Etude de I'expression des gènes nycthéméraux à la lumière de l'évolution

Laloum David

UNIL | Université de Lausanne
Faculté de biologie et de médecine

Swiss Institute of Bioinformatics

Département d'écologie et évolution

Etude de l'expression des gènes nycthéméraux à la lumière de l'évolution

Thèse de Doctorat en Médecine et ès Sciences (MD-PhD)

Présentée à la

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

par

David LALOUM

Médecin diplômé de France

Jury

Prof. Paul Franken, président et répondant MD-PhD
Prof. Marc Robinson-Rechavi, directeur de thèse
Prof. Andreas Wagner, expert
Prof. Olivier Delaneau, expert

Lausanne 2020

UNIL | Université de Lausanne
Faculté de biologie et de médecine

Ecole Doctorale
Doctorat MD-PhD

Imprimatur

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

Président·e	Monsieur	Prof.	Paul	**Franken**
Directeur·trice de thèse	Monsieur	Prof.	Marc	**Robinson-Rechavi**
Répondant·e	Monsieur	Prof.	Paul	**Franken**
Expert·e·s	Monsieur	Prof.	Andreas	**Wagner**
	Madame	Prof.	Olivier	**Delaneau**

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

Monsieur David LALOUM

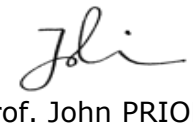
Docteur en médecine Université Claude Bernard - Lyon/F

intitulée

Etude de l'expression des gènes nycthéméraux à la lumière de l'évolution

Lausanne, le 24 novembre 2020

pour Le Doyen
de la Faculté de biologie et de médecine



Prof. John PRIOR
Vice-Director of the Doctoral School

À Mashka

Sub specie aeternitatis

À mes parents : Merci pour votre inépuisable soutien dans toutes ces studieuses aventures. Tout ceci n'aurait jamais pu être possible sans vous.

ACKNOWLEDGEMENTS

Un immense merci à Marc RR de m'avoir donné l'opportunité de faire cette thèse. Ce fût un plaisir de travailler avec toi. Merci pour ta disponibilité, ton ouverture et ta pertinence d'esprit.
I acknowledge all the great people that are working in Marc's lab for their availability and their useful discussions.

In addition, I would like to thank my dissertation committee for their evaluation of this work during my mid-thesis and final exam, Prof. Andreas Wagner, Prof. Oliver Delaneau, Prof. Ana Claudia Marques, and Prof. Ueli Schibler, and my thesis president, Prof. Paul Franken, for his support and availability during these years.

Enfin, je remercie en particulier :

Ma soeur
Tenace et téméraire, il y a quelque chose de très familial ! Merci d'être toujours la.

Jacques
14 ans de mariage ! Mon adolescence et ma jeune vie d'adulte furent heureuses en grande partie parce qu'elle fut avec toi. J'en garde des centaines de souvenirs très spéciaux. Merci ma poule pour ton soutien indéfectible. Reste gronchon, grognon et tout ce qui commence par gro. Oui ça peut être des mots composés.

Yannis
Tant de choses partagées, de Malaga à Berlin, des délicieux couteaux en passant par les soirées médecine, on en aura fait des choses ensemble mon poto. Reste comme tu es ma poule.

Mika
Quelle superbe médaille en or pour parer la poulette la moins matérialiste qui soit. A t-on déjà vu chose plus invraisemblable ? Reste comme tu es, BG.

Loic
Merci pour ton soutien ma poule. Aux vacances au portugal, espagne, autriche ... On en a parcouru du pays !

Papi
Les promesses sont les promesses. 41 ans (parce que je suis sympa), pas mariés, et à nous la baraque ! Animaux welcome, childrens tolerated. Waiting for it baby.

Léa
Ma grosse caille, tu es l'oreille de mes confidences. Merci pour ta bienveillance impérissable. Comme on l'aime ton répondant sans détours, bien direct, bim en pleine poire !

Nastasia
Droite, enthousiaste, pleine d'entrain et de curiosité de la vie. Qu'est qu'on aime ton énergie !

Margaux
J'ai toujours un plaisir immense à discuter des heures de tous ces sujets qui nous passionnent, comme de la folie, ou devrais-je dire de l'inadéquat ? Tu trônes juchée entre l'arbre déglingué et celui de la clarté d'esprit astucieusement aiguisée. Merci d'être toujours la.

Alex
Merci d'être la ma poule. Je passe toujours des moments en or avec toi, qu'est ce qu'on se marre. La vie est belle.

Ben
De très beaux souvenirs au grau, en mer et tous les endroits cabotés. Sacrés haut le cœur dans cette mauvaise mer ! Tu es très spécial ma poule, reste ainsi Capitaine.

Olivier
Alala, qu'est qu'on peut se marrer avec toi. C'est toujours un plaisir de passer du temps avec toi. J'espère qu'on pourra se voir plus souvent à l'avenir.

Julien et Fanny
Aux grosses déconnades, aux bonnes blagues bien noires, comme on les aime ! C'est toujours un plaisir de passer du temps avec vous, à discutailler de choses antiques et autres moeurs athéniennes. On en apprend des choses !

Clem
Toujours dans la diplomatie discrête, la déconne, j'aime passer des vacances avec toi. On se marre, c'est simple, authentique. La vie est parfaite telle quelle.

Sarah
Il en a de la chance ce Mehul ! C'est toujours un plaisir de te voir, on aimerait bien que ce soit plus souvent. Gros bisous ma belle

Laura
Sacré Ralotte, toujours dispo pour l'apéro et c'est ca qu'on aime. C'est toujours un grand plaisir de passer du temps avec toi.

Jano
Du matheux boutonneux au teufeur, du sans alcool à disons un peu plus que ça, de la cabane au fond du jardin à la maison de maître, sacré Jano ! Tu crois que tu vas passer de toubib à Skippy ? Si ya du pognon à se faire sur la bêtise humaine, je veux bien y participer.

Konstantina
Difficile de bosser avec toi pipelette. C'est un plaisir de compter parmis tes amis.

Mes cousins, Raphaël, Claire, Jordan, Max, Léa, Sarah, Diane
Aux nombreuses vacances partagées dans notre jeunesse et à toutes ces belles pêches !

André Bouillot
Des doigts en or pour un chirurgien en or. Continu d'être hors norme, c'est pour ça qu'on t'admire !

Haimara, Michel et François
À la grande finesse des blagues tellement bien ficelés qu'on ne les saisit pas toujours. Même si c'est souvent plutôt du rafistolage que de la grande couture. Merci pour votre bienveillance, c'est toujours un plaisir de vous voir.

Leblond et André
Belle et intense période de la petite salle à préparer ce fichu concourt. J'espère continuer à pouvoir vous voir de temps en temps mes poulettes.

Julien et Chloé
Toujours des supers moments partagés avec vous. Pas assez, je sais bien ...

Elza
Merci beaucoup pour m'avoir aidé à corriger les fautes d'anglais de ce manuscrit :-). Maintenant que tu es de retour d'Angleterre j'espère qu'on pourra se voir plus souvent !

≪Jadis l'esprit était Dieu, puis il devint homme, maintenant il s'est fait populace.≫

Ainsi parlait Zarathoustra, Nietzsche

RÉSUMÉ GRAND PUBLIC

Depuis des millions d'années, le vivant s'est progressivement synchronisé sur le rythme de l'alternance du jour et de la nuit. La terre met environ 24 heures pour réaliser une rotation complète, donnant une alternance quotidienne d'exposition à la lumière du soleil. Ainsi, la nature est gouvernée par des cycles énergétiques dus aux cycles lumière-obscurité.

L'expression des gènes, c'est-à-dire l'ensemble des processus qui transforment l'information génétique qui se trouve dans l'ADN en protéines biologiquement fonctionnelles, est une première étape de régulation des activités biologiques. Dans nos cellules, on constate que de très nombreux gènes sont concernés par une utilisation périodique. Au cours de cette thèse, je me suis particulièrement intéressé à ces gènes, c'est-à-dire à ceux qui sont exprimés de façon périodique, toutes les 24 heures, période appelée nycthémérale. En effet, certains gènes sont plus utilisés en début de journée, d'autres plus tard dans la journée, d'autres en début de nuit, d'autres encore à la fin de la nuit. Ainsi, la production de nombreuses protéines dans les cellules oscille avec une périodicité de 24h. Ces processus périodiques permettent à chaque cellule de produire en grande quantité, à certaines heures du jour ou de la nuit, des protéines qui assurent la même fonction ou des fonctions compatibles.
Le rythme circadien est un système autorégulé, capable de s'auto-générer, fournissant aux organismes la capacité d'anticiper les changements de leur environnement sur une échelle de temps de 24 heures. Un gène rythmique est un gène qui présente une variation quotidienne de l'abondance de sa molécule intermédiaire transcrite (l'ARN) ou de sa protéine traduite. Ces variations quotidiennes peuvent également être entraînées directement ou indirectement par des facteurs environnementaux tels que l'alternance des périodes de lumière-obscurité, de l'alimentation, des variations de température ou des activités sociales. C'est pourquoi il y a de nombreux gènes dont l'expression rythmique n'est pas autonome, mais directement ou indirectement entraînée par l'environnement lui-même. Une telle expression périodique des gènes est retrouvée presque partout dans la Nature : chez les animaux, les plantes, les bactéries et les champignons. 20% à 50% des protéines sont produites périodiquement (toutes les 24 heures) à partir d'ARN présent en quantité constante. Et inversement, il existe de nombreux ARN dont l'abondance est périodique alors que ce n'est pas le cas pour leurs protéines. Mais alors, pourquoi ? Mes résultats suggèrent que les variations quotidiennes concernent des protéines produites en quantité relativement abondante. La production en grande quantité de ces protéines est coûteuse pour la cellule. En effet, la fabrication de protéines nécessite une certaine somme d'énergie et de matériaux moléculaires que la cellule n'a pas forcément en quantité infinie. De surcroit, ces protéines périodiques seraient, en effet, encore plus coûteuses à produire si la cellule avait dû maintenir en permanence (i.e. de manière constante) un niveau suffisamment élevé de protéines pour assurer leur fonction biologique. En effet, les coûts de production des protéines sont suffisamment conséquents pour être soumis à la sélection naturelle. A contrario, les coûts de production des ARN sont probablement trop négligeables pour soutenir l'hypothèse que leur variation serait due à une stratégie d'économie pour la cellule. Par contre, pour un gène donné, la quantité de son ARN joue un rôle dans la variabilité du nombre de ses protéines entre les cellules. Ceci s'explique car la quantité de protéines n'est pas exactement la même dans chaque cellule. Plus la production de protéines par unité d'ARN est grande, plus la variabilité entre cellules est faible. Il semblerait que les gènes qui ont des ARN qui varient quotidiennement ont en moyenne une variabilité entre cellules plus faible que celle des autres gènes. Au cours des millions d'années d'évolution du vivant, des individus ont présenté, par hasard, des modifications qui, dans leur environnement, leur ont procuré un avantage par rapport aux autres individus. C'est le principe de sélection naturelle. Pour certains

gènes, leur fonction requiert un niveau élevé de protéines. Les gènes qui produisent périodiquement des protéines sont des gènes dont la quantité requise en protéines s'est avérée être coûteuse pour la cellule, et de fait, ce coût a donc été soumis à la sélection naturelle. Il est possible que la variation périodique des ARN dont les fonctions biologiques sont sensibles à de grandes différences de quantité de protéines entre les cellules (variabilité), ait également apporté un jour, un avantage suffisamment important à l'individu dans son environnement cyclique pour être sélectionné au cours de l'évolution. Comprendre les systèmes périodiques au sein du monde vivant nous aide à mieux appréhender les relations étroites que nous entretenons avec notre environnement. Cela passe par une compréhension des dynamiques temporelles qui s'opèrent au sein de nos cellules. La chrono-médecine, et peut-être la chrono-chirurgie, prendront un jour en compte le « tic-tac » qui règne dans l'expression de nos gènes ainsi que dans l'organisation temporelle des événements biologiques au cœur de nos cellules.

RÉSUMÉ

Les horloges circadiennes constituent aujourd'hui une partie importante de la compréhension des systèmes biologiques. Elles sont ubiquitaires, retrouvées dans un large éventail de processus biologiques, allant des systèmes moléculaires au comportement, et sont aussi trouvées presque partout dans la nature : chez les animaux, les plantes, les bactéries et les champignons. Cette thèse se concentre sur les systèmes biologiques qui répondent à des facteurs oscillant sur une échelle de temps de 24 heures.

La détection de tels gènes reste un aspect compliqué du travail d'analyse. Nous montrons que la plupart des méthodes de détection ne sont efficaces que pour des signaux intenses et qu'en-dehors de ceux-ci les algorithmes semblent détecter des gènes rythmiques de manière assez aléatoire.
Nous avons également cherché à comprendre pourquoi des gènes présentent des variations périodiques de la quantité de leur ARN ou de leurs protéines. En effet, 20% à 50% des protéines accumulées de manière cyclique (i.e. nycthéméraux) sont traduites à partir d'ARNm non-oscillants, et inversement, il existe de nombreux ARNm qui oscillent mais pas les protéines qu'elles codent. Mais alors, pourquoi ? Mes résultats suggèrent que la variation nycthémérale des protéines concerne des protéines hautement exprimées en moyenne, qui restent en moyenne plus coûteuses à produire pour la cellule comparativement aux protéines produites de façon non-rythmiques (en terme d'énergie et de matériel moléculaire). De surcroit, ces protéines rythmiques seraient en effet bien plus coûteuse à produire si la cellule avait dû maintenir en permanence (i.e. de manière constante) un niveau suffisamment élevé pour assurer la fonction. Les coûts de production de protéines sont suffisamment conséquents pour être soumis à la sélection naturelle, alors que les coûts de productions des ARNm ne le sont pas. Mais alors pourquoi les cellules produisent-elles périodiquement des ARNm ? Mes résultats suggèrent que l'oscillation périodique de la quantité des ARNm concerne des gènes qui ont en moyenne une variabilité (bruit) de cellule-à-cellule plus petite que les gènes avec des niveaux constants d'ARNm. La causalité n'étant pas établie, il est tout de même possible que la rythmicité des ARNm permette d'optimiser la précision d'expression pour des fonctions sensibles au bruit durant un certain laps de temps, et ce, de manière répétée, toutes les 24h. Enfin, la rythmicité des ARNm concerne des gènes qui ont subi une forte sélection purifiante. Cette forte sélection purifiante ne semble pas concerner des gènes qui ont des niveaux de protéines périodiques, bien que les données soient insuffisantes pour vraiment aller plus loin dans la formulation d'une explication évolutive.

Dans l'ensemble, il est possible que la rythmicité de l'expression des gènes ne fournisse un avantage adaptatif qu'aux espèces vivant dans des environnements très changeants (sur 24 heures). Dans de tels environnements, c'est-à-dire pour une grande partie des écosystèmes marins et terrestres, il est possible que la rythmicité de l'expression des gènes ait pu permettre la préservation de nouvelles propriétés biologiques complexes et coûteuses qui autrement auraient été éliminées. Les compromis évolutifs prennent en compte les avantages apportés par la fonction, ses coûts d'expression et la précision requise, et peut-être aussi la variabilité d'expression conduisant à une diversité phénotypique améliorant l'adaptabilité des individus dans des environnements fluctuants.

