UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA INDUSTRIALE

CORSO DI LAUREA MAGISTRALE IN INGEGNERIA CHIMICA E DEI PROCESSI INDUSTRIALI

Tesi di Laurea Magistrale in

Ingegneria Chimica e dei Processi Industriali

STUDY OF THE MULTI-LEVEL REGULATORY MECHANISMS IN THE CIRCADIAN RHYTHMS OF HUMAN LIVER CELLS

Relatore: Prof. Nicola Elvassore

Correlatore: Prof. Camilla Luni

Laureanda: FRANCESCA CENCI

ANNO ACCADEMICO 2019 – 2020

Riassunto esteso

I ritmi circadiani (dal latino *circa diem, "*intorno al giorno") sono variazioni cicliche delle attività biologiche con un periodo di 24 ore, la cui funzione è di sincronizzare comportamenti e fisiologia dell'organismo, come alternanza di attività e di riposo o variazioni cicliche di temperatura corporea e pressione sanguigna, ai cambiamenti dell'ambiente esterno dovuti alla rotazione della Terra. L'orologio biologico nei mammiferi ha una struttura gerarchica, in cui si distinguono: l'orologio centrale, costituito dal Nucleo Soprachiasmatico ("SCN") situato nell'ipotalamo, e gli orologi periferici, che sono situati in quasi tutte le cellule dell'organismo e sono coordinati dal primo, pur mantenendo un certo livello di indipendenza (ad oggi non del tutto noto). Inoltre, i ritmi circadiani sono fortemente influenzati da stimoli esterni definiti *Zeitgebers* (dal tedesco "che dà il tempo"): in particolare, il Nucleo Soprachiasmatico è sincronizzato dagli stimoli luminosi percepiti dalla retina dell'occhio, mentre gli orologi periferici (per esempio nel fegato, nel pancreas, nel tessuto muscolare o adiposo) sono sincronizzati in parte dai segnali umorali provenienti dal SCN e in parte da stimoli esterni, come alternanza di alimentazione e digiuno, ciclo sonno-veglia, dieta a restrizione calorica o ricca di lipidi e carboidrati.

Tuttavia, i ritmi circadiani di orologio centrale e periferico potrebbero non essere sincronizzati a causa di comportamenti scorretti dal punto di vista fisiologico, come lavorare, assumere cibo e/o essere esposti a luce artificiale durante la notte. A sua volta, questo fenomeno può portare a condizioni patologiche più o meno gravi: per esempio, il disallineamento tra ciclo sonno-veglia dell'organismo e ciclo giorno-notte esterno, con conseguente insonnia di notte e calo di attenzione di giorno; patologie legate all'alimentazione, come obesità, diabete e livelli alterati di trigliceridi e colesterolo, fino a un aumento di probabilità di sviluppare tumori.

Pertanto, risulta di fondamentale importanza comprendere quali siano gli stimoli esterni effettivamente in grado di alterare i ritmi circadiani e in che misura influiscano negativamente sulla salute dell'uomo. In particolare, sono necessarie misure quantitative e quanto più possibile esaustive sulla risposta dell'orologio biologico a determinati stimoli, pertanto l'analisi di dati di *omica* si è affermata come il principale strumento in questo ambito. Nello specifico, i dati di genomica riguardano la "lettura" degli elementi costitutivi delle molecole di DNA, ma queste sono uguali in tutte le cellule e non sottoposte a variazioni dinamiche indotte da stimoli esterni (escludendo fenomeni eccezionali, come mutazioni indotte da errori nella duplicazione del DNA o da agenti mutageni), pertanto non sono adatte allo studio dei ritmi circadiani, che richiede dati in grado di cogliere l'aspetto dinamico delle funzioni biologiche nell'arco delle 24 ore. Invece, altri due tipi di molecole dipendono strettamente dall'attivazione di geni in risposta a determinati stimoli: gli RNA messaggeri (o mRNA, o trascritti) e le proteine. Gli mRNA sono oggetto di studio della trascrittomica, che è uno dei settori più maturi dell'omica in termini sia sperimentali (cioè, di tecniche di sequenziamento), sia di analisi bioinformatica. Invece, le proteine, oggetto di studio della proteomica, sono importanti da un punto di vista biologico, in quanto sono i veri “attori” che svolgono le funzioni biologiche all’interno delle cellule, ma il loro studio è limitato dall’attuale mancanza di tecnologie *high-thoughput* che siano in grado di misurare ogni singola proteina estratta dalle cellule. Questo aspetto è particolarmente limitante per lo studio dei ritmi circadiani, perché le proteine più abbondanti, quindi più facilmente misurabili, sono coinvolte in funzioni strutturali prive di ciclicità, mentre quelle con ritmo circadiano sono verosimilmente quelle coinvolte in attività regolatorie e in minor quantità, quindi più difficili da misurare.

Nello specifico, questa Tesi si focalizza sullo studio dei ritmi circadiani di uno degli organi maggiormente influenzati dal ciclo giorno-notte, cioè il fegato: in particolare, i dati di trascrittomica e proteomica analizzati, forniti dalla ShanghaiTech University, sono raccolti da cellule derivate da fegato umano (HepG2) coltivate *in vitro*. In questo ambito, è di fondamentale importanza adottare un opportuno protocollo sperimentale per sincronizzare l’orologio biologico delle cellule e per simulare *in vitro* gli Zeitgeber che influiscono sugli organismi *in vivo*. In questo caso, due protocolli sono confrontati: uno standard, detto “DEX”, che somministra Dexamethasone per simulare gli effetti dell’ormone cortisolo, e uno innovativo, detto “PHY”, messo a punto dal Dott. Ross Eric Beaumont presso la ShanghaiTech University, il cui scopo è di riprodurre la variazione di concentrazione di insulina, glucosio e glucagone che avviene fisiologicamente passando da un regime di digiuno a uno di alimentazione.

Un ulteriore aspetto innovativo è la presenza di misure quantitative sia di trascrittomica sia di proteomica, a tempi corrispondenti e per la stessa durata totale: infatti, in entrambi i casi sono state misurate 4 repliche agli istanti 0 h, 4 h, 8 h, 12 h, 16 h e 20 h (sia DEX sia PHY, per un totale di 4 dataset).

In questo contesto, uno degli obiettivi principali della mia Tesi è di caratterizzare gli effetti dei protocolli DEX e PHY sulla dinamica di espressione di mRNA e proteine; infatti, il vantaggio di applicare questi protocolli a cellule coltivate *in vitro* consiste nella possibilità di studiare separatamente gli effetti di stimoli ormonali e metabolici, che sarebbero inscindibili *in vivo*. Questi risultati vengono raggiunti applicando tecniche di Bioinformatica e Data Analytics ai dataset di trascrittomica e proteomica separatamente (forniti già normalizzati e filtrati), considerando tre step principali: visualizzazione preliminare dei dati, caratterizzazione dei profili circadiani e analisi delle funzioni biologiche.

Infine, dati di trascrittomica e proteomica vengono integrati in un’unica analisi con l’obiettivo di comprendere meglio il meccanismo di regolazione dei ritmi circadiani nel fegato umano, che si presume coinvolga diversi livelli di omica. Infatti, il profilo oscillatorio di una proteina è reso possibile non solo da loop di attivazione e repressione della trascrizione di geni in mRNA, ma anche da modifiche post-trascrizionali degli mRNA e da modifiche post-traduzionali delle proteine, le cui conseguenze sono variazioni nella struttura, funzionalità e vita media delle proteine stesse.

Abstract

The major topic of this Thesis is circadian rhythms, meaning cyclic variations of biological functions with a period of 24 h, which allow mammals to anticipate environmental changes due to Earth rotation. The circadian machinery has a hierarchical structure, with a central clock in the brain and peripheral clocks in almost all cells, especially in liver, pancreas, muscle and adipose tissue: in particular, peripheral clocks are synchronized both by hormonal stimuli coming from the central clock and by external cues, like feeding and sleeping time.

In this Thesis, quantitative measurements of transcriptomics and proteomics, provided as already normalized and filtered data, are analysed with Bioinformatics and Data Analytics techniques in order to compare the effects of two synchronization protocols on human liver cells cultured *in vitro*: a standard protocol called "DEX" and an innovative one called "PHY", developed by Dott. R. E. Beaumont at ShanghaiTech University. The former employs Dexamethasone to mimic the effects of cortisol hormone, while the latter mimics physiological variations in insulin, glucose and glucagon concentration occurring when alternating fasting and feeding. The results of this Thesis prove that the innovative protocol PHY is able to activate a percentage of circadian rhythms that is of the same order of magnitude than the DEX one, both at transcriptome and proteome level. Moreover, the bimodal distribution of peak phases displayed in transcripts and proteins suggests a net distinction of liver functions carried out in the active and resting phase, which is confirmed also by the functional analysis of profiles peaking in the two phases. Moreover, from the biological point of view, both metabolic and hormonal stimuli are able to activate the most important liver biological functions encountered in literature, like metabolic processes, including glucose, lipid, and cholesterol/bile acid metabolism, together with metabolism of proteins pathways, e.g. Unfolded Protein Response (UPR) and post-translational protein modifications.

Finally, within the same protocol, the dynamics of transcripts and proteins differ considerably: indeed, a circadian profile for a transcript does not imply a circadian profile for the corresponding protein and vice versa and even when they are both circadian their peaks are usually shifted in time. This result suggests that many levels of regulatory functions, like post-transcriptional modifications of transcripts and post-translational modification of proteins, are involved in the determination of circadian rhythms of omics and, consequently, in the creation of cyclic behaviors and physiological changes in human liver.

Contents

Introduction

Circadian rhythms, from the Latin "circa diem", "about a day", are cyclic variations in behaviors and physiological status occurring in mammals with a period of 24 hours, like the alternation between activity (i.e. awareness, working, feeding) and rest (i.e. sleeping, fasting) or changes in blood pressure and body temperature. The circadian clock has a hierarchical structure, with the main pacemaker in the Suprachiasmatic Nucleus (SCN) of the Hypothalamus and peripheral clocks in many organs, e.g. liver, pancreas, muscle and adipose tissue. In particular, the central clock (i.e. SCN) is independently synchronized by external light captured by the retina of the eye, while peripheral clock are only partially independent: they are synchronized by the SCN through hormonal stimuli, but they are also influenced in an independent manner by external cues, like caloric intake and time of feeding and/or sleeping. However, the fact that SCN determines only partially the circadian rhythms of peripheral clocks means that they can be misaligned; for example, this may happen in humans when working, feeding and/or exposure to artificial light occurs at night. In turn, there is experimental evidence that circadian rhythms misalignment is associated to many diseases of different degrees of severity, from insomnia at night and drop in attention by day, to depression, weight-related diseases (like obesity, increased levels of triglyceride and cholesterol, diabetes) and even an increased probability to develop tumors. Therefore, it is important to ascertain which stimuli are the most influential for the circadian clock and which biological functions they activate in cells. To do so, quantitative measurements of omics data are the most appropriate. Indeed, the basic steps that allow cells to carry out their functions are: identification of specific sequences in the DNA that carry the information of interest, i.e. genes; transcription of genes into messenger RNAs (also called mRNAs or transcripts); translation of mRNAs into proteins, that are the real "actors" inside cells. However, DNA itself is static and does not change because of external stimuli (except for mutagen agents), therefore it is not adequate for studying dynamics; in contrast, mRNAs and proteins are expressed depending on regulatory mechanisms and external stimuli, thus they are taken into account for studying circadian profiles (i.e. sinusoidal profiles with one peak within 24 h).

In this context, the main objective of this Thesis is to study circadian rhythms of human liver cells (HepG2) cultured *in vitro*, by analysing quantitative *omics* data by means of Data Analytics and Bioinformatics techniques. Data are provided by ShanghaiTech University and involve two levels of omics: messenger RNAs (or mRNAs or transcripts) and proteins. Moreover, in order to entrain the circadian clock of the cells of the population, two synchronization protocols are employed: a standard one, "DEX", which provides cells with Dexamethasone for mimicking hormonal stimuli through cortisol, and an innovative one, "PHY", which reproduces physiological changes in insulin, glucose and glucagon when changing from feeding to fasting regime. The advantage of applying these perturbations *in vitro* is the possibility to consider separately the effects of hormonal and metabolic stimuli, which would be unfeasible *in vivo*. Another innovative aspect of this dataset is the presence of two types of omics data, transcriptomics and proteomics, which are collected at corresponding time points, meaning 0 h, 4 h, 8 h, 12 h, 16 h and 20 h (in DEX and PHY protocol, giving 4 datasets in total).

Therefore, Chapter 1 of this Thesis provides an overview of the current knowledge on circadian rhythms in terms of molecular organization and misalignment consequences on human health, focusing the attention on liver clock. Moreover, the main low- and high-throughput measurement technologies for transcripts and proteins are illustrated, highlighting advantages and limitations.

Chapter 2 is made of two main parts: the former describes the experimental setup used to generate the data of this Thesis, while the latter explains the Data Analytics and Bioinformatics techniques applied to analyse their dynamics and biological functions. These analyses are performed by means of algorithms implemented as MATLAB or R scripts, whose essential parts are shown in Appendix A and B.

The results of transcriptomics and proteomics analysis are shown and discussed in Chapter 3 and 4, respectively, making a comparison between DEX and PHY protocol performances. In both cases, the starting data were provided already normalized and filtered, while the analysis performed for this Thesis is made of three major steps: first, a preliminary analysis aiming at assessing the level of inter- and intra-variability between DEX and PHY protocols; then, different dynamics analysis for characterizing circadian profiles; finally, statistical tests for identifying the most important biological functions carried out by transcripts and proteins. In particular, it is interesting to distinguish between biological functions carried out in the active phase (day for humans) and in the resting phase (night for humans), which proves that there is a change in liver functionality throughout 24 hours. Moreover, also non-circadian profiles are analysed in case of transcriptomics data, more accurate than proteomics one, through statistical tests that assess whether some genes are constantly up-regulated in one protocol: indeed, different important functions may be constantly activated within cells by the two protocols.

While Chapter 3 and 4 focus on the comparison between DEX and PHY protocols for the same type of omic data, Chapter 5 compares profiles of transcriptomics and proteomics within the same protocol. These results have a limitation, because the two types of data are retrieved with different measurement techniques and are normalized independently from one another, thus they are not directly comparable. Nevertheless, this comparison allows to understand the degree of diversity between transcription and translation profiles, that is a fundamental step for improving the description of the multi-level regulatory mechanisms of human liver cells. Indeed, sinusoidal profiles of mRNAs can be obtained by activating the transcription of genes at specific hours of the day and repressing them for the remaining time. However, other levels of regulations are expected to intervene, like modifications of mRNA and proteins, that eventually modify the temporal profiles of proteins.

Finally, this Thesis concludes with an overall interpretation of the results and with some hints for future applications of the knowledge on liver circadian clock.

Chapter 1
Circadian Rhythms

This chapter provides an overview of the relevance of the circadian clock, how it is studied and what are the open questions, especially related to liver physiology and pathophysiology. First, the implications of circadian rhythms disruption on human health are provided, together with a simplified description of the molecular network that allows to generate oscillatory gene expression over the day. Then follows a presentation of the experimental strategies for collecting transcriptomics and proteomics datasets, by describing the most common low- and high-throughput technologies. Afterwards, the attention focuses on the interplay between metabolism and one of the most important peripheral clocks: the liver circadian clock. Finally, the open challenges in this field are explained, together with the aim of this Thesis.

1.1 Impact of circadian clock on human health

Time is a fundamental aspect of many cell, tissue and organ functions. In particular, the so-called 'biological clock' is an intrinsic timing system that coordinates cell behaviour by anticipating recurring environmental changes such as light-dark cycles, food availability and oscillations in temperature. The physiological functions that have a period of 24 hours are called *circadian* from the Latin *circa diem* ("about a day").

The spread of transmeridian air travels and untraditional working schedule (i.e. outside the interval 9 am – 5 pm), as well as unhealthy habits like nocturnal food intake, can cause a misalignment of the internal clock with respect to the external environmental rhythms. This, in turn, is demonstrated to cause physiological and psychological complications, even if the specific contribution given by circadian disruption is difficult to assess because it is potentially associated with other factors, like stress, unhealthy diet, tobacco and alcohol consumption. However, stricter control on single or combined factors can be achieved thanks to experiments on animal models, for example making well-designed changes on the environment (like duration of light and dark phases), on food availability (like a normal chow diet or *ad libitum* diet), and on the genome (like inducing mutations that affect the molecular oscillator). However, when animal models are employed one must always remember that extending the results to human physiology is not straightforward and it has many limitations. For example, mice models are widely-used, but rodents are nocturnal animals so circadian clock activities in the active phase and in the resting phase should be reversed when referring to human physiology. Then some implications of circadian disruption on human health are described, as a summary of many experimental evidence attained by employing human and/or animal models (Evans and Davidson, 2013).

An example of negative effects due to circadian disruption is the so-called *non-24 h sleep:wake disorder*, implying that people have an internal sleep/alertness cycle that is not aligned with the darkness/light external rhythm. This may lead to day-time sleepiness and night-time alertness and possibly to cognitive deficits.

Another important issue is to determine the relationship between circadian misalignment and aging and lifespan. This is quite difficult to be investigated for mammals (especially humans) due to their relatively long lifespan, however it has been proved that shift workers have an increased mortality risk. Moreover, experiments on mice proved that weekly light-darkness (LD) cycle inversions and/or weekly 6 h shifts of LD cycle induce a higher mortality. Even genetic models that generate defect on genes of the circadian clock decrease survival and accelerate the onset of age-related pathologies in mice.

Moreover, cancer and oxidative stress are serious complications that can be caused by dysfunction of the circadian machinery. For example, sighted woman exposed to artificial light at night have a higher rate of breast cancer, which has not been found for blind women. Also shift work and/or frequent air travel is associated with higher risks of various forms of cancer, including endometrial, colorectal, lymphatic, prostate and breast cancer.

As a consequence of the strict interplay between circadian clock and metabolism, circadian dysfunctions can cause many pathologies involving metabolism itself: for example, diabetes, increased weight gain and obesity, altered insulin resistance and altered levels of free fatty acid, cholesterol and triglyceride. Moreover, shift workers typically display a higher rate of gastrointestinal disorders, like ulcers and irritable bowel syndrome.

The list of disorders induced by circadian disruption is long: it includes cardiovascular diseases, like stroke, atherosclerosis and hypertension; reproduction issues, like irregular menstrual cycles, preterm births and spontaneous abortion; mood disorders, like depression, together with learning and memory deficits. Interestingly, many of the abovementioned diseases (e.g. cancer, diabetes and cardiovascular diseases) share a common factor: an uncontrolled or chronic inflammation status. Indeed, the immune system is regulated by the circadian clock and it has been proved that day workers display a higher risk for common infections, multiple sclerosis and other autoimmune disorders compared to day workers. Therefore, it was suggested that changes in the immune response may be a common factor influencing various diseases in shift workers as a consequence of circadian misalignment.

To conclude, the importance of the correct circadian clock function for human health justifies the efforts of many researchers attempting to characterize better its multi-level regulatory mechanism and to identify the most influential factors that can be efficiently manipulated in order to re-synchronize the circadian machinery efficiently.

1.2 Hierarchy in the circadian clock regulation

The circadian clock machinery is regulated at cellular level by the cyclic expression of clock genes which codify some key proteins involved in positive and negative transcriptional and translational feedback loops (Zanquetta et al., 2010).

The master pacemaker is located in the Suprachiasmatic Nucleus (SCN) in the brain, but peripheral oscillators are present in almost all cells, especially in liver, lung, pancreas, adipose tissue, gastrointestinal tract, skeletal muscle and fibroblasts. In particular, the Suprachiasmatic Nucleus contains 15 000 – 20 000 neurons and is entrained by light cues that are conveyed to the brain by the retino-hypothalamic tract. Afterwards, oscillations induced by the SCN orchestrate circadian rhythms in peripheral tissues thanks to humoral and nerve outputs. Besides stimuli coming from the central clock, peripheral clocks can be entrained by external cues in an independent way. However, this can lead to a misalignment between central and peripheral clocks (Figure 1.1), which is likely to be responsible for the onset of metabolic dysfunctions. This occurs, for example, if one sleeps during the day or if light exposure and/or food intake occur at night.

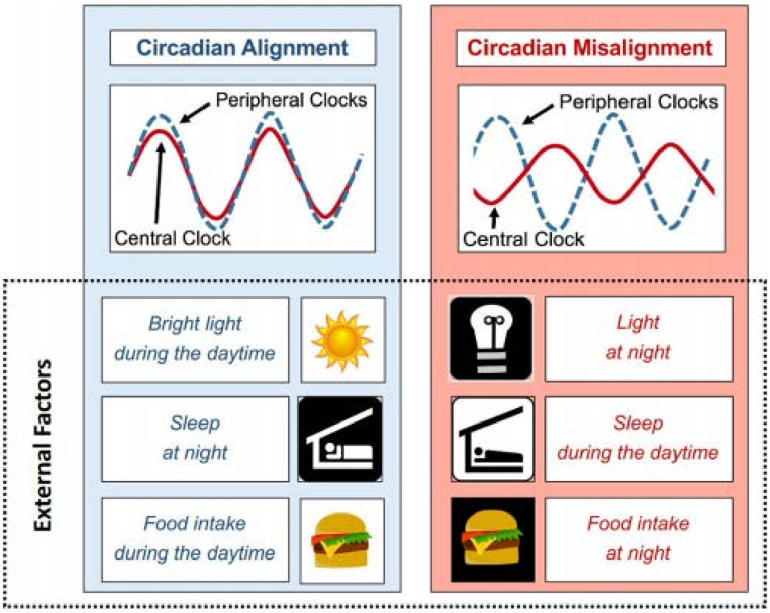


Figure 1.1 *Circadian Alignment vs Misalignment. On the left, the major sources of Circadian Alignment are explained: being exposed to light in the daytime, sleeping at night, eating during the active phase. On the right, the opposite behaviors are indicated as the main causes of Circadian Misalignment, meaning that Central Clock rhythms and Peripheral Clocks rhythms do not overlap anymore (taken from Poggiogalle et al., 2018).*

In particular, 9 key proteins make up the core of the circadian machinery:

- CLOCK
- BMAL1 (ARNTL)
- PER1
- PER2
- PER3
- CRY1
- CRY2
- REVERBα (NR1D1)
- RORα (RORA)

Their functions are discussed in the following section, by contextualizing them in the corresponding feedback loops.

1.2.1 The molecular organization of the circadian clock

The circadian clock in mammals is made up of many interconnected transcriptional and translational feedback loops. In general, a feedback loop is a set of events providing a certain output in response to a stimulus and the output itself can intensify the system (positive feedback loop) or inhibit the system (negative feedback loop). Specifically, in the circadian clock the main transcriptional activators are CLOCK and BMAL1, while the main transcriptional repressors are PERs and CRYs (Figure 1.2).

The transcriptional activators CLOCK and BMAL1 heterodimerize into CLOCK: BMAL1 complex and translocate into the nucleus, where they bind E-box sites (i.e. specific DNA binding regions), activating the expression of more than 300 clock-controlled genes (CCGs), including *per* and *cry*. Afterwards, PER and CRY proteins dimerize and translocate into the nucleus to repress CLOCK: BMAL1 transcriptional activity. Meanwhile, a variety of post-translation protein modifications of PERs and CRYs leads to their degradation and, consequently, to the reactivation of CCGs expression thanks to the binding of CLOCK: BMAL1 to E-box sites, thus generating diurnal cycles in circadian gene expression.

Other regulatory loops involve ROR and REV-ERB proteins that activate and repress transcription of *Bmal1* gene, respectively. In particular, REV-ERB α/β increase in level during the day and bind specific responsive promoter elements (RRE) to inhibit *Bmal1* transcription. In contrast, at night REV-ERB α levels are low enough to allow *Bmal1* transcription.

Notably, transcriptome studies have revealed that the number of common CCGs in different tissues is marginal, thus suggesting the contribution of tissue-specific factors to clock control.

Finally, it reveals that a large fraction of the genome (up to 50%) is potentially under clock control.

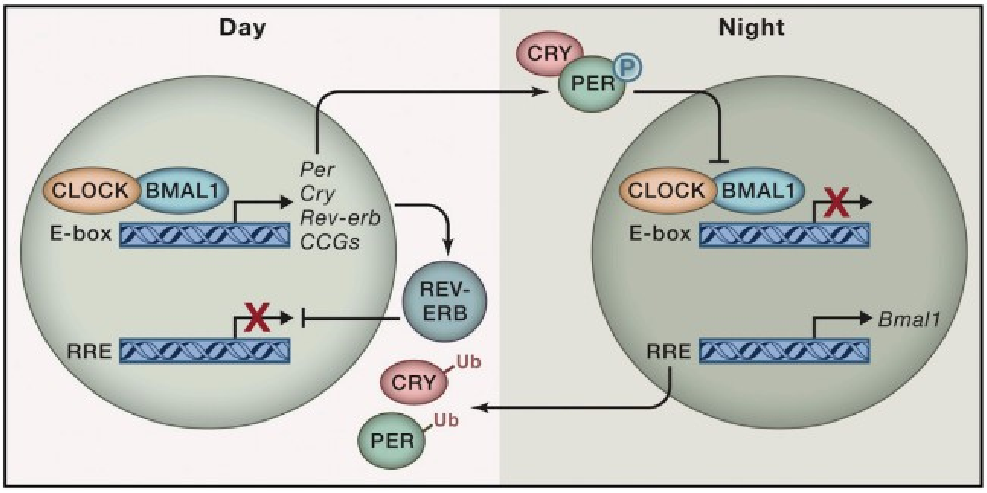


Figure 1.2 *Molecular organization of the circadian clock. The heterodimer CLOCK:BMAL1 binds to E-box sites and activates the transcription of circadian genes, among which Per and Cry. When PER and CRY proteins reach a peak in their concentration, they dimerize and repress CLOCK:BMAL1 transcription. REV-ERB α/β are involved in another transcription loop and they inhibit the transcription of Bmal1 gene. The alternation between activation and repression of clock-controlled genes generates circadian rhythms (from Asher et al., 2015).*

1.2.2 The specific case of the liver circadian clock

Even though it is commonly accepted that the Suprachiasmatic Nucleus is hierarchically dominant in the clock machinery, the degree of independency of tissue-specific clocks as well as their level of communication are not fully understood.

Koronowski et al., 2019, performed experiments that help clarify these aspects with respect to liver. Indeed, they compared transcriptomic and metabolomic profiles of *Wild type* mice ('WT'), i.e. mice as they are in nature, *Bmal1* null mice (or *knock out*, 'KO') and *Bmal1* null mice where BMAL1 expression was reconstructed exclusively in the liver ('Liver-RE').

The circadian behavior of BMAL1 was restored in Liver RE mice: it was subjected to periodic post-translational modifications, it was correctly recruited to chromatin at target genes and it activated circadian gene expression. As a further proof, Principal Component Analysis showed that Liver-RE mice tended to cluster closer to WT than to KO in both transcriptomics and metabolomics datasets. However, the recreation of BMAL1 circadian activity was just partial: indeed, only 10% of transcripts and 20% of metabolites had restored oscillations. Moreover, Liver-ER mice that were never exposed to light during the experiment were not able to restore any oscillations in transcriptome or metabolome, thus revealing the importance of light cues for the entrainment of liver circadian rhythms.

These results confirm that the liver circadian clock has a certain degree of independency, but at the same time it needs to receive stimuli from other clocks in order to ensure its fully correct functionality.

1.3 Central dogma of biology

The central dogma of molecular biology was formulated by Francis Crick in 1957 and it depicts the transfer of information from DNA to proteins. Two main biological processes are involved: transcription and translation. The former one refers to the synthesis of messenger RNA (mRNA) from genes, the latter one refers to the synthesis of proteins starting from mRNA.

However, biological systems are characterized by high variability and complexity, so the outcome of gene expression is not that straightforward. Indeed, the abundance of a protein at a specific moment can be interpreted as the result of a mass balance where the production term is given by mRNA translation and the consumption term is given by post-translational protein modifications. Both transcription and translation are subjected to variations: for example, gene expression can be altered because of methylation, that is the addition of methyl groups to DNA, and proteins can be edited through phosphorylation, acetylation, ubiquitination and SUMOylation.

In particular, expression of circadian genes is characterized by multiple layers of control, including temporal, transcriptional, post-transcriptional and post-translational modifications. Nevertheless, the detailed description of each phenomena and of their interconnections is far from complete and it is even less mature at proteomic level with respect to the transcriptomic one.

1.4 Experimental strategies to study the circadian clock

After perturbing the circadian clock *in vivo* or *in vitro* in a liver-specific manner, it is necessary to measure quantitatively the effects in order to understand how the clock reacts to the surrounding cues. The readout to measure the perturbation effects cannot be the genome itself (i.e. DNA), because it is static and independent from cells external stimuli, therefore two others molecule types are considered: mRNAs and proteins, that allow to determine which genes are activated and which biological functions are finally carried out within cells at a certain time point. They require different measurement techniques, which can be low- or high- throughput, as summed up in Table 1.1 and explained in section 1.4.1 and 1.4.2.

Table 1.1 *Summary of the low- and high- throughput measurement techniques in transcriptomics and proteomics filed that are described in section 1.4.1 and 1.4.2. In particular, RNA-Seq and LC-MS/MS are used to collect the data analysed in this Thesis.*

Field	Molecules	Low-throughput	High-throughput
Transcriptomics	Messenger RNAs (mRNAs)	Quantitative Polymerase Chain Reaction (qPCR)	RNA-Sequencing (RNA-Seq)
Proteomics	Proteins	Western Blot	Liquid Chromatography with Tandem Mass Spectrometry (LC-MS/MS)

1.4.1 Experimental techniques at RNA level

Historically, transcriptomics has been developed before proteomics and it is more mature both in terms of measurement techniques and of statistical methods to analyse it.

In particular, low-throughput technologies to measure transcriptomic data are *Reverse Transcriptase Polymerase Chain Reaction (RT-PCR)* and *Quantitative Polymerase Chain Reaction (qPCR).* In general, a *Polymerase Chain Reaction* (*PCR*) is a technique that allows to amplify DNA fragments (thus, double-stranded molecules) whose ends are known and it is performed in order to have enough signal during DNA measurements. Basically, the two DNA strands are separated by increasing temperature for few seconds (i.e. through *denaturation*), then temperature is lowered in order to allow specific probes, called *primers*, to identify the DNA sequence of interest and to start the enzymes activity (Figure 1.3): enzymes called *DNA polymerase* use nucleotides as building blocks for synthetizing new DNA strands that are complementary to the template ones. This cycle of amplification is repeated until a sufficient amount of duplicated DNA is obtained.

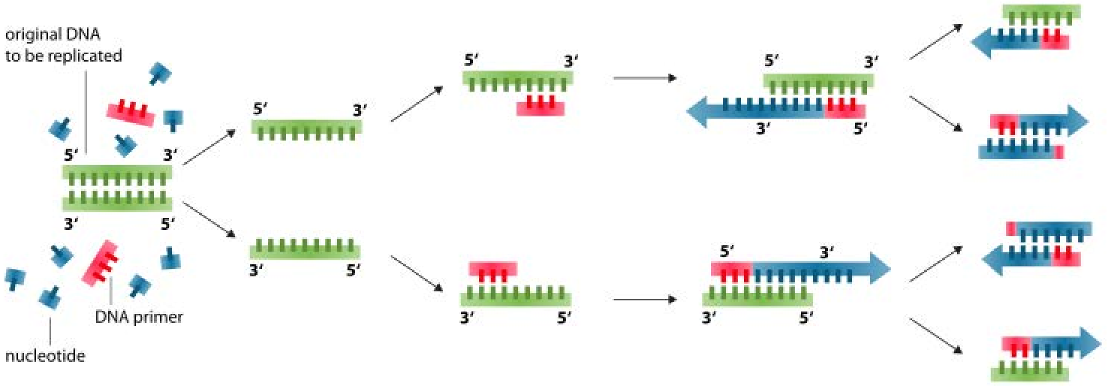


Figure 1.3 ***Polymerase Chain Reaction (PCR).*** *DNA is denatured by reaching a temperature of 94-98 °C for few seconds, then temperature is lowered in order to allow the binding between (known) DNA ends and the so-called 'primers'. These small molecules are required for the enzymes (i.e. DNA polymerases) to start the synthesis of new DNA molecules. In the figure, green DNA strands represent the original templates, while blue strands represent newly synthetized molecules (taken from upbiotech.wordpress.com).*

Also in transcriptomic field amplification is needed in order to have enough signal; however, the starting material is RNA, thus a single-stranded molecule. Therefore, PCR must be preceded by *Reverse Transcription* (RT): this is a reaction that involves a specific enzyme, i.e. *Reverse Transcriptase (RTase)*, for synthetizing a complementary strand to the template one, obtaining the so-called *complementary DNA* (*cDNA*). Finally, this double-stranded molecule is appropriate for being amplified by PCR, following the abovementioned procedure (Figure 1.4 a).

While RT PCR allows to study genes transcription qualitatively, *quantitative PCR (qPCR)* provides a quantitative measurement: indeed, fluorescent tags are added during DNA amplification in order to ensure sequences detectability. Therefore, qPCR output is an amplification curve with number of PCR cycles in the abscissa and fluorescence measurement in the ordinate: fluorescence is the quantity actually measured and it represents the cDNA abundance in the sample. This curve has typically a sigmoid trend: the first part overlaps with the baseline, because cDNA level is so low that it is confused with background noise; then, starting from a critical number of cycles (C_t), cDNA abundance is detected and increases exponentially, up to a point where it reaches a plateau (Figure 1.4 b).

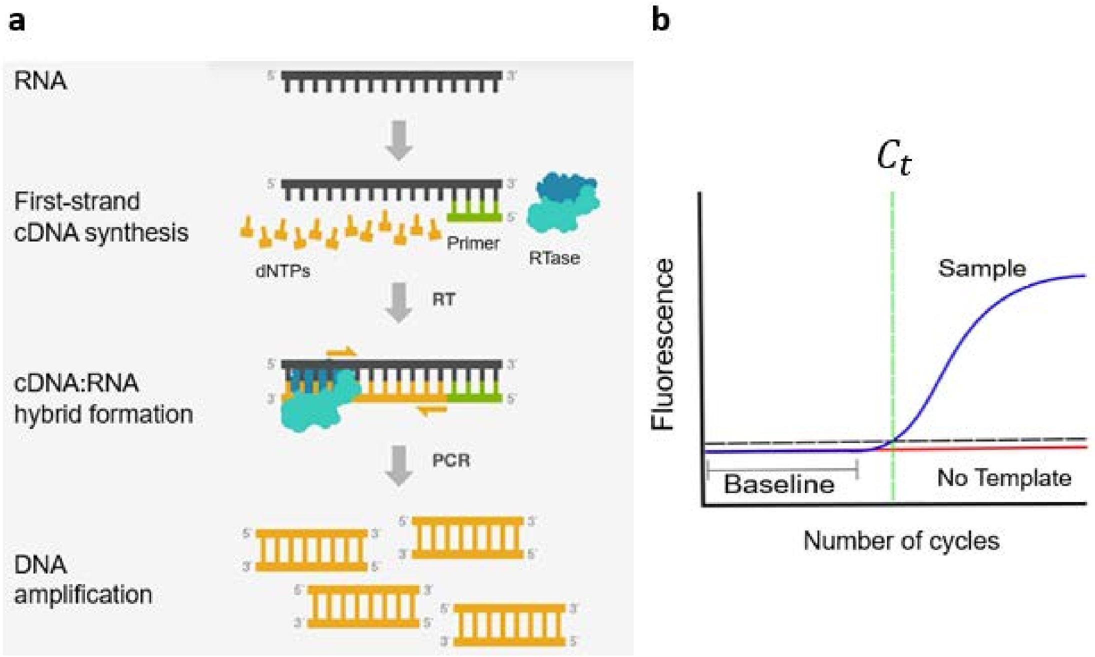


Figure 1.4 *These figures are related to two low-throughput techniques for analysing RNA:*

(a) ***RT-PCR.*** *When the starting material is (single-stranded) RNA instead of (double-stranded) DNA, the amplification by means of PCR must be preceded by a Reverse Transcription (RT), that allows to obtain the so-called 'complementary DNA' (cDNA). In this reaction, a 'probe' called 'primer' recognizes and binds specifically one of the RNA ends and it allows the enzyme (i.e. RTase) to start the synthesis of the complementary strand, by using dNTPs as building blocks. After RT, cDNA can be amplified though PCR (taken from www.thermofisher.com).*

(b) ***qPCR****: this plot shows a scheme of the amplification curves provided by qPCR. The first part of the curve overlaps with the baseline, because the RNA level is too low to be distinguished from noise. However, after a critical number of cycles (C_t) the fluorescence signal (that is related to the RNA abundance at each PCR cycle) is enough to be detected and the exponential part of the curve starts, until a final plateau (taken from bitesizebio.com).*

However, for profiling all transcripts, high-throughput technologies are preferred: they are the so-called 'Next Generation Sequencing' technologies (NGS, or *deep sequencing*). In particular, RNA-sequencing (or *RNA-Seq*) is one of the most common methods for generating transcriptomics datasets (Figure 1.5). A key step that makes RNA-Seq different from qPCR is the fragmentation of all transcripts into segments of desired (and controlled) lengths. At this point, fragments are reverse transcribed to cDNA, bound to specific oligonucleotides adapters (required for sequencing), amplified though PCR and finally detected by means of RNA-Seq platforms (e.g. Illumina). Finally, the obtained sequenced fragments, called *reads*, can be mapped to a reference genome or *de novo* assembled in order to 'reconstruct' the original transcripts (Kukurba and Montgomery, 2015), (Head et al., 2014). The final result is a matrix with thousands of genes in the rows and the corresponding mapped reads (called *counts*) organized in different columns, representing, for example, different subjects, treatments and/or time points.

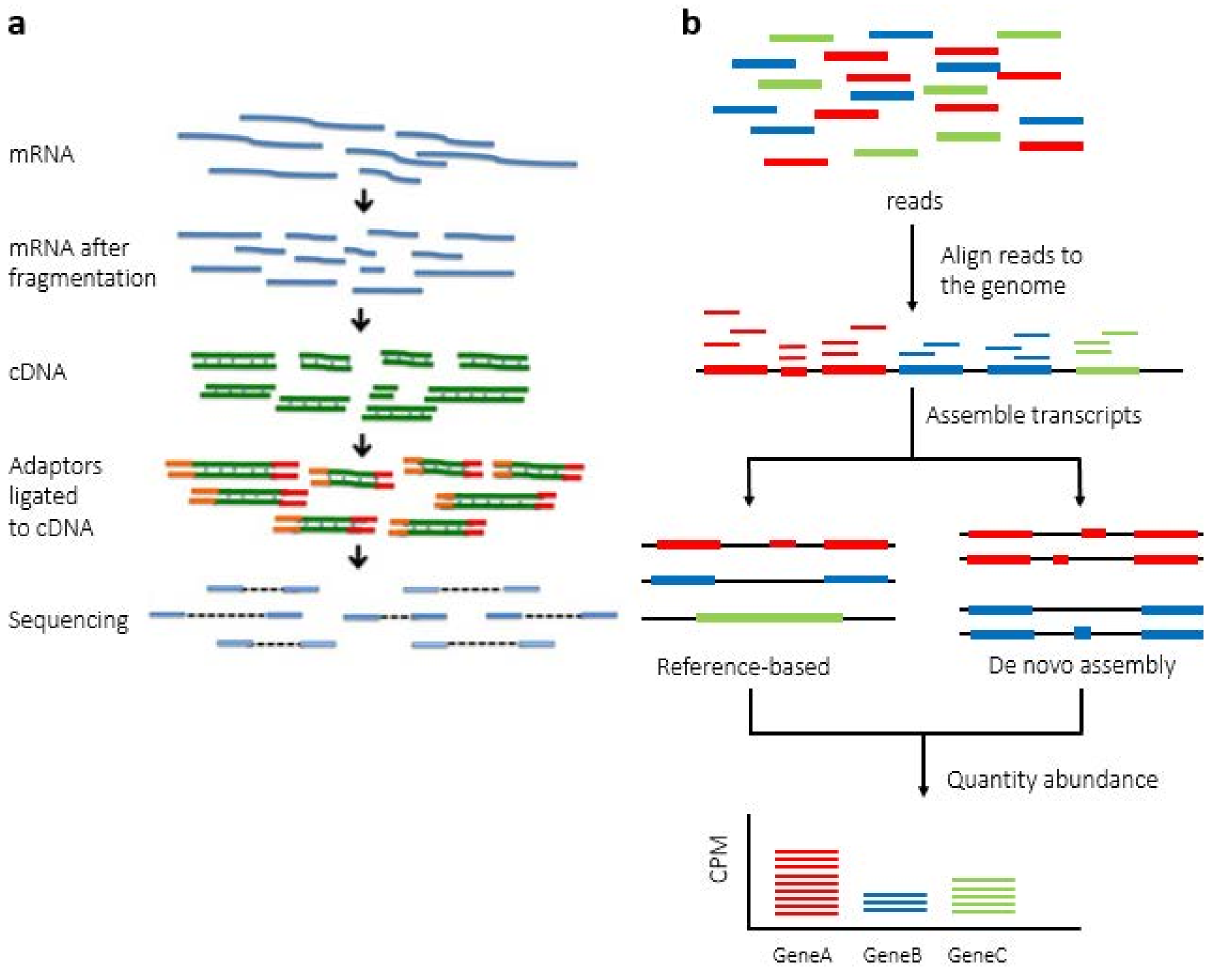


Figure 1.5 ***RNA-Seq. a)*** *Major steps from RNA extraction to the final measurement: mRNA (isolated from total RNA) is fragmented and converted to cDNA; then, the cDNA segments are bound to adapters and sequenced with a sequencing platform (e.g. Illumina), (adapted from the slides made available online by Steve Munger of The Jackson Laboratory, 29 September 2014).* ***b)*** *Main steps when reads obtained as output from RNA-Seq are available: they are aligned to a reference genome and combined to reconstruct the original transcripts, which can be done with or without a reference. This allows to quantify the abundance of the transcripts generated from different genes (in figure, genes A, B and C), which can be scaled in the form of Counts Per Million (CPM, defined in Chapter 2), (adapted from Kukurba and Montgomery, 2015).*

1.4.2 Experimental techniques at protein level

Western Blot is a low throughput technology for the identification and quantitation of specific proteins inside a multi-protein mixture, which is feasible thanks to the use of antibodies: indeed, they are Y-shaped proteins that are able to recognize and bind only specific proteins, called *antigens*, through a *key-to-lock* mechanism, without interacting with the rest of the protein mixture (Figure 1.6).

First, proteins extracted from cells are separated by gel electrophoresis, based on their different molecular weights. Then, they are transferred on a membrane, usually by applying an electric field orthogonal to the plane of the gel and of the membrane. However, this procedure does not saturate all the binding sites on the membrane, which can interact with the antibodies; to avoid that, the membrane is saturated thanks to the exposure to a protein solution (e.g. Bovine Serum Albumin). Then, the ‘primary’ antibody is added in order to identify and bind the proteins of interest. Finally, a ‘secondary’ antibody interacts specifically with the former one and carries an enzyme that reacts with appropriate substrates emitting colorimetric or chemiluminescent signals: therefore, image analysis on the membrane allows to quantify proteins abundance.

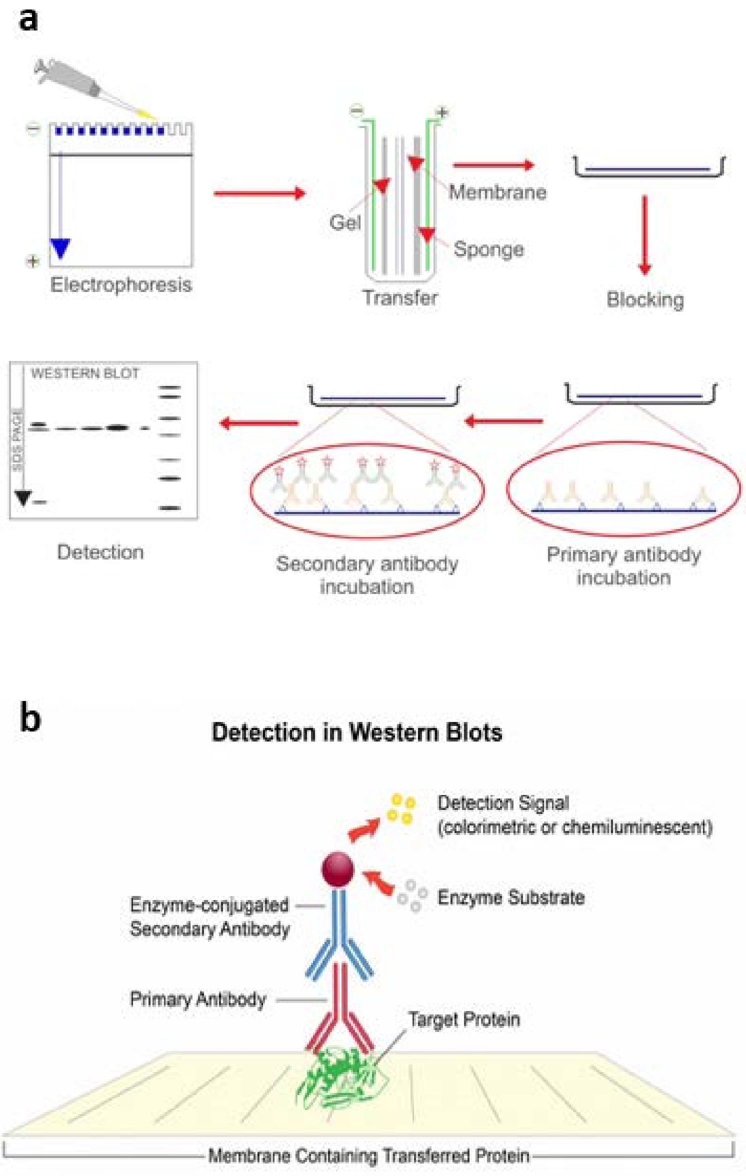

a
Electrophoresis
Gel
Membrane
Sponge
Transfer
Blocking
WESTERN BLOT
SDS PAGE
Detection
Secondary antibody incubation
Primary antibody incubation
b
Detection in Western Blots
Detection Signal (colorimetric or chemiluminescent)
Enzyme Substrate
Enzyme-conjugated Secondary Antibody
Primary Antibody
Target Protein
Membrane Containing Transferred Protein

Figure 1.6 ***Western blot****.* ***a)*** *Scheme of the main steps of the measurement: proteins are separated though gel electrophoresis and transferred on a membrane. However, not all membrane binding sites are saturated and they can bind the antibodies; to avoid that, the so-called ‘blocking’ is performed, meaning the saturation of free binding sites with appropriate protein solutions. Then, primary and secondary antibodies are added: the former binds specifically the protein(s) of interest, the latter binds the primary antibody and carries an enzyme that take part in colorimetric or chemiluminescent reactions: therefore, proteins identification and quantification is made by means of image analysis (taken from www.cusabio.com).* ***b)*** *A detail of the procedure, meaning the binding between proteins (fixed on the membrane) and primary and secondary antibodies, with the latter one carrying an enzyme that reacts with substrates emitting a colorimetric or chemiluminescent signal (taken from the website www.leinco.com).*

The main drawbacks of Western Blot are high costs and complicated technical execution: indeed, substantial expenses are required for labelled antibodies, for detection and imaging systems as well as for the employment of skilled analysts. Moreover, each step of the procedure must be carried out with high precision, otherwise the entire proteins measurement may fail. Moreover, this is a low-throughput technique, thus it usually takes more than four hours for detecting proteins in the order of tens. However, it has important advantages: it has high sensitivity, being able to detect even 0.1 nanograms of a specific protein within a sample, and high specificity, thanks to both the preliminary separation through electrophoresis and the highly specific interaction between antibodies and proteins.

However, the need to measure a high number of proteins in a time-efficient way leads to the development of high-throughput technologies also for proteomics. The most widely used method to measure directly protein abundance is Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS) (Angel et al., 2012), where 'bottom-up' Mass Spectroscopy is combined with Liquid Chromatography.

First, proteins extracted from cells are fragmented into peptides, through chemical reactions or using enzymes (Figure 1.7 a). This peptide mixture is then separated thanks to Liquid Chromatography: based on their chemical-physical properties, groups of peptides move through the chromatographic column with different retention times. Eluted peptides are detected by a first Mass Spectrometer, then fragmented further in order to make a second Mass Spectrometry analysis: therefore, the first MS identifies and quantifies peptides based on their molecular weight and charge, while the second MS identifies and quantifies peptides based on the specific fragments they generate (which depend on peptides molecular structure). In particular, in order to facilitate the quantitation and differential comparison of proteins, either at different times or different conditions, isotope-labelling methods are employed, introducing tags that are discriminated by the MS. For example, the experimental setup of this Thesis uses *Tandem Mass Tag* (or *TMT*), which involves isobaric tags (i.e. same mass and structure) to chemically label already isolated peptides (Figure 1.7 b). The advantage of using tags of the same mass consists in the possibility to analyse multiple samples within the same experiment, thus decreasing time and costs. Indeed, different samples employ variants of the tag that are distinguishable only by the heavy isotope distribution in their structure, while the overall mass is constant, meaning that all peptides and fragments are equally affected in terms of retention time in LC and of mass-to-charge ratio in MS. Finally, identified peptides are mapped to the corresponding proteins; indeed, even though MS actually measure peptides, results are interpreted in terms of proteins.

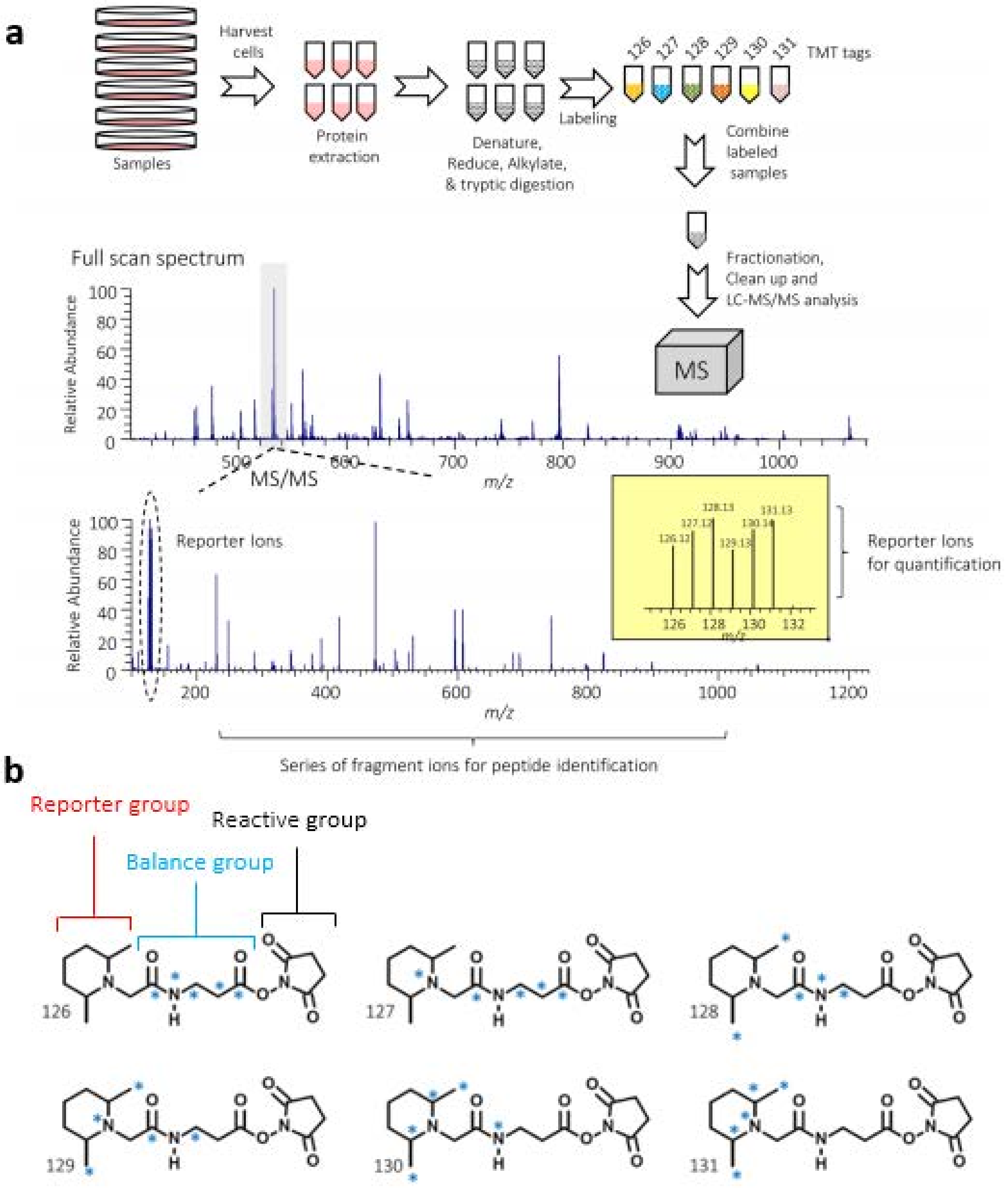


Figure 1.7 ***High-throughput measurement technique for proteomics: LC-MS/MS. a)*** *Scheme of a LC-MS/MS experiment: proteins are extracted from cells and fragmented into peptides chemically or enzymatically. Then, peptides are labelled with isobaric tags and pooled together: at this point, the labelled peptide mixture can be separated by LC and analysed by means of two MS: the first MS identifies peptides peaks, the second MS identifies the peptides composition and relative abundance by detecting peaks of peptide fragments and of reporter ions originated from TMT.* ***b)*** *Examples of isobaric tags used in a six-plex experiment. The three main components of tags are highlighted: a reporter group, which allows to determine the relative abundance of each peptide in the second Mass Spectrometry due to the presence of different number and/or positions of heavy isotopes (indicated by blue stars); a balance region, whose aim is to make the overall mass equal among all tags thanks to the presence of heavy isotopes; a reactive group which binds the peptides that are analysed (adapted from Rauniyar and Yates, 2014).*

1.5 Interplay between liver circadian clock and metabolism

Since datasets analysed in this Thesis are samples of human liver cells cultured *in vitro*, the attention now focuses on what is known about the liver circadian clock.

