EPFL

Thèse n° 8104

Circadian dynamics of RNA localisation in the mammalian liver

Présentée le 3 septembre 2021

Faculté des sciences de la vie
Unité du Prof. Naef
Programme doctoral en biotechnologie et génie biologique

pour l'obtention du grade de Docteur ès Sciences

par

Clémence Yumie Syloun HURNI

Acceptée sur proposition du jury

Prof. P. D. Barth, président du jury
Prof. F. Naef, directeur de thèse
Prof. D. Gatfield, rapporteur
Prof. S. Brown, rapporteur
Prof. E. Oricchio, rapporteuse

École
polytechnique
fédérale
de Lausanne

2021

Acknowledgements

I would like to express my gratitude to Felix for being an ever available supervisor. His attention to detail, his scientific rigour, and his original vision of biological systems are valuable elements that I take with me for my next steps in life.

The main reason that pushed me to do a Ph.D thesis is that I had not yet fulfilled my thirst for learning despite my years of study. I would like to thank all my colleagues who have shown pedagogy and patience when they had to explain theoretical concepts to me again and again. Special thanks to Nick, Jake, Cedric, Ben and Eric P. for being mentors and enriching my theoretical knowledge.

I would like to thank all the past and present members of the lab who have all in one way or another made life more pleasant whether in the lab, in the office or outside. I will never forget all our non-scientific spicy discussions. In addition to the previously listed colleagues, I heartfully thank Colas, Irene, Hugo, Nagammal, Lorenzo, Jérôme, Eric D., Daniel, and Ambroise for the amazing atmosphere in the lab. A special thanks to Sophie, the most caring and smiling administrative assistant.

This work would not have been possible without the support from the great facilites at EPFL: the Histology Core Facility, the BiOP, and the GECF, Thanks for your availabilty and reliability.

I am grateful for my family's support, even if I was often too lazy to explain what were my research topics.

Last but not least, the sweetest outcome of my 5 years in the Naef-lab: Damien. Thank you for always being by my side.

Contributions to this work

In this project, I performed all data analysis, microscope imaging and image processing, as well as all biochemical experiments with the exception of single-molecule RNA-FISH treatments and hybridisations that were done by the Histology Core Facility at EPFL, and the library preparation and RNA-sequencing that was done by the Gene Expression Core Facility at EPFL. I performed part of the bioinformatics analysis (RNA-seq mapping) with the help of my colleague in the Naef group, Cédric Gobet, Ph.D. I also would like to mention the crucial help from the following collaborators:

- Cédric Gobet helped me with the pipeline of bioinformatic analysis and data processing.
- Benjamin Weger, previously in the Naef lab and currently at the University of Queensland, Australia, contributed to the experimental design of the project on subcellular localisation of RNA and to the collection of the mouse liver samples.
- Colas Droin worked on the liver zonation project.
- Jérôme Mermet and Jake Yeung invited me to contribute to their project with the analysis of transcriptional bursting parameters.
- Jérôme Mermet and Daniel Mauvoisin contributed to the collection of mouse liver samples for the transcriptional bursting project and the liver zonation project.
- Romain Guiet and Olivier Buri from the BioP (EPFL) provided me with tips and tools for image processing.

Abstract

Gene expression in eukaryotes is a complex multi-step process. It starts in the nucleus with transcription, the synthesis of a mRNA copy from a DNA template. While in the nucleus, the RNA transcript is subject to multiple co- and post-transcriptional modifications, including splicing, capping, polyadenylation, and assembly into ribonucleoprotein complexes. Correctly processed mRNA is exported to the cytoplasm and translated by ribosomes to form a chain of amino acids: the protein. The mRNA lifetime in the cytoplasm is determined by the activity of a distinct set of RNA Binding Proteins, enzymes, and functional RNAs, which promote stability or degradation.

Each step of the RNA life cycle is tightly regulated to ensure proper cellular function. The kinetic rates governing these steps are dynamic and notably adapt to the daily fluctuations of the environment related to the alternance of day and night. Indeed, most organisms possess an internal timing system, called the *circadian clock*. The clock is a genetically encoded self-sustained transcriptional-translational feedback loop, which operates in almost every cell and tissue of the body. It controls the temporal gene expression program to synchronise cellular and physiological functions to the external world. In this thesis, I explore the transcriptome of the mouse liver by combining RNA-sequencing, mathematical modelling, and single-molecule RNA-FISH (smFISH), with an emphasis on the spatio-temporal organisation of RNA expression at the subcellular and tissue scales.

I first investigate how RNA are differentially localised at the scale of the liver tissue. Hepatocytes are arranged in structural and functional units called lobules, and carry out different physiological functions depending on their spatial position within the lobule. In this work, we characterised spatio-temporal gene expression profiles, and showed that while the expression of hundreds of genes is dually orchestrated by time and space, the circadian core clock is expressed uniformly within the liver lobule, and is therefore robust to the heterogeneous microenvironment.

Second, I explore the RNA localisation at the scale of a hepatocyte. The subcellular distribution of RNA in different compartments (here, the nucleus and the cytoplasm), are dictated by the balance of a synthesis term and a decay term. To quantify the kinetic parameters driving nuclear and cytoplasmic mRNA accumulation, I sequenced RNA from both cellular fractions from mouse livers sampled at different times of the day. Using a mathematical model describing rhythmic pre-mRNA and mRNA profiles, I could estimate the nuclear export rates and cytoplasmic degradation rates of $\sim$1400 genes. Nuclear export occurs on a much shorter time-scale than cytoplasmic degradation, and nuclear lifetime has only a minor contribution to the total RNA lifetime. However, a subset of metabolic genes remain in the nucleus for more than one hour (up to four hours), which accounts for the long phase delay between the peak times of transcription and of cytoplasmic accumulation. Furthermore, nuclear export contributes to the modulation and generation of rhythmic profiles of ~10% of the cycling nuclear mRNA. This study provides a comprehensive estimation of the nuclear and cytoplasmic life times in the liver and contributes to a better understanding of the dynamic regulation of the transcriptome during the feeding-fasting cycle.

Keywords: circadian rhythms, RNA processing, RNA-seq, liver zonation, single-molecule RNA FISH, mathematical modeling

Résumé

L’expression génique chez les eucaryotes est un processus complexe impliquant de multiples niveaux de régulation. Le processus commence dans le noyau avec la transcription, soit la synthèse d’une copie d’ARN messager (ARNm) à partir d’un segment d’ADN. Pendant son séjour dans le noyau, le transcrit d’ARN subit de multiples modifications co- et post-transcriptionnelles, notamment l’épissage, le coiffage, la polyadénylation et l’assemblage en complexe ribonucléoprotéique. L’ARNm est ensuite exporté vers le cytoplasme où il est traduit par les ribosomes pour produire une chaîne d’acides aminés : la protéine. La durée de vie de l’ARNm dans le cytoplasme est déterminée par l’activité d’un ensemble distinct de protéines de liaison à l’ARN, d’enzymes et d’ARN fonctionnels, qui favorisent sa stabilité ou, au contraire, sa dégradation.

Chaque étape du cycle de vie de l’ARN est étroitement régulée pour assurer le bon fonctionnement cellulaire. Les taux cinétiques qui régissent ces étapes sont dynamiques, et s’adaptent aux fluctuations journalières de l’environnement, dûes notamment à l’alternance du jour et de la nuit. En effet, la plupart des organismes possèdent un système interne permettant la synchronisation temporelle de leurs fonctions cellulaires et physiologiques au monde extérieur, appelé *horloge circadienne*. L’horloge est une boucle de rétroaction auto-régulée, encodée génétiquement, qui contrôle le programme temporel d’expression des gènes dans presque toutes les cefllules et tous les organes du corps. Dans cette thèse, j’étudie le transcriptome du foie de la souris en combinant du séquençage d’ARN à haut débit, de la modélisation mathématique et de l’hybridation *in situ* en fluorescence sur molécule d’ARN (smFISH), en mettant l’accent sur l’organisation spatio-temporelle de l’expression de l’ARN.

J’ai d’abord étudié comment les ARN sont différemment localisés à l’échelle du tissu hépatique. Les hépatocytes sont disposés en unités structurelles et fonctionnelles appelées lobules, et remplissent différentes fonctions physiologiques en fonction de leur position au sein du lobule. Dans ce projet, nous avons caractérisé les profils d’expression génique spatio-temporels, et montré que si l’expression de centaines de gènes est orchestrée à la fois dans le temps et l’espace, ce n’est pas le cas de l’horloge circadienne qui est exprimée uniformément dans le lobule du foie, et est donc robuste à ce micro-environnement hétérogène.

Dans un second temps, j’ai exploré la localisation de l’ARN à l’échelle d’un hépatocyte. Les distributions subcellulaires de l’ARN dans différents compartiments (ici, le noyau et le cytoplasme), découlent de l’équilibre entre le taux de synthèse et de dégradation. Pour quantifier les paramètres cinétiques conduisant à l’accumulation d’ARNm nucléaire et cytoplasmique, j’ai séquencé l’ARN des deux fractions cellulaires à partir de foies de souris échantillonnés à différents moments de la journée. À l’aide d’un modèle mathématique décrivant des profils rythmiques de pré-ARNm et d’ARNm, j’ai pu estimer les taux d'export nucléaire et de dégradation cytoplasmique de ~1400 gènes. L'export nucléaire se produit sur une échelle de temps beaucoup plus courte que la dégradation cytoplasmique, et la durée de vie nucléaire n'a qu'une contribution mineure à la durée de vie totale de l'ARN. Cependant, un sous-ensemble de gènes métaboliques reste dans le noyau pendant plus d'une heure (jusqu'à quatre heures), ce qui explique le long retard de phase entre les pics de transcription et d'accumulation cytoplasmique. En outre, l'export nucléaire contribue à la modulation et à la génération de profils rythmiques de ~10% des ARNm nucléaires rhythmiques.

Cette étude fournit une estimation complète des durées de vie nucléaires et cytoplasmiques dans le foie et contribue à une meilleure compréhension de la régulation dynamique du transcriptome au cours d'une journée.

Contents

Acknowledgements i

Contribution ii

Abstract iii

1 Introduction 1
- 1.1 Circadian clock system in mammals 1
  - 1.1.1 Overview 1
  - 1.1.2 The molecular circadian core clock 2
  - 1.1.3 The intricate interplay between the circadian clock and metabolism 3
  - 1.1.4 The clock in the mammalian liver 5
- 1.2 Circadian regulation of gene expression 7
  - 1.2.1 Circadian regulation of transcription 8
  - 1.2.2 Evidences for post-transcriptional circadian regulation 9
  - 1.2.3 Splicing 10
  - 1.2.4 Nuclear export 13
  - 1.2.5 Nuclear speckles and paraspeckles are hotspots of RNA processing factors 14
  - 1.2.6 N^6-methyladenosine (m^6A) RNA methylation 16
  - 1.2.7 Modulation of cytoplasmic mRNA stability 17
- 1.3 Liver zonation 18
  - 1.3.1 Structural organisation of the murine liver 18
  - 1.3.2 Zonated hepatic functions 19
  - 1.3.3 Factors regulating hepatic zonation 21
  - 1.3.4 Zonation of polyploidy 22
  - 1.3.5 Dynamic regulation of the liver zonation 23

2 Objectives of the thesis and presented work **25**

3 Spatio-temporal gene expression profile in the mammalian liver **29**
- 3.1 Background 29
- 3.2 Results 29
  - 3.2.1 Single-molecule RNA-FISH to explore spatio-temporal patterns 29
  - 3.2.2 Clock genes are uniformly expressed within the liver lobule 35

3.3 Concluding remarks 37
3.4 Space-time logic of liver gene expression at sublobular scale: Abstract 38

4 Comprehensive analysis of the circadian hepatic transcriptome at subcellular scale 39
4.1 Background 39
4.2 Experimental design 41
4.3 Transcript-based RNA-seq quantification 42
4.4 Characteristics and composition of nuclear and cytoplasmic transcriptomes in mouse liver 46
4.5 Relationship between nuclear and cytoplasmic mRNA reveals signatures of export and cytoplasmic half-life 48
  4.5.1 Localisation of Protein Coding transcripts matches the localisation of the encoded proteins 52
  4.5.2 Retained Intron as a class of nuclear-retained RNA 54
4.6 Relationship between nuclear pre-mRNA and mRNA reveals signatures of splicing and export times 57
  4.6.1 Nuclear pre-mRNA reveals the extent of co- versus post-transcriptional splicing 59
  4.6.2 Estimation of the relative variability of the kinetic rates 65
4.7 Model-based approach to quantify kinetic parameters from time-series profiles 67
4.8 Estimation of cytoplasmic degradation rates using time-series RNA-seq profiles 69
  4.8.1 Identification of distinct dynamic strategies driving nuclear and cytoplasmic mRNA abundances 71
4.9 Estimation of nuclear export rates rates using time-series RNA-seq profiles 76
4.10 Relationship between the kinetic rates of mRNA 83
  4.10.1 No global coordination of the export rates and cytoplasmic degradation rates . . 83
  4.10.2 Cellular fractionation reveals regulatory processing steps hidden in bulk analysis 84
  4.10.3 Subcellular RNA distribution versus estimated kinetic rates 88
4.11 Discussion 90
  4.11.1 Characterisation of the nuclear and cytoplasmic transcriptome revealed different subcellular localisation of distinct RNA classes 90
  4.11.2 Estimation of the co- versus post-transcriptional splicing times using nuclear pre-mRNA and mRNA 91
  4.11.3 Mathematical model to quantify kinetic parameters 92
  4.11.4 Concluding remarks 95
4.12 Supplementary Analysis 96
  4.12.1 Overview of temporal characteristics of the liver transcriptome 96
  4.12.2 Estimation of kinetic parameters in clock-deficient mouse model 102
  4.12.3 Comparison of temporal parameters from single-molecule RNA-FISH and RNA-seq 106
  4.12.4 Absence of apparent contamination by the endoplasmic reticulum in cuclear fractions 110

5 Circadian and chromatin contacts-dependent modulation of transcriptional bursting parameters 113

5.1 Background 113
5.2 Results 114
5.3 Concluding Remarks 116
5.4 Clock-dependent chromatin topology modulates circadian transcription and behavior: Abstract 118

6 Methods 119
6.1 Mouse Liver sampling 119
6.2 Cellular fractionation of liver tissue and RNA extraction 119
6.3 Single-molecule RNA-FISH 120
6.4 Microscope image acquisition and quantification of single-molecule RNA-FISH images 121
    6.4.1 Quantification of spatio-temporal mRNA profiles 121
    6.4.2 Modeling spatio-temporal profiles 122
    6.4.3 Quantification of nuclear and cytoplasmic temporal mRNA profiles 122
    6.4.4 Quantification of transcriptional bursting parameters 122
6.5 RNA-sequencing and mapping 123
    6.5.1 Library preparation 123
    6.5.2 Pseudo-alignment of RNA-seq reads with Kallisto 123
    6.5.3 Statistical analysis of rhythmic gene expression 124
6.6 Functional Enrichment Analysis 127
    6.6.1 Gene Ontology Analysis of nuclear and cytoplasmic genes 127
6.7 Mathematical modeling of rhyhthmic RNA processing rates 127
    6.7.1 Modeling temporal RNA profiles and estimation of kinetic parameters 128

Curriculum Vitae 151

1 Introduction

1.1 Circadian clock system in mammals

1.1.1 Overview

Most organisms, from plants, cyanobacteria, drosophila to mammals, have adapted to the daily environmental changes. They have developed a biological endogenous timekeeping system called the *circadian clock* (from the latin words *circa* and *dies*, "around a day") to anticipate and to coordinate their behaviour and physiology to the day/night cycles, to temperature and humidity variations, social interactions, etc. [1].

The circadian clock is organised in a hierarchical manner. In mammals, a "master clock" is located in the suprachiasmatic nucleus (SCN), a bilateral structure in the hypothalamus. These two small regions contain each ~20000 coupled neurons that integrate the photic signals perceived by the retina (cones, rods, and melanopsins containing ganglion cells) and transmitted via the retino-hypothalamic tract [2]. When cultured individually, these neurons retain their circadian gene expression, but with low amplitude and in poorly organized manner. In contrast, when their connectivity is maintained, which is the case in organotypic culture, individual rhythms are stable and phase-coherent [3]. The central pacemaker in the SCN aligns the phase of the "peripheral clocks" using various systemic routes such as the autonomic neuronal system, or hormones such as glucocorticoids and melatonin to generate rhythmic outputs such as the rest - activity cycle, feeding-fasting period or body temperature changes [4, 5]. Indeed, cells of virtually all the peripheral tissues and cells contain a self-sustained molecular clock [6, 7], particularly important in metabolic organs such as the liver, pancreas, kidney, muscle or adipocytes [8]. Importantly, these organs are not only entrained by the SCN, but also by other external cues like the food intake.

The circadian clock system is “entrained” by external stimuli, called *Zeitgebers* (ZT). In light-sensitive organisms, daylight is the dominant synchroniser. In absence of stimulus, a condition called “free-running”, the clock ticks at its endogenous period. In the 60’s, Aschoff examined endogenous rhythms in humans. Volunteers were isolated in bunkers without access to daylight nor information about external time. By monitoring their activity pattern, urine excretion and body temperature, it was shown that the endogenous period is highly variable, but with a mean around 25 hours, longer than the period of Earth rotation [9]. On the other hand, the activity of mice kept in constant darkness had a shorter period [10].

The genetic determinant of behavioral rhythmicity was first identified in 1971 by Ron Konopka in *Drosophila Melanogaster* mutant lines that had aberrant locomotor activity and eclosion rhythm [11]. They discovered a gene locus, named *period* (or *per*), that was later cloned for the first time in 1984 by the team of Jeffrey Hall and Michael Roshbash [12], and Michael Young [13]. The last three scientists were rewarded by the Nobel Prize in Physiology or Medicine in 2017 for their "discoveries of molecular mechanisms that control circadian rhythms". In mammals, the group of Joseph Takahashi discovered the first mammalian gene in a mutant mouse model: *Clock* [14], followed by other *core clock* genes, namely *Arntl* (or *Bmal1* [15]), *Cryptochromes* [16], and *Period* genes.

1.1.2 The molecular circadian core clock

At the molecular level, in mammals, the cell autonomous core clock is a transcriptional-translational feedback loop (TTFL) (Fig.1.1) [17, 18]. The positive limb of the network is constituted of two transcriptional activators, CLOCK and BMAL1. They activate the expression of their own repressors Cryptochromes (CRY1 and CRY2) and Periods (PER1, PER2 and PER3) by binding to regulatory elements containing *E-Boxes*. CRY and PER accumulate in the cytoplasm and dimerize before they translocate into the nucleus and interact with the CLOCK - BMAL1 complex, therefore downregulating their own expression. The decrease of BMAL1 and CLOCK levels results in a decrease of CRY and PER levels, which in turn lead to the accumulation of BMAL1 and CLOCK, thus starting a new cycle with a period of ~24 hours. A second loop is composed of the complex CLOCK-BMAL1 activating the expression of the nuclear receptor RORα, β and γ [19], REVERb α and β [20] (encoded by *Nr1d1* and *Nr1d2*), that compete to respectively activate and repress the expression of BMAL1 by binding to the ROR binding elements (ROREs). Finally, additional loops involving the ParB ZIP family members (DBP, TEF, HLF [21], and the repressors NFIL3 encoded by *E4bp4*), or the bHLH proteins (DEC1, DEC2) help to maintain the robustness of the core oscillator. Post-translational modifications also play a pivotal role [22]. For instance, casein kinase 1 (CK1) or F-box and leucine-rich repeat protein 3 (FBXL3) phosphorylate PER and CRY respectively, modifying the stability of PER/CRY complex, its nuclear translocation, and eventually promoting its proteasomal degradation, with opposite effects depending on which sites are phosphorylated. A mutation in a stabilising phosphorylation site of PER2 in human, and a mutation causing a decreased enzymatic activity of CKI δ both provoke the Familial advanced sleep phase syndrome (FASPS), shortening the intrinsic period and advancing the sleep onset [23, 24]. These examples demonstrate the importance of post-translational modifications of core clock proteins in regulating their stability and eventually circadian period length.

All together, these interlocked feedback loops generate intracellular cycles of 24 hours, and modulate the phases of expression of various target genes, based on the combination of *cis*-elements in the promoters and enhancers. There are additional layers of post-transcriptional and post-translational regulation that modulate the final rhythmic output, and will be covered later in this chapter.

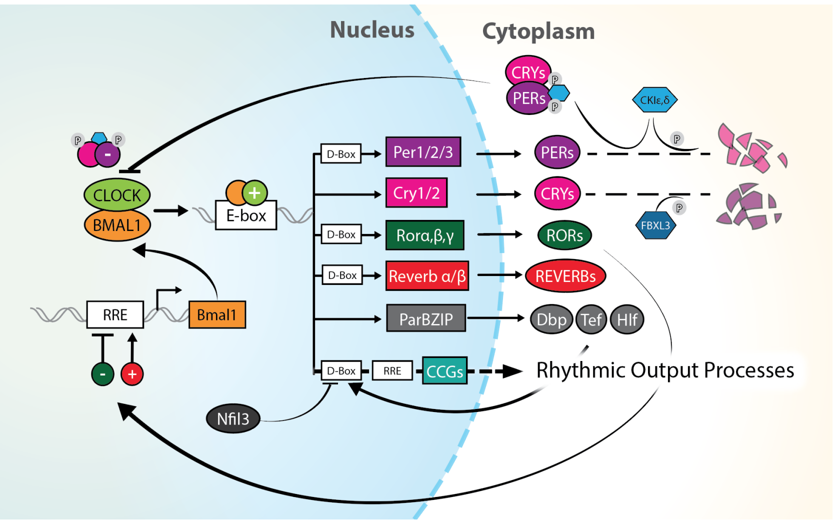


Figure 1.1 – Molecular architecture of the circadian clock in mammals. The clock is a self-sustained transcriptional-translational feedback loop. The network consists of two transcription factors, CLOCK and BMAL1, which activate the expression of their own repressors Cryptochromes (CRY1 and CRY2) and Periods (PER1, PER2 and PER3). CRYs and PERs accumulate and reduce the transcriptional activity of CLOCK - BMAL1 complex, which in turn leads to their own extinction. Then, BMAL1 and CLOCK activity rises again, generating a new cycle with a period of approximately 24 hours. Stability and localisation of PER and CRY are modulated by the kinases CKI and FBXL3. A second feedback loop is composed of the complex CLOCK-BMAL1 activating the expression of the target genes RORα, β, and γ, and REVERBα andvβ, which compete to respectively activate and repress the expression of BMAL1. The molecular clock also targets thousands of downstream clock-controlled genes to couple circadian cycles and physiological pathways.

1.1.3 The intricate interplay between the circadian clock and metabolism

The core mechanism of the transcriptional-translational feedback loop is universal in virtually all the cells in the body. In addition, precise combinations of tissue-specific transcription factors and promoters shape the rhythmic gene expression program in each organ [25]. For instance, liver-specific nuclear receptor HNF4α represses the activity of BMAL1:CLOCK and further regulate key hepatic functions in lipid, glucose and amino acid homeostasis [26]. A comparative study in the primate *Papio anubis* (baboon) compared the transcriptome in 64 tissues, and showed that more than 80% of the protein-coding genes were cycling in at least one tissue, however, the overlap between cycling genes was small across tissues [27]. Many genes were expressed in several tissues, but were rhythmic in some while constant in others, highlighting the fact that expression and rhythmic expression are tissue-specific. Similarly, studies in different mouse tissues showed that at least half of all the expressed genes cycles in at least one tissue. Again, the overlap of the rhythmic gene set between organs was very small, with less than 1% of the genes that oscillate in all the tissues [28]. Metabolically active tissues have the largest proportion of cycling genes, such as the liver (one the most studied organ in chronobiology), kidney, lung, brown fat, and heart. They carry out a large panel of catabolic and anabolic functions that require time coordination to prevent two opposite and incompatible processes from running simultaneously. The intimate connection between the clock system and cellular metabolic processes allows the organism to anticipate physiological needs and to respond to environmental changes [29].

In metabolic organs, food (un)availability is the dominant Zeitgeber. For instance, an inverted feeding regimen can uncouple the SCN oscillator from the peripheral tissues oscillators (liver, kidney, heart, and pancreas). After a week of restricted day-time feeding, the phases of clock genes in those tissues became antiphasic to the phase of the SCN. The shift was particularly fast in the liver [30], and shown to be entrained by glucocorticoids signalling pathway [31, 32]. When the feeding pattern is arrhythmic (mice are continuously provided with a small amount of food), 70% of the rhythmic genes under *ad libitum* feeding regimen lost their rhythm. Core clock gene expression, on the other hand, remained unaffected by the feeding rhythm. This suggested that in the liver, rhythmic food intake is the dominant Zeitgeber for a majority of genes [33]. One way that feeding cues are integrated by the core clock circuitry are insulin and insulin-like Growth Factor 1 (ILGF-1), which increase PER2 translation efficiency [34].

The importance of a healthy timing system is illustrated by metabolic dysfunctions arising in night-shift or rotation shift workers whose circadian rhythm is disrupted (prevalence of cardiovascular disease, obesity, prevalence of some cancers). In 2019, the International Agency for Research on Cancer (IARC) classified night-shift work as “probably carcinogenic to human”, even though more epidemiological studies are needed to confirm this claim in humans [35]. In rodents, metabolic disturbances have been described in mutant mice. For instance, the arrhythmic *Bmal1*-KO mice show signs of premature ageing like sarcopenia, loss of body weight and impaired glucose tolerance. CLOCK-KO mice are, on the other hand, obese and hyperphagic. Both mutant lines lost diurnal variations of triglycerides and glucose [36]. Conversely, a strict rhythmic pattern of feeding can improve metabolic functions. For example, mice fed with a high fat diet *ad libitum* exhibited dampened metabolic and physiological oscillations in addition to obesity, liver steatosis and other metabolic syndromes, while mice under a time-restricted feeding were protected against the adverse consequences of a high fat diet, and preserved robust circadian and metabolic oscillations [37]. The beneficial effects of a time-restricted feeding was observed even in fruit flies: those fed only during the day had a better sleep quality (less napping and longer sleep time during the night), and a slower decline in heart function as they aged [38].

The extensive crosstalk between the circadian cellular oscillator and the metabolic system relies on the overlap of some of their transcriptional networks [29]. One of these connections happens through the rhythmic production of NAD+, a central cofactor involved in redox reactions. CLOCK and BMAL1 cyclically transcribe *Nampt*, the rate-limiting enzyme of the NAD+ salvage pathway [39]. They thus regulate the rhythmic availability of NAD+, and by extension, the activity of NAD-dependent enzymes activities, such as SIRTUINS. SIRT1 is a histone deacetylase that regulates metabolic processes, including gluconeogenesis through the activation of PGC-1. It also binds CLOCK:BMAL1 and helps the remodelling of chromatin, promotes the degradation of PER2, thus creating a metabolic feedback loop tied to the circadian clockwork circuitry. Decline of NAD+ rhythm has been associated to ageing and metabolic stress [40]. Another way in which metabolism and the clock interact is through AMPK (AMP-activated protein kinase). When the cellular energy state is low, the high AMP/ATP ratio triggers a signalling cascade that leads to the activation of AMPK, which switches on catabolic pathways that produce ATP, while switching off ATP-consuming processes. In addition, AMPK phosphorylates and destabilses CRY1, promoting its degradation [41].

1.1.4 The clock in the mammalian liver

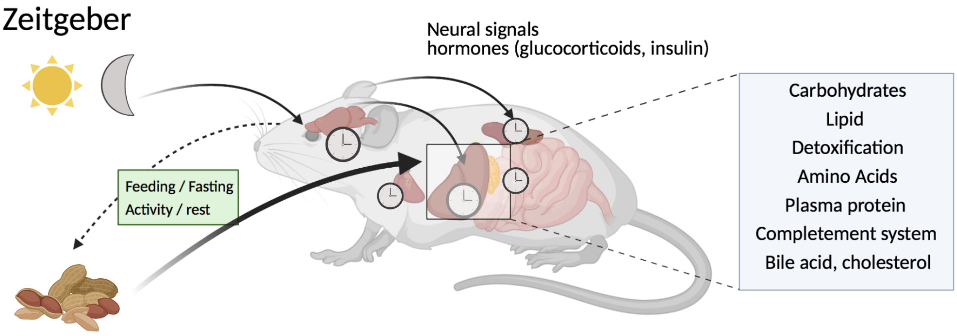


Figure 1.2 – Entrainment of the circadian clock in metabolic organs. Metabolic organs including liver, pancreas, kidneys, heart and adipose tissues are entrained by multiple Zeitgebers. Light/dark and temperature cycles indirectly affect peripheral organs by first entraining the central clock in the SCN, which systematically propagates the signals to peripheral tissues using neural and hormonal pathways. Circadian clocks of metabolic organs are also directly entrained by Zeitgebers such as feeding and fasting or physical activity. In these organs, the circadian clock regulates various metabolic processes (blue box). Scheme created with BioRender.

One of the most studied metabolic organs in chronobiology is the liver. Genome-wide investigations showed that 10 to 20% of the transcriptome is rhythmic [42], including many key rate-limiting enzymes involved in metabolic functions [43]. It maintains energy homeostasis throughout the day by balancing carbohydrate and lipid metabolism. It helps digestion by producing and recycling bile acids. It clears out blood from xenobiotics and old red blood cells. It is also the major site of protein synthesis, releasing in the bloodstream proteins such as Albumin, hormones, complement system proteins, and blood clotting factors. It has to adapt to very different physiological requirements along the day in order to maintain energy homeostasis. Changes are particularly important at the transition between the active and rest phases, to which the liver responds by two “peaks” of transcriptional activity, just before dawn and before dusk (ZT10 and ZT20) [44]. Additionally, proteomics [45], metabolomics and lipidomics [46] analysis revealed a widespread effect of post-transcriptional and post-translational regulation [47].

Maintenance of a constant level of blood glucose is a crucial role of the liver. This is achieved by a balance of breakdown and production of glucose [48]. When food consumption is maximal at the beginning of the active phase, the expression of glucose transporter and glucagon receptor increases. The excess glucose is stored as a form of glycogen polymers by a CLOCK-controlled gene *Gys2* [49], whose transcription peaks around ZT12 and activity late at night. Between meals, gluconeogenesis produces *de novo* glucose, a process catalysed by the rate-limiting enzyme PCK1 (or PEPCK1), and is stimulated by the circadian release of glucocorticoids from the adrenal cortex. CRY1 has been shown to interact with the glucocorticoid receptor, hence inhibiting the expression of genes containing glucocorticoid-responsive elements (GREs) like *Pck1*. CRY1 inhibits the phosphorylation of CREB, a transcription factor that promotes the expression of *Pck1* and *G6pc* when phosphorylated [50]. Finally, CRY1 also promotes the degradation of FOXO1, a positive regulator of *Pck1* and *G6pc*, further inhibiting gluconeogenesis [51].

The liver manages excessive carbohydrates, proteins, and fat by transforming these macromolecules into fatty acids (lipogenesis) and assembling them into triglycerides for long term storage, or by excreting them in the form of lipoproteins. In a fasted state, it consumes nonesterified fatty acids released by adipocytes, and produces energy and ketone bodies through β-oxidation. Many of the key enzymes in the lipid metabolism are rhythmic in the liver, such as *Elovl3*, *Elovl6*, *Fas*, *Agpat*, *Lpin*, and the nuclear receptor PPARs and the coactivator PGC-1, and SREBP (master regulator of lipid metabolism). The α isoform of PPAR is abundant in the liver, and is a direct target of BMAL1:CLOCK. Conversely, *Bmal1* promoter contains PPRE domain, and is thus under the control of PPARα [52]. The repressor REVERBα, together with the histone deacetylase HDAC3 and NCor1, is another central node linking the circadian clock and lipid metabolism. The complex rhythmically binds to many gene loci coding for proteins involved in lipogenesis [53]. The absence of either HDAC3 or REVERBα-/β provokes lipid accumulation in the liver (steatosis). REVERBα is also involved in bile acid metabolism. For instance, it regulates the activity of SREBP, a transcription factor that promotes insulin-driven lipogenic activity, through cyclic transcription of *Insig2*, a protein responsible for the sequestration of SREBP [54]. Importantly, REVERBα negatively regulates the rate-limiting enzyme converting cholesterol to bile acid, *Cyp7a1*.

Another crucial role of the liver is the removal of xenobiotic compounds from blood. Detoxification occurs in three phases. In Phase I, lipid-soluble compounds are modified and inactivated by oxidation, reduction or hydrolysis, mostly by Cytochrome P450 enzymes (CYPs). The products from Phase I are made more water-soluble during Phase II, so they are easier to be excreted into the bile acids, urine or feces. Phase III enzymes include transporters in charge of excretion of the metabolites. When a lipid-soluble xenobiotic crosses the cell membrane, it is detected by nuclear receptors such as CAR or SHP, both showing clear diurnal oscillations in the liver. CAR further regulates the expression of POR (Cytochrome P450 oxydoreductase) and ALAS1, a rate-limiting enzyme in the heme synthesis. Both proteins are necessary to activate the Cytochrome P450, as all CYPs need a heme as a prosthetic group. Proteomics data have revealed that many proteins of Phase I, II and III accumulated rhythmically in mouse liver [45]. The detoxification-related transcription factors *Dbp*, *Tef*, *Hlf* are direct output mediators of CLOCK/BMAL1 transcription [55]. They promote the transcription of downstream genes containing D-Boxes, in competition with the repressor NFIL3 (transcribed by REV-ERB / RORs). Some enzymes of Phase I, II and III (CYP2A4, CYP2A5, CYP3A4, ABCB1) are direct targets of the PARbZIP transcription factors. Moreover, they are thought to coordinate the expression of the other Cytochrome P450 enzymes by directly regulating the expression of CAR. The importance of the PARbZIP genes was demonstrated by the phenotypes of triple KO mouse (*Dbp*, *Tef* and *Hlf* knock-out) [56]. There was a high juvenile morbidity rate, mainly due to an epileptic seizure. Adult triple KO mice that survived showed signs of premature ageing and premature death, probably due to impaired detoxification processes.

Through these few examples, we see the intricate and bi-directional interplay between the circadian liver oscillator, and the complex physiological machinery. This link is particularly prominent in the liver, where most of its key functions are timely coordinated in order to respond and anticipate daily variation of nutrients availability.

1.2 Circadian regulation of gene expression

Core clock proteins are regulatory proteins driving the rhythmic transcription of thousands of genes, ultimately producing cycling physiological/ biochemical outputs. Daily fluctuations of RNA and protein levels do not only stem from a rhythmically regulated transcription, but also at post-transcriptional and post-translational level. In eukaryotes, RNA is synthesised in the nucleus, processed (5'capped, 3'-cleaved, polyadenylated, spliced, decorated by methyl-groups), and exported to the cytoplasm where it is translated and finally degraded [57]. Virtually any step of the RNA life cycle can be regulated in a circadian manner, generating *de novo* rhythms or modulating temporal patterns of circadian RNA / proteins [58]. For the last decade, many genome-wide studies quantified the extent of rhythmic post-transcriptional and post-translational regulation, with sometimes discrepant results that undeniably arise from different experimental design and data analysis methods, even when performed in the same tissue, such as the mouse liver [59, 54, 47, 60, 45, 61]. In the following paragraphs, I will focus on the main RNA processing steps known to be under circadian control, from chromatin accessibility to mRNA degradation. Translational and post-translational regulatory processes such as the acetylation [62] and phosphorylation [63] additionally generate rhythms at the level of protein accumulation, but will not be covered here.

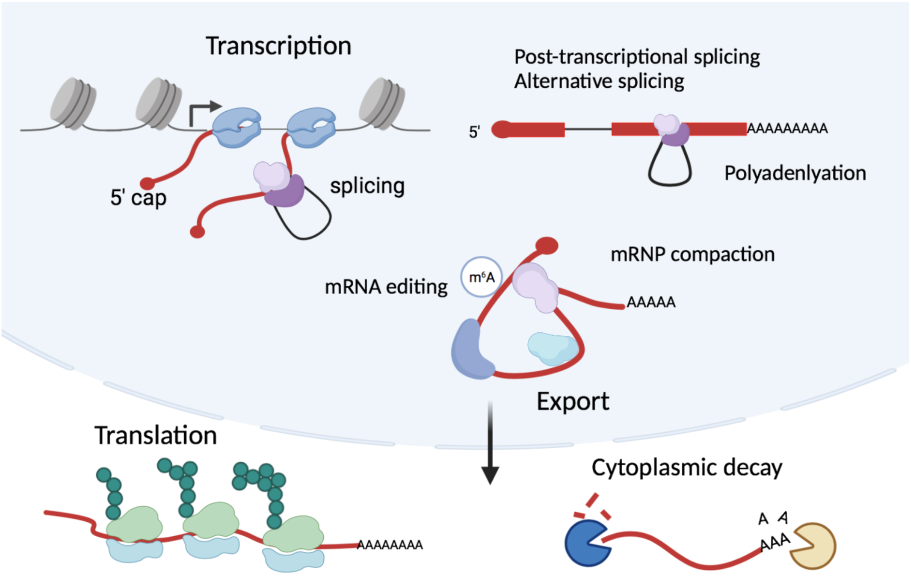


Figure 1.3 – Overview of gene expression regulatory steps putatively affected by the circadian clock (non-exhaustive). The RNA is transcribed by the RNA Polymerase II. Introns can be removed either during transcription (co-transcriptional splicing) or after completion of transcription and addition of the Poly(A) tail (post-transcriptional splicing). An intron can be alternatively spliced, sometimes resulting in the retention of the entire intronic region. The mRNA is bound by various RNA Binding Proteins (RBPs) and further modified, for example by the addition of methyl groups. Once fully mature, it is exported to the cytoplasm where it is translated by ribosomes. The mRNA is eventually degraded: deadenylases remove the Poly(A) tail and decapping enzymes remove the 5'cap, and the unprotected mRNA is cleaved by exonucleases.

1.2.1 Circadian regulation of transcription

Rhythmic transcription is the starting point of rhythmic gene expression, and has been extensively studied in the circadian context. Recruitment and loading of the transcriptional machinery on DNA at the transcription start site result from the combinatorial effect of core clock proteins and tissue-specific transcription factors [8]. Circadian changes of chromatin conformation also happens at a larger scale, involving enhancers networks [64, 65, 66]. For instance, Mermet et al. found a distal regulatory element (enhancer) making rhythmic contacts with the promoter of *Cry1* in mouse liver [64]. These interactions were abolished in *Bmal1*-KO mice. Knocking out 300 bp of the contacted regulatory region in mice was sufficient to compromise the rhythmic expression of *Cry1* and decrease the level of CRY1, and shortened the locomotor activity period by 15 minutes. Thus, even a short genomic region distant from the promoter - combined with other factors - is necessary to fine tune rhythmic behavior.

The main circadian transcription factors are BMAL1 and CLOCK. They participate in making the chromatin accessible with the help of numerous histone modifiers such as p300 [67] and CBP (acetylation of H3K9 and H3K27), and MLL1 and MLL3 (histone methyltransferases). In mouse liver, more than 2000 DNA locations are targeted by BMAL1:CLOCK heterodimer, and most of them are enriched around ZT6, as revealed by CHIP-seq experiments [68]. The accessibility of DNA has been assessed by mapping DNAse1 Hypersensitive Sites (DHSs) [69], showing that more than 8% of 65000 detected DHSs were diurnally cycling. Moreover, phases were matching those of chromatin opening histone modification (H3K27ac), and high density of RNA Polymerase 2 (PolII). Koike et al. performed an extensive analysis of the mouse liver cistrome using CHIP-seq of most of the core clock genes (BMAl1, CLOCK, NPAS2, PERs and CRYs) and mapped histone marks [47]. They revealed thousands of cyclic DNA-binding sites (up to 16’000 for the repressor CRY1), and drew a detailed circadian transcriptional landscape.

For many mammalian genes, transcription does not occur continuously, but rather in bursts [70]. Indeed, the gene promoter alternates from open to closed state based on several parameters [71], such as histone marks, nucleosome occupancy, DNA looping [64], and transcription factors availability. The rate of switching between the “on” and “off” state is called the burst frequency. During the “on” period, many PolII may load on the gene, producing a batch of transcripts. The number of produced transcripts per burst episode is called the burst size. This model of transcription explains why the distribution of number of RNA molecules per cell is likely to follow a negative binomial distribution rather than a Poisson distribution (expected for a continuous production) [72].

Bursting parameters vary greatly between genes [73], and are also tissue-specific. For instance, *Glul* is bursty in the mouse intestinal epithelial cells, but not in the liver [74]. In culture cells, transcriptional bursting was also shown to vary temporally, as the rhythmic transcription of core clock genes predominantly arose from daily changes in their burst frequencies (but not the burst sizes) [75, 76]. In the mammalian liver, it has been shown that the burst frequency but not the burst size is responsible for the increased expression of *Cry1* gene. Thus, stochastic gene expression is another layer of rhythmic gene expression.

1.2.2 Evidences for post-transcriptional circadian regulation

Rhythmic chromatin remodelling and transcription factors binding are not always sufficient to drive rhythmic transcription. Indeed, only a fraction of genes with a binding site targeted by BMAL1:CLOCK oscillate at the transcription level (26% in [77]), and their phase distribution is heterogeneous despite the tight time-window of DNA binding. To monitor *bona fide* transcription, intronic reads from whole tissue RNA-seq are often used as a suitable proxy for transcription, either coming from total RNA-seq [47, 60], Nascent-seq [59], PRO-seq, or GRO-seq. Several studies have shown a bimodal distribution of intron peaking at dawn and dusk (ZT9 and ZT21), driven by the coordinated transcription regulation by BMAL1-bound E/E' boxes and ROR bound ROR-elements [60, 69]. The peaks coincide with the two waves of global PolII occupancy, another indicator of transcriptional status [47]. However, in the study by Koike et al, 70% of the cycling introns were not followed by cycling exons, meaning that rhythmicity generated at the transcription level does not always propagate to the mRNA level. This pattern is likely due to a long mRNA half-life, which dampens the oscillation [54]. Actually, only 22% of cycling exons were preceded by cycling introns, meaning that the majority of rhythmic mRNAs is driven by post-transcriptional processes such as splicing, polyadenylation, export, and mRNA turnover. Another study in the mouse liver using nascent-RNA-seq also revealed that only 28% of rhythmic mRNA oscillations were driven by rhythmic transcription [59]. The extent of post-transcriptional regulation is still debated, as other groups showed smaller fraction of *de novo* rhythm in mRNA accumulation: 86% of mRNA rhythms were matching PolII loading profiles in [54], and 72% of rhythmic mRNA have rhythmic intron profile in total RNA-seq [60]. In the second study, interestingly, it seems like post-transcriptionally induced rhythmicity was more affected in *Bmal1*-KO than rhythmic transcription. Two other studies applied similar mathematical models to liver RNA-seq data and estimated the proportion of genes being post-transcriptionally regulated to be around 30% [78, 79]. 20% of the observed rhythm in mRNA accumulation was solely due to rhythmic degradation [78]. Discrepancies may come from different experimental designs (light-dark or dark-dark cycles, feeding paradigm, sampling density), type of data (total or nascent RNA-seq), or statistical methods (based on cutoff, model selection). Despite these differences, it is undeniable that post-transcriptional processes play a crucial role to tune the phase, boost the amplitude, or generate *de novo* rhythms. Indeed, a simple kinetic model with a rhythmic production and a constant degradation (mRNA half-life) describes that if transcription is the only rhythmic step, the amplitude dampens as rhythm propagates [80]. The more stable the transcript, the more the profile flattens and the phase is delayed. This behavior was observed in all the previously mentioned studies. Thus, additional layers of regulation are often necessary to maintain, or amplify these rhythms.

Therefore, mounting evidence supports the idea of an important role of the regulatory processes in controlling circadian rhythms, both during transcription (co-transcriptional splicing, recruitment of export machinery, SR proteins, and specific RNA-Binding proteins), and post-transcriptionally (RNA-editing, nuclear retention and nuclear export, regulation cytoplasmic stability and translation efficiency).

1.2.3 Splicing

One of the first steps during the processing of eukaryotic RNA is splicing, which is the removal of non-coding introns from the RNA precursor (pre-mRNA) to produce mature messenger RNA (mRNA). This step is carried out by the spliceosome, a multi-mega dalton complex consisting of five small nuclear RNAS (snRNAs: U1, U2, U4, U5 and U6) and their associated small nuclear ribonucleoproteins (snRNPs). The spliceosome assembles in a stepwise manner on each intron by recognising consensus sequences: 5' splice sites (ss), branchpoint sequence and 3' ss. It then catalyzes two transesterification steps to excise the intron, and ligates the upstream and downstream exons [81]. Spliced introns (lariat) are then quickly degraded. Gene architecture and splicing mechanisms differ between organisms. In lower complexity organisms such as yeasts, genes contain on average 1 intron. On the other hand, in mammals (notably humans), genes contain a median number of 8 introns, and the length is usually longer, from 1kb to 100 kb [82]. For comparison, the average length of exon in humans is 170nt [83]. Short and long introns are spliced with two different strategies. Spliceosome assembles across short introns (<250nt) by a model called "intron-definition" model, while longer introns found in mammals are removed by "exon-definition", where the spliceosomal assembly occurs across an exon [84].

Co- and post-transcriptional splicing

Splicing happens concurrently with PolII elongation in all studied organisms, including budding and fission yeasts [85, 86], Drosophila [87], and mammals [88, 89, 90]. However, some of the introns are also excised after the release of the transcript from chromatin, especially the terminal intron [91]. The extent of co- versus post-transcriptional splicing is still debated [92]. Transcription and splicing could be coordinated processes, or coincidentally happen at the same time due to similar kinetics [93]. Global analysis of splicing relies on purification of chromatin-bound nascent transcripts and subsequent sequencing, with or without metabolic labeling. With conventional RNA-seq methods, such as the Illumina platform, transcripts are fragmented to generate the library. Mapping short reads (~60 - 100 nt) does not allow a precise identification of isoforms. However, several metrics measuring the amount of co-transcriptional splicing have been developed, by focussing on reads mapping on spliced exon-exon junction versus unspliced exon-intron junction [93]. Several studies estimated co-transcriptional splicing frequencies: 75% in budding yeast [85], 83% in Drosophila S2 cells [87], 88% in mouse cell line (MEL) [94], around 75% in human cell lines [91], and 84 % in human tissues (adult and fetal brains and livers)[89]. Surprisingly, in the mouse liver, co-transcriptional efficiency was only 45% [90]. An independent group quantified again the mouse liver data, and confirmed the low co-transcriptional efficiency [93], showing that the extent of post-transcriptional splicing varies between tissues. If RNA PolII are uniformly distributed along a gene body, more reads would map on the 5'end than on the 3'end in nascent-seq. The same 5'-3' profile should be found within the individual intron if splicing is "perfectly" co-transcriptional. In human macrophages, many genes did not display the typical 5'-3' gradient along the gene body, and introns were present at a similar level than exon [95]. This observation indicates that not fully-spliced transcripts remain tethered to the chromatin for a while, and suggests post-transcriptional splicing events to be relatively frequent.

More recently, several studies used long-read sequencing in order to investigate splicing dynamics

[96, 94, 85]. Long-reads can identify different transcript isoforms, as well as determine the splicing order (“is the first transcribed intron spliced first?”). The distance between an exon-intron splice junction and the PolII position gives an estimation of the time-scale at which transcription and splicing occur. PolII position is given by the 3’end of nascent RNA, when reads are sufficiently long. Budding yeast *S.Cerevisiae* is a simple model to study splicing, as most genes contain only one intron and there is no alternative splicing. In this organism, for 50% of the genes, splicing is completed when PolII is 45 nucleotides downstream the 3’SS [85]. In a more complex, multi-introns model, such as murine erythroleukemia cells (MEL), splicing events were completed, on average, in a spatial window of 75 to 300 nt upstream PolII position [94]. Thus, spliceosomes are physically close to the transcription machinery, and splicing occurs in the same time-scale. An “all-or-none” behavior was also highlighted (either all intron spliced, or none), as also shown in fission and budding yeasts [86, 97]. Moreover, the authors of [94] showed a coordination between splicing and 3’end processing: unspliced transcripts tend to be poorly cleaved, while completely spliced transcripts also have an efficient 3’ cleavage. The estimation of physical proximity of RNA PolII and splicing events is discrepant between groups. Drexler et al. used Nanopore technology to sequence nascent RNAs from human cells (K562 and BL1184) and Drosophila S2 cells, labeled with 4sU and purified by cellular fractionation [96]. The vast majority of human transcripts were not spliced until after PolII had moved at least 4kb downstream a 3’SS. In Drosophila, however, splicing occurred on a much shorter scale (2kb for the majority of genes). The different kinetics are probably related to the different splicing mechanism (“exon-definition” for long mammalian intron, and “intron-definition” for short fly introns). Moreover, order of intron splicing did not follow the “first-come, first-served” model, thus , introns that are transcribed first are not necessarily spliced first.

To summarise, co-transcriptional splicing probably predominates for most introns. Transcription and splicing occur at similar rates and are physically closed, and splicing is coupled to other co-transcriptional processes such as 3’end cleaving and 5’ capping [98]. But it is also clear that a fraction of introns is removed after transcription completion, and that amount varies between organisms and tissues. Even within a gene, splicing does not affect all the introns the same way: terminal intron and introns flanking alternatively spliced exons are removed more slowly than the constitutive introns [87, 91], Co-transcriptional splicing allows a fast and efficient RNA processing, while post-transcriptional could offer extra time for further regulation [99].

Circadian regulation of splicing efficiency

Splicing efficiency is sensitive to various environmental stresses, such as heat shock, osmotic change, or genotoxic agents [100], but has not yet been described as a rhythmically regulated step genome-wide. However, a gene-specific example of altered splicing efficiency has been shown for *Cirbp* mRNA, encoding the cold-inducible RNA-binding protein CIRBP. In NIH3T3 fibroblasts, *Cirbp* mRNA shows robust oscillations that are driven by physiologically relevant temperature variations. However, pre-mRNA level stays constant at different temperatures. Gotic. et al. introduced the concept of splicing efficiency, or splicing “proneness”, which is the fraction of pre-mRNAs available for splicing. A mathematical model showed that splicing efficiency of *Cirbp* pre-mRNA was higher at low temperature, which results in a rhythmic accumulation of *Cirbp* and the CIRBP protein. CIRBP interacts with several clock mRNAs, and particularly *Clock*. Loss of CIRBP in NIH3T3 fibroblasts resulted in depletion of cytoplasmic *Clock*, but not in the nucleus, suggesting that CIRBP normally protects *Clock* mRNA in the cytoplasm, or alternatively, promotes its fast export.

Alternative splicing

A single gene locus can produce different transcripts based on the choice of splicing sites, a process called alternative splicing (AS). AS increases the variety of isoforms and proteins, but also generates premature termination codon (PTC), which results in a fast translation-dependent degradation, namely nonsense-mediated decay (NMD). It is a widespread phenomenon in mammals, as 80% of multi-exon genes are affected by AS [101]. It is regulated by splicing factors that either promote or block the access to splice sites from the spliceosomes.

Many splicing factors are robustly cycling [102, 103]. They include RNA-binding proteins (RBPs) such as heterogeneous RNA Proteins (hnRNPs), and SR proteins that can be additionally rhythmically phosphorylated [104]. Rhythmic alternative splicing events happen on a genome-wide scale in many organisms and metabolic tissues such as mouse liver [103] and pancreas (regulated by THRAP3)[105], mainly driven by temperature variations. AS can feedback to the circadian clock system: for example, *U2af26* has two rhythmic, light-inducible isoforms that interact with PER1 protein, modifying its stability and impacting the entire clock system in mice [104]. It was later shown that temperature variation was driving alternative splicing of *U2af26* and other mRNAs, thanks to rhythmic phosphorylation of some SR proteins [106].

Splicing efficiency and alternative splicing have thus been shown in specific cases to be a rhythmic regulatory step, mainly driven by temperature variations.

Intron retention

Fully-transcribed, polyadenylated RNA **r**etaining an **i**ntron are called RI-RNA. Retained introns are defined by weak splice sites, are often flanked by alternatively splicing exons, are shorter than the average intron, and have a high G/C content [107, 108]. They are particularly prevalent in neurons and immune cells and affect up to 80% of protein coding genes in all tissues [109]. The presence of an unspliced intron has different consequences on the transcript's fate [110]. Some RI-RNA are exported to the cytoplasm. Presence of the intron can influence translational initiation efficiency and mRNA stability, but also introduces a PTC that leads to NMD degradation [110]. But the role of intron retention is more often associated with the nucleus. Indeed, RI-RNA are usually enriched in the nucleus, and some introns are exclusively found there, in which case they are referred to as "detained intron" [108]. In the nucleus, RI-RNA are either degraded by the nuclear exosome, or serve as a pool of precursor RNAs, awaiting for an external signal to trigger their maturation, export, and translation. For instance, half of *Clk1/4* transcripts retain introns 3 and 4 [111], and are abundant in the nucleus. Heat-shock events or inhibition of Clk1/4 kinase activity by a drug both triggered the maturation of the unspliced transcripts by dephosphorylated SR proteins. The subsequent elevated level of CLl1/4 kinases in turn phosphorylated back SR proteins and downregulated the splicing event. This is an elegant model for homeostatic auto-regulation based on a reservoir of immature RNA transcripts in the nucleus. Another example is in mouse neural cells, where a pool of polyadenylated RI-RNA stably accumulates in the nucleus. Upon neural activity, these RI-RNA are spliced, exported, and actively transcribed [112]. Therefore, Intron Retention is a post-transcriptional mechanism for fine-tuning gene expression, for example by retaining mature, polyadenylated mRNAs in the nucleus [107].

1.2.4 Nuclear export

A newly synthesised pre-mRNA is never "naked", but is always associated with a spectrum of ribonuclear proteins, which mediate and coordinate different processing steps [113]. They ensure that only correctly processed mRNAs transit through nuclear pore and reache the cytoplasmic translation machinery, avoiding the translation of spurious transcripts with potential harmful side-effects [114]. The general mRNA export pathway requires the recruitment of the TRansport-EXport TREX protein complex, mainly composed of the THO subcomplex (THOC1/2/3/6/7), Alyref, UAP56, and plethora of other proteins, during transcription elongation [115]. The subunits of the complex are recruited by different elements, such as the C-terminal Domain of the RNA PolII and by the spliceosome machinery, thus coupling the process of transcription, splicing, and export. The complex then facilitates the loading of other factors and adaptors proteins such as SR proteins onto the mRNA and the packaging of the functional mRNA–protein complexes (mRNPs) for nuclear export. mRNA is then handed over to the transport receptors NXF1:NXT1, which physically interacts with the nuclear pore and promotes the shuttling of the transcript. The translocation through the nuclear pore itself is fast, occurring in about 0.5 second [116].

Variation of export rates depend on several factors. For instance, 5' capping, splicing and polyadenylation enhance the recruitment of the TREX complex and therefore, promote nuclear export [114]. However, splicing is not a necessary step for efficient export, as intron-less RNAs derived from cDNA are also well exported by the TREX pathway [117]. Interactions with RNA-Binding proteins, which recognise specific sequences (motifs), secondary structure, or RNA modifications, also dictate the fate of the transcripts. In search of *cis*-acting motifs explaining the long retention of some specific RNAs, two parallel high-throughput studies have identified sequences enriched for cytosins, which promote nuclear localisation of the transcripts [118, 119]. Specifically, sequences enriched in Alu repeats drive the nuclear localisation through binding of the hnRNPK (Heterogeneous nuclear ribonucleoproteins) on C-rich regions. Interestingly, hnRNPK has been shown to stabilise the circadian gene *Per3* [120]. Several other RBPs have been shown to interact with the circadian clock molecular machinery. For example, CIRBP interacts with the circadian clock machinery, potentially regulating the export of *Clock*: depletion of CIRBP led to a reduction of rhythms in cultured mouse fibroblasts [121], and *Clock* mRNAs were enriched in the nucleus, either because the lack of CIRBP resulted in an inefficient export, or because of a fast degradation in the cytoplasm. Another RBP, NONO, is a core component of nuclear paraspeckles (see 1.2.5). It binds hundreds of rhythmically expressed transcripts in mouse liver tissue [122]. Loss of NONO protein results in loss of rhythmicity in some target genes, but also in both phase advance and phase delay of the peak time of the remaining rhythmic genes in the cytoplasm.

This suggests that NONO can both slow down or fasten processing steps (including export). NONO additionally modulates circadian rhythm expression through its binding to PER1 protein. Knockdown of this protein in mouse fibroblasts and Drosophila cells caused attenuation of rhythms [123].

Importantly, the mechanisms regulating nuclear export differ depending on whether the RNA is coding, in which case the ultimate function is to be translated by ribosomes in the cytoplasm, or non-coding, and have therefore a functional role in specific location (ribosomal RNA, tRNA, small nuclear RNA, small nucleolar RNA, long non-coding RNA). Recently, a lot of effort has been put into finding *cis-* and *trans-* acting elements driving the subcellular localisation of long non-coding RNA, particularly through regulating their nuclear retention, or conversly, their export [114, 124, 125]. Long-non coding RNA (lncRNA) are transcripts > 200nt that are often capped with 7-methyl guanosine (m^7G) at their 5'end and polyadenylated at their 3'end. Because they do not produce any protein, their localisation is tightly linked to their functions [124]. For a long time, lncRNA were thought to be enriched in the nucleus, where they regulate chromatin structure (FIRRE), gene expression (XIST for X chromosome inactivation), or act as scaffold of nuclear condensates (MALAT1, NEAT1). However, many lncRNA are also exported to the cytoplasm, either through the NFX1 pathway for single- or few-exon transcripts with long exons or high A/U content, or through the TREX complex for G/C-rich transcripts [126]. Studying the localisation of lncRNA not only helps to understand their biological functions, but can also reveal specific motifs or features that could also be valid to regulate protein coding mRNAs localisation [126].

mRNA export is therefore one of the many important layers in the regulation of mammalian gene expression pathway, most likely coupled with other processes such as splicing and polyadenylation [113]. An additional role of the nuclear retention time is the buffering of noise arising from stochastic bursts of transcription [127]. The discontinuous synthesis of RNA generates large fluctuations of transcripts abundance, but such noise can be attenuated in the cytoplasm by imposing a (constant) time delay at the nuclear pore [128].

1.2.5 Nuclear speckles and paraspeckles are hotspots of RNA processing factors

Nuclear speckles

Nuclear speckles (NS) are phase-separated membraneless organelles found in the nucleus, in the interchromatin space. These compartments are condensates of RNA-Binding Proteins, built around the scaffold long-non-coding RNAs *Malat1*. A striking amount of poly(A)+ transcripts transit through NS, suggesting that they are a hotspots for mRNA processing [129].

Nuclear speckles are sometimes called “splicing speckles” due to the high concentration of splicing factors (snRNPS), SR and SR-like proteins such as SON, SRRM1, SRRM2, SRSF1 and SC35 among the hundred of proteins found in this compartment with distinct roles (transcription regulators, 3' end processing, mRNA modification, and mRNA packing and export) [130]. NS act as storage/ modification sites of splicing factors, but also as active splicing sites containing active spliceosomes [129]. In HeLA cells, 20% of active spliceosomes are not in the chromatin fraction (co-transcriptional splicing), but rather found in NS [131]. Moreover, the long-non coding RNA *Malat1* has been shown to indirectly interact with many nascent transcripts and to localise to chromatin near active transcription sites, further suggesting its role in pre-mRNA processing [132]. Liao et al. proposed an interesting model called “interfacial splicing model”: exonic sequences are sequestered “inside” the NS by SR proteins, while intronic sequences are in the nucleoplasm, bound by hnRNPs [133]. The partitioning of the transcript exposes the exon-intron splice site at the interface of the NS, where spliceosomes are preferentially located.

In addition to splicing factors, NS are enriched in export factors like components of the TREX complex and proteins of the exon-junction complex (EJC) that assemble on exon-exon junctions after splicing. TREX and EJC are recruited on the pre-mRNA during splicing [134]. Interestingly, even intronless transcripts transit through NS to gain export-competence by associating with the TREX complex, suggesting that even if splicing and export are coupled, they can also act independently [135]. Another role of NS is to enhance transcriptional activity. Indeed, several studies using smFISH showed that highly transcribed genes are in close proximity to nuclear speckles, and transcriptional bursting frequency is increased [136]. Highly expressed genes that tend to be close to NSs have a higher splicing efficiency, probably due to the higher concentration and availability of splicing factors [137]. Thus, nuclear speckles act as a hub of RNA processing factors, potentially coupling efficient transcription, splicing, and packing RNA into export-ready mRNA-protein complexes.

Nuclear paraspeckles

Like nuclear speckles, nuclear paraspeckles (NPS) are membrane-less, phase-separated organelles found in the interchromatin region, but distinct from nuclear speckles. They were discovered more recently than nuclear speckles (nuclear speckles: 2002 [138], nuclear paraspeckles: 1910 by Cajal y Ramos). They are present in all mammalian cells, except embryonic cells. They assemble around a scaffold long non-coding RNA *Neat1*. More than 40 proteins have been characterized in the complex. NONO, SFPQ, PSPC1, RBM14, HNRNPK, FUS and SWI/SNF are essential for the stable assembly of NPS [139].

Nuclear paraspeckles play multiple roles in gene expression regulation, particularly in retaining certain RNA species. For example, mRNAs containing inverted repeats of Alu sequences (IRAlus) are likely to form double-stranded RNA regions and are targeted to Adenosine-to-Inosine editing by the ADAR protein [140]. These hyperedited mRNAs can be bound by nuclear paraspeckles components, and thus be retained in the nucleus [141]. One well-identified example is the mouse *Cat2* mRNA, important for the cellular stress response [142]. One of the isoforms, CTN-RNA, contains inverted repeat elements in its 3’UTR. CTN-RNA are retained in nuclear paraspeckles, but upon cellular stress, the retention element in the 3’UTR is cleaved off and the shorter transcripts are released into the cytoplasm where it is translated. In rat pituitary cell line (GH4C1), components of the nuclear paraspeckles display rhythmic accumulation (NONO, SFPQ, PSPC1, RBM14, and *Neat1*), leading to rhythmic variations in paraspeckle number [143]. By fusing an IRAlu element in the 3’UTR of an EGFP reporter, Torres et.al showed that the reporter mRNA was retained in the nucleus and released in the cytoplasm in a circadian manner, ultimately producing cycling EGFP [143]. The disruption of paraspeckles by knocking down

Neat1 abolished rhythmic EGFP production. Therefore, nuclear retention by paraspeckles appears to be a post-transcriptional mechanism involved in circadian gene expression.

Interestingly, in mouse liver cells, the mRNA and protein level of NONO is constant throughout the day. However, the subnuclear localisation of NONO changes in response to glucose: during feeding time, or following an intraperitoneal injection of glucose, NONO localises in nuclear paraspeckles [122]. NONO binds around a thousand mRNAs in the liver, and about one third exhibits circadian accumulation. The majority of NONO binding sites were within introns, suggesting that NONO plays a role in processing pre-mRNAs. In a mouse model lacking NONO, a hundred of the cycling NONO targets become arrhythmic. The remaining rhythmic mature transcripts, mainly involved in glucose and lipid metabolism, were on average delayed by 2 hours, while their transcription phase was unaffected. This delay in mRNA accumulation in absence of NONO suggests that in normal conditions, NONO post-transcriptionally enhances RNA processing efficiency, generating robust in-phase oscillations. Nuclear paraspeckles are usually thought to prevent RNA export, therefore, this novel proposed role of NONO might be independent from its association with paraspeckles.

1.2.6 N^6-methyladenosine (m^6A) RNA methylation

mRNAs undergo various modification throughout the maturation process. The most common mRNA modification is N^6-methyladenosine (m^6A) methylation [144]. m^6A are deposited by "writers", a complex including the core methyltransferases METTL3, METTL14, and several auxiliary proteins such as WTAP, VIRMA, and RBM15. m^6A are preferentially deposited within long exons, near stop codons, and at the 3'UTR, but also in intronic regions [145]. The modification is reversible, and m^6A can be removed by demethylases (also called “erasers”) ALKBH5 and FTO. The marks are then decoded by “readers” proteins, linking mRNA to the correct downstream processing pathway. Readers include heterogeneous nuclear ribonucleoproteins HNRNPA2B1, HNRNPC, and YTHDC1 in the nucleus, and YTHDF1/2/3 in the cytoplasm [145]. m^6A methylation regulates gene expression by influencing a wide range of processing steps. The nuclear readers HNRNPA2B1 and YTHDC1 mediate alternative splicing [146]. The export is slowed down when METTL3 is depleted [147], and enhanced when the demethylase ALKBH5 is downregulated [148]. In the cytoplasm, YTHDF1 increases translational efficiency by recruiting translation initiation factors[149], while YTHDF2 promotes mRNA deadenylation and degradation by recruiting the deadenylase complex CCR4-NOT [150]. Additionally, it recruits RNaseP/MRP complex to promote endoribonucleolcytic cleavage [151].

The global coordination of methylation by writers, erasers, and readers could be a mechanism to sort transcripts into a “fast track” for processing, translation and decay [144]. m^6A methylation plays an interesting role in setting the correct circadian period length [147]. Global inhibition of RNA methylation by DAA, an inhibitor of methylation, lengthened the circadian periods of cultured cells and locomotor activity of mice by 3h and 1h. Moreover, the specific knockdown of METTL3 in U2OS and MEFS cells caused a prolonged nuclear retention of two clock genes *Per2* and *Bmal1*, which resulted in a lengthened circadian period. Thus, suppression of METTL3 was sufficient to slow down the clock by delaying RNA processing (potentially the export step).