ABSTRACT

Circadian clocks are now an important part of the understanding of biological systems. They are ubiquitous, found in a wide range of biological processes, from molecular systems to behavior, and are also found almost everywhere in nature: in animals, plants, bacteria and fungi. This thesis focuses on biological systems that respond to factors oscillating on a 24-hour time scale.

The detection of genes expressed with a periodicity of 24 hrs remains a complicated aspect of analytical work. We show that most detection methods are efficient only for strong signals and that outside of these genes, the algorithms seem to detect rhythmic genes in a rather random way.

We have also tried to understand why genes have periodic variations in the amount of their RNA or their protein they encode. Indeed, 20% to 50% of cyclically accumulated proteins (i.e. nycthemeral) are translated from non-oscillating mRNAs, and conversely, there are many mRNAs that oscillate but not the proteins they encode. Why is that? My results suggest that the nycthemeral variation of proteins concerns on average highly expressed proteins, which remain on average costlier to produce for the cell (in terms of energy and molecular material) compared to other proteins produced in a non-rhythmic way. Moreover, these rhythmic proteins would be even more expensive to produce if the cell had to maintain constantly a sufficient high effective level of these proteins to ensure the function. The costs of protein production are large enough to be under natural selection, whereas the costs of mRNA production are not. So, why do cells periodically produce some mRNAs? My results suggest that the periodic oscillation in mRNA quantity concerns genes that have on average weaker cell-to-cell variability (noise) than genes with constant mRNA levels. Since causality is not very clear, it is still possible that the rhythmicity of mRNAs may optimize the expression precision for noise-sensitive functions over a period of time, repeatedly, every 24 hours. Finally, mRNA rhythmicity concerns genes that have undergone a strong purifying selection. This strong purifying selection does not seem to concern genes that have periodic protein levels, although there is insufficient data to really go further in the formulation of an evolutionary explanation.

Overall, I suggest the hypothesis that rhythmicity of gene expression provides an adaptive advantage only to species living in highly changing environments (over 24 hours). In such environments, i.e. for a large part of marine and terrestrial ecosystems, it is possible that the rhythmicity of gene expression could have allowed the preservation of complex and costly new properties that would otherwise have been eliminated.
The evolutionary trade-offs take into account the advantages provided by the function, its expression costs and precision required, but maybe also the variability of expression leading to phenotypic diversity improving adaptability in a fluctuating environment.

Key words: Circadian rhythm, nycthemeral rhythm, gene expression, tissue-specificity, evolutionary conservation, rhythm detection algorithms, benchmark, expression costs, noise, fitness, phenotype

Contents

NOMENCLATURE

- Per1/PER1: Per1 for the core clock gene, PER1 for its protein.

UNITS

- Frequency: Hz (s^{-1})
- Expression Noise: The noise strength is based on the standard deviation σ and the mean gene expression μ. The unit depends on the method used (FPKM for the Fano factor method). We used the Barroso et al. method [1] which defines $F*$ as the ratio of the observed variance over the variance predicted by the mean expression level. Thus, $F*$ is unitless.
- Expression Cost: Here, we used the costs of making amino acids (AA) which is normally equal to the cost of making AA from the precursor and the cost to produce or to extract the precursor from the metabolism. This cost can be considered in energy lost in the form of the number of high-energy phosphate bonds (∼P) carried in ATP.

LIST OF TOOLS

Main softwares and utilities used in the current work:

Main Tools:

- R language, for integration, analysis, and visualization
- LaTeX, for papers and thesis writing, typesetting, formatting, and performing pdf output
- Overleaf, for cloud-based LaTeXedition
- BibTeX, for management of bibliography and reference database
- UNIL High Performance Computing clusters, for heavy or long processes
- GitHub, for managing and sharing data and code

Auxiliary Tools:

- Python language, for writing ancillary scripts (e.g. raw data)
- Perl, for some easier modifications of file contents.
- Microsoft Office
- Adobe Photoshop
- GNU Make, to automatically build libraries and files from the Unix-like system

INTRODUCTION

1 INTRODUCTION

1.1 Cyclic biological systems

1.1.1 Periodic systems

Among biological systems, a large number of oscillator systems are found, such as the cell cycle, circadian rhythms, calcium oscillations, or the oscillatory behaviour of the cardiac electrophysiological system. Such repetitive phenomena - attributed to rhythms, oscillations or cycles - are also retrieved at all functional scales of living organisms, from biochemical reactions to seasonal migratory behaviors. Just as there is a wide range of oscillatory systems, these oscillations cover a large frequency range. One can cite the high frequency rhythms detected by electroencephalogram such as alpha oscillations ($\sim$ 8-12 Hz) probably originated by thalamo-cortical interactions [2], periodic iterations of rapid eye movement sleep-like states which last around 2.42 minutes in cephalopods[3], or seasonal growth rates in Krill [4].

1.1.2 Circadian rhythms

a. Short history

Among biological periodic systems, the circadian rhythms (latin *circa* "around" and *diem* "day") is one of the most studied one. It is characterized by periodic biological processes that take around 24 hrs. The nyctinastic movement of some plants was the first circadian rhythm recorded around the 18th century. The nyctinasty is a nycthemeral (greek *nukthêmeron*, *nux*, *nuktos*, "night", and *hêmera*, "day") nastic movement of some plants in response to the day/night alternation. In 1729, astronomer De Mairan (Jean-Jacques d'Ortous, 1678-1771) found that such leaf movements of *Mimosa pudica* (also called the "sensitive") that normally track the sun during the course of a day were maintained in continuous darkness with a period that was approximately 24 hrs. The persistence of nyctinatic movements despite the lack of perception of night and day became the first signature of circadian rhythms as defined by Franz Halberg in 1959.

b. Definitions

`Nycthemeral rhythm` is defined as any biological cycle repeated every $\sim$24 hrs.

`Circadian rhythms` characterize any endogenous, autonomous, and entrainable oscillatory system with a period of about 24 hrs.

`The circadian clock` is the central mechanism that sets the timing for many circadian rhythms allowing to regulate processes such as sleep/wake cycles, hormonal activity, or feeding.

`"Circadian"`: Thus, usually when we use the term "circadian" we are referring to its endogenous nature. For instance, the "circadian period" of a peripheral (in one tissue) clock refer to the intrinsic period the system would display in constant condition and without the control of the central oscillator (SCN). This is also called the `"free-running"` period.

`Zeitgebers` ("time givers") are external or environmental cues able to entrain or synchronize circadian rhythms.

The period is the time after which a defined phase of the oscillation re-occurs. In practice, it is the time separating two consecutive peaks.

The phase is a time t of an oscillation within a period. In practice, we use the time-point of the peak of expression.

The amplitude is the measure of one half of the extent of the rhythmic change, i.e. the difference between the maximum and the half the value of the range of oscillation estimated from the best fitting curve (generally estimated by cosine curve model).

The entrainment is defined as the coupling of one biological rhythm to an environmental oscillator with the result that both oscillations have the same periods. Thus, in the case of an endogenous rhythm, its phase is affected by entrainment. For instance, in Light-Dark (LD) conditions, the circadian clock is entrained by light cues and consequently the period is exactly 24 hours. In contrast to the Dark-Dark (DD) condition, where the period (intrinsic circadian period) is longer or shorter (species and individual -dependant).

Core clock genes are genes whose products are necessary components for the generation and regulation of circadian rhythms.

Clock controlled genes (CCGs) are immediate downstream clock genes in the regulatory networks. These CCGs contain transcription factor binding sites (TFBS) associated with the clock in their promoter or nearby enhancers.

Rhythmic genes are all genes displaying a 24 hrs periodic variation of their mRNA level or their protein level, or both, constituting the nycthemeral transcriptome and proteome. The rhythmic expression of these genes can be entrained directly by the internal clock but also directly or indirectly by external inputs, such as the light-dark cycle or food-intake [5][6][7][8][9].

Oscillating, rhythmic, cycling, or periodic are used in the same meaning: fluctuation repeated every 24 hrs.

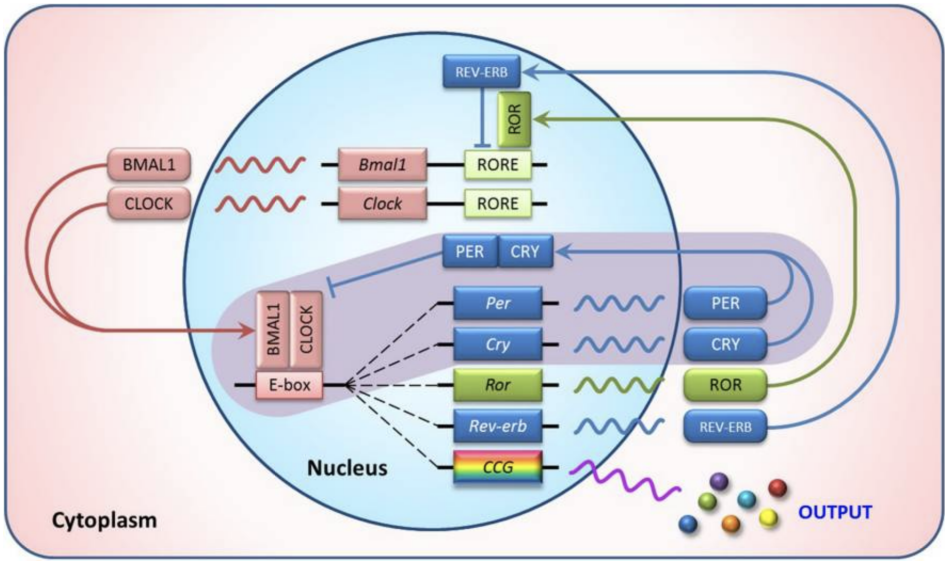
c. Specificity of circadian rhythms

Circadian rhythms are endogenously generated, i.e. the system is able to self-sustain the oscillations, allowing the rhythm to persist under constant, time-free conditions. Their phase can be altered by Zeitgebers and they are temperature compensated, i.e. the clock has a temperature-insensitive period, while retaining the ability to synchronize to temperature cycles. This last specificity might be explained by the fact that the circadian oscillator evolved from an adaptive temperature sensor (a gene circuit that responds only to temperature changes) as shown in Drosophila circadian clock [10].

d. The Core clock system

Core clock system: The Transcriptional-translational delayed feedback loop (TTFL)
The transcriptional-translational delayed feedback loop is a gene regulatory network motif. A gene expression oscillation can be produced and sustained. The principle of the TTFL is that a protein A activates a gene

b, protein B is produced and then represses gene a. After a time-delay as gene a is not transcribed and protein A is degraded, gene b is not activated anymore. Afterwards, protein B is degraded gene a is not repressed and can be transcribed again. Delays (transcription, splicing, post-transcriptional modification, translation, and post-translational modifications) permit to obtain an oscillatory system with a 24 hrs period. Thus, it consists of positive and negative elements such as BMAL1 and CLOCK proteins (positive regulators), and PER1, PER2, CRY1, and CRY2 proteins (negative regulators) in mammals. PER for "period", CRY for cryptochromes. In mammals, the circadian system consists of two loops (Figure 1.1).



Chen and Yang, 2015 [11]

Figure 1.1: Transcriptional feedback loops of the mammalian circadian clock. In the core loop (purple background), BMAL1/CLOCK heterodimer activates transcription of the Per and Cry genes via binding to the E-box elements in their promoter regions. The resulting PER and CRY proteins heterodimerize, translocate to the nucleus and interact with the BMAL1/CLOCK complex to inhibit their own transcription. In addition, ROR activates and REV-ERB represses RORE-mediated transcription, forming the secondary autoregulatory feedback loops. This clock mechanism also controls rhythmic expression of numerous genes, called clock controlled genes (CCG), to perform biochemical or physiological roles in a circadian manner. Source [11]

e. Characteristics of circadian rhythms

Interestingly, circadian rhythms are ubiquitous in scale - they exist in a broad array of biological processes (molecular, cellular, organs, and behavioural) - and in nature (they are found in animals, plants, bacteria, and fungi). In addition, core clock genes are essential for the maintenance of circadian rhythm in constant conditions. Nevertheless, each of them is not 100% indispensable, i.e. the knock out of one gene does not suppress rhythmicity, there seems to be a plasticity partly sustained by paralogs genes [12][13]. Furthermore, the Supra-Chiasmatic Nucleus (SCN) play the role of core pacemaker of circadian rhythms in mammals. It allows to maintain rhythmic expression of systemic network in constant conditions and to orchestrate peripheral tissue clocks (setting the phases of peripheral oscillators) through secretion of endogenous regulatory factors [14]. Until recently, the SNC was considered as the main or unique relay to synchronize physiology

with environmental changes. Finally, circadian rhythms can be entrained to shorter or longer periods, but not more than 28 hrs.

1.2 Rhythmic gene expression

1.2.1 Gene expression and function

Gene expression embodies all synthesis, degradation, and regulation processes that determine the levels of gene products: RNAs (transcriptome) and proteins (proteome). It's the most fundamental level at which the genome is active for many living organisms. Knowledge of gene expression patterns is now absolutely complementary with knowledge of the genome. Indeed, the detection of expression in specific conditions is indicative of function since gene expression makes the bridge between the transmitted genome and the macroscopic phenotype of individuals. The study of mutant organisms permits an understanding of genes and their expressions, their functions and their essentiality. In my work, I used comparative analysis between species which is another approach to explore gene function. I introduce these concepts in section 1.3.3.

1.2.2 Rhythmic genes: rhythmic mRNA or rhythmic protein abundances, or both

Rhythmic genes are usually rhythmically produced RNAs or proteins. To be more precise, we define a rhythmic gene as a gene which displays a nycthemeral change in the abundance of its mRNA or protein (or both), i.e. occurring over 24 hours and repeated every 24 hours. They represent the nycthemeral transcriptome and proteome. The rhythmic accumulation of these gene products can be induced by many potential regulatory factors (transcriptional activity, splicing, ...) as I will discuss in the following section (section 1.2.3). Their rhythmic regulation can be entrained directly by the internal clock or indirectly by external inputs, such as the light-dark cycle or food-intake (See section Per Se/Per Exter 1.2.5).

In this work, I use equivalently the terms: expression, abundance, or accumulation, since the rhythmicity of genes is based on time-series data of counts of mRNA or protein sequences. To be specific, I distinguish the following terms:
`Gene expression` means all the processes by which a gene sequence is read to produce an efficient quantity of functional molecules.
`Abundance` means the quantity of a given molecule at a given time t.
`Accumulation` means the number of a given molecule produced per time unit. It includes the accumulation due to the production rate (transcription, translation), compensated by the decay rate, and the molecule half-life (although the latter is sometimes defined using the first two).

1.2.3 From genes to proteins: Genesis and transmission of the rhythmic information

Such nycthemeral variations have been observed in almost all steps, from transcription initiation to post-translational modifications. For instance, certain histone modifications have been reported to follow nycthemeral changes [15][16][17][18]. The transcriptional activity [17][18], the alternative splicing [19][20], and the translational activity regulated for instance by rhythmic `polyadenylation` (poly-A) of RNA tails (role in the RNA stability) [21] have been reported to be rhythmic as well. One can also cite the possible major role of the translation of mRNAs upstream open reading frames (uORF) into the free-running period [22] that might explain the tissue-specific circadian periods [6] due to cell-type-specific uORF usage [23].

Finally, the occupancy and availability of ribosomes whose biogenesis is itself regulated by the circadian clock [24], the RNA or protein degradations [25][26], and the post-translational modifications (such as the phosphorylation or acetylation) [27][28][29], also play an important role in the rhythmic accumulation of transcripts and proteins.

