 2002; Vercauteren et al 2008; Ruan et al., 2012] and other regulatory proteins (i.e., co-activator FHL2 [Vogel and Kristie, 2006]; nuclear export protein HPIP [Mahajan et al., 2002], and PDCD2 [Scarr and Sharp, 2002]), respectively. Thus, HCF-1 forms complexes with diverse proteins most of which are transcription regulators.

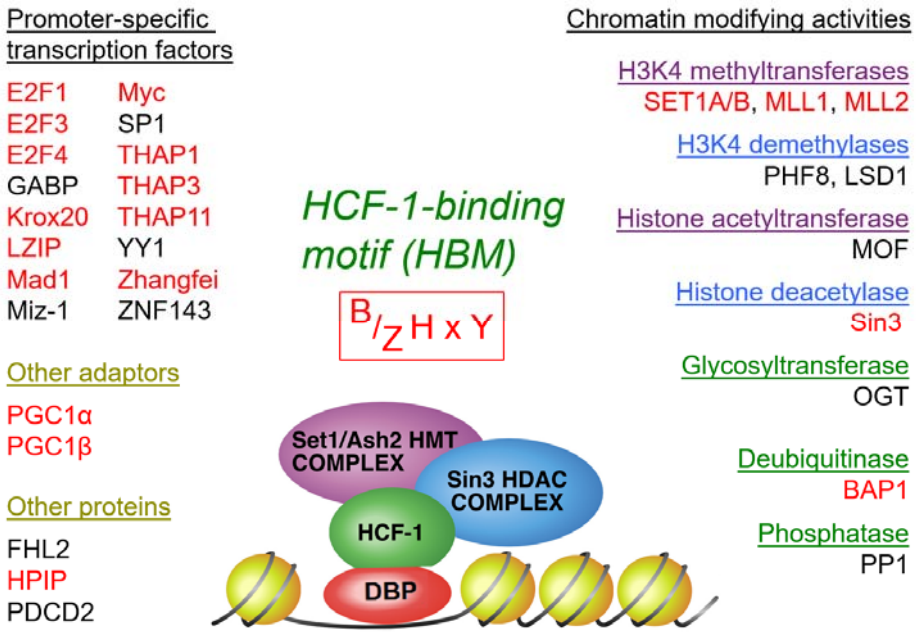


**Figure I-16**. Schematic of the proteins that associate with HCF-1 on promoters, their function and platform for association. Source: Figure from Winship Herr. Four categories of proteins are known to associate with HCF-1 in gene promoters: transcription factors, other protein adaptors, chromatin modifiers and proteins with other functions. The HCF-1 Binding Motif (HBM) found in many targeted proteins is defined in the middle of the figure (B denotes aspartate or asparagine, Z denotes glutamate or glutamine, H denotes Histidine, x means any amino acid and Y denotes Tyrosine). Those proteins in the list containing the HBM are highlighted in red.

HCF-1 is a dynamic adaptor protein. A well described example of this ability is the dynamicity of the interactions between HCF-1 and E2F family members during the progression of the cell cycle [Tyagi et al., 2007]. In the early G1-phase, HCF-1 associates with the repressor E2F4 in a complex with the repressor Sin3 deacetylase. Later on, for transitioning to S phase, HCF-1 associates to the activator E2F1 and this time it does it together with the H3K4 methylases Set1 and MLL, thus activating transcription. Therefore, HCF-1 interactions can have a broad range of functions that change with the progression of cellular processes such as cell proliferation.

## I.4.3 HCF-1 genome-wide chromatin association in HeLa cells

In non-synchronized HeLa cells, HCF-1 has been observed to bind to approximately 8,000 genomic locations. 67% of the binding sites fall in promoter regions, especially promoters close to CpG islands [Michaud et al., 2013]. The HCF-1 binding in promoters correlates with the transcriptional markers Pol2, H3K4me3 and H3K36me3. And upon depletion of HCF-1 the transcript levels of the HCF-1 target genes are largely modified in some cases in a cell-cycle stage specific manner [Michaud et al., 2013]. All together, these results suggest that HCF-1 is a broad transcription regulator in the HeLa chromatin.

An analysis of the DNA sequences underlying HCF-$1_C$ and HCF-$1_N$ promoter proximal binding sites reported potential interactions with the transcription factors THAP11, ZNF143, YY1 and GABP (Figure I-17). And those interactions were further substantiated with occupancy of these factors at HCF-1 binding sites by chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq). Moreover, the human GABP and YY1 and the mouse protein THAP11 were already shown to bind HCF-1 [Vogel and Kristie, 2000; Dejosez et al., 2008; Yu et al., 2010]. In the case of the proteins THAP11 and ZNF143 they tend to bind together in HCF-1-bound promoters (Figure I-17a). This latter phenomenon has also been reported in many

promoters by other researchers [Ngondo-Mbongo et al., 2013] where the so-called SBS2 DNA sequence motifs contains the motifs of both THAP11 and ZNF143.

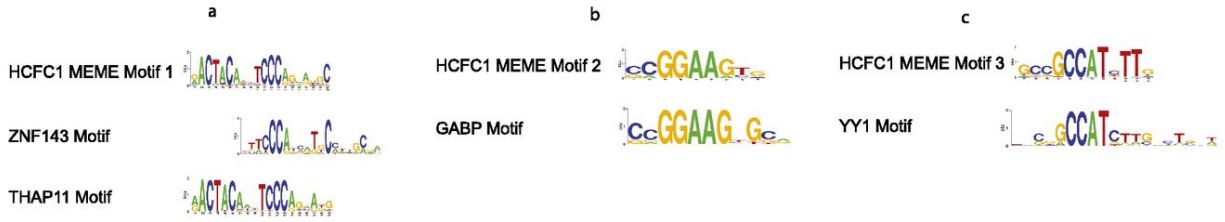



**Figure I-17**. Description of the DNA sequence motifs enriched in HCF-$1_C$ and HCF-$1_N$ binding sites. Source: [Michaud et al., 2013]. A search of motifs from 6 to 20 bp long was performed within the 200 bp sequences centered on each HCF-1c peak found in promoter proximal sites. The analysis was performed with the bioinformatic tool MEME [Bailey et al., 1994]. The analysis retrieved 3 motifs indicated in the top of each panel. In the bottom of each panel, motifs from transcription factors showing similarities with the provided HCF-1c motifs are included. a) HCF-1 motif showing similarities with the aggregation of the motifs from the transcription factors ZNF143 and THAP11. This motif suggests a cooperative binding between the transcription factors ZNF143 and THAP11. b) HCF-1 motif showing similarities with the DNA-binding motif for GABP. c) HCF-1 motif showing similarities with the DNA-binding motif for YY1.

The motif enrichment analyses in [Michaud et al., 2013] did not report a strong association with other transcription factors such as the E2Fs. But a targeted analysis of the E2F1 motif in HCF-1 promoters found the presence of E2F1 motifs, although in just 2% of the HCF-1 bound sequences. Other proteins that would be expected to bind together to HCF-1 in HeLa cells such as SP1 were not enriched at HCF-1 binding sites [Wysocka et al., 2003]. As SP1 binds to CpG islands as HCF-1 does and it can do it to secondary structures of the DNA [Raiber et al., 2012] it may be difficult to detect enrichments of SP1 binding sites under HCF-1 binding sites.

## I.5 The mouse liver and its regeneration

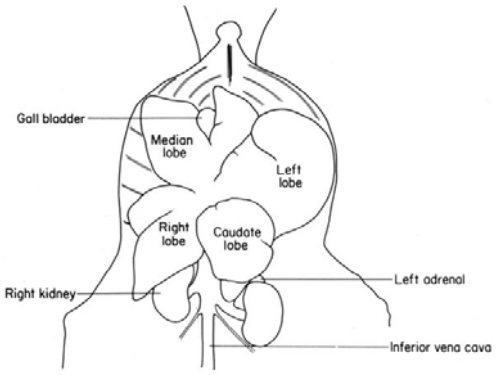
### I.5.1 The mouse liver

The liver performs essential metabolic, exocrine and endocrine functions. These include metabolism of dietary compounds, detoxification, regulation of glucose levels through glycogen storage and control of blood homeostasis by secretion of clotting factors and serum proteins.

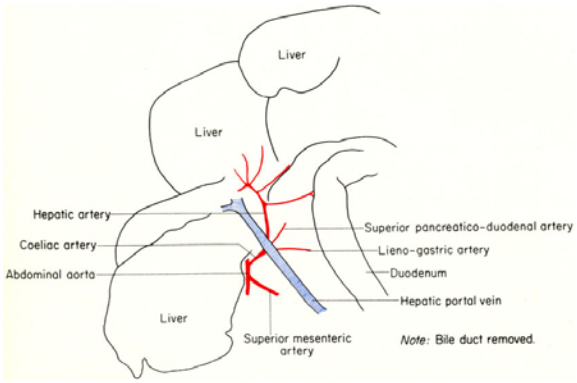
The liver is the largest internal organ in the mouse and it is formed of seven lobes with different sizes and shape (Figure I-18). The liver is connected to two large blood vessels, the hepatic artery and the portal vein. The hepatic artery brings oxygen-rich blood from the aorta, whereas the portal vein transports blood with digested nutrients from the entire gastrointestinal tract, the spleen and the pancreas. Once the blood is in the liver it is directed to the lobes by vessels. In addition, the liver produces, transports and stores bile, which aids in the emulsification and digestion of lipids in the small intestine. The bile is collected in the bile canaliculi, which merge to form bile ducts, giving name to the biliary tree. The gall bladder, sitting in the median lobe, is a small organ that stores bile temporarily to then release it into the small intestine.

The liver is a highly homogeneous organ where hepatocytes account for approximately 70% of the mass. The other 30% is composed by other non-parenchymal cell types that are highly organized. The liver is organized in lobules, which constitute its functional units (Figure I-19a). Each lobule is composed of a central vein, from which hepatocyte cords radiate towards portal triads. And the portal triad consists of a portal vein, hepatic artery and biliary duct. Hepatocytes close to the portal triad receive more oxygen as they are closer to the entering vascular supply, while hepatocytes close to the central vein have the poorest oxygenation and they receive nutrient-rich blood from the hepatic artery. Thus hepatocytes in different zones are specialized in different functions. The periportal hepatocytes specialize for gluconeogenesis, β-oxidation of fatty acids and cholesterol synthesis, while the central ones are more relevant for glycolysis, lipogenesis and cytochrome P-450-based drug detoxification [Schiff et al., 2007]. Hepatocyte cords are single-cell sheets of hepatocytes separated by sinusoids that carry blood from the portal triads to the central vein (Figure I-19b). The sinusoids are built from specialized endothelial cells of the liver. Stellate (or Ito) cells are located in the space of Disse between the hepatocyte cords and sinusoids. Also, Kupffer cells, which are the specialized macrophages of the liver, reside in sinusoids. Cholangiocytes are the epithelial cells lining the bile ducts

a)




b)




* Alimentary canal removed

**Figure I-18**. a) Front and b) back views of the mouse liver anatomy. Source: Website http://www.protocol-online.org/

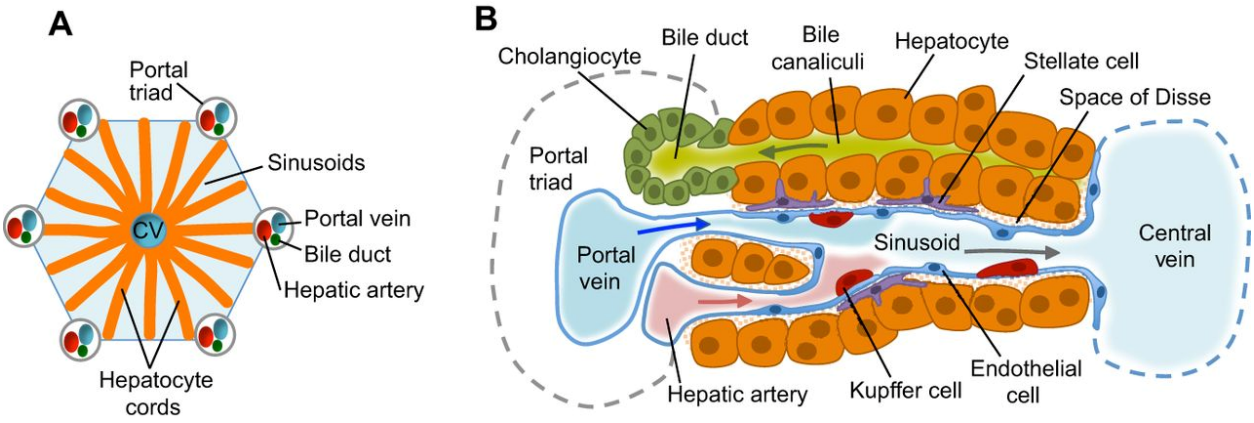



**Figure I-19**. Schematic of the liver lobules and the diversity of cell types composing the sinusoids. Source: [Gordillo et al., 2015]. a) The lobules are the functional structures of the liver. They have a hexagonal shape. In the center there is a central vein from which hepatocyte cords radiate towards the portal triads. The portal triad consists of a portal vein, hepatic artery and biliary duct. Hepatocyte cords are single-cell sheets of hepatocytes separated by sinusoids that carry blood from the portal triads to the central vein. b) Structure of the sinusoids and the diversity of cell types that form them.

## I.5.2 Mouse liver regeneration

Mouse liver regeneration induced by disease or injury is a compensatory hyperplasia, which means that an increase in the number of cells in the healthy liver happens in order to reestablish the proper liver size and functions. Liver regeneration after a 2/3 partial hepatectomy (PH) is a widely used model to study the mechanism of liver regeneration [Mitchell and Willenbring, 2008]. During the PH approximately 65-70% of the liver is removed, inducing the liver to regrow with tightly synchronous rounds of replication. Given that the remaining liver is not injured, it offers a physiological way to study mammalian cell proliferation in vivo.

The process of liver regeneration after PH does not rely on the action of stem cells. Although liver stem cells may contribute to the process, each cell type has the capacity to enter into a proliferative state [Widmann and Fahimi, 1975; Ukai et al., 1990; Malik et al., 2002]. Hepatocytes are the first cell types to proliferate during regeneration [Bouwens et al., 1986]. During the entire process the liver cells continue to perform crucial metabolic functions such as glucose regulation, synthesis of many blood proteins, secretion of bile, and biodegradation of toxic compounds required for homeostasis.

During the first 4 hrs after PH (priming phase) liver regeneration is induced by a series of signaling events. Paracrine and endocrine signals from non-parenchymal hepatic cells initiate liver regeneration in the early minutes after the surgery. In the portal blood there is an increase of gut-derived factors, inflammatory mediators of the immune response and immunoglobulin superfamily intracellular adhesion molecules (IgSF CAMs). This in turn activates Kupffer cells that subsequently induce hepatocyte proliferation by releasing the growth factor Tumor Necrosis Factor α (TNFα) and interleukin 6 (IL6) cytokines in a paracrine manner [Widmann and Fahimi, 1975; Strey et al., 2003; Fausto, 2006].

Transcriptional cascades are involved in the progression of mouse liver regeneration. Within 30 minutes post-PH the first transcription factors are activated by post-translational modifications, including NFκB, Stat3 and AP-1. And in parallel, new expression of transcription factors takes place such as for the highly responsive c-Jun and c-Fos, observed already 10 minutes after surgery, the growth factor EGR1, by 30 minutes post-PH, and the regulator of cell growth c-Myc, 40 minutes post-PH [Su et al., 2002].

Hepatocyte entry to cell proliferation is highly synchronized. On mice under light and food cycles of 12 hrs hepatocytes enter to the first round of G1 phase around 10 hrs post-PH. Synthesis phase takes place approximately at 36 hrs post-PH and it is followed by M phase at around 48 hrs post-PH (Figure I-20) [Minocha

et al., ms. in prep.]. In 1 week the normal function of the liver is almost reestablished. And the totality of the function is acquired in 3 weeks.

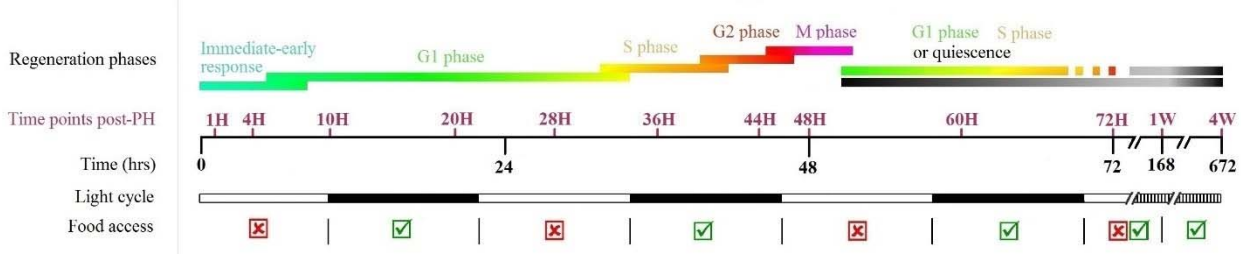



**Figure I-20**. Timing of the mouse liver regeneration phases. Description of the mouse liver regeneration stages observed during the study of liver samples collected at the specified time points post-PH. Mice were under light and food cycles of 12 hrs. Source: adapted from a figure created by Dominic Villeneuve.

## I.6 The Omics era

In the past decade many technological advances have taken place in the life sciences. A big effort has been made to develop technologies that allow the study of many molecules. Different research domains have appeared that study the totality of a specific type of molecule in cells or organisms. Thus the totality of molecules have been given names with the suffix '-ome', that refers to a totality. And the research domains studying these molecules are named with the suffix '-omics'. One example of these domains is genomics. This field intends to study the entire genome that is formed by the totality of the chromosomal sequences contained in cells. Another example is transcriptomics, which is the study of the transcriptome (ie. totality of transcripts contained in cells). These two domains have been revolutionized by the emergence of sequencing machines able to sequence the entire genome and transcriptome. Other Omic domains include lipidomics (for lipids), proteomics (for proteins) and metabolomics (for metabolites).

The life sciences are thus currently adapting in many ways to a new large-scale way of experimentation. There is consequently a need to educate new scientists with interdisciplinary expertise that combines experimental approaches with computation to handle the vast amounts of information generated. And better computational and data storage structures are necessary to support scientific research.

## I.6.1 Evolution of the sequencing technologies

The possibility to sequence the DNA revolutionized the research in biology. The very first approach for DNA sequencing was developed by Frederick Sanger and colleagues in 1977 [Sanger et al., 1977]. This technique was based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication. For 25 years Sanger sequencing was widely used by the scientific community. In the late 90's new methods for capillary DNA sequencing were patented, but it was during the early years of the 21st century when sequencing platforms reached the market with the appearance of the Next Generation Sequencing (NGS) that reduced the costs significantly (Figure I-21).

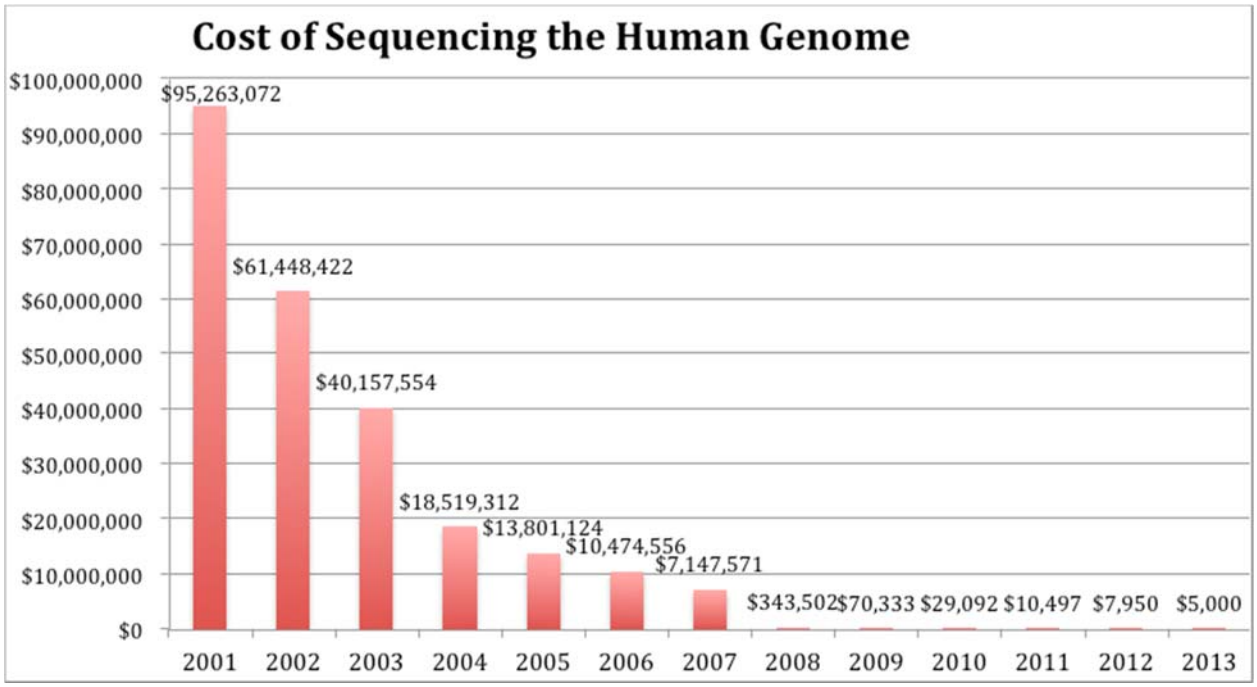
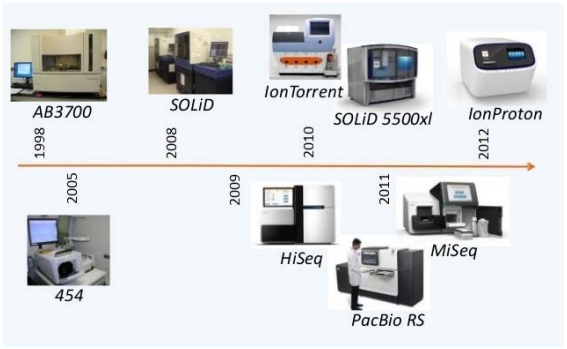



**Figure I-21**. Progression of the cost of sequencing an entire human genome since 2001. Source: NIH. The x-axis represents the years from 2001 to 2013. And the y-axis depicts the price in US dollars. Red bars illustrate the price of sequencing a human genome per year.

Multiple NGS machines have been developed by private companies (Figure I-22a). Broadly, previous to sequencing, a library needs to be prepared containing pieces of fragmented DNA with ligated sequencing adaptors at the ends. Library amplification takes place on some sort of solid bead or glass surface, depending on the platform. Then sequencing of the amplified molecules is performed [Metzker, 2010]. Two approaches were developed for this. The first is sequencing by synthesis were a polymerase synthesizes the DNA and

imaging detects which nucleotides are added at a time. In contrast, the second approach involves sequencing by ligation, where a DNA ligase identifies the nucleotide present at each position thanks to its mismatch sensitivity. Since 2010 platforms are available such as HiSeq from Illumina, IonTorrent, from Ion Torrent Systems Inc (now owned by Life Technologies) and SOLiD5500xl, from Applied Biosystems (now a Life Technologies brand). And in 2011 the new sequencers Miseq, from Illumina, and PacBio, from Pacific Biosciences, appeared. Miseq is a scaled-down version of the previous Illumina HiSeq machines. And PacBio combines nanotechnology with molecular biology and performs true single-molecule sequencing for either short or long DNA molecules. Currently, there is a lot of choice of technologies that perform with low sequencing error rates and are progressively more affordable (Figure I-22b). In addition, nowadays it is possible to multiplex the sequencing, by adding different libraries in a sequencing lane that were previously labeled with indexes differently. All these improvements in the last two decades are definitely pushing genomics and transcriptomics research forward permitting rapid and cost effective complete genome analyses.

a)




b)

| Feature | HiSeq2500 - Highoutput | HiSeq2500 – Rapid mode | MiSeq | PacBio RSII |
|---|---|---|---|---|
| Number of reads | 150-180M/lane | 100-150M/lane | 12-15M (v2) 20-25M (v3) | 50-80K/SMRT cell |
| Read length | 2 x 100 bp | 2 x 150 bp | 2 x 300 bp (v3) | ~ 10-20 kb |
| Yield per lane (PF data) | up to 35 Gb | up to 45Gb | up to 15 Gb | up to 0.4 Gb |
| Instrument Time | ~12-14 days | ~2 days | ~2 days | ~2 hours |
| Pricing per Gb | $59 (PE100) | $53 (PE150) | $108 (PE300) | $697 |

**Figure I-22**. Evolution of the sequencing technologies and the features of some of the current technologies. a) Names of the technologies appearing throughout the years since 1998. Source: Modified from SA Pathology. b) Features of widely used sequencing technologies nowadays (HiSeq2500 from Illumina in the highthroughput and rapid modes, MiSeq also from Illumina and PacBio RSII, from PacificBiosciences). Source: website of the UCDavis Genome Center dnatech.genomecenter.ucdavis.edu/.

## I.6.2 The ChIP-seq technique for the study of the genome-wide binding of transcription regulators

Chromatin Immunoprecipitation (ChIP) followed by sequencing (-seq) of the recovered pieces of DNA permits the study of transcription factors in the genome-wide regulation of transcription. This technique reveals protein

binding patterns along the entire genome in vivo. It gives the possibility to study RNA polymerase occupancy patterns along with the occupancy of other proteins like histones and chromatin binding proteins such as chromatin and transcription factors. This opens the possibility to identify transcriptional activities and to understand the mechanisms that govern them.

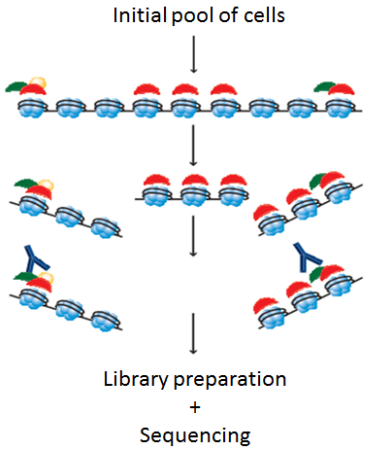



**Figure I-23**. Schematic of steps to follow in the ChIP-seq technique. From an initial pool of cells, the chromatin is fixed and then extracted from cells. The DNA is subsequently sheared and the pieces bound to the protein of interest are immunoprecipitated. The selected genomic fragments are de-crosslinked and prepared for sequencing.

ChIP experiments target the chromatin where a specific protein is bound and recovers the associated piece of DNA. The steps done in a ChIP are the following (Figure I-23):

1. Fixate the chromatin, typically with formaldehyde,
2. Extract the chromatin from cells,
3. Break the double stranded DNA into pieces (by sonication or DNA digestion),
4. Target the chromatin bound by the protein of interest with an antibody, and
5. De-crosslink the purified chromatin to obtain double stranded DNA.
6. Typically, library preparation includes the following steps:
    i. Addition of sequencing adapters to the ends, to allow fragment amplification and sequencing (Figure I-24).
    ii. PCR amplification of the starting material, to have the necessary amount of DNA to run the sequencing. There are a few drawbacks of doing this step: the amplification is known to be biased in favor of short and AT-rich fragments.
    iii. Size selection of the fragments.

7. Finally, the libraries are sequenced.

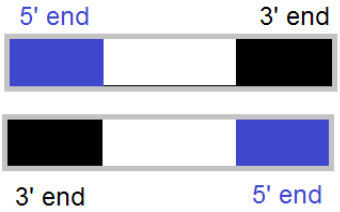



**Figure I-24**. Schema of a fragment of double-stranded DNA showing the 5' and 3' ends of each strand.

The sequencing step retrieves either the nucleotide sequence at the ends of the fragments or their entire sequence, depending on the technology used. For instance, one of the most frequently used technologies is the sequencing-by-synthesis, such as the one implemented by Illumina Inc. Illumina Inc has implemented a technology that sequences fragment ends. The traditional version of this technology, called single-end sequencing, is as follows:

- First, the two strands of DNA from the fragments are separated,
- They are randomly positioned in a solid phase and amplified creating clusters of single stranded DNA.
- Primers are added and extended with nucleotides linked to dyes of a specific color. The color of the dye bound to nucleotide added at each cycle is captured by digital imaging which can be translated into the nucleotide sequence of the 5' ends.

Recently, the paired-end sequencing technology has appeared in platforms such as Illumina GA IIx and Illumina HiSeq 2500. The paired-end sequencing is an extension of the traditional single-end sequencing technology. It sequences the 5' and 3' ends of a strand and keeps them paired. After doing all the single-end steps, the paired-end sequencer:

- Synthesizes in the same location of the solid phase the strand complementary to the strand already sequenced.
- Later, it gets rid of the previously sequenced strand and repeats the sequencing procedure for the complementary strand.
- At the end, the sequencing platform retrieves paired ends coming from the same fragment.

Usually, libraries from non-immunoprecipitated chromatin (Input) are sequenced as a control in ChIP-seq experiments. The Input fragments should in principle be random, but many times accumulations of fragments can be observed in certain genomic regions. So far, the reason for this is unknown.

The end sequences retrieved from sequencing, so-called reads, can be mapped onto a reference genome, thanks to the current availability of genome assemblies from many different species. To know where ChIPed DNA fragments come from helps the identification of protein binding sites. In the case of single end sequencing data, the fragment ends are not representative of all the genomic positions covered by the fragment because its size is not known. To gain resolution when inferring binding sites, the reads are typically shifted by an average distance towards the opposite end of the fragment. As with single-end data it is not possible to know the precise size of genomic fragments, the shift is set at half of the most frequent fragment size in the samples prepared for sequencing. There are different ways to estimate the most frequent fragment size. One way is to use one of the microfluidics based platforms for sizing the DNA of an extract of the library that will be sequenced. Another way of approximating the most frequent fragment size is by doing in-silico cross-correlations between corresponding peaks located on the positive and minus strand of the genome. Finally, preferential binding sites can be inferred by comparing the quantifications of fragments obtained from the ChIP with the Input ones in specific genomic locus.

The computational analysis of ChIP-seq data from RNA Polymerases, histones, transcription factors and co-transcription factors can be very useful to understand the mechanisms of transcriptional regulation and to identify transcriptional activities. Moreover, time course studies of ChIP-seq data can reveal insights on transcriptional activities occurring along cellular processes, such as cell proliferation.

### I.6.3 The RNA-seq technique for the study of the transcriptome

High-throughput RNA sequencing (RNA-seq) has become the method of choice for transcript profiling, discovery of new transcripts and identification of alternative splicing events. Its quantitative nature, reproducibility and depth has made it a useful tool for the study of gene expression compared with microarray-based methods [Wang et al., 2009].

RNA-seq takes advantage of reverse transcription to convert the RNA contained in a given sample to its corresponding cDNA sequence (Figure I-25). Finally, a sequencing library is prepared from the obtained cDNA and then sequenced. Typically the sequencing of RNA-seq libraries can be done with strand-specificity.

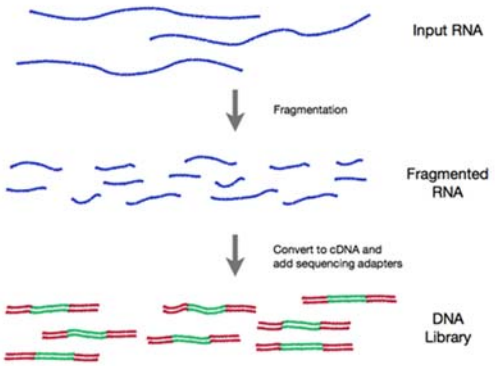



**Figure I-25**. Description of the preparation of a DNA library for RNA-seq. Source: website http://rnaseq.uoregon.edu/. The input RNA is fragmented and then reverse transcribed to obtain cDNA (green). Adapters are subsequently added for sequencing (red).

There exist different ways to prepare the RNA samples. In cells the ribosomal RNA (rRNA) transcripts, which originate from a couple of hundreds of genes, are very abundant (80% of the cellular transcriptome). If all the transcripts in cells are sequenced, because there is a limited number of sequences that can be obtained during sequencing, one would have very low or no coverage of transcripts other than rRNAs genes. To reduce the number of rRNA sequences different sample preparation protocols have been developed. The selection of PolyA transcripts is one option, as this would target all the protein coding mRNAs and a few types of non-protein coding RNAs that don't include rRNA. Alternatively, to allow the study of non-coding RNA's other than rRNA another solution is to deplete only ribosomal RNA. This is possible by competitive hybridization or other non-disclosed methods from private companies, such as Illumina.

To infer the likely origin of each fragment, the sequencing data can be mapped onto the genome of interest by using bioinformatic tools such as Bowtie [Langmead and Salzberg, 2012], TopHat [Kim et al., 2013] and Star [Dobin et al., 2012]. Each of them uses different mapping strategies some of them more adapted to specific questions, such as splice site discovery, isoforms or overall gene expression. Accordingly, specific expression analyses can be done to answer different biological questions, such as differential expression [Anders et al., 2013].

## Conclusions

The cell cycle is a process in living beings that produces great changes in cells. And these changes are highly controlled by waves of gene expression. A complete understanding of this fundamental process could contribute to the design of new medical approaches for diseases and disorders. To date a lot is known about key regulators of cell proliferation, such as cyclins, cyclin dependent kinases and transcription factors. The establishment of cell culture protocols in the second half of the last century facilitated the research about the cell division cycle on modified cells such as HeLa. Also, models from healthy organisms are used such as yeast or higher vertebrate organisms more genetically similar to humans such as mice. For example, mouse liver regeneration is an excellent physiological process for the study of cell proliferation. The mouse liver is one of the most homogeneous organs in mammalian bodies and upon injury cells can re-enter proliferation in a highly synchronized manner. Research on mouse liver regeneration has already provided many insights on the coordination of cell proliferation in a healthy organ where other key metabolic functions need to be maintained.

During cell proliferation a precise regulation of gene expression occurs. This system confers the possibility to learn about the regulation of transcription. The co-transcriptional factor Host Cell Factor 1 (HCF-1) has been shown to be required for a proper progression of the animal cell cycle. Indeed, its maturation by proteolytic cleavage and the non-covalent binding of the resulting pair of subunits is necessary for its transcriptional activity during the G1/S passage and mitosis. Interestingly, this protein interacts with diverse transcription factors and chromatin remodelers associated with different transcriptional outcomes. The way the HCF-1 adaptor protein regulates transcription is not yet well understood. Given the versatile nature of HCF-1 further studies in a genome-wide fashion may provide new insights about different mechanisms to regulate transcription.

# Chapter II. Evaluation of the paired-end sequencing technology for ChIP-seq studies

With single-end sequencing data, for a given fragment end that has been sequenced, it is possible to know its genomic origin by mapping it onto the reference genome. Nevertheless, there is no way to know where the other end of the same fragment maps. This introduces difficulties in the data analysis. First, there is no way to know the precise position of the entire ChIP-ed fragment, which complicates the precise identification of protein binding sites. Second, it is not possible to precisely identify the redundancy within a dataset. Fragment redundancy can be caused by natural repetitions present in the initial chromatin pool or by steps during the preparation of the sequencing library, such as PCR amplification. And this amplification is likely biased in favor of short and AT-rich fragments. Paired-end sequencing data, in contrast, links the two ends of a given fragment, and thus the size of fragments can be calculated and the repeated fragments can be precisely identified. In this Chapter I evaluate whether the information provided by paired-end sequencing is more useful than single-end data for ChIP-seq studies. This work was done in the context of the CycliX consortium and, although the studies presented are not complete, the conclusions made in this chapter were useful for the design of the CycliX PH experiments discussed in Chapters IV to VI.

## Results

To evaluate paired-end sequencing, ChIP-seq libraries were prepared from two pools of HeLa cells called Experiment 1 (E1) and Experiment 2 (E2) (Figure II-1). Cross-linked chromatin was extracted and sonicated. Subsequently, a ChIP with anti-H3K4me3 antibody was done, retrieving genomic fragments expected to be in nucleosomes. Three libraries (Lib1, Lib2 and Lib3) were prepared from sequential gel size selections (Lib1: 100-200 bp; Lib2: 200-300 bp, Lib3: 300-400 bp), as is typically done when performing single-end sequencing. From the Experiment 2, two technical replicates of the library preparations were done (E2-R1 and E2-R2) with size selections, as in E1. Further, one library was prepared from chromatin with DNA purification done with magnetic beads on sizes between 200-500 bp (E2-Beads). For further details, see the Methods section in this chapter.

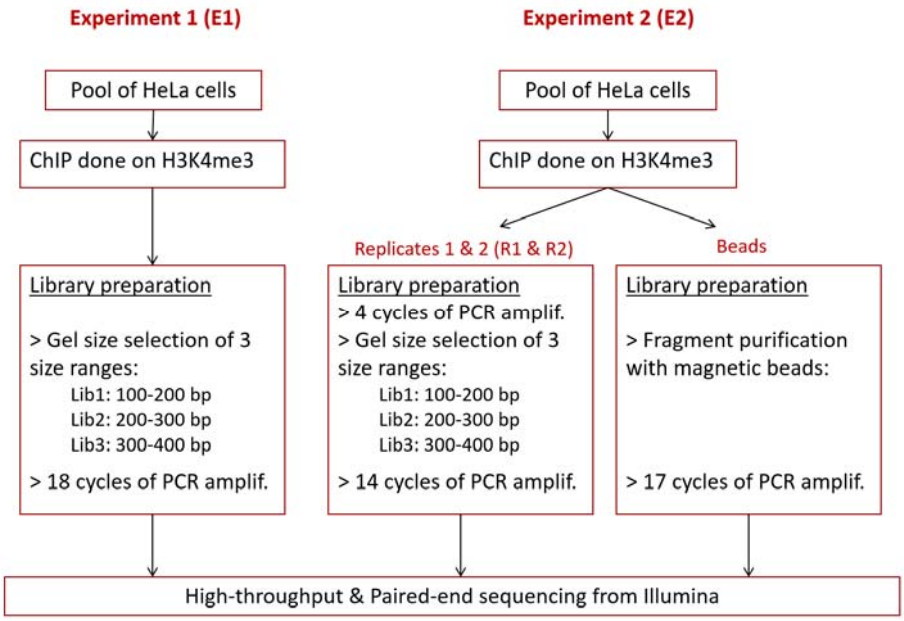



**Figure II-1**. Scheme for paired-end data preparation. Two independent pools of chromatin were prepared (E1 and E2). ChIPs were done on each chromatin sample with anti-H3K4me3 antibodies. Subsequent library preparations were done with different protocols. For E1 only one group of three sequential size-selected libraries was prepared. Whereas for E2, two library preparation replicates (R1 and R2) and one library with DNA purification with magnetic beads were prepared. Finally, fragments were sequenced with high-throughput and paired-end sequencing technology from Illumina.

## II.1 Study of the redundancy of sequenced fragments

After mapping the sequenced paired-end reads onto the human genome, it is possible to distinguish among the actual fragments those having a single copy (S1) in the entire data set and those repeated (S2) (Figure II-2a). Further, it is possible to identify the unique fragments contained within a given data set, some of them originating from single-copy fragments (S1), and others originating from repeated fragments (S3) (Figure II-2b).

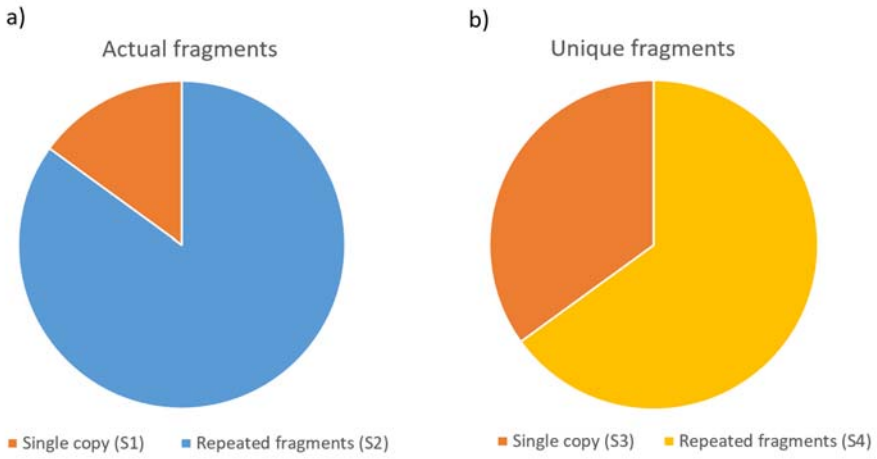



**Figure II-2**. Definition of distinct types of fragments identified with paired-end data. a) Among the entire set of sequenced fragments, some are observed as single copies (S1) while others are repeated (S2) to different degrees. b) The unique fragments within the entire data set can originate from either single-copy fragments (S3) or repeated fragments (S4).

Figure II-3a-c shows the fragment redundancy distribution for Lib1 in each experiment E1, E2-R1 and E2-R2. Note that in this section, to accelerate the analyses, only either chromosome 1 or 2, the two largest human chromosomes were analyzed; I did not observe any differences between these two chromosomes in these analyses. As shown in Figure II-3a-c, in these libraries approximately 30% of the unique fragments originate from single copies (S3), and about 70% of the unique fragments originate from repeated fragments (S4). The precise identification of single-copy and repeated fragments with paired-end data is important during the analysis of the data as the use of a specific degree of redundancy can help to diminish the effects of PCR amplification. In contrast, single-end data can only inform about the redundancy of the 5' or 3' ends of fragments. Thus single-end data doesn't allow a precise identification and use of a specific degree of redundancy to diminish the effects of PCR amplification. To evaluate how imprecise the identification of redundancy with single-end data is compared to paired-end data, I used the paired-end data from library E2-R1-Lib1 to simulate single-end data by separating the two sequence ends and counting the redundancy on the simulated 5' and 3' ends (Table II-1). With paired-end data, 29.1% of the unique fragments are identified as originating from single-copies (S3). In contrast, when simulating single-end data, only 4% of 5' ends are identified as single-copies and only 4% of the simulated 3' ends are identified as single-copies, adding up to a total of 8% of potential single-copies of ends identified. This shows that paired-end data improves by three-fold the identification of single-copy fragments (S3) and thus the repeated fragments (S4) and their degree of redundancy (S2).

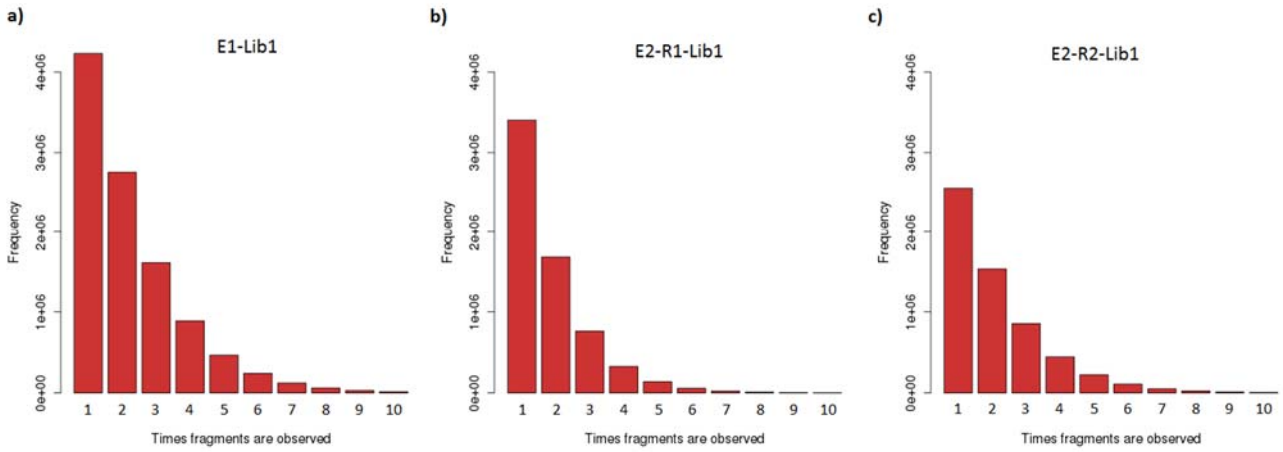



**Figure II-3**. Redundancy of unique fragments. a) Times unique fragments are observed in library E1 Lib1 (chromosome 1), b) library E2-R1 Lib1 (chromosome 2) and c) library E2-R2 Lib1 (chromosome 2). Only up to 10 repetitions are indicated. The first column represents unique fragments originating from single copies (S3) and from 2 to 10 represent the unique fragments originating from repeated fragments (S4).

| | PE fragments | SE 5' ends | SE 3' ends |
|---|---|---|---|
| Total events | 11,709,621 | 11,709,621 | 11,709,621 |
| Single-copy events | 29.1% | 4.2% | 4.2% |
| Repeated events | 70.9% | 95.8% | 95.8% |

**Table II-1**. Identification of redundancy with paired-end data versus with simulated single-end data. Data from E2-R1 Lib1 (chromosome 2) was analyzed. To simulate single end data, the paired 5' and 3' ends were separated. Thus three different events were analyzed: Paired-end fragments, single 5' ends and single 3' ends. A total of 11 mil events were counted in each case. And the percentages of single-copy and repeated events are indicated for each case.

I used the paired-end data to investigate the nature of redundancy. There are two possible sources for repeated fragments: identical fragments originating from the original chromatin sonication and linker ligation, and fragments arising after PCR amplification of a shared parent fragment. To study whether there can be natural repetitions, I compared replicate libraries originating from the same pool of cells (E2-R1 Lib1 and E2-R2 Lib1). As a result, approximately 10% of the unique fragments from each library are shared and 98% of them (data

not shown) are present in promoter regions of +/- 1kb around TSS. This shows that there can be naturally repeated fragments (Figure II-2f). Further, I investigated the size of single-copy fragments (S3) and repeated fragments (S4). Figure II-2d shows the size distribution for unique fragments (S3) or for fragments present up to 7 times in the data set. As the fragments are present in more copies, the size distribution gets shorter. This is indicative that shorter fragments are preferentially repeated, consistent with preferential PCR amplification of shorter fragments. Altogether, this shows that in the data there can be natural repetitions but also repeated fragments originating from biased PCR amplification that favors shorter fragments. Nevertheless, the ratio between natural and PCR amplified fragments could not be determined.

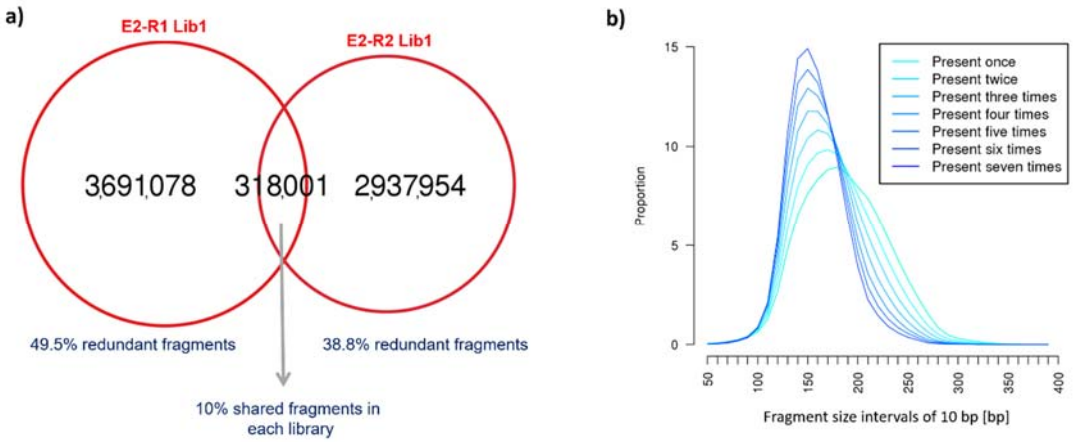



**Figure II-4**. Study of the nature of redundant fragments. a) Shared fragments between the E1-R1 Lib1 and E1-R2 Lib1 replicates on chromosome 2. Only unique fragments (S3+S4) with sizes 125-175 bp were taken for analysis. The percentage of repeated fragments (S4) is indicated below. Also, the percentage of shared fragments between libraries is indicated below. b) For specific number of redundancies (shown in blue lines) the proportion of fragments in a given size range (in 10 bp ranges) is shown. This analysis has been done with fragments in E1 Lib1 (chromosome 1).

A question to ask is how many copies of redundant fragments should be used in the analysis of this data to diminish the effects of PCR amplification. On this regard, I evaluated the use of one-copy of fragments (S3+S4) in an analysis similar to the one showed in Figure II-4b (data not shown). The proportional distribution of fragment sizes for one-copy of fragments (S3+S4) is identical to the distribution with fragments present twice, which tends to be shorter than the distribution for single-copy fragments (S3) as shorter fragments have been added (S4). Thus, the addition of more copies of repeated fragments would make the distribution to get shorter in turn, adding more PCR amplification effects in the analysis. Further, I evaluated the randomness of PCR

amplification and its effects in downstream analyses by comparing the effect of including different redundancy levels in the analysis (Figure II-5). Further, to discard potential effects due to the different fragment sizes only fragment sizes between 150-200 bp were used. I first did different selections of fragments with different levels of redundancy: i) only unique fragments (S3+S4), ii) unique fragments plus 1 more copy of the repeated ones (i.e., up to 2 copies of fragments), iii) up to 3 copies of fragments, iv) up to 4 copies of fragments, v) unique fragments plus the sequential addition of copies of repeated ones until reaching 95% of the total data, and vi) all the fragments in the data set. In each of these selections of fragments I did quantifications of the central 50 bp of fragments in promoters of annotated human genes (+/- 1 kb around TSS). Gene quantifications done when using all the fragments were sorted increasingly and displayed as a line in red. The rest of the quantifications done on each specific selection of fragments were sorted in the same order as for the quantifications done with all the fragments (red line) and displayed in colored lines (Figure II-5). Clearly, the ranking of genes changes when comparing the line with all the data and the line with only unique fragments, as the first draws a curve and the second traces a zigzag. Further, when comparing the line with only unique fragments and the line with up to 2 copies, it is visible that the amplitude of the zigzag is slightly smaller in the second. Thus addition of a second copy of repeated fragments also changes the ranking of genes, suggesting that promoters are differently affected by the redundancy. Thus, in this data, it seems that the best choice is to use only unique fragments for downstream analyses.