1.2.7 Modulation of cytoplasmic mRNA stability

In the cytoplasm, mature mRNAs interact with initiation factors that bind to the 5'cap and to the 5'UTR to induce the formation of the translation initiation complex, which in turn facilitates the recruitment and the assembly of the 40S and the 60S ribosomal subunit, and eventually starts the elongation of an amino acid chain (protein synthesis). The mRNA transcript is protected against degradation by the presence of the 5' 7-methylguanosine cap and of the 3'end poly(A) tail that interact with the cytoplasmic proteins eIF4E and with the poly(A)-binding protein (PABP) [152]. Degradation is initiated if the transcript is internally cleaved by endonucleases, but for the majority of genes, degradation occurs when one of the two protective features is missing, either through decapping, or through shortening of the poly(A) tail by deadenylases. One deadenylase, *Nocturnin* (*Ccnrl4*), is rhythmically transcribed in various mouse tissues [153]. Hundreds of genes in the mouse liver were found to show rhythmicity in poly(A) tail length (defined by the ratio of "long" versus "short" tail) [154], although the variation in length was not directly related to the deadenylase activity of NOCTURNIN [153]. Moreover, the poly(A) tail lengths correlate well with protein levels encoded by those mRNAs, further suggesting the role of the poly(A) tail, and particularly the deadenylation rate, in the regulation cytoplasmic mRNA stability [155].

Interaction with RBPs is another mechanism influencing rhythmic RNA degradation. For example, the three *Periods* genes in mice are targeted by different RBPs: *Per3* is stabilised by the heterogeneous ribonucleoprotein hnRNP K and destabilised by hnRNP D [120]. *Per2* is destabilised by PTB (hnRNP 1) [156]. Additionally, hnRNP proteins can also increase the translational efficiency, as shown for *Per1* with its interaction with hnRNP Q. *Per1* translational efficiency is also regulated by another RBP called LARK [157]. Regulation of the stability of *Per* gene has been first demonstrated in Drosophila [158], with *Per* transcript being more stable during the rising phase, and destabilised during the descending phase. Mathematical models showed that theoretically, varying the degradation rate allows fine-tuning the phases and amplifying amplitudes [54, 78, 79]. Two studies estimated that ~30% of the mouse liver rhythmic transcriptome is subject to rhythmic degradation [78, 79].

Collectively, all the RNA processing steps including transcription (chromatin conformation, transcription factors recruitment, histone modification, RNA Polymerase recruitment), co- and post-transcriptional modifications occurring in the nucleus (splicing, methylation, loading of RBPs, retention in nuclear and paranuclear speckles, export through the nuclear pores) and regulation of cytoplasmic mRNA stability, shape the circadian transcriptome. This complex regulatory network ensures a tight temporal coordination of biological functions.

1.3 Liver zonation

Most hepatic functions are not only temporarily partionned, but also spatially: specific subpopulations of hepatocytes carry out different biochemical functions based on their location within the liver, a phenomenon called **liver zonation**.

1.3.1 Structural organisation of the murine liver

Despite its homogeneous aspect, the liver is highly structured on the cellular scale. It is mainly made of hepatocytes (60% of the liver cell population, and 80% of the mass), and non-parenchymal cells: endothelial cells, cholangiocytes (lining bile ducts), Kupffer cells (resident macrophages) and Stellate cells (storage of fat and vitamin A). Hepatocytes are arranged in *lobules*, which are anatomical and functional subunits containing morphologically indistinguishable hepatocytes arranged in 10 to 15 concentric layers [159], often represented with a hexagonal shape, although it is not always the case, especially in rodents. Blood from two different sources enters into the lobules from the periphery (corner of the hexagon): 75% of the blood supply comes through the portal vein (PV), originating from the gastro-intestinal tract, and transports macromolecules (nutrients, toxins) absorbed in the intestines. The remaining 25% of the blood supply enters the lobule from hepatic arterioles, and provides oxygen to liver cells [159]. Once in the lobule, blood travels through small capillaries called sinusoids and drains to the central vein (CV). Bile flows in the opposite direction in bile canaliculi, from the center toward the periphery of the lobule. Bile is then excreted and stored in the gallbladder, awaiting for fat-containing bolus to enter the intestines. Historically, the lobules were divided in three discrete zones, with associated genes and proteins: the periportal zone (PP) around the portal node (portal veins, bile duct and hepatic arterioles), the pericentral zone (PC, around the central vein), and the midlobular zone (Mid), comprised between the PP and PC zones. Recently, the group of Itzkovitz refined the "porto-central" axis and subdivided the lobule in 9 continuous layers[160]. The concentration of oxygen, morphogens (particularly Wnt ligands [161]), nutrients, hormones (insulin, glucagon, thyroid hormones) and other biomolecules change as the blood runs along the porto-central axis. This spatial polarisation of the microenvironment dictates the functional roles of hepatocytes and their gene expression profile based on their position. Such compartmentalisation of metabolic pathways is thought to allow two (or more) opposite pathways using the same substrate to run in parallel, and avoid competition for the same substrates [162], much like the time-gating by the circadian clock. Furthermore, metabolic cascades can be spatially distributed, such that an intermediate metabolite can be transferred from one cell to another, similar to a “production line” [163].

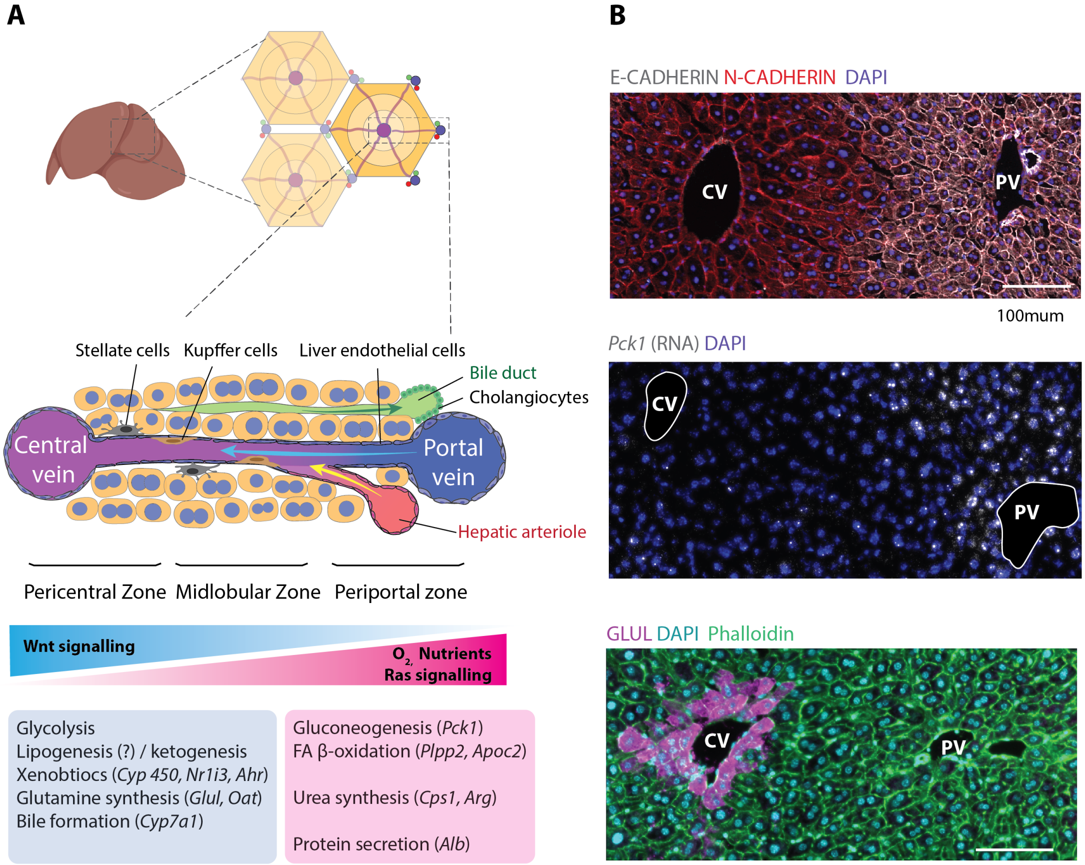


Figure 1.4 – Structure of the liver lobules and zonated liver functions. A: Classic hexagonal representation of a liver lobule with a central vein (CV) in the middle, et portal triad at the periphery. Portal triad contains a portal vein (PV), hepatic arteriole, and bile duct. Blood flows from the portal vein and hepatic arterioles toward the central vein through sinusoids, lined by fenestrated endothelial cells. The liver also contains Kupffer cells (resident macrophages) and Stellate cells in the small space between the sinusoid and the cord of hepatocytes, named Space of Disse (not represented). The cord of hepatocytes is divided in three zones: Pericentral zone (PC), Midlobular zone (Mid) and Periportal zone (PP). Gradients of oxygen, Ras, and Wnt determine the zonation of metabolic functions listed in the blue and red box. B: histological staining of mouse liver sections showing the zonated patterns of: E-CADHERIN and N-CADHERIN (immunofluorescence), *Pck1* (white dots, smRNA-FISH), and GLUL (immunofluorescence combined with membrane staining with Phalloidin). PV = portal vein, CV = central vein.

1.3.2 Zonated hepatic functions

Early studies of zonation used immunohistochemistry, immunofluorescence, and ISH techniques. Later, perfusion by digitonin followed by collagenase allowed the collection of a large amount of cells: portal, respectively retrograde perfusion of digitonin damages periportal, respectively pericentral cells. The undamaged cells are then collected by collagenase infusion. The first genome-wide study was performed by Brauening and colleagues, who quantified gene expression of periportal and pericentral hepatocytes with microarray, confirming the zonation of many previously known metabolic functions. In 2017, the group of Itzkovitz performed an elegant genome-wide reconstruction of spatial gene expression profiles. They combined two powerful methods - single-molecule RNA Fluorescent *in situ* hybridisation (smRNA-FISH) and single-cell RNA-sequencing (scRNA-seq) of dissociated hepatocytes –

to infer the position of each cell within the lobule. The transcriptome of each hepatocyte was compared to the spatial patterns of 6 zonated landmark genes determined by smRNA-FISH. Once the hepatocytes' coordinates were recovered, the spatial profiles of all the genes expressed in the liver were reconstructed. This study revealed an unexpected breadth of liver spatial heterogeneity, with ~ 50% of genes having a spatially non-uniform pattern. Among them, functions related to carbohydrate catabolic and anabolic processes, ammonia removal, xenobiotics detoxification, bile acid and cholesterol synthesis, fatty acid metabolism, Wnt and Ras targets, and hypoxia-induced genes were zonated, and will be briefly described later.

Because mRNA levels do not necessarily match protein levels (that ultimately dictate a cell function), the same group investigated again zonation profiles, but this time, at the level of proteins [163]. They sorted single hepatocytes by FACS based on known zonated membrane markers (CDH1, CD73) and reconstructed a proteomics zonation map. They demonstrated that 50% of the 3000 detected proteins were zonated, with a high correlation with their matching mRNA profiles. Interestingly, a key hepatic transcription factor, HNF4α, is uniformly expressed within a lobule but has a zonated protein profile (Periportal). Moreover, some miRNA such as miR-122-5p and miR-30a-5p were zonated with an inverted pattern compared to their target genes, suggesting that zonation is also regulated post-transcriptionally.

Spatial patterns can be "steep" and restricted to a few layers of hepatocytes, as seen with the expression of Glutamine Synthetase (*Glul*, Fig.1.4), or have smoother gradient-like patterns. Moreover, some genes such as *Hamp* are enriched in the midlobular zone [160]. Patterns are more or less dynamic in response to changes of nutrients, drugs, hormones, etc [164]. Metabolism of glucose was one of the first functions to be studied in the context of liver zonation, by several seminal works by Jungermann and coworkers. For instance, they showed that gluconeogenesis - driven by the rate-limiting enzyme *Pck1*- is carried out by periportal hepatocytes [165](Fig.1.4). Calculation of the glucose / glucose-6-phosphate flux in the periportal and pericentral zones in rat liver showed that glycolysis, the antagonist pathway of gluconeogenesis, mainly occurs in pericentral hepatocytes. This polarisation of gluconeogenesis / glycolysis was further confirmed by gene expression studies by Brauening et al. in 2006. However, glucokinase mRNA is not zonated, suggesting that regulation of its activity is post-translational [166].

Lipid metabolism is another zonated hepatic activity. The observation that steatosis (the accumulation of lipid droplets in hepatocytes that eventually provoke cirrhosis) primarily occurs around the central vein suggested the idea that *de novo* lipogenesis occurs more in the pericentral zone, while the consumption of lipid (fatty acid degradation), happens more in the periportal area. Expression *Plpp2*, and *Apoc2*, key enzymes in fatty acid β-oxidation, are mainly in the periportal zone [166]. The activity of CPT1, a rate-limiting enzyme involved in the translocation of fatty acid from the cytosol to the mitochondria, is also higher in the periportal zone [167], even if the gene expression profile is homogeneous within the lobule [168]. On the other hand, the activity of acetyl-CoA carboxylase (ACC), ATP citrate lyase (ACLY), and Fatty Acid Synthase (FASN), involved in lipogenesis, were higher in the pericentral zone [169]. Surprisingly, the mRNA expression levels of these three genes is higher in the periportal area[168]. Therefore, there are contradicting studies between the mRNA expression and enzymatic activity, suggesting that lipid metabolism may not be as clearly zonated as other hepatic functions, and depends on various parameters such as sex and fasted/fed state [170].

One main function of hepatocytes is the production of bile acids, which helps the absorption of fat in the intestines. In the classic pathway, cholesterol taken from the blood is converted into bile acids by a cascade of enzymatic steps. The rate-limiting enzyme, CYP7A1, is expressed pericentrally. The next enzyme in the chain, CYP8B1, is depleted in the first pericentral layer and peaks in the second layer of hepatocytes. This suggests that the intermediate metabolites produced by CYP7A1 may be transferred to the next layers of cells. CYP27A1 and BAAT, enzymes acting in later steps of the cascade, are also depleted in the most pericentral layer. This is a nice example of the “production line” pattern, where spatial order matches the cascade sequence [160].

The liver is in charge of whole-body drug clearance. Many of the enzymes involved in xenobiotics metabolism (members of the Cytochrome P450 family) are located around the central vein (CYP2E1, CYP1A2, CYP2B, etc.), except CYP2F2 that is periportal [171]. Other key enzymes involved in the early steps of detoxification were found pericentrally: Flavin-containing monooxygenases (*Fmo1, 2, 5*), the rate-limiting enzyme of heme biosynthesis *Alas1*, and the sensors of xenobiotics *Ahr* and *Nr1i3* (CAR) [168]. The spatial patterns of Phase II enzymes are more complexe: glucuronidation happens in the pericentral zone, while glutathione conjugation is periportal [171]. The zonation of xenobiotic metabolism can explain heterogeneous patterns found in drug-induced liver pathology. For instance, the intermediate molecules produced during acetaminophen degradation by Cyp450 are cytotoxic. Overdose of acetaminophen thus damage exclusively hepatocytes expressing CYP2E1 and CYP1A2 [172], resulting in tissue necrosis around the central vein.

A last zonated hepatic activity is the urea cycle: ammonia and glutamine are taken up and metabolised to urea by periportal hepatocytes (*Cps1, Arg1*). This process releases glutamate in the bloodstream. The excess ammonia that was not metabolised by periportal hepatocytes reaches the central zone. There, NH4+ is further transformed to glutamine by glutamine synthetase (GLUL), which additionally needs glutamate for this process. The glutamate-glutamine homeostasis is an example of "spatial recycling", where a metabolites produced in one zone (glutamate) is recycled in another zone [173]. Of note, Glutamine Synthetase is exclusively expressed around the central vein (1-3 layers of hepatocytes) and has been considered as a “stable” zonation marker (Fig.1.4).

1.3.3 Factors regulating hepatic zonation

The blood flowing in the sinusoids is a mixture of blood from the portal vein and the hepatic arteriole. Its composition gradually changes, as oxygen is consumed, metabolites are produced or eliminated and substrates are modified. Partial oxygen pressure is high in periportal blood, and drops by 50% in pericentral blood [174]. Thus, it is not surprising that periportal hepatocytes contain more oxygen-demanding mitochondria. Protein translation and subsequent secretion is another high ATP-demanding task. Thus, Albumin, complement system proteins, and blood clotting factors are preferentially produced by periportal hepatocytes, although these tasks are not strictly restricted (*Alb* mRNA is highly expressed in every layer). Cells adapt their transcriptional program to oxygen availability via hypoxia-inducible factor (HIF). The three isoforms of HIF are preferentially expressed in the pericentral area, where oxygen pressure is lower.

One of the most well-known master regulator of the zonation pattern is the Wnt/ β-catenin signalling pathway [161, 175]. In absence of a ligand, β-catenin is bound to the cell membrane and interacts with cadherins (of note, E-cadherin is exclusively periportal while N-cadherin is pericentral, Fig.1.4). β-catenin resides in a “destruction” multiprotein complex together with GSK3, APC, CK1, Axin, and DVL. When a Wnt ligand binds to Frizzled and to its co-receptors, such as LRP5 and LRP6, the Axin complex is dismantled, β-catenin is stabilised, translocates into the nucleus. There, it interacts with the DNA-bound TCF and LEF on Wnt Response Element (WRE), activating the expression of target genes. When β-catenin is absent, TCF4 preferentially binds HNF4α consensus sites in the opposite periportal region [176], positively regulating some genes involved in lipid and glucose metabolism.

Studies in liver tumors and other transgenic mice first suggested the idea that Wnt/ β-catenin signalling pathway was associated with the pericentral zone. In liver-specific APC-KO mice, where β-catenin is not repressed anymore, investigators observed a “pericentralisation” of the lobule, resulting in a reduced expression of *Pck1* and urea cycle enzymes, and the expansion of the Glutamine Synthetase positive area [161]. On the other hand, transgenic mice with liver-specific loss of β-catenin resulted in the absence of the pericentral *Cyp1a2* and *Cyp2e1* [177]. Finally, in the genome-wide study by the group of Itzkovitz [160], about 30% of the zonated genes were known Wnt target genes. Positively regulated genes were more pericentral, whereas negatively regulated genes were periportal. Liver endothelial cells lining the central vein secrete some of the Wnt ligands (WNT2, WNT9B [178, 179]), creating short-range gradients. β-catenin acts in concert with Hedgehog signalling pathway (Hh). Hh signalling components have the opposite spatial pattern (periportal enrichment). The mutual repression of the two pathways could shape the homeostatic liver functions [180]. A mathematical model explained how a shift of balance of one of them, as it is the case in mutant mice, lead to a pericentralisation or periportalisation of the lobule [181].

Altogether, these studies show that zonation of liver function is orchestrated by the tight balance of nutrients, oxygen and hormones availability (e.g. glucagon [182]), and interacting signalling pathways, including local gradients of Wnt ligands expressed by endothelial cells and Hedgehog. Additional pathways modulate gene expression, such as Ras / MAPK/ ERK, are thought to be periportal [166].

1.3.4 Zonation of polyploidy

Hepatocytes have a distinguishable morphological feature: polyploidy, which is an increase of the number of chromosome sets per cell, and of the number of nuclei per cell. At birth, most hepatocytes are mono-nucleated with a standard two sets of genome copies. After weaning, bi-nucleated hepatocytes are generated following cytokinesis failure [183], probably in response to insulin signalling. Ploidy increases following mitotic failure, producing nuclei with 2, 4, 8, up to 16 copies of the genome. The majority of the hepatocytes are mononucleated with 2n or 4n, or binucleated with 2x2n or 2x4n. These proportions vary with age, stress, or surgery [184]. A study based on 3D reconstruction of mouse liver showed an enrichment of 2n cells around the portal vein, and a depletion of 2x2n cells in the midlobular zone [185]. Another study confirmed some of these zonation profiles by establishing a spatial map of polyploidy in the liver of mice aged between 2 weeks and 12 months [186]. Polyploidisation progresses with age, but the speed varies depending on the zone. Finally, the ploidy might be diurnally orchestrated

[62]: by quantifying tissue sections, the authors observed that the percentage of 2x2n hepatocytes was maximal around ZT6, while the proportion of 1x4n cells was antiphasic. Cells with other combinations of nuclei and ploidy were not rhythmic. This observation suggests a possible link between ploidy, the liver zonation and circadian rhythm.

1.3.5 Dynamic regulation of the liver zonation

Many of the key zonated hepatic functions are also temporarily regulated, orchestrated by the interplay of the circadian clock, feeding cycles, and systemic signals. For instance, the rate-limiting enzyme of gluconeogenesis *Pck1*, a well-characterized portally expressed gene is maximally transcribed toward the end of the fasting time (ZT10) for *de novo* glucose synthesis. However, the spatial zonation profiles have always been analysed as a static phenomenon. Few groups studied the dynamical aspect of the zonation, but always in the frame of nutrient changes (especially for carbohydrate metabolism [162]) or in relation with ageing [186]. To bridge the gap between the fields of the circadian rhythms and the liver zonation, Naef laboratory at EPFL and the Itzkovitz laboratory at the Weizmann Institute of Science published a collaborative work [168]. Single-cell RNA-seq of hepatocytes was performed at four different time-points of the day (early morning, mid-day, early night and mid-night), as in [187]. Spatial profiles were reconstructed based on 27 zonated, non-rhythmic landmark genes from the previously published dataset [187]. A mixed-effect model describing both spatial (position along the 8 layers) and temporal (phase and amplitude) features were fitted to the reconstructed profiles. In total, ~5000 genes were classified based on their probability of being zonated (Z), rhythmic (R), independent rhythmic-zonated (Z + R; rhythmic parameters are the same in all layers) or interacting (Z × R; rhythmic parameters depend on the layer). 30% of the genes expressed in the liver are significantly zonated, 20% are rhythmic, and 7% are both temporally and spatially controlled. Dually regulated genes include the well known key hepatic functions such as carbohydrate, lipid, and amino acid metabolism, but also previously unknown zonated functions such as protein synthesis, proteasomal activity and mitophagy. This study revealed the broad richness of the spatio-temporal gene expression dynamics in the mammalian liver.

2 Objectives of the thesis and presented work

The liver is a central hub for whole-body metabolic homeostasis. It is tightly regulated by the endogenous circadian clock and the feeding cycles to perform the right functions at the right time of the day [188]. As one of the organs with the highest number of cycling genes, the mammalian liver has been extensively studied by chronobiologists [28]. Several high-throughput time-course studies have investigated the circadian rhythms of the hepatic cistrome, transcriptome, proteome, acetylome, and lipidome, revealing the richness and the complexity of temporal coordination of hepatic functions and the extensive crosstalk between the circadian clock and metabolism (to cite only few studies: [47, 59, 60, 45, 62, 46]).

In this thesis, I explore different aspects of temporal gene expression regulation in the mouse liver, with an emphasis on *in situ* localisation of RNA transcripts.

Spatio-temporal gene expression profile in the mammalian liver

First, I investigated how RNA transcripts are differentially expressed at the tissue scale. In the liver, cells are arranged in a structural unit called a lobule, in which blood flow generates gradients of oxygen and morphogens. This polarised microenvironment leads to a differential gene expression and dictates the physiological status of hepatocytes, a phenomenon called liver zonation. Many zonated genes are additionally modulated by the circadian clock or the feeding/fasting cycle, showing robust rhythmic patterns at the tissue level. Until now, circadian gene expression studies have been systematically performed at the bulk level, while single-cell spatial studies have neglected the temporal regulation of the liver. Thus, how rhythms are implemented in the different zones within the lobule and at the single-cell level remained unknown. Moreover, the spatial pattern of core clock genes, which regulate the rhythmic expression of many zonated genes, have never been described. In this project, I quantified spatial gene expression profiles using single-molecule RNA Fluorescent *in situ* hybridisation (smFISH). I developed an image analysis pipeline to detect, quantify, localise, and model smFISH signals in liver tissues. This work became part of the collaborative project between the Naef group and the Itzkovitz group from the Weizmann Institute of Science (Rehovot, Israel). My work contributed to the paper “Space-time logic of liver gene expression at sub-lobular scale”, published in Nature Metabolism by the following authors (co-first authors in bold): **C. Droin, J. El Kholtei, K. Bahar Halpern**, C.Hurni, M.

Rozenberg, S. Muvkadi, S. Itzkovitz, F. Naef [168]. I validated spatio-temporal gene expression profiles in liver tissue by smFISH. Particularly, I showed that the circadian clock is largely non-zonated, and is therefore robust to heterogeneous niche signals in the different lobular zones.

Comprehensive analysis of the circadian hepatic transcriptome at subcellular scale

Second, I explored the localisation of RNA transcripts at the scale of liver cells, comprising a majority of hepatocytes. The eukaryotic RNA is subject to extensive regulation from transcription to degradation. The interplay of the kinetic parameters governing each step between these two endpoints influences the transcript abundance in different subcellular compartments. Moreover, each of these steps provides an opportunity to regulate gene expression in a circadian manner, in order to tune the phase, modulate the amplitude, or generate *de novo* post-transcriptional rhythms. In this work, I investigated the differential accumulation of nuclear and cytoplasmic transcripts by combining time course RNA-seq experiments of fractionated mouse liver cells, mathematical modelling, and smFISH. I investigated how the combination of rates of different RNA processing steps, specifically the nuclear export rate, cytoplasmic degradation rate, and the extent of co- and post-transcriptional splicing, can shape the subcellular RNA distribution at different times of the day.

First, I explored the relationship between two RNA populations (nuclear pre-mRNA and nuclear mRNA, or nuclear mRNA and cytoplasmic mRNA) without the temporal dimension. I showed that while the majority of protein coding transcripts are more abundant in the cytoplasmic fraction relative to the nuclear fraction, those that are nuclear-enriched often code for nuclear proteins, suggesting a concordance of RNA and protein localisation. Correlation with transcript length showed that short protein coding mRNAs are more abundant in the cytoplasm, while longer transcripts are preferentially found in the nucleus. Additionally, transcript length also influences the extent of co- versus post-transcriptional splicing: long transcripts are more co-transcriptionally spliced, while short pre-mRNA are already polyadenylated before splicing completion.

Then, I used a mathematical model previously developed in the Naef laboratory in order to estimate nuclear export rates and cytoplasmic degradation rates. By comparing rhythmic profiles (phases and relative amplitudes) of nuclear pre-mRNA, nuclear mRNA, and cytoplasmic mRNA, I could infer the nuclear and cytoplasmic lifetime of ~1400 genes. The median half-life of a rhythmic transcript in the cytoplasm is 2.5h, while the median nuclear lifetime is shorter than 30 minutes. For a majority of the genes, nuclear lifetime has only a minor contribution to the total RNA lifetime. However, a subset of metabolic genes remain in the nucleus for more than one hour (up to four hours), which accounts for the long phase delay between the peak times of transcription and of cytoplasmic accumulation. Additionally, we observed rhythmic patterns in the nucleus that most likely originate from a rhythmic regulation of the nuclear export rate, affecting ~10% of the oscillations of nuclear transcriptome, and driving the rhythms of ~100 nuclear mRNAs. This work suggests that mRNA oscillations can be post-transcriptionally regulated at the level of nuclear export, and provides a global and quantitative view of these processes.

Circadian and chromatin contacts-dependent modulation of transcriptional bursting parameters

Finally, I estimated transcriptional bursting parameters of a core clock gene, *Cry1*, in mouse liver tissue. I used smFISH probes targeting intronic and exonic region of *Cry1*, and inferred the burst frequency (rate of switching between periods of transcriptional inactivity and activity) and burst intensity (average number of RNAs transcribed per burst episode) at two times of the day. I also estimated these parameters in a mouse model with a deletion of an enhancer that is rhythmically recruited to the *Cry1* promoter (*Cry1*Δe mice, [64]). I showed that the burst fraction (proportional to the burst frequency), but not the burst size, regulates the rhythmic gene expression level of *Cry1*. Moreover, the burst fraction of *Cry1* is modulated by rhythmic promoter-enhancer contacts. This project is a contribution to the paper of my colleagues Jérôme Mermet and Jake Yeung: "Clock-dependent chromatin topology modulates circadian transcription and behavior", published in Genes and Development by the following authors (first authors in bold): **Jérôme Mermet, Jake Yeung**, Clémence Hurni, Daniel Mauvoisin, Kyle Gustafson, Céline Jouffe, Damien Nicolas, Yann Emmenegger, Cédric Gobet , Paul Franken, Frédéric Gachon, Félix Naef [64].

3 Spatio-temporal gene expression profile in the mammalian liver

3.1 Background

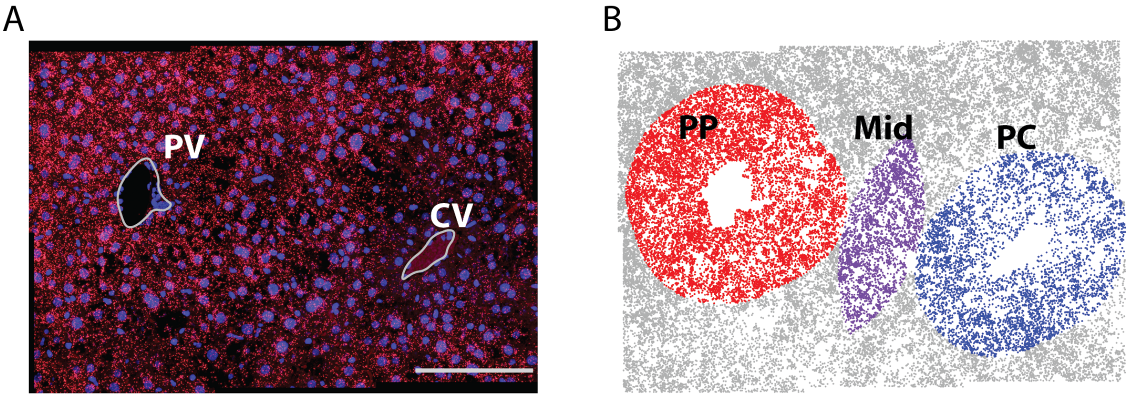
Many physiological functions of the liver are temporally orchestrated by the interplay of the endogenous circadian clock, systemic signals, and feeding rhythms [189]. In addition to being regulated in time, the liver is also structured in space [162]. It is composed of repeating anatomical units termed lobules in which blood flow creates gradients of oxygen and morphogens from the portal vein to the central vein. These variations of the microenvironment lead to differential gene expression and dictate the physiological status of hepatocytes. This phenomenon is called liver zonation (see 1.3.1). Despite the known tight temporal regulation and the extensive zonation of the liver functions, how time and space act in concert is still unknown. Studying chronobiology at the scale of a lobule could reveal new diurnally regulated liver functions that are usually hidden in bulk analysis, and unveil spatial parameters of circadian genes.

3.2 Results

3.2.1 Single-molecule RNA-FISH to explore spatio-temporal patterns

In order to characterise spatio-temporal mRNA profiles, we performed single-molecule RNA-FISH (smFISH) on liver sections at different times of the day. Each microscopy image contains a central vein (CV) and a portal vein (PV) that were manually detected based on the presence or absence of the bile ducts. For some smFISH experiments, we also used an immunostaining against Glutamine Synthetase, a well-characterised marker of the central vein. The Euclidean distance between the CV and PV was calculated, and each detected mRNA dot was assigned either the pericentral zone (PC), midlobular zone (Mid), or periportal zone (PP) based on its distance from the closest vein (Fig.3.1). Because the smFISH protocol was not compatible with immunostaining of the hepatocyte membrane, we could not count the exact number of RNA molecules per cell. Instead, we quantified a “density per nucleus”: we divided the total number of detected dots by the number of segmented nuclei. To assess the rhythmicity of mRNA profiles in the three zones, we fitted the mRNA counts with a harmonic regression within a generalised linear model (GLM) framework. We modelled mRNA counts with a negative binomial distribution where the mean of the distribution is also a function of the number of nuclei (Fig.3.1.D). The rhythmicity parameters a and b and the mean level m can be shared between PC, PP and Mid (same phase and amplitude), or can be independent. The most parsimonious model is selected based on the Bayesian Information Criterion (BIC). The dispersion parameter θ was assumed to be shared across all zones and time points, and was also estimated as part of the GLM framework.

We subsequently compared the smFISH profiles to the spatial profiles reconstructed from single-cell RNA-seq (scRNA-seq). The scRNA-seq dataset was provided by collaborators from the Itzkovitz lab (Weizmann Institute of Science, Israel). In 2017, the group of Itzkovitz reconstructed the spatial mRNA expression profiles along the porto-central axis of thousands of genes in the liver ([160], see 1.3.1). They performed single-cell RNA-sequencing of dissociated hepatocytes, and compared the expression level of 6 landmark genes whose spatial profiles were previously characterised by smFISH. They could thus infer the position of each hepatocyte within the layers of the liver lobule, and revealed that ~50% of genes have a spatially non-uniform pattern. They further extended this work by performing the same experiment at four times of the day (ZT0, ZT6, ZT12 and ZT18). The reconstructed spatio-temporal patterns were fitted by a mixed-effect linear model developed by C.Droin, co-first author of [168]. Briefly, the static spatial profile is described by a polynomial up to degree 2, while temporal profile is represented by a sine and cosine function (harmonic regression). Additionally, interactions between time and space are described as space-dependent oscillatory functions. Then, the BIC is used to select the most parsimonious model for each gene. RNA profiles are classified in 5 groups based on the retained parameters. The simplest model is the flat (F) model, when only the intercept is retained. Purely rhythmic genes (R) have only the rhythmic space-independent parameters. Purely zonated genes (Z) have only the zonation parameters (mean and slope). If the model comprises the rhythmic and the zonation parameters but not the interaction term, the gene is classified as "Z+R", where zonation and time independently modulate the spatio-temporal pattern. The most complex model comprises all the parameters with a time-space interaction term and is referred to as "ZxR", where the spatial profile varies with time (or the phase and amplitude vary depending on the layer). Dually regulated genes represented 7% of the expressed genes, and mostly consist in Z + R genes, with only a minority of Z x R patterns.



C Spatio-temporal profiles: Single-cell RNA-seq

$$y_{x,t,i} = \mu_i + \mu(x) + a(x)\cos(\omega t) + b(x)\sin(\omega t) + \varepsilon_{x,t,i}$$

$$\begin{cases} \mu(x) = \mu_0 + \mu_1 P_1(x) + \mu_2 P_2(x) & \text{| static zonation} \\ a(x) = a_0 + a_1 P_1(x) + a_2 P_2(x) & \text{| zone-dependent} \\ b(x) = b_0 + b_1 P_1(x) + b_2 P_2(x) & \text{| rhythmicity} \end{cases}$$

D Spatio-temporal profiles: smFISH

$$y_{z,t} \sim NB(\mu_{z,t}, \theta)$$

$$log_2(\mu_{z,t}) = m_z + a_z cos(\omega t) + b_z sin(\omega t) + log_2(N_z)$$

Figure 3.1 – Characterisation of spatio-temporal gene expression profiles based on smFISH. A: smFISH targeting *Agxt* at ZT4. Nuclei are stained in blue (DAPI staining). Scale bar: 100μm. B: detected RNA transcripts appear as grey dots. Those assigned to the PC, Mid of PP zones are colored in red, purple and blue, respectively. C: Extended harmonic regression model for spatio-temporal expression profiles (scRNA-seq). y_t is the log-transformed gene expression, x denotes the layer index, t denotes time, and i denotes the biological replicates. The model includes a static but zonated layer-dependent mean $\mu(x)$, as well as layer-dependent harmonic coefficients $a(x)$ and $b(x)$. All layer-dependent coefficients are modelled as second order polynomials. Temporal dependency is modelled with 24-h periodic harmonic functions. μ_i are random effects needed since the data is longitudinal in space (8 layers measured in each animal). D: Harmonic regression model for spatio-temporal expression profiles (smFISH). y_t is the number of mRNA per zone and per time-point, which follows a negative Binomial distribution. $\mu_{z,t}$ is mRNA counts in zone z, m_z is the average mRNA level in zone z, a_z and b_z the coefficient of the cosine and sine function, N_z the number of nuclei in the zone z and is used as an offset of the model, ω the frequency ($\frac{2*pi}{24}$). a_z, b_z and m_z can be shared between the three zones, or can be independent.

To validate the spatio-temporal profiles predicted from single-cell RNA-seq data, we performed smFISH on representative liver genes. We first targeted the liver-specific gene *Agxt* (Alanine-Glyoxylate And Serine-Pyruvate Aminotransferase), which codes for an enzyme that converts glyoxylate and L-alanine to glycine. We selected this gene because it is rhythmic at the whole tissue level [60], and because it was previously shown to be enriched in the periportal zone [160]. We performed smFISH on liver tissue every 4 hours and quantified the spatio-temporal pattern as explained above (Fig.3.2.A). The spatio-temporal pattern is best fitted with the model that shares the same rhythmic parameters for all three zones, with a zone-specific mean expression level. *Agxt* is maximally expressed in all three zones at ZT6.6, with a $\log_2$FC of 1.1. It is more expressed around the periportal zone, and decreases around the pericentral vein. Thus, according to smFISH data, the pattern of *Agxt* corresponds to a “Z+R” model: the zonated profile oscillates in time while keeping its slope (no zone-dependent rhythmicity). The mean expression level in the PP zone is very close to the mean of the Mid zone, as seen by the “plateau” on the spatial profile (Fig.3.2). Of note, the first and the second best models have a very close BIC value. The second best model is the one with the same mean, phase and amplitude for the PP and Mid zone (phase: ZT 6.6 and amplitude: 1.2), while PC has its own independent parameters (phase: ZT 7.0, amplitude = 0.78). By scRNA-seq, *Agxt* was also classified as a "Z+R" pattern, peaking at ZT5. Additionally, the spatial profile is described by a polynomial of degree 2, therefore, we observe the same "curved" profile in both smFISH and scRNA-seq. Thus, the reconstructed scRNA-seq and smFISH profiles were consistent.

We also quantified the spatio-temporal profiles of *Actb* (Fig.3.3.A). *Actb* is maximally expressed at ZT1, with the same mean, amplitude and phase in all three zones, suggesting that it is a purely rhythmic gene (R) (Fig.3.3.B). The second best fitted model is the one with the same rhythmic parameters, but with a lower mean expression level in PP zone compared to the PC and Mid zone, suggesting that *Actb* could be slightly zonated. By scRNA-seq, *Actb* was classified as a purely zonated gene (Z). Even if ZT0 appears to be higher than ZT12 in the raw data (Fig. 3.3.D, shaded line), a model without rhythmic parameters was preferred. The low temporal resolution of the scRNA-seq dataset (sampled every 6 hours) could explain why *Actb* oscillations were not detected by scRNA-seq. Moreover, the expression at the PC is predicted to be only 25% higher than in the PP zone. As a comparison, *Agxt* increases by 42% between the PC and PP zones. Thus, such small slopes might be difficult to be accurately detected by smFISH.

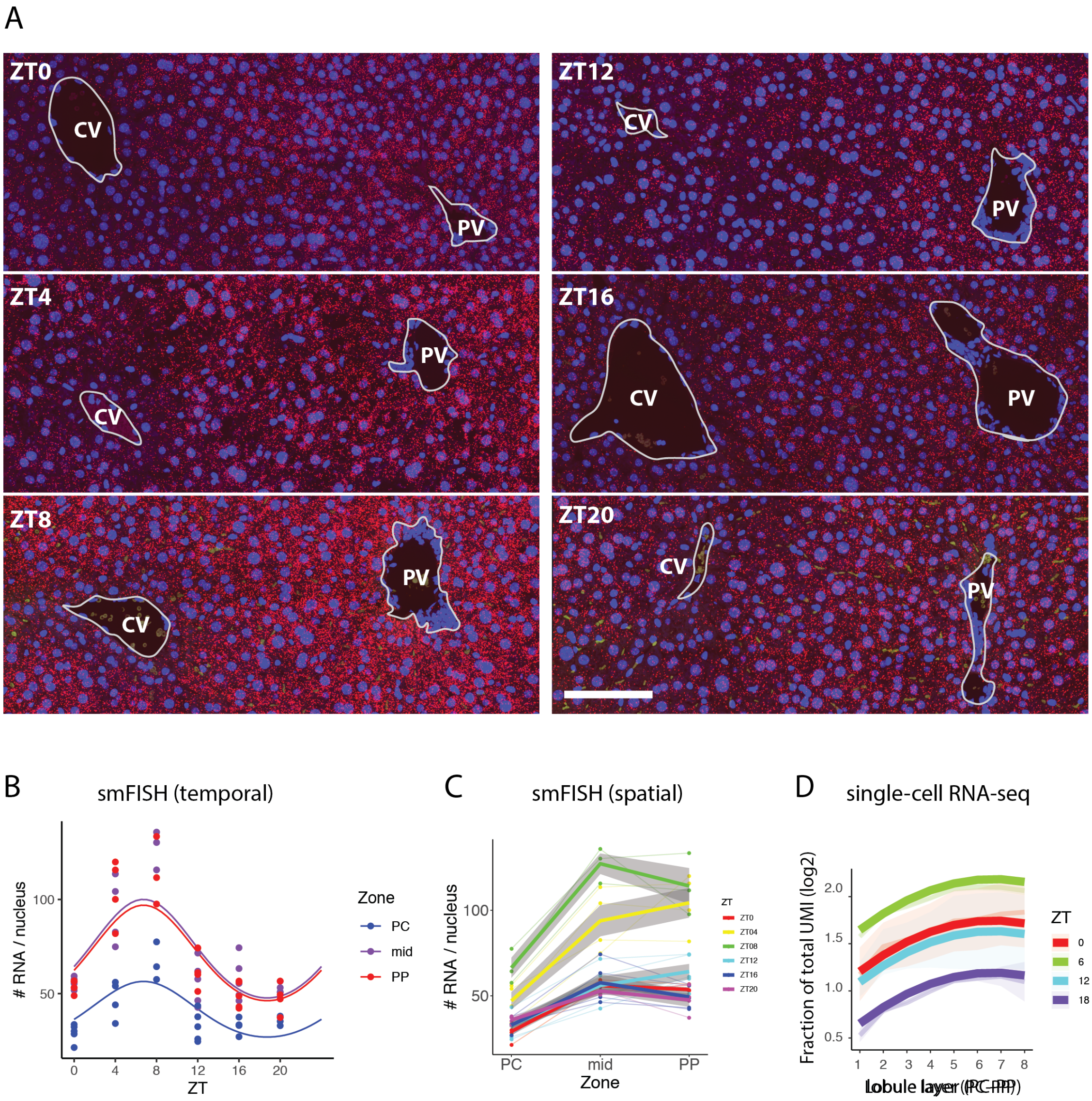


Figure 3.2 – Spatio-temporal profile of *Agxt*. A: smFISH of *Agxt* on liver FFPE tissue (red dots). Representative images at ZT0, 4, 8, 12, 16 and 20. Maximal projection of all z-stacks. CV = central vein, PV = portal vein. Blue: nuclei stained with DAPI. scale bar: 100μm B: Quantification of mRNA transcripts on smFISH images. A harmonic generalised linear model assuming a negative binomial distribution is fitted to the number of mRNA in each three zones and divided by the total number of nuclei per image (see 6.4.1). Here, the best model includes a zone-specific mean, and shared rhythmic parameters (same amplitude and phase for the three zones). Only one mouse per time point were used(n = 6). 5-6 images were taken per animal (technical replicates) and the average number of mRNA molecules per number of nuclei per image is represented by a datapoint. Quantification was done on a total of 4174 nuclei for PC zone, 5064 nuclei for PP zone, and 1357 for the Mid zone. C: same quantification as in B, but represented as a spatial profile. D: Reconstructed spatio-temporal profile based on scRNA-seq data. X-axis: lobule layers from PC to PP. Y-axis: $\log_2$(Expression) expressed as the fraction of total UMI per cell. *Agxt* is classified as "Z+R".

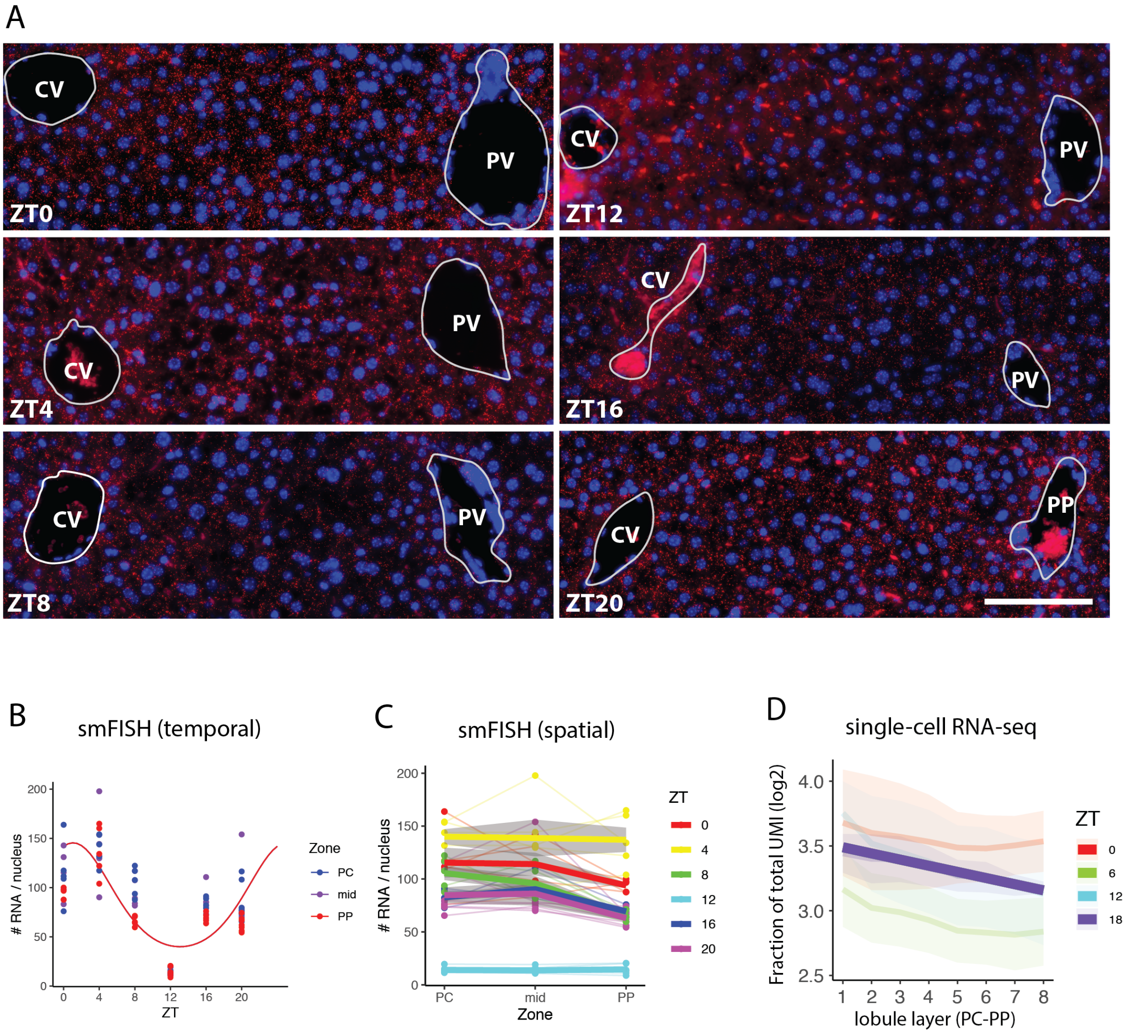


Figure 3.3 – Spatio-temporal profile of *Actb*. A: smFISH of *Actb* on liver FFPE tissue (red dots). Representative images at ZT0, 4, 8, 12, 16 and 20. Maximal projection of all z-stacks. CV = central vein, PV = portal vein. Blue: nuclei stained with DAPI. scale bar: 100μm. Large bright oval shapes are erythrocytes. B: Quantification of mRNA transcripts on smFISH images. A harmonic generalized linear model assuming a negative binomial distribution is fitted to the number of mRNA in each three zones and divided by the total number of nuclei per image (see 6.4.1). Here, the best model includes shared rhythmic parameters (same amplitude and phase for the three zones). Only one mouse per time point were used (n = 6). 5-6 images were taken per animal (technical replicates) and the average number of mRNA molecules per number of nuclei per image is represented by a datapoint. Quantification was done on a total of 6184 nuclei for PC zone, 9038 nuclei for PP zone and 2616 for the Mid zone. C: same quantification as in B, but represented as a spatial profile. D: Reconstructed spatio-temporal profile based on scRNA-seq data. X-axis: lobule layers from PC to PP. Y-axis: log$_2$(Expression) expressed as the fraction of total UMI per cell. *Actb* is classified as "Z".

3.2.2 Clock genes are uniformly expressed within the liver lobule

Given that many of the zonated genes are also transcribed in a circadian manner, we assessed whether core clock genes were sensitive to spatial regulation and were also differentially expressed within the lobule. We selected *Per1*, a core clock gene involved in repressing the transcriptional activity of BMAL1/CLOCK, which peaks in the early night. *Per1* was previously shown to be slightly enriched in the periportal area [160]. We also targeted *Bmal1* (also known as *Arntl*), which peaks at the end of the night time. *Arntl* was classified as purely rhythmic (R) by both scRNA-seq and smFISH, peaking at ZT23 in smFISH and at ZT21.5 in scRNA-seq (Fig.3.5.A-D). *Per1* was also classified as R despite the slight periportal enrichment observed in the raw data (Fig. 3.5.A, E-G). By smFISH, the best model was also the one with the same mean, amplitude and phase for all three zones (Fig.3.5.E, black line). The second best model, where all three zones have different means but same rhythmic parameters, has an almost equivalent BIC value (1946.732 versus 1946.737). In this model, *Per1* is more expressed in the periportal zone, consistent with the raw data of scRNA-seq. We also verified the profiles of two core clock genes *Cry1* and *Rorγ*, and members of the PARbZip family (the activators *Tef* and *Hlf* and the repressor *Nfil3*) by smFISH. Consistent with their predicted models by scRNA-seq, no zonation patterns were detected for these genes (data not shown). Thus, the temporal profiles of clock genes are consistently detected by both smFISH and scRNA-seq, but the classification of profiles with a small slope is very delicate, and these genes can often be attributed to different models with comparable scores. The profiles of other reference core clock genes were also assigned to the rhythmic category (Fig.3.4). This suggests that the circadian clock is largely non-zonated, as confirmed by smFISH. Therefore, the circadian clock is robust to the heterogeneous hepatic microenvironment.

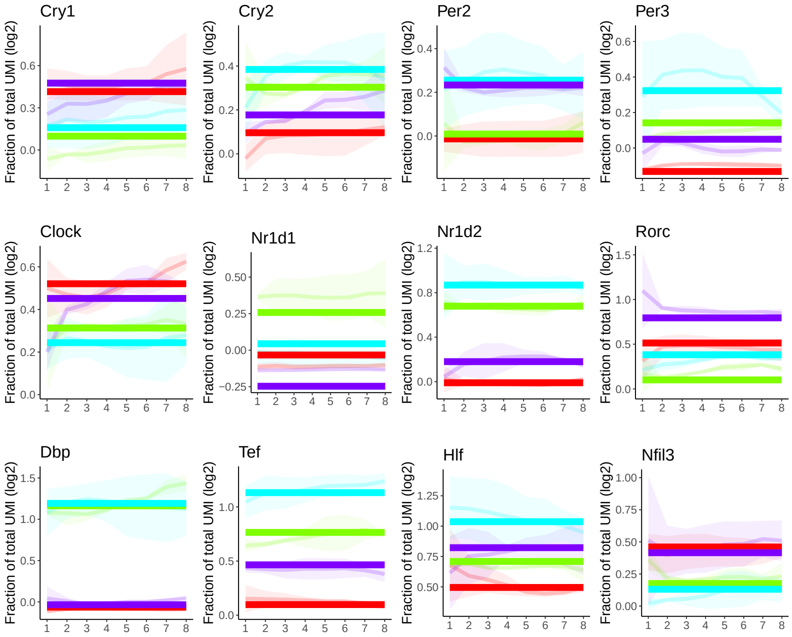


Figure 3.4 – Spatial and temporal profiles and fits of circadian core-clock genes from scRNA-seq data. Layer 1 is the most pericentral layer, layer 8 is the most periportal layer. Shaded line: raw data with corresponding standard error, in bold line: fitted data. All circadian clock genes have been assigned to the purely rhythmic model (R).

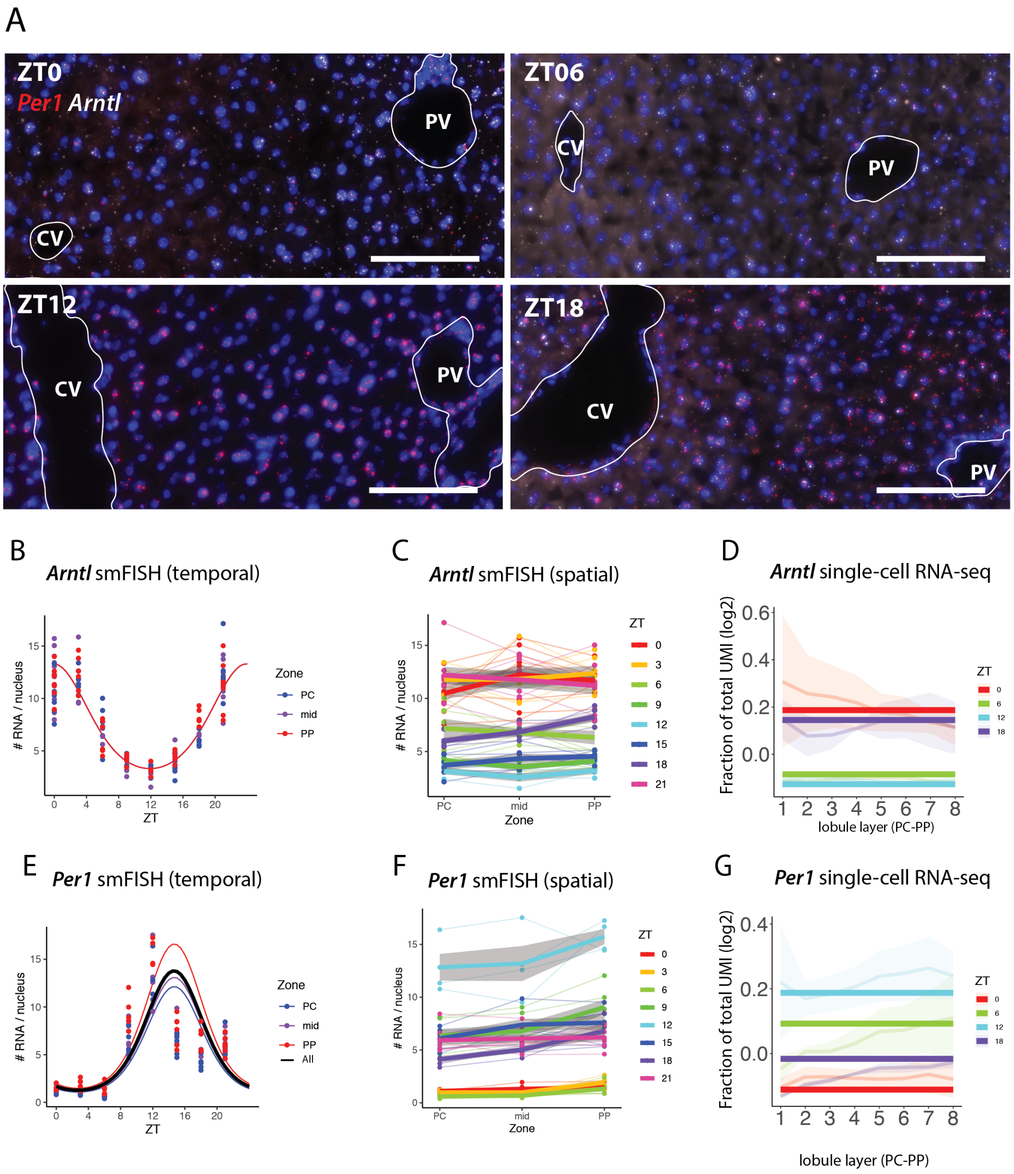


Figure 3.5 – Spatio-temporal profiles of *Arntl* and *Per1*. A: smFISH of *Arntl* in white and *Per1* in red on liver cryosections. Livers were sampled every 3 hours, and were not from the same animals as for *Agxt* and *Actb*. Representative images at ZT0, 6, 12 and 18. Maximal projection of all z-stacks. CV = central vein, PV = portal vein. Blue: nuclei stained with DAPI. scale bar: 100μm. B: Quantification of mRNA transcripts on smFISH images. A harmonic generalized linear model assuming a negative binomial distribution is fitted to the number of mRNA in each three zones and divided by the total number of nuclei per image. Here, the best model includes shared rhythmic parameters (same amplitude and phase for the three zones). 5-8 images were taken per animal (technical replicates) and the average number of mRNA molecules per number of nuclei per image is represented by a datapoint. Quantification was done on a total of 6400 nuclei for PC zone, 6179 nuclei for PP zone and 2419 for the Mid zone. C: same quantification as in B, but represented as a spatial profile. D: Reconstructed spatio-temporal profiles based on scRNA-seq data. X-axis: lobule layers from PC to PP. Y-axis: $\log_2$(Expression) expressed as the fraction of total UMI per cell. *Arntl* is classified as "R". E: Quantification of *Per1*. The black line shows the fit for the model where all the parameters are shared between the three zones. Additionally, the red, purple and blue line represent the fit by the model where the amplitude and phase are common, but the mean is different. This corresponds to a "Z+R" model. Note that the phase fitted by the model (ZT14) is not optimal by visual inspection: this is because the temporal profile does not follow a cosinus. A different shape (function) may better fit the measurements. F: Spatial representation of *Per1* quantified by smFISH. G: Reconstructed spatio-temporal profile based on scRNA-seq data. *Per1* is classified as R.

3.3 Concluding remarks

smRNA-FISH is a sensitive, albeit low-throughput method to accurately detect, localise, and quantify individual RNA molecules [190]. Thus, in addition to generating appealing and colorful images, smFISH can be used to quantify gene expression profiles in time and space. We used smFISH as an orthogonal approach to validate spatio-temporal profiles predicted from scRNA-seq data, from purely rhythmic to dually regulated Z+R profiles. Overall, the profiles were consistent between smFISH and scRNA-seq, but we also observed some limitations, particularly when two competing models result in close BIC values. These discrepancies most likely reflect uncertainties in the spatial analysis of smFISH in tissues. We also used a less refined spatial resolution, defining only three zones instead of eight layers as in [160]. Moreover, the low temporal resolution of the scRNA-seq dataset decreases its power to detect rhythmic patterns. Additionally, mice used for the smFISH experiments and the scRNA-seq were under different feeding regimens, which could affect their physiological status [60] (night-restricted feeding for the smFISH, *ad libitum* for the scRNA-seq). But for clock genes as for other genes expressed in the liver, the predicted spatio-temporal profiles were noticeably similar using both approaches.

By combining smFISH on intact liver tissue and zonation profiles from scRNA-seq, we showed that while core clock genes are temporally oscillating, their spatial patterns are homogeneous between the liver zones. This suggests that the clock is robust to zonated signals in the different lobular zones. Thus, the spatial patterns of rhythmic metabolic functions (carbohydrate metabolism, xenobiotic metabolism, lipid metabolism, etc.) are regulated by independent factors.

3.4 Space-time logic of liver gene expression at sublobular scale: Abstract

This work was published in Nature Metabolism by the following authors (first authors in bold): **C. Droin, J. El Kholtei, K. Bahar Halpern**, C.Hurni, M. Rozenberg, S. Muvkadi, S. Itzkovitz, F. Naef, [168].

The mammalian liver is a central hub for systemic metabolic homeostasis. Liver tissue is spatially structured, with hepatocytes operating in repeating lobules, and sub-lobule zones performing distinct functions. The liver is also subject to extensive temporal regulation, orchestrated by the interplay of the circadian clock, systemic signals and feeding rhythms. However, liver zonation has previously been analysed as a static phenomenon, and liver chronobiology has been analysed at tissue-level resolution. Here, we use single-cell RNA-seq to investigate the interplay between gene regulation in space and time. Using mixed-effect models of messenger RNA expression and smFISH validations, we find that many genes in the liver are both zonated and rhythmic, and most of them show multiplicative space-time effects. Such dually regulated genes cover not only key hepatic functions such as lipid, carbohydrate and amino acid metabolism, but also previously unassociated processes involving protein chaperones. Our data also suggest that rhythmic and localized expression of Wnt targets could be explained by rhythmically expressed Wnt ligands from non-parenchymal cells near the central vein. Core circadian clock genes are expressed in a non-zonated manner, indicating that the liver clock is robust to zonation. Together, our scRNA-seq analysis reveals how liver function is compartmentalized spatio-temporally at the sub-lobular scale.

4 Comprehensive analysis of the circadian hepatic transcriptome at subcellular scale

4.1 Background

Every step of the RNA life cycle is tightly regulated to ensure proper cellular function. Gene expression begins with transcription, where transcription factors and co-activators bind to DNA gene regulatory elements and recruit RNA Polymerase II (PolII) to form the RNA transcription complex, and synthesise a pre-mRNA copy of the DNA template. While in the nucleus, the RNA transcript is subject to multiple co- and post-transcriptional modifications, including splicing, 5' capping, 3' end processing, polyadenylation, methylation, assembly into ribonucleoprotein (RNP) complexe, before being exported through the nuclear pores [191]. After reaching the cytoplasm, the RNP interacts with translation initiation factors to start the production of a functional protein. The mRNA lifetime in the cytoplasm is determined by the activity of a distinct set of RNA Binding Proteins (RBPs), enzymes, and functional RNAs, which promotes stability, or conversely degradation through deadenylation, endonucleolytic cleavage [152], and silence their translation (miRNA) [192].

Collectively, the balance of production and decay rates determines the abundance of RNA in a given form (nascent, pre-mRNA, mature mRNA) in a given subcellular compartment (chromatin-bound, nucleus, cytoplasm, organelles) [193]. The localisation of the RNA transcript also depends on its function. For example, many non-coding RNAs such as long non-coding RNAs (lncRNAs) are found in the nucleus [194], where they remodel chromatin architecture (*Xist*, *Firre*) or act as structural scaffold for nuclear bodies (*Malat1*, *Neat1*). On the other hand, protein coding transcripts, which are meant to be translated, are exported into the cytoplasm [114]. Interestingly, a study in the mouse liver showed that a significant proportion of protein coding transcripts were enriched in the nucleus (13%), including *Mlxipl* and *Nlrp6*, suggesting that the nuclear to cytoplasmic export rate is also a regulatory node for protein coding genes [74].

The kinetic rates regulating the RNA processing steps are not constant, but adapt to environmental stimuli such as the presence of hormones, temperature shifts [195], nutrient levels [196], inflammation [95], or time of the day [78, 79]. In principle, every step during the RNA life-cycle could be regulated in a temporally rhythmic manner, contributing to circadian gene expression. Rhythmic transcription is the starting point of rhythmic gene expression, and has been extensively studied in the circadian context, particularly in the mouse liver [47, 59, 60]. The liver has the highest number of cycling mRNA among all the tissues [28], and the rhythmic patterns are driven both by the endogenous circadian clock, and by rhythmic (un)availability of food [30, 197]. Interestingly, rhythmic synthesis alone does not account for the oscillations of all RNAs, implying a rhythmic regulation at the post-transcriptional level. The extent to which post-transcriptional regulation contribute to circadian gene expression is still debated, varying from ~15% [60, 54], up to 70% [47]. Post-translational regulation also plays a role in generating rhythmic profiles, as 50% of the cycling proteins did not have matching cycling mRNA [61, 45]. Conversely, many rhythmically transcribed genes do not have corresponding cycling mRNA [47] or protein [63], suggesting that amplitudes are dampened as rhythm propagates. There is therefore more and more evidence underlining the importance of post-transcriptional regulation in driving mRNA rhythmicity. Mathematical models showed that temporal regulation of mRNA stability (i.e. degradation rate) offer the flexibility to boost amplitudes and fine-tune phases, [54, 79, 78, 158]. Rhythmic mRNA stability can be modulated through the activity of plethora of RNA-binding proteins [198, 156], through silencing by miRNA [199], or by regulating poly(A) tail length [154] and regulating deadenylase activity [200]. In the nucleus, processing rates, including export rates, can be modulated by m^{6}A methylation [147], or through rhythmic retention in nuclear paraspeckles [143].

In this study, we aim at understanding the circadian dynamics of RNA regulatory programs in the mouse liver. More specifically, we focus on nuclear export, cytoplasmic degradation, and to some extent on splicing. We performed RNA-sequencing of polyadenylated and total RNA from nuclear and cytoplasmic fractions of mouse liver samples collected every 4 hours along a full daily cycle. This dataset provides a genome-wide and temporal inventory of RNA subcellular localisation. We used these measurements to feed our simplified model describing the RNA processing pathway: 1) in the nucleus, pre-mRNA is co- and post-transcriptionally spliced to become a mature nuclear mRNA and 2) after spending some time in the nucleus, the mRNA is transported to the cytoplasm where it is degraded.