Thus, it becomes easy to observe the huge amount of constraints on transmission of rhythmic information across these multiple regulatory layers of gene expression for which two parameters must be taken into consideration: the damping (of the relative amplitude), and the delay (of the rhythm). For instance, gene products with shorter half-lives preserve rhythmic information more efficiently than those with longer half-lives. This is partly the reason why, for instance, only a limited overlap was found between the rhythmic accumulation of pre-mRNA and that of mature mRNA [17][18]. This is also believed to be the reason why many proteins, encoded by rhythmic mRNAs, are not rhythmic, due to long protein half-lives [26]. Given a moderate half-life, the strongest oscillations can be expected if production and degradation occur in anti-phase (i.e., if they have a phase difference of 12 hr in a 24 hr period) [25]. Two recent studies have indicated that 20% to 50% of cyclically accumulating proteins are expressed from nonoscillating mRNAs [29][26], i.e. they are produced by translation or protein degradation. One must note that clock genes showed constant translational rate (many papers use the "translational efficiency" terms), indicating exclusion from time of day–dependent translational control [22]. The delays between mRNA and protein accumulations that have been reported for several core clock components might have post-translational origins and be linked to the translation of mRNAs upstream open reading frames (uORF) [22]. Most cyclically transcribed RNAs are translated at one of two major times in a 24-h day [30][22], their amplitudes is damped and their phases are shifted by an average of 5.5 h (in mouse liver) [26][31]. Finally, it seems that all proteins accumulated rhythmically and encoded by non-oscillating transcripts are phased to a single time of day (this peak time seems, however, experiment-dependant since it was not the same one for the three studies that follow)[31][30][22] which seem to be associated with specific pathways such as iron metabolism (through the rhythmic translation of transcripts containing iron responsive elements), protein biosynthesis machinery (including ribosomal proteins), and poly(A) binding proteins [22].

Note: Here and in the rest of the Introduction, I will introduce my thesis work in italics.
Thus, genes can be separated in four sets (cycling transcripts and cycling proteins, cycling transcripts and non-cycling proteins, non-cycling transcripts and cycling proteins, and non-cycling transcripts and non-cycling proteins) whose contents are individual- and conditions-dependant. I propose some evolutionary explanations to explain why some genes have rhythmic transcripts abundances or rhythmic proteins abundances, or both. Several parameters can be explored to better understand the presumed evolutionary advantages brought by their rhythmic regulation rather than non-rhythmic expression.

1.2.4 Tissue-specificity

The nycthemeral transcriptome has long been known to be tissue-specific [32][33][34], i.e. a given gene can be rhythmically accumulated in some tissues, and constantly or not expressed in others. The rhythmic proteome is no exception to this rule.
Only recent studies seem to provide some mechanistic explanations. Yeung et al. have shown that regulatory mechanism behind this tissue-specificity of the rhythmic transcriptome seems to be due to precise chromatin loops recruiting clock- and tissue-specific transcription factors (TFs) which generate these tissue-specific

rhythms [35]. Beytebiere et al. have also shown that DNA binding of the core clock gene BMAL1 is largely tissue-specific, likely because of differences in chromatin accessibility between tissues and co-binding of tissue-specific transcription factors that would allow BMAL1 to have the ability to drive tissue-specific rhythmic transcription [36]. Finally, other recent studies suggest that tissue-specific rhythmic oscillations are controlled at the translational level with a role of the translation of cell-type-specific uORF mRNAs [37][38][22].
Apart from these mechanistic explanations, part of my work has been to propose an evolutionary reason justifying why there is a tissue-specific rhythmic expression. What are the characteristics that make the rhythmic expression of these genes advantageous for the tissue? Are these tissue-specific expressed genes?

1.2.5 Per se / Per exter

De Mairan (Jean-Jacques d'Ortous, 1678-1771) suggested that diurnal variations in temperature, likewise in light, could synchronise circadian rhythms. Indeed, these two oscillating environmental factors have since been considered as the two main `zeitgebers` of circadian rhythms. Many other factors can be considered as zeitgebers such as social interactions, feeding, or systemic signals. Each nycthemeral intermediate or product could be theoretically a zeitgeber.
Autonomous rhythmicity has been monitored at different scales. Individual cells (in vitro cultured fibroblasts) have been shown to be able to harbor self-sustained and cell-autonomous circadian clocks suggesting that potentially every cell in an organism can resonate with environmental time [8]. Peripheral tissues have shown autonomous rhythms as well (shown in mouse) [6]. Inter-cellular coupling are complex, cells are neither fully independent of each other, nor an entirely homogeneous population [39]. In addition, Cyanobacteria possess one of the simplest known circadian clocks consisting of a cluster of three tandemly located genes called: kaiA, kaiB, and kaiC. An even simpler model is found in many prokaryotic genomes where kaiA is missing but still display circadian cycles [40].
However, the last $\sim$15 years of research show how complex the regulation of nycthemeral rhythms is at cellular scale. The rhythmic expression concerns a poorly defined set of genes whose rhythmic entrainment seems to be regulated directly or indirectly by a huge amount of factors. Feeding for instance, has been shown to be the main driver of rhythmic translation in the liver [41], for which half of rhythmic proteins did not come from rhythmic mRNAs, suggesting a translational responsiveness to feeding. From a macroscopic phenotypic view, recent research studied the direct light effects independent from the circadian process and shows that a quarter of the nycthemeral sleep-wake cycle is directly sustained by the light (SCN-independent) implying a non-circadian function for the central structure comprising the master circadian clock (SCN)[5]. This means that the SCN should be seen not only as the orchestrator of circadian rhythms, but also as a critical intermediate of the direct photic regulation independent of the central clock. Finally, peripheral clocks are also capable of sustaining synchronized rhythmicity under zeitgeber-free conditions in the absence of the SCN pacemaker [42], showing how complex the network is; who regulates whom, and from what? Each product can potentially be a rhythmic driver for another. These results illustrate how difficult it is to know if an observed rhythm is generated by the circadian clock or by one of the innumerable external factors such as feeding, light, temperature, social interactions, sleep-wake behavior, atmospheric pressure, etc. Thus, many nycthemeral systems display non-autonomous rhythms, directly or indirectly controlled by the local clock or mainly by the environment itself, or both [5][6][7][8][9]. Many environmental factors can be seen as strong and regular regulators.
This is one of the reasons why we considered the entirety of observed rhythms in the transcriptome and the proteome to be biologically relevant. Especially if the data came from experiments carried out in normal

conditions (light-dark cycles, i.e. not in free-running conditions).

1.3 Rhythmicity and Evolutionary conservation

1.3.1 Adaptability in cyclic environments

a. Adaptive value of circadian rhythms

One of the major experiment showing how circadian rhythms improve the fitness of organism has been done in Cyanobacteria. The strains that had a circadian intrinsic period similar to that of the light/dark cycle (introduced in the lab) were favored under competition [43]. In wild type and in long- and short-circadian period mutants of Arabidopsis thaliana, plants with a clock period matched to the environment contain more chlorophyll, fix more carbon, grow faster, and survive better than plants with circadian periods differing from their environment [44].
Thus, circadian rhythms provide advantages that improve fitness by improving survival or fecundity due to a better adaptation to the environment changes [45][43][44]. Indeed, this endogenously generated rhythm provides a time framework allowing organisms to synchronize physiological processes to their cyclic environment and to anticipate its changes.

b. 24hrs, the strongest influence on Earth

The ubiquity of nycthemeral rhythms in life can be seen as a high adaptability of living organisms to one of the most robust external pattern in nature: the 24-hours cyclicality. It is due to the periodic earth rotation on itself, exposing it to sunlight. That is why, light/dark cycles and nycthemeral variations of the temperature govern the energetic cycles on Earth. Darwin and Wallace rapidly realized that natural selection always operates within an environment [46], which means that it is not possible to separate the process of evolution from the surroundings in which it occurred.
Nycthemeral timing systems have evolved through adaptation to periodic factors in the geophysical environment. The light-dark cycle have been operating as a rhythmicity source operational for billions of years (although the period was shorter due to the shorter moon-earth distance).

1.3.2 Rhythmicity and Fitness

Apart from the circadian system that confers evident advantages due to autonomy and time-anticipation, the responsiveness of some pathways to oscillating environmental factors with day-night time-scales should improve the fitness as well. Indeed, this responsiveness is relevant in the context of behavior, neural plasticity, physiology, sleep, navigation, sociobiology, migration, hibernation, life history, adaptation, etc. Fitness means the ability to survive, find a mate, and reproduce. Basically, the more offspring an organism produces during its lifetime (or of higher quality), the greater its biological fitness is. Organisms with good responsiveness to changes in their environment should improve their fitness. Indeed, organisms with physiological systems able to adapt to changing circumstances are expected to have a higher stability of survival and reproduction [47]. Fitness is thus conceivable in the context of nycthemeral gene expression since it improve their adaptation in environments with nycthemeral changes. Adaptation is the process by which populations of organisms evolve in such a way as to become better suited to their environments as advantageous traits become predominant [48]. This thesis focuses on systems that respond to factors oscillating on a 24-hour

time scale. Which is not so obvious at the molecular level considering that most biological oscillators have a generally shorter period, i.e. ultradian (∼3h for the p53-Mdm2 system for instance).

1.3.3 Evolutionary conservation of rhythmic gene expression

Evolutionary conservation provides a valuable filter through which we can highlight functional biological networks, such as, for instance, for clock-controlled functions [49]. Since function tends to be conserved between orthologs [50][51], nycthemeral genes which are biologically relevant can be highlighted by the evolutionary conservation signal based on comparative approaches between species. Structures can be compared if they are homologous, i.e. if they descend from a structure in a common ancestor. Homology is thus a tool that can be used to better understand gene expressions, their functionality level in organisms. Despite the pervasiveness of biological clocks among species, functional properties of the broader set of all oscillating genes remain largely unexplored [52].
One of the main topic of this thesis has been to apply evolutionary conservation, representing selection on function, to cyclically expressed genes.

1.3.4 Rhythmic gene expression as an evolving phenotype

Similarly as "genotype", some phenotypes are now conceived in quantitative and measurable terms on a comprehensive molecular level. Expression profiles are seen as new phenotypic traits that extend the classical concept of the phenotype [53]. Thus, mRNA levels can be seen as phenotypic traits. That is why we consider the rhythmic transcriptome and the rhythmic proteome as phenotypic traits (Figure 1.2), whose modifications are subject to neutral, positive, or purifying selection in a given environment.

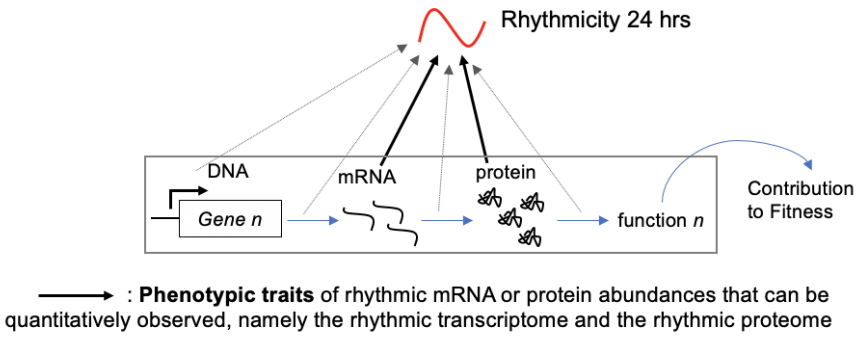


Figure 1.2: Nycthemeral transcriptome / Nycthemeral proteome can be seen as phenotypic traits

1.4 Benchmarking

The detection of rhythmic genes has been my first challenge during this thesis. Many algorithms exist but it has been difficult to sort them out. What do they really detect as rhythmic and which smart cutoff should we use? (Figure 1.3)
This is the question we asked ourselves. We performed a benchmark comparing seven known rhythm detection methods based on a comparative approach of rhythmic orthologs (transcripts) in animals, allowing to provide a functional sense (biologically) in rhythm's detection (RESULTS 1).
Furthermore, I developed a pipeline available from GitHub repository laloumdav/rhythm_detection_benchmark

containing data and reproducible code for this paper, and especially, that can easily integrate new datasets or algorithms.

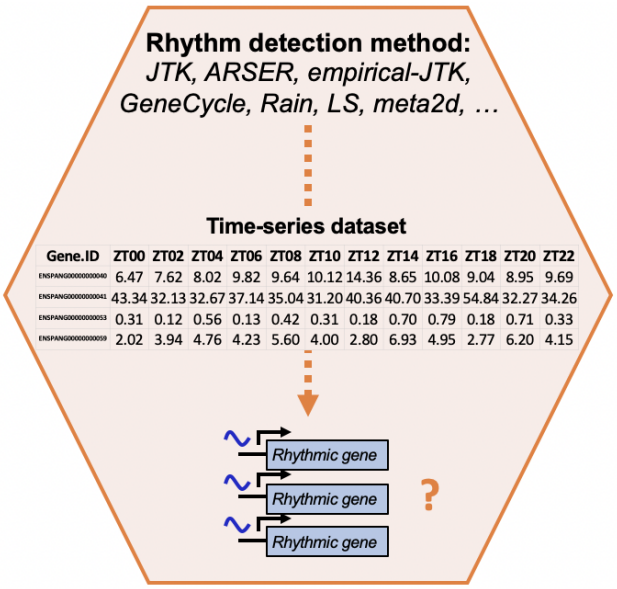


Figure 1.3: Methods detecting nycthemeral oscillations in gene expression time-series datasets

RESULTS 1

2 RESULTS 1: PUBLISHED ARTICLE, *Methods detecting rhythmic gene expression are biologically relevant only for strong signal*

2.1 Main text and figures

For this section, the bibliography is at the end of the article.

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Methods detecting rhythmic gene expression are biologically relevant only for strong signal

David Laloum[1,2], **Marc Robinson-Rechavi**[1,2]*

1 Department of Ecology and Evolution, Batiment Biophore, Quartier UNIL-Sorge, Université de Lausanne, Lausanne, Switzerland, **2** Swiss Institute of Bioinformatics, Batiment Génopode, Quartier UNIL-Sorge, Université de Lausanne, Lausanne, Switzerland

* marc.robinson-rechavi@unil.ch

OPEN ACCESS

Citation: Laloum D, Robinson-Rechavi M (2020) Methods detecting rhythmic gene expression are biologically relevant only for strong signal. PLoS Comput Biol 16(3): e1007666. https://doi.org/10.1371/journal.pcbi.1007666

Editor: Attila Csikász-Nagy, King's College London, UNITED KINGDOM

Received: August 22, 2019

Accepted: January 18, 2020

Published: March 17, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1007666

Data Availability Statement: Data are publicly available from the NCBI GEO database, as specified in the Materials/Datasets section.