Circadian rhythms in human liver are entrained by both feeding/ fasting cycles and light cues (through stimuli coming from the Suprachiasmatic Nucleus) and they influence key phenomena involving the liver, such as:

- nutrient homeostasis
- autophagy
- glucose homeostasis
- lipid homeostasis
- bile acid homeostasis
- Endoplasmic Reticulum (ER) stress

1.5.1 Circadian clock and nutrient homeostasis

Within human liver, many metabolic regulators are able to sense feeding/fasting states and to control the expression of hundreds of proteins involved in metabolic processes.

For example, AMPK is a fasting-sensitive protein kinase that regulates energy homeostasis in cells and it influences, and is influenced by, the circadian clock mechanisms. Indeed, AMPK can induce CRY degradation, thus affecting the length of circadian cycling. Moreover, the activation of the circadian RORα activates AMPK, which in turn modulates lipid metabolism and avoids hepatic steatosis, or *Fatty Liver* disease, i.e. the accumulation of an excessive amount of fat in the liver.

In addition, glucagon (i.e. the hormone that induces liver to convert stored glycogen into glucose) is able to activate the expression CREB protein, which in turn up-regulates the transcription of the circadian *Per1* gene.

1.5.2 Circadian clock and Autophagy

Sensing the nutrient status is important for maintaining homeostasis: for example, if starvation is sensed, autophagy is activated in order to use glycogen and lipid components within the cell to generate fuel. However, this is not a rare event: assuming a normal feeding/fasting cycle, the low nutrients level during the resting phase and its high level during the active phase generates a periodic activation of autophagy.

An example of the interaction between proteins belonging to the circadian machinery and proteins activating autophagy is given by SIRT1. Indeed, SIRT1 is a nutrient-sensing protein that can up-regulate the expression of many genes involved in autophagy and energy metabolism thanks to the formation of a complex with CLOCK:BMAL1 that binds to E-box elements. Moreover, that complex creates circadian oscillations in the rate-determining enzyme of the NAD^+ salvage pathway, whose goal is to synthesise NAD, a cofactor involved in many metabolic redox reactions as electron carrier. During regulation of NAD^+, SIRT1 contributes to circadian rhythms by deacetylating BMAL1 and PER2. Moreover, SIRT1 cyclically deacetylates histones leading to a repressive configuration of chromatin, thus causing a down-regulation of the final targets. Therefore, both SIRT1 interaction with BMAL1:CLOCK and its deacetylation of key regulator proteins generate day-night rhythmicity of many genes.

1.5.3 Circadian clock and Glucose Homeostasis

Circadian clock and glucose metabolism are strictly related. For example, insulin (an important metabolic hormone) is a regulator of *Per2* and *Rev-Erbα* and both glucose and insulin can affect the expression of *Dec1* and *Dec2*.

In order to investigate the relationship between circadian clock and glucose metabolism, some gene knockout experiments have been performed, in which the genome of an organism (usually mouse) is engineered to block the activity of one or more circadian genes. In particular, liver-specific *Bmal1*-knockout (KO) mice helped determining the relationship between the circadian clock and GLUT2, the main hepatic glucose transporter. The physiological expression of *Glut2* in mice has a peak in the resting phase and a trough in the active phase, but when *Bmal1* is knocked-out both transcripts and proteins of *Glut2* are constantly low. Moreover, liver-specific deletion of *Bmal1* in mice lead to fasting hypoglycaemia, hypoinsulinemia and loss of circadian rhythms in the expression of genes involved in hepatic glucose metabolism. All these results prove that an appropriate glucose homeostasis cannot be achieved without the contribution of the liver circadian clock.

1.5.4 Circadian clock and Lipid and Bile Acid Homeostasis

The circadian clock interacts with regulatory transcription factors of lipid metabolism (e.g. SREBP1c) and with enzymes of cholesterol and bile acid metabolism. The alteration of their regulatory pathways leads potentially to Non-Alcoholic Fatty Liver Disease (NAFLD).

Indeed, lipid accumulation in the liver of patients affected by NAFLD is due mainly to an altered-regulation of *de novo* lipogenesis, which in turn is strictly related to the activity of circadian proteins. For example, SREBP1c is a transcription factor that activates lipogenic genes and both its transcripts and proteins display circadian rhythmicity. Its activity depends on feeding/ fasting cycles but also on the circadian machinery, since it is regulated by circadian proteins like BMAL1, REV-ERBα, RORα, RORγ and DEC1. The importance of the correct rhythmic functionality of SREBP1c is proved by the fact that its overexpression causes fatty liver, hyperglycemia and hyperinsulinemia.

1.5.5 Circadian clock and Endoplasmic Reticulum Stress

Within hepatocytes, protein folding and lipid synthesis occur in the Endoplasmic Reticulum (ER), which is under circadian control.

Anomalies like changes in the redox state or a high protein demand can lead to the so-called *ER stress*, i.e. the inability of ER to fold proteins correctly. The cellular response to an accumulation of unfolded or misfolded proteins is given by *Unfolded Protein Response* (UPR), that restores cellular homeostasis by interrupting protein synthesis and by increasing the production of specific proteins involved in protein folding. Only in case that the normal ER functionality is not restored within a certain time-lapse, UPR activates the programmed cell death (*apoptosis*).

Finally, since ER stress is more likely at specific times of the day, its regulatory proteins are under circadian control. This has been proved through experiments employing animal models: in mice, a correctly functioning circadian clock controls UPR regulators (i.e. IRE1α), thus determining a biphasic 12 hour periodicity for genes that are regulated by UPR itself; however, if *Cry1/Cry2* genes are knock-out in mice, the hepatic *Ire1α* gene is constitutively expressed. This change of expression profile of UPR regulators induced by the deletion of circadian genes prove that the UPR pathway is under clock control in mice.

1.6 Open challenges

Given the highly interconnected structure of circadian regulation in the cell, the study of circadian clock by more traditional low-throughput techniques is very limiting for its holistic understanding. Experimental high-throughput technologies overcome this limitation by giving a comprehensive picture but come at the price of a lower accuracy and noisier data. Thus, the data analysis needs to be strongly statistically based and robust to catch overall coordinated changes rather than single-molecule precision.

Moreover, the size of the data we are dealing with is about 20000 protein-coding genes (i.e. genes that are actually used for synthetizing proteins) and even more proteins: indeed, mRNAs and proteins are subjected to a large number of modifications, thus allowing to synthetize final molecules with different biological functions starting from the same gene (see Figure 1.8). In addition, the so-called “big data” are usually characterized by collinear data that provide redundant information, which complicates the understanding of the most influential variables and of the main classes of observations in the system under study. This opens the problem of how to visualize and synthetize this information and requires the combination of many competencies, starting from data analytics and bioinformatics, that allow to handle data, and biomedical and chemical engineering, biotechnology and systems biology in order to interpret the biological meaning of the results.

In particular, the main challenge from the data analysis point of view is to develop a robust methodology to identify circadian profiles and to characterize their main parameters, like period, phase, amplitude and mean expression level. Moreover, even the most traditional bioinformatics techniques, like Differential Gene Expression Analysis, need to evolve in order to take advantage of the correlation between time samples that is typical of time-course datasets.

However, characterizing the dynamics of *omics* data is just the starting point: once transcripts or proteins displaying the profiles of interest are identified, they need to be associated to their specific biological processes or molecular location. Moreover, since circadian genes are typically up-regulated at a given time interval and then they are repressed, it is important to combine the analysis of the biological behaviour with the profile characterization.

Finally, the most challenging task is to improve the understanding of the nested loops of transcription, post-transcriptional modifications, translation and post-translational modifications that generate periodic oscillations in central and peripheral clocks. This objective requires to study and integrate a large number of biological phenomena, so it is complicated to design an experimental campaign that is able to cover each aspects of these regulatory mechanisms. Therefore, it is essential not only to integrate different levels of *omics* data (e.g. transcriptomics and proteomics), but also to integrate the current results to those found in literature in order to confirm the results themselves and to complete the description of the neglected phenomena. This will ultimately provide a clear understanding of the gene activity over time, which may be useful, for example, for future development of treatments and drugs with targeted action.

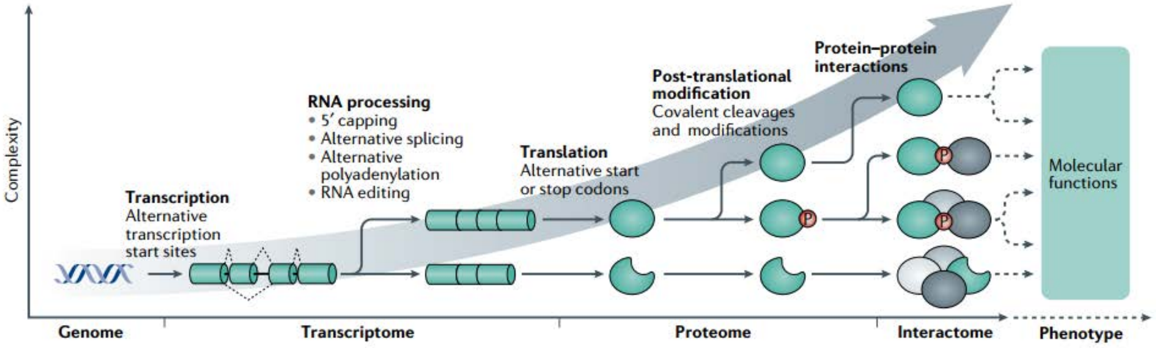

Complexity
Transcription
Alternative transcription start sites
RNA processing
• 5′ capping
• Alternative splicing
• Alternative polyadenylation
• RNA editing
Translation
Alternative start or stop codons
Post-translational modification
Covalent cleavages and modifications
Protein–protein interactions
P
Molecular functions
Genome
Transcriptome
Proteome
Interactome
Phenotype

Figure 1.8 *Omics data analyses allowed to understand that there is not a biunivocal relation between a protein and the corresponding gene, due to the large number of phenomena that are interposed between genotype and phenotype. Indeed, the biological diversity in an organism is primarily related to the number of protein-coding genes in the genome, but the genome itself is stable, therefore the cells ability to change their phenotype as a response to external stimuli is due to differences at RNA and proteome level. In particular, it is estimated that about 56-84% of protein variability within cells is due to variations in messenger-RNA, which in turn can differ because of many phenomena: for example, starting regions of transcription may change (humans have more than 4 starting sites in their genes) and mRNAs may be subjected to post-transcriptional modifications ("RNA processing" in the figure). Moreover, the most common post-translational modification is alternative splicing, which in some cases can lead to the formation of more than 10 differently spliced transcripts from the same gene. Something similar occurs with proteins, too: translation can start from different sites and proteins change due to post-translational modifications, whose number exceeds 400. Finally, the variety of possible biological functions in cells is increased by the possibility for proteins to create multi-molecular complexes through protein-protein interactions (the so-called interactome), whose contribution in determining cell diversity is still not known completely and will be assessed more precisely in the future thanks to the technologies advances (from Bludau et al., 2020).*

1.7 Thesis Aim

The aim of this Thesis is to improve the understanding of different aspects of circadian rhythms in human liver cells, thanks to an integrated analysis of transcriptomic and proteomic data.

First of all, the main advantage of the experimental dataset of this Thesis is the use of *in vitro* cell cultures: indeed, it allows to investigate independently different synchronization stimuli, which would not be feasible *in vivo*. In particular, two synchronization protocols are adopted: an innovative physiological protocol ('PHY'), where cells are synchronized by means of glucose, insulin and glucagon administration, and a standard protocol that uses *Dexamethasone* hormone ('DEX'). The former one mimics the feeding/fasting cycle throughout 24 hours, the latter mimics systemic hormonal control by tissue-tissue in vivo cross-talk. Therefore, characterizing the effects of the two protocols is essential for understanding the level of independence of this peripheral clock (i.e. liver) from the systemic control, thus the ability of the liver to induce oscillatory dynamics in biological functions disregarding the alignment with the SCN.

Another innovative aspect of this dataset is the presence of two types of *omics* data, i.e. transcriptomics and proteomics, both of which are measured by means of high-throughput technologies: RNA-Sequencing (RNA-Seq) and Liquid-Chromatography Mass Spectrometry (LC-MS/MS), respectively. Moreover, in both transcriptomics and proteomics experiments, cells are subjected to the same experimental protocols and biological samples are collected at corresponding time points: 6 time points evenly spaced from 0 h to 20 h, with 4 replicates each. Even though an integrated analysis of the two datasets is complicated by the intrinsic differences between mRNA and proteins and between RNA-Seq and LC-MS/MS, the comparison among their dynamics and biological functions may provide useful insight into the interconnected phenomena at the basis of the circadian clock machinery, like post-transcriptional and post-translational modifications.

In synthesis, the aim is to understand whether the hepatic circadian clock has just a predetermined set of functions that need a correct synchronization in order to work or if synchronization depends on the synchronizer used (e.g. diet or hormones) and how this affects hepatic functions.

Chapter 2
Methods

This chapter summarizes the methods used throughout this Thesis. It is divided in two main parts. The first part gives a description of the experimental setup and methodologies; indeed, experimental data collection was previously performed by Dr. Ross Beaumont at ShanghaiTech University, but it is briefly described here in order to understand the data at the basis of this analysis. The second more extensive part focuses on the computational methods for analysing data. It includes suitable techniques for data normalization (even if normalization and filtering were performed before this Thesis), for preliminary visualization of data variability and for selection and characterization of circadian rhythms. Then follows a description of differential expression analysis, a tool used to identify dynamic differences of circadian and non-circadian profiles, and of enrichment analysis, a method applied to study the biological implications of the identified circadian genes/proteins in human liver cells. Finally, the analysis are carried out through algorithms implemented in MATLAB or R, therefore the basic parts of corresponding codes are shown in Appendix A and B.

2.1 Experimental methods

2.1.1 Cell culture and synchronization protocols

Since the main objective is to study circadian rhythms in human liver, the experiments employ HepG2 cells, i.e. a human hepatic-derived cell line commonly used as an *in vitro* model for human hepatic cells. Cells are cultured at 37°C in a culture media with 5% CO_2 and the species shown in Table 2.1. In particular, *DMEM* stands for *Dulbecco's Minimal Essential Medium*, it is made of salts, amino acids and vitamins with no -, low- or high glucose concentration; *FBS* stands for *Fetal Bovine Serum* and it is a mixture of many important proteins for the cells survival and proliferation; finally, penicillin and streptomycin are used to avoid bacterial contamination.

Table 2.1 ***Cell culture medium.*** *It is changed every 48-72 h and frequently tested for mycoplasma contamination.*

Cell culture medium	Concentration
DMEM	Low glucose (5.55 mM)
FBS	10% volumetric
Penicillin and Streptomycin	100 U/mL

In *in vivo* human cells, circadian rhythms are generated and maintained by the synchronization of cells biological functions during the circadian cycle (about 24 h), which in turn is possible thanks to the entrainment by means of different types of stimuli. Therefore, in order to study how cells are coordinated *in vivo*, the experimental method involves two synchronization protocols *in vitro*: they are called *DEX* and *PHY* protocol. The former is a reference protocol widely used in literature, based on the *Dexamethasone* (Balsalobre et al., 2020), the latter has been developed *ad hoc* in this study and it involves glucose, insulin and glucagon. Finally, the advantage of *in vitro* experiments is that it is possible to decouple different synchronization stimuli, meaning hormones and food intake, which would be unfeasible *in vivo*.

2.1.1.1 DEX protocol

The first synchronization protocol is a standard in the circadian field and it is inspired by the cortisol activity in the circadian clock: indeed, cortisol is a hormone mainly secreted by adrenal glands and it is part of a known *in vivo* systemic synchronization due to tissue-tissue cross-talk. Actually, DEX protocol employs Dexamethasone, which is a synthetic drug that mimics the cortisol biological function, by following an experimental procedure explained in Table 2.2 and Figure 2.1: first, cells are pre-treated for almost 3 days (70 h), then they are synchronized with 200nM of Dexamethasone for 2 h, let in free-running conditions for 24 h and finally sampled every 4h for a total of 24 h.

2.1.1.2 PHY protocol

The PHY protocol in an innovative synchronization method that mimics the change of glucose, insulin and glucagon concentrations *in vivo*, passing from feeding to fasting regimes. Table 2.3 shows all the reagents employed in these experiments, together with the corresponding concentrations in human blood. In particular, blood concentrations are not exactly equal to the ones employed in PHY protocol, but this does not influence the reliability of the results because values that are measured *in vivo* are not directly translatable into an *in vitro* system: for example, blood concentration is not the same as at the surface of liver cells; moreover, cultures of HepG2 cells are not a perfect model of human liver cells, which is also a heterogeneous spatially organized population of cells. However, the most realistic aspect of PHY protocol is the fold change in insulin and glucagon concentrations when the regime passes from feeding to fasting: insulin concentration decreases 10-fold from feeding to fasting regime (Parry et al., 2017), while glucagon increases 4-fold (T.Y. Yue et al., 2012). Moreover, 25 mM of glucose represent the feeding regime, while 2 mM represents the fasting regime: coherently, when the highest concentration of insulin is administered, 25 mM of glucose is added, while when cells are stimulated by the highest concentration of glucagon, 2 mM of glucose is provided. Thus, preliminary experiments have been carried out in order to determine the final concentrations shown in Table 2.3, respecting these fold-changes.

Finally, in PHY protocol cells are pre-treated for 3 days before being stimulated with feeding and fasting cycles for other 3 days. Afterwards, cells are cultured in free-running conditions for 24 h and then collected every 4 hours for 24 h (Table 2.2 and Figure 2.1).

Table 2.2 ***Experimental protocols***. *Sequence and duration of the main steps of the Dexamethasone (DEX) and physiological (PHY) protocols.*

Protocol	Pre-treatment	Treatment	Free-running	Sampling
DEX	70 h	2 h	24 h	24 h
PHY	72 h	72 h	24 h	24 h

Table 2.3 ***Physiological protocol.*** *Glucose, insulin and glucagon concentrations in feeding and fasting conditions.*

Species	Feeding concentration (8 h)	Fasting concentration (16 h)	Human blood concentration
Glucose	25 mM	2 mM	4.4-7.7 mM (Baura et al. 2012)
Insulin	5 nM	0.5 nM	100 – 2000 pM
Glucagon	25 nM	100 nM	< 40 pM

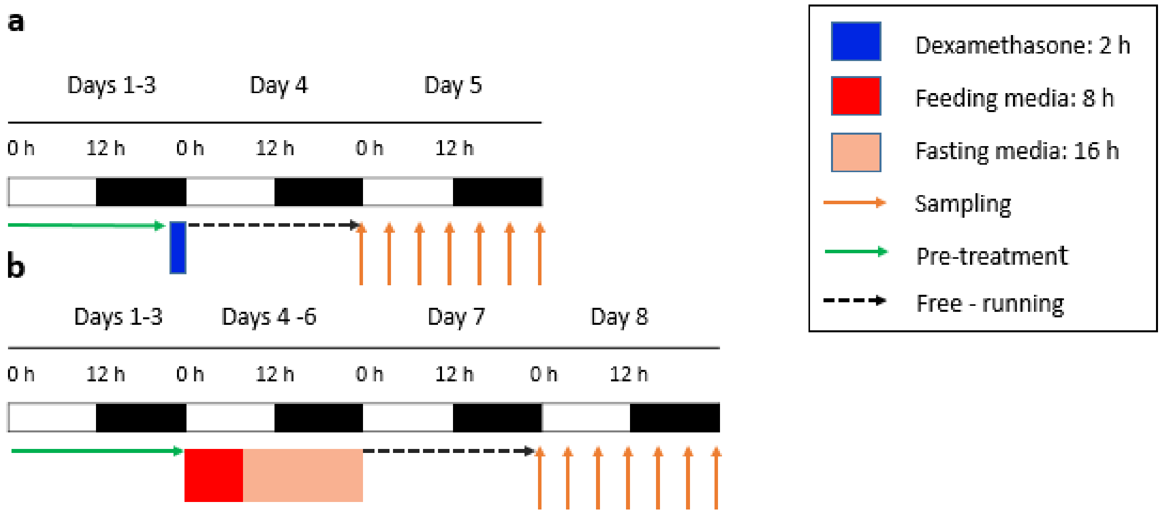


Figure 2.1 ***Final experimental setup.*** *Illustration of the main experimental steps of Dexamethasone and physiological protocols: 3 days of pre-treatment, then treatment, followed by one day of free-running and finally one day of sampling.*

2.1.2 Bulk RNA-sequencing

After 0 h, 4 h, 8 h, 12 h, 16 h and 20 h from the free-running period, cells are washed in warm DPBS-/- and total RNA is isolated thanks to the addition of 0.5 mL of Trizol (ThermoFisher Scientific, 15596026) at room temperature for 5 min. The lysate is then transferred to 1.5 mL RNase-and DNase-free tubes (Quality Scientific Plastics, 509- GRDS-Q). Chloroform (200 µL, Merk, 650498) is added and tubes are shaken vigorously for 15 s and left at room temperature for a further 5 min. After centrifugation at 12000 g for 15 min at 4°C, the aqueous phase is transferred to a fresh 1.5 mL RNase-and DNase-free tube and an equal volume of ethanol (Sigma, 459836) diluted to 70% in UltraPure distilled H_2O (ThermoFisher, 10977015) is added and mixed gently via pipetting. Samples are then loaded onto mini spin columns and the RNA is washed and eluted in accordance with the RNeasy kit protocol (Qiagen, 74106). RNA quantification is measured at 260 nm and purity 11 with the 260/280nm ratio using the NanoDrop 2000 spectrophotometer (ThermoFisher Scientific). Sequencing is conducted by Genewiz using the NovaSeq platform. Each sequencing library provides 6 G data per sample and is constructed from at least 500 ng RNA, with 260/280 ratios > 2.0, using pair-end primes at 150 bp and 8 M clean reads per sample on average.

2.1.3 LC-MS/MS for proteomics

After 0 h, 4 h, 8 h, 12 h, 16 h and 20 from the free-running period, cells are collected in order to extract peptides and to TMT label them. At the indicated times, media is removed from the wells and cells are washed once in icecold DPBS-/- and lysed in 0.5 mL 4% SDS (Sigma, 71736) supplemented with 1x phosphatase (ThermoFisher Scientific, 78420) and protease (ThermoFisher Scientific, 78430) inhibitors. Samples are left on ice for 30 min, homogenised with frequent pipetting, and centrifuged at 10000 g for 10 min at 4°C. The supernatant is transferred to a fresh 1.5 mL tube and protein concertation is determined with the BCA assay kit (ThermoFisher, 23227). Protein lysates for samples (100 ug) and internal standards (100 ug, equal mix of all samples) are reduced in 0.1 M DDT (Sigma, 43816) at 56°C for 60 min. Amicon Ultra centrifugal filter units (Merk, UFC500396) are used for protein purification and concentration. Samples are alkylated in the dark with 55 mM iodoacetamide (Sigma, I1149) for 30 min at room 13 temperature. Samples receive 3x 100 uL TEAB (100 mM) washes (pH 8.0), with subsequent trypsin digestion (Promega, V5111) in a water bath at 37°C for 16-hr. Peptides are subsequently desalted with C18 spin columns (ThermoFisher Scientific, 89873) and dried under vacuum at 4°C (Labconco, 7310039). Samples are reconstituted in 100 µL TEAB (50 mM) and labelled with TMT 10-plex reagents (ThermoFisher Scientific, 90110). Following a 60 min incubation period, 8 µL of 5% hydroxylamine (Sigma, 159417) was added to each sample and left to incubate for 15 min at room temperature to quench the reaction. Samples were then desalted and dried under vacuum and stored at -80°C until analysis.

The next step is LC-MS/MS analysis: samples are prefractionated using a Shimadzu Nexera X2 LC-30AD HPLC system with a uplc, BEH-C18 column (2.1 x 150 mm, 1.7 µm) separating peptides based on gradient elution. Mobile phase A was ACN- H_2O (2:98, v/v, pH 10) and mobile phase B was ACN- H_2O (98:2, v/v, pH 10). The gradient profile is as follows: 0-2 min, 5% B; 2-7 min, 5-8% B; 7-42 min, 8-18% B; 42-64 min, 18-32% B; 64-66 min, 32-90% B; 66-70 min, 90% B; 70-71 min, 90-5% B; 71-75 min, 5% B. The flow rate is 250 µL/min and elutants were collected every two minutes and separated into 15 fractions, dried under vacuum, and stored at -80°C until MS/MS analysis. Fractions were analysed by a Fusion mass spectrometer (ThermoFisher Scientific) equipped with a Nanospray Flex source (ThermoFisher Scientific). Samples are reconstituted in 10 µL of 0.1% formic acid and loaded onto a capillary C18 trap column (2 cm x 100 µm) and separated by a C18 column (50 cm x 75 µm) on an EASY-nLC 1200 system (ThermoFisher Scientific). Mobile phase A is 2% ACN/0.1% formic acid and mobile phase B is 100% ACN/0.1% formic acid. The linear gradient is: 0-90 min, 5-22% B; 90-110 min, 22-40% B; 115 min, gradient increased to 90% B; 115-120 min, held at 90% B. The flow rate is 300 nL/min. Full MS scans are acquired in the mass range of 300 - 1800 m/z with a mass resolution of 120K. The MS/MS are isolated by Quadrupole and tected by Ion trap, with resolution set to 60K. Peak list files were searched against UniProt human reference proteome by MaxQuant v. 1.6. Peptide-spectrum matches (PSMs) were adjusted to 1% and assembled to a final protein-level false discovery rate (FDR) of 1%.

In particular, the high-throughput LC-MS/MS with isobaric labelling (TMT) allows to analyse up to 10 samples in a single experiment. Therefore, 6 sets of 10 samples (called 'TMT samples') are analysed in separated mass spectrometry experiments, as indicated in Table 2.4 and 2.5.

Table 2.4 ***DEX proteomics normalization.*** *Three sets of 10 samples that are analysed in three mass spectrometry experiments thanks to isobaric labelling (TMT). 'Standards' are polled samples with equal amounts of single samples and it is useful for normalizing samples that belong to different TMT samples.*

TMT sample 1			TMT sample 2			TMT sample 3		
Protocol	**Time**	**Replicate**	**Protocol**	**Time**	**Replicate**	**Protocol**	**Time**	**Replicate**
DEX	0 h	1	DEX	0 h	2	DEX	0 h	3
DEX	4 h	1	DEX	4 h	2	DEX	4 h	3
DEX	8 h	1	DEX	8 h	2	DEX	8 h	3
DEX	12 h	1	DEX	12 h	2	DEX	12 h	3
DEX	16 h	1	DEX	16 h	2	DEX	16 h	3
DEX	20 h	1	DEX	20 h	2	DEX	20 h	3
DEX	0 h	4	DEX	8 h	4	DEX	16 h	4
DEX	4 h	4	DEX	12 h	4	DEX	20 h	4
Standard			Standard			Standard		
Standard			Standard			Standard		

Table 2.5 ***PHY proteomics normalization.*** *Each set is made of 10 samples that can be analysed in the same mass spectrometry experiment thanks to isobaric labelling (TMT). 'Standards' are polled samples with equal amounts of single samples and it is useful for normalizing samples that belong to different TMT samples.*

TMT sample 4			TMT sample 5			TMT sample 6		
Protocol	**Time**	**Replicate**	**Protocol**	**Time**	**Replicate**	**Protocol**	**Time**	**Replicate**
PHY	0 h	1	PHY	0 h	2	PHY	0 h	3
PHY	4 h	1	PHY	4 h	2	PHY	4 h	3
PHY	8 h	1	PHY	8 h	2	PHY	8 h	3
PHY	12 h	1	PHY	12 h	2	PHY	12 h	3
PHY	16 h	1	PHY	16 h	2	PHY	16 h	3
PHY	20 h	1	PHY	20 h	2	PHY	20 h	3
PHY	0 h	4	PHY	8 h	4	PHY	16 h	4
PHY	4 h	4	PHY	12 h	4	PHY	20 h	4
Standard			Standard			Standard		
Standard			Standard			Standard		

Finally, these 10-plex TMT labels allow a relative quantitation among the 10 conditions within the same TMT sample. However, in order to make comparable all replicates at all time samples (a total of 2x4x6 = 48 samples), the so-called *standards* are added: they mix equal amounts of proteins and from each one of the 48 samples.

2.2 Computational Methods

2.2.1 Data pre-processing

High-throughput technologies like RNA-Seq provide data characterized by systematic variability, e.g. due to different library sizes (i.e. number of reads) among samples. Therefore, it is important to correct for this type of variability before performing any analysis, in order to make samples comparable.

2.2.1.1 Normalization and filtering of RNA-seq data

As regards the transcriptomic data of this Thesis, they are provided as already normalized through the so called 'Trimmed Mean of M values' (TMM) method.

As described in Chapter 1, section 1.4.1, RNA-Seq provides sequenced fragments of transcripts, called *reads*, which are converted into 'counts': the term count (r_{ik}) refers to the number of reads that are mapped to transcript i in sample k. However, the total number of reads usually varies from sample to sample, introducing a bias that needs to be deleted (which is the main objective of normalization); in general, the expected value of the counts (r_{ik}) of gene i in sample k can be defined as:

$$E(r_{ik}) = n_k(T_{ik}L_i)/(\sum_{j=1}^{N_k} T_{jk}L_j) = n_k(T_{ik}L_i)/(S_k) \qquad (2.1)$$

where n_k is the total number of reads in sample k, T_{ik} is the number of copies of transcript i in sample k, L_i is the length of transcript i, N_k is the number of transcripts in sample k.

Moreover, $S_k = \sum_{j=1}^{N_k} T_{jk} L_j$ is the so-called ‘size factor’, it varies from sample to sample and it is unknown because it depends on N_k and T_{jk}, which are unknown. Therefore, biases are introduced by the size factors: samples with higher S_k tend to under-estimate many genes with respect to samples with smaller S_k.

The basic assumption that allows to normalize data is that the majority of genes inside each sample is not differentially expressed (for details in differential expression, see section 2.2.6). Considering two samples (indicated as 1 and 2), the M_i value is calculated as:

$$M_i = \log_2\left(\frac{E(r_{i1})}{E(r_{i2})}\right) = \log_2((n_1 \frac{T_{i1}L_i}{S_1})/(n_2 \frac{T_{i2}L_i}{S_2})) = \log_2(n_1 \frac{T_{i1}L_i}{S_1}) - \log_2(n_2 \frac{T_{i2}L_i}{S_2}) =$$

$$= \log_2(T_{i1}) - \log_2(T_{i2}) + \log_2\left(n_1 \frac{L_i}{S_1}\right) - \log_2(n_2 \frac{L_i}{S_2}) \qquad (2.2)$$

In particular, assuming that the majority of genes is not differentially expressed between condition 1 and 2, $\log_2(T_{i1}) - \log_2(T_{i2})$ of equation (2.2) becomes:

$$\log_2(T_{i1}) - \log_2(T_{i2}) = 0 \qquad (2.3)$$

Thus, when the scaling factor (SF) is calculated as the trimmed mean of all M_i values (one per each transcript), the difference $\log_2(n_1\, L_i/S_1) - \log_2(n_2 L_i/S_2\,)$ is the term of M_i actually used to calculate SF. In particular, indicating with $\bar{r}_1$ the vector of counts of all transcripts in sample 1 and with $\bar{r}_2$ the vector of counts of all transcripts in sample 2, meaning the vectors:

$$\begin{pmatrix} r_{11} \\ \dots \\ \dots \\ r_{N_k 1} \end{pmatrix} \quad \begin{pmatrix} r_{12} \\ \dots \\ \dots \\ r_{N_k 2} \end{pmatrix} \qquad (2.4)$$

SF results to be:

$$SF = TMM\{\log_2(\bar{r}_1) - \log_2(\bar{r}_2)\} \qquad (2.5)$$

where TMM stands for trimmed mean, i.e. the mean of all values excluding the most extreme ones in order to reduce the influence of possible outliers on the result. This leads to:

$$SF = \log_2\left(\frac{n_1}{S_1}\right) - \log_2\left(\frac{n_2}{S_2}\right) = \log_2\left(\frac{n_1 S_2}{S_1 n_2}\right) \qquad (2.6)$$

which is used to normalize data following this procedure:

- all counts $\bar{r}_1$ are multiplied by $1/(\sqrt{2^{SF}})$
- all counts $\bar{r}_2$ are multiplied by $(\sqrt{2^{SF}})$

Finally, the expected values of normalized counts become:

$$E(r_{i1}) = n_1 \frac{T_{i1}L_i}{S_1} / \sqrt{(n_1 S_2)/(S_1 n_2)} \; = \; T_{i1} L_i \cdot \sqrt{(n_1 n_2)/(S_1 S_2)} \qquad (2.7)$$

$$E(r_{i2}) = n_2 \frac{T_{i2}L_i}{S_2} \cdot \sqrt{(n_1 S_2)/(S_1 n_2)} \; = \; T_{i2} L_i \cdot \sqrt{(n_1 n_2)/(S_1 S_2)} \qquad (2.8)$$

Now the expression values of transcript i is comparable between sample 1 and sample 2. However, in real datasets there are more than two samples and an extension of this algorhiythm is already implemented in the bioinformatic tool used in this Thesis: edgeR v. 3.26 written in R v. 3.6 (Robinson et al., 2010).

Before proceeding with the analysis, counts are converted into the so-called *Counts Per Million* (or *CPM*), i.e. counts scaled by the number of million fragments $\left(\frac{N}{10^6}\right)$ that are sequenced:

$$CPM = \frac{r_{ik}}{\frac{N}{10^6}} = \frac{r_{ik}}{N} 10^6 \tag{2.9}$$

Then, low count genes are filtered: for this Thesis, only genes that have at least 1 CPM in at least 2 replicates of a condition (i.e. at least 2 replicates with 1 CPM in one time point of DEX or PHY protocol) are retained for the analysis, the others are excluded.

2.2.1.2 Filtering and normalization of LC-MS/MS data

As described in the previous section, the experimental design for measuring protein abundance with LC-MS/MS is based on TMT samples, with 10 sub-samples each.

First, proteomics data are filtered with the following method: if one protein has no measurements in at least one standard in the three TMT samples of DEX protocol (Table 2.4) and neither in at least one standard in the three TMT samples of PHY protocol (Table 2.5), it is deleted.

Afterwards, the normalization is carried out through two different steps: first, considering only the sub-samples within one TMT sample, then considering sub-samples belonging to different TMT samples.

The first step aims at normalizing the sub-samples of a TMT sample in order to have the same mean of the distribution of all proteins. This procedure is justified by the assumption that the majority of proteins is involved in basic biological functions, thus they are not expected to be up-regulated or down-regulated because of the synchronization protocol.

Instead, the second step is based on the multiplexing strategy described by Plubell et al., 2017, that allows to normalize also sub-samples belonging to different TMT samples. This is possible thanks to the two ‘standards’ (i.e. pooled samples) that are added to each experimental set: indeed, the abundance of each protein in the two standards are averaged in order to have a reference value for every protein. Then, the three reference values (one per TMT experiment) for each protein are averaged through a geometric mean and scaling factors are calculated in order to adjust each reference value to the geometric mean.

Finally, Figures 2.2 and 2.3 show the difference of data variability before and after normalization considering PHY protocol at time point 0 h (the other time points are similar, as well as DEX protocol).

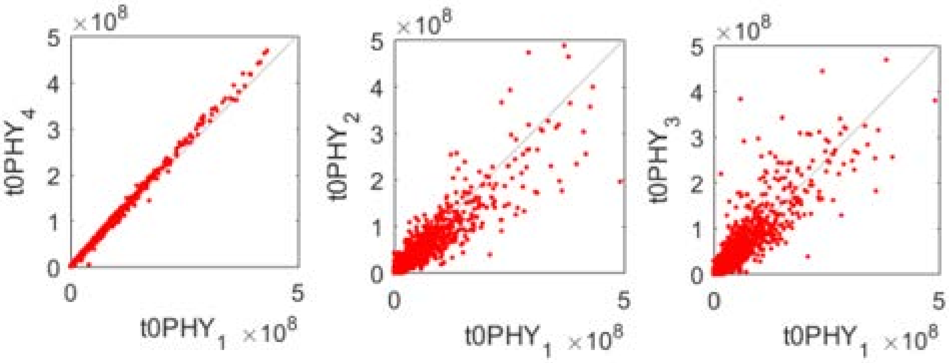


Figure 2.2 ***Data before normalization***. *Dots represent PHY proteomic measurements at time point 0 h and they allow to compare pairs of replicates as indicated by the axes. Moreover, the diagonal represent equal measurements in the two replicates compared. The remaining data (i.e. DEX protocol and the other time points of PHY protocol) are omitted because they give similar results.*

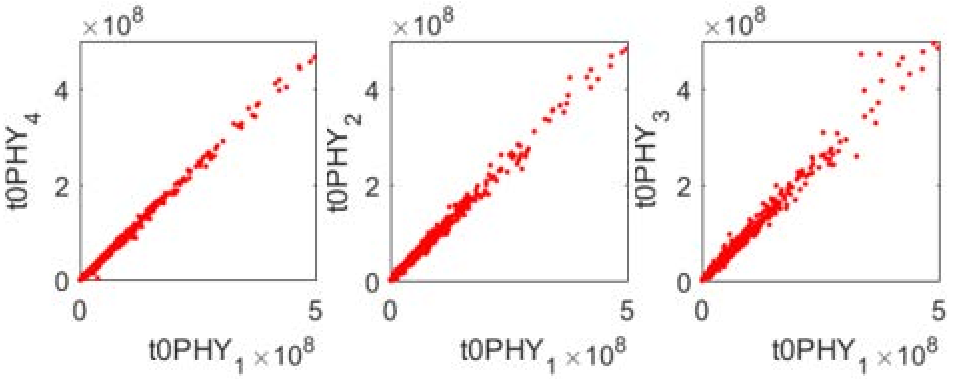


Figure 2.3 ***Data after normalization.*** *Dots represent PHY proteomic measurements at time point 0 h and they allow to compare pairs of replicates as indicated by the axes. Moreover, the diagonal represent equal measurements in the two replicates compared. The remaining data (i.e. DEX protocol and the other time points of PHY protocol) are omitted because they give similar results.*

As shown by Figure 2.2 and 2.3, the normalization performs well in making the measurements comparable: while before normalization a significant number of dots tends to depart from the diagonal, after the normalization the majority of them lies on the diagonal itself (as it is expected from replicates of the same time point in the same protocol). The only exception is the pairwise comparison between replicate 1 and replicate 4, which is similar both before and after normalization: this is due to the fact that they belong to the same TMT sample (as shown in Table 2.5).

2.2.2 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a multivariate statistical method that allows to summarize the information of a high-dimensional dataset, while reducing the loss of information. The key principle is to project a given matrix X, with N observations and M variables, into a new coordinate space made of the so-called ‘Principal Components’ (PCs). These latent variables are built by solving an optimization problem, which aims at maximizing the percentage of variability explained by each principal component (see Figure 2.4) and at minimizing the residuals of data in the new coordinate space. Moreover, one of the objectives of PCA is to find correlations among variables, in order to reduce data dimensionality eliminating redundant information. Therefore, instead of retaining M variables, Q latent variables are selected, with Q that is usually lower than the matrix apparent rank (i.e., Q is lower than M and N).

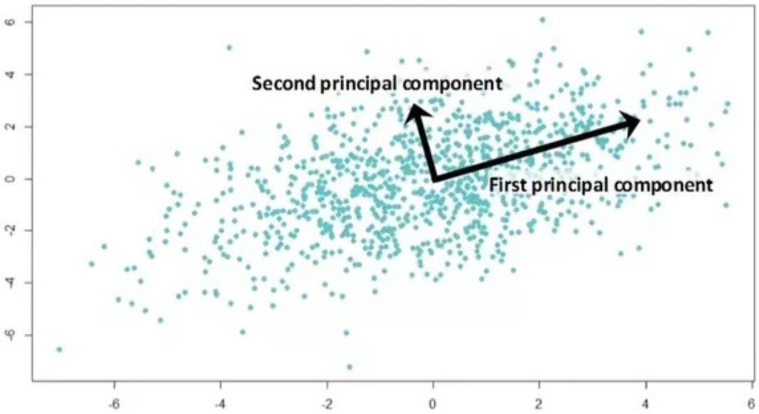


Figure 2.4 ***PCA.*** *The cloud of data extends mainly along a direction that corresponds to the first Principal Component (PC1). Instead, PC2 captures the direction of the highest remaining variability. Since two PCs are retained, Q=2 (taken from www.medium.com).*

The coordinates of data projections on the PCs space are called ‘scores’, while the coefficients of the linear combination of the M original variables giving the PCs are collected in Q vectors called ‘loadings’. The outer products of scores and loadings becomes the new representation of the original matrix X:

$$X = \sum_{a=1}^{R} t_a p_a^T = \sum_{a=1}^{Q} t_a p_a^T + \sum_{a=Q+1}^{R} t_a p_a^T = TP^T + E \qquad (2.10)$$

where R is the rank of X, t_a represents the scores of each observation on the a^th PC, p_a^T is the transpose of the loading vector for the a^th PC, E is the residual matrix that has to be minimized in the least-square sense.

In particular, score plots are useful for studying similarities and/or differences among observations: opposite score values with respect to one principal component reveals anti-correlation between two observations, while getting similar scores for two observations means that they are positively correlated (an example is provided in Figure 2.5). Similar interpretations on positive and/or negative correlations of variables can be made based on their location on the loading plot.

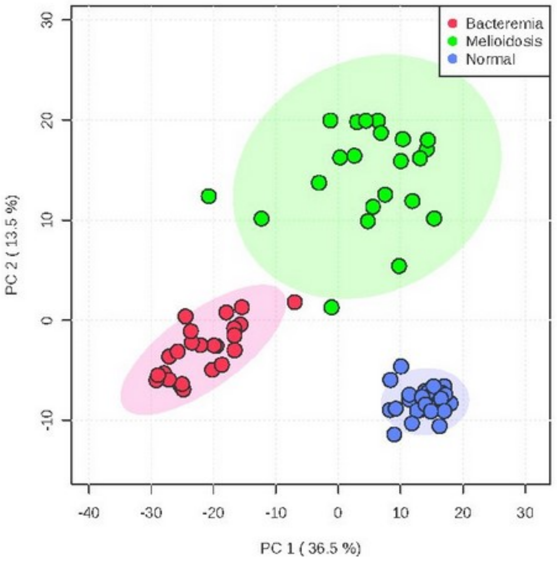


Figure 2.5 ***Score plot.*** *This is an example of 2D - score plot representing human plasma of Bacteremia (red cluster), Melioidosis (green cluster) and control (blue cluster) subjects. Observations that are similar are projected close to each other in the PCA score plot, generating three different clusters of data (taken from Lau et al.(2016), 'Metabolomic Profiling of Plasma from Melioidosis Patients Using UHPLC-QTOF MS Reveals Novel Biomarkers for Diagnosis', International Journal of Molecular Sciences 17(3):307).*

As regards the applications of PCA, it can be used to perform an explorative analysis on a given dataset, for example to reveal the level of inter- and intra-variability among different groups of data. Additionally, it can be calibrated on historical datasets of a manufacturing process in order to project new observations on the PCA model itself, revealing the conformity of new products with the historical ones. In this Thesis, PCA is used for two applications: for an exploratory analysis of transcriptomic and proteomic data aiming at assessing the level of inter- and intra-variability between DEX and PHY protocol, and for selecting circadian profiles after calibrating a PCA model with *in silico* periodic profiles (see section 2.2.3).

2.2.3 Circadianity: PCA model

An innovative tool developed to identify circadian profiles is used in this Thesis: a PCA model that was developed and validated by Matteo Bicego while working in the group of Prof. Nicola Elvassore at the Industrial Engineering Department of University of Padova and Vimm (Venetian Institute of Molecular Medicine). As it is unpublished it is briefly described below.

As anticipated in section 2.2.2, Principal Component Analysis can be used both for a preliminary data visualization and for calibrating a model. The major difference is that in the former case new data are used to calculate both scores and loadings, therefore data themselves are used to create the latent space where they are projected. In contrast, in the latter case a specific dataset, called calibration dataset, is used to calculate the loadings and to select the adequate number of PCs for the latent space, while the scores of new experimental data are calculated by using the loadings obtained though calibration. Therefore, all new data are projected into the same latent space, whose characteristics depend only on the calibration dataset used.

More specifically, Matteo Bicego's algorithm is a 3-step process: first, it calibrates a PCA model based on an *in silico* set of profiles; then, it performs a permutation of all the replicates available at each time point for every transcript (or protein) in the experimental data; last, it projects all the permutated profiles and calculates their barycentre projections on the two-dimensional PCA plot.

2.2.3.1 PCA calibration

The calibration profiles that are used to build the model are made of 6 values that mimic experimental samplings between 0 h and 20 h, with intervals of 4 h between samples. The overall sampling time is 20 h and not 24 h because circadian profiles are oscillating with a period of 24 h, therefore time 0 and 24 h are overlapping. Three main types of profiles make up the calibration matrix:

- the first one represents randomly fluctuating profiles. Specifically, random profiles are built by making all the possible combinations of ordinate values between 0 and 1, with steps of 0.25, for all time points (Figure 2.6);
- oscillating profiles with a period of 2π given by the following equation:
  $$y = \sin(x + s + m \cdot (\sin(x + s))) \qquad (2.11)$$
  where $x \in (0,6)$, while s and m are parameters that shift the overall profile and modify the profile shape, respectively. In particular, s is between 0.1 and 6.3 with steps of 0.1, while m is between -1 and 1, with steps of 0.05. Figure 2.7 a and b show some of those profiles as an example;
- oscillating profiles with a period of 2π given by the following equation:
  $$y = \sin\big(x + s + m \cdot (\cos(x + s))\big) \qquad (2.12)$$
  whose parameters have the same meaning of equation (2.11) and that generates different shapes due to the cosine function inside the sine one. In particular, s is between 0.1 and 6.3 with steps of 0.1, while m is between -1 and 1, with steps of 0.05. Figure 2.7 c and d show some profiles as an example.

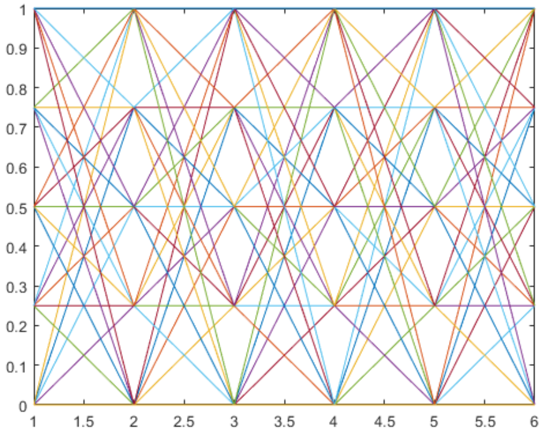


Figure 2.6 *Random profiles generated by the PCA model, through all the possible combinations of 0, 0.25, 0.5, 0.75, 1 values at each of the 6 values of the independent variable (that represents the sampling time).*

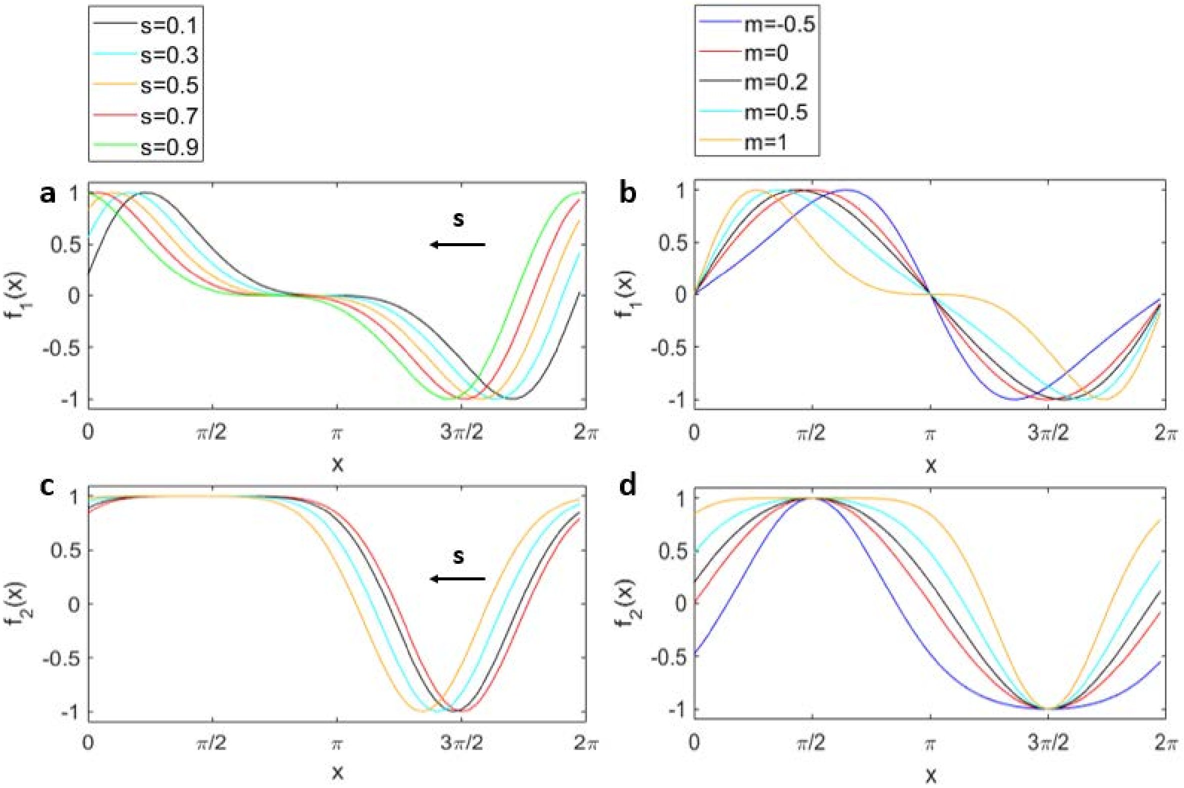


Figure 2.7 *Effects of s and m parameters on the periodic functions of equations (2.11) and (2.12), which in turn are indicated in the ordinate as $f_1(x)$ and $f_2(x)$, respectively. In particular, a and c subplots show the shifting effects of s, setting m=0, while b and d show the effects of m parameter on the final shape, setting s=0.1.*

Then, the calibration matrix is pre-processed by means of Standard Normal Variate (*snv*, meaning that the mean is subtracted to each row, which is then divided by its standard deviation) and used as input for a Principal Component Analysis (PCA). The percentages of explained variance for each principal component are equal to, in order: 26.34%, 26.28%, 16.13%, 16.13%, 15.12%, 0%. Therefore, 2D-score plots allow to represent more than 50% of the total data variability.