As a first step of initial analysis, we explore the differences between RNA populations without the temporal dimension. We analyse the relationships between nuclear pre-mRNA and nuclear mRNA, and between nuclear mRNA and cytoplasmic mRNA, which, at steady-state, reflect the ratio of the production rate and the decay rate. This first round of analysis revealed distinct signatures of export, cytoplasmic degradation, and splicing, affecting specific classes of RNA. Notably, protein coding transcripts are more abundant in the cytoplasm than in the nucleus, however, many nuclear-enriched transcripts code for nuclear proteins associated with gene expression regulation (DNA modification, RNA processing and export), suggesting a concordance of RNA and protein localisation. Moreover, nuclear-enriched transcripts are shorter on average, while longer transcripts are more abundant in the cytoplasm. Transcript length also influences the extent of co- versus post-transcriptional splicing: short nuclear polyadenylated transcripts contain on average more unspliced introns, while longer transcripts are more co-transcriptionally spliced. Next, we estimate the nuclear export rates and the cytoplasmic degradation rates using a mathematical model developed by Wang et al.[201]. By comparing the rhythmic profiles (amplitudes and phases) of nuclear pre-mRNA, nuclear mRNA, and cytoplasmic mRNA, we estimate the export rates and cytoplasmic degradation rates of ~1400 genes. The median half-life of rhythmic transcripts in the cytoplasm is ~2.5h, while the median nuclear lifetime is shorter than 30 minutes. The mathematical model is able to detect rhythmic degradation processes. This analysis uncovered rhythmic patterns in the nucleus that most likely originate from a rhythmic regulation of the nuclear export rate. This post-transcriptional regulatory step contributes to the increase of the amplitude of highly rhythmic genes such as *Dbp*, *Pck1*, or *Lpin1*, but also to the generation of *de novo* rhythms, and affects the temporal profiles of ~10% of the rhythmic nuclear transcriptome. Finally, by combining the estimated export rates and cytoplasmic degradation rates, we reveal that for a majority of genes, the nuclear mRNA are processed within minutes, and the total transcript lifetime is mainly determined by the cytoplasmic mRNA stability. However, some metabolic genes remain in the nucleus for a few hours (up to 4 -5 hours), delaying the peak phase in the cytoplasm. Therefore, we suggest that regulation of the nuclear export rate is a post-transcriptional tool to regulate the timing of cytoplasmic mRNA accumulation by generating *de novo* rhythm or by tuning the phase and amplitude of the nuclear mRNA.

4.2 Experimental design

In order to characterise the kinetics of RNA regulatory processes occurring in different cellular compartments, we isolated different RNA populations in the mouse liver. We sampled livers from individual C57BL/6j and Cry1/Cry2 double-KO mice (CryKO) housed in a light-dark cycle and under a restricted feeding regime every four hours (n = 2 per time point) (Fig.4.1.A). We fractionated liver cells from intact tissue in sucrose gradient and isolated nuclear (Nuc or N) and cytoplasmic (Cyt or C) RNA. We sequenced ribo-depleted total RNA (T), and polyadenylated RNA (A). In addition, we also sequenced total RNA from the unfractionated liver tissue (U or Unf) (Fig.4.1.A). Together, we have three RNA populations (N, C, U), and five types of RNA: Nuclear Total RNA “NT", Nuclear Polyadenylated RNA “NA”, Cytoplasmic Total RNA “CT”, Cytoplasmic Polyadenylated RNA “CA", and Unfractionated Total RNA “UT". We further quantified pre-mRNA (intron “I”) and mRNA (exon “E”) (see Methods: 4.3).

The RNA processing program is extremely rich and complex. We simplified the events occurring along the RNA lifecycle and include the following steps in our model (Fig.4.1.B, C): we assume that the pre-mRNA (p), which is transcribed at a rate T, is spliced and polyadenylated at a rate of s to produce mature nuclear RNA (m). Then, this transcript is exported at a rate e into the cytoplasm (M), where it is finally degraded at a rate γ (Fig.4.1.A). We use NTI (nuclear total pre-mRNA) to approximate the level of pre-mRNA (p). NAE (nuclear polyadenylated mRNA) and CAE (cytoplasmic polyadenylated mRNA) are used to describe the accumulation of mature RNA in the nucleus (m) and in the cytoplasm (M). We focus particularly on nuclear export rate and cytoplasmic degradation rate. We will often refer to these rates as cytoplasmic half-life ($\frac{log(2)}{\gamma}$), and as nuclear export time or nuclear retention time ($\frac{log(2)}{e}$), with the multiplicative factor of log(2). The sequential transformation of the RNA transcript can be written as a system of simple ordinary differential equations (ODEs) (Fig.4.1.D). At steady, the equations are set to 0 and each species is described by the ratio of its production and decay term.

In our study, CT (cytoplasmic total) is redundant with CA (cytoplasmic poly(A), therefore, it is not included in the model, but we used it as an internal control. RNA from the unfractionated samples (Unf) are also not used in the model, but serve as a reference dataset.

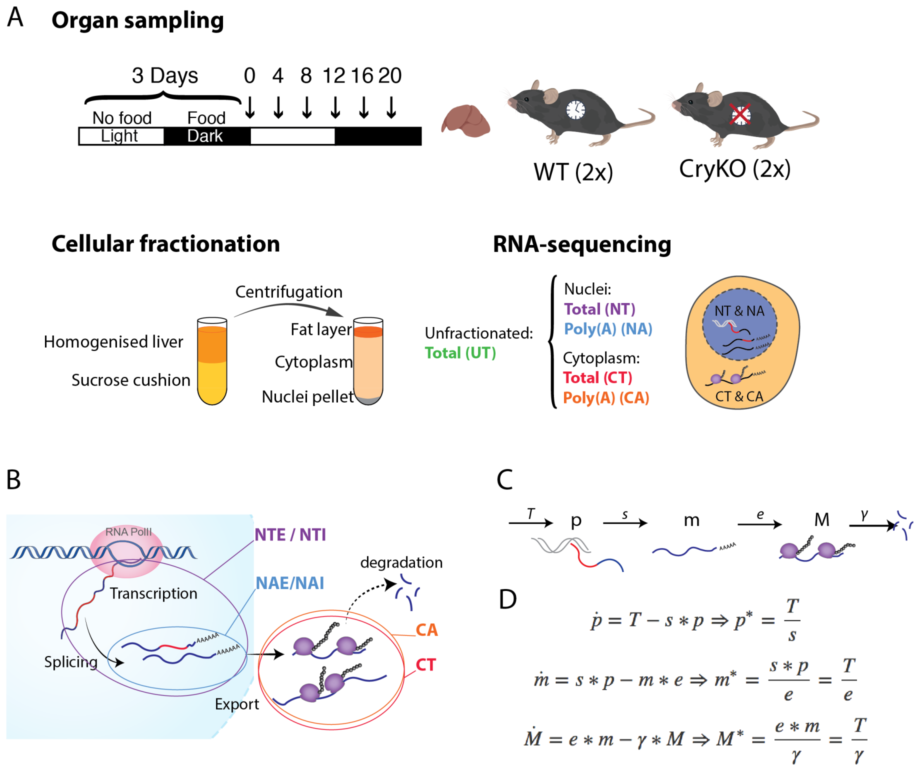


Figure 4.1 – Kinetic model describing the RNA processing steps from the nucleus to the cytoplasm. A: Scheme of the experimental design: 12 WT and 12 CryKO mice were kept in a 12:12 light-dark cycle. 3 days prior to the experiment, food was given only during the dark phase (active phase). Livers were collected every 4 hours (n = 2), homogenised, and centrifuged in sucrose gradient. RNA from cytoplasmic fraction (supernatant in sucrose) and nuclei (pellet) were extracted. Ribo-depleted RNA (T) or Poly(A) pull-down RNA (A) from nuclear and cytoplasmic fractions were sequenced. RNA from unfractionated liver tissues was also sequenced (ribo-depletion). B: Scheme representing the journey of an RNA transcript starting with the transcription by RNA Polymerase II, splicing of the introns (red) and addition of a poly(A) tail at the 3' end, export to the cytoplasm, and degradation. C: Linear representation of each RNA processing step modeled in this study. D: A system of first-order linear equations describes the transformation of pre-mRNA to nuclear mRNA to cytoplasmic mRNA: T = Transcription rate [transcript × min^{-1}], s= splicing and polyadenylation rate [min^{-1}], e = export rate [min^{-1}], γ = degradation rate [min^{-1}]. p = pre-mRNA, m = nuclear mRNA, M = cytoplasmic mRNA. The dot indicates the approximation at steady-state.

4.3 Transcript-based RNA-seq quantification

We quantified both the expression at the gene level and at the transcript level. Indeed, genes often have several expressed isoforms that are not processed the same way [114]: the final destination of a protein coding gene is the cytoplasm where it is translated, while a target of nonsense-mediated decay is not expected to be long-lived. Thus, in this context, it could be biologically more relevant to study individual isoforms and take into account the RNA biotype. We additionally distinguished pre-mRNA from mRNA. A commonly used procedure when analysing introns and exons from total RNA-seq is the "union exon model", in which all the isoforms of a given gene are merged [202]. A stringent annotation is then used: if the genomic region is defined as "intron" in *every* isoform, then it is considered as an intronic region. Otherwise, it is considered an "exon". While this method is easier to implement [202], it creates some artifacts. Many intronic regions are masked, with the following consequence: when introns are efficiently spliced before being exported and translated, the density of reads mapping on an intronic region is lower than on an exonic region, especially in the cytoplasm, and to a lesser extent in the unfractionated sample. However, an intron might be annotated as “exon” because of one specific isoform, which may not even be expressed in the studied cell type. When counts are normalised by the gene length (RPKM, FPKM, TPM), the exon length is artificially longer, resulting in a “dilution” of the reads. Thus, the intron/exon ratio is biased such that splicing seems to occur less frequently (Fig.4.2 B and C, example with *Eif1*). One solution is to create a genome annotation specific to the liver transcriptome, by merging only the isoforms expressed above a certain threshold. Even if this method corrects the annotation of some genes, it does not solve all the cases, especially when there are still several isoforms expressed (*Eif1*: 3 out of 5 isoforms are expressed in the liver). This is particularly the case when comparing two very different RNA populations. For instance, Retained Introns are mainly present in the nucleus [107, 110] and are annotated as “exon”. In the cytoplasm, however, the retained intron isoform may not be present and the corresponding genomic region is a spliced intron. In the union-exon model, the exon length will be again too long, and the resulting normalised count will be underestimated. Therefore, we used the pseudo-alignment algorithm implemented in Kallisto for RNA-seq quantification in order to quantify the expression level per transcript rather than per gene [203]. We provided as a reference the annotations of both mRNA and pre-mRNA (Fig.4.2.A). Technically, Kallisto’s algorithm does not discriminate intronic from exonic regions, but only a “pre-mRNA transcript” from a “mRNA transcript”. However, for sake of clarity, we name mature RNA “E” as for exon, and pre-mRNA “I” as for intron. We acknowledge that a transcript-based approach is intrinsically more difficult and generates more noise, especially when we include both premature and mature transcripts, because different isoforms often have a high proportion of genomic overlap. Moreover, a nascent RNA may temporarily have the same structure as a specific isoform because of the yet incomplete splicing. Moreover, when the algorithm cannot discriminate between similar isoforms, reads are distributed equally and are “diluted”.

Genes with a high number of annotated isoforms tend to express more isoforms, and do not follow a minimalistic strategy [194] (Fig. 4.2.C). On average, 2-3 isoforms are expressed per gene, but some outliers have up to 16 expressed isoforms (e.g. *Aopep* 16 isoforms, *Ncor1* 15 isoforms, Fig.4.2.D). The dominant isoform usually captures at least 25% of the total gene expression (Fig.4.2.E). We only selected isoforms that were contributing to at least 1/n to the total gene expression (n = number of annotated isoforms), but with an additional minimal threshold of 0.2. To buffer noise associated with the high number of isoforms, we summed the counts estimated by Kallisto if several isoforms belonged to the same RNA biotype (based Ensembl classification). We then normalised counts by the average transcript length, weighted by the relative expression of each isoform in each RNA population (Nuc PolyA, Nuc Total, Cyt PolyA, Cyt Total, Unf Total). The normalisation by the transcript length was performed using DESeq2 [204]. To reduce the number of biotypes, we grouped the biologically related biotypes into 8 supergroups (see Methods 6.5.2). Protein coding (PC) transcripts contain an open reading frame (ORF). Retained Intron (RI) is an alternatively spliced isoform that keeps an intron in its final form [110]. Nonsense-mediated decay transcript (NMD) contains a premature stop codon and is likely to be targeted to degradation [152]. Long non-coding RNA (lncRNA) is a non-coding transcript longer than 200bp. Processed Transcripts (PT) is a broad class that contains transcripts without an open reading frame, including both long (> 200nt) and short non-coding RNA. Therefore, some well-known long non-coding RNA such as *Firre* are annotated as PT. Small nuclear RNAs (snRNA) are short non-coding transcripts involved in the processing of pre-mRNA. Small nucleolar RNA (snoRNA) are small non-coding RNAs, in charge of ribosomal RNA modifications (2'-O-methylation modifications and pseudouridylation [205]). Here, "snoRNA" also contain a subclass of snoRNA: the small Cajal Body-specific RNA (scaRNA), which are “guides RNA” in charge of post-transcriptional modifications of the spliceosomal U RNA (snRNA). Pseudogenes, whether they are being processed or not, are non-coding transcripts that resemble functional genes. Finally, micro-RNAs (miRNA), are small non-coding RNA (~22bp) that repress the translation of the target mRNA. Of note, the quantification of small RNA (miRNA, snoRNA) are relatively unreliable due to their size [206]. Finally, we also performed the RNA-seq quantification at the gene level by summing all the counts assigned to the different isoforms of the gene, and by normalising the counts by the gene length (average transcript length weighted by the relative expression).

RNA-seq normalisation methods such as TPM or RPKM estimate the fraction of RNA species in the sample, but do not allow a comparison of the absolute transcript abundance across samples with different RNA composition, such as nuclear, cytoplasmic, or unfractionated RNA [207]. It only allows a *relative* comparison of concentrations. In order to adjust the ratio between nuclear and cytoplasmic RNA as well as possible, we took advantage of the dataset published by [196], where they converted the counts of RNA-seq dataset of nuclear Poly(A) and cytoplasmic Poly(A) RNA from mouse liver into number of molecules based on single-molecule RNA-FISH quantification (see Methods 6.5.3). By applying a simple linear regression, we rescaled the RPKM values of NAE, NAI, NTE, NTI, CAE, and CTE estimated by DESeq2 such that the ratio between nuclear and cytoplasmic RPKM is less biased, although still arbitrary.

In total, we have 10 measurements for each of the twelve mice (NTE, NTI, NAE, NAI, CTE, CTI, CAE, CAI, UTE, UTI), quantified with three different level of stratification: gene-level, biotype-level, and isoform-level (with 10'000 to 12'000 unique genes and ~35'000 isoforms). We will not use isoform-level quantification, as the diversity of isoforms makes the quantification noisier and adds an unwanted layer of complexity. Since the proportion of introns in the cytoplasm is low (Fig.4.3) and is not biologically relevant in our study, we also ignore CTI and CAI samples.

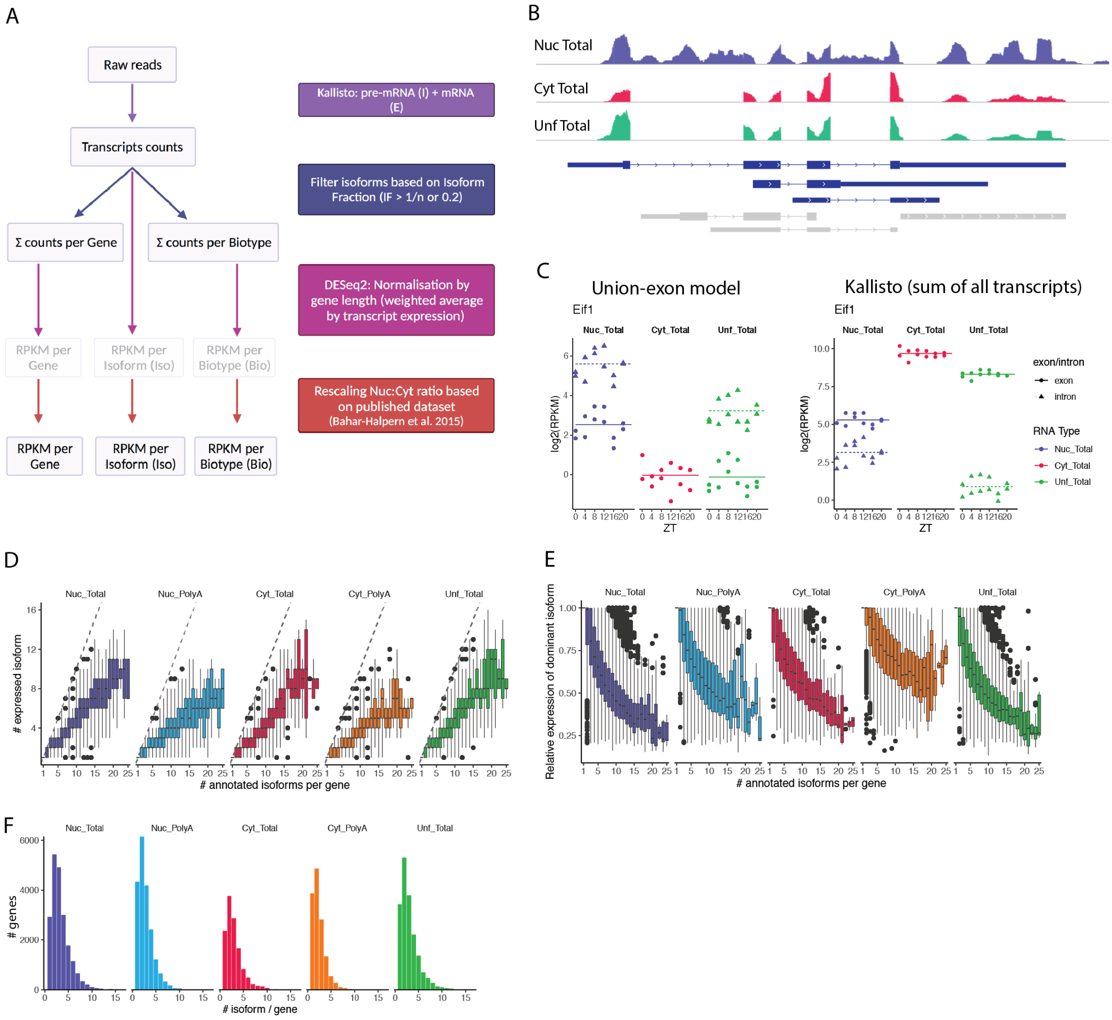


Figure 4.2 – A: Pipeline of RNA-seq quantification using Kallisto. Estimated counts per isoform are further summed by Gene or by Biotype, and normalised by DESeq2. Additional filtering steps based on expression level are done (see Methods 6.5.3). B: Reads from nuclear total, cytoplasmic total, and unfractionated total RNA population, aligned to the mm10 genome using RNASTAR, displayed on IGV browser. Each RNA population is auto-scaled. Below are the isoforms annotated by Ensembl for the gene *Eif1*. In blue: the three expressed isoforms: ENSMUST00000049385, ENSMUST00000132618, ENSMUST00000152521. In grey: the other isoforms not expressed according to Kallisto. C: Quantification of *Eif1* with the union-exon model and by Kallisto (sum of the three isoforms in order to have a quantification "per gene"). In the union-exon model, because of the overlapping isoforms, only a small portion of the gene is considered as “intron”. Thus, the exonic counts are artificially lowered, the pre-mRNA seems to be enriched in Unf Total. The isoform-specific quantification with Kallisto (right) corrects that artefact. D: Distribution of number of isoforms per gene after filtering (based on expression level and isoform fraction). E: Fraction of total gene expression of the dominant isoform, ordered by number of annotated isoforms per gene. F:Distribution of number of expressed isoforms per gene.

4.4 Characteristics and composition of nuclear and cytoplasmic transcriptomes in mouse liver

We first characterise the different RNA populations in mouse liver cells without the temporal dimension.

The nucleus is the cellular compartment where transcription and splicing occurs. The Nuclear Total RNA population (NT) includes all the transcripts found in the nucleus, from nascent transcripts to fully transcribed, ready to be exported mature RNAs. As expected, the proportion of intron versus exon reads is the highest in Nuclear Total sample, and the lowest in the cytoplasmic fraction (Fig.4.3). In terms of the number of reads (reflecting the mass of the corresponding RNA), around 75% of the reads align to pre-mRNA sequences. These proportions are similar to what has been previously described in mouse liver for nuclear total RNA [122]. In mammals, introns are much longer than in lower organisms, ranging from 1kb to 100 kb [82], which explain why intronic reads represent such a large fraction of the RNA population.

Nuclear PolyA RNA population (NA) represents the population of fully-transcribed and polyadenylated transcripts. Interestingly, 45% of the reads mass still map on pre-mRNA. Splicing is thought to mainly occur co-transcriptionally, with 75 to 85% of the introns being removed while RNA Polymerase II is still transcribing [92], from budding yeast [85] to human tissues [89]. However, the mouse liver is the only example where co-transcriptional splicing has been shown to be less efficient, with only 45% of introns being co-transcriptionally removed [90].

When these fractions of mass are normalized by the length of the gene features (exons and introns) and thus reported as a fraction of the number of molecules (in RPKM, normalised by the length of the transcript), about 1 in 8 polyadenylated transcripts in the nucleus is still a pre-mRNA, suggesting that further splicing is indeed occurring after transcription is completed. Note that in Fig.4.3.B, we only report fractions of protein coding transcripts in order to avoid biases associated with the different compositions of RNA biotypes (e.g. snoRNA present almost exclusively in Nuclear Total samples consume 20% of the exonic reads, see Fig.4.3.C). Finally, as expected, almost no pre-mRNA is detected in the cytoplasmic RNA populations. The few percent that are still counted as pre-mRNA could come from contamination by nuclear RNA during fractionation, or by an error during the pseudo-alignment with Kallisto's algorithm.

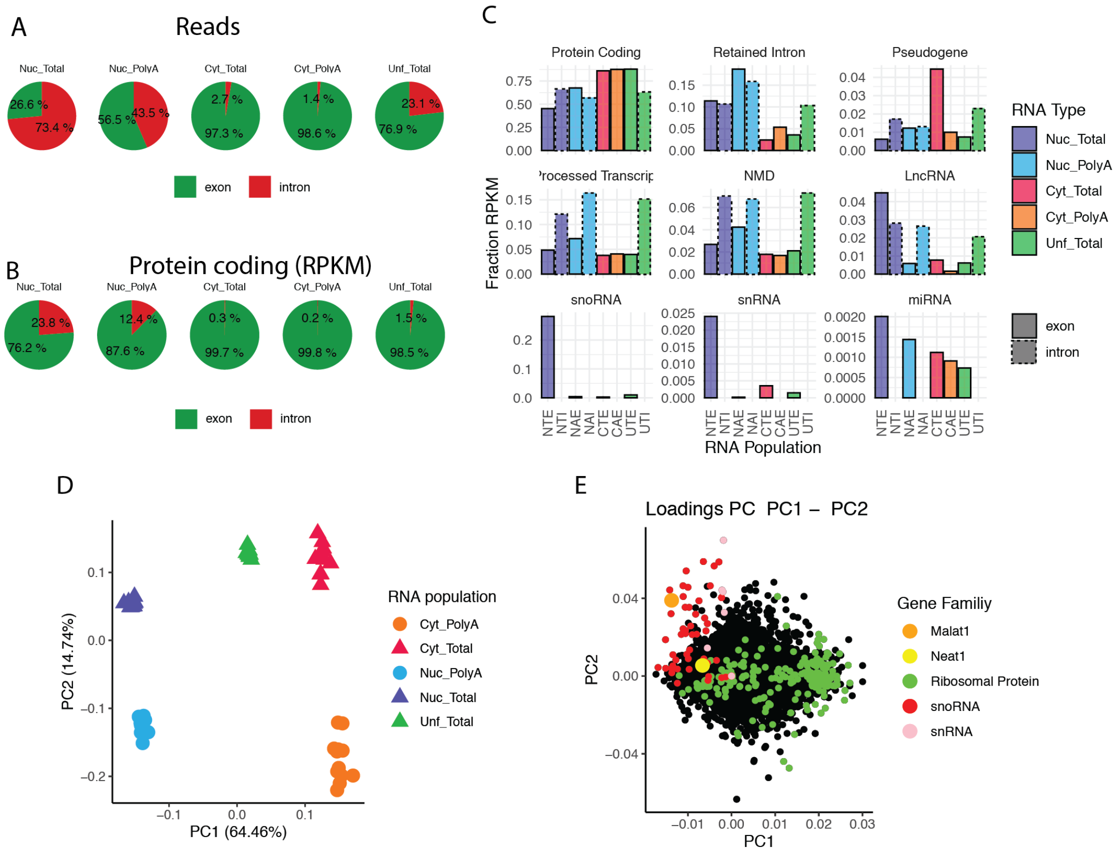


Figure 4.3 – Global characteristics of PolyA and Total RNA populations in nuclear and cytoplasmic fractions. A: Percentage of reads mapping on pre-mRNA or mRNA transcripts in each condition. B: Proportion of RPKM (normalised by length) of protein coding transcripts mapping on pre-mRNA and mRNA. C: Fraction of each biotype in each RNA sample. Dotted lines represent pre-mRNA. Pre-mRNA from cytoplasmic fraction are not shown. D: Principal Component Analysis using mature mRNA only. PC 1 separates Nuclear from Cytoplasmic samples, and PC2 separates Total from Polyadenylated samples. E: Loadings factors of PC1 and PC2, color-coded by biotypes. *Neat1* and *Malat1* are highlighted (yellow and orange dots).

The vast majority of detected RNAs are protein coding transcripts, making up to 85% of mRNA molecules in the cytoplasm and in the whole cells (Fig.4.3.B). All the other biotypes account for less than 10% of the total RNA population in the cytoplasm. The nucleus is enriched for functional small non-coding RNA. In fact, the most abundant small non-coding RNA are small nucleolar RNA (snoRNA) found in the nucleolus and scaRNA, found in Cajal-bodies. SnoRNA, like snRNA, are not polyadenylated, and are thus found almost exclusively in the Nuclear Total fraction. Retained Intron isoforms (RI) are the most abundant in the nucleus (10 to 20%). RI are usually enriched in the nucleus, where they are either exported, degraded, or spliced post-transcriptionally [110]. Of note, the true composition of nuclear and cytoplasmic RNA populations are different from what we report here as the extremely abundant ribosomal RNAs have been depleted before sequencing. Also, we excluded abundant tRNA that are difficult to sequence and map due to their secondary structure. We also excluded unannotated RNAs (*Gm*) or abundant RNA labelled as “miscellaneous” RNA.

Principal Component Analysis (PCA) on the mRNA measurements shows that the subcellular localisation explains most of the variability between samples (64%, Fig.4.3.D). Unfractionated samples lie between nuclear and cytoplasmic samples on PC1, although a little closer to cytoplasmic RNA. Differences due to the type of RNA (PolyA versus Total) account for 17% of the variability. Genes mostly expressed in the nucleus (upper left quadrant of Fig.4.3.D) belong to the family of snoRNA and snRNA as previously seen in Fig.4.3.C. The long non-coding genes *Malat1* and *Neat1*, hallmarks of nuclear-enriched transcripts, also drive the difference between nuclear and cytoplasmic samples [74]. On the other side of the plot, genes coding for ribosomal proteins (*Rpl*, *Rps*) are abundant in the cytoplasm. Overall, PCA confirms the previous observations with the proportions of biotypes in each compartment. Together with pre-mRNA and mRNA distribution, we recapitulated previously demonstrated characteristics of nuclear and cytoplasmic RNA distribution, and verified that our cellular fractionation of the liver cells performed as expected [194, 95, 208, 209].

4.5 Relationship between nuclear and cytoplasmic mRNA reveals signatures of export and cytoplasmic half-life

We hypothesised that the ratios of RNA from two different populations could reveal the variability in the parameters dictating RNA dynamics. We first considered the relationship between Nuclear Polyadenylated mRNA (NAE) and Cytoplasmic Polyadenylated mRNA (CAE) at steady-state. In the simplest scenario (no loss in the nucleus), the ratio of NAE over CAE reflects the ratio of two parameters: the nuclear export rate and the cytoplasmic degradation rate (Fig.4.1). Namely, a fast export from the nucleus to the cytoplasm would deplete the amount of RNA in the nucleus, and a long cytoplasmic half-life in the cytoplasm would lead to an accumulation of mRNA in the cytoplasm. Both processes decrease the ratio between nuclear and cytoplasmic RNA and results in an enrichment of RNA in the cytoplasm. On the contrary, RNAs with a long export time (or long nuclear retention) and short half-life in the cytoplasm would increase the nuclear to cytoplasmic ratio, resulting in an apparent nuclear accumulation (Nuc/Cyt ratio, later calculated as the difference of $\log_2$(RPKM) of NAE and CAE). Therefore, we investigated the relative subcellular levels.

Across all genes, the expression levels of NAE span 2.6 orders of magnitude in $\log_{10}$ (from -2.5 to 6.3 in $\log_2$, whiskers of the boxplot, Fig.4.4.A). CAE has a wider distribution of expression level, from -2.5 to 10 in $\log_2$, which corresponds to about 4 orders of magnitude. The most abundant genes code for secreted proteins such as Murine Urinary Proteins *Mups*, Albumine *Alb*, and Apolipoproteins (*Apoc1, Apoa2*), and are enriched in the cytoplasm. The NAE/CAE ratio varies by a factor of almost 2500 (outliers excluded), with a median $\log_2$ around -2.3, indicating that RNA transcript counts are more abundant in the cytoplasm than in the nucleus. However, a fraction of transcripts are also found enriched in the nucleus. A previous study by Bahar-Halpern et al. [196] showed that in the mouse liver, some protein coding genes were indeed enriched in the nucleus, such as *Nlrp6* (here $\log_2$-ratio = 1.6), and glucose metabolism-related genes *Mlxipl* ($\log_2$-ratio = 3.4), *Gcgr* ($\log_2$-ratio = -0.1), and *Gck* ($\log_2$-ratio =-1.2). Indeed, we verified the subcellular localisation of *Mlxipl* by single-molecule RNA-FISH (smFISH), and the accumulation of nuclear transcripts was evident, although not quantifiable due to its high abundance and crowdedness of mRNA molecules (Fig.4.4.C). On the other side of the spectrum, *Actb* is a typical example of a cytoplasmically enriched transcript ($\log_2$-ratio = -3.64). By smFISH, the $\log_2$ ratio varied between time points, ranging from -1.42 to -3.8, with a median of -2.7.

We next aimed to determine which molecular processes (rate constants) were most responsible for determining the ratio of the mRNA between the two measurements. In our simplified model, the steady state concentration of $\log_2$(NAE) = $\log_2$(Transcription rate) + $\log_2$(export time), and the steady state concentration of $\log_2$(CAE) = $\log_2$(Transcription rate) + $\log_2$(half-life). Fig.4.4.B shows the $\log_2$-ratio of NAE/CAE as a function of either $\log_2$(NAE) or $\log_2$(CAE), and the correlation coefficient is higher for CAE than for NAE ($\rho = 0.67$ for CAE, $\rho = 0.06$ for NAE). Given that the $\log_2$(Transcription rate) is common to both CAE and NAE, the strong correlation coefficient suggests that the cytoplasmic half-life has a higher influence on the ratio than the export time. Thus, the main driver of heterogeneity in the nuclear-cytoplasmic ratio is the cytoplasmic half life rather than the nuclear retention time.

We examined several genomic features that have been proposed to modulate kinetic rates (transcription rate, export, decay) and thus mRNA localisation, such as gene length, transcript length, number of exons, 3'UTR and 5'UTR length [208, 209, 126]. At the gene-level, the 5'UTR length and the gene length do not correlate with the NAE/CAE ratio (Fig.4.4.F, J). The 3'UTR length has a positive correlation ($\rho = 0.39$), indicating that transcripts with longer 3'UTR sequences are less cytoplasmic (Fig.4.4.G). The length of the 3'UTR is often proposed as a predictor of mRNA stability, because of the presence of many regulatory elements. Indeed, 3' UTR acts as a binding site for various regulatory mechanisms, most of which involved in decreasing the cytoplasmic stability of the transcript [210]. This is notably the case of element promoting mRNA decay such as AU-rich elements, GU-rich elements (GREs) and PUF protein-binding elements [152], and miRNAs binding sites involved in translation repression and mRNA cleavage [192]. However, both stabilising and destabilising RNA-Binding Proteins often compete for the same mRNA substrate [152], which in the end results in conflicting or undetectable effect of the 3'UTR length on processing rates [209]. Thus, this positive correlation between 3'UTR length and nuclear localisation could reflect both a slow export or a short half-life due to an increased presence of destabilising regulatory elements in the 3'UTR [210].

Interestingly, we found that transcript length (corresponding to the sum of all exon lengths) has a significant positive correlation with nuclear localisation ($\rho = 0.5$, Fig.4.4.D). Likewise, the number of exons also explain part of the NAE/CAE ratio ($\rho = 0.4$, Fig.4.4.E). Transcripts of hundreds kB take several hours to be fully transcribed, which could explain the longer residence time in the nucleus [211]. However, in that case, we would expect a correlation with the gene length (exons + introns), which was not observed. One possible explanation would be that rather than transcription time, the number of splicing events needed per transcript is the limiting step before export.

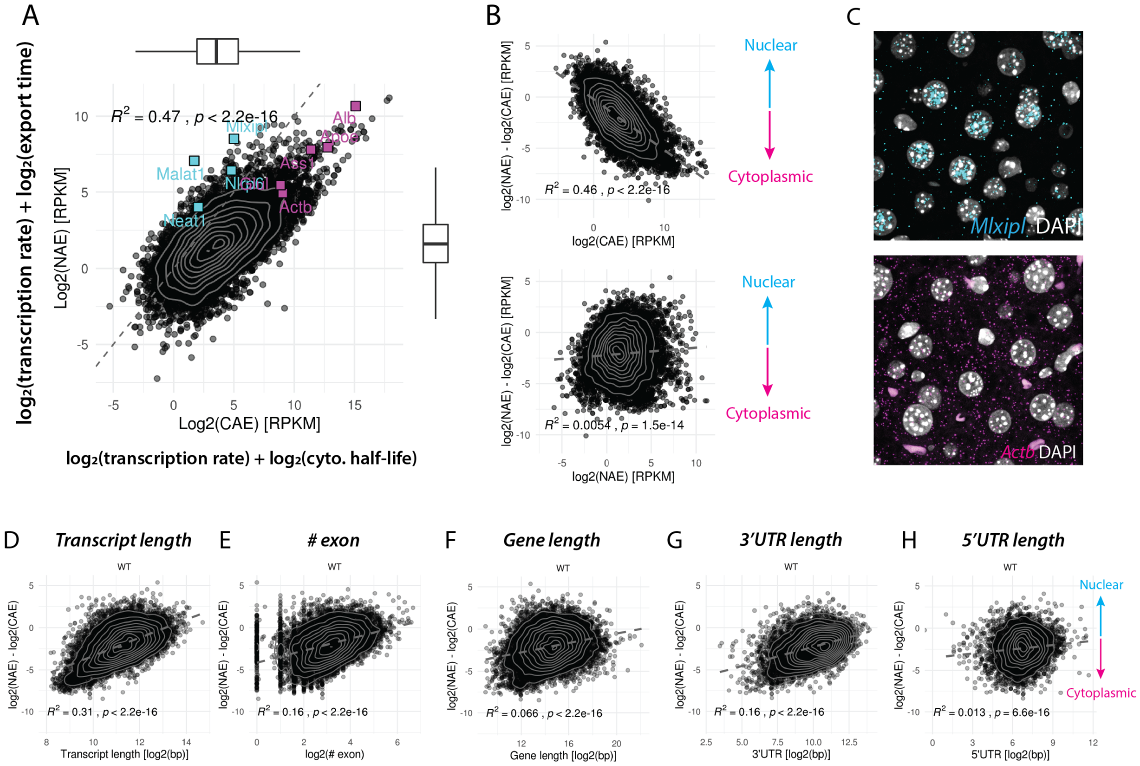


Figure 4.4 – The ratio of nuclear and cytoplasmic mRNA reflects the relationship between export time and cytoplasmic half-life. A: NAE versus CAE, in $\log_2$(RPKM), averaged over 6 time points and 12 animals. WT animals only. Cyan squares represent known nuclear genes, and magenta squares cytoplasmic genes. Dashed line indicates the 1:1: ratio. We note here that the abline has been scaled with the dataset from [196], and does not represent the absolute ratio. In grey: Two-dimensional kernel density estimation. Number of genes = 10914. B: NAE / CAE ratio in $\log_2$-scale against the average expression in the cytoplasm (up) or in the nucleus (bottom). C: smFISH of the nuclear transcript *Mlxipl* and the cytoplasmic transcript *Actb* in liver FFPE sections, at ZT12. Nuclei are stained with DAPI (grey). E,F,G,H: NAE/CAE $\log_2$-ratio against the following genomic features: Transcript length, number of exons 5'UTR length, 3'UTR length, gene length. These features are transcript-specific, and were averaged based on the relative expression level of each transcript to obtain a value per gene. Grey dashed line is the regression line

Because different classes of RNA are potentially processed by distinct regulatory programs, particularly in the context of localisation [114], we performed a similar analysis as above, but stratified by RNA biotypes (Fig.4.5).

Protein coding (PC) transcripts are mostly located in the cytoplasm (median $\log_2$(NAE/CAE): -2.3, Fig.4.5.B). The cytoplasmic localisation of PC transcripts is not surprising, because their ultimate purpose is to be translated and to produce proteins in the cytoplasm. The most nuclear-enriched transcripts are the retained intron (RI), with a median $\log_2$(NAE/CAE) ratio of -1.3 (Fig. 4.5.B). RI transcripts are exported to the cytoplasm where they can be translated, but many are rapidly degraded because the remaining intron introduced a Premature Termination Coding (PTC), which triggers the degradation by nonsense-mediated decay [110, 152]. Some retained introns are *exclusively* found in the nuclear compartment, in which case they are referred to as "Detained Introns" [108]. The specific case of Detained Introns would not be captured in the NAE/CAE ratio, because the ratio implies that the

transcripts are detected in both compartments. Thus, the actual nuclear enrichment of RI transcripts might be even stronger.
For a long time, long non-coding RNA (lncRNA) were thought to be mainly found in the nucleus, where they regulate chromatin structure (*Firre*), gene expression (*Xist*, X chromosome inactivation), or act as scaffold of nuclear condensates (*Malat1*, *Neat*). However, many lncRNA are also exported to the cytoplasm [212]. Recently, the subcellular localisation of lncRNA has been extensively studied [126, 119, 124]. Because they do not produce any protein, their localisation determines their functions [124]. Here, the well-known *Malat1* and *Neat1* are the most nuclear transcripts. *Dreh*, a lncRNA regulating cytoskeleton, is one of the most cytoplasmically enriched lncRNA. The median NAE/CAE $\log_2$-ratio is similar to that of protein coding transcripts. Of note, some lncRNA such as *Firre* are annotated as “Processed Transcript” (PT) in the Ensembl database. PT are even more cytoplasmically enriched, suggesting that indeed, long non-coding RNA are not only restricted to the nucleus, but found in a wide range of locations [212].

The positive correlation between cytoplasmic expression level and cytoplasmic localisation previously observed in Fig.4.4.B is valid for all biotypes, suggesting that cytoplasmic stability rather than export influences the most subcellular distribution (Fig.4.5.A). This is particularly the case for PC transcripts, where no correlation is found with the expression level of NAE. Interestingly, we find now a significant positive correlation between the nuclear expression level and the NAE/CAE $\log_2$-ratio for RI-RNA and lncRNA, accounting for 20% of the variance ($R^2 = 0.22$ and $R^2 = 0.2$). Therefore, the modulation of the export rate can partially determine the subcellular localisation of lncRNA and RI-RNA, and these differences could be the signature of the different mechanisms regulating processing of protein coding transcripts or lncRNA and RI-RNA. Finally, the positive correlation of the NAE/CAE $\log_2$-ratio with the (mature) transcript length is detected for all biotypes, and is striking for PC genes compared to that observed at the all gene-level ($R^2 = 0.41$. Fig.4.5.D). Therefore, the subcellular localisation related to transcript length is not a biotype-specific feature.

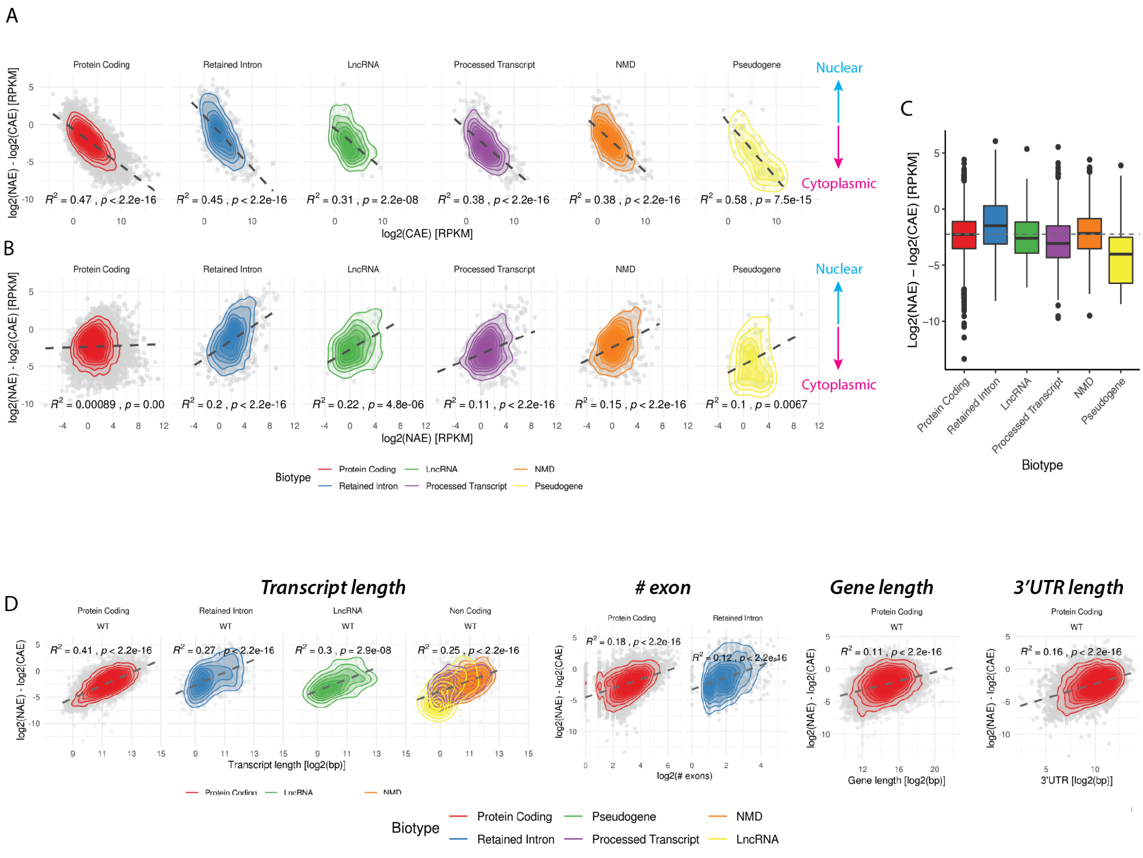


Figure 4.5 – Relationship between NAE and CAE per biotype reveals biotype-specific and biotype-unspecific patterns. A: $\text{Log}_2(\text{NAE/CAE})$ shows a strong correlation with the expression level of CAE, defined as $\log_2(\text{Transcription rate}) + \log_2(\text{half-life})$. B: Correlation of $\log_2(\text{NAE/CAE})$ against the expression level of NAE, defined as $\log_2(\text{Transcription rate}) + \log_2(\text{export time})$. C: boxplot of the $\log_2(\text{NAE/CAE})$ ratio per biotype. Dashed line indicates the median level when considering all the transcripts together. D: $\log_2(\text{NAE/CAE})$ against several genomic features: Transcript length, number of exon, 3'UTR exon, and gene length. When R^2 is less than 0.1, scatterplot is not shown. Transcripts length positively correlates with the $\log_2(\text{NAE/CAE})$ in all RNA biotypes. In all figures, number of gene per biotype is: PC = 10104, RI = 1093, LncRNA = 88, PT = 868, NMD = 883, Pseudogene = 72.

4.5.1 Localisation of Protein Coding transcripts matches the localisation of the encoded proteins

Although most protein coding transcripts are mainly found in the cytoplasm, and that this enrichment most likely results from a short cytoplasmic half-life, additional mechanisms may retain PC transcripts in the nucleus [114, 113, 126]. For instance, one proposed role of nuclear retention is to buffer noise associated with stochastic transcription [196, 127]. To characterise the biological functions of these nuclear-enriched protein coding transcripts, we performed a functional enrichment analysis using Gene Ontology (GO) database. We compared differentially expressed protein coding transcripts in NAE and CAE (absolute value of $\log_2(\text{FC}) > 2$, 1600 genes in both groups out of 8925). We could confirm an interesting and somewhat puzzling observation previously reported by Fazal et al. [208]: they showed a concordance between protein and RNA localisation in human cells, using "APEX-seq", a proximity

RNA-labelling and sequencing method. Here, we found that nuclear transcripts code for nuclear proteins involved in epigenetic modifications such as DNA methylation (KMT2, SETD1, EHMT1) and DNA demethylation (KDM1-4) (Fig. 4.6.A). GO Terms related to mRNA processing (*Iws1*, *Pabpn1*, *Cpsf6*) and particularly export (THO complex, *Nxf1*, Nucleoporins NUP) are also significantly enriched in the nucleus. Additionally, factors involved in transcription by PolII and PolII (*Ice1* and *Ice2*, *Snapc4*, *Cdh8*, *Ell*), were also enriched in the nucleus.

On the other side of the spectrum, cytoplasmic RNA are mainly involved in translation, from ribosomal RNA biogenesis and assembly (RPL, RPS), translation initiation (EIF), to elongation (EEF), and also involved in mitochondrial translation. Transcripts coding for proteins involved in oxidative phosphorylation are also enriched in the cytoplasm: subunits of the mitochondrial respiratory complexes I-IV (NDUF, SDH, UQCR and COX), Electron carrier (ETF and CYCS), and subunits of ATP synthase (ATP5). Two metabolic functions mainly carried out by hepatocytes are also enriched in the cytoplasm: fatty acid β-oxidation and detoxification. Proteins involved in the four reactions of the fatty acid β-oxidation cycle (acyl CoA dehydrogenase, enoyl CoA hydratase, hydroxyacyl-CoA dehydrogenase and acetyl-CoA acyltransferase) and other auxiliary enzymes involved in fatty acid oxidation in mitochondria are coded by cytoplasmic RNA, as well as Glutathione-S-transferases (GST) from all three superfamilies (cytosolic, mitochondrial, and microsomal), involved in Phase-II of detoxification of xenobiotics.

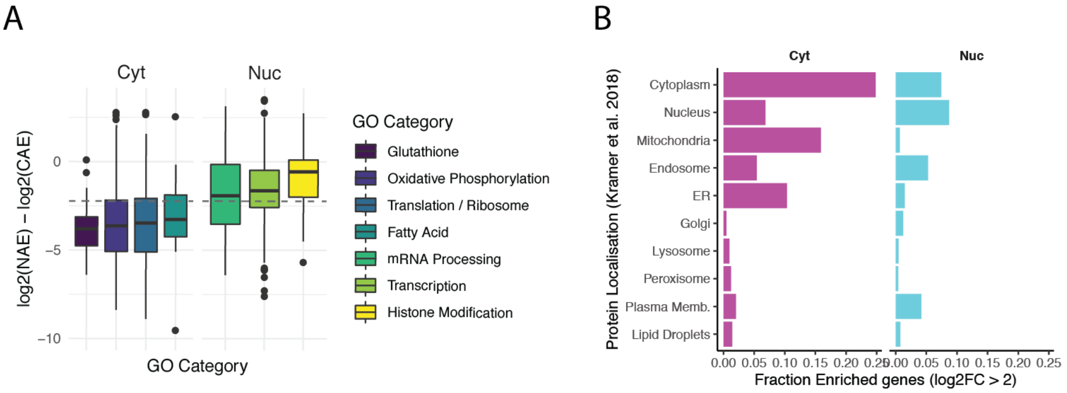


Figure 4.6 – Localisation of protein coding transcripts matches the localisation of the encoded protein. Boxplots of the $\log_2$(NAE/CAE) ratio of protein coding transcripts belonging to GO Terms significantly enriched in the cytoplasm or in the nucleus, based on EnrichR (Adjusted P-value < 0.1 and Combined Score > 50). Differential gene expression was calculated with DESeq2. Transcripts (grouped by biotype) with a $\log_2$FC > 2 of < -2 were used (1685 in NAE and 1651 in CAE). GO Terms were grouped based on semantic similarity. Boxplots are constructed using all the genes annotated in the significant GO Terms, whether these genes are enriched in one compartment or not. Dotted line indicates the median $\log_2$(NAE/CAE) ratio of all protein coding genes. B: Localisation of transcripts enriched in the cytoplasm (magenta) and in the nucleus (cyan) based on a published dataset of mouse liver tissue [213].

We further verified the overall concordance between RNA and protein localisation using a published dataset, in which the authors quantified proteins in different cellular fractions of mouse liver cells by MASS-spectrometry, and assigned them to an organelle by protein correlation profiling [213]. In our dataset, more than one third of the transcripts enriched in the cytoplasm (35%) were coding for cytoplasmic proteins, and one fourth for proteins localised in mitochondria, compared to only 10% for nuclear protein. On the other hand, a similar proportion of nuclear transcripts were coding for nuclear or cytoplasmic proteins (28% and 25%, Fig.4.6.B). Together, this analysis suggests that transcripts localisation matches to some extent protein localisation. Transcripts coding for highly expressed functions (respiration, translation) are enriched in the cytoplasm. Because these are constantly needed house-keeping functions, which need little regulation, transcripts are presumably stable and accumulate in the cytoplasm. On the other hand, nuclear enriched transcripts often perform regulatory processes (RNA processing, epigenetic modifications), which need to be dynamic. A rapid response and adaptation to a stimulus could not be achieved with long-lived transcripts, except if a particularly efficient destructive system is in place. Moreover, since nuclear retention acts as a passive filter reducing transcriptional noise, retaining these transcripts could allow a better control of the amount of transcripts available for translation [127].

4.5.2 Retained Intron as a class of nuclear-retained RNA

Retained Intron (RI) is a class of transcripts mainly found in the nucleus (Fig.4.5.C, [208], [214]). From there, they can be exported into the cytoplasm, where they are often targeted for nonsense-mediated decay [110]. Some are also degraded in the nucleus by nuclear exosomes [110]. Interestingly, RI transcripts can be retained in the nucleus as stable transcripts and act as a reservoir of precursor mRNA, awaiting for a signal that triggers their splicing and subsequent export [112]. To gain more insight about the maturation (splicing) of Retained Intron, and to assess if RI indeed shows signatures of nuclear enrichment due to their presumably slow export rate, we analysed pairs of RI and PC transcripts that differ only by the presence of the retained intron (Fig.4.7.A). By restricting the analysis to pairs of RI-PC, we make the assumption that the export (or degradation) rate of the RI transcript is the removal of the retained intron, which then produces the PC mRNA. We compared the relative expression of nuclear and cytoplasmic RI mRNA (NAE-RI and CAE-RI) to their corresponding spliced PC isoforms (NAE-PC and CAE-PC). We summarised the different expression patterns in a heatmap and grouped genes in 5 clusters (Fig.4.7.B).

First, as expected, we observed that RI are far less abundant in the cytoplasm than in the nucleus (CAE-RI < NAE-RI), and that CAE-RI are also less abundant than their corresponding PC isoform (CAE-RI < CAE-PC), suggesting that RI transcripts are indeed mainly nuclear RNA. Out of the 260 pairs analysed, only 14 CAE-RI were more expressed than their corresponding CAE-PC (cluster 5). When we took a closer look at these genes, we noticed one limitation of the isoform-specific quantification by Kallisto, as illustrated by *Clptm1* (Fig. 4.7.G). In CAE, almost no reads map on the genomic region corresponding to the retained intron, suggesting that the main cytoplasmic isoform is the protein coding one. But the 3'UTR regions differ between PC and IR isoforms. Because most of the reads map on short 3'UTR region corresponding to RI, but barely on the longer 3'UTR corresponding to the PC isoform, Kallisto's algorithm still assigns a majority of the reads to the RI isoform, despite the intron being spliced. Thus, in this specific case, it is wrong to assume that the intron is still retained in the cytoplasm.

The majority of the PC transcripts are more abundant in the cytoplasm than in the nucleus. These cytoplasmically-enriched transcripts are in cluster 1 and 2. The main difference between cluster 1 and 2 resides in the relative abundance of the nuclear isoforms. In cluster 1 (n = 100), NAE-PC transcripts are slightly higher than NAE-RI, whereas in cluster 2, the RI isoforms are more abundant than the PC isoform (n = 87). In the second case, the lifetime of the RI is longer than the PC isoform, either because the post-transcriptional removal of the intron is a slow step, or because once fully-spliced, the PC is efficiently exported to the nucleus. In the end, these PC transcripts mainly reside in the cytoplasm, therefore, the splicing time of the remaining intron may only have a limited impact on the global subcellular distribution compared to the effect of cytoplasmic half-life.

Nuclear-enriched genes are clustered in two groups. In cluster 3 (n = 35), the RI and PC isoforms are present at comparable levels, suggesting similar relative lifetimes of NAE-PC and NAE-RI. In cluster 4 (n = 24), there is a clear predominance of NAE-RI over NAE-PC. A possible explanation for cluster 4 (high NAE-RI) is that the RNA transcript is retained in the nucleus in its immature form, and the splicing of the last remaining intron (maturation) is a rate-limiting step. Once matured, the PC transcript is rapidly exported. Genes in cluster 4 includes the splicing modulator *Arglu1*, a subunit of the pre-alpha-trypsin inhibitor complex *Itih3*, the Glucokinase Regulatory Protein *Gckr*, and *Shfl*. At the gene-level (when biotypes are not considered), they all appear as nuclear-enriched genes, and this is explained by the predominance of the RI isoform, while the PC is in fact more cytoplasmic. Transcripts in cluster 3 were significantly longer and had more exons compared to cluster 1 and 2 (pairwise t-test, p-value <0.05, data not shown), in agreement with the previous observation that longer transcripts are more nuclear (Fig.4.4.E), but no other genomic features, nor common functional roles (Gene Ontology, KEGG) that could differentiate the five clusters were found.

Despite the small number of genes analysed, we could reveal distinct signatures. For example, some transcripts that appear to be nuclear-enriched at the gene-level are in fact retained as their "immature" RI form, while the spliced mature PC form does not accumulate in the nucleus but is rather rapidly exported. This suggests intron retention is a potential post-transcriptional mechanism that modulates the nuclear residence time of protein coding transcripts.

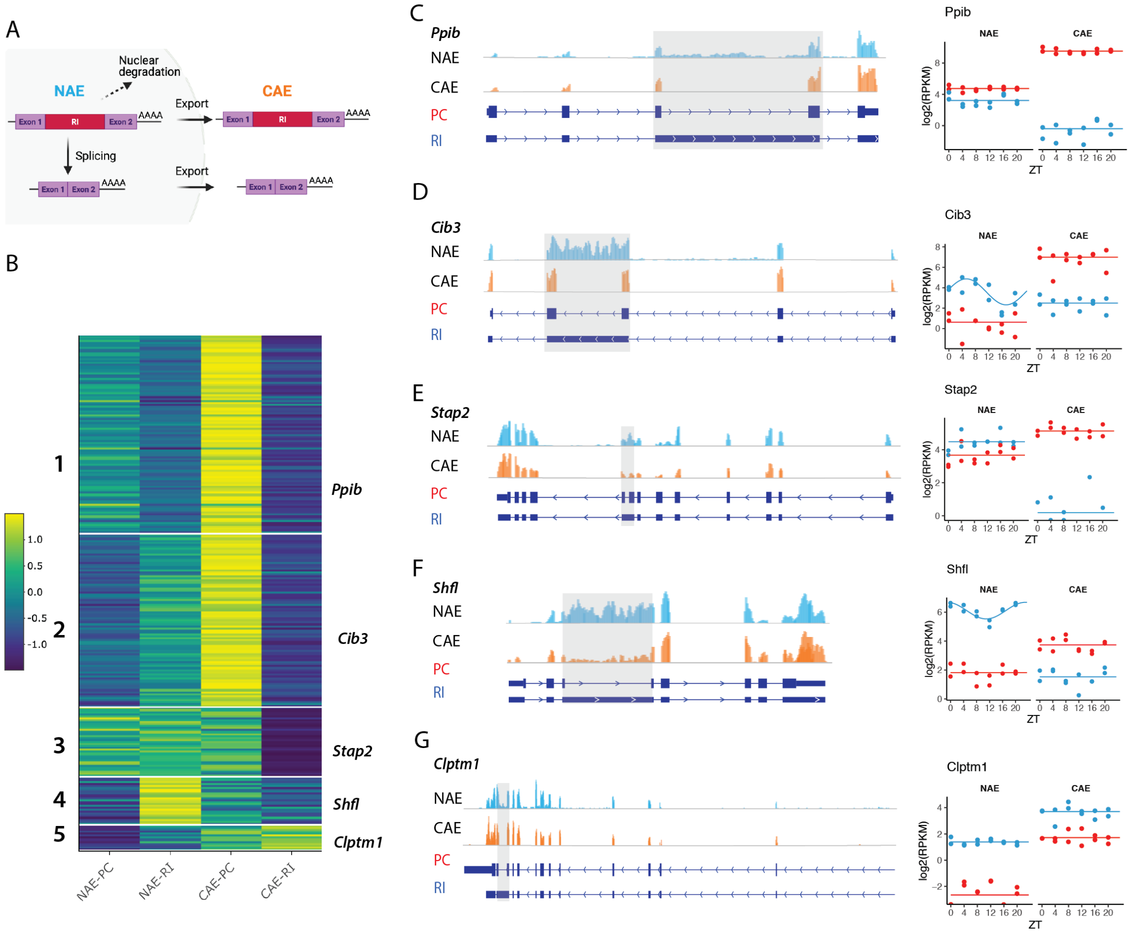


Figure 4.7 – Retained Intron is a nuclear-enriched RNA biotype. A: Fates of an RI transcript: in the nucleus, RI is either degraded by nuclear exosomes, exported as such into the cytoplasm, or spliced post-transcriptionally to become the final protein coding transcript. B: Expression profiles of the RI isoform in the nucleus and in the cytoplasm (NAE-RI, CAE-RI) and the protein coding isoform in the nucleus and in the cytoplasm (NAE-PC, CAE-PC) of 260 genes. Only isoforms that differ by the presence of one additional exon were selected. Expression levels were normalised per gene (per row, z-score). Genes were further clustered in 5 groups using hierarchical clustering. C-G: Genome tracks of representative genes from each cluster (1-5). Grey box indicates the retained intron. On the left, temporal profile of the corresponding RI and PC genes. Ensembl transcript ID (RI and PC) *Ppib*: ENSMUST00000213785, ENSMUST00000034947, gene length: 6.39 kb , *Cib3*: ENSMUST00000211946, ENSMUST00000098630, gene length: 8.5 kb, *Stap2*: ENSMUST00000233494, ENSMUST00000043785, gene length: 8.5kb, *Shfl*: ENSMUST00000043911, ENSMUST00000175820, gene length: 5.6 kb, *Stap2*: ENSMUST00000227181, ENSMUST00000006697, gene length: 7.78 kb, *Clptm1*: ENSMUST00000208846, ENSMUST00000055242, gene length: 31.77kb. (Blue: RI, Red: PC).

4.6 Relationship between nuclear pre-mRNA and mRNA reveals signatures of splicing and export times

We now analyse the first step of our simple model of RNA processing (Fig.4.1.C, D) in order to quantify the relative contribution of splicing and nuclear export rates on the ratio of pre-mRNA to nuclear mRNA.

If we assume that most introns are spliced before the end of transcription and polyadenylation (i.e. co-transcriptional splicing), NTI (Nuclear Total pre-mRNA) corresponds to the nascent transcripts [215, 216, 47]. At steady-state, the ratio of NTI versus NTE (or versus NAE if one neglects exons of the nascent transcript) represents the ratio of splicing time over export time. The ratio is high if the transcript is slowly spliced and / or quickly exported. On the contrary, the ratio is low if splicing is fast and / or if the mRNA is retained in the nucleus. If splicing is not purely co-transcriptional, a fraction of pre-mRNA are polyadenylated and are thus captured in NAI samples. In this case, the relationship between NTI and NAE still represents the ratio of splicing and export times, but the rate s would be a combination of co-transcriptional splicing (nascent transcript → NAE), and post-transcriptional splicing (NAI → NAE). In order to estimate the proportion of polyadenylated pre-mRNA (NAI) in the total population of pre-mRNA (NTI), we first recall here the observation that ~12% of the nuclear polyadenylated transcriptome encoding for proteins is composed of pre-mRNA (Fig.4.3). At the gene-level, the median $\log_2$(NAI/NAE) is -3.05, meaning that the majority of polyadenylated mRNA are 8 times more abundant than the pre-mRNA isoform (Fig.4.8). As a comparison, in the nuclear total RNA population, mRNA is only 2.6 times more abundant than pre-mRNA ($\log_2$(NTI/NTE) = -1.4). The two nuclear mRNA populations (NAE and NTE) correlate with a Pearson's correlation ρ of 0.9 (Fig.4.11.B), and the $\log_2$-ratio of NTI/NAE of -1.32 is very close to the -1.41 found for NTI/NTE. These two observations suggest that there is a large overlap between NAE and NTE populations, apart from the differences noted in Fig.4.3.C (sn(o)RNA, etc.), and the differences corresponding to the nascent transcripts are most likely minor. By assuming that NAE equals NTE, and by comparing $\log_2$(NAI/NAE) and $\log_2$(NTI/NAE), we deduce that NTI consists of 30% of polyadenylated pre-mRNA (NAI). This is a rough estimation, but gives an idea of the composition of NTI genome-wide, and suggests that post-transcriptional splicing might play a role as an mRNA maturation step.

To start analysing the relationship between splicing and export rates, we first neglect post-transcriptional splicing and compare NTI and NAE. We find a strong correlation between the two measurements (Fig.4.8.A, $\rho = 0.77$), suggesting that transcription rate, which is the common factor to both NTI and NAE (Fig.4.1.D), has a larger effect on the expression level of NTI and NAE than export and splicing rates. By performing a similar comparison as in Fig.4.4.B, we find that the $\log_2$-ratio of NTI/NAE is more correlated with the export time (NAE) than with the splicing time (NTI) (Fig.4.8.C and D, $\rho = 0.53$ for NAE, $\rho = 0.12$ for NTI). This observation suggests that, on average, the abundance of spliced nuclear mRNA (NAE) is more strongly determined by the variation of nuclear retention time rather than by a fast splicing of the pre-mRNA. Note, however, that because transcription, splicing and export are not always independent processes but can be functionally coupled, the interpretation of the correlations may be more complicated. For instance, splicing factors are recruited along with conserved mRNA

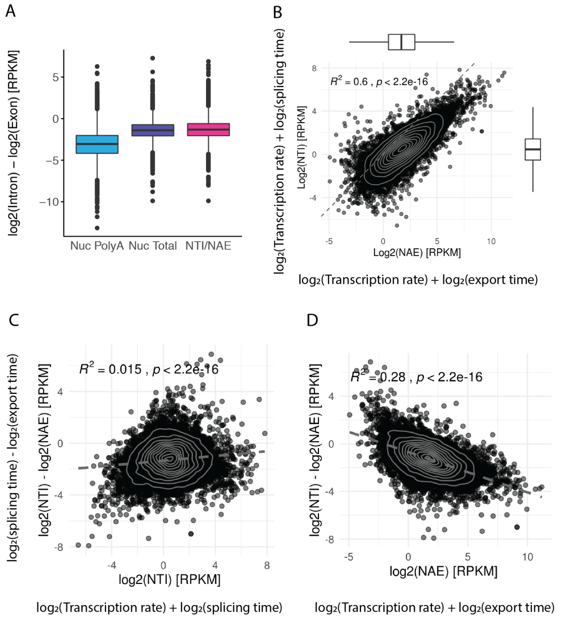


Figure 4.8 – The ratio of nuclear pre-mRNA and nuclear mRNA reflects the relationship between transcription rate, splicing time, and export time. A: $\log_2$ ratio of pre-mRNA versus mRNA in Nuclear PolyA (NAI/NAE), in Nuclear Total (NTI/ NTE) and in NIT/NAE. B: NTI versus NAE, in $\log_2$(RPKM), averaged over 6 time-points. According to our model, $\log_2$(NTI) is defined as $\log_2$(Transcription rate) + $\log_2$(splicing time) and $\log_2$(NAE) is defined as $\log_2$(Transcription rate) + $\log_2$(export time). In grey: kernel 2d density. Boxplots on top and on the right show the distribution of NTI and NAE. C : NTI / NAE ratio in $\log_2$-scale, defined as $\log_2$(splicing time) - $\log_2$(export time), against the average expression of NTI. D : NTI / NAE ratio in $\log_2$-scale against the average expression of NTI. Grey line indicates the linear regression.

export machinery (TREX) and Exon Junction Complex during transcription, coupling splicing with export [217]. The local environment associated with highly transcribed genes has a high concentration of splicing factors [218, 219, 137], further linking transcription rate with efficient splicing. Therefore, the anticorrelation of NAE (defined as $\log_2$(transcription rate) + $\log_2$(export time)) with NTI / NAE (defined as $\log_2$(splicing time) - $\log_2$(export time)) could also be interpreted as a high transcription rate promoting efficient splicing (short splicing time).