Abstract

The nycthemeral transcriptome embodies all genes displaying a rhythmic variation of their mRNAs periodically every 24 hours, including but not restricted to circadian genes. In this study, we show that the nycthemeral rhythmicity at the gene expression level is biologically functional and that this functionality is more conserved between orthologous genes than between random genes. We used this conservation of the rhythmic expression to assess the ability of seven methods (ARSER, Lomb Scargle, RAIN, JTK, empirical-JTK, GeneCycle, and meta2d) to detect rhythmic signal in gene expression. We have contrasted them to a naive method, not based on rhythmic parameters. By taking into account the tissue-specificity of rhythmic gene expression and different species comparisons, we show that no method is strongly favored. The results show that these methods designed for rhythm detection, in addition to having quite similar performances, are consistent only among genes with a strong rhythm signal. Rhythmic genes defined with a standard *p*-value threshold of 0.01 for instance, could include genes whose rhythmicity is biologically irrelevant. Although these results were dependent on the datasets used and the evolutionary distance between the species compared, we call for caution about the results of studies reporting or using large sets of rhythmic genes. Furthermore, given the analysis of the behaviors of the methods on real and randomized data, we recommend using primarily ARS, empJTK, or GeneCycle, which verify expectations of a classical distribution of *p*-values. Experimental design should also take into account the circumstances under which the methods seem more efficient, such as giving priority to biological replicates over the number of time-points, or to the number of time-points over the quality of the technique (microarray vs RNAseq). GeneCycle, and to a lesser extent empirical-JTK, might be the most robust method when applied to weakly informative datasets. Finally, our analyzes suggest that rhythmic genes are mainly highly expressed genes.

Author summary

To be active, genes have to be transcribed to RNA. For some genes, the transcription rate follows a circadian rhythm with a periodicity of approximately 24 hours; we call these genes "rhythmic". In this study, we compared methods designed to detect rhythmic genes

Funding: Funding was received from Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung (173048, to MR-R). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

in gene expression data. The data are measures of the number of RNA molecules for each gene, given at several time-points, usually spaced 2 to 4 hours, over one or several periods of 24 hours. There are many such methods, but it is not known which ones work best to detect genes whose rhythmic expression is biologically functional. We compared these methods using a reference group of evolutionarily conserved rhythmic genes. We compared data from baboon, mouse, rat, zebrafish, fly, and mosquitoes. Surprisingly, no method was particularly effective. Furthermore, we found that only very strong rhythmic signals were relevant with each method. More precisely, when we use a usual cut-off to define rhythmic genes, the group of genes considered as rhythmic contains many genes whose rhythmicity cannot be confirmed to be biologically relevant. We also show that rhythmic genes mainly contain highly expressed genes. Finally, based on our results, we provide recommendations on which methods to use and how, and suggestions for future experimental designs.

This is a *PLOS Computational Biology* Benchmarking paper.

Introduction

The nycthemeral transcriptome is characterized by the set of genes that display a rhythmic change in their mRNAs levels with a periodicity of 24 hours. These include, but are not limited to, circadian genes whose rhythm is endogenous and entrainable. In baboon, 82% of protein-coding genes have been reported to be rhythmic in at least one tissue [1]. The nycthemeral rhythmicity of these transcripts can be driven by the internal oscillator clock or by other circadian input such as food-intake, the light-dark cycle, sleep-wake behavior, or social activities. Moreover, the nycthemeral transcriptome is tissue-specific [2, 3]. Given the importance of biological rhythms in understanding biology and medicine, many algorithms have been proposed to detect such rhythms. Some were developed specifically for biological data, while others were adapted from other fields where periodicity is important, such as Lomb Scargle (LS). Most methods are based on non-parametric models that search for referenced patterns, classically sinusoid, called time-domain methods, while some are frequency-domain methods based on spectral analysis [4]. Some of them have been designed to detect more diverse waveforms, including asymmetric patterns, such as RAIN [5] or empirical_JTK (empJTK) [6]. For instance, RAIN outperformed the original JTK_CYCLE algorithm for simulated data consisting of sinusoidal and ramp waveforms [6]. Thus, methods differ in the conception of their algorithm and in how they take into account features of the dataset such as curve shapes, period, noise level, presence of missing data, phase shifts, sampling rates [7], asymmetry of the waveform, or the number of cycles (total period length of the experiment). Each method has in principle different strengths and weaknesses for some features of the dataset. In *Arabidopsis*, HAYSTACK identified 45% more cycling transcripts than COSOPT, mainly due to the inclusion of a 'spike' pattern in its model [8]. Deckard et al. [7] studied how four methods (LS, JTK_CYCLE, de Lichtenberg, and persistent homology) performed across a variety of organisms and periodic processes. Based on synthetic data, they investigated the algorithms' ability to distinguish periodic from non-periodic profiles, to recover period, phase and amplitude, and they evaluated their performance for different signal shapes, noise levels, and sampling

rates. They proposed a decision tree to recommend one of these four algorithms based on these features of datasets [7].

The performance of algorithms to identify such periodic signal has been assessed so far based on synthetic (i.e., simulated) data, or on benchmark sets of known cycling genes. Hughes et al. [9] recently published guidelines for the analysis of biological rhythms and proposed a web-based application (CircaInSilico) for generating synthetic genome biology data to benchmark statistical algorithms for studying biological rhythms. While such benchmarks are useful to explore the behavior of methods in a set of cases, the applicability of results to real data is limited. For example, simulations need to impose an a priori fluctuation pattern, typically cosine. The fluctuation of transcript abundance of core clock genes does seem to follow a cosine shape [10], but sometimes follows non-sinusoidal periodic patterns in mouse liver (e.g., Nr1d1 or Arntl) [11] (based on the data from [12]). The fluctuations of the nycthemeral transcriptome are entrained by a complex network involving external cues [13–16], as simplified in Fig 1a, which might yield non-sinusoidal periodic patterns among rhythmic genes even if

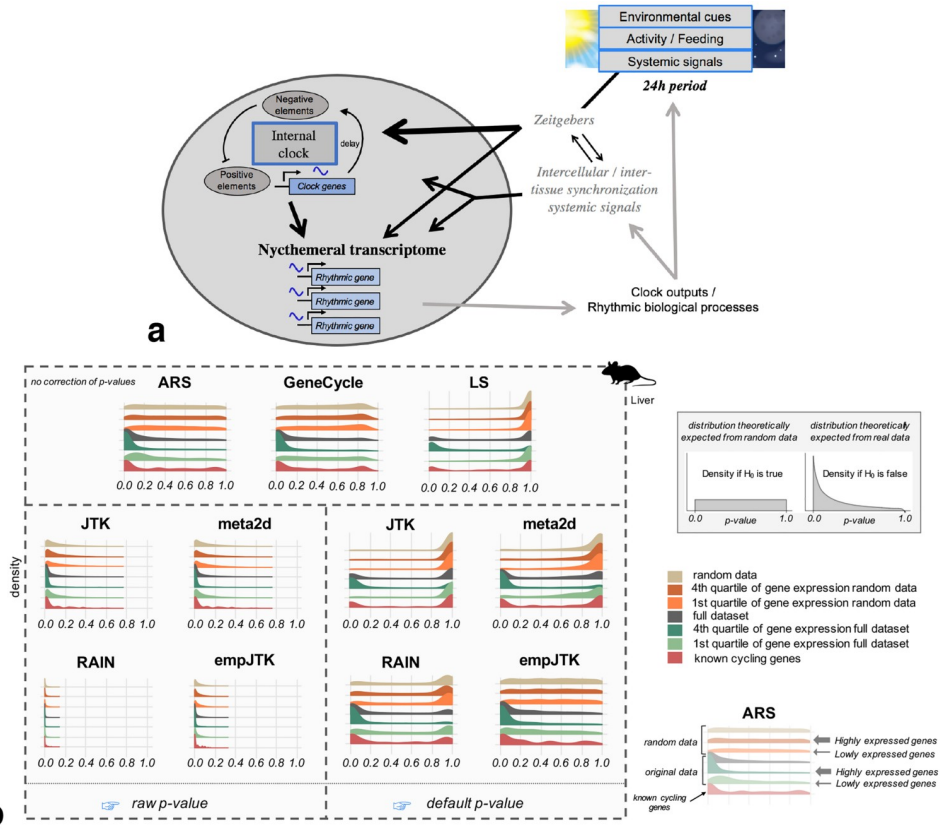


Fig 1. The nycthemeral transcriptome is the group of genes whose mRNAs have periodic variations with a 24h period, called rhythmic genes. To detect these rhythmic genes, we applied seven methods to time-series datasets that produced different density distribution of *p*-values. a) Simplified diagram of the entrainment of nycthemeral gene expression. Environmental cues include the light-dark cycle, food-intake, sleep-wake behavior, social activities, or any other 24h periodic event. **b)** Density distribution of *p*-values obtained before (raw) and after the default correction (software) for the seven methods applied to mouse liver data (microarray) sub-categorized in: i. randomized data which represents the null hypothesis; ii. randomized data restricted to the first and fourth quartiles of the median gene expression level, to check for the impact of expression level under the null; iii. the full original dataset; iv. the first and fourth quartiles of the median gene expression level of the original data; and v. a subset of known cycling genes (99 genes from KEGG "circadian entrainment" among which we expect a large proportion of rhythmic mRNA accumulation). The default *p*-values of ARS, GeneCycle, and LS are uncorrected. Mouse image credit to Anthony Caravaggi (license CC BY-NC-SA 3.0).

https://doi.org/10.1371/journal.pcbi.1007666.g001

circadian genes were sinusoidal. But the biological relevance of these waveforms is still not clear. This raises two issues: benchmarks based on simulations are biased towards methods that detect the same types of patterns as simulated; and when an algorithm detects more rhythmic genes, it could be more true positives or more false positives. When pattern constraints are released this increases the number of genes detected as rhythmic, but is not necessarily informative on the capacity of the algorithm to detect genes whose rhythmicity is biologically relevant.

Using real data and randomization tests, we compared seven methods: JTK_CYCLE (**JTK**) [17], **LS** [18, 19], ARSER (**ARS**) [4], and **meta2d** (Fisher integration of LS, ARS, and JTK), are frequently used by many studies and are all included within the MetaCycle R package [20]. We also included empirical_JTK (**empJTK**) [6] and **RAIN** [5], which have been recently developed to deal with more non cosine patterns and with asymmetric waveforms. empJTK and RAIN aim to improve the original JTK algorithm which assumed that any underlying rhythms have symmetric waveforms (more precisely, only the waveform coded into the JTK algorithm will be detected, which is the sine curve by default) [5]. Finally, robust.spectrum [21] extents a robust rank-based spectral estimator to the detection of periodic signals. It is integrated in the GeneCycle R package [22] and called **GeneCycle** in this paper. We excluded de Lichtenberg [23], Persistent Homology [24], COSOPT [25], Fisher's G test [26], MAPES, Capon, and other algorithms for reasons such as i. difficult accessibility of the software which limit their use by researchers, ii. their higher sensitivity to certain features of the data such as the sampling density, the number of replicates and/or periods, noise level, and waveform, iii. their weaker efficiency on simulated data or known cycling genes, or iv. their previously reported less good detection of non-sinusoidal periodic patterns [4, 6, 7, 27–29]. We first analysed the behavior of these seven methods applied to a variety of real datasets in animals, and within each dataset, we compared results between representative gene subsets such as highly and lowly expressed genes, known cycling genes, and randomized data. Contrary to real data, randomized data is not expected to show any signal of rhythmicity, which we used to test proper statistical behavior under the null hypothesis. Secondly, as function tends to be conserved between orthologs [30], true rhythmic genes are expected to be enriched in orthologs that are themselves rhythmic in other species. Indeed, evolutionary conservation provides a valuable filter through which to highlight functional biological networks, notably for clock-controlled functions [31]. The biological relevance of rhythmic genes is expected to be higher for rhythmic orthologs. An unknown proportion of the genes reported as rhythmic but not conserved will be true positives, whose rhythmicity evolved recently in one species or was lost in the other. This would only be a problem if a method would somehow favour these non-conserved ones while reporting true positives; we do not see any reason to expect such a behavior. On the other hand, errors in the prediction of rhythmicity by each method are not expected to be conserved between orthologs. Rather than benchmarking rhythm detection methods based on a profile, we used the biological relevance of genes detected rhythmic. Notably we considered that, among orthologs, those which conserved their rhythmic expression can formed a suitable reference group of rhythmic genes. Thus, the best methods are expected to report rhythmic genes with a high proportion of rhythmic orthologs. We used this approach to compare the algorithms based on their ability to capture biologically relevant evolutionary conservation signal within nycthemeral genes, and compared them to a Naive method.

Results

We used gene expression time-series datasets that come from circadian experiments and kept the data from healthy, wild-type individuals for seven species (S1 Table), allowing comparisons

among vertebrates and among insects. We benchmarked methods on animal data since organ homology allowed to compare datasets for which we expect conservation of functional patterns (tissue-specific rhythms). For readability, we present vertebrate results in the main figures and insect results in supplementary results (S6 File). Apart from the rat and Anopheles datasets, data with several biological replicates were obtained already normalized over replicates (one value per time-point).

We define a rhythmic gene as a gene which displays a nycthemeral change in its mRNA abundance, i.e. occurring over 24 hours and repeated every 24 hours. All these rhythmic genes represent the nycthemeral transcriptome. Different organs have been reported to have transcriptomes which are more or less rhythmic [2]. The rhythmic expression of these genes can be entrained directly by the internal clock or indirectly by external inputs, such as the light-dark cycle or food-intake [13–16] (Fig 1a). We consider the entirety of these rhythms to be a biologically relevant signal to detect. That is why we preferred data from light-dark and ad-libitum experimental conditions whenever possible (S1 Table), as providing a better representation of wild conditions.

Some methods are distinguished by their higher sensitivity to alternative patterns such as peak, box, or asymmetric profiles. A visual inspection of the KEGG "Circadian entrainment" gene set (see Methods) provides indeed informal confirmation that such patterns can be observed among known cycling genes, such as Npas2, Nr1d1, or Bhlhe41 (S1 File).

Analysis of statistical behaviors of methods applied to real data

***p*-values distribution analysis.** First, a good method should produce a uniform distribution of *p*-values when there is no structure in the data, in contrast to the distribution obtained from empirical data, which is expected to be skewed towards low *p*-values because of the presence of rhythmic genes. We investigated the properties of the different methods applied to randomized vs real data. We also investigated to what extent the density distribution of *p*-values of each method was affected by gene expression levels. Indeed, higher expression provides more power for detecting rhythmic patterns—highly expressed genes have more chance to shape rhythmic patterns because the variations of expression levels are relative to the general expression level—but this should not be the main driver of results. I.e., a method to detect rhythmicity should not be essentially reporting high expression levels. Even if true rhythmic genes were enriched in high expressed genes, we expect a good method to report both high and low *p*-values, at each expression level.

Fig 1b shows the density distribution of raw *p*-values obtained for the seven methods applied to mouse liver data (microarray) sub-categorized in: i. randomized data which represents the null hypothesis; ii. randomized data restricted to the first and fourth quartiles of the median gene expression level, to check for the impact of expression level under the null; iii. the full original dataset; iv. the first and fourth quartiles of the median gene expression level of the original data; and v. a subset of known cycling genes (8 to 99 genes according to species, see Methods). Results from the other datasets are provided in S2 and S3 Files. Surprisingly, only ARS and GeneCycle displayed close to the expected uniform raw *p*-value distribution for randomized data (Fig 1b). The adjustment by default of empJTK (minimum of the *p*-value calculated from an empirical null distribution, and of Bonferroni) recovered the expected uniform distribution, suggesting that this correction allows recovering proper *p*-values (Fig 1b). We used each software output "*p*-values" for calls, which we call "default *p*-value". In some software, these values result from an internal *p*-value adjustment, so we also analysed "raw *p*-values" (uncorrected, see Methods and Table 1 for JTK). For ARS, GeneCycle, and LS, the default *p*-values are uncorrected. Under the null hypothesis, LS has an abnormal peak near *p*-value = 1

Table 1. Raw, default, and BH.Q in JTK algorithm.