This calibrated model is particularly suitable for studying circadian rhythms because of the main features of both its loading plot and its score plot (for details on loading and score plot, see section 2.2.2). Indeed, the six variables of the *in silico* dataset (representing the six time points between 0 h and 20 h) are ordered along a circumference in the loading plot, resembling a clock (Figure 2.8). Moreover, the 2D-score plot has some useful peculiarities:

- some profiles of equation (2.11) and (2.12) are projected on the external circumference in 2D and on the equator of the 3D score plot (black dots of Figure 2.9, corresponding to the temporal profiles of Figure 2.10);
- the rest of the periodic profiles are projected inside that circumference, but still in the most external region (green and orange dots of Figure 2.9);
- random profiles are projected on its central area.

Finally, 'calibrating' a PCA model means choosing the number of Principal Components for the latent space and calculating the loadings for projecting future observations. For this application, the first two PCs are retained, allowing to visualize a 2D- clock that represent the whole circadian cycle of 24 h, while the scores (t_{new}) of new experimental data (x_{new}) are calculated as:

$$\mathrm{t}_{new} = x_{new}\mathrm{P}_{cal} \tag{2.13}$$

where P_{cal} is the loading matrix calculated during model calibration. When scores of experimental data are calculated with this procedure, the result is that circadian profiles are projected toward the external part of the clock, while random profiles are distributed around the centre.

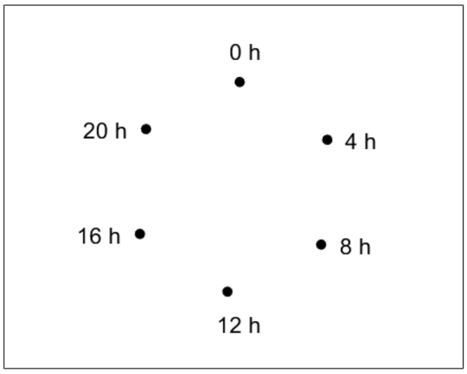


Figure 2.8 ***Loading plot.*** *This is a plot that represents the new coordinates of the original matrix variables: in this context, 6 time points between 0h and 20h.*

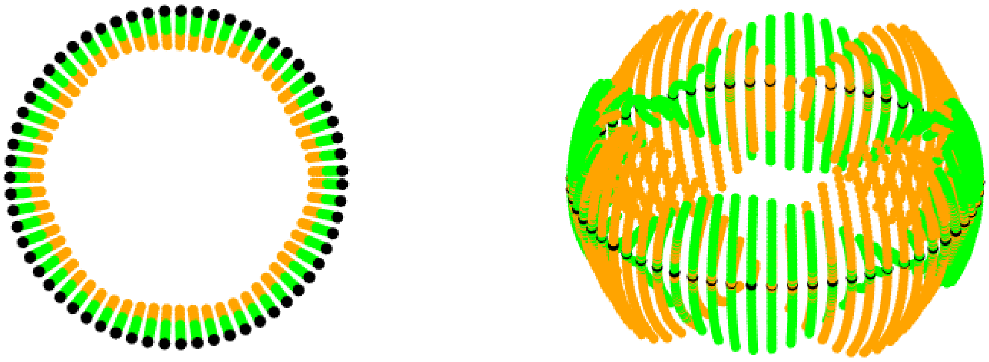


Figure 2.9 *Score plot of the periodic functions: equation (2.11) is represented with orange dots, equation (2.12) is represented with green dots. Black dots highlights the profiles that are projected along the external circumference of the 2D plot (on the left) and on the equator of the 3D plot (on the right).*

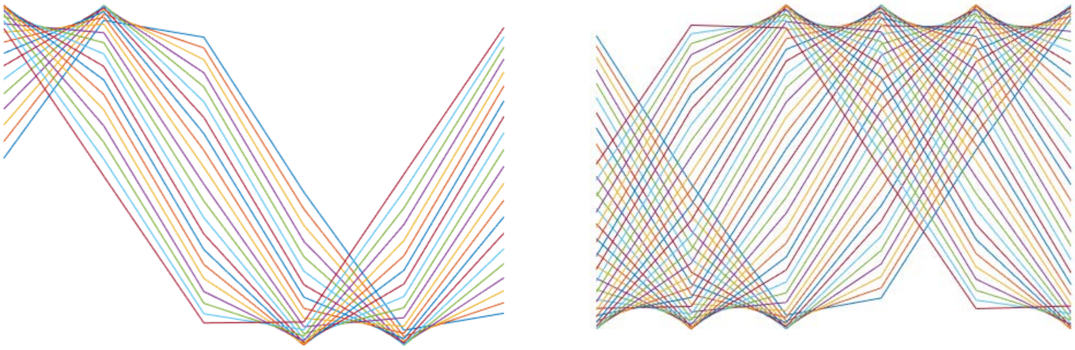


Figure 2.10 *Visualization of the in silico profiles that are projected along the circumference (or equator in the 3D plot). They are divided into two sub-plots just for achieving a better visualization.*

2.2.3.2 Experimental profiles projection

For studying circadian rhythms, temporal profiles of mRNAs and proteins are needed; however, both RNA-Seq and LC-MS/MS are destructive analyses, therefore at each sampling time (meaning 0 h, 4 h, 8 h, 12 h, 16 h, 20 h) different cells are actually used for isolating transcripts or proteins, leading to pseudo-temporal profiles. Since cells evolutions are never represented for the entire duration of the experiment, all the permutated profiles (calculated with the 4 replicates per time point) are taken into account.

Therefore, all the permutated profiles are projected into the PCA model, then the barycentre coordinates corresponding to each gene are calculated in terms of PC1 and PC2. The next step consists in the determination of a relative radius for each transcript: the Cartesian coordinated of the barycentres are converted into polar ones, thus giving the radial distance from the centre of the PCA clock; then, this absolute radius is divided by the maximum one, that is found with calibration profiles (for details in the calculation, see Appendix A.1). The higher the relative radius (*Rr*), the more marked is the circadian behavior of the profile.

To obtain a rigorous threshold of *Rr* to define a profile as circadian, the probability of the in silico generated profiles of being circadian or not as a function of *Rr* is plotted in Figure 2.11. The critical *Rr* = 0.7, when the probability of being circadian is higher than not being circadian, is defined as a threshold.

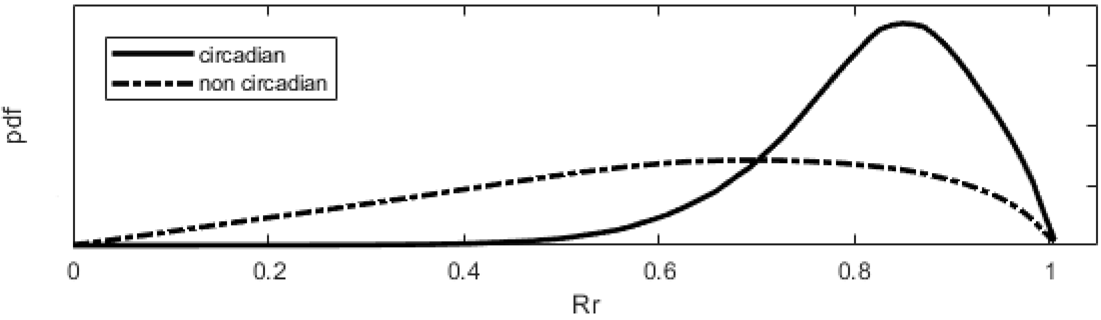


Figure 2.11 *Probability Density Function (pdf) of circadian (solid line) and not circadian (dotted line) profiles, calculated at each relative radius value (between 0 and 1).*

This procedure allowed to determine the proper threshold to select oscillatory profiles with a period of 24 h: a minimum relative radius of 0.7 (corresponding to a probability of approximately 95% of being circadian), which is used in the next chapters to select circadian transcripts and proteins.

2.2.4 Estimation of phase and amplitude

The relative radius is a useful parameter to define circadianity, but it is not enough to characterize a circadian dynamics; therefore, other two parameters are estimated: peak phases and amplitudes.

As anticipated, phases are calculated by the PCA model; this is possible because periodic profiles that display one peak within 24 h are projected on the PCA model close to the loading (meaning, sampling time) corresponding to their time of peak. Therefore, both the barycentres scores and the loadings are converted from Cartesian coordinates into polar coordinates and phases are calculated as the difference between the angular coordinate of the profile barycentre and the angular coordinate of the first loading (i.e. time point 0 h, see Figure 2.12) . This calculation provides a peak phase between 0 and 2 π, which can be easily converted into a phase between 0 h and 24 h through a proportion.

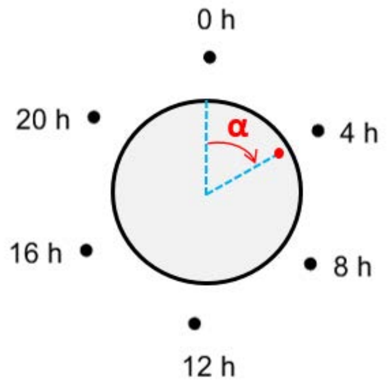


Figure 2.12 *Graphical representation of the phase calculation: considering the projection of the profile barycentre (the red dot in the figure), the phase alpha in radians units is given by the difference between the angular coordinate of the barycentre and the angular coordinate of t=0h.*

Instead, amplitudes are not calculated by the PCA model. After different trials involving cosine and sine fitting on the experimental data of this Thesis, the following formula has proved to be the most robust for amplitude definition given the noise in the data:

$$\mathrm{A} = (\max(x_i) - \min(x_i))/2 \tag{2.14}$$

where x_i represents the mean measurements of gene i across 20 h.

2.2.5 Heatmaps

Heatmaps are a common way of visualizing gene expression data (Figure 2.13). They can be described as grids in which each row represents a profile of gene expression (or transcription, or protein abundance) and each column represents a sample, while the color of each position in the grid represents the value of the property of interest. For a proper visualization, a key step before generating heatmaps is to order the genes (or proteins) according to their profile similarity: the most widely used method in omics applications is hierarchical clustering. To this aim, a distance measure between profiles needs to be calculated. For example, the Pearson correlation coefficient ρ_{XY} can be used:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{2.15}$$

where, given two statistical variables X and Y, σ_{XY} represents the covariance between X and Y, while σ_X and σ_Y represent their standard deviations.

This way, all values are between -1 and 1: -1 represents perfect negative correlation, 1 represents perfect positive correlation, 0 indicates absence of correlation. Afterwards, this distance matrix is ordered to place similar profiles (meaning with small distance between each other) next to each other, and this is the final input then plotted as a heatmap.

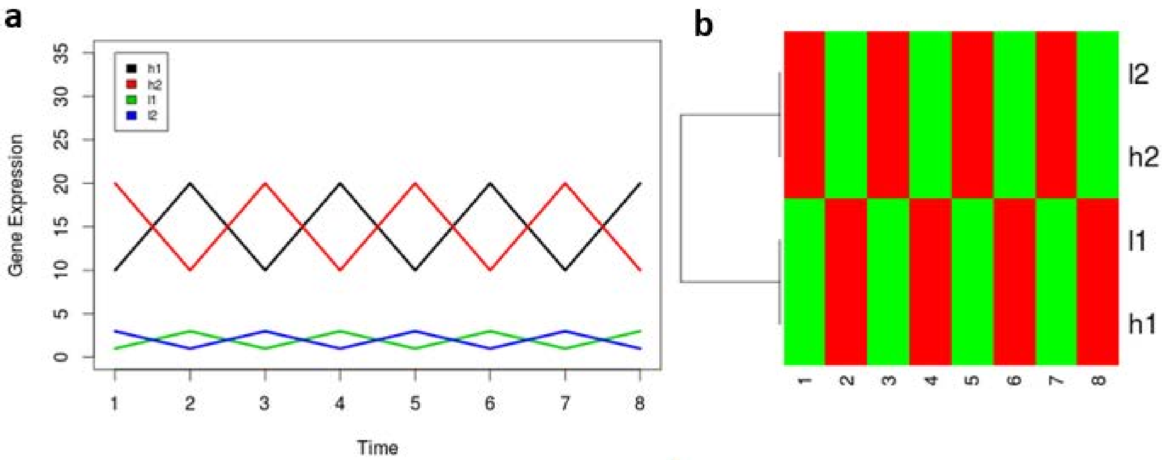


Figure 2.13 ***Heatmaps.*** *These figures illustrate how heatmaps allow to visualize common patterns of gene expression. Indeed, Figure 2.13 a shows two profiles with high absolute values, i.e. h1 and h2, and two profiles with low absolute values, i.e. l1 and l2. However, the relative change in gene expression is more similar between h1 and l1 and between h2 and l2, because peaks and valleys occur at the same sampling time. Thus, Figure 2.13 b represents a heatmap with clustering that highlights the similarity between l2 and h2 and between l1 and h1 (taken from www.opiniomics.com).*

However, in the circadian field heatmaps are usually built in a slightly different manner: indeed, instead of calculating the distance matrix and performing clustering, profiles are ordered based on peak phases.

In particular, the original matrix is first scaled in order to take on values between -1 and 1, by using the formula:

$$x_{scaled} = 2\left(\frac{x-\min(\bar{x})}{\max(\bar{x})-\min(\bar{x})}\right) - 1 \tag{2.16}$$

where x_{scaled} is the expression value between -1 ad 1, x is the original value, $\bar{x}$ is the vector of expression values within the same profile (in other terms, one row of the original matrix). Then, scaled profiles are ordered based on their phase estimation and the matrix is represented as an image of different colors and/or intensities, giving the final heatmap (an example is provided in Figure 2.14).

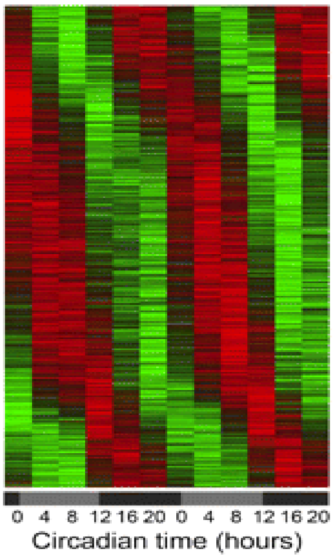


Figure 2.14 ***Heatmaps of circadian data.*** *This figure highlights the presence of expression peaks distributed within 0 h and 24 h (taken from Keller et al., A circadian clock in macrophages controls inflammatory immune responses, PNAS December 15, 2009 106 (50) 21407-21412).*

2.2.6 Differential Expression Analysis

Differential expression analysis is a method for evaluating the statistical significance of differences in gene expression among different conditions, e.g. treatment and control. This is done by carrying out hypothesis tests, which are based on the calculation of a test-statistics and on its comparison with an adequate probability distribution in order to assign a p-value to each test.

For example, a t-test allows to determine the statistical significance of the difference between two means (Figure 2.15). Indeed, given two sets of observations belonging to group A and B respecting the hypothesis of equal variances ($\sigma_A = \sigma_B$), of independence of their sampling and of Gaussianity of their true populations, in a t-test these groups are compared in order to test whether the true means of their populations (μ_A and μ_B, respectively) are equal (*null hypothesis*, H_0) or statistically different (*alternative hypothesis,* H_1):

$$H_0: \quad \mu_A = \mu_B \tag{2.17}$$

$$H_1: \quad \mu_A \neq \mu_B \tag{2.18}$$

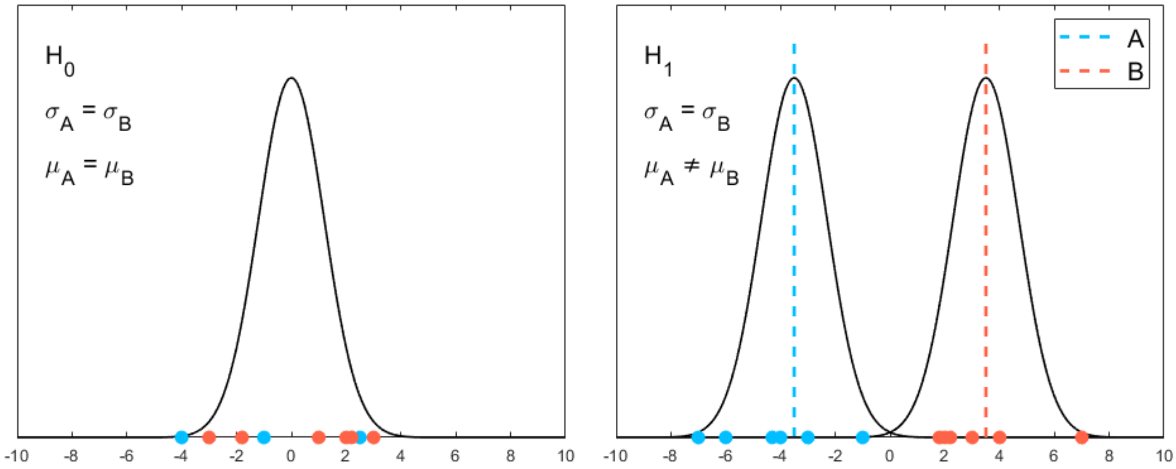


Figure 2.15 *Graphical visualization of the meaning of null hypothesis (H_0) and alternative hypothesis (H_1): in the former case, the two groups A and B belong to the same true population (Gaussian distribution, equal variances and means), while in the latter case observations of group A and B belong to different populations (even though the distributions are Gaussian with equal variances, the mean is different).*

The t-test statistics is calculated and the corresponding p-value is retrieved from tables of Student's t-distributions (available, for example, in software with statistics tools, like MATLAB). If the p-value is smaller than a threshold fixed *a priori*, usually 0.05 or 0.01, the null hypothesis is rejected and the population means are considered statistically different.

Many other statistical tests have been developed, by testing different null hypothesis (like the equality of population variances instead of means), by using different test-statistics and/or probability distributions, but the rationale is always the same.

Finally, when a statistical test is performed, one of the four cases shown in Table 2.6 may occur, without having the possibility to determine precisely which one has occurred. The goal is to maximize the identification of true positive and minimize the false positive and negative. A particularly challenging task, especially given the small number of measurements for each transcript (usually less than 5) and the high number of transcript types (of the order of 10^4). Adding to this, in the case of temporal series, multiple time-point comparisons are required.

Table 2.6 *The four possible outcomes of a hypothesis test.*

	$\boldsymbol{H_0}$ **accepted**	$\boldsymbol{H_0}$ **rejected**
$\boldsymbol{H_0}$ **true**	True negative	False positive
$\boldsymbol{H_0}$ **false**	False negative	True positive

The next sections describe the two algorithms applied in this thesis to find Differentially Expressed Genes (DEGs): edgeR and FunPat.

2.2.6.1 edgeR

A widely-used R package to select DEGs is edgeR, whose name is an acronym for "empirical analysis of DGE in R". In particular, DGE stands for *digital gene expression* and refers to the fact that this method is suitable for 'digital', meaning discrete, expression data, as in case of counts in RNA-Seq (opposed to the continuous distribution of values that was typical of microarray data).

In edgeR, raw counts are normalized by means of TMM ("Trimmed Mean of M values", see section 2.2.1) and then their variance is estimated by fitting a Negative Binomial (NB) model. Indeed, this model is appropriate for RNA-Seq that are usually overdispersed (Zhou et al., 2011), meaning with a variance that is bigger than the mean; therefore, counts of transcript i in sample k depend on the two parameters that define a NB distribution:

$$r_{ik} \sim NB(\mu_{ik}, \varphi_i) \tag{2.19}$$

i.e., they depend on the mean μ_{ik} and the dispersion φ_i. In addition, these parameters are linked by the equation:

$$\varphi_i = \mu_{ik}(1 + \mu_{ik}\varphi_i) \tag{2.20}$$

where $\mu_{ik}\varphi_i$ is the overdispersion term; when this term is equal to 0, the NB distribution is equivalent to the Poisson one (with $\mu_{ik} = \varphi_i$) .

In order to estimate the dispersion, edgeR maximizes the negative binomial likelihood.
In general, given that a random variable X has a probability function p depending on one parameter and given that its outcome is equal to x, the likelihood is defined as the probability that X takes on a certain value x when the true value of the parameter is ϑ:

$$\mathcal{L}(\vartheta \mid x) = p_\vartheta(x) = P_\vartheta(X = x) \tag{2.21}$$

Moreover, Maximum Likelihood Estimation (MLE, that is the method used by edgeR) is a statistical technique that allows to estimate the parameters of a distribution by maximizing the corresponding likelihood function. However, the Negative Binomial dispersion may be inaccurate: indeed, it tends to estimate high dispersions for genes with low counts, therefore edgeR can implement the Quasi-Likelihood (QL) method to take into account gene specific variability.

The next step of the DE analysis is linear modelling, which requires a design matrix as input: each row of this matrix correspond to a sample, while columns indicate the groups to which a given sample is associated. In other terms, the columns of the design matrix represent the conditions that are compared in the DE analysis: in the case of this Thesis, the two synchronization protocols are compared at each time point. However, this leads to 6 pairwise comparisons (at time 0h, 4h, 8h, 12h, 16h, 20h) that are independent from one another, thus edgeR does not take advantage of the correlation between subsequent time points in time course data that may make DEGs selection more robust.

For fitting experimental data, edgeR uses generalized linear models (GLM). In general, linear models (LM) estimate the dependent variable y, or 'response variable', as a linear combination of the input variables, or 'factors' or 'regressors' (x_i), through the parameters β_i:

$$\mu_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i + \cdots + \beta_n x_n \quad (2.22)$$

Where the assumption is that the response variable has a normal distribution (N) with mean μ_i and error ε (Figure 2.16):

$$y \sim N(\mu_i, \varepsilon) \quad (2.23)$$

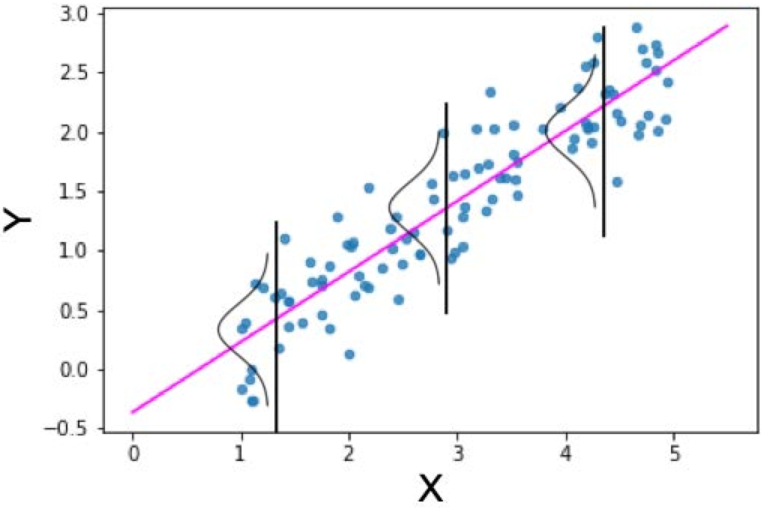


Figure 2.16 *Normal distribution of the dependent variable y (taken from www.towardsdatascience.com).*

Instead, when the gaussianity of y is not respected, the relationship between response variable and regressors is not linear anymore. Therefore, another type of regression is appropriate: generalized linear models (GLM). In particular, GLM are characterized by:

- a probability distribution for y;
- a linear predictor η_i;
- a link function $g(\cdot)$.

Indeed, in the original formulation of GLM, y needed to have a probability distribution belonging to the so-called 'exponential family' (e.g. normal, binomial or Poisson distributions), but then it has been extended to other types of distributions.

As regards the linear predictor, it is defined as:

$$\eta_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i + \cdots + \beta_n x_n \tag{2.24}$$

Finally, the link function $g(\cdot)$ relates the expectation of the response variable ($E(y) = \mu_i$), to the linear predictor:

$$g(\mu_i) = \eta_i \tag{2.25}$$

In particular, edgeR fits a quasi-likelihood negative binomial GLM.

After fitting, a quasi-likelihood F test is performed. In general, an F test is a statistical technique that assesses the statistical significance of the difference between two populations variances. Considering two independent samples, both normally distributed, and given their calculated variances (s_1^2 and s_2^2, respectively), the F test aims at verifying that the true variances of the corresponding populations (σ_1^2 and σ_2^2, respectively) are equal or different. In particular, the null hypothesis (H_0) is defined as:

$$H_0: \sigma_1^2 = \sigma_2^2 \tag{2.26}$$

while the alternative hypothesis (H_1) is:

$$H_1: \sigma_1^2 \neq \sigma_2^2 \tag{2.27}$$

In order to evaluate the statistical significance of the variances difference, the F statistics must be calculated:

$$F_{(df_{max}, df_{min})} = s_{max}^2 / s_{min}^2 \tag{2.28}$$

The numerator is always the highest number between s_1^2 and s_2^2 and the denominator is always the lowest one, thus F is higher than 1. Then, this calculated statistics is compared to the values that are tabulated for the Fisher-Snedecor distribution, based on the degrees of freedom df_i:

$$df_i = n_i - 1 \tag{2.29}$$

where n_i is the number of observations in each sample ($i = max, min$ referring to the maximum and minimum variances, respectively). Finally, if the F statistic is higher than the tabulated one, the null hypothesis is rejected and the two variances can be defined as statistically different. The last step of the selection of differentially expressed genes is to correct for multiple testing (see section 2.2.6.3).

2.2.6.2 FunPat

The development of high-throughput measurement techniques has allowed to investigate temporal profiles of gene expression, therefore new statistical methods have been developed in order to take into account the correlation between subsequent time points for a more robust selection of DEGs.

FunPat is an algorithm for DE Analysis available as R package, developed by Sanavia, Finotello and Di Camillo (Sanavia et al., 2015); it is developed for RNA-Seq data with a temporal profile and it is able to handle data with few replicates (e.g. data with only two replicates at one time point).

The input data need to be already normalized, because this R package has no functions to perform normalization (unlike edgeR). Moreover, all the following discussions refer to logarithmic data (even though it is not explicit in the symbols used, for ease of visualization). Then, a p-value is attributed to each gene based on the Bounded Area method developed by Di Camillo (Di Camillo et al., 2006). This method considers the area A between two profiles, that may be either two experimental conditions to be compared (e.g. treatment and control, as shown in Figure 2.17) or one experimental condition versus a baseline, and compares A with an area derived from a null distribution (A_0).

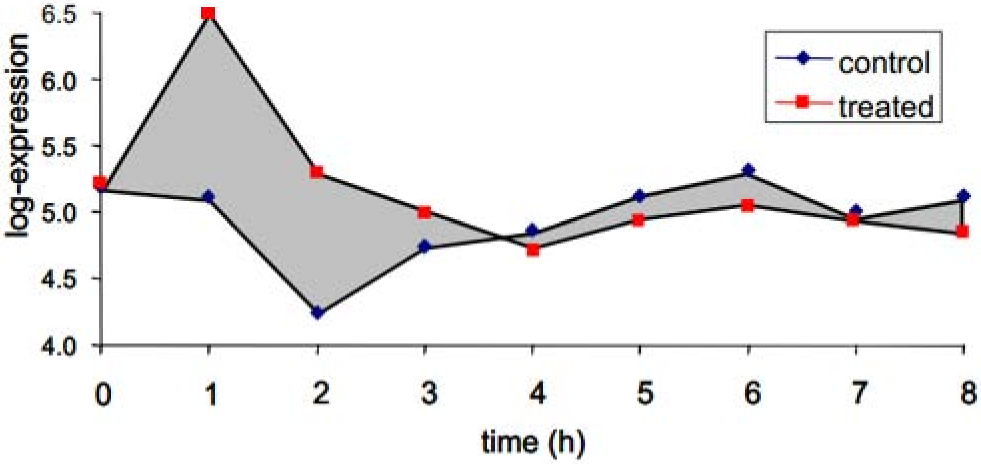


Figure 2.17 *This figure highlights the bounded area calculated between the profiles of control and treatment conditions for the same gene, whose expression values are in logarithmic scale.*

In particular, considering a gene X with two expression profiles $x^T(t_k)$ and $x^C(t_k)$ (representing treatment and control, respectively), the method defines at each time point t_k, k=1,…, M (with M total number of time points) the deviation between the two measurements:

$$d(t_k) = x^T(t_k) - x^C(t_k) \tag{2.30}$$

Then, each contribution to the total area bounded between the two profiles is calculated as:

$$A_k = (d(t_{k+1}) + d(t_k)) \cdot (t_{k+1} - t_k)/2 \tag{2.31}$$

when $sign\big(d(t_{k+1})\big) = sign(d(t_k))$, otherwise it is calculated as:

$$A_k = (d(t_{k+1})K_1 + d(t_k)K_2) \cdot (t_{k+1} - t_k)/2 \tag{2.32}$$

where $K_1 = d(t_{k+1})/(d(t_{k+1}) + d(t_k))$ and $K_2 = d(t_k)/(d(t_{k+1}) + d(t_k))$.

Finally, the bounded area can be calculated:

$$A = \sum_{k=1}^{M-1} A_k \tag{2.33}$$

Once A has been calculated for each gene, DEGs are selected when

$$A > \vartheta_A \tag{2.34}$$

where ϑ_A is a threshold corresponding to a significance level α, which in turn is determined based on a null hypothesis of the bounded areas. In particular, the null hypothesis distribution can be calculated when replicates are available in at least one time point. For example, considering two replicates x_a and $x_b$under the null distribution:

$$x_a = \mu + \varepsilon_a \tag{2.35}$$

$$x_b = \mu + \varepsilon_b \tag{2.36}$$

where μ is the real (unknown) gene expression and ε is the error, therefore the deviation between the two replicates is given by the difference between the two errors:

$$d = x_a - x_b = \varepsilon_a - \varepsilon_b \tag{2.37}$$

Therefore, $d(t_k)$ calculated with the available replicates are fitted in order to have a null distribution d^{H_0}, from which B profiles with M time points are sampled: they are used to calculate bounded areas under the null hypothesis (A^{H_0}). Finally, Gamma, Log-normal and Weibull distributions are fitted on A^{H_0} and the best model is chosen based on the goodness of fit and parameter precision.

As for edgeR, p-values are corrected for multiple testing with Benjamini-Hochberg method (described in section 2.2.6.3), providing the final list of Differentially Expressed Genes.

2.2.6.3 Correction for multiple testing

As described at the beginning of section 2.2.6, the null hypothesis is rejected when the p-value is lower than a threshold, e.g. 0.05. If only one test was performed, this would mean that the probability of rejecting the null hypothesis when it is true is 0.05 at most: indeed, this probability can never be equal to zero (as desired) because when statistical tests are carried out, some p-values fall below the threshold just by chance. The rejection of the null hypothesis when it is true is called ‘false positive’ or ‘Type I error’(as shown in Table 2.6).

However, in the context of genomics and transcriptomics, thousands of tests are carried out in each experiment, one for each gene/transcripts. This increases the false positive rate to a level that is not acceptable anymore, therefore statistical methods have been developed to limit this drawback.

For example, a widely used method, which is used also in this Thesis, is the Benjamini-Hochberg correction for controlling the False Discovery Rate (FDR), which is defined as the expected value of the proportion of Type I errors among the number of rejections of the null hypothesis. First of all, the estimated p-values for each gene are sorted in ascending order and ranked, then DEGs are selected by setting:

$$p(j) \leq \delta \frac{j}{m} \tag{2.38}$$

Where m is the total number of tests, j is the rank of the jth p-value and δ is a number fixed a priori (equivalent to the 0.05 threshold previously described).

2.2.7 Enrichment Analysis

When mRNAs or proteins are analysed, it is important to characterize not only their dynamics and relative abundances among different conditions, but also their biological functions. Usually, some gene lists of interest are extracted from data, which is the most critical aspect to identify the associated meaningful biological functions: for example, in Chapter 3, gene lists referring to differentially expressed genes or circadian genes peaking at specific times are identified for the DEX or PHY protocol. Genes biological function can be characterized by the so-called Enrichment Analysis, whose key steps are:

- selection of a biological database as a reference;
- implementation of a statistical test;
- choice of an informative and meaningful way to visualize the results.

2.2.7.1 Biological Databases

Many biological databases are available, containing a large number of *gene sets*, meaning groups of genes with similar biological functions and/or similar locations inside a cell; genes that are collected in a gene set are said to be *annotated* to that gene set. However, some databases are not accurate enough, because they are generated automatically without being revised by experts in the field. Therefore, more accurate results are obtained with curated databases, although for more explorative applications non-curated database play an important role for a first screening.

A widely-used database is the so-called Gene Ontology (GO), in which genes are organized in a hierarchical structure with the more general terms as ‘roots’, to which more specific terms are connected through two types of relations: 'is-a' or 'part-of' (Figure 2.18 a). This is the result of an integrated effort of different levels of manual curation and automatic annotation, and users can select the level of accuracy for their specific application.

Moreover, three main types of GO terms are defined: Cellular Component (CC), indicating the location inside a cell where the gene product is active; Biological Process (BP), i.e. a set of molecular activities that usually lead to a chemical or physical transformation; Molecular Function (MF), meaning activities carried out by single gene products (i.e. proteins or RNAs) or by the combinations of gene products (i.e. molecular complexes).

In this Thesis, another widely-used database is employed: REACTOME, which is an open-source and manually curated pathways database. In general, a pathway is a sequence of biological activities inside a cell that lead to a certain product (like the synthesis of a molecule) or a certain biological change inside a cell. In order to organize the information, REACTOME creates a hierarchical structure among pathways: for example, metabolism of lipids and metabolism of carbohydrates are a sub-category of the same metabolism pathway (another example is provided in Figure 2.18 b).

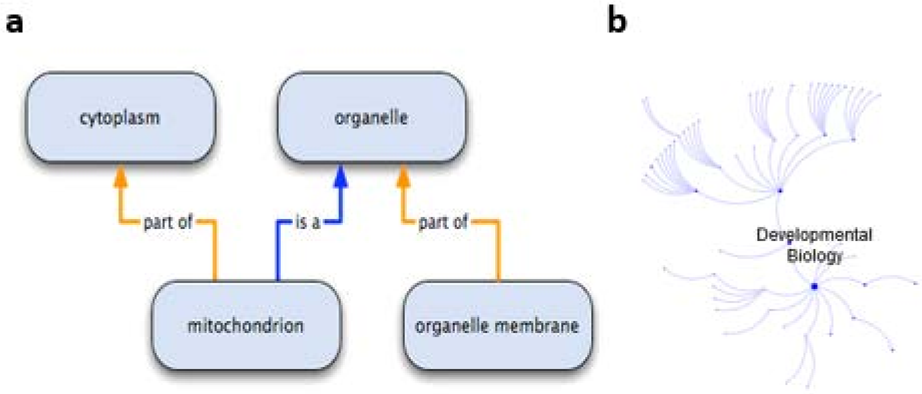


Figure 2.18 ***a)*** *Example of relations between generic nodes, or 'parents' (here, mitochondrion and organelle membrane), and more specific terms, or 'children' (here, cytoplasm or organelle) in GO: mitochondrion is part of cytoplasm and is a type of organelle, organelle membrane is part of an organelle (taken from www.geneontology.org).* ***b)*** *Example of the REACTOME pathway 'Developmental Biology': the biggest blue node at the centre represents the most general term, the nodes directly connected to it are the most general ones remaining, and so on towards the external region of the diagram.*

2.2.7.2 Statistical Test

In this Thesis, an enrichment analysis is performed though a right-sided hypergeometric test. In general, a hypergeometric distribution describes the probability of k successes in n draws without replacements from a population of size N, in which K subjects have the feature of interest (that determine a successful draw). When a gene list of interest is selected and a specific pathway appears among the annotations of those genes, a hypergeometric test assesses whether this pathway is over-represented or not.

This hypergeometric test is equivalent to the one-tailed Fisher Exact Test (Rivals et al, 2007); as an example of Fisher Exact Test, a test of over-representation of a certain Pathway i in a list of DEGs is considered (Table 2.7). In this example, a DEGs are annotated to Pathway i, while b DEGs are not; moreover, c genes that are not selected as DEGs are annotated to Pathway i, while d are not. The main question to be answered is: is there a statistical evidence that DEGs tend to be annotated to that Pathway more than non-DEGs? The Null Hypothesis states that the numbers of enriched DEGs and non-DEGs are different only by chance, while the Alternative Hypothesis states that Pathway i is over-represented in the list of DEGs.

Table 2.7 *Table of observed occurrences that are used to calculate p-values in a one-tailed Exact Fisher Test. Two gene lists are considered: genes selected as Differentially Expressed and those that are not selected. In particular, a DEGs are annotated to the pathway of interest, while b of them are not. Moreover, c genes that are not selected as DEGs are annotated to the GO term, while d are not. The main feature of this type of tables is that the total number of elements of each row and column is fixed; only the specific a, b, c and d can change.*

	∈ Pathway i	**∉ Pathway i**	**Rows Total**
DEGs	a	b	$\boldsymbol{a+b}$
Non-selected as DEGs	c	d	$\boldsymbol{c+d}$
Columns Total	$\boldsymbol{a+c}$	$\boldsymbol{b+d}$	$\boldsymbol{n=a+b+c+d}$

Fisher demonstrated that the probability of occurrence of the values in Table 2.7 is:

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!} \tag{2.39}$$

Therefore, to assess whether the Null Hypothesis should be rejected or not, the probability of having the observed values of Table 2.7 (i.e. a, b, c and d) is summed by the probability of having most extreme values. In particular, the probability of more extreme values correspond to all the possible numbers of DEGs annotated to Pathway i that are higher than the one observed in Table 2.7, i.e. the cases where $a_{Table\ 2.7} < a < (a+b)_{Table\ 2.7}$. In other terms, if the total number of DEGs is 15 and 12 of them are annotated to the pathway of interest, the most extreme cases correspond to a number of annotated DEGs equal to 13, 14 and 15; the other numbers (b, c and d) can be calculated thanks to the fact that the total number of elements in rows and columns is fixed. In this example, the probability that the values of Table 2.7 or more extreme ones occur is given by:

$$p - value = p_{a=12} + p_{a=13} + p_{a=14} + p_{a=15} \tag{2.40}$$

As in DE Analyses, multiple tests are performed, thus p-values are corrected with Benjamini-Hochberg method; finally, significant terms are selected (i.e. Null Hypothesis is rejected) based on a threshold on the p-value fixed *a priori* by the user.

2.2.7.3 Functional network visualization

The output from an enrichment analysis is often a long list of significant gene sets which is difficult to interpret due to the hierarchical structure of most of the databases containing annotated gene sets. Indeed, esides the main issue, i.e. how accurate is the compilation of the database, often created in an automatic way, other obstacles are the presence of redundant terms and the fact that many genes may be shared among them, especially within a hierarchy. Therefore it is important to highlight only the most relevant ones and to visualize the resulting information to understand the biological processes really implicated in the data.

A powerful software for network visualization is Cytoscape v.3.7.2 (Shannon et al., 2003), an open access software enriched by many user-contributed tools. In particular, ClueGO v2.5.6 (Bindea et al. 2009) integrates Cytoscape with the functionality of enrichment analysis. Many biological databases are interrogated through Cytoscape interface and REACTOME Pathway is one of them, therefore this software is used in this Thesis for both performing the statistical tests and visualizing the final gene network.

In particular, Cytoscape facilitates the results visualization by organizing gene sets into nested networks (or *enrichment maps*), in which nodes represent gene-sets and edges link gene-sets that are biologically related. In particular, enrichment maps are characterized by different levels of meaning: the position of the node in the tree layout represents the node specificity, with the most specific terms as 'leaves' and the most general ones as 'roots'; the symbol dimension of the nodes represents their statistical significance, meaning that most significant terms (i.e. lower p-values) have bigger nodes; different colors allow to visualize different clusters of categories, identified by the overlap of identified genes present (this is independent from the hierarchy of the categories).

Finally, networks can be created by using clueGO, i.e. a Cytoscape App that calculates a statistics, called *kappa-score*, to link the network nodes based on the number of shared genes among nodes themselves. First of all, to calculate the kappa score between two pathways, the number of genes annotated to them is taken into account (as shown in Table 2.8).

Table 2.8 *Number of genes annotated to Pathway 1 and Pathway 2: they are used in clueGO in order to calculate kappa-score statistics, which in turn can be used to create edges among nodes in the enrichment networks.*

		Pathway 2	
		yes	*no*
Pathway 1	*yes*	a	b
	no	c	d

The numbers of annotated genes in Table 2.8 are used to calculate the following parameters:

$$p_0 = \frac{(a+d)}{(a+b+c+d)} \tag{2.41}$$

$$marginal_a = \frac{(a+b)(a+c)}{(a+b+c+d)} \tag{2.42}$$

$$marginal_b = \frac{(c+d)(b+d)}{(a+b+c+d)} \tag{2.43}$$

$$p_e = \frac{(marginal_a + marginal_b)}{(a+b+c+d)} \tag{2.44}$$

Then, kappa-score (k) is calculated as:

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{2.45}$$

The kappa-score statistics is calculated for all the biological terms and these values are collected in a term-term similarity matrix, which is finally used to identify the clusters.

An alternative is to display networks that connect nodes based on the hierarchy of the specific biological database used; in this Thesis, for example, networks are built by using REACTOME hierarchy (an example is provided in Figure 2.19).

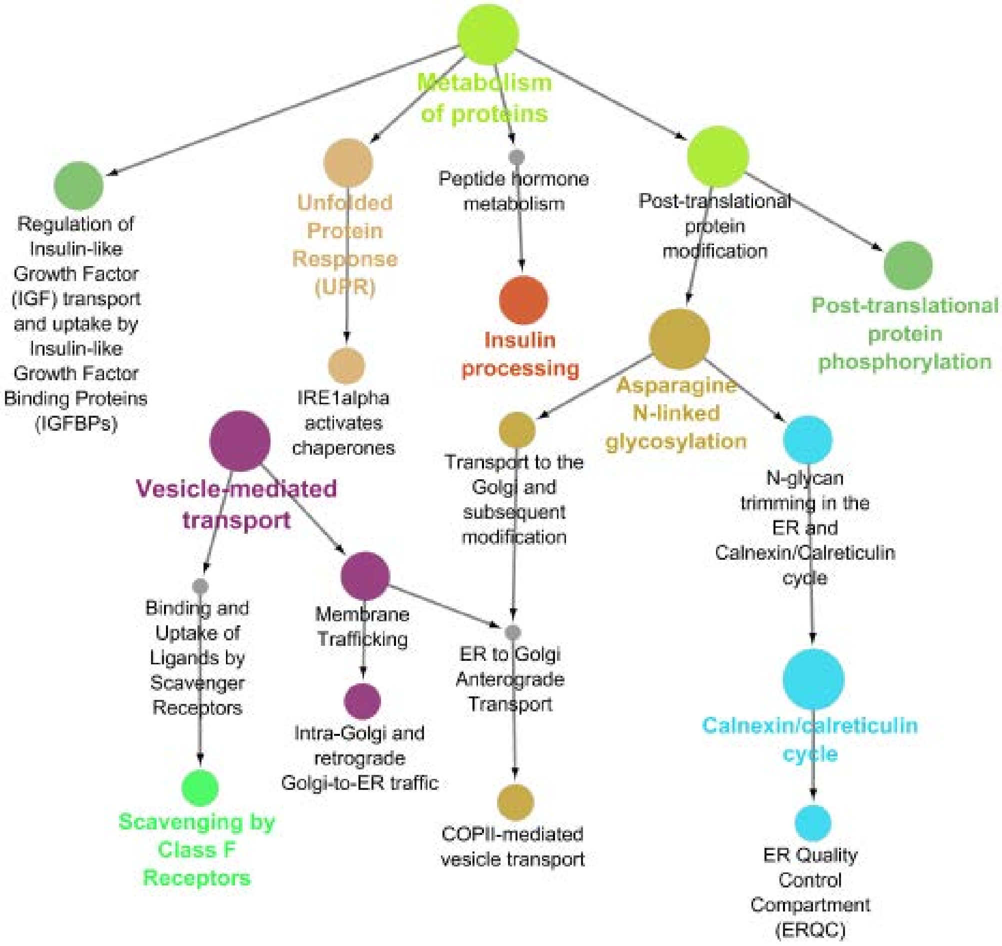

Metabolism of proteins
Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)
Unfolded Protein Response (UPR)
Peptide hormone metabolism
Post-translational protein modification
Post-translational protein phosphorylation
Insulin processing
IRE1alpha activates chaperones
Asparagine N-linked glycosylation
Vesicle-mediated transport
Transport to the Golgi and subsequent modification
N-glycan trimming in the ER and Calnexin/Calreticulin cycle
Binding and Uptake of Ligands by Scavenger Receptors
Membrane Trafficking
ER to Golgi Anterograde Transport
Calnexin/calreticulin cycle
Intra-Golgi and retrograde Golgi-to-ER traffic
Scavenging by Class F Receptors
COPII-mediated vesicle transport
ER Quality Control Compartment (ERQC)

Figure 2.19 *Example of REACTOME network. It shows the enrichment analysis on the PHY transcriptomic data of this Thesis peaking in the resting phase: two major pathways are nested, i.e. metabolism of proteins and vesicle-mediated transport.*

Chapter 3
Transcriptomic data analysis

The aim of this chapter is to show and discuss the analysis of transcriptomic data. First of all, data collected after synchronization through dexamethasone stimuli (i.e. DEX protocol) and feeding-fasting stimuli (i.e. PHY protocol) are visualized by means of an exploratory analysis. Then, the PCA model is used to select circadian transcripts in both protocols, whose dynamics is characterized further by comparing their peak phases and amplitudes. However, also transcripts having flat profiles may carry out important differences in biological functions in human liver cells under the two synchronization conditions; in this case, parameters like relative radius, amplitude or phase have little meaning, so they are studied through another statistical technique: differential expression analysis. Finally, after characterizing the dynamics of circadian and flat profiles in both protocols, they are interpreted from the biological point of view thanks to an enrichment analysis.

3.1. Exploratory analysis

The high-throughput RNA-Sequencing allowed to measure mRNAs over 6 time points for both DEX and PHY protocols: 0 h, 4 h, 8 h, 12 h, 16 h and 20 h with 4 replicates each. After normalization and filtering, 12940 transcripts are retained for further analyses; moreover, two types of normalized datasets are provided: one with linear transcriptomic data and another one with logarithmic data. In particular, the former type is useful when dynamic profiles need to be characterized, while the latter one has as one of the main advantages the reduction of the number of orders of magnitude spanning the data. To make DEX and PHY datasets comparable for the preliminary analysis, logarithmic data have been scaled just by subtracting the mean of each transcript across time points, while linear data have been scaled by means of Standard Normal Variate (i.e., for each transcript, subtraction of the mean and division by the standard deviation). Then, scaled data are analysed through an explorative PCA in order to assess similarities among samples of the same protocol and differences between the two protocols themselves (Figure 3.1).

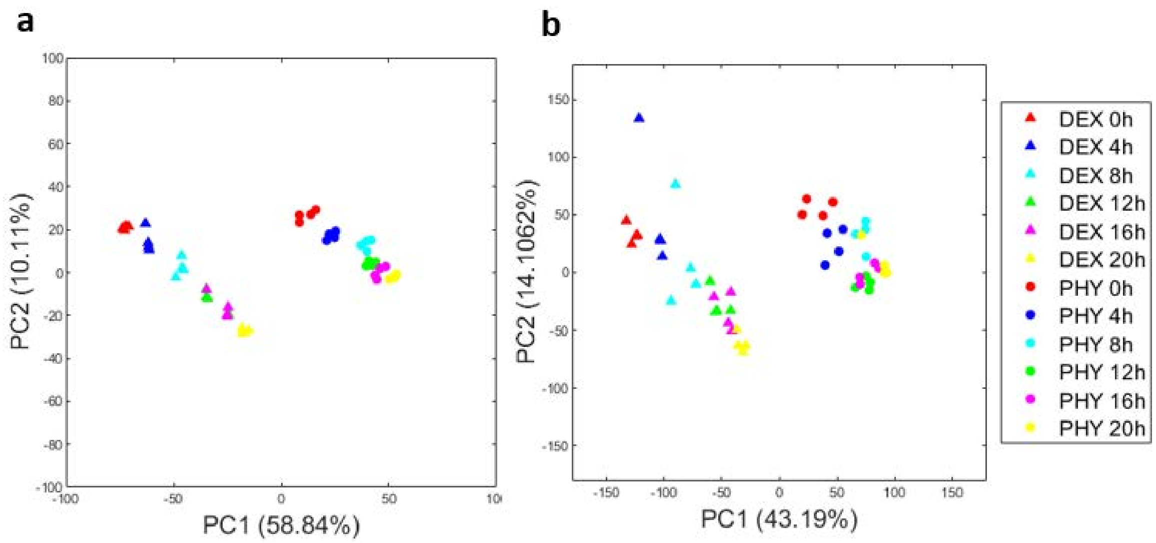


Figure 3.1 *PCA analysis of normalized and filtered transcriptomics data: a) refers to log2 data that are mean centered; b) refers to linear data that are scaled through Standard Normal Variate.*

In both cases, samples from the same protocol, DEX or PHY, are clearly separated. Within each of the two protocols, variability of samples within the same time point is lower than that between different time points. Thus, data are accurate enough to follow temporal changes of transcription within the 20 hours. In particular, DEX and PHY samples are anti-correlated with respect to the first principal component, which is the one that explains almost half of the total variability (li59 % in case of logarithmic data, 43% in case of linear data). Interestingly, not only there is an ordered sequence of time points from 0 h to 20 h, but also the four replicates of DEX and those of PHY corresponding to the same time point are positively correlated with respect to the second principal component. These results are consistent and independent from the normalization method used.

3.2. Dynamics characterization

The first objective of this section is the identification of circadian profiles in liver transcriptome induced by hormonal stimuli (i.e. DEX protocol) and by feeding-fasting stimuli (i.e. PHY protocol), by using the PCA model with a minimum relative radius of 0.7 for the circadian profiles selection (Figure 3.2).

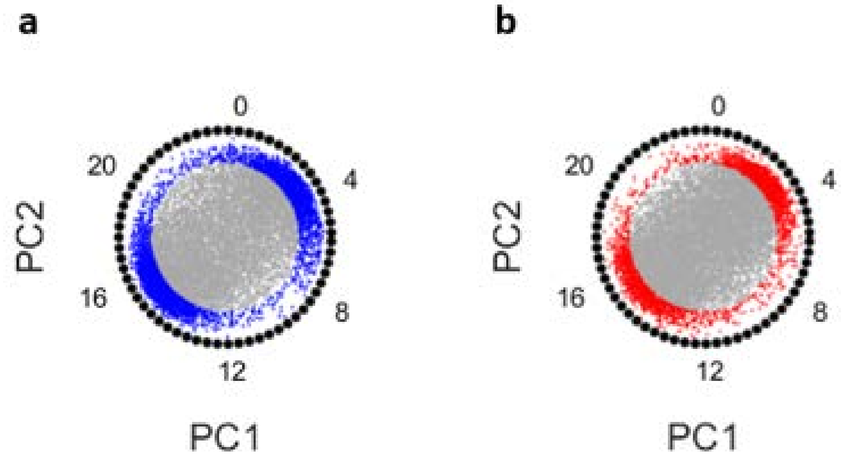


Figure 3.2 *PCA score plots of the barycentres of DEX profiles (a) and PHY profiles (b). Grey dots refer to not circadian profiles, while blue and red dots refer to DEX and PHY circadian profiles, respectively*

The results of the calibrated PCA model show that the dexamethasone synchronization is able to induce a higher percentage of circadian transcripts than the physiological one, even though the two percentages have the same order of magnitude:

- 4570 circadian transcripts in DEX protocol (35.32% with respect to the total number of detected mRNAs);
- 2812 circadian transcripts in PHY protocol (21.73% with respect to the total).

Although the PCA model had been previously tested after its development, it is important to evaluate its performance in selecting circadian profiles also on the experimental data we used. However, a direct visualization of each single dynamics is unfeasible due to the high dimension of the dataset, therefore another strategy is employed: heatmaps that order expression profiles based on the peak phases estimated by the model itself (for details, see Chapter 2, section 2.2.5). In particular, Figures 3.3 ad 3.4 show heatmaps for different transcripts lists: the former one refers to circadian profiles in both DEX and PHY protocols, while the latter one refers to profiles that are not circadian in either protocol.