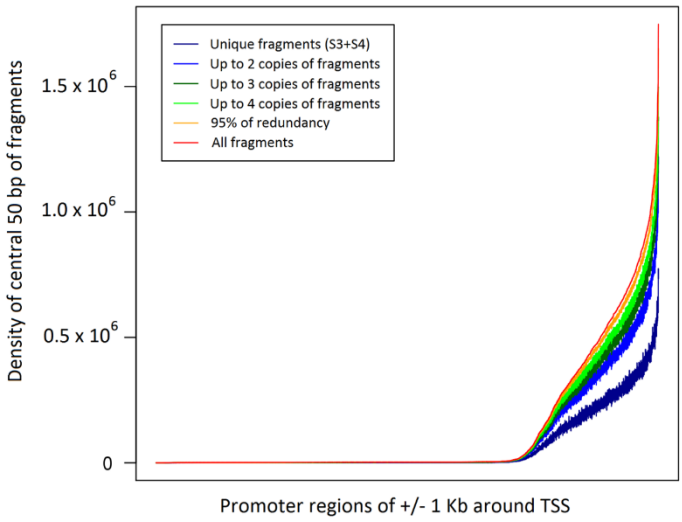



**Figure II-5**. Ranking of genes when using different degrees of redundancy for quantifications. Different pools of fragments were analyzed: i) only unique fragments (S3+S4) (dark blue), ii) unique fragments plus 1 more copy of the repeated ones (i.e., up to 2 copies of fragments) (light blue), iii) up to 3 copies of fragments (dark green), iv) up to 4 copies of fragments (light green), v) unique fragments plus the sequential addition of copies of repeated ones until

reaching 95% of the total data (orange), and vi) all the fragments in the data set (red). In each of these selections of fragments, densities of central 50 bp of fragments were quantified in the promoters of annotated human genes (+/- 1 kb around TSS). Only data on chromosome 1 and fragment sizes between 150-200 bp were used to discard effects due to the different fragment sizes. Gene quantifications done when using all the fragments were sorted increasingly and displayed as a line in red. The rest of the quantifications done on each specific pool were sorted in the same order as for the quantifications done with all the fragments and displayed in lines.

Furthermore, I studied the distribution of redundant fragments in the randomly selected *ANKRD20A11P* gene promoter (Figure II-6). The accumulation of 50 central bp of all the fragments draws two peaks, which suggest the existence of two positioned nucleosomes with H3K4me3 (N1 and N2). This is also observed when visualizing only unique fragments (S3+S4). Interestingly, the visualization of only single-copy fragments (S3), the inter N1-N2 region reaches similar levels as compared to N1 and N2, which disappears when displaying those fragments present twice in the dataset. As the fragments present twice may contain natural repetitions, this would seem to suggest that nucleosome fragments are more often natural repetitions. Because of the nature of the epitope targeted (H3K4me3), I'm looking at distinct chromatin structures that may favor natural repetitions in the data.

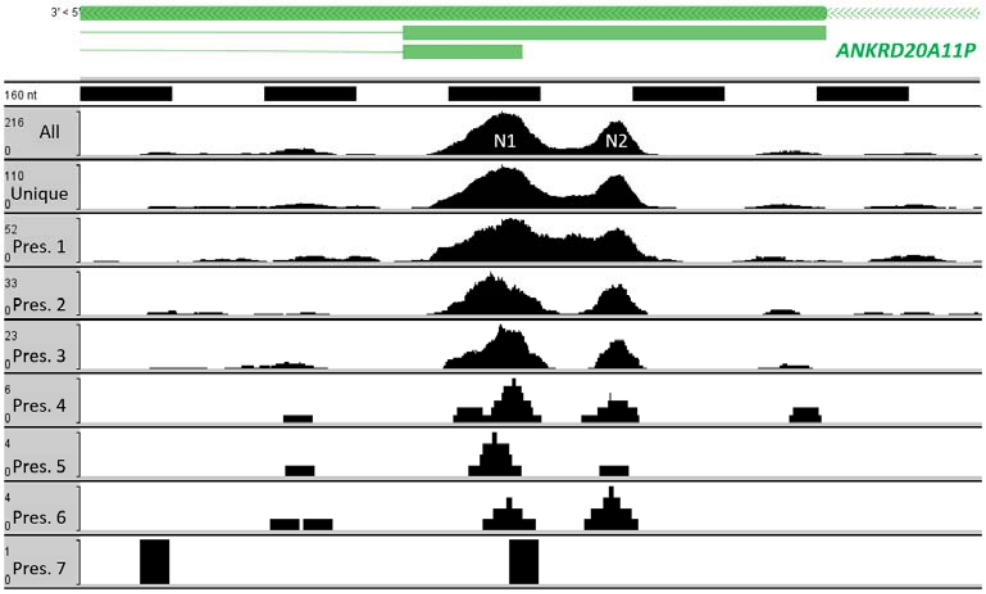



**Figure II-6**. Genomic view of the *ANKRD20A11P* gene promoter. On top, the promoter structure is described. The annotated transcription unit is described with a long green box on

the top, and the annotated exon-intron structures appear below depicted with rectangles and lines, respectively. A 1.6kb region in the promoter is shown. Data from the E2-Beads library were used for this figure. Data tracks display the accumulated central 50 bp of fragments after doing different selections of fragments: from top to bottom, i) all the fragments, ii) unique fragments (S3+S4), iii) only fragments present once (S3), iv) only fragments present twice, v) only fragments present three times, vi) only fragments present four times, vii) only fragments present five times, viii) only fragments present six times and ix) only fragments present seven times. Two nucleosomes containing H3K4me3 are labeled (N1 and N2).

## II.2 Analysis of the sequenced fragment sizes

Figure II-3 depicts de distributions of fragment sizes across all the libraries analyzed. The sizes included in each library of a given experiment replicate largely follow the size selection, although not as precisely as theoretically imagined (Figure II-7). Such overlaps could be owing in part to diffusion of fragments in the gel or, for example, imprecise cutting of the agarose either sideways or top down. Also, the size ranges are wider and smaller than originally desired (compare the results in Figure II-7 and the size selections described in Figure II-1), which shows that the gel size selections are not completely precise. Indeed, the bead preparation procedure which was meant to give a broad size selection gave a relatively precise size distribution of between 100-200 bp (see description of E2-Beads in Figure II-1).

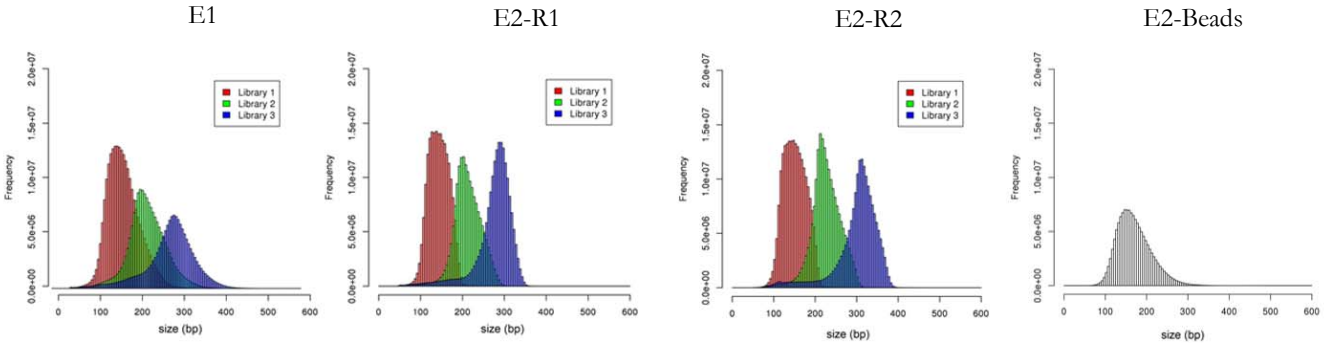



**Figure II-7**. Distributions of fragment sizes in the libraries from E1, E2-R1, E2-R2 and E2-Beads, as determined from paired-end sequencing results. For E1, E2-R1 and E2-R2, three libraries (Lib1 in red, Lib2 in green and Lib3 in blue) were prepared from the each ChIP, from sequential gel size selections (Lib1: 100-200 bp; Lib2: 200-300 bp, Lib3: 300-400 bp). And for E2-Beads, one library was prepared with DNA purification done with magnetic beads on sizes between 100-400 bp (E2-Beads). From the mapped paired-end fragments, fragment sizes were

calculated, one-copy of redundant fragments was selected and histograms were plotted with a resolution of 10 bp.

Theoretically, the sizes of sequenced fragments from a ChIP against H3K4me3 will vary randomly if the DNA shearing done during library preparation is random. A question to ask is whether these randomness would affect the identification of binding sites when using single-end data as in that situation it is not possible to shift the ends precisely towards the center of the fragments. The analysis of paired-end data from a ChIP against H3K4me3 provides insights on this regard. This analysis initially focuses on the study of the data from library 'E1-Lib1' on two cases of gene promoters: 2.5 kb promoter region of the circadian *CLOCK* gene and of the transmembrane protein *TMEM165* gene (Figure II-8). Track 'Lib1, t1' from the figure shows how both promoters have an accumulation of H3K4me3 that extends approximately 2.5 kb. In the case of the *CLOCK* gene, the accumulation of H3K4me3 extends in the entire promoter region displayed and has in its center one exon (green box), whereas in the *TMEM165* gene there is one peak upstream of the annotated TSS (leftmost position of the blue box) and it extends 2 kb downstream the TSS. Track 'Lib1, t3' in the figure showing centered 50 bp sequences displays with more resolution the potential localization of the H3K4me3, and thus H3K4me3-containing nucleosomes, which is nicely shown in the form of peaks in some cases clearly separated from each other. A good example of such separation is the 'x' peak on the *CLOCK* promoter, which contain a median fragment size of 145 bp, the size of a nucleosome, and probably represents an isolated nucleosome. The analysis of fragment sizes within the peaks reveal that they have diverse distributions of fragment sizes as displayed by their median and standard deviation (Figure II-8).

Interestingly, as shown in the tracks labelled 'Lib1, t2 Shifted 48 bp, cumulative', a simulation of shifted single-end data (the average half size of the fragments in the CLOCK region) results in a profile similar to the one obtained with paired-ends. This shows that single-end data of H3K4me3 after doing a size selection, although not ideal owing to different fragment sizes in different peaks, can provide similar information as the paired-end data. The study of the E2-R1 and E2-R2 replicate libraries lib1 in the *CLOCK* promoter further confirms the variability of fragment sizes under the peaks. Even though the peaks are quite maintained the peaks are built with diverse fragment sizes (Figure II-9).

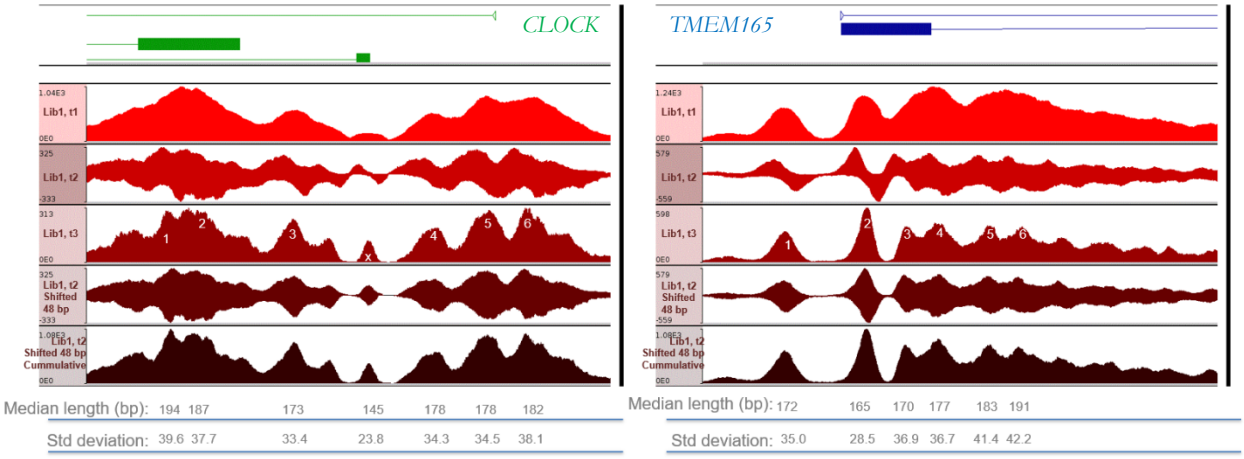



**Figure II-8**. Diversity of fragment sizes in the *CLOCK* and *TMEM165* gene promoters in E1-Lib1, and its effect in the analysis with paired-end vs simulated single-end data. The top of the figure shows the annotated genic regions with a thin line on the top, and the annotated exon-intron structures below drawn with rectangles and lines, respectively. The *CLOCK* gene promoter is antisense (green) and the *TMEM165* is sense (blue). Regions of 2.5 kb are shown. The sequenced ends are displayed in different ways in each track: type 1 (t1), the entire DNA sequence displayed; type 2 (t2), terminal 50 bp of + and – strands displayed; type 3 (t3), central 50 bp of each fragment displayed. In the tracks 'Lib1, t3' specific peaks have been numbered (1-6, and 'x' in the CLOCK region), and the median and standard deviation of fragment sizes is shown in the bottom of the figure for the fragments in each peak. This figure shows fragment-length variation between sites within and among different promoters.

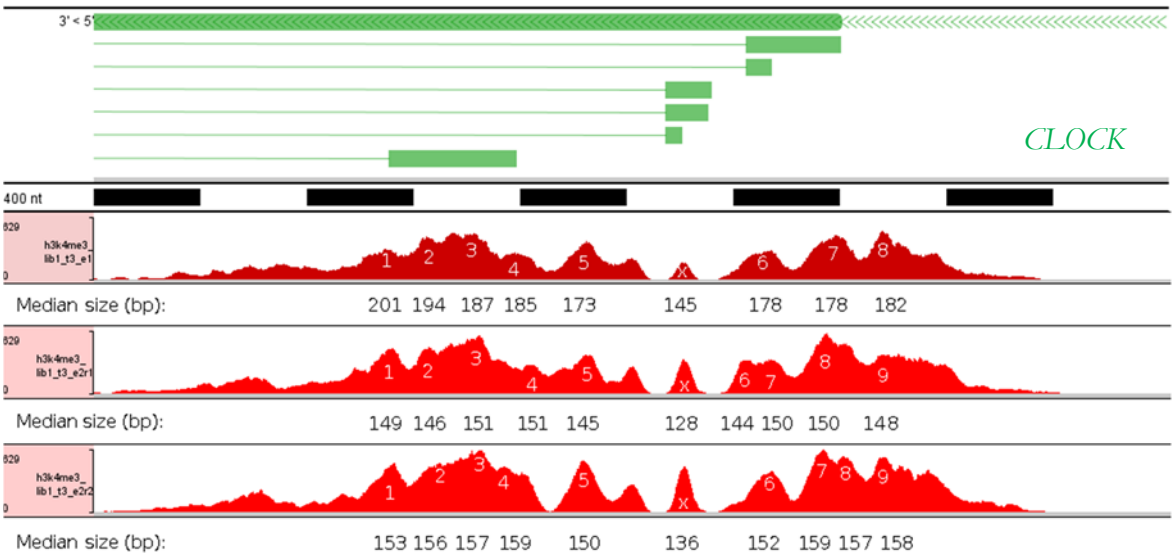


**Figure II-9**. Diversity of fragment sizes in the *CLOCK* gene promoter across libraries. The top of the figure shows the annotated genic region with a line on the top, and the annotated exon-intron structures below drawn with rectangles and lines, respectively. The *CLOCK* gene promoter is antisense (green). A 4kb region in the *CLOCK* promoter is shown. Tracks with the central 50 bp of fragments (t3) of a Library 1 are displayed. The first track displays data from E1-Lib1, the second shows data from E2-R1-Lib1, and the later shows data from E2-R2-Lib1. Under each track a space has been left to show the median fragment size calculated in the precise position of the labeled peaks.

Although within a specific size selection there is observed diversity, a more thorough study of the *CLOCK* promoter suggests that specific size ranges could be linked to specific chromatin compaction levels. Thus, the shearing of the DNA would not be completely random. Indeed the fragments associated to different size selections display different genomic profiles (Figure II-10). The 'x' peak, which could potentially represent an individual nucleosome, disappears when long fragment sizes are used for the analysis (E1-Lib3). Also the longer fragments are more focused on the left peaks and the shorter fragments show cleaner peaks on the right side. As the ChIPs were done against H3K4me3 and thus the fragments should fall within at least one nucleosome, these observations suggest that there is some restriction on the shearing of the DNA. This restriction could potentially be the degree of packaging of the nucleosomes, which at high density may be resistant to shearing. Thus, the fact that there are longer fragment sizes in the left region may reflect together nucleosome packing in this region. And this seems possible because the average fragment size in library 3 is 298 bp, which can fit two nucleosomes.

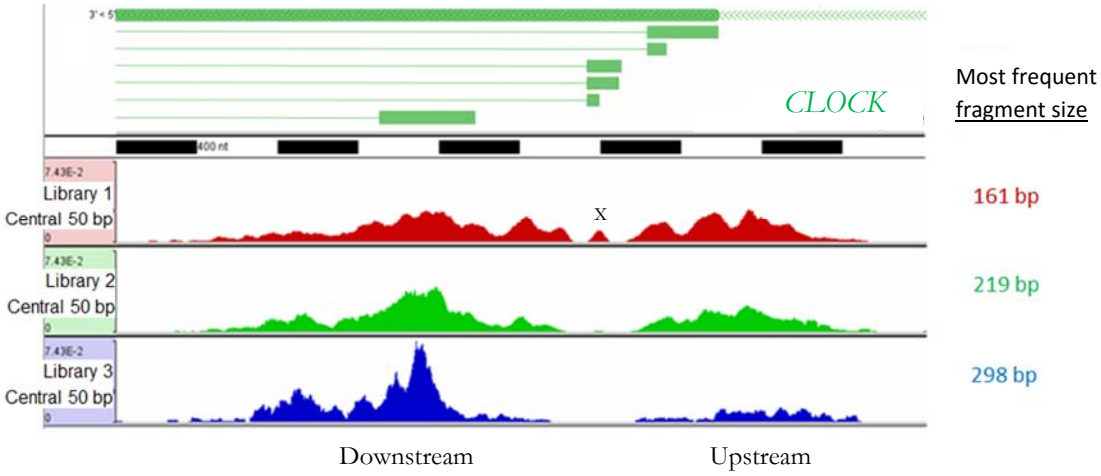



**Figure II-10**. Profiles in the *CLOCK* gene promoter for the three E1 libraries. The top of the figure shows the annotated genic region with a long box on the top, and the annotated exon-

intron structures below drawn with rectangles and lines, respectively. The *CLOCK* gene promoter is antisense. A 4kb region in the *CLOCK* promoter is shown. Tracks with the central 50 bp of fragments (t3) of a Library Lib1 are displayed. Data tracks display the accumulated central 50 bp of fragments in E1 libraries Lib1 (red), Lib2 (green) and Lib3 (blue). The 'x' peak is labeled in the Lib1 track and the downstream and upstream regions flanking the 'x' peak are indicated in the bottom.

To further investigate the nature of the *CLOCK* promoter, I used all the different E1, E2-R1 and E2-R2 libraries Lib1 and Lib3, as they display notable differences in profiles (Figure II-11). Libraries Lib1 have a similar profile across replicates. In contrast, libraries Lib3 display differences across replicates for unknown reasons. To further study the possible relation between fragment sizes and the underlying nucleosome organization, different displays of the fragments were prepared showing the centers of the fragments and the fragment ends, the later informing where the DNA was cut (Figure II-12). The information provided by the fragment centers and the cuts was used to infer the position of nucleosomes in the *CLOCK* promoter. Notably, frequent cuts occurred in 'Part A' of library Lib3 which interestingly leave space enough for nucleosomes of precisely 145 bp (Figure II-12). These potential positions of nucleosomes were further supported by the position of peaks in library Lib3. The centers of long fragments coincide with the positions between nucleosomes, suggesting that the long fragments contained two nucleosomes. 'Part B', in contrast, displays clear peaks composed of shorter fragments. Thus, nucleosomes could be inferred by centering them on the peaks. Interestingly, the inferred nucleosomes organization shows two distinct nucleosome packaging levels in the same promoter. Part A, has more compacted nucleosomes, whereas Part B tends to be less condensed (Figure II-12).

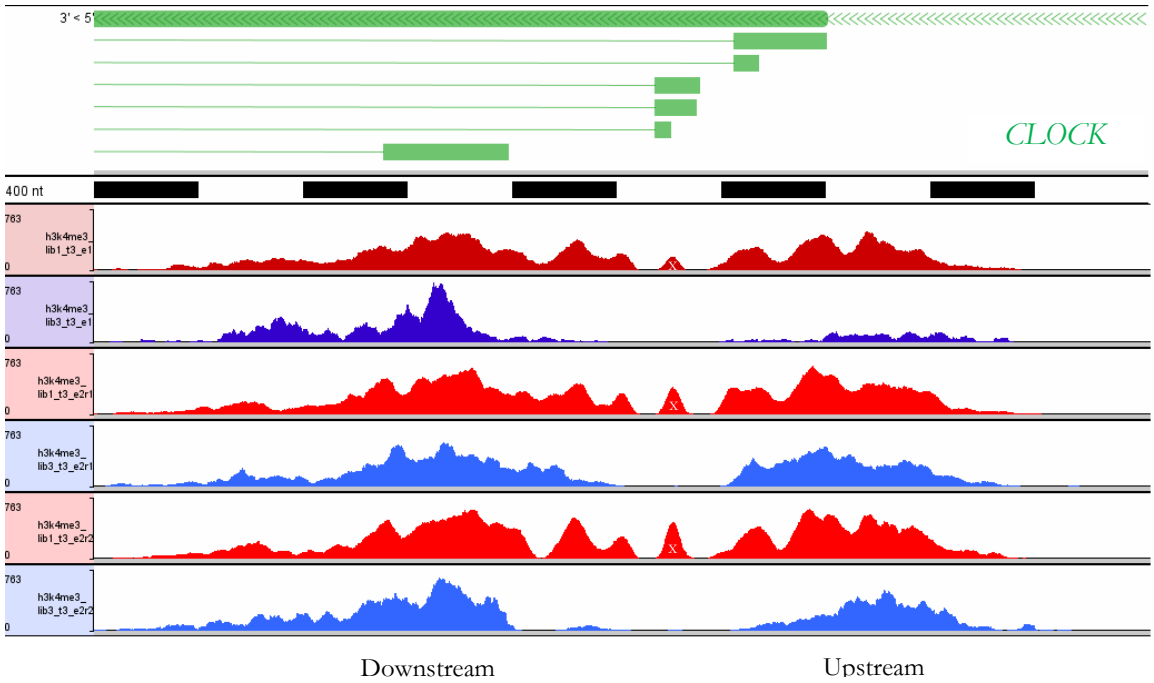


**Figure II-11**. H3K4me3 profiles in the *CLOCK* gene in the replicate Lib1 and Lib3 libraries. The top of the figure shows the annotated genic region with a long box on the top, and the annotated exon-intron structures below drawn with rectangles and lines, respectively. The *CLOCK* gene promoter is antisense (green). A 4kb region in the *CLOCK* promoter is shown. Data from library Lib1 (red) and Lib3 (blue) of E1, E2-R1 and E2-R2 are shown. The 'x' peak is labeled in the Lib1 tracks and the downstream and upstream regions flanking the 'x' peak are indicated in the bottom.

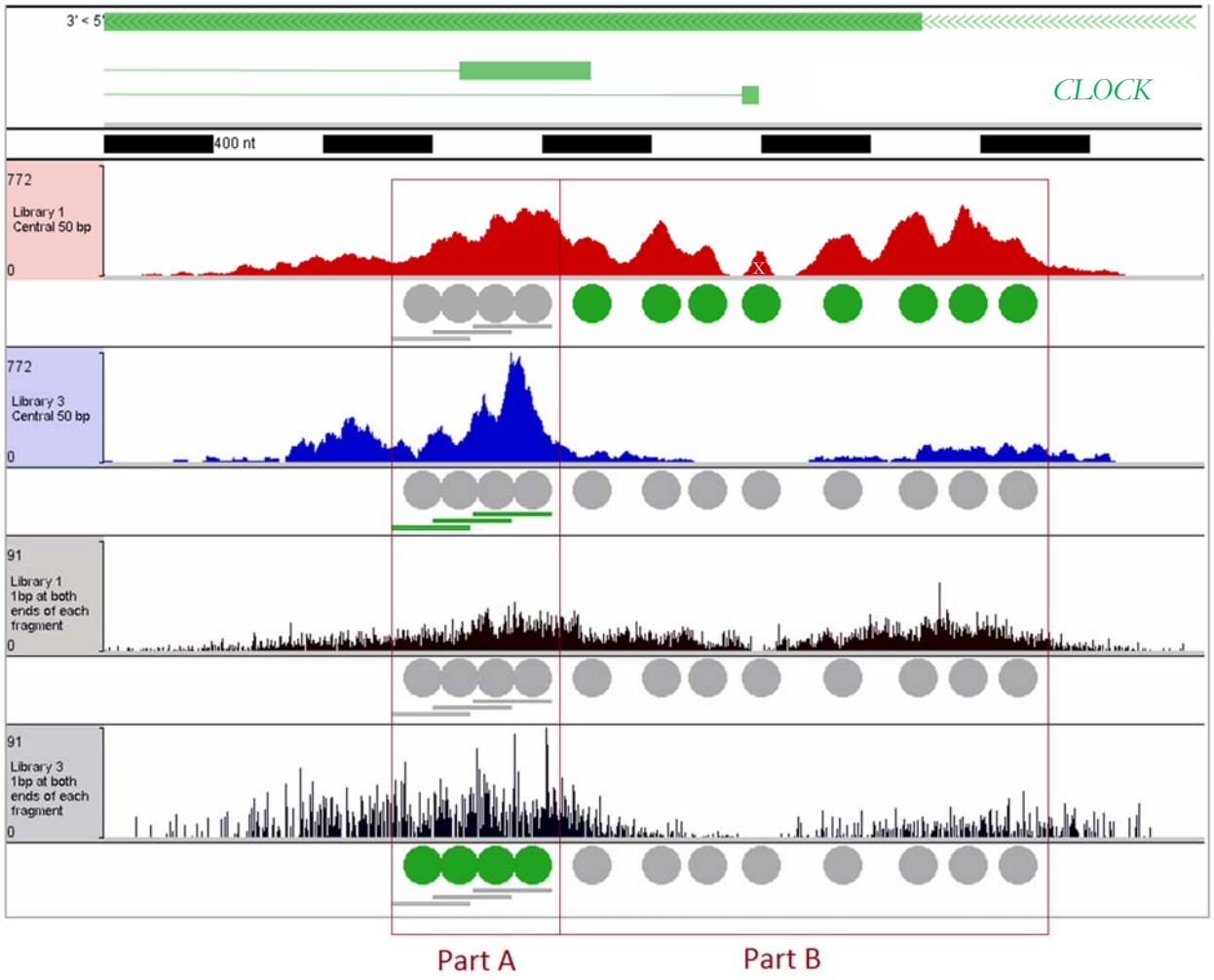



**Figure II-12**. Analysis of different group sizes in the *CLOCK* promoter unravelling the underlying organization of nucleosomes. The top of the figure shows the annotated genic region with a long box on the top, and the annotated exon-intron structures below drawn with rectangles and lines, respectively. The *CLOCK* gene promoter is antisense (green). A 4kb

region in the *CLOCK* promoter is shown. Data from E1 was used for the analysis. The first two tracks show the central 50 bp of fragments in E1 libraries Lib1 and Lib3, whereas the last two tracks display the profiles of the single nucleotide at the ends of fragments, for both E1 libraries Lib1 and Lib3. The information in Lib1 was used to infer mono-nucleosomes and the information in Lib3 was used to infer di-nucleosomes. A white space is left between tracks where circles are drawn depicting nucleosomes of 145 bp. Green circles refer to nucleosomes inferred in the track above. Grey circles refer to nucleosomes inferred with the information from other tracks. Green lines under a pair of nucleosomes indicate di-nucleosomes that could be inferred from the peaks in Lib3. The green lines were added to other tracks in grey. The 'x' nucleosome is labeled in the 1st track. Part A and B of the *CLOCK* promoter are specified, displaying different inferred nucleosome compaction.

The previous analysis uses libraries containing a mix of fragment sizes, and yet different size ranges may inform about different compaction levels. To gain more resolution an analysis similar to the above mentioned on parts A and B was done, although with some modifications (See the Methods section in this chapter). Briefly, Lib1-Lib3 libraries within single experiments were pooled and an in-silico size selection was done to keep fragments potentially originating from partial nucleosomes, mono-nucleosomes and di-nucleosomes. Additionally, different displays of fragments were prepared: the 5 leftmost and rightmost nucleotides, and the 5 central nucleotides of fragments were displayed in different tracks, to facilitate the identification of the inner part of the fragments (See Figure II-13 for an example in E1). As a result of this analysis in experiments E1, E2-R1 and E2-R2, similar nucleosome positions are observed on the region downstream of the 'x' peak across replicate experiments (Figure II-14), although, different nucleosome positions are inferred upstream of the 'x' peak, for unknown reasons. Remarkably, at least in Part A of the *CLOCK* promoter, the in-silico selection of sizes and the use of different displays of fragments proves to be useful to the study of protein associations with the DNA.

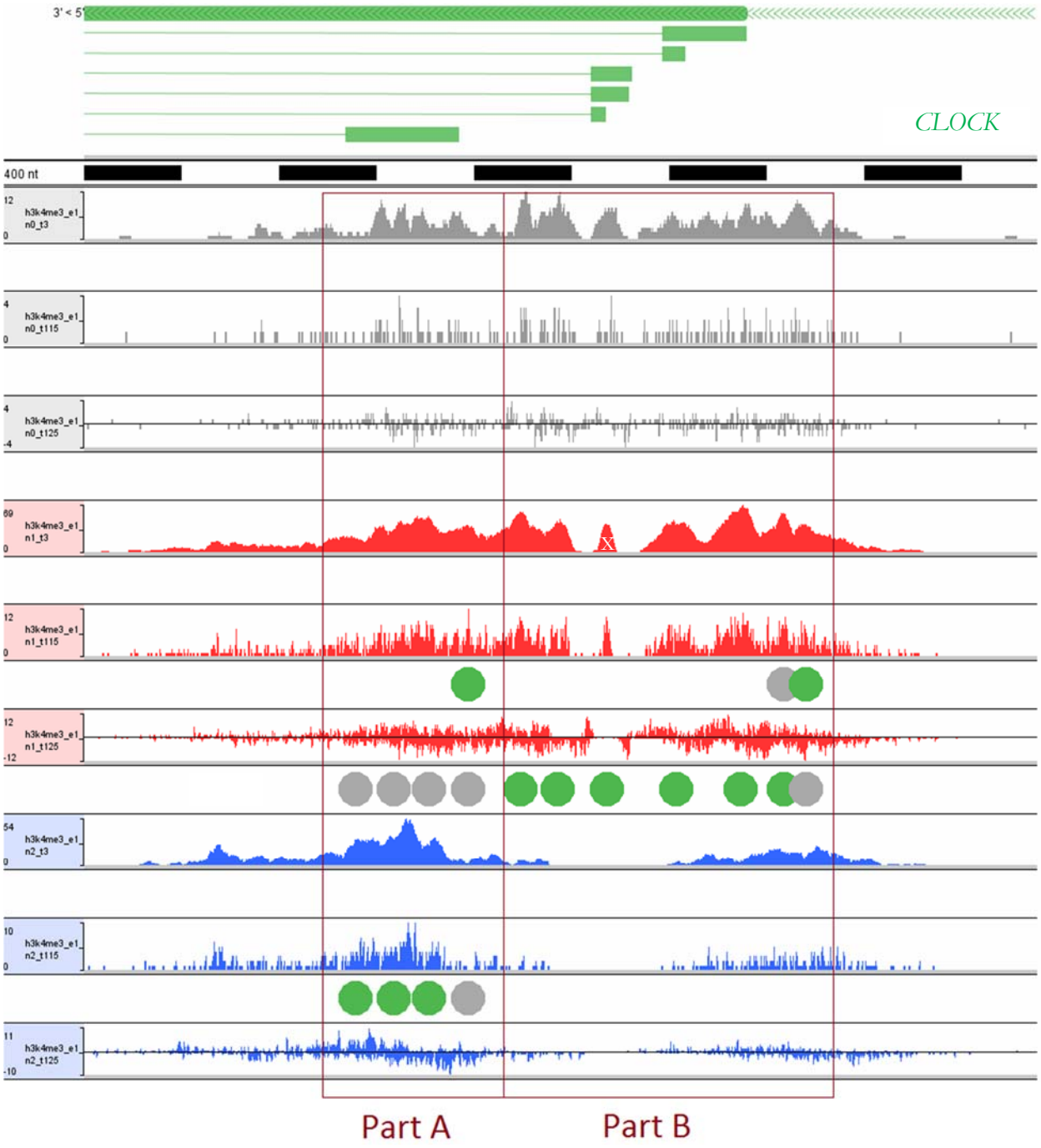



**Figure II-13**. Analysis of the nucleosome organization in the *CLOCK* promoter after an in-silico size selection in the E1 pooled libraries. The top of the figure shows the annotated genic region with a long box on the top, and the annotated exon-intron structures are depicted below with rectangles and lines, respectively. The *CLOCK* gene promoter is antisense. A 4kb region in the *CLOCK* promoter is shown. Libraries Lib1, Lib2 and Lib3 were pooled in-silico and sequential in-silico size selections were done: n0) fragments with sizes [100, 124] bp potentially

representing partial nucleosomes (grey), n1) fragments with sizes [150, 174] bp originating from mono-nucleosomes (red) and n2) fragments with sizes [290, 339] bp originating from di-nucleosomes (blue). Three data tracks are displayed for each case of size selection. The first track displays the central 50 bp of fragments, the second track displays the central 5 bp of fragments. In contrast, the third track displays the 5 nucleotides in the 5' end and 3' end of fragments that are displayed in the upper and lower part of the track, respectively, separated by a black line. A white space is left between tracks where circles are drawn depicting nucleosomes of 145 bp. Green circles refer to nucleosomes inferred in the track above, although they could be further confirmed by information in other tracks. Grey circles refer to nucleosomes inferred with the information from other tracks. The 'x' nucleosome is labeled in white in the 4th track. Part A and Part B in the promoter are also labeled in the bottom of the figure.

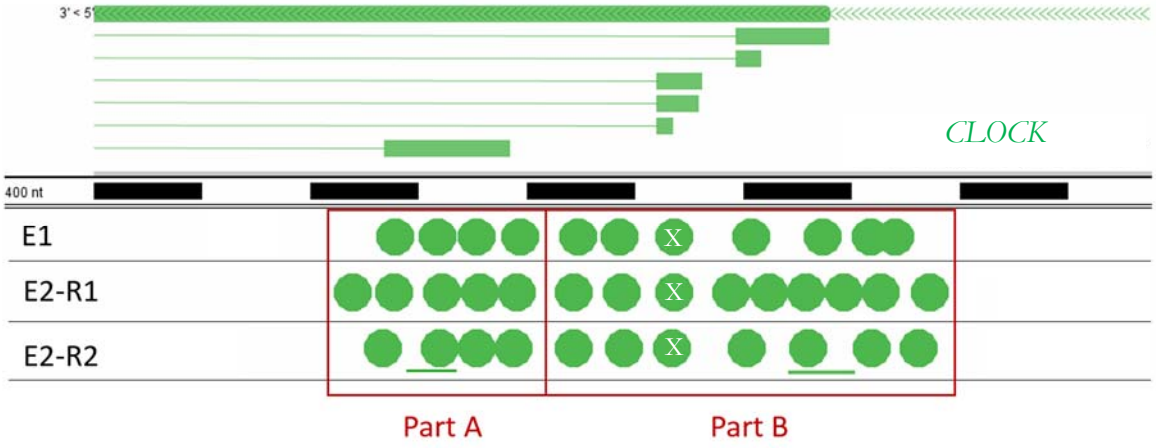



**Figure II-14**. Representation of the inferred nucleosome organization across experiments and replicates. The top of the figure shows the annotated genic region with a long box on the top, and the annotated exon-intron structures are depicted below with rectangles and lines, respectively. The *CLOCK* gene promoter is antisense. A 4kb region in the *CLOCK* promoter is shown. Green circles depict nucleosomes of precisely 145 bp inferred in the CLOCK promoter. Green lines indicate that the associated nucleosome could be assigned to the area delimited by the line. The 'x' nucleosome, Part A and B are labeled. The two areas have different chromatin compactions across samples. Part A has more compacted nucleosomes, whereas part B tends to be less condensed, although variability is observed among samples for unknown reasons.

# Discussion

## Is paired-end data more useful than single-end data for ChIP-seq studies?

Table II-3 summarizes the advantages of each technology. Single-end data is cheaper (currently, approximately 1,400 CHF is the cost for single-end sequencing of 100-bp long reads versus a price of 2,400 CHF for paired-end sequencing of 100-bp long reads). Further, single-end sequencing is slightly faster to sequence, as the paired-end mode includes the single-end steps plus an extension to sequence the opposite end in the complementary strand. Further, single-end data is useful as it can approximate the position of ChIPed fragments (Figure II-8). Nevertheless, paired-end data is precise. Given that different peaks can have fragments with different underlying sizes, it is not possible to apply a universal shift to all fragments without having discriminatory effects on some peaks over others. Thus paired-end data is more useful for the identification of centers of fragments and therefore the genomic localization of epitopes. Interestingly, as paired-end data allows the precise in-silico size selection, that is useful for the identification of binding sites. Furthermore, as no gel-size selection is needed during library preparation, no material is lost. Therefore, when using paired-end sequencing, less starting material is required for the ChIP-seq experiment. This is an important issue. For example, in the context of the CycliX project, where resected mice livers were used, the loss of fewer material with paired-end data is very useful. Moreover, paired-end data can offer further advantages when histones are the epitope studied. As shown in this work, in-silico size selections of fragments can help to reveal chromatin structures (Figure II-13). Further, paired-end data allows to precisely identify the degree of redundancy. This allows to observe the nature and effects of redundancy as shown in Figures II-4-6. Additionally, the use of paired-end data has also allowed the precise identification of three times more single-copy fragments compared to simulated single-end data (Table II-1). With single-end data, potentially the non-correctly identified single-copy fragments would be discarded as redundant fragments.

Although not in depth, collectively, the analysis of paired-end sequencing presented here was sufficiently convincing to argue for its use in the CycliX PH experiments described in Chapters IV to VI.

| **Single-end data** | **Paired-end data** |
|---|---|
| Cheaper | Precise knowledge of the ChIPed fragments |
| Faster | Precise identification of redundancy |
| | It avoids size selection, thus experiments can start with less material |
| | Precise in-silico size selection |
| | Chromatin structure can be revealed by studying in-silico size selections of ChIP-seq data from histones |

**Table II-3**. Advantages of using single-end and paired-end sequencing data.

## Paired-end ChIP-seq data reveals chromatin structure in promoters

I have shown that the analysis of different fragment size ranges in H3K4me3 ChIP-seq data sets can be used to reveal potential chromatin structure in promoter regions. In the promoter of an active gene (i.e., *CLOCK*), some positions of the DNA are more frequently sheared, most likely because they are not protected by nucleosomes (Figure II-12). Nevertheless, one should take such an analysis with care because when I performed identical analyses on technical replicates which should have equal chromatin structure I got different results. It would be interesting to compare these observations with micrococcal nuclease (MNase) experiments that specifically target nucleosomes by digesting internucleosomal regions to further confirm the results obtained with ChIP-seq data with sonicated chromatin.

Two different levels of nucleosome compaction within the same promoter are revealed by the analysis of fragment sizes in the *CLOCK* promoter (Figure II-12): part A shows a more compacted chromatin compared to part B. This observation can give interesting insights about the regulation of transcription of the core circadian *CLOCK* gene in HeLa cells. Interestingly, Pol2 ChIP-seq data in this promoter shows an increased Pol2 occupancy in Part B where nucleosomes are less condensed (data not shown). Further, no visible Pol2 is observed through the gene body, which could suggest that in HeLa cells Pol2 is paused in the promoter, supported by a compacted chromatin downstream the Pol2 site. Nevertheless, these results should be taken carefully as they have been obtained from a heterogeneous population of HeLa cells, where cells may be at different stages of the cell-division cycle and can also have different copies of chromosomes plus translocations, complicating the analysis further. Nevertheless, it could be interesting to perform studies of this kind on other kinds of cells with more stable genomes and at similar cellular states.

# Methods

Experimental procedures done by Romain Groux.

## Cell culture

Human HeLa cells were grown at 37°C in Joklik's Modified Eagles's Medium (JMEM) with 5% fetal bovine serum (FBS).

## Chromatin immunoprecipitation (ChIP)

Different rounds of chromatin immunoprecipitation were done. For E1, $2.7x10^7$ cells were used, and for E2, $4.6x10^8$ cells were used. Cells were cross-linked for 8 min using 1% formaldehyde. DNA was isolated and sonicated to approximately 100–300 bp using a Bioruptor (Bioruptor UCD-200 from Diagenode) and 10 cycles of 30-sec on and off pulses at maximum power. Sonicated DNA was immunoprecipitated, washed, and eluted as described (Le Martelot et al. 2013). Immunoprecipitations were done with different antibodies. In E1, the anti-H3K4me3 pAB-003-050 from Diagenode was used. For E2, the anti-H3K4me3 ab1012-100 from Abcam was used.

## Library preparation and Ultra-high-throughput sequencing

In E2, three aliquots were taken as technical replicates for ChIP and were prepared differently for sequencing. Samples in E1 and E2 containing 12 to 20 ng of ChIP-DNA were used to prepare paired-end sequencing libraries. The NEXTflex Illumina ChIPseq Library preparation kit from Bioo Scientific was used to prepare paired-end sequencing libraries. Additionally, libraries from Total input DNA were prepared. Fragment size selections were done in each experiment. In E1, E2-R1 and E2-R2, three fragment size ranges were selected after agarose gel. The selection targeted the size ranges: 100-200 bp, 200-300 bp, and 300-400 bp. For the E2-Beads experiment, DNA fragments were not subjected to agarose electrophoresis size selection but rather purified with magnetic beads as described in the library preparation kit; this purification gives a broad size selection of about 100-400 bp. Libraries containing 6 to 20 ng/ul were used for sequence analysis. 50-bp paired-end reads were sequenced with the Illumina platform Genome Analyzer IIx. For an outline, see Figure II-1.

## Processing of the data

Fifty-bp sequence paired-end reads were mapped onto the human genome (NCBI37/hg19) using Elandv2e from Illumina. Fragments with good quality in both reads as specified by default in the Illumina software (i.e., the quality of the first 25 bp is higher than Q20) were selected. Furthermore, fragments mapping onto unique positions of the genome and with sizes between 50 to 1,000 bp were used for analysis. For quantification and visualization purposes, the data were converted into AV format files [Martin et al., unpublished], either AV1 or AV2. An AV1 file is used for quantification and contains tab separated information of genomic regions (chromosome, start position, length of the region, strand and a number that can indicate a quantification). Whereas, an AV2 file is a compressed and indexed AV1 file used for visualization with the CycliX viewer [Martin et al., unpublished]. For the analysis of the redundancy of fragments all the copies of repeated fragments were used. In contrast, for the analysis of fragment sizes, only one copy of repeated fragments was used. Further, different displays of the sequenced fragments were prepared on AV2 files. For quantification purposes the central 50 bp of each fragment were calculated.

For the final analysis of nucleosome positions in the *CLOCK* promoter the three libraries from each E1, E2-R1 and E2-R2 samples were pooled in-silico and three in-silico size selections were done: [100-124] bp, potentially originated in partial nucleosomes, [150-174], mono-nucleosomes, and [290-339] bp, di-nucleosomes. Additionally, the extreme 5 nucleotides and the central 5 nucleotides of fragments were displayed in different tracks.

All the handling of the data was done with the UNIX shell, Perl and the R software [R Core Team, 2013].

## Transcription unit annotation and quantification

The human gene list used for the analyses was obtained from the Encode project [ENCODE Project Consortium, 2012]. Version v16 of the annotation was used. For quantification of promoter regions, the most 5' annotated gene transcription start site in genes was used, and a region of +/- 1 kb around the selected TSS was quantified. The AV1regionsum tool associated to the AV format was used to quantify genomic regions. Quantifications were done on ChIP and Input datasets and scaled to the number of fragments analyzed per library. log2(ChIP/Input) ratios were calculated on each promoter quantification to correct for potential biases also present in the Input.

## Genomic displays of ChIP-seq data

The CycliX viewer [Martin et al., unpublished] was used to visualize ChIP-seq tracks, especially AV2 files that can be handled very efficiently on the viewer. Among other functions, the genomic viewer allows the selection of the desired genome stored in the NCBI database and the visualization of genomic data in a cumulative view. The density of accumulated data can also be scaled to the total number of fragments per library and tracks can be displayed with any desired scale.

# Chapter III: HCF-1 chromatin binding in HeLa cells along the cell division cycle

The co-transcriptional regulator HCF-1 is required for proper progression of the mammalian cell cycle [Goto et al., 1997; Julien and Herr, 2003] at least in part via interactions with the E2F family of transcription factors [Tyagi et al., 2007]. Nevertheless, little is known about the genome-wide activities of HCF-1 during cell proliferation. [Michaud et al., 2013] showed interesting insights of the genome-wide chromatin association of HCF-1 in a heterogeneous population of dividing HeLa cells. In these cells HCF-1 is a common component of active CpG-island promoters, and it coincides with the promoter occupancy of transcription factors such as ZNF143, THAP11, YY1 and GABP. In this chapter, I describe further insights about the chromatin association of HCF-1 in HeLa cells, by analyzing data (J. Michaud, personal communication) from a population of synchronized HeLa cells via double thymidine block and release.

## Results

For this research, dividing HeLa cells were cell-cycle synchronized via double-thymidine block and release. FACS shows that the cells were highly synchronized (Figure III-1), as the cells progressively transit from G1 (0 hrs) to S (4 hrs), G2/M (8+10 hrs) and G1 (12+16 hrs). Thus, these samples were used for ChIP-seq. ChIPs were done targeting HCF-1 and the transcriptional marks Pol2, H3K4me3 and H3K36me3. As paired-end sequencing was not available at the time that these experiments were performed, samples were subjected to single-end sequencing. Figure III-2 indicates the number of reads analyzed for each sample. Except for two Input samples, all samples had at least 5 million sequenced fragments.

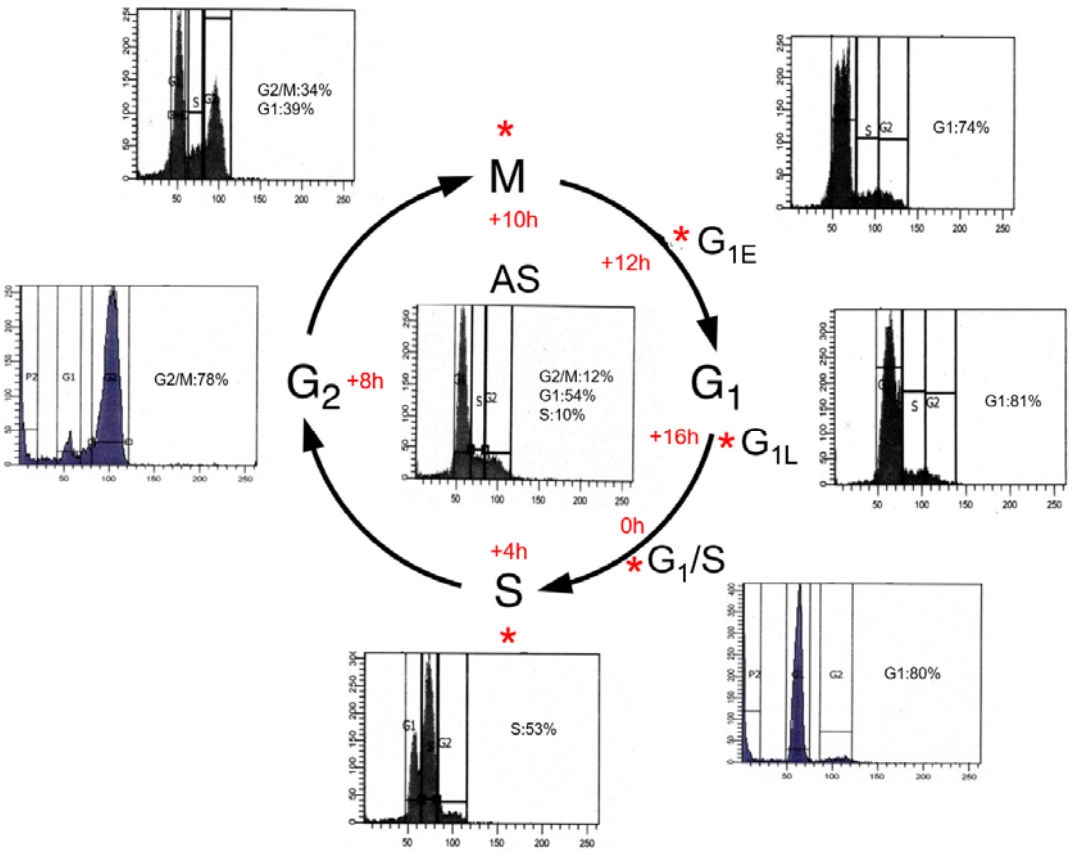



**Figure III-1**. Description of the time points after release of HeLa cells from double thymidine block. Source: Joëlle Michaud. The time after release is indicated in red. The associated panels represent the FACS results per time point, with an asynchronous (AS) population shown in the center. The FACS graphics show the amounts of DNA per cell on the x axis, and number of events in the y axis. All the FACS results show a good synchrony of HeLa cells at each time point.

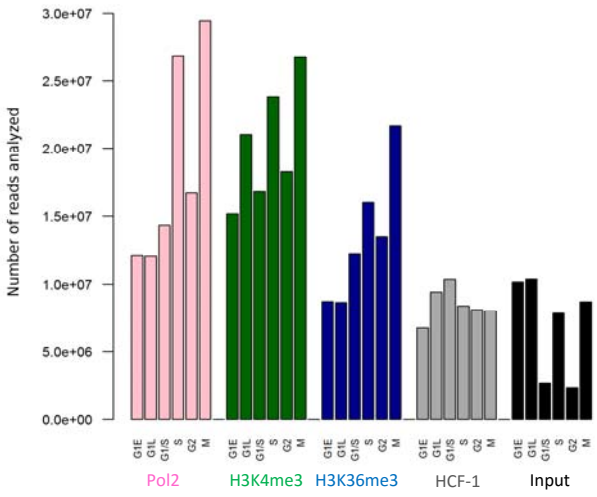



**Figure III-2**. Number of reads analyzed in each ChIP-seq library prepared from synchronized HeLa cells. Data for Pol2 (pink), H3K4m3 (green), H3K36me3 (blue), HCF-1 (grey) and Input (black) libraries are shown, for each of the collected stages of the cell cycle: G1E, G1L, G1/S, S, G2 and M phases.

## III.1 HCF-1 binds to the chromatin of HeLa cells differently along the cell cycle

To investigate the genome-wide binding events of HCF-1 along the cell cycle I did a genome-wide analysis of sequence reads with a sliding window of 100 bp (referred to as bins) in 50 bp steps. The analysis detected 100bp-bins where there was a higher density of HCF-1 ChIPed reads compared to the Input sample. If such bins were closer than 200 bp to each other, they were aggregated. The resulting aggregated bin sets thus contain potential HCF-1 binding regions. As shown in Figure III-3, there is a progression in the number of binding events during the cell cycle. A lower number of regions is observed at G1E, with 619 bound regions of which two-thirds (402) are regions that are HCF-1 bound in all time points and one-third (217), although not bound at all times, are also bound in at least one or the other adjacent time points. Only three regions were specific to the G1 time point. While cells go through the cell cycle the number of HCF-1 bound regions increases. At G1/S phase, 3,400 regions were bound and the progression culminates between S and G2 phases when the number of regions has increased up to over 5,000. Finally at M phase, the number decreases, although still showing 2,437 regions bound. This completes a cell cycle of HCF-1 activity.