We next stratified the analysis by biotype. As expected, we still do not observe a relationship between NTI/NAE $\log_2$-ratio and NTI for protein coding transcripts (Fig.4.9.A), since this group makes up the majority of genes. However, we observe a weak positive correlation for Retained Intron, lncRNA, and non-coding RNA (Pseudogenes, NMD, Processed transcripts). This suggests that splicing time influences the overall abundance of pre-mRNA over mRNA, or that transcription rate and splicing rate are not as coordinated as for PC genes. Even if the biogenesis of most RNAs included in this analysis follows the same steps (transcription by RNA Polymerase II, 5'm^{7}G capping, 3'polyadenylation), the processing patterns could differ from one biotype to another [124]. For instance, lncRNA are notoriously known to be poorly spliced, as seen in Fig.4.9.B [220]. Also, in contrast to PC genes, many lncRNA are transcribed by an RNA PolII with a differentially phosphorylated C-terminal domain [221], which may also contribute to the increased intron levels. Concerning the relationship of the NTI/NAE $\log_2$-ratio with NAE, Retained Intron (RI) and non-coding RNAs (ncRNA) are the most strongly correlated with the export time (Fig.4.9.C). Here, note that when we compare NTI and NAE, the export rate e reflects not only the nucleo-cytoplasmic shuttle of the RNA transcript, but also any process that makes the nuclear mRNA "disappear". In case of an RI transcript, e therefore includes the post-transcriptional splicing of the remaining intron, after which the transcript is no longer an RI isoform. Additionally, RI, and other aberrant lncRNA and ncRNA are subject to RNA surveillance mechanisms and degraded if necessary by RNA exosomes [222]. Therefore, the signatures of export and splicing processes are visible in the pre-mRNA/mRNA levels of the purified nuclear RNA, and could additionally reflect distinct mechanisms regulating mRNA levels of protein coding transcripts and other RNA biotypes.

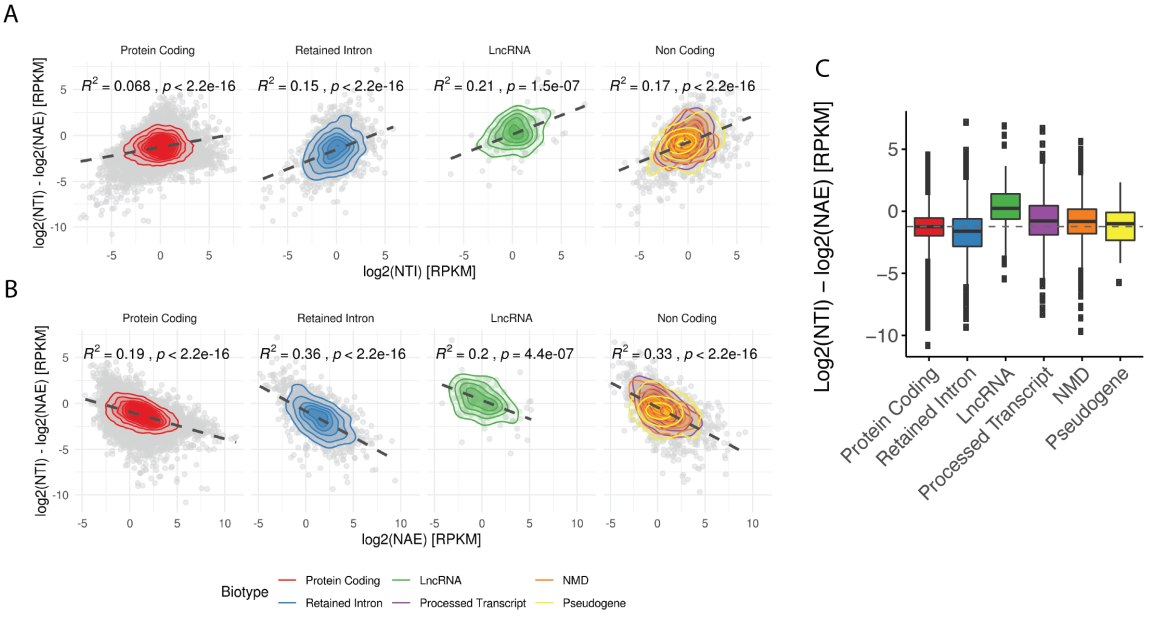


Figure 4.9 – Relationship between splicing and export times per biotype. A: NTI / NAE ratio in $\log_2$-scale, defined as $\log_2$(splicing time) - $\log_2$(export time), against the average expression of $\log_2$(NTI). B: NTI / NAE ratio in $\log_2$-scale against the average expression of $\log_2$(NAE). Gray line indicates the linear regression. Non-Coding RNA consists of NMD, Pseudogenes and Processed Transcripts that show similar patterns. C: $\log_2$(NTI / NAE) of each biotype, showing that LncRNA are the least well spliced RNA.

4.6.1 Nuclear pre-mRNA reveals the extent of co- versus post-transcriptional splicing

On average, polyadenylated nuclear RNA consists of 10% - 15% of pre-mRNA (Fig.4.3.B). Thus, although the average expression level of NAI is low (Fig.4.8.A), a non-negligible fraction of introns is spliced post-transcriptionally [90].

To further distinguish post-transcriptional splicing events, we compared the relative expression level of polyadenylated pre-mRNA (NAI), defined in our model as $\log_2$(Transcription time) + $\log_2$(post-transcriptional splicing time), and nuclear mRNA (NAE), defined as $\log_2$(Transcription time) + $\log_2$(export time) (Fig.4.10). Analogously to previous analysis, the ratio at steady-state results from different combinations of rates: when NAI/NAE is high, either the post-transcriptional removal of intron is slow, or fully spliced mRNA are quickly exported. On the other hand, the NAI/NAE $\log_2$-ratio is low when mature mRNA are retained in the nucleus, or alternatively, when post-transcriptional splicing is fast, which can also be interpreted as the fact that most introns have already been removed co-transcriptionally. We find that the $\log_2$-ratio of NAI/NAE correlates better with NAI than with NAE ($\rho = 0.65$ for NAI, $\rho = 0.005$ for NAE). Given that the $\log_2$(Transcription rate) is common to both NAI and NAE, the strong correlation coefficient suggests that the post-transcriptional splicing time has a larger influence on the ratio than the export time (Fig.4.10.A, B). All RNA biotypes have similar correlation strength (Fig.4.10.C, D). However, they differ by the splicing ratio: lncRNAs are again the least well spliced RNA class, while protein coding transcripts have the lowest NAI/NAE $\log_2$-ratio (Fig.4.10.C). Moreover, polyadenylated pre-mRNA is never enriched over mRNA ($\log_2(\mathrm{NAI} / \mathrm{NAE}) < 0$ for >98% of all the protein coding transcripts), thus, post-transcriptional removal of introns always happens on a shorter time-scale than export, and protein coding transcripts are not retained in the nucleus in the unspliced form. However, as shown in the previous section (see 4.5.2), hundreds of protein coding transcripts contain a retained intron. It is therefore possible that rather than being retained as a “protein coding pre-mRNA”, the transcript is retained as its unspliced Retained Intron isoform.

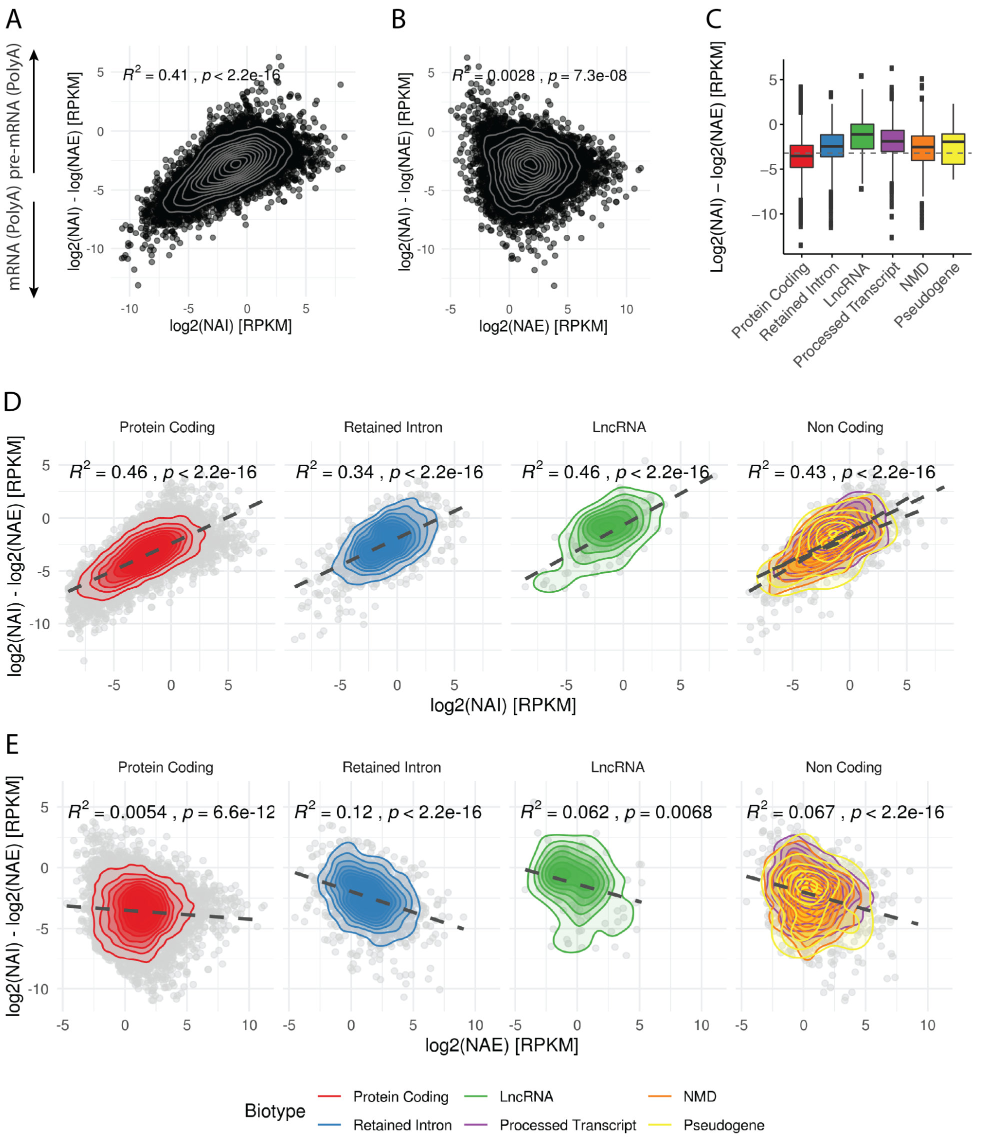


Figure 4.10 – Comparison of polyadenylated nuclear pre-mRNA and mRNA reveals signatures of post-transcriptional splicing. A and B: $\log_2$-difference between nuclear pre-mRNA (NAI) and nuclear mRNA (NAE) against $\log_2$(NAI) in A and against $\log_2$(NAE) in B. In grey: kernel 2d density. The NAI / NAE $\log_2$-ratio has a stronger correlation with the expression level of NAI than NAE. C: NAI / NAE $\log_2$-ratio for each biotype. LncRNA are the least well-spliced mRNA. D and E: same plot as in A and B, but for each biotype. Grey line is the linear regression line.

To further distinguish different scenarios explaining the relationship between NAI and NAE, we compared the $\log_2$-ratios with the Nuclear Total RNA population. NTI consists of both nascent pre-mRNA and NAI, while NTE consists of both nascent mRNA and NAE. We estimated above that NTI consists of ~30% of NAI, but this ratio varies for each gene depending on the extent of co- versus post-transcriptional splicing. Therefore, by integrating both Nuclear Total and Nuclear PolyA RNA population in the same analysis, we aim to estimate the extent of co- and post-transcriptional splicing.

As expected, the level of NTI is higher than NAI, because NTI represents the sum of both NAI and nascent pre-mRNAs (Fig.4.11.A). The situation when NAI value approaches NTI is when the population of nuclear pre-mRNAs mainly consists of fully transcribed and polyadenylated transcripts, and the proportion of nascent pre-mRNA is very small. This situation can happen if the gene length is short: transcription and 3' end processing are already terminated by the time the spliceosome assembles and splices the newly transcribed intron. If splicing and transcription are two processes occurring concurrently, the longer the gene, and particularly the longer the downstream exon, the higher the probability for the upstream intron to be detected and spliced. We indeed observe a strong correlation between the expression of NAI and the gene length, while NTI, representing the total pool of pre-mRNA in the nucleus, is not influenced by the gene length (Fig.4.12.A, B). This suggests that short genes are less co-transcriptionally spliced than long genes. Among genes that are poorly co-transcriptionally spliced due to their short length, we find for instance members of different classes of apolipoproteins (APOA, APOC, APOE, APOF, APOL, APOM, Fig.4.11). On the other side of the spectrum, we find genes with a much higher value of NTI compared to NAI, suggesting that for those genes, NTI is mainly composed of nascent pre-mRNA and that polyadenylated pre-mRNA are relatively less abundant. For these genes, technically, we cannot determine if the introns are removed while transcription is still ongoing, or removed on an extremely fast time-scale post-transcriptionally. Co-transcriptional splicing creates a typical sawtooth pattern of decreasing density of reads mapping on intron, with higher signal toward the 5' end, described in nascent-seq dataset [89, 59]. We observed the sawtooth patterns in long genes such as *Egfr* or *Plcxd2* (Fig.4.11FD), supporting the hypothesis that when NTI level is much larger than NAI, NTI indeed reflects the population of nascent RNA. Therefore, the NTI/NAI $\log_2$-ratio represents the extent of co- versus post-transcriptional splicing. Moreover, the strong correlation of the $\log_2$-ratio and gene length (Fig.4.12) further supports the idea that the rate of co-transcriptional splicing is higher for long genes than short genes, potentially related to the total time required for the completion of transcription. This correlation was observed only for protein coding transcripts, and to some extent for NMD (R^2 of 0.15 compared to R^2 of 0.4 for PC transcript), suggesting that the mechanisms regulating splicing differ between distinct classes of RNA.

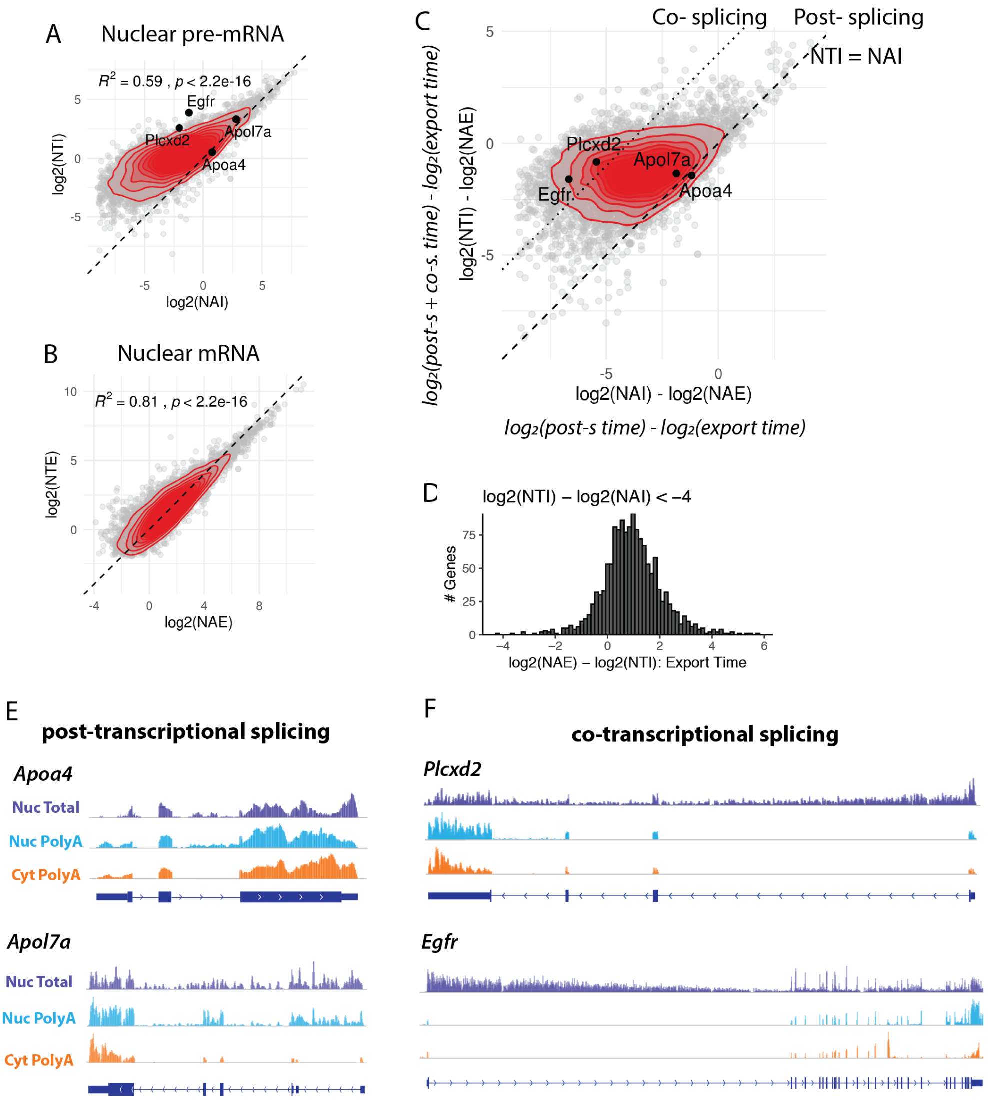


Figure 4.11 – Relationship between NTI and NAI reveals the extent of co- versus post- transcriptional splicing of protein coding transcripts. A: Comparison of pre-mRNA in PolyA et Total RNA population: $\log_2(\mathrm{NTI})$ against $\log_2(\mathrm{NAI})$. NTI, which consists of NAI and nascent pre-mRNA, is thus higher than NAI. Genes whose genome track are shown in D and E are labelled. Dashed line indicates the identity line. Red: 2D kernel density. Only protein coding genes are shown. B: Comparison of nuclear mRNA from PolyA and Total RNA population (NAE versus NTE). C: x-axis: the $\log_2$-ratio of NAI over NAE is defined as the $\log_2$(post-transcriptional splicing time (post-s)) minus $\log_2$(export time). y-axis: the $\log_2$-ratio of NAI over NAE is defined as the $\log_2$(post-s + co-transcriptional splicing (co-s)) minus the $\log_2$(export time). The dashed line indicates when splicing is mainly post-transcriptional (NTI = NAI). Dotted line is $\log_2(\mathrm{NAI}) + 4$. The more the transcript is located above the identity line (toward the dotted line), the higher the extent of co-transcriptional splicing. If a transcript is on the identity line, we cannot determine the relative contribution of the export time and the post-transcriptional splicing. If a transcript is mostly co-transcriptionally spliced, the value of the post-transcriptional splicing approaches 0. Assuming that co-transcriptional splicing time does not vary much, $\log_2(\mathrm{NTI}/\mathrm{NAE})$ is directly proportional to the export time. D: distribution of estimated $\log_2$export times (plus constant time) of co-transcriptionally spliced transcripts ($\log_2(\mathrm{NTI}/\mathrm{NAI}) < -4$). E: Genome track view of short genes that are mainly post-transcriptionally spliced: *Apoa4*: gene length 2.76 kb. *Apol7a*: 11kb. No reads map on the intronic region in the cytoplasm. F: Genome track view of long genes that are mainly co-transcriptionally spliced: *Plcxd2*: gene length 51kb. *Egfr*: 166 kb. Genome tracks are not to scale.

Finally, we analyse again the NTI/NAE $\log_2$-ratio using the assumption that if NTI values are close to NAI, splicing is mostly post-transcriptional, while if NTI is much greater than NAI, the splicing regime is mainly co-transcriptional. We plot the NTI/NAE $\log_2$-ratio against the NAI/NAE $\log_2$-ratio (Fig.4.11.C): in this way, genes on the identity line are presumably post-transcriptionally spliced, and genes further apart from the diagonal are more co-transcriptionally spliced. If an RNA transcript goes through extensive post-transcriptional splicing (on the identity line, Fig.4.11.C), the relative contribution of the export time and the post-transcriptional splicing time cannot be separately estimated. However, in the most extreme case where splicing is purely co-transcriptional, post-transcriptional splicing is virtually null. Because we estimate co-transcriptional splicing at the gene-body level, and not specifically for each intron, we make the simplifying assumption that the elongation rate is constant, and do not take into account local variation of rates, for instance around the transcription start sites, and the termination sites [**?**]. Assuming that co-transcriptional splicing time, which is directly related to the elongation rate [92], does not vary among genes, the NTI/NAE $\log_2$-ratio presumably reflects the export time plus a constant.
We defined genes as co-transcriptionally spliced when the $\log_2$(NTI/NAE) is < -4 (arbitrary threshold), including 1400 genes (~15% of the protein coding transcripts). The $\log_2$ export times of co-transcriptionally spliced genes are normally distributed (Fig.4.11.D). Among genes that are putatively slowly exported, we find two members of the 17β-Hydroxysteroid dehydrogenases (*Hsd17b7* and *Hsd17b12*), and *Fkbp5*, a co-chaperone that modulates glucocorticoid receptor activity. Among genes that are putatively quickly exported, we find several transcription factors such as *Foxo1*, active during the fasted state and inactive in response to insulin during the fest state, *Foxo3*, regulating liver lipid metabolism, or the deacetylase *Sirt1*, which acts as a sensor of cellular NAD+ level and also interacts with the molecular circadian clock system. Preliminary analysis did not reveal any common biological functions, nor enrichment for binding of specific RNA-Binding Proteins.

The comparison of the expression level of nuclear pre-mRNA in both Total and Poly(A) conditions is therefore informative about the extent of co- and post-transcriptional splicing. The splicing regime is strongly influenced by the gene length, and by extension, by the elongation time. This is particularly the case for protein coding transcripts, while differences in splicing and export signatures observed for the other biotypes suggest that regulatory programs may differ.

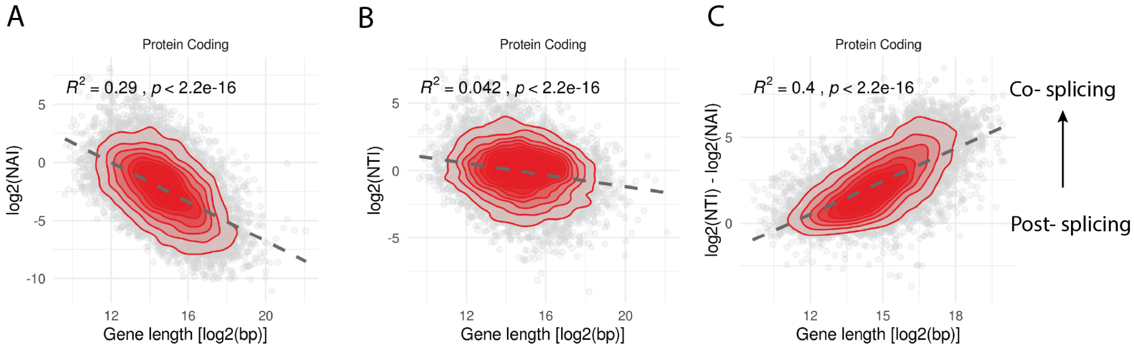


Figure 4.12 – Long protein coding genes are more co-transcriptionally spliced than short genes. A: Correlations of the gene length (in $\log_2$(bp)) with the $\log_2$(NAI) (A), $\log_2$(NTI) (B) or the $\log_2$-ratio of NTI and NAI (C). The R^2 were all < 0.1 for other biotypes, except NMD with a R^2 of 0.17 with NTI/NAE $\log_2$-ratio.

4.6.2 Estimation of the relative variability of the kinetic rates

We finally combined all the variables (NTI, NAI, NAE, CAE) and analysed if the subcellular distribution could be explained by the combination of splicing and export rates. For instance, splicing efficiency has been shown to be the main predictor of the subcellular localisation of lncRNA, so that poorly spliced RNA were mainly found in the nucleus [126]. By comparing log-ratios, we get rid of the influence of the transcription rate (Fig.4.13). Here, we did not observe a correlation between the subcellular localisation ($\log_2$(NAE/CAE) and between the total splicing time (log2_2(NTI/NAE)), the post-transcriptional splicing time (log2_2(NAI/NAE)) nor the proportion of co- versus post-transcriptional splicing frequency ($\log_2$(NAI/NAE)) for any biotype. We suggested in the previous section that cytoplasmic half-life has the largest influence on the $\log_2$-ratio of NAE versus CAE (Fig.4.5, Fig.4.4), therefore, the influence of other regulatory processes could be smaller, explaining the absence of correlation. To estimate the relative contribution of each process (transcription time, splicing time, export time, cytoplasmic half-life), we estimated their variance across all genes (and in log-scale). Under the simplifying assumption that these parameters are all independent, we can estimate the variance of each parameter by computing the matrix of covariance. For example, the matrix of covariance of Fig.4.8.B is computed as follow: cov(NTI, NAE) = cov(($\log_2$(Splicing time) - $\log_2$(Transcription rate), $\log_2$(Export time) - $\log_2$(Transcription rate)). Because of the assumption of independence and of the relationship: $cov(Y,Y) = var(Y)$, the covariance matrix reduces to the variance of the common factor, here, transcription rate. The variance of the transcription rate is 3.5 when comparing NTI to NAE (Fig.4.8.B), and 2.6 when using a different pair (NAE versus CAE, Fig.4.4.A). By computing the covariance matrix of the scatterplots presented in Fig.4.13 (A and B), the estimated variance of cytoplasmic half-life (CAE) is in the same range as transcription time (3.1 in A, 2.4 in B). In comparison, the variance of the export time is smaller: 0.4 when estimated by comparing $\log_2$(NTI/NAE) to $\log_2$(NAE/CAE) as in A, and 0.8 when comparing $\log_2$(NAI/NAE) to $\log_2$(NAE/CAE) as in B. The variance of the total splicing time can also be estimated from the variance of NTI (from A), and is 1.4, while the variance of the post-transcriptional splicing time is in the same range as cytoplasmic half-life and transcription rate (2.8, from B). The estimated variances are different depending on which pair of comparisons is used, but it is difficult to determine whether this inconsistency is due to incorrect assumptions of independence between the parameters. Nevertheless, this analysis suggests that transcription, cytoplasmic degradation times and post-transcriptional splicing times have larger variances relative to the mean compared to the export and splicing times. This supports the hypothesis that variation of the stability of cytoplasmic mRNA transcript explains more the subcellular localisation than the variation of the export step (Fig.4.4.B), and thus, the regulation of the export process does not significantly impact the global distribution of subcellular RNA transcripts.

With the quantification of nuclear pre-mRNA and mRNA, we could investigate some regulatory processes occurring in the nucleus, namely the export and splicing, and additionally separate the global splicing term into co- and post-transcriptional splicing. It is difficult to numerically evaluate the rates, because we analyse the steady-state levels, which always reflects the contribution of two factors. However, how one variable correlates (or not) with another variable, or with genomic features such as gene length reveals specific patterns which often differ between protein coding transcripts and other biotypes. Analysis of the relative variance helps to evaluate which process influences the most the

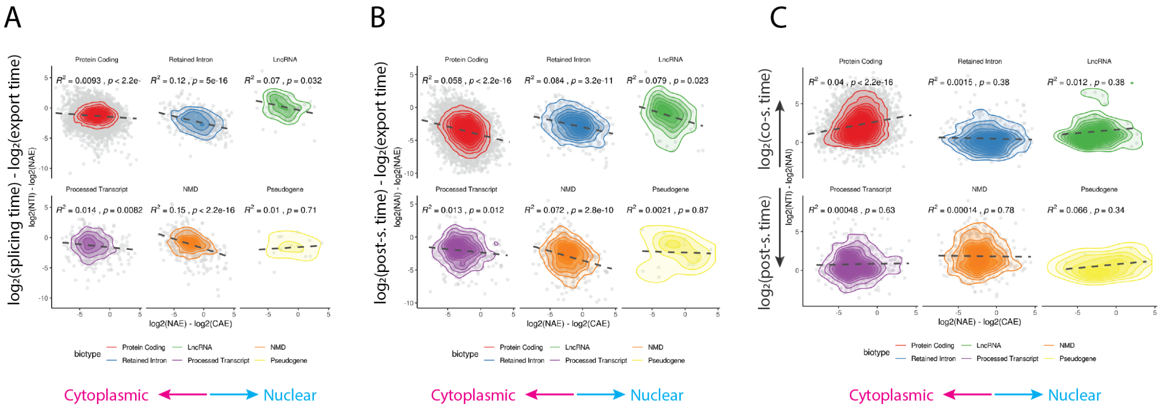


Figure 4.13 – Subcellular distribution of RNA transcripts is not explained by the relationship between splicing and export times. A: Scatterplot showing the relationship between export time and splicing time ($\log_2$(NTI) - $\log_2$(NAE)) versus export time and cytoplasmic half-life ($\log_2$(NAE) - $\log_2$(CAE)) for each biotype. B: same plot as A, but with the ratio of post-transcriptional splicing and export time on the y axis ($\log_2$(NAI/NAE)). C: comparison of the extent of co- versus post- transcriptional splicing ($\log_2$(NTI / NAI) against $\log_2$(NAE) - $\log_2$(CAE)). Number of transcripts per biotype: PC: 8337. RI: 509. LncRNA: 66. PT: 494. NMD: 536. Pseudogenes: 16.

RNA ratios. In the next section, we will use time-series RNA-seq profiles and focus on rhythmic genes in an attempt to quantify these rates and explain the patterns observed in the analysis performed at steady-state.

4.7 Model-based approach to quantify kinetic parameters from time-series profiles

The accumulation of RNA transcripts in any of the subcellular compartments results of the balance between synthesis and decay rates. At steady-state (Fig.4.4, Fig.4.8), it was generally not possible to deduce their respective contributions. However, as our lab showed before, it is sometimes possible to determine these rates from time-series profiles when genes are rhythmic, using a mathematical modelling approach developed by Wang et al.[201]. In essence, by comparing phase delays and relative amplitude of causally related and oscillating RNA species, the approach can infer production and degradation rates. Intuitively, the oscillations of long-lived and stable transcripts tend to dampen and the phase (time of maximum expression level) will be delayed. On the contrary, the oscillations of short-lived transcripts propagate with a minimal loss of amplitude and short phase delay.

In the initial study by Wang et al.[201], the production of mature mRNA was assumed to be equal to the pre-mRNA multiplied by a parameter k that represents the rate of pre-mRNA processing, including splicing and nuclear export. Intronic reads were used to represent pre-mRNA, and exonic reads from the same biological samples used for mRNA. The two kinetic rates regulating the mRNA level were the transcription and degradation rates. The authors were thus able to determine half-lives (log(2)/degradation rate) for thousands of genes in mouse liver. Moreover, they uncovered the contributions of rhythmic post-transcriptional regulation (specifically degradation) in modulating temporal patterns of ~35% of the rhythmic hepatic transcriptome. In that study, all post-transcriptional regulation processes (co- and post- transcriptional splicing, and export) were pooled in a single rate k. Using our fractionated liver cells data, we aim to exploit the quantification of mRNA in different cellular compartments to add more details to the kinetic parameters of mRNA dynamics. More specifically, we want to focus on the nucleocytoplasmic transportation, which was previously overlooked and whose contribution to the 24 hours rhythmic gene expression has not been studied. Therefore, we split the RNA processing into two distinct steps (hereafter referred to as NTI-NAE or step 1, and NAE-CAE or step 2) in order to apply the mathematical model of Wang et al. The first process describes the accumulation of nuclear mRNA (NAE) due to splicing of pre-mRNA NTI and to nuclear export, the second describes the accumulation of mRNA in the cytoplasm (CAE) due to export of NAE and cytoplasmic degradation (Fig.4.14.A). Each step is described by an ordinary differential equation, where the first variable (NTI in step 1, NAE in step 2) and the rate of degradation (export e in step 1, cytoplasmic degradation γ in step 2) are either constant or rhythmic (for more details see 6.7). The combination of constant or rhythmic production and degradation terms generates four different kinetic models (Model 1 to Model 4, Fig.4.14.B). The optimal model is selected by combining a maximum-likelihood approach with the Bayesian information criterion (BIC) to control for model complexity. An arbitrary threshold of 0.6 is set on the BIC weight. For genes with constant levels of pre-mRNA and mRNA (Model 1 or M1), only the ratio between the production and degradation can be determined, similar to the analysis without the temporal component. So the degradation rate γ is structurally non identifiable ([223], see 6.7). The degradation rate γ in Model 2 (rhythmic production, constant degradation) depends on the relationship of the relative amplitudes and the phase shift between the first and second variable (NTI and NAE, or NAE and CAE). The rhythmic pattern of the RNA is described by a cosinor function with a period of 24 hours, and is raised to a power β ranging from 1 to 2 in order to account for the deformed, peaked oscillatory profiles that deviate from a symmetric cosinus function. When β is 1 (simple cosinus), an analytical solution exists (Fig.4.14.C) [79]. The phase delay cannot exceed 6 hours, and the relative amplitude of the second RNA species cannot be higher than the relative amplitude of the first RNA species (the amplitude always dampens) [54]. If the shape of the temporal profile is more peaked ($\beta > 1$), the solution has to be found by numerical integration. Model 3 (M3) and model 4 (M4) include rhythmic degradation, modeled as a simple cosinor with a mean, a relative amplitude, and a phase. For genes in Model 3 (M3, constant production, rhythmic degradation), the kinetic parameters can be particularly difficult to determine: because the first RNA species is constant, there is no phase shift nor difference in amplitude, and the parameters are estimated from the shape of the temporal profile. In model 4 (M4), rhythmic degradation can amplify the relative amplitude of the second RNA species, or shift peak time beyond what is observed with a constant degradation. In all models, the production term (splicing s in step 1, or export e in step 2) is defined as the ratio of the first and second RNA species. Because NTI, NAE, and CAE come from samples processed and sequenced separately, the relative abundance is not defined, therefore, these parameters do not have a biologically interpretable value and will not be discussed.

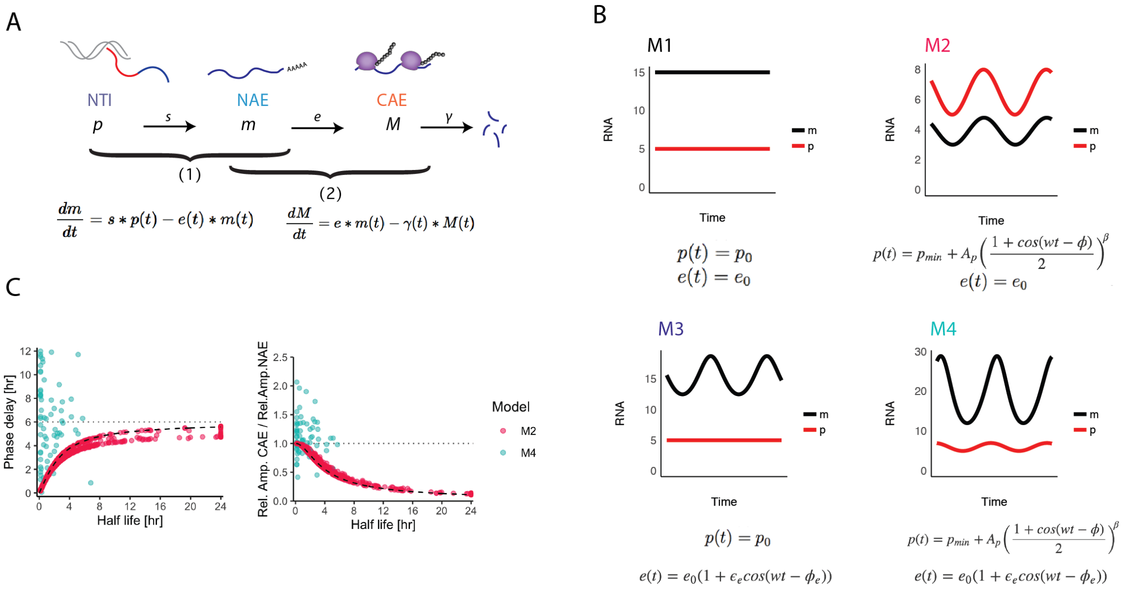


Figure 4.14 – Kinetic model describing the temporal accumulation of pre-mRNA and mRNA. A: RNA processing steps are split in two steps: first, we describe the temporal accumulation of nuclear mRNA in function of splicing rate s and export rate e, and second, the temporal accumulation of cytoplasmic mRNA in function of export rate e and degradation rate γ. B: The combination of constant or rhythmic pre-mRNA (p) and constant or rhythmic degradation generates four models. C: Relationship between the ratio of relative amplitudes, phase delay, and half-life. The data points represent estimated half-lives at step 2 ($\frac{log(2)}{\gamma}$). If the degradation term is constant (M2), and if $\beta = 1$ (simple cosinus), then the analytical solution is: $arctan(\frac{\omega}{\gamma})$ (left panel), and $\frac{\gamma}{\sqrt{\gamma^2+\omega^2}}$ (rightpanel). If β is not equal to 1, the solution is found by numerical integration. If the degradation term is not constant, M4, this relationship is not valid anymore, generating temporal profiles with strongly increased or decreased amplitudes, or large phase shift.

4.8 Estimation of cytoplasmic degradation rates using time-series RNA-seq profiles

We first applied the mathematical model to step 2 (NAE-CAE) in order to estimate cytoplasmic half-lives. We start with step 2 since this concerns cytoplasmic half-lives, and has been addressed before to some extent by Wang et al.[201]. To facilitate identification and take into account biologically plausible ranges, we set the range of possible half-lives from 10 minutes to 24 hours. During the optimisation, the degradation rate γ sometimes reaches our fixed upper or lower limit, and is therefore only identifiable on one side (Left or Right, see Methods: 6.7). Even if the exact value cannot be determined, it informs us about the stability of the gene in the cytoplasm and therefore we decide to keep those genes for discussion (Fig.4.15.B). Out of the 11406 genes expressed in the liver, we confidently classified 1424 genes as having a rhythmic accumulation of nuclear RNA (M2 and M4), of which 79 are additionally rhythmically degraded (M4). The rhythmic profile of 386 genes in the cytoplasm is due solely to the temporal regulation of the cytoplasmic half-life (Fig.4.15.A, M3). Thus, ~15% of expressed genes have a rhythmic temporal profile in at least one cellular compartment. Additionally, 25% are degraded rhythmically, in line with previous studies in the mouse liver (28% in [201], 30% in [79]). The median half-lives were calculated including only values that do not reach the upper or lower boundary, and are 2.45h for M2, 3.88h for M3 and 2.36h for M4. These estimations are consistent with the median of 2.5h estimated by [201] in mouse liver. In general, mRNA half-lives range from minutes to several hours, and the estimation varies depending on the model (cell type, tissue) and method (total RNA-seq, transcription inhibition, metabolic labeling [224]). In NIH3T3 cells, using pulse-chase labeling method, the estimated median mRNA half-life was 9h [225]. In other systems, the median of estimated mRNA half-life was 3.9h in mouse ESCs [226], 4h in HEK293 cells [224], 3.4h in HELA cells [227], and 2.1h in another experiment with NIH3T3 fibroblasts [155]. Our estimations are shorter, probably because our method focuses only on rhythmic genes that are known to be particularly short-lived in order to sustain rhythms [79].

In agreement with the previous study by Wang and colleagues in mouse liver, M4 are more abundantly expressed, and have larger amplitude compared to M2 (Fig.4.15.C). Contrary to what had been previously reported, the distribution of peak times in the cytoplasm differed between the models. Indeed, while M2 phases are more or less distributed throughout the days, we observe two enrichments of phases in M3. The first wave of peak expression is between ZT3 and ZT9, with genes involved in breakdown of proteins such as *Psmd3*, *Psmf1*, *Adrm1*. A second enrichment of M3 peak time is between ZT14 and ZT18, including genes related to mRNA processing (Gene Ontology Term GO:0006397) such as *Srsf1*, *Srsf3*, *Sfpq*, *Thoc7*, and *Hnrnph1*. Preliminary analysis did not reveal any common RNA-Binding Proteins motifs in these groups of genes.

In Fig.4.14.D, we show some representative genes to illustrate the typical behavior of each model. When degradation is constant (M2), the cytoplasmic half-life is determined by the ratio of relative amplitudes and the phase delay. The more stable the transcript, the more the oscillations dampen, and the longer the phase difference between NAE and CAE. Note that the phase delay never exceeds 6 hours (Fig.4.14.C) [54]. Transcripts with a short half-life, have similar temporal profiles in terms of peak and relative amplitude in the nucleus and in the cytoplasm (Fig.4.14.D). On the other hand, a stable transcript such as *Uox*, which has an estimated half-life of 6.1h, has a damped amplitude in the cytoplasm compared to the nucleus ($\log_2$FC of NAE: 1.2 , $\log_2$FC of CAE: 0.5). Incidentally, *Uox* is sometimes classified as non-rhythmic when the transcriptome is analysed at bulk-level (no fractionation nor intron / exon), and its rhythmic transcription is sometimes overlooked in circadian studies [228]. Genes classified in model 3 are rhythmic only in the cytoplasm. *Fus* is a well-known example of a gene rhythmically regulated at the post-transcriptional level [201, 54]. Finally, genes in M4 are rhythmically regulated both at the transcription and degradation level, resulting in more complex patterns, for instance a higher relative amplitude in the cytoplasm (*Ddo*, 7h) or a large phase delay (*Cbs*, 1.4h). Rhythmic degradation can also advance the peak time of CAE, thus fine-tuning the phase of short-lived transcripts [201]. Stability of *Period* genes are post-transcriptionally regulated, as first demonstrated in Drosophila, with *Per* transcript being more stable during the rising phase, and destabilised during the descending phase [158]. Similarly, the three *Periods* genes in mice are targeted by different RBPs, such as the stabiliser hnRNP K and destabiliser by hnRNP D affecting *Per3*, or hnRNP 1 regulating the stability of *Per2* [120, 156]. However, in our dataset, all three *Periods* genes were classified as M2 in all RNA comparisons (NAE-CAE, NIT-NAE, or UTI-UTE). Fitting the pattern with a particularly short half-life was sufficient to explain the temporal patterns (*Per1* and *Per2*: 10 minutes, *Per3*: 30 min).

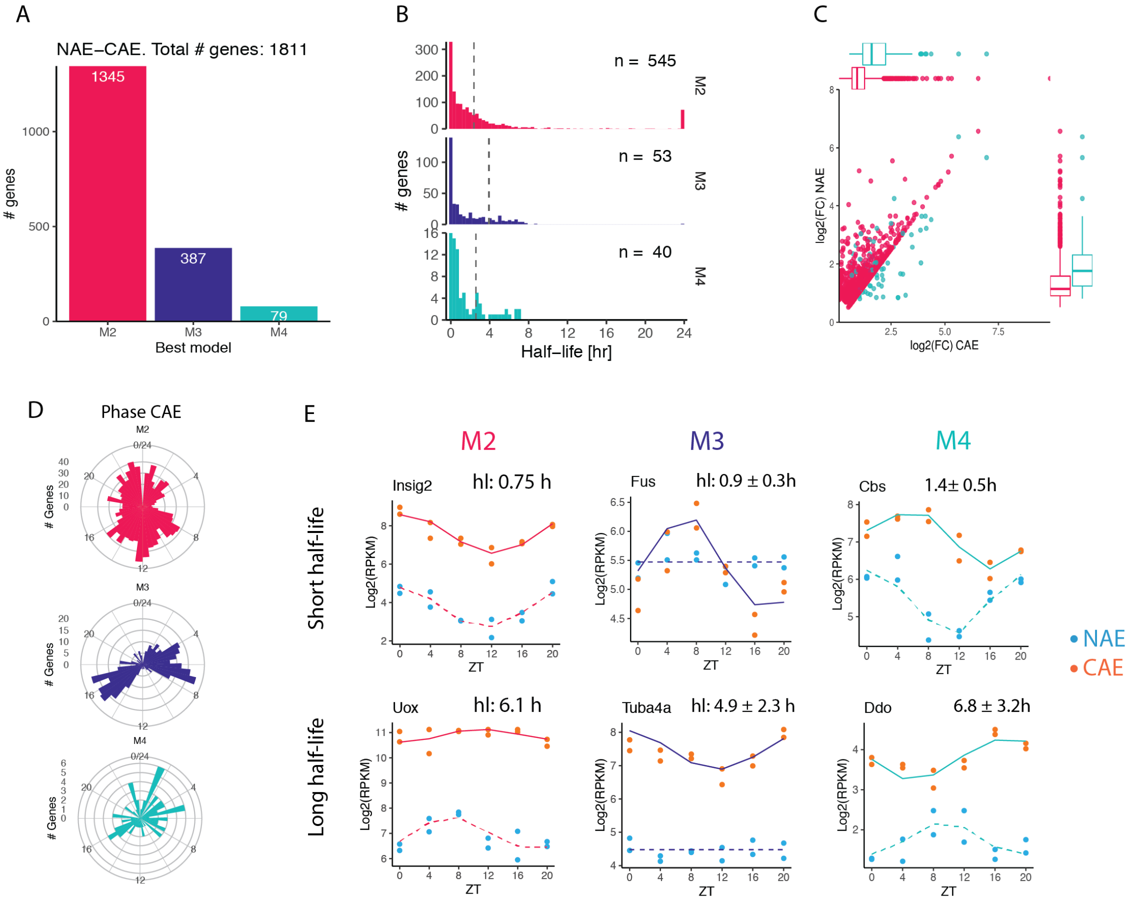


Figure 4.15 – Estimation of cytoplasmic half-life using time-serie RNA-seq profiles. A: Classification of genes in M2 (rhythmic NAE, constant degradation), M3 (constant NAE, rhythmic degradation), or M4 (rhythmic NAE, rhythmic degradation). B: Distribution of cytoplasmic half-lives. The median (dotted line) is calculated excluding half-lives reaching the lower or upper boundary (10 minutes and 24 hours). C: Distribution of $\log_2$FC of NAE and CAE of genes classified in M2 and M4. Note that if degradation is constant (M2), the $\log_2$FC of CAE cannot be higher than of the NAE. D: Circular histogram showing the phases (peak times) of CAE in each model. E: Representative temporal profiles of NAE and CAE of genes in M2, M3 and M4 genes with short half-life (up) or long half-life (bottom).

4.8.1 Identification of distinct dynamic strategies driving nuclear and cytoplasmic mRNA abundances

Our mathematical model uses the relationship of the phases and amplitudes between RNA species in order to estimate the degradation rates. However, the relative expression level of NAE and CAE is not taken into consideration. In order to integrate the relative subcellular abundance in our analysis of rhythms, we clustered the 1345 genes classified as M2 in 8 groups according to the mean absolute expression of CAE and of NAE, of the ratio of NAE and CAE, the $\log_2$ fold-change of NAE and of CAE, and the phase difference between NAE and CAE (Fig.4.16.A). The last three parameters are an indirect estimation of the half-life.

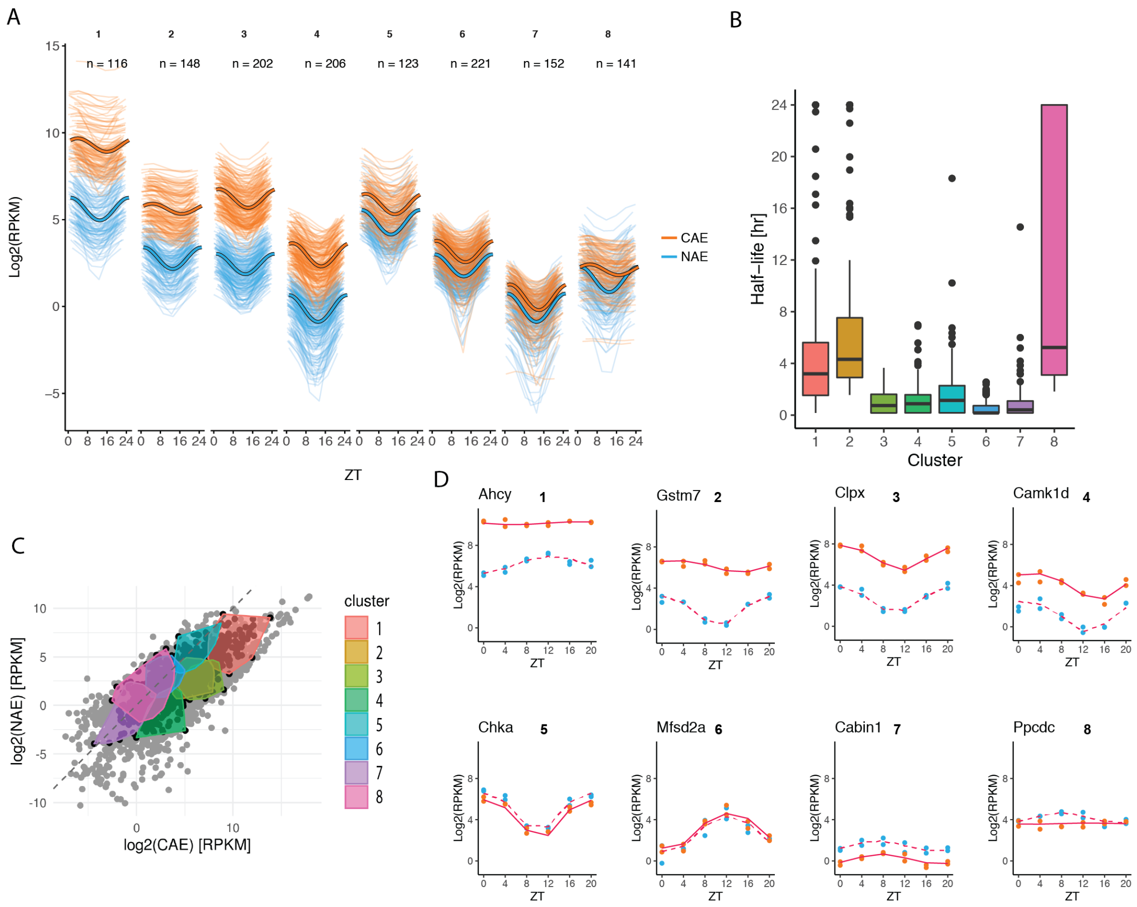


Figure 4.16 – Clustering of M2 genes according to the mean expression level reveals different dynamic strategies to reach the same steady-state level. A: Temporal profiles of NAE and CAE in each of the 8 clusters (hierarchical k-means clustering). The phase of all genes have been aligned to ZT0 (NAE). The profiles drawn with thick lines are computed based on the centers (mean) of each parameter. B: Boxplots of predicted half-lives in each cluster. C: $\log_2$(NAE) versus $\log_2$(CAE), with colored polygons indicating the location of genes in each of the 8 clusters. Grey dots show all the genes, while black dots show only rhythmic genes classified in model 2. D: Temporal RNA-seq profiles fitted by the model of representative genes of each cluster (1 to 8 from top left to bottom right)

Cluster 1 contains genes that are the most abundant in the cytoplasm and in the nucleus. Cytoplasmic temporal profiles are relatively dampened compared to the large amplitude in the nucleus and have a large phase delay, reflecting a long half-life (Fig.4.16.B). Moreover, RNA transcripts are enriched in the cytoplasm compared to the nucleus (median $\log_2$(NAE/CAE is -3.5)). Therefore, the most abundant transcripts tend to be relatively stable compared to other rhythmic genes. A few genes do not match this pattern and still show large cytoplasmic amplitudes ($\log_2$FC > 2) and short half-life such as *Thrsp*, *Upp2*, and *Fgl1* despite their high cytoplasmic abundance. The members of the Inter-Alpha-Trypsin Inhibitor Heavy 1 to 4 are all in cluster 1 (except *Itih5* that is almost not expressed in the liver). While *Ith2* and *Ith4* are mainly cytoplasmic ($\log_2$(NAE/CAE) of -2.6 and -3.9 respectively), *Ith3* is more abundant in the nucleus, despite its slow predicted degradation rate (half-life: 11 hours, $\log_2$(NAE/CAE) is 0.4). However, we showed in a previous section (section 4.5.2) that the nuclear enrichment of *Itih3* is in fact due to the presence of a retained intron (RI) isoform. The RI isoform is short-lived (half-life of 30 min

when estimated at the biotype-level), and does not accumulate in the cytoplasm. On the other hand, the protein coding isoform is long-lived and enriched in the cytoplasm, matching the pattern of genes in cluster 1 ($\log_2$(NAE/CAE) of -2.5, half-life 24h). *Itih1* also appears to be more nuclear compared to the other genes in cluster 1 ($\log_2$ ratio of -1). *Itih1* has no annotated Retained Intron, however, on the genome track, we clearly see an increased density of reads mapping on the intron 19 compared to other intronic regions that are likely better spliced. The presence of the poorly spliced, putative retained intron could also explain why *Itih1* appears to be nuclear-enriched despite the long half-life.
Genes in cluster 2 have a similar pattern to cluster 1 (cytoplasmic enrichment, long half-life), but at a lower expression level. Thus, overall, the relatively long half-life could explain the preferential cytoplasmic localisation for genes in these two clusters.

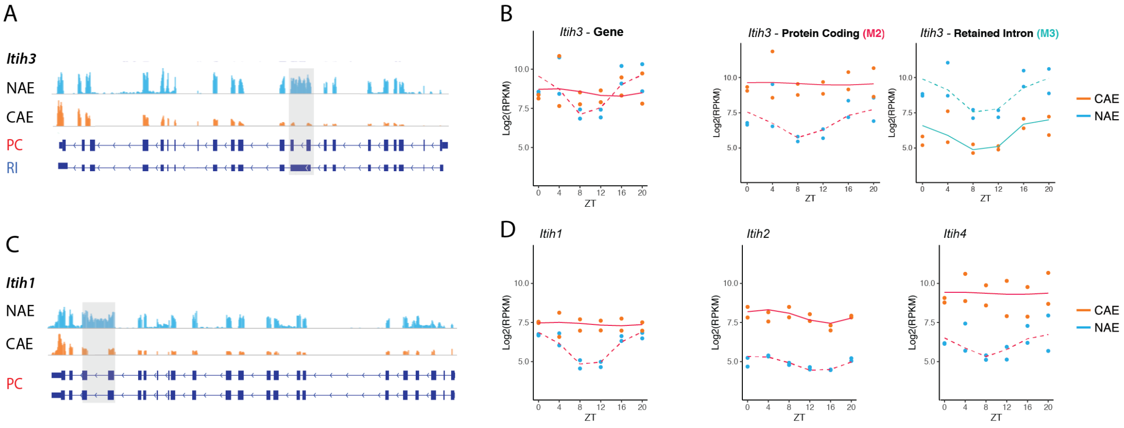


Figure 4.17 – *Itih1* and *Itih3* are enriched in the nucleus, potentially through an intron mechanism. A: Genome track view of *Itih3* of NAE and CAE at ZT16. Grey box indicates the retained intron. Gene length: 15kb isoform: PC Isofom: ENSMUST00000006697, RI isoform: ENSMUST00000227181. B: Fitted RNA-seq profiles of *Itih3* at the gene-level or at the biotype-level. PC isoforms: sum of counts mapping on ENSMUST00000006704 and ENSMUST00000163118. C: Genome track view of *Itih1* of NAE and CAE at ZT16. Grey box indicates the putative retained intron. Gene length: 14.1kb. Genome tracks are not on scale. D: Fitted RNA-seq profiles of *Itih1*, *Itih2* and *Itih4* at the gene-level.

Genes in clusters 3 and 4 are also enriched in the cytoplasm, however, as implied by the large amplitudes of CAE and similar phases between in the nucleus and in the cytoplasm, these genes are short-lived (Fig.4.16.B). In this case, the cytoplasmic localisation cannot be entirely explained by the transcript stability, but instead could result from a particularly efficient export. When we report the mean expression levels of all the clusters on the NAE-CAE space (similar to the scatterplot of Fig.4.4.A), cluster 2 and 3 indeed overlap due to their equivalent expression level. Thus, different combinations of kinetic rates (slow degradation or fast export) lead to similar mean levels at steady-state (Fig.4.16.C). Fast RNA export can be promoted by m^6A methylation, a common post-transcriptional mRNA modification [148]. Based on a published dataset of Methylated RNA immunoprecipitation sequencing (MeRIP-seq) in the mouse liver at ZT13 [229], we compared the number of methylation sites and the size of the peaks, but preliminary analysis did not reveal any difference between clusters.

Genes in clusters 5, 6 and 7 are also predicted to be short-lived (Fig.4.16.B). In contrast to clusters 3 and 4, the NAE/CAE ratio tends toward a nuclear enrichment, so the high turnover rate could explain the cytoplasmic depletion. We typically find core clock genes in these three clusters (Fig.4.18.A, note that *Dbp* and *Nr1d1* were classified as M4). As expected, they are all short-lived (< 2.5h). Moreover, they are more enriched in the nucleus compared to all cycling genes (the median value of $\log_2$(NAE/CAE) of all rhythmic genes is -2.3). The subcellular localisation of RNA transcripts was verified *in situ* by smFISH on the same liver samples used for the RNA-seq (Fig.4.18.D, E). *Bmal1* (also called *Arntl*) has a $\log_2$(NAE/CAE) ratio of -1.44. At the peak (ZT0), there was on average 1.6 times more cytoplasmic mRNA than nuclear mRNA (8.3 cytoplasmic transcript for 5 nuclear transcripts "per nucleus" (see Methods 6.3). In RNA-seq data, the cytoplasmic enrichment is larger (~2.7 times). However, as discussed in section 4.12.3, the nuclear to cytoplasm ratio quantified by smFISH tends to be biased toward higher nuclear values. Because the thickness of the tissue section is 8μm, we sample a higher proportion of the nuclear volume compared to the hepatocyte volume (nuclear diameter: $\sim 9\mu$m, hepatocyte volume: $\sim 7000\mu\text{m}^3$ [185]). Thus, one should always keep in mind that only the relative nuclear to cytoplasmic ratio among different genes can be compared.

The three *Periods* isoforms stand out by their strong nuclear enrichment, despite having similarly short half-lives compared to the other core clock genes, suggesting that they could be potentially slowly exported. We compared *Per1* transcripts localisation by smFISH on the same sample as *Bmal1*. In RNA-seq data, nuclear *Per1* transcripts are 2.6 times more abundant than cytoplasmic transcripts. By smFISH, we found a 1:1 ratio (~5 nuclear and cytoplasmic mRNA in the imaged cytoplasmic and nuclear volumes at ZT12). Despite the different ratios quantified by RNA-seq and smFISH, it is clear that *Per1* RNAs are more nuclear-enriched compared to *Bmal1*. Thus, this suggests that *Periods* RNAs may contain signals (e.g. motifs) or are bound by RBPs that cause them to be more slowly exported compared to other core-clock and cyclic genes. Several genes have similarly short half-life, nuclear enrichment, and high cytoplasmic amplitude as *Periods* genes. These genes, including *Mthfr*, *Chrna*, *Chka*, and *Ciart* (Fig.4.18.C), and might be processed in a similar way to *Periods*. Future work may identify these mechanisms.

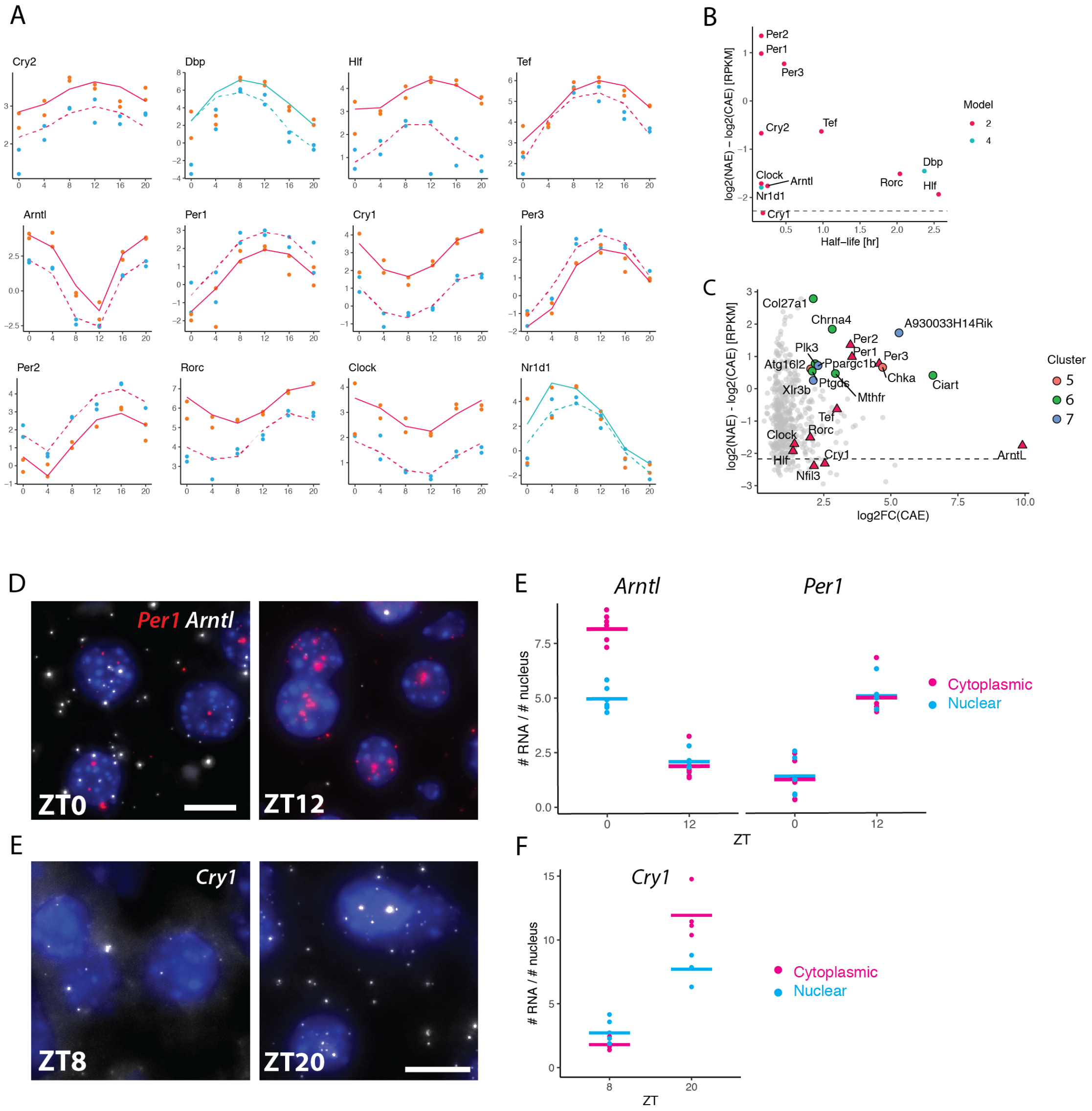


Figure 4.18 – Core clock genes are short-lived, and *Period* genes show nuclear localisation A: RNA-seq temporal profiles of core clock genes and some clock-controlled genes (here ParbZIP family members) in NAE and CAE. B: Estimated half-lives against the log$_2$(NAE/CAE) ratio. As expected, clock genes are all short-lived and more nuclear compared to the median ratio of all rhythmic genes (median indicated by dashed horizontal line). C: log$_2$(NAE/CAE) against the amplitude in the cytoplasm of all the genes in cluster 5, 6 and 7. Core clock genes are marked with triangles. Genes with a similar pattern to *Periods* genes are labelled with a dot. These genes have a high cytoplasmic amplitude (log$_2$FC CAE >2) and are enriched in the nucleus. D: smFISH of *Per1*, *Bmal1* (also named *Arntl*) at their respective peak and trough. Scale bar: 10μm. E: Quantification of mRNA transcripts on smFISH images. 5-6 images were taken per animal (technical replicates) and the average number of mRNA molecules per number of nuclei per image is represented by a point. The mean is represented by the bar. Because n=1 per time point, no error bar is shown.

Finally, cluster 8 contains genes with a rather flat cytoplasmic profile, suggesting a long half-life. Contrary to clusters 1 and 2, transcripts are not particularly enriched in the cytoplasm. Therefore, the observed NAE/CAE could result from a slow export. However, because of the overall low expression, we cannot exclude that noise hides cytoplasmic rhythms. Still, some genes have reliable patterns, such as *Ppcdc* (Fig.4.16.D), and could be candidates for being regulated by a slow export.

The combined analysis of rhythms, absolute and relative levels, and half-lives pointed toward genes that may be rapidly exported (clusters 3 and 4), rapidly degraded in the cytoplasm (cluster 5-7), or slowly exported from the nucleus (cluster 8). Genes with similar cytoplasmic expression levels are controlled by distinct combinations of rates. Among core clock genes in clusters 5 to 7, the three *Periods* isoforms stood out as potentially slowly exported. In the next section, we will provide an independent analysis of nuclear export times applying the same mathematical framework on NTI and NAE RNA samples.