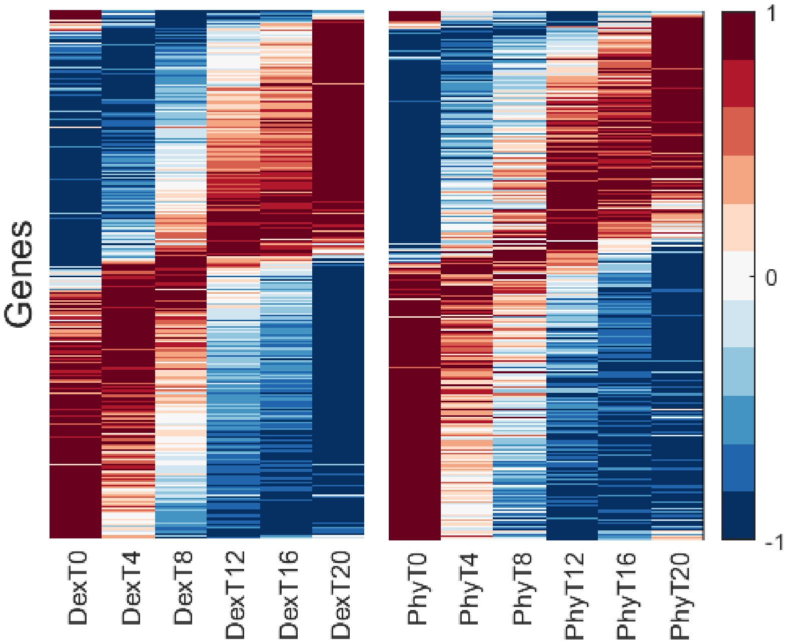


Figure 3.3 *Heatmaps of transcripts profiles that are circadian both in DEX and PHY. DEX (left) and PHY (right) linear data are scaled between -1 and 1 and plotted separately. The order of genes is different between left and right heatmap, because it is based on the phases of DEX and PHY protocols, respectively. Color bar indicates the level of gene expression scaled between -1 and 1.*

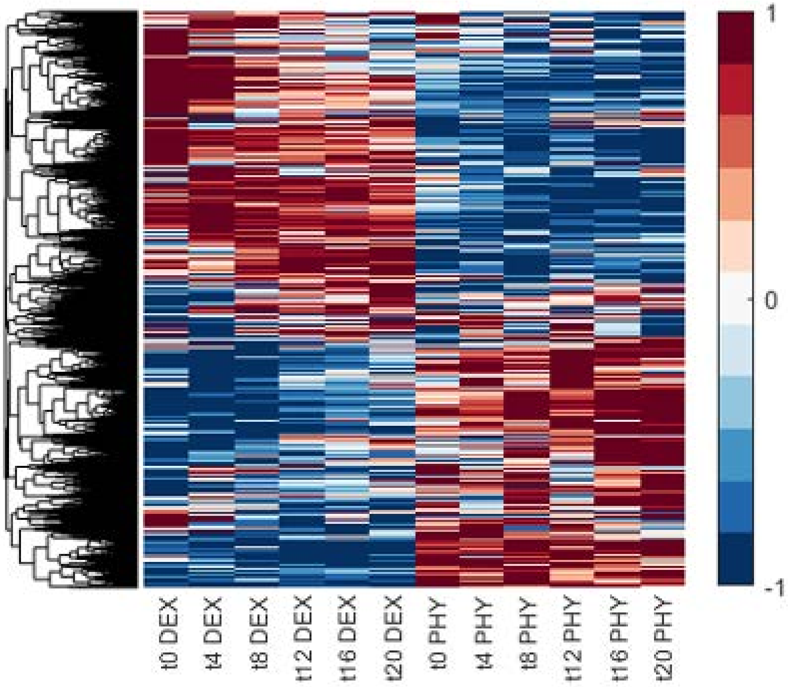


Figure 3. 4 *Heatmaps of transcripts profiles that are not circadian in either protocol. In this case, data are scaled between -1 and 1 by using both DEX and PHY data, but instead of being ordered with respect to phases (which have little meaning with non-circadian profiles), they are ordered with hierarchical clustering based on profiles similarity, setting Euclidean distance and complete linkage. The dendrogram obtained through clustering is shown on the left, the meaning of colors/color intensities is explained by the legend on the right.*

These results give confidence on the accuracy of the model-selected and excluded (i.e. circadian and non-circadian) genes, respectively. Indeed, in Figure 3.3 it is evident that the ordered profiles are up-regulated around the sampling time indicated by the peak phases and then they are down-regulated, while in Figure 3.4 there are no profiles showing a specific peak during the 20 hours. However, two main groups of data can be identified in this case: one with constantly higher values in PHY than in DEX protocol and one with constantly higher values in DEX. This trend is analysed further thanks to a Differential Expression Analysis (section 3.2.2).

3.2.1. Circadian profiles

Based on the model estimation of relative radii, circadian profiles in DEX and PHY are selected and compared, in order to understand if the liver circadian clock acts in a similar way under hormonal or feeding synchronization. First of all, it is interesting to determine whether the genes that are expressed rhythmically in the two protocols are the same or not (Figure 3.5); then, circadian profiles are characterized in terms of peak phases and amplitudes (section 3.2.1.1 and 3.2.1.2).

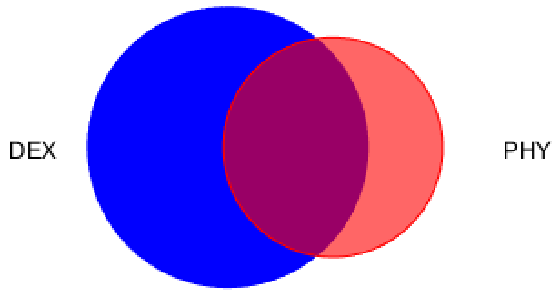


Figure 3.5 *Venn diagram of the transcriptomics dataset: blue area represents the number of DEX circadian mRNAs, while the red circle represents PHY circadian mRNAs.*

The Venn diagram (Figure 3.5) shows that the list of circadian transcripts in PHY is not a subset of that of DEX: indeed, there is an intersection between the two, but there is also a significant number of genes that are periodic exclusively in one protocol (especially in case of DEX data).

3.2.1.1 Circadian profiles: peak phases

The aim of this section is to compare DEX and PHY dynamics in terms of peak phases, which indicates the time of the day when the transcript abundance is higher. However, a higher mRNA abundance does not mean necessarily that the corresponding biological function is already active. Indeed, the biological activity of protein-coding genes is carried out by their final product, i.e. the encoded protein, whose synthesis may occur with some delay (see Chapter 5).

However, studying these genes at mRNA level still gives a direction towards specific biological functions.

The phases are first visualized as 3D histograms considering circadian transcripts for DEX and PHY protocol (Figure 3.6) in an independent manner; then, DEX and PHY phases are compared considering circadian profiles in both protocols or only in DEX (Figure 3.7 a) and circadian profiles in both protocols or only in PHY (Figure 3.7 b).

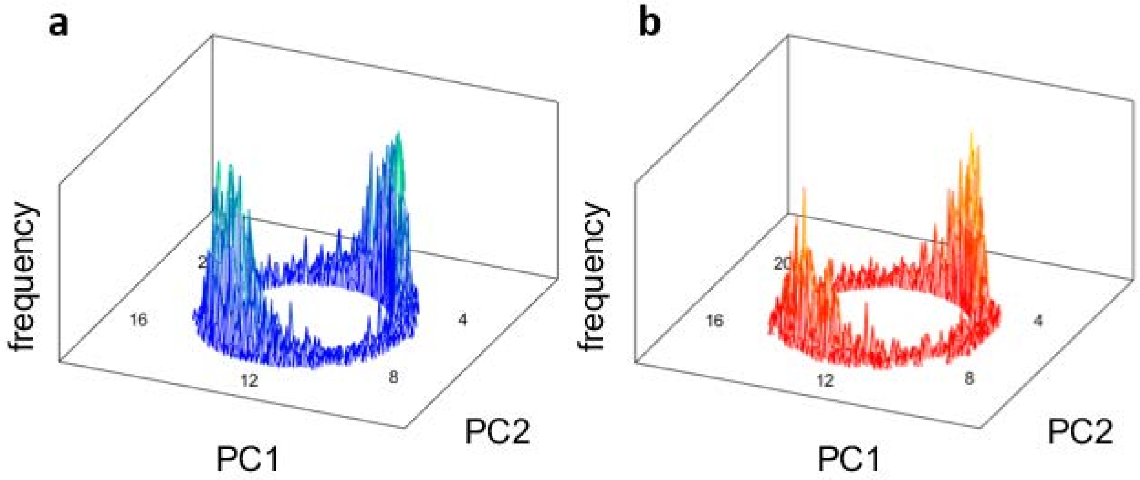


Figure 3.6 *3D histograms representing peak phases of circadian transcripts in DEX (a) and PHY (b) protocol. The reference for calculating phases is the projection of profile barycentres on the 2D-PCA score plot (as explained in Chapter 2, section 2.2.4).*

Histograms in Figure 3.6 show that in both protocols there are two main peaks of transcriptional activity: around 3 h and around 15 h. This suggests a potentially clear distinction between mRNAs peaking in the active and resting phase, which in turn is useful for distinguishing among biological functions carried out by circadian transcripts in the two phases (section 3.4). While the definition of active and resting phase is not straightforward in an *in vitro* system, section 3.3 more thoroughly discusses how to relate *in vitro* time with *in vivo* diurnal or nocturnal activities.

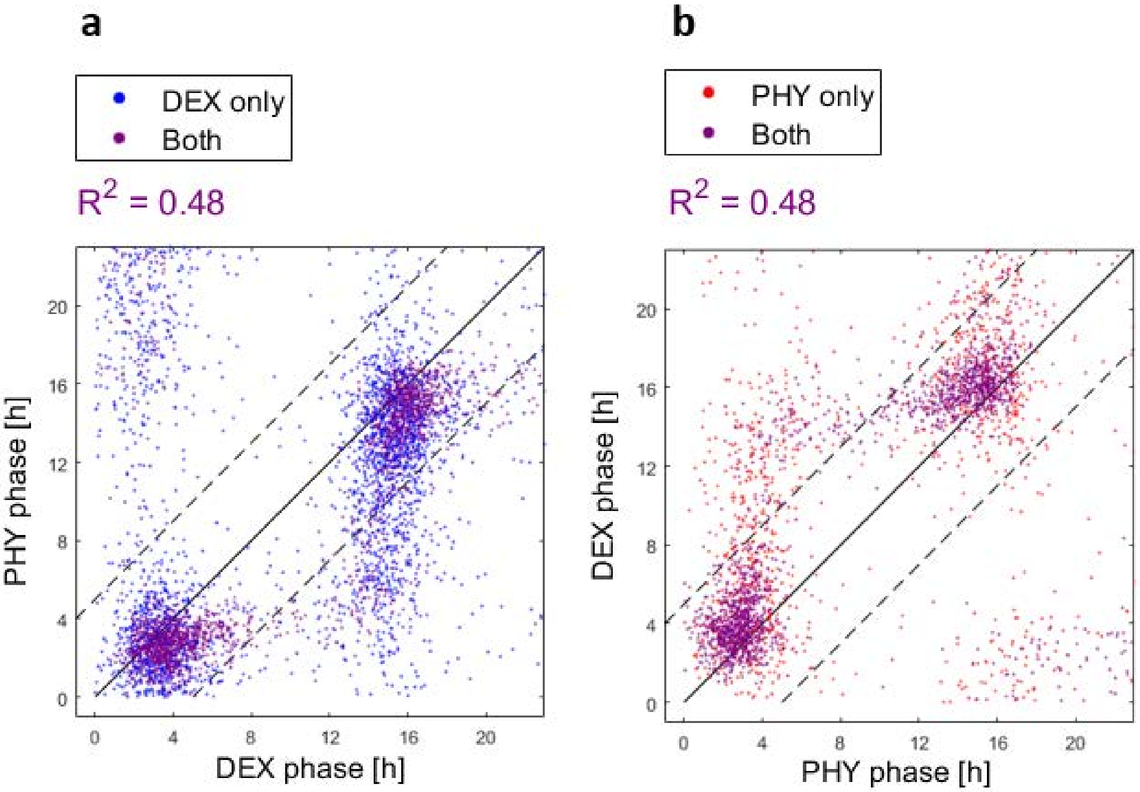


Figure 3.7 *Comparison of phases between the two protocols. Each dot represents a measured transcript with circadian profile in the conditions indicated in the legend. Solid line indicates exact phase equality, dotted lines indicate a deviation of 5h from equality. The R2 is calculated considering only profiles that are circadian in both conditions.*

The R^2 calculated with transcripts that are circadian in both protocols is relatively low (0.48), meaning that there is not a high correlation between the phases of the two protocols. However, it can be noticed from Figure 3.7 that the majority of profiles have peak phases that differ for less than 5 h between the two protocols (i.e., the majority of points lie within the two dotted diagonals).

3.2.1.2 Circadian profiles: amplitudes

Analogous plots are made to compare amplitudes between the two protocols: Figure 3.8 a represents amplitude values for transcripts that are circadian in both DEX and PHY (purple dots) and only in DEX (blue dots), while Figure 3.8 b represents amplitude values for transcripts circadian in both (purple dots) or only in PHY protocol (red dots).

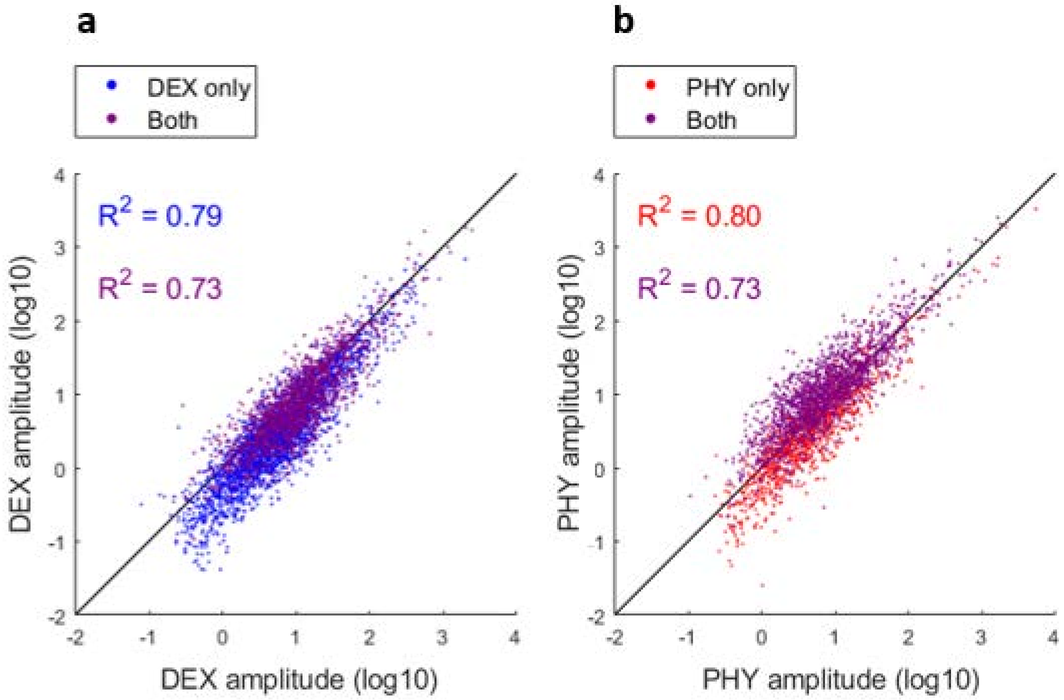


Figure 3.8 *Comparison of the amplitudes (in logarithmic scale) of circadian transcripts of the two protocols. Dots represent log10 amplitudes estimated for the profiles that are circadian in the protocol(s) indicated by the legend, while the diagonal represent amplitudes that are equal between DEX and PHY protocol. R2 of the amplitudes of DEX and PHY profiles is calculated considering profiles that are circadian only in DEX (Figure 3.8 a), only in PHY (Figure 3.8 b) or in both protocols (Figure 3.8 a and b).*

These results differ from those of peak phases: indeed, almost all points are distributed around the diagonal, meaning that there is a strong positive correlation between amplitudes of DEX and amplitudes of PHY profiles. This is proved by the relatively high R^2 values for all three comparisons: 0.73 when profiles are circadian in both protocols, 0.79 for profiles that are circadian only in DEX and 0.80 for profiles that are circadian only in PHY. Therefore, DEX and PHY synchronization conditions seem to have a higher impact on circadian transcript phase rather than on their amplitude.

3.2.2. Not circadian profiles

The dynamics analysis of paragraph 3.2.1 employed parameters, e.g. relative radius, phase and amplitude, that are suitable for the description of circadian profiles only.

However, there might be transcripts that carry out essential biological functions for the liver physiology even without displaying a circadian rhythmicity and that are different between the two protocols due to the perturbation itself; thus they need different analysis strategies to be characterized.

For example, if heatmaps are built with circadian transcript only in DEX (Figure 3.9 a) or only in PHY protocol (Figure 3.9 b), a general trend can be noticed. In the former case, when DEX profiles have peaks between 0h and 8 h, PHY profiles tend to take on constantly low values, while when DEX profiles peak between 12 h and 20 h, PHY profiles tend to take on constantly high values. In contrast, when profiles are circadian only in PHY protocol, the qualitative trend is less clear and requires a more detailed analysis to select genes with different behavior.

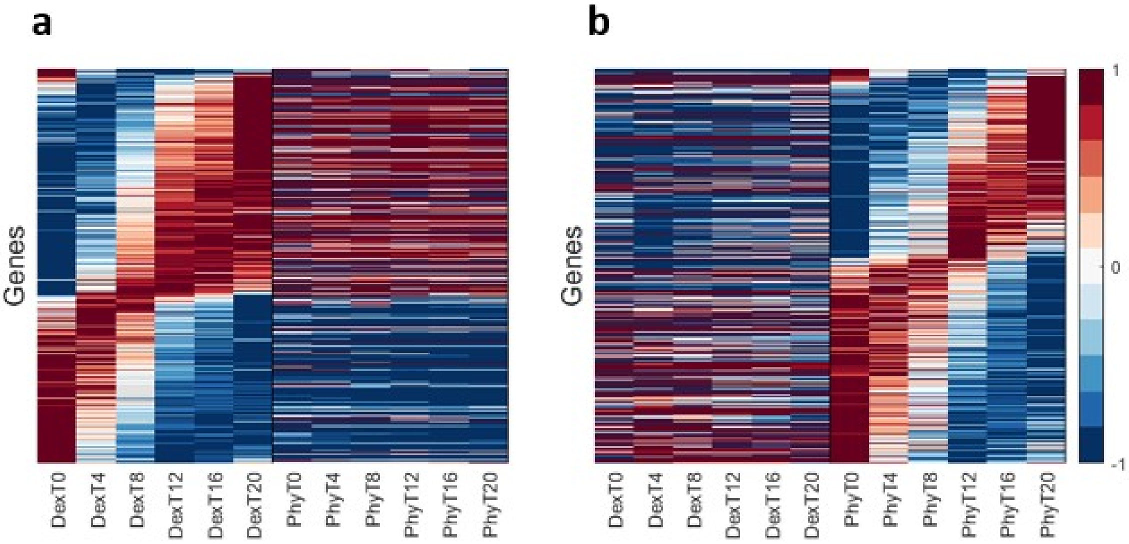


Figure 3.9 *Heatmaps considering transcripts profiles that are circadian only in DEX (a) or only in PHY (b) protocol. In both cases, the profiles matrix is scaled between -1 and 1 considering DEX and PHY data for a and b, respectively, and the color and/or color intensity represent the scaled value as indicated by the legend. Finally, profiles are sorted with respect to the phases estimated for DEX and PHY profiles in a and b, respectively.*

To rigorously confirm this qualitative observation and capture other potential differences between the two conditions, a Differential Expression (DE) analysis is performed.

3.2.2.1 Differential Expression Analysis set-up

In order to select differentially expressed genes (‘DEGs’) between DEX and PHY protocols, two different methods are adopted. The former allows to perform pairwise comparisons between DEX and PHY expression values at each time point independently, while the latter compares pairs of ordered profiles directly, thus taking advantage of the correlation between subsequent time points in order to improve its performance (for details, see Chapter 2, section 2.2.6). They are implemented in two R packages, edgeR and FunPat, respectively.

In the first method, edgeR package requires an input matrix given by raw counts of time points from 0 h to 20 h, 4 replicates each; then, after specifying the design matrix and estimating the dispersion, differentially expressed transcripts are identified by setting a threshold fold change of 2 and a FDR of 0.05. Finally, profiles are selected as DE when they are DE in at least one time point.

As regards FunPat package, three matrices need to be used as input: one representing DEX measurements, one representing PHY measurements and a third one (‘replicates matrix’) that allows to build the error model. In contrast to edgeR, FunPat need normalized data as input in order to eliminate batch effects and, therefore, to model the error variability correctly. Moreover, as suggested by the authors, the two matrices representing DEX and PHY measurements are created by calculating the mean of the corresponding 4 replicates at each time point: indeed, mean expression values ensure a more robust performance of the model, which is less influenced by potential measurements errors at a certain time point.

Instead, there is no precise information on how to build the replicates matrix, just some peculiar possibilities are illustrated; however, the basic principle is to use data that may differ only because of measurements errors and not due to different treatment conditions. Therefore, the experimental setup (Chapter 2, section 2.1) is taken into account in order to model noise variability: the main assumption that guides the creation of the replicates matrix is that the four replicates of the same time point, within the same protocol, differ among each other only because of measurement errors. Therefore, DEX and PHY replicates at each time points are compared (Table 3.1) to build the error model, in order to avoid confounding between this random error variability among equivalent samples with statistically different expression values.

Table 3.1 *Scheme of the pairwise comparison for creating the two columns, i.e. A and B, of the replicate matrix of FunPat. This example shows only the comparisons at time 0 h in DEX protocol, but all time points and all protocols are used in the algorithm.*

Protocol	Time sample	Replicate (A)	Replicate (B)
DEX	0 h	1	2
		1	3
		1	4
		2	3
		2	4
		3	4

Finally, DEGs are selected with FunPat when an adjusted p-value lower than 0.01 is found.

3.2.2.2 Differential Expression Analysis results

The results of edgeR and FunPat analyses are compared in Figure 3.10 as Venn diagrams representing the candidate lists of DEGs.

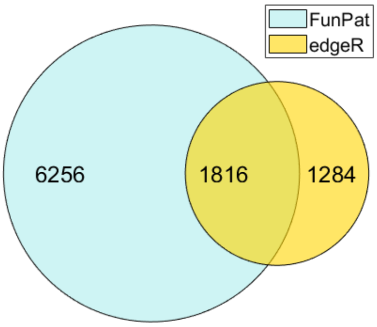


Figure 3.10 *Differential Expression Analysis results. This Venn diagram represents the number of transcripts that are selected as Differentially Expressed by FunPat and/or by edgeR (indicated by the colors in the legend).*

The list of candidate DEGs selected by FunPat is more than double than the one selected by edgeR: 8072 with FunPat and 3100 with EdgeR. As proved in the review of Spies et al. (2017), intersecting candidate lists from different tools is useful for extracting real differentially expressed genes, since true positives tend to lie within the intersection of lists, while false positives tend to fall among the genes that are uniquely identified by a single tool. Therefore, the next analyses of DEGs take into account of only the 1816 genes in the intersection.

To verify the trend that was qualitatively observed in Figure 3.9, heatmaps are built with genes that are circadian only in one protocol and identified as DEGs (Figure 3.11 a and b).

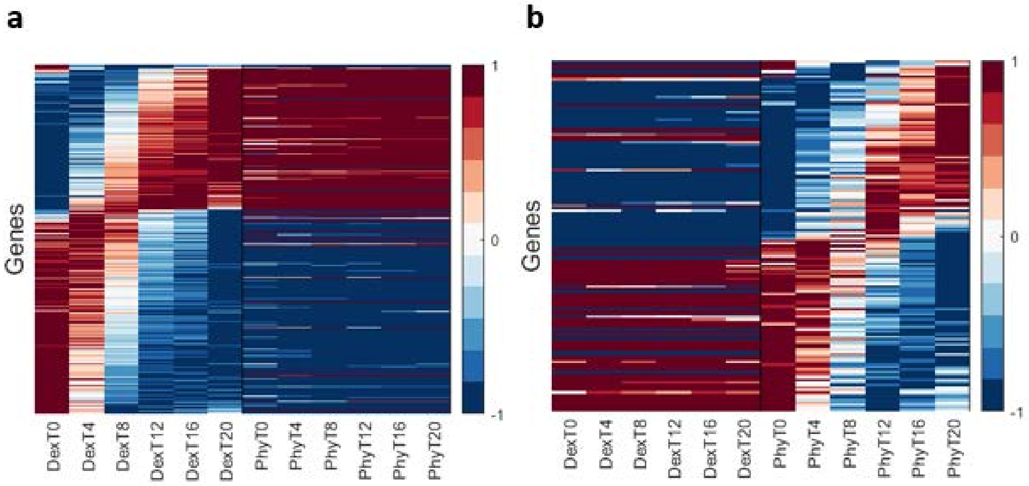


Figure 3.11 *Heatmaps of DEGs that are circadian in only one protocol: DEX and PHY protocol in figure a and b, respectively. Data are scaled considering DEX and PHY measurements, respectively, in order to take on values between -1 and 1. Different colors and/or different color intensities indicate different numerical values as shown in the legend. Finally, profiles are ordered with respect to the phase calculated for the circadian protocol.*

The percentage of DEGs profiles circadian only in DEX protocol (Figure 3.11 a) with respect to the total number of profiles that are circadian only in that protocol is 19.7%, while in case of PHY it is one order of magnitude smaller, i.e. 4.9%. Moreover, the observed trend of constantly low or high values in PHY protocol when DEX profiles are circadian and peaking in 0-8 h or in 12-20 h interval, respectively, is confirmed when DEGs are considered and it is even more marked. As regards transcripts profiles that are circadian only in PHY, the visualization of only DEGs highlights that the majority of these DEX non-circadian profiles also tend to follow a similar behavior, taking on constantly high values or low values when PHY transcripts peak in 0 h-8 h or 12 h-20 h interval, respectively. Finally, the percentage of non circadian profiles in either protocol that are selected as DEGs is 8.9%; they are visualized through a heatmap in Figure 3.12.

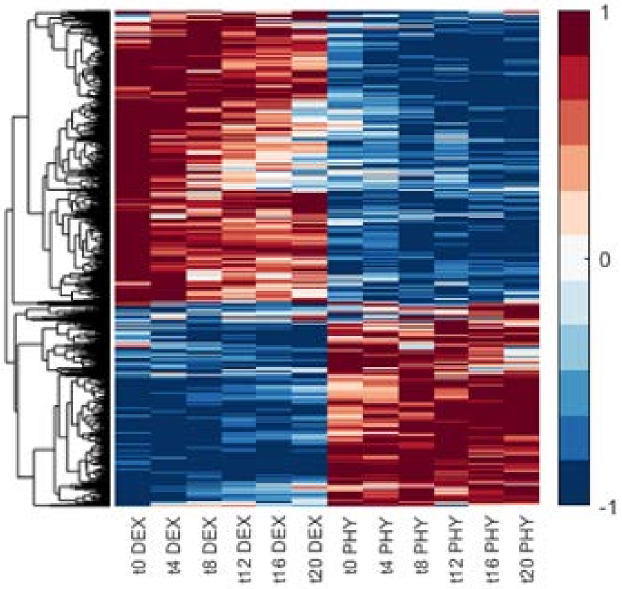


Figure 3.12 *Heatmap of the 645 DEGs selected by both edgeR and FunPat that are not circadian in either protocol. Each row is scaled considering both DEX and PHY measurements across 20 h (corresponding to the same Gene Symbol) in order to take on values between -1 and 1. In the abscissa, the variable names are specified (i.e. the sampling times); on the left, a dendrogram is built with hierarchical clustering (Euclidean distance, complete linkage).*

The non circadian DEGs heatmap confirms the presence of two main groups of data (which is even more evident than in section 3.2.1): one with constantly lower values in DEX than in PHY protocol and another with constantly higher values in DEX than in PHY protocol. This result suggests to perform separated functional analyses for the two groups, in order to understand the main biological functions that are carried out mainly by DEX and PHY non circadian transcripts throughout the day.

In conclusion, the two synchronization protocols affect the identity of the circadian genes that oscillate in the two conditions, but also for transcripts that oscillate in both the phase of peaking may be different; instead it has a minor impact on the amplitude of oscillation. Moreover, also non-oscillating genes or oscillating only in one condition are impacted by showing different levels of expression. Finally, sections 3.3 and 3.4 analyse the functional biological implications of these phenomenological differences.

3.3 The active phase and the resting phase

First, given the experimental data used in this work are from an *in vitro* culture system, it is essential to understand how results relate to normal definitions in human physiology of ‘day’, or ‘active phase’, and ‘night’ or ‘resting phase’. Moreover, this is important also to compare our results with literature (especially in Chapter 4), where most studies are performed in mouse models: indeed, mouse are nocturnal animals, therefore physiologic functions of the active phase are performed during the day in humans, but during the night in mouse. For this reason, from this point on the terminology 'active phase' and 'resting phase' refers to day and night in humans, and night and day in mouse, respectively.

Experimental data from the physiological protocol (PHY) mimic the active phase with a feeding regime (high insulin and glucose) and the resting one with a fasting regime (high glucagon and low glucose). In DEX protocol, a stimulus of dexamethasone is given to the cells; the physiologic analogue of Dexamethasone is cortisol, a hormone secreted *in vivo* in the second half of the night, reaching a peak in the early morning (Tsigos et al., 2002). Therefore, in both protocols, the interval between 0 h and 8 h is expected to be the ‘active phase’, while the one between 12 h and 20 h to be the ‘resting phase’ (Figure 3.13).

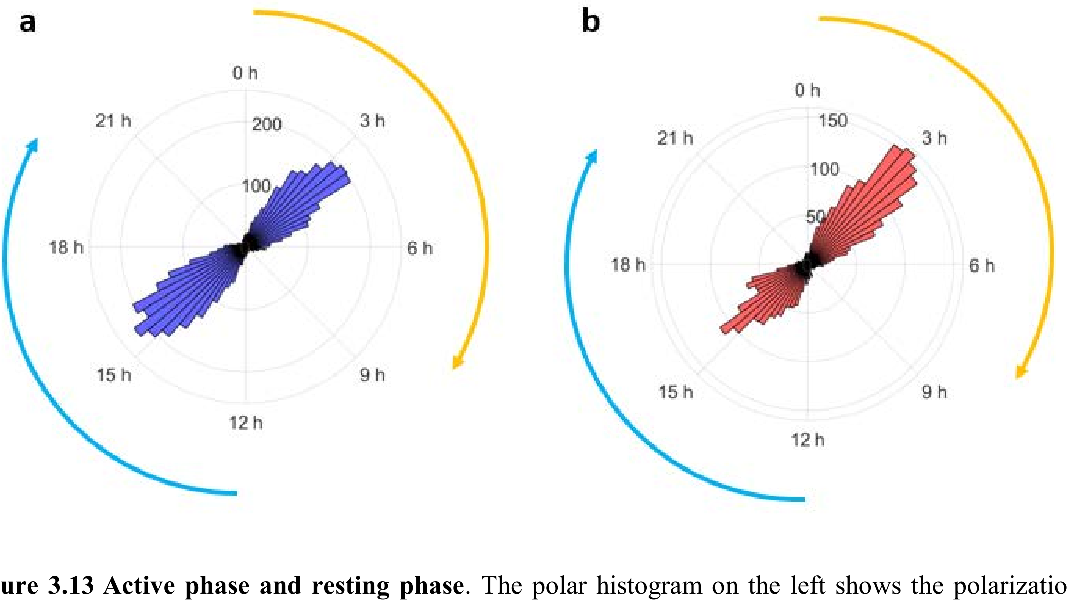


Figure 3.13 Active phase and resting phase. The polar histogram on the left shows the polarization of peak phases in circadian transcripts of DEX protocol, while the one of the right is similar but refers to the PHY protocol. Moreover, orange arrows (0 h - 8 h) indicate the active phase and light blue arrows indicate resting phase (12 h - 20 h).

Moreover, a polar histogram is shown in Figure 3.14 in order to visualize the difference of peak phases of DEX and PHY profiles that are circadian in both protocols (in total, 1714 transcripts): interestingly, even though the two protocols have different synchronization efficiencies, when transcripts oscillate in both they tend to have very similar peak phases, with shifts between 0 h and 3 h.

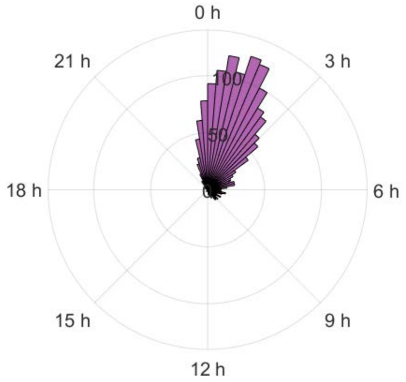


Figure 3.14 *Circadian mRNAs in both protocols: this polar histogram shows the difference of peak phases between DEX protocol and PHY protocol, when mRNAs are circadian in both.*

As explained in section 1.2.1, CLOCK and BMAL1 activate the transcription of their target genes by binding their E-box or D-box binding sequences, increasing their transcription during the active phase. Even if CLOCK and BMAL1 bind only a subset of genes having E-Box and D-Box, there should be a higher amount of these genes peaking during the active phase, and we checked it to verify the results are consistent with the definition of active and resting phases above.

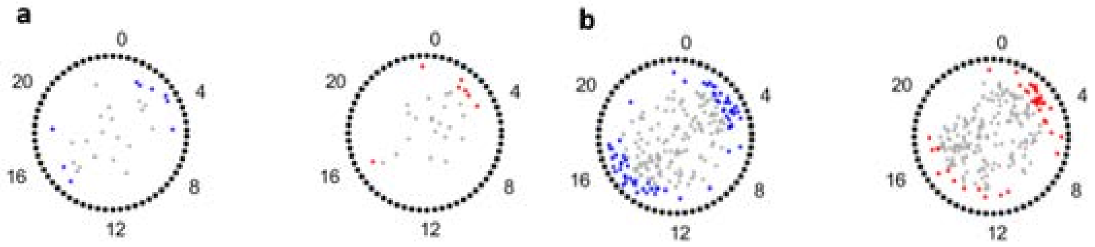


Figure 3.15 *PCA score plot representing the barycentres of genes annotated to D-BOX (a) and E-BOX(b) and belonging to DEX or PHY protocol. Blue dots represent DEX circadian profiles, red dots represent PHY circadian profiles, grey dots represent non circadian profiles, while the external black dots are the scores of the periodic profiles used during the calibration of the PCA model.*

The score plots in Figure 3.15 representing DEX profiles do not help defining the active and resting phases, because circadian profiles have phases that are evenly distributed between the two time intervals. On the contrary, the physiological synchronization makes the majority of circadian transcripts peak between 0 h and 8 h, thus the initial assumption that this interval corresponds to ‘active phase’ seems to be confirmed.

We further confirmed this looking at PER2 mRNA, which is one of the most important E-Box-containing circadian gene, and ARNTL (also known as BMAL1) whose protein dimerizes with CLOCK and should be up-regulated during the resting phase (Figure 3.16). Both PER2 and ARNTL oscillate only in PHY data, peaking at the beginning of the active phase and of the resting phase, respectively. Thus, in PHY protocol the intervals 0-8h and 12-20h can be strongly associated with the physiological active and resting phases; more uncertainty is present in the DEX protocol, potentially due to a sub-optimal cell entrainment by dexamethasone.

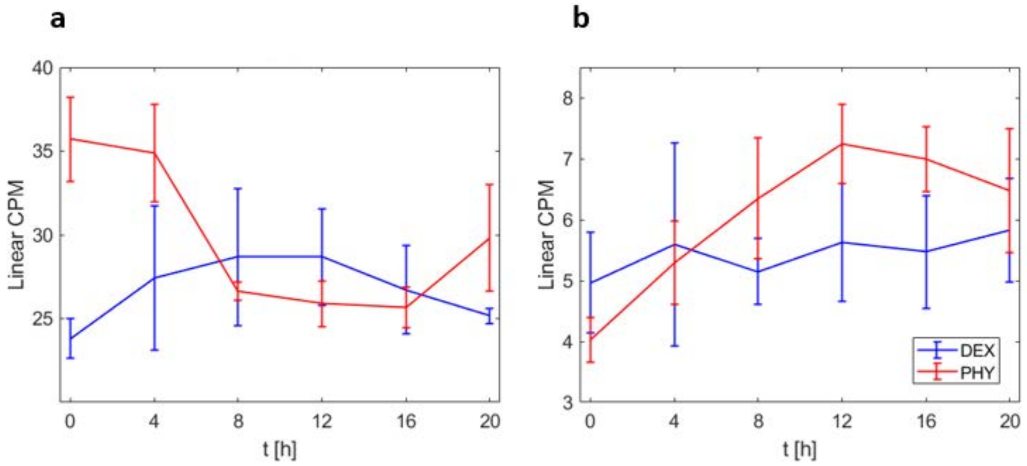


Figure 3.16 *Mean profiles of PER2 (a) and ARNTL (b), also known as BMAL1, expressed as CPM in linear scale. Blue lines refer to DEX protocol, red lines refer to PHY protocol and in both cases the vertical segments represent the standard deviation calculated at each time point with the four replicates.*

3.4. Biological interpretation

As explained in section 3.2, the majority of peak phases is distributed into two time intervals: between 0 h and 8 h and between 12 h and 20 h. Since peak phases correspond to the time when transcripts exercise their function (even though one must remember that proteins are the real 'actors' within a cell, therefore a delay is inevitable between mRNAs synthesis and their biological function), the enrichment analysis is performed for the two time intervals separately. This allows to define the major pathways during the 'active' phase and during the 'resting' phase, which is useful especially for organs like liver that have a change of functions throughout the 24 hours.

First, circadian profiles are analysed as shown in Table 3.2.

Table 3.2 *Groups of circadian transcripts used for the enrichment analysis.*

Circadian protocol	Peak phase
DEX	0-8h
DEX	12-20h
PHY	0-8h
PHY	12-20h

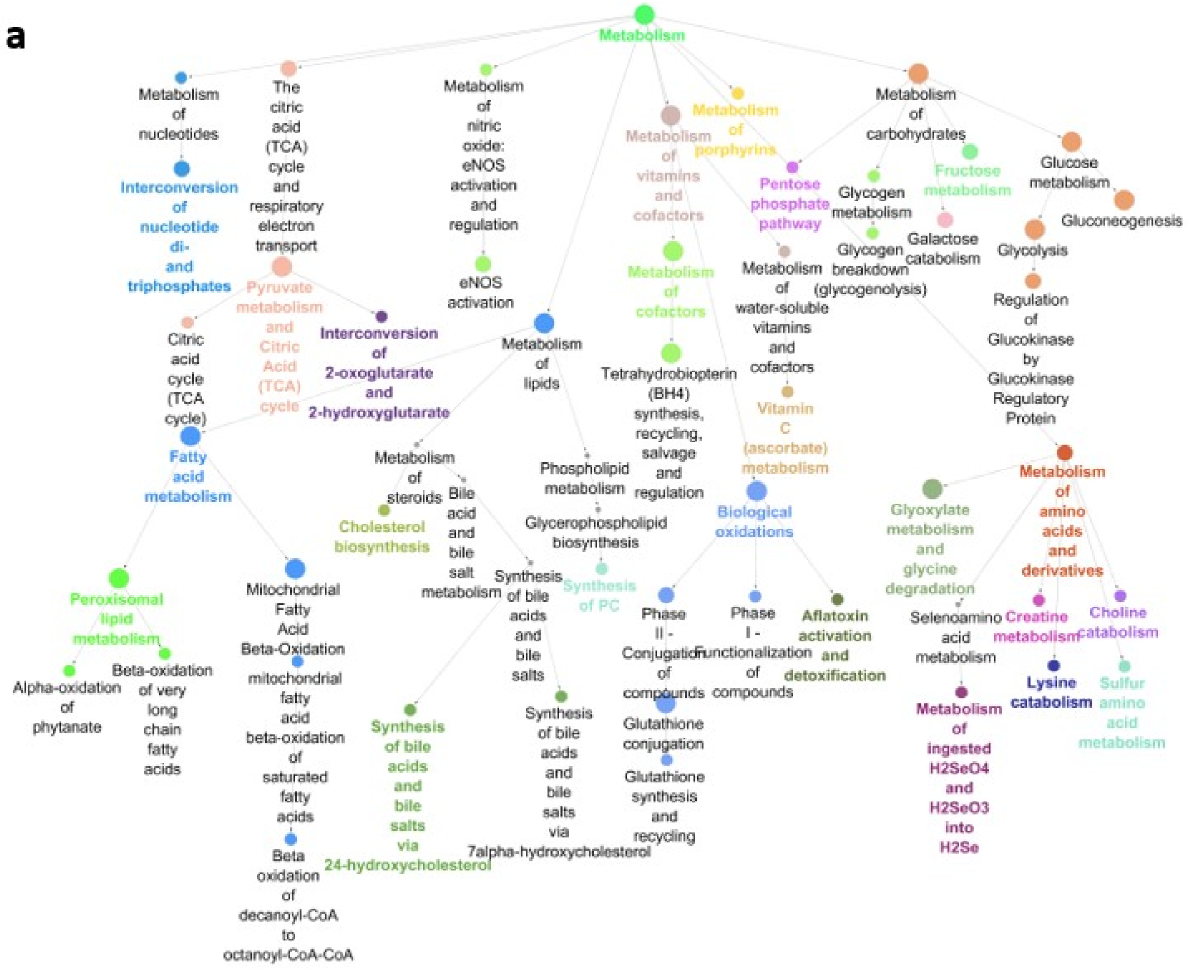
Then, the biological function of non circadian profiles is analysed: to do so, Differentially Expressed Genes (DEGs) are taken into account, by distinguishing between those with constantly lower values in DEX than in PHY from those having constantly higher values in DEX than in PHY protocol (two main clusters in Figure 3.12).

The tool used to analyze these gene sets is the enrichment analysis using REACTOME as biological database, i.e. a curated collection of biological pathways. The hierarchical structure of Reactome distinguishes between more general terms (the ‘roots’ of the hierarchical tree) and more specific ones (the ‘leaves’ of the tree). Within the hierarchy, clusters of nodes sharing many genes (according to the so-called kappa-score statistics) are identified by different colors. Only nodes with a FDR-corrected p-value lower than 0.05 are displayed (for details, see Chapter 2, section 2.2.7).

3.4.1 Enrichment analysis of circadian profiles

Gene sets of circadian profiles from the four lists mentioned in Table 3.2 are analysed by enrichment analysis. Because of the high number of identified pathways in the results, a sub-selection of these is presented here. The retained pathways are selected based on their biological relevance for liver biology (see Chapter 1) and correspond to the following main branches in Reactome hierarchy: Metabolism, Metabolism of Proteins, Vesicle-mediated transport, Immune system, Signal Transduction, Cellular response to external stimuli.First, the graphical results of the four lists are reported, then follows an overall discussion of these data.

3.4.1.1 DEX circadian profiles peaking in the active phase


a
Metabolism
Metabolism of nucleotides
Interconversion of nucleotide di- and triphosphates
The citric acid (TCA) cycle and respiratory electron transport
Pyruvate metabolism and Citric Acid (TCA) cycle
Citric acid cycle (TCA cycle)
Interconversion of 2-oxoglutarate and 2-hydroxyglutarate
Metabolism of nitric oxide: eNOS activation and regulation
eNOS activation
Metabolism of vitamins and cofactors
Metabolism of porphyrins
Metabolism of cofactors
Tetrahydrobiopterin (BH4) synthesis, recycling, salvage and regulation
Metabolism of water-soluble vitamins and cofactors
Vitamin C (ascorbate) metabolism
Pentose phosphate pathway
Metabolism of carbohydrates
Glycogen metabolism
Fructose metabolism
Glycogen breakdown (glycogenolysis)
Galactose catabolism
Glucose metabolism
Gluconeogenesis
Glycolysis
Regulation of Glucokinase by Glucokinase Regulatory Protein
Metabolism of lipids
Fatty acid metabolism
Peroxisomal lipid metabolism
Alpha-oxidation of phytanate
Beta-oxidation of very long chain fatty acids
Mitochondrial Fatty Acid Beta-Oxidation
mitochondrial fatty acid beta-oxidation of saturated fatty acids
Beta oxidation of decanoyl-CoA to octanoyl-CoA-CoA
Metabolism of steroids
Cholesterol biosynthesis
Bile acid and bile salt metabolism
Synthesis of bile acids and bile salts
Synthesis of bile acids and bile salts via 24-hydroxycholesterol
Synthesis of bile acids and bile salts via 7alpha-hydroxycholesterol
Phospholipid metabolism
Glycerophospholipid biosynthesis
Synthesis of PC
Biological oxidations
Phase II - Conjugation of compounds
Glutathione conjugation
Glutathione synthesis and recycling
Phase I - Functionalization of compounds
Aflatoxin activation and detoxification
Glyoxylate metabolism and glycine degradation
Metabolism of amino acids and derivatives
Selenoamino acid metabolism
Metabolism of ingested H2SeO4 and H2SeO3 into H2Se
Creatine metabolism
Choline catabolism
Lysine catabolism
Sulfur amino acid metabolism

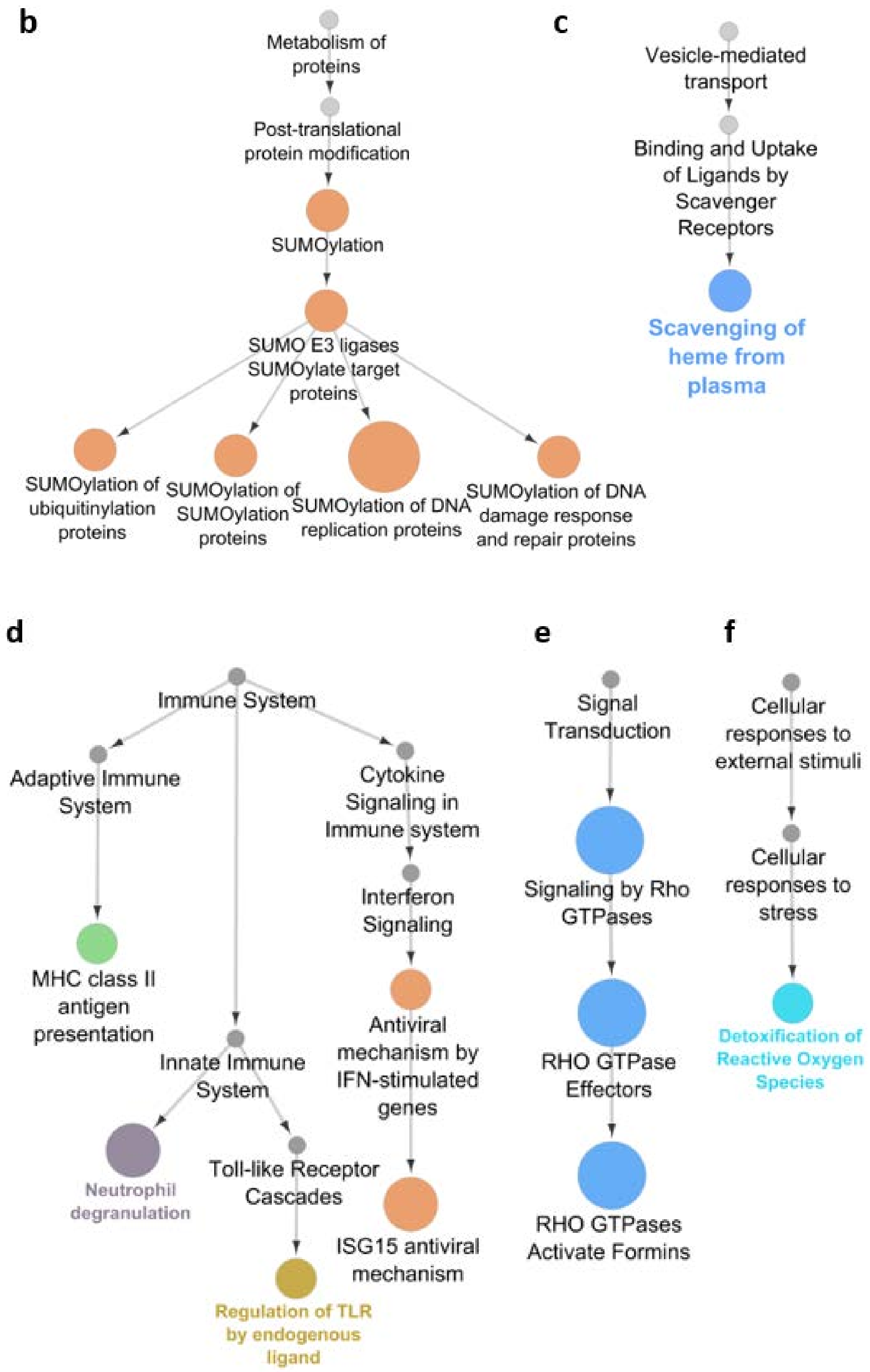
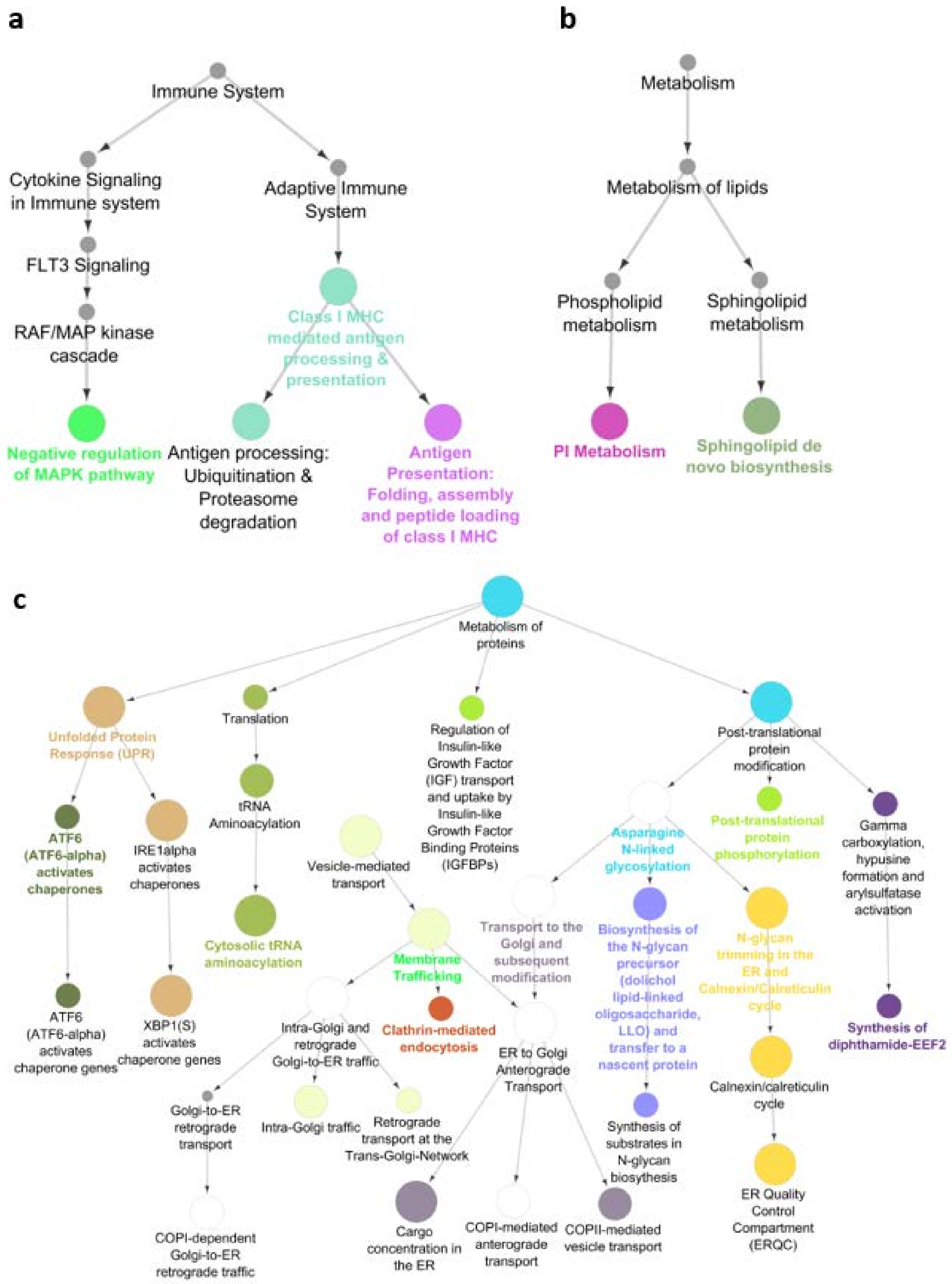

b
Metabolism of proteins
Post-translational protein modification
SUMOylation
SUMO E3 ligases SUMOylate target proteins
SUMOylation of ubiquitinylation proteins
SUMOylation of SUMOylation proteins
SUMOylation of DNA replication proteins
SUMOylation of DNA damage response and repair proteins
c
Vesicle-mediated transport
Binding and Uptake of Ligands by Scavenger Receptors
Scavenging of heme from plasma
d
Immune System
Adaptive Immune System
MHC class II antigen presentation
Innate Immune System
Neutrophil degranulation
Toll-like Receptor Cascades
Regulation of TLR by endogenous ligand
Cytokine Signaling in Immune system
Interferon Signaling
Antiviral mechanism by IFN-stimulated genes
ISG15 antiviral mechanism
e
Signal Transduction
Signaling by Rho GTPases
RHO GTPase Effectors
RHO GTPases Activate Formins
f
Cellular responses to external stimuli
Cellular responses to stress
Detoxification of Reactive Oxygen Species