Interestingly, approximately 80% of the regions identified fall within promoters. Even though further analyses could inform about the precise gene activities of HCF-1, these provides a hint about the link between HCF-1

with transcription regulation in promoters, as observed in [Michaud et al., 2013] in non-synchronized HeLa cells.

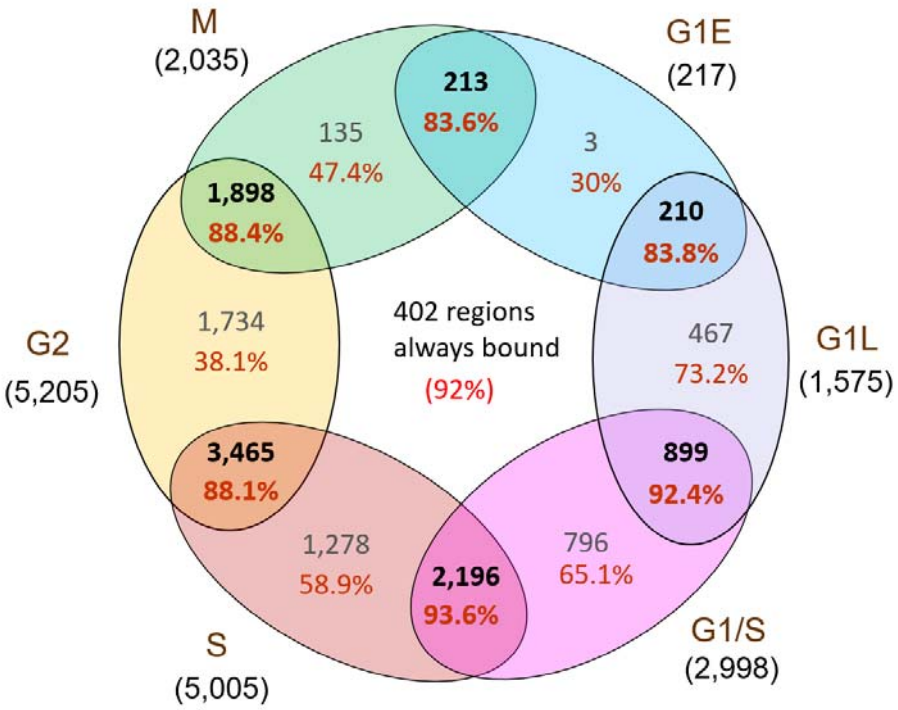



**Figure III-3**. Distribution of HCF-1 binding regions across the synchronized time points of the HeLa cell cycle. Dark blue numbers in the exterior of the chart represent the number of identified regions at each time point. Grey numbers inside of the ovals indicate the number of regions within each specific time point that are not shared with the adjacent time points. Black numbers shows the regions shared between adjacent time points. Red numbers indicate the percentage of the indicated regions that fall within a promoter region (+/- 1 kb around TSS). Center, regions that were bound by HCF-1 in all time points is shown. These regions were not included in the analysis described above.

## III.2 HCF-1 has diverse ways of binding to the chromatin

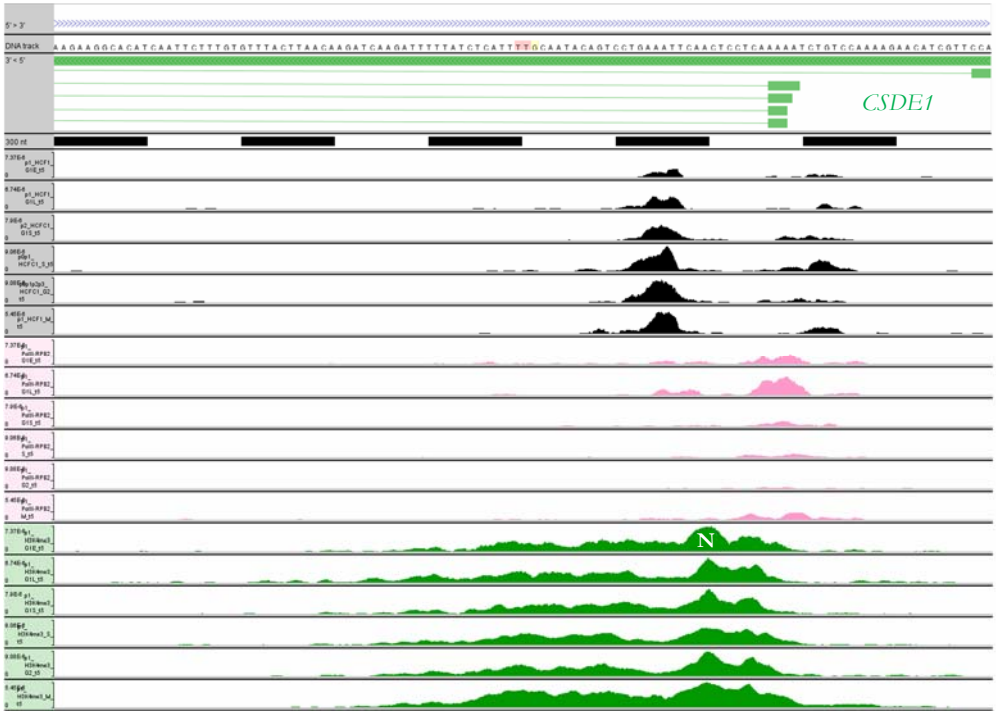
In total there were 7,809 regions were HCF-1 binds at at least one of the analyzed time points. Ninety-two% of the regions have lengths between 100 and 300 bp with longer regions reaching lengths up to 2 kb. As shown in Figure III-4a-b, HCF-1 displays a binding region longer than 300 bp in the promoter of the *CSDE1* gene and the bidirectional promoter between the *FBX19* and the *FBX19-AS* antisense gene. In the *CSDE1* promoter a long binding region of approximately 800 bp is observed (Figure III-4a), which, interestingly, is composed of

two HCF-1 binding regions flanking a Pol2 occupied region. Further, the downstream HCF-1 binding region localizes with what appears to be a nucleosome free region, indicated by the H3K4me3 data. This gives new insights about the mode of action of HCF-1 in promoters. So far it was observed that HCF-1 generally binds approximately 40 bp upstream of annotated TSSs, just upstream of Pol2 peaks. The upstream peak in the *CSDE1* promoter does likewise. But in contrast, the downstream peak is observed after the Pol2 peak. This promoter has a special chromatin structure. It could be that the linker DNAs upstream and downstream the nucleosome labeled 'N' in the figure interact with each other by means of a complex formed by HCF-1. Alternatively, it could also be that the downstream peak is associated to a non-annotated transcription start site.

The bidirectional promoter between *FBXL19* and *FBXL19-AS* also displays a long region of HCF-1 binding (Figure III-4b), this time of approximately 1 kb, and again with two main binding sites, but now more extended than in the *CSDE1* promoter and yet associated with apparently long depleted nucleosome regions. The presence of a protein complex formed by HCF-1 seems to be creating an open chromatin structure or HCF-1 binding is possible because of an open structure.

The *FBXL19* and *FBXL19-AS* bidirectional promoter is not the only case of bidirectional promoter where HCF-1 binds. In other cases of long HCF-1 binding regions, a double binding site is often present in annotated bidirectional promoters in proliferating HeLa cells. Interestingly, HCF-1 binds to 1,262 bidirectional promoters (with two divergent annotated TSSs separated by a maximum distance of 1 kb; see Methods in this chapter for further details), which is 25% of all the bidirectional promoters in the human genome (there is a total of 4,918 cases of bidirectional promoters). This finding shows that, HCF-1 is not only a common component of unidirectional promoters in HeLa cells [Michaud et al., 2013], but also displays a wide association with bidirectional promoters.

a)




b)

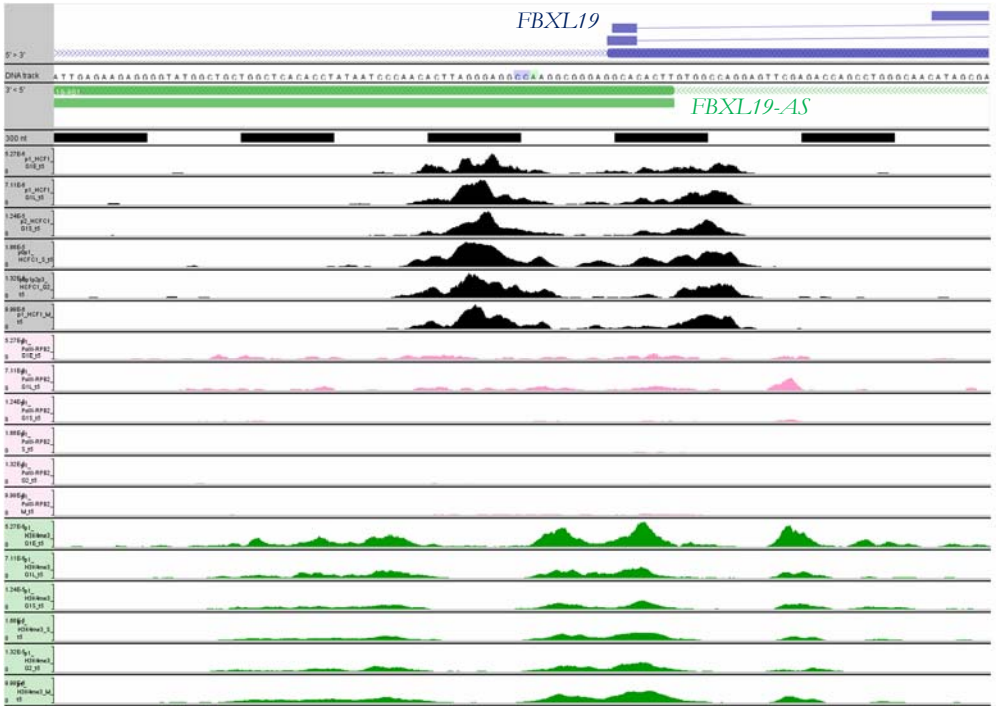



**Figure III-4**. Gene promoter cases displaying different lengths of HCF-1 binding sites of more than 300 bp. On the top of each panel the structure of sense (blue) and anti-sense (green)

genes and associated transcripts is described. Below, black and white boxes indicate the scale in 300bp segments. Data tracks are displayed from the ChIP-seq libraries for HCF-1 (black), Pol2 (pink) and H3K4me3 (green). For each ChIP-seq time point there is one track displayed (from top to bottom: G1E, G1L, G1/S, S, G2 and M). The density of shifted 35 bp of sequenced reads is displayed. a) *CSDE1* gene promoter. This gene is known to be involved in male gonad development, nuclear-transcribed mRNA catabolic process and regulation of transcription. In this promoter the nucleosome 'N' observed in the H3K4me3 data between the HCF-1 binding sites is specified. b) Bidirectional promoter between the genes *FBXL19-AS1* and *FBXL19*, from left to right. They are involved in the regulation of the cell cycle by means of regulation of protein degradation.

## III.3 HCF-1 association with promoters involved in cell-cycle regulation: MCM3 and CDC6 examples

As HCF-1 is involved in cell-cycle progression, particularly the G1 to S phase transition, I next examined HCF-1 association with two promoters involved in the G1 phase progression: the DNA helicase subunit MCM3 and the CDC6 regulatory protein, both members of the pre-replication complex at origins of DNA replication. For the MCM3 promoter (Figure III-5), the HCF-1 binding region is of 300 bp. In this case it is interesting to see that the HCF-1 binding density reaches a maximum at S phase, where Pol2 decreases. Thus, in these promoter there are different HCF-1 profiles across cell cycle phases. At S phase there is a peak towards the right of the region. Whereas at G2, there are two peaks separated by approximately 150 bp approximately. It would be interesting to further investigate these dynamic patterns as HCF-1 seems to change its binding mode at different times of the cell cycle. Perhaps interactions with different DNA-binding proteins occur at different phases.

The CDC6 promoter displays a binding region of approximately 150 bp (Figure III-6). This promoter shows one similarity with the *MCM3* case in that it displays a double peak at S phase that is anti-correlated with Pol2 occupancy. Further, the peak at G1/S phase is to the left, whereas the peak at G2 phase is to the right. This could show that the double peak observed at S phase, is indeed showing different binding sites of HCF-1 at different times. Interestingly, HCF-1 and Pol2 are anticorrelated, suggesting an association between HCF-1 and Pol2 activity. To investigate further this double peak I used additional ChIP-seq data targeting HCF-1 in an asynchronous population of HeLa cells prepared in the context of the CycliX project (for further details see Chapters IV to VI). In the CycliX project, ChIP-seq libraries were prepared with mouse liver chromatin as main material but also a 5% of HeLa chromatin was added as an internal 'spike' control. The advantage of this data is that it is paired-end sequencing data which, as shown in Chapter II, offers the possibility to do a more precise

analysis of the position of the epitopes by adjusting the sizes and displays of the genomic fragments. I could compare the sequencing data with the human genome and thus extract those sequences that were specific to the human genome with a very low loss of sequences (1%) being shared between the human and the mouse mappings. I pooled the sequencing data from the different spike controls from ChIPs against HCF-1. Then I selected fragments with sizes between 50 and 150 bp and created different displays of the mapped reads: 50 central bp, 5 central bp and 5 bp on the extremes of the fragments. The double peak observed in the synchronized HeLa cells was still obvious in the non-synchronized HeLa cells when using a resolution of 5 bp in the center of the short fragments. Interestingly, those two peaks lie over the position of binding motifs of transcription factors known to interact with HCF-1 for transcription regulation, the complex between ZNF143 and THAP11 (associated to the SBS1 and SBS2 motifs), and the transcription factor E2F1. Combined with the data from synchronized HeLa cells (Figure III-6c), these data indicate that HCF-1 can bind to the same promoter via interactions with different proteins at different times.

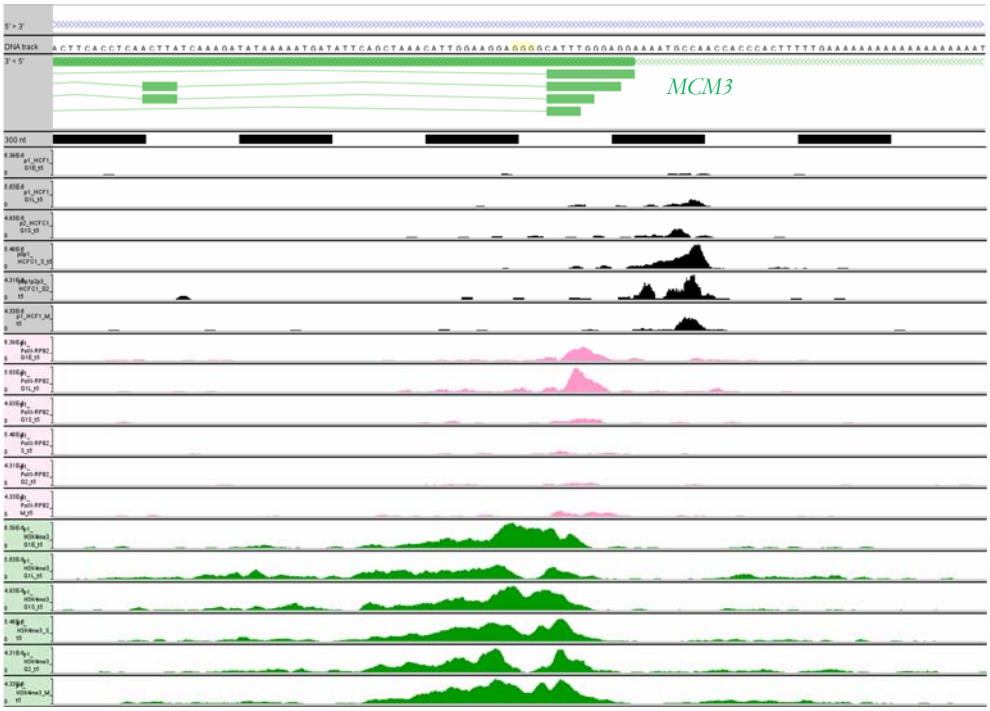



**Figure III-5**. Gene promoter of the *MCM3* gene displaying a binding region of HCF-1 of 300 bp. On the top, the structure of anti-sense (green) genes and its associated transcripts are shown. Below, black and white boxes define scale in 300 bp segments. Data tracks are displayed from the ChIP-seq libraries for HCF-1 (black), Pol2 (pink) and H3K4me3 (green). For each set of ChIPs there is one track displayed for each time point (from top to bottom: G1E, G1L, G1/S, S, G2 and M). The density of centered 35 bp sequence reads is displayed.


**Figure III-6**. *CDC6* gene promoter displaying a 150 bp binding region for HCF-1. A) Double binding site of HCF-1 observed with paired-end sequence data. On the top of the panel the *CDC6* promoter structure and associated transcripts are described. Below, black and white boxes define the scale in 300 bp segments. Data tracks are displayed for the analyzed fragment sizes of 50 to 150 bp displaying the 50 central bp (t3) and the 5 bp at the centers (t11-5) to gain resolution. B) Nucleotide sequence in the positive strand underlying the two observed peaks with E2F and SBS1 (ZNF143) and SBS2 (ZNF143/THAP11) binding motifs indicated. C) HCF-1(green) and Pol2 (pink) profiles across the cell cycle on the *CDC6* promoter. The density of centered 35 bp of sequenced reads (t5) is displayed. The profiles show an anti-correlated density of Pol2 and HCF-1 suggesting a repressive function of HCF-1.

# Discussion

## HCF-1 has a preference to bind to promoters in synchronized HeLa cells

In [Michaud et al., 2013] it was shown that HCF-1 is a common component of gene promoters in 67% of the annotated promoters in non-synchronized HeLa cells. Consistent with this, in this work I have shown that also in synchronized HeLa cells HCF-1 tends to bind to promoters, as the 80% of the regions bound by HCF-1 lie within promoters. Further, I have shown that HCF-1 sits on 25% of the bidirectional promoters that can be calculated from the Encode v12 annotation of human genes. Interestingly, HCF-1 co-binds with ZNF143 in 33% of promoters in asynchronous HeLa cells [Michaud et al., 2013]. ZNF143 is a transcription factor that binds to HCF-1 [Michaud et al., 2013] and is known to control the expression of divergent protein-protein and protein-non-coding RNA gene pairs [Anno et al, 2011; Ngondo-Mbongo et al., 2013]. A further step on the role of HCF-1 in bidirectional promoters would be to study the overlap of ZNF143 and HCF-1 in bidirectional promoters.

## HCF-1 could repress transcription in promoters

In a heterogeneous population of HeLa cells, the presence of HCF-1 in promoters correlated with both Pol2 activity and mRNA transcript levels [Michaud et al., 2013]. In this work, cells have been synchronized at specific stages of the cells cycle, allowing for a more precise comparison of Pol2 and HCF-1 activities at the same times. Interestingly, specific gene cases such as *CDC6* and *MCM3* display an anti-correlated activity of Pol2 and HCF-1 that could only be seen after synchronizing cells at specific stages of the cell cycle. These observations, which where only possible because the cells were synchronized, suggest that HCF-1 could perform repressing functions in promoters. A genome-wide comparison of Pol2 occupancy and HCF-1 binding density may inform more widely about the roles of HCF-1 in promoters.

## HCF-1 might have different modes of binding to the chromatin in different promoters

Previously [Michaud et al., 2013], HCF-1 was observed principally as single peaks near TSS (40 bp upstream the TSS). Here, I examined some TSS where HCF-1 displays more complex binding patterns. This suggests that HCF-1 has different ways to associate to the chromatin across promoters. Indeed, HCF-1 is known to interact with a wide range of promoter-specific transcription regulators (see Introduction section I.4.2). The

way HCF-1 binds to the chromatin in an array of interactions may differ. Further studies on the underlying sequences of promoters and binding sites could provide insights about the mechanisms of interaction of HCF-1 in promoters.

## HCF-1 could bind to E2F1 and THAP11/ZNF143 in the same promoter

In cycling HeLa cells, HCF-1 is recruited to promoters involved in promoting cell proliferation. This recruitment is done partly by the E2F proteins E2F1 and E2F4 [Tyagi et al., 2007]. Nevertheless, the analysis of HCF-1 binding sites in HeLa cells did not display an enrichment of E2F motifs [Michaud et al., 2013]. This could be due to the nature of the analysis that searches for motifs present in many promoters, which perhaps is not the case for E2F. Moreover, the E2F1 binding motif is not very well defined which hinders the identification of potential binding sites. The CDC6 promoter shown in Figure III-7 provides a new vision about how HCF-1 could be tethered to promoters by both E2F1 and THAP11/ZNF143. HCF-1 may bind this promoter in two different positions in association with E2F1, on the one hand, and THAP11/ZNF143, on the other hand. A question to ask now is how this double binding is done, whether at the same time or in different times. Interestingly, Figure III-7c shows that at the synchronized time points G1/S and G2 HCF-1 ChIPed fragments tend to align with one or the other of the two peaks. It is a possibility that HCF-1 is binding in different peaks at different times, to regulate transcription in response to different signals depending on cell cycle states.

# Methods

Experimental procedures were performed by Joëlle Michaud.

## Cell culture and synchronization

Human HeLa-S cells were grown at 37°C in Juklik's Modified Eagles's Medium (JMEM) with 5% fetal calf serum. HeLa-S cells were synchronized using double thymidine block as previously described (Tyagi et al. 2007). Briefly, exponentially growing cells in suspension ($3x10^5$ to $4\ x10^5$ cells/ml) were grown in the presence of excess thymidine (2 mM) for 12 hours. Cells were centrifuged at low speed and the medium containing thymidine removed. Cells were subsequently resuspended in fresh media lacking supplemental thymidine and allowed to grow for 11 more hours. A second 2 mM thymidine treatment was done for 12 hrs. The G1/S population of cells was subsequently harvested or transferred to normal media for harvesting at subsequent cell cycle stages: 4 hrs (S phase), 8 hrs (G2 phase), 10 hrs (M phase), 12 hrs (early G1 phase) and 16 hrs (late G1 phase). Synchronization was checked by FACS sorting (Figure III-1).

## Chromatin immunoprecipitation (ChIP)

Between $2 \times 10^7$ and $8 \times 10^7$ HeLa-S cells were used per ChIP. Synchronized HeLa-S cells were cross-linked for 8 min using 1% formaldehyde. DNA was isolated and sonicated to 100–300 bp using a Bioruptor machine (Bioruptor UCD-200; Diagenode) and 5 rounds of 30-sec pulses on and off at maximum power. Sonicated DNA was immunoprecipitated, washed, and eluted as described in [Tyagi et al., 2007]. The following antibodies were used: polyclonal anti-HCF-$1_C$ (H12) targeting the C-terminal subunit of HCF-1 [Wilson et al., 1993a], polyclonal anti-Pol2 (POLR2B, sc-67318), polyclonal anti-H3K4Me3 (Abcam ab-8580), polyclonal anti-H3K36Me3 (Abcam ab-9050). Input was also taken from the samples. 18 cycles of PCR were done and gel size selections between 100-300 bp were done.

## Library preparation and high-throughput sequencing

Single-end sequencing libraries were prepared from 5 to 10 ng of ChIP-DNA. Libraries were prepared with the ChIP-seq DNA Sample Prep Kit (Illumina). Total input libraries were also prepared. Thirty-five bp at the end of the genomic fragments were sequenced on an Illumina Genome Analyzer 2. Single end sequencing mode was used.

## Processing of the data

Thirty-five-bp single-end sequences were mapped onto the human genome (NCBI37/hg19) using Bowtie 0.12.7. All reads with good quality, as indicated by Illumina quality scores, were kept for analysis (from the first 12 bp of reads, the intensity at the highest nucleotide divided by the sum of the intensities at the two highest base pairs should be higher or equal to 0.6). Those reads mapping at a unique genomic location and with no mismatches were kept. For the HCF-1 datasets, ChIP-seq libraries from different chromatins were pooled due to the low number of reads that would not allow the proper identification of HCF-1 binding sites. On average, 34% of the reads were repetitions. To diminish the potential effect of PCR amplification artifacts, only up to two extra copies of all repeated reads were used for analysis. For quantification and visualization, 35 bp from each read were shifted towards the center of fragments and accumulated in the positive strand to gain more resolution when inferring the actual binding sites. The shifting was done based on the most frequent fragment length that was resolved by using Bioanalyzer and further confirmed by a cross-correlation analysis performed in R [R Core Team, 2013]. The information of the shifted reads was prepared in AV format, in either AV1 or AV2 files. An AV1 file used for quantification contains tab separated information of genomic regions (chromosome, start position, length of the region, strand and a number that can indicate a quantification). Instead, an AV2 file is a compressed and indexed AV1 file used for visualization with the CycliX viewer [Martin et al., unpublished].

All the handling of the data was done with the UNIX shell, Perl and the R software [R Core Team, 2013].

## Genome annotation and quantification

The human gene list used for the analyses was obtained from the Encode project [ENCODE Project Consortium, 2012]. Version v16 of the annotation was used. Quantifications on genomic bins of 100 bp were done. Genomic bins of 100 bp were defined in chromosomes starting at position 1 and sliding 50 bp at a time. The remaining regions smaller than 100 bp at the end of each chromosome were not used for analysis. The AV1regionsum tool associated with the AV format was used to quantify genomic regions. Quantifications were done on ChIP and Input datasets and scaled to the number of fragments analyzed per library. The average input scores were added to the quantifications as pseudo-counts to stabilize the variance in low scores. log2(ChIP/Input) ratios were calculated on each bin to correct for potential biases also observed in the Input.

Two transcription start sites in opposite directions closer than 1 kb were considered bidirectional promoters. As for a given gene, multiple TSS can be annotated, distinct combinations of TSS from a pair of genes were counted as one bidirectional promoter.

## Identification of HCF-1 binding sites

To investigate the chromatin binding events by HCF-1 along the cell cycle, genomic bins of 200 bp were scanned. Bins within 'blacklisted' genomic regions by the Encode Consortium were discarded, as they typically show high quantifications also in the Inputs for unknown reasons (http://hgwdev.cse.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeMapability) [Fujita et al., 2011]. The bins with resulting log2(ChIP/Input) ratios higher than 1 were kept for further analysis. And those bins closer than 100 bp to any other selected region were merged into one region.

## Genomic visualization of ChIP-seq data

The CycliX viewer [Martin et al., unpublished] was used to visualize ChIP-seq tracks, especially AV2 files that can be handled very efficiently on the viewer. Among other functions, this genomic viewer allows the visualization of genomic data in a cumulative view. The density of accumulated data can also be scaled to the total number of fragments per library. And tracks can be displayed with any desired scale.

# Chapter IV: Genome-wide transcriptional responses to partial hepatectomy in the mouse liver

Often, to investigate cell proliferation, research is done in cell culture samples from immortalized cells because they are easier to handle in the laboratory. The scientific community has learnt a lot from this work. Nevertheless, there are important differences between the division cycle in normal cells and cultured cells, which need to adapt to the cell culture environment. In this regard, the mouse liver is a robust model to investigate cell proliferation in a healthy organism that is genetically very similar to humans. As already introduced (see Chapter I, Introduction), after resection of 2/3 of the liver, referred to as 2/3 partial hepatectomy or 2/3 PH, the organ rapidly initiates a compensatory regrowth via cell division to acquire its initial mass and functionality. Although the mouse liver is not completely homogeneous, approximately 80% of the genomes are from hepatocytes, making it one of the most homogeneous organs in the mammalian body. Also, upon resection, the dividing hepatic cells respond in a highly synchronized manner. This together with the recent advent of the high-throughput sequencing technologies provide an excellent opportunity to investigate the genome-wide transcriptional responses to partial hepatectomy, including cell proliferation.

The research described in this chapter was done in the context of the CycliX program where labs from Geneva and Lausanne joined efforts to study the interconnections of three cellular cycles in the mouse liver: the circadian cycle, the nutrient response cycle and the cell-division cycle. My role in the CycliX project has been to understand the genomic and transcriptomic responses to partial hepatectomy. This research was done in collaboration with Dominic Villeneuve, technician in the laboratory of Winship Herr. He performed 2/3 PH and collected liver samples throughout the regeneration of the liver.

## Results

To investigate the genome-wide transcriptional responses, two different kinds of sequencing libraries were prepared to have different perspectives of the process of gene transcription: ChIP-seq (genomic level) and RNA-seq (transcriptomic level) libraries. ChIP-seq libraries were prepared from ChIPs against Pol2, H3K4me3 and H3K36me3. And RNA-seq libraries were prepared from polyadenylated (PolyA) transcripts.

## IV.1 The changes in the transcriptome of post-PH livers show early responses similar to sham control mice but different responses during regrowth

To reveal the genome-wide changes in gene expression that occur after 2/3 PH, the tissue remaining after PH was collected as indicated in Figure IV-1. All PH were performed at two hours after the beginning of the 12 hrs light phase (ZT02), which also corresponds to 2 hrs of fasting. Some of the samples were collected at the priming phase after PH (1 hr and 4 hrs post-PH). Samples were also collected at the the time when the first round of cell proliferation initiates (10 hrs and 20 hrs) passing through the different major cell-proliferation cell cycle phases S, G2 and M (28 hrs, 36 hrs, 44 hrs and 48 hrs). Samples were also collected during the second round of cell proliferation (60 hrs) and later (72 hrs and 1 week) until full recovery (4 weeks). As a reference of the resting liver, samples were also collected directly at 0 hrs and after 4 weeks (collected at ZT02).

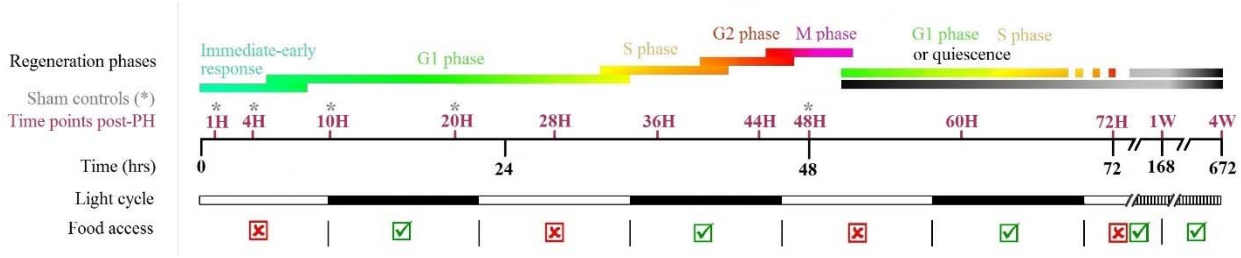



**Figure IV-1**. Description of the mouse liver regeneration stages, the time points at which liver samples were collected and the food and light conditions of the mice. Source: adapted from a figure created by Dominic Villeneuve.

After PH, livers may not only initiate regenerative responses because of the resection but may also be sensitive to other steps of the procedure. For example, the mice are always anesthetized with isoflurane. This relieves them from the pain of cutting the abdomen and the resection of the liver. The isoflurane is known to be detoxified by proteins that are specifically produced in the liver. And other proteins in the liver may be responding to the operation such as the proteases in the coagulation cascade or other signal transduction pathways sensing changes in the environment such as temperature, etc. To identify responses specific to regeneration, liver samples from Sham operated mice (Sham) were also collected. These mice were anesthetized and all steps of surgery except resection were performed. Their livers were collected at a collection of time points as the resected livers. To be certain to cover early events caused by a surgical procedure, the Sham time-point selection favored early time points. As shown in Figure IV-1, Sham samples were collected at 1 hr, 4 hrs, 10 hrs, 20 hrs and 48 hrs.

The gene expression landscape during mouse liver regeneration is not well described yet. To have a broad picture of the genes that may be transcriptionally active, I first detected those genes with transcripts in the RNA-seq results of any PolyA RNA sample. Transcripts from 12,032 annotated genes were detected (Table IV-1). Among the 25,959 gene transcripts that were not detected, some may not have passed the quantitation thresholds because of very low or no expression. Others might not be detected because they were not recovered during the sample preparation which included PolyA selection. Consistent with this hypothesis 93% of the detected transcripts are protein coding. Furthermore, among transcripts not detected 52% of them are non-protein coding.

| | % Protein coding |
|---|---|
| 12,032 gene transcripts detected | 93% |
| 25,959 gene transcripts not detected | 48% |
| | |
| 37,991 total annotated gene transcripts | 60% |

**Table IV-1**. Number of genes with detected or non-detected transcripts. The percentage of protein coding genes in each group is specified. This provides an indication of how many of them may produce PolyA transcripts that could be targeted with the PolyA selection in the RNA-seq experiments. Also, similar information is provided from the total annotated genes in the Ensembl67 mouse genome that was used for this analysis.

For each time point, RNA-seq transcriptome analyses were performed on separate mice replicates. To evaluate the similarity of replicates I calculated correlations among replicates which resulted in Pearson correlation coefficients of 0.99, reflecting almost perfect correlations. Nevertheless, I did observe slightly lower correlations between two sets of 0-hr replicates, called replicates 1-3 and replicates 4-6, where the Pearson correlation coefficient was 0.97. I do not know the reason for this increased variance, but may reflect slightly different environmental conditions for the mice used in the two sets of replicates.

For an overview of the differences in gene expression among conditions and replicates, I did a hierarchical clustering of the RNA-seq libraries (Figure IV-2). The result of the hierarchical clustering is represented in a tree structure, so-called dendogram, where leaves represent the analyzed libraries, and increased length of the branches represents the increased difference between sample nodes. The clustering algorithm forced samples with similar gene transcript patterns to group together. In this way, two big clusters of libraries were obtained. The biggest cluster groups Sham controls (labeled S), early post-PH (labeled X) time points and the control livers where no surgery was done (labeled C) at either 0 hrs or 4 weeks. The second cluster contains those

samples collected during post-PH cell proliferation (36-72 hrs). Notably, there are differences among libraries from resting livers. Replicates 1 to 3 (C0H1-3) are similar to the 4 week sample (X4W1-3) after PH where livers have completed regeneration, and to the control 4 week livers (C4W1-2) where no surgery was performed. Instead, replicates 4 to 6 at 0 hr (C0H4-6) are more alike to the Sham 1 hr samples. This difference between 0 hr replicates is consistent with the abovementioned poorer correlation results for these samples. This suggests that replicates 4 to 6 from resting livers may have been affected by some external stimuli, potentially modifying expression of genes. This shows how biological replicates can differ, but as the correlation between replicates was not strongly affected, I used all the samples for further analyses.

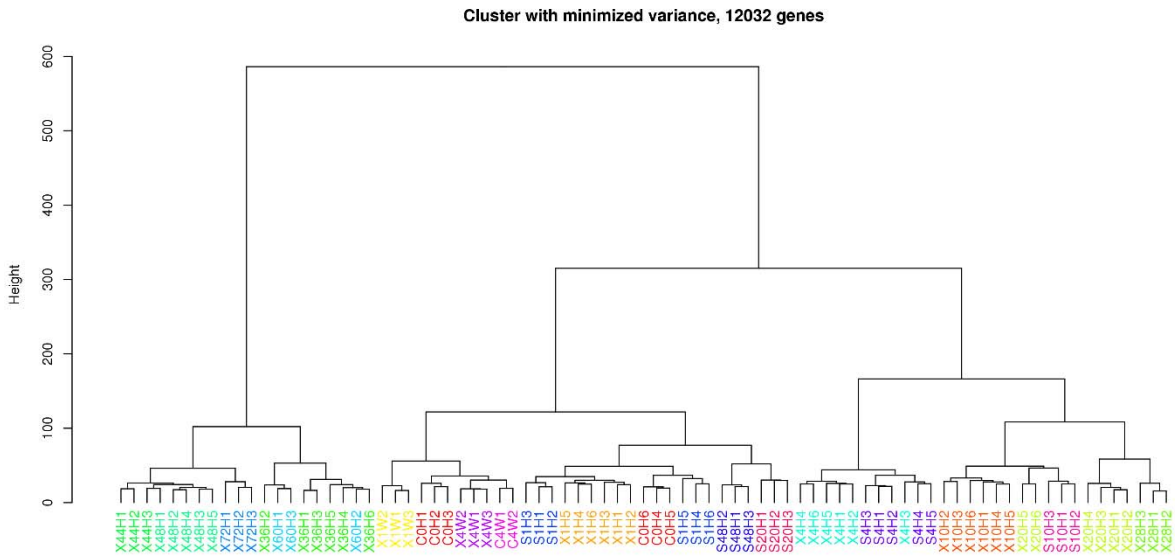



**Figure IV-2**. Classification of RNA-seq libraries based on the differences in expression of genes. Hierarchical clustering of the RNA-seq replicate libraries was done using the quantifications of the 12,490 expressed genes across samples. Control 0-hrs and 4-weeks samples are labeled 'C', post-PH experiment samples are labeled 'X' and Sham control sample libraries are labeled 'S'. All the RNA-seq libraries are represented as nodes of the tree and the length of the branches indicate the difference in gene expression changes across samples. The Ward's minimized variance criterion was used to minimize the total variance within-cluster [Ward, 1963]. All the replicates of a given condition have been assigned a similar color.

A Principal Component Analysis (PCA) informs further about the changes in gene expression occurring across conditions (Figure IV-3). For this analysis the calculated Reads Per Kilobase of transcript per Million mapped reads (RPKM) from the list of 12,032 expressed gene transcripts was used. The direct comparison of samples based on the differences in gene expression is highly complex to visualize and interpret, especially because there

are many gene transcripts, technically called dimensions. The principle of the PCA here is to summarize the changes in expression, technically called dimension reduction. As a result of the PCA, 'Principal Components' (PC) are created, as many as the original number of dimensions. Among them, the 1st PC (PC1) contains the biggest shared change observed among transcripts, the 2nd PC (PC2) takes the biggest change when removing the one contained in PC1, and so on for the rest of the components. In this way, the inspection of only the first principal components can inform about the biggest shared changes among samples. As a result of the PCA in the controls and regenerating livers, the PC1 contains 26% of the total variation and PC2 includes 19%, which gives a total of 45% of the variation represented in the PC1 and PC2 (Figure IV-3a). For ease of inspection, I averaged the positions of the different replicates in PC1 and PC2 (Figure IV-3a) and then show the standard deviations in a second plot (Figure IV-3b). Consistent with the hierarchical clustering, an inspection of the PC1 and PC2 plot shows that PH livers display two different cycles of gene expression. The first cycle is a Sham-like cycle. It occurs during the early time points post-PH until about 20 hrs. During this period, post-PH livers change similarly to the Sham controls, and they progress in a clock-wise manner back towards the 0 hr state. This shared Sham and PH cycle could be owing to normal variation (e.g. circadian cycle and/or feeding) and/or the surgical procedure. The relative contribution of these different effects was not determined. The second PH cycle is evident after 20 hrs. Then, the changes in gene expression diverge compared to the Sham samples, with a large difference evident at 48 hrs for PH and Sham samples. The second cycle is extended until one week in the analysis. Interestingly, the second cycle includes time points where cell division occurs (i.e., time points between 36 and 72 hrs) and thus may represent the post-PH proliferative response.

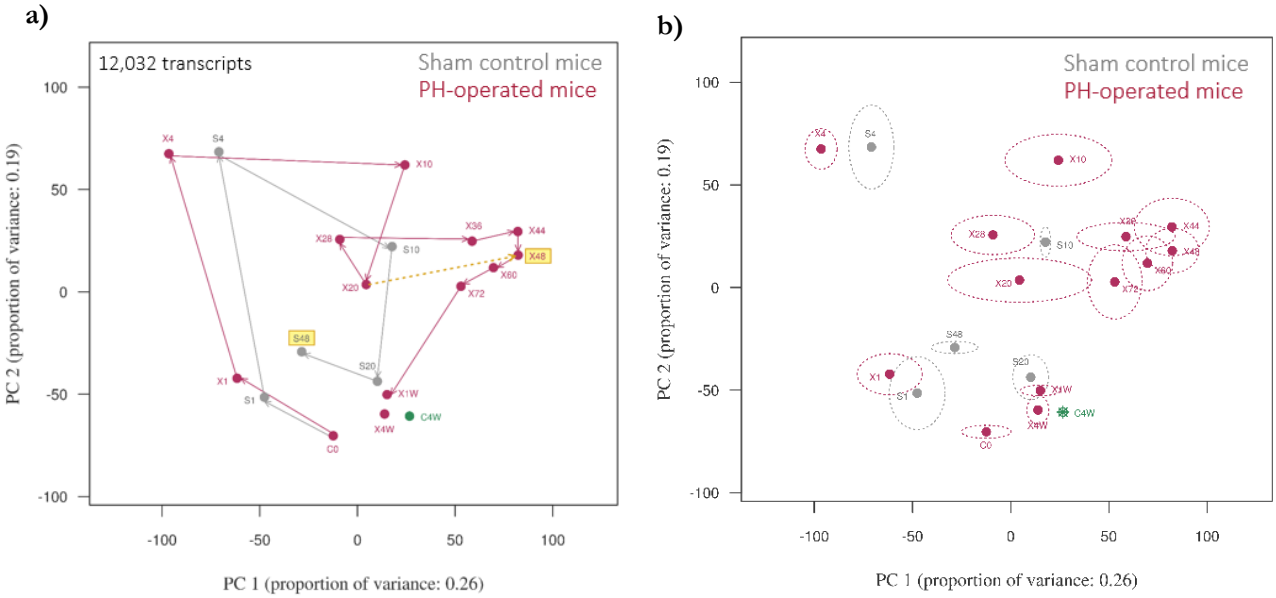


**Figure IV-3**. Two-dimensional plot displaying the coordinates of the collected samples in the PC1 and PC2 of the PCA using the set of 12,032 expressed genes. The samples displayed are the resting liver (C0 dot in red), post-PH samples (X dots in red), Sham controls (S dots in grey) and the control mice at 4 weeks (C4W dot in green). The replicates were averaged and displayed as single dots. a) Paths followed by the post-PH and the Sham liver transcriptomes. The respective 48 hrs time points are highlighted in yellow indicating that they display the biggest observed distance for the same time point between PH and Sham conditions. b) Standard deviations of the replicates for each condition displayed as ovals. The width of the ovals represents the +/- standard deviation for PC1, and the height represents the +/- standard deviation for PC2.

The interesting results showing two different response cycles post-PH encouraged me to look in more detail at the genes changing expression. I did a differential expression analysis between 0 hr resting livers and each time point during regeneration. In each pair of conditions, genes showing significant differential expression and with log2 fold-changes higher than 0.5 or lower than -0.5 were classified as changing expression post-PH. Three sets of genes were defined as a result based on their expression dynamics (Table IV-2): 1) genes not expressed, 2) genes expressed with stable levels across time points and 3) genes differentially expressed post-PH. 6,530 genes were detected with stable expression after PH. Interestingly, this set of genes is enriched for basic metabolic and gene expression functions (with a p-value lower than 0.05) that may be key for the survival of the organism. Also, healing functions of the complement and coagulation cascade were observed, and 8 genes out of 22 annotated core circadian clock genes. On the other hand, a total of 5,502 genes were classified as changing expression post-PH. An enrichment analysis of the associated genes provided a list of statistically significant functions that illuminates the functional changes in the liver after PH (data not shown). Genes in diverse metabolic pathways change their expression. Even though diverse sets of genes participate in metabolic pathways and their functions are highly interconnected, it is possible that the liver needs to adjust metabolic processes to allow or support regrowth or to recover from the surgery. This new metabolic state is also accompanied by significant differential expression of proliferative genes. Under normal conditions hepatic cells remain in a quiescent state, but upon PH the cell proliferation programs need to be initiated after a long resting period. Thus, new expression of cell cycle genes is necessary for cell division to occur. The enrichment analysis also showed that ribosome biogenesis function is enriched consistent with the necessity for new ribosomes to support the increased production of proteins for the two daughter cells. Further, the signaling of a significant number of elements in pathways such as PPAR, MAPK and ErbB change their expression. Signal transduction is very important for regeneration. Some key proteins in these pathways are known to be already expressed in livers before hepatectomy, such as NFkB [Fausto et al., 2000]. Yet, other intermediary proteins need to be generated consistent with the signaling pathway obtained in the functional enrichment analysis. Interestingly,

12 out of the 22 annotated core circadian clock genes are included in the set of differentially expressed genes. Indeed, part of the variation observed during regeneration could constitute normal variation in cells that may include the circadian cycle, where expression of genes changes in a cyclic mode.

| Set | Number |
|---|---|
| 1. Transcripts not detected | 25,959 |
| 2. Transcripts detected & stable | 6,530 |
| 3. Transcripts detected & changing | 5,502 |

**Table IV-2**. Classification of genes in three sets based on their expression dynamics. The table displays the number of gene transcripts with non-detected expression, with stable expression and with changing expression after PH.

For a better understanding of the expression changes occurring after resection I grouped the genes in the changing group by using the clustering method Partitioning Around Medoids (PAM). Here, this method groups together genes with similar profiles of RPKMs to any number of desired groups. To start, I calculated the 1-Pearson correlation as a measure of dissimilarity between gene quantifications, and then the PAM clustering put genes together by minimizing the dissimilarities within groups. The algorithm chooses one representative gene for each group, called 'Medoid', and iteratively adds genes to each group minimizing the dissimilarities among genes within clusters. Further, during the clustering process another gene may be assigned as the new Medoid of a given group if this decreases the overall dissimilarity among genes within the group. I ran the PAM clustering for 2 to 30 different groups.

The grouping of genes into two clusters already reveals interesting insights about the expression changes (Figure IV-4). On the left panel of the figure, I display a heatmap plot that displays with a color gradient the relative changes of expression for each gene across time points. Each row of the heatmap represents one gene and each column represents one time point starting with the resting liver in the first column, and the subsequent post-PH time points in the following columns. The heatmap displays relative changes of RNA-seq RPKMs per gene with a color gradient. The white color represents the lowest quantification for that gene and the red color is assigned to the highest quantification observed on genes. Therefore intermediate colors represent in a scaled manner the values in between the white and red extremes. The two-cluster grouping separates genes into two very distinct patterns: 2,415 genes have decreased transcript expression after PH that recovers when the regeneration is completed, and 3,087 genes have increased transcript levels during regeneration showing the highest levels in two stages: around 20 hrs and between 36 and 72 hrs. Given that these two groups are detected

by the clustering as the most different groups, perhaps these two groups may generally define the two cycles observed in the results of the PCA and hierarchical clustering described above.

The PAM clustering algorithm provides information about the quality of the clustering by means of a 'Silhouette score'. This score describes how well integrated a gene is in its group compared to the second most similar group, being a score of 1 being the maximum that a gene can achieve. The right panel of Figure IV-4 displays the sorted silhouette scores within clusters, with the top gene being the assigned Medoid. Genes in the left panel were identically sorted according to their silhouette scores. With this sorting many genes appear to have similar profiles. Nevertheless, the genes at the bottom of each group display low silhouette scores and their profiles differ marginally compared to the Medoids. As the PAM algorithm forces genes to be a group, genes at the bottom, having to be assigned to the group that best suited them, can be in fact anti-correlated; these appear with negative silhouette scores of which there are very few in the sorting shown in Figure IV-4. They are listed as 'outliers' in the figure. As abovementioned, it is possible to break the data into more groups to obtain further information about transcription changes during post-PH.

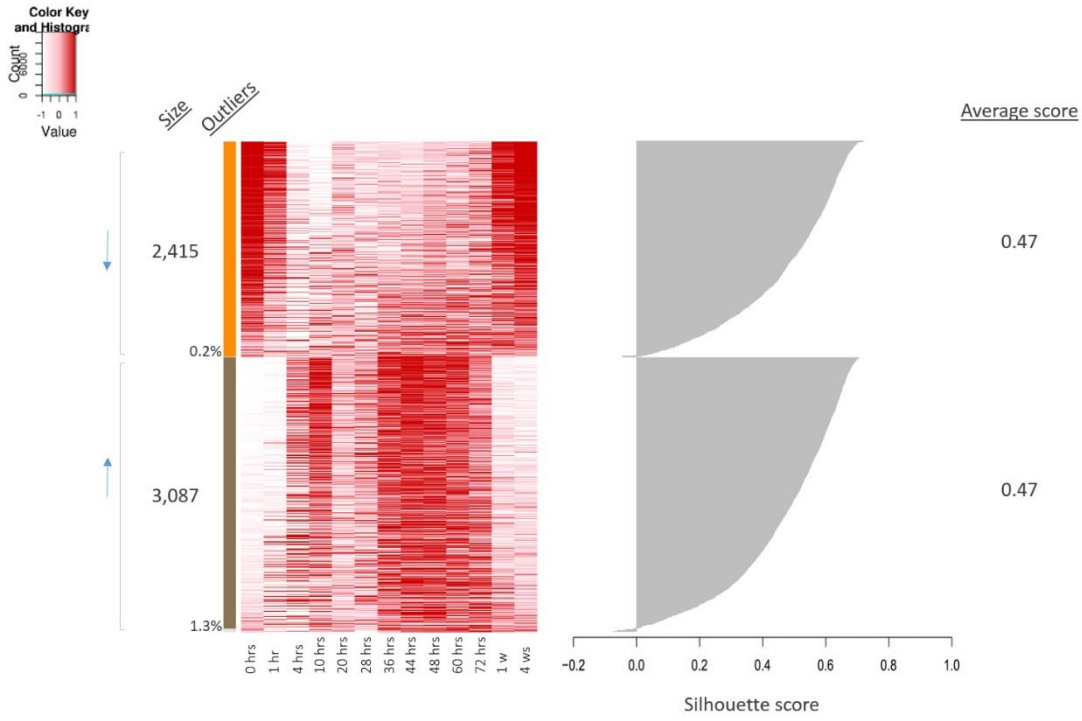



**Figure IV-4**. Clustering of genes changing expression into two groups. The PAM algorithm was performed for two groups with the list of 5,502 genes (Table IV-2) changing expression levels post-PH. Left panel: heatmap displaying the relative changes of expression of the genes in the two resulting clusters. The size of each group and the number of outliers are indicated

to the left. Additionally, arrows indicate the sense of the response post-PH: down-regulation, and up-regulation. Right panel: Silhouette score distributions within cluster and the average cluster score.

The clustering of genes into two groups distinguishes two very different patterns that are shared by most of the genes within clusters. The first cluster of genes displays higher expression values when no regeneration occurs, whereas the second cluster includes higher expression values during regeneration. When genes are split into more than two groups, the new clustering redistributes all genes into the new groups, even if genes share a very prominent pattern observed in the previous clustering with two groups. Figure IV-5 displays silhouette scores resulting for 'k' clusterings for 3 to 10 groups. The different clusterings show a generalized decrease of the average score per group compared to the clustering into two groups. This is a consequence that genes with a strong shared pattern found in the 2-group clustering are being separated and that genes are not completely correlated. Silhouette scores are calculated from two features from a gene assigned to a specific group: the average distance from that gene to the other genes in the same group, and the average distance of that gene to genes in the most similar group among the remaining groups. Thus, if genes among clusters share a big similarity, as it is the case here, the silhouette scores will tend to decrease, especially if their RPKM quantifications are not very well correlated. Nevertheless, the majority of the Silhouette scores are positive, suggesting that more detailed patterns arise from the new clusterings. Interestingly, there is always one group that stays with a high average score ranging between 0.46 and 0.51 irrespective of the number of groups (k=5 has 0.43, k=6 has 0.41 and k=7 has 0.51), suggesting that genes within this cluster are highly correlated and that their profile can be well distinguished from the other clusters. This group with higher silhouette values displays a profile (data not shown) with higher expression between 36 and 72 hrs across clusterings. This gene cluster is cleanest in the clustering of 7 groups, when the highest average score of 0.51 is acquired. On the other hand, the rest of the groups continuously tend to decrease their scores. Thus, to investigate the nature of different groups of genes (e.g. functional enrichment analysis) I chose the clustering with 7 groups to look further into the expression profiles (Figure IV-6).