4.9 Estimation of nuclear export rates rates using time-series RNA-seq profiles

The life cycle of an RNA transcript starts in the nucleus, where it undergoes various processing steps. Introns are co- and post-transcriptionally spliced, a m^7G-cap is added to the 5' end, a Poly(A) tail is added to the 3'end, and the transcript is coated with RNA-binding proteins, forming export competent ribonucleoproteins [230]. The combination of these processes ultimately modulates the time spent by an RNA in the nucleus. Moreover, any of these steps can in principle be regulated in a circadian manner. In order to estimate the nuclear lifetimes and to uncover potentially rhythmic patterns generated by rhythmic nuclear export, we applied the mathematical model used above and compared the rhythmic profiles of nuclear pre-mRNA (NTI) and nuclear mRNA (NAE). Here, the export time refers to the time spent by a fully transcribed and spliced polyadenylated transcript in the nucleus. Note that therefore, nuclear degradation, although not specifically included in the model, acts to increase the effective export rate.

We confidently classified 1632 genes in one of the three models with rhythmic NAE (BIC <0.6, Fig.4.19). The vast majority (1428) were rhythmically transcribed and exported with a constant rate (M2). 13% were exported rhythmically (M3 + M4, Fig.4.20), suggesting that nuclear export is a post-transcriptional mechanism contributing to modulation and generation of rhythmic profiles. In Fig.4.20 we show some representative profiles of genes classified as M3, whose rhythmic accumulations of nuclear mRNA are only attributed to rhythmic mRNA export. Proteins encoded by these genes cover a large variety of biological functions, such as deacetylation (*Lypla2*, *Sirt5*), nuclear speckles assembly and splicing (*Srrm2*, *Akap8l*), hypermethylation of sn(o)RNA m^7G cap (*Tgs1*), or NAD(P)H)-dependent oxidoreduction (*Rdh13*). However, Gene Ontology Terms enrichment analysis did not reveal any common biological function for M3 genes.

The nuclear rhythms generated by rhythmic export propagate until the cytoplasm. However, the median $\log_2$FC of CAE (cytoplasmic mRNA) is lower than those originating from M2 or M4 genes (median $\log_2$FC: 1.1 (M2), 0.76 (M3), 1.5 (M4), data not shown). While some oscillations are completely lost in the cytoplasm because of the long stability of cytoplasmic mRNA, others keep oscillating with relatively large amplitudes (Fig.4.21).

The distribution of NAE phases in M3 shows an enrichment around ZT8 (Fig.4.19.C, 33% of the M3 genes are in the time window between ZT6 and ZT10). By contrast, peak times of M2 genes, whose rhythms are generated by rhythmic transcription, show the typical bimodal distribution of phases peaking at the end of the light phase (ZT22) and the dark phase (ZT10), reflecting the two “waves” of transcriptional activity in the liver [54, 69]. Phases of M4 genes, whose nuclear accumulation is also regulated by rhythmic export, are distributed throughout the day. The phase of shortest export time (highest rate) is enriched in the second part of the night phase (half of the M4 genes have the highest export rate between ZT18 to ZT 24/0). The difference between the peak of export rate and peak of transcription rate modulates the resulting pattern of NAE: temporal profiles of metabolic transcripts such as *Pck1*, *Lpin1* or *Ppard* are amplified by antiphasic rhythmic export, while genes such as *Sqle* or *Lipc* are dampened while maintaining a phase coherence because export peak time matches transcription peak time (Fig.4.19.D). Thus, the temporal profiles of several key metabolic enzymes are boosted by regulating the export process in a timely manner.

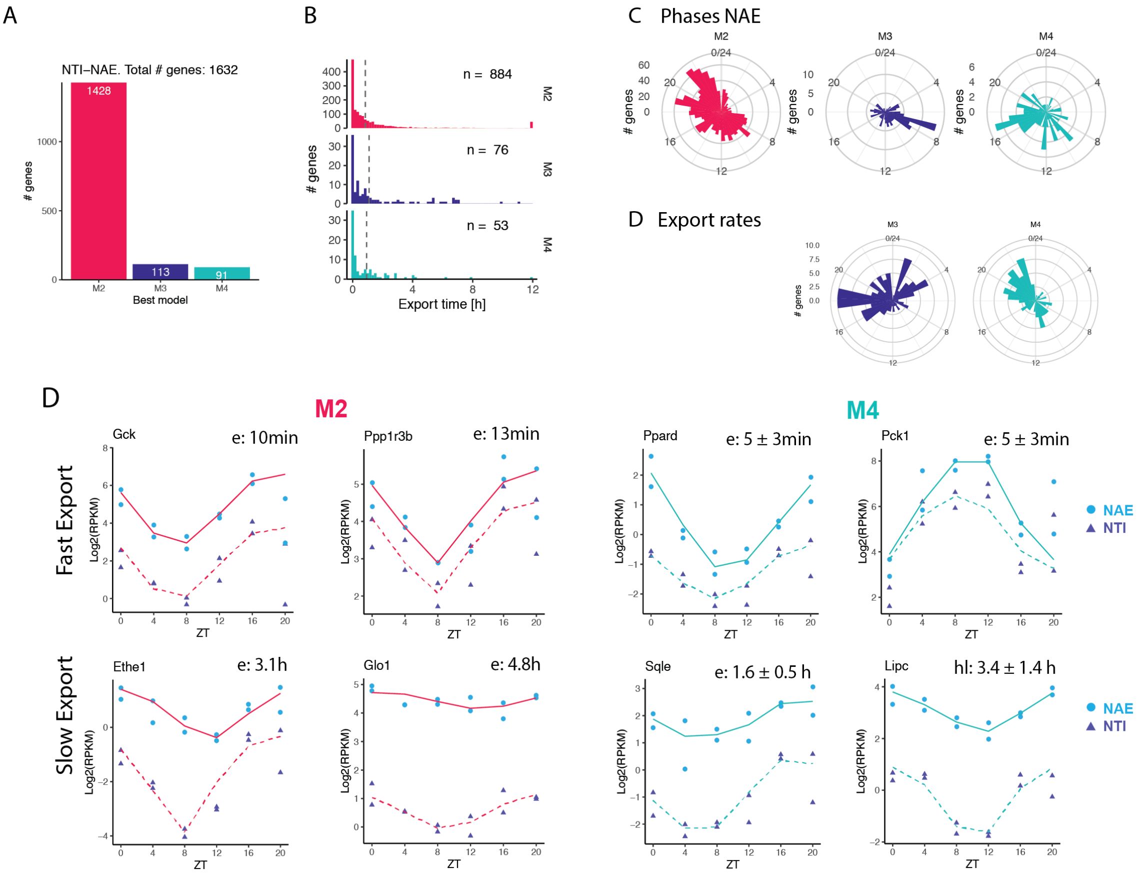


Figure 4.19 – Estimation of export times using time-series RNA-seq profiles. A: Classification of genes in M2 (rhythmic transcription, constant export), M3 (constant transcription, rhythmic export), or M4 (rhythmic transcription, constant export). B: Distribution of estimated nuclear lifetimes (export times) of the three models. The median (dotted line) is calculated excluding export times reaching the lower or upper boundary (5 minutes and 12 hours). C: Distribution of peak times of rhythmic nuclear mRNA (NAE). D: Distribution of peak time of rhythmic export rates (shortest export times) of M3 and M4 (in M2, export is constant). D: Representative temporal profiles of nuclear pre-mRNA (NTI) and nuclear mRNA (NAE) in M2, M3 and M4 genes with short export time (up) or long export time (bottom).

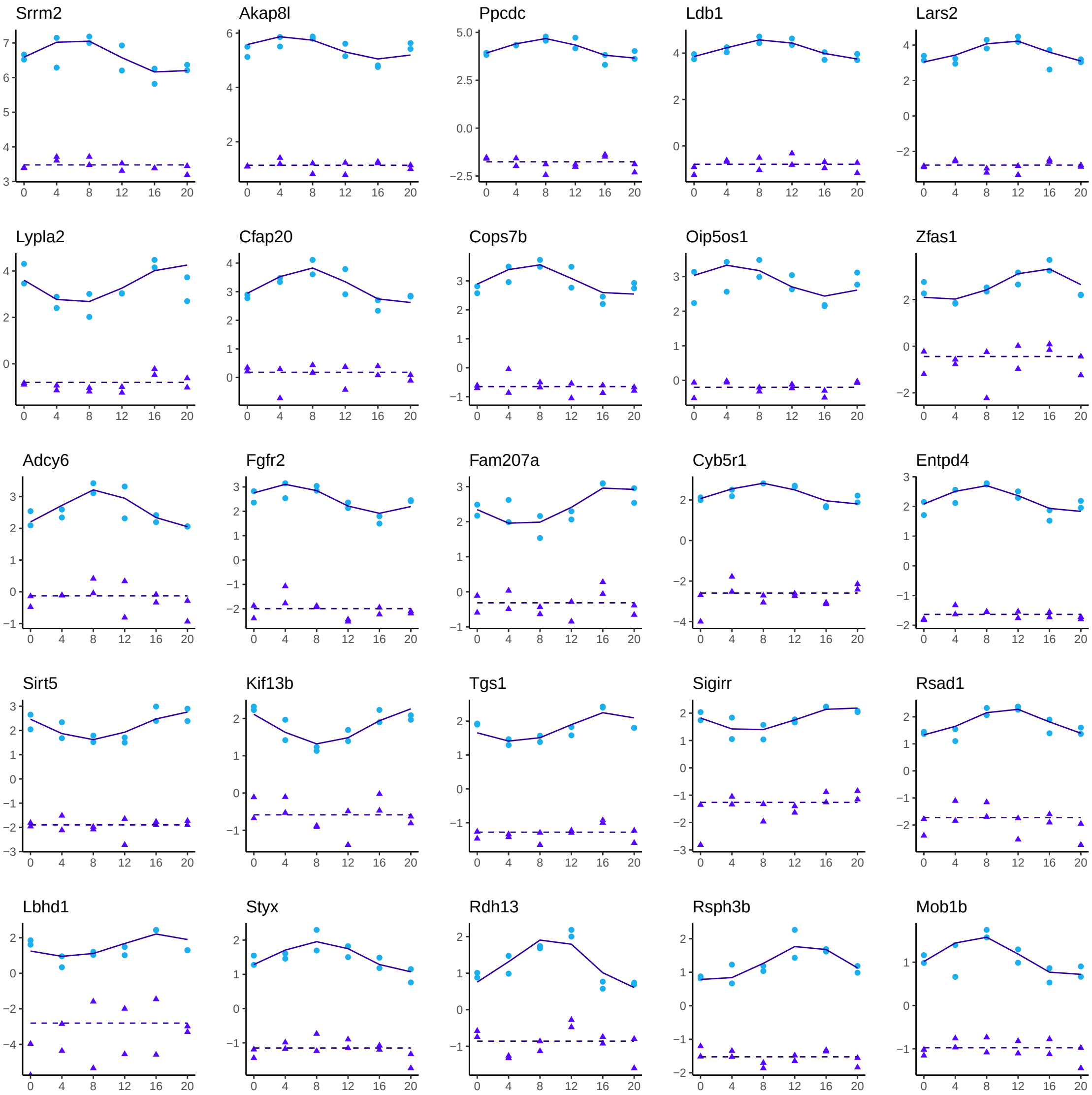


Figure 4.20 – Representative temporal profiles of NTI-NAE genes classified as M3 (constant transcription, rhythmic degradation). All the genes all have a $\log_2$FC > 0.8. Genes were all classified as M2 or M4 in the analysis of step 2 (NAE vs CAE). Plots are arranged by expression level of NAE (highest NAE *Srrm2* in the upper left corner). Light blue dots and solid line: NAE, purple dots and dotted line: NTI.

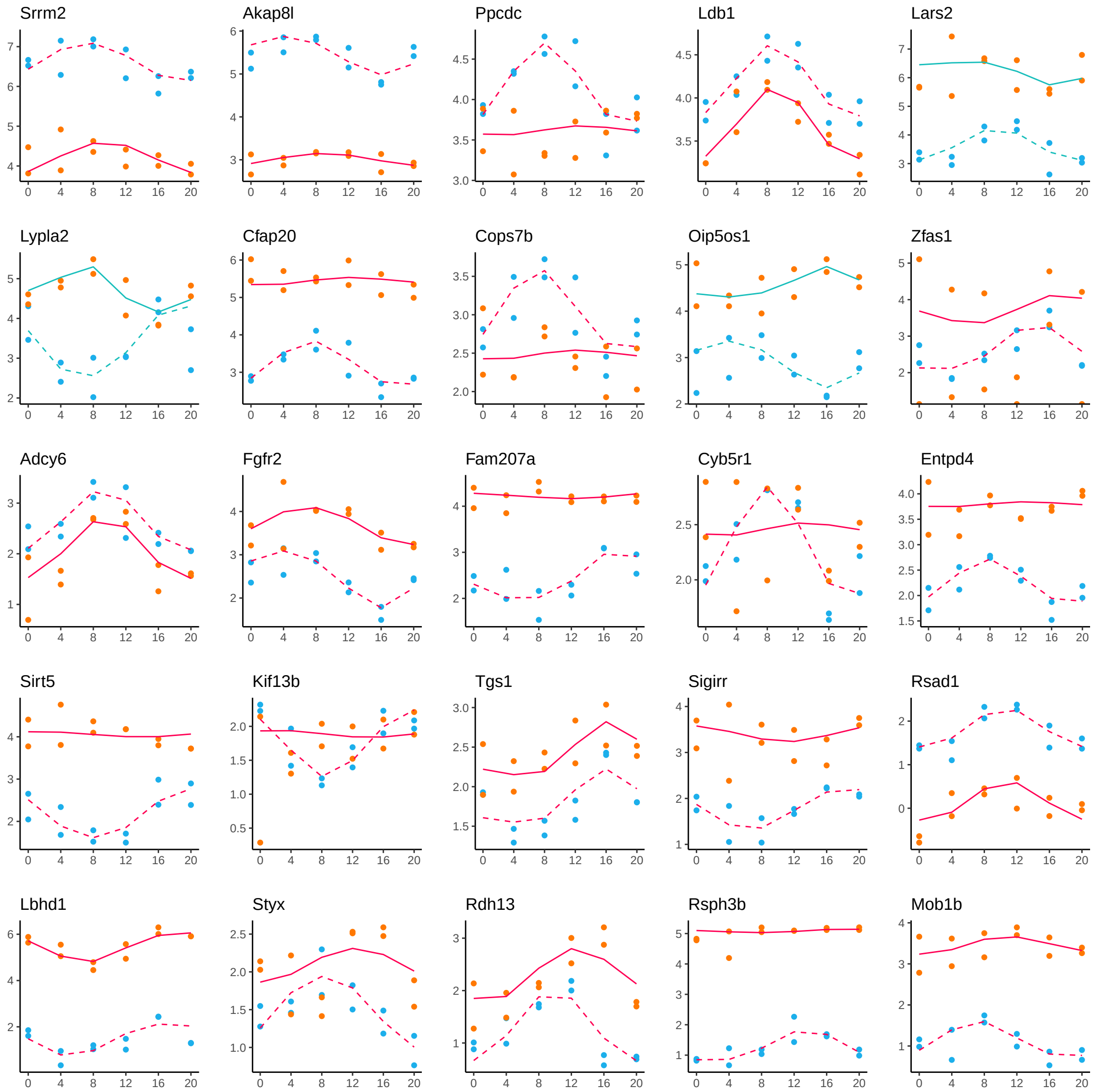


Figure 4.21 – Representative temporal profiles of NAE and CAE of genes classified as M3 in step 1 (constant transcription, rhythmic degradation). Plots are arranged by expression level of NAE (highest NAE *Srrm2* in the upper left corner). Light blue dots and dashed line: NAE, orange dots and solid line: CAE. All the genes were classified as M2 in step 2 (constant degradation), except *Lars2* and *Lypla2* that were classified as M4 (rhythmic degradation)

From the literature, the nuclear residence time is usually shorter compared to the cytoplasmic half-life. Indeed, previous studies estimated that nuclear lifetime ranges from 5 minutes to less than a couple of hours (maximal estimated nuclear residence time: 40 minutes in [116], 90 min in [127], 2 hours in [196]). In our model of step 1, we set the range of possible export times from 5 minutes to 12 hours. More than one third of the genes in our dataset reaches the lower boundary of 5 minutes (600 genes), suggesting that indeed, rhythmic transcripts are usually exported on a short time scale. The median export time, if we include those genes, is 25 minutes for M2 and M3 genes, and 12 minutes for M4 genes. If we exclude export times reaching the boundaries (lower and upper limits), the median of estimated export time is about 50 to 60 minutes in all three models (Fig.4.19.B).

We further stratified the analysis of the 1438 genes classified as M2 by clustering genes in function of their temporal profiles (phases, relative amplitudes) and mean expression levels, similar to the clustering in Fig.4.16. This allows a better identification of long-lived genes from rapidly exported genes (Fig.4.22).
Genes in cluster 5 and 6 have the longest retention time in the nucleus (shortest export time: 1h and 1.6h, median: 2.3h and 4.6 h). As a consequence, the nuclear mRNA amplitude is strongly reduced, such that a majority of these transcripts are no longer considered rhythmic in step 2, and are classified as M1 (Fig.4.22). We looked for functional enrichments (GO Terms) in cluster 5 and 6 (potentially retained in the nucleus) using the rhythmic liver transcriptome as the background. A relevant metabolic function enriched in cluster 5 is related to sterol metabolism (hormone biosynthesis GO:0042446) with *Srd5a1*, *Hsd17b2* and *Dhcr7*, and SREBP cleavage-activating protein *Scap*, and all have relatively long nuclear lifetime between 2h and 3.5h. However, major genes involved in the directly related function of cholesterol synthesis (GO:0042632) are processed extremely rapidly in the nucleus (*Insig1*, *Cyp7a1*, *Hmgcr*, export time of 5 minutes). This suggests that despite being rhythmically transcribed, biologically related genes have distinct dynamic patterns in the nucleus. We additionally noticed that genes involved in fructose metabolism (*Aldob*: 6.8h, *Khk*: 4.8h, *Sord*: 1.4h, *Fbp1*: 6h) were also particularly long-lived, but were classified in cluster 1 due to their high expression level.

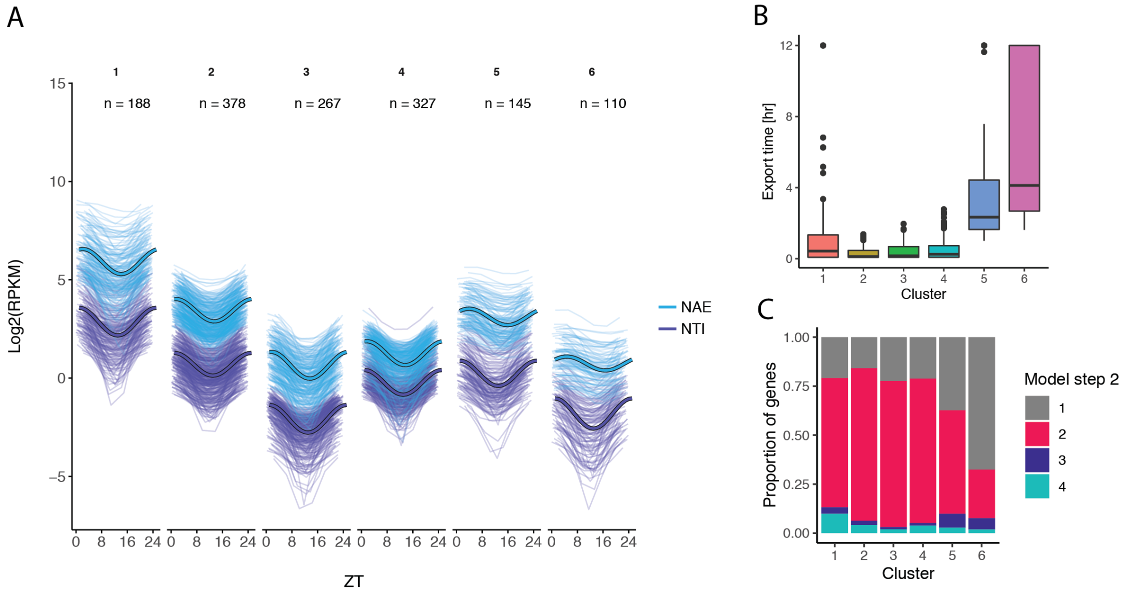


Figure 4.22 – A: Traces of the temporal profiles of NTI and NAE grouped in 6 clusters based on the mean expression of NTI and NAE, their relative expression level, the phase delay, and the log$_2$FC (hierarchical k-means clustering). The phase of all genes have been aligned to ZT0 (NTI). The profiles drawn with thick lines are computed based on the centers (mean) of each parameter. B: Boxplots of predicted export time in each cluster. C: Classification of genes in each cluster in step (NAE-CAE). Due to the long export time, the amplitudes in cluster 6 are damped and are not detected in step 2 anymore, resulting in classification as M1 in step 2.

Clock genes are typically quickly exported, and are therefore in cluster 2, 3 and 4. Nuclear residence times range from 5 minutes to 1.1 hour (longest: *Clock*. Note that *Nr1d1* and *Dbp* are classified in M4, and the NAE relative amplitude is amplified, see Fig.4.23). The three *Periods* mRNA were particularly enriched in the nucleus, and we hypothesised in the previous section that they could have (relatively) long nuclear residence times (Fig.4.18). Here, all three *Period* genes are classified as M2, indicating that their export time is not gated in function of time of day. In the analysis, we find that *Per1* has an export time of 12 min, while *Per2* and *Per3* are more retained, with a nuclear half-life of 43 min and 39 min respectively.

In Fig.4.16.D, we showed a group of genes that were enriched in the nucleus despite their long cytoplasmic half-lives (cluster 8). We hypothesised that those mRNAs might be slowly exported, explaining the relative nuclear abundance. However, the median export time of genes in cluster 8 was only 15 minutes (median computed using only genes classified as M2 in step 1). Genes in cluster 5, 6 and 7, which include the clock genes, were equivalently quickly exported (medians of 12, 8 and 15 minutes, respectively). Median values of clusters 1, 2, 3 and 4 are respectively 21 min, 25 min, 28 min (longest export time) and 10 min (shortest export time). Our mathematical model has a limited resolution in the estimation of export times, and therefore might not be able to discriminate between different dynamic strategies when involving particularly rapidly exported and degraded transcripts. Nevertheless, we were able to uncover rhythmic patterns in the nucleus, which strongly suggests that the nuclear lifetime varies around the day for ~ 10% of the rhythmic nuclear mRNA.

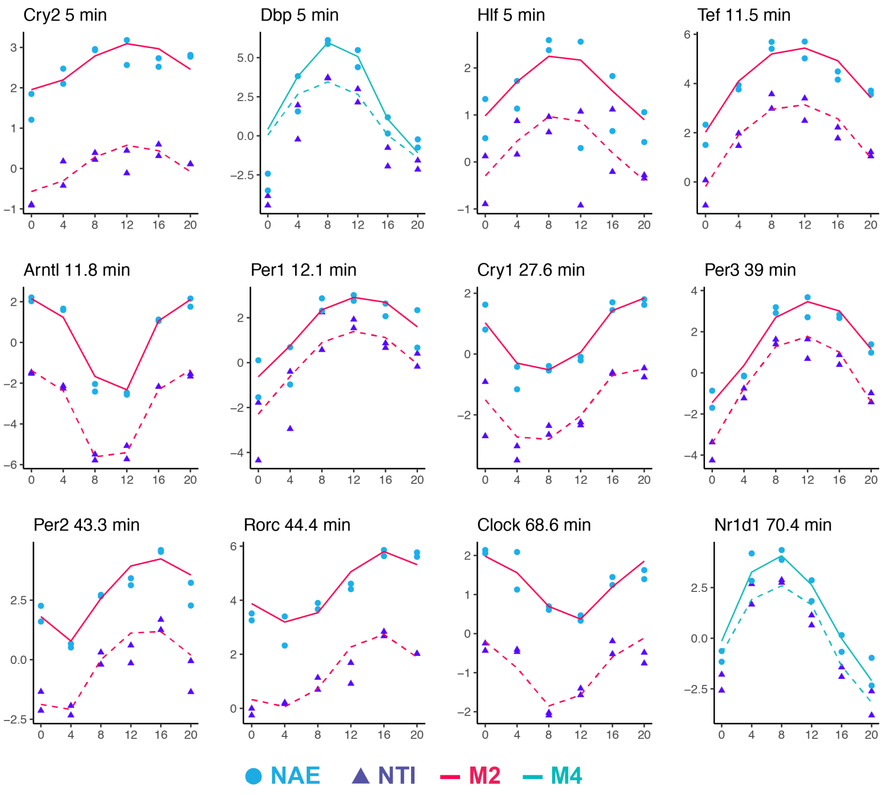


Figure 4.23 – Fitted temporal profiles of clock genes, with their corresponding export time. All the genes are classified as M2, except *Dbp* and *Nr1d1*, whose amplitudes are boosted with a rhythmic export rate.

4.10 Relationship between the kinetic rates of mRNA

4.10.1 No global coordination of the export rates and cytoplasmic degradation rates

So far, we have investigated the two steps of the RNA life cycle independently: the pre-mRNA (NTI) is spliced to produce a mature and polyadenylated mRNA (NAE), which resides in the nucleus for a certain time until being exported (step1), and once in the cytoplasm (CAE), is degraded at a specific rate γ (step 2). We now explore whether these two rates are related to each other, to get a global picture of the entire process.

Globally, we find virtually no correlation between the nuclear export time and cytoplasmic half-life for rhythmically transcribed genes (M2 in both step 1 and step 2). One situation where the two processes could be coordinated is for rhythmic genes transcribed with high amplitudes: in order to propagate rhythmic temporal patterns without major loss between cellular compartments, the transcript might be processed efficiently from the transcription site in the nucleus to the cytoplasm. For instance, all the clock genes are exported within 1 hour and have short cytoplasmic half-lives of less than one hour, except for *Rorγ* and *Hlf* (2h and 2.5h respectively) (*Arntl*, *Clock*, *Cry1*, *Cry2*, *Per1*, *Per2*, *Per2*, *Nr1d1*, *Nr1d2*, *Rorc*, and PARbZip family members *Dbp*, *Tef*, *Hlf*). However, there was no notable relationship between the export rate or the degradation rate with the amplitude nor with the relative amplitude of NTI, NAE or CAE, which suggests that the control mechanism that sets the relative amplitude could be largely independent of the export and degradation rate, and is therefore predominantly controlled at the transcriptional level (data not shown). We further compared the processing rates with the rhythmicity of their encoded proteins. We used the dataset published by Wang and Mauvoisin et al. [231], who assessed the rhythmicity of ~5000 nuclear proteins in mouse liver. While there was no significant difference between export rates of transcripts coding for cycling or constant nuclear proteins, those coding for proteins displaying a circadian pattern in the nucleus were overall degraded faster in the cytoplasm (Fig.4.24.B,C). Thus, our data did not reveal a coordination of export and degradation rates neither globally, nor specifically for high-amplitude rhythmic transcripts, although the proteomic data revealed a systematic decrease in the cytoplasmic half-life for proteins with nuclear oscillations.

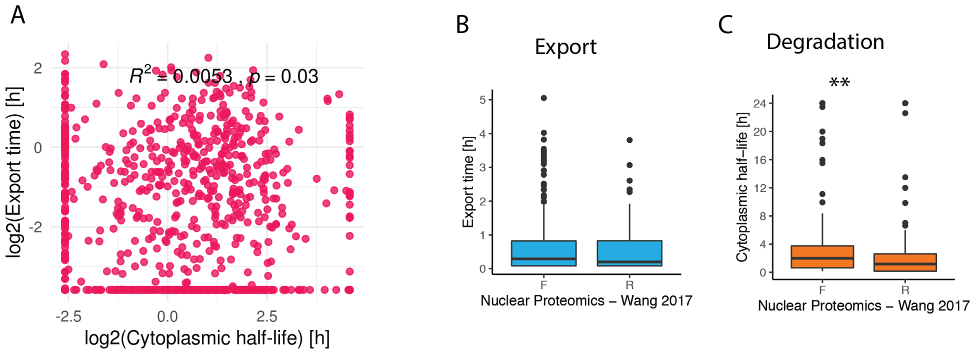


Figure 4.24 – A: Scatterplot shows no correlation between export time and cytoplasmic degradation half-life. Only genes classified as M2 in both step 1 (NTI-NAE) and step 2 (NAE-CAE) are shown. R^2 when excluding estimated parameters reaching the upper and lower boundaries: 0.0038, p-value = 0.17. B: Boxplot of export times estimated from step 1 of genes whose corresponding proteins were shown to be cycling (R) or flat (F) in nuclear fraction of mouse liver (Wang, Mauvoisin et.al 2017 [231]). The two populations of export times were not significantly different (Mann-Whitney U test, p-value: 0.56). C: Boxplots of cytoplasmic half-lives estimated from step 2. The cytoplasmic half-lives of transcripts with cycling protein (R) are significantly shorter than those who are not cycling (p-value: 0.003). Number of R genes = 147, F genes = 253.

4.10.2 Cellular fractionation reveals regulatory processing steps hidden in bulk analysis

In the present study, we fractionated liver cells and we estimated separately the nuclear and cytoplasmic lifetimes. In principle, their sum represents the total lifetime of the RNA transcript. We thus compare the sum of the export and the degradation times (“combined”) to the total lifetime estimated using the unfractionated sample ("total", using pre-mRNA UTI against mRNA UTE). When comparing UTI to UTE, as originally done in [201], the estimated degradation rate γ encompasses all the post-transcriptional processes from splicing, nuclear export, and cytoplasmic degradation.

We observed a significant positive correlation between the combined lifetimes with the total lifetime (Pearson’s correlation ρ = 0.54, Fig.4.25.A). The combined lifetime of 60% of the genes are within a margin of error of a factor 2 compared to the total lifetime (341 / 561 genes, genes classified as M2 in all three comparisons). Some discrepancies arise when the cytoplasmic RNA level is relatively low compared to the nuclear RNA. As shown with the example of *Pnpla6* (Fig.4.25.B), the rhythmic temporal profile detected in UTE mainly reflects the oscillations of nuclear mRNA, while the low abundance and low amplitude cytoplasmic profile makes almost no contribution to the overall signal. In this case, the export time is 50 minutes and the cytoplasmic half-life is 24 hours, while the total lifetime estimated by UTI-UTE is 30 minutes, and thus, at the bulk level, we predominantly measure the export time, and not the sum of the two rates. On the other hand, the same scenario (short export time, long cytoplasmic half-life) is well captured by the unfractionated samples if CAE level is relatively high. For example, *Slco1a4* is exported very quickly (5 min), and is long-lived in the cytoplasm (4.5h, Fig.4.25.C). At the bulk level (UTI-UTE), the estimated total lifetime is 4.6h, accurately representing the total lifetime of the RNA transcript. We verified whether the total lifetime was only reflecting the cytoplasmic half-life due to the overall higher abundance of CAE, or if the nuclear retention time was also contributing to the total lifetime. The cytoplasmic half-life alone explains 20% of the total lifetime, while nuclear half-life explains 15% (Fig.4.25.D, E). Thus, even if the correlation of the total lifetime is stronger with the cytoplasmic lifetime than with nuclear lifetime, the nuclear residence time significantly contributes to the total lifetime.

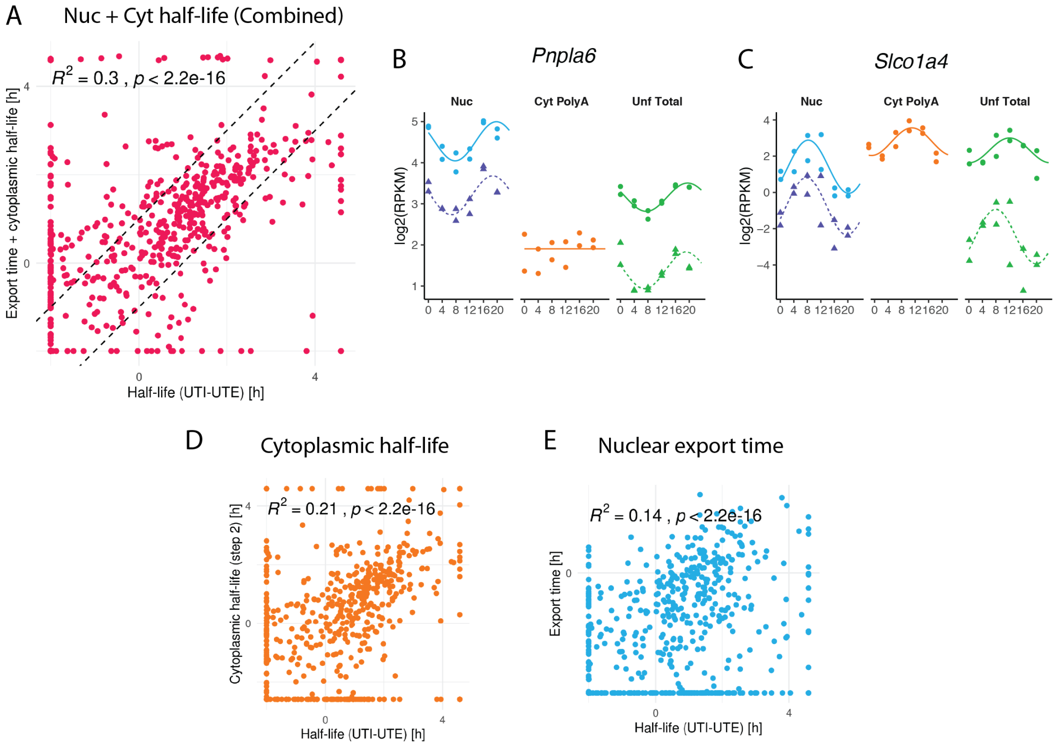


Figure 4.25 – The RNA lifetime estimated from unfractionated RNA population is the sum of nuclear and cytoplasmic lifetimes. A: Correlation between the total half-life estimated by comparing the temporal profiles of UTI and UTE against the sum of the export time estimated from step 1 and the cytoplasmic degradation time estimated from step 2, in $\log_2$. Only genes classified as M2 in all three comparisons are used (number of genes = 561). If excluding genes reaching the upper and lower boundaries: $R^2 = 0.44$. 60% of the genes are between the two dashed lines (intercept at +1 and -1). B: Temporal RNA-seq profiles of *Pnpla6* of NTI (dashed line), NAE (solid line), CAE, UTI (dashed) and UTE (solid green). The temporal profiles of UTI and UTE most likely reflect the temporal profiles of nuclear RNA, but not CAE. Thus, the total estimated lifetime (30 minutes) does not match the sum of nuclear export time (50 minutes) plus the cytoplasmic half-life (24h). C: Temporal RNA-seq profiles of *Slco1a4*. The sum of the nuclear export imte (5 minutes) plus the cytoplasmic half-life (4.5h) matches to the total half-life estimated by UTI-UTE (4.6h). D and E: Scatterplots of the total estimated half-life (UTI-UTE) against cytoplasmic half-life estimated in step 2 (B) or nuclear export time from step 1 (C).

We next sought to determine the relative contribution of the nuclear and cytoplasmic lifetime to the total lifetime. We looked at the relative difference between the nuclear and cytoplasmic lifetime in function of the total lifetime, including only genes whose combined times matched the total lifetime by at most a factor 2 (Fig.4.26.A, n = 341). Overall, RNA transcripts spend more time in the cytoplasm than in the nucleus (median of the $\log_2$(nuclear/cytoplasmic lifetime): -1.3). Out of the 561 genes, 119 (20%) have a longer export time than cytoplasmic degradation time. Focussing on long-lived transcripts revealed different combinations of rates. Genes such as *Acox2*, *Uox* and *Ahcy* are rapidly exported (within 15 minutes), and their long total lifetime is predominantly determined by their cytoplasmic stability (6.1h, 11.1h and 13.5h), and thus, nuclear export is not a limiting step. Some genes spend approximately the same amount of time in both compartments, resulting in an equal contribution to the long total lifetime. For example *Gckr*, which regulates the activity of Glucokinase by sequestering the enzyme in the nucleus when glucose level is low, has a nuclear lifetime of 2.9h and a cytoplasmic lifetime of 3.2h (total: 7.3h). *Dio1*, which catalyses the conversion of the inactive thyroid hormone T4 to its active T3 form, also spends an equal amount of time in the nucleus and in the cytoplasm (2.5h, cytoplasmic: 2.5h, total: 4h). Another protein related to thyroid hormone, *Thrsp* (thyroid hormone responsive), has similar nuclear and cytoplasmic lifetime (1.4h in the nucleus, 1h in the cytoplasm, total: 1.9h). *Abcb11*, the bile salt export pump, also shows a similar pattern (nuclear lifetime: 2.6h, cytoplasmic: 2.4h, total: 6.4h). The peak time of these transcripts in the cytoplasm is thus delayed because of the relatively long nuclear retention time. On the rightmost side of the x-axis, where transcripts spend more time in the nucleus than in the cytoplasm, we show as examples *Cutal* (nuclear lifetime: 2.5h, cytoplasmic: 10min), *Gramd4* (nuclear lifetime: 1.2h, cytoplasmic: 12min), or the circadian gene *Clock* (nuclear lifetime: 1.14h, cytoplasmic: 10min). These genes are particularly unstable in the cytoplasm, which results in very similar temporal profiles in both the nucleus and in the cytoplasm. In this case, the total lifetime is predominantly determined by the nuclear export time.

In our model, the nuclear residence time cannot fully explain the longest total lifetime because we set the maximal export time to 12 hours. Moreover, when the transcript is retained in the nucleus for a long period of time, the amplitude of the oscillations of NAE is reduced to a point that the gene can no longer be classified as "M2" in step 2, but is instead classified as M1 (constant nuclear mRNA). Here, the longest nuclear retention time that still allows rhythms to propagate into the cytoplasm is 3.8h if we consider only genes whose combined lifetime matches well the total lifetime estimated, like *Pltp* (nuclear lifetime: 3.8h, cytoplasmic: 1.1h, total: 5.6h) or *Hsd17b2* (nuclear lifetime: 3.5h, cytoplasmic: 0.9h, total: 4.9h). Of note, *Hsd17b2* was indeed shown as a potentially retained transcript in the analysis of NTI - NAE at steady (Fig.4.11). If we include genes that are less well estimated, a nuclear retention of up to 5 hours still generate rhythmic profiles that are detectable at step 2, for instance *Pdia3* (nuclear lifetime: 5.1h, cytoplasmic: 10min, total: 2.5h), *Fdft1* (nuclear lifetime: 4.8h, cytoplasmic: 2.0h, total: 13.8h) or *Oat* (nuclear lifetime: 4.0h, cytoplasmic: 1.1h, total: 15.3). Therefore, there is an upper threshold on the retention time above which downstream rhythm propagation is severely impaired. There are few genes with export time above 5 hours and that were not included in the global analysis (Fig.4.25) because not classified as M2 in step2, but still show good concordance between nuclear and total lifetime, such as *Wwox* (nuclear lifetime: 8.6, total: 6.4h), *Lpl* (nuclear lifetime: 8.1, total: 6.9h), *Glo1* (nuclear lifetime: 4.76, total: 5.5h), or *Egfr* (nuclear lifetime: 12h, total: 7.3h, also highlighted as a potentially retained transcript in Fig.4.11).

Together, this analysis confirms that isolating nuclear and cytoplasmic transcriptomes can be used to separately estimate regulatory processing steps that are indistinguishable at the bulk level. Moreover, while time spent in the nucleus has a minor contribution to the overall lifetime for a majority of the transcript, nuclear retention plays an important role by delaying the phase in the cytoplasm for several genes with important metabolic functions related to carbohydrate, thyroid hormone, and bile acids.

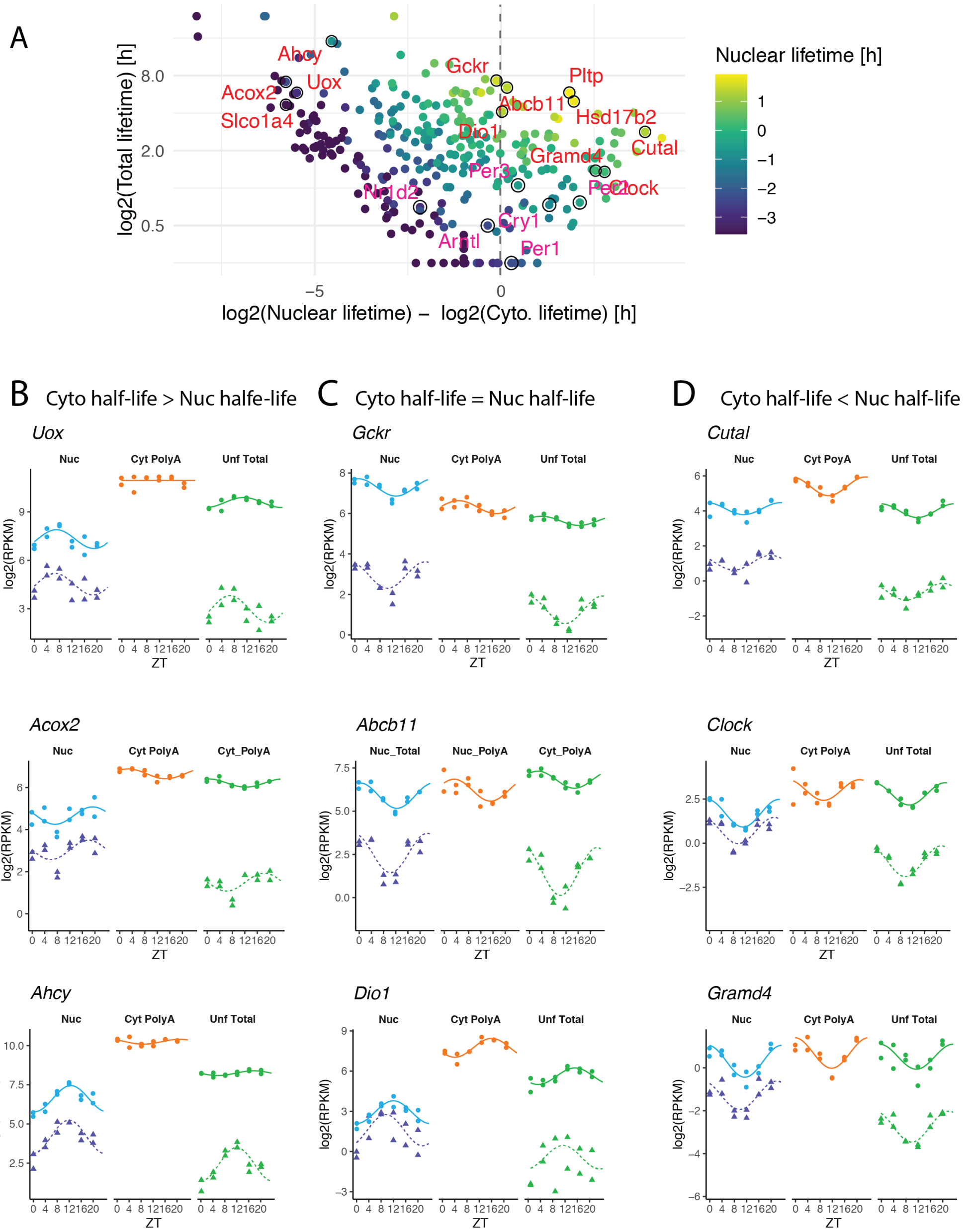


Figure 4.26 – Relative contribution of the export time and cytoplasmic half-life to the total lifetime. A: Scatterplot showing the relationship between the total lifetime estimated from UTI and UTE against the log$_2$-ratio of the export time (nuclear lifetime) and the cytoplasmic degradation time (cyto. Lifetime=. The y-axis is log$_2$-transformed, but values on the axis are linear. Color code indicates long export time (yellow) or fast export (dark blue). Clock genes are highlighted in pink. Only genes classified as M2 in all three comparisons (NTI-NAE, NAE-CAE, and NAE-CAE) are shown. Additionally, the sum of the nuclear and cytoplasmic lifetime is within a margin of error of 2 compared to the total lifetime. B, C and D: Temporal profiles of NTI (dotted blue line), NAE (solid blue), CAE (orange) and UTI and UTE (green lines) of selected genes. Temporal profiles were fitted independently for each RNA with a harmonic linear regression, not with the mathematical model. In B: Cytoplasmic lifetime is much longer than the nuclear lifetime, therefore, contributes predominantly to the total lifetime estimated by comparing UTI and UTE. C: RNA transcript spends an equivalent amount of time in the nucleus and in the cytoplasm. D: Cytoplasmic half life is shorter than nuclear half-life

.

4.10.3 Subcellular RNA distribution versus estimated kinetic rates

At steady-state, the ratio of nuclear versus cytoplasmic RNA levels is directly related to the ratio of the time spent in the nucleus (nuclear retention time) over the time spent in the cytoplasm (cytoplasmic half-life). However, different combinations of rates lead to the same steady-state level, and their relative contribution is unknown. In order to quantify this, in the previous sections, we independently estimated the nuclear export rates and the cytoplasmic degradation rates of rhythmically transcribed genes, based on their temporal patterns (phase delay and relative amplitude). In principle, the ratio of these two rates should explain the subcellular distribution of transcripts observed at steady-state. Unfortunately, the difference between the $\log_2$-nuclear lifetime and the cytoplasmic lifetime did not correlate with the $\log_2$(NAE/CAE) (Fig.4.27.A $\rho = 0.02$, p-value = 0.6). Even after restricting the analysis to genes whose export and degradation rates that matched the total degradation rates estimated from UTI-UTE (from the section above, n = 341 genes, "well-estimated"), the correlation only marginally improved ($\rho = 0.16$, p-value = 0.0028, data not shown). When analysing the $\log_2$(NAE/CAE) ratio without the temporal dimension (Fig.4.4, section 4.6.2), we suggested that the cytoplasmic degradation is the process with the largest variance along with transcription rates, and that differential RNA distribution in the nucleus and in the cytoplasm is most likely explained by the variation in cytoplasmic stability (Fig.4.4). When investigated separately, the cytoplasmic degradation rates could not explain more than 2% of the $\log_2$(NAE/CAE) ratio variability (Fig.4.27.B). However, if we restrict the analysis to the 341 well-estimated genes, the correlation improved, with a Pearson's correlation $\rho = 0.33$ (p-value = 3.2 x 10^{-10}). The export rate is a poor predictor of the ratio (Fig.4.27.C), independently of the analysed genes set. Therefore, for a subset of ~ 300 genes, the degradation rate is a more potent driver of the localisation of RNA transcripts than nuclear export rate. Stable transcripts tend to accumulate in the cytoplasm, while short-lived transcripts appear to be enriched in the nucleus because of the high turnover rate.

We additionally investigated the relationship between the kinetic rates and the transcript length, because we showed that longer transcripts tend to be more enriched in the nucleus compared to the cytoplasm (Fig.4.4, Fig.4.5). The transcript length (but not the gene length) has a significant negative correlation with the cytoplasmic half-life in both gene sets (Fig.4.27.D, $\rho = -0.3$ for all genes classified as M2, $\rho = -0.34$ if restricted to well-estimated genes, p-value $< 10^{-10}$). We found no significant correlation with the export rate (Fig.4.27.E). This suggests that long transcripts are associated with faster cytoplasmic degradation rate, which results in their apparent enrichment in the nucleus.

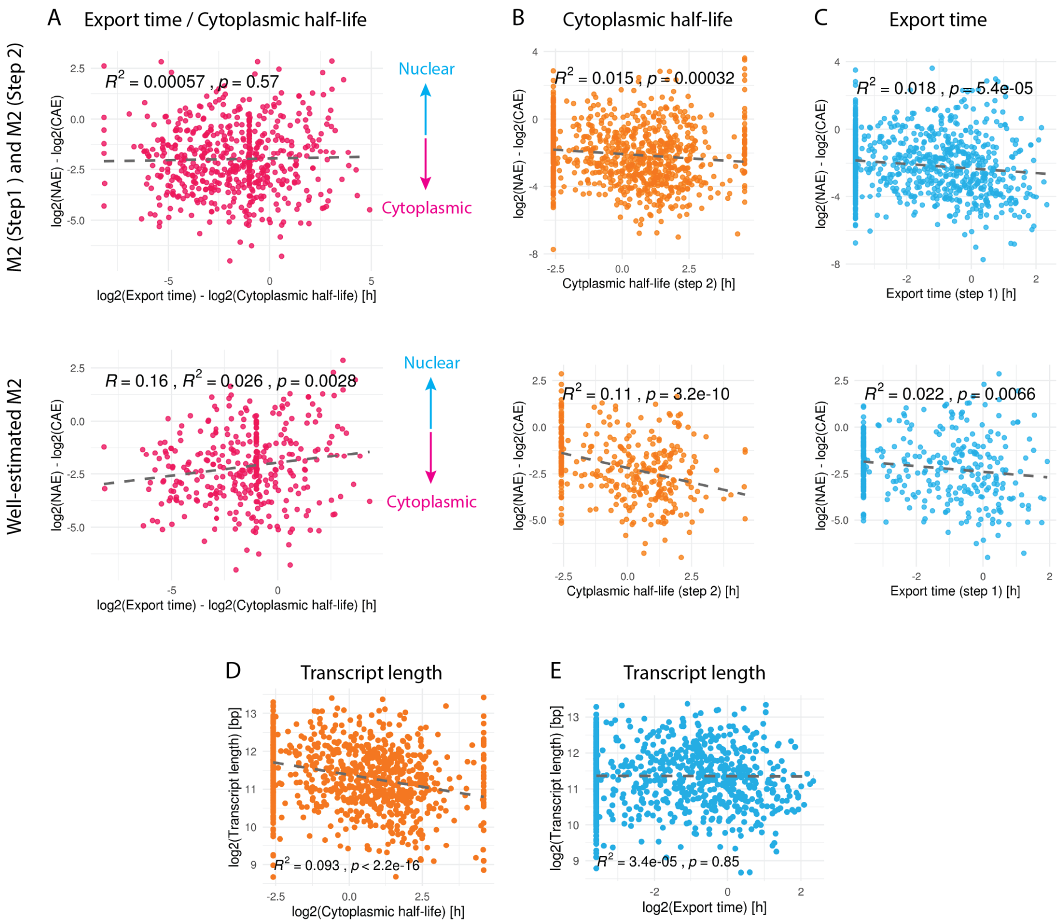


Figure 4.27 – The ratio of nuclear and cytoplasmic lifetimes does not explain the subcellular distribution of rhythmic genes. A: Scatterplot showing the relationship between the Top row: Genes that were classified as M2 in both step 1 and step 2. n = 884 genes. Bottom row: genes that were classified as M2 in step 1, step 2 and UTI-UTE, and whose sum of lifetimes matches the total lifetime estimated with UTI-UTE. N = 341 genes. Excluding parameters that reach the upper or lower boundary only marginally affected the R^2 by less than 1% in all comparisons. D and E: correlation of the transcript length (y-axis) with the cytoplasmic half-life or with the export times ($\log_2$-transformed). If only well-estimated parameters are used (341 genes): $R^2 = 0.12$ for cytoplasmic half-life and $R^2 = 0.023$ for export time.

4.11 Discussion

4.11.1 Characterisation of the nuclear and cytoplasmic transcriptome revealed different subcellular localisation of distinct RNA classes

As an initial exploratory analysis of our dataset from nuclear and cytoplasmic fractions, we first characterised the differences between RNA populations, and investigated two classes of RNA with specific subcellular distributions: protein coding transcripts and Retained Intron.

Protein coding transcripts are the most abundant RNA biotype, particularly in the cytoplasmic fraction. These PC transcripts are on average more abundant in the cytoplasm than in the nucleus. Still, a non-negligible fraction of protein coding transcripts are found at greater levels in the nucleus. A functional enrichment analysis (GO terms) of differentially localised transcripts revealed a concordance between the preferential localisation of the RNA and its encoded protein [208]. We found that nuclear transcripts code for gene expression regulatory functions such as epigenetic modifications, mRNA processing, and export (THO complex, *Nxf1*, Nucleoporins NUP), while proteins coded by cytoplasmic mRNA are mainly involved in translation, oxidative phosphorylation, fatty acid β-oxidation and detoxification. The concordance between RNA and protein localisation was further verified with a published dataset from mouse liver [213].
Transcripts coding for housekeeping functions are often stabilised, while those coding transcription regulatory proteins could be kept unstable in order to facilitate rapid adaptation of protein production in response to external stimuli [152]. Here, because we analysed the steady-state expression level, we could not determine whether the relative subcellular abundance is a consequence of the long half-lives of housekeeping functions versus the presumably short half-lives of genes involved on regulatory functions [232], or if an active nuclear retention mechanism restrict the nuclear export, therefore acting as a passive filter lowering the noise associated with stochastic transcription [127, 128, 196].

Transcripts with a retained intron (RI) make up to 15% of the nuclear transcriptome, and are particularly enriched in the nucleus compared to the cytoplasm. RI transcripts are either rapidly degraded in the nucleus, or exported to the cytoplasm, but also serve as substrates of post-transcriptional splicing [111, 129, 110]. We thus asked whether the relationship between RI and the corresponding PC had a specific signature reflecting a mechanism of nuclear retention. Among the 260 pairs defined, ~10% were associated with a greater nuclear abundance at the gene-level, and this enrichment was predominantly due to the higher abundance of the RI isoform over the PC isoform. This pattern could reflect the scenario where RI is slowly post-transcriptionally spliced, and the resulting PC mRNA is rapidly exported. This suggests that post-transcriptional splicing could slow down the export of a PC transcript by retaining the RNA as an immature form in the nucleus. Notably, the nuclear-enriched gene *Gckr* followed that pattern and has a long estimated nuclear retention time (3h). However, we were not able to estimate the kinetic parameters for the retained intron isoform because it was not cycling, and our method is limited to rhythmic profiles. Moreover, even if we restrict the analysis to pairs of transcripts that differ only by the presence of one intron in PC, we are not able to tell if these pairs represent *bona fide* precursor-product relationships. Nevertheless, RI is a class of RNA biotypes with distinct signatures related to the preferential nuclear localisation, which is influenced by both the nuclear lifetime and cytoplasmic lifetime. What triggers the post-transcriptional splicing and release of the mRNA includes heat shock events [111], or neural activity [112], but have only been demonstrated on a case-by-case basis, and remain largely unknown.

4.11.2 Estimation of the co- versus post-transcriptional splicing times using nuclear pre-mRNA and mRNA

Splicing is thought to mainly occur co-transcriptionally in most studied organisms [87], [86, 91, 89, 87]. However, the proportion of introns that are removed concurrently with transcription is still under debate, depending on the model organism and on experimental method [92].

In this study, we investigated the question of the proportion of co- versus post-transcriptionally spliced genes by comparing the expression level of pre-mRNA from isolated nuclei. We specifically compared the population of total pre-mRNA (NTI) to the polyadenylated pre-mRNA (NAI), which represents transcripts that have been fully transcribed and terminated, but still contain intronic regions. We first verified that the NAI/NTI $\log_2$-ratio could reflect the splicing regime. A value of NAI close to NTI means that most of the pre-mRNA in the nucleus are terminated pre-mRNA, thus, a large fraction of intron still needs to be post-transcriptionally removed. If NTI is much larger than NAI, most of the captured pre-mRNA are nascent transcripts, meaning that almost no introns remain in the polyadenylated form. We visually verified if the small NAI/NTI $\log_2$-ratio reflects a co-transcriptional splicing regime by inspecting the distribution of reads mapping on intronic regions. We observed sawtooth patterns with the 5' enrichment and 3' depletion of reads, particularly on long intron, which is typically described in nascent-seq datasets, and reflect the (uniform) distribution of RNA PolII along the gene body [59, 89]. Therefore, the relationship of nuclear polyadenylated pre-mRNA versus the total population of nuclear pre-mRNA informs us about the extent of co- versus post-transcriptional splicing. We noticed a strong correlation of the NAI/NTI $\log_2$-ratio with the gene length: long pre-mRNA tends to be more co-transcriptionally spliced, while short pre-mRNAs are more post-transcriptionally spliced. These observations support the view that transcription and splicing are two processes occurring in parallel, and that splicing frequency depends on the time required for transcription. When the gene is short, transcription is already completed by the time the intron is recognised and spliced. If the gene is long, there is more time allocated to the spliceosome to recognise the splice sites, assemble on the pre-mRNA, and cleave the intron. However, it is also known that the machineries involved in the pre-mRNA processing interact with each other, for instance by recruiting splicing and processing factors directly on the site of transcription [233], or because all these factors are concentrated in one location, such as in nuclear speckles, or in phase-separated condensates [218]. Moreover, coordination patterns of RNA processing steps differ whether it is for protein coding transcripts or other RNA classes, such as lncRNA, and depends on the phosphorylation status of the RNA PolII [221], or on the recruited RNA-binding proteins [119]. Here, we observed different signatures of export and splicing for different RNA biotypes. For instance, lncRNA has the lowest splicing frequency, with a particularly high abundance of polyadenylated pre-mRNA. Retained Intron is more influenced by the export time compared to other RNA biotypes. Here, we emphasise again that the process of export does not discriminate between the transport of RNA to the cytoplasm, and the nuclear degradation or, in case of Retained Intron, the post-transcriptional splicing of the remaining intron. Moreover, the correlation between the extent of co- versus post-transcriptional splicing with gene length was absent for all the other biotypes except by Nonsense-mediated decay NMD (NMD: R^2 0.17, Protein Coding: $R^{0.4}$). Because the degradation mediated by Nonsense-mediated decay occurs in the cytoplasm after the initiation of translation, it would not be surprising that NMD transcripts are processed in a similar way as protein coding transcripts in the nucleus. Together, these observations further suggest that different biotypes are regulated by different processing programs.

In this analysis, we investigated the extent of co- versus post-transcriptional splicing at the transcript-level, without taking into consideration different splicing efficiencies among introns of the same gene body. But each intron has different splicing kinetic rates depending on its position along the gene [91], on the length of the downstream exon, ans on whether the flanking exons are alternatively or constitutively spliced [107]. One method is to quantify reads mapping on an intron - exon junctions versus reads mapping on exon-exon junctions, and to define a co-transcriptional splicing index for each intron (CoSI) [194], however, this would not allow to discriminate different isoforms like with we did with the pseudo-alignment using Kallisto. A more sophisticated but complex method is the long-read sequencing, for instance using Nanopore technology, which additionally allows to uncover patterns of splicing order [234].

4.11.3 Mathematical model to quantify kinetic parameters

For years, the gold standard to study RNA dynamics has been through transcriptional blockage methods (Amanitin, Actinomycin D [235]) and fitting an exponential decay function. Pulse-chase strategy with uracil analogs (4sU [225], EU [236]) label nascent transcripts during a short period of time. These labelled transcripts are then chased at different time points, and again, the decay curve allows an estimation of the half-life. Snapshot images of subcellular RNA distribution, either obtained by smRNA-FISH or by RNA-seq of fractionated cells allows a characterisation of nuclear and cytoplasmic transcriptomes [214, 74, 209, 95, 194]. Recently, proximity-labelling based sequencing (APEX-seq, [208]) could achieve a high spatial resolution, mapping the localisation of RNA transcripts in several subcellular compartments such as the nucleolus, endoplasmic reticulum, cytosol, outer and inner membrane of the mitochondria. These studies revealed broad patterns of localisation for diverse RNA classes.

Mathematical modelling methods using ordinary differential equations (ODEs) to describe RNA-seq signals can infer genome-wide the kinetic parameters of RNA processing steps in a label-free manner [216]. Temporally dynamic datasets reflecting different cellular states are needed in order to solve the system, for instance during cell cycle, stem cell differentiation [237], or at different circadian time-points [78, 79]. Here, we applied the model previously developed in the Naef laboratory. A system of two ODEs describes the rhythmic profiles of pre-mRNA and mRNA, which uses the relationship between the phases and relative amplitudes in order to infer the production and the decay terms. Additionally, the decay term is described as a cosine function, and thus identifies temporal patterns regulated by rhythmic degradation. In order to apply the ODE system as such, we splitted our model of an RNA lifecycle in two distinct steps (NTI →NAE, and NAE→CAE). This allowed us to estimate the cytoplasmic degradation rate of ~1300 genes and the nuclear export rate of ~1400 rhythmic genes. Additionally, modulating the nuclear lifetime in a rhythmic manner contributes to generating oscillations of nuclear mRNA of ~100 genes, and also to boosting the amplitude of several key metabolic genes (*Pck1*, *Lpin1*, *Por*). We estimated that rhythmic nuclear export is a post-transcriptional regulatory step affecting ~10% of the rhythmic profiles of mature nuclear mRNA.

Limitations of the model

The model can be extended by fitting both processes together, such that the term common to both step 1 and step 2 (the export rate *e*) is fitted only once. Additionally, in the current model, if the nuclear mRNA (NAE) is rhythmically exported in step 1, the temporal profile may have a shape that deviates from a cosinor function. But that same NAE is then modeled with a cosinor in step 2, which may not be the most appropriate function, resulting in a high error of the fit, and consequently, the gene would be unclassified in step 2. Modeling the two steps together would however increase the number of parameters and complexify the model, while affecting the fit of only a small fraction of the genes: less than 5% of the genes classified in step 1 were transcribed and exported rhythmically (91 genes in M4), therefore, the complexification of the model may only bring minor improvements.

One major limitation of the model is that the degradation rate (or export rate) is estimated using only the relationship of the phases and relative amplitudes, but the relative mean expression level of the two RNA species are not taken into account. In this study, we were not able to explain the subcellular distribution of transcripts using the estimated export rates and cytoplasmic degradation rates. Even if we could confidently quantify the relative contribution of the nuclear and cytoplasmic lifetimes to the total RNA lifetime for ~300 genes, the relative amount of time spent by a transcript in each compartment did not reflect the relative subcellular abundance of the transcript. Therefore, a major improvement to the model would be to integrate the information contained in the ratio of the mean expression level, and use it to constrain the parameters. We attempted to include the mean expression levels and stratified the analysis by clustering temporal patterns in function of the relative subcellular abundance (Fig.4.16). While this analysis revealed that the same steady-state levels of nuclear and cytoplasmic RNA can be achieved by different dynamic strategies (different combination of rates), our predictions concerning the export rates that should explain the observed patterns, for instance, that *Period* genes should be retained in the nucleus, were not confirmed with the estimation from step 1.
The absence of correlation between the NAE/CAE $\log_2$-ratio and the ratio of the export and degradation rates could also be because the errors on the export times are larger than anticipated. The model reliably detects long export times on the order of our sampling time (in the range of hours). While cytoplasmic mRNA lifetime is indeed in the range of hours (median of 2.5h), the nuclear lifetime is much shorter (median of 25 min), and many genes reached our lower boundary of 5 minutes. Therefore, the temporal resolution may not allow a precise quantification of processes happening on extremely short time-scales. While this is still informative to discriminate rapidly exported transcripts from nuclear retained transcripts, and also to discover rhythmic patterns (M3, M4), the resulting ratio of rates might be too affected to predict the final subcellular localisation. The $\log_2$(NAE/CAE) was partially explained by the cytoplasmic half-lives when restricting the analysis to ~300 well-estimated parameters (~10% of the total variance, Fig.4.27.D), while no correlation was found when comparing the export rate

of the same genes. This further suggests that cytoplasmic degradation rates might be better estimated than export rates.
A last reason that may account for part of the inability to explain the NAE/CAE $\log_2$-ratio is that long-lived nuclear transcripts are not taken into account when comparing the two rates. Long nuclear retention is associated with damped amplitudes. Subsequently, these transcripts are not detected as rhythmic in step 2, but classified as M1, and the degradation rate is unidentifiable. We found that the maximum nuclear lifetime for the rhythm to be propagated into the cytoplasm is around 4 to 5 hours.

Identification of long-lived nuclear RNA

Despite not being able to fully explain the subcellular distribution of RNA, our dataset still allowed us to obtain biologically relevant estimations of cytoplasmic half-lives for rhythmic transcripts (~2.5h). Additionally, even if we were not able to precisely quantify short export times, we were able to distinguish nuclear mRNA that are processed within minutes from those that have longer half-lives (in the range of hours). Long-lived nuclear mRNA are of particular interest, because they can only be detected by using cellular fractionation. In the original work by Wang et al. [201], pre-mRNA and mRNA from total unfractionated liver cells were used in order to estimate the cytoplasmic degradation rates, but nuclear lifetime was neglected. For a majority of genes, this assumption is valid because of the little contribution of nuclear lifetime to the total lifetime. However, when nuclear lifetime significantly contributes to the total RNA lifetime, our dataset provides a different perspective to the global view of the kinetic processes. Typically, we could quantify the relative contribution of nuclear and cytoplasmic lifetimes, and highlight some genes whose cytoplasmic mRNA phase is delayed because of the slow nuclear export (*Gckr*, *Dio1*, *Ppa1*, *Hsd17b2*, *Abcb11*).

In this study, we described distinct dynamic patterns and highlighted genes with intriguing profiles, but were not able to find mechanisms that could explain the different regulatory modes. We performed many functional enrichment analyses interrogating several databases (GO, KEGG, Wikipathway), and usually found enrichments for functions related to the metabolism of lipid, cholesterol, carbohydrates, xenobiotics, and complement system. Enrichment of these functions is expected since we are studying the rhythmic transcriptome of the liver. However, in most cases, the enriched biological functions were not specific enough to clearly separate the different groups of genes. Searching for common biological functions to explain similar temporal patterns is probably a too naive method.