Figure 3.17 *REACTOME pathways that are enriched in DEX circadian profiles peaking between 0 h and 8 h: a) metabolism, b) metabolism of proteins, c) vesicle-mediated transport, d) immune system, e) signal transduction, f) cellular response to external stimuli.*

3.4.1.2 DEX circadian profiles peaking in the resting phase


a
Immune System
Cytokine Signaling in Immune system
FLT3 Signaling
RAF/MAP kinase cascade
Negative regulation of MAPK pathway
Adaptive Immune System
Class I MHC mediated antigen processing & presentation
Antigen processing: Ubiquitination & Proteasome degradation
Antigen Presentation: Folding, assembly and peptide loading of class I MHC
b
Metabolism
Metabolism of lipids
Phospholipid metabolism
Sphingolipid metabolism
PI Metabolism
Sphingolipid de novo biosynthesis
c
Metabolism of proteins
Unfolded Protein Response (UPR)
ATF6 (ATF6-alpha) activates chaperones
IRE1alpha activates chaperones
ATF6 (ATF6-alpha) activates chaperone genes
XBP1(S) activates chaperone genes
Translation
tRNA Aminoacylation
Cytosolic tRNA aminoacylation
Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)
Vesicle-mediated transport
Membrane Trafficking
Clathrin-mediated endocytosis
Intra-Golgi and retrograde Golgi-to-ER traffic
Golgi-to-ER retrograde transport
COPI-dependent Golgi-to-ER retrograde traffic
Intra-Golgi traffic
Retrograde transport at the Trans-Golgi-Network
ER to Golgi Anterograde Transport
Cargo concentration in the ER
COPI-mediated anterograde transport
COPII-mediated vesicle transport
Transport to the Golgi and subsequent modification
Post-translational protein modification
Asparagine N-linked glycosylation
Biosynthesis of the N-glycan precursor (dolichol lipid-linked oligosaccharide, LLO) and transfer to a nascent protein
Synthesis of substrates in N-glycan biosythesis
Post-translational protein phosphorylation
N-glycan trimming in the ER and Calnexin/Calreticulin cycle
Calnexin/calreticulin cycle
ER Quality Control Compartment (ERQC)
Gamma carboxylation, hypusine formation and arylsulfatase activation
Synthesis of diphthamide-EEF2

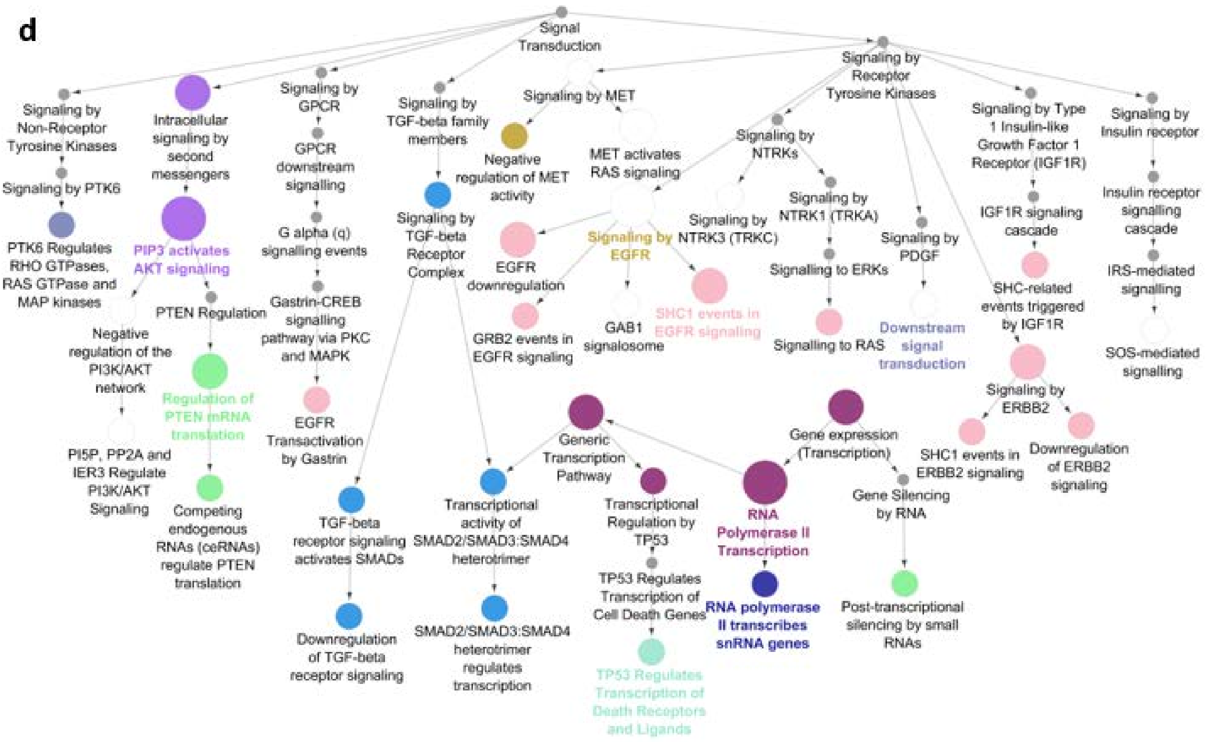
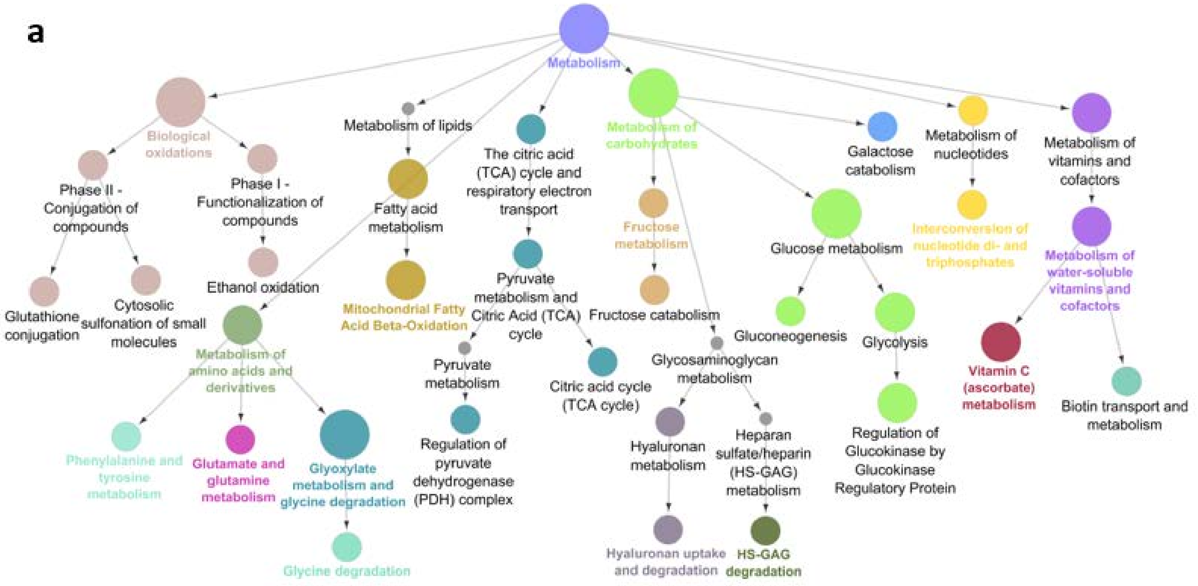

d
Signal Transduction
Signaling by Non-Receptor Tyrosine Kinases
Signaling by PTK6
PTK6 Regulates RHO GTPases, RAS GTPase and MAP kinases
Intracellular signaling by second messengers
PIP3 activates AKT signaling
PTEN Regulation
Negative regulation of the PI3K/AKT network
PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling
Regulation of PTEN mRNA translation
Competing endogenous RNAs (ceRNAs) regulate PTEN translation
Signaling by GPCR
GPCR downstream signalling
G alpha (q) signalling events
Gastrin-CREB signalling pathway via PKC and MAPK
EGFR Transactivation by Gastrin
Signaling by TGF-beta family members
Signaling by TGF-beta Receptor Complex
TGF-beta receptor signaling activates SMADs
Downregulation of TGF-beta receptor signaling
Transcriptional activity of SMAD2/SMAD3:SMAD4 heterotrimer
SMAD2/SMAD3:SMAD4 heterotrimer regulates transcription
Signaling by MET
Negative regulation of MET activity
MET activates RAS signaling
EGFR downregulation
GRB2 events in EGFR signaling
Signaling by EGFR
GAB1 signalosome
SHC1 events in EGFR signaling
Generic Transcription Pathway
Transcriptional Regulation by TP53
TP53 Regulates Transcription of Cell Death Genes
TP53 Regulates Transcription of Death Receptors and Ligands
RNA Polymerase II Transcription
RNA polymerase II transcribes snRNA genes
Gene expression (Transcription)
Gene Silencing by RNA
Post-transcriptional silencing by small RNAs
Signaling by Receptor Tyrosine Kinases
Signaling by NTRKs
Signaling by NTRK3 (TRKC)
Signaling by NTRK1 (TRKA)
Signalling to ERKs
Signalling to RAS
Signaling by PDGF
Downstream signal transduction
Signaling by Type 1 Insulin-like Growth Factor 1 Receptor (IGF1R)
IGF1R signaling cascade
SHC-related events triggered by IGF1R
Signaling by ERBB2
SHC1 events in ERBB2 signaling
Downregulation of ERBB2 signaling
Signaling by Insulin receptor
Insulin receptor signalling cascade
IRS-mediated signalling
SOS-mediated signalling

Figure 3.18 *REACTOME pathways that are enriched in DEX profiles peaking between 12 h and 20 h: a) immune system; b) metabolism; c)metabolism of proteins and vesicle-mediated transport; d) signal transduction.*

3.4.1.3 PHY circadian profiles peaking in the active phase


a
Metabolism
Biological oxidations
Phase II - Conjugation of compounds
Glutathione conjugation
Cytosolic sulfonation of small molecules
Phase I - Functionalization of compounds
Ethanol oxidation
Metabolism of amino acids and derivatives
Phenylalanine and tyrosine metabolism
Glutamate and glutamine metabolism
Glyoxylate metabolism and glycine degradation
Glycine degradation
Metabolism of lipids
Fatty acid metabolism
Mitochondrial Fatty Acid Beta-Oxidation
The citric acid (TCA) cycle and respiratory electron transport
Pyruvate metabolism and Citric Acid (TCA) cycle
Pyruvate metabolism
Regulation of pyruvate dehydrogenase (PDH) complex
Citric acid cycle (TCA cycle)
Metabolism of carbohydrates
Fructose metabolism
Fructose catabolism
Glycosaminoglycan metabolism
Hyaluronan metabolism
Hyaluronan uptake and degradation
Heparan sulfate/heparin (HS-GAG) metabolism
HS-GAG degradation
Glucose metabolism
Gluconeogenesis
Glycolysis
Regulation of Glucokinase by Glucokinase Regulatory Protein
Galactose catabolism
Metabolism of nucleotides
Interconversion of nucleotide di- and triphosphates
Metabolism of vitamins and cofactors
Metabolism of water-soluble vitamins and cofactors
Vitamin C (ascorbate) metabolism
Biotin transport and metabolism

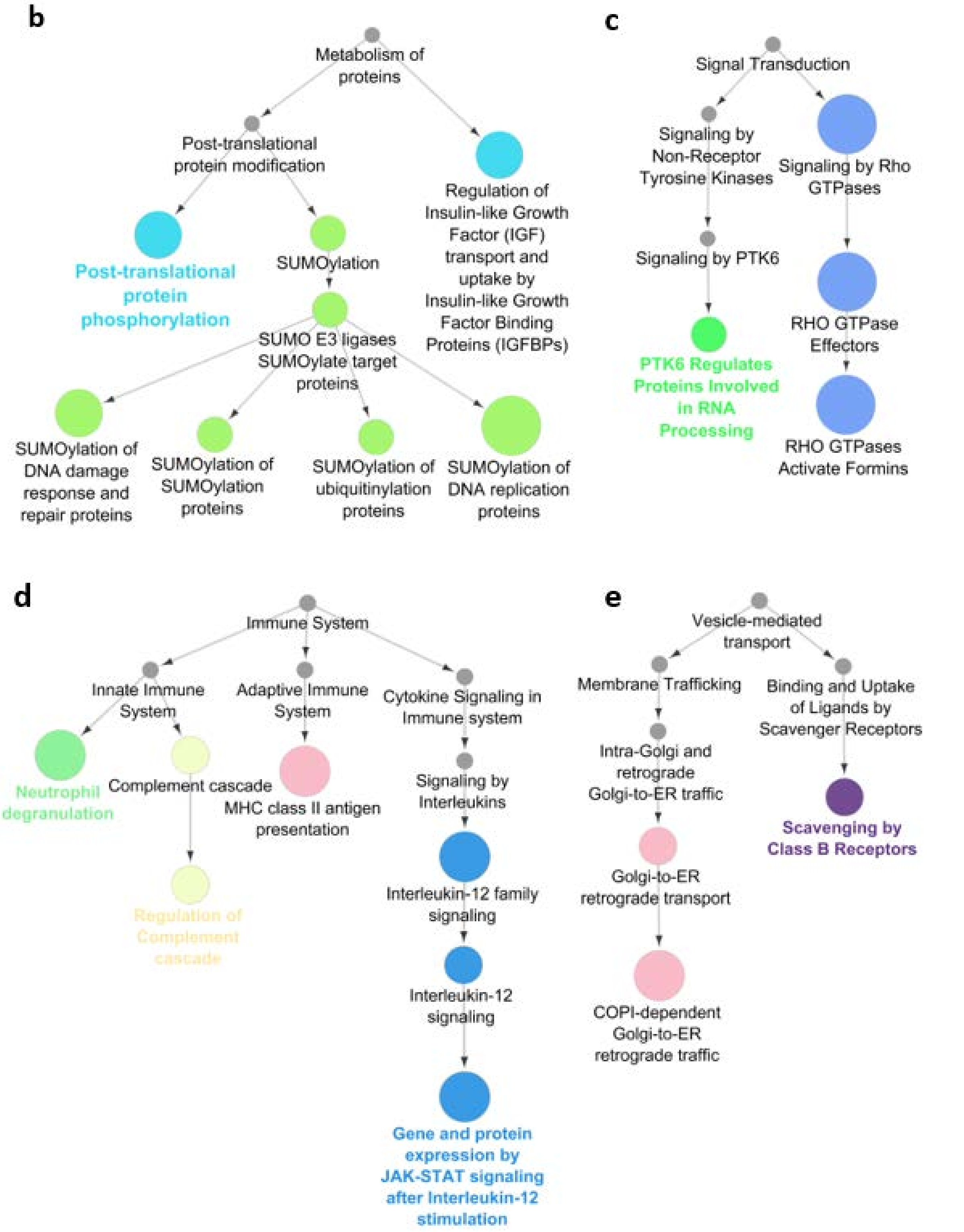

b
Metabolism of proteins
Post-translational protein modification
Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)
Post-translational protein phosphorylation
SUMOylation
SUMO E3 ligases SUMOylate target proteins
SUMOylation of DNA damage response and repair proteins
SUMOylation of SUMOylation proteins
SUMOylation of ubiquitinylation proteins
SUMOylation of DNA replication proteins
c
Signal Transduction
Signaling by Non-Receptor Tyrosine Kinases
Signaling by Rho GTPases
Signaling by PTK6
PTK6 Regulates Proteins Involved in RNA Processing
RHO GTPase Effectors
RHO GTPases Activate Formins
d
Immune System
Innate Immune System
Adaptive Immune System
Cytokine Signaling in Immune system
Neutrophil degranulation
Complement cascade
MHC class II antigen presentation
Signaling by Interleukins
Regulation of Complement cascade
Interleukin-12 family signaling
Interleukin-12 signaling
Gene and protein expression by JAK-STAT signaling after Interleukin-12 stimulation
e
Vesicle-mediated transport
Membrane Trafficking
Binding and Uptake of Ligands by Scavenger Receptors
Intra-Golgi and retrograde Golgi-to-ER traffic
Scavenging by Class B Receptors
Golgi-to-ER retrograde transport
COPI-dependent Golgi-to-ER retrograde traffic

Figure 3.19 *REACTOME pathways that are enriched in PHY circadian profiles peaking between 0 h and 8 h: a) metabolism; b) metabolism of proteins; c) signal transduction; d) immune system; e)vesicle-mediated transport*

3.4.1.4 PHY circadian profiles peaking in the resting phase

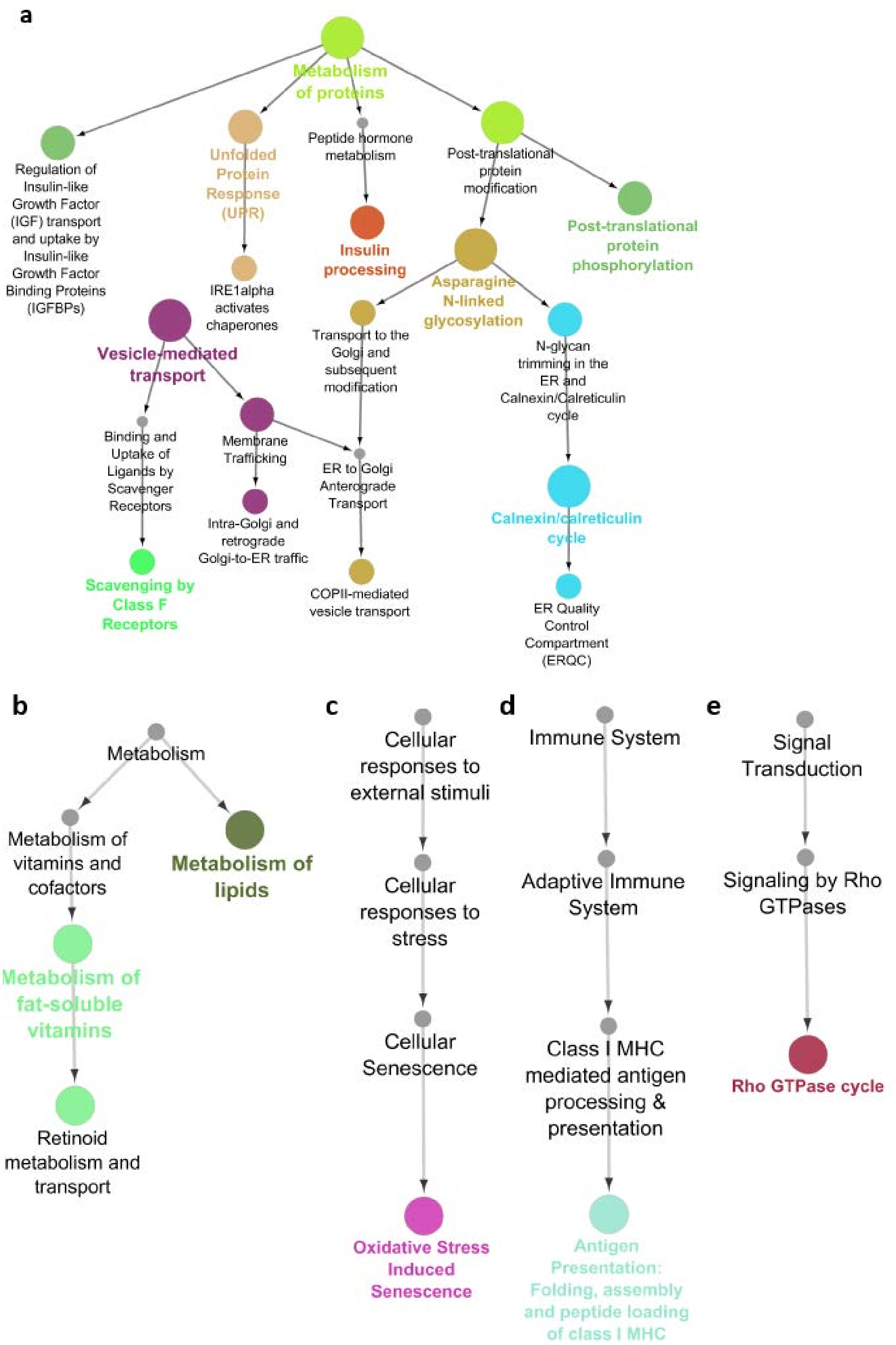

a
Metabolism of proteins
Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)
Unfolded Protein Response (UPR)
IRE1alpha activates chaperones
Peptide hormone metabolism
Insulin processing
Post-translational protein modification
Post-translational protein phosphorylation
Asparagine N-linked glycosylation
Transport to the Golgi and subsequent modification
N-glycan trimming in the ER and Calnexin/Calreticulin cycle
Calnexin/calreticulin cycle
ER Quality Control Compartment (ERQC)
Vesicle-mediated transport
Binding and Uptake of Ligands by Scavenger Receptors
Scavenging by Class F Receptors
Membrane Trafficking
Intra-Golgi and retrograde Golgi-to-ER traffic
ER to Golgi Anterograde Transport
COPII-mediated vesicle transport
b
Metabolism
Metabolism of vitamins and cofactors
Metabolism of lipids
Metabolism of fat-soluble vitamins
Retinoid metabolism and transport
c
Cellular responses to external stimuli
Cellular responses to stress
Cellular Senescence
Oxidative Stress Induced Senescence
d
Immune System
Adaptive Immune System
Class I MHC mediated antigen processing & presentation
Antigen Presentation: Folding, assembly and peptide loading of class I MHC
e
Signal Transduction
Signaling by Rho GTPases
Rho GTPase cycle

Figure 3.20 *REACTOME pathways that are enriched in PHY circadian profiles with phases between 12 h and 20 h: a) metabolism of proteins and vesicle-mediated transport; b)metabolism; c) cellular response to stimuli; d) immune system; e) signal transduction.*

3.4.1.5 Discussion

As regards DEX circadian profiles peaking in the active phase, metabolism pathway has many enriched sub-categories, while metabolism of proteins is present but to a lower extent: only SUMOylation (a type of post-translational modification) is enriched. Moreover, vesicle mediated transport, signal transduction, cellular response to external stimuli have only one enriched sub-category, while immune system has a higher number of enriched pathways, but still lower with respect to metabolism.

Instead, when DEX circadian profiles peak in the resting phase, only metabolism of lipids remains enriched among the metabolism pathways, while metabolism of proteins has a higher number of enriched sub-categories and it is nested to vesicle-mediated transport. There are other differences from DEX profiles peaking during in the active phase, like the fact that cellular response to external stimuli is absent and that signal transduction is enriched to a higher extent than in the previous case. The only common aspect is that immune system is enriched but it does not show a high number of sub-categories.

Similarly to DEX profiles, PHY circadian transcripts peaking in the active phase have a relatively high number of metabolism sub-categories enriched, while only post-translational protein phosphorylation and SUMOylation are enriched in the metabolism of proteins. Moreover, vesicle-mediated transport and signal transduction have only two enriched sub-categories, while the immune system has three.

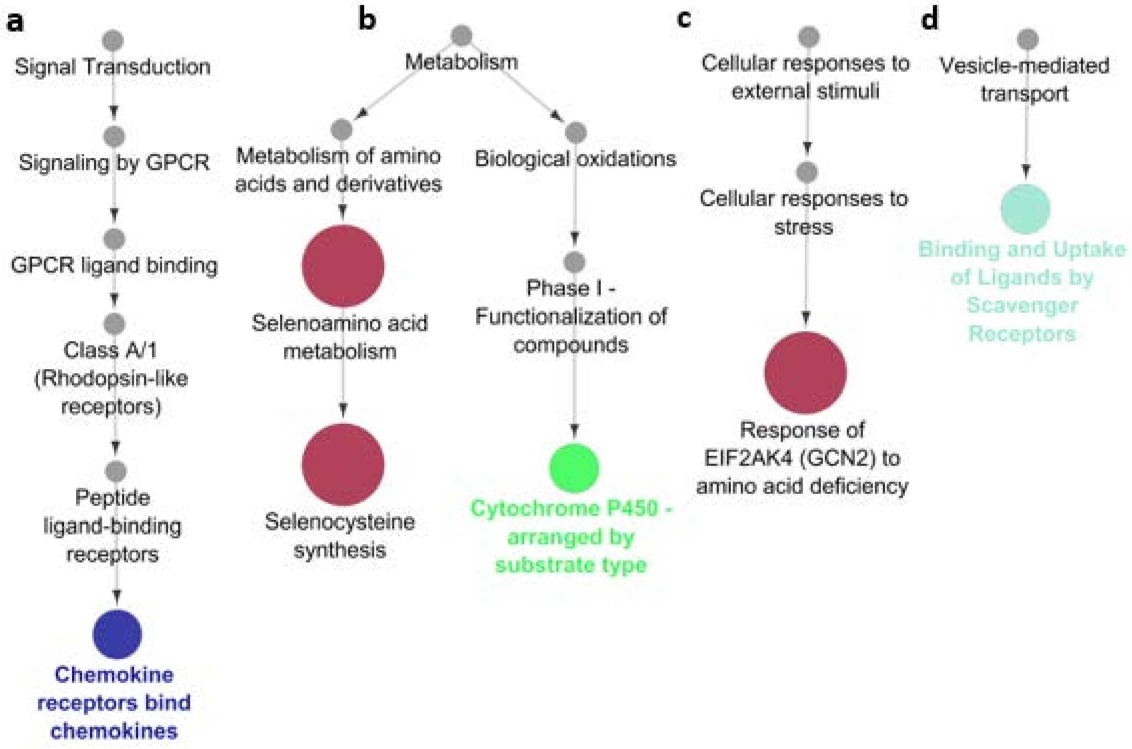
Moreover, in the resting phase the most enriched pathway in PHY protocol is metabolism of proteins, while metabolism has only two enriched sub-categories and immune system, signal transduction and cellular response to external stimuli have only one.

These results match the most relevant liver pathways encountered in literature; for example, in our analyses the main hepatic metabolic processes, including glucose, lipid, and cholesterol/bile acid metabolism, are enriched when PHY circadian genes are analysed and this is in accordance with the findings of many researches in the circadian field. Moreover, circadian rhythms in cells are possible thanks to many regulatory mechanisms, that are not completely clear up to date, but that likely involve dynamic post-translational modifications of key proteins, like phosphorylation, acetylation, ubiquitination, and SUMOylation (Gad Asher and Sassone-Corsi, 2015), (Anand R. Saran et al, 2020), (Lee et al., 2015), (Mehra et al., 2009).

Finally, an interesting result is that even though the hormonal synchronization (i.e. DEX protocol) makes a higher percentage of profiles oscillate than feeding-fasting stimuli (i.e. PHY protocol), the evolution of liver biological functions throughout the day is similar in both cases: for example, both in DEX and PHY protocols there is a higher number of metabolism pathways enriched during the active phase and a higher number of metabolism of proteins terms enriched in the resting phase. Moreover, in both cases metabolism of lipids is the only metabolism pathway which is enriched all day long. Interestingly, even though many sub-categories of metabolism of proteins are enriched when profiles peaking in the resting phase are analysed, the SUMOylation pathway is enriched only when DEX and PHY circadian profiles are peaking during the active phase.

3.4.2 Enrichment analysis of non-circadian DEGs

While the previous section analysed circadian profiles, this one considers not circadian profiles in either protocol that are selected as Differentially Expressed by both FunPat and edgeR. In particular, DEGs list is divided into the two main clusters obtained by hierarchical clustering in Figure 3.12: flat profiles constantly up-regulated in PHY and those in DEX.


a
Signal Transduction
Signaling by GPCR
GPCR ligand binding
Class A/1 (Rhodopsin-like receptors)
Peptide ligand-binding receptors
Chemokine receptors bind chemokines
b
Metabolism
Metabolism of amino acids and derivatives
Biological oxidations
Selenoamino acid metabolism
Selenocysteine synthesis
Phase I - Functionalization of compounds
Cytochrome P450 - arranged by substrate type
c
Cellular responses to external stimuli
Cellular responses to stress
Response of EIF2AK4 (GCN2) to amino acid deficiency
d
Vesicle-mediated transport
Binding and Uptake of Ligands by Scavenger Receptors

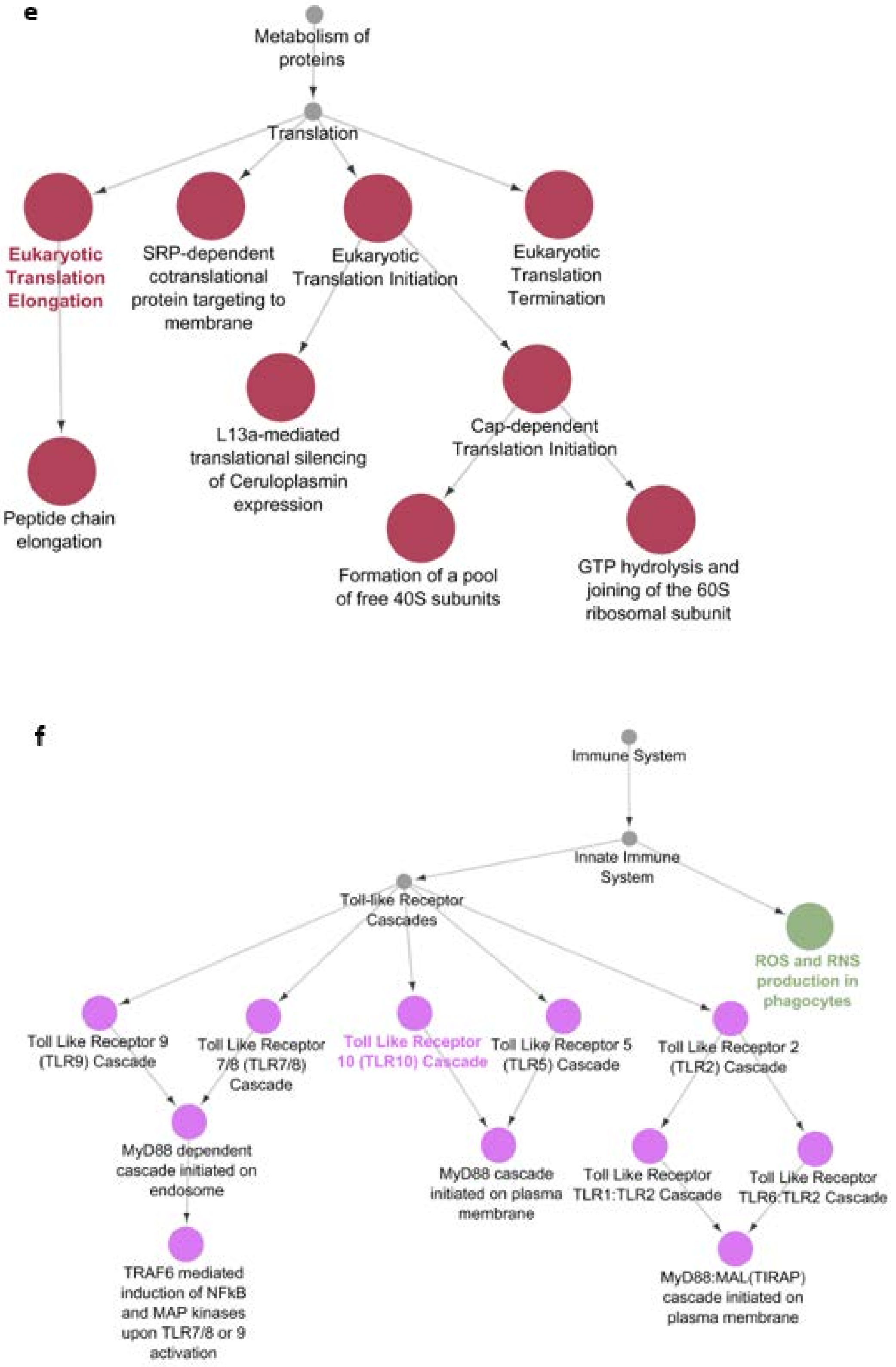

e
Metabolism of proteins
Translation
Eukaryotic Translation Elongation
SRP-dependent cotranslational protein targeting to membrane
Eukaryotic Translation Initiation
Eukaryotic Translation Termination
Peptide chain elongation
L13a-mediated translational silencing of Ceruloplasmin expression
Cap-dependent Translation Initiation
Formation of a pool of free 40S subunits
GTP hydrolysis and joining of the 60S ribosomal subunit
f
Immune System
Innate Immune System
Toll-like Receptor Cascades
ROS and RNS production in phagocytes
Toll Like Receptor 9 (TLR9) Cascade
Toll Like Receptor 7/8 (TLR7/8) Cascade
Toll Like Receptor 10 (TLR10) Cascade
Toll Like Receptor 5 (TLR5) Cascade
Toll Like Receptor 2 (TLR2) Cascade
MyD88 dependent cascade initiated on endosome
MyD88 cascade initiated on plasma membrane
Toll Like Receptor TLR1:TLR2 Cascade
Toll Like Receptor TLR6:TLR2 Cascade
TRAF6 mediated induction of NFkB and MAP kinases upon TLR7/8 or 9 activation
MyD88:MAL(TIRAP) cascade initiated on plasma membrane

Figure 3.21 *REACTOME pathways enriched in not circadian genes differentially expressed, with constantly lower values in DEX than in PHY protocol: a) signal transduction; b) metabolism; c) cellular responses to external stimuli; d) vesicle-mediated transport; e) metabolism of proteins; f) immune system.*

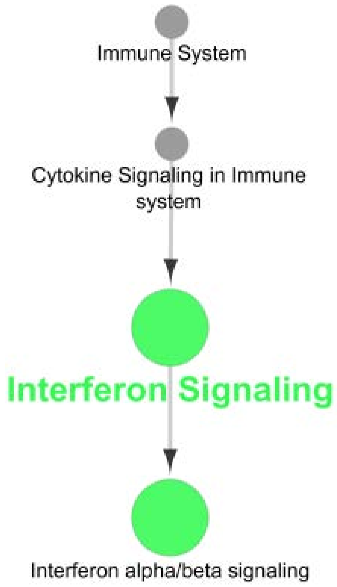

Immune System
Cytokine Signaling in Immune system
Interferon Signaling
Interferon alpha/beta signaling

Figure 3.22 *REACTOME pathway enriched in differentially expressed non circadian genes, with constantly higher values in DEX than in PHY protocol: immune system.*

3.4.2.1 Discussion

As anticipated, these results are obtained by considering all DEGs that are not circadian, in order to understand their biological functions when they are constantly up-regulated in one protocol.

In particular, transcripts that are constantly up-regulated in PHY protocol are enriched in a higher number of pathways, especially in metabolism of proteins, metabolism and immune system. In contrast, among the list of pathways included in the discussion of this thesis, genes that are up-regulated in DEX protocol are enriched only in immune system. Therefore, these results suggest that, even though PHY protocol generates a lower percentage of circadian profiles than DEX one, it still maintains the essential biological functions within the liver (e.g. metabolism and metabolism of proteins) even when the circadianicity is not fully attained.

3.5 Summary of transcriptomics results

The 12940 transcripts profiles of the reference protocol, i.e. DEX, and of the innovative one, i.e. PHY, are first visualized through PCA score plots, using both logarithmic and linear data. The latter displays a higher variability with respect to the former one, even if Standard Normal Variate scaling is applied. However, two main aspects can be highlighted in both results: first, the major source of variability is given by the synchronization protocol itself; then, especially with logarithmic data, four replicates of the same time point tend to form a cluster that is distinguishable from those of the other time points. These clusters suggest that the majority of transcripts tends to change considerably across time points, which in turn suggests that there might be a relevant percentage of circadian profiles (i.e. sinusoidal profiles with one peak within 24 hours), instead of randomly fluctuating profiles.

After this preliminary analysis, the dynamics of all profiles are characterized and compared between the two protocols. In particular, circadian rhythms are identified, revealing that the percentage of circadian oscillations of PHY profiles is of the same order of magnitude of DEX profiles (22 % and 35%, respectively). This means that the perturbation of cells cultured *in vitro* by using feeding stimuli is as much effective as the hormonal one for the purpose of activating the circadian clock of human liver cells. Then, circadian profiles in DEX and PHY protocol are compared considering two key parameters for sinusoidal profiles: phase and amplitude. The former, defined as the time point where the peak occurs, displays a low level of correlation (R^2 below 0.5), even though profiles circadian in both protocols are shifted only between 0 h and 3 h. Moreover, circadian profiles are characterized by a bimodal distribution of phases with peaks concentrated around 3 h and 15 h in both protocols, indicating a net separation between functions activated in the two halves of the 24 h. Instead, the comparison between DEX and PHY amplitudes shows a relatively high level of correlation (R^2 between 0.73 and 0.80), meaning that if the level of oscillation is high in one protocol, it is likely to be high also in the other. Moreover, the characterization of non-circadian profiles by means of Differential Expression Analysis reveals that two main trends occur within transcripts that are circadian in only one protocol. Indeed, DEGs that are circadian only in DEX peaking between 0 h and 8 h have corresponding profiles in PHY that are constantly down-regulated (i.e. with constant low values), while when they peak between 12 h and 20 h the corresponding PHY profiles are constantly up-regulated. Similarly, DEGs profiles that are circadian only in PHY with phases between 0 h and 8 h are up-regulated in DEX, while in the other cases they are down-regulated in DEX. In addition, DEGs that are not circadian in either protocol can be divided into two main groups: transcripts that are constantly up-regulated in DEX and those that are constantly up-regulated in PHY, meaning that, when gene transcription is constant throughout 24 hours, hormonal and feeding stimuli are able to activate different biological functions.

Finally, the objective is to determine which biological functions are activated in liver by hormonal and metabolic stimuli, distinguishing between activities that are carried out at different times of the day (due to circadian regulations) or constantly throughout 24 h. In the former case, it is important to define the active phase (i.e., day for humans and night for nocturnal animals like mice) and resting phase (i.e., night for humans), because liver is expected to carry out different functions in the two phases, in order to adapt the organisms to the changing level of activity, awareness and food availability. Based on the experimental setup and on the peak phases of reference circadian genes, it can be concluded that the interval 0-8 h corresponds to the active phase, while 12-20 h to the resting one. Moreover, the majority of circadian profiles have peak phases in 0-8 h and 12-20 h intervals, thus including only them in the discussion does not lead to a relevant loss of information. When the enrichment analysis is performed considering these two groups of circadian genes, both DEX and PHY results are in accordance with the main liver functions found in literature: for example, hepatic metabolic processes, including glucose, lipid, and cholesterol/bile acid metabolism, together with dynamic post-translational modifications of key proteins, like phosphorylation, acetylation, ubiquitination, and SUMOylation. Many other pathways are enriched, but only the most relevant in terms of number of enriched sub-categories and of biological relevance are shown for both DEX and PHY: vesicle-mediated transport, immune system, signal transduction, cellular response to external stimuli. Finally, the biological functions of DEGs constantly up-regulated in PHY and in DEX are analysed separately: the former has more pathways enriched in common with circadian profiles, i.e. signal transduction, metabolism, cellular responses to external stimuli, vesicle-mediated transport, metabolism of proteins and immune system, instead of only immune system as the latter. This suggests that the slightly lower percentage of circadian profiles generated by the innovative protocol with respect to the reference one does not lead to a loss of biological functions, but just to a non-circadian behaviour for some of the genes involved.

Chapter 4
Proteomic data analysis

The aim of this chapter is to analyse the dynamics in human liver proteome induced by the dexamethasone and physiological protocols. First of all, since there are still many technical limitations in the field of shotgun proteomics, it is important to compare this Thesis results with some other sources in literature in order to assess their comparability. However, this comparison is complicated by the scarcity of papers on human liver circadian proteome; thus, four papers are used even though there are some limitations in the comparison, like the differences in measurement technologies and/or in the organisms studied (mice instead of humans, for example).

This allows to compare the inter-variability among different literature sources and among literature and the data of this Thesis, in order to assess whether they have the same order of magnitude or not, which in turn may support the good quality of the DEX and PHY data themselves. Together with the estimation of percentages of common detected proteins and common circadian proteins, a comparison is made among peak phases of circadian proteins shared between literature and DEX and PHY data, in order to evaluate the differences in synchronization due to different experimental protocols. Afterwards, the attention focuses on the data of this Thesis only, by comparing the main parameters (i.e. phases and amplitudes) in the two protocols. Finally, a biological interpretation is provided thanks to an enrichment analysis of the circadian proteins, that are divided into those peaking in the active phase or in the resting phase, thus allowing to characterize the evolution of the human liver functions throughout the day.

4.1 Comparison with state of the art

The aim of this section is to deepen the understanding of data variability regarding Proteomics, both among literature sources and between literature and the proteomic data of this Thesis. This is useful for understanding how our findings position respect to the state of the art.

Proteomics field is less mature than the transcriptomics one and this reflects also in the scarcity of published researches, especially in the case of human liver circadian proteome. Therefore, it is necessary to take into consideration some experimental data that are not directly comparable to the ones of this Thesis, due to differences in the synchronization protocol, in the organisms considered (e.g. mice instead of humans), in the measurement technologies and in the algorithms for circadian rhythms identification. It is not clear yet how all these factors may influence the study on circadian rhythms, so this comparison is useful only for getting a general picture of the major findings about circadian rhythms in liver proteome up to date.

The first paper considered is *Circadian Orchestration of the Hepatic Proteome* by Reddy et al., 2006 (Current Biology 16, 1107–1115). It studies mouse liver proteome thanks to *in vivo* experiments, by taking samples from groups of mice on the second cycle after the transfer from synchronization to free-running conditions. In particular, the former condition refers to 12 hours of light (*L*, 220 mW/cm^2) followed by 12 hours of dim red light (*DR*, less than 5 mW/cm^2), while the latter refers to constant dim red light for 24 hours. The technology used to measure protein abundance is two-dimensional difference gel electrophoresis (2D-DIGE), where three replicates for each sample can be separated at the same time, based on the molecules dimension and charge. Finally, circadian proteins, whose abundance (determined by image analysis) was detected as having circadian behavior by ANOVA, were identified by mass spectrometry.

The second paper considered is *Circadian clock-dependent and -independent rhythmic proteomes implement distinct diurnal functions in mouse liver* by Mauvoisin et al., 2014 (PNAS, 111 (1) 167-172). Experimental data are retrieved thanks to *in vivo* experiments on mice subjected to physiological light-dark cycles. Mice are maintained with free access to water and food, but starting from four days before the experiments the food access is restricted to night. Moreover, protein abundance is assessed by means of SILAC Mass Spectrometry, a technique where proteins are metabolically labelled *in vivo* followed by mass spectrometry. Respect to the method of labeling of this Thesis, SILAC labeling substitutes light isotopes of carbon, nitrogen or hydrogen with heavier isotopes during protein synthesis, thus samples to be compared are mixed immediately after protein collection with less technical variability between samples. However, the mass spectrometry coverage is reduced in this case. Finally, circadian rhythms are identified through a harmonic regression, i.e. a non linear regression employing sinusoidal terms for the fitting, while the significance of the peak values during the 24 hours is assessed with an F test followed by a Benjamini-Hochberg correction.

Then, the analysis considers *In-vivo Quantitative Proteomics Reveals a Key Contribution of Post-Transcriptional Mechanisms to the Circadian Regulation of Liver Metabolism* by Robles et al., 2014 (PLOS Genetics, 10(1):e1004047). In this case, *in vivo* experiments are performed on mice that are subjected to 12 hours of light, 12 hours of darkness, with free access to water and food, before being subjected to 24 hours of constant darkness. Also in this case, SILAC Mass Spectrometry is used, while the bioinformatics analysis is different with respect to the previous paper. Indeed, Perseus algorithm is employed: it fits a cosine with a period of 23.6 h to each protein abundance profile, setting amplitude and phase as free parameters.

The fourth and last paper considered in the comparison is *A proteomics landscape of circadian clock in mouse liver* by Wang et al., 2018 (Nature Communications, 9(1):1553). These *in vivo* experiments employ mice subjected to 12 h: 12 h light-dark cycles, with free access to food and water for 2 weeks. Different types of data are collected: transcription factors (*TFs*), transcription co-regulators (*TCs*), DNA-binding proteins (*DBPs*) by means of catTFRE followed by Liquid Chromatography Mass Spectrometry (LC - MS) and nuclear proteome and whole liver proteome by means of label free LC – MS. In particular, transcription factors are proteins that bind only specific tracts of the DNA, called TF response elements (TFREs), and they are difficult to be detected because of their low amounts within cells. Therefore, before being measured through LC-MS, it is useful to increase their concentration thanks to the so-called ‘catTFRE’, meaning artificial DNA molecules containing *ad hoc* TF response elements that are repeated in adjacent positions. After collecting the biological samples, circadian rhythms in protein expression profiles are identified thanks to JTK-Cycle (an algorithm previously developed in the literature).

As described in the experimental setup section of this Thesis, DEX and PHY data are collected during *in vitro* experiments on human liver cells, after being subjected to hormonal or feeding stimuli, respectively. Moreover, the measurement technology is Liquid Chromatography Tandem Mass Spectrometry (LC – MS/MS), using Tandem Mass Tag (TMT) as labelling. Finally, circadian profiles are identified with the PCA model described in Chapter 2, section 2.2.3.

For sake of conciseness, the following nomenclature is adopted:

- *Reddy 2006* refers to Reddy et al., 2006;
- *Mauvoisin 2014* refers to Mauvoisin et al., 2014;
- *Robles 2014* refers to Robles et al., 2014;
- *Wang 2018* refers to Wang et al., 2018;
- *DEX* refers to the dataset collected under the Dexamethasone protocol of this Thesis;
- *PHY* refers to the dataset collected under the physiological protocol of this Thesis.

Table 4.1 *Experimental setup of the four proteomics articles. In particular, 'Exp. Cond.' Stands for 'experimental conditions ', 'Protein ident.' stands for 'protein identification'.*

Study	Species	Exp. Cond.	Synchronizer/ Desynchronizer	Protein labeling	Protein ident.	Circadianity identification
Reddy 2006	Mouse	in vivo	12h light+12h dark, constant darkness	label-free	MS	In gel identification (ANOVA)
Mauvoisin 2014	Mouse	in vivo	12h light+12h dark, food at night	SILAC	MS	harmonic regression
Robles 2014	Mouse	in vivo	12h light+12h dark, constant darkness, food ad libitum	SILAC	MS	harmonic regression
Wang 2018	Mouse	in vivo	12h light+12h dark, food ad libitum	label-free	MS	JTK
DEX	Human	in vitro	Dexamethasone	TMT	MS	PCA model
PHY	Human	in vitro	insulin/ glucagon/ glucose	TMT	MS	PCA model

4.1.1 Intersection of detected proteins

It is important to remind that proteomics is not as mature as transcriptomics, not only at the data analysis level, but also at measurement level: the available technologies, indeed, do not provide full coverage, thus some proteins may be expressed without being detected. Therefore, different experiments usually detect a different total number of proteins (Figure 4.1).

Indeed, Reddy 2006 measures 642 proteins overall, while Mauvoisin 2014 measures the profiles of 5827 proteins, taking samples every 3 hours for two days (8 time samples per day). However, only 4408 of these protein profiles have measurements in at least 8 of 16 time points, so they are the only ones to be included in the analysis in order to have more robust results.

Moreover, Robles 2014 detects 3132 proteins in total, while Wang 2018 provides many supplementary datasets: indeed, different independent experiments have been carried out in order to investigate different levels of the molecular clock network. In particular, the interesting data for our purposes are transcription factors (*TFs*), transcription co-regulators (*TCs*), DNA-binding proteins (*DBPs*), the nuclear sub-proteome and the whole proteome of mouse liver. In principle, the first four datasets should be a subset of the fifth one, but due to the lack of full coverage in proteomics experiments there are some proteins in the sub-proteome datasets that are not measured in the whole proteome dataset, therefore a merged matrix of 8074 proteins is considered.

Finally, the number of proteins detected in DEX experiments of this Thesis is 3691, while the one in PHY experiments is 5939.

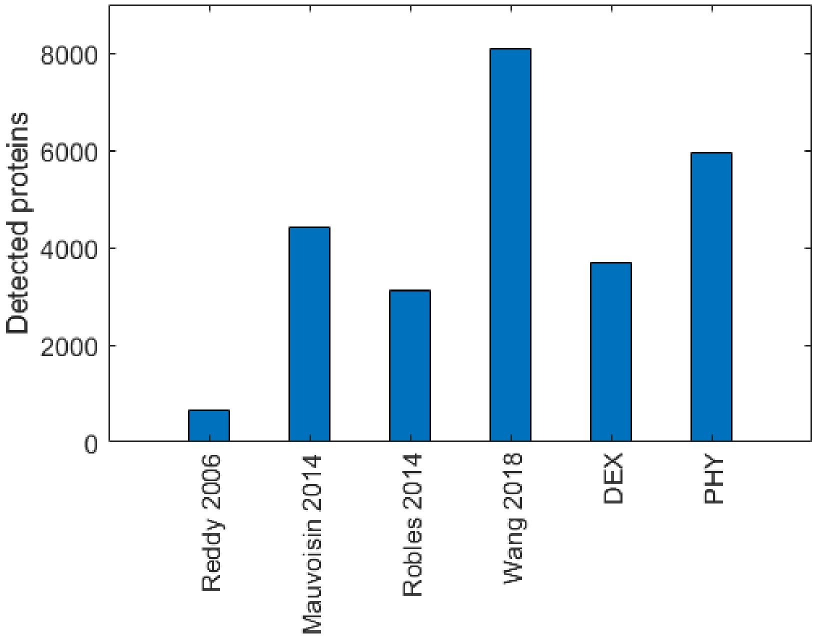


Figure 4.1 *Number of detected protein profiles in the four literature sources and in the datasets of this Thesis.*

The next step of the analysis is to assess whether shorter lists of detected proteins are just subsets of longer lists or they show a consistent number of different proteins. Unfortunately, Reddy 2006 provides only the matrix of circadian proteins, therefore it cannot be included in the analysis. The results of the pairwise comparisons between datasets are shown in Table 4.2 and in the Figures 4.2 – 4.5.