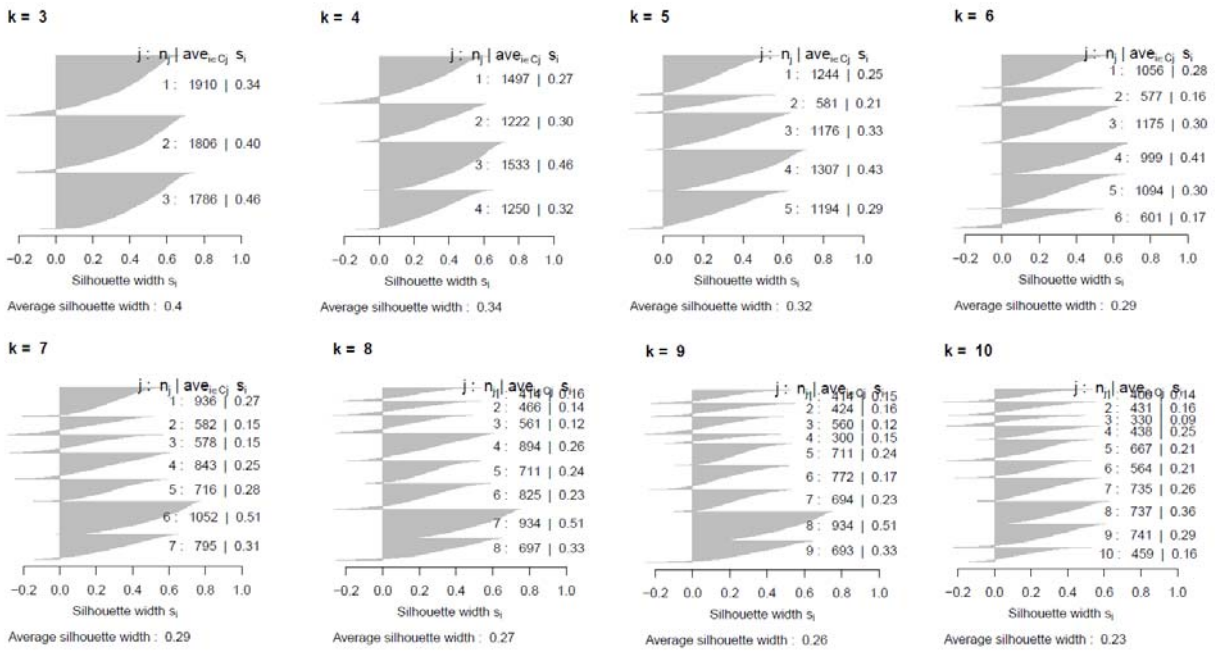



**Figure IV-5**. Distributions of silhouette scores across PAM clustering with the number of clusters 'k' varying from 3 to 10. On the right side of each cluster distribution, the cluster number (j) is indicated along with the number of genes included ($n_j$) and the average silhouette score (ave$i\epsilon_j$ $S_i$). On the bottom, the average cluster silhouette score is provided.

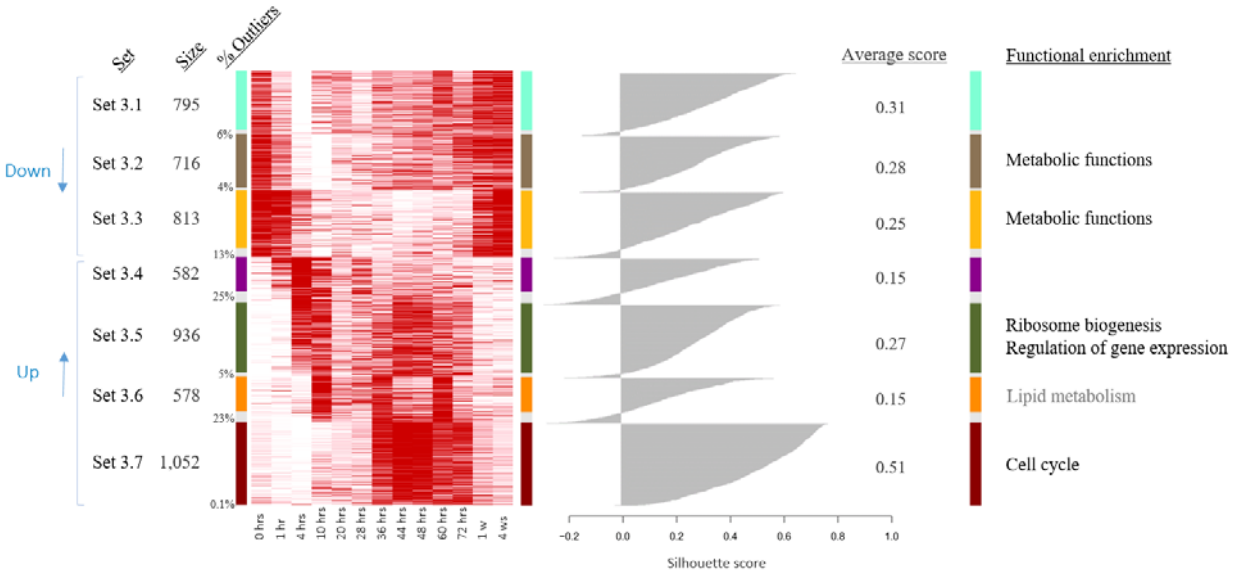



**Figure IV-6**. Clustering of genes changing expression into seven groups. PAM clustering was performed for 7 groups with the list of 5,502 genes changing expression levels post-PH. Left

panel: heatmap displaying the relative changes of expression of the clustered genes. Central panel: Silhouette score distributions within each cluster. On the left side arrows indicate the sense of the response post-PH: down-regulation, and up-regulation. Right panel: Results of a functional enrichment analysis for each gene cluster.

To better understand the results of the 7-group clustering, I sorted the groups depending on when they show the earliest peak of altered expression, and I named the sorted groups Set 3.1 to Set 3.7. The clustering with 7 groups still shows two main profiles: up- and down-regulation after the resection takes place, and yet it is possible to see more specific profiles with different waves of transcriptional responses to PH. The groups Set 3.1 to set 3.3 that show a gradual decrease of expression after PH and a gradual increase for Sets 3.1 and 3.2 starts already at 10 hrs after PH, whereas for the Set 3.3 it starts prominently much later at 1 week post-PH, near the end of the regeneration process. This gradual way of changing expression could be related to a homeostatic mechanism to adjust gene expression during the long process of recovery. The genes in sets 3.2 and 3.3 show a functional enrichment in metabolic pathways. This result suggests that the metabolism adapts after the surgery by means of gene expression changes. The genes up-regulated after hepatectomy, instead, show sharper profiles compared to the down-regulated Sets 3.1 to 3.3. Genes in Set 3.4 show an early response. They do not have any significant functional enrichment but, interestingly, among these genes there is the *Hypoxia-inducible factor 1-alpha* (*Hif1a*), which is known to respond to partial hepatectomy and helps sensing the need for sinusoidal endothelial reconstruction [Maeno et al., 2005]. Other genes included in this set that are also involved in signaling necessary for regeneration are *Jun*, *Cd14*, *Map3k6*, *Gadd45g*, *Mapk14*, *Gadd45b*, *Egfr*, *Map4k3*, *Mapkapk2*, *Rap1b*, *Nfkb1*. Set 3.5 also has this early response, but also shows a second round of expression increase during the cell division time points between 36 and 72 hrs. These genes are enriched for ribosome biogenesis and regulation of gene expression, such as mRNA surveillance, RNA transport, spliceosome and RNA degradation. This expression profile could be explained by a preparation of cells for proliferation where an increase of ribosome abundance is necessary. Set 3.6 has a profile showing three peaks at 10, 36 and 60 hrs. These times coincide with the times of the day and night when the mice are about to have (10 hrs) or have had 2 hrs access to food (36 and 60 hrs). For this relatively small set of genes there is no highly significant enrichment of functions but the more highly enriched pathways are related to lipid metabolism, which are pathways that are rapidly activated in the liver during food intake. Thus, this set of genes could be related to re-feeding responses. The last group of genes, Set 3.7, shows higher expression levels between 36 and 72 hrs, which are the times when cell proliferation is expected to occur. This group, which was retained over the different numbers of groupings, is consistently linked to annotated cell-cycle pathways.

The k=7 clustering of transcript levels in Figure IV-6 shows different waves of expression linked to liver and cell proliferation functions. Nevertheless, it could be that the responses observed are due to normal liver

function or more specifically to the manipulation of the mice during the experiment that could induce stress or healing, as the Sham-like cycle in the PCA suggests. To identify the regeneration-specific responses, I compared the post-PH responses to the Sham samples available by calculating log2 fold-changes between the Post-PH and the Sham quantifications (Figure IV-7). At each time point, to reduce the noise, I removed those genes showing little to no expression in the Sham and post-PH samples (i.e., RPKM quantification lower than the threshold of expression 0, higher than the previously applied expression threshold of -1.3 to avoid big fold-changes originating from low scores). This analysis shows that at 1 hr there is a general tendency for genes to stay close to the line of zero fold difference between PH and Sham samples. Yet, the genes in Set 3.4 show a slightly increased expression post-PH relative to Sham mice. This is consistent with the fact that this group includes very responsive and regeneration-specific genes such as *Hif1a*. Later on, at 20 hrs, the genes in Set 3.3 start to show a higher Sham expression compared to post-PH livers, which increases progressively until the 48 hrs time point. The top 10 genes showing the biggest differences are the stearoyl-Coenzyme A desaturase 1, 3 and 4 (*Scd1*, *Scd3* and *Scd4*), the predicted gene *Gm4956*, the *Hsd3b5* gene that is involved in steroid hormone biosynthesis, the pseudogene *Serpina4-ps1*, the regulator of iron entry into the blood *Hamp2*, the indolethylamine N-methyltransferase (*Inmt*), the thyroid hormone responsive protein *Thrsp* involved in lipid biosynthesis, and *Dsg1c* involved in cell adhesion. Also at 48 hrs one gene case shows a higher expression in the Sham samples than in the regenerating livers: the regulator of G-protein signaling 16 (*Rgs16*). These gene identification indicate a decreased expression of some metabolic and biosynthetic activities in the liver during preparation and progression of cell proliferation. Finally, during liver regeneration, group Set 3.7 displays the most significant differences between PH livers and Sham controls, and these differences are not early responses but responses occurring at 48 hrs after the resection. This shows that the cell proliferation functions occurring at 48 hrs are regeneration specific.

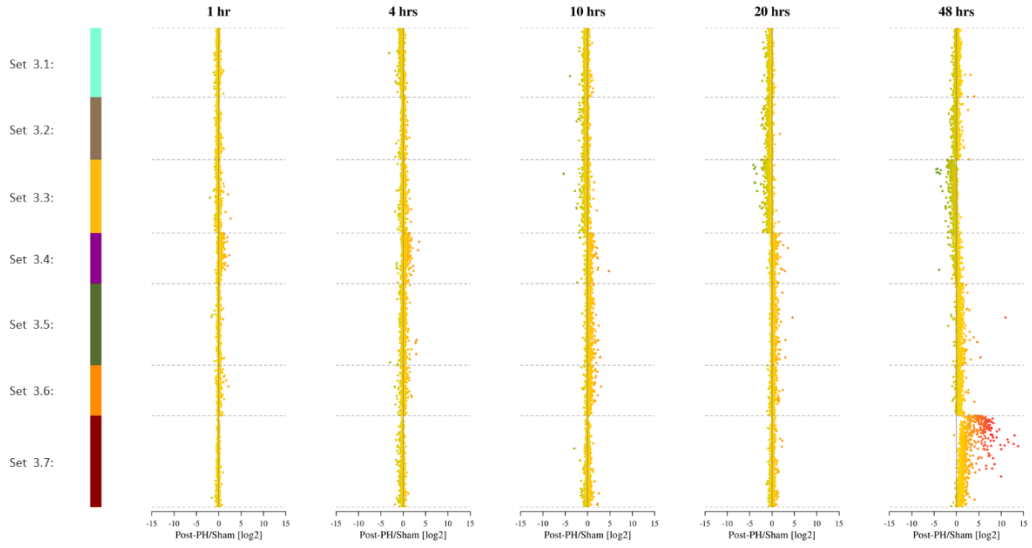


**Figure IV-7**. Ratio of the averaged replicate log2(RPKM) that compares Post-PH vs Sham samples at the time points 1 hr, 4 hrs, 10 hrs, 20 hrs and 48 hrs. The genes are sorted in a similar way as in Figure IV-6. A color gradient is applied to color the dots which describes how extreme their associated values are. Green colors represent higher scores in the Sham sample, while red scores represent higher scores in the post-PH samples. Gene transcripts displaying log2(RPKM) quantifications lower than 0 at both the Sham or Post-PH samples are not shown.

## IV.2 The transcription of 1/3 of the genes annotated as cell cycle regulators shows regeneration-specific responses in the mouse liver

I studied further the transcripts involved in cell proliferation. For that I used the list of 125 genes annotated in the KEGG pathway describing the murine cell division cycle and I checked where the genes fall in the gene sets I defined from 1 to 3, including Set 3.1 to Set 3.7 (Figure IV-8). The 40% of the cell cycle genes is included in Set 3.7. Some of them are very different at 48 hrs after resection, maybe because they are required for mitosis at 48 hrs. Other cell-cycle genes in Set 3.7 that show less differences with the Sham 48 hrs, would maybe show higher differences with a Sham 36 hrs or other times when proliferation is occurring. Unfortunately, more time points are not available. Interestingly, not all the regeneration-specific responses at 48 hrs are involved in the cell division cycle. They are included in diverse pathways such as gene expression or Glycerolipid metabolism. This suggests that the liver not only regrows but also adapts other relevant functions.

Not all the cell cycle genes are included in Set 3.7. Indeed, 13 of the annotated cell cycle genes are not detected as expressed at any of the studied time points. This could be due to different reasons. Perhaps some cell cycle genes are tissue specific and express in other tissues or cell lines. Further, thirty-nine genes belong to Set 2 and are thus already expressed but not varying during regeneration. These cell-cycle genes could already be expressed in the liver because of their alternative participation in other processes. This could further support the hypothesis that some cell-cycle genes are also required for other functions in the liver.

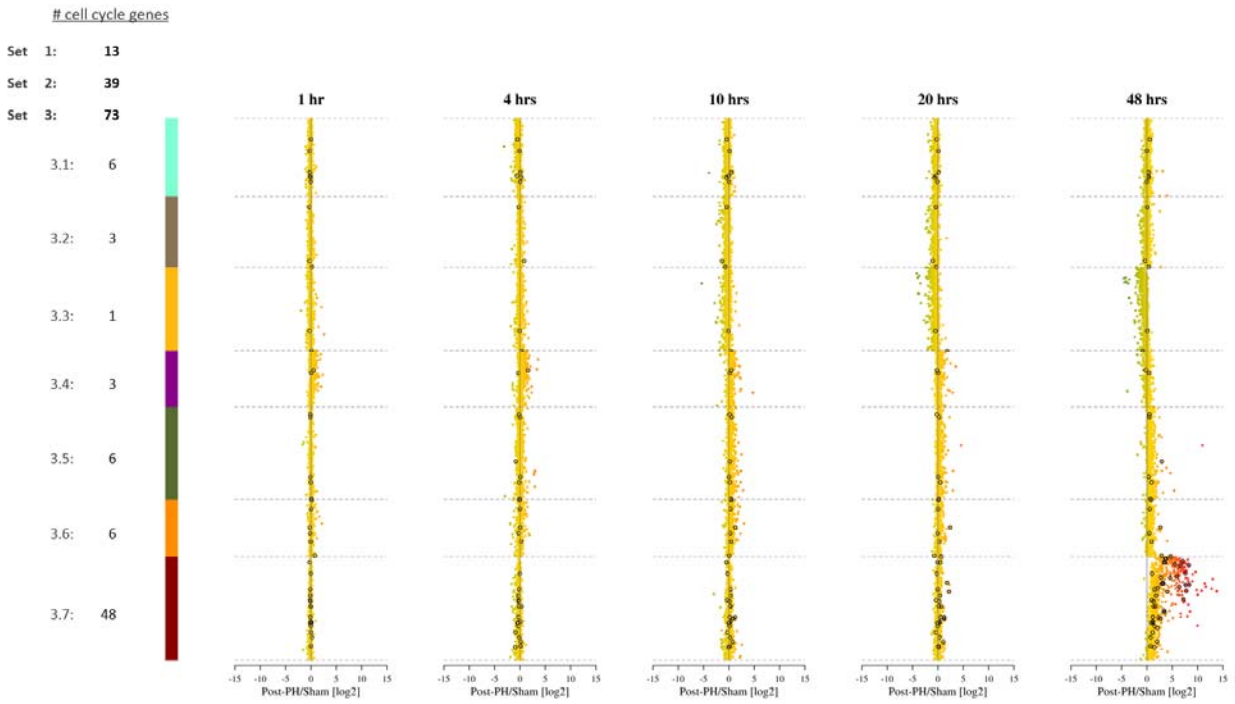



**Figure IV-8**. Ratio of the averaged replicate log2(RPKM) that compares Post-PH vs Sham samples at the time points 1 hr, 4 hrs, 10 hrs, 20 hrs and 48 hrs. The difference compared to figure IV-7 is that the genes annotated as cell-cycle regulators have been highlighted with blue circles. Gene transcripts displaying log2(RPKM) quantifications lower than 0 at both the Sham or Post-PH samples are not shown. The number of cell-cycle genes in each of Sets 1-3 (Table IV-2) and Set 3 groups 3.1 to 3.7 are listed to the left of the figure.

Some cell-cycle genes with differential expression were classified into Sets 3.1 to 3.6. Remarkably, most of these transcripts change similarly in the Sham livers, but a few of them are slightly more expressed after resection, such as the *Ccnd1* gene, encoding Cyclin D1, which shows the highest fold-change in Set 3.6. The synthesis of this cyclin gene is initiated during G1 and drives G1 phase progression. Figure IV-10a shows that the expression of Cyclin D1 is activated in both the Sham and the resected livers by 4 hrs, and the early expression progresses in an impressively correlated manner between post-PH and Sham controls. Later on, after 4 hrs, the expression of *Ccnd1* in Sham livers and post-PH livers diverges with higher levels in the post-PH samples. This profile suggests that the 1st G1 phase after resection could be initiated somewhere between 4 and 10 hrs post-PH.

Figure IV-9 also displays the expression profiles of other classical cell cycle genes, such as those encoding Cyclin E, Cyclin A, Cyclin B and CDK1. The expression profile of these other cyclins in the Sham livers stays stable at very low or no expression levels, whereas in the regenerating livers there is an increase of expression after

the early stages. Interestingly, the maximum transcript levels in regenerating livers are reached close to the time when their protein expression is expected. Cyclin E is required for the transition from G1 to S phase and it consistently starts to be expressed between 28 and 36 hrs, at the time of initiation of the first synthesis phase post-PH (see Figure IV-1). The somatic Cyclin A form cyclin A2 is necessary for different phases of the cell cycle, but it is first required for the progression of the synthesis phase. Consistently, Cyclin A2 is expressed between 28 and 48 hrs (Figure IV-9). Cyclin B is a protein that is expressed by S phase where it will bind to the Cyclin Dependent Kinase 1 (CDK1) and become active by phosphorylation. It will then promote entry and exit from mitosis. During mouse liver regeneration the expression of this protein is quite similar to what I just described. Cyclin B2 starts to be expressed just before S phase and peaks by M-phase entry at 44 hrs. CDK1, the partner protein of both Cyclin A and B, it is always expressed in immortalized cells. Here, in livers that just entered a proliferative state, it shows renewed expression before S phase, and peaks by its end when it is required to bind either Cyclin A or B. The expression of these classical cell cycle regulators seems to be very precise in time. This is also confirmed by studying the Pol2 occupancy in those genes (Figure IV-10). In all the gene examples the mRNA levels and the Pol2 occupancy in the promoter and the gene body tend to be highly coordinated on the times when there are detectable mRNA levels. In the genes *Ccna2* and *Ccnb2* there is an accumulation of Pol2 previous to the its elongation that is observable at 10 hrs post-PH and lasts for at least 28 hrs before Pol2 is released. It is interesting to see how the promoter of those genes has already responded. Probably there are chromatin changes that have occurred in them already by 10 hrs post-PH. And the gene is somehow waiting during 28 hrs to initiate its transcription.

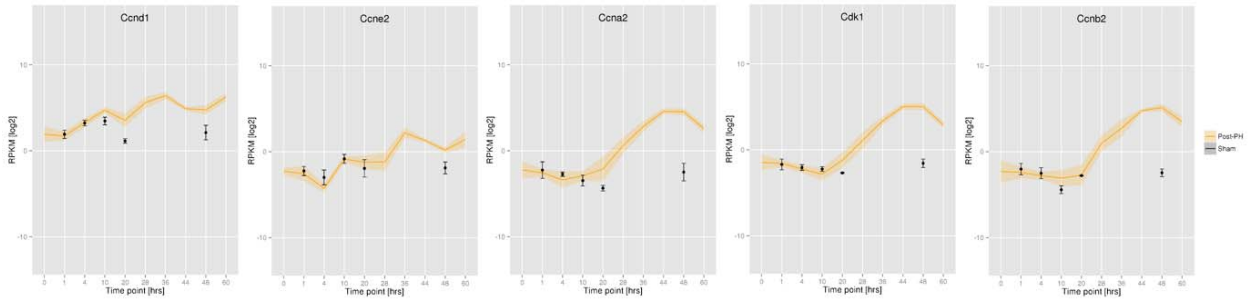



**Figure IV-9**. Expression profiles of the classic cell-cycle genes *Ccnd1*, *Ccne2*, *Ccna2*, *Cdk1* and *Ccnb2*. Quantifications from both post-PH (yellow shaded line) and Sham (black error bars) replicate samples are shown until the 60 hrs time point.

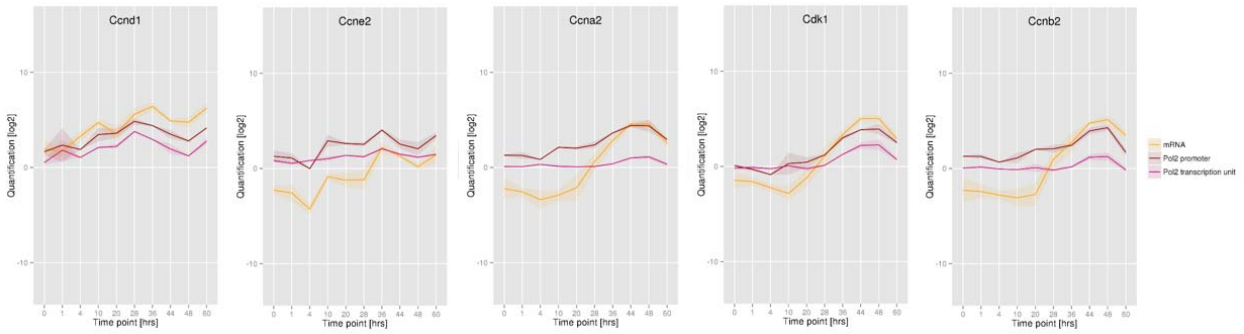



**Figure IV-10**. Transcription profiles of the classic cell-cycle genes *Ccnd1*, *Ccne2*, *Ccna2*, *Cdk1* and *Ccnb2*. The transcript levels are displayed in yellow, and in brown and pink the Pol2 occupancy in the promoter and transcription unit is displayed, respectively. Post-PH samples are shown until the 60 hrs time point.

## IV.3 Early responding genes are already occupied by Pol2

I have shown above that after PH there is new transcription occurring on many genes, and for other genes there is an adjustment of their expression in response to the PH resection. A question to ask is how coordinated are the different steps of transcription throughout regeneration. To answer that question I collected those genes with up- or down- regulated transcript levels at any transition between two adjacent time points after PH. Figure IV-11 displays in yellow the sorted fold-changes in the transitions from 0 to 1 hr and 28 to 36 hrs, and this information is compared to the corresponding Pol2 fold-changes in the promoter and body. In most of the transitions available I could observe very coordinated Pol2 promoter and body occupancy and transcript abundance fold-changes, as in the case of the transition between 28 and 36 hrs. But there is one exception, the transition from 0 hr quiescent state to the early response at 1 hr. At 1 hr there are already 300 genes with differential transcript levels, and Pol2 occupancy in the body of the genes that tends to follow this up regulation. In contrast, for those cases with up-regulated expression, the Pol2 occupancy in the promoter changes little. This group of up-regulated genes is significantly enriched for functions such as regenerative signaling and acute response and includes genes such as *Jun*, *Jund*, *Mapkapk2*, *Tnfrsf1b*, *Gadd45b*, *Gadd45g*. These genes are known to respond very rapidly in the priming phase of regeneration [Fausto et al., 2000; Li et al., 2001; Cressman et al., 1995; Taub et al., 1996; Su et al., 2002]. As illustrated in Figure IV-12, these genes are already transcribed with low density of Pol2 across the body. Perhaps their expression is responsive to changes in the environment. Further, Pol2 also occupies the promoter of those genes in the 0 hrs resting liver, and there is a slight change in occupancy at 1 hr. As their promoters are already prepared for transcription, this could facilitate a fast increase of transcription later in response to the removal of 2/3 of the liver. Further, in both cases of the

*Gadd45g* and the *Tnfrsf1b* promoters it is observable that there is a pair of peaks in the promoter that could represent the process of formation of the preinitiation complex and pausing. In the case of *Tnfrsf1b*, however, there is an extra middle peak at 0 hrs of unknown significance. Altogether, this suggests that Pol2 is poised at the promoter in the 0 hrs resting liver for rapid activation that is observed at the mRNA level already at 1 hr post-PH.

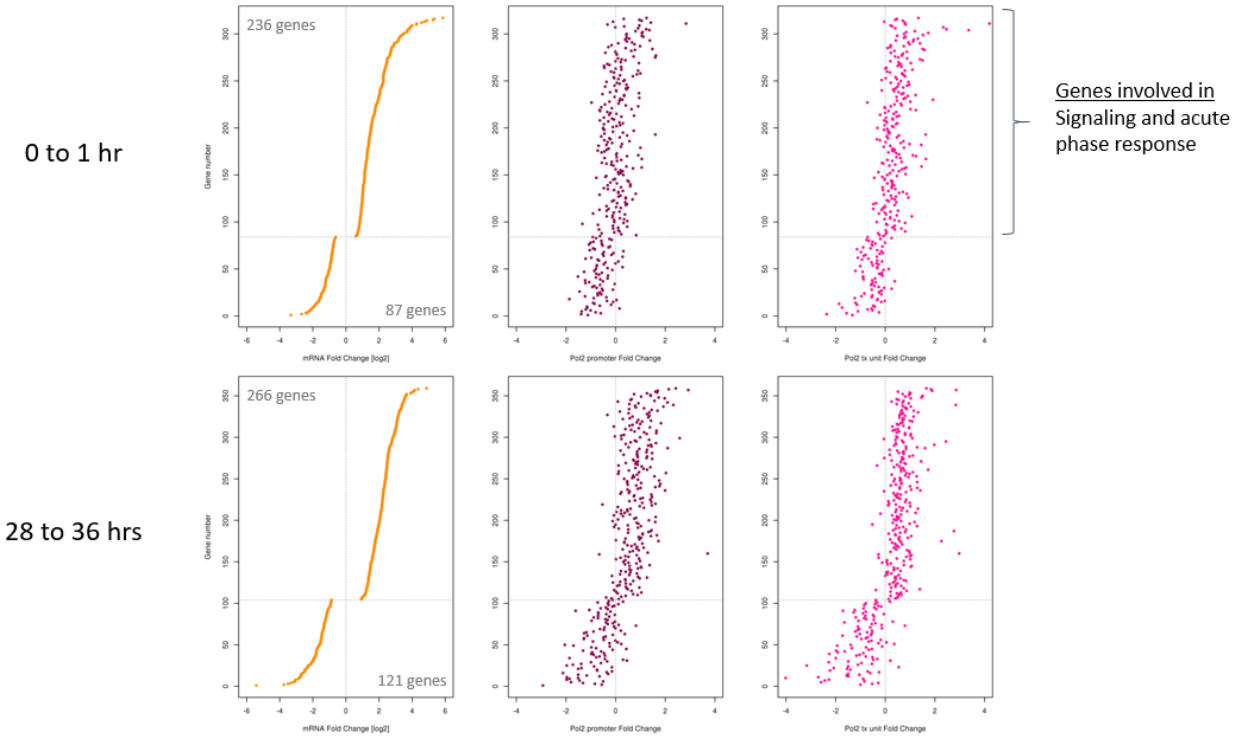



**Figure IV-11**. Transcript and Pol2 occupancy fold-changes in genes whose transcript levels are up- or down- regulated in a given transition between time points. The three upper panels display data for the transition between 0 and 1 hr post-PH, and the bottom panels display data for the transition between 28 and 36 hrs post-PH. In each set of three panels, the left-most panel displays transcript fold-changes of the genes detected as changing transcript levels. The genes are sorted by their fold-change starting in the bottom with the more negative fold-change between the first time point and the second, and increasing until they reach positive fold-changes. Significant up- or down-regulation of transcripts were selected having a log2 fold change higher than 0.5 or lower than -0.5. In the central and right-most panel the corresponding Pol2 fold-changes in the promoter and the transcription unit are displayed, respectively. Additionally, the function of the genes in the upper quadrant of the transition at 1 hr post-PH is indicated.

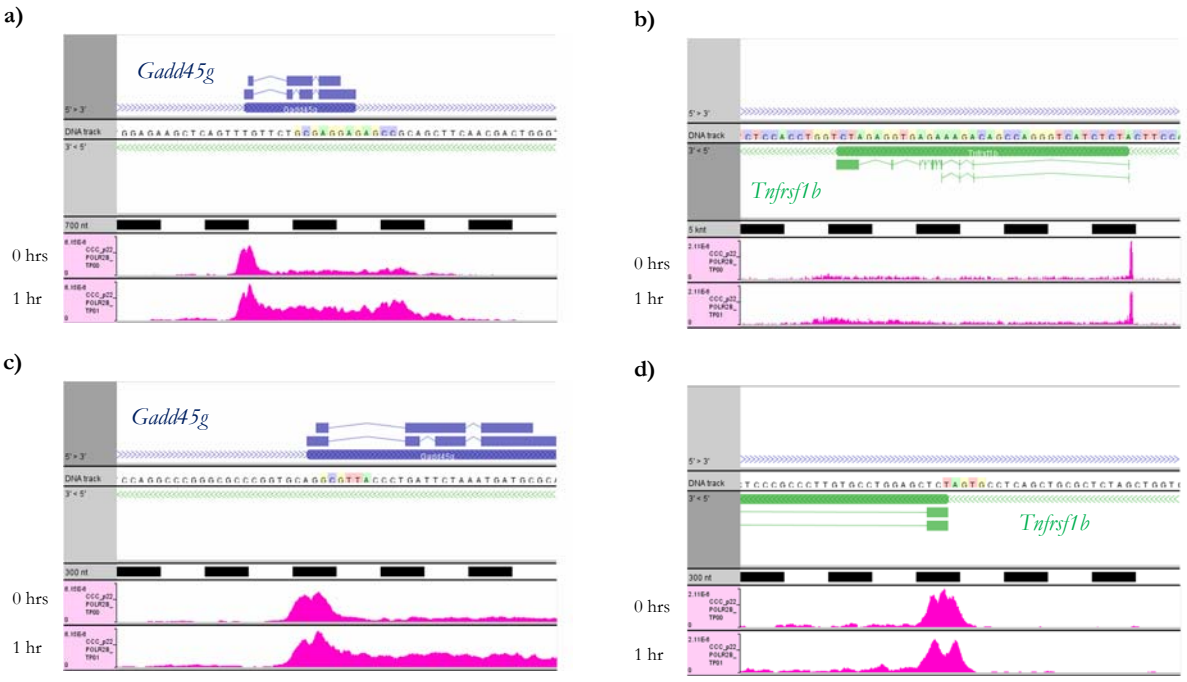



**Figure IV-12**. Pol2 profiles of the *Gadd45g* and *Tnfrsf1b* genes in the 0 hrs resting liver and at 1 hr post-PH. In the top of each panel the structure of the annotated gene is shown (in green the genes on the negative strand, and in blue the genes on the positive strand). Pol2 ChIP-seq profiles are shown in pink. The profiles displayed have been created by the accumulation of the central 50 bp of the sequenced paired-end fragments with sizes between 50-500 bp. a, b) Pol2 occupancy in the entire gene. c, d) Pol2 occupancy in the promoter region.

## IV.4 Overall transcription of genes and their transcript levels are highly correlated, although there are interesting exceptions

After a careful inspection at the fold-changes of transcript abundance and Pol2 occupancy between transitions I could identify three different cases of transcription dynamics. 98% of the differentially expressed genes show a coordinated change in Pol2 occupancy and transcript levels. Nevertheless, in approximately 70 cases out of 5,502 genes a slight delay in the transcript production can be observed, like in the case of the metabolic Acsm5 gene, which is involved in the regulation of fatty acid metabolism (Figure IV-13). Between 4 and 10 hrs there

is an increase of Pol2 occupancy in the promoter and body of the gene, while increased mRNA levels are observed only after 10 hrs post-PH. In these cases of genes, the transcription seems to take a long time.

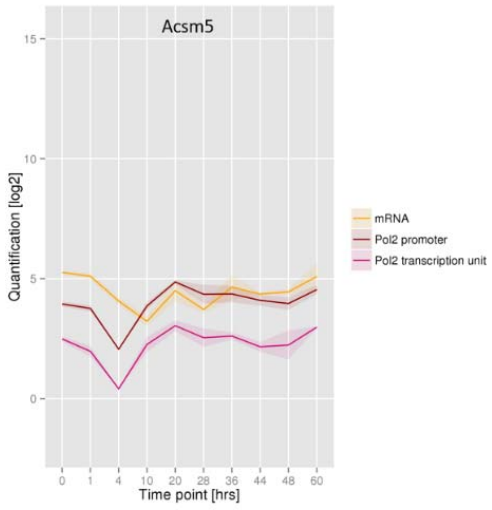



**Figure IV-13**. Transcription profiles of the *Acsm5* gene involved in the regulation of fatty acid metabolism. The transcript levels are displayed in yellow, and in brown and pink the Pol2 occupancy in the promoter and transcription unit are displayed, respectively. Post-PH samples are shown until the 60 hrs time point.

There is a unique case where the transcript and Pol2 occupancy are anti-correlated: the long non-coding Neat1 lncRNA gene (Figure IV-14). Neat1 is a highly abundant 4 kb ncRNA that is unspliced and since it is polyadenylated it could be selected during sample preparation [Clemson et al., 2009]. Neat 1 is involved in the formation of the nuclear compartments called 'paraspeckles' whose function is not well-known. Nevertheless, the paraspeckles are very dynamic structures that accumulate specific proteins and ncRNAs, such as Neat1, and whose structure can be altered in response to changes in cellular metabolic activity and Pol2 transcription [Fox et al., 2002], like the ones occurring during liver regeneration.

The Pol2 and mRNA profiles of *Neat1* show how when Pol2 reaches a minimum, a peak of mRNA takes place and vice versa. This profile could be explained as a delay of Neat1 lncRNA production compared to Pol2 transcription. However, there is a difference compared to the previous example Acsm5. In the Acsm5 gene the Pol2 activity starts when the mRNA levels have already decreased, thus there could be a homeostatic mechanism to recover the *Acsm5* levels. Instead, Pol2 occupancy in the *Neat1* gene begins when there are already high transcript levels. Curiously, the transcript levels are drawing sort of two phases: early phase (around 10 hrs) and cell cycle phase (around 36 hrs). Since this lncRNA is very responsive to changes in cellular metabolic activity and Pol2 transcription it could be that the structure of paraspeckles is affected during the

two response cycles to PH, first due to metabolic changes in the early stages of regeneration and later on due to cell proliferation. In order to confirm this, it would be interesting to do immunostaining experiments of paraspeckle-specific components during liver regeneration.

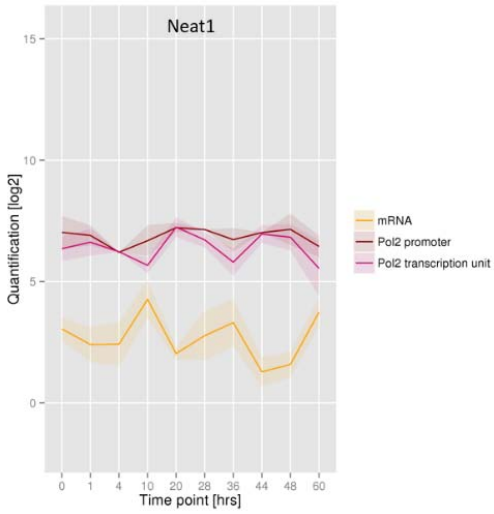



**Figure IV-14**. Transcription profiles of the *Neat1* lncRNA gene. Neat1 is a PolyA RNA that has been shown to regulate the formation of paraspeckles. The transcript levels are displayed in yellow, and in brown and pink the Pol2 occupancy in the promoter and transcription unit are displayed, respectively. Post-PH samples are shown until the 60 hrs time point.

The *Saa* cluster of genes are another interesting case of transcriptional response as they display a special and big response in both the PH and Sham samples (Figure IV-15). These genes form a cluster that encode for apoliproteins (proteins that bind lipids), and they are major acute-phase proteins in mice [Lowell et al., 1986]. In response to PH each Saa gene display a different transcriptional response, although they are known to respond to the same inflammatory stimulus [Koj 1974, Kushner 1982]. The *Saa4* and *Saa1-like* genes display a profile with an increased Pol2 transcriptional activity accompanied with increased H3K4me3 levels and high H3K36me3 levels. In contrast, *Saa1*, *Saa2* and *Saa3*, show a switch on Pol2 activity (very correlated with transcript levels, data not shown) with none or very low H3K4me3 and H3K36me3 accumulation. Indeed, the transcription and histone modification patterns along with their organization suggest that the promoters of the *Saa1* and *Saa2* genes are related. Further, although very little there is Pol2 and H3K4me3 signal in between the divergent promoters. These unusual chromatin profiles may represent novel regulation mechanisms of transcription for these strongly and quickly activated promoters.

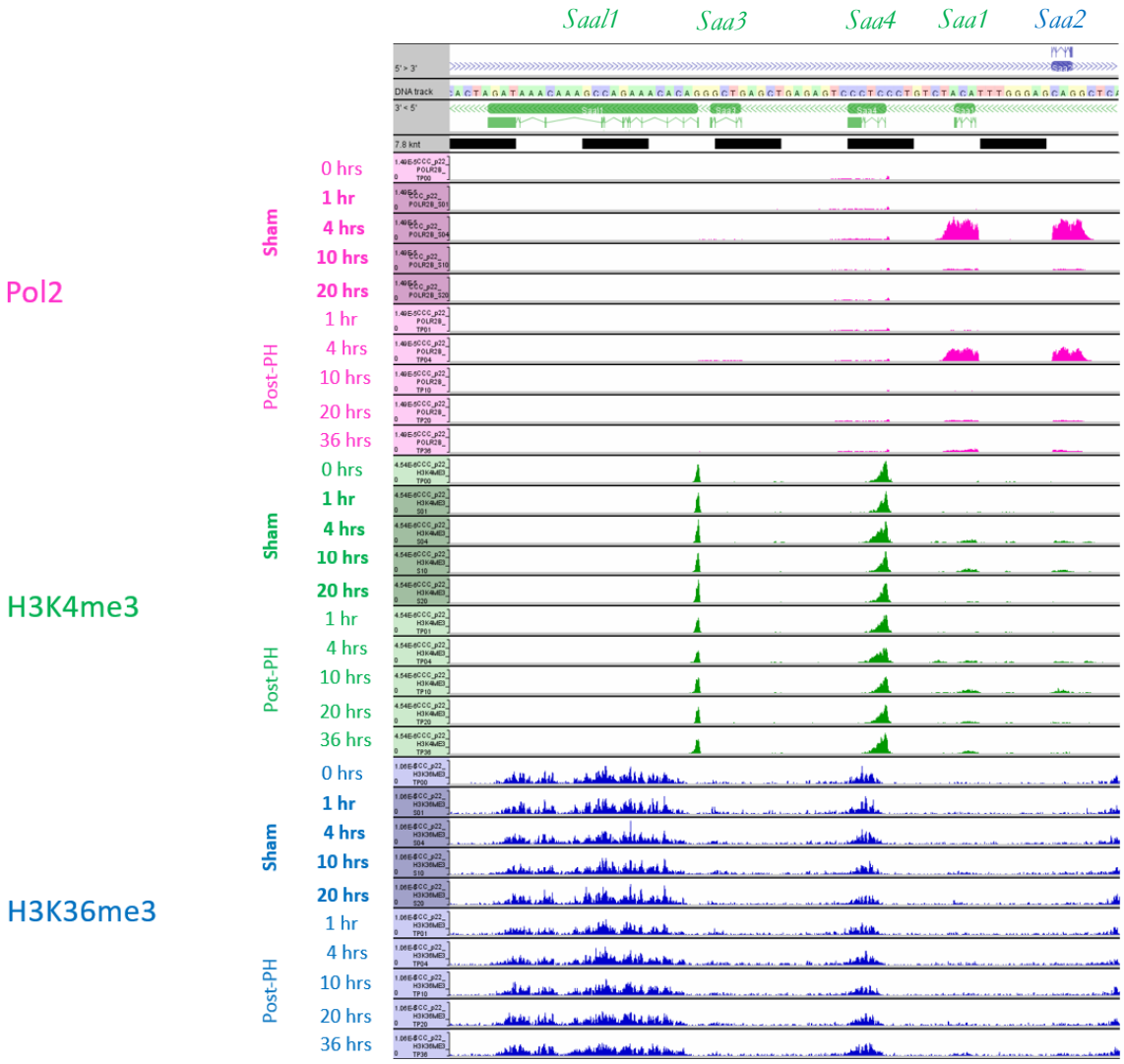



**Figure IV-15**. ChIP-seq profiles of the *Saa* gene cluster. These are acute-phase response genes. In the top panel the structure of the annotated genes is shown (in green the genes on the negative strand, and in blue the genes on the positive strand). Tracks are displayed containing ChIP-seq profiles from different epitopes: Pol2 (pink), H3K4me3 (green) and H3K36me3 (blue). For each epitope different conditions are displayed. The first is the resting liver (0 hrs), followed by the Sham samples at 1 hr, 4 hrs, 10 hrs and 20 hrs, and then the post-PH samples at: 1 hr, 4 hrs, 10 hrs, 20 hrs and 36 hrs. The profiles displayed have been created by the accumulation of the central 50 bp of the sequenced paired-end fragments with sizes between 50-500 bp.

# Discussion

## The mouse liver as a model to study cell proliferation in vivo

A major goal of the study of murine liver regeneration is to better understand cell growth and proliferation in vivo so that it can be translated into better treatments for uncontrolled tumorigenic cell proliferation or to apply this knowledge in liver regenerative medicine. A potential approach for treatments could be based on the mechanisms regulating gene transcription. In this chapter I have identified potential gene candidates whose expression is required for regeneration, even though they are not annotated to be genes regulating cell growth or proliferation. One thing that in-vivo studies like this teaches us is that different physiological functions are interconnected or may be tissue specific. The liver is a very important metabolic organ and its function is highly regulated by processes such as the circadian cycle. All these interconnections may be taken into account for the design of treatments in an organ-specific manner.

## The Sham control mice are very important for the research of liver regeneration

Mouse liver regeneration in response to 2/3 PH is a well stablished system to investigate the regeneration process that the human liver is also able to undergo. Nevertheless, the process of the operation alone can induce molecular changes in liver cells. Furthermore, normal daily processes such as circadian cycle or feeding can influence gene expression profiles in a time course. Thus, to understand liver regeneration it is key to have a control for the operation. The samples from Sham operated mice have thus been very important for this project. The comparison of Sham versus Post-PH samples shows an overlap of responses that may be introduced by the operation. Further, this comparison has permitted the detection of regeneration-specific gene expression responses. I must acknowledge Professor Ueli Schibler for his advice on utilizing Sham controls for this project.

## Different Pol2 responses might be associated to the early and late responses to partial hepatectomy

The genome-wide analysis of Pol2 occupancy has revealed different responses to PH. I have shown that during regeneration there are two types of transcriptional responses: 1) rapidly increased transcription on promoters already poised to act and 2) progressive responses associated with an increased Pol2 promoter and gene

occupancy. Additionally, in some cases such as the *Saa* genes, Pol2 occupancy is not associated with histone marks traditionally observed during active transcription. Indeed, different transcription initiation mechanisms might be associated with the different responses. Further studies on those promoters with unusual patterns could give more insights into transcriptional regulation. For example, sequence features such as TATA boxes or CpG islands could be differentially associated with different mechanisms of initiation.

## About the hepatocyte-specificity of this study

The liver is composed by different types of cells where approximately 70-80% of the genomes originate from hepatocytes, and yet each cell type participates in different ways and times in the process of liver regeneration [Widmann and Fahimi, 1975; Ukai et al., 1990; Malik et al., 2002]. Something that needs to be evaluated is to what extent the results observed in this research describe transcriptional activities in hepatocytes or in the other cell types of the liver.

Probably, the proliferative responses observed between 36 and 48 hrs that are regeneration-specific are occurring in hepatocytes, as they are the first cell types to proliferate and they do so at that time [Minocha et al., ms. in prep.]. For example, the increase of *Ccna2* expression that is observed at 36 hrs, is hepatocyte-specific, as also observed by immunostaining [Minocha et al., ms. in prep.].

The ChIP-seq data could be useful to answer this question, as its contents are influenced by the percentage of genomes present in the liver. Given that about 80% of the genomes in the liver originate from hepatocytes, high Pol2 scores probably originate from hepatocytes. On this regard, it would be interesting to identify highly expressed genes specific to non-parenchymal liver cells and see how represented they are in the Pol2 data.

Furthermore, not all hepatocytes proliferate during regeneration. Some of them do it, and they may initiate new expression of cell proliferation genes, whereas others keep on doing metabolic functions [Gebhardt and Matz-Soja, 2014]. Potentially, the regeneration-specific responses observed at 48 hrs that are not originating from cell cycle genes, could be produced in non-proliferating hepatocytes.

## Homeostatic response coordinated transcriptionally

The liver performs essential metabolic, synthetic and detoxification functions. Still an open question is how these important functions are maintained during regeneration and their role to support regrowth of the liver. Total liver function likely changes because of the loss of functional cell mass owing to resection. In response,

alterations in hepatic metabolism have been observed in livers after partial hepatectomy that change differently compared to Sham control mice [Rudnick and Davidson, 2012; Huang and Rudnick, 2014].

In this chapter I show interesting results that show a regeneration-specific decrease of expression of metabolic and biosynthetic genes while an increase occurs at cell proliferation genes. Between 20 and 48 hrs post-PH, genes in Set 3.3 and Set 3.4 related to metabolic functions display a decreased expression that is specific for regenerating livers. In contrast at 48 hrs there is a bigger number of genes in cluster 3.7 that increase their expression during regeneration. These genes are involved in cellular functions such as the cell division cycle or metabolic pathways like Glycerolipid metabolism or lysosome. All together these results might inform about a balanced metabolic and proliferative function regulated at the gene expression level during liver regeneration.

# Methods

## Animals

C57/BL6 12-14 week old male mice were housed. The experimental mice were entrained for two weeks on a 12 hrs on/12 hrs off feeding and dark-light cycle with food access between ZT12 and ZT24 (ZT0 is defined as the time when the lights are turned on and ZT12 as the time when the lights are turned off).

## Partial hepatectomies and Sham operations

Mice were subjected to 2/3 PH surgery under isoflurane anesthesia at ZT2, and sacrificed together by cervical dislocation at 1, 4, 10, 20, 28, 36, 44, 48, 60, 72 hrs and one and three weeks post-surgery. PH was performed as described [Mitchell and Willenbring, 2008]. Sham-operated controls were subjected to laparotomy and livers were collected at 1, 4, 10, 20 and 48 hrs after operation.

## Chromatin immunoprecipitation (ChIP)

Protocol adapted from Le Martelot et al. (2012). For ChIP, three mice livers were used per time point and the nuclei were pooled. Two biological replicates were prepared for each replicate. Full livers or livers post-PH from mice were used for ChIP. Livers were immediately homogenized in 4 ml per liver of 1x PBS including 1% formaldehyde, and the homogenate was kept for 5 min at room temperature. Cross-linking reactions were stopped by the addition of 25 ml of ice-cold 2.2 M sucrose buffer (150 mM glycine, 10 mM HEPES pH 7.6, 15 mM KCl, 2 mM EDTA, 0.15 mM spermine, 0.5 mM spermidine, 0.5 mM DTT and 0.5 mM PMSF). The homogenate was layered on top of a 10 ml cushion of 2.05 M sucrose (containing the same ingredients and including 10% glycerol and 125 mM glycine) and centrifuged for 1 hr at 24,000 rpm (100,000g) at 4 °C in a Beckmann SW28 rotor. The nuclei were resuspended in 1.4 ml of ice-cold Buffer A (20 mM Tris, pH 7.5, 150 mM NaCl, 2 mM EDTA), transferred to a 1.5-ml centrifuge tube and sedimented at 2000 rpm in a benchtop centrifuge for 30 sec. The pellet was kept and resuspended in Buffer A and sedimented similarly a second time. Nuclei from three livers were pooled for all time points. Fort this, the nuclei were resuspended in 1.2 ml per liver of Nuclear Lysis Buffer (50 mM Tris, pH 8.1, 10 mM EDTA, 1% SDS, 0.15 mM spermine, 0.5 mM spermidine, 0.5 mM DTT and 0.5 mM PMSF), and sonicated with a Bioruptor-Pico from Diagenode for 9 rounds of 30 seconds of sonication and 90 seconds off. Chromatin has been diluted with 4 volumes of Immunoprecipitation Dilution Buffer (20mM Tris-HCl pH 8.1, 150 mM NaCl, 2mM EDTA, 1% Triton X-

100, 0.01% SDS, 50 µg/ml PMSF, 1 µg/ml leupeptin, 0.15 mM, spermine, 0.5 mM spermidine). Pre-clearing of the chromatins was done by adding 100 µl of pre-immune serum, incubation for 1hr at 4°C on a rotating wheel and adding 200 µl of a homogeneous protein A-agarose suspension (100 µl bed volume). This was incubated for 3 hrs and centrifuged in a microfuge during 2 minutes at a speed of 3,000 rpm. The supernatant was collected and DNA content was quantified. The concentration of fragmented cross-linked chromatin was adjusted with a mix 1:4 of NLB:IPDB. 5% of human chromatin was added (HeLa spinner cells) as a control for ChIP-seq. Chromatins were incubated overnight at 4°C on a rotating wheel with the following antibodies: anti-RPB2 (Santa Cruz Biotechnology, sc-673-18), anti-H3K4me3 (Abcam, ab8580), and anti-H3K36me3 (Abcam, ab9050). For chromatin immunoprecipitation, 40 µl of protein A bead suspension were added to the chromatin and incubated for 3 hrs at 4°C on a rotating wheel. The beads were then washed twice with 1ml of IPWB1 (20 mM Tris-HCl pH 8.1, 50 mM NaCl, 2 mM EDTA, 1% Triton X-100, 0.1% SDS) and centrifuged 2 min at 3,000 rpm at 4°C. Later the beads were washed with 1 ml of IPWB2 (10 mM Tris-HCL pH 8.1, 250 mM LiCl, 1 mM EDTA, 1% NP-40, 1% sodium deoxycholate) and centrifuged for 2 minutes at 3,000 rpm at 4°C. And the beads were also washed twice with 1 ml of 1x TE pH 8.0 and centrifuged for 2 minutes at 3,000 rpm at 4°C. Protein–DNA complexes were eluted from the beads, de-cross-linked, and treated with RNase A and, subsequently, with proteinase K, as described [O'Geen et al., 2006]. The DNA concentration was determined by fluorometry on the Qubit system (Invitrogen). A total of 10 ng DNA were used for the library preparation.