One genomic feature associated with the estimated kinetic rates is the transcript length. At steady-state, we observed that the long transcripts were preferentially located in the nucleus, while short transcripts were more abundant in the cytoplasm. We found a weak yet significant association with the cytoplasmic degradation rates, such that short transcripts tend to be associated with long half-life, while long transcripts are more prone to degradation. The link between transcript length and decay rate was already hinted at by a previous study in *Drosophila* Kc167 cells [209]. Degradation predominantly occurs through the deadenylation of the poly(A) tail followed by exonuclease attack [152], and has not been shown to be specifically related to the transcript length. The higher instability of long transcripts may reflect an increased probability to be stochastically attacked by endonucleases [209].

In our search for possible mechanisms discriminating between rapidly and slowly exported RNA, we looked for differential m^6A RNA methylation patterns using a published dataset of MeRIP-seq in mouse liver [229]. m^6A RNA methylation is a very common mRNA modification, which is usually associated with fast processing, export, translation, and decay [147, 144]. Preliminary analysis did not reveal specific enrichment of m^6A counts or sites that could differentiate rapidly exported from slowly exported nuclear mRNA, but we may need to refine the analysis.
RNA processing rates such as splicing, packing, and export are modulated by various RBPs, which are recruited on specific RNA sequences / structures [193]. We specifically looked at the RNA bound by NONO, a multifunctional nuclear protein associated with nuclear paraspeckles, which has been shown to modulate the maturation of RNA, notably those related to glucose and lipid metabolism in mouse liver [122]. Using the published NONO-RIP seq in WT and NONOgt mice (lacking a functional NONO protein)[122], we separately examined rhythmic target RNA whose phase were delayed in NONOgt compared to WT (suggesting that NONO promotes fast processing) and those who were advanced in NONOgt (suggesting a retention role of NONO). However, at this stage of the analysis, we were unable to draw any conclusion concerning the potential role of NONO regarding our estimated export times.

The default pathway of a correctly processed protein coding mRNA, i.e. 5'capped, 3'polyadenylated, spliced, coated with the right set of RBPs, is most likely the export to the cytoplasm [114]. The presence of retention-promoting features such as specific motifs, hyperedited regions, high GC content, or unspliced intron may compromise the export of the transcript [238]. In our study, a more in-depth search for such features should be implemented in the future in order to explain the different combination of kinetic rates that we observed. Finally, it would be interesting to distinguish the cases when nuclear enrichment is a consequence of the accumulation of slow processing steps, or if mRNAs are actively retained, and therefore creates a reservoir of mRNA ready to be released in response to a signal. In the context of the liver transcriptome, feeding-related signals (insulin, glucagon), could trigger such response.

4.11.4 Concluding remarks

From birth to death, every step during the life cycle of an RNA transcript is tightly regulated to ensure proper cellular function. In this study, we made a simple model of the complex network of RNA processing steps, including splicing, nuclear export, and cytoplasmic degradation. We provide a genome-wide and temporal catalogue of RNA subcellular localisation in the liver, along with a comprehensive estimation of the nuclear and cytoplasmic life times of cycling transcripts. This work suggests that mRNA oscillations can be post-transcriptionally regulated at the level of nuclear export, and contributes to a better understanding of the dynamic regulation of the transcriptome over the 24h day.

4.12 Supplementary Analysis

In this chapter, we provide additional analyses related to the Chapter 4. However, these analyses are not necessary for the understanding of the project.

4.12.1 Overview of temporal characteristics of the liver transcriptome

Rhythmic transcriptome of WT mouse liver

Circadian rhythms of the liver transcriptome have already been extensively studied [60, 47, 228]. It is one of the organs with the highest number of rhythmic genes, with an estimate of 10 to 20% of cycling mRNA [239, 8, 42]. However, this proportion of rhythmic genes varies from one study to another. This can be due to different experimental conditions (light-dark, dark-dark cycle, age, constant or feeding regimen [37, 33, 228], but also to statistical analysis, whether it is based on a cutoff (p-value, minimal amplitude), or on model selection [60, 228]. In this section, we briefly present basic temporal characteristics of the dataset, and recapitulate some known results from the liver chronobiology. We analysed rhythmic genes in different RNA populations without taking into consideration any modelling aspect. We fitted a cosinor function with a period of 24 hours to the read counts during the normalisation procedure of RNA-seq with DESeq2 [204] (see Methods 6.5.3). We tested the fitted function against a model with an intercept only, and genes are considered rhythmic if the p-value is less than 0.01. We additionally set a threshold on the amplitude so that a rhythmic gene has to be at least twice as expressed at the peak than at the trough ($\log_2$ fold-change > 1, Fig.6.3.A). We found between 500 and 850 rhythmic genes in each of the eight RNA populations, which represents ~5% of the liver transcriptome (11'000 genes) (Fig.6.3.B). Even if the proportions are lower than what has been previously reported, we actually found similar values from another published total liver RNA-seq dataset [60] when applying the same fitting method and cutoffs (around 500 rhythmic genes in Total RNA from whole liver tissue). 115 genes were rhythmic in all RNA populations, and 179 in at least 7 out of 8 conditions.

When rhythm propagates from one compartment to another, the temporal features (namely phase and amplitude) vary depending on the stability (half-life) of the transcript [54]. If a transcript is long-lived, the oscillations dampen and the phase peaks later. On the contrary, short-lived transcripts have similar phases and relative amplitudes in each compartment. If the degradation is not constant but rhythmic, other temporal patterns can be observed, for example: the relative amplitude increases, the phase shift is much larger than expected (theoretical limit is 6 hours), or rhythm can be generated *de novo* [79]. The mathematical framework describing how degradation and export rates modulate rhythmic patterns are explained in the main results section (See Method: 6.7). Therefore, in section, we only provide a simple overview of the rhythmic dataset.

The amplitude range is the largest at the transcriptional level (introns) and the lowest in the cytoplasm (Fig.6.3.A), suggesting a gradual damping of oscillations as rhythms propagate. The higher number of rhythmic genes detected in the cytoplasm compared to the nucleus (pre-mRNA and mRNA) suggests that post-transcriptional mechanisms generate rhythmic mRNA accumulation [47, 59]. However, these oscillations are less strong than those generated at the transcriptional level, such as the core clock genes. Particularly, *Arntl, Nr1d1* and *Per3*, together with *Dbp* and *Rgs16*, are amongst the genes with the largest fold change in all RNA types (Fig.6.3.C). We show some representative temporal profiles of core clock genes in Fig.4.29. They are all rhythmically transcribed (rhythmic intron) and accumulate with similar phases and amplitudes in all RNA populations, in agreement with their estimated short half-lives [78]. The only exception is *Cry2*, with p-value of 0.02 did not pass our cutoff for rhythmicity. Because the liver is a metabolic organ that responds to the circadian food (un)availability, genes related to lipid (*Lpin1*), carbohydrates metabolism(*Gys2*), and energy homeostasis (*Nampt*) are also rhythmic.

We observe the typical bimodal distribution of transcriptional activity (depicted by NTI and UTI) peaking at dusk and dawn, around ZT10 and ZT21 [44] (Fig.6.3.D). Accumulation of mRNA in the nucleus (NAE and NTE) also follow the typical bimodal distribution with the same phases as pre-mRNA, while the two peaks of cytoplasmic mRNA are shifted by ~4 to 5 hours in the cytoplasm, particularly in CTE. In CAE, the distribution is more uniform. The phase distribution in UTE is also bimodal, however, the profile is more homogeneous overall, probably because Unf Total is a mixture of both nuclear and cytoplasmic transcripts, peaking at different times of the day.
We specifically looked at the phase difference of each gene between two RNA populations: NTI - NAE (nuclear pre-mRNA that is polyadenylated and spliced), and NAE-CAE (nuclear mRNA exported to the cytoplasm). The average phase shift is 1 hour between NTI and NAE, and 2 hours between NAE and CAE (Fig.6.3.E and F). This suggests that the export rate, which governs the phase delay between NTI and NAE, happens on a shorter time-scale than the cytoplasmic degradation rate, which controls the NAE-CAE phase shift. Therefore, with a simple analysis of rhythmic profiles, we can already discern genes with atypical parameters. *Gstm7* has a long phase delay between the peak time in the nucleus and in the cytoplasm (4.7h) and we indeed estimated a long half-life of 5h (Fig.6.3.F). In the nucleus, *Aqp8* also stood out because of the long phase delay between NTI and NAE (2.7h), and has a long nuclear lifetime of 2h, longer than the median of 20 minutes estimated in the main section.

In the main section, we used a more sophisticated mathematical model to analyse rhythmic patterns with underlying assumptions. For instance, here, the cosinor is fitted independently on each RNA species, therefore, the phase of pre-mRNA is allowed to peak before the mRNA, which is biologically not relevant. Alternatively, we could use a model-selection based approach as in [60] and [228], where rhythmic parameters are fitted on all the conditions at the same time, and are allowed to be shared (same phase and amplitude in all conditions, a parsimonious model), or independent. The model of Wang et al. [201] was preferred, as it allows to estimate the kinetic parameters dictating the mRNA profiles, and additionally uncovers rhythmic degradation processes.

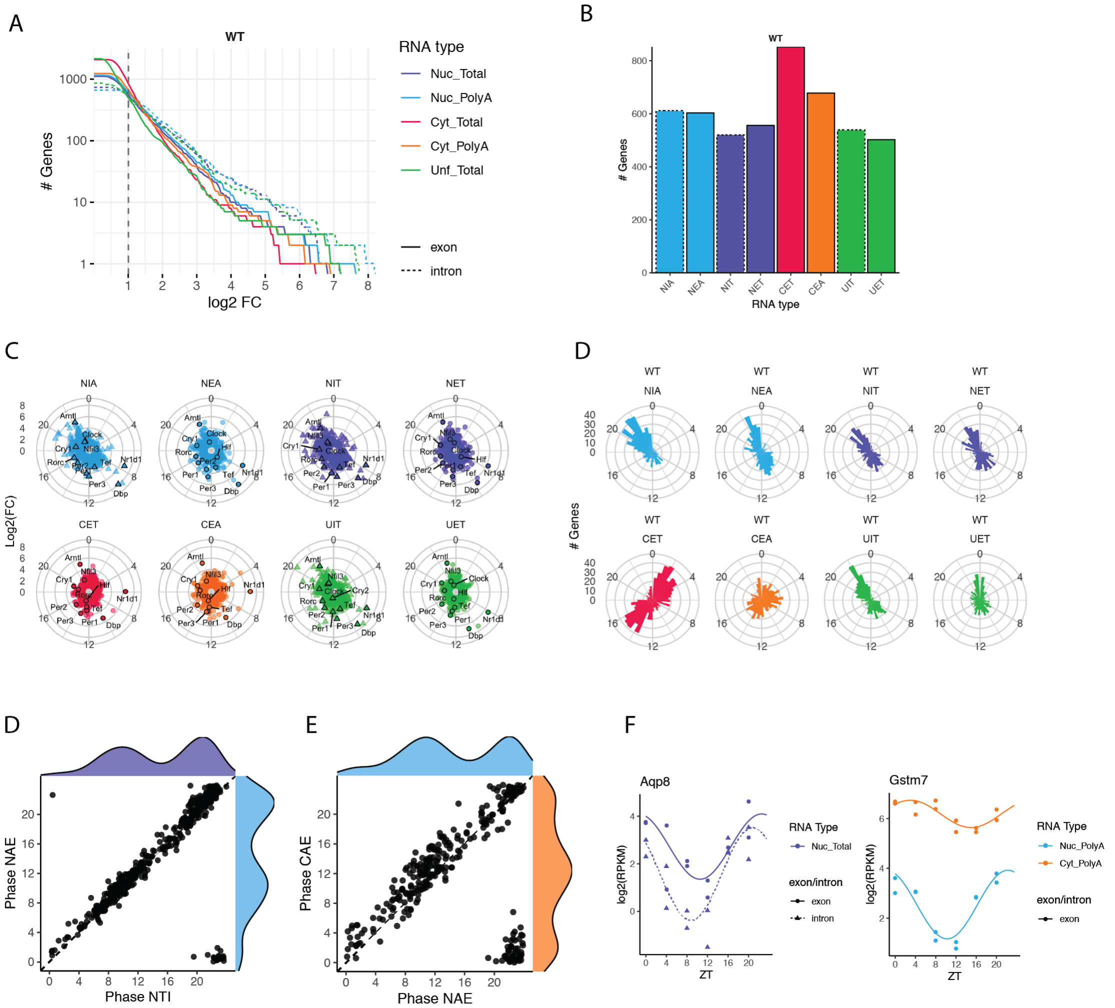


Figure 4.28 – Temporal parameters of different RNA populations. A: Cumulative number of genes with a $\log_2$FC larger than a threshold value (on the x-axis) in each RNA population. Dotted line indicates the threshold at $\log_2$FC of 1. B: Number of rhythmic genes with a $\log_2$FC >= 1 per population. C: Circular plot showing the Distribution of phase and amplitude of all rhythmic genes. Clock genes are highlighted. D: Distribution of phases of rhythmic genes. E: Comparison of phases between NTI and NAE (337 genes, mean phase shift 0.9 hours) and NAE and CAE (299 genes, mean phase shift of 2.1h) F: RNA-seq profiles of two example genes with large phase delay: *Aqp8* with a phase shift of 2.5 hours between pre-mRNA and mRNA in the nucleus, and *Gstm7* with a phase delay of 4.7h.

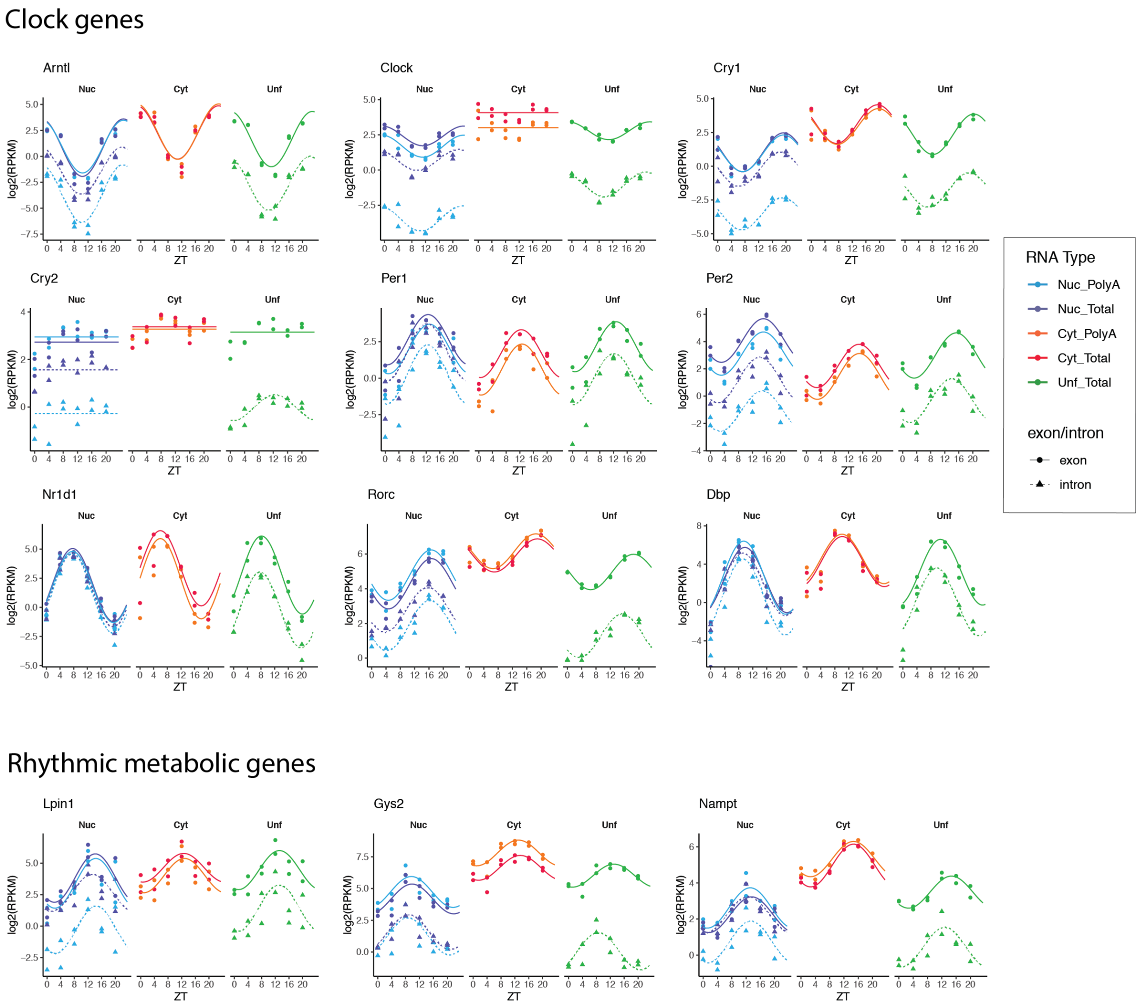


Figure 4.29 – Representative RNA-seq profiles ($\log_2$(RPKM)) of core clock genes and metabolic genes in 8 RNA populations: Nuc Total, Nuc PolyA, Cyt Total, Cyt PolyA, Unf Total. pre-mRNA profiles are shown in dashed lines. Raw counts were fitted with a cosinor function and tested against a model with an intercept only, using DESeq2. A gene is considered rhythmic if p-value < 0.01 and $\log_2$FC > 1.

Rhythmic transcriptome of clock-deficient mouse liver (Cry1/Cry2 KO)

In addition to WT mice, we use a clock-disrupted mouse model which lacks functional core clock proteins CRY1 and CRY2 [240]. These mice have disrupted patterns of locomotor activity and of feeding. In a recently published comparative study in mouse liver [228], 40% of the rhythmic genes in WT became arrhythmic in Cry1/2 KO, and overall, amplitudes were damped. Among the remaining cycling genes, half of them were "food-driven", which means that they lose rhythmic expression in the absence of rhythmic food intake *ad libitum*. For a more in-depth analysis of the rhythmic phenotypes of the Cry1/Cry2-KO mice, please refer to the work of Weger et al., co-published by the Naef group [228].

In our dataset, with a threshold on the amplitude ($\log_2$FC > 1), we found around 300 rhythmic genes in the nucleus, and only around a hundred in the cytoplasm, for a total of 769 genes rhythmic at least in one RNA population (Fig.4.30). All core clock genes (*Cry1*, *Cry2*, *Per1*, *Per2*, *Per3*, *Arntl*, *Clock*, *Nr1d1*, *Rorc*) were arrhythmic in CryKO mice. It was previously reported that *Per2* is driven by systemic cues and continues to cycle even in the absence of a functional clock [241]. In our dataset, even if *Per2* did not pass our set threshold for rhythmicity, it has the lowest p-value among all the core clock genes (p-value of 0.028 in UTI), suggesting that there were still some remaining rhythmicity.

Overall, there is a decrease of amplitudes between CryKO and WT (Fig.4.30.B), in line with the previous observation that genes whose transcription is driven by the clock have higher amplitudes than those entrained by systemic signals [228]. The amplitude range was the largest at the transcriptional level (UTI, NTI) (Fig.4.30.A). Only 48 genes have an amplitude above 8-fold, compared to a 138 in WT. *Rgs16* is the gene with the highest amplitude in all compartments, similar to WT. Other high amplitude genes include genes involved in hepatic metabolic function, such as *Lpin1* , *Scd1* , *Angptl4* and *Txnip* (lipid metabolism), *G6pc* and *Pck1* (glucose metabolism), and *Hmcgr*, *Hmgcs2 Sqle*, *Insig1*, *Srebf1* (cholesterol metabolism). We also observed a shift of peak times between WT and CryKO (Fig.4.30.D). Overall, genes in CryKO tend to be slightly phase-advanced compared to WT, a phenomenon also observed in [228].

Only 327 rhythmic genes in CryKO were also cycling in WT. Therefore, more than half of the rhythmic genes were previously not cycling in WT, despite both groups of animals being held in the same conditions (light-dark condition, night-restricted feeding). The average mean expression level of the "*de novo*" rhythms were 2 to 4 times lower than the mean expression level of genes that were rhythmic in both datasets, thus, these rhythms could be due to noise inherent to low expression. However, some genes were convincingly rhythmic only in CryKO, such as the lipogenic genes *Me1* and *Retsat*, and the cholesterogenic *Msmo1* (Fig.4.30.D). All three genes are targets of the nuclear hormone receptor PPARα, a central regulator bridging the core clock machinery to lipid, cholesterol, and ketone bodies metabolism [52]. Globally, genes that were rhythmic only in CryKO were enriched for KEGG Pathway "PPARα signalling", (p-value $< 10^{-7}$). However, the mRNA level of *Pparα* was not rhythmic in CryKO, while peaking during fasting time in WT (ZT8). Therefore, it seems that the transcriptional program regulating lipid homeostasis and induced by fasting might be stronger in CryKO than in WT, although the regulation cannot be explained by the transcription of the PPARα.

Globally, there are less rhythmic genes than in WT, and the remaining cycling genes have on average a lower amplitude, as expected from a mouse model with no running clock [228]. However, because the mice are no in free-running conditions, but kept in an environment with cycling stimuli (12:12 light-dark cycle, restricted access to food during the active phase), many functions, particularly those related to feeding / fasting cycles, are still rhythmically entrained.

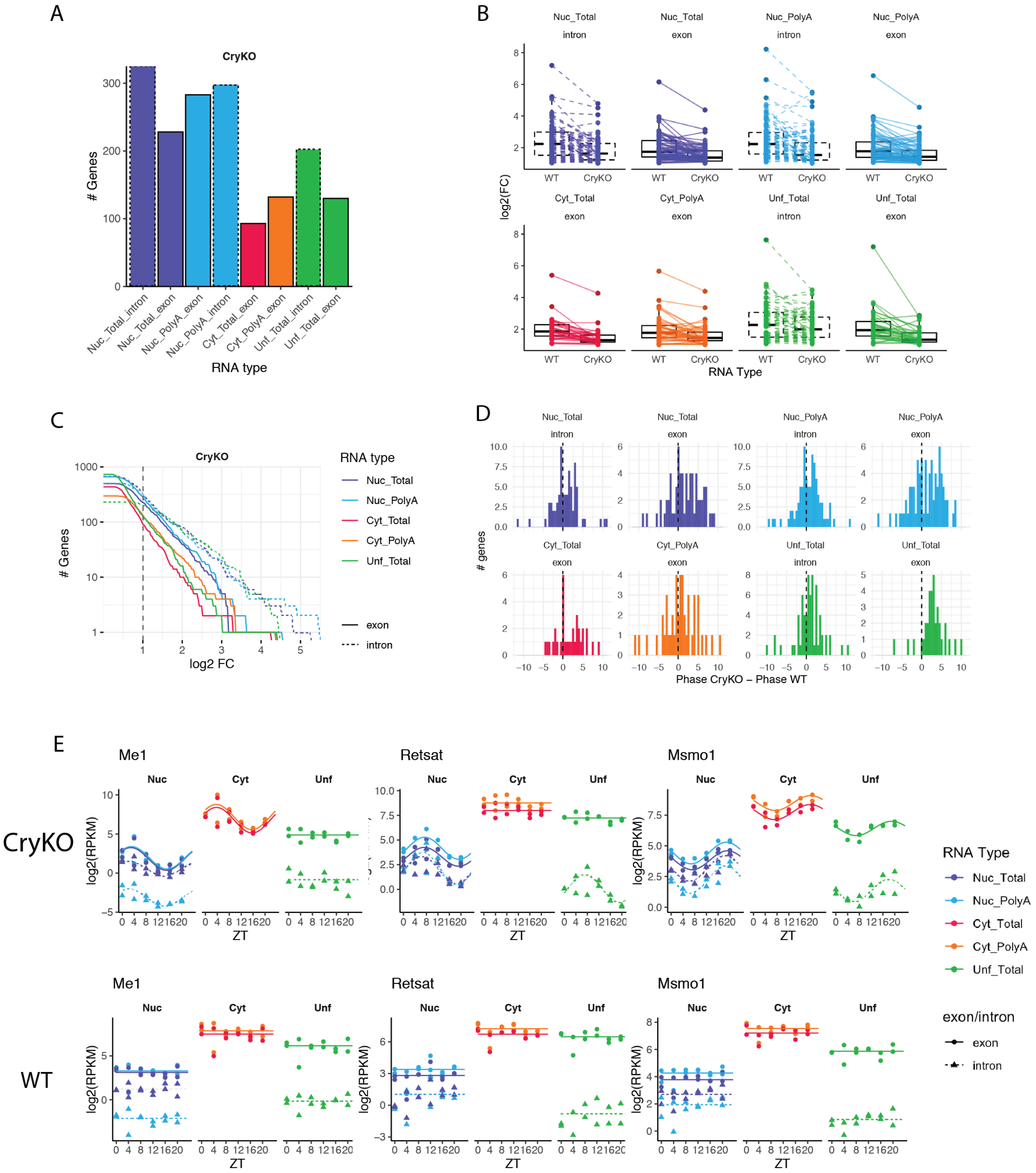


Figure 4.30 – Temporal characteristic of CryKO liver transcriptome. A: Number of rhythmic genes with a $\log_2 FC > 1$ in each RNA Type. B: $\text{Log}_2 FC$ of rhythmic genes in WT and CryKO. The average $\log_2 FC$ is lower in CryKO than in WT (Paired t-test, p-value at least $< 10^{-4}$ for all RNA types. Number of genes per RNA type: NTI = 74, NTE = 68, NAI = 92, NAE =78, CTE = 33 ,CAE = 46, UTI = 68, UT = 35. C: Cumulative number of genes with a $\log_2 FC$ larger than a threshold value (on the x-axis) in each RNA population. Dotted line indicates the threshold at $\log_2 FC$ of 1. D: Phase delay between CryKO and WT. If the phase delay is positive, CryKO peaks earlier than WT. E: Temporal RNA-seq profiles of three genes that were flat in WT but rhythmic in CryKO.

4.12.2 Estimation of kinetic parameters in clock-deficient mouse model

In the main section, we investigated the role of the clock in regulating rhythmic RNA accumulation patterns in the cellular fractions, as well as their roles in the regulation of rhythmic nuclear export and rhythmic mRNA degradation by analysing rhythmic patterns of NTI and NAE, and NAE and CAE. We performed the same analysis in Cry1/Cry2-KO mice (step 1: NTI-NAE, and step 2: NAE-CAE). However, due to the lack of robust temporal patterns involving rhythmic degradation or rhythmic export, we only provide here a preliminary round of analysis.

We first compare nuclear pre-mRNA and mRNA (NTI versus NAE) to explore rhythmic regulation of the export rate. We classified 941 in one of the three models (Fig.4.31.A). 77% were classified as M2 (rhythmic transcription, constant export). We found a surprisingly high number of genes classified as M3 (rhythmic export), but with only little overlap with the M3 genes in WT (genes in common: 10). 41 genes have their amplitudes or phases additionally tuned by rhythmic export (M4). Out of the 726 M2 genes, 429 were also rhythmically transcribed in WT. We noticed that overall, genes that were cycling only in CryKO were enriched for functions related to lipid and sterol metabolism. Particularly the target genes regulated by the lipid-activated nuclear receptors LXR and PPARα, master regulators of the energy homeostasis, have more robust rhythmic patterns in CryKO than in WT, suggesting a stronger response of the transcriptional program regulating lipid homeostasis, most likely in response to fasting and feeding pattern. We focussed the analysis on genes that were common in both genotypes. 60% of the rhythmically transcribed genes classified as M2 in WT became arrhythmic (M1, n = 756) in CryKO, and ~30% are also classified as M2 (n = 366, Fig.4.31.B). Genes that are still transcribed rhythmically in CryKO (M2) have lower amplitudes both in pre-mRNA (NTI) and mRNA (NAE) compared to WT (Fig.4.30.B). As mentioned above, some genes, mainly related to sterol and fructose metabolism, have larger amplitudes in CryKO than WT, for example *Srebf1*, *Cyp39a1*, *Khk*, or *Adlob* (Fig.4.31.D). Phases are also shifted (Fig.4.30.D), with some extreme cases like *Tymp* (phase difference of 6 hours, Fig.4.31.D). These differences in temporal profiles result in a low concordance of estimated half-lives ($R^2 = 0.15$, Fig.4.31. F). Still, some profiles remain similar in both genotypes, as exemplified by *Acat2* (Fig.4.31.G) The majority of genes classified as M3 in WT (constant transcription, rhythmic export) became completely arrhythmic in the absence of functional CRY1 and CRY2 proteins (65%, Fig.4.31.A). 18 genes were classified as M2 in CryKO, 3 as M4 (*Mup14*, *Egr1* and *Slc15a2*), and 10 genes remained classified as M3 in CryKO suggesting that the rhythmic profiles detected in nucleus is not driven by the clock, but by other systemic signals (food, activity). These genes include *Fgfr2*, *Kif13b*, *Zfp871*, *Rdh13*, *Dnajc12*, *Rnf169*, *Rsad1*, *Sh3pxd2a*, *2310022B05Rik*, and *Txndr2*. The overlap of genes classified as M3 is small, but not due to random sampling (hypergeometric test, p-value =0.0006). Surprisingly, we found 174 genes classified as M3 in CryKO. Among these genes, half were classified as M1 in WT (n = 76), such as *Bcr1*, and 54 were classified as M2, such as *Acox2* (Fig.4.31.H). *Sntg2* and *Vmp1* are the only genes that were classified as M4 in WT and became M3 in absence of rhythmic transcription, which would reflect a scenario where a gene remains rhythmically exported despite the loss of rhythmic transcription (Fig.4.31.I).

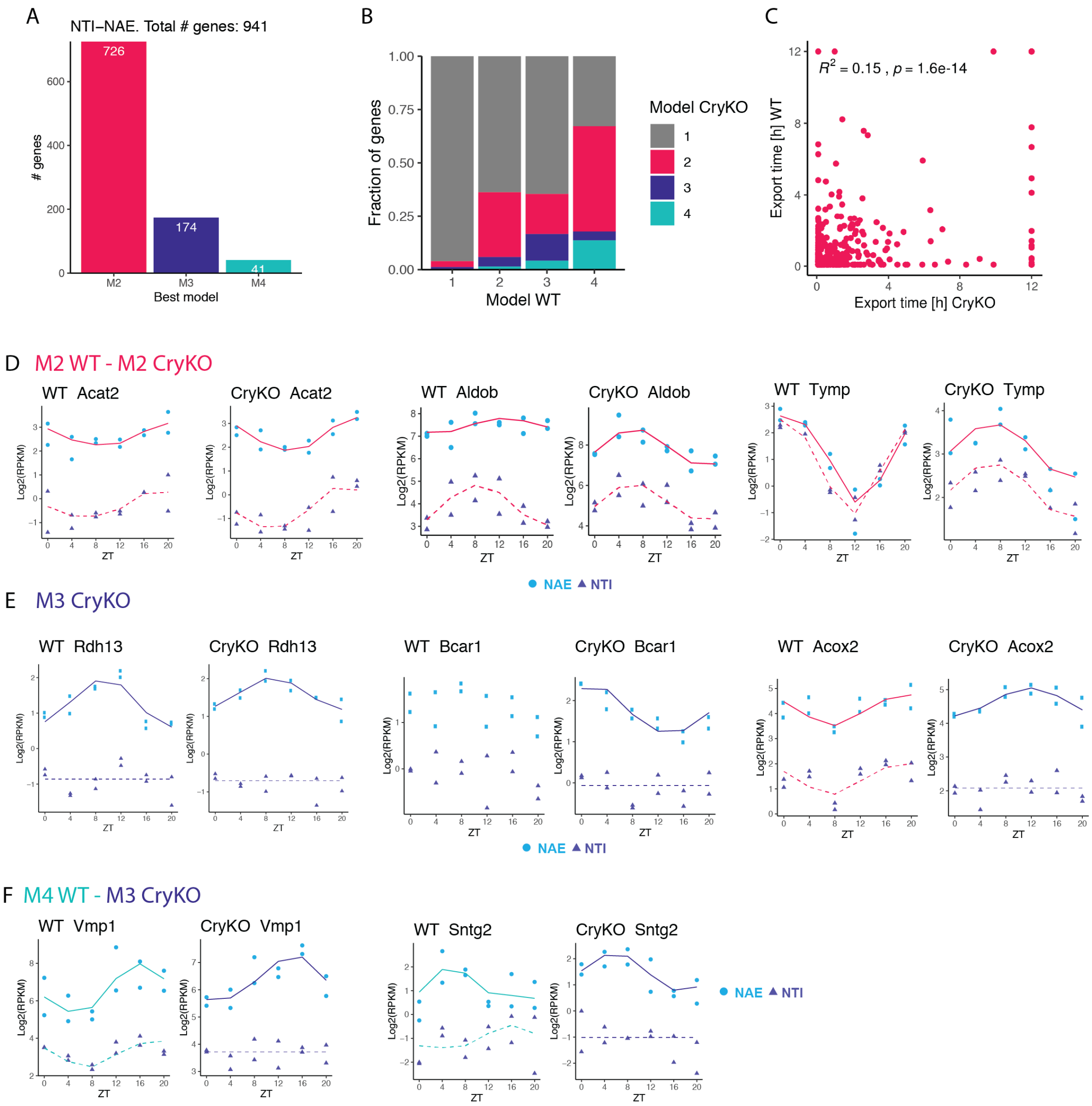


Figure 4.31 – A: Classification of NTI and NAE from CryKO in M2 (rhythmic transcription, constant export), M2 (constant transcription, rhythmic export), or M4 (combination of rhythmic transcription and export). B: Proportion of genes classified as M1, M2, M3 or M4 in CryKO, compared to their classification in WT. C: Distribution of the amplitudes ($\log_2$FC) of NTI and NAE in WT and CryKO shows a global decrease of amplitude in CryKO. P-value NTI: 0.0017, p-value NAE: $< 10^{-10}$, paired t-test. Only genes classified as M2 in both genotypes are shown. D: Correlation of amplitudes in WT and CryKO (left: NTI, right: NAE). E: Phases of WT and CryKO of genes classified as M2 in both genotypes (left: NTI, right: NAE). Circulation correlation: $\rho = 0.13$ and p-value = 0.005, $\rho = 0.16$ for NTI, $8*10^{-4}$ for NAE. F: Correlation of export time estimated for genes classified as M2 in both genotypes (n = 366). G: Temporal fits of NTI and NAE in WT and CryKO. *Acat2* has a similar profile in both genotypes, and similar estimated export time (1.8h in WT and 2.1h in CryKO). *Aldob* has a larger amplitude in CryKO than in WT. Estimated export time: 6.8h in WT, 5 min in CryKO. *Tymp* is delayed by 6 hours in CryKO. Estimated export time: 1h in WT, 5 min in CryKO. H: Genes classified as M3 in CryKO. *Rdh13* has a similar temporal profile in both genotypes. *Bcar1* was classified as M1 in WT, M3 in CryKO. *Acox2* was M2 in WT, M3 in WT. NAE peaks at ZT19. 19.6 in WT, and at ZT15.4 in CryKO. I: Genes classified as M4 in WT, but due to the loss of rhythmic transcription, become M3 in WT.

We next compared the nuclear mRNA and cytoplasmic mRNA to explore rhythmic cytoplasmic degradation.

Out of the 1811 genes classified in M2, M3 or M4 (rhythmic in either nucleus, cytoplasm, or both according to the model classification) in WT animals, 433 were still rhythmic in CryKO. Additionally, 281 genes were rhythmic only in CryKO but not in WT. These genes had a similar range of amplitudes in CryKO than the 433 that were rhythmic in both WT and CryKO, but the median expression level was 2 times lower, hinting that this might be a consequence of more noisy patterns. Out of the 714 rhythmic genes in CryKO, the vast majority (87%) were classified in M2 (Fig.4.32.A). Only 48 genes were classified as M3, and 44 genes as M4. The proportion of rhythmically degraded genes in CryKO is thus two times lower compared to WT (12% versus 25% in WT). As with the phases of NTI and NAE, the distribution of cytoplasmic phases shows a bimodal distribution, enriched around ZT8 and ZT20. The median of estimated half-lives of M2 genes is 2.75h, similar to what has been estimated for WT genes. Globally, these half-lives only poorly correlate with those estimated in WT (Fig.4.32.D), as it was the case in step1. In Fig.4.32.E, we show some representative examples of genes found in M2: *Por* peak at the same time of day in both genotypes (CryKO: ZT11, WT: ZT12). The estimated half-lives are similar (CryKO: 0.75h, WT: 1.0h), and they only differ by the reduced amplitude in the CryKO liver ($\log_2$FC CryKO: 0.9, $\log_2$FC WT: 2.3). The temporal profiles of *Oat* are similar in both genotypes ($\log_2$FC CryKO: 0.8, $\log_2$FC WT: 0.7) and the stabilities are in the same range (CryKO: 1.1h, WT: 1.7h). The main difference is in their peak times, which are almost antiphasic (CryKO: ZT19, WT: ZT9). Finally, *Ttc23* has a shorter half-life in CryKO than in WT.

We next focussed on M3 genes, whose rhythmicity is solely due to rhythmic degradation. Only 8 genes were commonly classified as M3 in both WT and CryKO mice (*Nop53*, *Slc25a42*, *Sox5*, *BC023105*, *Tardbp*, *Enox2*, *Apoo-ps*, *Fam214b*). We also looked for genes that were classified as M4 in WT (rhythmic transcription and rhythmic degradation) and were still rhythmically degraded despite being constitutively transcribed (M3 in CryKO). Only 4 genes met the condition: *Bhlhe40*, *Ppp1r3b*, *Mir6236* and *Scp2-ps2*, but only the first two genes show convincing temporal patterns (Fig.4.32.E).

A clock-deficient mouse model is in theory an excellent way to check whether findings involving rhythms are due to the circadian clock or to other rhythms, e.g. due to the day/night cycle or the feeding/fasting cycle. Here, as there was little overlap between the rhythmic profiles and the classifications between WT and CryKO, and as we had no particular findings to test, we decided not to pursue the analysis of CryKO. However, additional rounds of analysis may reveal post-transcriptional regulation directly related to the clock.

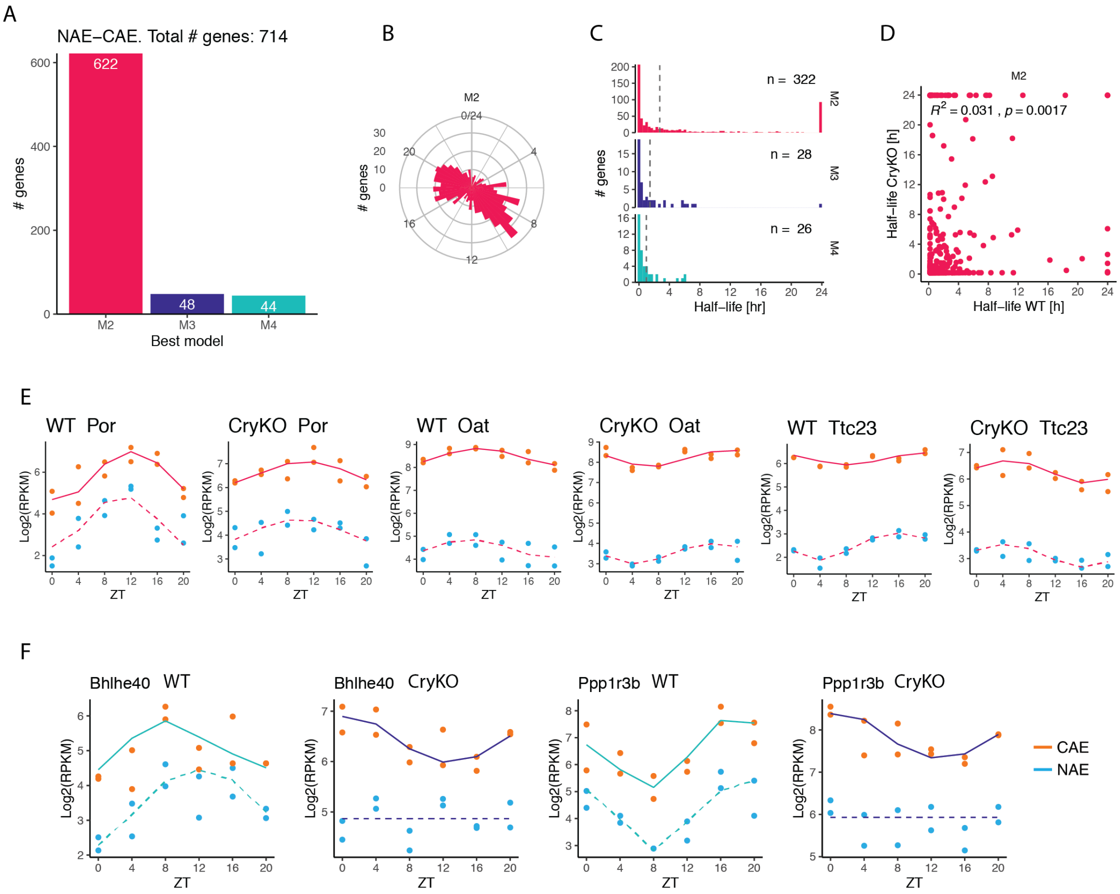


Figure 4.32 – Classification of rhythms in Cry1/Cry2-KO mice. A: Number of genes classified in M2, M3 and M4 based on the comparison of NAE and CAE. B: Distribution of CAE phase of genes classified as M2. C: Distribution of estimated half-lives. The median half-lives is 2.75h for M2, 1.5h for M3 and 1h for M4. Half-lives reaching the upper or lower boundaries were not included to compute the medians. D: Comparison of estimated half-life of genes classified in M2 in both WT and CryKO (n = 317). E: Representative temporal fit of genes classified in M2 in both WT and CryKO. *Por*: similar phase and half-life, but reduced amplitude in CryKO. Half-life WT: 1.0h, CryKO: 0.75h. *Oat*: A large phase delay between WT and CryKO is observed. Half-life WT: 1.05h, CryKO: 1.75h. *Ttc23*: Half-life in WT is much longer than in CryKO (5.1h against 0.4h). F: Temporal RNA-seq profiles of two genes that were rhythmically transcribed and degraded in WT (M4), but constitutively transcribed and still rhythmically degraded in CryKO (M3).

.

4.12.3 Comparison of temporal parameters from single-molecule RNA-FISH and RNA-seq

To substantiate the RNA-seq profiles, we performed smFISH experiments on the same liver tissue and quantified the rhythmic parameters of nuclear and cytoplasmic mRNA.

Because the smFISH protocol was not compatible with immunostaining of the hepatocyte membrane, we could not count the exact number of RNA molecules per cell. Instead, we quantified a "density per nucleus": we divided the total number of detected dots by the number of segmented nuclei. The number of RNA / number of nuclei is only a proxy of the actual number of molecules per cell, because many hepatocytes are binucleated [242]. To assess the rhythmicity of nuclear and cytoplasmic mRNA, we fitted the mRNA counts with a harmonic generalised linear model. We assumed a Negative Binomial distribution of the mean mRNA counts per time-point and per localisation (nucleus, cytoplasm). Additionally, we use a model selection approach (See Methods 6.4.3). In short, the parameters describing rhythmic profiles (a and b) can be either constant or null. Moreover, they can be the same for nuclear and cytoplasmic RNA or independent, and the best model is chosen best on BIC. If a and b are independent, temporal patterns are not the same in the nucleus and in the cytoplasm.

We targeted *Agxt*, a liver-specific gene with a strong diurnal pattern with a large phase delay between nuclear and cytoplasmic RNA accumulation (Fig.4.33). In RNA-seq, the phase and amplitude ($\log_2$FC) of nuclear mRNA (NTE) are ZT 1 and 1.55, and the phase and amplitude of cytoplasmic mRNA (CTE) are ZT 5 and 1.52. By smFISH, the number of RNA transcripts per number of nuclei ranges from 30 to almost 100 in the cytoplasm, and from 10 to 20 in the nucleus. The model with the lowest BIC was the one with different parameters for nuclear and cytoplasmic profiles. Fitted phases were late by 2 hours in smFISH compared to RNA-seq (nuclear: ZT 3, cytoplasmic: ZT 7), and while the cytoplasmic amplitude was similar to the one in RNA-seq (1.3), the nuclear amplitude was 3 times lower (0.5). The low amplitude of nuclear RNA could be because of the particularly low ZT0 sample, which looks like an outlier that acts to decrease the amplitude. Thus, cytoplasmic amplitude was consistent with RNA-seq, but nuclear amplitude was not. Despite the difference of phases of two hours, the robust rhythms and the phase differences between nuclear and cytoplasmic mRNA are well detected by smFISH.

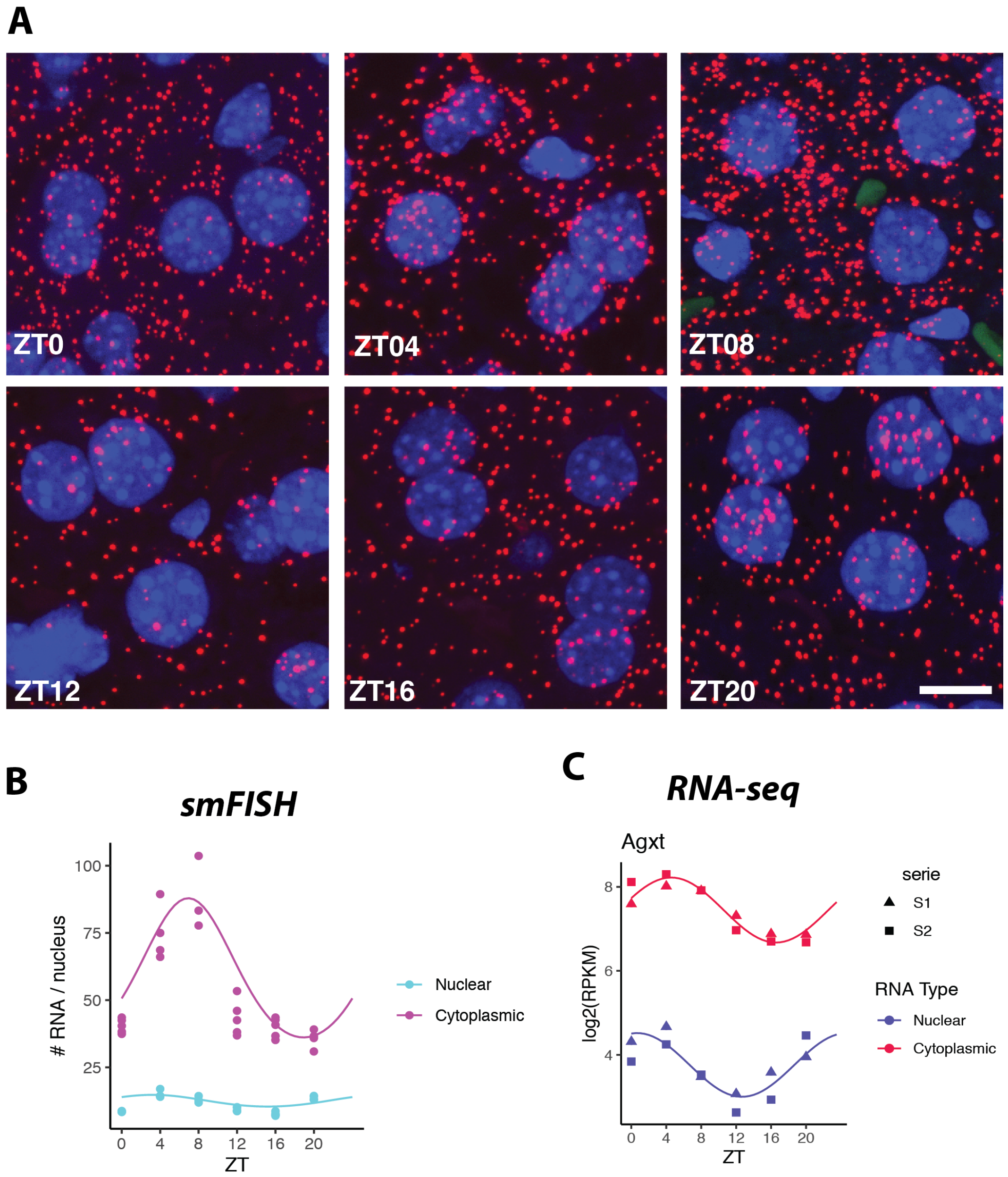


Figure 4.33 – Comparison of temporal characteristics of RNA-seq and smFISH. A: smFISH of *Agxt* on liver FFPE tissue. Representative images at ZT0, 4, 8, 12, 16 and 20. Maximal projection of all z-stacks (8μm). Blue: nuclei stained with DAPI. scale bar: 10μm B: Quantification of mRNA transcripts on smFISH images. A harmonic generalised linear model assuming a negative binomial distribution of the mean mRNA count is fitted to the number of nuclear and cytoplasmic mRNA. (Methods). Only mice from serie 2 were used (n = 6). 5-6 images were taken per animal (technical replicates) and the average number of mRNA molecules per number of nuclei per image is represented by a datapoint. Between 300 and 700 nuclei were segmented by time-point. Only hepatocytes from the midlobular zone were quantified, in order to avoid biases due to zonation. Images were taken with a spinning disk confocal microscope. C: RNA-seq temporal profile of NAE and CAE.

We also quantified the temporal expression of *Actb*, a well known cytoplasmically enriched mRNA (Fig.4.34, Fig.4.4). In RNA-seq, the phase and amplitude of nuclear mRNA (NTE) are ZT22 and 0.6, and the phase and amplitude of cytoplasmic mRNA (CTE) are ZT1 and 0.9. By smFISH, the number of nuclear mRNA per number of nuclei varies from 2 to 15, and from 10 to 150 in the cytoplasm. The temporal profile of nuclear mRNA peaks at ZT0 with a fitted amplitude of 1.3. The cytoplasmic profile peaks at ZT1 with fitted amplitude of 2. The concordance of the phases between RNA-seq and smFISH is in a window of 2 hours. However, the amplitude in smFISH is larger than in RNA-seq, presumably because of the ZT12 sample that has a particularly low number of mRNA.

With these two examples, we showed that rhythmic patterns of nuclear and cytoplasmic mRNA can be detected and quantified by smFISH. The phases could be estimated within a 2 hours window compared to RNA-seq, and the fitted amplitudes were sometimes sensitive to outliers. However, the "conversion rate" of the number of molecules to RPKM varies between *Actb* and *Agxt*. The average CTE expression level of *Agxt* is 7.5 $\log_2$(RPKM) and NTE is 3.75 $\log_2$(RPKM), which corresponds to 50 cytoplasmic and 11 nuclear mRNA molecules per nucleus (median over 6 time points). The average *Actb* expression level is 9.7 $\log_2$(RPKM) in the cytoplasm and 5.9 $\log_2$(RRKM) in the nucleus, corresponding to an average of 70 and 10 cytoplasmic and nuclear mRNA. The RNA-seq quantification suggests that nuclear and cytoplasmic *Actb* RNA are ~4.5 times more abundant than nuclear and cytoplasmic *Agxt* RNA. However, by smFISH, *Actb* is only 1.4 times more abundant in the cytoplasm than *Agxt*. Potential sources of discrepancy include different hybridisation efficiencies for different probes, and experimental batch effects affecting the hybridisation amplification steps, and eventually the imaging setting and image analysis parameters. Importantly, *Agxt* is a zonated gene enriched in the periportal area. Even if we only quantified mRNA from the midlobular zone of both *Agxt* and *Actb*, we actually showed that the expression level of *Agxt* already reaches its highest value in the midlobular zone. Thus, we most likely overestimated the expression level by smFISH compared to the bulk-level quantification by RNA-seq (Fig.4.33). Therefore, while smFISH can be used as an alternative method to confirm rhythmic patterns, one should keep in mind the spatial heterogeneity of a tissue, and that bulk-level quantification may not be representative of *in situ* measurements.

We calibrated the nuclear and cytoplasmic RNA-seq data based on a previously published dataset [74] (see Methods 6.5.3), in order to obtain a nuclear to cytoplasmic ratio as "realistic" as possible, even if the relationship between NAE and CAE is still not absolute. Here, for *Agxt*, there are ~5 times more mRNA detected in the cytoplasm than in the nucleus. By RNA-seq, the $\log_2$ ratio of 2.25 suggests a cytoplasmic enrichment of ~13 times. For *Actb*, cytoplasmic mRNA are ~7 times more abundant than nuclear mRNA, and the cytoplasmic enrichment estimated by RNA-seq is ~14 times. The nuclear to cytoplasmic ratio estimated by RNA-seq and by smFISH differ by a factor 2 to 3. The ratio quantified by smFISH is also not absolute, because we do not sample the same proportion of the nucleus and the cytoplasm. The average nuclear diameter of a tetraploid hepatocyte is 9 μm, while the average hepatocyte volume (including mono- and bi-nucleated cells of 2n and 4n) is around 7000μm^3 [185], with length varying from 15 to 25 μm (personal measurements). With a tissue section of 8μm, we never count mRNA from an entire hepatocyte, and the sampling proportion differs between the nucleus and the hepatocyte. If we roughly assume that a hepatocyte is a cube with an edge of 20μm containing a nucleus of a diameter of 9μm, and that we uniformly sample a section of 8μm, on average, a proportion

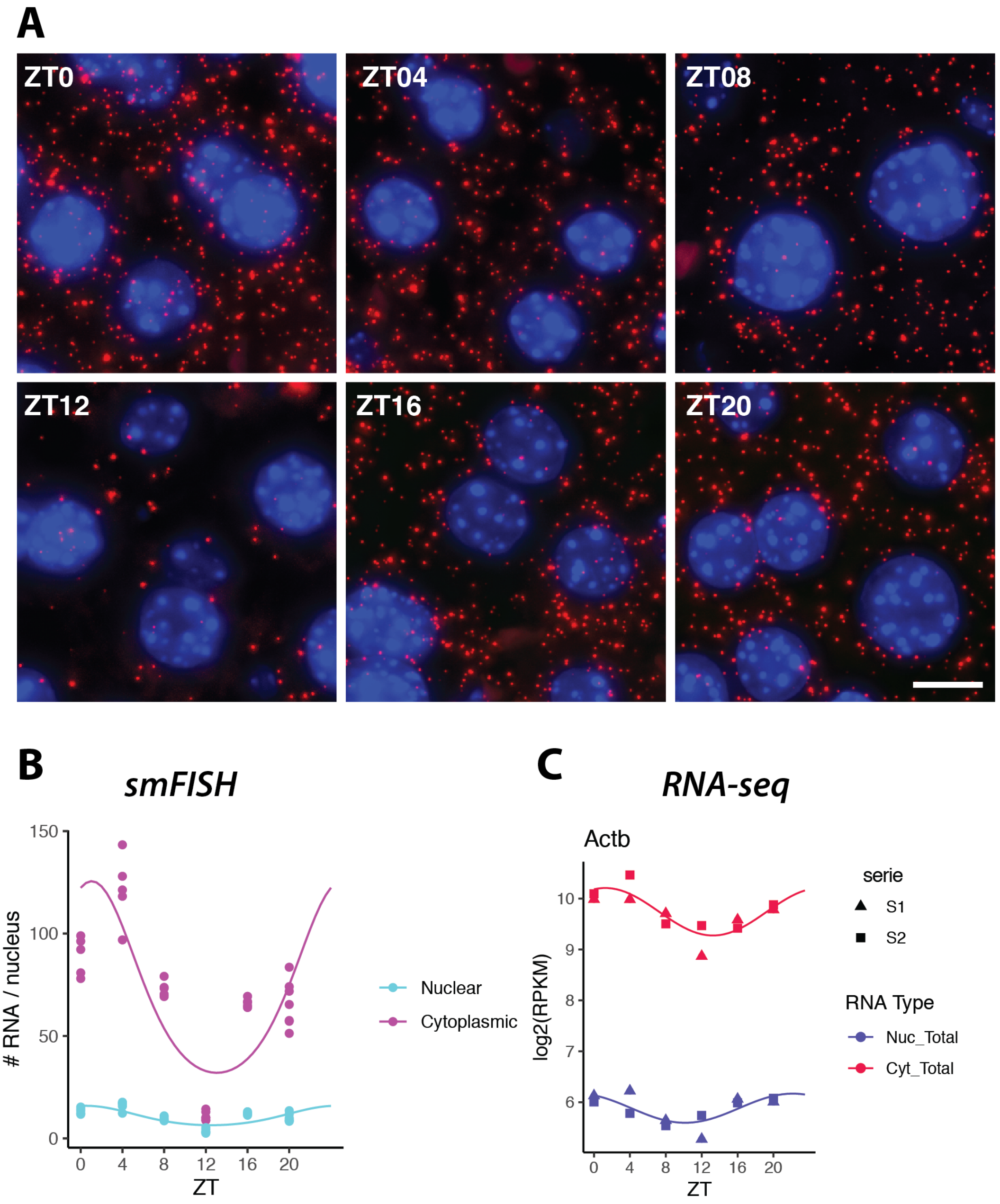


Figure 4.34 – Comparison of temporal characteristics of RNA-seq and smFISH. A: smFISH of *Actb* on liver FFPE tissue. Representative images at ZT0, 4, 8, 12, 16 and 20. Maximal projection of all z-stacsk (8μm). Blue: nuclei stained with DAPI. scale bar: 10μm B: Quantification of mRNA transcripts on smFISH images. Only mice from serie 2 were used (n = 6). 5 to 6 images were taken per animal and per time point. Only hepatocytes from the midlobular zone were quantified, in order to avoid biases due to zonation. C: RNA-seq temporal profile of NAE and CAE

of 0.5 of the nuclei are sampled, but only 0.37 of the cytoplasm. Thus, different sampling proportions of the nuclear and the cytoplasmic content could account for a difference of 30%, and therefore, the nuclear to cytoplasmic mRNA ratio is biased toward slightly higher nuclear value. Moreover, there is a risk that mRNA that are actually above or below the nucleus appear like they are inside the nucleus when projecting all the z-stack images on a 2D plane. It would be interesting to use a 3D reconstruction image software to quantify the error rate when assigning a transcript to the nucleus or to the cytoplasm. Finally, we cannot exclude different probe binding efficiency in the nucleus and in the cytoplasm, as well different backgrounds fluorescent intensities that impair the signal-to-noise ratio. For all the reasons listed above, one should always keep in mind that the nuclear to cytoplasmic ratio, either by smFISH or RNA-seq, are arbitrary. Only the “relative ratio” among different genes can be compared.

4.12.4 Absence of apparent contamination by the endoplasmic reticulum in cuclear fractions

The outer face of the nuclear envelope forms a continuum with the membrane of the endoplasmic reticulum (ER). Therefore, during cellular fractionation, there is a risk that the ER and the associated ribosomes are co-purified with the nuclear fraction [95]. These ER-bound ribosomes specifically translate mRNAs coding for transmembrane and secreted proteins, although transcripts coding for cytosolic proteins have been found to be translated by the same ER-bound ribosomes [243]. Transcripts, especially those part of the secretome, might be wrongly assigned to the nuclear fraction, while they are actually located in the cytoplasm. Therefore, we verified whether ER-translated genes were enriched in the nuclear fractions. To this end, we compared the Nuc/Cyt ratio in both PolyA and Total RNA populations with two published datasets. In the first dataset, Chen et al.[244] fractionated J558 murine plasmacytoma cells by sequential detergent extraction method, and isolated cytoplasmic and ER-bound RNAs. The expression level of 68 selected genes were measured in each population by qPCR, and a cytosolic vs ER enrichment score was attributed to each gene. We did not observe any correlation between (Fig.4.35.A and B). We did not observe any significant correlation between the nuclear localisation and the ER-enrichment score in both PolyA and Total RNA populations, indicating that transcripts translated by ER-bound ribosomes are not enriched in the nucleus. In the second dataset (HEK293 cells), ER-bound ribosomes were biotinylated and pulled down, followed by the sequencing of the associated mRNA (proximity-specific ribosome profiling) [245]. They assigned an enrichment value based on the $\log_2$ ratio of ribosome footprint in the biotinylated versus whole-cell sample. Again, no correlation between nuclear enrichment and ER-enrichment was found (Fig.4.35.C, D). Finally, we also looked at the subcellular localisation of genes coding for protein annotated as “highly secreted” by MetazSecKB, based on prediction by SignalP4, Phobius, TargetP and WoLF PSORT. The $\log_2$ ratio of NAE/CAE of genes annotated as “highly secreted” was not significantly different from those not annotated as such in both PolyA and Total population (Fig.4.35.E, F). Based on these results, we conclude that there is no detectable contamination of the nuclear RNA population by the mRNA translated on the endoplasmic reticulum.

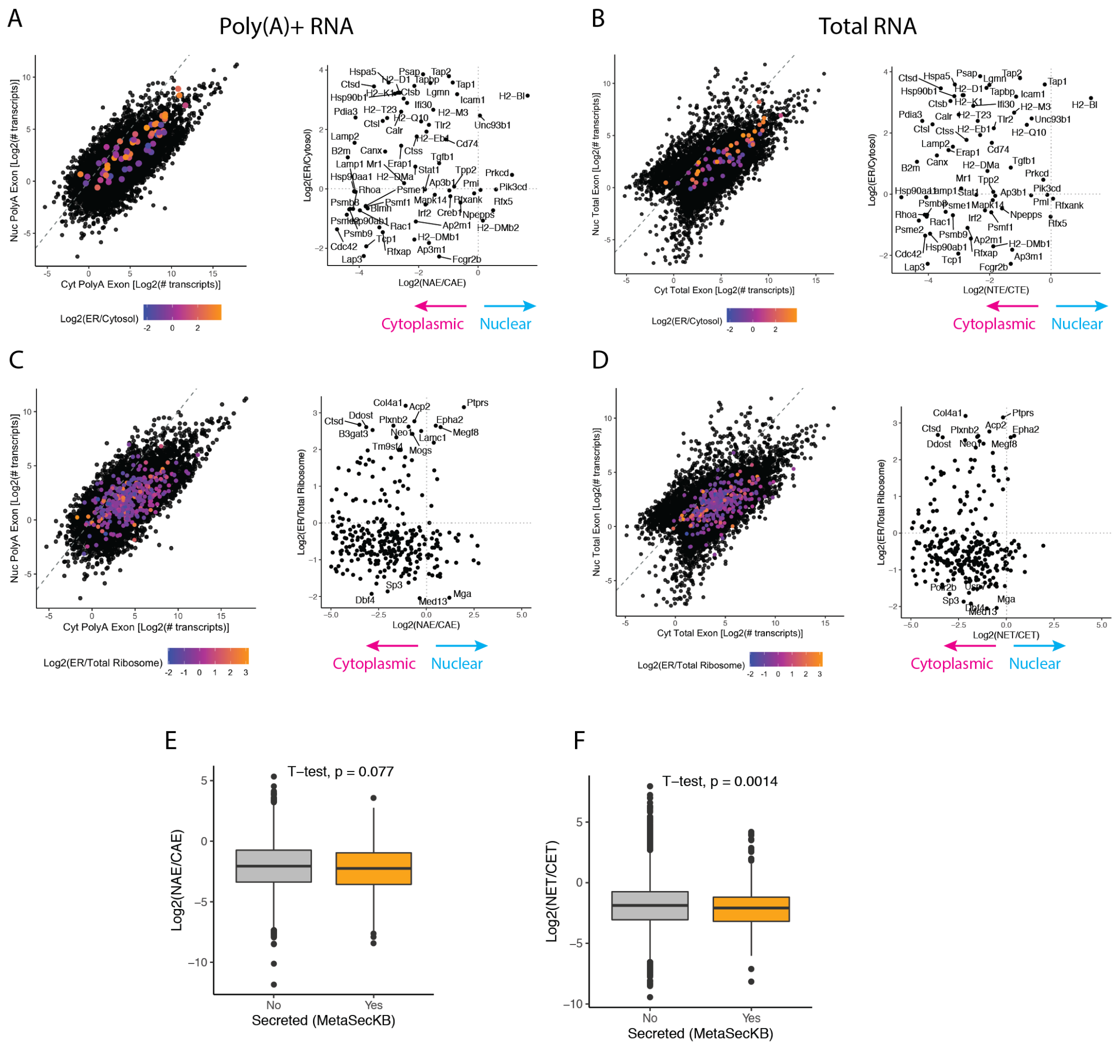


Figure 4.35 – Nuclear fractions are not enriched for transcripts translated by ER-bound ribosomes. A: Left: scatterplot of $\log_2$(NAE) and $\log_2$(CAE). Colored dots represent genes from [209], color-coded by their enrichment score (orange = cytosolic, blue = ER-enriched). No correlation was found between the ER enrichment score and the nuclear localisation (p-value: 0.2, Pearson's correlation R^2: 0.02). B: same as A, but comparing Total RNA (NTE and CTE). p-value = 0.2, R^2 = 0.02. C: Comparison with the dataset published by Jan et al. in HEK243 cells [245]. Again, no correlation is found between the ER-enrichment score and the Nuc/Cyt ratio in both PolyA and Total RNA populations (p-value = 0.6 and p-value = 0.9). E,F: $\log_2$(NAE/CAE) and $\log_2$(NTE/CTE) ratio of genes predicted as "Highly secreted" by MetaSecKB database ("YES"). Mean ratios were not significantly different between genes annotated as "highly secreted" ("YES") and the other genes ("NO").