Table 4.2 *Comparison of detected proteins among different literature sources and this Thesis datasets.*

	REDDY 2006	**MAUVOISIN 2014**	**ROBLES 2014**	**WANG 2018**	**DEX**	**PHY**
REDDY 2006	**642**					
MAUVOISIN 2014	-	**4408**				
ROBLES 2014	-	2662	**3132**			
WANG 2018	-	3979	2906	**8074**		
DEX	-	2464	2022	3027	**3691**	
PHY	-	3223	2442	4393	3691	**5939**

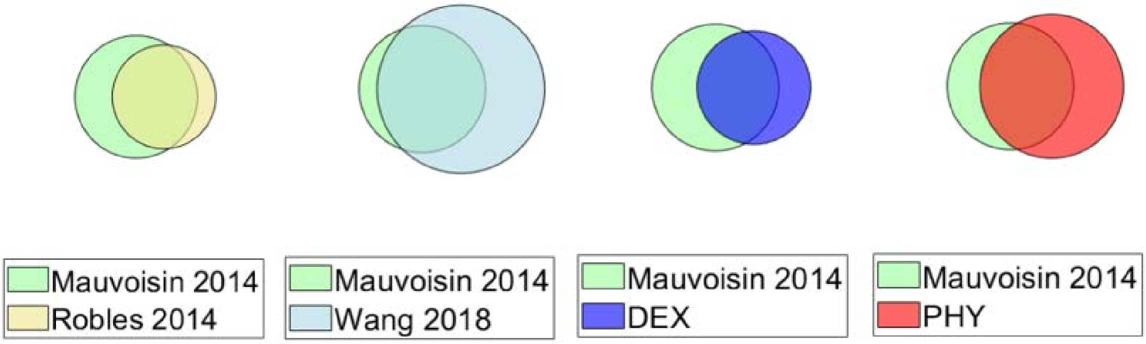


Figure 4.2 *Comparisons of detected proteins in Mauvoisin 2014 and other sources.*

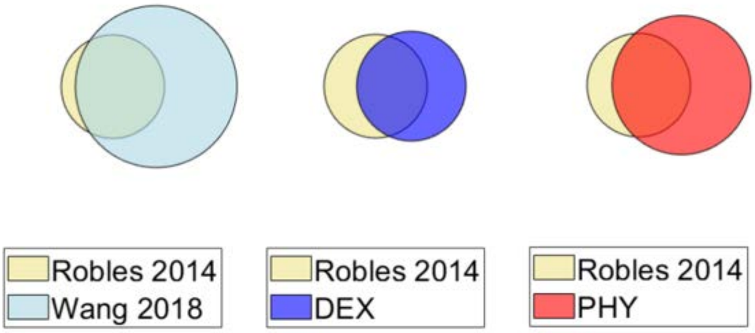


Figure 4.3 *Comparisons among detected proteins in Robles 2014 and other sources.*

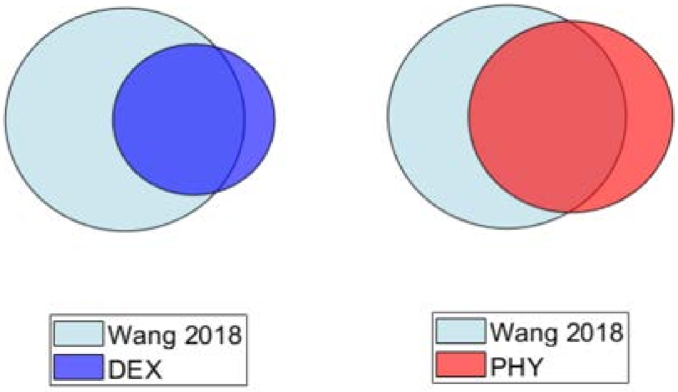


Figure 4.4 *Comparisons among detected proteins in Wand 2018 and in the data of this Thesis.*

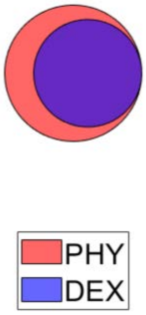


Figure 4.5 *Comparison of detected PHY and DEX.*

The level of intersection between different laboratories results is not negligible since it exceeds 50% almost in all comparisons: for example, Robles 2014 shares 85% of its detected proteins with Mauvoisin 2014, 93% with Wang 2018 and 65% and 78% with DEX and PHY data. However, there is not a perfect overlap between studies, which may be due to the fact that biological system, experimental design and measurement technologies are not the same. Moreover, even when two measurements are carried out through mass spectrometry they are not directly comparable: indeed, many factors are different, like the labelling strategy, the type of ionization or the type of detector.

The only percentages of intersection that are lower than 50% are those calculated with respect to the total number of detected proteins in Wang 2018 (they are between 36 % and 54.4 %). However, this is reasonable if one considers that the corresponding dataset is the merge of 4 different experimental campaigns, thus the total number of detected proteins (which is used to calculate those percentages) is almost double than in the other cases.

Finally, the percentages of intersection of DEX and PHY data with the literature matrices are completely similar to those of literature datasets among each other, thus supporting the good outcome of the experiments analysed in this Thesis. Moreover proteins identified in DEX data are a subset of proteins identified in PHY data, which may be due to the fact that biological system and measurement technologies are the same.

4.1.2 Proportion of circadian proteins within the same experiment

The identification of circadian proteins among all detected proteins may be influenced by many factors: for example, by the efficiency of experimental protocols in attaining synchronization, by the accuracy of measurement technologies and by the ability of algorithms to distinguish between random noise and statistically significant changes in protein abundance. Therefore, this section aims at providing only an order of magnitude of the percentages of circadian proteins in different experiments (Figure 4.6).

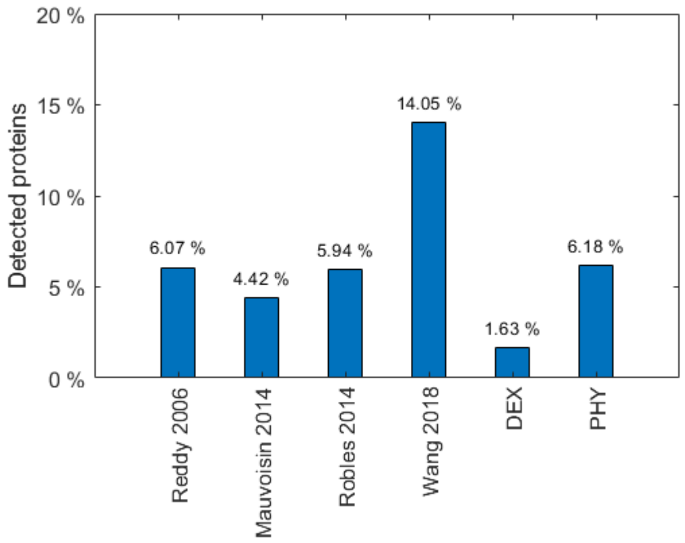


Figure 4.6 *Percentages of circadian proteins identified in each source.*

As can be seen from Figure 4.6, in general the percentage of circadian proteins is between 4% and 6%, considering each experiment separately. Some exceptions are the lower percentage in DEX, which may be due to lower efficiency of the experimental protocol, or more probably to technical limitations of the measurement, and the higher percentage in Wang 2018, which is probably due to the merge of different circadian proteins dataset as previously explained. In this case, also data coming from Reddy 2006 are available, which employ 2D difference gel electrophoresis for circadian protein selection and mass spectrometry only for protein identification, and these results suggest that the measurement technology does not impact significantly on the percentage of circadian proteins detected. Finally, these relatively low percentages of circadian proteins may be increased in the future by improving of measurement technologies, thus revealing the presence of new circadian proteins that have not been detected yet due to technological resolution limitations.

4.1.3 Intersection of circadian proteins

The previous section highlighted the relatively low percentages of circadian proteins that are identified within the same experiment; instead, this section aims at understanding whether those subsets of circadian proteins are similar across experiments or they are unique for each study. In particular, all comparisons are organized considering three approaches:

1) pairwise comparisons between two different papers (Table 4.3 and Figure 4.7),
2) pairwise comparisons between a paper and DEX (or PHY) dataset (Table 4.4 and Figure 4.8),
3) pairwise comparisons between DEX and PHY data.

In particular, when DEX and PHY data are considered, two groups are taken into account: the proteins that are actually identified as circadian by the PCA model (thus with a relative radius equal to or higher than 0.7), but also proteins that are close to this threshold (thus including profiles with a relative radius between 0.5 and 0.7).

Table 4.3 *Common circadian proteins among different literature sources.*

	REDDY 2006	MAUVOISIN 2014	ROBLES 2014	WANG 2018
REDDY 2006	**39**			
MAUVOISIN 2014	3	**195**		
ROBLES 2014	1	33	**186**	
WANG 2018	2	41	32	**1134**

Table 4.4 *Common circadian proteins in literature and in the DEX and PHY datasets of this Thesis.*

	DEX ($Rr \geq 0.5$)	DEX ($Rr \geq 0.7$)	PHY ($Rr \geq 0.5$)	PHY ($Rr \geq 0.7$)
REDDY 2006	5	0	5	1
MAUVOISIN 2014	26	0	52	20
ROBLES 2014	20	5	51	19
WANG 2018	63	5	169	38

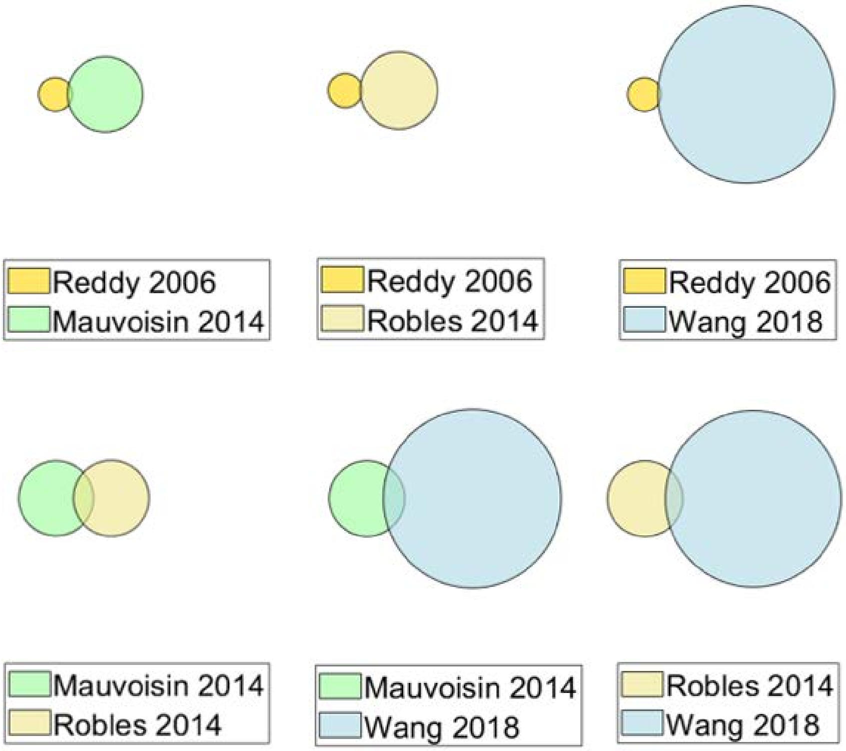


Figure 4.7 *Comparisons among circadian proteins of different literature sources.*

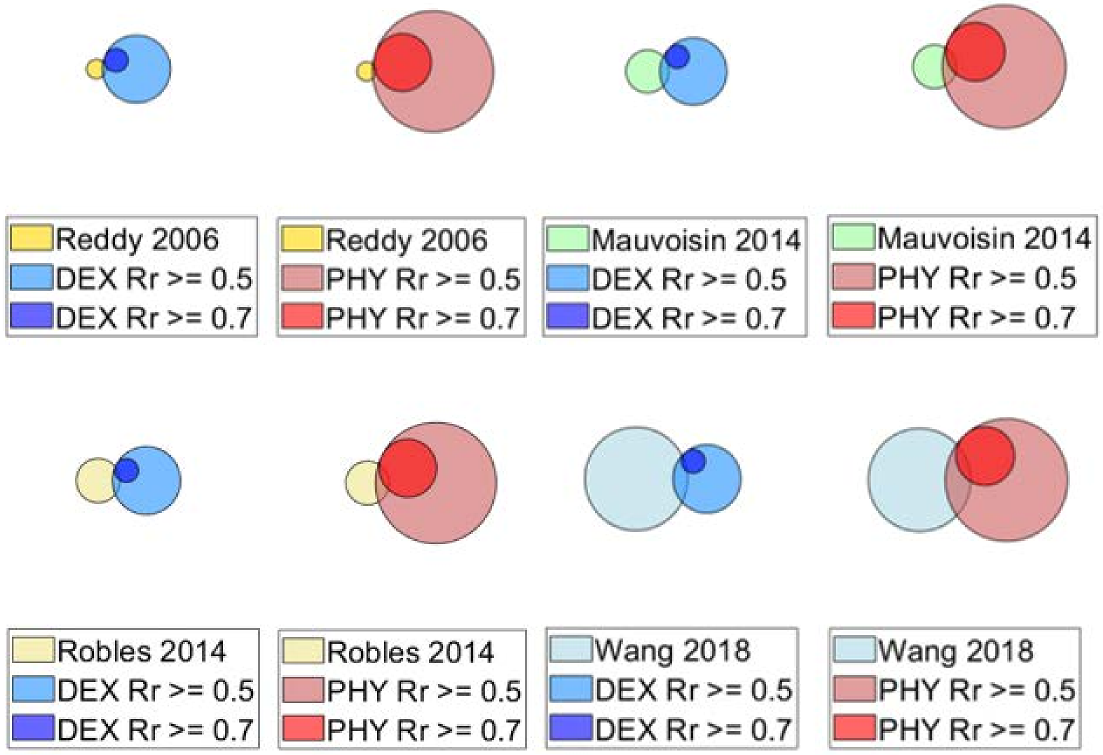


Figure 4.8 *Comparisons between circadian proteins of literature and of DEX and PHY datasets.*

As regards the comparison between circadian proteins of two literature datasets, the papers that share the highest percentages of circadian proteins with other sources are Mauvoisin 2014 and Robles 2014. Indeed, the former shares about 21 % of its circadian proteins with Wang 2018 and almost 17 % with Robles 2014, while the latter shares almost 18% of its circadian proteins with Mauvoisin 2014 and about 17% with Wang 2018. Instead, Wang 2018 seems to have the lowest percentages of circadian proteins shared with other literature sources (they are all below 3.62%), but from the absolute values in Table 4.3 it is clear that those percentages are lowered by its high total number of circadian proteins.

As regards the second approach, DEX protein profiles that are circadian, therefore with a relative radius (*Rr*) higher than 0.7, yield relatively low percentages of intersection when compared to each one of the four papers. In contrast, when DEX proteins list is increased by taking into account also profiles that are close to the threshold of circadian rhythms, meaning with a relative radius between 0.5 and 0.7, the percentages of intersection are significantly increased. Also PHY circadian profiles ($Rr \geq 0.7$) and PHY profiles close to the threshold (*Rr* higher than 0.5) lead to percentages of intersection with the four papers that resemble those of the first approach, thus proving further the good quality of the experimental data of this Thesis. Finally, in both DEX and PHY data, lowering the threshold to 0.5 for the identification of circadian rhythms lead to a significant increase in the number of shared proteins, therefore these protein lists are employed for comparing peak phases of common circadian proteins (see section 4.1.3.2).

Moreover, when DEX and PHY data of this Thesis are compared with each other, the number of circadian proteins that are identified in both protocols with a minimum relative radius of 0.5 is equal to 208, that corresponds to 13 % of the total circadian proteins identified in PHY protocol and to 42 % of those identified in DEX protocol (that is a higher percentage than those of the previous approaches). However, the number of shared proteins between the two protocols decreases dramatically when the stricter threshold of 0.7 is considered for *Rr*: indeed, 17 circadian proteins are shared, which corresponds to 4.63% of the total circadian proteins in PHY and to 28.33 % of the circadian proteins identified in DEX. Even though the percentages are lower in this case, they are still comparable to those of the first and second approach and this may be explained by the fact that these data are directly comparable because they are generated with similar experimental setup (except for the synchronization protocol) and they are analysed with the same algorithm.

Overall, these results show that the inter-variability across laboratories is relatively high, without revealing particular differences based on different experimental conditions. For example, the percentage of circadian proteins is not significantly lower in PHY protocol with respect to the literature data, proving that *in vitro* models not necessarily lead to less efficiency in the synchronization with respect to *in vivo* models. The only source that provides low percentages (below 1.5%) in all comparisons is Reddy 2006; this diversity is likely due to the measurement method, i.e. 2D difference gel electrophoresis followed by mass spectrometry, since the remaining experimental conditions are similar to those of the other three papers: they all use *in vivo* mice models, with synchronization protocols based on light-darkness cycles. In contrast, the method used to identify circadian rhythms seems not be the major source of similarity, because both Reddy 2006 and Mauvoisin 2014 use analysis of variance to identify significant changes in the expression profiles, i.e. circadian rhythms, without identifying a greater number of circadian proteins with respect to the other sources.

4.1.3.1 Novel detection of proteins and novel identification of circadian profiles

The previous section considered pairwise comparisons only, so this section summarizes the results by specifying whether a detected protein in DEX or PHY has already been detected in one or more papers or it is identified for the first time within the liver circadian studies.

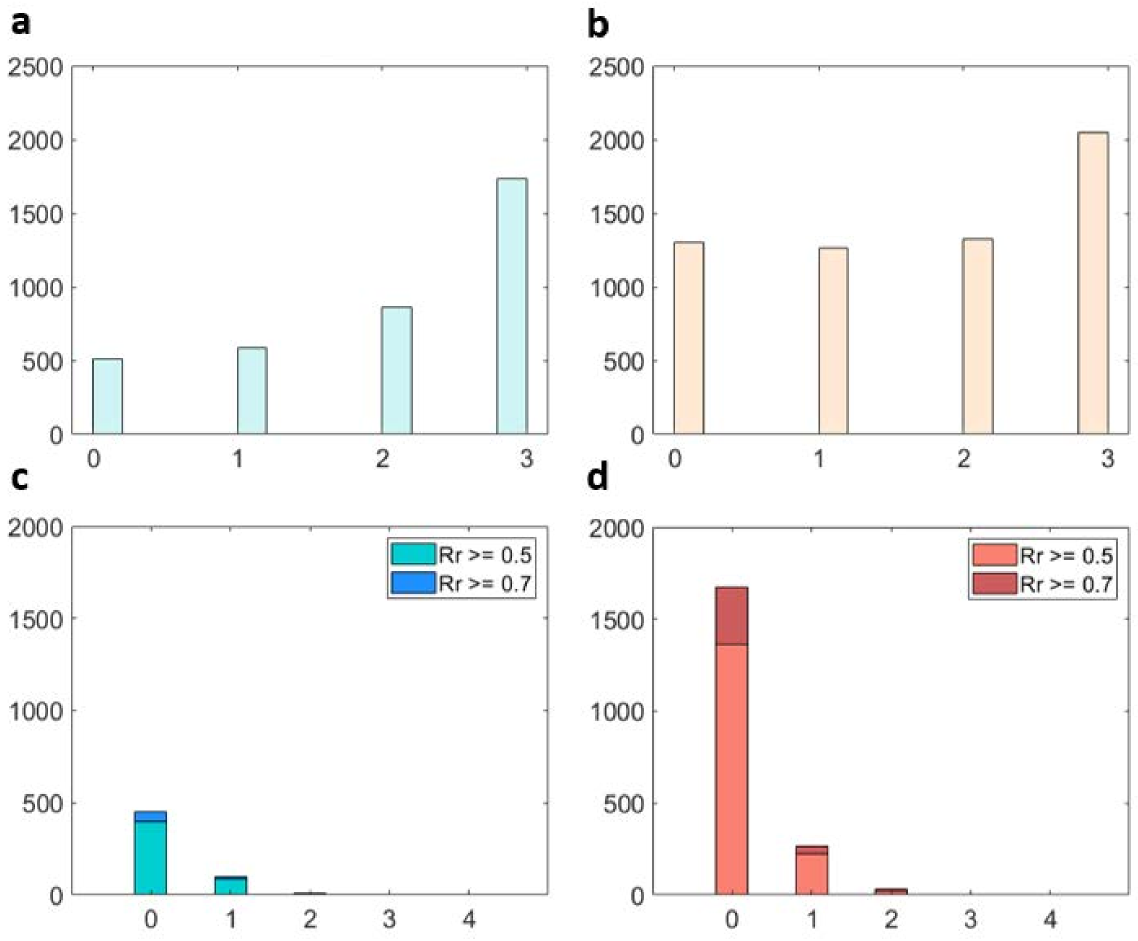
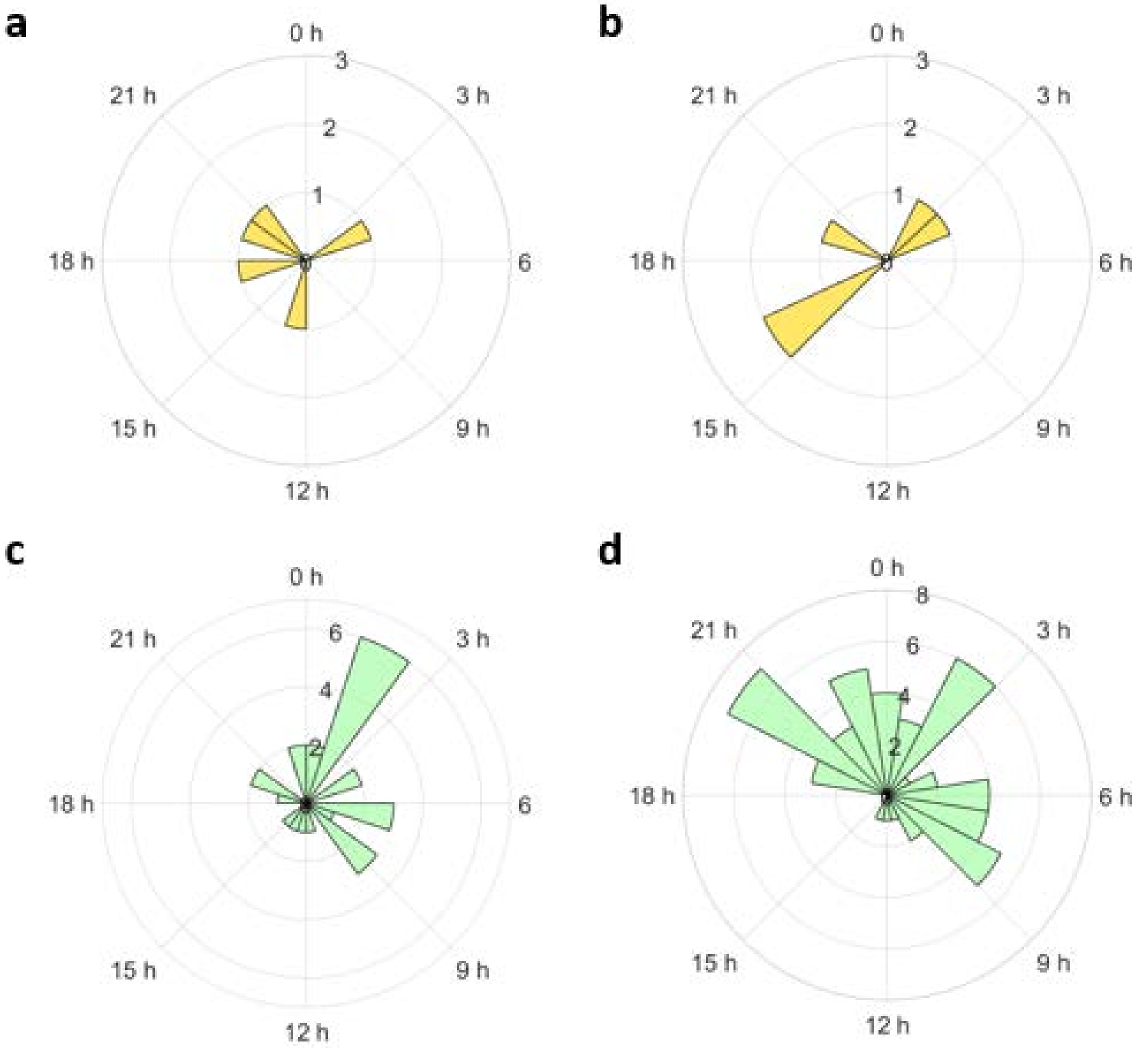


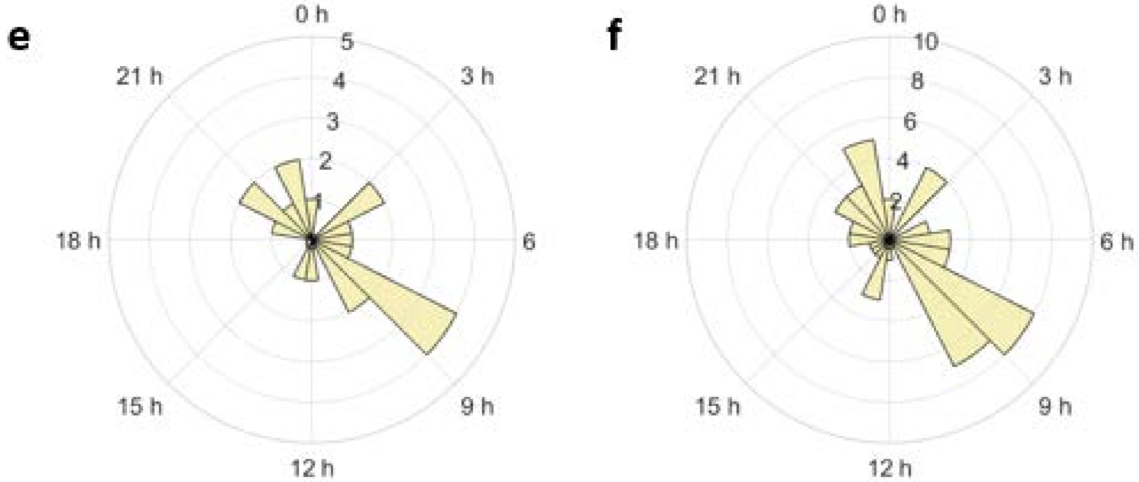
Figure 4.9 *Histograms showing the number of times a given protein (DEX protocol on the left, PHY protocol on the right) has been detected in literature. In particular, a) and b) refer to the whole datasets of detected proteins and it considers 0, 1, 2 or 3 number of citations because Reddy 2006 is not considered (only circadian proteins are available in that case). On the contrary, histograms c) and d) refer to DEX and PHY circadian proteins that have already been identified as circadian in literature sources. In particular, stacked bar plots are used to represent both circadian profiles (Rr >= 0.7) and profiles that are close to the threshold (Rr >= 0.5).*

As is shown in the histograms of Figure 4.9 a and b, when the whole set of measured proteins (circadian or not) is considered for DEX and PHY, there is a majority of proteins that have been already detected in at least one paper with respect to those that have never been measured before. In contrast, when circadian proteins are taken into account (Figure 4.9 c and d), both in the case of a relative radius above 0.5 and in the case of a relative radius above 0.7 there is a net predominance of proteins that are identified as circadian for the first time.

4.1.3.2 Common circadian proteins: phases comparison

After determining the intersection among circadian proteins in literature and in the DEX and PHY datasets of this Thesis, the corresponding phases are compared. In order to have a significant number of elements for the comparison, protein profiles with $Rr \geq 0.5$ are considered (instead of the more restrictive $Rr \geq 0.7$).


a
0 h
3
2
1
0
21 h
3 h
18 h
6
15 h
9 h
12 h
b
0 h
3
2
1
0
21 h
3 h
18 h
6 h
15 h
9 h
12 h
c
0 h
6
4
2
21 h
3 h
18 h
6
15 h
9 h
12 h
d
0 h
8
6
4
2
21 h
3 h
18 h
6 h
15 h
9 h
12 h


e
0 h
5
4
3
2
1
21 h
3 h
18 h
6
15 h
9 h
12 h
f
0 h
10
8
6
4
2
21 h
3 h
18 h
6 h
15 h
9 h
12 h

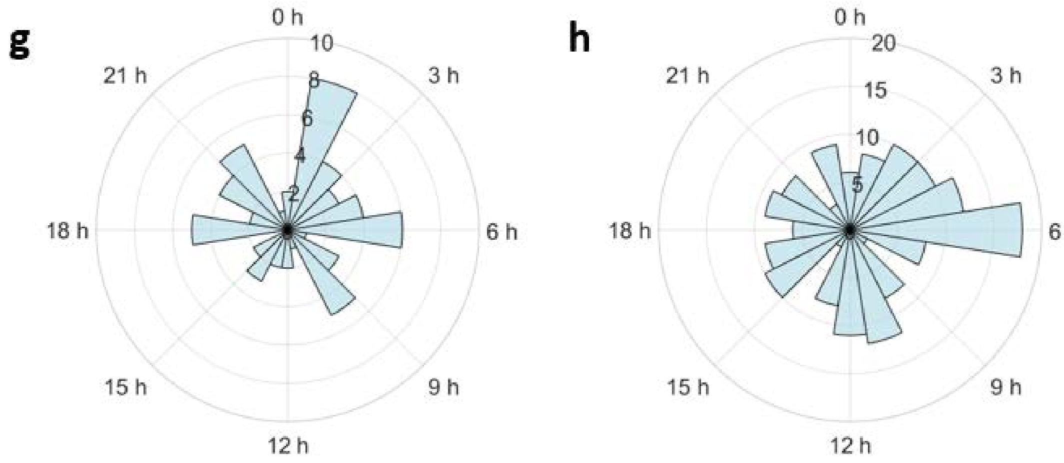


Figure 4.10 *Comparison of peak phases among proteins that have been identified as circadian in DEX or PHY protocol of this Thesis and in literature: on the left, polar histograms show the difference of peak phases between DEX circadian proteins and circadian proteins in Reddy 2006 (a), Mauvoisin 2014 (c), Robles 2014 (e) and Wang 2018 (g). On the right, polar histograms show the difference of peak phases between PHY circadian proteins and circadian proteins in Reddy 2006 (b), Mauvoisin 2014 (d), Robles 2014 (f) and Wang 2018 (h).*

The results of the comparisons between DEX data and literature are similar to those of the comparison between PHY data and literature. Indeed, all the eight pairwise comparisons of Figure 4.10 show that proteins that are identified as circadian both by the PCA model used for this Thesis and by the literature algorithms are characterized by phases estimation that differ greatly across different sources. It is not clear up to date what causes all these differences in peak phases for the same circadian proteins; therefore, in future researches it may be interesting to investigate in detail the role of different synchronization protocols and/or experimental setups in determining the timing of major functions of the liver circadian clock.

4.1.4 Conclusions on data variability across proteomics literature

When the four papers are compared with each other and with DEX and PHY data of this Thesis, it is important to remind that they are not directly comparable. Indeed, they are characterized by many differences:

- the studied organisms, e.g. mice or humans;
- the types of experiments, e.g. *in vivo* or *in vitro*;
- the synchronization protocols, e.g. light-darkness alternation, hormonal or feeding stimuli;
- the measurements techniques, e.g. 2D difference gel electrophoresis or mass spectrometry;
- the algorithms for the identification of circadian profiles, e.g. analysis of variance, Perseus software, JTK_Cycle or the PCA model described in this Thesis.

Nevertheless, all the pairwise comparisons presented in the previous sections are useful for providing a general picture of the available knowledge on the liver circadian proteome and for supporting the good quality of the experimental data analysed in this Thesis. Indeed, all the different steps of the analysis highlight how the inter-variability across different laboratories is relatively high, but it is entirely comparable between literature and DEX (or PHY) data or between pairs of literature sources among each other.

First, there is a high inter-variability with respect to the number of detected proteins during the experiment, due to the lack of full coverage in the high-throughput measurement technologies that are currently available in the proteomic field. Indeed, even Mauvoisin 2014 and Robles 2014 that employ the same technology, meaning SILAC mass spectrometry, measure different proteins, as well as the DEX and PHY experiments of this Thesis.

Moreover, a higher inter-variability is found when circadian proteins are compared: indeed, the identification of circadian proteins is even more influenced by the experimental setup than the proteins measurement itself. For example, it may vary across laboratories not only because of the technical limitations during the measurements (for example, when circadian proteins are not identified because their measurement failed), but also because of the influences of the synchronization protocol or because of the different performances of the algorithm in the identification of circadian rhythms. However, also in this case the percentages of intersection between DEX and PHY data with literature sources are analogue to those obtained by comparing literature sources among each other.

Finally, also when the same proteins are identified as circadian in two or more other sources, their peak phases change considerably (they may be even in antiphase), which could be related to the difficulty of determining active and resting phase within a cell *in vitro* system and to differences between human cells used in this study and mouse animal model, because mouse is a nocturnal animal (see Chapter 3, section 3.3). Last, discrepancy between phases also highlight the different role of different biological time keepers and are studied in more detail.

4.2 Proteomics analysis

After a critical evaluation of our results with literature in section 4.1, next sections only focus on the dynamics of the proteomics data of this Thesis. The first step consists of an explorative analysis of already normalized and filtered data, in order to evaluate the level of inter- and intra-variability between DEX and PHY profiles. Then, circadian proteins are identified by means of the PCA model, whose performance is evaluated by visualizing the selected proteins through ordered heatmaps. To complete the dynamics characterization, peak phases and amplitudes of DEX and PHY circadian profiles are compared. Finally, the attention focuses on the biological function of the selected circadian proteins, distinguishing between the main biological activities carried out in the active and resting phase.

4.2.1 Exploratory analysis

For the proteomics analysis, two datasets are provided: one for DEX and one for PHY, with the former showing a lower number of total proteins than the latter one due to the lack of full coverage during the measurements. Therefore, only the intersection (3691 proteins) between the two datasets is considered whenever DEX and PHY are compared. Moreover, the sampling is made of 4 replicates every 4 hours, starting at time 0 h and ending at time 20 h.

In order to visualize the level of intra- and inter- variability between DEX and PHY data, a PCA analysis is performed, considering also the standard samples for DEX and PHY normalization (Figure 4.11).

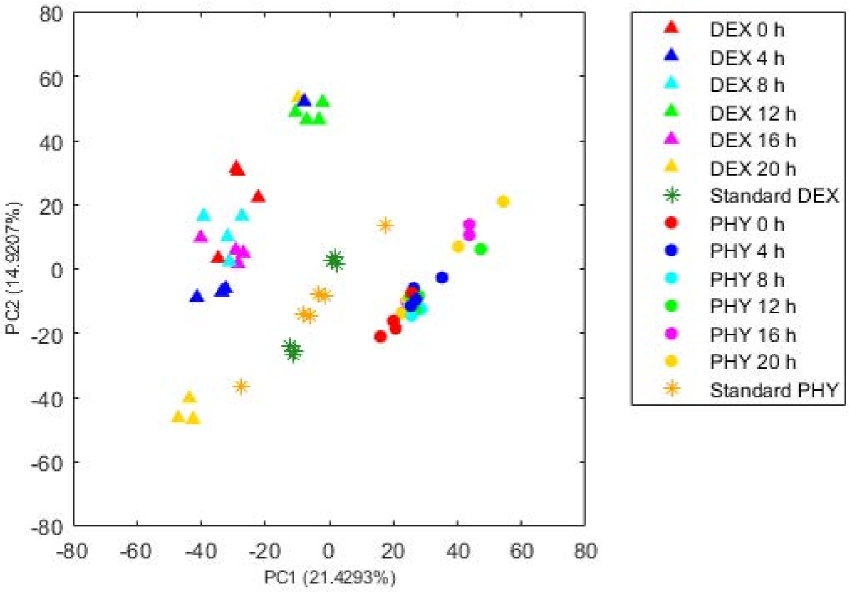


Figure 4.11 *PCA of normalized and filtered linear proteomics data, after scaling with Standard Normal Variate the matrix containing DEX, PHY and standards (i.e. pooled samples).This plot shows the scores of: DEX data (triangles), PHY data (circles), Standard data within DEX samples (green stars) and Standard data within PHY samples (orange stars). Only proteins in the intersection between DEX and PHY (3691 proteins) are considered.*

With respect to the corresponding PCA analysis performed with transcriptomic data (see Chapter 3, section 3.1) there is a lower level of intra-variability, because symbols corresponding to different time points within the same protocol tend to be mixed together in this case, while they formed clearly separated groups with transcriptomic data. This result may be due to the lower number of circadian profiles (as shown in section 4.2.2), which leads to a predominance of profiles that do not show significant changes in the expression profile across different time points. Moreover, the lower number of circadian profiles may be caused by a buffering effect at protein level due to post-translational modifications or degradation phenomena involving proteins themselves.

However, the major result found in transcriptomics is confirmed: indeed, also with proteomics the two protocols are separated with respect to the first principal component. Since the major source of variability in the experimental data is due to the synchronization method, it means that hormonal stimuli and feeding-fasting stimuli are able to generate an overall phenotypic difference.

Finally, another expected result is the fact that DEX and PHY standard data are projected around the origin of the score-plot, coherently with the fact that they are made up by mixing of all the other samples (as described in Chapter 2, section 2.1.3).

4.2.2 Circadianity of human liver proteome

The PCA model is used to estimate the relative radius (*Rr*) of each profile in DEX and PHY protocol and, consequently, their circadianity. The results are visualized as two-dimensional score plots, inside a clock representing the 24 hours of a complete circadian cycle (Figure 4.12). Since the percentage of circadian profiles is relatively low (as discussed in section 4.1), also profiles slightly below to the threshold of 0.7 are considered (*Rr* between 0.5 and 0.7).

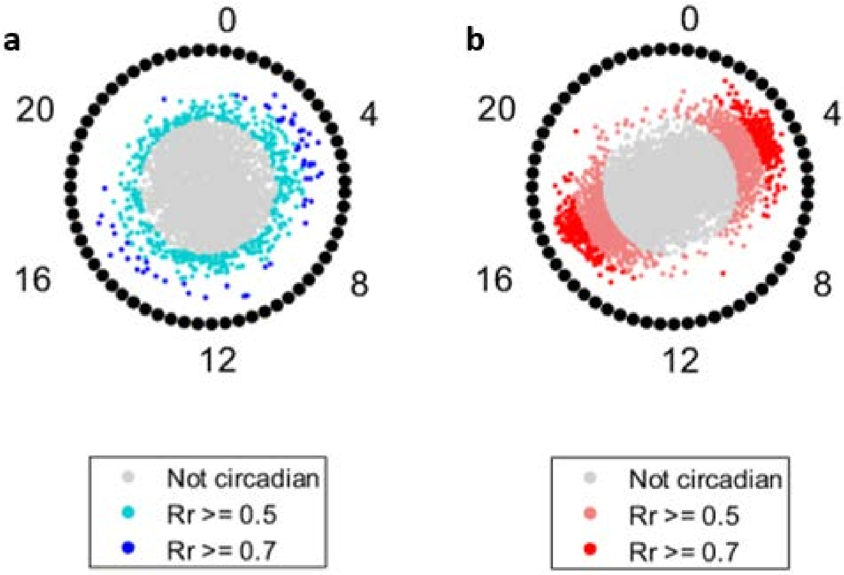


Figure 4.12 ***PCA model results***. *Two-dimensional score plot of the barycentres of DEX proteomic data (a) and PHY proteomic data (b), distinguishing among non-circadian profiles, circadian profiles and profiles slightly below the threshold.*

As anticipated in section 4.1, these results show that, relative to the total number of proteins detected, the physiological protocol induces an oscillatory behavior of more proteins than the Dexamethasone one: 6.2% and 1.6%, respectively, when proper circadian proteins are considered (*Rr* ≥ 0.7), and 27.2% and 13.5%, respectively, when also profiles close to the threshold are included (*Rr* ≥ 0.5). Moreover, PHY protocol causes a higher level of phase polarization: the majority of PHY profiles peak close to 4 h and 16 h, while phases of DEX profiles still display a certain polarization, but with a higher dispersion. Finally, the shorter list of circadian proteins (i.e. the DEX one) is not a subset of the longer one (i.e. the PHY list), but there is an intersection corresponding to the 28.3% of the total number of circadian proteins in DEX and to 4.6% of the total in PHY, or 41.8% and 12.9% for DEX and PHY when *Rr*≥0.5, as shown in the Venn diagrams of Figure 4.13.

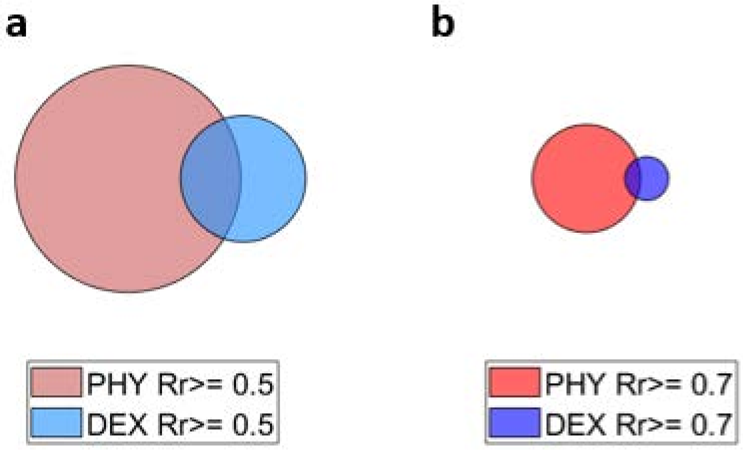


Figure 4.13 *Circadian proteins. Venn diagrams showing the intersection of proteins in DEX and PHY protocols slightly below the threshold (a) and of properly circadian proteins in the two protocols (b). In particular, the two Venn diagrams have the same axes, therefore bigger circles of (a) actually represent a higher number of proteins with respect to those of (b).*

In order to evaluate the quality of circadian profile selection through the PCA model, different ordered heatmaps are built (Figure 4.14) considering the specific cases of Table 4.5.

Table 4.5 ***Heatmaps description.*** *Groups of data that are used to build the heatmaps of Figure 4.14.*

Figure 4.14	Circadian PHY	Circadian DEX	Scaling
a	no	yes	DEX
b	yes	no	PHY
c	yes	yes	DEX and PHY separately
d	no	no	DEX and PHY

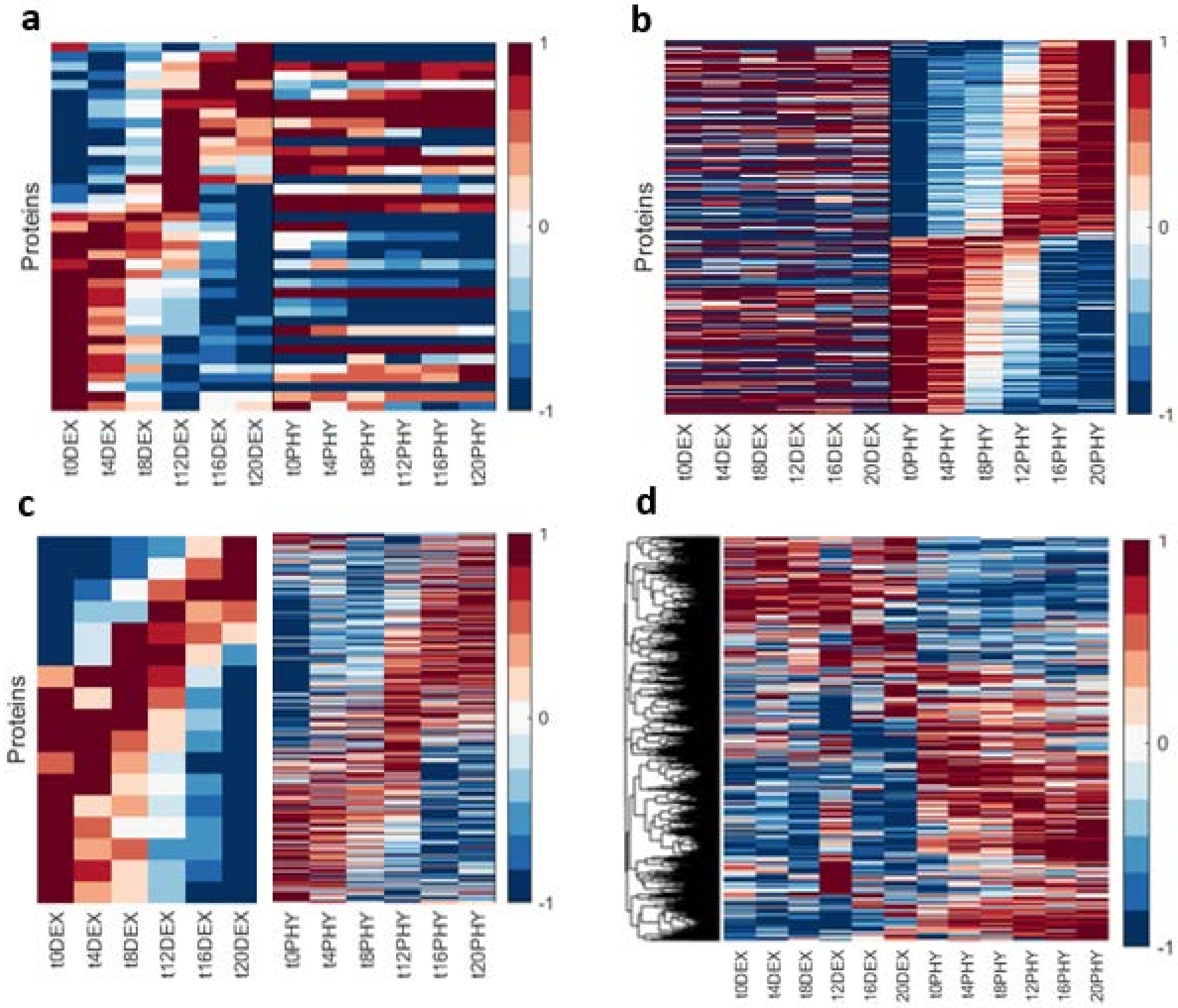


Figure 4.14 *Heatmap of proteins classified as circadian or not in DEX and PHY conditions by the PCA model. Heatmaps a, b, c and d refer to the 4 conditions in Table 4.5. In particular, data are scaled between -1 and 1 and ordered with respect to phases by using DEX data in a, PHY data in b, DEX data on the left half of c, PHY data on the right half of c (thus, the two halves displayed in c have different genes orderings). Instead, heatmap d is built by scaling DEX and PHY data between -1 and 1 and by ordering proteins with hierarchical clustering, because phase has little meaning for non-circadian profiles. In all heatmaps, different colors/color intensities represent different values as indicated by the legend on the right.*

Figure 4.14 a and b show that circadian profiles in only one protocol are correctly identified, while the corresponding random profiles of the other condition do not show a specific pattern. In particular, there is not a majority of random profiles being constantly high or constantly low depending on the phases of the corresponding circadian profiles, as happened with both DEX and PHY circadian profiles in the transcriptomics analysis (see Chapter 3, section 3.2.2).

Moreover, Figure 4.14 c, besides confirming the good selection of circadian profiles, shows that with PHY measurements there is a higher synchrony, confirming the result shown in the polar diagram. Finally, Figure 4.14 d represents non-circadian profiles in DEX and PHY protocols: the majority of profiles seem to take on higher values in PHY protocol, but no particular patterns are present.

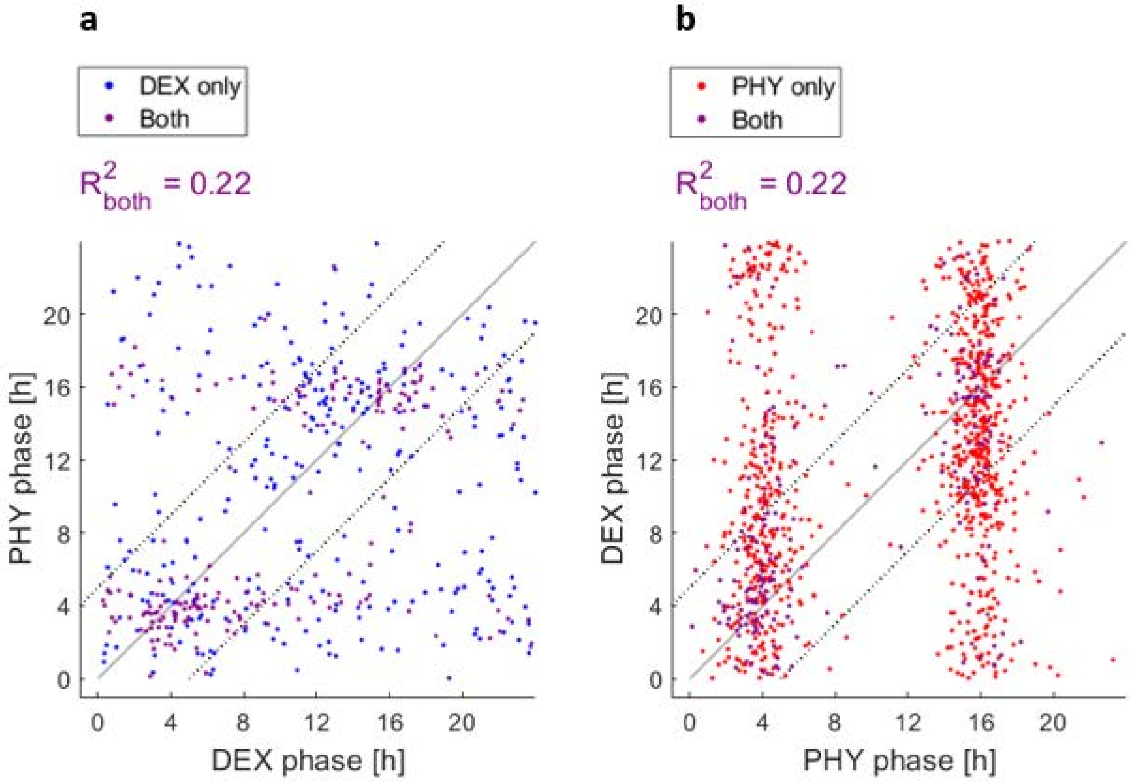
4.2.3 Phases and Amplitudes

The aim of this section is to characterize the main differences in the dynamics induced by DEX and PHY synchronization protocols, which in turn guide the division of the whole dataset into different protein lists for the enrichment analysis. Therefore, two parameters are compared between the two protocols, phase and amplitude, considering the cases of Table 4.6.

Table 4.6 *Scheme of the comparisons between DEX and PHY data, considering two key parameters for circadian profiles: phase and amplitude. Figures 4.15 shows the corresponding results.*

Figure 4.15	Parameter	Circadian data
a	Phase	Only DEX or both
b	Phase	Only PHY or both
c	Amplitude	Only DEX or both
d	Amplitude	Only PHY or both

Moreover, in order to have a relevant number of proteins for the comparison, not only circadian proteins are considered ($Rr \geq 0.7$), but also proteins slightly below the threshold of circadian rhythms ($Rr \geq 0.5$).



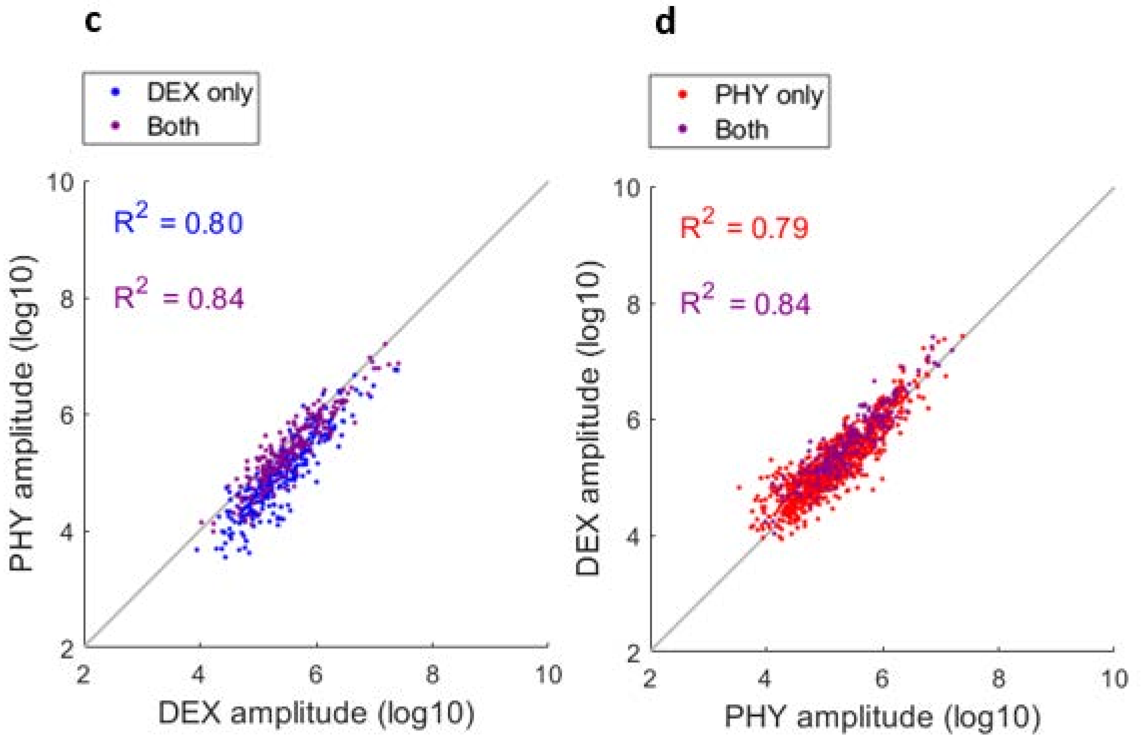


Figure 4.15 *Comparison between DEX and PHY. Phases and amplitudes are compared between DEX and PHY protocols, considering circadian profiles and profiles slightly below the threshold (Rr ≥ 0.5) as indicated in Table 4.6.*
In all subplots, each dot represents the parameter value (i.e. phase or amplitude) in the two protocols, the solid diagonal represents equal values, while dotted lines in the phase dot plots represent differences of 5 h from equality. Finally, R^2 are shown, whose colors refer to the data indicated by the legend.

The predominant trend in Figure 4.15 a and b is the polarization of peak phases at 4 h and 16 h in PHY protocol, while DEX phases are still concentrated between 0 h and 8 h and between 10 h and 18 h (especially in b), but a relevant number of them is evenly distributed outside those ranges. Therefore, another phase visualization is made at the beginning of section 4.2.4 in order to assess whether it is reasonable or not to divide DEX data into the same time intervals of PHY data.