## ChIP-seq library preparation and ultra-high-throughput sequencing

10 ng of immunoprecipitated material was used for sequencing library preparation. Total input was also prepared for sequencing at each time point and replicate. Paired-end sequencing libraries were prepared with the 'MicroPlex Library Preparation' kit, from Diagenode, following the instructions of the manufacturer (Diagenode, catalog no C05010011). No size selection was performed and the chromatin was amplified with 14 cycles of PCR. 100 nucleotides at the fragment ends were sequenced with paired-end sequencing technology from HiSeq 2100 (Illumina).

## RNA-seq library preparation and sequencing

Total RNA from individual livers was used for RNA-seq. Strand-specific libraries were prepared with the TruSeq Stranded mRNA Library Prep kit from Illumina. 100 nucleotides at the fragments ends were sequenced with single-end sequencing technology from HiSeq 2100 (Illumina).

## ChIP-seq data preparation and quantification

The extreme 50 bp of the sequenced paired-ends were mapped onto the mouse (mm9) and human genome (hg19) with Elandv2e. Only fragments whose two reads had good sequencing quality (i.e., quality of the first 25 bp higher than Q20, as specified by Illumina) and mapping on unique genomic locations were kept. For further analysis fragment sizes between 50 and 500 bp were taken and only one copy of redundant fragments was kept.

To quantify the density of epitopes in the mouse genome, the 50 bp in the center of the sequenced fragments were used. The Ensembl 67/NCBI 37 annotation of transcription units was used.

Pol2 quantifications were done in the promoter proximal region of +/- 250 bp around annotated TSS. H3K4me3 quantifications were done in the promoter regions of +/- 1 kb around TSS. Also Pol2 and H3K36me3 quantifications in the body of annotated transcription units were done on a region of 500 bp downstream of the TSS to 2 kb downstream the PolyA signal. The same quantifications were done on the Input libraries as a control. All quantifications were scaled to the number of fragments analyzed per library. 500 pseudo-counts (i.e., 10 fragments with 50 bp displayed) were added to the quantifications to stabilize the variance in low scores. Finally, the log2 ratios between ChIP and Input quantifications were calculated. Among the multiple transcription units associated to a given gene, the one containing the maximum Pol2 occupancy across time points in the promoter proximal region was used for analysis. All the handling of the data was done with the UNIX shell, Perl and the R software [R Core Team, 2013].

## RNA-seq data preparation and quantification

Sequenced reads with low quality were discarded (i.e., quality of the first 25 bp lower than Q20, as specified by Illumina). Adapters were removed with the Cutadapt tool [Martin, 2011]. Sequences shorter than 40 bp and sequences with low complexity were discarded with the tool 'prinseq-lite' [Schmieder and Edwards, 2011]. For the removal of low complexity sequences the method 'dust' was used. The selected reads were mapped by using Tophat 2.0.13 [Kim et al., 2013], which in turn reuses the Bowtie2 mapper [Langmead et al., 2012]. Reads were first mapped onto coding sequences annotated in the mm9 mouse annotation and, the non-mapped, were aligned onto the genome. The mapped reads were sorted by using samtools 0.1.19 [Li et al., 2009] and they were assigned to genes by using the 'featureCounts' function from the 'Rsubread' R package [Liao et al., 2013; R Core Team, 2013]. Reads falling within genes with strand-specificity were assigned to the genes. Those reads mapping on unique genomic positions were counted as 1, whereas those reads mapping onto multiple genomic

positions where assigned a score of 1 divided by the number of different positions they were mapped. Finally Reads Per Kilobase of transcript per Million mapped reads (RPKM) were calculated on a log2 scale and used for analysis. All the handling of the data was done with the UNIX shell, Perl and the R software [R Core Team, 2013].

## Identification of genes expressed and differentially expressed post-PH

The identification of expressed genes was assessed to detect the expression activities in the mouse liver. The identification of significant gene expression was done on each library by implementing the method proposed by [Hart et al., 2013]. The maximum threshold obtained across libraries was taken as final threshold (log2(RPKM)=-1.3). The final number of genes expressed at some library was 12,032. The analysis was done on the R software [R Core Team, 2013].

The genes detected as expressed were further classified into differentially expressed or not compared to the resting liver. For this analysis, RNA-seq replicate libraries for a differential expression analysis done with the edgeR library implemented on R [Robinson et al., 2010]. Tag-wise dispersion was estimated and gene-wise exact tests were computed to identify differences in means between time points, defined as negative binomial random variables [Robinson et al., 2008]. P-values were adjusted by using the Benjamini method. And very strict adjusted p-values were selected (lower than 0.0000001) to retain log2 fold-changes higher than 0.5 or lower than -0.5.

## Clustering of the transcriptome

Genes expressed in some sample were grouped based on their dynamics between 0 and 60 hrs post-PH. For this analysis, the log2(RPKM) quantifications per replicate samples were averaged. The grouping of genes was done with a hierarchical clustering strategy implemented in R [R Core Team, 2013]. The Ward's minimized variance criterion was used on Euclidean distances between pairs of samples. And the R library 'dendextend 0.17.1' was used to handle the resulting dendogram [Galili, 2014].

Partitioning Around Medoids (PAM) was done on the 5,502 genes displaying significant differential expression. For that analysis, the 1-Pearson correlation was used as distance measure. The log2(RPKM) quantifications were centered and scaled to obtain z-scores. The PAM clustering implemented in the 'cluster 1.14.4' library from R was used [Maechler et al., 2013]. For display of the clustering the heatmap.2 function from the 'ggplot 2.14.1' R library was used [Moeller et al., 2014]. For displaying, clusters were sorted based on the time when the

genes contained in the cluster showed an altered expression. The earlier the quantification alterations appeared across time points, the higher the clusters were positioned in the final heat map.

## Comparison of differentially expressed genes between post-PH and Sham samples

The log2(RPKM) quantifications across replicate samples were averaged for each gene classified as differentially expressed post-PH. Genes were sorted in the same way as for the PAM clustering (see above). A comparison was done at each time point where both Sham and post-PH samples of RNA were collected (1 hr, 4 hrs, 10 hrs, 20 hrs and 48 hrs). For the comparison log2 ratios were calculated between the post-PH and Sham averaged quantifications. Fold-changes on genes with scores lower than 0 at both conditions were discarded. Plots were created with the R statistical software [R Core Team, 2013].

## Display of gene expression profiles

Line plots were created with the profiles of mRNA and Pol2 log2 quantifications, with for the Pol2 quantifications both the proximal promoter and body of gene profiles are shown. Shaded lines or error bars are displayed through-out to describe the standard deviation across samples. All the plots were created using the 'ggplot2' library [Wickham, 2009] used in the R statistical software [R Core Team, 2013].

## Functional enrichment of gene lists and pathway annotation

The Webgestalt tool was used [Wang et al., 2013] where enrichments were done on the GeneOntology (GO) [The Gene Ontology Consortium, 2015], KEGG pathways [Kanehisa and Goto, 2000] and Wikipathways [Kutmon et al., 2016] databases.

The annotation of cell-cycle genes was obtained from the KEGG database (entry mmu04110). Entrez gene identifiers were converted to Ensembl identifiers by using the biomaRt tool in R [Durinck et al., 2005; Durinck et al., 2009] that in turn accessed the org.Mm.eg.db 2.10.1 database [Carlson].

## Genomic displays of ChIP-seq data

The CycliX viewer [Martin et al., unpublished] was used to visualize ChIP-seq tracks, especially AV2 files, which can be handled very efficiently on the viewer. Among other functions, the genomic viewer allows the selection of the desired genome stored in the NCBI database and the visualization of genomic data in a cumulative view. The density of accumulated data can also be scaled to the total number of fragments per library. And tracks can be displayed with any desired scale.

# Chapter V: Insights on the accumulation of H3K36me2 and H3K36me3 in the mammalian genome

The di- (H3K36me2) and tri- (H3K36me3) methylation of H3K36 have been observed to accumulate in transcribed genes. Previous studies show that H3K36me3 accumulates especially by the 3' end of the transcription unit. In yeast and chicken, H3K36me2 and H3K36me3 colocalize [Bannister et al., 2005; Pokholok et al., 2005; Rao et al., 2005], while in Drosophila H3K36me2 has been observed closer to the 5' end of genes [Bell et al., 2007], suggesting a different regulation of H3K36 di- and trimethylation across species and thus transcription. Further studies have observed a link between H3K36me3 and splicing regulation [Li et al., 2003; Krogan et al., 2003; Xiao et al., 2003; Kizer et al., 2005; Schaft et al., 2003]. Nevertheless, the mechanism of regulation is still not well understood, especially in mammals, and studies have only been done on specific genes. In this chapter I characterize the accumulation of di- and tri- methylation in the chromatin of the regenerating mouse liver, where many changes in gene expression take place. For that I have used the data described in Chapter IV, particularly the H3K36me3 ChIP-seq data. Also a H3K36me2 ChIP-seq library was prepared at the 60 hrs time point post-PH to allow the comparison between the two marks and the library was sequenced with multiplexing, in contrast to the H3K36me3 library that was sequenced in a whole lane.

## Results

### V.1 Characterization of the genome-wide accumulation of H3K36me2 and H3K36me3 in mouse liver chromatin

Consistent with what has been observed in previous studies, both H3K36me2 and H3K36me3 marks are observed in genic regions of mouse live chromatin (Figure V-1a). Both marks are not observed in similar positions within genes (Figure V-1b). In the regenerating mouse liver H3K36me2 is observed close to the 5' end of genes, as in Drosophila [Bell et al., 2007], while H3K36me3 is observed across the 3' end. Indeed, there tend to be a transition between H3K36me2 and H3K36me3 close to the 3' end of the first intron. This was previously observed for H3K36me3 [Huff et al., 2010] and thus it is believed that it has a link to the regulation of splicing [Kim et al., 2011]. In the 100 Mb genomic view and in the surroundings of genic regions it is also observable that dimethylation of H3K36 is present at intergenic regions as well [Figure V-1], although at lower levels, which suggests that this mark could also be linked to other processes outside of genes. H3K36me2 was previously observed during DNA repair events [Fnua et al., 2011]. Perhaps one of the functions for this mark

in intergenic regions during regeneration is DNA repair as during regeneration cells proliferate and the replicated chromosomes often need to be repaired during this process.

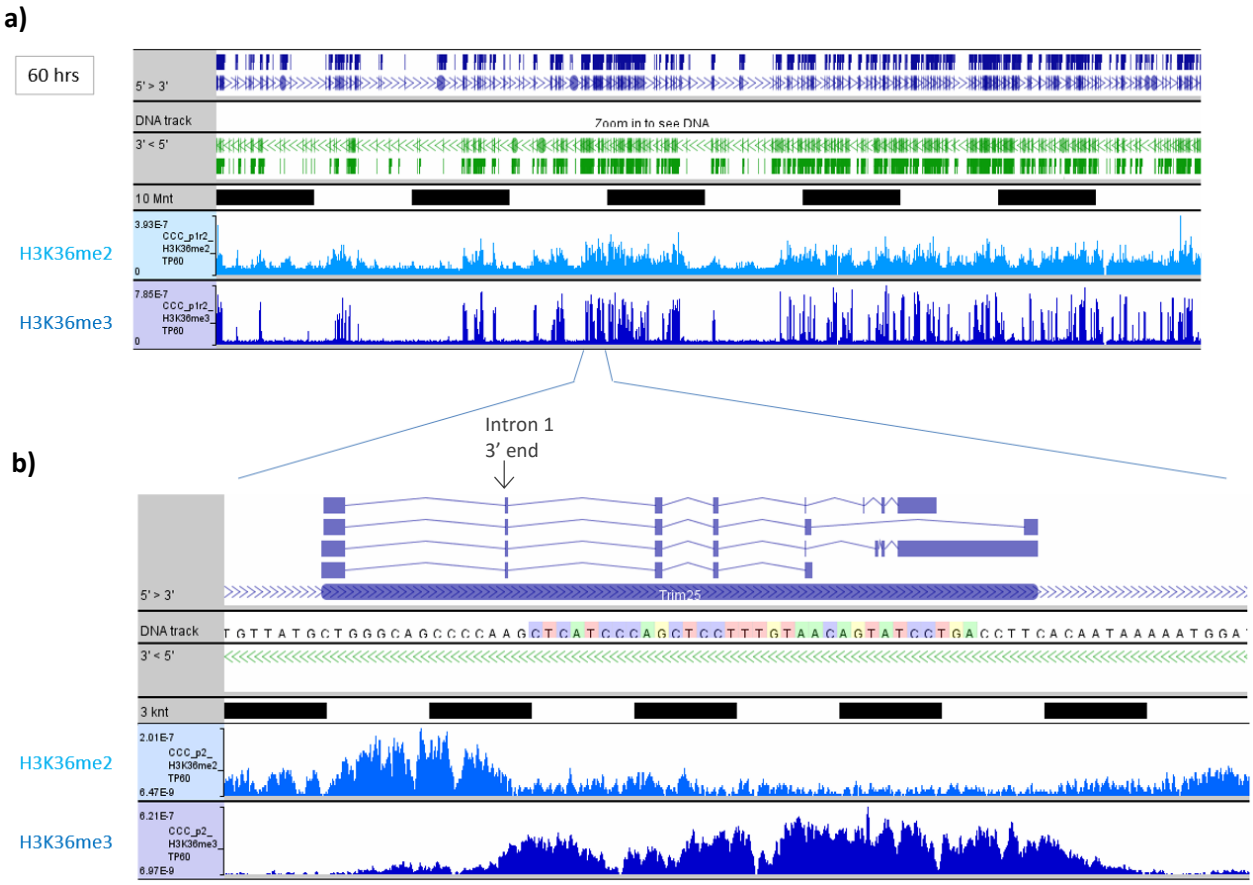



**Figure V-1**. ChIP-seq profiles of the H3K36me2 and H3K36me3 libraries at 60 hrs post-PH. a) 100 Mb section of Chromosome 11. b) Region of the *Trim25* gene on Chromosome 11. In both panels, the top part depicts the annotated genes in the positive and negative strands (blue and green boxes, respectively) and their intron/exon structures depicted with lines and boxes, respectively. Below, bars in black and white depict sections of the genomic region displayed and the size of each bar is indicated to the left. Two tracks containing ChIP-seq data are shown. The first ChIP-seq track displays the H3K36me2 ChIP-seq profiles and the second track displays data from the H3K36me3 library at 60 hrs post-PH. The profiles displayed have been created by the accumulation of the central 50 bp of the sequenced paired-end fragments with sizes between 50-500 bp.

To characterize the profiles of H3K36me2 and H3K36me3 genome-wide, I did an initial selection of 13,438 annotated genes that showed significant transcript levels. As the annotation of mouse genes includes diverse intron/exon structures that can be expressed in different tissues, for each selected gene I chose for further analyses the transcription unit with the highest Pol2 quantification on its promoter at 60 hrs post-PH (see Methods in Chapter IV). Second, I classified genes according to their H3K36me3 levels (high, low and zero) (see Methods) along their body (region of [500 bp after TSS, -500 bp before PAS]) as illustrated in Figure V-2. For further analyses I only kept those transcription units with high trimethylation levels, adding up to a total of 8,967 transcription units.

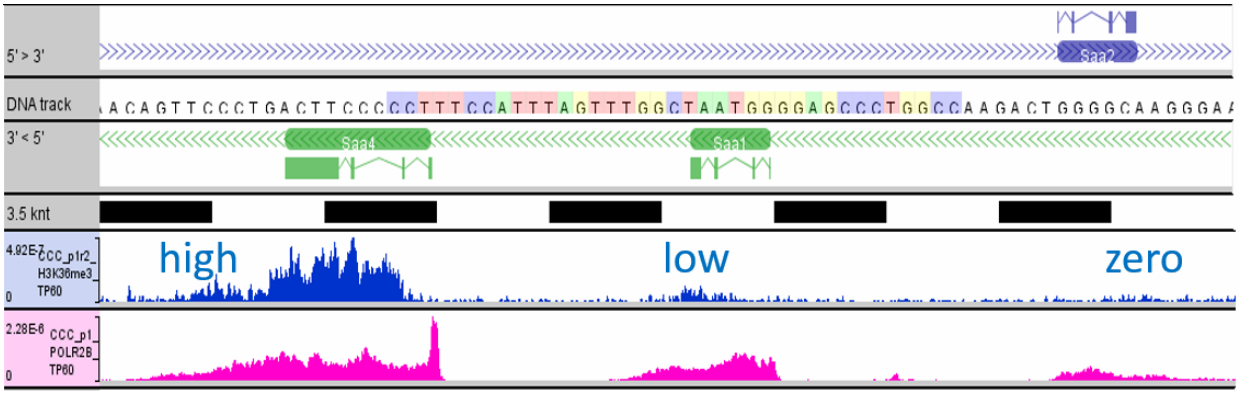



**Figure V-2**. ChIP-seq profiles of the H3K36me2 and H3K36me3 libraries at 60 hrs post-PH for the *Saa1*, *Saa2* and *Saa4* genes. The profiles displayed have been created by the accumulation of the central 50 bp of the sequenced paired-end fragments with sizes between 50-500 bp. These three genes, exemplify the classification of genes with high, low and zero levels of H3K36me3 accumulation. For further studies I took only the genes showing high levels of H3K36me3.

The first analysis aimed to identify transcripts with significant levels of H3K36me3 around the 3' end of the first intron and visualize how the accumulation of H3K36me3 distributes compared to H3K36me2 around the first splicing site. (Figure V-3). For that I selected the transcription units with at least one intron and that had significant H3K36me3 accumulation in a region of 2 kb around the 3' end of the first intron. This provided a list of 6,838 transcription units (75% of the transcription units with high H3K36me3 levels), suggesting that for most of the transcription units accumulating H3K36me3, this mark is already present around the 3' end of the 1st intron. Then I quantified at a nucleotide resolution the H3K36me2 and H3K36me3 accumulation in the region [-4, +1] kb around the 3' end of the second exon. At this point two data matrices were prepared: one with H3K36me2 quantifications and the other with H3K36me3 quantifications. Each row contained data for each transcription unit and the columns contained the absolute quantifications at each of the 5,000 quantified

nucleotides. To be able to see the relative changes of methylation across genes the quantifications were centered and scaled. In the resulting heatmap (Figure V-3), the relative quantifications within transcription units are represented by different colors in a gradient between blue (lowest signal), white (intermediate signal) and red (highest signal). Also, the transcription units were sorted by the length of their second exon. This sorting makes evident a transition from H3K36me2 to H3K36me3. Further, the position of the transition draws a shape similar to the length of the second exon (compare position of the transition in the panels for H3K36me2 and H3K36me3 with the panel to the right, which displays the distribution of sizes of the second exon). This suggests a link between the H3K36me2 to H3K36me3 transition and the 3' end of the first intron. Altogether, this shows that there is a tendency to have higher H3K36me2 levels upstream the 3' end of the first intron and an increase of H3K36me3 levels downstream.

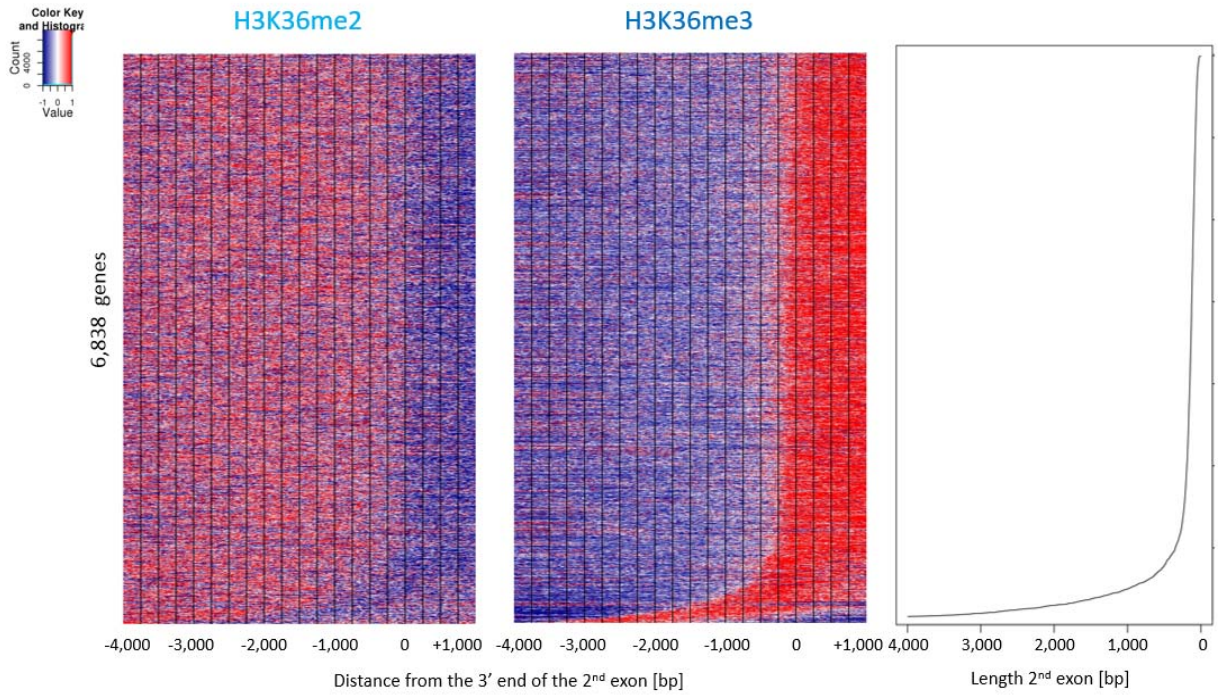



**Figure V-3**. Density heat maps of the accumulation of H3K36me2 (left) and H3K36me3 (middle) in a set of 6,838 transcription units with high levels of H3K36me3 along their selected transcription unit and the displayed region. The methylation densities are shown with a nucleotide resolution in the region [-4 kb, +1 kb] around the annotated 3' end of the second exon. The quantifications have been done on the accumulation of the central 50 bp of the sequenced paired-end fragments with sizes between 50-500 bp. Further, they have been centered and scaled so that the relative densities within the region are displayed for each gene, with red color assigned to the highest score, white to the middle score and blue to the lowest.

Finally the transcription units have been sorted by the length of their second exon, whose distribution is displayed in the right-most panel.

## V.2 In HeLa cells there is also transition between H3K36me2 and H3K36me3 by the 5' end of the second exon

To investigate further the H3K36me2 to H3K36me3 transition at the 3' end of the first intron, I examined HeLa cells chromatin. In the context of the CycliX project, a 5% of HeLa chromatin was added to the mouse chromatin as an internal spike control (see chapter IV for further details). The HeLa-specific reads from H3K36me2 and H3K36me3 ChIP-seq libraries at 60 hrs were taken and a similar analysis to the one shown in Figure V-3 was done. As a result, a clear increase of H3K36me3 (middle panel) is observed by the 3' end of the first intron. Also, a diffuse curve is observed in the H3K36me2 data (left panel) suggesting higher levels of H3K36me2 upstream of the 3' end of the first intron. Notably, however, the methylation density of H3K36me2 is very diffuse. The reason for this may be the lower number of reads in the H3K36me2 library, owing to the multiplexing done during sequencing, which was not done for the H3K36me3 library sequencing. Nevertheless, genomic views of gene cases clearly show a transition from di- to tri-methylation that is linked to the 5' end of the second exon (Figure V-5).

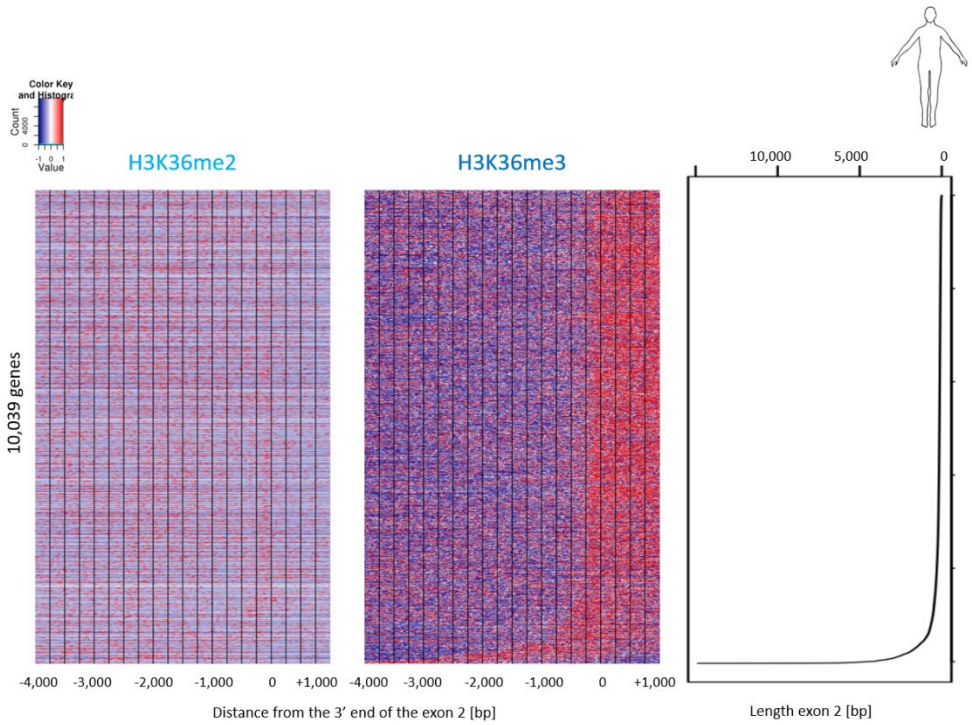


**Figure V-4**. Density heat maps of the accumulation of H3K36me2 (left) and H3K36me3 (middle) in a set of 10,039 HeLa cell transcription units with high levels of H3K36me3 along their body and the displayed region. The methylation densities are shown with a nucleotide resolution in the region [-4 kb, +1 kb] around the annotated 3' end of the second exon. The quantifications have been done on the accumulation of the central 50 bp of the sequenced paired-end fragments with sizes between 50-500 bp. Further, they have been centered and scaled so that the relative densities within the region are displayed for each gene, with red color assigned to the highest score, white to the middle scores and blue to the lowest. Finally the transcription units have been sorted by the length of their second exon, whose distribution is displayed in the right-most panel.

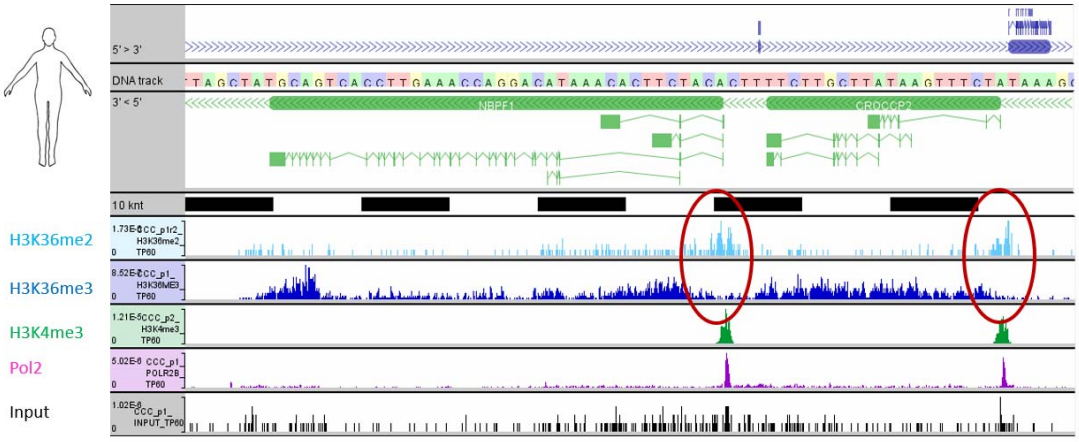



**Figure V-5**. ChIP-seq profiles of the human H3K36me2, H3K36me3, H3K4me3, Pol2 and Input libraries at 60 hrs post-PH for the candidate tumor suppressor in neuroblastoma *NBPF1* and the pseudogene *CROCCP2*. The ChIP-seq profiles displayed have been created by the accumulation of the central 50 bp of the sequenced paired-end fragments with sizes between 50-500 bp.

In order to compare the mouse liver and the HeLa chromatin I calculated the average H3K36me2 and H3K36me3 scores at a nucleotide resolution in a region of +/- 1 kb around the 3' end of the first intron and displayed them as lines (Figure V-6). In the mouse case, it is obvious that there is a decay of H3K36me2 and an increase of H3K36me3 around the second exon. There is also a slight decay of H3K36me2 in the HeLa chromatin and an increase of H3K36me3 around the 3' end of the first intron, although the transition in the HeLa chromatin is less obvious, but curiously there is a very prominent spike in the H3K36me2 signal at the

3’ end of the first intron. These results suggest that in the mouse and HeLa chromatin it tends to occur a transition from H3K36me2 to H3K36me3 in transcription units at the 3’ end of the first intron, similar to the observations in Drosophila.

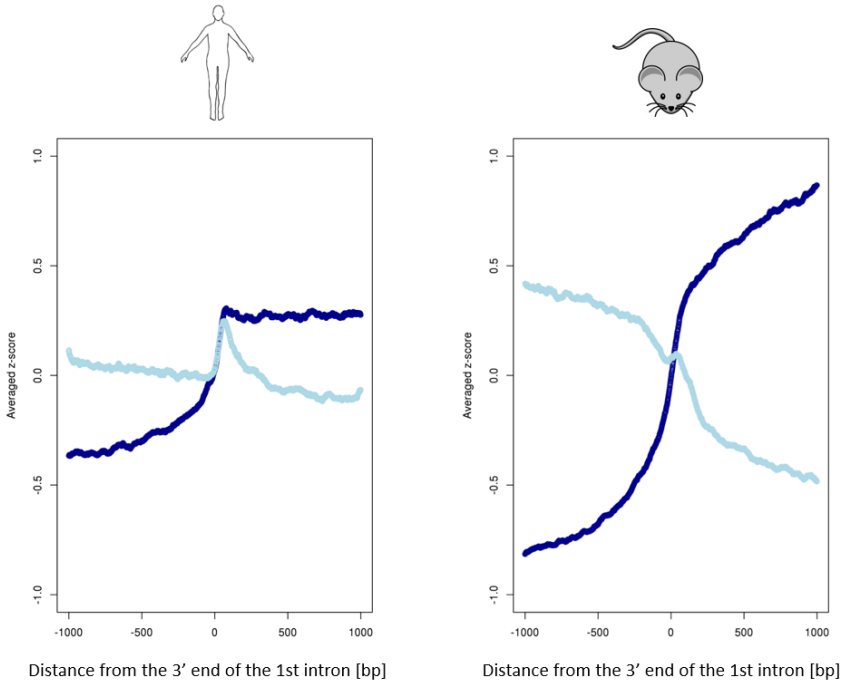



**Figure V-6**. Cumulative density profiles of H3K36me2 (light blue lines) and H3K36me3 (dark blue lines) around the 3’ end of the first intron, in the HeLa (left panel) and mouse liver (right panel) chromatins. The densities are displayed in a region of +/- 1 kb around the annotated 3’ end of the first intron. Further, they have been calculated at a nucleotide resolution from the accumulation of the central 50 bp of the sequenced paired-end fragments with sizes between 50-500 bp.

## V.3 Genes where multiple transcription units are expressed during regeneration do not exhibit obvious effects on the H3K36me2/3 accumulation

The abovementioned results are consistent with the hypothesis that H3K36me3 accumulation is associated with the appearance of the complete first intron during transcription which could be linked to splicing regulation [Huff et al., 2010; Kim et al., 2011]. To further test a possible link between H3K36me3 accumulation and splicing, I identified genes where different splicing events occur during regeneration. For this purpose the RNA-seq reads were mapped to transcription units. Gene cases were selected where different transcription

units were transcribed at different time points post-PH. In particular, the genes *ZFP281* and *SH3BP1* were inspected because their alternative transcription units display very different 5' end positions of their second exon (i.e., further than 500 bp). Later, visual inspection was done to investigate whether the H3K36me3 profiles were notably affected. Clearly, the profiles do not differ during regeneration (Figures V-7 and V-8).

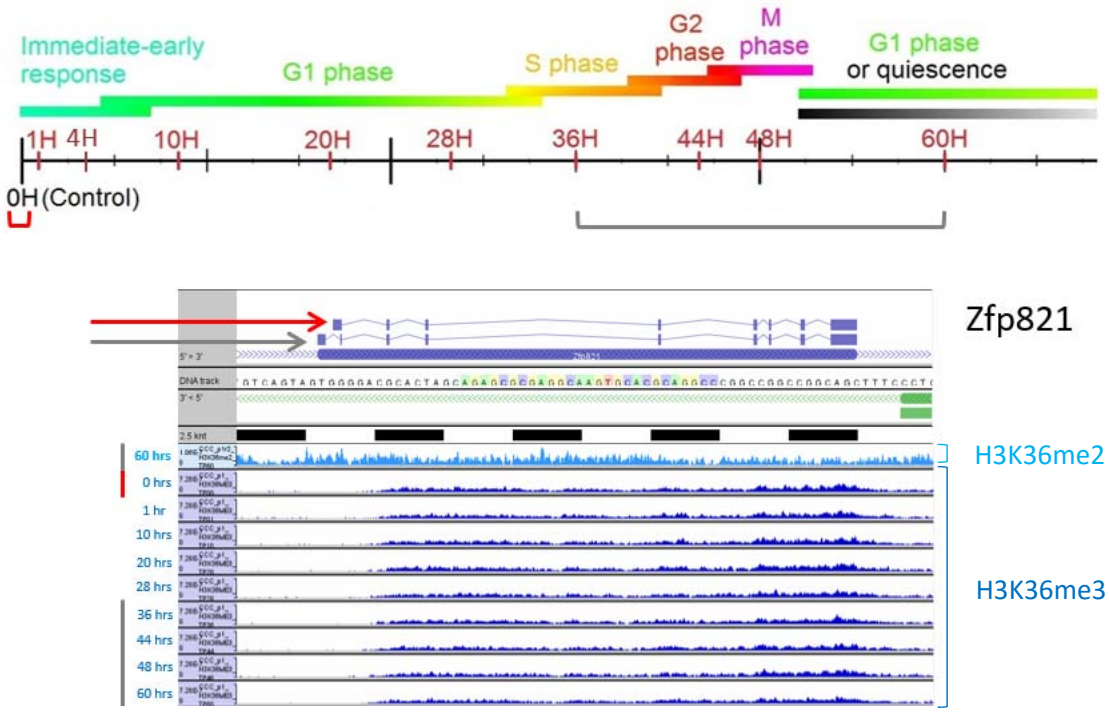



**Figure V-7**. ChIP-seq data for the *Zpf821* gene. The first track shows the profile of the H3K36me2 library at 60 hrs. The others contain the profiles from the H3K36me3 data from 0 hrs to 60 hrs post-PH. The ChIP-seq profiles displayed have been created by the accumulation of the central 50 bp of the sequenced paired-end fragments with sizes between 50-500 bp. Based on the transcript mappings of the RNA-seq data, two annotated transcripts from this gene are expressed at different time points. The red transcript is expressed only in the resting liver and the grey transcript is expressed between the time points at 36 and 60 hrs.

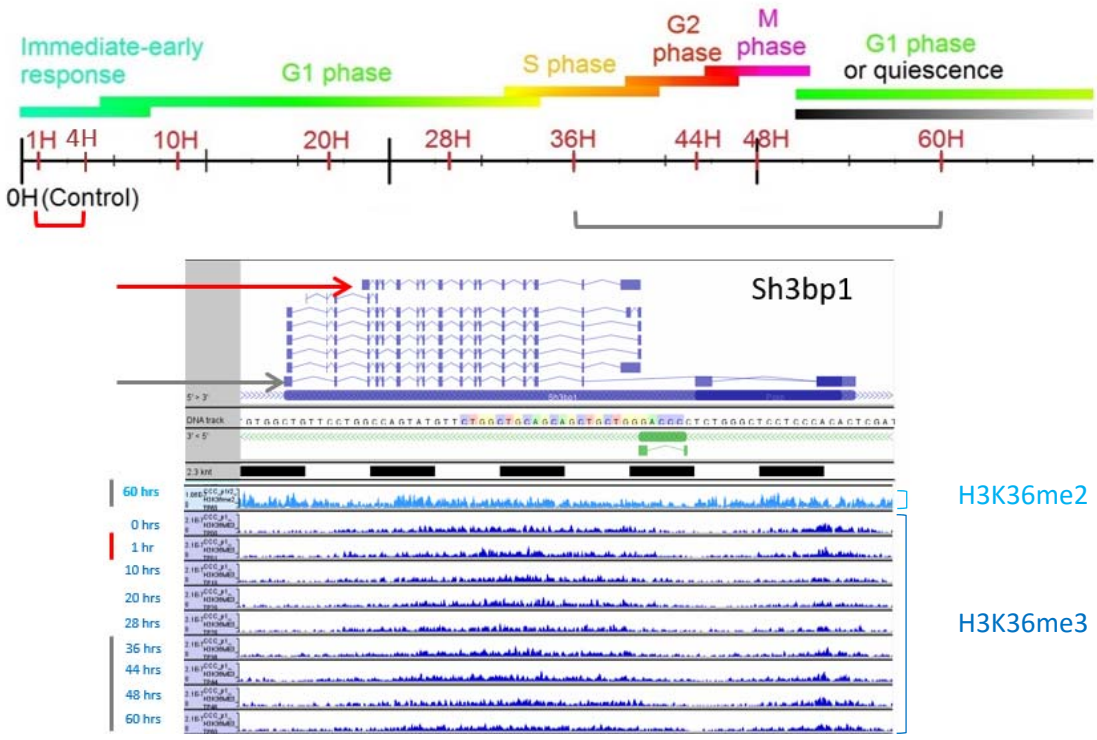



**Figure V-8**. ChIP-seq data for the *Sh3bp1* gene. The first track shows the profile of the H3K36me2 library at 60 hrs. The others contain the profiles from the H3K36me3 data from 0 hrs to 60 hrs post-PH. The ChIP-seq profiles displayed have been created by the accumulation of the central 50 bp of the sequenced paired-end fragments with sizes between 50-500 bp. Based on the transcript mappings of the RNA-seq data, two annotated transcripts from this gene are expressed at different time points. The red transcript is expressed only in the resting liver and the grey transcript is expressed between the time points at 36 and 60 hrs.

## V.4 H3K36me3 accumulation is also observed in murine genes where no splicing is required

Among the transcription units expressed in the mouse liver showing high H3K36me3, there are 164 cases that don't have introns. The product of these 164 transcription units are proteins in 81 cases, pseudogenes in 70 cases, 7 lncRNAs, 2 miRNAs, 2 snRNAs and 2 of other kinds of ncRNAs. This shows that H3K36me3 is also present in transcription units not requiring splicing events. I further looked into these cases of genes (Figures V-9 to V12). The first of the examples is the tRNA methyltransferase Trmt12 (Figure V-9). This gene is transcribed, as shown by the Pol2 occupancy along the gene and the resulting transcripts observed in the RNA-seq data. After the promoter of this gene there is an increased H3K36me2 accumulation that is followed by a

transition towards H3K36me3 by 1/3 of the body length of this gene. Curiously, the H3K36me2 and H3K36me3 transition occurs at a point in which RNA-seq tags decrease. Furthermore, in contrast to the majority of the cases I could see, the H3K36me3 accumulation is terminating earlier than the 3' end of the gene.

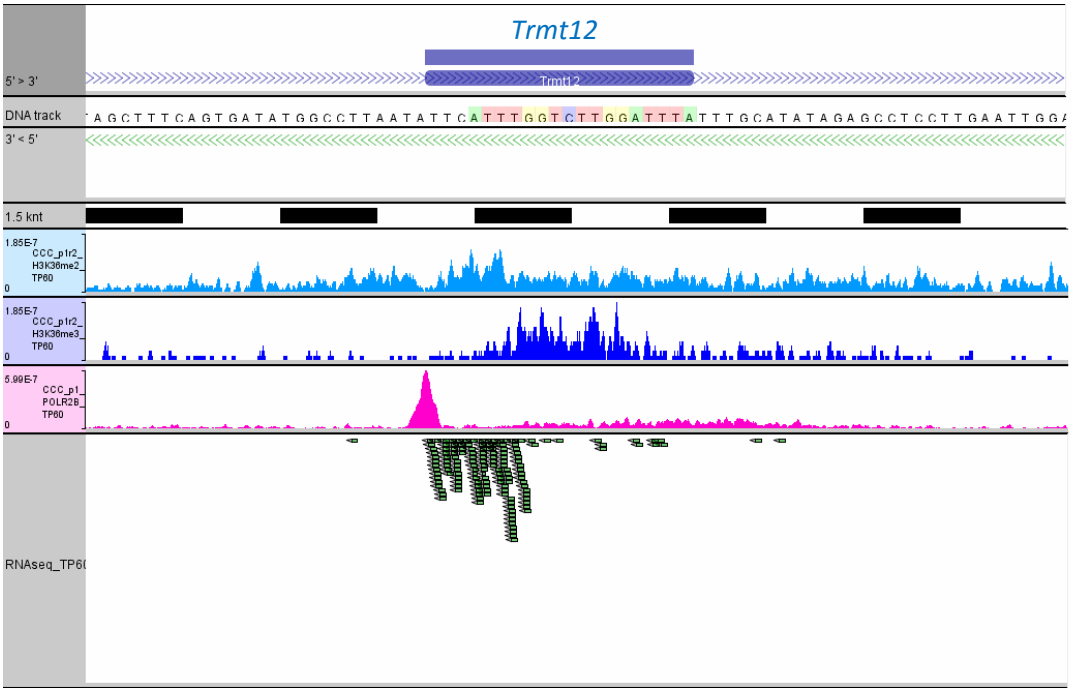



**Figure V-9**. ChIP-seq profiles of the H3K36me2 and H3K36me3 libraries and RNA-seq reads at 60 hrs post-PH for the *Trmt12* gene. The ChIP-seq profiles displayed have been created by the accumulation of the central 50 bp of the sequenced paired-end fragments with sizes between 50-500 bp. The RNA-seq profile displays the entire sequenced single-end and strand-specific reads.

The second example is the bidirectional promoter created by the Zinc Finger Protein gene *ZFP830*, also called *OMCG1*, and the *CCT6b* gene (Figure V-10). The encoded protein from the *ZFP830* gene has been shown to be required for mitosis progression of mouse intestinal progenitor and embryonic stem cells. Consistently, it seems to be transcribed at 60 hrs post-PH, during the second round of cell division, as Pol2 is present through its body. In this case, there is also a transition between H3K36me2 and H3K36me3 downstream of the promoter, and now the H3K36me3 accumulation terminates downstream of the 3' end of the gene.

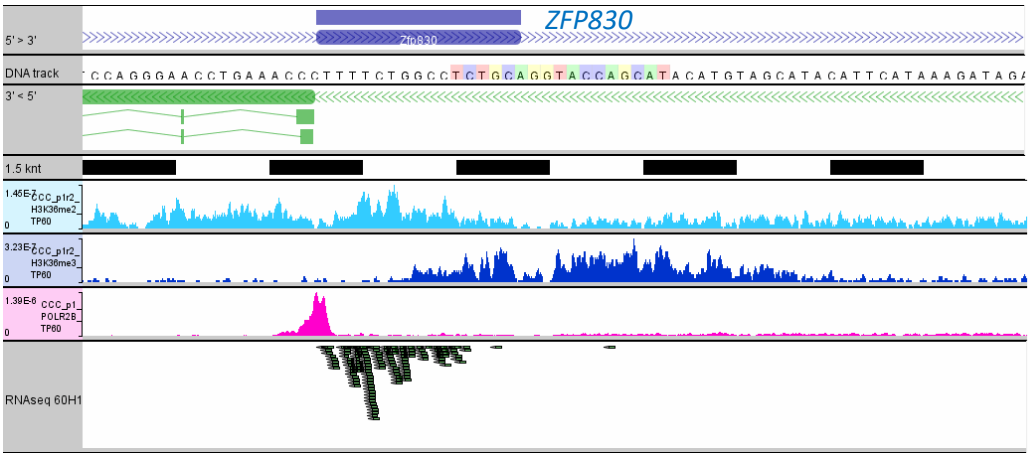



**Figure V-10**. ChIP-seq profiles of the H3K36me2 and H3K36me3 libraries and RNA-seq reads at 60 hrs post-PH for the bidirectional promoter composed of the *Zfp830* (blue) and the *CCT6b* (green) genes. The ChIP-seq profiles displayed have been created by the accumulation of the central 50 bp of the sequenced paired-end fragments with sizes between 50-500 bp. The RNA-seq profile displays the entire sequenced single-end and strand-specific reads.

The third case is the CEBPa transcription factor (Figure V-11). The encoded protein functions in homodimers and heterodimers with CEBPb and CEBPg which modulate the expression of cell cycle regulators. This example shows a different profile compared to the previous cases. There is no transition between H3K36me2 and H3K36me3. Instead, there is a rapidly appearing accumulation of H3K36me3, which again terminates past the 3' end of the gene, which interestingly coincides with a slightly increased dimethylation towards the 3' end, as it was observed in yeast and chicken [Bannister et al., 2005; Pokholok et al., 2005; Rao et al., 2005]. This alternative pattern is apparently still conserved in some genes.

The last example I show is the case of the *miRNA122* gene (Figure V-12). This is a very abundant miRNA that is specifically expressed in the liver of vertebrates. Its transcription is regulated by the liver-specific HNF4a transcription factor [Li et al., 2011] and its activity is partly controlled by the circadian Rev-ErbA alpha [Gatfield et al., 2009]. According to its high abundance in cells this gene is one of the genes accumulating the highest density of the Pol2. The transcription of miRNA is typically done by Pol2 which often binds to a promoter near the DNA sequence associated to the mature miRNA that produces the pre-miRNA. The pre-miRNA is later exported to the cytoplasm and cleaved by the RNAse III enzyme Dicer. In the case of the *miRNA122* gene, the final product is the annotated gene depicted in Figure V-12. There is an evident transition between H3K36me2 and H3K36me3 by the promoter of the transcribed DNA, coinciding with the beginning of Pol2 occupancy and RNA-seq tags, and H3K36me3 decreases by the region of the final miRNA product, even

though Pol2 and RNA-seq tags are present downstream. These profiles, suggest a special regulation of H3K36me2 and H3K36me3 deposition in this miRNA.

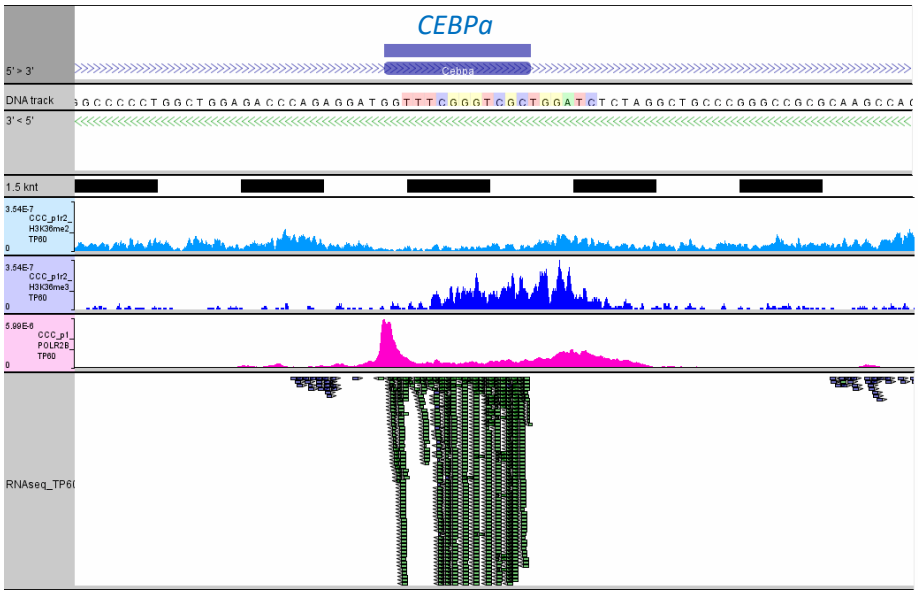



**Figure V-11**. ChIP-seq profiles of the H3K36me2 and H3K36me3 libraries and RNA-seq reads at 60 hrs post-PH for the *CEBPa* gene. The ChIP-seq profiles displayed have been created by the accumulation of the central 50 bp of the sequenced paired-end fragments with sizes between 50-500 bp. The RNA-seq profile displays the entire sequenced single-end and strand-specific reads.