JTK	description	R
raw *p*-value	No correction	-
default *p*-value	Bonferroni correction of raw *p*-values	p.adjust(raw.pvals, method="bonf")
BH.Q (this paper)	Benjamini-Hochberg correction of raw *p*-values	p.adjust(raw.pvals, method="BH")
BH.Q (software)	Benjamini-Hochberg correction of default *p*-values	p.adjust(default.pvals, method="BH")

https://doi.org/10.1371/journal.pcbi.1007666.t001

(Fig 1b), implying an issue with its definition of the null hypothesis, or maybe a one-sided test when a two-sided test would be appropriate. The three other algorithms (RAIN, JTK, and meta2d) seem to have issues with false positives, displaying large proportions of low *p*-values even for randomized data. This issue was also recently reported by Hutchison and Dinner [32] who in addition showed that a combined method, such as meta2d which integrates results from ARS, JTK, and LS, under-perform the individual methods for low *p*-values [32].

Before analysing the impact of expression levels, we checked that the data follow a typical bimodal density distribution of gene expression (S1 File) and that using the median of time-points for gene expression gives similar results to using the minimum or the mean value (S1 File). Unsurprisingly, higher expression levels imply a higher power to detect rhythmic patterns (S1 File). The *p*-values distributions imply that most methods detect almost all highly expressed genes, and almost no lowly expressed genes, as "rhythmic" (Fig 1b). The normalization of gene expression values (Z-score transformation) did not change the *p*-values distributions within highly expressed genes, and particularly did not recover rhythmicity within lowly expressed genes (S1 File). This was not due to sampling biases of microarray data since results are consistent with RNAseq data (S1 File). Thus, the differences obtained between highly and lowly expressed genes either reflect true biology or a lower signal to noise ratio in lowly expressed genes. We think that (i) a method which is able to detect at least some lowly expressed genes as rhythmic is preferable, and (ii) a method should not detect almost all highly expressed genes as rhythmic. Overall, ARS, empJTK, and GeneCycle had the best behavior, producing a uniform distribution under the null hypothesis, and a skew towards low *p*-values for all empirical data.

Much more rhythmic signal is detected among genes with high amplitude (S1 File). This does not necessarily imply that the rhythmicity of the low amplitude genes isn't biologically relevant. From data of the same mouse experiment [2], we observed differences of *p*-value density distributions between microarray and RNAseq, with the skew towards low *p*-values less marked for RNAseq data (S1 File). This can be due to the more precise temporal resolution of the microarray time-series dataset, or to differences in the detection of gene expression by RNAseq vs microarrays (Fig 2a). When we restricted the microarray time-series to the same time-points as in the RNAseq series, we obtained a *p*-value distribution very similar to that of the RNAseq data (Fig 2b). The same time-series restriction applied to known cycling genes produced comparable results (S1 File). This supports a major role of the temporal resolution for method results, relative to a minor role for the difference between RNAseq and microarrays. That is why for the next steps, we only considered the microarray dataset for the mouse.

This observation can be generalized to diverse datasets. We see that each method loses in efficiency when the number of 24h cycles decreases, or when the number of time-points sampled decreases (Fig 3a). We show only results of this comparison for ARS, GeneCycle, and empJTK because they were the only methods with correct behavior in their *p*-value distributions (Fig 1b). For the same number of time-points, performance seems better with two cycles than only one cycle, as shown comparing zebrafish and baboon data which have both twelve time-points (Fig 3a). But this observation could be confused by the comparison of different species or different samples' quality. ARS performed better with a smaller total number of

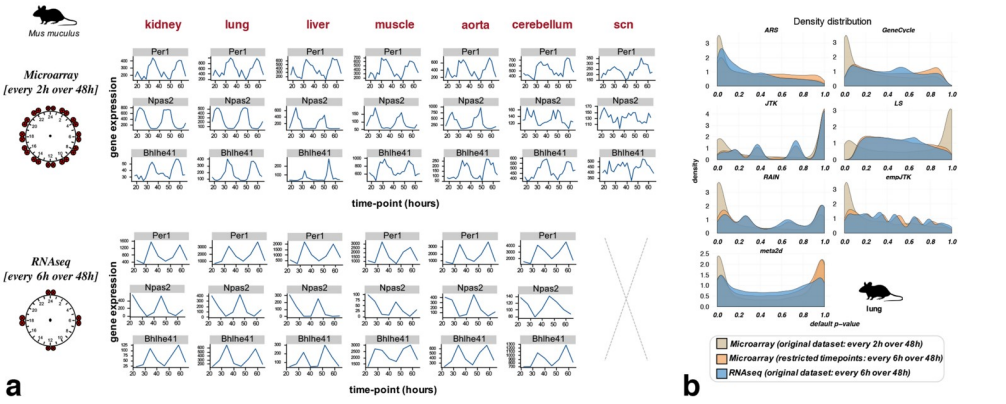


Fig 2. Fewer time points per cycle lead to a weaker detection of rhythmic patterns even if the transcriptome profiling quality is better. a) Bhlhe41, Npas2, and Per1 expression over time from data of the same mouse experiment [2] using two transcriptome profiling techniques: microarray vs RNAseq. The number of time-points with data is 24 for microarray and 8 for RNAseq. **b)** The restriction of microarray time-series to the same time-points as in the RNAseq series produces similar *p*-value distributions to those obtained with RNAseq. This supports a major role of the temporal resolution for method results, relative to a minor role for the difference between RNAseq and microarrays. Mouse image credit to Anthony Caravaggi (license CC BY-NC-SA 3.0).

https://doi.org/10.1371/journal.pcbi.1007666.g002

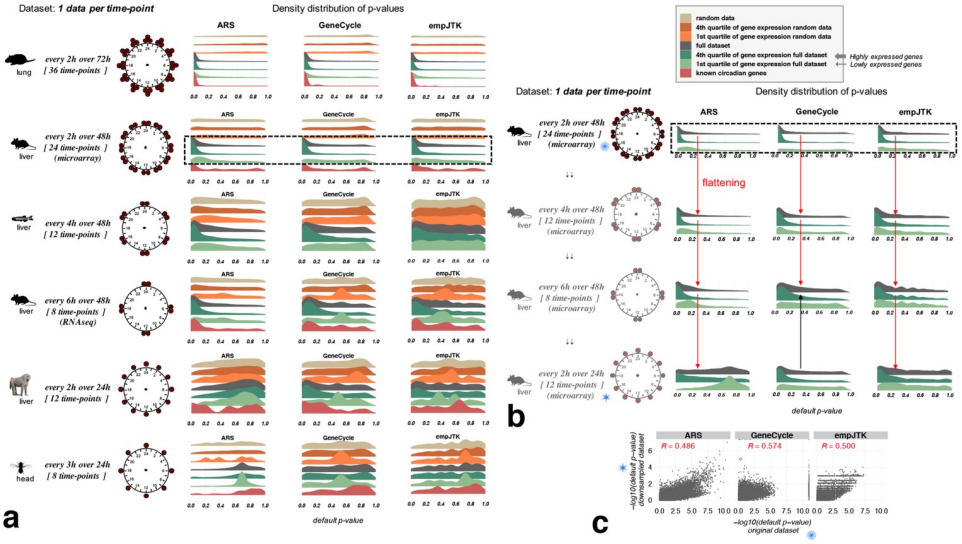


Fig 3. Datasets with one replicate per time-point over a unique cycle of 24 hours do not provide enough information to detect rhythmicity. Methods lose in statistical power for detecting rhythmic patterns in gene expression when the number of 24h cycles decreases, or when the number of time-points sampled decreases. **a)** Default *p*-value distributions obtained for ARS, GeneCycle, and empJTK applied to different datasets and sub-categorized in: i. randomized data which represents the null hypothesis; ii. randomized data restricted to the first and fourth quartiles of the median gene expression level, to check for the impact of expression level under the null; iii. the full original dataset; iv. the first and fourth quartiles of the median gene expression level of the original data; and v. a subset of known cycling genes (8 to 99 genes according to species, see Methods). For each dataset, the number of time-points with data and the temporal resolution is illustrated around a 24h clock. For the same number of time-points, performance seems better with two cycles than only one cycle (zebrafish vs baboon). **b)** The reduction of the number of time-points of the mouse liver microarray dataset shows increasingly weak rhythm detection by ARS, GeneCycle, and empJTK, shown by a flattening of the *p*-value distribution on the full dataset (red arrow). GeneCycle showed no difference between a few time-points over two cycles or more time-points over a single cycle (black arrow). **c)** Scatter-plots of *p*-values obtained before and after down-sampling (every 2h over 48h vs. every 2h over 24h) for the full dataset. Each point is a gene. *R* is the Pearson correlation; *p*-value $< 2.2e{-}16$ in all cases. After down-sampling, the rhythmic signal is retrieved for the same genes. Images credit: Anthony Caravaggi (mouse), Ian Quigley (zebrafish), wikipedia GNU GPL Muhammad Mahdi Karim (baboon), and Public Domain for other images (from PhyloPic).

https://doi.org/10.1371/journal.pcbi.1007666.g003

time-points but over two cycles than with more total time-points over a single cycle (mouse RNAseq vs baboon in Fig 3a), indicating that ARS is very dependant on the repetitive nature of profiles. The reduction of the number of time-points of the mouse microarray dataset shows similar effects on the rhythm detection by ARS, GeneCycle, and empJTK (Fig 3b). Of note, GeneCycle presented more or less no differences between having a few time-points over two cycles and having more time-points over a single cycle (black arrow Fig 3b).

Overlap between methods. Among genes called rhythmic, we analysed the number of those called in common by the different methods. For *p*-value thresholds of 0.05 or 0.01, we found a large proportion of genes called rhythmic by only one or few methods (Fig 4a and S1 File which shows the Jaccard index heatmap for mouse liver). Nevertheless, the overlap between all methods was the largest category for the mouse liver data (Fig 4a). Using a very low false positive tolerance with FDR thresholds of 0.5%, all methods except LS overlap largely (S1 File). If we ignore *p*-value thresholds and consider the first 6000 genes detected rhythmic for each method, the overlap becomes stronger (Fig 4d). We obtained similar results from the most informative dataset (S1 File). Indeed, the rat lung dataset has 36 time-points spread over

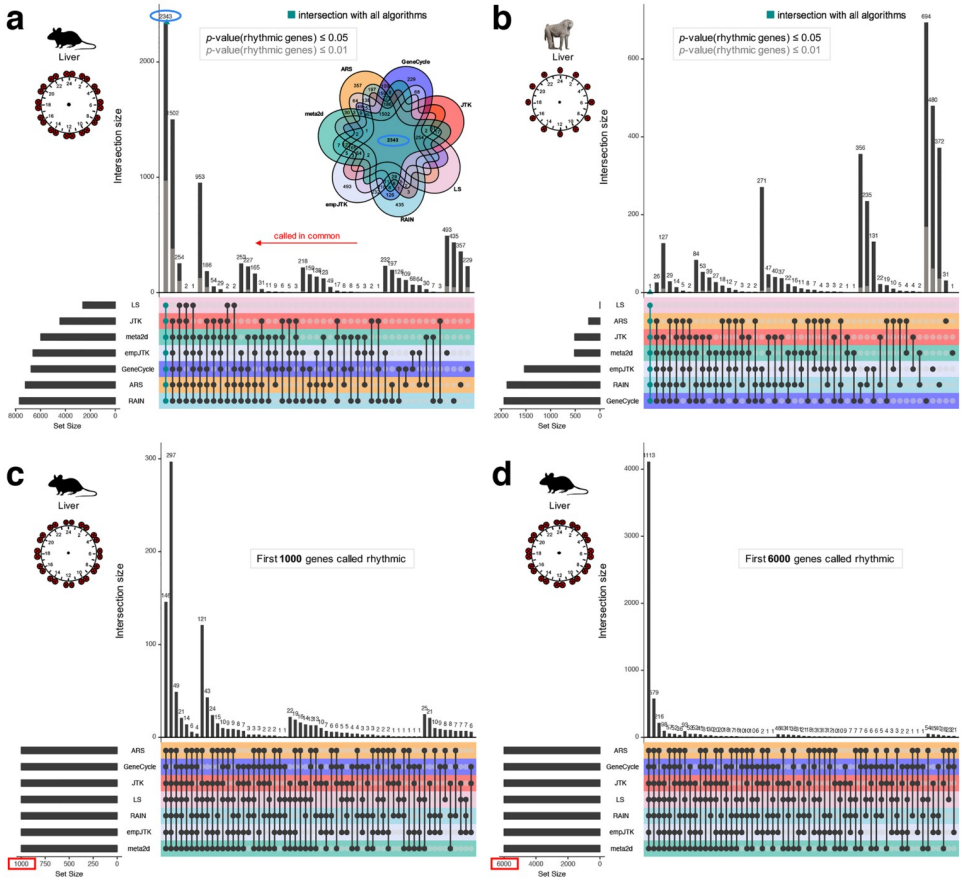


Fig 4. Methods detect the same first top rhythmic genes, but with inconsistencies in the meaning of their *p*-values. Upset diagrams show the number of rhythmic genes called in common by the methods. Each intersection is exclusive, i.e. one gene can appear in only one intersection. **(a,b)** Upset diagram for mouse liver dataset (microarray) (**a**) and baboon liver dataset (**b**) for the *p*-value thresholds of 0.05 (black) or 0.01 (grey) for calling genes rhythmic. The Venn diagram (**a**) illustrates the upset diagram with, for instance, 2343 genes called rhythmic by all methods. **(c,d)** Upset diagram for mouse liver dataset (microarray) for the first 1000 (**c**) or 6000 (**d**) genes detected rhythmic for each method. With a smaller number of top rhythmic genes, the overlap between methods is weaker. Images credit: Anthony Caravaggi (mouse) and wikipedia GNU GPL Muhammad Mahdi Karim (baboon).

https://doi.org/10.1371/journal.pcbi.1007666.g004

three 24h cycles (Fig 3a). Thus, the same genes seem to be called rhythmic by all methods but the threshold of significance appears inconsistent. Some methods are expected to produce different *p*-values because their underlying assumptions are different, i.e. other than sinusoidal for RAIN and empJTK. But the bulk of the methods are designed to find sinusoidal patterns and thus should ideally produce similar *p*-values, or at least similar ordering of results. Thus, our observations suggest an issue with the significance of *p*-value thresholds for the methods. While in principle effect size is often more relevant than *p*-value, these methods are all used in practice to produce *p*-values, define a threshold, and provide a list of "rhythmic genes", thus consistency of these *p*-values is important. With a smaller number of top rhythmic genes, the overlap between methods was weaker (Fig 4c and 4d). Thus the methods agree on a large number of rhythmic genes, but not necessarily on the order of significance among them. Finally, for baboon liver data there was less overlap of methods (Fig 4b; S1 File), which might be due to the low information in that data (Fig 3a).