5 Circadian and chromatin contacts-dependent modulation of transcriptional bursting parameters

5.1 Background

Due to the stochastic nature of transcription, a majority of mammalian genes are transcribed in bursts [246, 70, 73]. Indeed, most RNAs are produced during limited time periods followed by longer periods of transcriptional inactivity. Although it is a ubiquitous phenomenon, genes are characterised by transcriptional bursting parameters that vary considerably from gene to gene, in a tissue-specific manner [196]. Typically, these parameters are the *burst frequency*, which is the rate of switching between periods of transcriptional inactivity and activity, and the burst *size*, which is the average number of RNAs transcribed per burst episode.

smFISH distributions can be used to precisely describe the mode of transcription of a gene and notably provides information regarding its transcriptional bursting properties [72, 75]. If transcription is constitutive, the expected distribution of mRNAs per cell follows a Poisson distribution, while if occurring in bursts, the variance of the distribution is often greater than expected with a Poisson distribution, and the bursting model, which can be approximated with a negative binomial distribution, can increase the amount of variance (known as overdispersion). The number of actively transcribing loci per nucleus can be estimated using an intronic probe signal and is known as the burst frequency, while the intensity of the dot at the transcription site presents a good approximation of the burst size [72]. The burst frequency is inversely proportional to the expression noise (the noise is reduced if the frequency is high).

Transcriptional bursting parameters depend on many aspects of gene regulation such as local chromatin environment, histone modification and DNA looping [71]. Also, circadian genes are known to display changes in bursting parameters throughout the day. Notably, in cultured cells, rhythmic variation of burst frequency but not burst size has been shown to modulate the rhythmic expression of various circadian reporters [73, 75]. Moreover, in the *Bmal1* core clock gene, the daily increase of burst frequency was positively correlated with the presence of acetylated histones at the promoter.

For this project, I collaborated with Jérôme Mermet and Jake Yeung, first co-authors of [64], to measure expression and transcriptional bursting parameters of *Cry1*, a core clock component. They had previously found that in this gene, BMAL1-dependent oscillatory promoter-enhancer interactions participated in modulating rhythmic expression in the mouse liver. Deletion of an enhancer that was rhythmically recruited to the *Cry1* promoter was sufficient to decrease the rhythmic chromatin contacts. The deletion also reduced the peak of *Cry1* mRNA expression profile, and shortened the circadian period of locomotor activity of mice (*Cry1*Δe strain).

Since *Cry1* was (1) differentially expressed throughout the day, (2) displaying rhythmic histone acetylation levels at its promoter and (3) regulated by promoter-enhancer contacts, we thought of assessing the variations in its bursting pattern at two time-points corresponding to its expression peak and trough. By using smFISH probes targeting intronic and exonic region of *Cry1* and inferring bursting parameters in WT and *Cry1*Δe mice at two times of the day, I showed that the burst frequency, but not the burst size, regulates the rhythmic gene expression level of *Cry1*. The frequency is also modulated by rhythmic promoter-enhancer contacts. Thus, as for reporter expression in cultured cells, the rhythmic expression of an endogenous gene in the liver is also exclusively linked to modulations of the burst frequency, and these bursting properties are directly linked to the presence of dynamic promoter-enhancer contacts. Taken together, the results presented in this study established oscillating and clock-controlled promoter-enhancer looping as a regulatory layer underlying circadian transcription (through modulation of burst frequency) and behavior.

5.2 Results

To quantify the expression of *Cry1* mRNA and estimate its bursting features, we first performed smFISH in liver cryosections at the peak of expression (ZT20) and trough (ZT8), and in WT and *Cry1*Δe animals (Fig.5.1). Because the simultaneous staining of hepatocyte membranes was not compatible with the smFISH protocol, we did not retrieve the absolute copy number of mRNA per cell, nor did we take into account binucleation. In addition, the thickness of the cryosection (8μm) is smaller than the diameter of hepatocytes (~25μm), and does not capture whole cells. Instead, we measure a “concentration” of mRNA related to the number of segmented nuclei per microscopy image. We were thus able to compare the relative variation of RNA concentration between conditions. On average, mRNA concentration varies from 4 to 20 mRNA/nucleus in WT, and from 7 to 13.7 mRNA/nucleus in *Cry1*Δe (Fig.5.1.b). The peak accumulation is reduced by 31% in *Cry1*Δe compared to WT. At ZT8, *Cry1*Δe level is slightly higher than WT level. RNA-seq was performed on the same liver samples used for smFISH (Fig.5.1.c), and shows very similar patterns: a reduction of 27% of expression between the two genotypes at their peak expression time (ZT20), and a slightly higher level at ZT8 in *Cry1*Δe. The latter is explained by the phase advance of *Cry1* expression in *Cry1*Δe compared to WT (on average, clock genes are phase advanced by 30 minutes). Thus, we were able to recapitulate RNA-seq quantification using smFISH as an alternative approach.

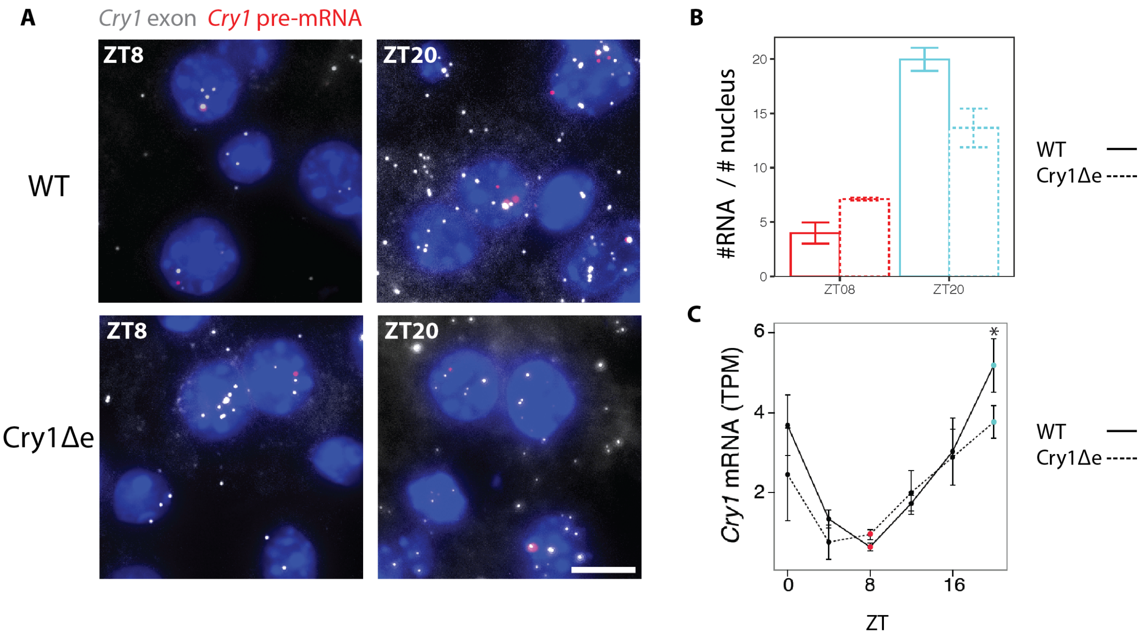


Figure 5.1 – smFISH against *Cry1* pre-mRNA in the liver of WT (top) and *Cry1*Δe (bottom) animals at ZT8 (left) and ZT20 (right). B: Quantification of Cry1 transcripts (exon) in smFISH images. Unit is the number of RNA transcripts divided by the total number of nuclei in each microscopy image. Shown are mean and standard errors over two animals. C: RNA-seq *Cry1* mRNA profiles in *Cry1*Δe and littermates WT animals (Transcripts per Million, TPM). Bars indicate standard deviation over 3 animals (on courtesy of Jake Yeung).

Next, we asked whether time of the day and the *Cry1* intronic enhancer could modulate transcriptional bursting parameters. While counting mRNAs in the cytoplasm uses exonic probes, analysis of transcriptional bursting is best done with intronic probes. We designed smFISH probes targeting *Cry1* pre-mRNA to detect active transcription sites (TSs), and estimated the transcriptional bursting parameters at ZT20 and ZT8. We defined the *burst fraction* as the fraction of active TSs in each nucleus, proportional to the burst frequency per allele. We used the *burst intensity*, which is the mean intensity of a TSs instead of the burst *size* (the actual number of RNA being transcribed), which we were able to compare across conditions. Because hepatocytes are polyploid cells harbouring 2, 4, 8 or even 16 copies of their genome, we first estimated the ploidy of each nucleus based on the size [185]. We fitted a Gaussian mixture model with four clusters to represent diploids, tetraploid and octoploids nuclei (Fig.5.2.a). The last cluster with an extremely large variance accounts for the outliers and is not used in the analysis. Ploidy distribution changes in function of age [186], thus, the proportions vary from one study to another. However, for mice older than 8 weeks old, the majority of nuclei is tetraploid [196, 186, 247]. In our case, the most abundant hepatocytes are indeed tetraploids (67%, average diameter of $9.38 \pm 0.89\mu$m), followed by octoploids (21%, $13 \pm 0.81\mu$m) and diploid cells (11%, $6.80 \pm 0.38\mu$m) (Fig. 5.2).

We modeled the number of active TSs per nucleus with genotype-dependent slopes and compared it to a reduced model without a genotype effect (Fig.5.3.A), and showed a significant difference at ZT20. We divided the number of active TSs by the estimated ploidy to obtain the burst fraction (Fig.5.3.B). The estimated burst fraction was 2.4 times higher at ZT20 compared to ZT8 in WT and 1.92 in *Cry1*Δe

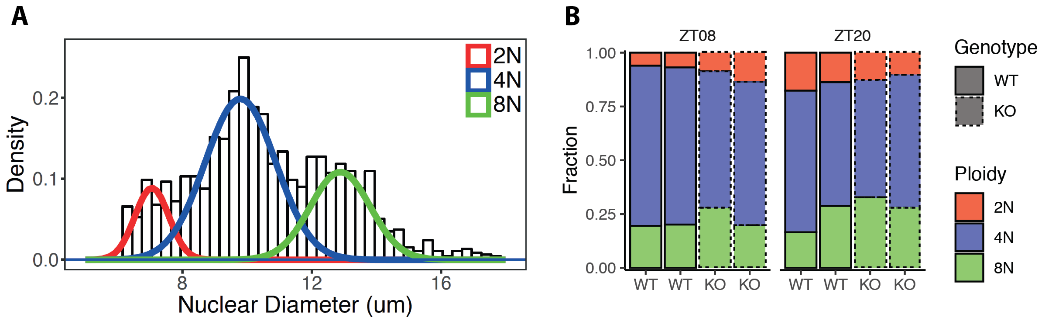


Figure 5.2 – Ploidy distribution. A: Distribution of nucleus diameters for a representative animal. Colored curves show Gaussian mixture model corresponding to diploids (2N), tetraploid (4N) and octoploid (8N) nuclei. The fourth curve catching outliers is not shown. B: percentage 2N, 4N and 8N nuclei in each animal (n=2 per time point and genotype).

(Fig.5.3), indicating a time-dependent modulation of gene expression by varying burst frequency. Importantly, the burst fraction was reduced by 28% in *Cry1*Δe, a value matching the observed reduction of mRNA accumulation in RNA-seq and smFISH. For the burst size, as for the burst fraction, we modeled the mean intensity of intronic dots with genotype-dependent intercepts and compared it to a reduced model. By contrast to the burst fraction, the burst intensity was similar in all conditions (Fig.5.3.D,E). Thus, the decreased *Cry1* mRNA levels in *Cry1*Δe at ZT20 can be quantitatively explained by the lowered burst fraction. In summary, dynamic DNA loops involving clock enhancers modulate transcriptional bursting in mammalian tissues, specifically, rhythmic enhancer control burst frequency while maintaining burst size.

5.3 Concluding Remarks

Most studies assessing transcriptional bursting parameters from smFISH data either fit a negative binomial to the distribution of transcripts per cell [72, 73], or quantify the number and signal intensity of TSs together with the number of transcripts per cell [248, 249]. In our imaging conditions, these approaches could not be considered due to the impossibility to simultaneously perform smFISH and mark cell boundaries. Consequently, in this study, we estimated the burst size (or burst intensity) from the intronic signal intensity at TSs, and the burst fraction from the number of active TS per loci (after assigning every nuclei to a ploidy). The burst size was estimated from the relative change in intronic signal intensity at TS rather than the absolute number of transcripts at TS. It is theoretically possible to identify active TS from the intronic signal and quantify the transcription intensity with the co-localising exonic signal normalised to the intensity of a single RNA molecule, for instance in the cytoplasm, but the approach is imprecise notably because of the important background variations in different regions of the cell. Also, our smFISH labelling strategy uses an amplification step that makes it difficult to estimate the fluorescence intensity of a single mRNA. It is also worth noting that estimating the number of transcripts at TSs is also challenging using more conventional smFISH techniques lacking signal amplification steps, as in these conditions, the normalised signal of a single transcript also varies by 2 to 3 folds [250].

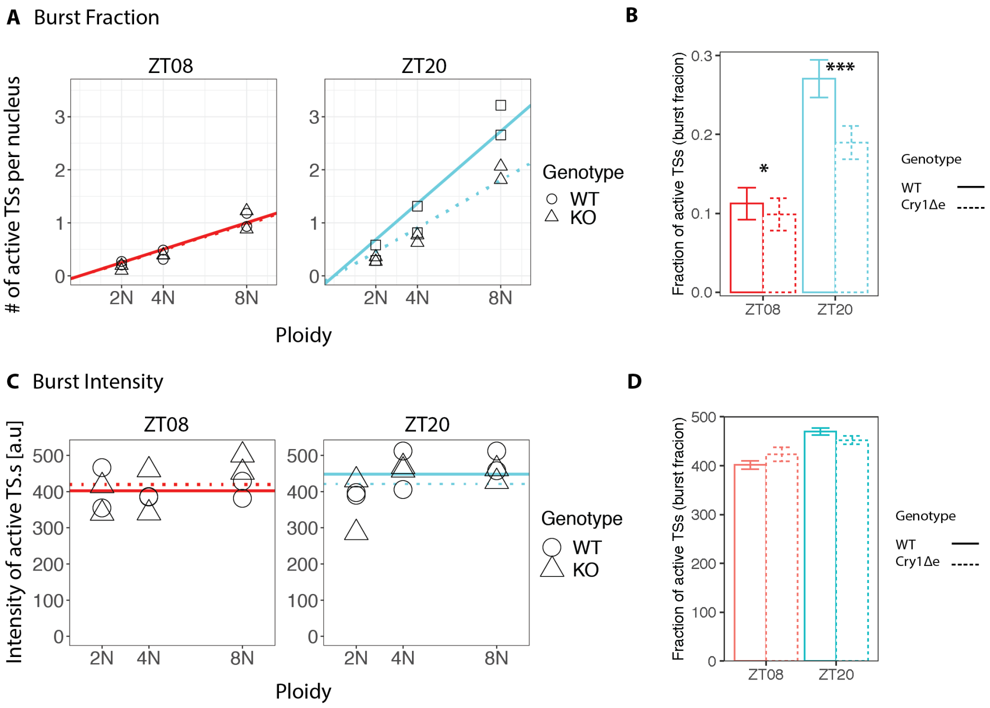


Figure 5.3 – Burst frequencies and burst intensities measured from images of smFISH performed against *Cry1* pre-mRNA in *Cry1*Δe (dashed) and WT (solid) at ZT8 (red) and ZT20 (blue). A: Number of active transcription sites (TSs) averaged per animal increases with ploidy. At ZT20, *Cry1*Δe animals show reduced number of TSs compared to WT: lines show mixed-effect linear model with genotype-dependent slopes. P-value testing the null hypothesis that slopes are equal: 0.00014 (F-test). At ZT8, the slopes are not different (p=0.84). B: Number of active TSs in each nucleus divided by the estimated ploidy. Shown are means and standard errors over nuclei collected from two animals in each of the four conditions. Number of nuclei: WT-ZT8 n=2191; *Cry1*Δe-ZT8 n=983; WT-ZT20 n=2150; *Cry1*Δe-ZT20 n=1473. In B, * and *** show $p<0.05$ and $p<0.001$, respectively, T-test. C: Active TS intensity averaged per animal shows comparable intensities across ploidy and conditions: lines show mixed effect model with genotype-dependent intercepts, intercept comparisons at both ZT8 (p(H_0:equal intercept)=0.53, F-test) and ZT20 (p=0.41) are not significant. In D, differences between genotypes are not significant.

Despite these limitations, our approach provides a reliable estimate of the bursting properties of an endogenous core clock gene in mouse liver, and how they vary between time points and experimental conditions. These bursting properties of *Cry1* are likely to vary in other systems, since bursting parameters are tissue-specific [196]. Our analysis confirmed that changes in expression levels in circadian genes throughout the day mainly arise from modulation of burst frequency. While this tendency had already been observed in cultured cells [71, 76], it was so far never confirmed in tissues. Also, our data highlighted the importance of DNA looping in this process. The formation of enhancer-promoter contacts influences the burst frequency together with other factors such as histone acetylation [71] or nucleosome density [251, 252].

5.4 Clock-dependent chromatin topology modulates circadian transcription and behavior: Abstract

This work was part of the paper published in Genes and Development by the following authors (first authors in bold): **Jérôme Mermet, Jake Yeung**, Clémence Hurni, Daniel Mauvoisin, Kyle Gustafson, Céline Jouffe, Damien Nicolas, Yann Emmenegger, Cédric Gobet , Paul Franken, Frédéric Gachon, Félix Naef [64].

The circadian clock in animals orchestrates widespread oscillatory gene expression programs, which underlie 24-h rhythms in behavior and physiology. Several studies have shown the possible roles of transcription factors and chromatin marks in controlling cyclic gene expression. However, how daily active enhancers modulate rhythmic gene transcription in mammalian tissues is not known. Using circular chromosome conformation capture (4C) combined with sequencing (4C-seq), we discovered oscillatory promoter–enhancer interactions along the 24-h cycle in the mouse liver and kidney. Rhythms in chromatin interactions were abolished in arrhythmic Bmal1 knockout mice. Deleting a contacted intronic enhancer element in the Cryptochrome 1 (Cry1) gene was sufficient to compromise the rhythmic chromatin contacts in tissues. Moreover, the deletion reduced the daily dynamics of *Cry1* transcriptional burst frequency and, remarkably, shortened the circadian period of locomotor activity rhythms. Our results establish oscillating and clock-controlled promoter–enhancer looping as a regulatory layer underlying circadian transcription and behavior.

6 Methods

6.1 Mouse Liver sampling

All mice experiments were performed on male mice aged from 8 to 12 weeks old, and housed in 12:12 light:dark cycle. 3 days prior sacrifice, mice were given access to food only during the active phase (night-restricted feeding). WT mice were C57/BL6J mice and Cry1/Cry2 double-KO (later referred to as CryKO) are described in [240]. Breeders genotype was Cry1 -/- and Cry2+/-. For the project on Transcriptional Bursting, Cry1Δe mice were generated by the EPFL Transgenic core facility and described in [64].

For the project on liver zonation (Chapter 3), 4 WT mice were sacrificed by decapitation every 3 hours (ZT0, 3, 6, 9, 12, 15, 18 and 21). ZT0 is the beginning of the light phase, and ZT12 is beginning of the dark phase. For the project on subcellular localisation (Chapter 4), 2 WT mice and 2 CryKO mice were sacrificed by decapitation every 4 hours (ZT0, 4, 8, 12, 16, 20).

Pieces of liver were immediately collected after sacrifice for histological analysis (single-molecule RNA-FISH and immunostaining) and for RNA extraction.

6.2 Cellular fractionation of liver tissue and RNA extraction

For nuclear and cytoplasmic RNA extraction, a piece of ~300mg was dissected from the big lobe. Cellular fractionation was done as described in [253]. Liver pieces were put in 15ml Falcon tube filled with ice-cold PBS up to a total volume of 2.9ml, then transferred into a Potter-elvehjem homogeniser, and homogenised with a PTFE pestle with few milliliters of a 2.2M sucrose homogenisation buffer (total volume of homogenisation buffer: 18ml). The homogenate was transfered into the remaining homogeneisation buffer and layered on top of 7.2ml of a 2.05M sucrose cushion buffer in ultra-centrifuge tubes. The liver homogenate was ultra-centrifuged in a Beckman SW28 rotor at a speed of 24K rotation per minute, at 4C° for 1 hour. 10 ml of the supernatant (cytoplasmic extract) was taken. From the 10ml of the cytoplasmic extract, 2 ml was poured in 8ml of RNAse-free H2O and 30ml ethanol (final EtOH concentration of 70%) and frozen overnight at -80C°. The remaining cytoplasmic was stored at

-80C°. The next day, the cytoplasmic extract in ethanol was centrifuged at 5000 rpm for 15 min at 4C° to precipitate RNA. Ethanol was discarded, and the pellet was resuspended in 800μl of Qiazol (Qiagen). RNA was extracted on spin columns using the miRNEasy Mini Kit (Qiagen) according to manufacturer's instructions.

After taking 10 ml of the cytoplasmic fraction following the ultracentrifugation, the remaining sucrose buffer was discarded. The wall of the tubes were cleaned with a tissue to remove all the remaining sucrose buffer. The pellet (nuclei) was resuspended in 3ml of Nuclear RNA extraction buffer, and was mechanically homogeneised with a syringe. 3ml of water-saturated phenol (pH 4.5) and 1200μl of chloroform:isoamyl alcohol 24:1 were added to the resuspended pellet, and were well shaken before resting for 15 minutes at room temperature. The mix was centrifuged at 4C° at 4000rpm for 20min. The supernatant was resuspended in ethanol (700 μl of supernatant for 1000μl of ethanol 100%). RNA was then extracted on spin columns using miRNEasy kit from Qiagen according to manufacturer's instructions, including a gDNAse treatment (Qiagen).

For total liver RNA extraction, pieces from the main liver lobes were frozen in liquid nitrogen. About 20mg were grounded and were used for RNA extraction using the miRNEAsy kits (Qiagen).

Composition of 2.2M homogenisation buffer: 2.2M sucrose, 15mM KCl, EDTA, 10mM Hepes pH 7.6, 0.15 mM spermine, 0.5 mM spermidine, in RNAse-free water. Just before use, add 1/200 PMSF, 1/100 DTT 0.1M, and 1/100 Protease inhibitor cocktail (aprotinin, leupeptin, pepstatin). Composition of the 2.05M sucrose cushion buffer: 2.05M sucrose, 10% glycerol, 15mM KCl, EDTA, Hepes Ph.7.6, 0.15mM spermine, 0.5mM spermidine, in RNAse-free water. Just before use, add: 1/100 Protease Inhibitor (aprotinin, leupeptin, pepstatin), 1/200 PMSF, 1/100 DTT 0.1M. Composition of the Nuclear RNA extraction buffer : guanidium thiocyanate (50% of the final weight), Na citrate 0.75M pH7, in RNAse-free water. Just before use, add: 1/10 NaAcO 2M pH4 and 1/100 β-mercaptoethanol.

The quality of extracted RNA was assessed with the Nanodrop (concentration, 260/280 and 230/260 ratio) and with the Agilent Tapestation 4200 (automated electrophoresis). All the RNA had a RIN value ranging from 7.5 to 9.5. Cytoplasmic RNA, particulary in WT, had the lowest quality.

6.3 Single-molecule RNA-FISH

Dissected liver pieces were immediately embedded in O.C.T Compound (Tissue-Tek; Sakura-Finetek USA), snap-frozen in isopentane cooled with dry ice, and stored at -80C° (fresh-frozen samples). Other pieces of liver were fixed in 10% Neutral buffered Formalin (NBF) at 4C° for 24 to 36 hours. Fixed samples were then washed in PBS 1x for 30 minutes, and dehydrated in standard ethanol series followed by xylene bath, and finally embedded in paraffin (formalin fixed paraffin embedded, FFPE tissues). Sections were 8μm thick.

Single-molecule RNA Fluorescent *in situ* Hybridization (smFISH) experiments were all performed using the RNAScope technology (Advanced Cell Diagnostics). smFISH experiments were performed by the Histology Core Facility at EPFL, according to manufacturer's instructions. Housekeeping genes *Ppib*, *Ubc* and *Polr2A* were used as positive controls, and *DapB*, from the *Bacillus subtilis* strain SMY was used as an internal negative control. Nuclei were counter-stained with DAPI or with Dracq5.

smFISH experiments with *Arntl* and *Per1* was performed together with an immunofluorescence against Glutamine Synthetase, a marker of the pericentral vein (ab49873, Abcam, diluted 1:2000 in PBS/BSA and 0.5%/Triton-X0.01%). Sections were mounted with ProLong Gold Antifade Mountant. For fresh-frozen liver cryosections, the RNAScope Fluorescent Multiplex Assay was used. For FFPE samples, the RNAScope Fluorescent Multiplex V2 Assay was used.
Catalog number of the probes: *Cry1* pre-mRNA: 500231. *Cry1*: 500031. *Per1*: 438751. *Arntl*: 438741. *Actb*: 316741. *Agxt*: 525261. *Mlxipl*: 558141.

6.4 Microscope image acquisition and quantification of single-molecule RNA-FISH images

Two different microscopes were used depending on the project. For the Transcriptional Bursting project and the Liver Zonation project (Chapter 3 and Chapter 5), cryosections were imaged with the Leica DM5500 widefield microscope and a motorised-stage. Z-stacks were acquired with a distance of 0.2μm between each Z position, (~40 images per frame) with an oil-immersion x63 objective. For the Subcellular Localisation project (Chapter 4), all samples were FFPE. Sections were imaged with the Visitron Spinning Disk CSU W1 with a motorized stage.Z-stacks were acquired with a distance of 0.2μm between each Z position, (~40 images per frame) with an oil-immersion x63 objective.
All z-stacks were maximally projected for analysis. The image processing pipeline was developed using ImageJ. mRNA spots are detected by first applying a Gaussian blur, followed by the edge-detector Laplacian filter. The local maxima - corresponding to the spots - are computed and counted. Nuclei are detected by applying a median blur filter, the Otsu method for automatic thresholding, and the watershed algorithm for segmentation. We were not able to stain the hepatocytes membrane together with the smFISH protocol: none of the antibodies usually employed to stain the membranes were compatible with the smFISH protocole (E-Cadherin (ab76055), N-Cadherin (ab76057), Pan-cadherin (ab6529, ab16505), β-catenin (ab32572, D10A8), F-actin binding Phalloidin (A12379)). Therefore, we did not quantify the number of mRNA molecules per cell. Instead, we quantified a density of mRNA per nuclei: we counted all the mRNAs present in one microscopy image and divided by the number of segmented nuclei.

6.4.1 Quantification of spatio-temporal mRNA profiles

mRNA and nuclei were detected and counted as mentionned above. Each microscopy image contains one central vein (CV) and one portal vein (PV). PV and CV were manually detected based on the presence or absence of bile ducts, detected by DAPI staining. To facilitate the identifcation of the veins when targeting putatively non-zonated genes (*Per1* and *Arntl*), we additionally performed an immunostaining of Glutamine Synthetase (Abcam ab49873, dilution 1:2000), a marker of CV. The contour of the veins were manually drawn in ImageJ. Endothelial cells lining the veins and cholangiocytes forming the bile ducts were excluded from the analysis. The Euclidean distance between two veins (d) and the distance from the veins of each mRNA transcript and of each nuclei were calculated. mRNA transcripts were assigned to the periportal (PP) or pericentral (PC) zone when the distance from the corresponding vein was smaller than one-third of d. If the mRNA and the nuclei were at a distance of $\frac{2}{3}d$ from both the CV and the PV, they were assigned to the midlobular zone (Mid). d ranges from 50 to 130μm (Fig.3.1).

6.4.2 Modeling spatio-temporal profiles

To assess the rhythmicity in the three liver zones (PC, PP and Mid), we built a simple model selection framework based on generalised linear model. We fitted a harmonic regression on the mRNA counts per image with a log-link function, and used the number of detected nucleus per zone as an offset in order to get the density of "number of mRNAs per number of nuclei". We assumed that y_t is the number of mRNA per zone and per time-point, which follows a negative Binomial distribution. We used the relation:

$$y_{z,t} \sim NB(\mu_{z,t}, \theta)$$
$$log_2(\mu_{z,t}) = m_z + a_z cos(\omega t) + b_z sin(\omega t) + log_2(N_z)$$

with $m_{z,t}$ the number of mRNA in the zone z per time point t, N_z the number of nuclei in the zone z, a_z and b_z the coefficient of the cosine and sine functions, ω the frequency ($\frac{2*pi}{24}$). The phase (peak of expression) is defined as arctan($\frac{b}{a}$) and the amplitude (or $\log_2$ fold change) as $\sqrt{a^2+b^2}$. To compare rhythmicity parameters and the mean expression level in the three zones (m$_z$, a_z and b_z), we used a model selection approach. We allowed a_z and b_z to be zero (non-rhythmic), nonzero (rhythmic), or to be shared between PP, PC and/or Mid. If a_z and b_z are shared between PC, PP or Mid, then the phase and log$_2$FC are the same in the three zones, and the model is more parsimonious. The mean mRNA count m_z can also be shared (no difference of expression level between zone) or independent, in which case the profile is zonated. The Baysian Information Criterion (BIC) is then calculated to account for model complex, and the model with the lowest BIC is chosen. BIC = $-log(L) + K * log(N)$, with $log(L)$ the log-likelihood, K the number of parameters, and N the number of data points.

6.4.3 Quantification of nuclear and cytoplasmic temporal mRNA profiles

We detected mRNA and nuclei as mentioned above. Additionally, each mRNA was labelled as "nuclear" if it belongs to a segmented nucleus, or cytoplasmic otherwise. To assess the rhythmicity of nuclear and cytoplasmic mRNA profiles, we used the same model selection framework as for the zonation profiles. Instead of using zone-specific parameters to describe the temporal profile (m_z, a_z, b_z), we used "localisation-specific" parameters (nuclear or cytoplasmic). These parameters can be either zero (flat profile) or non-zero (rhythmic), and can be shared (same phase and amplitude in the nucleus and in the cytoplasm), or independent. The best model is chosen best on the BIC.

6.4.4 Quantification of transcriptional bursting parameters

For the estimation of *Cry1* transcriptional bursting parameters (burst size and burst fraction), we counted the number of active transcription sites (TSs) per nucleus. To detect TSs, we designed smRNA-

FISH probes targeting *Cry1* pre-mRNA (RNAScope catalog number: 500231, targeting the 1047-2454 region of intron 1). To estimate burst size, transcription site intensities were quantified on the sum projection of the ten best focused stacks per image. Total transcription site signals were computed using a mask of 3x3 pixels. Burst fraction was calculated as the number of active transcription sites in each nucleus divided by its estimated ploidy, and these fractions were then averaged over the entire populations of nuclei. We assigned the ploidy (2N, 4N, 8N) based on the nuclear diameter. Typical values of nuclear diameters reported in the literature are approximately 7μm (2N), 9μm (4N) and 11μm (8N) [254, 185]. 16N nuclei also exist, but are rare and were not included in the analysis [185]. A four-components Gaussian mixture model was fitted to the diameter distribution (R Package “mixtools”). Nuclei with a probability of >0.7 to belong to one of the 3 inferred populations with the smallest mean were assigned to 2N, 4N, 8N, respectively. The Gaussian distribution with the largest variance captured outliers in nuclei diameters (>15-18 μm) and were discarded.

6.5 RNA-sequencing and mapping

We sequenced three different RNA populations: from isolated nuclei (N), from the cytoplasmic extract (C), and from total liver tissue (unfractionated, U). We sequenced polyadenylated RNA by Poly(A)+ selection, and total RNA after ribo-depletion. In total, we have five types of RNA: Unfractionated total (UT), Nuclear Total (NT), Nuclear PolyA (NA), Cytoplasmic Total (CT), and Cytoplasmic PolyA (CA).

6.5.1 Library preparation

Library preparation and sequencing were done by the Gene Expression Core Facility (GECF) at EPFL. To sequence polyadenylated nuclear and cytoplasmic RNA, we used the "TruSeq stranded mRNA LT" kit from Illumina, starting from 650 ng RNA, according the the manufacturer’s protocol. To sequence total RNA, we used the "Kapa RNA hyperPrep with Riboerase" prep combined with "KAPA Unique Dual-Indexed Adapter Kit" from Illumina, starting from 650 ng RNA. Libraries were sequenced on a Hiseq4000. We used a SBS 50 cycles and SR cluster kit, performing a single read sequencing of 65 nucleotides. We sequenced on average 70M reads for NA and NT, 25M reads for CA and CT, and 50M reads for UT.

6.5.2 Pseudo-alignment of RNA-seq reads with Kallisto

Quantification of pre-mRNA and mRNA expression levels was performed using Kallisto version 0.46.0 [203]. A combined index was built using reference fasta files from Ensembl for mus musculus (mm10, GRCm38). In particular, pre-mRNA (introns and exons) and mRNA (exons) reference sequences were used as an input and named accordingly. Pseudoalignment and quantification were run on the aforementioned combined kallisto index with parameters "–single –rf-stranded -l 100 -s 30 -s 30". TPM (transcript per million) and estimated counts at (pre-) transcript level were used for further analysis.

We filtered lowly expressed transcripts. We only used TPM of mRNA in Unf Total RNA population, because the estimation of mature transcript is more robust than of pre-mRNA. We kept transcripts that have a sum of TPM > 0.5 in either WT or CryKO. We then calculated the Isoform Fraction (IF) for each transcript, which is the TPM of the isoform divided by the sum of TPM of all isoforms of a given gene. We defined a gene-specific threshold that depends on the number of annotated isoforms (n): the IF has be greater than $\frac{1}{n}$ to be kept. Additionally, we set the maximal threshold to 0.2, such that when a gene has only two isoforms, if one contributes to more than 20% of the total expression, it is not filtered out. If a transcript meets the condition in at least 6 samples out of the 120 samples (5 RNA populations, 2 genotypes), it is kept. Again, we filter transcripts based on the mRNA TPM only, because it is more difficult to obtain robust isoform-specific quantification of pre-mRNA, that share most of their intron between isoforms. In addition to the isoform-specific quantification, we pooled the isoforms belonging to the same biotype in order to have a biotype-specific quantification. We summed the TPM and the estimated counts. As for the gene length used for normalisation, we calculated the averaged effective gene length (estimated by Kallisto), weighted by the expression level of each isoform in each condition (therefore, gene length may differ between RNA population). We created 7 biotypes, based on the Ensembl description. Protein coding (PC), Retained Intron (RI), snRNA, and Processed Transcripts are defined by Ensembl. Nonsense mediated decay and non stop decay are grouped as "NMD". We grouped lincRNA, bidirectional promoter lncRNA, sense overlapping, sense intronic, 3prime overlapping ncRNA, lncRNA and in one biotype named "long non-coding RNA (lncRNA) . Small Nucleolar RNA (snoRNA) comprises the biotypes "snoRNA" and "scaRNA". All the biotypes containing the keyword "pseudogenes", whether they are processed, transcribed or not are grouped as "Pseudogene".

Finally, we summed all the transcripts by gene, independently of their biotype, in order to have gene-wise quantification. The gene length was calculated as the weighted average based on the relative expression of all the isoforms in each RNA population.

Defining pairs of Protein Coding and Retained intron

To define pairs of protein coding transcript and retained intron, we matched the exon start and exon end based on the gene annotation provided by Ensembl (release 102, November 2020). We allowed a difference of 1 nucleotide. We also considered as "pairs" when the retained intron overlapped two consecutive exons.

6.5.3 Statistical analysis of rhythmic gene expression

We normalised the counts estimated by Kallisto using the R package "DESeq2" [204]. The normalisation procedure by DESeq2 models raw counts with a negative binomial distribution with two parameters: a mean μ_g and a dispersion parameter α_g. The gene-specific mean μ_g is the number of counts multiplied by a sample-specific "size factor" estimated by the median-of-ratio method, which rescales the library size and aligns the median expression level of all the samples to a "pseudo-reference sample". The dispersion parameter α_g is estimated for each gene. To overcome the low number of replicates, DESeq2 algorithm assumes that genes with similar level of expression have a similar dispersion. Thus, it fits a curve to the dispersion estimates in function of expression values. Gene-specific α_g are then shrunk toward this curve.

The fit is performed using a Generalized Linear Model (GLM). Here, we directly implement a rhythmic analysis during the normalisation process, using the following model:

$$log_2(\mu_{g,s,t}) = a_{g,s} * cos(wt(s)) + b_{g,s} * sin(wt(s)) \tag{6.1}$$

where $N_{g,s,t}$ is the mean counts fitted by the negative binomial distribution for gene g, for sample s at circadian time-point (ZT) t. $a_{g,s}$ and $b_{g,s}$ the coefficients of the cosine and sine functions, and $\omega = \frac{2\pi}{24}$. The phase (peak of expression) is defined as $arctan(\frac{b}{a})$ and the amplitude is (or log $_2$ fold change) as $\sqrt{a^2 + b^2}$.

The full model with rhythmic parameters is tested with a Likelihood Ratio test against the reduced model with an intercept only. The intercept fitted by DESeq is used as the mean expression level over the 6 time points. Genes with a p-value < 0.01 and a $\log_2$ FC > 1 were considered as rhythmic.

The normalisation procedure was carried out on pre-mRNA and mRNA count sets separately, and for each RNA type (Nuc Total, Nuc PolyA, Cyt Total, Cyt Poly, Unf Total). We used the combined library size (pre-mRNA + mRNA) per sample to adjust for sequencing depth. The size factor was estimated using mRNA only, such that the median level of mRNA is aligned for all the samples. The same size factor is applied to the pre-mRNA, such that the relationship between pre-mRNA and mRNA is conserved.

Normalisation of counts by gene length

For statistical testings such as differential gene expression, DESeq2 uses raw counts. However, for visualisation or exploratory purposes (e.g. PCA), we use counts normalised by gene length, in a similar way to RPKM (Reads Per Kilobase Million), using the estimated size factor, library size of the pseudo-reference sample, and gene length. The gene length is specific to each RNA fraction. It is the weighted average based on the relative expression level of each isoform. Our RPKM-like value is thus defined as follow:

$$RPKM = \frac{Counts}{size\ factor * pseudo\ lib.size * gene\ length} * 10^9 \tag{6.2}$$

Filtering RNA-seq data

We removed transcripts assigned to one of the following Ensembl biotype (release 102): T-cell receptor genes, Immunoglobin genes, TEC, mitochondrial RNA, miscelanneous RNA, ribozyme, antisense RNA, mitochonrdial RNA, rRNA, miRNA, and tRNA.

We filtered genes based on their expression level. Since the three RNA types have been processed and normalised separately, we cannot directly compare the expression values. Therefore, we defined a threshold value specific to each RNA type per genotype. We arbitrarily set the threshold such that at least 50% of the genes are kept, based on the mean value fitted by DESeq2 (Intercept). We only considered the mRNA expression level, because the estimation of the pre-mRNA level is less accurate. If the mean expression level of a gene is higher than the threshold in at least one RNA type, but not in others, it is kept. We applied the same procedure to filter genes from the biotype-specific quantification and transcripts for the isoform-specific quantification. In total, the number of expressed genes in the gene-wide quantification is 11251, 12081 in the biotype-specific quantification, and 40991 transcripts from 12219 genes in the isoform-specific quantification.

Calibration of the nuclear and cytoplasmic ratio

The expression levels in RPKM are not comparable across the different RNA types due to independent RNA extraction, sequencing, and normalisation procedures. In order to readjust the nuclear-cytoplasmic ratio, we took advantage of the dataset published by Bahar-Halpern et.al. [196]. The authors converted the counts obtained from sequencing nuclear and cytoplasmic PolyA RNA from mouse liver into number of molecules per cells using smFISH. This allowed the authors to calibrate the RNA-seq data, and to compare the ratio of nuclear and cytoplasmic mRNA. In order to rescale the RPKM of our nuclear and cytoplasmic fractions, we applied a linear regression model such that: $log_2(Nuc.\ transcripts\ BH) \sim \alpha_n + log_2(Cyt.\ RPKM)$ and $log_2(Cyt.\ transcripts\ BH) \sim \alpha_c + log_2(Nuc.\ RPKM)$. The offset is then our calibration factor. Here, Nuc refers to either NAE or NET, and Cyt to CTE or CAE. We used the mean RPKM over all the time points, fitted by DESeq2. BH stands for Bahar-Halpern, first author of [196]. We decided to apply the same calibration to the CryKO dataset, assuming that relationship between nuclear and cytoplasmic transcript abundance is globally similar between genotypes. The $\log_2$(NAE) RPKM was shifted by a factor -1.34, the $\log_2$(CAE) RPKM by a factor 1.2, the $\log_2$(NTE) by a factor -0.15, and the $\log_2$(CTE) by a factor 0.7.

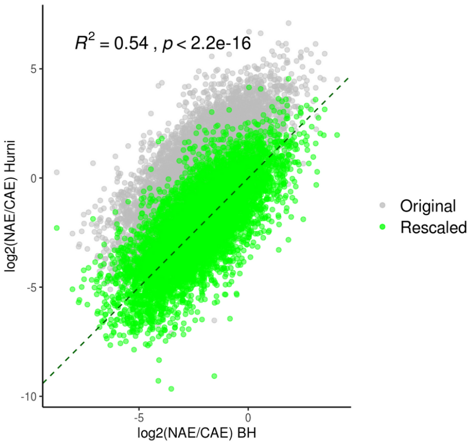


Figure 6.1 – Correlation between the nuclear-cytoplasmic ratio from our dataset (Hurni) and the dataset published in [196]. NAE and CAE are RPKM averaged over the 6 time-points. In grey: correlation before calibration. In green: after calibration. The $\log_2$ratio was shifted by a factor [196]. In grey:

6.6 Functional Enrichment Analysis

Every time we tested for enrichment of biological functions, we combined several sources including Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Wikipathways databases. We used both the EnrichR R package [255] or TopGO package using *weight01* algorithm [256].

6.6.1 Gene Ontology Analysis of nuclear and cytoplasmic genes

We tested for gene set enrichment of protein coding transcripts in nuclear and cytoplasmic fractions using Gene Ontology Terms derived from "Biological Processes". We used the Fisher's exact test implemented in EnrichR package to test for significance [255]. GO Terms with an adjusted p-value < 0.1 and a Combined Score > 30 were considered as enriched. We grouped together Terms that were semantically similar. Nuclear and Cytoplasmic enrichment were assessed using DESeq2 [204]. Raw counts of NAE and CAE samples were tested for enrichment using the cellular localisation (Nuclear of Cytoplasmic) as a variable in the design formula. Transcripts with a $\log_2$ FC > 2 and an adjusted p-value >0.01 were considered as enriched.

6.7 Mathematical modeling of rhyhthmic RNA processing rates

The processes regulating RNA expression rate are numerous, including transcription regulation steps (chromatin conformation, histone marks), PolII elongation, termination (polyadenylation of 3' tail and addition of 5' cap), co- and post-transcriptional splicing, m^6A methylation, nucleo-cytoplasmic export, binding of various RBPs, deadenylation of polyA tail, and final degradation. All these processes affect the stability of the mRNA transcript, and therefore its level of accumulation.

To describe the accumulation of each RNA specie in the nucleus or in the cytoplasm, we assume the following model (Fig.6.2): pre-mRNA p, transcribed with a rate T, is spliced and polyadenylated at a rate of s to produce mature nuclear RNA m. Then, this transcript is further processed and exported at a rate e into the cytoplasm M, where it is finally degraded at a rate γ. This model ignores nuclear degradation.

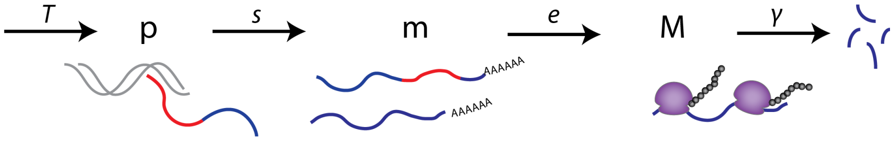


Figure 6.2 – Scheme describing the journey of an RNA from transcription, accumulation in the nucleus, accumulation in the cytoplasm, and degradation

This model is summarised with a simple set of first-order differential equations:

$$\dot{p} = T - s * p \tag{6.3a}$$

$$\dot{m} = s * p - m * e \tag{6.3b}$$

$$\dot{M} = e * m - \gamma * M \tag{6.3c}$$

where: p = pre-mRNA, m = nuclear mRNA, M = cytoplasmic mRNA. The parameters are: T = Transcription rate $[molecule * time^{-1}]$, s = splicing and polyadenylation rate $[time^{-1}]$, e = export rate $[time^{-1}]$, γ = cytoplasmic degradation rate $[time^{-1}]$.

Since the system is under-determined, and that transcription rate is unknown, we first focused on the analysis at steady-state, where the variables are described by the ratio of the transcription and degradation term (Eq.6.4). We can describe the subcellular enrichment of RNA transcripts in the nucleus or in the cytoplasm, by using the ratio $\frac{p}{m}$ and $\frac{m}{M}$. However, it is nearly impossible to determine which term drives the most that distribution. For instance, there could be a higher number of nuclear RNA transcripts than cytoplasmic transcripts because they are retained in the nucleus, or they could be extremely short-lived in the cytoplasm, which results in an apparent nuclear enrichment.

$$p^* = \frac{T}{s} \tag{6.4a}$$

$$m^* = \frac{s * p}{e} \approx \frac{T}{e} \tag{6.4b}$$

$$M^* = \frac{e * m}{\gamma} = \frac{T}{\gamma} \tag{6.4c}$$

We use nuclear pre-mRNA NTI as a proxy for p, nuclear polyadenylated mRNA NAE for nuclear transcripts m, and cytoplasmic polyadenylated mRNA CAE for M. When these variables are injected in the equation 6.4 and are $\log_2$-transformed, they become :

$$log_2(NTI) = log_2(Transcription\ rate) - log_2(splicing\ rate) \tag{6.5a}$$

$$log_2(NAE) = log_2(Transcription\ rate) - log_2(export\ rate) \tag{6.5b}$$

$$log_2(CAE) = log_2(Transcription\ rate) - log_2(degradation\ rate) \tag{6.5c}$$

6.7.1 Modeling temporal RNA profiles and estimation of kinetic parameters

In order to estimate the kinetic parameters (splicing and polyadenylation s, export e, and cytoplasmic degradation γ), we used a mathematical modeling approach developed by two former post-doctoral researchers from the Naef gorup (Jingkui Wang and Laura Symul, co-authors of [78]). In the original model, a first-order differential equation describes the temporal variation of mature RNA m (exonic reads from total liver RNA-seq) in function of pre-mRNA p (intronic reads), processing rate k, and degradation rate γ. Pre-mRNA is described either by a cosinor function with a period of 24 hours if rhythmic, or by a constant if expressed constitutively. In addition, the degradation term can also be constant and rhythmic, and the latter case would explain some temporal patterns such as gain of relative amplitude or a particularly large phase-delay otherwise not possible in case of constant degradation [79].

Here, we split the RNA processing journey (Fig. 6.2) into two successive but independent steps: step 1 describes the transformation of pre-mRNA p into nuclear polyadenylated mRNA m, and step 2 described the transformation of m to cytoplasmic mRNA M.

For step 1, the equations are the following:

$$\frac{dm}{dt} = s * p(t) - e(t) * m(t) \tag{6.6a}$$

with

$$p(t) = p_{min} + A_p(\frac{1 + cos(wt - \phi_p)}{2})^{\beta} \tag{6.6b}$$

$$e(t) = e_0(1 + \epsilon_e * cos(wt - \phi_e)) \tag{6.6c}$$

p_{min} is the minimum value of p, A_p is the amplitude (peak to trough), ϕ_p is the phase (peak time), and ω is the angular frequency $\frac{2\pi}{24}$. If the gene is constitutively expressed, then $p(t) = p_0$. e_0 is the mean export rate. If the export is rhythmic, the relative amplitude of export ϵ_e varies between $]0,1]$. ϕ_e is the phase of export rate.

We apply the same model to step 2 between nuclear RNA m and cytoplasmic RNA M, with γ being the cytoplasmic degradation rate.

$$\frac{dM}{dt} = e * m(t) - \gamma(t) * M(t) \tag{6.7a}$$

with

$$m(t) = m_{min} + A_m(\frac{1 + cos(wt - \phi_m)}{2})^{\beta} \tag{6.7b}$$

$$\gamma(t) = \gamma_0(1 + \epsilon_\gamma * cos(wt - \phi_\gamma)) \tag{6.7c}$$

We used a cosine function with a period of 24 hours to describe circadian oscillations. To account for patterns that are more peaked, an exponent β is added to the cosine function, which ranges from 1 to 2 (Fig. 6.3). We do not allow a higher value of β, as we empirically observed that a sharper function tends to catch outlier data points and artificially increase the number of rhythmic genes.

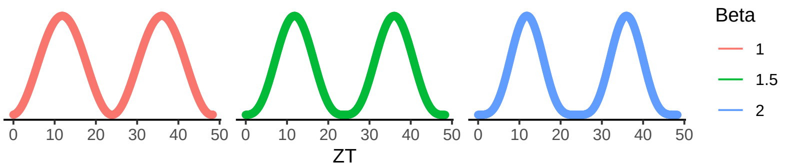


Figure 6.3 – Shape of cosinor function for $\beta = 1, 1.5$ or 2

According to the model, pre-mRNA p and export e can be either rhythmic or constant (m and γ in step 2). The combination of constant or rhythmic production and degradation terms generates 4 possible models (Fig.6.4).

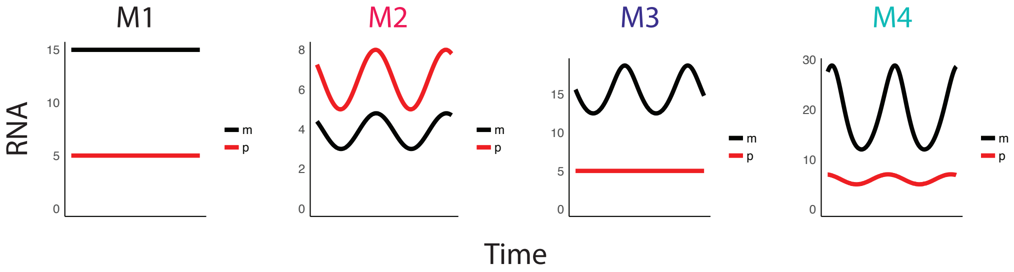


Figure 6.4 – Combination of rhythmic or constant transcription (depicted by pre-mRNA p) and rhythmic or constant degradation generates four different models. m = mRNA.

- Model 1 describes the case when transcription is constant, thus profiles of pre-mRNA and mRNA accumulation are flat. The only parameter that can estimated is the ratio of production and degradation rates, which is the ratio of the mean expression level of pre-mRNA over mRNA (steady-state).
- Model 2 represents the case when rhythmic accumulation of both species is driven solely by the first variable. The relationship of phases and amplitudes of the two species are derived from the equations, and can be summarised in two properties: first, the relative amplitude (peak minus trough, divided by the mean) of the second specie cannot be higher than the relative amplitude of the first specie. In other terms, the oscillations dampen as rhythm propagates. The dampening is more important as the half-life is long (Fig. 6.5). Second, the phase difference is always less than 6 hours. In model 2, when $\beta = 1$, the phase delay is given by $arctan(\frac{w}{\gamma})$, which is not defined above $\pi/2$ (6 hours in circadian time). If β >1, there is not analytical solution, and the solution is found by numerical integration.
- Model 3 represents the case when a gene is constitutively expressed, but a rhythmic post-transcriptional step (export or degradation) generates rhythmic accumulation. Since the first variable is flat, only the ratio of degradation over production can be confidently estimated. The

mean half-life is often non-identifiable, and depends on the shape of the second variable (peaked shape are more easily fitted).

- Model 4 is when both p and e (m and γ in step 2) are oscillating. The combination of rhythmic input and post-transcriptional processing allows patterns that do not follow the equation described in Fig.6.5 for model 2, such as an amplification of relative amplitude, or a large phase-shift.

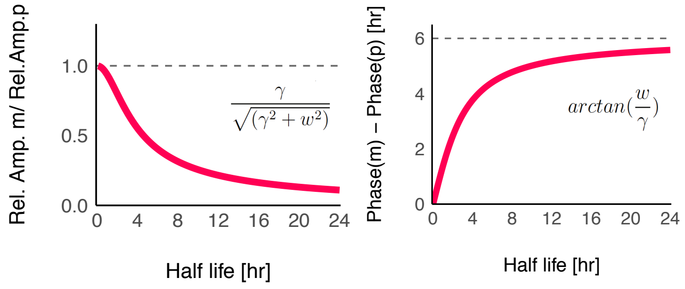


Figure 6.5 – Relationship between half-life and relative amplitudes (left) and phase delay (right) for M2 model with $\beta = 1$.

We fitted each of the four models to experimental RNA-seq profiles. We used NTI for unspliced pre-mRNA p, NAE for nuclear mature RNA m, and CAE for cytoplasmic RNA M. We applied the mathematical model to step 1 and step 2 independently. Each gene is fitted with the four models, and the one with best the Bayesian information criterion (BIC) is chosen. We only selected genes with a probability of BIC > 0.6.

Parameters estimation

We fitted the model to experimental profiles of step 1 and step 2 independently. The following explanation and equations (6.8 and 6.9) refers to step 1. To adapt the equation to step 2, nuclear RNA m has to be replaced by cytoplasmic RNA M, and pre-mRNA p by m.

We inferred the rates of 8 parameters (for M4): A_p, Min_p, ϕ_p, β, s, e_0, ϵ_e, and ϕ_e. Depending on the gene, s, e_0, ϵ_e and ϕ_e can be difficult to identify, especially for M3 and M4. To alleviate the parameter non-identifiability, we estimated instead the following combinations of parameters: $s' = s/e_0$, $\epsilon_e' = \epsilon_e * e_0/\sqrt{(e_0^2 + w^2)}$) and $\phi_e' = \phi : e + atan(\frac{w}{e_0})$ [79]. We used the function *optim* from the R package "stats" with the method "L-BFGS-B". The optimisation was done using R code developed by Jingkui Wang, author of [78], with some adjustments.

We assumed that RNA-seq counts data follows a binomial distribution [204]. Thus, the log-likelihood function to be minimised for each gene is:

$$logL = \sum_{n} logNB(n_m(t)|\mu_m(t), \alpha_m(t)) + logNB(n_p(t)|\mu_p(t), \alpha_p(t)) \tag{6.8a}$$

with

$$\mu_p(t) = p(t)S_p(t)L_p \tag{6.9a}$$

$$\mu_m(t) = m(t)S_m(t)L_m \tag{6.9b}$$

$$\tag{6.9c}$$

Here, $n(t)$ is the read counts. The subscripts p and m stand for pre-mRNA and nuclear mRNA. α is the dispersion parameter of the negative binomial distribution, specific to each gene and time-point. It is estimated by the R package "DESeq2" [204], which uses an empirical Bayes shrinkage method to make more robust estimation of α (fit type: "parametric"). $\mu(t)$ is the expected mean of counts, which is the concentration of transcripts ($p(t)$ or $m(t)$) multiplied by the gene length, and a sample-specific scaling factor S, and further multiplied by 10^9 to obtain convenient values. Thus, the "concentration of tanscripts" can be understood as an equivalent of RPKM (reads per kilo per million base pairs). The scaling factor S, also estimated by DESeq2, is a normalisation term specific to each sample multiplied by the library size of a "pseudo-sample" (mean of the twelve library sizes, each scaled by their specific normalisation factor). The scaling factor was first estimated separately on each RNA population, including only mRNA, such that the median expression level of mRNA are aligned between samples (Nuc.Total, Nuc.PolyA, and Cyt.PolyA). $m(t)$ was calculated by numerical integration, because there is no analytical solution.

Boundaries on parameters

We bounded the parameters with physiologically relevant values:

- Nuclear retention time ($\frac{log(2)}{e_0}$): 5 minutes to 12 hours [196, 127].
- Cytoplasmic half-life ($\frac{log(2)}{\gamma_0}$): 10 minutes to 24 hours.
- Relative amplitude of rhythmic export and rhythmic degradation (ϵ_e and ϵ_γ): between 0 and 1. It was noticed by Wand et al. that relative amplitude close to 1 renders the optimisation very sensitive. Therefore, ϵ_e and ϵ_γ higher than 0.8 are penalized by a sigmoid function.

Identifiability Analysis for kinetic parameters

Depending on the dataset (number of observations, sampling density), parameters cannot always be identified unambiguously. *Structural non-identifiability* results from the model and parametrisation, while *practical non-identifiability* results from the amount and quality of the experimental data. For instance, genes in M1 (constant expression and accumulation), only the ratio of production and degradation term can be determined, and γ (cytoplasmic degradation rate in NAE-CAE, and export rate in NIT-NAE) is thus structurally non-identifiable. To determine the practical identifiability of the parameter γ, we used the Profile Likelihood (PL) as described in [223] and as implemented in [78].

We sampled 10 different values of γ within the boundaries and maximized the likelihood for each γ. If the likelihood is constant for all the values, the parameter is structurally non-identifiable. If the parameter is practically non-identifiable, the likelihood varies for different values of γ and reaches a minimum, but the variation is smaller than a defined threshold. If the PL varies sufficiently, then the parameter is considered as identifiable.

Here, we defined two thresholds: the 0.68 quantile and 0.95 quantile of a χ^2 distribution with degree of freedom of 1 [257]. If the variation of PL is higher than the 0.95 quantile, the identifiability of γ is considered "good". If it is only higher than the 0.68 quantile, is it considered as "pass". If lower, it is defined as practically non-identifiable.

We looked at the PL variation for γ values lower than the estimated γ ("left identifiability") and for values higher than the estimated γ ("right identifiability"). Therefore, if the estimated γ reached the upper or lower boundary, the parameter would necessarily be non identifiable from the right, respectively left.

In our analysis, we selected genes whose estimated degradation / export rates was at least identified as "passed" either from the left, from the right, of from both sides.

Bibliography

[1] M. H. Hastings, A. B. Reddy, and E. S. Maywood, “A clockwork web: Circadian timing in brain and periphery, in health and disease,” *Nature Reviews Neuroscience*, vol. 4, no. 8, pp. 649–661, 2003.

[2] C. Liu, S. Li, T. Liu, J. Borjigin, and J. D. Lin, “Transcriptional coactivator PGC-1α integrates the mammalian clock and energy metabolism,” *Nature*, vol. 447, no. 7143, pp. 477–481, may 2007.

[3] S. Yamaguchi, H. Isejima, T. Matsuo, R. Okura, K. Yagita, M. Kobayashi, and H. Okamura, “Synchronization of Cellular Clocks in the Suprachiasmatic Nucleus,” *Science*, vol. 302, no. 5649, pp. 1408–1412, nov 2003.

[4] J. A. Mohawk, C. B. Green, and J. S. Takahashi, “Central and peripheral circadian clocks in mammals.” *Annual review of neuroscience*, vol. 35, pp. 445–62, 2012.

[5] U. Albrecht, “Timing to Perfection: The Biology of Central and Peripheral Circadian Clocks,” *Neuron*, vol. 74, no. 2, pp. 246–260, 2012.

[6] S.-H. Yoo, S. Yamazaki, P. L. Lowrey, K. Shimomura, C. H. Ko, E. D. Buhr, S. M. Siepka, H.-K. Hong, W. J. Oh, O. J. Yoo, M. Menaker, J. S. Takahashi, Y. Okuno, M. Doi, H. Okamura, K. Horikawa, T. Kudo, and S. Shibata, “PERIOD2::LUCIFERASE real-time reporting of circadian dynamics reveals persistent circadian oscillations in mouse peripheral tissues,” *Proceedings of the National Academy of Sciences*, vol. 101, no. 15, pp. 5339–5346, apr 2004.

[7] E. Nagoshi, C. Saini, C. Bauer, T. Laroche, F. Naef, and U. Schibler, “Circadian Gene Expression in Individual Fibroblasts,” *Cell*, vol. 119, no. 5, pp. 693–705, nov 2004.

[8] J. Yeung, J. Mermet, C. Jouffe, J. Marquis, A. Charpagne, F. Gachon, and F. Naef, “Transcription factor activity rhythms and tissue-specific chromatin interactions explain circadian gene expression across organs,” *Genome Research*, vol. 28, no. 2, pp. 182–191, feb 2018.

[9] J. Aschoff and R. Wever, “Human circadian rhythms: a multioscillatory system,” *Federation Proceedings*, vol. 35, no. 12, pp. 2326–2332, 1976.

[10] K. Eckel-Mahan and P. Sassone-Corsi, “Phenotyping Circadian Rhythms in Mice,” *Current protocols in mouse biology*, vol. 5, no. 3, pp. 271–281, sep 2015.

[11] R. J. Konopka and S. Benzer, “Clock mutants of Drosophila melanogaster.” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 68, no. 9, pp. 2112–2116, 1971.

[12] P. Reddy, W. A. Zehring, D. A. Wheeler, V. Pirrotta, C. Hadfield, J. C. Hall, and M. Rosbash, “Molecular analysis of the period locus in Drosophila melanogaster and identification of a transcript involved in biological rhythms,” *Cell*, vol. 38, no. 3, pp. 701–710, 1984.

[13] W. A. Zehring, D. A. Wheeler, P. Reddy, R. J. Konopka, C. P. Kyriacou, M. Rosbash, and J. C. Hall, “P-element transformation with period locus DNA restores rhythmicity to mutant, arrhythmic drosophila melanogaster,” *Cell*, vol. 39, no. 2 PART 1, pp. 369–376, 1984.

[14] M. H. Vitaterna, D. P. King, A. M. Chang, J. M. Kernhauser, P. L. Lowrey, J. D. McDonald, W. F. Dove, L. H. Pinto, F. W. Turek, and J. S. Takahashi, “Mutagenesis and mapping of a mouse gene, clock, essential for circadian behavior,” *Science*, vol. 264, no. 5159, pp. 719–725, apr 1994.

[15] N. Gekakis, D. Staknis, H. B. Nguyen, F. C. Davis, L. D. Wilsbacner, D. P. King, J. S. Takahashi, and C. J. Weitz, “Role of the CLOCK protein in the mammalian circadian mechanism,” *Science*, vol. 280, no. 5369, pp. 1564–1569, jun 1998.

[16] K. Kume, M. J. Zylka, S. Sriram, L. P. Shearman, D. R. Weaver, X. Jin, E. S. Maywood, M. H. Hastings, and S. M. Reppert, "mCRY1 and mCRY2 are essential components of the negative limb of the circadian clock feedback loop." *Cell*, vol. 98, no. 2, pp. 193–205, jul 1999.

[17] P. L. Lowrey and J. S. Takahashi, "Genetics of the mammalian circadian system: Photic entrainment, circadian pacemaker mechanisms, and posttranslational regulation," pp. 533–562, nov 2000.

[18] L. P. Shearman, S. Sriram, D. R. Weaver, E. S. Maywood, I. Chaves, B. Zheng, K. Kume, C. C. Lee, G. T. Van Der Horst, M. H. Hastings, and S. M. Reppert, "Interacting molecular loops in the mammalian circadian clock," *Science*, vol. 288, no. 5468, pp. 1013–1019, may 2000.

[19] T. K. Sato, S. Panda, L. J. Miraglia, T. M. Reyes, R. D. Rudic, P. McNamara, K. A. Naik, G. A. FitzGerald, S. A. Kay, and J. B. Hogenesch, "A functional genomics strategy reveals Rora as a component of the mammalian circadian clock." *Neuron*, vol. 43, no. 4, pp. 527–37, aug 2004.

[20] N. Preitner, F. Damiola, L. Lopez-Molina, J. Zakany, D. Duboule, U. Albrecht, and U. Schibler, "The orphan nuclear receptor REV-ERBalpha controls circadian transcription within the positive limb of the mammalian circadian oscillator." *Cell*, vol. 110, no. 2, pp. 251–60, jul 2002.

[21] F. Gachon, E. Nagoshi, S. Brown, J. Ripperger, and U. Schibler, "The mammalian circadian timing system: from gene expression to physiology," *Chromosoma*, vol. 113, no. 3, pp. 103–112, sep 2004.

[22] A. Hirano, Y. H. Fu, and L. J. Ptáek, "The intricate dance of post-translational modifications in the rhythm of life," pp. 1053–1060, dec 2016.

[23] K. L. Toh, C. R. Jones, Y. He, E. J. Eide, W. A. Hinz, D. M. Virshup, L. J. Ptáček, and Y. H. Fu, "An hPer2 phosphorylation site mutation in familial advanced sleep phase syndrome," *Science*, vol. 291, no. 5506, pp. 1040–1043, feb 2001.

[24] Y. Xu, Q. S. Padiath, R. E. Shapiro, C. R. Jones, S. C. Wu, N. Saigoh, K. Saigoh, L. J. Ptáček, and Y. H. Fu, "Functional consequences of a CKIδ mutation causing familial advanced sleep phase syndrome," *Nature*, vol. 434, no. 7033, pp. 640–644, mar 2005.

[25] J. Yeung, J. Mermet, C. Jouffe, J. Marquis, A. Charpagne, F. Gachon, and F. Naef, "Transcription factor activity rhythms and tissue-specific chromatin interactions explain circadian gene expression across organs," *Genome Research*, vol. 28, no. 2, pp. 182–191, feb 2018.

[26] M. Qu, T. Duffy, T. Hirota, and S. A. Kay, "Nuclear receptor HNF4A transrepresses CLOCK: BMAL1 and modulates tissue-specific circadian networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, no. 52, pp. E12 305–E12 312, dec 2018.

[27] L. S. Mure, H. D. Le, G. Benegiamo, M. W. Chang, L. Rios, N. Jillani, M. Ngotho, T. Kariuki, O. Dkhissi-Benyahya, H. M. Cooper, and S. Panda, "Diurnal transcriptome atlas of a primate across major neural and peripheral tissues," *Science*, vol. 359, no. 6381, mar 2018.