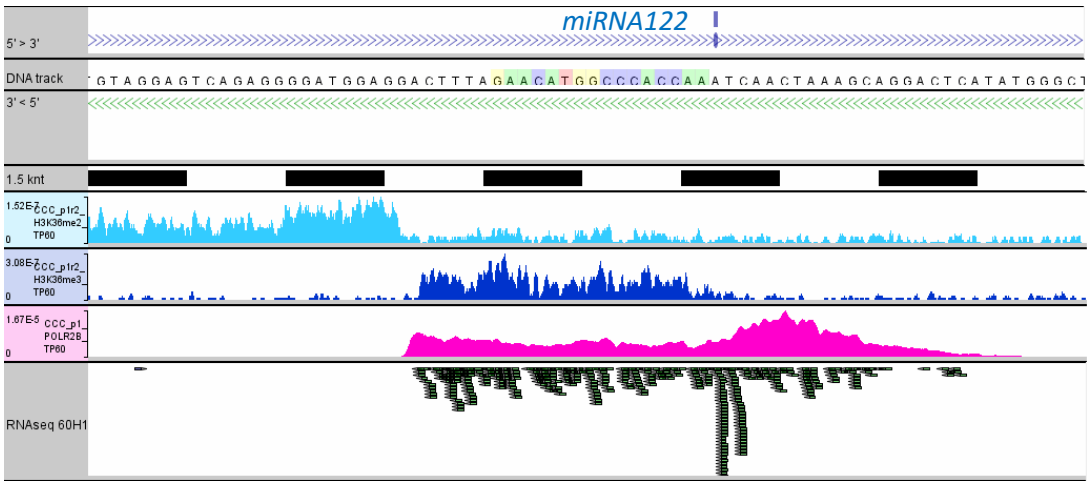



**Figure V-12**. ChIP-seq profiles of the H3K36me2 and H3K36me3 libraries and RNA-seq reads at 60 hrs post-PH for the *miRNA122* gene. The ChIP-seq profiles displayed have been

created by the accumulation of the central 50 bp of the sequenced paired-end fragments with sizes between 50-500 bp. The RNA-seq profile displays the entire sequenced single-end and strand-specific reads. The figure displays that Pol2 transcribes a long gene (non-annotated) that includes the DNA sequence that will give rise to the mature miRNA (depicted with a blue box on the genes track). This is further supported by the RNA-seq tags that extend in a similar region as Pol2 occupancy. The pre-miRNA is exported to the cytoplasm and cleaved by the RNAse III enzyme Dicer. Consistent with the maturation of the transcript, more RNA-seq reads are observed in the region originating the mature miRNA.

All these four examples show an accumulation of H3K36me3 without evident splicing. This indicates that the H3K36me3 methyltransferase Set2 may still be attached to Pol2 even if there are no required splicing events in the gene expression. Interestingly, different profiles of H3K36me2 and H3K36me3 methylation are observed, which could be associated to different regulations of transcription.

# Discussion

## In a cancerous human cell line and the healthy mouse liver, H3K36me2 accumulates close to the 5' end of transcribed genes, as in Drosophila

In yeast and chicken, H3K36 methylation resides promoter distal at transcribed regions [Bannister et al., 2005; Pokholok et al., 2005; Rao et al., 2005]. In contrast, in this work I have observed similarities in the genic position where H3K36me2 accumulates in the healthy mouse liver, the cancerous HeLa cells and the previously reported Drosophila chromatin [Bell et al., 2007]. Thus, perhaps similarities exist in the regulation of transcript elongation in those species. Curiously, evolutionarily speaking, the chicken genome is closer to the human and murine ones than the Drosophila genome, thus one could expect also more similarities on the cellular mechanisms. But it has to be noted that in [Bannister et al., 2005] where it is concluded that in chicken erythrocytes the distribution of H3K36me2 coincides with H3K36me3, the analysis was only done on two gene cases. Perhaps a genome-wide study on those cells would provide a different view.

In Drosophila, only two enzymes have been identified that methylate H3K36, and interestingly they split tasks in the promoter and the body of genes, as one is responsible for the dimethylation observed close to the 5' end and the other is responsible for the trimethylation observed towards the 3' end of genes [Bell et al., 2007]. For human and mouse different H3K36 dimethylases may be involved in the addition of a second methyl group in different genic regions.

As shown in [Bell et al., 2007], in Drosophila, reduction of H3K36me3 is lethal and leads to increased acetylation levels, specifically at lysine 16 of histone H4 (H4K16ac), a mark typically associated with chromatin compaction [Corona et al., 2002; Dorigo et al., 2003; Shogren-Knaak et al., 2006]. In contrast, reduction of both H3K36me2 and H3K36me3 decreases H4K16ac. Thus, it showed that di- and trimethylation of H3K36 have opposite effects on H4K16 acetylation, and it was proposed that these differences enable dynamic changes in chromatin compaction during transcript elongation, through different pathways. Further studies on this in mammalian genomes could give insights about the regulation of elongation.

## About the link between H3K36me3 and splicing

There is a general belief that H3K36me3 accumulation is related to splicing events that might occur co-transcriptionally [Huff et al., 2010; Kim et al., 2011]. Consistent with this hypothesis I could observe an increase

of H3K36me3 by the 3' end of the first intron in 80% of the expressed genes having H3K36me3 accumulation through their body. A novelty in this chapter is that 20% of the genes do not follow this pattern apparently for two reasons. First, some of the genes, apparently displaying alternative splicing at very separate times of the liver regeneration process, do not display any obvious change in the accumulation of H3K36me3, such as a modified position of the increase of the accumulation (Figures V-7-8). Perhaps, in these cases there is indeed a modified accumulation, but it is just not possible to discern it as less cells proliferate at 60 hrs compared to the early time points. Thus, less new accumulation might take place. The second case of genes where no link is observed between H3K36me3 and splicing is indeed those genes where no splicing is known to occur. Even though they accumulate H3K36me3, they do not display splicing events during their expression (Figures V-9-12). Presumably, in these genes, for unknown reasons, the H3K36me3 mark may not be read by effector proteins involved in splicing. Nevertheless, it is very interesting to see that they still accumulate H3K36me3. Perhaps, evolutionarily speaking, these genes are prepared for accommodating splicing complexes when other conditions are fulfilled, such as specific DNA sequences typically observed in splicing sites.

# Methods

## Animals and partial hepatectomies

The details about the animals used and the partial hepatectomies are the same as described in Chapter IV.

## ChIP, ChIP-seq library preparation and ultra-high-throughput sequencing

The details are the same as described in Chapter IV.

For this chapter it has to be added that a ChIP was done against H3K36me2 on mouse liver chromatin at 60 hrs post-PH.

## RNA, RNA-seq library preparation and ultra-high-throughput sequencing

The details are the same as described in Chapter IV.

## ChIP-seq data preparation and quantification

The details are the same as described in Chapter IV.

Additionally, for the work described in this Chapter V, quantifications on intergenic bins of 200 bp were done. Bins of 200 bp were defined by sliding 100 bp at a time on intergenic regions further than 5 kb of any annotated gene. The regions smaller of 200 bp at the end of each intergenic region was not used for analysis. The bins were defined by using Perl.

Also, in contrast to the studies shown in Chapter IV, among the multiple transcription units associated to a given gene, the one containing the maximum Pol2 occupancy at 60 hrs in the promoter proximal region was used for analysis.

## RNA-seq data preparation, quantification of transcript isoforms and identification of expressed isoforms

Sequenced reads with low quality were discarded (i.e., quality of the first 25 bp higher than Q20, as specified by Illumina). Adapters were removed with the Cutadapt tool [Martin, 2011]. Some sequences were discarded with the tool 'prinseq-lite' [Schmieder and Edwards, 2011]: sequences shorter than 40 bp and sequences with low sequence complexity. For the last case the method 'dust' was used. The selected reads were mapped onto the transcriptome annotation by using RSEM 1.2.19 [Li., 2011]. The mm9 version of the transcriptome was used for the mapping. An isoform counts table was retrieved and used for further analyses. All the handling of the data was done with the UNIX shell. The quantified counts per isoform were converted to Counts Per Million (CPM). Those isoforms with CPM higher than 1 were selected as expressed.

## Identification of genes with high, low or zero levels of H3K36me3 accumulation

For the identification of H3K36me3 accumulation a background distribution was defined to compare to quantifications in genic regions. The background distribution was defined by log2(ChIP/Input) quantifications on bins of 200 bp. The background distribution and the distribution of genic quantifications of H3K36me3 were compared with a two-sided statistical test. The p-values were adjusted by the Benjamini method. Genic regions with adjusted p-values lower than 0.05 were selected. The non-selected genes were classified as having 'zero' level of accumulation. Among the selected genes, the ones displaying a log2(ChIP/Input) ratio higher than 0.5 were classified as having 'high' accumulation. Otherwise, genes were classified as having 'low' levels of H3K36me3 accumulation. All the calculations were done with the R statistical software [R Core Team, 2013].

## Density profiles of methylation accumulation on multiple genes at a one nucleotide resolution

Nucleotide positions around the TSS of selected transcripts were taken. All the genomic regions taken were aligned together from 5' to 3'. All the handling of genomic regions was done on the UNIX shell.

On the nucleotide positions, densities of the central 50 bp of sequenced fragments were quantified at the previously defined nucleotide positions. A matrix of quantifications per Transcription Unit (TU) was built and analyzed on the R statistical software [R Core Team, 2013]. Z-scores were calculated from the quantifications

and displayed with the heatmap.2 function included on the 'gplots 2.14.1' [Warnes et al., 2014]. Further, the z-scores at each TU were averaged and displayed as lines with the R software.

## Density profiles of methylation accumulation on exons and introns scaled to 1 kb

Selected transcription units with at least two intronic regions were analyzed. Nucleotide positions from 1kb upstream the selected TSS's to the 3' end of the second intron were put together from 5' to 3'. All the handling of genomic regions was done on the UNIX shell.

On the nucleotide positions, densities of the central 50 bp of sequenced fragments were quantified at the previously defined nucleotide positions. A matrix of quantifications per TU was built and analyzed on the R statistical software [R Core Team, 2013]. All the introns and exons in that region were scaled to a size of 1 kb. This means that if either an intron or an exon is bigger than 1 kb, the regions were divided into 1,000 sub-regions (leaving the non-integer remaining to the last region) and to each new sub-region the average quantification among the respective nucleotide positions was assigned. Instead, if either an exon or intron was shorter than 1 kb, the individual nucleotide positions were duplicated with the same quantifications to extend to 1 kb. For this, new extended positions were calculated by dividing 1,000 by the length of the region. The non-integer part of the division was assigned to the first new extended positions one by one until using them all. Z-scores were calculated from the quantifications and displayed with the heatmap.2 function included on the 'gplots 2.14.1' R package [Warnes et al., 2014].

## Genomic displays of ChIP-seq and RNA-seq data

The CycliX viewer [Martin et al., unpublished] was used to visualize ChIP-seq and RNA-seq tracks, especially AV2 files that can be handled very efficiently on the viewer. Among other functions, the genomic viewer allows the selection of the desired genome stored in the NCBI database and the visualization of genomic data in a cumulative view or in a tag-wise view. The density of accumulated data can also be scaled to the total number of fragments per library, and tracks can be displayed with the desired scale.

# Chapter VI: The role of HCF-1 in the mouse liver chromatin

In human cells, HCF-1 is proteolytically cleaved to produce HCF-$1_N$ and HCF-$1_C$ subunits that associate non-covalently and thus become fully functional [Wilson et al., 1993a; Wilson et al., 1995b; Kristie et al., 1995]. Mature HCF-1 interacts with a plethora of proteins to co-regulate transcription (Figure I-16). This makes HCF-1 a very interesting protein to understand different mechanisms for transcription regulation. Yet, little is known about the role of this highly conserved protein in the chromatin of a healthy organism. The mouse liver confers an excellent system to study the chromatin associations of HCF-1 and its role there. In this chapter, I discuss where in the genome HCF-1 binds under conditions that induce gene expression changes, paying special attention to the chromatin association of each subunit.

## Results

### VI.1 As in HeLa cells, the murine HCF-1 is proteolytically cleaved producing the HCF-$1_N$ and HCF-$1_C$ subunits that associate with each other

In HeLa cells, the co-transcriptional factor HCF-1 is a family of polypeptides generated from a large full length precursor protein [Wilson et al., 1993a]. [Wilson et al., 1993a; Wilson et al., 1993b] discovered the existence of multiple HCF-1 peptides ranging in a mass of 110 kD to 300 kD (Figure I-14). Currently, different widely used antibodies against HCF-1 exist that can target both human and murine HCF-1 and can be used to evaluate the presence of HCF-1 peptides. I used several of the existing HCF-1 antibodies to study the presence of different polypeptides, at first in whole HeLa cell extracts, in order to compare with the previously published results. The profile of immunoblotted whole HeLa cell lysates (Figure VI-1) is similar to the bands of HCF-1 obtained after purification of nuclear HCF-1 (Figure I-14) [Wilson et al., 1993a]. I used antibodies targeting specific residues pertaining to the HCF-$1_N$ and HCF-$1_C$ subunits. For the HCF-$1_N$ subunit I used the N961 antibody [Machida et al., 2009], targeting residues 967 to 1011. On the HCF-$1_C$ subunit, I could target residues in the acidic region by using the in-house H12 antibody [Wilson et al., 1993a], that was also used in [Michaud et al., 2013] and was used for the experiments in Chapter III. Further, as there is few H12 antibody left, I used an alternative antibody to target the HCF-$1_C$ subunit that recognizes a similar peptide sequence as the H12 antibody that was produced by Bethyl Laboratories. In Figure VI-1, whole protein extracts from HeLa cells were immunoblotted using the N961 antibody and the antibodies targeting HCF-$1_C$ (Bethyl and H12). The N961 identifies very clearly HCF-1 peptides with atomic masses similar to the peptides of 116, 125, 127, 150 and 300 kD. The C-terminal-related lanes recover very similar bands with similar intensities associated to the

peptides of 110, 125, 127, 150 kD and slightly some of full length HCF-1 at 300 kD. If I focus only on the bands corresponding to the polypeptides reported in [Wilson et al., 1993a] it can be noted that the N961 antibody misses the 110 kD-like peptide, whereas the HCF-$1_C$ antibodies can't detect the 116 kD-like peptide. Notably, within all the lanes new bands appear with sizes lower than 100 kD. They could originate from degradation products, non-specific binding or new peptides that were not previously observed. Furthermore, a difference stands between the work in [Wilson et al., 1993a] and these experiments. In the first case, only nuclear HCF-1 was analyzed whereas these new results have been obtained from whole cell lysates. Perhaps the new bands have a cytosolic origin.

The murine HCF-1 sequence is 98% similar to the human protein and is also cleaved by their HCF-$1_{PRO}$ repeats, producing a distribution of HCF-1 peptides similar to HeLa cells [Kristie, 1997]. Nevertheless, it is not known whether they associate. To answer this question, I made denaturing immunoblots comparing HeLa and mouse liver cell lysates when adding different antibodies: Bethyl and N961 (Input lanes of the membranes shown in Figure VI-2). Input lanes from mouse liver cell lysates show that the murine HCF-1 is also proteolytically cleaved and is a family of peptides that constitute the HCF-$1_C$ and HCF-$1_N$ subunits. Among the murine peptides some have similar sizes to the ones found in the HeLa Input lanes and in [Wilson et al., 1993a] between 110 and 300 kD. But a few differences arise in the bands associated to smaller masses. Perhaps due to different degradation mechanisms or different cytosolic peptides.

Furthermore, the human HCF-1 amino- and carboxy-terminal subunits associate non-covalently [Wilson et al., 1993a; Wilson et al., 1995b; Kristie et al., 1995]. As the murine HCF-1 is also proteolytically cleaved, the produced subunits might also associate. To investigate whether both subunits associate, I did Immunoprecipitations (IP) in HeLa and mouse liver cell extracts by using antibodies targeting the HCF-$1_N$ subunit (N961 and N13 antibodies) and the HCF-$1_C$ subunit (H12 and Bethyl antibodies). Then, I electrophoresed the IPed and Input materials in two SDS-polyacrylamide gels, and I blotted each gel with one of the Bethyl and N961 antibodies (Figure VI-2, IP lanes). HCF-1-specific bands are visible in both human and mouse lanes. Further, HCF-1-specific bands are observed when blotting a specific subunit, regardless of the subunit that was immunoprecipitated, suggesting that the mouse HCF-$1_C$ and HCF-$1_N$ subunits associate. Nevertheless, a negative IP control should be done to further confirm that the bands are not background.

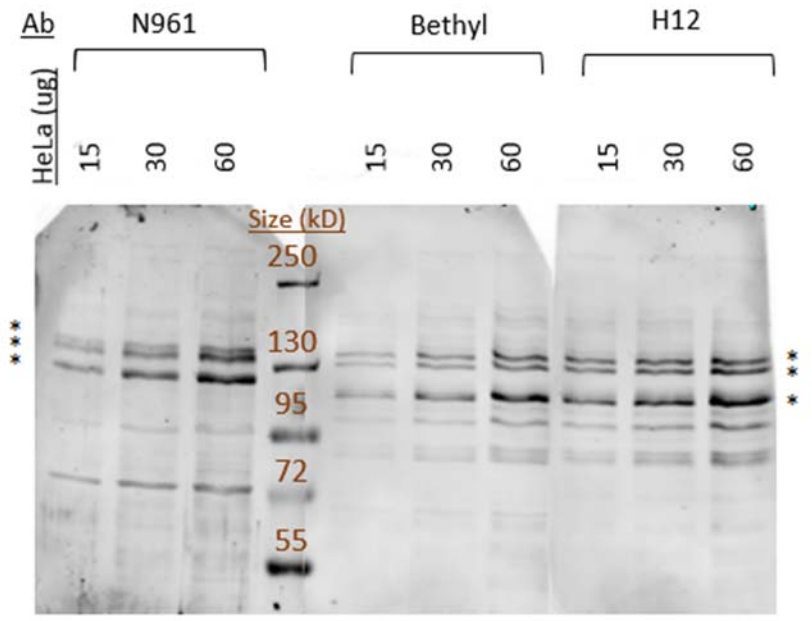



**Figure VI-1**. Probing HCF-1 family of polypeptides in HeLa cells with multiple HCF-$1_N$ and HCF-$1_C$ antibodies. Different amounts of whole HeLa protein extracts (15ug, 30ug, and 60ug) were electrophoresed for 1 hr in a denatured 7.5% SDS-polyacrylamide gel and HCF-1 was immunoblotted by using different primary antibodies: N961, to target HCF-$1_N$, and Bethyl and H12 antibodies, to target HCF-$1_C$. *, HCF-1 fragments.

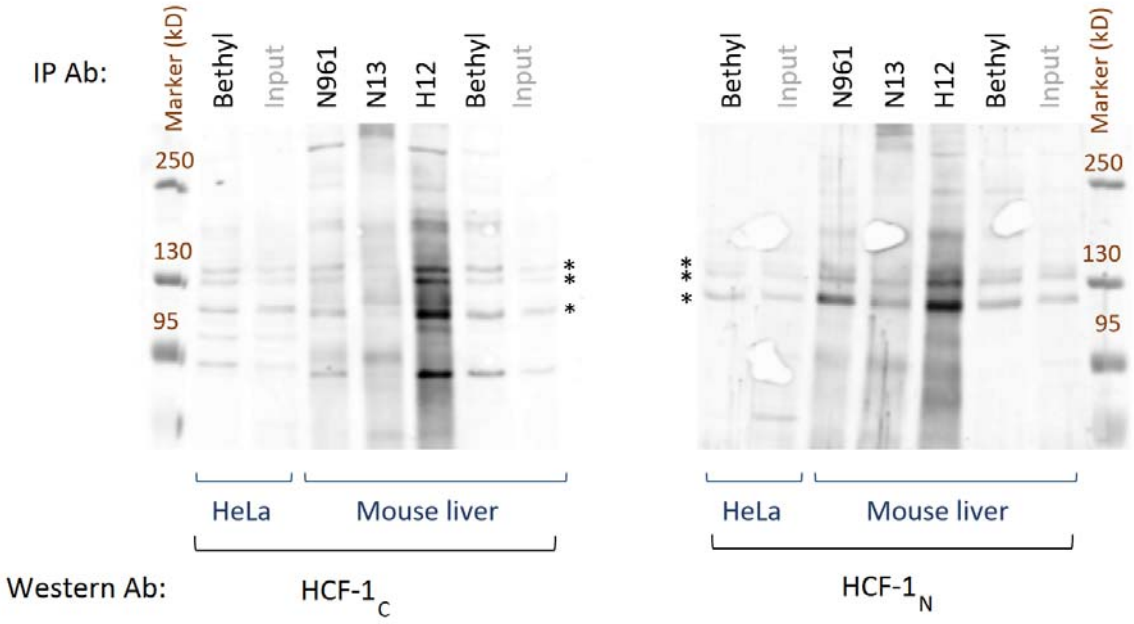


**Figure VI-2**. The HCF-1 family of polypeptides from HeLa and mouse liver cells associate. IPs were done against HCF-1 in HeLa and mouse liver cells by using the antibodies targeting the HCF-$1_N$ subunit (N961 and N13) and HCF-$1_C$ subunit (H12 and Bethyl). Two denaturing 7.5% SDS-polyacrylamide gels were electrophoresed during 1.5 hrs after loading IP and input materials from HeLa (lanes 1 and 2 in each gel) and mouse liver (lanes 3 to 7 in each gel) cell extracts. Left, immunoblot with Bethyl anti-HCF-$1_C$ antibody. Right, immunoblot with N961 anti-HCF-$1_N$ antibody. *, HCF-1 fragments.

## VI.2 Murine HCF-1 binds to the chromatin

Wysocka et al (2001) demonstrated that HCF-1 binds naturally to the chromatin of uninfected human cells. In the mouse liver it is unknown whether HCF-1 can bind to the chromatin. To investigate this open question ChIPs against HCF-1 were performed and followed by high-throughput sequencing. For that purpose, liver chromatin samples were prepared from mice under specific food and light conditions (with food access every 12 hrs and light only during the hours when no food was available, which is when they tend to sleep). Livers were collected at ZT02 (two hrs after food removal). As a control for the ChIP-seq, a 5% human spike control from HeLa cells was added to the mouse chromatin prior to ChIP. The genomic sites displaying significant ChIP-ed fragments over Input in the human chromatin can be compared with the genomic binding sites observed in a previous study by [Michaud et al., 2013]. A match between both datasets would indicate a good ChIP-seq on the HeLa but also on the mouse liver chromatins. After mixing the chromatins a ChIP was done, and a paired-end sequencing library was prepared and sequenced. The retrieved data were mapped onto the human and mouse genomes. Figure VI-3, shows the numbers of fragments with good quality mapping onto the Human and Mouse genomes. Input data from 0 hr livers already described in Chapter IV were also used to support the detection of significant ChIP fragments over Input throughout the genome. Fragments with good sequencing quality, mapping to unique positions of the mouse or human genomes and with sizes between 50 and 500 bp were taken for analysis. Among the selected fragments, there was a very small percentage mapping on both genomes (0.5%) and these fragments were assigned only to the mouse dataset. Low redundancy was obtained and only one-copy of redundant fragments was used for analysis.

To compare the human spikes data and the data from Michaud et al. (2013), I defined genomic bins of 200 bp sliding by 100 bp and I quantified the density of central 50 bp of the sequenced fragments in the human spikes. Approximately 90% of the bins with a log2 ratio ChIP over Input higher than 1 fall within proximal promoter regions. This analysis identifies approximately 5,000 genes bound by HCF-1 in the HeLa spikes control while about 5,400 genes were found to be bound by HCF-1 in Michaud et al. (2013). Figure VI-4, displays an example

of a genomic region in the human Chromosome 16 where gene promoters are bound by HCF-1, and they do it similarly in the spikes control data and the data from Michaud et al. (2013). This match between HeLa datasets suggests that the mouse-human spike ChIP-seq experiment worked properly for both the human and mouse chromatins.

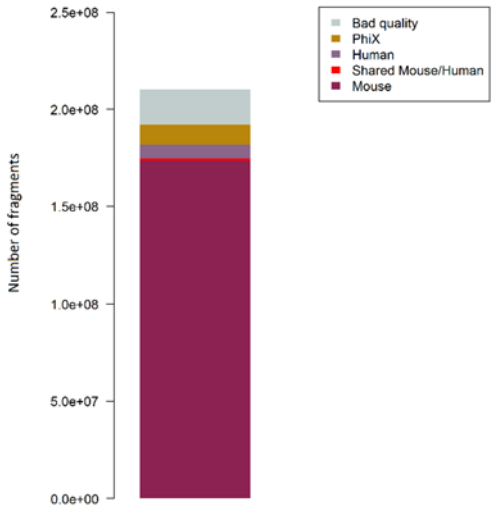



**Figure VI-3**. Number of fragments obtained after mapping the ChIP-seq reads onto the human and mouse genomes (hg19 and mm9 versions, respectively). The numbers of fragments are indicating, specifying whether their reads were of bad quality (grey) (see Methods sections in this chapter), mapping onto the PhiX genome (yellow) (5% of PhiX sample was added by the sequencing facility), mapping only onto the Human genome (light purple), mapping only onto the Mouse genome (violet) or mapping onto both the Human and Mouse genomes (red).

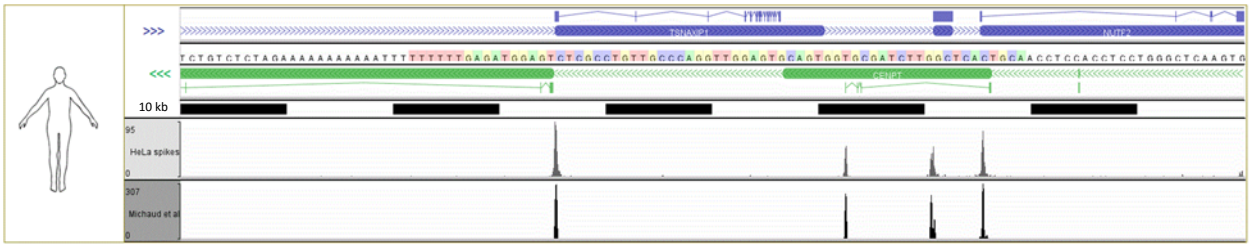
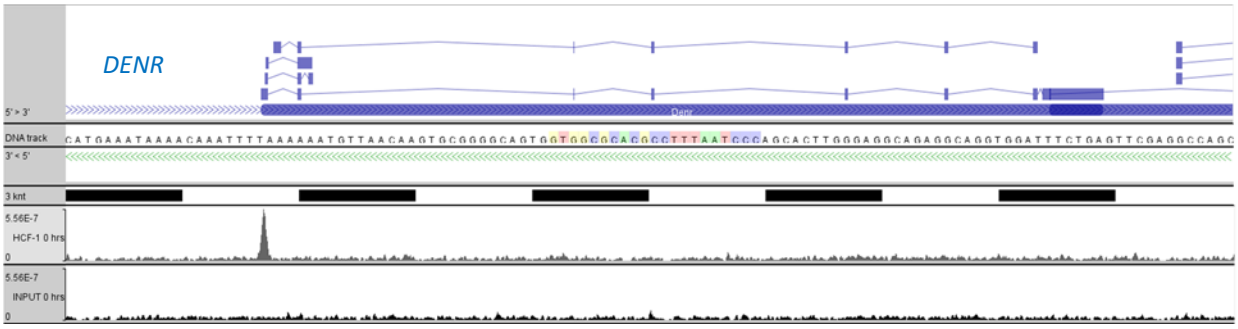



**Figure VI-4**. HCF-1 binding sites in a human genomic region of 100 kb. The figure displays the genomic view of the genes in this genomic region and the associated human ChIP-seq data. On the top the structure of sense (blue) and anti-sense (green) genes and associated transcripts are described. Below, black and white boxes define 10 kb segments of the region.

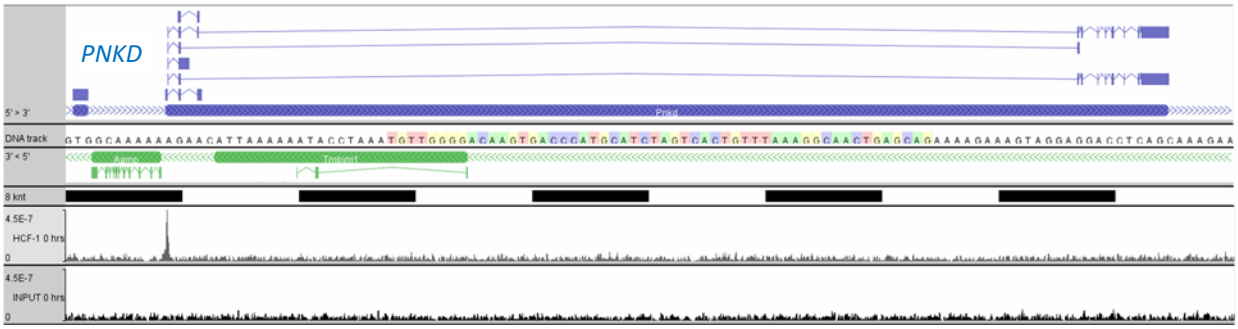
Two data tracks are displayed from the ChIP-seq library prepared on a resting liver and from the data in [Michaud et al., 2013]. On the first tack, the density of the central 50 bp of analyzed fragments is displayed. The second shows the centered single ends 38 bp long.

The sequencing data mapping onto the mouse genome displays regions of high densities of ChIP-ed fragments over Input. HCF-1 binds to promoters of genes such as the *Denr*, *Pnkd*, *Qrich1* and *Syt3*, as shown in Figure VI-5. To further test whether HCF-1 binds to these promoters, I did ChIPs on mouse liver chromatin samples using the HCF-$1_N$ N961 and HCF-$1_C$ Bethyl antibodies from mice under entrainments of i) 12-hrs light-dark cycles and ad-libitum food access, and ii) 12-hrs light-dark cycles and 12-hrs dark-cycle food access. I tested whether ChIPed fragments could be detected by using primers on the four aforementioned promoters and quantifying the amplicons of quantitative Polymerase Chain Reactions (qPCR). As a negative reference for the ChIP, I used Immunoglobulin G (IgG). As a negative control, I also tested quantifications on a tRNA gene (MS6) that didn't display binding by HCF-1 in the ChIP-seq data. Surprisingly, none of the four selected promoters displayed HCF-1 association in the ad-libitum fed sample, whereas two of the four promoters showed significant HCF-1 binding in the restricted food access sample. This unexpected result illustrates a dynamic association of HCF-1 with promoters and the importance of the feeding regimen for the study of HCF-1 chromatin association. Indeed, HCF-1 binding on those gene promoters may be in response to metabolic cues.

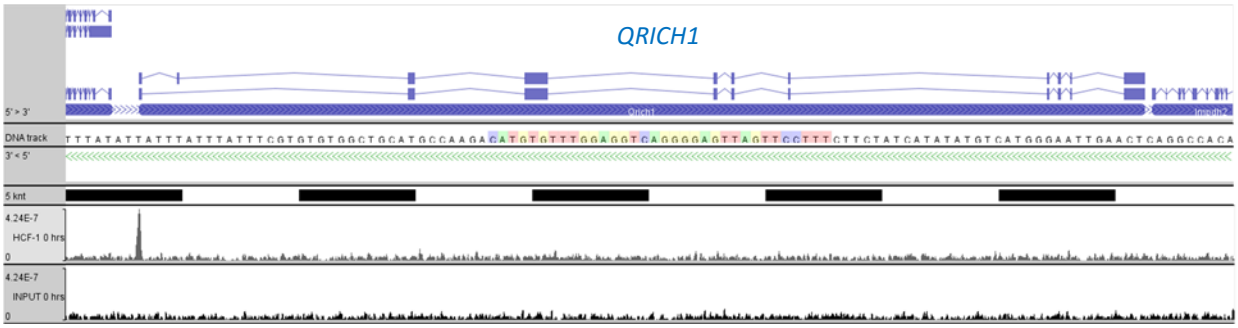
a)


DENR
5' > 3'
Denr
DNA track
3' < 5'
3 knt
5.56E-7
HCF-1 0 hrs
0
5.56E-7
INPUT 0 hrs
0

b)


PNKD
5' > 3'
Pnkd
DNA track
3' < 5'
8 knt
4.5E-7
HCF-1 0 hrs
0
4.5E-7
INPUT 0 hrs
0

c)


QRICH1
5' > 3'
Qrich1
DNA track
3' < 5'
5 knt
4.24E-7
HCF-1 0 hrs
0
4.24E-7
INPUT 0 hrs
0

d)

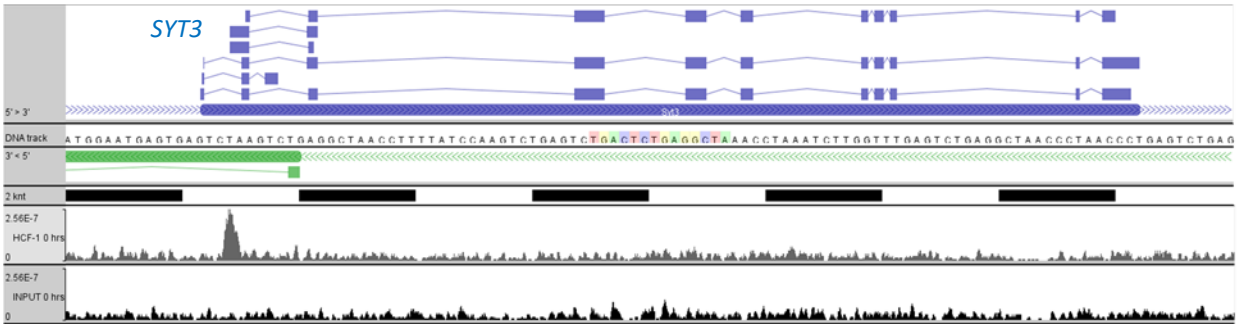
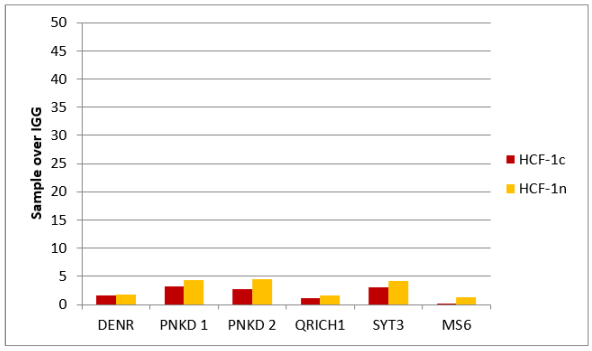

SYT3
5' > 3'
Syt3
DNA track
3' < 5'
2 knt
2.56E-7
HCF-1 0 hrs
0
2.56E-7
INPUT 0 hrs
0

**Figure VI-5**. HCF-1 binds to the chromatin of the mouse resting livers. Genomic views of the genes a) *Denr*, b) *Pnkd*, c) *Qrich1* and d) *Syt3* and associated ChIP-seq data are displayed. On the top the structure of sense (blue) and anti-sense (green) genes and associated transcripts are described. Below, black and white boxes define size-scaled segments of the region, with the size specified on the left. Two data tracks are displayed: the first being the HCF-1 library data and the second the Input data. In both data tracks, the central 50 bp of analyzed fragments are accumulated and displayed.

a)




b)

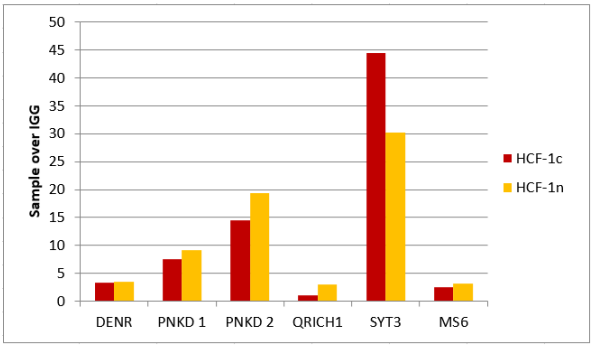



**Figure VI-6**. Quantitative PCR analysis of TSS-proximal HCF-1 peaks in the resting liver. Samples were collected at ZT02. ChIPs were done against HCF-1 with the HCF-$1_C$ Bethyl (red) and HCF-$1_N$ N961 (yellow) antibodies. Primers on the promoters of the genes *Denr*, *Pnkd*, *Qrich1*, *Syt3* and *Ms6* were tested on the ChIPed material. In the case of *Pnkd*, two primers were tested, as HCF-1 binding site was more extended than in the other genes. The tRNA gene *MS6* was taken as a negative control, as HCF-1 did not show binding there. The analysis has been done under two different conditions: a) mice with 12-hrs light-dark cycles and ad libitum access to food and b) mice entrained with 12-hrs light-dark cycles and 12-hrs dark-cycles food cycles. It has to be noted that the food entrainment of 12 hrs was 1 hr shifted compared to the light cycle (i.e., food was removed at ZT01 and the light was turned on at ZT00). Sample over IgG ratios are displayed.

## VI.3 HCF-1 tends to bind to annotated gene promoters in the mouse liver chromatin

HCF-1 has been shown to be a common component of transcriptional regulatory complexes at active gene promoters in the human genome. I investigated on the ChIP-seq data whether in the mouse liver there are similar patterns and I could observe that in the resting mouse liver, HCF-1 indeed tends to bind promoter regions, as it is already evident in Figure VI-5. Additionally, a computational analysis of the HCF-1 ChIP densities versus Inputs in genomic bins, identified bins with significant binding of HCF-1. When aggregating those bins closer than 200 bp, 80% of the detected genomic regions fall within a TSS proximal region (+/- 250 bp around the TSS). In the resting liver, 173 promoters were bound by HCF-1. This number of genes is considerably reduced compared to the over 5,000 HCF-1 bound promoters found in proliferating HeLa cells. The decrease is likely the result of the different nature of cells. HeLa cells are genetically modified cells that are adapted for constant replication. There, HCF-1 acts in genes regulating cell proliferation. In contrast, the resting mouse liver is of a healthy organism that in this conditions does not proliferate and instead performs metabolic, synthesis and detoxification functions.

Nevertheless, similarities arise between the mouse and the HeLa chromatin. In the mouse liver, approximately 92% of the bins with significantly more HCF-1 fragments than in the Input fall within annotated CpG islands. This result was previously observed in the HeLa chromatin [Michaud et al., 2013]. Thus, HCF-1 displays a CpG-island promoter preference in both the mouse liver and HeLa chromatins.

The presence of HCF-1 in promoters is highly associated to Pol2 occupancy. 98% of the genes bound by HCF-1 are also occupied by Pol2. To illustrate this co-occupancy in promoters of the resting liver, an automated visualization of HCF-1 and Pol2 densities in HCF-1 bound promoters displays an overall coordination between their binding in unidirectional and bidirectional promoters (Figure VI-7). A PAM clustering of the HCF-1 promoters into five groups according to the position of the TSS relative to the HCF-1 binding site revealed that in the major group HCF-1 is bound just upstream of the TSS and yet it can also be found downstream and relatively far upstream, as shown in Figure VI-7a.

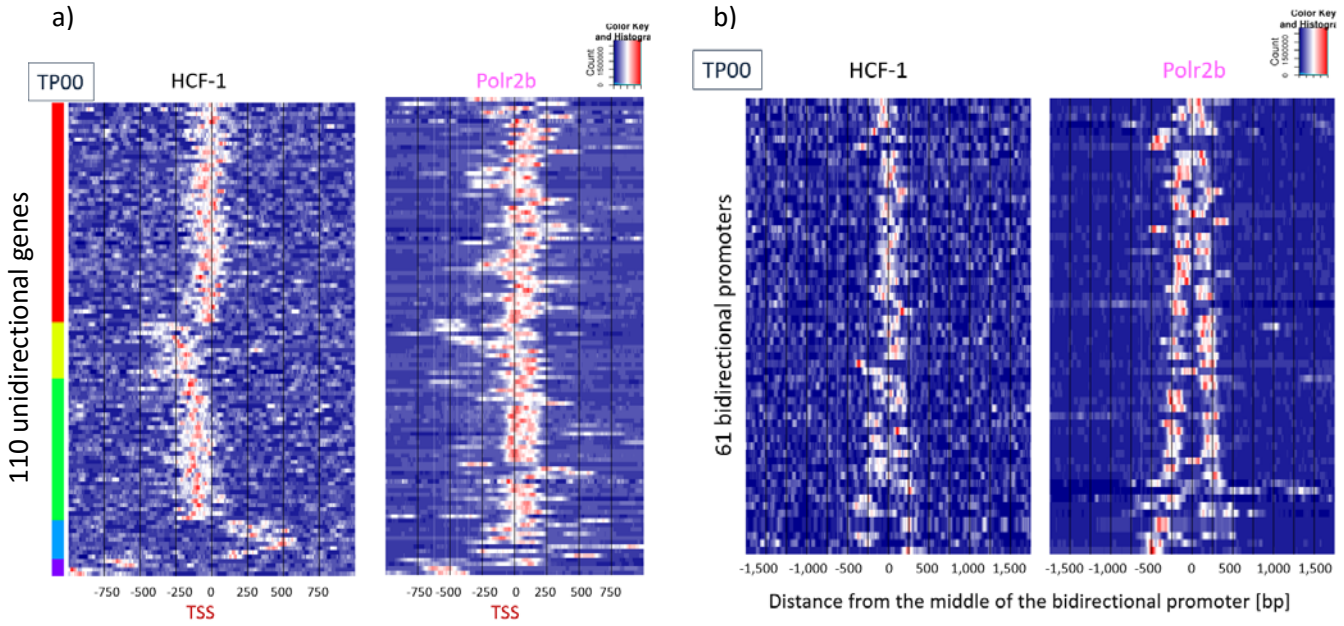



**Figure VI-7**. View of the HCF-1 mouse liver chromatin binding densities at 0 hr in unidirectional and bidirectional promoters. A set of 110 unidirectional promoters and 61 bidirectional promoters displaying HCF-1 binding were selected. In bidirectional promoters, HCF-1 can display either one or two binding sites associated to the different TSSs. HCF-1 and Pol2 densities of the central 50 bp of the sequenced fragments were calculated at nucleotide resolution and genes were piled up and displayed in a color heat map, where each row represents one gene and each column one nucleotide. Density quantifications were centered and scaled in order to display the relative profiles per gene. In this way the meaning of the color code is as follows: the redder the color in a given position the higher the density at that nucleotide within the promoter, whereas the bluer the color, the lower the density. a) HCF-1 density profiles within the 110 unidirectional promoters bound by HCF-1 in the resting liver. Densities are displayed in a region of +/- 1 kb around the selected TSS. All promoters have been displayed in a sense direction with the direction of transcription to the right. Because HCF-1 binds at different positions with respect to the TSS, genes were clustered with the PAM algorithm into five groups to see more clearly the binding sites. Each gene cluster has been colored differently in the bar to the left. b) HCF-1 density profiles within the 61 bidirectional promoters bound by HCF-1 in at least one TSS in the resting liver. Bidirectional promoters were sorted by the distance between the TSSs. Densities are displayed in a region of +/- 1.75 kb around the center position between the selected TSSs.

## VI.4 HCF-1 displays diverse functions in non-dividing differentiated cells

As abovementioned, HCF-1 binds to few genes in the resting liver compared to HeLa cells which could be due to the different proliferative and differentiated states of HeLa cells and resting liver cells. Thus, HCF-1 in HeLa cells binds to genes highly associated with cell proliferation and division. In contrast, the genes bound by HCF-1 in the resting liver are not involved in cell proliferation. They are associated to very diverse functions. HCF-1 bound genes such as *glycogen synthase kinase 3 alpha* (*Gsk3a*) or the *fat mass and obesity associated* (*Fto*) regulate glycogen and lipid metabolism, respectively. Also, genes such as the *peroxisome proliferative activated receptor, gamma, coactivator-related 1* (*Pprc1*) can regulate mitochondrial biogenesis. Further, HCF-1 binds to the methyltransferase *Mll1* gene. And genes related to gene expression functions are also bound by HCF-1. Examples of this are the *Snapc3* gene, which encodes for a subunit of the snRNA-activating protein complex required for transcription by Pol2 and Pol3, the gene of the *Car1* transcription factor and the translation initiation factor gene *Eif2s2*.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *1110004F10Rik* | *AU040320* | *Denr* | *Hcfc1r1* | *Mir3960* | *Pnkd* | *Smcr8* | *Ubr2* |
| *1110019D14Rik* | *Banp* | *Deptor* | *Hexim1* | *Mll1* | *Ppme1* | *Snapc3* | *Uhrf2* |
| *1700003G18Rik* | *Bcorl1* | *Dffa* | *Hmbox1* | *Mob1a* | *Pppde1* | *Snx16* | *Unk* |
| *1700008O03Rik* | *Btrc* | *Dhx30* | *Ing3* | *Morf4l1* | *Pprc1* | *Snx27* | *Uqcr10* |
| *1700030K09Rik* | *C2cd3* | *Dnajb12* | *Inpp5b* | *Mrpl34* | *Prcc* | *Spink10* | *Uqcrh* |
| *1810006K21Rik* | *Calcoco1* | *Dus3l* | *Ints9* | *Mrpl50* | *Prmt6* | *Srprb* | *Usp48* |
| *2500004C02Rik* | *Calr3* | *Dync1i2* | *Invs* | *Napb* | *Qrich1* | *Stau1* | |
| *2610001J05Rik* | *Cant1* | *E230015B07Rik* | *Itpr1* | *Ncdn* | *Ralbp1* | *Styxl1* | |
| *7SK* | *Car11* | *Eif2s2* | *Lamp1* | *Ncln* | *Rangap1* | *Supt5h* | |
| *9430015G10Rik* | *Ccdc132* | *Eny2* | *Lamtor2* | *Ncoa3* | *Rbm38* | *Sympk* | |
| *Aak1* | *Ccdc150* | *Erp44* | *Lcorl* | *Nedd8* | *Rbm39* | *Syt3* | |
| *Aamp* | *Cdk5rap2* | *Fbxo38* | *Lrrc41* | *Nfxl1* | *Rexo2* | *Tfeb* | |
| *Abcb8* | *Cdk9* | *Fbxw11* | *Lrsam1* | *Nr6a1* | *Rnf121* | *Thoc6* | |
| *Acad8* | *Chchd4* | *Fem1a* | *Lsmd1* | *Nudcd1* | *Rpgrip1l* | *Thyn1* | |
| *Ado* | *Chmp4b* | *Fen1* | *M6pr* | *Nudt13* | *Rpia* | *Timm13* | |
| *Ahdc1* | *Clec16a* | *Fto* | *Map3k10* | *Nup153* | *Rpl12* | *Tipin* | |
| *Ap2m1* | *Clpb* | *Fyco1* | *March5* | *Nutf2* | *Rpp40* | *Tmem101* | |
| *Apba3* | *Cpeb3* | *Gm10615* | *Max* | *Ormdl3* | *Rps7* | *Tmem115* | |
| *Aplf* | *Ctu1* | *Gmpr2* | *Mdh1* | *Pa2g4* | *Rraga* | *Tmem43* | |
| *Arv1* | *Cwc22* | *Gsk3a* | *Mdh2* | *Pcif1* | *Rsbn1* | *Top3a* | |
| *Asxl1* | *Dedd* | *Gtf2h4* | *Mfsd11* | *Pemt* | *Sec11c* | *Trpc2* | |
| *Atp5f1* | *Dennd1a* | *Hbp1* | *Mir2861* | *Phf23* | *Smarcal1* | *Ttc13* | |

**Table VI-1**. List of 173 mouse genes displaying HCF-1 binding sites in their promoter proximal region. The gene symbols are indicated and sorted numerically followed by alphabetically.

## VI.5 The HCF-$1_C$ and HCF-$1_N$ subunits associate with mouse chromatin

It is not completely understood how the two subunits of HCF-1 bind the chromatin. To provide insights about this on a genome-wide scale, I asked where each subunit binds to the chromatin and whether there are differences between them. As described in Figure VI-8, to answer this question I prepared mice entrained under 12 hrs light-dark and food access cycles. As in Chapter IV, I had shown that in the Sham livers there are major changes in gene expression compared to the resting livers (and their use avoids hepatectomy and its associated loss of liver sample), I investigated changes in HCF-1 subunit chromatin association in that context. Hence, samples were collected at 0 hr, 1 hr and 4 hrs post Sham operations. In total, three replicates were prepared for each condition. For each replicate sample different ChIPs were performed: targeting the HCF-$1_C$ and HCF-$1_N$ with the Bethyl and N961 antibodies, respectively, and H3K4me3 as a positive control. Inputs were also collected to identify significant densities of ChIP-ed fragments over Input. Sequencing libraries were prepared and were subjected to Illumina high-throughput sequencing.

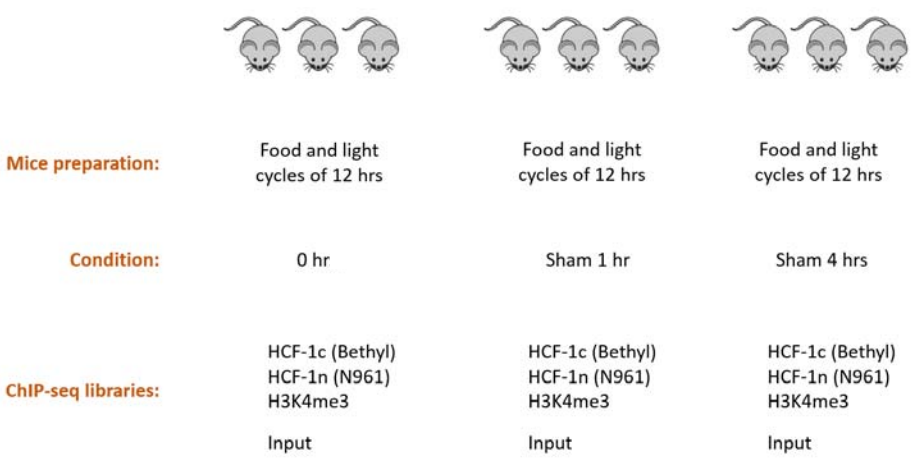



**Figure VI-8**. Design of ChIP-seq experiments to investigate the chromatin association of the two subunits of HCF-1. Three replicate mice were prepared for each condition: resting liver, Sham 1 hr and Sham 4 hrs. Mice were prepared under 12-hrs light-dark and food access cycles. From each liver sample, crosslinked chromatin was collected and ChIPs were performed targeting HCF-$1_C$ (Bethyl antibody), HCF-$1_N$ (N961 antibody) and H3K4me3 as a positive control. Input samples were prepared directly from the crosslinked chromatin.

As shown in Figure VI-9, previous to sequencing, the concentration of the ChIP libraries was on average 3ng/ul. The Inputs tended to have higher concentrations, as expected, although they were quite variable for unknown reasons. All the libraries were sequenced in the same run with paired-end sequencing and multiplexed

to different degrees: the 18 libraries from HCF-1 ChIPs were sequenced in a total of 6 lanes, and the 18 Input and H3K4me3 libraries were sequenced in only 2 lanes, to allow for more sequenced fragments on the HCF-1 libraries.