Use of evolutionary conservation as a benchmark

Signal of evolutionary conservation. We expect biologically relevant rhythmic activity of genes to be more conserved between species than putative false positives from detection methods. For each condition (species and tissue), we defined the group of genes whose orthologs are called rhythmic in the homologous tissue of another species (Fig 5a). For example, starting with all mouse genes, we only kept mouse-zebrafish one-to-one orthologs. Considering the liver, these orthologs were separated into two groups: genes for which the ortholog is detected as rhythmic in zebrafish liver, called rhythmic orthologs; and the remaining one-to-one orthologs (Fig 5b). Mouse-zebrafish orthologs, that are detected rhythmic in zebrafish liver, were significantly more enriched in small *p*-values in mouse liver, for all methods (Kolmogorov-Smirnov test *p*-values < 0.001 with Kolmogorov's D statistic around 10-15% of maximum deviation, Fig 5d). Similar results were obtained using different methods and/or a different threshold to call orthologs as rhythmic in zebrafish liver (S1 File). This result obtained for distant species (S1 File) shows that the conservation of rhythmicity at the transcriptomic level is informative. Similar results were obtained in other species comparisons (S1 File), with a stronger signal for evolutionarily close species such as mouse and rat (with Kolmogorov's D statistic around 10-15% of maximum deviation, S1 File), although we found no consistent correlations of the orthologs *p*-values between the rat and the mouse (S1 File). Of note, the comparison of species under different conditions (light-dark versus dark-dark) is a limitation in itself since the overlap of the rhythmic transcriptome between these two conditions has been shown to be low [33–35] (although this interpretation remains limited by the thresholds used). However, we found a good correlation of *p*-values obtained between these two conditions in the head of *Anopheles gambiae* (R=0.605, S1 File) suggesting that this limitation does not hide most of the conserved signal. Thus, for the same homologous organ, rhythmic orthologs have a stronger statistical signal of rhythmicity than non-rhythmic orthologs. We are going to use this evolutionary conservation of the rhythmicity of gene expression in order to compare the performance of methods. We expect that a method which detects more genes with biologically relevant rhythmicity should also detect more conservation of rhythmicity. This is both justified in principle, because evolutionary conservation implies relevance to the functioning of the organism, and in practice, since orthologs of rhythmic genes have smaller *p*-values (Fig 5d).

Only strong rhythmic signals of gene expression are relevant. In this last part, we compared the performances of methods to detect the rhythmic orthologs. For a given dataset, the best method is expected to report rhythmic genes with the highest proportion of rhythmic orthologs. It should be noted that this does not imply that we expect all rhythmic behavior to

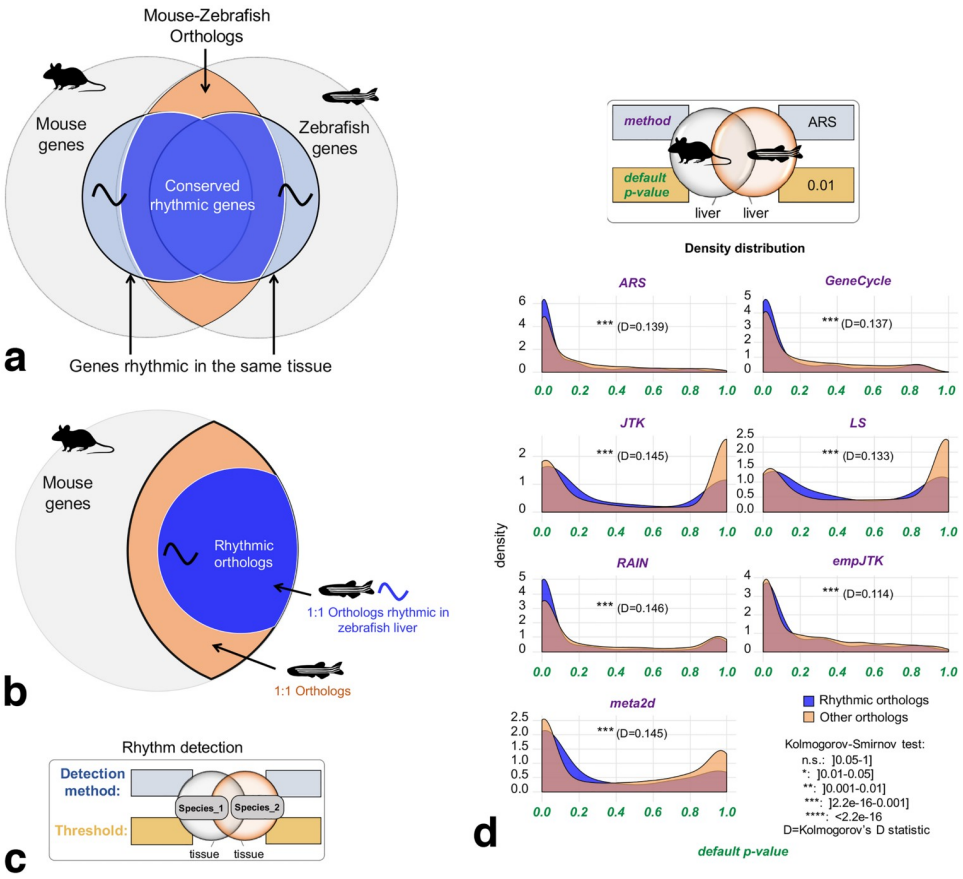


Fig 5. Signal of evolutionary conservation of rhythmic gene expression. Orthologous genes detected as rhythmic in the same organ of two species have a stronger statistical signal of rhythmicity than those not detected as rhythmic in at least one species. **a)** Mouse and zebrafish share orthologous genes, some of which are rhythmic in the homologous tissues. **b)** Method used for ortholog benchmarking, as in panel **d**: From all mouse genes, only mouse-zebrafish one-to-one orthologs are kept. Considering the liver, these orthologs are separated into two groups: genes for which the ortholog is detected as rhythmic in zebrafish liver, called rhythmic orthologs; and the remaining one-to-one orthologs. **c)** Chart providing the legends to inform about the method and the threshold used to call genes rhythmic for each condition (species and tissue). **d)** *p*-values density distribution of rhythmic orthologs vs non-rhythmic orthologs obtained for the seven methods applied to mouse liver data. Mouse-zebrafish orthologs, that are detected rhythmic in zebrafish liver, are significantly more enriched in small *p*-values in mouse liver, for all methods (Kolmogorov-Smirnov test *p*-values < 0.001). Images credit: Anthony Caravaggi (mouse), Ian Quigley (zebrafish).

https://doi.org/10.1371/journal.pcbi.1007666.g005

be conserved between orthologs, but rather that true rhythmic genes should have more rhythmic orthologs than false-positive predictions. For a given *p*-value threshold, each method detects a certain number of rhythmic genes (genes with *p*-value under the threshold). At each threshold we calculated the proportion of orthologs rhythmic in species2 among one-to-one species1-species2 orthologs, as defined in Fig 6a. This proportion allows to assess how each method is able to detect the conservation of rhythmicity and can be calculated for each *p*-value threshold. The benchmark set is composed of orthologs detected rhythmic in the second species, called rhythmic orthologs. To define this set, we chose a rhythmicity detection method among ARS, empJTK, and GeneCycle, in agreement with results of previous sections, and a *p*-value threshold of 0.01.

A risk is that orthologs have conservation of gene expression levels and that there is a bias towards calling highly expressed genes "rhythmic". To control for this in the benchmarking, we added a "Naive" method based only on expression levels. This Naive method simply orders

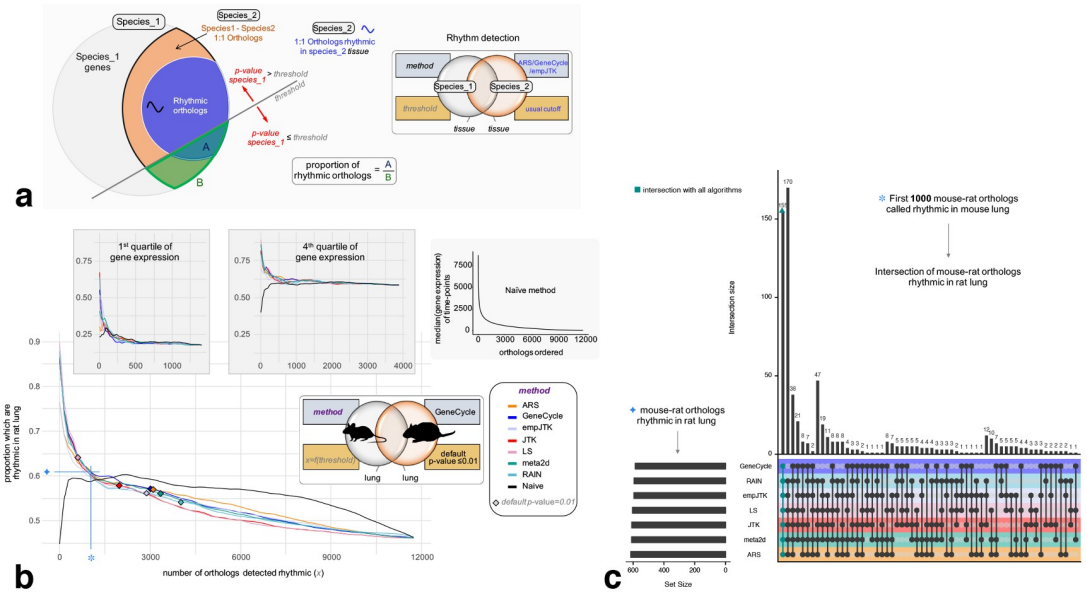


Fig 6. Only strong rhythmic signals of gene expression are relevant. Methods designed for rhythm detection in gene expression show an advantage only for the genes with a strong rhythmic signal, i.e. related to very small *p*-values. For a fixed number of top genes called rhythmic, all the methods, despite their design differences, retrieve approximately the same proportion of biologically functional rhythmic genes and the same genes themselves. **a)** Method to obtain figure b: For a given *p*-value threshold, each method detects a certain number of rhythmic genes (genes with *p*-value ≤ threshold). At each threshold, we calculate the proportion of orthologs rhythmic in species2 (A) among one-to-one species1-species2 orthologs (B). The benchmark set is composed of one-to-one orthologs detected rhythmic in the second species (using method ARS, GeneCycle, or empJTK), called rhythmic orthologs. **b)** Variation of the proportion rhythmic orthologs/all orthologs in mouse as a function of the number of mouse orthologs detected rhythmic, for each method applied to the mouse lung dataset. The benchmark gene set is composed of mouse-rat orthologs, detected rhythmic in rat lung by the GeneCycle method with default *p*-value ≤ 0.01. The black line is the Naive method which orders genes according to their median expression levels (median of time-points), from highest expressed to lowest expressed gene, then, for each gene, calculates the proportion of rhythmic orthologs among those with higher expression. The proportion of the benchmark set among one-to-one orthologs is higher for highly expressed genes (4th quartile) than for lowly expressed genes (1st quartile) (∼60% vs ∼20% respectively). Diamonds correspond to a *p*-value threshold of 0.01. **c)** Upset diagram showing the number of rhythmic orthologs (figure **a**) called in common by the methods among the first 1000 mouse-rat orthologs that are called rhythmic in mouse lung. Images credit: Anthony Caravaggi (mouse) and Public Domain for other images (from PhyloPic).

https://doi.org/10.1371/journal.pcbi.1007666.g006

genes (orthologs here) according to their median expression levels (median of time-points), from highest expressed to lowest expressed gene, then, for each gene, we calculated the proportion of rhythmic orthologs among those with higher expression. We also present results for subsets obtained from the division in four quartiles of expression levels. Fig 6b shows the variation of the proportion defined above as a function of the number of orthologs detected rhythmic, obtained for each method applied to the mouse lung dataset. The benchmark gene set was defined by mouse-rat orthologs, detected rhythmic in rat lung by the GeneCycle method (default *p*-value ≤ 0.01). Genes are given by order of their detection by the methods. The genes with small *p*-values, i.e. with a strong signal of rhythmicity, had a high proportion of rhythmic orthologs. Importantly, for all methods, this proportion was higher than that obtained from the Naive method (Fig 6b). Results are consistent in almost all species comparisons, with exceptions for cerebral tissues (S4 File). However, the thresholds of 0.01 are to the right of the intersection between the curves of rhythm detection methods and the Naive, except for LS. This means that, for an apparently reasonable threshold (*p*-value ≤ 0.01), ranking genes by expression level performed "better" than all methods designed specially for rhythm detection. We made the same observation using an FDR-based threshold (FDR≤ 0.01 or FDR≤ 0.1, S1 File). These results

imply that even with a stringent *p*-value or FDR threshold, such as 0.01, the rhythmic nature of some of the genes considered rhythmic is not relevant. These rhythm detection methods were relevant only for genes with very high signal of rhythmicity, where they performed better than a Naive method. Finally, for the top 1000 mouse-rat orthologs detected as rhythmic in mouse lung, all the methods reported a similar proportion of rhythmic orthologs, around 62%, mainly highly expressed genes (fourth quartile of gene expression) (Fig 6b). And the overlap between these orthologs was largely detected by all methods (Fig 6c). Thus, for genes with a high signal of rhythmicity, all methods performed similarly to detect the tissue-specific conservation of gene expression rhythmicity. Similar results were obtained for other species comparisons (S5 File).

Discussion

The methods designed for rhythm detection in gene expression perform similarly and only for strong rhythmic signal. In this study, we show that orthologous genes detected as rhythmic in the same organ of two species have a stronger statistical signal of rhythmicity than those detected as not-rhythmic in at least one species. These results support our hypothesis that the nycthemeral rhythmicity at the gene expression level is biologically functional, and that this functionality is more conserved between orthologous genes than between random genes. We define the nycthemeral transcriptome as all genes displaying a rhythmic expression repeated every 24 hours. In order to assess the performance of seven methods to detect these rhythms, we used this concept of conservation of the rhythmicity between species for benchmarking. We employed genes whose orthologs had a rhythmic expression called in the same homologous organ as a proxy for a true positive set, as done in some previous benchmarks. For instance, Rosikiewicz et al. [36] assessed the quality of microarrays quality control methods based on evolutionary conservation of expression profiles, and Kryuchkova et al. [37] benchmarked tissue-specificity methods in the same way. This approach based on real data, also used by Boyle et al. [3] to solve the issue of weak overlap between the same tissues from the same species from different experiments, avoids relying on simulations which tend to favor methods using the same model, e.g. the same patterns, and has the advantage of not being based on specific assumptions, other than general evolutionary conservation of function. By taking into account the tissue-specificity of rhythmic gene expression and different species comparisons, we show that no method is strongly favoured. For instance, one would have expected that the added features of RAIN and empJTK allowing then to detect more diverse patterns than a classical sinusoidal would have favored them. But this flexibility did not provide them any advantage in the benchmark. Furthermore, the comparison of the methods with a 'Naive' one, uninformed about rhythmicity, shows an advantage for informed methods only for the genes with a strong rhythmic signal. Thus, only genes with a strong rhythmic signal, i.e. the top genes called rhythmic, can be considered as relevant. Even if the threshold of "relevance" of these genes is dependant on the evolutionary distance of the species compared, these results suggest a call for caution about the results of previous studies reporting or based on large sets of rhythmic genes. For the same number of genes called rhythmic, all the methods, despite their design differences, retrieved approximately the same proportion of biologically functional rhythmic genes (Fig 6b) and the same genes themselves (Fig 6c).

The issue of significance

For the same *p*-value threshold, the number of genes called rhythmic is different from one method to another, with a large proportion of these genes detected rhythmic by only one or a few methods. But, if we consider the top genes called rhythmic for each method, without

taking into account any *p*-value threshold, the overlap of rhythmic genes become strong between the methods (Fig 4c and 4d). This highlights an issue with the meaning of the *p*-value and the associated thresholds used. This is directly related to the issue of correction that needs to be improved in this field. When a smaller number of top rhythmic genes is used, the overlap between methods becomes weaker (Fig 4c and 4d). Thus, the order of calling genes rhythmic is different from one method to another. Finally, since methods performed better than a Naive method only for genes with a strong rhythmic signal, we can not conclude for the relevance of the other genes called rhythmic, even when they have very low nominal *p*-values.