Instead, the comparison among amplitudes (in *log$_{10}$* scale) leads to opposite results: indeed, amplitudes of protein profiles in DEX and PHY protocols are highly correlated, with a R^2 of 0.84 in case of proteins with *Rr* ≥ 0.5 in both protocols and 0.80 and 0.79 when *Rr* ≥ 0.5 only in DEX or PHY protocol, respectively. Therefore, amplitude is not a useful parameter to characterize the differences between DEX and PHY protein profiles.

4.2.4 Biological interpretation

The aim of this paragraph is to analyse the biological functions of DEX and PHY circadian proteins. To do so, the whole data are divided into sub-groups depending on their peak of expression, that potentially corresponds to their time of action within the cell. Moreover, the enrichment analysis is performed on DEX and PHY data independently, considering strictly circadian proteins with $Rr \geq 0.7$. First of all, to better quantify the protein polarization across 24h mentioned above and to select what time points of peaking to cluster together for functional analysis, the polar histograms for the two protocols are shown in Figure 4.16. Moreover, Figure 4.17 shows the difference in peak phases of circadian proteins in both protocols (with $Rr \geq 0.5$ for having a relevant number of proteins).

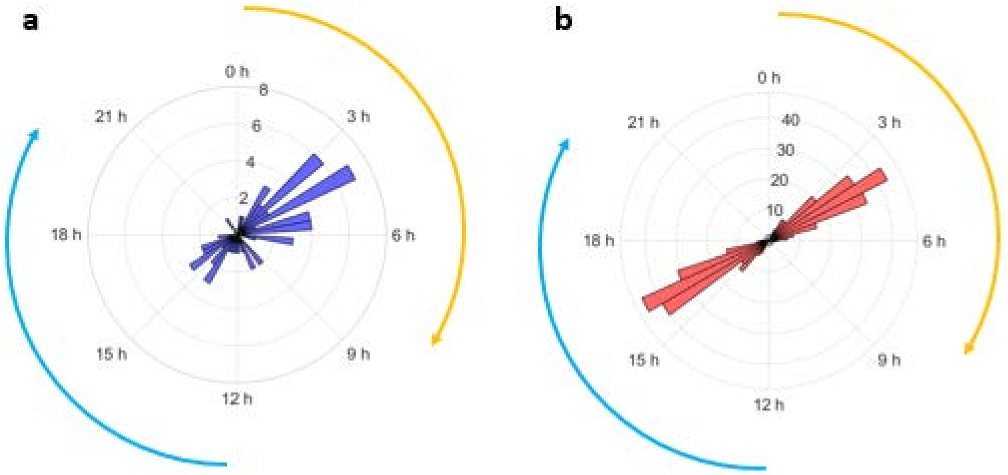


Figure 4.16 *Peak phases of all the circadian proteins (i.e. Rr ≥ 0.7) in DEX measurements (a) and in PHY measurements (b). In both cases, the orange arrows represent the active phase, while the light blue arrows represent the resting phase.*

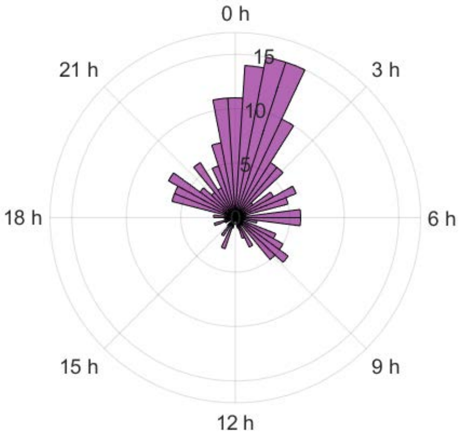


Figure 4.17 *Difference of peak phases between DEX and PHY proteins, considering the common proteins having Rr ≥ 0.5.*

As shown in the polar histograms, DEX phases are less polarized with respect to PHY ones, but still selecting only proteins with phases between 0 h and 8 h and between 12 h and 20 h does not cause too much loss of information. Moreover, when proteins have $Rr \geq 0.5$ in both protocols, they tend to be synchronized: indeed, as shown in Figure 4.17, the difference in peak phases is mainly below 3h.

To understand the role of proteins in these temporal intervals, a functional enrichment analysis is performed using the list of proteins shown in Table 4.7. The enrichment analysis is performed within Reactome database, which is a manually-curated characterization of genes/proteins involved in specific cellular functions and is organized according to a hierarchical structure. For ease of result interpretation, graphical results include the whole hierarchy even when not all levels are enriched (non-enriched ones are indicated in grey): this is useful for clustering biological functions and for making comparisons at different levels of biological function specificity.

Table 4.7 *Scheme of the enrichment analysis.*

Figure	Circadian Protocol	Phase
4.18 a	DEX	active
4.18 b and c	DEX	resting
4.19	PHY	active
4.20	PHY	resting

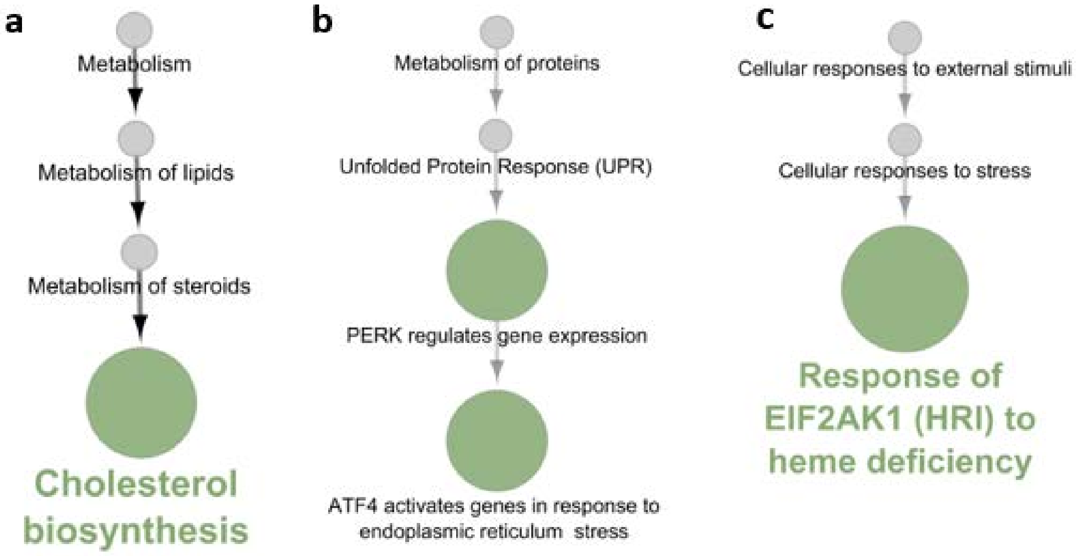
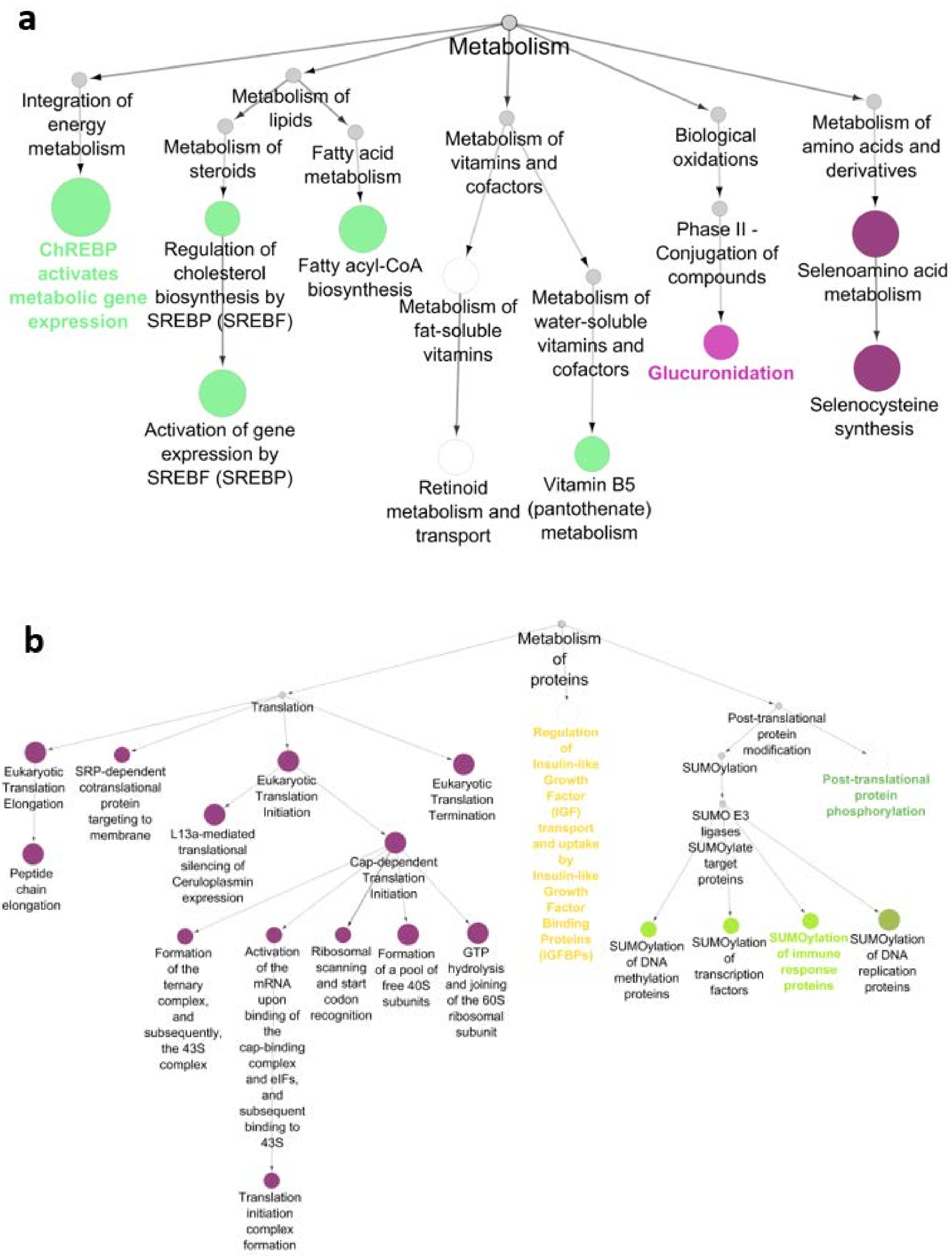

a
Metabolism
Metabolism of lipids
Metabolism of steroids
Cholesterol biosynthesis
b
Metabolism of proteins
Unfolded Protein Response (UPR)
PERK regulates gene expression
ATF4 activates genes in response to endoplasmic reticulum stress
c
Cellular responses to external stimuli
Cellular responses to stress
Response of EIF2AK1 (HRI) to heme deficiency

Figure 4.18 ***DEX proteins results.*** *REACTOME pathways that are enriched in DEX proteins peaking in the active phase (a) and in the resting phase (b, c): a) metabolism; b) metabolism of proteins; c) cellular response to external stimuli.*


a
Metabolism
Integration of energy metabolism
ChREBP activates metabolic gene expression
Metabolism of lipids
Metabolism of steroids
Regulation of cholesterol biosynthesis by SREBP (SREBF)
Activation of gene expression by SREBF (SREBP)
Fatty acid metabolism
Fatty acyl-CoA biosynthesis
Metabolism of vitamins and cofactors
Metabolism of fat-soluble vitamins
Retinoid metabolism and transport
Metabolism of water-soluble vitamins and cofactors
Vitamin B5 (pantothenate) metabolism
Biological oxidations
Phase II - Conjugation of compounds
Glucuronidation
Metabolism of amino acids and derivatives
Selenoamino acid metabolism
Selenocysteine synthesis
b
Metabolism of proteins
Translation
Eukaryotic Translation Elongation
Peptide chain elongation
SRP-dependent cotranslational protein targeting to membrane
Eukaryotic Translation Initiation
L13a-mediated translational silencing of Ceruloplasmin expression
Cap-dependent Translation Initiation
Formation of the ternary complex, and subsequently, the 43S complex
Activation of the mRNA upon binding of the cap-binding complex and eIFs, and subsequent binding to 43S
Translation initiation complex formation
Ribosomal scanning and start codon recognition
Formation of a pool of free 40S subunits
GTP hydrolysis and joining of the 60S ribosomal subunit
Eukaryotic Translation Termination
Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)
Post-translational protein modification
SUMOylation
SUMO E3 ligases SUMOylate target proteins
SUMOylation of DNA methylation proteins
SUMOylation of transcription factors
SUMOylation of immune response proteins
SUMOylation of DNA replication proteins
Post-translational protein phosphorylation

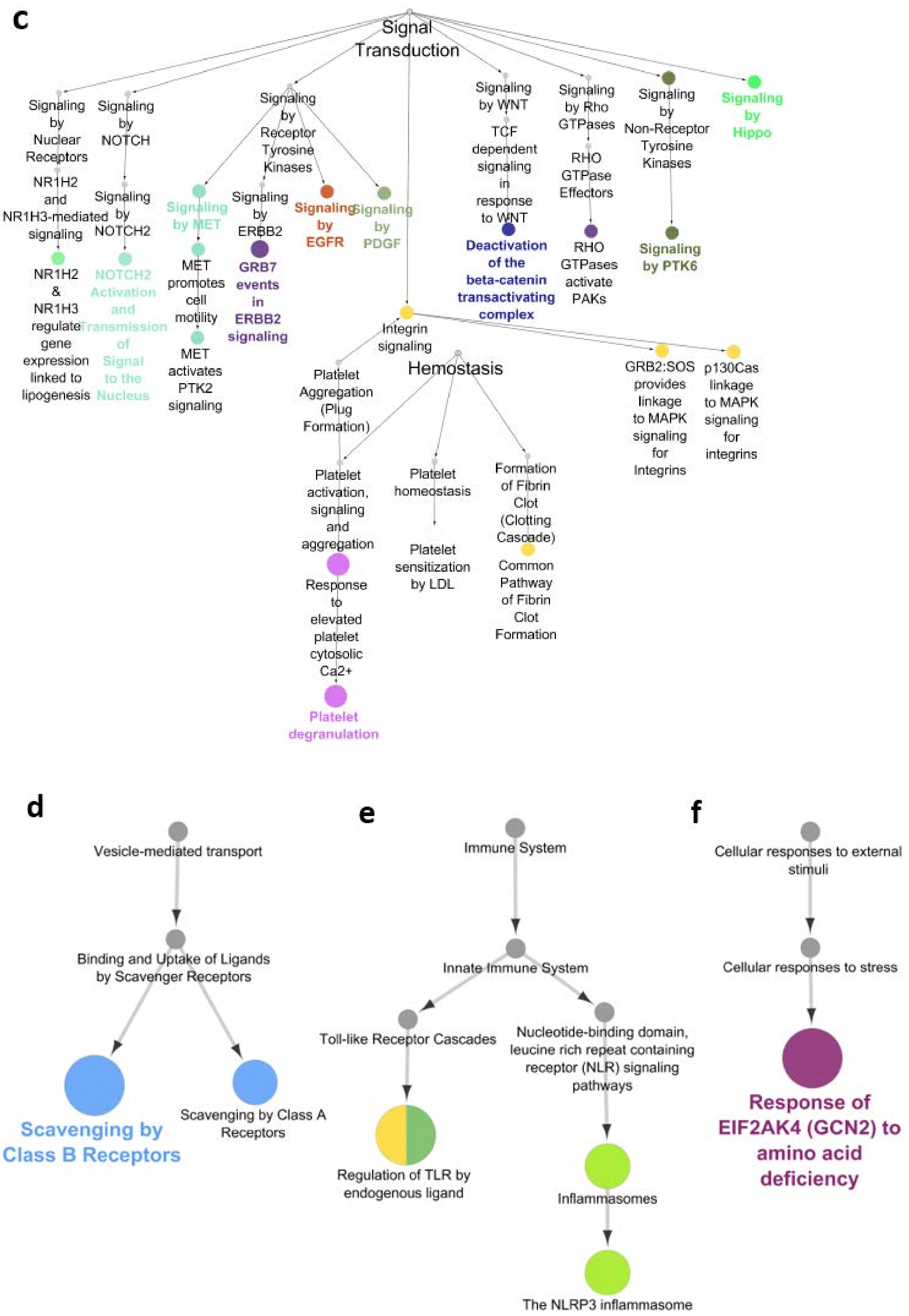
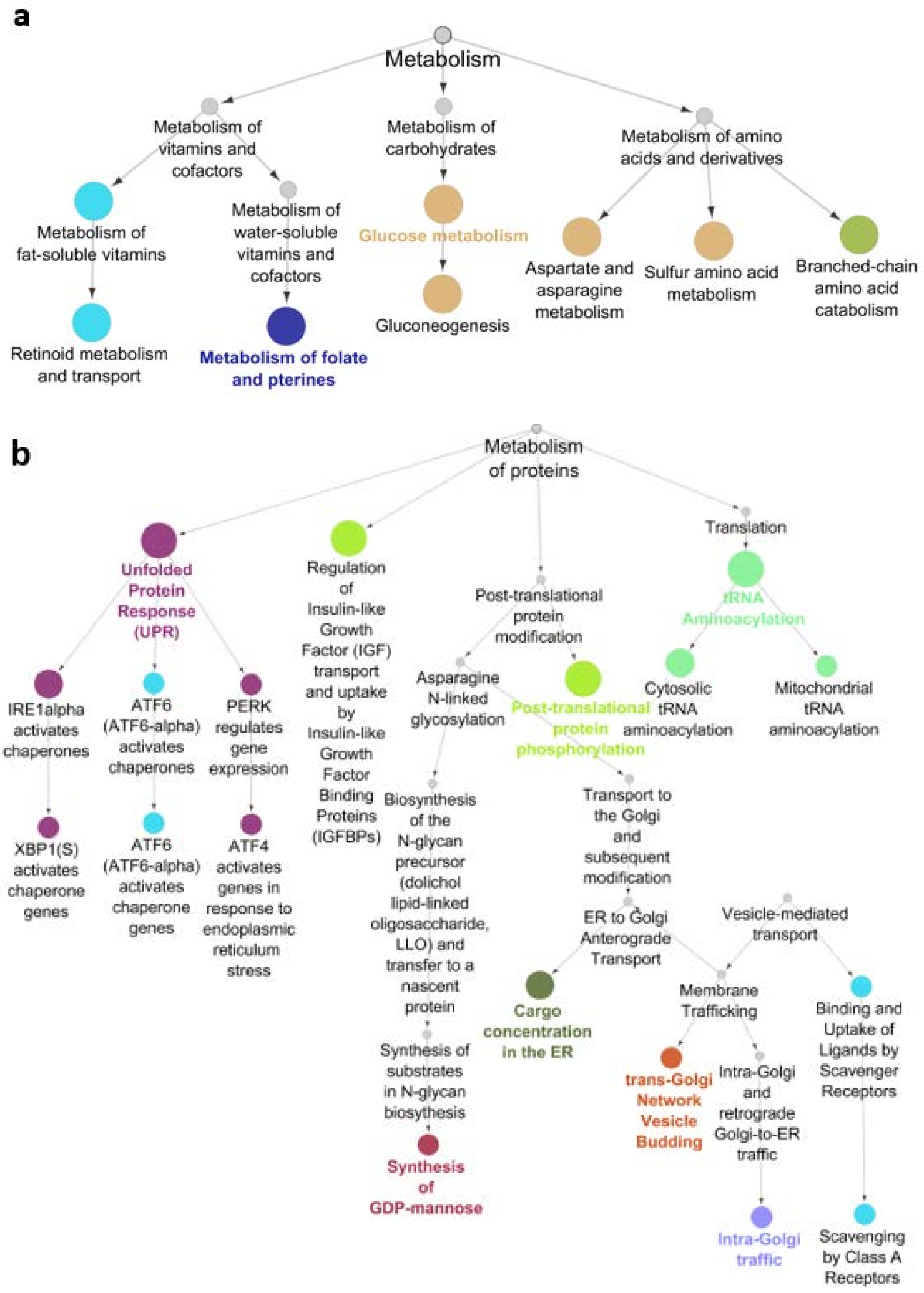

c
Signal Transduction
Signaling by Nuclear Receptors
Signaling by NOTCH
Signaling by Receptor Tyrosine Kinases
Signaling by WNT
Signaling by Rho GTPases
Signaling by Non-Receptor Tyrosine Kinases
Signaling by Hippo
NR1H2 and NR1H3-mediated signaling
Signaling by NOTCH2
Signaling by MET
Signaling by ERBB2
Signaling by EGFR
Signaling by PDGF
TCF dependent signaling in response to WNT
RHO GTPase Effectors
Signaling by PTK6
NR1H2 & NR1H3 regulate gene expression linked to lipogenesis
NOTCH2 Activation and Transmission of Signal to the Nucleus
MET promotes cell motility
GRB7 events in ERBB2 signaling
Deactivation of the beta-catenin transactivating complex
RHO GTPases activate PAKs
MET activates PTK2 signaling
Integrin signaling
Hemostasis
Platelet Aggregation (Plug Formation)
GRB2:SOS provides linkage to MAPK signaling for Integrins
p130Cas linkage to MAPK signaling for integrins
Platelet activation, signaling and aggregation
Platelet homeostasis
Formation of Fibrin Clot (Clotting Cascade)
Platelet sensitization by LDL
Common Pathway of Fibrin Clot Formation
Response to elevated platelet cytosolic Ca2+
Platelet degranulation
d
Vesicle-mediated transport
Binding and Uptake of Ligands by Scavenger Receptors
Scavenging by Class B Receptors
Scavenging by Class A Receptors
e
Immune System
Innate Immune System
Toll-like Receptor Cascades
Nucleotide-binding domain, leucine rich repeat containing receptor (NLR) signaling pathways
Regulation of TLR by endogenous ligand
Inflammasomes
The NLRP3 inflammasome
f
Cellular responses to external stimuli
Cellular responses to stress
Response of EIF2AK4 (GCN2) to amino acid deficiency

Figure 4.19 *Some of the major REACTOME pathways that are enriched in PHY circadian proteins with phases between 0 h and 8 h: a) metabolism; b) metabolism of proteins; c) signal transduction and hemostasis; d)vesicle-mediated transport; e) immune system; f) cellular responses to external stimuli.*


a
Metabolism
Metabolism of vitamins and cofactors
Metabolism of carbohydrates
Metabolism of amino acids and derivatives
Metabolism of fat-soluble vitamins
Metabolism of water-soluble vitamins and cofactors
Glucose metabolism
Aspartate and asparagine metabolism
Sulfur amino acid metabolism
Branched-chain amino acid catabolism
Gluconeogenesis
Retinoid metabolism and transport
Metabolism of folate and pterines
b
Metabolism of proteins
Unfolded Protein Response (UPR)
Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)
Post-translational protein modification
Translation
tRNA Aminoacylation
IRE1alpha activates chaperones
ATF6 (ATF6-alpha) activates chaperones
PERK regulates gene expression
Asparagine N-linked glycosylation
Post-translational protein phosphorylation
Cytosolic tRNA aminoacylation
Mitochondrial tRNA aminoacylation
XBP1(S) activates chaperone genes
ATF6 (ATF6-alpha) activates chaperone genes
ATF4 activates genes in response to endoplasmic reticulum stress
Biosynthesis of the N-glycan precursor (dolichol lipid-linked oligosaccharide, LLO) and transfer to a nascent protein
Transport to the Golgi and subsequent modification
ER to Golgi Anterograde Transport
Vesicle-mediated transport
Cargo concentration in the ER
Membrane Trafficking
Binding and Uptake of Ligands by Scavenger Receptors
Synthesis of substrates in N-glycan biosythesis
trans-Golgi Network Vesicle Budding
Intra-Golgi and retrograde Golgi-to-ER traffic
Synthesis of GDP-mannose
Intra-Golgi traffic
Scavenging by Class A Receptors

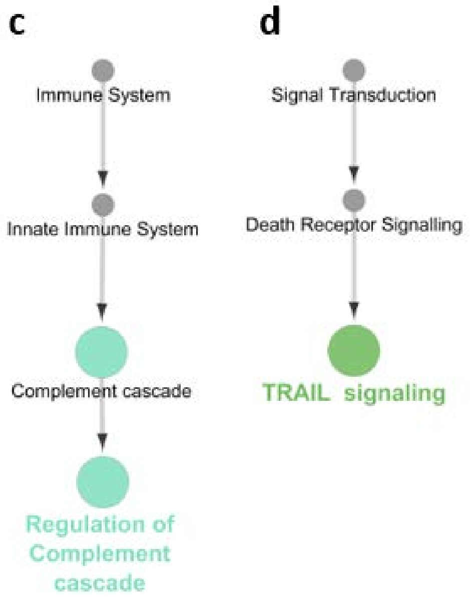


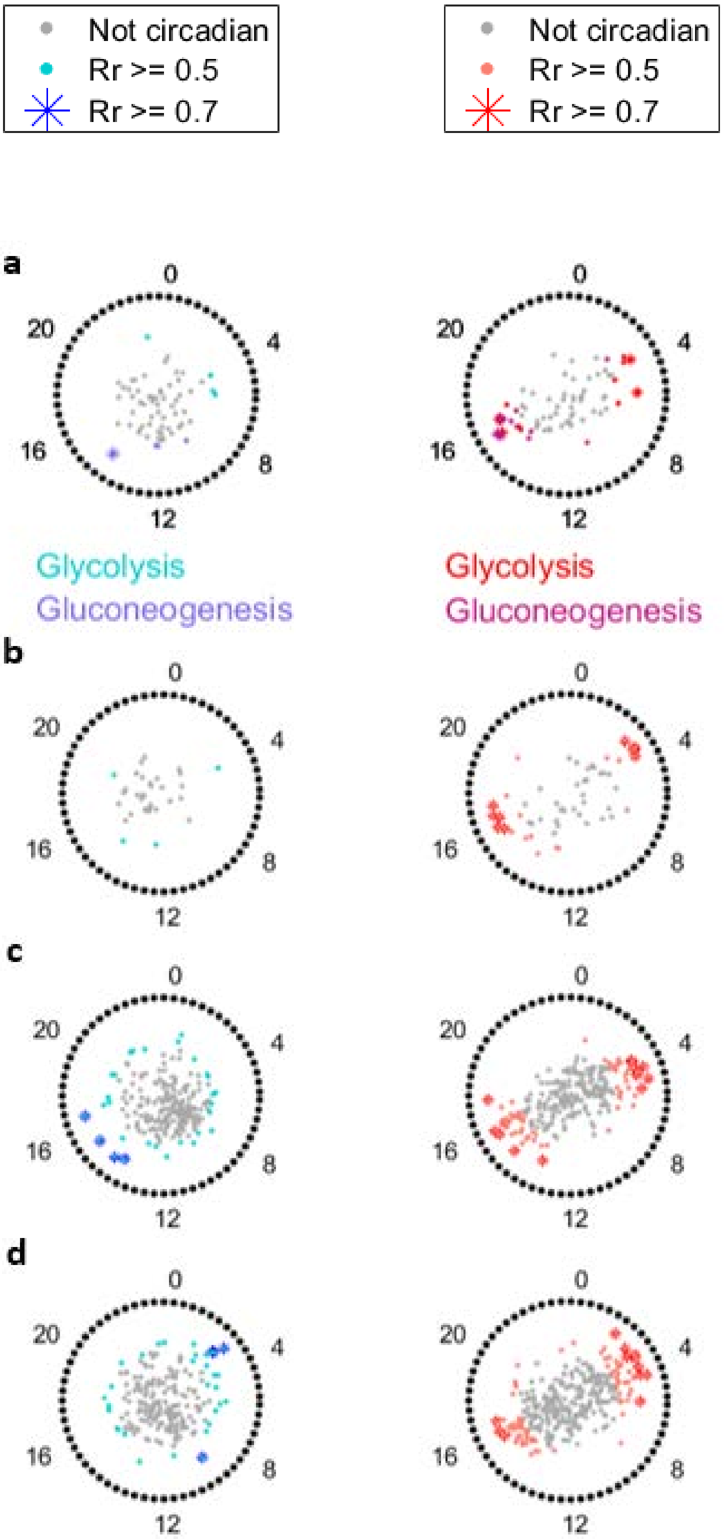
Figure 4.20 *Some of the major REACTOME pathways that are enriched in the PHY circadian proteins with phases between 12 h and 20 h: a) metabolism; b) metabolism of proteins and vesicle-mediated transport; c) immune system; d) signal transduction.*

The enrichment analysis of DEX data provides results that are consistent with those of transcriptomics (Chapter 3, section 3.4), both in terms of the major enriched pathways and in terms of the time intervals in which they are activated. The main difference is the lower number of enriched pathways due to the lower number of circadian proteins identified. Indeed, only one pathway is enriched in DEX protocol peaking between 0 h and 8 h, that is metabolism of lipids and, specifically, of cholesterol-related processes. While between 12 h and 20 h there are two enriched pathways: metabolism of proteins and cellular response to stress. The activation of metabolism pathways during the active phase and of metabolism of proteins during the resting phase is consistent with the results of transcriptomic data, which supports the assumption that the definition of ‘active phase’ (i.e. 0-8h) and ‘resting phase’ (i.e. 12-20h) overlaps for transcriptomics and proteomics, meaning that the delay due to protein translation is negligible for this purpose.

As regards PHY data, a higher number of enriched pathways is present, due to the higher number of circadian proteins. In particular, pathways of metabolism and metabolism of proteins are present both in the active and resting phase, even though the number of pathways of metabolism is still higher in 0-8 h; moreover, signal transduction is enriched all day long, with a higher number of functions activated during the active phase.

Since metabolism and metabolism of proteins are enriched in both protocols and are quite generic functions, they are analysed further. In particular, a series of PCA score plot is built by considering the genes that are annotated to those pathways in the REACTOME hierarchy

(Figure 4.21 and 4.22, which retain only pathways with a significant number of circadian proteins).



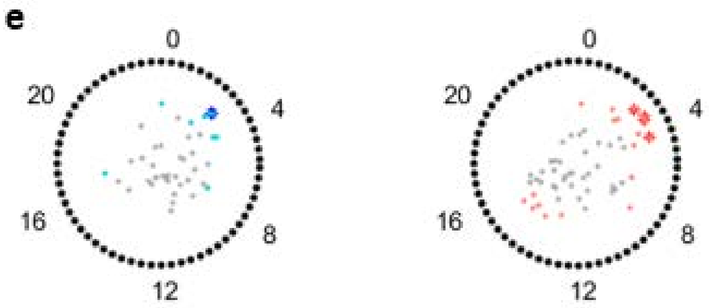
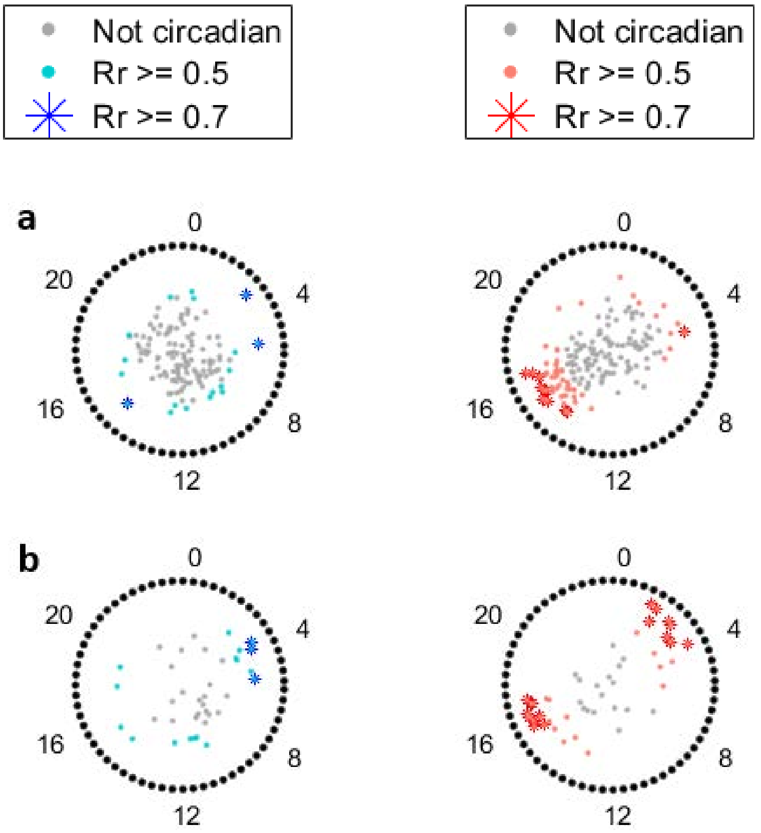


Figure 4.21 ***REACTOME pathway: Metabolism***. *Each dot represents genes annotated to the following pathways:*

a) glycoslysis and gluconeogenesis, two categories of glucose metabolism;
b) metabolism of water-soluble vitamins and cofactors;
c) metabolism of amino acids and derivatives;
d) metabolism of lipids;
e) metabolism of steroids.



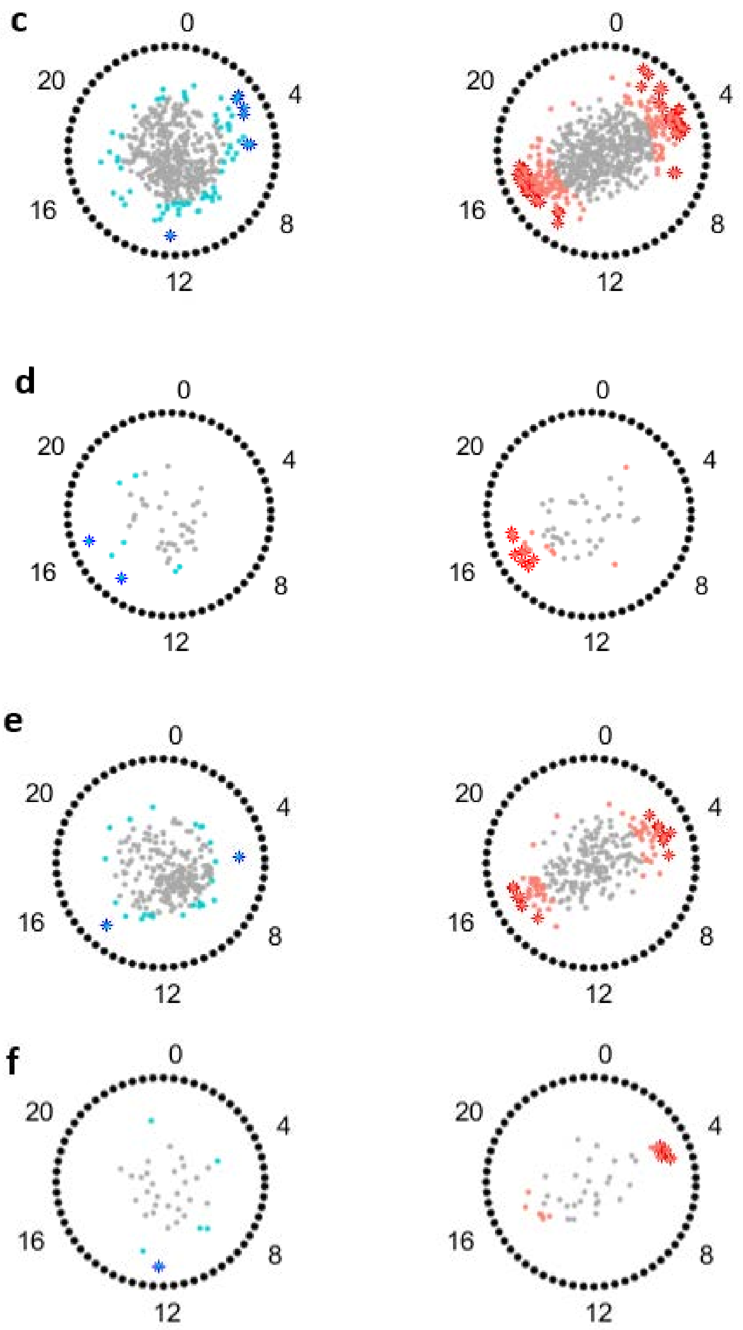


Figure 4.22 ***REACTOME pathway: Metabolism of Proteins.*** *PCA–score plots are built considering genes annotated to the pathways:*

a) Asparagine N-linked glycosylation;
b) Post-translational protein phosphorylation;
c) Post-translational protein modification;
d) Unfolded Protein Response (UPR);
e) Translation;
f) SUMOylation of DNA replication proteins.

As regards metabolism pathways, the number of circadian proteins in PHY protocol is higher than in DEX protocol; in particular, two major trends are found with PHY peak phases:

- phases evenly distributed between active and resting phase (i.e. glucose metabolism, metabolism of water-soluble vitamins and co-factors, metabolism of amino acids and derivatives, Metabolism of lipids). Notably, looking at specific functions within glucose metabolism, PHY protocol clearly shows glycolysis function during the active phase and gluconeogenesis during the resting phase, what exactly matches liver glucose metabolism *in vivo;*
- phases concentrated during the active phase (i.e. metabolism of steroids). Of note, metabolism of steroids is a sub-category of metabolism of lipids and results show how specific subfunctions are highly temporally regulated.

Considering metabolism of proteins, DEX has a low number of annotated proteins that are circadian and their peak phases are partly in the active and partly in the resting phase, without a clear trend. In contrast, a higher number of PHY proteins involved in metabolism of proteins are circadian in PHY protocol and their peak phases may be:

- concentrated at both 4 h and 16 h (i.e. translation, post-translational protein modification, post-translational protein phosphorylation);
- concentrated in the resting phase (i.e. Asparagine N-linked glycosylation, Unfolded Protein Response);
- concentrated in the active phase (i.e. SUMOylation of DNA replication proteins).

4.2.5 Summary of proteomics results

The exploratory analysis of proteomics reveals that the major source of variability is given by the synchronization protocol itself; in contrast, there is no net separation among observations due to the six time points between 0 h and 20 h, suggesting that the majority of profiles could be random (i.e., non-circadian). This assumption is confirmed by the PCA model results, because only 6.2% and 1.6% are circadian in PHY and DEX protocols, respectively (considering $Rr \geq 0.7$). The main reason of these low percentages is likely the fact that circadian proteins are involved in regulatory functions, so they are fewer in quantity than proteins that carry out basic functions, like maintaining cell structure.

Moreover, when DEX and PHY profiles are compared in terms of amplitude, a high correlation is found (R^2 between 0.79 and 0.84), meaning that if a protein displays high oscillations in one protocol, it is likely that it oscillates in the other protocol, too.

As regards peak phases, the interesting result is that proteins with $Rr \geq 0.5$ in both protocols have a phase difference of 0 h - 3 h between DEX and PHY. Therefore, even though the two synchronization protocols activate different percentages of circadian profiles, proteins that are circadian in both tend to be synchronized, meaning that they carry out their biological functions in the same phase (i.e. active or resting). Moreover, the innovative protocol (i.e. PHY) not only activates a higher percentage of circadian rhythms, but also determines a more marked polarization of peak phases, with a main peak at 4 h and another one at 16 h. However, this conclusion needs to be further confirmed by future experiments using full-coverage measurement technologies: indeed, some DEX circadian proteins may be present but not detected and they may also influence the bimodal distribution of peak phases in this protocol.

As regards proteins biological functions, the enrichment analysis of circadian profiles peaking in the active (i.e. 0-8 h) and resting phase (i.e. 12-20 h) highlight the presence of the same pathways encountered in literature, like metabolism pathways, e.g. metabolism of lipids, and metabolism of proteins, e.g. Unfolded Protein Response (UPR) and post-translational protein modifications. This is a further proof that the innovative protocol is able to ensure the key biological functions of liver circadian clock, without worsening the performance of the *in vitro* system with respect to the reference protocol. From the applicative point of view, these results can be interpreted as the ability of metabolic stimuli to synchronize the proteome of the human liver clock in an efficient way and independently from hormonal stimuli.

Chapter 5
Circadian proteins and transcripts

The aim of this chapter is to integrate the analysis of the dynamics of DEX and PHY transcripts with their corresponding proteins. In particular, transcriptomic and proteomic results are compared in terms of circadianity and oscillatory properties (phase and amplitude).

Analysing proteins directly and understanding if and to what extent circadian proteins are linked to circadian mRNAs is useful for providing a general picture of circadian rhythms in human liver cells. Indeed, even though transcriptomics alone may be easier to study thanks to more mature measurement as well as bioinformatics techniques, proteins are the real 'actors' within cells and they need to be described properly. However, this cannot be done simply by extrapolating the results of transcriptomics, because there are relevant differences in their dynamics in terms of percentage of circadian profiles and in terms of shifted peak phases, as demonstrated in this chapter.

5.1 Circadian profiles

In the datasets of Chapter 3 and 4, the total number of measured profiles (after filtering) was:

- 12940 transcripts in both DEX and PHY protocols;
- 3691 proteins in DEX protocol;
- 5895 proteins in PHY protocol.

However, all the comparisons of this chapter are made considering only the intersection between measured transcripts and proteins in DEX (i.e. 3665 genes/proteins) and in PHY protocol (i.e. 5841).

In particular, some possibilities are explored, for example the fact that circadian mRNAs may induce oscillations also in their proteins or that downstream processes, such as protein degradation, may give rise to a buffering effect dampening oscillation. In addition, the opposite relation between proteins and transcripts is considered, i.e. the possibility that proteins oscillations are due to the oscillatory behavior of the corresponding transcripts. Finally, in case of transcripts and proteins that are both circadian, the maintenance of the phase of peaking is verified.

First of all, the percentages of circadian profiles in the four datasets of this Thesis are compared (Table 5.1). These results show that the percentages of circadian profiles in transcriptomics is one order of magnitude higher than in proteomics: indeed, circadian transcripts are 40 % in DEX and 25 % in PHY, while circadian proteins are 1.6 % in DEX and 6.2 % in PHY. Interestingly, the hormonal synchronization (i.e. DEX protocol) has opposite effects with mRNAs and with proteins: indeed, in the former case it is able to induce oscillatory dynamics twice as much as the feeding-fasting stimuli (PHY protocol), while in the latter case the percentage of oscillatory dynamics is four times lower.

However, the comparisons between transcripts and proteins and between the two protocols may be influenced by measurements limitations: indeed, proteins that are more abundant have more possibilities of being detected, e.g. structural proteins that allow cells to maintain their shape, but they are not expected to be circadian. In contrast, it is likely that proteins displaying a circadian behaviour are those having regulatory functions, which are usually less abundant and therefore more difficult to be measured.

As regards the possibility of obtaining circadian rhythms in final proteins when their corresponding transcripts are circadian, it seems to be just the minority of cases: indeed, only 2.3% of DEX circadian transcripts generates circadian proteins and only 11% of PHY circadian transcripts generates circadian proteins. In contrast, an intermediate situation is found when circadian proteins are compared with their initial transcripts: indeed, 56 % and 44 % of circadian proteins in DEX and PHY protocol, respectively, are generated by circadian transcripts. However, both results will become more grounded in the future if they are confirmed by full-coverage proteomics analyses. Indeed, since this discussion does not involve all 12940 transcriptomic profiles, but only the intersection of DEX mRNAs and proteins and the intersection of PHY mRNAs and proteins, as shown in Table 5.2, the total number of circadian transcripts results to be 1468 over 4570 total circadian transcripts (i.e. 32%) in DEX and 1467 over 2812 total circadian transcripts (i.e. 52%) in PHY. Therefore, the results may change (or may be further confirmed) if the proteins corresponding to the excluded circadian transcripts will be available.

Table 5.1 *Percentages of circadian mRNAs and proteins in DEX and PHY protocols, considering only genes and proteins identified both by RNA-Seq and mass spectrometry in each of the two protocols (DEX and PHY).*

	CIRCADIAN TRANSCRIPTS (%)	CIRCADIAN PROTEINS (%)
DEX PROTOCOL	40.05	1.63
PHY PROTOCOL	25.12	6.23

Table 5. 2 *Absolute values that are used to calculate percentages in the Table 5.3. 'T' stands for 'transcripts', 'P' stands for 'proteins', 'circ' stands for 'circadian', 'not circ' stands for 'not circadian'; in particular, the table columns refer to (in order):*

- *total number of transcripts and proteins,*
- *total number of circadian transcripts,*
- *total number of circadian proteins,*
- *number of circadian transcripts that have circadian proteins,*
- *number of circadian transcripts that have not circadian proteins,*
- *number of not circadian transcripts that have circadian proteins,*
- *number of not circadian transcripts that have not circadian proteins.*

	INTERSECTION T AND P	**CIRC T**	**CIRC P**	**CIRC T AND CIRC P**	**CIRC T AND NOT CIRC P**	**NOT CIRC T AND CIRC P**	**NOT CIRC T AND NOT CIRC P**
DEX	3665	1468	59	33	1435	26	2171
PHY	5841	1467	365	161	1306	204	4170

Table 5.3 *Percentages of the comparisons between transcriptomics dynamics and proteomics dynamics. 'T' stands for 'transcripts', 'P' stands for 'proteins', 'circ' stands for 'circadian', 'not circ' stands for 'not circadian'. As regards the percentages of this table, they are calculated by dividing the number indicated by the column by the number indicated by the row (e.g., 2.25 % is the percentage of circadian transcripts that have also circadian proteins, calculated with respect to all circadian transcripts, while 55.93% is the percentage of circadian transcripts that have also circadian proteins, calculated with respect to all circadian proteins).*

		CIRC T AND CIRC P	**CIRC T AND NOT CIRC P**	**NOT CIRC T AND CIRC P**	**NOT CIRC T AND NOT CIRC P**
DEX	% of tot circ T	2.25	97.75	-	-
	% of tot not circ T	-	-	1.18	98.82
	% of tot circ P	55.93	-	44.07	-
	% of tot not circ P	-	39.79	-	60.21
PHY	% of tot circ T	10.97	89.03	-	-
	% of tot not circ T	-	-	4.66	95.34
	% of tot circ P	44.11	-	55.89	-
	% of tot not circ P	-	23.85	-	76.15

Finally, in the next sections important parameters for the dynamics characterization are estimated and compared, excluding from the analyses gene symbols corresponding to non-circadian transcripts and non-circadian proteins, i.e. having barycentres that are distributed in the central area of the two-dimensional PCA clock (Figure 5.1).

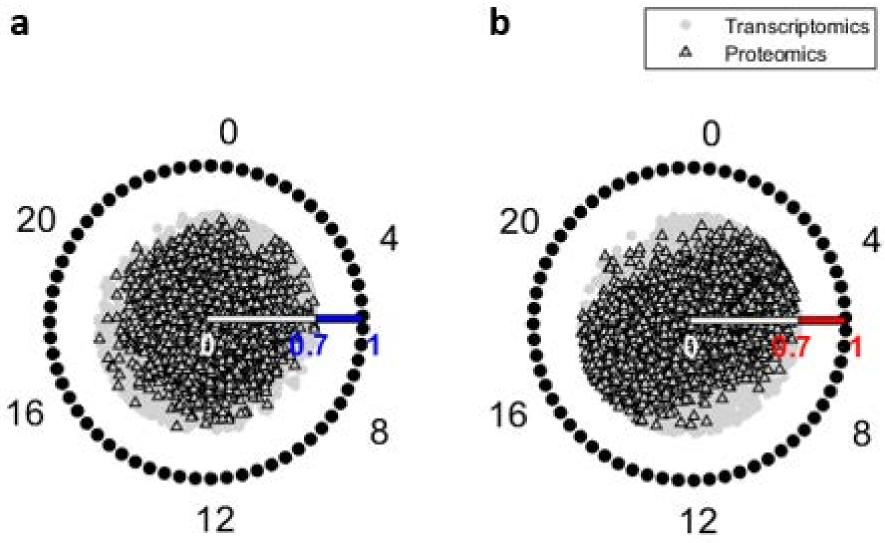


Figure 5.1 *PCA score plot of:*

a) barycentres of DEX transcripts and of their corresponding proteins, when they are both not circadian, (to be circadian, they need be projected into the external area corresponding to the blue part of the radius);

b) barycentres of PHY transcripts and of their corresponding proteins, when they are both not circadian (to be circadian, they need be projected into the external area corresponding to the red part of the radius).

5.2 Peak phases comparison

The objective is to compare the dynamics of transcripts and proteins in terms of peak phases, which is useful for understanding whether the presence of a peak in mRNA causes a peak in the corresponding protein in a short time-lapse or shifted in time. In particular, the more peaks are shifted, the more it is likely that other regulatory mechanisms (e.g. post-transcriptional modifications) are activated within cells in order to ensure the right timing of biological functions in human liver cells.

The analysis is general at first, then it goes into detail:

- polar histograms representing the phases of all circadian transcripts and proteins, in DEX and PHY protocols, considering the four datasets in an independent manner in order to include also profiles that are circadian in one dataset and not in the others (Figure 5.2 and 5.3);
- PCA score plots of profiles barycentres, considering DEX and PHY protocols in an independent manner (Figure 5.4 and 5.5, respectively). In particular, two cases are considered: circadian proteins having also circadian transcripts and circadian proteins originated from not circadian transcripts (where circadianity means $Rr \geq 0.7$). In the former case, also polar histograms with difference in peak phases are shown.

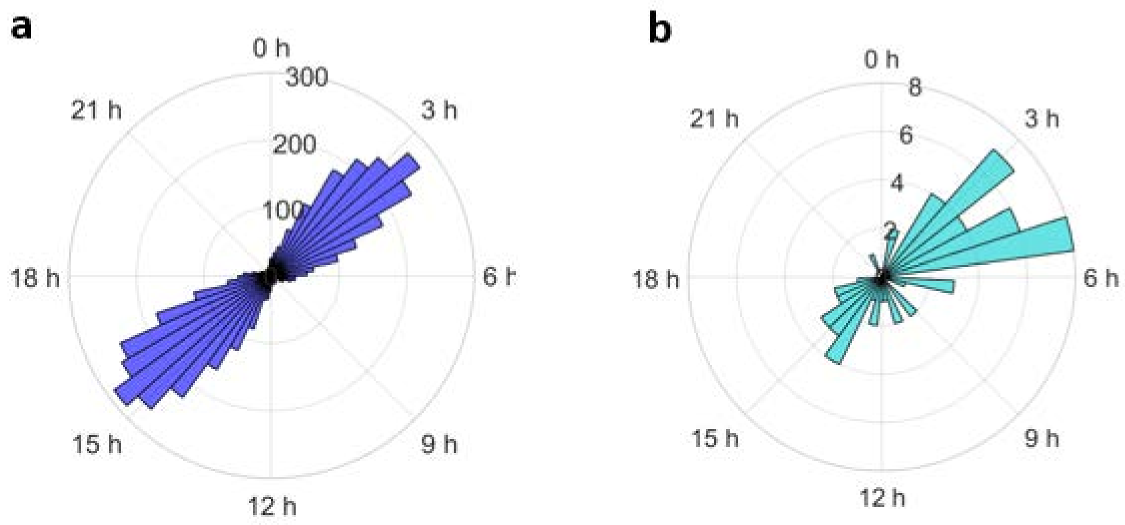


Figure 5.2 ***DEX data***. *Polar histograms showing the distribution of phases of:*

a) all (4570) circadian transcripts in DEX protocol;

b) all (60) circadian proteins in DEX protocol.

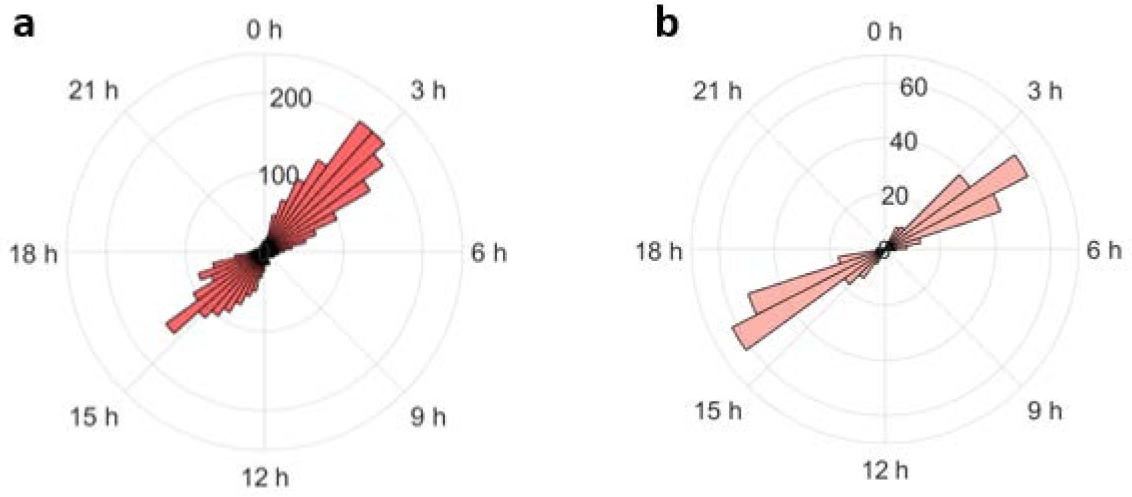


Figure 5.3 ***PHY data***. *Polar histograms showing the distribution of phases of:*

a) all (2812) circadian transcripts in PHY protocol;

c) all (367) circadian proteins in PHY protocol.