[28] R. Zhang, N. F. Lahens, H. I. Ballance, M. E. Hughes, and J. B. Hogenesch, "A circadian gene expression atlas in mammals: implications for biology and medicine." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 45, pp. 16 219–24, nov 2014.

[29] J. Bass and J. S. Takahashi, "Circadian Integration of Metabolism and Energetics," *Science*, vol. 330, no. 6009, 2010.

[30] F. Damiola, N. Le Minh, N. Preitner, B. Kornmann, F. Fleury-Olela, and U. Schibler, "Restricted feeding uncouples circadian oscillators in peripheral tissues from the central pacemaker in the suprachiasmatic nucleus." *Genes & development*, vol. 14, no. 23, pp. 2950–61, dec 2000.

[31] N. Le Minh, F. Damiola, F. Tronche, G. Schütz, and U. Schibler, "Glucocorticoid hormones inhibit food-induced phase-shifting of peripheral circadian oscillators," *EMBO Journal*, vol. 20, no. 24, pp. 7128–7136, dec 2001.

[32] A. Balsalobre, S. A. Brown, L. Marcacci, F. Tronche, C. Kellendonk, H. M. Reichardt, G. Schütz, and U. Schibler, "Resetting of circadian time in peripheral tissues by glucocorticoid signaling." *Science (New York, N.Y.)*, vol. 289, no. 5488, pp. 2344–7, sep 2000.

[33] B. J. Greenwell, A. J. Trott, J. R. Beytebiere, S. Pao, A. Bosley, E. Beach, P. Finegan, C. Hernandez, and J. S. Menet, "Rhythmic Food Intake Drives Rhythmic Gene Expression More Potently than the Hepatic Circadian Clock in Mice," *Cell Reports*, vol. 27, no. 3, pp. 649–657.e5, apr 2019.

[34] P. Crosby, R. Hamnett, M. Putker, N. P. Hoyle, M. Reed, C. J. Karam, E. S. Maywood, A. Stangherlin, J. E. Chesham, E. A. Hayter, L. Rosenbrier-Ribeiro, P. Newham, H. Clevers, D. A. Bechtold, and J. S. O'Neill, “Insulin/IGF-1 Drives PERIOD Synthesis to Entrain Circadian Rhythms with Feeding Time,” *Cell*, vol. 177, no. 4, pp. 896–909.e20, may 2019.

[35] T. C. Erren, P. Morfeld, J. V. Groß, U. Wild, and P. Lewis, “IARC 2019: "Night shift work" is probably carcinogenic: What about disturbed chronobiology in all walks of life?” p. 29, nov 2019.

[36] R. D. Rudic, P. McNamara, A.-M. Curtis, R. C. Boston, S. Panda, J. B. Hogenesch, and G. A. FitzGerald, “BMAL1 and CLOCK, Two Essential Components of the Circadian Clock, Are Involved in Glucose Homeostasis,” *PLoS Biology*, vol. 2, no. 11, p. e377, nov 2004.

[37] M. Hatori, C. Vollmers, A. Zarrinpar, L. DiTacchio, E. Bushong, S. Gill, M. Leblanc, A. Chaix, M. Joens, J. Fitzpatrick, M. Ellisman, and S. Panda, “Time-Restricted Feeding without Reducing Caloric Intake Prevents Metabolic Diseases in Mice Fed a High-Fat Diet,” *Cell Metabolism*, vol. 15, no. 6, pp. 848–860, jun 2012.

[38] S. Gill, H. D. Le, G. C. Melkani, and S. Panda, “Time-restricted feeding attenuates age-related cardiac decline in Drosophila,” *Science*, vol. 347, no. 6227, pp. 1265–1269, mar 2015.

[39] Y. Nakahata, M. Kaluzova, B. Grimaldi, S. Sahar, J. Hirayama, D. Chen, L. P. Guarente, and P. Sassone-Corsi, “The NAD+-dependent deacetylase SIRT1 modulates CLOCK-mediated chromatin remodeling and circadian control.” *Cell*, vol. 134, no. 2, pp. 329–40, jul 2008.

[40] D. C. Levine, H. Hong, B. J. Weidemann, K. M. Ramsey, A. H. Affinati, M. S. Schmidt, J. Cedernaes, C. Omura, R. Braun, C. Lee, C. Brenner, C. B. Peek, and J. Bass, “NAD+ Controls Circadian Reprogramming through PER2 Nuclear Translocation to Counter Aging,” *Molecular Cell*, vol. 78, no. 5, pp. 835–849.e7, jun 2020.

[41] K. A. Lamia, U. M. Sachdeva, L. DiTacchio, E. C. Williams, J. G. Alvarez, D. F. Egan, D. S. Vasquez, H. Juguilon, S. Panda, R. J. Shaw, C. B. Thompson, and R. M. Evans, “AMPK regulates the circadian clock by cryptochrome phosphorylation and degradation.” *Science (New York, N.Y.)*, vol. 326, no. 5951, pp. 437–40, oct 2009.

[42] L. Sun, J. Ma, C. W. Turck, P. Xu, and G. Z. Wang, “Genome-wide circadian regulation: A unique system for computational biology,” pp. 1914–1924, jan 2020.

[43] S. Panda, M. P. Antoch, B. H. Miller, A. I. Su, A. B. Schook, M. Straume, P. G. Schultz, S. A. Kay, J. S. Takahashi, and J. B. Hogenesch, “Coordinated Transcription of Key Pathways in the Mouse by the Circadian Clock,” *Cell*, vol. 109, no. 3, pp. 307–320, may 2002.

[44] H. R. Ueda, S. Hayashi, W. Chen, M. Sano, M. Machida, Y. Shigeyoshi, M. Iino, and S. Hashimoto, “System-level identification of transcriptional circuits underlying mammalian circadian clocks.” *Nature genetics*, vol. 37, no. 2, pp. 187–92, feb 2005.

[45] D. Mauvoisin, J. Wang, C. Jouffe, E. Martin, F. Atger, P. Waridel, M. Quadroni, F. Gachon, and F. Naef, “Circadian clock-dependent and -independent rhythmic proteomes implement distinct diurnal functions in mouse liver.” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 1, pp. 167–72, jan 2014.

[46] Y. Adamovich, L. Rousso-Noori, Z. Zwighaft, A. Neufeld-Cohen, M. Golik, J. Kraut-Cohen, M. Wang, X. Han, and G. Asher, “Circadian clocks and feeding time regulate the oscillations and levels of hepatic triglycerides,” *Cell Metabolism*, vol. 19, no. 2, pp. 319–330, feb 2014.

[47] N. Koike, S.-H. Yoo, H.-C. Huang, V. Kumar, C. Lee, T.-K. Kim, and J. S. Takahashi, “Transcriptional Architecture and Chromatin Landscape of the Core Circadian Clock in Mammals,” *Science*, vol. 338, no. 6105, pp. 349–354, oct 2012.

[48] H. Reinke and G. Asher, “Crosstalk between metabolism and circadian clocks,” *Nature Reviews Molecular Cell Biology*, p. 1, jan 2019.

[49] R. Doi, K. Oishi, and N. Ishida, “CLOCK regulates circadian rhythms of hepatic glycogen synthesis through transcriptional activation of Gys2,” *Journal of Biological Chemistry*, vol. 285, no. 29, pp. 22 114–22 121, jul 2010.

[50] E. E. Zhang, Y. Liu, R. Dentin, P. Y. Pongsawakul, A. C. Liu, T. Hirota, D. A. Nusinow, X. Sun, S. Landais, Y. Kodama, D. A. Brenner, M. Montminy, and S. A. Kay, “Cryptochrome mediates

circadian regulation of cAMP signaling and hepatic gluconeogenesis." *Nature medicine*, vol. 16, no. 10, pp. 1152–6, oct 2010.

[51] H. Jang, G. Y. Lee, C. P. Selby, G. Lee, Y. G. Jeon, J. H. Lee, K. K. Y. Cheng, P. Titchenell, M. J. Birnbaum, A. Xu, A. Sancar, and J. B. Kim, "SREBP1c-CRY1 signalling represses hepatic glucose production by promoting FOXO1 degradation during refeeding," *Nature Communications*, vol. 7, p. 12180, jul 2016.

[52] L. Chen and G. Yang, "PPARs Integrate the Mammalian Clock and Energy Metabolism." *PPAR research*, vol. 2014, p. 653017, 2014.

[53] D. Feng, T. Liu, Z. Sun, A. Bugge, S. E. Mullican, T. Alenghat, X. S. Liu, and M. A. Lazar, "A circadian rhythm orchestrated by histone deacetylase 3 controls hepatic lipid metabolism." *Science (New York, N.Y.)*, vol. 331, no. 6022, pp. 1315–9, mar 2011.

[54] G. Le Martelot, D. Canella, L. Symul, E. Migliavacca, F. Gilardi, R. Liechti, O. Martin, K. Harshman, M. Delorenzi, B. Desvergne, W. Herr, B. Deplancke, U. Schibler, J. Rougemont, N. Guex, N. Hernandez, and F. Naef, "Genome-Wide RNA Polymerase II Profiles and RNA Accumulation Reveal Kinetics of Transcription and Associated Epigenetic Changes During Diurnal Cycles," *PLoS Biology*, vol. 10, no. 11, pp. e1 001 442–e1 001 442, nov 2012.

[55] F. Gachon, "Physiological function of PARbZip circadian clock-controlled transcription factors," *Annals of Medicine*, vol. 39, no. 8, pp. 562–571, 2007.

[56] F. Gachon, F. F. Olela, O. Schaad, P. Descombes, and U. Schibler, "The circadian PAR-domain basic leucine zipper transcription factors DBP, TEF, and HLF modulate basal and inducible xenobiotic detoxification." *Cell metabolism*, vol. 4, no. 1, pp. 25–36, jul 2006.

[57] P. Cramer, "Eukaryotic Transcription Turns 50," pp. 808–812, oct 2019.

[58] J. S. Takahashi, N. Lahens, H. Ballance, M. Hughes, J. Hogenesch, L. Dayon, F. Sizzano, A. Palini, M. Kussmann, P. Waridel, and E. Al., "Enriching the Circadian Proteome," *Cell Metabolism*, vol. 25, no. 1, pp. 1–2, jan 2017.

[59] J. S. Menet, J. Rodriguez, K. C. Abruzzi, and M. Rosbash, "Nascent-Seq reveals novel features of mouse circadian transcriptional regulation." *eLife*, vol. 1, p. e00011, nov 2012.

[60] F. Atger, C. Gobet, J. Marquis, E. Martin, J. Wang, B. Weger, G. Lefebvre, P. Descombes, F. Naef, and F. Gachon, "Circadian and feeding rhythms differentially affect rhythmic mRNA transcription and translation in mouse liver." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 47, pp. E6579–88, nov 2015.

[61] M. S. Robles, J. Cox, M. Mann, R. Scheltema, and J. Olsen, "In-Vivo Quantitative Proteomics Reveals a Key Contribution of Post-Transcriptional Mechanisms to the Circadian Regulation of Liver Metabolism," *PLoS Genetics*, vol. 10, no. 1, p. e1004047, jan 2014.

[62] D. Mauvoisin, F. Atger, L. Dayon, A. Núñez Galindo, J. Wang, E. Martin, L. Da Silva, I. Montoliu, S. Collino, F. P. Martin, J. Ratajczak, C. Cantó, M. Kussmann, F. Naef, and F. Gachon, "Circadian and Feeding Rhythms Orchestrate the Diurnal Liver Acetylome," *Cell Reports*, vol. 20, no. 7, pp. 1729–1743, aug 2017.

[63] J. Wang, D. Mauvoisin, E. Martin, F. Atger, A. N. Galindo, L. Dayon, F. Sizzano, A. Palini, M. Kussmann, P. Waridel, M. Quadroni, V. Duli?, F. Naef, and F. Gachon, "Nuclear Proteomics Uncovers Diurnal Regulatory Landscapes in Mouse Liver," *Cell Metabolism*, vol. 25, no. 1, pp. 102–117, jan 2017.

[64] J. Mermet, J. Yeung, C. Hurni, D. Mauvoisin, K. Gustafson, C. Jouffe, D. Nicolas, Y. Emmenegger, C. Gobet, P. Franken, F. Gachon, and F. Naef, "Clock-dependent chromatin topology modulates circadian transcription and behavior," *Genes & Development*, vol. 32, no. 5-6, pp. 347–358, mar 2018.

[65] J. R. Beytebiere, A. J. Trott, B. J. Greenwell, C. A. Osborne, H. Vitet, J. Spence, S. H. Yoo, Z. Chen, J. S. Takahashi, N. Ghaffari, and J. S. Menet, "Tissue-specific BMAL1 cistromes reveal that rhythmic transcription is associated with rhythmic enhancer–enhancer interactions," *Genes and Development*, vol. 33, no. 5-6, pp. 294–309, mar 2019.

[66] J. Mermet, J. Yeung, and F. Naef, "Oscillating and stable genome topologies underlie hepatic

physiological rhythms during the circadian cycle," *PLOS Genetics*, vol. 17, no. 2, p. e1009350, feb 2021.

[67] J. P. Etchegaray, C. Lee, P. A. Wade, and S. M. Reppert, "Rhythmic histone acetylation underlies transcription in the mammalian circadian clock," *Nature*, vol. 421, no. 6919, pp. 177–182, jan 2003.

[68] G. Rey, F. Cesbron, J. Rougemont, H. Reinke, M. Brunner, and F. Naef, "Genome-wide and phase-specific DNA-binding rhythms of BMAL1 control circadian output functions in mouse liver." *PLoS biology*, vol. 9, no. 2, p. e1000595, feb 2011.

[69] J. A. Sobel, I. Krier, T. Andersin, S. Raghav, D. Canella, F. Gilardi, A. S. Kalantzi, G. Rey, B. Weger, F. Gachon, M. Dal Peraro, N. Hernandez, U. Schibler, B. Deplancke, and F. Naef, "Transcriptional regulatory logic of the diurnal cycle in the mouse liver," *PLoS Biology*, vol. 15, no. 4, apr 2017.

[70] A. Raj and A. van Oudenaarden, "Nature, nurture, or chance: stochastic gene expression and its consequences." *Cell*, vol. 135, no. 2, pp. 216–26, oct 2008.

[71] D. Nicolas, N. E. Phillips, and F. Naef, "What shapes eukaryotic transcriptional bursting?" pp. 1280–1290, 2017.

[72] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi, "Stochastic mRNA Synthesis in Mammalian Cells," *PLoS Biology*, vol. 4, no. 10, p. e309, sep 2006.

[73] D. M. Suter, N. Molina, D. Gatfield, K. Schneider, U. Schibler, and F. Naef, "Mammalian genes are transcribed with widely different bursting kinetics," *Science*, vol. 332, no. 6028, pp. 472–474, apr 2011.

[74] K. Bahar Halpern, S. Tanami, S. Landen, M. Chapal, L. Szlak, A. Hutzler, A. Nizhberg, and S. Itzkovitz, "Bursty gene expression in the intact mammalian liver," *Molecular Cell*, vol. 58, no. 1, pp. 147–156, apr 2015.

[75] D. Nicolas, B. Zoller, D. M. Suter, and F. Naef, "Modulation of transcriptional burst frequency by histone acetylation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, no. 27, pp. 7153–7158, jul 2018.

[76] N. E. Phillips, A. Hugues, J. Yeung, E. Durandau, D. Nicolas, and F. Naef, "The circadian oscillator analysed at the single-transcript level," *Molecular Systems Biology*, vol. 17, no. 3, mar 2021.

[77] A. J. Trott and J. S. Menet, "Regulation of circadian clock transcriptional output by CLOCK:BMAL1," *PLOS Genetics*, vol. 14, no. 1, p. e1007156, jan 2018.

[78] J. Wang, L. Symul, J. Yeung, C. Gobet, J. Sobel, S. Lück, P. O. Westermark, N. Molina, and F. Naef, "Circadian clock-dependent and -independent posttranscriptional regulation underlies temporal mRNA accumulation in mouse liver," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, no. 8, pp. E1916–E1925, 2018.

[79] S. Lück, K. Thurley, P. F. Thaben, and P. O. Westermark, "Rhythmic Degradation Explains and Unifies Circadian Transcriptome and Proteome Data," *Cell Reports*, vol. 9, no. 2, pp. 741–751, 2014.

[80] J. Mermet, J. Yeung, and F. Naef, "Systems Chronobiology: Global Analysis of Gene Regulation in a 24-Hour Periodic World," *Cold Spring Harbor Perspectives in Biology*, vol. 9, no. 3, p. a028720, mar 2017.

[81] M. C. Wahl, C. L. Will, and R. Lührmann, "The Spliceosome: Design Principles of a Dynamic RNP Machine," pp. 701–718, feb 2009.

[82] T. Nojima, K. Rebelo, T. Gomes, A. R. Grosso, N. J. Proudfoot, and M. Carmo-Fonseca, "RNA Polymerase II Phosphorylated on CTD Serine 5 Interacts with the Spliceosome during Co-transcriptional Splicing," *Molecular Cell*, vol. 72, no. 2, pp. 369–379.e4, oct 2018.

[83] M. K. Sakharkar, B. S. Perumal, K. R. Sakharkar, and P. Kangueane, "An Analysis on Gene Architecture in Human and Mouse Genomes," *In Silico Biology*, vol. 5, no. 4, pp. 347–365, jan 2005.

[84] L. De Conti, M. Baralle, and E. Buratti, "Exon and intron definition in pre-mRNA splicing," pp. 49–60, jan 2013.

[85] F. Carrillo Oesterreich, S. Preibisch, and K. M. Neugebauer, "Global analysis of nascent rna reveals transcriptional pausing in terminal exons," *Molecular Cell*, vol. 40, no. 4, pp. 571–581, nov 2010.

[86] L. Herzel, K. Straube, and K. M. Neugebauer, "Long-read sequencing of nascent RNA reveals coupling among RNA processing events," *Genome Research*, vol. 28, no. 7, pp. 1008–1019, jul 2018.

[87] Y. L. Khodor, J. Rodriguez, K. C. Abruzzi, C. H. A. Tang, M. T. Marr, and M. Rosbash, "Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila," *Genes and Development*, vol. 25, no. 23, pp. 2502–2512, dec 2011.

[88] T. Nojima, T. Gomes, A. R. F. Grosso, H. Kimura, M. J. Dye, S. Dhir, M. Carmo-Fonseca, and N. J. Proudfoot, "Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing," *Cell*, vol. 161, no. 3, pp. 526–540, apr 2015.

[89] A. Ameur, A. Zaghlool, J. Halvardson, A. Wetterbom, U. Gyllensten, L. Cavelier, and L. Feuk, "Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain," *Nature Structural and Molecular Biology*, vol. 18, no. 12, pp. 1435–1440, dec 2011.

[90] Y. L. Khodor, J. S. Menet, M. Tolan, and M. Rosbash, "Cotranscriptional splicing efficiency differs dramatically between Drosophila and mouse," *RNA*, vol. 18, no. 12, pp. 2174–2186, dec 2012.

[91] H. Tilgner, D. G. Knowles, R. Johnson, C. A. Davis, S. Chakrabortty, S. Djebali, J. Curado, M. Snyder, T. R. Gingeras, and R. Guigó, "Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs," *Genome Research*, vol. 22, no. 9, pp. 1616–1625, sep 2012.

[92] K. M. Neugebauer, "Nascent RNA and the coordination of splicing with transcription," *Cold Spring Harbor Perspectives in Biology*, vol. 11, no. 8, p. a032227, aug 2019.

[93] L. Herzel and K. M. Neugebauer, "Quantification of co-transcriptional splicing from RNA-Seq data," *Methods*, vol. 85, pp. 36–43, sep 2015.

[94] K. A. Reimer, C. A. Mimoso, K. Adelman, and K. M. Neugebauer, "Co-transcriptional splicing regulates 3 end cleavage during mammalian erythropoiesis," *Molecular Cell*, jan 2021.

[95] D. M. Bhatt, A. Pandya-Jones, A.-J. Tong, I. Barozzi, M. M. Lissner, G. Natoli, D. L. Black, and S. T. Smale, "Transcript Dynamics of Proinflammatory Genes Revealed by Sequence Analysis of Subcellular RNA Fractions," *Cell*, vol. 150, no. 2, pp. 279–290, jul 2012.

[96] H. L. Drexler, K. Choquet, and L. S. Churchman, "Human co-transcriptional splicing kinetics and coordination revealed by direct nascent RNA sequencing," *bioRxiv*, p. 611020, apr 2019.

[97] T. Alpert, K. Straube, F. Carrillo Oesterreich, and K. M. Neugebauer, "Widespread Transcriptional Readthrough Caused by Nab2 Depletion Leads to Chimeric Transcripts with Retained Introns," *Cell Reports*, vol. 33, no. 4, p. 108324, oct 2020.

[98] L. Herzel, D. S. Ottoz, T. Alpert, and K. M. Neugebauer, "Splicing and transcription touch base: Co-transcriptional spliceosome assembly and function," pp. 637–650, oct 2017.

[99] M. Sahebi, M. M. Hanafi, A. J. van Wijnen, P. Azizi, R. Abiri, S. Ashkani, and S. Taheri, "Towards understanding pre-mRNA splicing mechanisms and the role of SR proteins," pp. 107–119, aug 2016.

[100] G. Biamonti and J. F. Caceres, "Cellular stress and RNA splicing," pp. 146–153, mar 2009.

[101] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge, "Alternative isoform regulation in human tissue transcriptomes," *Nature*, vol. 456, no. 7221, pp. 470–476, nov 2008.

[102] R. El-Athman, L. Fuhr, and A. Relógio, "A Systems-Level Analysis Reveals Circadian Regulation of Splicing in Colorectal Cancer," *EBioMedicine*, vol. 33, pp. 68–81, jul 2018.

[103] N. J. McGlincy, A. Valomon, J. E. Chesham, E. S. Maywood, M. H. Hastings, and J. Ule, "Regulation of alternative splicing by the circadian clock and food related cues," *Genome Biology*, vol. 13, no. 6, jun 2012.

[104] M. Preussner, I. Wilhelmi, A. S. Schultz, F. Finkernagel, M. Michel, T. Möröy, and F. Heyd, "Rhythmic U2af26 Alternative Splicing Controls PERIOD1 Stability and the Circadian Clock in Mice," *Molecular Cell*, vol. 54, no. 4, pp. 651–662, may 2014.

[105] B. Marcheva, M. Perelis, B. J. Weidemann, A. Taguchi, H. Lin, C. Omura, Y. Kobayashi, M. V. Newman, E. J. Wyatt, E. M. McNally, J. E. Manning Fox, H. Hong, A. Shankar, E. C. Wheeler, K. M. Ramsey, P. E. MacDonald, G. W. Yeo, and J. Bass, "A role for alternative splicing in circadian

control of exocytosis and glucose homeostasis," *Genes and Development*, vol. 34, no. 15-16, pp. 1089–1105, jul 2020.

[106] M. Preussner, G. Goldammer, A. Neumann, T. Haltenhof, P. Rautenstrauch, M. Müller-McNicoll, and F. Heyd, "Body Temperature Cycles Control Rhythmic Alternative Splicing in Mammals," *Molecular Cell*, vol. 67, no. 3, pp. 433–446.e4, aug 2017.

[107] U. Braunschweig, N. L. Barbosa-Morais, Q. Pan, E. N. Nachman, B. Alipanahi, T. Gonatopoulos-Pournatzis, B. Frey, M. Irimia, and B. J. Blencowe, "Widespread intron retention in mammals functionally tunes transcriptomes," *Genome Research*, vol. 24, no. 11, pp. 1774–1786, nov 2014.

[108] P. L. Boutz, A. Bhutkar, and P. A. Sharp, "Detained introns are a novel, widespread class of post-transcriptionally spliced introns." *Genes & development*, vol. 29, no. 1, pp. 63–80, jan 2015.

[109] R. Middleton, D. Gao, A. Thomas, B. Singh, A. Au, J. J. Wong, A. Bomane, B. Cosson, E. Eyras, J. E. Rasko, and W. Ritchie, "IRFinder: Assessing the impact of intron retention on mammalian gene expression," *Genome Biology*, vol. 18, no. 1, mar 2017.

[110] A. G. Jacob and C. W. Smith, "Intron retention as a component of regulated gene expression programs," pp. 1043–1057, sep 2017.

[111] K. Ninomiya, N. Kataoka, and M. Hagiwara, "Stress-responsive maturation of Clk1/4 pre-mRNAs promotes phosphorylation of SR splicing factor," *Journal of Cell Biology*, vol. 195, no. 1, pp. 27–40, oct 2011.

[112] O. Mauger, F. Lemoine, and P. Scheiffele, "Targeted Intron Retention and Excision for Rapid Gene Regulation in Response to Neuronal Activity," *Neuron*, vol. 92, no. 6, pp. 1266–1278, dec 2016.

[113] M. Stewart, "Polyadenylation and nuclear export of mRNAs," *Journal of Biological Chemistry*, vol. 294, no. 9, pp. 2977–2987, mar 2019.

[114] A. F. Palazzo and E. S. Lee, "Sequence determinants for nuclear retention and cytoplasmic export of mRNAs and lncRNAs," oct 2018.

[115] J. Katahira, "MRNA export and the TREX complex," pp. 507–513, jun 2012.

[116] A. Mor, S. Suliman, R. Ben-Yishay, S. Yunger, Y. Brody, and Y. Shav-Tal, "Dynamics of single mRNP nucleocytoplasmic transport and export through the nuclear pore in living cells," *Nature Cell Biology*, vol. 12, no. 6, pp. 543–552, jun 2010.

[117] A. F. Palazzo and A. Akef, "Nuclear export as a key arbiter of "mRNA identity" in eukaryotes," pp. 566–577, jun 2012.

[118] C. J. Shukla, A. L. McCorkindale, C. Gerhardinger, K. D. Korthauer, M. N. Cabili, D. M. Shechner, R. A. Irizarry, P. G. Maass, and J. L. Rinn, "High-throughput identification of <scp>RNA</scp> nuclear enrichment sequences," *The EMBO Journal*, vol. 37, no. 6, mar 2018.

[119] Y. Lubelsky and I. Ulitsky, "Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells," *Nature*, vol. 555, no. 7694, pp. 107–111, mar 2018.

[120] D. Kim, B. Langmead, and S. L. Salzberg, "HISAT: a fast spliced aligner with low memory requirements." *Nature methods*, vol. 12, no. 4, pp. 357–60, apr 2015.

[121] J. Morf, G. Rey, K. Schneider, M. Stratmann, J. Fujita, F. Naef, and U. Schibler, "Cold-inducible RNA-binding protein modulates circadian gene expression posttranscriptionally," *Science*, vol. 338, no. 6105, pp. 379–383, oct 2012.

[122] G. Benegiamo, L. S. Mure, G. Erikson, H. D. Le, E. Moriggi, S. A. Brown, and S. Panda, "The RNA-Binding Protein NONO Coordinates Hepatic Adaptation to Feeding," *Cell Metabolism*, vol. 27, no. 2, pp. 404–418.e7, feb 2018.

[123] S. A. Brown, J. Ripperger, S. Kadener, F. Fleury-Olela, F. Vilbois, M. Rosbash, and U. Schibler, "Cell biology: PERIOD1-associated proteins modulate the negative limb of the mammalian circadian oscillator," *Science*, vol. 308, no. 5722, pp. 693–696, apr 2005.

[124] C. J. Guo, G. Xu, and L. L. Chen, "Mechanisms of Long Noncoding RNA Nuclear Retention," pp. 947–960, nov 2020.

[125] L. Statello, C. J. Guo, L. L. Chen, and M. Huarte, "Gene regulation by long non-coding RNAs and its biological functions," pp. 96–118, feb 2021.

[126] B. Zuckerman and I. Ulitsky, "Predictive models of subcellular localization of long RNAs," *RNA*, vol. 25, no. 5, pp. 557–572, feb 2019.

[127] N. Battich, T. Stoeger, and L. Pelkmans, “Control of Transcript Variability in Single Mammalian Cells,” *Cell*, vol. 163, no. 7, pp. 1596–1610, dec 2015.

[128] T. Stoeger, N. Battich, and L. Pelkmans, “Passive Noise Filtering by Cellular Compartmentalization,” *Cell*, vol. 164, no. 6, pp. 1151–1161, mar 2016.

[129] J. M. Gordon, D. V. Phizicky, and K. M. Neugebauer, “Nuclear mechanisms of gene expression control: pre-mRNA splicing as a life or death decision,” pp. 67–76, apr 2021.

[130] D. L. Spector and A. I. Lamond, “Nuclear speckles,” *Cold Spring Harbor Perspectives in Biology*, vol. 3, no. 2, pp. 1–12, 2011.

[131] C. Girard, C. L. Will, J. Peng, E. M. Makarov, B. Kastner, I. Lemm, H. Urlaub, K. Hartmuth, and R. Lührmann, “Post-transcriptional spliceosomes are retained in nuclear speckles until splicing completion,” *Nature Communications*, vol. 3, no. 1, p. 994, jan 2012.

[132] J. M. Engreitz, K. Sirokman, P. McDonel, A. A. Shishkin, C. Surka, P. Russell, S. R. Grossman, A. Y. Chow, M. Guttman, and E. S. Lander, “RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent pre-mRNAs and chromatin sites,” *Cell*, vol. 159, no. 1, pp. 188–199, sep 2014.

[133] S. E. Liao and O. Regev, “Splicing at the phase-separated nuclear speckle interface: a model,” *Nucleic Acids Research*, vol. 49, no. 2, pp. 636–645, jan 2021.

[134] A. P. Dias, K. Dufu, H. Lei, and R. Reed, “A role for TREX components in the release of spliced mRNA from nuclear speckle domains,” *Nature Communications*, vol. 1, no. 7, pp. 1–10, oct 2010.

[135] K. Wang, L. Wang, J. Wang, S. Chen, M. Shi, and H. Cheng, “Intronless mRNAs transit through nuclear speckles to gain export competence,” *Journal of Cell Biology*, vol. 217, no. 11, pp. 3912–3929, nov 2018.

[136] J. H. Su, P. Zheng, S. S. Kinrot, B. Bintu, and X. Zhuang, “Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin,” *Cell*, vol. 182, no. 6, pp. 1641–1659.e26, sep 2020.

[137] F. Ding and M. B. Elowitz, “Constitutive splicing and economies of scale in gene expression,” *Nature Structural & Molecular Biology*, p. 1, may 2019.

[138] A. H. Fox, Y. W. Lam, A. K. Leung, C. E. Lyon, J. Andersen, M. Mann, and A. I. Lamond, “Paraspeckles: A novel nuclear domain,” *Current Biology*, vol. 12, no. 1, pp. 13–25, jan 2002.

[139] T. Hirose, T. Yamazaki, and S. Nakagawa, “Molecular anatomy of the architectural NEAT1 noncoding RNA: The domains, interactors, and biogenesis pathway required to build phase-separated nuclear paraspeckles,” *Wiley Interdisciplinary Reviews: RNA*, vol. 10, no. 6, p. e1545, nov 2019.

[140] B. L. Bass, “RNA editing by adenosine deaminases that act on RNA,” pp. 817–846, nov 2002.

[141] L. L. Chen, J. N. DeCerbo, and G. G. Carmichael, “Alu element-mediated gene silencing,” *EMBO Journal*, vol. 27, no. 12, pp. 1694–1705, jun 2008.

[142] K. V. Prasanth, S. G. Prasanth, Z. Xuan, S. Hearn, S. M. Freier, C. F. Bennett, M. Q. Zhang, and D. L. Spector, “Regulating gene expression through RNA nuclear retention,” *Cell*, vol. 123, no. 2, pp. 249–263, oct 2005.

[143] M. Torres, D. Becquet, M. P. Blanchard, S. Guillen, B. Boyer, M. Moreno, J. L. Franc, and A. M. François-Bellan, “Circadian RNA expression elicited by 3’-UTR IRAlu-paraspeckle associated elements,” *eLife*, vol. 5, no. JULY, jul 2016.

[144] B. S. Zhao, I. A. Roundtree, and C. He, “Post-transcriptional gene regulation by mRNA modifications,” pp. 31–42, dec 2016.

[145] S. Lesbirel, N. Viphakone, M. Parker, J. Parker, C. Heath, I. Sudbery, and S. A. Wilson, “The m6A-methylase complex recruits TREX and regulates mRNA export,” *Scientific Reports*, vol. 8, no. 1, p. 13827, dec 2018.

[146] W. Xiao, S. Adhikari, U. Dahal, Y. S. Chen, Y. J. Hao, B. F. Sun, H. Y. Sun, A. Li, X. L. Ping, W. Y. Lai, X. Wang, H. L. Ma, C. M. Huang, Y. Yang, N. Huang, G. B. Jiang, H. L. Wang, Q. Zhou, X. J. Wang, Y. L. Zhao, and Y. G. Yang, “Nuclear m6A Reader YTHDC1 Regulates mRNA Splicing,” *Molecular Cell*, vol. 61, no. 4, pp. 507–519, feb 2016.

[147] J.-M. Fustin, M. Doi, Y. Yamaguchi, H. Hida, S. Nishimura, M. Yoshida, T. Isagawa, M. S. Morioka, H. Kakeya, I. Manabe, and H. Okamura, “RNA-methylation-dependent RNA processing controls the speed of the circadian clock.” *Cell*, vol. 155, no. 4, pp. 793–806, nov 2013.

[148] I. A. Roundtree, G. Z. Luo, Z. Zhang, X. Wang, T. Zhou, Y. Cui, J. Sha, X. Huang, L. Guerrero, P. Xie, E. He, B. Shen, and C. He, “YTHDC1 mediates nuclear export of N6-methyladenosine methylated mRNAs,” *eLife*, vol. 6, oct 2017.

[149] X. Wang, B. S. Zhao, I. A. Roundtree, Z. Lu, D. Han, H. Ma, X. Weng, K. Chen, H. Shi, and C. He, “N6-methyladenosine modulates messenger RNA translation efficiency,” *Cell*, vol. 161, no. 6, pp. 1388–1399, jun 2015.

[150] X. Wang, Z. Lu, A. Gomez, G. C. Hon, Y. Yue, D. Han, Y. Fu, M. Parisien, Q. Dai, G. Jia, B. Ren, T. Pan, and C. He, “N 6-methyladenosine-dependent regulation of messenger RNA stability,” *Nature*, vol. 505, no. 7481, pp. 117–120, 2014.

[151] O. H. Park, H. Ha, Y. Lee, S. H. Boo, D. H. Kwon, H. K. Song, and Y. K. Kim, “Endoribonucleolytic Cleavage of m6A-Containing RNAs by RNase P/MRP Complex,” *Molecular Cell*, vol. 74, no. 3, pp. 494–507.e8, may 2019.

[152] N. L. Garneau, J. Wilusz, and C. J. Wilusz, “The highways and byways of mRNA decay,” pp. 113–126, feb 2007.

[153] S. Kojima and C. B. Green, “Circadian genomics reveal a role for post-transcriptional regulation in mammals,” *Biochemistry*, vol. 54, no. 2, pp. 124–133, jan 2015.

[154] S. Kojima, E. L. Sher-Chen, and C. B. Green, “Circadian control of mRNA polyadenylation dynamics regulates rhythmic protein expression,” *Genes and Development*, vol. 26, no. 24, pp. 2724–2736, 2012.

[155] T. J. Eisen, S. W. Eichhorn, A. O. Subtelny, K. S. Lin, S. E. McGeary, S. Gupta, and D. P. Bartel, “The Dynamics of Cytoplasmic mRNA Metabolism,” *Molecular Cell*, vol. 77, no. 4, pp. 786–799.e10, feb 2020.

[156] K. C. Woo, T. D. Kim, K. H. Lee, D. Y. Kim, W. Kim, K. Y. Lee, and K. T. Kim, “Mouse period 2 mRNA circadian oscillation is modulated by PTB-mediated rhythmic mRNA degradation,” *Nucleic Acids Research*, vol. 37, no. 1, pp. 26–37, 2009.

[157] S. Kojima, K. Matsumoto, M. Hirose, M. Shimada, M. Nagano, Y. Shigeyoshi, S. I. Hoshino, K. Ui-Tei, K. Saigo, C. B. Green, Y. Sakaki, and H. Tei, “LARK activates posttranscriptional expression of an essential mammalian clock protein, PERIOD1,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 6, pp. 1859–1864, feb 2007.

[158] W. V. So and M. Rosbash, “Post-transcriptional regulation contributes to Drosophila clock gene mRNA cycling,” *EMBO Journal*, vol. 16, no. 23, pp. 7146–7155, dec 1997.

[159] S. Colnot and C. Perret, “Liver Zonation,” 2011, pp. 7–16.

[160] K. B. Halpern, R. Shenhav, O. Matcovitch-Natan, B. Tóth, D. Lemze, M. Golan, E. E. Massasa, S. Baydatch, S. Landen, A. E. Moor, A. Brandis, A. Giladi, A. Stokar-Avihail, E. David, I. Amit, and S. Itzkovitz, “Single-cell spatial reconstruction reveals global division of labour in the mammalian liver,” *Nature*, vol. 542, no. 7641, pp. 1–5, feb 2017.

[161] S. Benhamouche, T. Decaens, C. Godard, R. Chambrey, D. S. Rickman, C. Moinard, M. Vasseur-Cognet, C. J. Kuo, A. Kahn, C. Perret, and S. Colnot, “Apc tumor suppressor gene is the "zonation-keeper" of mouse liver.” *Developmental cell*, vol. 10, no. 6, pp. 759–70, jun 2006.

[162] K. Jungermann and T. Kietzmann, “Zonation of parenchymal and nonparenchymal metabolism in liver.” *Annual review of nutrition*, vol. 16, pp. 179–203, 1996.

[163] S. Ben-Moshe, Y. Shapira, A. E. Moor, R. Manco, T. Veg, K. Bahar Halpern, and S. Itzkovitz, “Spatial sorting enables comprehensive characterization of liver zonation,” *Nature Metabolism*, vol. 1, no. 9, pp. 899–911, sep 2019.

[164] T. Kietzmann, “Metabolic zonation of the liver: The oxygen gradient revisited,” *Redox Biology*, vol. 11, pp. 622–630, 2017.

[165] K. Jungermann and N. Katz, “Functional specialization of different hepatocyte populations,” pp. 708–764, 1989.

[166] A. Braeuning, C. Ittrich, C. K?hle, S. Hailfinger, M. Bonin, A. Buchmann, and M. Schwarz, "Differential gene expression in periportal and perivenous mouse hepatocytes," *FEBS Journal*, vol. 273, no. 22, pp. 5051–5061, nov 2006.

[167] J. Schleicher, C. Tokarski, E. Marbach, M. Matz-Soja, S. Zellmer, R. Gebhardt, and S. Schuster, "Zonation of hepatic fatty acid metabolism — The diversity of its regulation and the benefit of modeling," *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, vol. 1851, no. 5, pp. 641–656, 2015.

[168] C. Droin, J. E. Kholtei, K. Bahar Halpern, C. Hurni, M. Rozenberg, S. Muvkadi, S. Itzkovitz, and F. Naef, "Space-time logic of liver gene expression at sub-lobular scale," *Nature Metabolism*, vol. 3, no. 1, pp. 43–58, jan 2021.

[169] N. Katz, J. Thiele, Giffhron-Katz, and Susanne, "Zonal distribution of fatty acid synthase in liver parenchyma of male and female rats," *European Journal of Biochemistry*, vol. 180, no. 1, pp. 185–189, mar 1989.

[170] J. Schleicher, U. Dahmen, R. Guthke, and S. Schuster, "Zonation of hepatic fat accumulation: insights from mathematical modelling of nutrient gradients and fatty acid uptake," *Journal of The Royal Society Interface*, vol. 14, no. 133, 2017.

[171] K. O. Lindros, "Zonation of cytochrome P450 expression, drug metabolism and toxicity in liver," *General Pharmacology: The Vascular System*, vol. 28, no. 2, pp. 191–196, feb 1997.

[172] M. R. McGill and H. Jaeschke, "Metabolism and disposition of acetaminophen: Recent advances in relation to hepatotoxicity and diagnosis," pp. 2174–2187, sep 2013.

[173] D. Häussinger, "Liver glutamine metabolism." pp. 56S–62S, 1990.

[174] K. Jungermann and T. Kietzmann, "Oxygen: Modulator of metabolic zonation and disease of the liver," pp. 255–260, 2000.

[175] R. Gebhardt, K. S. Lerche, F. Götschel, R. Günther, J. Kolander, L. Teich, S. Zellmer, H.-J. Hofmann, K. Eger, A. Hecht, and F. Gaunitz, "4-Aminoethylamino-emodin–a novel potent inhibitor of GSK-3beta–acts as an insulin-sensitizer avoiding downstream effects of activated beta-catenin." *Journal of cellular and molecular medicine*, vol. 14, no. 6A, pp. 1276–93, jun 2010.

[176] A. Gougelet, C. Torre, P. Veber, C. Sartor, L. Bachelot, P.-D. Denechaud, C. Godard, M. Moldes, A.-F. Burnol, C. Dubuquoy, B. Terris, F. Guillonneau, T. Ye, M. Schwarz, A. Braeuning, C. Perret, and S. Colnot, "T-cell factor 4 and ?-catenin chromatin occupancies pattern zonal liver metabolism in mice," *Hepatology*, vol. 59, no. 6, pp. 2344–2357, jun 2014.

[177] S. Sekine, P. J. A. Gutiérrez, B. Y.-A. Lan, S. Feng, and M. Hebrok, "Liver-specific loss of beta-catenin results in delayed hepatocyte proliferation after partial hepatectomy." *Hepatology (Baltimore, Md.)*, vol. 45, no. 2, pp. 361–8, feb 2007.

[178] B. Wang, L. Zhao, M. Fish, C. Y. Logan, and R. Nusse, "Self-renewing diploid Axin2(+) cells fuel homeostatic renewal of the liver." *Nature*, vol. 524, no. 7564, pp. 180–5, aug 2015.

[179] M. Preziosi, H. Okabe, M. Poddar, S. Singh, and S. P. Monga, "Endothelial Wnts regulate β-catenin signaling in murine liver zonation and regeneration: A sequel to the Wnt-Wnt situation," *Hepatology Communications*, vol. 2, no. 7, pp. 845–860, jul 2018.

[180] M. Matz-Soja, C. Rennert, K. Schönefeld, S. Aleithe, J. Boettger, W. Schmidt-Heck, T. Weiss, A. Hovhannisyan, S. Zellmer, N. Klöting, A. Schulz, J. Kratzsch, R. Guthke, and R. Gebhardt, "Hedgehog signaling is a potent regulator of liver lipid metabolism and reveals a GLI-code associated with steatosis," *eLife*, vol. 5, no. MAY2016, may 2016.

[181] E. Kolbe, S. Aleithe, C. Rennert, L. Spormann, F. Ott, D. Meierhofer, R. Gajowski, C. Stöpel, S. Hoehme, M. Kücken, L. Brusch, M. Seifert, W. von Schoenfels, C. Schafmayer, M. Brosch, U. Hofmann, G. Damm, D. Seehofer, J. Hampe, R. Gebhardt, and M. Matz-Soja, "Mutual Zonated Interactions of Wnt and Hh Signaling Are Orchestrating the Metabolism of the Adult Liver in Mice and Human," *Cell Reports*, vol. 29, no. 13, pp. 4553–4567.e7, dec 2019.

[182] X. Cheng, S. Y. Kim, H. Okamoto, Y. Xin, G. D. Yancopoulos, A. J. Murphy, and J. Gromada, "Glucagon contributes to liver zonation." *Proceedings of the National Academy of Sciences of the United States of America*, p. 201721403, mar 2018.

[183] G. Gentric, S. Celton-Morizur, and C. Desdouets, “Polyploidy and liver proliferation,” pp. 29–34, feb 2012.

[184] A. Miyajima, M. Tanaka, and T. Itoh, “Stem/progenitor cells in liver development, homeostasis, regeneration, and reprogramming.” *Cell stem cell*, vol. 14, no. 5, pp. 561–74, may 2014.

[185] H. Morales-Navarrete, F. Segovia-Miranda, P. Klukowski, K. Meyer, H. Nonaka, G. Marsico, M. Chernykh, A. Kalaidzidis, M. Zerial, and Y. Kalaidzidis, “A versatile pipeline for the multi-scale digital reconstruction and quantitative analysis of 3D tissue architecture,” *eLife*, vol. 4, no. DECEMBER2015, dec 2015.

[186] S. Tanami, S. Ben-Moshe, A. Elkayam, A. Mayo, K. Bahar Halpern, and S. Itzkovitz, “Dynamic zonation of liver polyploidy,” *Cell and Tissue Research*, vol. 368, no. 2, pp. 405–410, may 2017.

[187] K. B. Halpern, R. Shenhav, O. Matcovitch-Natan, B. Tóth, D. Lemze, M. Golan, E. E. Massasa, S. Baydatch, S. Landen, A. E. Moor, A. Brandis, A. Giladi, A. Stokar-Avihail, E. David, I. Amit, and S. Itzkovitz, “Single-cell spatial reconstruction reveals global division of labour in the mammalian liver,” *Nature*, 2017.

[188] U. Schibler, “The daily timing of gene expression and physiology in mammals.” *Dialogues in clinical neuroscience*, vol. 9, no. 3, pp. 257–72, 2007.

[189] C. B. Green, J. S. Takahashi, and J. Bass, “The Meter of Metabolism,” pp. 728–742, sep 2008.

[190] A. M. Femino, F. S. Fay, K. Fogarty, and R. H. Singer, “Visualization of single RNA transcripts in situ,” *Science*, vol. 280, no. 5363, pp. 585–590, apr 1998.

[191] D. Grunwald and R. H. Singer, “In vivo imaging of labelled endogenous beta actin mRNA during nucleocytoplasmic transport,” *Nature*, vol. 467, no. 7315, pp. 604–607, sep 2010.

[192] A. Wilczynska and M. Bushell, “The complexity of miRNA-mediated repression,” pp. 22–33, jan 2015.

[193] A. Chin and E. Lécuyer, “RNA localization: Making its way to the center stage,” pp. 2956–2970, nov 2017.

[194] S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Röder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, D. Robyr, M. Sammeth, L. Schaffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guig, and T. R. Gingeras, “Landscape of transcription in human cells,” *Nature*, vol. 489, no. 7414, pp. 101–108, sep 2012.

[195] I. Gotic, S. Omidi, F. Fleury-Olela, N. Molina, F. Naef, and U. Schibler, “Temperature regulates splicing efficiency of the cold-inducible RNA-binding protein gene Cirbp,” *Genes and Development*, vol. 30, no. 17, pp. 2005–2017, sep 2016.

[196] K. Bahar Halpern, I. Caspi, D. Lemze, M. Levy, S. Landen, E. Elinav, I. Ulitsky, and S. Itzkovitz, “Nuclear Retention of mRNA in Mammalian Tissues,” *Cell Reports*, vol. 13, no. 12, pp. 2653–2662, dec 2015.

[197] C. Vollmers, S. Gill, L. DiTacchio, S. R. Pulivarthy, H. D. Le, and S. Panda, “Time of feeding and the intrinsic circadian clock drive rhythms in hepatic gene expression,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 50, pp. 21 453–21 458, dec 2009.

[198] E.-J. Kim, Y.-S. Yoon, S. Hong, H.-Y. Son, T.-Y. Na, M.-H. Lee, H.-J. Kang, J. Park, W.-J. Cho, S.-G. Kim, S.-H. Koo, H.-g. Park, and M.-O. Lee, “Retinoic acid receptor-related orphan receptor α-induced activation of adenosine monophosphate-activated protein kinase results in attenuation of hepatic steatosis.” *Hepatology (Baltimore, Md.)*, vol. 55, no. 5, pp. 1379–88, may 2012.

[199] N.-H. Du, A. B. Arpat, M. De Matos, and D. Gatfield, “MicroRNAs shape circadian hepatic gene expression on a transcriptome-wide scale.” *eLife*, vol. 3, p. e02510, may 2014.

[200] J. E. Baggs and C. B. Green, “Nocturnin, a deadenylase in Xenopus laevis retina: A mechanism for posttranscriptional control of circadian-related mRNA,” *Current Biology*, vol. 13, no. 3, pp. 189–198, feb 2003.

[201] J. Wang, L. Symul, J. Yeung, C. Gobet, J. Sobel, S. Lück, P. O. Westermark, N. Molina, and F. Naef, “Circadian clock-dependent and -independent posttranscriptional regulation underlies temporal mRNA accumulation in mouse liver,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, no. 8, pp. E1916–E1925, feb 2018.

[202] S. Zhao, L. Xi, and B. Zhang, “Union Exon Based Approach for RNA-Seq Gene Quantification: To Be or Not to Be?” *PLOS ONE*, vol. 10, no. 11, p. e0141910, nov 2015.

[203] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, “Near-optimal probabilistic RNA-seq quantification,” *Nature Biotechnology*, vol. 34, no. 5, pp. 525–527, may 2016.

[204] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” *Genome Biology*, vol. 15, no. 12, p. 550, dec 2014.

[205] W. A. Decatur and M. J. Fournier, “RNA-guided nucleotide modification of ribosomal and other RNAs,” pp. 695–698, jan 2003.

[206] A. Helwak, G. Kudla, T. Dudnakova, and D. Tollervey, “Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding,” *Cell*, vol. 153, no. 3, pp. 654–665, apr 2013.

[207] S. Zhao, Z. Ye, and R. Stanton, “Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols,” *RNA*, vol. 26, no. 8, p. rna.074922.120, aug 2020.

[208] F. M. Fazal, S. Han, K. R. Parker, P. Kaewsapsak, J. Xu, A. N. Boettiger, H. Y. Chang, and A. Y. Ting, “Atlas of Subcellular RNA Localization Revealed by APEX-Seq,” *Cell*, vol. 178, no. 2, pp. 473–490.e26, jul 2019.

[209] T. Chen and B. van Steensel, “Comprehensive analysis of nucleocytoplasmic dynamics of mRNA in Drosophila cells,” *PLoS Genetics*, vol. 13, no. 8, p. e1006929, aug 2017.

[210] L. W. Barrett, S. Fletcher, and S. D. Wilton, “Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements,” pp. 3613–3634, nov 2012.

[211] D. A. Jackson, A. Pombo, and F. Iborra, “The balance sheet for transcription: an analysis of nuclear RNA metabolism in mammalian cells,” *The FASEB Journal*, vol. 14, no. 2, pp. 242–254, feb 2000.

[212] J. Carlevaro-Fita and R. Johnson, “Global Positioning System: Understanding Long Noncoding RNAs through Subcellular Localization,” pp. 869–883, mar 2019.

[213] N. Krahmer, B. Najafi, F. Schueder, F. Quagliarini, M. Steger, S. Seitz, R. Kasper, F. Salinas, J. Cox, N. H. Uhlenhaut, T. C. Walther, R. Jungmann, A. Zeigerer, G. H. H. Borner, and M. Mann, “Organellar Proteomics and Phospho-Proteomics Reveal Subcellular Reorganization in Diet-Induced Hepatic Steatosis,” *Developmental Cell*, vol. 47, no. 2, pp. 205–221.e7, oct 2018.

[214] C. Xia, J. Fan, G. Emanuel, J. Hao, and X. Zhuang, “Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 116, no. 39, pp. 19 490–19 499, sep 2019.

[215] D. Gaidatzis, L. Burger, M. Florescu, and M. B. Stadler, “Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation,” *Nature Biotechnology*, vol. 33, no. 7, pp. 722–729, jul 2015.

[216] M. Furlan, E. Galeota, N. Del Gaudio, E. Dassi, M. Caselle, S. de Pretis, and M. Pelizzola, “Genome-wide dynamics of RNA synthesis, processing and degradation without RNA metabolic labeling,” 2019.

[217] J. Katahira, “Nuclear Export of Messenger RNA,” *Genes*, vol. 6, pp. 163–184, 2015.

[218] B. R. Sabari, A. Dall’Agnese, A. Boija, I. A. Klein, E. L. Coffey, K. Shrinivas, B. J. Abraham, N. M. Hannett, A. V. Zamudio, J. C. Manteiga, C. H. Li, Y. E. Guo, D. S. Day, J. Schuijers, E. Vasile, S. Malik, D. Hnisz, T. I. Lee, I. I. Cisse, R. G. Roeder, P. A. Sharp, A. K. Chakraborty, and R. A. Young, “Coactivator condensation at super-enhancers links phase separation and gene control,” *Science*, vol. 361, no. 6400, p. eaar3958, jul 2018.

[219] A. M. Ishov, A. Gurumurthy, and J. Bungert, “Coordination of transcription, processing, and export of highly expressed RNAs by distinct biomolecular condensates,” *Emerging Topics in Life Sciences*, apr 2020.

[220] M. Melé, K. Mattioli, W. Mallard, D. M. Shechner, C. Gerhardinger, and J. L. Rinn, “Chromatin

environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs," *Genome Research*, vol. 27, no. 1, pp. 27–37, jan 2017.

[221] M. Schlackow, T. Nojima, T. Gomes, A. Dhir, M. Carmo-Fonseca, and N. J. Proudfoot, "Distinctive Patterns of Transcription and RNA Processing for Human lincRNAs," *Molecular Cell*, vol. 65, no. 1, pp. 25–38, jan 2017.

[222] S. L. Wolin and L. E. Maquat, "Cellular RNA surveillance in health and disease," pp. 822–827, nov 2019.

[223] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer, "Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood," vol. 25, no. 15, pp. 1923–1929, 2009.

[224] A. Lugowski, B. Nicholson, and O. S. Rissland, "Determining mRNA half-lives on a transcriptome-wide scale," *Methods*, vol. 137, pp. 90–98, mar 2018.

[225] B. Schwanhüusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach, "Global quantification of mammalian gene expression control," *Nature*, vol. 473, no. 7347, pp. 337–342, may 2011.

[226] V. A. Herzog, B. Reichholf, T. Neumann, P. Rescheneder, P. Bhat, T. R. Burkard, W. Wlotzka, A. Von Haeseler, J. Zuber, and S. L. Ameres, "Thiol-linked alkylation of RNA to assess expression dynamics," *Nature Methods*, vol. 14, no. 12, pp. 1198–1204, dec 2017.

[227] H. Tani and N. Akimitsu, "Genome-wide technology for determining RNA stability in mammalian cells: Historical perspective and recent advantages based on modified nucleotide labeling," pp. 1233–1238, 2012.

[228] B. D. Weger, C. Gobet, F. P. David, F. Atger, E. Martin, N. E. Phillips, A. Charpagne, M. Weger, F. Naef, and F. Gachon, "Systematic analysis of differential rhythmic liver gene expression mediated by the circadian clock and feeding rhythms," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 118, no. 3, jan 2021.

[229] S. Jabs, A. Biton, C. Bécavin, M. A. Nahori, A. Ghozlane, A. Pagliuso, G. Spanò, V. Guérineau, D. Touboul, Q. Giai Gianetto, T. Chaze, M. Matondo, M. A. Dillies, and P. Cossart, "Impact of the gut microbiota on the m6A epitranscriptome of mouse cecum and liver," *Nature Communications*, vol. 11, no. 1, pp. 1–16, dec 2020.

[230] W. Garland and T. H. Jensen, "Nuclear sorting of RNA," *WIREs RNA*, vol. 11, no. 2, mar 2020.

[231] Y. Wang, Z. Kuang, X. Yu, K. A. Ruhn, M. Kubo, and L. V. Hooper, "The intestinal microbiota regulates body composition through NFIL3 and the circadian clock," *Science*, vol. 357, no. 6354, 2017.

[232] M. Rabani, R. Raychowdhury, M. Jovanovic, M. Rooney, D. J. Stumpo, A. Pauli, N. Hacohen, A. F. Schier, P. J. Blackshear, N. Friedman, I. Amit, and A. Regev, "High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies," *Cell*, vol. 159, no. 7, pp. 1698–1710, dec 2014.

[233] T. Saldi, M. A. Cortazar, R. M. Sheridan, and D. L. Bentley, "Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing," pp. 2623–2635, jun 2016.

[234] H. L. Drexler, K. Choquet, and L. S. Churchman, "Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores," *Molecular Cell*, vol. 77, no. 5, pp. 985–998.e8, mar 2020.

[235] L. T. Lam, O. K. Pickeral, A. C. Peng, A. Rosenwald, E. M. Hurt, J. M. Giltnane, L. M. Averett, H. Zhao, R. E. Davis, M. Sathyamoorthy, L. M. Wahl, E. D. Harris, J. A. Mikovits, A. P. Monks, M. G. Hollingshead, E. A. Sausville, and L. M. Staudt, "Genomic-scale measurement of mRNA turnover and the mechanisms of action of the anti-cancer drug flavopiridol." *Genome biology*, vol. 2, no. 10, p. research0041.1, sep 2001.

[236] C. Miller, B. Schwalb, K. Maier, D. Schulz, S. Dümcke, B. Zacher, A. Mayer, J. Sydow, L. Marcinowski, L. Dölken, D. E. Martin, A. Tresch, and P. Cramer, "Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast," *Molecular Systems Biology*, vol. 7, no. 1, p. 458, jan 2011.

[237] G. L. Manno, R. Soldatov, H. Hochgerner, A. Zeisel, V. Petukhov, M. E. Kastriti, P. Lönnerberg, A. Furlan, J. Fan, Z. Liu, D. Van Bruggen, J. Guo, E. Sundström, G. Castelo-Branco, I. Adameyko, S. Linnarsson, and P. V. Kharchenko, "RNA velocity in single cells," 2017.

[238] A. F. Palazzo and Y. M. Kang, "GC-content biases in protein-coding genes act as an "mRNA identity" feature for nuclear export," *BioEssays*, vol. 43, no. 2, p. 2000197, feb 2021.

[239] M. E. Hughes, L. DiTacchio, K. R. Hayes, C. Vollmers, S. Pulivarthy, J. E. Baggs, S. Panda, and J. B. Hogenesch, "Harmonics of Circadian Gene Transcription in Mammals," *PLoS Genetics*, vol. 5, no. 4, p. e1000442, apr 2009.

[240] G. T. Van Der Horst, M. Muijtjens, K. Kobayashi, R. Takano, S. I. Kanno, M. Takao, J. De Wit, A. Verkerk, A. P. Eker, D. Van Leenen, R. Buijs, D. Bootsma, J. H. Hoeijmakers, and A. Yasui, "Mammalian Cry1 and Cry2 are essential for maintenance of circadian rhythms," *Nature*, vol. 398, no. 6728, pp. 627–630, apr 1999.

[241] B. Kornmann, O. Schaad, H. Bujard, J. S. Takahashi, and U. Schibler, "System-driven and oscillator-dependent circadian transcription in mice with a conditionally active liver clock." *PLoS biology*, vol. 5, no. 2, p. e34, feb 2007.

[242] G. Gentric and C. Desdouets, "Polyploidization in liver tissue," pp. 322–331, feb 2014.

[243] D. W. Reid and C. V. Nicchitta, "Comment on "principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling"," *Science*, vol. 348, no. 6240, pp. 1217–a, jun 2015.

[244] Q. Chen, S. Jagannathan, D. W. Reid, T. Zheng, and C. V. Nicchitta, "Hierarchical regulation of mRNA partitioning between the cytoplasm and the endoplasmic reticulum of mammalian cells," *Molecular Biology of the Cell*, vol. 22, no. 14, pp. 2646–2658, jul 2011.

[245] C. H. Jan, C. C. Williams, and J. S. Weissman, "Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling," *Science*, vol. 346, no. 6210, nov 2014.

[246] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, "Stochastic gene expression in a single cell," *Science*, vol. 297, no. 5584, pp. 1183–1186, aug 2002.

[247] P. D. Wilkinson, E. R. Delgado, F. Alencastro, M. P. Leek, N. Roy, M. P. Weirich, E. C. Stahl, P. A. Otero, M. I. Chen, W. K. Brown, and A. W. Duncan, "The Polyploid State Restricts Hepatocyte Proliferation and Liver Regeneration in Mice," *Hepatology*, vol. 69, no. 3, pp. 1242–1258, mar 2019.

[248] A. Senecal, B. Munsky, F. Proux, N. Ly, F. E. Braye, C. Zimmer, F. Mueller, and X. Darzacq, "Transcription factors modulate c-Fos transcriptional bursts," *Cell Reports*, vol. 8, no. 1, pp. 75–83, jul 2014.

[249] S. O. Skinner, H. Xu, S. Nagarkar-Jaiswal, P. R. Freire, T. P. Zwaka, and I. Golding, "Single-cell analysis of transcription kinetics across the cell cycle," *eLife*, vol. 5, no. JANUARY2016, jan 2016.

[250] A. Raj and S. Tyagi, "Detection of Individual Endogenous RNA Transcripts In Situ Using Multiple Singly Labeled Probes," in *Methods in Enzymology*. Academic Press Inc., jan 2010, vol. 472, pp. 365–386.

[251] C. R. Brown, C. Mao, E. Falkovskaia, M. S. Jurica, and H. Boeger, "Linking Stochastic Fluctuations in Chromatin Structure and Gene Expression," *PLoS Biology*, vol. 11, no. 8, p. 1001621, aug 2013.

[252] S. S. Dey, J. E. Foley, P. Limsirichai, D. V. Schaffer, and A. P. Arkin, "Orthogonal control of expression mean and variance by epigenetic features at different genomic loci," *Molecular Systems Biology*, vol. 11, no. 5, p. 806, may 2015.

[253] J. A. Ripperger and U. Schibler, "Rhythmic CLOCK-BMAL1 binding to multiple E-box motifs drives circadian Dbp transcription and chromatin transitions," *Nature Genetics*, vol. 38, no. 3, pp. 369–374, mar 2006.

[254] N. C. Martin, C. T. McCullough, P. G. Bush, L. Sharp, A. C. Hall, and D. J. Harrison, "Functional analysis of mouse hepatocytes differing in DNA content: Volume, receptor expression, and effect of IFNγ," *Journal of Cellular Physiology*, vol. 191, no. 2, pp. 138–144, 2002.

[255] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen, and

A. Ma’ayan, “Enrichr: a comprehensive gene set enrichment analysis web server 2016 update,” *Nucleic acids research*, vol. 44, no. W1, pp. W90–W97, jul 2016.

[256] A. Alexa, J. Rahnenführer, and T. Lengauer, “Improved scoring of functional groups from gene expression data by decorrelating GO graph structure,” *Bioinformatics*, vol. 22, no. 13, pp. 1600–1607, jul 2006.

[257] W. Q. Meeker and L. A. Escobar, “Teaching about Approximate Confidence Regions Based on Maximum Likelihood Estimation,” *The American Statistician*, vol. 49, no. 1, p. 48, feb 1995.

Clémence Yumie **Hurni**

Bioengineer (ing. bioing. dipl. EPF)

Bvd. de Grancy 14
1006 Lausanne

☎ +41 79 830 89 04

clemencehh@hotmail.com

Age: 29 years old
Marital status: single
Swiss / Japanese

www.linkedin.com/in/cyhurni

Education

Ph.D Candidate in Bioengineering
EPFL 2016 - 2021

Master of Science in Bioengineering
EPFL 2013- 2015
Award of Excellence
Exchange year at Kyoto Prefectural University of Medicine, Japan

Bachelor of Life Sciences and Technology
EPFL 2010-2013

Languages

French	Mother tongue
English	Fluent
Japanese	Mother tongue (spoken)
German	Intermediate

Softwares/IT

R (excellent)
ImageJ (very good)
Matlab (intermediate)
Python (basic)
LaTeX
Word / Excel / Powerpoint
Adobe Illustrator

Skills

Multidisciplinary background in both computational and experimental biology

Computational skills:
Bioinformatics (RNA-seq), data analysis (R), image processing (FIJI)

Technical skills:
Molecular biology (RNA and DNA processing), single-molecule RNA-FISH, antibody-based detection, *microscopy* (confocal, brightfield), *histology*, *animal experimentation* (mouse, Module 2 RESAL), Mammalian cell culture

Scientific experiences

2016 - 2021

Ph.D Candidate - Laboratory of Computational Systems Biology at EPFL, Lausanne

- *Circadian dynamics of RNA localisation in the mouse liver*
- Interdisciplinary approach combining next-gen sequencing and imaging (smFISH) to model sublobular and subcellular RNA distribution in mouse liver along a full daily cycle
- Production and analysis of large biological datasets
- Study director for animal experimentation, management of licenses and transgenic animals in the lab
- Management of purchases of lab reagents and equipments

2015 - 2016
(6 months)

Research Assistant - Laboratory of Stem Cells Dynamics, EPFL

- Study of the biomechanical mechanisms of the corneal epithelial regeneration in mouse model

2014 - 2015
(1 year)

Master Student - Department of Ophthalmology, Kyoto Prefectural University of Medicine, Japan

- *Lineage tracing of Lrig1-expressing stem/progenitor cells in the murine cornea*
- Handling of pre- and post-surgery care, setting up new protocols for live animal imaging, supervision of breedings

Professional experiences

2014
(2 months)

Administrative assistant - Human Brain Project, Lausanne

- Implementation of a database for patents portfolio
- Presenting the database to Tech Transfer Offices of international collaborating universities

Jul - Sep
(2 months)

Internship - Centre Hospitalier Universitaire vaudois (CHUV) Lausanne

- Generalisation and cleaning of a Matlab code used to evaluate PET-SCAN images

2010 - present

Teaching Assistant at EPFL
General Physics, Mathematical and Computational modeling of System Biology

Scientific publications

ResearchGate *https://www.researchgate.net/profile/Hurni_Clemence*

Ce document a été imprimé au Centre d’impression EPFL,
imprimerie climatiquement neutre, certifiée myClimate.