ARS, empJTK, and GeneCycle produce consistent *p*-values

ARS, empJTK, and GeneCycle were the methods that showed the best behavior on real and randomized data (single species tests). They were the only methods displaying both a uniform distribution of their *p*-values under the null hypothesis, and a left-skewed distribution when applied to real data. For empJTK, its default correction allowed to produce these expected results. However, each of these three methods is conceptually completely different, which indicates that there is not one conceptual framework which dominates rhythmic gene detection. ARS combines time-domain and frequency-domain analyses. GeneCycle, which is the robust spectrum function of the R package, is based on a robust spectral estimator which is incorporated into the hypothesis testing framework using a so-called g-statistic together with correction for multiple testing. And, empJTK improves the original JTK including additional reference waveforms in its rhythm detection. The other methods all presented major issues. LS has a right-skewed distribution of its initial uncorrected *p*-values suggesting an invalid null hypothesis. JTK, RAIN, and meta2d had also issues with their null hypothesis displaying left-skewed distributions of their uncorrected *p*-values. Their default adjustment was excessive, favoring high *p*-values obtained after correction. Hutchison and Dinner [32] observed this on simulated data, and proposed that it was due to non independence of measurements from the same time series.

Biological insight into gene rhythmicity in animal tissues

Our results support the hypothesis that rhythmic genes are largely enriched in highly expressed genes (Table 2). Experimental noise that would mask the rhythmic signal of lowly expressed genes could also explain this result in part, especially considering that the datasets with good sampling used microarray technology. BooteJTK compares the noise to the amplitude of a time series, in addition to evaluating the rank order of the values, and thus might provide a more relevant rhythm detection by improving the variance estimation from biological replicates [38]. The observation of known cycling genes in different organs seems to indicate different profiles of rhythmicity possible for the same gene. For instance, Npas2 displays a cosine shape in kidney and lung, and a peak/box shape in liver and muscle (Fig 2a). This observation suggests that methods might perform differently depending on the organ studied. This is also one of the reasons why all our analyses were made for homologous organs.

In mouse-baboon comparisons, there were no significant differences of *p*-value density between rhythmic and non-rhythmic orthologs in cerebral tissues: brain stem, cerebellum, and supra-chiasmatic nucleus, except for the hypothalamus (S4 File). This could be explained by the fact that there are only low amplitudes of expression of clock genes and few rhythmic genes in almost all brain regions. This is assumed to be due to an inefficient synchronization of individual cellular oscillators in brain cells to avoid noise into the synchronizator element [39]. In addition, it could also be an essential aspect for intrinsic brain processes which could require a constant expression of most genes.

Table 2. t-test comparing the expression levels between rhythmic (p-value $\leq$ 0.005) and non-rhythmic genes (randomly chosen same number of genes among those with p-value $>$ 0.01), in mouse liver dataset (microarray).

Method	group	*n*	Mean
ARS	rhythmic	4019	1151.2
	non-rhythmic	4019	383.6
t-test t = 25.4 p <2.2e–16 df = 7021.4			
meta2d	rhythmic	4520	1113.2
	non-rhythmic	4520	398.5
t-test t = 24.8 p <2.2e–16 df = 8050.1			
empJTK	rhythmic	3373	1113.9
	non-rhythmic	3373	442.5
t-test t = 20.1 p <2.2e–16 df = 6260.8			
RAIN	rhythmic	5044	1066.3
	non-rhythmic	5044	384.2
t-test t = 25.4 p <2.2e–16 df = 8935.3			
JTK	rhythmic	2646	1214.4
	non-rhythmic	2646	454.3
t-test t = 19.6 p <2.2e–16 df = 4742.9			
LS	rhythmic	736	1500.3
	non-rhythmic	736	526.5
t-test t = 12.6 p <2.2e–16 df = 1351.2			
GeneCycle	rhythmic	4145	1082.8
	non-rhythmic	4145	425.6
t-test t = 22.0 p <2.2e–16 df = 7622.8			

https://doi.org/10.1371/journal.pcbi.1007666.t002

The importance of having an informative dataset

Because of the cost and complexity of circadian experiments, time-series datasets of gene expression in animals are rare, especially in the same experimental conditions. Algorithms must be able to deal with little data, but importantly experiments should take into account the algorithms' sensitivity. All algorithms appeared to produce relatively poor p-values distributions when applied to the available Drosophila or baboon datasets, and, for the baboon dataset, were almost always less efficient than the Naive method (S1 File). This baboon dataset is probably not very informative, which raises questions about the biological conclusions from the associated study [1]. With only one replicate per time-point, over only one cycle of 24 hours, the algorithms are unable to detect repetitive patterns. Variations over a single 24 hours cycle appear to be insufficient to detect rhythmic signal, when there is no evidence of repetition over several cycles. Moreover, each data comes from different outbred individuals. The variations of gene expression between two time-points can be due to individual variations or real oscillation within the population. It is possible that sinusoidal patterns with a continuous trend over successive time-points could be detected without replicates, although power will be lacking, but patterns such as the peak pattern will be extremely sensitive to inter-individual variation. Fig 3 generally suggests that datasets with one replicate per time-point over a unique cycle of 24 hours do not provide enough information that would allow to correctly detect the rhythmicity. It seems that ARS in peculiar is very sensitive to the repetitive nature of profiles. Of note, for time-series with low sampling frequency, a recent improvement of empJTK, called BooteJTK, allows to detect rhythms robustly relative to sampling frequency [38]. Thus, if only one 24h cycle is feasible, several biological replicates must be favored. Our results support the conclusions of Hutchison et al. [6] who indicate that for a fixed number of samples, better sensitivity and specificity are achieved with higher numbers of replicates than with higher

sampling density. We propose that future experiments should produce data with two biological replicates per time-points as a strict minimum. Obviously, we suggest considering biological replicates as new cycles within one replicate, as proposed in recent guidelines [9]. GeneCycle, and to a lesser extent empJTK, were the most robust methods when applied to weakly informative datasets. Thus, the performance of the algorithms is dependent on techniques and experimental designs used for the experiments. The optimization of experimental plans (see section Recommendations) could improve the methods' performance for the detection of rhythmically expressed genes. Moreover, we recommend producing data over at least two cycles to be sure of the repetitive nature of profiles, and to avoid a potential random influence of the shared environment, which might be considered rhythmic since it affects all replicates. Finally, contrary to the mouse experiment, the rat experiment has been done under zeitgeber conditions which have most likely resulted in more genes being expressed rhythmically, so in proportion, more periodic patterns. This might explain the higher density of small *p*-values obtained for the rat dataset (Fig 3a). Comparison between these two datasets is not expected to have removed the signal, since we found a good correlation of *p*-values obtained between two conditions, light-dark versus dark-dark, in data produced from the same experiment (S1 File).

Limitations and improvement of methods

ARS and GeneCycle need complete chronological data and cannot deal with biological replicates. Except for LS, RAIN, and empJTK, all other methods studied here assume equally spaced time-points. Furthermore, ARS needs an integer sampling interval with regular time-series datasets and cannot deal with missing values, or with several replicates per time-point. In this study, ARS appeared to be efficient only for the dataset with at least two cycles of data. Indeed it produced aberrant *p*-value distributions when applied to datasets restricted to one cycle of 24 hours. But, for these datasets, all algorithms behaved poorly. The improvement of JTK by empJTK produced much better results than the original JTK algorithm. It is possible that the improvement of RAIN suggested by Hutchison and Dinner [32], which allows to produce uniform *p*-values distribution under the null, might similarly improve the results of RAIN. We believe that LS could be a very interesting method if its null hypothesis could be clarified and would thus provide *p*-values with proper behavior. LS has advantages that other algorithms don't. For instance, it can deal with irregular intervals, missing data, and has been shown to stay efficient on small sample size [27], which constitutes one of the big issues of circadian transcriptomic data. On the other hand, relative to JTK, ARS, or MICOP methods, LS has also been shown to be highly sensitive to the increase of sampling intervals and to noise for proteomic data [40].

A good method must, at least, display a uniform distribution under the null hypothesis, and a classic skewed distribution when applied to full dataset or even more to known cycling genes. It should also be able to detect efficiently rhythmic orthologs, which represent an important part of the functionally relevant nycthemeral rhythmicity. In this study, we did not assess the amplitudes, phases, and precise period provided by the algorithms. We only analysed the performance of methods for nycthemeral or circadian rhythms in gene expression data, and cannot conclude directly for ultradian or seasonal rhythms, and for other types of datasets which are not gene expression data.

Recommendations

Experimental design. 1. Always use at least 2 biological replicates per time-point.

2. One full period sampled is the minimum required. Two periods are to be preferred.

3. Favor time-points number (small temporal resolution) over transcriptome profiling quality (e.g., microarray vs RNAseq).
4. Favor regular sampling because only few algorithms can deal with irregular interval time-series.
5. For a fixed number of samples, favor higher numbers of replicates over higher sampling density (see also [6]).

Recommendations about the choice of rhythm detection method, the arrangement of the time-series dataset, and the interpretation of results based on these seven methods studied.

1. Only genes with a strong rhythmic signal should be considered as relevant. By "strong" we mean the top genes called rhythmic, knowing that the threshold of p-value ≤ 0.01 is already not stringent enough for some methods.
2. Take into account that detected rhythmic genes are strongly enriched in highly expressed genes.
3. LS could be a good candidate to improve.
4. Favor ARS, GeneCycle, or empJTK with default parameters.
5. Consider biological replicates as new cycles with one replicate.
6. Check by eye for rhythms of known circadian genes.
7. Never duplicate and concatenate data before running algorithms [9].
8. Never consider technical replicates as biological replicates [9].

Methods

Pre-processing

For each time-series dataset, only protein coding genes were kept. For microarrays, we removed probIDs which were assigned to several GeneIDs. ProbIDs or genes which contained one or several missing values have been removed, allowing comparison between all methods even those which can not deal with missing values. Genes with no expression (= 0) at all time-points were also removed. For each species dataset, we only kept comparable conditions to other species of reference. Tissues separated in sub-tissues such as adrenal gland in adrenal cortex and adrenal medulla in baboon experiment were removed.

For each condition (species and tissue), several datasets have been built: i. the full original dataset; ii. the first and fourth quartiles of the median gene expression level of the original data; iii. randomized data (time-points redistributed randomly); iv. randomized data restricted to the first and fourth quartiles of the median gene expression level; and v. a subset of known cycling genes when such data was available (8 to 99 genes according to species).

Normalization by Z-score

The normalization of gene expression values by Z-score transforms the pre-processed data such that for gene i with the original expression value at time-point j is $gene.ij$, we have:

$$gene.ij.normalized = gene.ij - xi$$

with $xi = mi - \frac{Zi}{j}$. mi is the mean expression of gene i: $mi = \frac{\sum gene.ij}{j}$; and $Zi = \frac{mi-m}{sd}$; m and sd being the mean and the standard deviation of the original full dataset.

Orthology relationships

For each species comparison, orthologs relationships have been downloaded from OMA [41]. For simplicity, we only considered one-to-one orthologs. In species comparisons, we only kept orthologous genes that had available data in both species.

Algorithms and packages

`MetaCycle` R package was performed with parameters: minper = 20h and maxper = 28h. This package incorporates the 3 algorithms to detect rhythmic signals from time-series datasets: ARSER (ARS), JTK_CYCLE (JTK), and Lomb-Scargle (LS). It also provides meta2d that integrates analysis results from multiple methods based on an implementation strategy (see "Introduction to implementation steps of MetaCycle" in MetaCycle documentation for more details). ARS does not deal with several replicates per time-point. To not introduce biases, we only kept one replicate for ARS performing when the dataset was provided with several replicates per time-point.

`Rain` R package was performed with parameters: period = 24h, period.delta = 4h (width of period interval), and method = 'independent'. In order to obtain unadjusted pvalues as output, we modified the source code of the rain and MetaCycle R packages.

`Empirical-JTK` (empJTK) was executed by running bash commands with parameters: cosine waveform, 24 hours' period, look for phases every 2 hour from 0 to 22 hours and look for asymmetries every 2 hour from 2 to 22 hours (GitHub alanlhutchison/empirical-JTK_CYCLE-with-asymmetry). It is important to run empJTK with python version 2.7.11. Raw *p*-value correspond to `P` output (P-value corresponding to Tau, uncorrected for multiple hypothesis testing), and default *p*-value correspond to `empP` output (min(*p*-value calculated from empirical null distribution, Bonferroni)).

`GeneCycle` R package [22] was downloaded from CRAN. We used the robust.spectrum function developped by [21] that computes a robust rank-based estimate of the periodogram/correlogram.

Plots have been created using `ggplot2` R package (version 3.1.0); Upset diagrams using `UpSetR` R package (version 1.3.3) [42]; and Venn diagram using `venn` R package (version 1.7).

Statistical analysis of rhythmic gene expression

All the rhythm detection methods (See Materials) were applied to each pre-processed dataset, producing a list of *p*-values as output. Then, for each gene having several results (ProbIDs or transcripts), we combined *p*-values by Brown's method using the EmpiricalBrownsMethod R package. Thus, for each dataset, we obtained a unique *p*-value per gene. Whenever the per-gene normalization was not necessary (unique data for all genes), we obtained the original *p*-value for each gene. FDR is the false discovery rate adjustment of default *p*-values using `p.adjust` R function.

Naive method

The Naive method is only based on expression levels of genes and is not informed about rhythm detection. It simply orders genes according to their median expression levels (median of time-points), from highest expressed to lowest expressed gene. Then, for each gene *i*, we

calculate the proportion of rhythmic orthologs among those with higher expression, i.e. among the genes from the highest expressed one to the gene *i*.

Availability of data and scripts. The data and scripts for reproducing plots and analysis are available at https://github.com/laloumdav/rhythm_detection_benchmark.

Materials

Ethics statement

We had ethical issues to use olive baboon data since these data needed the sacrifice of twelve baboons. We would like to remind that such data would have been impossible to get in Switzerland where the primate research is prohibited. We still support Switzerland ethical considerations in matter of animal research and think that the scientific knowledge can not justify an irresponsible employment of life on earth. While being aware that our results would have been less robust without these data and that these considerations on primate could also be generalized to other living organisms.

Datasets

***Mus musculus* (13 tissues).** Raw microarray and RNA-seq data, from [2], was downloaded from GEO accession (GSE54652). Microarray gc-rma normalized data was sent by Katharina Hayer from CircaDB database [43]. Expression values were already normalized between biological replicates to average out both biological variance between individual animals and technical variance between individual dissections. RNA-seq data was already normalized using DESeq2. Data was obtained for adrenal gland, aorta, brain stem, brown adipose, cerebellum, heart, hypothalamus, kidney, liver, lung, muscle, SCN (only microarray), and white adipose. Probesets on the Affymetrix MoGene-1.0-ST-V1 array were cross-referenced to best-matching gene symbols by using Ensembl BioMart software.

***Papio anubis* [olive baboon] (11 tissues used).** RNA-seq data from [1] was downloaded already normalized by using DESeq2. Read counts per gene were calculated using FeatureCounts. We kept data for aorta, brain stem, cerebellum, heart, hypothalamus, kidney, liver, lung, muscle, SCN, and white adipose tissues. Data were already provided with Ensembl gene symbols.