These results show that there is a high level of polarization in the phases of circadian transcripts/proteins, even though with some differences between transcriptomics and proteomics. Indeed, in the former case both protocols are characterized by the same bimodal distribution with maxima at 3 h and 15 h. Instead, DEX proteomics still displays a certain polarization at 3 h and 15 h, but with a higher dispersion of phases. Moreover, PHY protein phases are highly polarized, but both of the two main peaks are shifted one hour forward with respect to the transcriptomics ones. Finally, it is interesting to notice that the circadian transcription and translation of liver biological functions is concentrated mainly in 3 hours within the active phase and 3 hours within the resting phase.

This polarization in the four datasets may suggest that there is a net distinction between human liver functionality in the active and resting phase, which is supported by the enrichment analyses of transcriptomics (Chapter 3, section 3.4) and proteomics (Chapter 4, section 4.2.4). Indeed, even though some pathways resulted to be activated all day long, like metabolism of lipids in DEX and PHY transcriptomics, some clear distinctions were highlighted, like the majority of metabolism pathways enriched in the active phase and the majority of metabolism of proteins pathways enriched in the resting phase.

Then, DEX and PHY results are studied in deeper detail: Figure 5.4 and 5.5 show PCA score plots of circadian proteins with circadian or non-circadian transcripts, together with polar histograms representing the intersection of transcriptomic and proteomic circadian data.

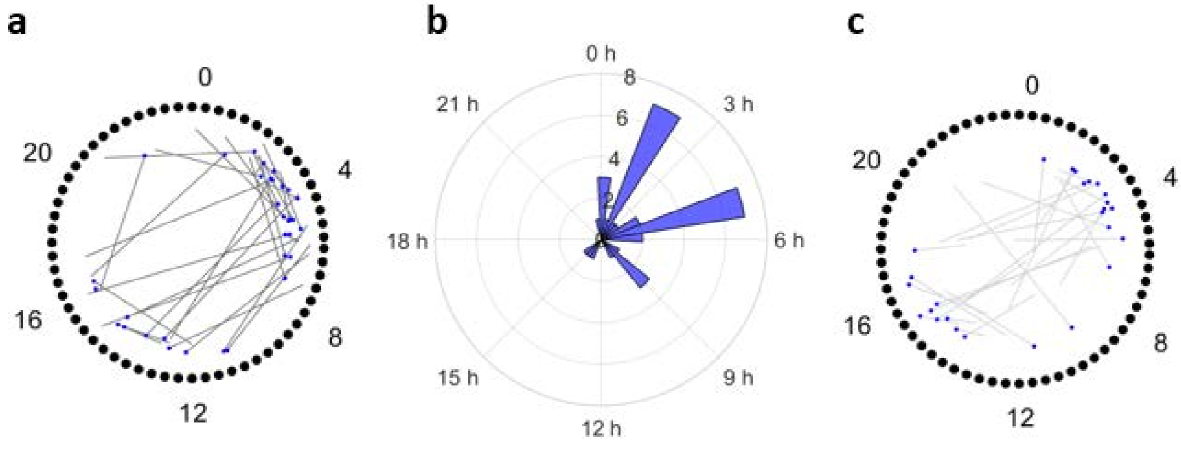


Figure 5.4 ***DEX results.*** *Figures a and b refer to profiles that are circadian both in the transcriptomic and proteomic dataset and they show, respectively, the 2D-PCA score plot of their barycentres and the difference between peak phases of proteins and of transcripts. Finally, c is the PCA score plot of the barycentres of profiles that are circadian in the proteomic dataset, but not in the transcriptomic one. Blue dots represent circadian proteins, dark grey lines connect them to circadian transcripts (a), while light grey lines connect them to non circadian transcripts (c).*

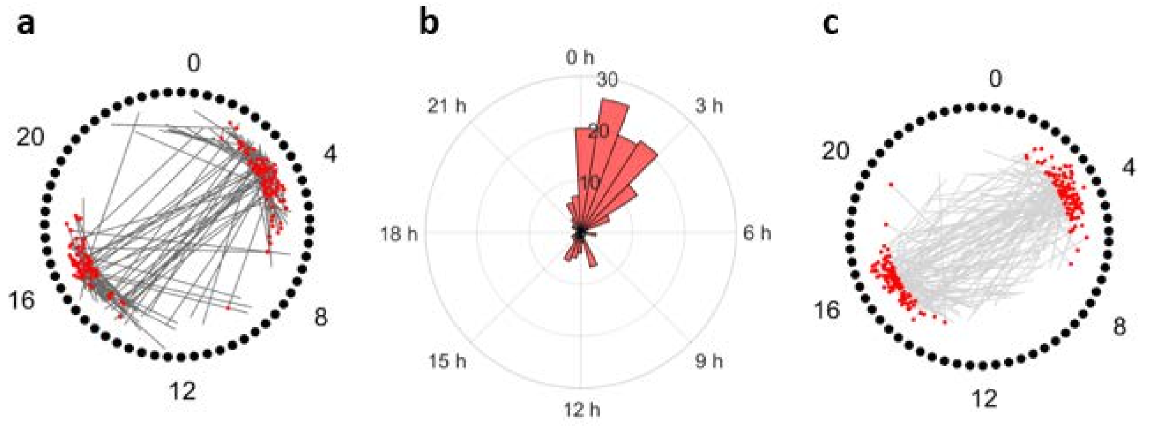


Figure 5.5 ***PHY results.*** *Figures a and b refer to profiles that are circadian both in the transcriptomic and proteomic dataset and they show, respectively, the 2D-PCA score plot of their barycentres and the difference between peak phases of proteins and of transcripts. Finally, c is the PCA score plot of the barycentres of profiles that are circadian in the proteomic dataset, but not in the transcriptomic one. Red dots represent circadian proteins, dark grey lines connect them to circadian transcripts (a), while light grey lines connect them to non circadian transcripts (c).*

Two different trends are found with DEX and PHY data: in the former case, the number of circadian proteins translated from circadian transcripts are almost the same as those originated from not circadian transcripts. Instead, in the latter case, the number of circadian proteins not having circadian transcripts is predominant.

The differences of peak phases between circadian proteins and transcripts is different in the two protocols, too. As regards DEX protocol, there is not a predominant shift of peak phases and shifts are mainly between 0 h and 9 h. In contrast, PHY proteomics can be either synchronized with PHY transcriptomics (with phases shifted of 0-3 h) or be in antiphase with it (shifted forward of 12-15 h), with the former case that is markedly more frequent than the latter.

5.3 Amplitudes comparison

In this section, amplitudes are estimated and compared order to understand if there is high correlation between the level of oscillation around the mean in transcriptomics and proteomics and if circadian profiles tend to be those having higher amplitudes. In particular, since the measurement technologies are different, transcriptomic and proteomic amplitudes estimations are made comparable by using the relative amplitude, meaning the amplitude of each profile divided by the mean expression value for each (Figure 5.6 and 5.7).

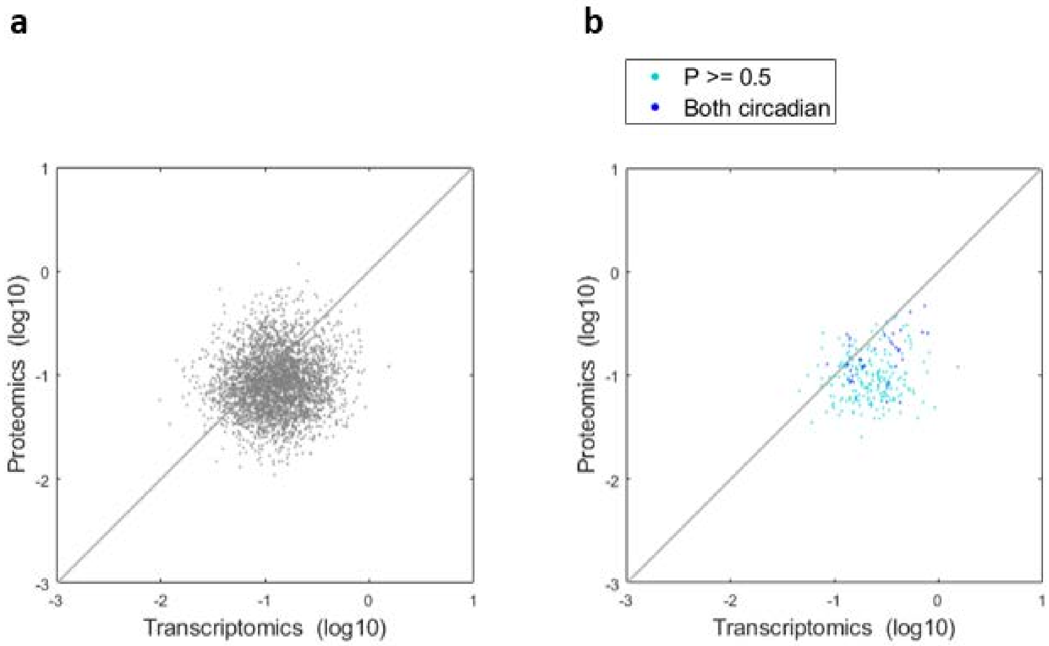


Figure 5.6 ***DEX data.*** *Comparison of the relative amplitudes of transcriptomics and proteomics, considering:*

a) all DEX profiles in common;

b) profiles that are circadian in both DEX transcriptomic and DEX proteomic datasets (blue dots) or profiles that are circadian in the transcriptomic dataset, while they are slightly below the threshold of circadianity in the proteomic one (light blue dots, indicated by 'P' in the legend).

In both cases, the diagonal represents equal relative amplitudes.

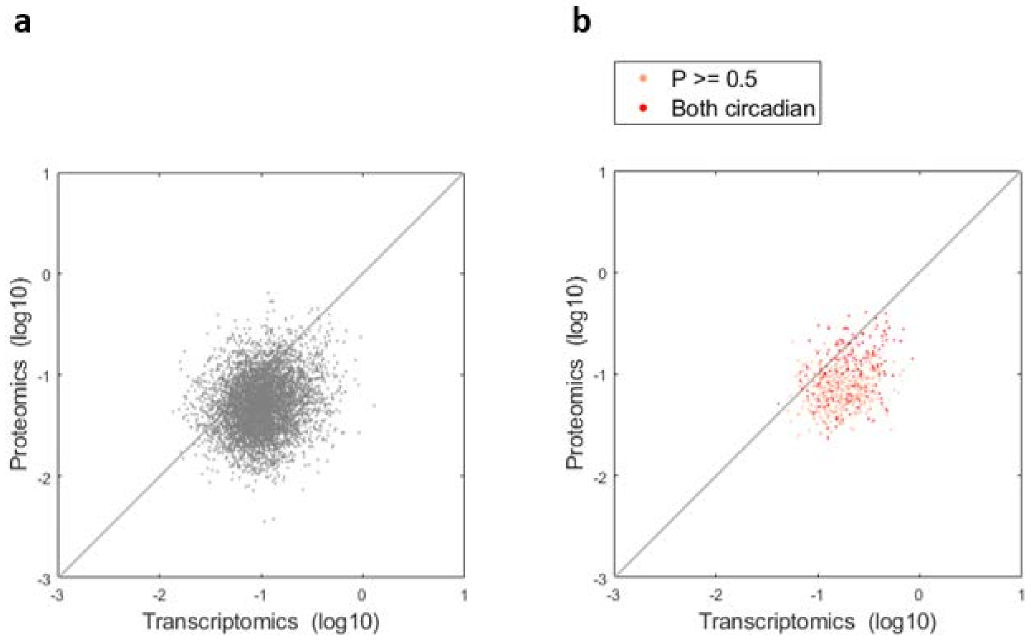


Figure 5.7 ***PHY data.*** *Comparison of the relative amplitudes of transcriptomics and proteomics, considering:*

a) all DEX profiles in common;

b) profiles that are circadian in both DEX transcriptomic and DEX proteomic datasets (red dots) or profiles that are circadian in the transcriptomic dataset, while they are slightly below the threshold of circadianity in the proteomic one (pink dots, indicated by 'P' in the legend).

In all cases, the diagonal represents equal relative amplitudes.

The R^2 values corresponding to all data subsets of Figure 5.6 and 5.7 are shown in Table 5.4.

Table 5.4 *R^2 calculated with log10 values of the relative amplitudes of transcriptomic and proteomic data. In particular, this index is calculated with the whole datasets (when both proteins and transcripts are measured, shown in 'ALL' column), with transcripts that are circadian and proteins that are slightly below the circadianity threshold ('$R_{r,T} \geq 0.7$, $R_{r,P} \geq 0.5$' column), with transcripts and proteins that are both circadian ('$R_{r,T} \geq 0.7$, $R_{r,P} \geq 0.7$' column).*

PROTOCOL	ALL	$R_{r,T} \geq 0.7$ $R_{r,P} \geq 0.5$	$R_{r,T} \geq 0.7$ $R_{r,P} \geq 0.7$
DEX	0.00	0.03	0.17
PHY	0.03	0.08	0.08

These results show that there is no correlation between relative amplitudes of transcripts and proteins: indeed, just a minority of points in Figure 5.6 and 5.7 lie on the diagonal and all the calculated R^2 indexes are much lower than 1. This means that higher amplitudes for the transcripts does not mean necessarily higher amplitudes for the corresponding proteins, which may be due to the fact that mRNAs and proteins have different stabilities, that mRNAs can be translated many times and/or that proteins are subjected to post-translational modifications.

Moreover, by comparing Figure 5.6 a with 5.6 b and Figure 5.7 a with 5.7 b, it can be noted that the selected circadian proteins tend to have higher amplitudes, which is evident in particular for DEX proteins with a relative radius higher than or equal to 0.7.

5.4 Conclusions on mRNAs and proteins comparison

The comparison between transcriptomic and proteomic data highlights the presence of many differences in their dynamics, in both protocols. First of all, the percentages of circadian profiles in transcriptomics is one order of magnitude higher than in proteomics and also when they are both circadian the phase shift can be relevant (up to 12 h in DEX and between 12 h and 15 h in PHY). Moreover, there is a low correlation between the relative amplitude of transcripts and that of proteins. This lack of overlap between mRNA and protein profiles suggests that post-transcriptional and post-translational modifications on transcripts and proteins, respectively, may have a key role in shaping circadian rhythms of the human liver clock.

Moreover, both the standard synchronization protocol based on hormonal stimuli (i.e. DEX protocol) and the innovative one based on feeding-fasting cycles (i.e. PHY protocol) are able to induce a bimodal distribution of phases, with one peak in the active phase and one in the resting phase, in both transcriptomic and proteomic data. However, the two protocols have different performances with transcriptomics and proteomics in terms of circadianity: indeed, DEX protocol has a higher percentage of circadian transcripts than PHY (40% over 25% of PHY), while it produces a lower percentage of circadian proteins than PHY protocol (1.6% versus 6.2% of PHY).

Finally, the comparison highlighted that only a small percentage of circadian transcripts is translated into circadian proteins (2.5 % and 11% for DEX and PHY, respectively), while almost half of the circadian proteins are derived from circadian transcripts (56% and 44% for DEX and PHY, respectively). This is in line with the state of the art, because also in case of Reddy 2006 and Mauvoisin 2014 only half of the circadian proteins are translated from circadian mRNAs. However, these results are affected by the incomplete coverage of proteomic measurements and may become more robust with future technological advances in protein mass spectrometry.

Conclusions and future perspectives

The main objective of this Thesis is to study circadian rhythms in human liver cells, i.e. liver biological functions that are activated at a given hour during the day, thus with a period of about 24 hours. To do so, *omics* data are collected over time from human liver cells cultured *in vitro* and are analysed through Data Analytics and Bioinformatics techniques; in particular, these data are characterized by two main innovations:

1) the presence of two sets of data, one extracted from cells perturbed by means of a reference synchronization protocol, called "DEX", and another specifically developed for these data, called "PHY" or "physiological"; the former uses Dexamethasone to mimic cortisol hormone, while the latter provides cells with physiological changes in insulin, glucose and glucagon;
2) the presence of two different levels of omics data, i.e. transcriptomics and proteomics, which are collected at corresponding time points (0 h, 4 h, 8 h, 12 h, 16 h and 20 h); in particular, these data are biologically related because proteins are synthetized from transcripts.

Therefore, two specific objectives of this Thesis are:

1) comparing the performance of DEX and PHY protocols in order to understand how hormonal and metabolic stimuli are able to synchronize the liver clock in an independent manner, which would be unfeasible to study *in vivo* because these two effects could not be decoupled;
2) comparing the dynamics of transcripts and corresponding proteins, in order to assess whether their temporal profiles overlap or they vary significantly, which in turn suggests the influence of post-transcriptional and post-translational modifications on the final circadian rhythms in human liver.

As regards transcriptomics, the major source of variability is the synchronization protocol itself, but also the time of collection is relevant since it determines the formation of different clusters on a PCA score plot. Not only different time points are clearly separated, but they are also arranged chronologically, suggesting that there might be a relevant percentage of circadian profiles, i.e. sinusoidal profiles with one peak within 24 hours, instead of randomly fluctuating profiles. In particular, the innovative protocol is able to induce a percentage of circadian profiles that is of the same magnitude as the reference one: 22% and 35% in PHY and DEX, respectively. Moreover, both hormonal and metabolic stimuli determine a bimodal distribution of peak phases with two main peaks, i.e. 3 h and 15 h, suggesting that there is a net distinction between liver functions carried out in the active and resting phase. However, also non-circadian profiles are characterized further, since they may reveal the ability of DEX and PHY protocols to activate different biological functions all day long: indeed, statistical tests allow to extract two main groups of flat profiles, i.e. transcripts that are constantly up-regulated in DEX and others in PHY protocol. The final step is the biological interpretation; however, for the purpose of describing the physiology of circadian clocks, it is important to distinguish between active (i.e. day in humans) and resting phase (i.e. night in humans). Based on the experimental setup, i.e. time and duration of cells synchronization, and on the peak phases of some reference genes, it is possible to conclude that the interval 0-8 h corresponds to the active phase, while 12-20 h corresponds to the resting phase. Therefore, the transcriptomic dataset to be analysed from the biological point of view can be divided into: profiles that are circadian in DEX or PHY protocol and peaking in active or resting phase; non circadian profiles up-regulated in DEX or in PHY protocol. Overall, the results match the most relevant pathways encountered in the literature about liver circadian clock, like the main hepatic metabolic processes, including glucose, lipid, and cholesterol/bile acid metabolism. Interestingly, circadian profiles in DEX and PHY protocols have a similar behavior throughout the day: for example, metabolism pathways, e.g. carbohydrates, lipids and vitamins metabolism, are more enriched in the active phase, while metabolism of proteins, e.g. Unfolded Proteins Response (UPR) and post-translational modifications, are more enriched in the resting phase. Finally, non circadian profiles that are constantly up-regulated in PHY protocol are enriched in a higher number of pathways, especially in metabolism of proteins, metabolism and immune system. Therefore, the slightly lower percentage of circadian rhythms in PHY protocol does not lead to a loss of essential liver functions, but just to a non-oscillatory behaviour for some of them.

Compared to transcriptomics, proteomic field is less mature because of technical limitations: indeed, up to date there is not a high-throughput measurement technique allowing to achieve a full-coverage, therefore part of the information is unavoidably lost. For this reason, proteomic data analysis is preceded by a comparison with literature, which proves the reliability of our dataset: indeed, the level of inter-variability between our data and published data is comparable to that among different literature sources, in terms of measured proteins, circadian profiles and phase shifts.

As shown in the preliminary PCA score plot, the major source of variability in the proteomic data of this Thesis is the synchronization method, while within one protocol different time points tend to mix together. This means that subsequent time points are not significantly different, suggesting a low percentage of circadian profiles. This assumption is confirmed by the identification of circadian rhythms, giving only 6.2% and 1.6% in PHY and DEX, respectively: therefore, the two percentages of circadian rhythms are of the same magnitude, with the innovative protocol even outperforming the reference. Moreover, the major trend in circadian profiles phases is a bimodal distribution with two main peaks at 4 h and 16 h, suggesting a net distinction between liver functions carried out in the active and resting phase. Also in this case, the biological analysis of circadian profiles peaking in 0-8 h or 12-20 h highlights the most important liver functions encountered in literature, with metabolism pathways that are more enriched in the active phase and metabolism of proteins that is more enriched in the resting phase, in both protocols.

Finally, there is experimental evidence in literature that circadian rhythms are created and sustained by a multi-level regulatory mechanism, involving the regulation of genes transcription, modifications of mRNAs and the regulation of synthesis and modifications of proteins. In this Thesis, this theory is supported by the results of the integrated analysis of transcriptomics and proteomics. Indeed, if there were no regulatory mechanisms modifying mRNAs and proteins, their profiles would be similar since proteins are synthetized from mRNAs. Instead, their dynamics are significantly different. Indeed, there is not a biunivocal correspondence between circadian mRNAs and proteins, because a relatively low percentage of circadian transcripts are translated into circadian proteins (2.5 % and 11% for DEX and PHY, respectively), while only half of the circadian proteins are derived from circadian transcripts (56% and 44% for DEX and PHY, respectively). Moreover, even when profiles are circadian both in transcriptomics and proteomics, they display relevant shifts in peak phases (between 0 h and 12 h in DEX and between 12 h and 15 h in PHY). All these differences between mRNAs and proteins dynamics suggest that post-transcriptional and post-translational modifications, acting on transcripts and proteins respectively, may have a key role in shaping circadian dynamics of the liver biological functions.

Future improvements of protein measurements techniques may help confirming these results thanks to full-coverage proteomics, thus giving more complete conclusions on the performance of the innovative protocol and on the similarities between transcripts and proteins.

Moreover, this study on the liver circadian clock can potentially be applied for future researches in the biomedical and pharmaceutical field. For example, one of the most promising application is pharmacokinetics, because there are circadian rhythms also in the absorption, distribution and excretion of drugs. Therefore, understanding how an organ (e.g. liver) can be stimulated in order to activate precise biological responses at specific times of the day may help developing more efficient treatments, therefore improving patients life and reducing waste of human efforts and economical resources due to poorly effective therapeutics.

Nomenclature

Abbreviations and acronyms

DEX	Dexamethasone (protocol)
PHY	Physiological (protocol)
RNA-Seq	RNA-Sequencing
TMT	Tandem Mass Tag
LC-MS/MS	Liquid Chromatography – Mass Spectrometry
SF	Scaling factor
TMM	Trimmed mean of M values
CPM	Counts Per Million
PCA	Principal Component Analysis
PCs	Principal Components
DEGs	Differentially Expressed Genes
MLE	Maximum Likelihood Estimation
GLM	Generalized Linear Models
GO	Gene Ontology
CC	Cellular Component
BP	Biological Process
MF	Molecular Function

Symbols

$E(\cdot)$	Expected value
r_{ik}	Counts of gene i in sample k
n_k	total number of reads in sample k
T_{ik}	number of copies of transcript i in sample k
L_i	length of transcript i
N_k	number of transcripts in sample k
S_k	size factor
X	Matrix of data
R	Rank of X
$\mathrm{t_a}$	Scores on the ath PC
$\mathrm{p_a}$	Loading vector for the ath PC
Q	Number or retained PCs
T	Scores matrix
P	Loadings matrix
E	Residual matrix
Rr	Relative radius
A	Amplitude
ρ_{XY}	Pearson correlation coefficient
σ_{XY}	Covariance between X and Y

σ	Standard deviation of X
μ	Mean
H_0	Null hypothesis
H_1	Alternative hypothesis
$NB(\mu_{ik}, \varphi_i)$	Negative Binomial distribution with mean μ_{ik} and dispersion φ_i
$\mathcal{L}$	Likelihood
p	probability
N	Normal distribution
η_i	Linear predictor (of the Generalized Linear Model)
$g(\cdot)$	Link function (of the Generalized Linear Model)
F	F statistics

References

Ahmad, Y., Lamond, A. I. (2014). A perspective on proteomics in cell biology. *Trends Cell Biol.,* **24(4)**, 257–264.

Angel, T. E., Aryal1, U. K., Hengel, S. M., Baker, E. S., Kelly, R. T., Robinson, E. W., and Smith, R. D. (2012). Mass spectrometry based proteomics: existing capabilities and future directions. *Chem Soc Rev.*, **41(10),** 3912–3928.

Asher, G., Sassone-Corsi, P. (2015). Time for food: the intimate interplay between nutrition, metabolism, and the circadian clock. *Cell,* **161(1)**, 84-92.

Balsalobre, A., Brown, S. A., Marcacci, L., Tronche, F., Kellendonk, C. Reichardt, H. M., Schütz, G., Schibler, U. (2020). Resetting of circadian time in peripheral tissues by glucocorticoid signaling. *Science*, **289(5488)**, 2344-7.

Baura, G. D. (2012). *Medical Device Technologies*. Academic Press, p. 423-452.

Beaumont, R. E. (2019). Physiologically-relevant in vitro circadian cell synchronization protocol. *Post-doctoral Thesis*, Shanghai Institute for Advanced Immunochemical Studies (SIAIS), ShanghaiTech University.

Bludau, I., Aebersold, R. (2020). Proteomic and interactomic insights into the molecular basis of cell functional diversity. *Molecular cell Biology,* **21**, 327–340.

Chaix, A., Zarrinpar, A., Panda, S. (2016). The circadian coordination of cell biology. *JCB*, **215**, 1, 15-25.

Di Camillo B, Toffolo G, Nair SK, Greenlund LJ, Cobelli C. (2007). Significance analysis of microarray transcript levels in time series experiments. *BMC Bioinformatics*, **8**(Suppl 1), S10.

Evans, J. A., Davidson, A. J. (2013). Health Consequences of Circadian Disruption in Humans and Animal Models. In: *Chronobiology: Biological Timing in Health and Disease* , Volume 119, (Martha U. Gillette, Ed.), Academic Press, p. 283-323.

Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., and Ordoukhanian, P. (2014). Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*, **56**, 61-77.

Kim, K. J. (2019). The Role of Circadian Clocks in Metabolism. *Chronobiol. Med.*, **1(3)**, 107-110.

Koronowski, K. B., Kinouchi, K., Welz, P. S., Smith, J. G., Zinna, V. M., Shi, J., Samad, M., Chen, S., Magnan, C. N., Kinchen, J. M., Li, W., Baldi, P., Benitah, S. A., Sassone-Corsi, P. (2019). Defining the Independence of the Liver Circadian Clock. *Cell,* **177(6)**, 1448–1462.e14.

Kukurba, K. R. and Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harb Protoc.,* **11**, 951-69.

Lee, Y., Chun, S. K., Kim, K. (2015). Sumoylation controls CLOCK-BMAL1-mediated clock resetting via CBP recruitment in nuclear transcriptional foci. *Biochimica et Biophysica Acta* **1853 (10, A)**, 2697–2708.

Lee-Liu, D., Almonacid, L. I., Faunes, F., Melo, F., Larrain, J. (2012).Transcriptomics using next generation sequencing technologies. *Methods Mol Biol*, **917**, 293-317.

Mauvoisin, D., Atger, F., Dayon, L., Kussmann, M., Naef, F., Gachon, F. (2017). Circadian and Feeding Rhythms Orchestrate the Diurnal Liver Acetylome. *Cell Reports*, **20**, 1729–1743.

Mauvoisina, D., Wangc, J., Jouffea, C., Martina, E., Atgera, F., Warideld, P., Quadronid, M., Gachona, F., Naefc, F. (2014). Circadian clock-dependent and -independent rhythmic proteomes implement distinct diurnal functions in mouse liver. *PNAS,* **111 (1),** 167-172.

Mehra, A., Baker, C.L. , Loros, J. J., Dunlap, J. C. (2009). Post-Translational Modifications in Circadian Rhythms. *Trends Biochem Sci.*, **34(10)**, 483–490.

Musiek, E. S., FitzGerald, G. A. (2013). Molecular Clocks in Pharmacology. *Handb Exp Pharmacol,* **217**, 243–260.

Parry, A. S., Woods, R. M., Hodson, L., Hulston C. J. (2017). A Single Day of Excessive Dietary Fat Intake Reduces Whole-Body Insulin Sensitivity: The Metabolic Consequence of Binge Eating. *Nutrients,* **9(8)**, 818.

Poggiogalle, E., Jamshed, H., Peterson, C. M. (2018). Circadian regulation of glucose, lipid, and energy metabolism in humans. *Metabolism,* **84**, 11-27.

Rauniyar, N. and Yates, J. R. (2014). Isobaric Labeling-Based Relative Quantification in Shotgun Proteomics. J. *Proteome Res.,* **13**, 5293 −5309.

Reddy, A. B., Karp, N. A., Maywood, E. S., Sage, E. A., Deery, M., O'Neill, J. S., Wong, G.K.Y., Chesham, J., Odell, M., Lilley, K. S., Kyriacou, C. P., Hastings, M. H. (2006). Circadian Orchestration of the Hepatic Proteome. Current Biology, **16**, 1107–1115.

Rivals, I., Personnaz, L., Taing, L., Potier, M.C. (2007). Enrichment or depletion of a GO category within a class of genes: which test?. *Bioinformatics, 23 (4), 401–407.*

Robinson, M. D., McCarthy, D. J., Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26(1)**, 139-40.

Robles, M. S., Cox, J., Mann, M. (2014). In-Vivo Quantitative Proteomics Reveals a Key Contribution of Post-Transcriptional Mechanisms to the Circadian Regulation of Liver Metabolism. *PLOS Genetics*, **10(1)**:e1004047.

Sanavia, T., Finotello, F., Di Camillo, B. (2015). FunPat: function-based pattern analysis on RNA-seq time series data. *BMC Genomics*, **16** (Suppl 6), S2.

Saran, A. R., Dave, S., Zarrinpar, A. (2020).Circadian Rhythms in the Pathogenesis and Treatment of Fatty Liver Disease. *Gastroenterology*, **158 (7)**, p. 1948-1966.e1.

Spies, D., Renz, P. F., Beyer, T. A., Ciaudo, C. (2017). Comparative analysis of diff erential gene expression tools for RNA sequencing time course data. *Briefings in Bioinformatics,* 1-11.

Tsigos, C., Chrousos, G. P. (2002). Hypothalamic-pituitary-adrenal axis, neuroendocrine factors and stress. *J Psychosom Res.* **53** (4), 865–71.

Wang, Y., Song, L., Liu, M., Ge, R., Zhou, Q., Liu, W., Li, R., Qie, J., Zhen, B., Wang, Y., He, F., Qin, J., Ding, C. (2018). A proteomics landscape of circadian clock in mouse liver. *Nature Communications*, **9**, 1553.

Yue J, Burdett E, Coy D, et al. Somatostatin receptor type 2 antagonism improves glucagon and corticosterone counterregulatory responses to hypoglycaemia in streptozotocininduced diabetic rats. *Diabetes*, 2012, **61**, 197-207.

Zanquetta, M. M., Corrêa-Giannella, M. L., Monteiro, M. B., Villares, S. M .F. (2010). Body weight, metabolism and clock genes. *Diabetology & Metabolic Syndrome,* **2**, 53.

Sitography

https://cytoscape.org/

https://emedicine.medscape.com

https://reactome.org/

https://sciencing.com

https://www.thermofisher.com

Appendix A

This appendix aims at showing part of the codes used in this Thesis, implemented as MATLAB scripts and/or functions. All codes are built specifically for this Thesis, except for those in appendices A.1 and A.2 that refer to the algorithm created by Matteo Bicego, while working at the Venetian Institute of Molecular Medicine and at the Industrial Engineering Department of the University of Padova.

A.1 Calculation of barycentres projections and relative radii

This is part of the algorithm developed by Matteo Bicego for calculating the scores of each profile barycentre (`sB`), which are projected inside the 2-D PCA clocks (with the external circumference given by the calibration profiles). Moreover, the polar coordinates of `sB` are useful for calculating the relative radius (`Rr`), which in turn allows to identify circadian profiles setting a threshold of $Rr \geq 0.7$.

In particular, `loads` refers to the matrix of loadings obtained by calibrating the PCA model; `snv` performs a Standard Normal Variate scaling; `N_P(j)` is the number of non-zero measurements for gene `j`. Finally, `Y` is a 3-D matrix with: 4096 elements in the first dimension representing all possible permutations of the 4 replicates at each time point; 6 elements in the second dimensions representing time point 0 h, 4 h, 8 h, 12 h, 16 h and 20 h; as many elements in the third dimension as the number of transcripts (or proteins).

```
For j=1:size(Y,3)
    sR(:,:,j)=snv(Y(:,:,j))*loads;
    sB(j,:)=sum(sR(:,:,j))./ N_P(j);
end

[T,R]=cart2pol(sB(:,1),sB(:,2));
Rr=R./Rm;
```

A.2 Calculation of circadian profiles phases

This is an algorithm for the characterization of circadian profiles in terms of phase (`Y1`). After calibrating the PCA model and projecting experimental data on the PCA score plot, phases can be calculated. The inputs are `XP` and `XG`: the former is the matrix of barycentres scores, while the latter has the 2-D coordinates of time point 0 h on the loading plot.

```
XP = sB;
XG = repmat(loads(1,1:2),size(sB,1),1);
[T_G,~]=cart2pol(XG(:,1),XG(:,2));
[T_P,~]=cart2pol(XP(:,1),XP(:,2));
```

```
pos_T_G=find(T_G<0);
T_G(pos_T_G,1)=2*pi+T_G(pos_T_G,1); % T_G goes from 0 to 2pi
pos_T_P=find(T_P<0);
T_P(pos_T_P,1)=2*pi+T_P(pos_T_P,1); % T_P goes from 0 to 2pi

Y1=T_G-T_P;
pos_Y1=find(Y1<0);
Y1(pos_Y1,1)=2*pi+Y1(pos_Y1,1);  % Positive delay
```

A.3 Calculation of amplitude, relative amplitude and mean expression

This is the code for calculating amplitude (`A_vett`), relative amplitude (`Ar_vett`) and mean expression (`mean_expression_vett`) of profiles. In particular, `MEAN` is the matrix having as many rows as the number of profiles and 6 columns with the mean of the 4 replicates for each time point (i.e. 0 h, 4 h, 8 h, 12 h, 16 h and 20 h).

```
for g=1:size(MEAN,1)
    y=MEAN(g,:);
    A=(max(y)-min(y))/2;
    A_vett(g,1)=A;
    mean_expression_vett(g,1)=mean(y);
    Ar_vett(g,1)=A/mean(y);
end
```

A.4 Exploratory PCA

This section shows the code for the PCA of transcriptomic data in $\log_2$ scale (`log_count_matrix`), while the codes for analysing linear transcriptomic data and linear proteomic data are omitted because they are analogue: indeed, the only difference is the use of Standard Normal Variate for scaling data, instead of mean centering. Finally, "[…]" is used for ease of visualization and represents commands that are similar to the rows immediately above and below (thus, they refer to time points 4 h, 8 h,12 h and 16 h).

```
%% Mean-centering of log2 data (already filtered and normalized)
log_gene_mean = mean(log_count_matrix,2);
log_count_matrix_centered = log_count_matrix-log_gene_mean;

%% Principal Component Analysis
X = log_count_matrix_centered';
[pc,zscores,pcvars,~,explained] =
pca(X,'Centered',false,'NumComponents',2);

% PCA score plot
figure1 = figure;
axes('Parent',figure1,'FontSize',8), hold on

h0=plot(zscores([1:4],1),zscores([1:4],2),'^','MarkerSize',5,'MarkerFaceCol
or','r','Color','r');
[…]
h20=plot(zscores([21:24],1),zscores([21:24],2),'^','MarkerSize',5,'MarkerFa
ceColor','y','Color','y');
```

```
ph0=plot(zscores([29:32],1),zscores([29:32],2),'o','MarkerSize',5,'MarkerFa
ceColor','r','Color','r');
[…]
ph20=plot(zscores([49:52],1),zscores([49:52],2),'o','MarkerSize',5,'MarkerF
aceColor','y','Color','y');

xlabel(['PC1 (',num2str(explained(1)),'%)'],'FontSize',8);
ylabel(['PC2 (',num2str(explained(2)),'%)'],'FontSize',8);
axis equal, axis square, box on, xlim([-180 180]),ylim([-180 180])
legend([h0,h4,h8,h12,h16,h20,h24,...
    ph0,ph4,ph8,ph12,ph16,ph20,ph24],...
    {'DEX 0h ','DEX 4h','DEX 8h','DEX 12h','DEX 16h','DEX 20h',...
    'PHY 0h ','PHY 4h','PHY 8h','PHY 12h','PHY 16h','PHY 20h'})
```

A.5 PCA model: 2D score plot

The calibrated PCA model is used to project new permutated profiles into a 2D score plot; in particular, the code of this section deals with the proteomic data; transcriptomics is plotted with a similar code, even if only *Rr* ≥0.7 is considered in this case. In the following code, `loads` refers to the calibration loadings; `sc` refers to the calibration scores (used to plot the external clock); `sBx` and `sBy` refer to the scores on PC1 and PC2, respectively, of the barycentre of each protein (or transcript).

```
figure
pl1=plot(sBx,sBy,'.','Color',rgb('lightgrey')); hold on % all data
pl2=plot(sBx(Rr>= 0.5),sBy(Rr>= 0.5),'.','Color',rgb('DarkTurquoise')); %
close to the threshold
pl3=plot(sBx(Rr>= 0.7),sBy(Rr>= 0.7),'.','Color','b'); % properly circadian
scatter(sc(1261:1323,1),sc(1261:1323,2),'ok','Filled') % external clock
hhh=[0:4:20];
for z=1:6
    text(loads(z,1)*4.8,loads(z,2)*4.8,mat2str(hhh(z)),'Fontsize',16), hold
on
end
axis off,axis equal,box on, axis ([-3.5 3.5 -3.5 3.5])
xlabel('PC1','Fontsize',16),ylabel('PC2','Fontsize',16)
legend([pl1,pl2,pl3],{'Not_circadian','Rr>=0.5','Rr>=0.7'},'FontSize',12,
'location','southoutside')
```

Moreover, in Chapter 5 barycentres of both circadian transcripts and proteins are projected into the same 2D-PCA score plot and linked by segments in order to highlight their phase shifts. The corresponding code is shown below, where `TABLE` is a table containing the results (i.e., the parameters calculated for each profile), while `colRt`, `colRp`, `colBar1t`, `colBar2t`, `colBar1p`, `colBar2p` refer, respectively, to the column indexes of relative radius of transcriptomics, relative radius of proteomics, scores on PC1 and PC2 of transcriptomics and proteomics.

```
% relative radius
RrT=table2array(TABLE(:,colRt));
RrP=table2array(TABLE(:,colRp));

% all barycentres
xBt=table2array(TABLE(:,colBar1t));
yBt=table2array(TABLE(:,colBar2t));
```

```
xBp=table2array(TABLE(:,colBar1p));
yBp=table2array(TABLE(:,colBar2p));

% barycentres of circadian transcripts and proteins
Xt= xBt(RrT >= 0.7 & RrP >= 0.7);
Yt= yBt(RrT >= 0.7 & RrP >= 0.7);
Xp= xBp(RrT >= 0.7 & RrP >= 0.7);
Yp= yBp(RrT >= 0.7 & RrP >= 0.7);

% phases of circadian transcripts and proteins
ht=ht(RrT >= 0.7 & RrP >= 0.7);
hp=hp(RrT >= 0.7 & RrP >= 0.7);

% plot
figure
% lines that connect data
for j= 1: length(Xt)
    plot([Xt(j) Xp(j)],[Yt(j) Yp(j)],
'color',rgb('DimGray'),'MarkerSize',0.5);hold on
end

p=plot(Xp,Yp,'o','Color',COL,'MarkerSize',2,'MarkerFaceColor',COL); hold on
% P data
scatter(sc(1261:1323,1),sc(1261:1323,2),'ok','Filled') % external clock
hhh=[0:4:20];
for z=1:4
    text(loads(z,1)*4.8,loads(z,2)*4.8,mat2str(hhh(z)),'Fontsize',16), hold
on
end
for z=5:6
    text(loads(z,1)*4.8-0.4,loads(z,2)*4.8,mat2str(hhh(z)),'Fontsize',16),
hold on
end
axis equal, box on, xlim([-3.5 3.5]), ylim([-3.5 3.5]), axis off
xlabel('PC1','Fontsize',16),ylabel('PC2','Fontsize',16)
```

A.6 Heatmaps

To build heatmaps of circadian genes, the total matrix of normalized and filtered data is divided into gene lists, i.e. profiles circadian in one protocol or in both. Then, the selected sub-matrix (`cpm`) is scaled between -1 and 1 and sorted with respect to peak phases (`h`); finally, the sorted heatmap is plotted by using `imagesc`.

```
% scaling between -1 and 1
cpmMax = max(cpm(:,index),[],2);
cpmMin = min(cpm(:,index),[],2);
cpmScaled = (2*(cpm-cpmMin)./(cpmMax-cpmMin))-1;

% sort profiles with respect to phases
[~,index] = sort(h);
cpmScaled = cpmScaled(index,:);

% generate the heatmap
imagesc(cpmScaled)
```

Instead, the heatmap of non circadian profiles (cpm_noncirc) sorts transcripts or proteins based on hierarchical clustering, with complete linkage and Euclidean distance as settings. This needs to be done because peak phases cannot be defined for random profiles.

```
H=clustergram(cpm_noncirc,'standardize','row','linkage','complete',...
'columnPdist','euclidean','cluster','column','Displayrange',1,'OptimalLeafO
rder','true','colormap','redbluecmap',...
    'columnlabels',["t0 DEX","t4 DEX","t8 DEX","t12 DEX","t16 DEX","t20
DEX",...
    "t0 PHY","t4 PHY","t8 PHY","t12 PHY","t16 PHY","t20 PHY"]);
```

A.7 Polar histograms for phases visualization

The code of this section generates polar histograms used in this Thesis to visualize the distribution of peak phases (h_2pi). A similar code is used for visualizing the difference of peak phases between DEX and PHY transcripts (or proteins) or between transcripts and proteins within the same protocol (the only difference is the input vector).

```
figure
polarhistogram(h_2pi,50), hold on
rticks([0 200 400 600])
thetaticks([0 45 90 135 180 225 270 315])
thetaticklabels({'0 h','3 h','6 h','9 h','12 h','15 h','18 h','21 h'})
pax = gca;
pax.ThetaDir = 'clockwise';
pax.ThetaZeroLocation = 'top';
hold off
```

A.8 Dot-plots of phases and amplitudes

The performances of DEX and PHY protocols are compared in terms of phases and amplitudes of circadian profiles; in particular, this section shows the code for comparing DEX and PHY phases through dot-plots, while the corresponding one for amplitudes is omitted because it is similar.

In the code below, number 1 refers to one protocol (i.e. DEX or PHY), while 2 refers to the other protocol. First, phases estimated in radians (phase1_h) are converted into 0-24 h scale, then they are divided into two groups: phases of profiles that are circadian only in one protocol (phase1(index1_0)) or in both protocols (phase1(index1)). Finally, correlation indexes R^2 are calculated for both groups.

```
% 0-24 h scale
phase1=24*phase1_h/(2*pi);
phase2=24*phase2_h/(2*pi);
thr = 5; % dotted diagonals: difference of 5 h from equality

% phase dot-plot
figure
plot([0 24],[0 24],'k','LineWidth',1), hold on % diagonal
plot([0 24]-thr,[0 24],'--','Color','k','LineWidth',0.7) % dotted line
plot([0 24]+thr,[0 24],'--','Color','k','LineWidth',0.7) % dotted line

% circadian in only one protocol
h1 = plot(phase1(index1_0),phase2(index1_0),'.','MarkerSize',MS,
'Color',color1);

% circadian in both
h2 = plot(phase1(index1),phase2(index1),'.','MarkerSize',MS,
'Color',color2);

% R^2 in only one protocol
c=corrcoef(phase1(index1_0),phase2(index1_0));
R2_honly=c(1,2).^2;

% R^2 in both protocols
c=corrcoef(phase1(index1),phase2(index1));
R2_hboth=c(1,2).^2;
```

A.9 Venn diagrams

Venn diagrams are used to find the proportion of the intersection between pairs of datasets. In particular, in the following code `Labels1` and `Labels2` are string vectors containing Gene Symbols of two groups, e.g. Gene Symbols of DEX and PHY circadian profiles, or of detected proteins in this Thesis and in literature.

```
% number of elements
N1total=size(Labels1,1);
N2total=size(Labels2,1);

% intersection between the two datasets
C=intersect(Labels1,Labels2);
Nintersect=size(C,1);
clear C

% Venn diagrams
figure
venn([N1total N2total],Nintersect,'FaceColor',{col1,col2},
'EdgeColor','black');
axis square
axis off
```

Appendix B

This appendix contains part of the algorithms for Differential Expression Analysis implemented as R scripts and functions. In particular, two R packages are employed: edgeR (edgeR v. 3.26 written in R v. 3.6), for making pairwise comparisons between the two protocols at each time point, and the more recent FunPat (Sanavia et al., 2015), that was specifically built for analysing temporal profiles.

B.1 edgeR

After RNA-Seq reads have been aligned to a reference genome and filtered, edgeR performs a TMM normalization (with `calcNormFactors`) and specifies a design matrix (`design`) for the DE analysis. Then, gene dispersion is estimated (`estimateDisp`) and the Generalized Linear Model is fitted (`glmQLFit`) ; finally, the contrasts are specified (i.e., the two protocols at each time points) and statistical tests are performed in order to assign a p-value to each test.

```
## GENE NORMALIZATION

y <- calcNormFactors(y) # normalization: TMM
y$norm_cpm <- cpm(y, normalized.lib.sizes=TRUE, log=F)
# normalized count matrix
y$norm_logcpm_p1 <- cpm(y, normalized.lib.sizes=F, prior.count=1, log=TRUE)
# log2(normalized pseudo-count)

## DEG CONTRAST DESIGN

time<-factor(c(rep("t0h",4),rep("t4h",4),rep("t8h",4),rep("t12h",4),
rep("t16h",4),rep("t20h",4))
Synchronization <- factor(c(rep("DEX",24),rep("PHY",24)))
Group <- factor(paste(time,Synchronization,sep="."))

design <- model.matrix(~0+Group)
rownames(design) <- colnames(y)

## GENE DISPERSION ESTIMATION

y <- estimateDisp(y, design, robust=TRUE)
commonDisp<-y$common.dispersion
#plotBCV(y)


## GLM FIT

fit <- glmQLFit(y, design)

## DEG SCORING

my.contrasts <- makeContrasts(
  DEXvsPHYt0h = Groupt0h.PHY-Groupt0h.DEX,
  DEXvsPHYt4h = Groupt4h.PHY-Groupt4h.DEX,
  DEXvsPHYt8h = Groupt8h.PHY-Groupt8h.DEX,
  DEXvsPHYt12h = Groupt12h.PHY-Groupt12h.DEX,
  DEXvsPHYt16h = Groupt16h.PHY-Groupt16h.DEX,
  DEXvsPHYt20h = Groupt20h.PHY-Groupt20h.DEX,
  levels=design)

## STATISTICAL TESTS

qlf.DEXvsPHYt0h <- glmQLFTest(fit, contrast=my.contrasts[,"DEXvsPHYt0h"])
qlf.DEXvsPHYt4h <- glmQLFTest(fit, contrast=my.contrasts[,"DEXvsPHYt4h"])
qlf.DEXvsPHYt8h <- glmQLFTest(fit, contrast=my.contrasts[,"DEXvsPHYt8h"])
qlf.DEXvsPHYt12h <- glmQLFTest(fit, contrast=my.contrasts[,"DEXvsPHYt12h"])
qlf.DEXvsPHYt16h <- glmQLFTest(fit, contrast=my.contrasts[,"DEXvsPHYt16h"])
qlf.DEXvsPHYt20h <- glmQLFTest(fit, contrast=my.contrasts[,"DEXvsPHYt20h"])
```

Once p-value are obtained through this code, they are corrected for multiple testing, for example with Benjamini Hochberg method, and a cut-off is selected a priori by the user in order to extract a list of candidate DEGs: in this Thesis, a fold-change of 2 and an FDR of 0.05. The simplified code is shown below for time point 0 h, but is the same for all the remaining time points.

```
## CUT-OFF SELECTION

THRESHOLD = 2
FDR = 0.05

## CANDIDATE DEGs FOR t=0h

name1 <- "DEXvsPHYt0h"
qlf <- eval(parse(text=paste("qlf.",name1,sep="")))

DEGs_index <- decideTests(qlf, adjust.method="BH", p.value=FDR,
lfc=log2(THRESHOLD)) # to get the indexes of DEGs inside the matrix
```

B.2 FunPat

This is part of the algorithm used for selecting DEGs with the Bounded Area method (see Chapter 2); in this case, the input matrix contains already normalized data in logarithmic scale (`ws1`). This R package requires three matrices: two for the conditions to be compared, i.e. DEX and PHY, and a third one with replicates for fitting the error model. The former two are obtained by calculating the mean transcription value at each time point, as shown in the code below obtaining `DEXmean`, `PHYmean`; instead, the latter is omitted here for ease of visualization, but the final matrix is the one explained by Table 3.1 in Chapter 3. Finally, corrected p-values are obtained for each test through the function `SEL.TS.AREA`.

```
## INPUT

load("input/ws1.RData")
EntrezID <- y$genes$EntrezGene
logCPM <- y$norm_logcpm_p1
```

```
logCPM_DEX <- logCPM[,1:24]
logCPM_PHY <- logCPM[,29:52]
remove("logCPM")

timePoints<-c("0h","4h","8h","12h","16h","20h")

## DEX MATRIX

DEXmean <-
data.frame(rowMeans(logCPM_DEX[,1:4]),rowMeans(logCPM_DEX[,5:8]),
rowMeans(logCPM_DEX[,9:12]),rowMeans(logCPM_DEX[,13:16]),
rowMeans(logCPM_DEX[,17:20]),rowMeans(logCPM_DEX[,21:24]))
colnames(DEXmean) <- timePoints
rownames(DEXmean)<- EntrezID

## PHY MATRIX

PHYmean <-
data.frame(rowMeans(logCPM_PHY[,1:4]),rowMeans(logCPM_PHY[,5:8]),
rowMeans(logCPM_PHY[,9:12]),rowMeans(logCPM_PHY[,13:16]),
rowMeans(logCPM_PHY[,17:20]),rowMeans(logCPM_PHY[,21:24]))
colnames(PHYmean) <- timePoints
rownames(PHYmean)<- EntrezID

## REPLICATES MATRIX: called “replic”, obtained as explained in Chapter 3,
## Table 3.1

[…]

## FUNPAT test

rank.res<-
SEL.TS.AREA(replicates=replic,data1=DEXmean,data2=PHYmean,takelog=FALSE)
```

At this point, p-values corrected for multiple testing are attributed to each test and DEGs can be selected with user-defined cut-offs: in this Thesis, a corrected p-value of 0.01.

Acknowledgements

Ringrazio il gruppo di ricerca di Paolo De Coppi presso la University College of London per avermi accolta nel proprio laboratorio durante il mio soggiorno a Londra e per avermi consentito di collaborare a distanza anche dopo il rientro in Italia.

Ringrazio il Prof. Nicola Elvassore per avermi dato la possibilità di conoscere un ambito di ricerca innovativo e di alto livello, permettendomi così di acquisire un approccio integrato ai vari campi del sapere che sarà di grande utilità per la mia crescita professionale.

Un ringraziamento particolare lo dedico alla Prof.ssa Camilla Luni, perché nonostante il disagio dovuto all'impossibilità di frequentare i laboratori di persona nei mesi di emergenza sanitaria, la sua constante disponibilità, la sua competenza e capacità di organizzare il lavoro mi hanno aiutata a superare le difficoltà, a inserirmi in un contesto multidisciplinare e a rendere i mesi di tirocinio molto più produttivi.

Last but not the least, ringrazio mio fratello Giovanni e i miei genitori per essere stati sempre il mio punto di riferimento, per avermi supportata in tutto il mio percorso scolastico e accademico, per aver creduto in me anche nei momenti di sconforto e per il grande affetto che mi hanno sempre dimostrato. Inoltre, ringrazio i famigliari e gli amici più stretti, per essere stati presenti nei momenti più importanti della mia vita, facendomi sentire apprezzata e circondata da affetti.