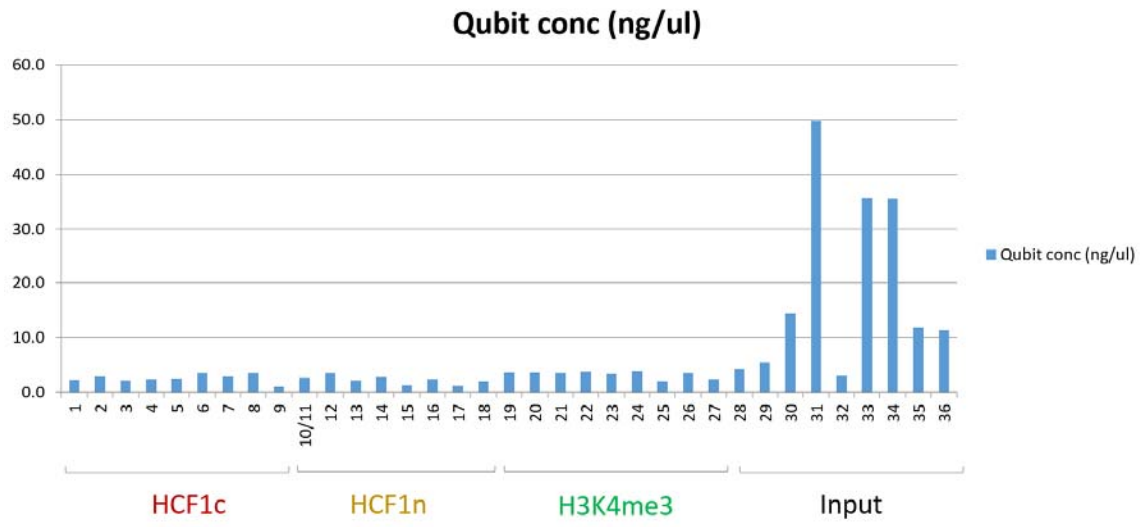



**Figure VI-9**. Distribution of library concentrations before sequencing. During library preparation, libraries were numbered from 1 to 36. Libraries 1 to 9 contain ChIPs of HCF-$1_C$. Libraries 10 to 18 contain ChIPs against HCF-$1_N$. Libraries 19 to 27 contain ChIPs against H3K4me3. Libraries 28 to 36 contain Inputs. Within each set of libraries the first set of three contains replicates 1-3 for time point 0 hr. The second set of three contains replicates 1-3 for Sham 1 hr livers. The last set of three libraries contains replicates 1-3 for Sham 4 hrs livers. Libraries 10 and 11, which contained different multiplexing indexes, were pooled together during library preparation, and the concentration of the mix was obtained. After sequencing, demultiplexing of these two libraries was performed.

As expected because of the different degrees of multiplexed sequencing, the HCF-$1_C$ and HCF-$1_N$ libraries retrieved more sequenced fragments than other libraries (Figure VI-10). Approximately, 80 million fragments were obtained in the HCF-1-subunit libraries versus 20 million obtained from the H3K4me3 and Input libraries. An exception was the Replicate 3 library at Sham 4 hrs from a ChIP against H3K4me3. This library showed higher number of fragments even though it was similarly multiplexed. Some of the sequence data was removed because of either bad sequencing quality, not possible to map, originating from the PhiX spike control added by the sequencing facility or fragments mapping to multiple locations. After this removal, high numbers of fragments were retained: 20 million in the H3K4em3 and Input libraries, and 40 million in the HCF-1 libraries. Some Inputs showed a reduced number of fragments to analyze (Sham 1 hr replicate 2, and replicates 2 and 3

of Sham 4 hrs). Curiously, these Input libraries with reduced number of final fragments for analysis, had 30% less fragments mapping onto the mouse genome than usual (data not shown), suggesting that they may be contaminated with some additional chromatin. Because of that, I pooled the 6 good quality Input libraries for further data analyses.

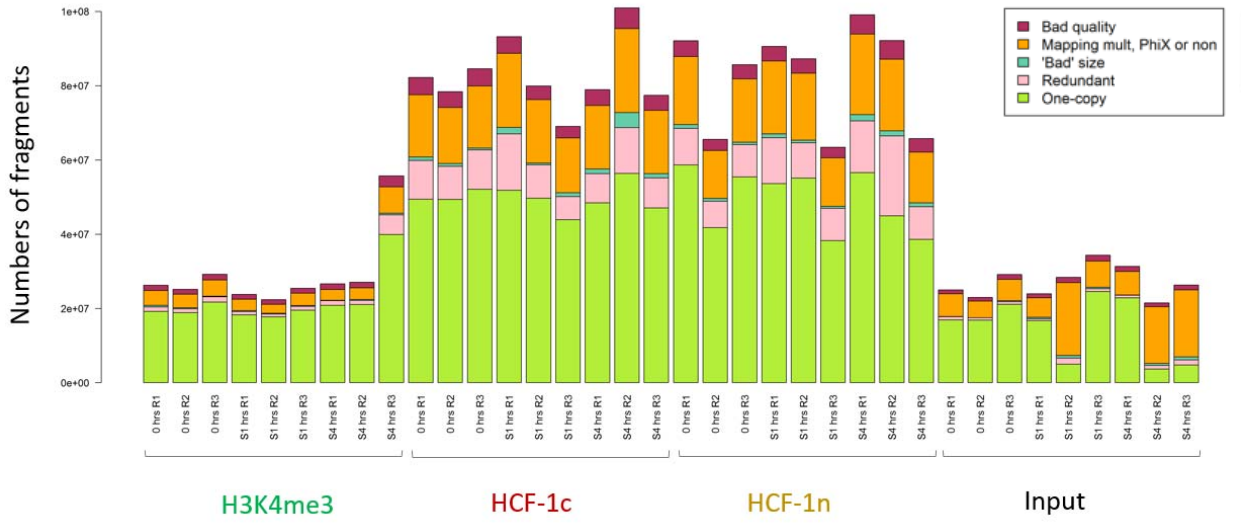



**Figure VI-10**. Number of human and mouse fragments obtained after mapping the ChIP-seq reads of all the libraries prepared targeting HCF-1. Reads were mapped onto the mouse mm9 genome assembly. The red areas represent the numbers of fragments removed from the analysis because of bad sequencing quality. Further, the areas in orange represent the number of fragments either mapping on multiple genomic locations or with origin on the PhiX control genome or not mapping onto the mouse genome. In light blue, the number of fragment sizes out of [50, 500] bp are indicated. The pink areas depict the number of fragments with good quality, desired size range and uniquely mapped onto the mouse genome. Finally, light green areas represent the numbers of fragments used for analysis, which contain fragments present only once in the dataset plus one copy of the redundant fragments.

The visualization of the 50 central nucleotides of sequenced fragments from 50 to 500 bps displays HCF-1 binding sites also observed when using the H12 antibody. For example, on the promoter of the *PNKD* gene (Figure VI-11) it can be seen that the replicates obtained with the N961 and Bethyl antibodies show a similar binding site to the HCF-$1_C$ H12 seen previously. Notably some variation can be observed among replicates. For example, the HCF-$1_N$ Replicate 2 of the Sham 4 hrs condition displays a very low profile compared to the other replicates.

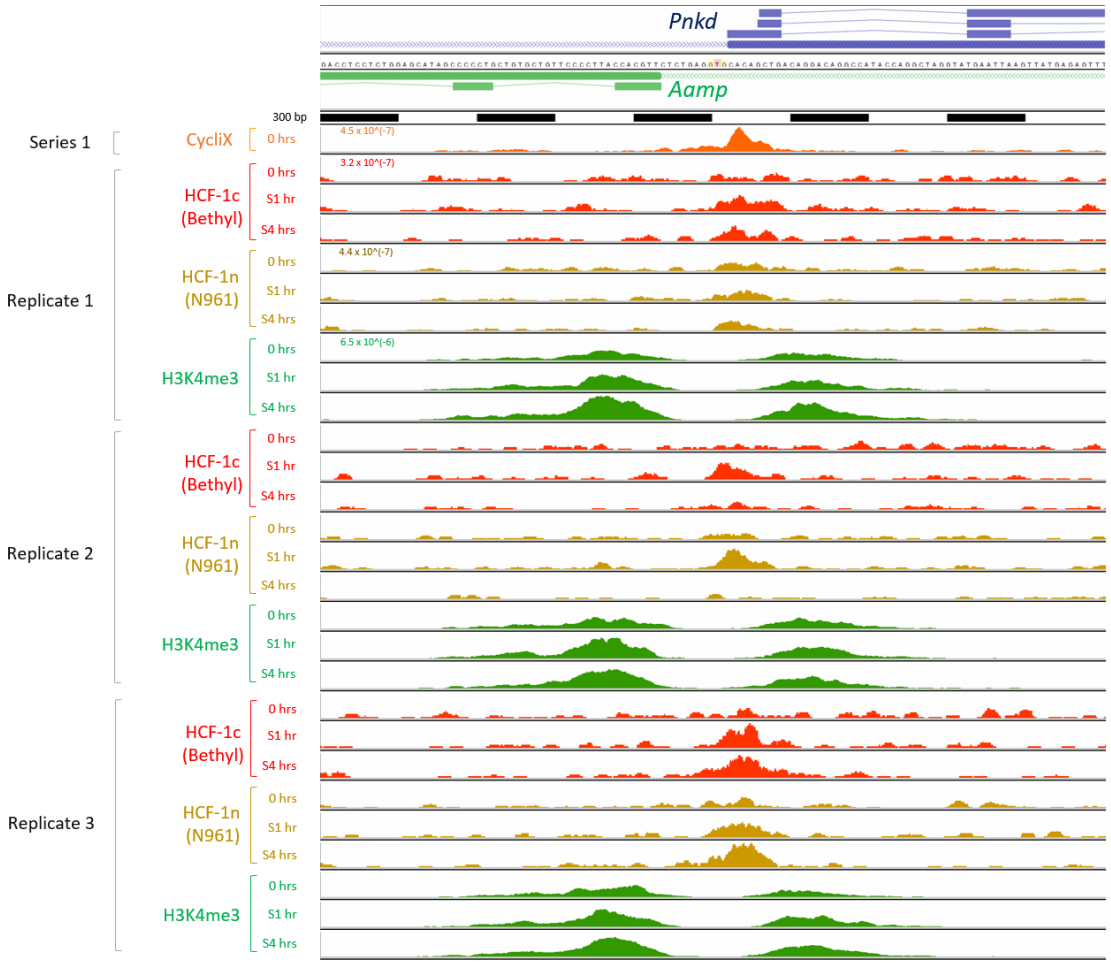



**Figure VI-11**. Genomic profiles of HCF-$1_N$ and HCF-$1_C$ binding in the *PNKD* and *AAMP* bidirectional promoter. The figure displays the genomic view of the *PNKD* and *AAMP* bidirectional promoter and the associated ChIP-seq data. On the top the structure of sense (blue) and anti-sense (green) genes and associated transcripts are shown. Below, black and white boxes define 300 bp segments of the region. Data from three time points are shown: 0 hr, Sham 1 hr and Sham 4 hrs. The first track, displays HCF-1 ChIP-seq data prepared from resting livers in the first series of mice prepared in the CycliX project (see Chapter IV). The ChIP was done using the H12 antibody. All the other tracks display data for the ChIPs against HCF-$1_N$ (N961 antibody), HCF-$1_C$ (Bethyl antibody and H12 with only one sample) and H3K4me3. In each data track, the central 50 bp of sequenced fragments are accumulated and displayed. Both subunits of HCF-1 bind to this bidirectional promoter in a position closer to the transcriptional start site of *PNKD*.

## VI.6 The chromatin association of both subunits of HCF-1 increase after the Sham operation but they may influence cellular states differently through transcription regulation

Here, I studied where in the genome the HCF-1 subunits bind and whether there are differences between their binding patterns. As a first approach, I did an analysis in Chromosome 1 by using the CycliX viewer software. With this tool I detected genomic regions with more density of HCF-1 fragments than the background. To do so, I defined thresholds for each of the ChIP and Input libraries by visually identifying the maximum fragment densities at a nucleotide resolution in intergenic regions of Chromosome 1. Then I selected genomic regions with nucleotide densities higher than the intergenic-region threshold within the HCF-1 ChIP libraries. I did similarly for the Input data to identify potential biases that could also be present in the ChIP tracks. After that, I removed from the selected ChIP-associated regions those regions also obtained in the Input tracks. To know in which type of genomic region these binding sites were located, I aggregated regions closer than 200 bp within tracks and I then merged all the selected regions across replicates and time points for each specific ChIP (i.e., HCF-$1_N$ or HCF-$1_C$). As shown in Table VI-2, this global analysis of HCF-1 binding across time points and replicates detects in Chromosome 1 a total of 16 binding regions retrieved by the HCF-$1_C$ antibody and 120 binding regions retrieved by the HCF-$1_N$ antibody, with an overlap of the 80% of the HCF-$1_C$ sites with the HCF-$1_N$ sites. Notably, the HCF-$1_C$ subunit (targeted by the Bethyl antibody) is observed in few genomic locations compared to the HCF-$1_N$ subunit, in Chromosome 1 across time points and replicates. Interestingly, in the CycliX data from resting livers prepared with the H12 anti-HCF-$1_C$ antibody, only 6 binding sites are observed in Chromosome 1 (data not shown), and all of them are observed in the newly prepared libraries, although not necessarily at 0 hr; 3 out of the 6 cases are observed at time points post Sham operation. This comparison suggests that the new ChIPs worked correctly and emphasizes the variability of HCF-1 ChIPs across replicates also observed in Figure VI-11. Additionally, almost 90% of the HCF-$1_N$ and HCF-$1_C$ binding sites them fall within promoter proximal sites, which is consistent with what was observed in HeLa cells [Michaud et al., 2013; chapters IV and VI of this thesis]. This further confirms the good quality of the data obtained from the ChIP-seq experiments and shows remarkable differences between subunits. The fact that the chromatin immunoprecipitation retrieved these differences in promoters might be explained by differences in the complexes formed there. Nevertheless, these differences would need to be further confirmed by qPCR.

| Subunit | # binding sites | % in promoters |
|---|---|---|
| HCF1c | 16 | 88% |
| HCF1n | 120 | 87% |

**Table VI-2**. Summary of the genomic regions bound by each subunit of HCF-1 in Chromosome 1. The number of binding regions are indicated for each subunit and the percentage among them that fall within promoter proximal regions (+/- 250 bp around TSS). From the HCF-$1_C$ binding sites 10 have been considered false positives, very close to background and located in intergenic regions, and were not included in the table above. 13 of the binding sites are shared by both subunits.

As a quality control, I compared replicate libraries by doing a correlation analysis of the log2(ChIP/Input) ratios from promoter proximal regions (Figure VI-12). The HCF-$1_C$ libraries tend to display low ratios although, there is a clear increase produced in the Shams that is correlated among replicates. The HCF-$1_N$ libraries are more highly correlated, and as it has been already discussed, the HCF-$1_N$ replicates show more binding sites than the HCF-$1_C$ libraries, which, as shown in Figure VI-12, tend to correlate across replicates. It can also be noted that the second HCF-$1_N$ replicate at Sham 4 hrs displays ratios lower than the other corresponding replicates. This has also been noted in the visualizations of the ChIP-seq data (Figure VI-11). The fact that the ratios are lower for this Replicate 2, may be due to a low efficiency of the ChIP. Finally, the H3K4me3 replicate libraries correlate very well. Concluding, the ChIP libraries are of good quality and can be used for further studies, with exception of Replicate 2 of the HCF-$1_N$ ChIP on a Sham 4 hrs liver.

The careful study of the HCF-1 binding sites further shows the differences between the HCF-$1_N$ and HCF-$1_C$ chromatin binding. In promoters bound by the two subunits of HCF-1, each subunit tend to peak in a different position as in the case of the *Smarcal1* gene promoter (Figure VI-13a,b). Across replicates (Figure VI-13a), there is a tendency for HCF-$1_C$ to position upstream of HCF-$1_N$ within the promoter that can be better observed in an overlay of both subunits (Figure VI-13b). Nevertheless, H3K4me3 can be observed around the two subunits of HCF-1, suggesting that both subunits are associated in a nucleosome free region. Further, the accumulation of this histone modification typically associated with active transcription shows that HCF-1 in this promoter may participate in the regulation of transcription.

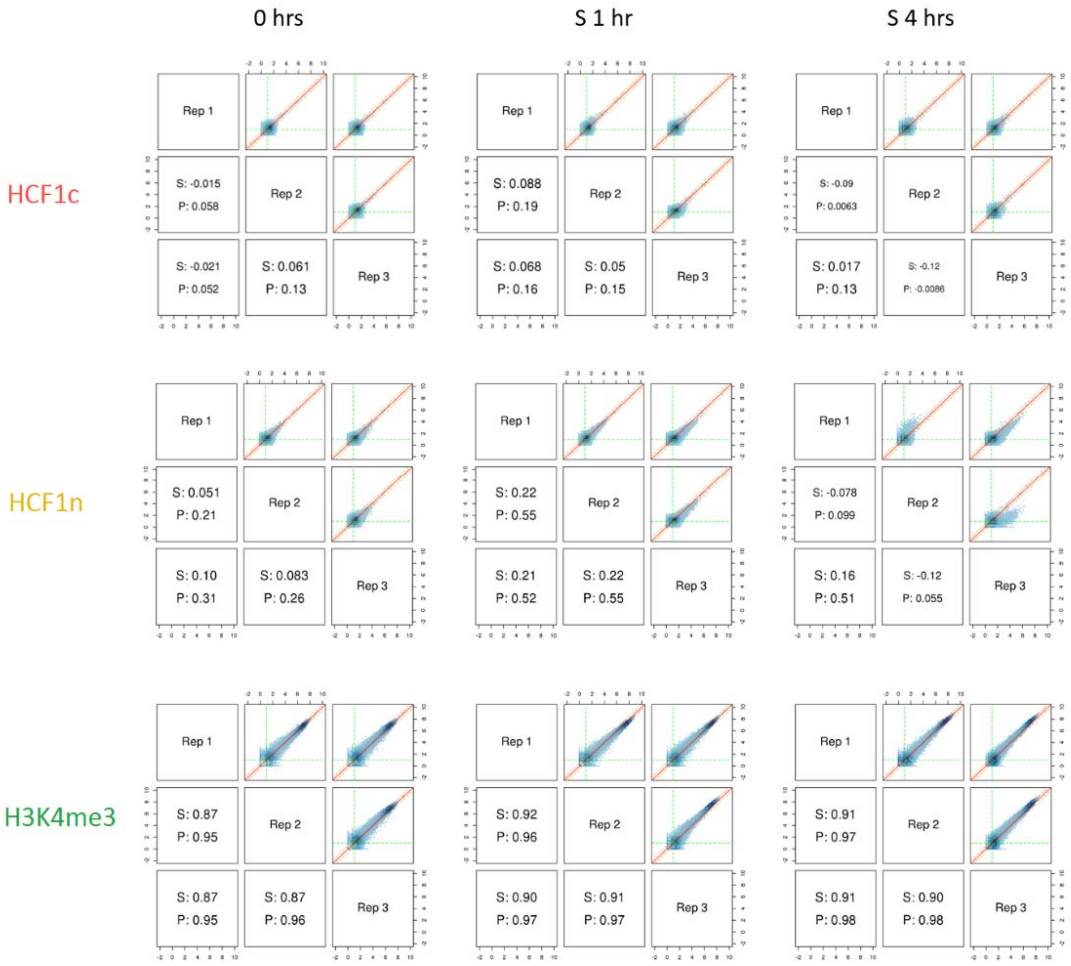
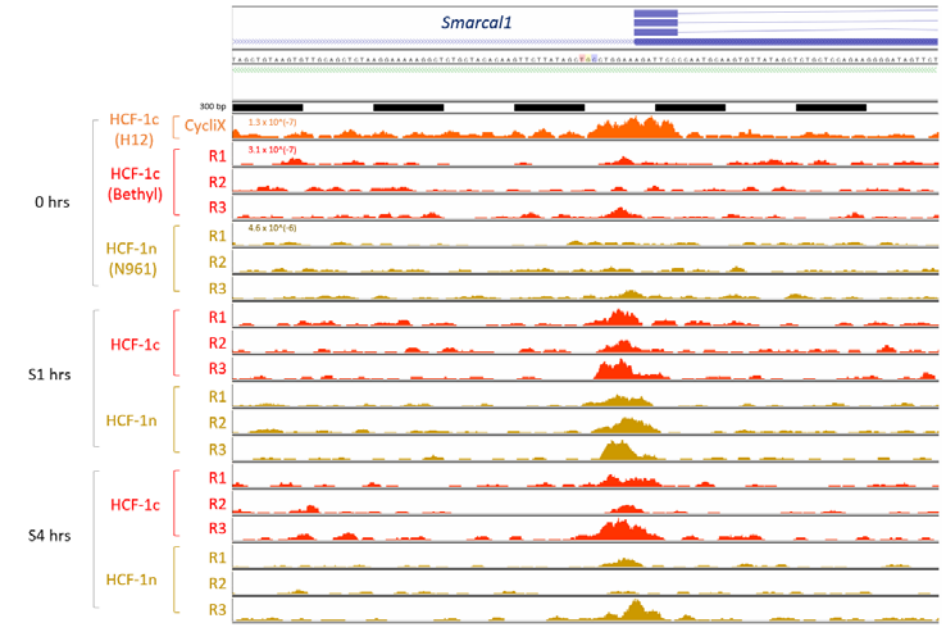



**Figure VI-12**. Correlation analysis among replicates. Nine panels show comparisons among replicates at specific time points (0 hr, Sham 1 hr and Sham 4 hrs) and ChIPs (HCF-$1_C$ with Bethyl, HCF-$1_N$ with N961 and H3K4me3). Promoter regions were quantified for HCF-1 libraries (+/- 250 bp around TSS) and H3K4me3 (+/- 500 bp around TSS). All Inputs were pooled and quantifications were done on both defined promoter regions. Log2(ChIP/Input) ratios were calculated for each sample. In each panel the diagonal quadrants indicate the name of the samples, and the quantifications from pairs of samples were plotted in scatterplots (upper diagonal quadrants) and their Spearman and Pearson correlation coefficients were calculated (lower diagonal quadrants). The size of the correlation coefficients inform about the degree of correlation.

a)




b)

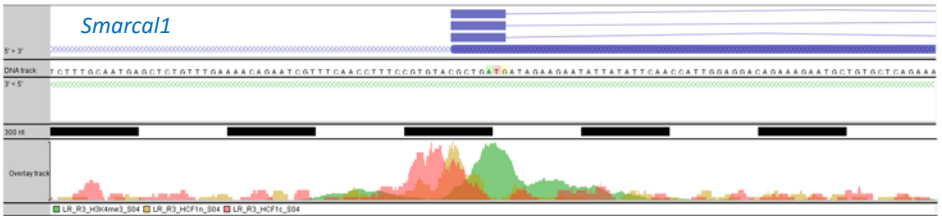
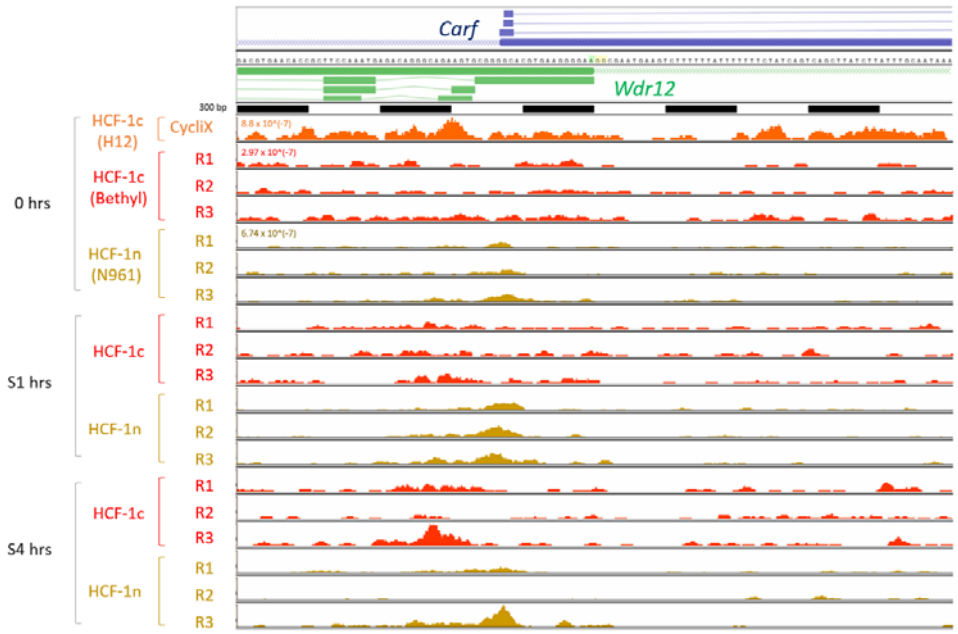



**Figure VI-13**. Genomic profiles of HCF-$1_N$ and HCF-$1_C$ binding in different positions within promoters. a) The figure displays the genomic view of the *Smarcal1* gene promoter and the associated ChIP-seq data from both subunits. On the top of the figure the structure of sense (blue) and anti-sense (green) genes and associated transcripts is described. Below, black and white boxes define segments of the region, which have a size that is specified on the left legend. Data from three time points is shown: 0 hrs, Sham 1 hr and Sham 4 hrs. For each time point replicate data tracks are displayed for the ChIPs against HCF-$1_N$ in yellow (N961 antibody), HCF-$1_C$ (Bethyl antibody in red and H12, 0 hr CycliX series 1, in orange) and H3K4me3, in green. In each data track, the central 50 bp of sequenced fragments are displayed and piled up. b) Overlay of the data from the HCF-$1_N$ and HCF-$1_C$ subunits at the Sham 4 hrs time point in the promoter of the *Smarcal1* gene. Data from Replicate 3 was used. Also the H3K4me3 ChIP-seq profile has been overlaid to depict the position of nucleosomes in the promoters. The color code is as in a).

HCF-1 binds to a big percentage of bidirectional promoters (chapter III and this chapter). Among the selected genes in Chromosome 1, there is one case of bidirectional promoter between the genes *Wdr12* and *Carf* where the two subunits of HCF-1 bind (Figure VI-14a). Notably, the HCF-$1_N$ subunit is positioned on the right of HCF-$1_C$. Similarly to the *Smarcal1* promoter (Figure VI-13a,b), H3K4me3 is flanking the binding sites, but, this time, there is one nucleosome in between both subunits (Figure VI-14b). In this bidirectional promoter both subunits could have a separate activity, perhaps to regulate transcription of each of the two genes. Alternatively, they could be associated with the entrance and exit from the nucleosome.

a)




b)

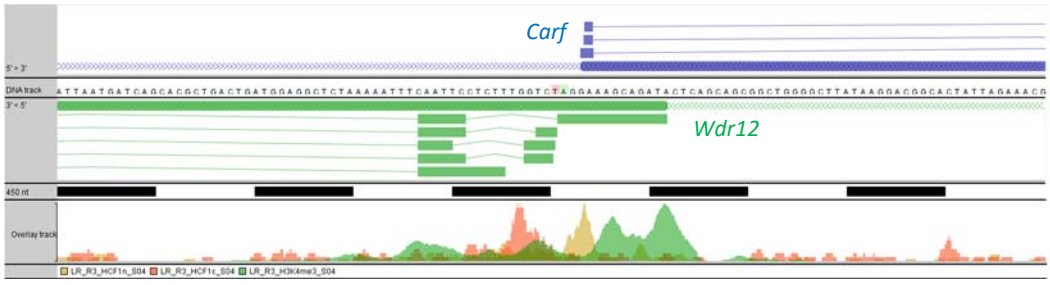



**Figure VI-14**. Genomic profiles of HCF-$1_N$ and HCF-$1_C$ binding in the bidirectional promoter between the genes *Wdr12* and *Carf*. a) The figure displays the genomic view of the bidirectional promoter between the genes *Wdr12* and *Carf* and the associated ChIP-seq data from both subunits. On the top of the figure the structure of sense (blue) and anti-sense

(green) genes and associated transcripts is described. Below, black and white boxes define 450 bp segments of the region. Data from three time points are shown: 0 hr, Sham 1 hr and Sham 4 hrs. For each time point replicate, data tracks are displayed for the ChIPs against HCF-$1_N$ in yellow (N961 antibody), HCF-$1_C$ (Bethyl antibody, in red, and H12, 0 hr CycliX series 1, in orange) and H3K4me3, in green. In each data track, the central 50 bp of sequenced fragments are accumulated and displayed. b) Overlay of the Replicate 3 data from the HCF-$1_N$ and HCF-$1_C$ subunits at the Sham 4 hrs time point in the bidirectional promoter. Also the H3K4me3 ChIP-seq profile has been overlaid to depict the position of nucleosomes in the promoters. The color code is as in a).

Genes with binding of only one subunit of HCF-1 were observed. The neighboring promoters of the genes *Nvl* and *Cnih4* are bound by the HCF-$1_N$ and HCF-$1_C$ subunits, respectively (Figure VI-15). Both subunits bind at 1 hr in the Sham operated mouse livers and seem to correlate with an increased transcription as shown by the increase of H3K4me3 when HCF-1 binds. Interestingly, in each promoter transcription is enhanced under different settings. In the *Nvl* promoter nucleosomes with H3K4me3 can be observed around the position where HCF-$1_N$ is sitting. In contrast, in the *Cnih4* promoter there is H3k4me3 only downstream of the HCF-$1_C$ binding. These differences in H3K4me3 patterns suggest that different protein complexes and different mechanisms are associated to the activation of transcription in these two genes separated by 7 kb. As a side note, it is also interesting to note that trimethylation levels are already observed in these genes in the resting liver along with Pol2 (data not shown). HCF-1 in these genes may be responsible for boosting transcription of genes that are already transcribed possibly in response to the Sham operation.

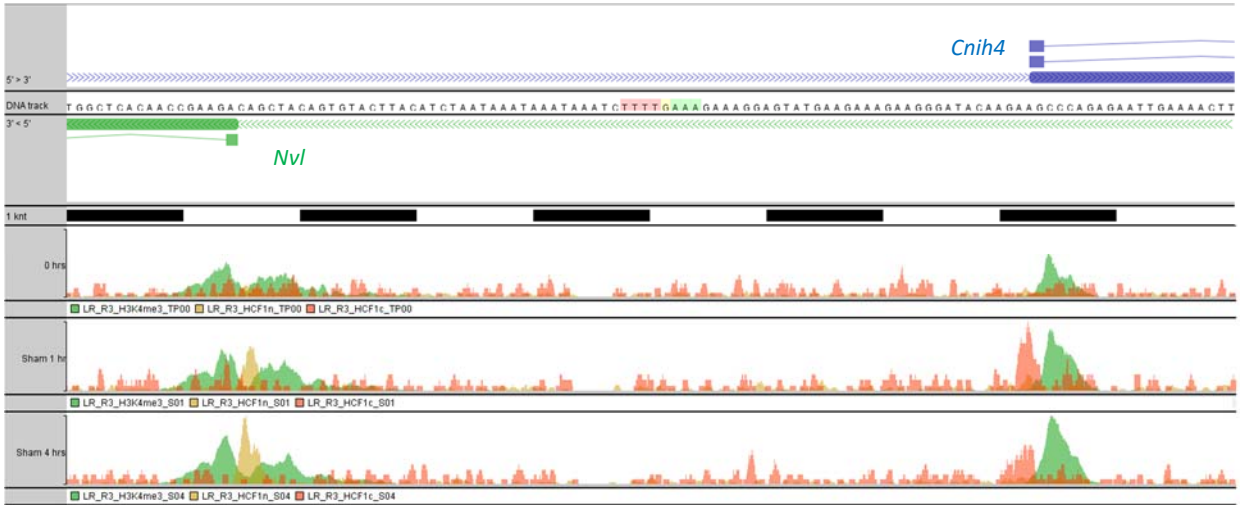


**Figure VI-15**. Progression of HCF-1 binding in the promoters of the genes *Nvl* and *Cnih4*. On the top of the figure the structure of sense (blue) and anti-sense (green) genes and associated transcripts is described. Below, black and white boxes define 1 kb segments of the region. Data from three time points is shown: 0 hr, Sham 1 hr and Sham 4 hrs. For each time point overlaid data tracks are displayed for the ChIPs against HCF-$1_N$ (N961 antibody) in yellow, HCF-$1_C$ (Bethyl antibody) in red and H3K4me3 in green. In each data track, the central 50 bp of sequenced fragments are displayed and piled up.

# Discussion

In this chapter I have investigated the role of HCF-1 in the mouse liver chromatin. The study of HCF-1 chromatin association in the liver reveals how HCF-1 is a versatile and rapid promoter component.

## Heterologous spike chromatin can be used as a double quantitative and qualitative internal control for ChIP-seq experiments

Different methods have been proposed as a control of the ChIP-seq experiment and support the analysis of the sequencing data. These methods intend to control for technical issues during the ChIP-seq experiment and that could be used to correct quantifications during the in-silico analysis [Bonhoure et al., 2014; Hu et al., 2014; Orlando et al., 2014; Hu, Petela et al., 2015; Grzybowski et al., 2015]. These methods typically involve the addition of an independent chromatin to the sample that can be targeted by the ChIP and sequencing.

For the research of the mouse liver, a 5% of HeLa chromatin was added to the mouse chromatin as described in the methods section in chapter IV [Bonhoure et al., 2014]. Interestingly, this spike-in method offered an alternative utility for the study of HCF-1 chromatin association. The HCF-1 binding events in the spiked HeLa chromatin could be compared with a previous study of HCF-1 binding sites in the HeLa chromatin [Michaud et al., 2013]. Thus, a match between both chromatins demonstrate a proper ChIP-seq on the chromatin of study. In this research, HCF-1 binds similarly in the spiked chromatin as in the mouse liver chromatin, which confirms that HCF-1 binds to few promoters in the resting mouse liver. Nevertheless, it should not be discounted that the antibody used may not efficiently recognize the mouse epitope as the human epitope.

## The ways of action of HCF-1 in gene promoters

In both the mouse liver and HeLa cells, the precursor HCF-1 is cleaved producing the HCF-$1_N$ and HCF-$1_C$ subunits that associate and bind to the chromatin, especially in promoter proximal regions [Michaud et al., 2013; this chapter]. At least in HeLa cells, HCF-1 interacts with a wide range of proteins for transcription regulation (see section I.4.2 of the Introduction) showing that HCF-1 is indeed a versatile protein that is able to mediate cellular functions in different ways. Furthermore, in this chapter I have observed cases of genes where the two subunits of HCF-1 display different binding sites or independent binding. These results should be further confirmed by ChIP-qPCR as such observations could be due to inefficient immunoprecipitation or alternatively, to epitope unavailability. Nevertheless, interestingly, Julien and Herr (2003) showed that the

independent action of the HCF-$1_N$ and HCF-$1_C$ through the cell cycle is sufficient for proper passage through G1 phase and proper exit of mitosis, respectively. This suggests, that each subunit of HCF-1 could have independent functions, and perhaps could bind independently in promoters, which would be consistent with the observations in this chapter if they are further confirmed.

Moreover, even though, previous studies showed that the majority of HCF-1 subunits remain associated after cleavage, alternative pre-mRNA processing of HCF-1 can produce a variant HCF-1, called HCF-1Δ382–450 as a result of the removal of an exon encoding the amino-terminal HCF-1 residues. This smaller variant undergoes proteolytic cleavage, but the resulting subunits don't remain associated [Wilson et al., 1995b]. Yet, it is not known whether these independent subunits are functional. Further studies about this variant could give insights about different mechanisms of action of HCF-1 in promoters.

The action of HCF-1 in promoters seems to be very rapid and dynamic. In my studies of independent replicates of Sham operated mouse livers I observe high variability, which is also evident in the comparison with the CycliX data at 0 hr. Replicate livers were collected in a row and the time between operations was not more than 10 minutes. Notably, this time is enough to observe different profiles across replicates suggesting that HCF-1 may be mediating activities in those promoters that require little time.

# Methods

## Cell culture

Human HeLa cells cells were cultured at 37°C under 5% CO2 in DMEM supplemented with 10% fetal bovine serum (FBS), 100 U/mL penicillin, and 100 μg/mL streptomycin unless otherwise specified.

## Animals

C57/BL6 10-15 week old male mice were used. For the study of the association of the two subunits of HCF-1 mice were entrained during two weeks on 12 hrs light-dark cycles and ad libitum access to food. For the study of the chromatin binding of HCF-1 in the resting liver different mice were prepared under different food-access conditions. First, mice were entrained during two weeks on 12 hrs light-dark cycles and ad libitum access to food. Later, new mice were entrained during two weeks on a 12 hrs on/12 hrs off feeding and light-dark cycle with food access between ZT13 and ZT01 (ZT0 is defined as the time when the lights are turned on and ZT12 as the time when the lights are turned off). Further, experimental mice used for ChIP-seq of HCF-1 were entrained for two weeks on a 12 hrs on/12 hrs off feeding and light-dark cycle with food access between ZT12 and ZT24 (ZT0 is defined as the time when the lights are turned on and ZT12 as the time when the lights are turned off). In the last case, three replicate mice were prepared per time point.

## Sham operations

Sham-operated controls were subjected to laparotomy and livers were collected at 1 and 4 hrs after operation. Three biological replicates were prepared for each time point.

## Protein extraction from HeLa cells

Cells were washed with PBS and incubated during 1 hr at 4°C on a rotating wheel in 30 μl of lysis buffer (1% NP40, 50 mM Tris pH 8.0, 300 mM NaCl, and one complete mini-tablet from Roche). Lysate were centrifuged at 14,000 rpm for 5 minutes at 4°C.

## Nuclear protein extraction from mouse livers

(Adapted from personal communication from Prof. Ueli Schibler). One mouse was anesthetized with isoflurane and killed by cervical dislocation. The livers were perfused with 5 mL of 1X PBS through the spleen and immediately collected. The tissue was put in a glass dounce. 0.3 M Homogenization buffer was added to a final volume of 4 ml (10 mM HEPES-KOH pH 7.6, 15 mM KCl, 2 mM EDTA, 10% Glycerol 0.3 M sucrose, 0.15 mM Spermine, 0.5 mM Spermidine, 0.5 mM PMSF and 0.5 mM DTT). 5 ml (out of 25ml) of 2.2 M Homogenization buffer was also added (10 mM HEPES-KOH pH 7.6, 15 mM KCl, 2 mM EDTA, 10% Glycerol, 2.2 M Sucrose, 0.15 mM Spermine, 0.5 mM Spermidine, 0.5 mM PMSF and 0.5 mM DTT). The tissue was homogenized up and down three times with a Teflon pestle. The rest of the 2.2 M homogenization buffer was added to the homogenate, mixed well and kept on ice. The homogenate was carefully overlaid on a 10 ml sucrose cushion in an ultracentrifuge tube (10 mM HEPES-KOH pH 7.6, 15 mM KCl, 2 mM EDTA, 10% Glycerol, 2.05 M Sucrose, 0.15 mM Spermine, 0.5 mM Spermidine, 0.5 mM PMSF and 0.5 mM DTT). The tubes were equilibrated before centrifugation. Centrifugation was for 1 hr, at a speed of 24,000 rpm and at 4°C. The top lipid layer was removed with a spatula. The sucrose solution was aspirated and the tube flipped upside down at the end of the aspiration. The tube wall was rinsed with water while holding it upside down and then put back on ice. 500 µl of RIPA buffer was added (including 1 mini-complete tablet added before use) to resuspend the nuclei and transfer the suspension to a new microfuge 1.5 ml tube. 500 µl were added to clean nuclei off the centrifuge tube and transferred to the 1.5 ml microfuge tube. The tube was vortexed and incubated 20 minutes on ice and subsequently centrifuged for 10 minutes at maximum speed at 4°C. The supernatant containing all soluble proteins was removed and stored.

## Antibodies

List of antibodies used:

- The anti-rabbit polyclonal anti-HCF-$1_N$ N961 [Machida et al., 2009]
- The anti-rabbit polyclonal anti-HCF-$1_N$ N13 [Capotosti et al., 2011].
- The anti-rabbit polyclonal anti-HCF-$1_C$ H12 [Wilson et al., 1993a]
- The anti-rabbit polyclonal anti-HCF-$1_C$ (Bethyl laboratories inc. cat. # A301-399A)

## Immunoprecipitation

For immunoprecipitation of human HeLa cells, whole cell extract two 10-cm dishes with $1x10^7$ cells each were used. For immunoprecipitation of nuclear protein extract of mouse liver one liver was used.

100 µl of each sample were diluted with IP dilution buffer/RIPA 2:1. The composition of the IP dilution buffer was: 50 mM Tris-HCl pH 7.4, 150 mM NaCl, 1 mM EDTA, 2.5 µl/ml PMSF and 0.5 µl/ml DTT. The dilutions from mouse liver extracts were incubated overnight at 4°C with 10 µl of polyclonal anti-HCF-$1_C$ (Bethyl), 10 µl of polyclonal anti-HCF-$1_C$ H12, 10 µl of polyclonal anti-HCF-$1_N$ N961 and 10 µl of polyclonal anti-HCF-$1_N$ N13. Samples were incubated overnight at 4°C with the antibodies. For immunoprecipitation, protein A-agarose beads were prepared. 100 µl of a 50% slurry of protein A-agarose beads per sample were washed three times with 1 ml of IP buffer and 5 min incubation on a rotating wheel between washes. The supernatant was removed and resuspended on 50 µl of IP buffer per sample. 100 µl of protein A-agarose beads solution were added to each sample, incubated for 3 hrs at 4°C and centrifuged three min at 4,000 rpm. The pellet was kept, washed three times with 1 ml of IP buffer and centrifugation during 3 min at 4,000 rpm. The beads were resuspended in 100 µl of IP buffer and 50 µl of Red loading buffer, and incubated for 10 min at 95°C. The tubes were centrifuged during 3 min at 4,000 rpm and the supernatant transferred to new tubes ready to use.

## Immunoblot analysis

For immunoblot analysis, nitrocellulose membranes were incubated for 30 min with 5 ml of LI-COR blocking buffer, followed by incubation with the antibodies rabbit polyclonal HCF-$1_C$ (1:1000, Bethyl) and polyclonal anti-HCF-$1_N$ (N961; 1:2000) in 50% LI-COR blocking buffer and 50% PBST (PBS containing 0.1% Tween 20), at 4°C overnight. The membranes were washed four times in the same buffer and incubated with appropriate secondary antibody (donkey anti-rabbit used at a dilution 1:10,000) in 50% LI-COR blocking buffer and 50% PBST at room temperature for 1 hr. The membranes were extensively washed in PBS containing 1% Tween20 and scanned with an Odyssey infrared imager (LI-COR).

## Quantitative PCR

Quantitative PCR has been performed using the following primers:
1) DENR F (5'-TCG CGA CAT CCT CCT AGC-3') and DENR R (5'-ACT CCT CTT CCC GGC CAC-3')
2) PNKD_1 F (5'-GAG CCC AAG CTC CAC CTT AC-3') and PNKD_1 R (5'-CTT CCG TCC TGC CTA CAG AG-3')
3) PNKD_2 F (5'-CAG CTG AAG GGA AAC AAG C-3') and PNKD_2 R (5'-CCG TCG GGA ATT GTA GTT TG-3')
4) QRICH1 F (5'-GTT CCC TTT AAG CCT GGC AG-3') and QRICH1 R (5'-CTT ACA TTC AGC CCT CAG GG-3')

5) SYT3 F (5'-CCA TAT TCT GCA GAC CGG AG-3') and SYT3 R (5'-GCT GGT CTT GCT GCT AGT GC-3')

Quantitative PCR on ChIP samples was performed using KAPA SYBR fast qPCR universal kit from Kapa Biosystems and a Rotorgene RG300A sequence detector (Corbett Research). The ChIP samples were normalized with the total Input DNA amount using the ΔCt method as in [Tyagi and Herr, 2009].

## Chromatin immunoprecipitation from mouse liver chromatin

Protocol adapter from Le Martelot et al. (2012). For ChIP, individual mice livers were used per time point. Two biological replicates were prepared for each replicate. Full livers or livers post-PH from mice were used for ChIP. Livers were immediately homogenized with a dounce in 4 ml per liver of 1x PBS including 1% formaldehyde, and the homogenate was kept for 5 min at room temperature. Cross-linking reactions were stopped by the addition of 25 ml of ice-cold 2.2 M sucrose buffer (150 mM glycine, 10 mM HEPES pH 7.6, 15 mM KCl, 2 mM EDTA, 0.15 mM spermine, 0.5 mM spermidine, 0.5 mM DTT and 0.5 mM PMSF). The homogenate was layered on top of a 10 ml cushion of 2.05 M sucrose (containing the same ingredients and including 10% glycerol and 125 mM glycine) and centrifuged for 1 hr at 24,000 rpm (100,000 g) at 4 °C in a Beckmann SW28 rotor. The nuclei were resuspended in 1.4 ml of ice-cold Buffer A (20 mM Tris, pH 7.5, 150 mM NaCl, 2 mM EDTA), transferred to a 1.5-ml centrifuge tube and sedimented at 2000 rpm in a benchtop centrifuge for 30 sec. The pellet was kept and resuspended in Buffer A and sedimented similarly a second time. Nuclei from three livers were pooled for all time points. Fort this, the nuclei were resuspended in 1.2 ml per liver of Nuclear Lysis Buffer (50 mM Tris, pH 8.1, 10 mM EDTA, 1% SDS, 0.15 mM spermine, 0.5 mM spermidine, 0.5 mM DTT and 0.5 mM PMSF), and sonicated with a Bioruptor-Pico from Diagenode for 9 rounds of 30 seconds of sonication and 90 seconds off. Chromatin has been diluted with 4 volumes of Immunoprecipitation Dilution Buffer (20mM Tris-HCl pH 8.1, 150 mM NaCl, 2mM EDTA, 1% Triton X-100, 0.01% SDS, 50 μg/ml PMSF, 1 μg/ml leupeptin, 0.15 mM, spermine, 0.5 mM spermidine). Pre-clearing of the chromatins was done by adding 100 μl of pre-immune serum, incubation for 1 hr at 4°C on a rotating wheel and adding 200 μl of a homogeneous protein A-agarose suspension (100 μl bed volume). This was incubated for 3 hrs and centrifuged in a microfuge during 2 minutes at a speed of 3,000 rpm. The supernatant was collected and DNA content was quantified. The concentration of fragmented cross-linked chromatin was adjusted with a mix 1:4 of NLB:IPDB. 5% of human chromatin was added (HeLa spinner cells) as a control for ChIP-seq. Chromatins were incubated overnight at 4°C on a rotating wheel with the following antibodies: anti-RPB2 (Santa Cruz Biotechnology, sc-673-18), anti-H3K4me3 (Abcam, ab8580), and anti-H3K36me3 (Abcam, ab9050). For chromatin immunoprecipitation, 40 μl of protein A bead suspension were added to the

chromatin and incubated for 3 hrs at 4°C on a rotating wheel. The beads were then washed twice with 1ml of IPWB1 (20 mM Tris-HCl pH 8.1, 50 mM NaCl, 2 mM EDTA, 1% Triton X-100, 0.1% SDS) and centrifuged 2 min at 3,000 rpm at 4°C. Later the beads were washed with 1 ml of IPWB2 (10 mM Tris-HCL pH 8.1, 250 mM LiCl, 1 mM EDTA, 1% NP-40, 1% sodium deoxycholate) and centrifuged for 2 minutes at 3,000 rpm at 4°C. The beads were also washed twice with 1 ml of 1x TE pH 8.0 and centrifuged for 2 minutes at 3,000 rpm at 4°C. Protein–DNA complexes were eluted from the beads, de-cross-linked, and treated with RNase A and, subsequently, with proteinase K, as described [O'Geen et al., 2006]. The DNA concentration was determined by fluorometry on the Qubit system (Invitrogen). A total of 10 ng DNA were used for each library preparation.

## ChIP-seq library preparation and ultra-high-throughput sequencing

10 ng of immunoprecipitated material was used for sequencing library preparation. Total input was also prepared for sequencing at each time point and replicate. Paired-end sequencing libraries were prepared with the 'MicroPlex Library Preparation v2' kit, from Diagenode, following the manufacturer's instructions (Diagenode, catalog no C05010013). No size selection was done and the chromatin was amplified with 12 cycles of PCR. 50 nucleotides at the fragments ends were sequenced with paired-end sequencing technology from HiSeq 2500 (Illumina).

## ChIP-seq data preparation and quantification

On the libraries containing mouse and human chromatin, the extreme 50 bp of the sequenced paired-ends were mapped onto the mouse (mm9) and human genome (hg19) with Elandv2e. On those libraries containing only mouse chromatin, mapping was only done on the mouse genome. Only fragments whose two reads had good sequencing quality (i.e., quality of the first 25 bp lower than Q20, as specified by Illumina) and mapping on unique genomic locations were kept. For further analysis fragment sizes between 50 and 500 bp were taken and only one copy of redundant fragments was kept. Good quality inputs (see Figure VI-10) from individual livers were pooled to have enough fragments for analysis.

To quantify the density of ChIPed epitopes the 50 bp in the center of the sequenced fragments were used. Different quantifications were done. On the one hand, HCF-1 binding was quantified in genomic bins of 200 bp in the libraries from HeLa spikes and in resting liver libraries. A similar procedure was done for the respective Inputs. For that, chromosomes were divided into genomic bins of 200 bp defined by sliding 100 bp, starting by the first nucleotide, and any incomplete bin left at the end of chromosomes was not used for analysis.

Moreover, mouse HCF-1 and Input quantifications were done in the promoter proximal region of +/- 250 bp around annotated TSS. For this purpose the Ensembl67/NCBI 37 annotation of transcription units was used.

All quantifications were scaled to the number of fragments analyzed per library. 500 pseudo-counts (i.e., 10 fragments with 50 bp displayed) were added to the quantifications to stabilize the variance in low scores. Finally, the log2 ratios between ChIP and Input quantifications were calculated. Among the multiple transcription units associated to a given gene, the one containing the maximum Pol2 occupancy during regeneration was taken (see chapter IV).

Transcription start sites closer than 1 kb in opposite directions were considered bidirectional promoters. Distinct combinations of TSS from a pair of genes were counted as one bidirectional promoter.

All the handling of the data was done with the UNIX shell, Perl and the R software [R Core Team, 2013].

## Genomic displays of ChIP-seq data

The CycliX viewer [Martin et al., unpublished] was used to visualize ChIP-seq tracks, especially AV2 files, which can be handled very efficiently on the viewer. Among other functions, the genomic viewer allows the selection of the desired genome stored in the NCBI database and the visualization of genomic data in a cumulative view. The density of accumulated data can also be scaled to the total number of fragments per library. And tracks can be displayed with any desired scale.

## Density profiles of HCF-1 accumulation on multiple genes at a one nucleotide resolution

Nucleotide positions around the TSS of selected transcripts were taken. All the genomic regions taken were put together from 5' to 3'. All the handling of genomic regions was done on the UNIX shell.

Densities of the central 50 bp of sequenced fragments were quantified at the previously defined nucleotide positions. A matrix of quantifications per TSS was built and analyzed on the R statistical software [R Core Team, 2013]. Z-scores were calculated from the quantifications and displayed with the heatmap.2 function included on the 'gplots 2.14.1' [Warnes et al., 2014]. Further, the z-scores at each TSS were averaged and displayed as lines with the R software [R core Team, 2013].

# Chapter VII: Conclusions and reflections

# Conclusions

Mammalian transcription is a highly complex mechanism where many diverse elements participate to allow the reading of the genetic program. In this thesis I have investigated different aspects of mammalian gene transcription in both cancerous HeLa cells and the mouse liver, and I had the privilege to do this research by integrating computational and experimental approaches. In this chapter, I present what I have learnt in my research put into a larger context and that are a part of manuscripts currently in preparation. There are two major publications that will arise from my doctoral research: a methodological manuscript including the studies of paired-end sequencing data in Chapter II, and a manuscript about the transcriptional responses to partial hepatectomy in the mouse liver included in Chapter IV. Nevertheless, I am also involved in the preparation of other co-authorship publications: one is a methodological article about redundancy and the second is about the definition of responses to PH.

## Towards improved ways to analyze the association of proteins to the chromatin

Chromatin immunoprecipitation techniques have been used since 1983 for the detection of chromatin association of proteins in vivo. A big step on the research of these interactions has been done in the early 21st century with the possibility to sequence the pieces of DNA where proteins associate and thus the identification of genome-wide interaction sites. The technique for ChIP followed by sequencing (ChIP-seq) has been progressively improved. For example, the library preparation step allows a better recovery of data although it can also introduce biases in the final data. During library preparation PCR amplification is a necessary step to achieve enough coverage after sequencing but can create redundancy in the final data that is not equally distributed among ChIP-ed fragments. This in turn introduces biases in the downstream quantification analysis of the fragment densities. In this work I have shown that the paired-end mode of sequencing offers advantages compared to the single-end mode during the in-silico analysis of the ChIP-seq data. Within paired-end data the redundancy can be precisely identified and controlled, hence diminishing the effects of unequal PCR amplification. Additionally, I have also described new opportunities for downstream analysis of the recovered paired-end fragments (e.g., in-silico fragment size classification) to better identify and characterize the sites of interaction in the chromatin. This has provided insights about the nucleosome positioning in promoters and about the multiple binding positions of the co-transcriptional factor HCF-1 within the same promoter. Thus, paired-end sequencing offers advantages compared to single-end sequencing for ChIP-seq.

Other techniques exist that aim similar purposes as ChIP-seq. One alterative is the so-called MNase-seq technique that is widely used for the study of nucleosome positioning. In this case chromatin samples are treated with the endo-exonuclease Micrococcal Nucleoase (MNase) that preferentially digests single-stranded and exposed DNA and thus DNA compacted around nucleosomes can be preferentially recovered. With this technique, however, some bias is introduced in the analysis especially due to the underlying DNA sequence and the way reactions are done [Meyer and Liu, 2014]. Another alternative technique is ChIP-exo that appeared recently, in 2011 [Rhee et al., 2011]. This technique improves the identification of precise binding sites that ChIP-seq cannot reach. ChIP-exo takes advantage of nucleases to digest the nucleic acids surrounding binding sites, thus, it has the potential to identify binding sites at 20-95 bp resolution. Nevertheless, this technique right now is not so well established and its performance seems to vary across transcription factors [Source: SEQanswers online forum]. Furthermore, it is also debatable how useful these two techniques can be for further analysis of the density of sequenced reads as the protocols also include library preparation with PCR amplification steps. This, as for ChIP-seq, would introduce effects on the analysis of quantifications, although perhaps less in the case of ChIP-exo as the fragments may be of very similar sizes after digestion. It would be interesting to compare the performance of the use of ChIP-seq with paired-end sequencing and these alternative technologies.

## The transcriptional responses to partial hepatectomy

The healthy liver is an organ that plays an important role in the adaptation to inflammation and other conditions that modify energy demands of the organism [reviewed in Diehl and Rai, 1996]. For that, hepatocytes are very sensitive to signals generated under those conditions and they exhibit great plasticity. Extracellular signals can rapidly initiate responses within individual hepatocytes that modulate existing gene expression to adapt to the new environment [reviewed in Diehl and Rai, 1996]. Partial hepatectomy produces a huge change in the liver structure and function and, in part, this leads to very rapid gene expression responses by signal transduction, within minutes after resection [Su et al., 2002]. In this thesis I have shown that these very early responses can be performed by RNA polymerases already positioned at the promoter and by massive transcription that does not involve the deposition of histone marks as in the Saa cluster of genes. Indeed, the mouse liver is highly responsive to changes. Interestingly, these rapid responses are also observed in the Sham control mice, suggesting that these genes may be frequently activated. At least more frequently than the regeneration-specific genes, which have in many cases not been transcribed since a long period and thus may require more involved rearrangements to be activated in response to PH.

Partial hepatectomy leads to cell proliferation and division that is first observed in hepatocytes. By 36 hrs, S phase occurs in the hepatocytes followed by the M phase around 48 hrs, leading to cell division [Apte, 2015; Minocha et al., 2016]. Consistent with this timing, I have observed regeneration-specific responses at 48 hrs post-PH that in part include the transcription of cell cycle genes. Interestingly, not all of the genes activated in a regeneration-specific manner have been previously related to the cell division cycle, as they are involved in other pathways such as metabolic pathways (for example, glycerolipid metabolism). This could be explained by a heterogeneous adaptation of gene expression programs across hepatocytes. Indeed, in the late 90's a heterogeneous activity of hepatocytes in response to environmental cues was already described [for a review see Diehl and Rai, 1996; Gebhardt and Matz-Soja, 2014]. The structure of the lobule functional units contribute to these heterogeneity. Hepatocytes, depending on their position, may respond differently, as each conduct provides different signaling and nutrients to hepatic cells. In this work, I have observed the alteration of genes with diverse functions, including cell proliferation and metabolic pathways. Given the existing hepatocyte heterogeneity, perhaps these different functions are modified in different hepatocytes. Further research on the altered genes could give more details about the homeostatic and regenerative responses of the liver.

## Genome-wide characterization of H3K36me2 and H3K36me3 states

Histone methylation has been widely described across species and they are believed to be marks that do not modify the structure of the chromatin per se [for a review see Bannister and Kouzarides, 2011]. These marks can be written by methylases and can be read by downstream effectors that drive molecular processes such as gene transcription [Jenuwein and Allis, 2001]. In previous studies done in yeast and chicken, dimethylation and trimethylation of H3K36 were observed in similar genic positions towards the 3' end of genes [Bannister et al., 2005; Pokholok et al., 2005; Rao et al., 2005]. In contrast, in the work of [Bell et al., 2007] done in Drosophila and in this work done in mice and humans, it has been observed that these two marks accumulate at different positions; H3K36me2 tends to accumulate towards the 5' end of genes, whereas H3K36me3 tends to accumulate towards the 3' end. These results suggest potential similarities in the regulation of Pol2 elongation on genes in Drosophila and mammals.

The increase of H3K36me3 tends to occur by the 3' end of the first intron of expressed genes, which is the position where the first splicing event occurs in those genes. Interestingly, not all the expressed genes containing H3K36 methylation show a dependency of these marks with a splicing site. I could observe that this accumulation also exists in genes whose transcripts do not undergo splicing, suggesting that this mark may also be involved in other mechanisms of gene transcription.

## The cellular function of Host Cell Factor 1

Host Cell Factor 1 is a protein that is well conserved among metazoan species. Interestingly, the activity of HCF-1 is also well conserved during evolution as VP16 induced complex formation is observed in extracts from nematodes and insects as well as vertebrates [Kristie et al., 1989; Wilson et al., 1993a]. This conservation suggests that the function of HCF-1 in cells is important. Indeed, HCF-1 plays a relevant role during cell proliferation [Julien and Herr, 2003; Goto et al., 1997]. Consistent with that, the expression of HCF-1 is correlated with proliferating cells. The expression is higher in embryonic tissues where proliferation occurs for development, and is low in examined adult tissues, where cells are generally in a more quiescent state [Wilson et al., 1995a]. Further, in this work done in cycling HeLa cells, I could observe that HCF-1 binds to the chromatin especially between G1/S passage and M phase, where it has been shown to be essential [Julien and Herr, 2003; Goto et al., 1997].

Nevertheless, it is thought that HCF-1 is involved in functions other than regulation of cell division. Studies in worms show that HCF-1 is not essential for viability while it is in flies and mice [Lee et al., 2007; Rodriguez-Jato et al., 2011; Minocha et al., 2016], suggesting different roles of HCF-1 in different species. It seems to be also the case in the mouse liver. There, where it is very abundant [Shilpi Minocha, personal communication], I have shown that HCF-1 associates to promoter proximal regions of genes not necessarily regulating cell proliferation. HCF-1 also binds to genes associated with metabolic functions, which are essential physiological functions of the liver. It is not a unique case of a cell cycle regulator that is involved in the regulation of metabolism. Fajas and colleagues have shown that the cyclin-cdk-Rb-E2F1 pathway not only regulates the cell cycle but also adipogenesis; these transcriptional regulators are activated by insulin and glucose, even if cells are not proliferating [Fajas, 2013]. Interestingly, HCF-1 has been shown to participate in this pathway for G1/S transition. Perhaps, HCF-1 is also part of this mechanism that modulates metabolism.

# My journey from computational to experimental biology

In this section I share my experience through these five years of doctoral studies where I aimed at integrating experimental and computational biology. Perhaps this can inspire other students.

Almost five years ago I arrived to Lausanne to join Winship Herr's lab to pursue my doctoral studies in the Integrated Experimental and Computational Biology program (IECB). I was completely excited about this. Although I did my bachelor in computer sciences I was also captured by the beauty of biology. Perhaps Winship Herr remembers the first thing I said to him when we met on my first day which is that I was very excited about learning biology. Now that I look back in time, I realize that 'biology' means much more than concepts one can learn from a book, as I could think at that time.

During the first steps of my journey I was exposed to many new concepts and techniques that at first were difficult to digest. Thus, I started to build from the very basics. I attended lectures at the Faculty of Biology and Medicine on topics such as cell biology or regulation of gene expression. Learning about those topics helped me to better understand the projects I was involved in and the projects of the people around me. This in turn, allowed me to start asking more focused questions and to start looking at the data with a better understanding and more detail. This was a big and very important step for me towards a better involvement in the research I was doing.

From the computational point of view, the research I started to do required more dynamicity of the code than the kind of projects that I was involved in before. As a computer scientist, I learnt to develop well-structured software. In contrast, during my doctoral studies, I was involved in data mining projects where every day new scientific questions were asked or modification of previous analyses were done. Because reproducibility is very important in science, I adapted the way to work by keeping versions of scripts, results and parameters used to be able to retrieve them. In other terms, I created my computational lab book.

As part of the IECB doctoral program I could participate in many different courses organized by different institutions such as the Swiss Institute of Bioinformatics or the Staromics doctoral program from the Conference Universitaire de Suisse Occidentale (CUSO). These courses, especially oriented to doctoral students and post-docs, were an invaluable support that improved the way I participated in projects and gave me a wider perspective about topics related to my research. For example, I could attend courses and workshops on statistics or the analysis of data such as ChIP-seq and RNA-seq that provided me with more computational tools.

Half way into my journey, two years and a half after beginning my doctoral studies, I started to feel more confident and able to ask more relevant research questions. This coincided with the beginning of a new project to investigate transcriptional programs in response to partial hepatectomy in the mouse liver, integrated into the CycliX consortium. At that time, I could work closely with a very talented and motivated team composed of Nicolas Guex, Winship Herr, Viviane Praz and Dominic Villeneuve, and always with the support of Nouria Hernandez, coordinator of the CycliX consortium. That was a very creative and productive time that was coordinated with weekly meetings. The mouse liver proved to be a very nice system to investigate transcription in a healthy organism. And this opened many opportunities to start new projects. This encouraged me later to go to the wet lab and do my own experiments. I wanted to investigate further whether HCF-1 binds to the mouse liver chromatin and where, paying special attention to both subunits of the mature HCF-1. I was very fortunate that Winship Herr accepted the idea that a computer scientist learned to work in the lab and that I conducted my own experimental research. During this period, I was supervised by Fabienne Lammers, a very experienced technician in the lab, and especially by Dominic Villeneuve, also a very experienced technician with expertise in mouse work and multiple techniques to investigate gene expression, such as ChIP-seq. I could start by learning very basic skills such as pipetting, calculating concentrations or preparing buffers. And little by little, I built on top of that. With the support of protocols provided by Fabienne Lammers and Dominic Villeneuve, I could investigate HCF-1 by doing western blots, IPs, ChIP-qPCR and ChIP-seq, as well as taking care of mice that were entrained for light-dark and food access cycles of 12 hrs. On the one hand, this was a big challenge for me. I had to learn to organize myself to do experiments, which requires a lot of mental energy and adaptation for somebody who never worked at the bench. Nevertheless, this was indeed a very exciting and inspiring experience. I started to think about experiments in a more detailed way, which has been very important for me to better understand the projects I was involved in and to start asking and answering biological questions. In the end, all this experience has been very useful to learn to be an independent scientist. Importantly, the time I spent in the lab taught me how to better collaborate with biologists. I could see how experimental biologists work, which needs and challenges they have, and I started to interact more deeply with people in the lab and labs in the same department. This is also an important step towards the integration of computational and experimental biologists.

Nowadays the research of life sciences requires an integration of experimental and computational approaches. Thus the opportunity that the IECB doctoral program gave me and other students is very useful I believe for the future of the life sciences. To finish, I just would like to encourage students with computational or experimental backgrounds to join programs like this. This experience is a platform that opens many exciting research possibilities.

# References


- Ajuh, P.M., Browne, G.J., Hawkes, N.A., Cohen, P.T.W., Roberts, S.G.E., and Lamond, A.I. (2000). Association of a protein phosphatase 1 activity with the human factor C1 (HCF) complex. Nucleic Acids Res. 28, 678–686.
- Anders, S., McCarthy, D.J., Chen, Y., Okoniewski, M., Smyth, G.K., Huber, W., and Robinson, M.D. (2013). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat. Protoc. 8, 1765–1786.
- Apte, U.M. (2015). Liver Regeneration: Basic Mechanisms, Relevant Models and Clinical Applications (Academic Press).
- Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M., and Teichmann, S.A. (2004). Structure and evolution of transcriptional regulatory networks. Curr. Opin. Struct. Biol. 14, 283–291.
- Bähler, J. (2005). Cell-cycle control of gene expression in budding and fission yeast. Annu. Rev. Genet. 39, 69–94.
- Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. 2, 28–36.
- Bannister, A.J., Schneider, R., Myers, F.A., Thorne, A.W., Crane-Robinson, C., and Kouzarides, T. (2005). Spatial distribution of di-and tri-methyl lysine 36 of histone H3 at active genes. J. Biol. Chem. 280, 17732–17736.
- Bannister, A.J., and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. Cell Res. 21, 381–395.
- Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. Cell 129, 823–837.
- Bell, O., Wirbelauer, C., Hild, M., Scharf, A.N.D., Schwaiger, M., MacAlpine, D.M., Zilbermann, F., van Leeuwen, F., Bell, S.P., Imhof, A., et al. (2007). Localized H3K36 methylation states define histone H4K16 acetylation during transcriptional elongation in Drosophila. EMBO J. 26, 4974–4984.
- Bentley, D.L. (2014). Coupling mRNA processing with transcription in time and space. Nat. Rev. Genet. 15, 163–175.
- Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas, E.J., Gingeras, T.R., et al. (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. Cell 120, 169–181.

- Berridge, M.J. (2012). Cell Signalling Biology: Module 9 - Cell Cycle and Proliferation. Biochem. J. 1–42.
- Bonhoure, N., Bounova, G., Bernasconi, D., Praz, V., Lammers, F., Canella, D., Willis, I.M., Herr, W., Hernandez, N., Delorenzi, M., et al. (2014). Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. Genome Res. 24, 1157–1168.
- Bouwens, L., Baekeland, M., de Zanger, R., and Wisse, E. (1986). Quantitation, tissue distribution and proliferation kinetics of Kupffer cells in normal rat liver. Hepatology 6, 718–722.
- Briggs, S.D., Bryk, M., Strahl, B.D., Cheung, W.L., Davie, J.K., Dent, S.Y.R., Winston, F., and David Allis, C. (2001). Histone H3 lysine 4 methylation is mediated by Set1 and required for cell growth and rDNA silencing in Saccharomyces cerevisiae. Genes Dev. 15, 3286–3295.
- Cai, Y., Jin, J., Swanson, S.K., Cole, M.D., Choi, S.H., Florens, L., Washburn, M.P., Conaway, J.W., and Conaway, R.C. (2010). Subunit composition and substrate specificity of a MOF-containing histone acetyltransferase distinct from the male-specific lethal (MSL) complex. J. Biol. Chem. 285, 4268–4272.
- Capotosti, F., Hsieh, J.J.-D., and Herr, W. (2007). Species selectivity of mixed-lineage leukemia/trithorax and HCF proteolytic maturation pathways. Mol. Cell. Biol. 27, 7063–7072.
- Capotosti, F., Guernier, S., Lammers, F., Waridel, P., Cai, Y., Jin, J., Conaway, J.W., Conaway, R.C., and Herr, W. (2011). O-GlcNAc transferase catalyzes site-specific proteolysis of HCF-1. Cell 144, 376–388.
- Clemson, C.M., Hutchinson, J.N., Sara, S.A., Ensminger, A.W., Fox, A.H., Chess, A., and Lawrence, J.B. (2009). An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Mol. Cell 33, 717–726.
- Conaway, R., and Conaway, J. (1997). General transcription factors for RNA polymerase h. Prog Nucleic Acid Res Mol Biol 56, 327N346.
- Conaway, J.W., and Conaway, R.C. (1999). Transcription elongation and human disease. Annu. Rev. Biochem. 68, 301–319.
- Cook, P.R. (1999). The organization of replication and transcription. Science (80-. ). 284, 1790–1795.
- Corona, D.F. V, Clapier, C.R., Becker, P.B., and Tamkun, J.W. (2002). Modulation of ISWI function by site-specific histone acetylation. EMBO Rep. 3, 242–247.
- Cramer, P., Bushnell, D.A., Fu, J., Gnatt, A.L., Maier-Davis, B., Thompson, N.E., Burgess, R.R., Edwards, A.M., David, P.R., and Kornberg, R.D. (2000). Architecture of RNA polymerase II and implications for the transcription mechanism. Science 288, 640–649.
- Cressman, D.E., Diamond, R.H., and Taub, R. (1995). Rapid activation of the Stat3 transcription complex in liver regeneration. Hepatology 21, 1443–1449.

- Daou, S., Mashtalir, N., Hammond-Martel, I., Pak, H., Yu, H., Sui, G., Vogel, J.L., Kristie, T.M., and Affar, E.B. (2011). Crosstalk between O-GlcNAcylation and proteolytic cleavage regulates the host cell factor-1 maturation pathway. Proc. Natl. Acad. Sci. 108, 2747–2752.
- Dejosez, M., Krumenacker, J.S., Zitur, L.J., Passeri, M., Chu, L.-F., Songyang, Z., Thomson, J.A., and Zwaka, T.P. (2008). Ronin is essential for embryogenesis and the pluripotency of mouse embryonic stem cells. Cell 133, 1162–1174.
- Dejosez, M., Levine, S.S., Frampton, G.M., Whyte, W.A., Stratton, S.A., Barton, M.C., Gunaratne, P.H., Young, R.A., and Zwaka, T.P. (2010). Ronin/Hcf-1 binds to a hyperconserved enhancer element and regulates genes involved in the growth of embryonic stem cells. Genes Dev. 24, 1479–1484.
- Diehl, A.M., and Rai, R.M. (1996). Liver regeneration 3: Regulation of signal transduction during liver regeneration. FASEB J. 10, 215–227.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21.
- Dorigo, B., Schalch, T., Bystricky, K., and Richmond, T.J. (2003). Chromatin fiber folding: requirement for the histone H4 N-terminal tail. J. Mol. Biol. 327, 85–96.
- Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. Bioinformatics 21, 3439–3440.
- Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. 4, 1184–1191.
- ENCODE Project Consortium, and others (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.
- Elaine R. Mardis. (2012). GENOMICS/DNA SEQUENCING: DNA sequencing technologies: The next generation and beyond. Website BioOptics World.
- Fajas, L. (2013). Re-thinking cell cycle regulators: the cross-talk with metabolism. Front. Oncol 3.
- Fausto, N. (2000). Liver regeneration. J. Hepatol. 32, 19–31.
- Fausto, N. (2006). Involvement of the innate immune system in liver regeneration and injury. J. Hepatol. 45, 347–349.
- Fnu, S., Williamson, E.A., De Haro, L.P., Brenneman, M., Wray, J., Shaheen, M., Radhakrishnan, K., Lee, S.-H., Nickoloff, J.A., and Hromas, R. (2011). Methylation of histone H3 lysine 36 enhances DNA repair by nonhomologous end-joining. Proc. Natl. Acad. Sci. 108, 540–545.
- Fox, A.H., Lam, Y.W., Leung, A.K.L., Lyon, C.E., Andersen, J., Mann, M., and Lamond, A.I. (2002). Paraspeckles: a novel nuclear domain. Curr. Biol. 12, 13–25.

- Freiman, R.N., and Herr, W. (1997). Viral mimicry: Common mode of association with HCF by VP16 and the cellular protein LZIP. Genes Dev. 11, 3122–3127.
- Gatfield, D., Le Martelot, G., Vejnar, C.E., Gerlach, D., Schaad, O., Fleury-Olela, F., Ruskeepää, A.-L., Oresic, M., Esau, C.C., Zdobnov, E.M., et al. (2009). Integration of microRNA miR-122 in hepatic circadian gene expression. Genes Dev. 23, 1313–1326.
- Gebhardt, R., and Matz-Soja, M. (2014). Liver zonation: Novel aspects of its regulation and its impact on homeostasis. World J. Gastroenterol. WJG 20, 8491–8504.
- Gonzales D, Lammers F, Galliot B and Herr W, unpublished results
- Gordillo, M., Evans, T., and Gouon-Evans, V. (2015). Orchestrating liver development. Development 142, 2094–2108.
- Goto, H., Motomura, S., Wilson, A.C., Freiman, R.N., Nakabeppu, Y., Fukushima, K., Fujishima, M., Herr, W., and Nishimoto, T. (1997). A single-point mutation in HCF causes temperature-sensitive cell-cycle arrest and disrupts VP16 function. Genes Dev. 11, 726–737.
- Gregory R. Warnes, Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz and Bill Venables (2014). gplots: Various R programming tools for plotting data. R package version 2.14.1. http://CRAN.R-project.org/package=gplots
- Grzybowski, A.T., Chen, Z., and Ruthenburg, A.J. (2015). Calibrating ChIP-Seq with nucleosomal internal standards to measure histone modification density genome wide. Mol. Cell 58, 886–899.
- Gunther, M., Laithier, M., and Brison, O. (2000). A set of proteins interacting with transcription factor Sp1 identified in a two-hybrid screening. Mol. Cell. Biochem. 210, 131–142.
- Hahn, S. (2004). Structure and mechanism of the RNA polymerase II transcription machinery. Nat. Struct. Mol. Biol. 11, 394–403.
- Hart, T., Komori, H.K., LaMere, S., Podshivalova, K., and Salomon, D.R. (2013). Finding the active genes in deep RNA-seq gene expression studies. BMC Genomics 14, 778.
- Hirose, Y., and Manley, J.L. (2000). RNA polymerase II and the integration of nuclear events. Genes Dev. 14, 1415–1429.
- Hirose, Y., and Ohkuma, Y. (2007). Phosphorylation of the C-terminal domain of RNA polymerase II plays central roles in the integrated events of eucaryotic gene expression. J. Biochem. 141, 601–608.
- Hu, B., Petela, N., Kurze, A., Chan, K.-L., Chapard, C., and Nasmyth, K. (2015). Biological chromodynamics: a general method for measuring protein occupancy across the genome by calibrating ChIP-seq. Nucleic Acids Res. 43, e132–e132.

- Hu, M., Sun, X.-J., Zhang, Y.-L., Kuang, Y., Hu, C.-Q., Wu, W.-L., Shen, S.-H., Du, T.-T., Li, H., He, F., et al. (2010). Histone H3 lysine 36 methyltransferase Hypb/Setd2 is required for embryonic vascular remodeling. Proc. Natl. Acad. Sci. 107, 2956–2961.
- Hu, Z., Chen, K., Xia, Z., Chavez, M., Pal, S., Seol, J.-H., Chen, C.-C., Li, W., and Tyler, J.K. (2014). Nucleosome loss leads to global transcriptional up-regulation and genomic instability during yeast aging. Genes Dev. 28, 396–408.
- Huang, J., and Rudnick, D.A. (2014). Elucidating the metabolic regulation of liver regeneration. Am. J. Pathol. 184, 309–321.
- Huff, J.T., Plocik, A.M., Guthrie, C., and Yamamoto, K.R. (2010). Reciprocal intronic and exonic histone modification regions in humans. Nat. Struct. Mol. Biol. 17, 1495–1499.
- Jackson, D.A., Hassan, A.B., Errington, R.J., and Cook, P.R. (1993). Visualization of focal sites of transcription within human nuclei. EMBO J. 12, 1059.
- Jenuwein, T., and Allis, C.D. (2001). Translating the histone code. Science. 293, 1074–1080.
- Jothi, R., Cuddapah, S., Barski, A., Cui, K., and Zhao, K. (2008). Genome-wide identification of in vivo protein--DNA binding sites from ChIP-Seq data. Nucleic Acids Res. 36, 5221–5231.
- Julien, E., and Herr, W. (2003). Proteolytic processing is necessary to separate and ensure proper cell growth and cytokinesis functions of HCF-1. EMBO J. 22, 2360–2369.
- Julien, E., and Herr, W. (2004). A switch in mitotic histone H4 lysine 20 methylation status is linked to M phase defects upon loss of HCF-1. Mol. Cell 14, 713–725.
- Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S.L., and others (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14, R36.
- Kim, S., Kim, H., Fong, N., Erickson, B., and Bentley, D.L. (2011). Pre-mRNA splicing is a determinant of histone H3K36 methylation. Proc. Natl. Acad. Sci. 108, 13564–13569.
- Kim, T., and Buratowski, S. (2007). Two Saccharomyces cerevisiae JmjC domain proteins demethylate histone H3 Lys36 in transcribed regions to promote elongation. J. Biol. Chem. 282, 20827–20835.
- Kizer, K.O., Phatnani, H.P., Shibata, Y., Hall, H., Greenleaf, A.L., and Strahl, B.D. (2005). A novel domain in Set2 mediates RNA polymerase II interaction and couples histone H3 K36 methylation with transcript elongation. Mol. Cell. Biol. 25, 3305–3316.
- Knez, J., Piluso, D., Bilan, P., and Capone, J.P. (2006). Host cell factor-1 and E2F4 interact via multiple determinants in each protein. Mol. Cell. Biochem. 288, 79–90.
- Kobayashi, T., and Horiuchi, T. (1996). A yeast gene product, Fob1 protein, required for both replication fork blocking and recombinational hotspot activities. Genes to Cells 1, 465–474.

- Koj, A. (1974). Acute-phase reactants. In Structure and Function of Plasma Proteins, (Springer), pp. 73–131.
- Kooistra, S.M., and Helin, K. (2012). Molecular mechanisms and potential functions of histone demethylases. Nat. Rev. Mol. Cell Biol. 13, 297–311.
- Kornberg, R.D. (1999). Eukaryotic transcriptional control. Trends Biochem. Sci. 24.
- Krishnamurthy, S., and Hampsey, M. (2009). Eukaryotic transcription initiation. Curr. Biol. 19.
- Kristie, T.M., Pomerantz, J.L., Twomey, T.C., Parent, S.A., and Sharp, P.A. (1995). The cellular C1 factor of the herpes simplex virus enhancer complex is a family of polypeptides. J. Biol. Chem. 270, 4387–4394.
- Kristie, T.M. (1997). The Mouse Homologue of the Human Transcription Factor C1 (Host Cell Factor) CONSERVATION OF FORMS AND FUNCTION. J. Biol. Chem. 272, 26749–26755.
- Kristie, T.M., LeBowitz, J.H., and Sharp, P.A. (1989). The octamer-binding proteins form multi-protein--DNA complexes with the HSV alpha TIF regulatory protein. EMBO J. 8, 4229.
- Krogan, N.J., Kim, M., Tong, A., Golshani, A., Cagney, G., Canadien, V., Richards, D.P., Beattie, B.K., Emili, A., Boone, C., et al. (2003). Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II. Mol. Cell. Biol. 23, 4207–4218.
- Kushner, I. (1982). THE PHENOMENON OF THE ACUTE PHASE RESPONSE. Ann. N. Y. Acad. Sci. 389, 39–48.
- Kutmon, M., Riutta, A., Nunes, N., Hanspers, K., Willighagen, E.L., Bohler, A., Mélius, J., Waagmeester, A., Sinha, S.R., Miller, R., et al. (2015). WikiPathways: capturing the full diversity of pathway knowledge. Nucleic Acids Res. gkv1024.
- Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.
- Le Martelot, G., Canella, D., Symul, L., Migliavacca, E., Gilardi, F., Liechti, R., Martin, O., Harshman, K., Delorenzi, M., Desvergne, B., et al. (2012). Genome-wide RNA polymerase II profiles and RNA accumulation reveal kinetics of transcription and associated epigenetic changes during diurnal cycles. PLoS Biol 10, e1001442.
- Lee, J.S., and Shilatifard, A. (2007). A site to remember: H3K36 methylation a mark for histone deacetylation. Mutat. Res. - Fundam. Mol. Mech. Mutagen. 618, 130–134.
- Lee, S., Horn, V., Julien, E., Liu, Y., Wysocka, J., Bowerman, B., Hengartner, M.O., and Herr, W. (2007). Epigenetic regulation of histone H3 serine 10 phosphorylation status by HCF-1 proteins in C. elegans and mammalian cells. PLoS One 2, e1213.
- Lee, T.I., and Young, R.A. (2000). TRANSCRIPTION OF EUKARYOTIC PROTEIN -CODING GENES. Annu. Rev. Genet. 34, 77–137.

- Li, B., Howe, L., Anderson, S., Yates, J.R., and Workman, J.L. (2003). The Set2 histone methyltransferase functions through the phosphorylated carboxyl-terminal domain of RNA polymerase II. J. Biol. Chem. 278, 8897–8903.
- Li, B., Carey, M., and Workman, J.L. (2007). The Role of Chromatin during Transcription. Cell 128, 707–719.
- Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 1.
- Li, H., Ilin, S., Wang, W., Duncan, E.M., Wysocka, J., Allis, C.D., and Patel, D.J. (2006). Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. Nature 442, 91–95.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and others (2009). The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079.
- Li, M., Phatnani, H.P., Guan, Z., Sage, H., Greenleaf, A.L., and Zhou, P. (2005). Solution structure of the Set2--Rpb1 interacting domain of human Set2 and its interaction with the hyperphosphorylated C-terminal domain of Rpb1. Proc. Natl. Acad. Sci. U. S. A. 102, 17636–17641.
- Li, W., Liang, X., Kellendonk, C., Poli, V., and Taub, R. (2002). STAT3 contributes to the mitogenic response of hepatocytes during liver regeneration. J. Biol. Chem. 277, 28411–28417.
- Li, Z.-Y., Xi, Y., Zhu, W.-N., Zeng, C., Zhang, Z.-Q., Guo, Z.-C., Hao, D.-L., Liu, G., Feng, L., Chen, H.-Z., et al. (2011). Positive regulation of hepatic miR-122 expression by HNF4α. J. Hepatol. 55, 602–611.
- Liang, Y., Vogel, J.L., Narayanan, A., Peng, H., and Kristie, T.M. (2009). Inhibition of the histone demethylase LSD1 blocks $α$-herpesvirus lytic replication and reactivation from latency. Nat. Med. 15, 1312–1317.
- Liao, Y., Smyth, G.K., and Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. 41, e108–e108.
- Lin, J., Puigserver, P., Donovan, J., Tarr, P., and Spiegelman, B.M. (2002). Peroxisome proliferator-activated receptor γ coactivator 1β (PGC-1β), a novel PGC-1-related transcription coactivator associated with host cell factor. J. Biol. Chem. 277, 1645–1648.
- Liu, W., Tanasa, B., Tyurina, O. V, Zhou, T.Y., Gassmann, R., Liu, W.T., Ohgi, K.A., Benner, C., Garcia-Bassets, I., Aggarwal, A.K., et al. (2010). PHF8 mediates histone H4 lysine 20 demethylation events involved in cell cycle progression. Nature 466, 508–512.
- Lowell, C.A., Potter, D.A., Stearman, R.S., and Morrow, J.F. (1986). Structure of the murine serum amyloid A gene family. Gene conversion. J. Biol. Chem. 261, 8442–8452.

- Lu, R., Yang, P., Padmakumar, S., and Misra, V. (1998). The herpesvirus transactivator VP16 mimics a human basic domain leucine zipper protein, luman, in its interaction with HCF. J. Virol. 72, 6291–6297.
- Lu, R., and Misra, V. (2000). Zhangfei: a second cellular protein interacts with herpes simplex virus accessory factor HCF in a manner similar to Luman and VP16. Nucleic Acids Res. 28, 2446–2454.
- Luciano, R.L., and Wilson, A.C. (2003). HCF-1 functions as a coactivator for the zinc finger protein Krox20. J. Biol. Chem. 278, 51116–51124.
- Luger, K., Mäder, a W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 389, 251–260.
- Machida, Y.J., Machida, Y., Vashisht, A.A., Wohlschlegel, J.A., and Dutta, A. (2009). The deubiquitinating enzyme BAP1 regulates cell growth via interaction with HCF-1. J. Biol. Chem. 284, 34179–34188.
- Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. (2013). cluster: Cluster Analysis Basics and Extensions. R package version 1.14.4.
- Maeno, H., Ono, T., Dhar, D.K., Sato, T., Yamanoi, A., and Nagasue, N. (2005). Expression of hypoxia inducible factor-1α during liver regeneration induced by partial hepatectomy in rats. Liver Int. 25, 1002–1009.
- Mahajan, S.S., Little, M.M., Vazquez, R., and Wilson, A.C. (2002). Interaction of HCF-1 with a cellular nuclear export factor. J. Biol. Chem. 277, 44292–44299.
- Malik, R., Selden, C., and Hodgson, H. (2002). The role of non-parenchymal cells in liver growth. In Seminars in Cell & Developmental Biology, pp. 425–431.Martin, C., and Zhang, Y. (2005). The diverse functions of histone lysine methylation. Nat. Rev. Mol. Cell Biol. 6, 838–849.
- Marc Carlson. org.Mm.eg.db: Genome wide annotation for Mouse. R package version 2.10.1
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. 17, pp – 10.
- Mazars, R., Gonzalez-de-Peredo, A., Cayrol, C., Lavigne, A.-C., Vogel, J.L., Ortega, N., Lacroix, C., Gautier, V., Huet, G., Ray, A., et al. (2010). The THAP-Zinc Finger Protein THAP1 Associates with Coactivator HCF-1 and O-GlcNAc Transferase A LINK BETWEEN DYT6 AND DYT3 DYSTONIAS. J. Biol. Chem. 285, 13364–13371.
- Metzker, M.L. (2010). Sequencing technologies – the next generation. Nat. Rev. Genet. 11, 31–46.
- Meyer, C.A., and Liu, X.S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nat. Rev. Genet. 15, 709–721.

- Michaud, J., Praz, V., Faresse, N.J., JnBaptiste, C.K., Tyagi, S., Schütz, F., and Herr, W. (2013). HCFC1 is a common component of active human CpG-island promoters and coincides with ZNF143, THAP11, YY1, and GABP transcription factor occupancy. Genome Res. 23, 907–916.
- Minocha, S., Sung, T.-L., Villeneuve, D., Lammers, F., and Herr, W. (2016). Compensatory embryonic response to allele-specific inactivation of the murine X-linked gene Hcfc1. Dev. Biol. 412, 1–17.
- Misaghi, S., Ottosen, S., Izrael-Tomasevic, A., Arnott, D., Lamkanfi, M., Lee, J., Liu, J., O'Rourke, K., Dixit, V.M., and Wilson, A.C. (2009). Association of C-terminal ubiquitin hydrolase BRCA1-associated protein 1 with cell cycle regulator host cell factor 1. Mol. Cell. Biol. 29, 2181–2192.
- Mitchell, C., and Willenbring, H. (2008). A reproducible and well-tolerated method for 2/3 partial hepatectomy in mice. Nat. Protoc. 3, 1167–1170.Morgan, D.O. (2007). The Cell Cycle: Principles of Control (New Science Press).
- Myers, L.C., and Kornberg, R.D. (2000). Mediator of transcriptional regulation. Annu. Rev. Biochem. 69, 729–749.
- Nagai, T., Matsumoto, N., Kurotaki, N., Harada, N., Niikawa, N., Ogata, T., Imaizumi, K., Kurosawa, K., Kondoh, T., Ohashi, H., et al. (2003). Sotos syndrome and haploinsufficiency of NSD1: clinical features of intragenic mutations and submicroscopic deletions. J. Med. Genet. 40, 285–289.
- Nakayasu, H., and Berezney, R. (1989). Mapping replicational sites in the eucaryotic cell nucleus. J. Cell Biol. 108, 1–11.
- Ng, H.H., Robert, F., Young, R.A., and Struhl, K. (2003). Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. Mol. Cell 11, 709–719.
- Ngondo-Mbongo, R.P., Myslinski, E., Aster, J.C., and Carbon, P. (2013). Modulation of gene expression via overlapping binding sites exerted by ZNF143, Notch1 and THAP11. Nucleic Acids Res. gkt088.
- Nimura, K., Ura, K., Shiratori, H., Ikawa, M., Okabe, M., Schwartz, R.J., and Kaneda, Y. (2009). A histone H3 lysine 36 trimethyltransferase links Nkx2-5 to Wolf--Hirschhorn syndrome. Nature 460, 287–291.
- Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 27, 29–34.
- O'Geen, H., Nicolet, C.M., Blahnik, K., Green, R., and Farnham, P.J. (2006). Comparison of sample preparation methods for ChIP-chip assays. Biotechniques 41, 577.
- Okitsu, C.Y., Hsieh, J.C.F., and Hsieh, C.-L. (2010). Transcriptional activity affects the H3K4me3 level and distribution in the coding region. Mol. Cell. Biol. 30, 2933–2946.

- Orlando, D.A., Chen, M.W., Brown, V.E., Solanki, S., Choi, Y.J., Olson, E.R., Fritz, C.C., Bradner, J.E., and Guenther, M.G. (2014). Quantitative ChIP-Seq normalization reveals global modulation of the epigenome. Cell Rep. 9, 1163–1170.
- Pandya-Jones, A. (2011). Pre-mRNA splicing during transcription in the mammalian system. Wiley Interdiscip. Rev. RNA 2, 700–717.
- Pasero, P., Bensimon, A., and Schwob, E. (2002). Single-molecule analysis reveals clustering and epigenetic regulation of replication origins at the yeast rDNA locus. Genes Dev. 16, 2479–2484.
- Piluso, D., Bilan, P., and Capone, J.P. (2002). Host cell factor-1 interacts with and antagonizes transactivation by the cell cycle regulatory factor Miz-1. J. Biol. Chem. 277, 46799–46808.
- Pokholok, D.K., Harbison, C.T., Levine, S., Cole, M., Hannett, N.M., Lee, T.I., Bell, G.W., Walker, K., Rolfe, P.A., Herbolsheimer, E., et al. (2005). Genome-wide map of nucleosome acetylation and methylation in yeast. Cell 122, 517–527.
- Pray-Grant, M.G., Daniel, J. a, Schieltz, D., Yates, J.R., and Grant, P. a (2005). Chd1 chromodomain links histone H3 methylation with SAGA- and SLIK-dependent acetylation. Nature 433, 434–438.
- Proudfoot, N.J. (2000). Connecting transcription to messenger RNA processing. Trends Biochem. Sci. 25, 290–293.
- Raiber, E.-A., Kranaster, R., Lam, E., Nikan, M., and Balasubramanian, S. (2012). A non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro. Nucleic Acids Res. 40, 1499–1508.
- Rao, B., Shibata, Y., Strahl, B.D., and Lieb, J.D. (2005). Dimethylation of histone H3 at lysine 36 demarcates regulatory and nonregulatory chromatin genome-wide. Mol. Cell. Biol. 25, 9447–9459.
- Renaud, M., Praz, V., Vieu, E., Florens, L., Washburn, M.P., l'Hôte, P., and Hernandez, N. (2014). Gene duplication and neofunctionalization: POLR3G and POLR3GL. Genome Res. 24, 37–51.
- Rhee, H.S., and Pugh, B.F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell 147, 1408–1419.
- Richmond, R.K., Sargent, D.F., Richmond, T.J., Luger, K., and Mader, A.W. (1997). Crystal structure of the nucleosome resolution core particle at 2.8 A. Nature 389, 251–260.
- Robinson, M.D., and Smyth, G.K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics 9, 321–332.
- Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140.
- Roeder, R.G. (1996). The role of general initiation factors in transcription by RNA polymerase II. Trends Biochem Sci 21, 327–335.

- Roguev, A., Schaft, D., Shevchenko, A., Pijnappel, W.W.M.P., Wilm, M., Aasland, R., and Stewart, A.F. (2001). The Saccharomyces cerevisiae Set1 complex includes an Ash2 homologue and methylates histone 3 lysine 4. EMBO J. 20, 7137–7148.
- Rousseeuw, J., P., and Kaufman, L. (1990). Finding groups in data (Wiley Online Library).
- Ruan, H.-B., Han, X., Li, M.-D., Singh, J.P., Qian, K., Azarhoush, S., Zhao, L., Bennett, A.M., Samuel, V.T., Wu, J., et al. (2012). O-GlcNAc transferase/host cell factor C1 complex regulates gluconeogenesis by modulating PGC-1α stability. Cell Metab. 16, 226–237.
- Rudnick, D.A., and Davidson, N.O. (2012). Functional relationships between lipid metabolism and liver regeneration. Int. J. Hepatol. 2012.
- Ruthenburg, A.J., Allis, C.D., and Wysocka, J. (2007). Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. Mol. Cell 25, 15–30.
- Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. 74, 5463–5467.
- Santos-Rosa, H., Schneider, R., Bannister, A.J., Sherriff, J., Bernstein, B.E., Emre, N.C.T., Schreiber, S.L., Mellor, J., and Kouzarides, T. (2002). Active genes are tri-methylated at K4 of histone H3. Nature 419, 407–411.
- Scarr, R.B., and Sharp, P.A. (2002). PDCD2 is a negative regulator of HCF-1 (C1). Oncogene 21, 5245–5254.
- Schaft, D., Roguev, A., Kotovic, K.M., Shevchenko, A., Sarov, M., Shevchenko, A., Neugebauer, K.M., and Stewart, A.F. (2003). The histone 3 lysine methyltransferase, SET2, is involved in transcriptional elongation. Nucleic Acids Res. 31, 2475–2482.
- Schatten, H., Schatten, G., Mazia, D., Balczon, R., and Simerly, C. (1986). Behavior of centrosomes during fertilization and cell division in mouse oocytes and in sea urchin eggs. Proc. Natl. Acad. Sci. U. S. A. 83, 105–109.
- Schiff, E.R., Sorrell, M.F., and Maddrey, W.C. (2007). Schiff's Diseases of the Liver (Lippincott Williams & Wilkins).
- Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863–864.
- Schneider, R., Bannister, A.J., Myers, F. a., Thorne, A.W., Crane-Robinson, C., and Kouzarides, T. (2004). Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. Nat. Cell Biol. 6, 73–77.
- Shi, Y., Lan, F., Matson, C., Mulligan, P., Whetstine, J.R., Cole, P.A., Casero, R.A., and Shi, Y. (2004). Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. Cell 119, 941–953.
- Shilatifard, A. (1998). Factors regulating the transcriptional elongation activity of RNA polymerase II. Faseb J 12, 1437–1446.

- Shilatifard, A. (2012). The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis. Annu. Rev. Biochem. 81, 65.
- Sims, R.J., Millhouse, S., Chen, C.-F., Lewis, B.A., Erdjument-Bromage, H., Tempst, P., Manley, J.L., and Reinberg, D. (2007). Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. Mol. Cell 28, 665–676.
- Smith, E.R., Cayrou, C., Huang, R., Lane, W.S., Côté, J., and Lucchesi, J.C. (2005). A human protein complex homologous to the Drosophila MSL complex is responsible for the majority of histone H4 acetylation at lysine 16. Mol. Cell. Biol. 25, 9175–9188.
- Shogren-Knaak, M., Ishii, H., Sun, J.-M., Pazin, M.J., Davie, J.R., and Peterson, C.L. (2006). Histone H4-K16 acetylation controls chromatin structure and protein interactions. Science (80-. ). 311, 844–847.
- Spector, D.L., Fu, X.-D., and Maniatis, T. (1991). Associations between distinct pre-mRNA splicing components and the cell nucleus. EMBO J. 10, 3467.
- Strey, C.W., Markiewski, M., Mastellos, D., Tudoran, R., Spruce, L.A., Greenbaum, L.E., and Lambris, J.D. (2003). The proinflammatory mediators C3a and C5a are essential for liver regeneration. J. Exp. Med. 198, 913–923.
- Su, A.I., Guidotti, L.G., Pezacki, J.P., Chisari, F. V, and Schultz, P.G. (2002). Gene expression during the priming phase of liver regeneration after partial hepatectomy in mice. Proc. Natl. Acad. Sci. 99, 11181–11186.
- Tal Galili (2014). dendextend: Extending R's dendrogram functionality. R package version 0.17.1. (http://CRAN.R-project.org/package=dendextend)
- Taub, R. (1996). Liver regeneration 4: transcriptional control of liver regeneration. FASEB J. 10, 413–427.
- Team, R.C. (2013). R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- The Gene Ontology Consortium (2015). Gene Ontology Consortium: going forward. Nucleic Acids Res. 43, D1049–D1056.
- Thomas, L.R., Foshage, A.M., Weissmiller, A.M., Popay, T.M., Grieb, B.C., Qualls, S.J., Ng, V., Carboneau, B., Lorey, S., Eischen, C.M., et al. (2015). Interaction of MYC with host cell factor-1 is mediated by the evolutionarily conserved Myc box IV motif. Oncogene.
- Tsukada, Y., Fang, J., Erdjument-Bromage, H., Warren, M.E., Borchers, C.H., Tempst, P., and Zhang, Y. (2006). Histone demethylation by a family of JmjC domain-containing proteins. Nature 439, 811–816.

- Tyagi, S., Chabes, A.L., Wysocka, J., and Herr, W. (2007). E2F activation of S phase promoters via association with HCF-1 and the MLL family of histone H3K4 methyltransferases. Mol. Cell 27, 107–119.
- Tyagi, S., and Herr, W. (2009). E2F1 mediates DNA damage and apoptosis through HCF-1 and the MLL family of histone methyltransferases. EMBO J. 28, 3185–3195.
- Ukai, K., Terashima, K., Imai, Y., Shinzawa, H., Okuyama, Y., Takahashi, T., and Ishikawa, M. (1990). Proliferation Kinetics of Rat Kupffer Cells after Partial Hepatectomy Immunohistochemical and UItrastructural Analysis. Pathol. Int. 40, 623–634.
- Vercauteren, K., Gleyzer, N., and Scarpulla, R.C. (2008). PGC-1-related coactivator complexes with HCF-1 and NRF-2β in mediating NRF-2 (GABP)-dependent respiratory gene expression. J. Biol. Chem. 283, 12102–12111.
- Vermeulen, M., Mulder, K.W., Denissov, S., Pijnappel, W.W.M.P., van Schaik, F.M.A., Varier, R.A., Baltissen, M.P.A., Stunnenberg, H.G., Mann, M., and Timmers, H.T.M. (2007). Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. Cell 131, 58–69.
- Vogel, J.L., and Kristie, T.M. (2000). The novel coactivator C1 (HCF) coordinates multiprotein enhancer formation and mediates transcription activation by GABP. EMBO J. 19, 683–690.
- Vogel, J.L., and Kristie, T.M. (2006). Site-specific proteolysis of the transcriptional coactivator HCF-1 can regulate its interaction with protein cofactors. Proc. Natl. Acad. Sci. 103, 6817–6822.
- Wagner, E.J., and Carpenter, P.B. (2012). Understanding the language of Lys36 methylation at histone H3. Nat. Rev. Mol. Cell Biol. 13, 115–126.
- Wang, J., Duncan, D., Shi, Z., and Zhang, B. (2013). WEB-based gene set analysis toolkit (WebGestalt): update 2013. Nucleic Acids Res. 41, W77–W83.
- Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63.
- Ward Jr, J.H. (1963). Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58, 236–244.
- Wei, X., Samarabandu, J., Devdhar, R.S., Siegel, A.J., Acharya, R., and Berezney, R. (1998). Segregation of transcription and replication sites into higher order domains. Science (80-. ). 281, 1502–1505.
- Wickham, H. (2009). ggplot2: elegant graphics for data analysis (Springer Science & Business Media).
- Widmann, J.-J., and Fahimi, H.D. (1975). Proliferation of mononuclear phagocytes (Kupffer cells) and endothelial cells in regenerating rat liver. A light and electron microscopic cytochemical study. Am. J. Pathol. 80, 349.
- Wilson, A.C., LaMarco, K., Peterson, M.G., and Herr, W. (1993a). The VP16 accessory protein HCF is a family of polypeptides processed from a large precursor protein. Cell 74, 115–125.

- Wilson, A.C., Cleary, M.A., Lai, J.-S., LaMarco, K., Peterson, M.G., and Herr, W. (1993b). Combinatorial control of transcription: the herpes simplex virus VP16-induced complex. In Cold Spring Harbor Symposia on Quantitative Biology, pp. 167–178.
- Wilson, A.C., Parrish, J.E., Massa, H.F., Nelson, D.L., Trask, B.J., and Herr, W. (1995a). The gene encoding the VP16-accessory protein HCF (HCFC1) resides in human Xq28 and is highly expressed in fetal tissues and the adult kidney. Genomics 25, 462–468.
- Wilson, A.C., Peterson, M.G., and Herr, W. (1995b). The HCF repeat is an unusual proteolytic cleavage signal. Genes Dev. 9, 2445–2458.
- Wysocka, J., Reilly, P.T., and Herr, W. (2001). Loss of HCF-1--chromatin association precedes temperature-induced growth arrest of tsBN67 cells. Mol. Cell. Biol. 21, 3820–3829.
- Wysocka, J., and Herr, W. (2003). The herpes simplex virus VP16-induced complex: the makings of a regulatory switch. Trends Biochem. Sci. 28, 294–304.
- Wysocka, J., Myers, M.P., Laherty, C.D., Eisenman, R.N., and Herr, W. (2003). Human Sin3 deacetylase and trithorax-related Set1/Ash2 histone H3-K4 methyltransferase are tethered together selectively by the cell-proliferation factor HCF-1. Genes Dev. 17, 896–911.
- Wysocka, J., Swigut, T., Xiao, H., Milne, T.A., Kwon, S.Y., Landry, J., Kauer, M., Tackett, A.J., Chait, B.T., Badenhorst, P., et al. (2006). A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. Nature 442, 86–90.
- Xiao, T., Hall, H., Kizer, K.O., Shibata, Y., Hall, M.C., Borchers, C.H., and Strahl, B.D. (2003). Phosphorylation of RNA polymerase II CTD regulates H3 methylation in yeast. Genes Dev. 17, 654–663.
- Yokoyama, A., Wang, Z., Wysocka, J., Sanyal, M., Aufiero, D.J., Kitabayashi, I., Herr, W., and Cleary, M.L. (2004). Leukemia proto-oncoprotein MLL forms a SET1-like histone methyltransferase complex with menin to regulate Hox gene expression. Mol. Cell. Biol. 24, 5639–5649.
- Yu, H., Mashtalir, N., Daou, S., Hammond-Martel, I., Ross, J., Sui, G., Hart, G.W., Rauscher, F.J., Drobetsky, E., Milot, E., et al. (2010). The ubiquitin carboxyl hydrolase BAP1 forms a ternary complex with YY1 and HCF-1 and is a critical regulator of gene expression. Mol. Cell. Biol. 30, 5071–5085.