***Rattus norvegicus* (lung).** Raw microarray data from [44] was downloaded from GEO accession (GSE25612). Over 3 days, 54 samples were extracted in light-dark condition with a temporal resolution closer for some time-points (See paper for more details). Contrary to the study, we still considered the 3 successive days samples as successive days measurements. ARS, JTK and RAIN methods don't operate with irregular time-series. We normalized time-series by calculating the mean value of irregular time-points to obtain regular time-series. rma normalization was performed using the rma R-package. Probesets on the Affymetrix 230-2-probe array were cross-referenced to best-matching gene symbols by using Ensembl BioMart software.

***Dano rerio* (liver).** Raw microarray data from [3] was downloaded from GEO accession (GSE87659). Data was already rma-normalized, averaged gene-level signal intensity, and already cross-referenced to best-matching transcript symbols.

***Anopheles gambiae* (head and body).** Raw microarray data from [33] was downloaded from GEO accession (GSE22585). Non-blood fed female mosquito heads and bodies were collected under light dark and constant dark conditions. We only used data collected in LD condition, except for the comparison of both conditions (LD versus DD). We normalized data using the rma R package and cross-referenced to best-matching gene symbols by using VectorBase software.

***Aedes aegypti* (head).** Raw microarray data from Ľeming was downloaded from GEO accession (GSE60496). Non-blood fed female mosquito heads were collected under light dark and constant dark conditions. We only used data collected in LD condition. NimbleGen Aedes aegypti 12plex array already rma normalized were provided with VectorBase geneIDs.

***Drosophila melanogaster* (head and body).** RNA-seq data from [45] was downloaded from GEO accession (GSE64108). They measured RNA concentrations in the head and body of 3-, 5-, and 7-week-old adult flies in ad libidum feeding or 12-hour time-restricted feeding conditions. We only used data from ad libidum feeding condition of 5-week-old adult flies with best temporal resolution.

Cross-referenced gene IDs and known cycling genes

GeneID, protein coding status, ProbSetID, transcriptsID were downloaded from Ensembl [46] or VectorBase [47] using BioMart.

Known cycling genes were obtained from the KEGG [48] or FlyBase [49] database:

- KEGG circadian entrainment entry pathway for the mouse (mmu04713) and the rat (rno04713). This is the pathway by which light activates SCN neurons and the resulting signaling cascade that leads to a phase resetting of the circadian rhythm generated in these neurons. Most of these genes are not involved in generating the rhythm itself and as such cannot be called 'clock genes'.
- KEGG circadian rhythm entry pathway for the baboon (human hsa04710), and Anopheles (aga04711)
- FlyBase circadian rhythm entry pathway for Drosophila (GO:0007623).

Supporting information

S1 Table. Gene expression time-series datasets. Gene expression time-series datasets that come from circadian experiments. We kept data from healthy, wild-type individuals for these seven species, allowing comparisons among vertebrates and among insects. We preferred data from light-dark (LD) and ad-libitum experimental conditions whenever possible as providing a better representation of wild conditions. LD for regular alternation of light and darkness each 24h; and DD for continuous darkness usually after an entrainment to a 12h:12h light: dark.
(XLSX)

S1 File. Supplementary results.
(PDF)

S2 File. Density distribution of raw and default *p*-values obtained for the seven rhythm detection methods applied to vertebrate datasets. Density distribution of *p*-values obtained before (raw) and after the default correction (software) for the seven methods applied to each vertebrate dataset, sub-categorized in: i. randomized data which represents the null hypothesis; ii. randomized data restricted to the first and fourth quartiles of the median gene expression level, to check for the impact of expression level under the null; iii. the full original dataset; iv. the first and fourth quartiles of the median gene expression level of the original data; and v. a subset of known cycling genes when such data was available (8 to 99 genes according to species). The default *p*-values of ARS, GeneCycle, and LS are uncorrected.
(PDF)

S3 File. Following S2 File.
(PDF)

S4 File. Signal of evolutionary conservation of rhythmic gene expression in vertebrates. *p*-values density distribution of rhythmic orthologs vs non-rhythmic orthologs obtained for the seven methods applied to different vertebrate datasets. Orthologous genes detected as rhythmic in the same organ of two species have a stronger statistical signal of rhythmicity than those detected as not-rhythmic in at least one species. From all species_1 genes, only species_1-species_2 one-to-one orthologs are kept. Considering homologous tissues, these orthologs are separated into two groups: genes for which the ortholog is detected as rhythmic in this tissue of species_2, called rhythmic orthologs; and the remaining one-to-one orthologs.
(PDF)

S5 File. Variation of the proportion A/B as a function of the number of orthologs detected rhythmic, obtained for each method applied to different vertebrate datasets. The benchmark gene set is composed of species_1-species_2 orthologs, detected rhythmic in the homologous tissue of species_2 by the ARS, GeneCycle, or empJTK method with default p-value ≤ 0.01 or 0.05. See Fig 6 for definitions of sets A and B. The black line is the Naive method which orders genes according to their median expression levels (median of time-points), from highest expressed to lowest expressed gene, then, for each gene, calculates the proportion of rhythmic orthologs among those with higher expression.
(PDF)

S6 File. Results in insects.
(PDF)

Acknowledgments

We thank Paul Franken for useful discussions, as well as all members of the Robinson-Rechavi lab.

Author Contributions

Conceptualization: David Laloum, Marc Robinson-Rechavi.

Data curation: David Laloum.

Formal analysis: David Laloum.

Funding acquisition: Marc Robinson-Rechavi.

Methodology: David Laloum, Marc Robinson-Rechavi.

Writing – original draft: David Laloum.

Writing – review & editing: Marc Robinson-Rechavi.

References

1. Mure LS, Le HD, Benegiamo G, Chang MW, Rios L, Jillani N, et al. Diurnal transcriptome atlas of a primate across major neural and peripheral tissues. Science. 2018; 359 (6381). https://doi.org/10.1126/science.aao0318
2. Zhang R, Lahens NF, Ballance HI, Hughes ME, Hogenesch JB. A circadian gene expression atlas in mammals: Implications for biology and medicine. Proceedings of the National Academy of Sciences. 2014; 111(45):16219–16224. https://doi.org/10.1073/pnas.1408886111

3. Boyle G, Richter K, Priest HD, Traver D, Mockler TC, Chang JT, et al. Comparative Analysis of Vertebrate Diurnal/Circadian Transcriptomes. PLOS ONE. 2017; 12(1):1–18. https://doi.org/10.1371/journal.pone.0169923
4. Yang R, Su Z. Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation. Bioinformatics. 2010; 26(12):i168–i174. https://doi.org/10.1093/bioinformatics/btq189
5. Thaben PF, Westermark PO. Detecting Rhythms in Time Series with RAIN. Journal of Biological Rhythms. 2014; 29(6):391–400. https://doi.org/10.1177/0748730414553029
6. Hutchison AL, Maienschein-Cline M, Chiang AH, Tabei SMA, Gudjonson H, Bahroos N, et al. Improved Statistical Methods Enable Greater Sensitivity in Rhythm Detection for Genome-Wide Data. PLOS Computational Biology. 2015; 11(3):1–29. https://doi.org/10.1371/journal.pcbi.1004094
7. Deckard A, Anafi RC, Hogenesch J, Haase S, Harer J. Design and Analysis of Large-Scale Biological Rhythm Studies: A Comparison of Algorithms for Detecting Periodic Signals in Biological Data. Bioinformatics (Oxford, England). 2013; 29. https://doi.org/10.1093/bioinformatics/btt541
8. Michael TP, Mockler TC, Breton G, McEntee C, Byer A, Trout JD, et al. Network Discovery Pipeline Elucidates Conserved Time-of-Day–Specific cis-Regulatory Modules. PLOS Genetics. 2008; 4(2):1–17. https://doi.org/10.1371/journal.pgen.0040014
9. Hughes ME, Abruzzi KC, Allada R, Anafi R, Arpat AB, Asher G, et al. Guidelines for Genome-Scale Analysis of Biological Rhythms. Journal of Biological Rhythms. 2017; 32(5):380–393. https://doi.org/10.1177/0748730417728663
10. Korenčič A, Bordyugov G, Košir R, Rozman D, Goličnik M, Herzel H. The Interplay of cis-Regulatory Elements Rules Circadian Rhythms in Mouse Liver. PLOS ONE. 2012; 7(11):1–13.
11. Chudova D, Ihler A, Lin K, Andersen B, Smyth P. Bayesian detection of non-sinusoidal periodic patterns in circadian expression data. Bioinformatics (Oxford, England). 2009; 25:3114–20. https://doi.org/10.1093/bioinformatics/btp547
12. Miller BH, McDearmon EL, Panda S, Hayes KR, Zhang J, Andrews JL, et al. Circadian and CLOCK-controlled regulation of the mouse transcriptome and cell proliferation. Proceedings of the National Academy of Sciences. 2007; 104(9):3342–3347. https://doi.org/10.1073/pnas.0611724104
13. Yoo SH, Yamazaki S, Lowrey PL, Shimomura K, Ko CH, Buhr ED, et al. PERIOD2::LUCIFERASE real-time reporting of circadian dynamics reveals persistent circadian oscillations in mouse peripheral tissues. Proceedings of the National Academy of Sciences. 2004; 101(15):5339–5346. https://doi.org/10.1073/pnas.0308709101
14. Boothroyd CE, Wijnen H, Naef F, Saez L, Young MW. Integration of Light and Temperature in the Regulation of Circadian Gene Expression in Drosophila. PLOS Genetics. 2007; 3:1–16. https://doi.org/10.1371/journal.pgen.0030054
15. Nagoshi E, Saini C, Bauer C, Laroche T, Naef F, Schibler U. Circadian Gene Expression in Individual Fibroblasts: Cell-Autonomous and Self-Sustained Oscillators Pass Time to Daughter Cells. Cell. 2004; 119(5):693–705. https://doi.org/10.1016/j.cell.2004.11.015 PMID: 15550250
16. Gerber A, Saini C, Curie T, Emmenegger Y, Rando G, Gosselin P, et al. The systemic control of circadian gene expression. Diabetes, Obesity and Metabolism. 2015; 17(S1):23–32. https://doi.org/10.1111/dom.12512 PMID: 26332965
17. Hughes ME, Hogenesch JB, Kornacker K. JTK_CYCLE: An Efficient Nonparametric Algorithm for Detecting Rhythmic Components in Genome-Scale Data Sets. Journal of Biological Rhythms. 2010; 25(5):372–380. https://doi.org/10.1177/0748730410379711 PMID: 20876817
18. Lomb NR. Least-squares frequency analysis of unequally spaced data. Astrophysics and Space Science. 1976; 39(2):447–462. https://doi.org/10.1007/BF00648343
19. Scargle J. Studies in astronomical time series analysis. II—Statistical aspects of spectral analysis of unevenly spaced data. The Astrophysical Journal. 1983; 263.
20. Wu G, Hogenesch JB, Anafi RC, Hughes ME, Kornacker K. MetaCycle: an integrated R package to evaluate periodicity in large scale data. Bioinformatics. 2016; 32(21):3351–3353. https://doi.org/10.1093/bioinformatics/btw405 PMID: 27378304
21. Ahdesmäki M, Lähdesmäki H, Pearson R, Huttunen H, Yli-Harja O. Robust detection of periodic time series measured from biological systems. BMC Bioinformatics. 2005; 6(1):117. https://doi.org/10.1186/1471-2105-6-117 PMID: 15892890
22. Ahdesmaki M, Fokianos K, Strimmer K. GeneCycle: Identification of Periodically Expressed Genes; 2012. Available from: https://CRAN.R-project.org/package=GeneCycle.
23. de Lichtenberg U, Wernersson R, Jensen TS, Nielsen HB, Fausbøll A, Schmidt P, et al. New weakly expressed cell cycle-regulated genes in yeast. Yeast. 2005; 22(15):1191–1201. https://doi.org/10.1002/yea.1302 PMID: 16278933

24. Cohen-Steiner D, Edelsbrunner H, Harer J, Mileyko Y. Lipschitz Functions Have Lp-Stable Persistence. Foundations of Computational Mathematics. 2010; 10(2):127–139. https://doi.org/10.1007/s10208-010-9060-6

25. Straume M. DNA Microarray Time Series Analysis: Automated Statistical Assessment of Circadian Rhythms in Gene Expression Patterning. In: Numerical Computer Methods, Part D. vol. 383 of Methods in Enzymology. Academic Press; 2004. p. 149—166. Available from: http://www.sciencedirect.com/science/article/pii/S0076687904830076.

26. Fokianos K, Strimmer K, Wichert S. Identifying periodically expressed transcripts in microarray time series data. Bioinformatics. 2004; 20(1):5–20. https://doi.org/10.1093/bioinformatics/btg364 PMID: 14693803

27. Zhao W, Agyepong K, Serpedin E, Dougherty ER. Detecting Periodic Genes from Irregularly Sampled Gene Expressions: A Comparison Study. EURASIP Journal on Bioinformatics and Systems Biology. 2008; 2008(1):769293.

28. Dequéant ML, Ahnert S, Edelsbrunner H, Fink TMA, Glynn EF, Hattem G, et al. Comparison of Pattern Detection Methods in Microarray Time Series of the Segmentation Clock. PLOS ONE. 2008; 3(8):1–9.

29. Wu G, Zhu J, Yu J, Zhou L, Huang JZ, Zhang Z. Evaluation of Five Methods for Genome-Wide Circadian Gene Identification. Journal of Biological Rhythms. 2014; 29(4):231–242. https://doi.org/10.1177/0748730414537788 PMID: 25238853

30. Gabaldón T, Koonin EV. Functional and evolutionary implications of gene orthology. Nature Reviews Genetics. 2013; 14:360 EP –. https://doi.org/10.1038/nrg3456 PMID: 23552219

31. Gerhart-Hines Z, Lazar MA. Circadian Metabolism in the Light of Evolution. Endocrine Reviews. 2015; 36(3):289–304. https://doi.org/10.1210/er.2015-1007 PMID: 25927923

32. Hutchison AL, Dinner AR. Correcting for Dependent P-values in Rhythm Detection. bioRxiv. 2017;.

33. Rund SSC, Hou TY, Ward SM, Collins FH, Duffield GE. Genome-wide profiling of diel and circadian gene expression in the malaria vector Anopheles gambiae. Proceedings of the National Academy of Sciences. 2011; 108(32):E421–E430. https://doi.org/10.1073/pnas.1100584108

34. Leming MT, Rund SS, Behura SK, Duffield GE, O'Tousa JE. A database of circadian and diel rhythmic gene expression in the yellow fever mosquito Aedes aegypti. BMC Genomics. 2014; 15(1):1128. https://doi.org/10.1186/1471-2164-15-1128 PMID: 25516260

35. Wijnen H, Naef F, Boothroyd C, Claridge-Chang A, Young MW. Control of Daily Transcript Oscillations in Drosophila by Light and the Circadian Clock. PLOS Genetics. 2006; 2:1–18. https://doi.org/10.1371/journal.pgen.0020039

36. Rosikiewicz M, Robinson-Rechavi M. IQRray, a new method for Affymetrix microarray quality control, and the homologous organ conservation score, a new benchmark method for quality control metrics. Bioinformatics. 2014; 30(10):1392–1399. https://doi.org/10.1093/bioinformatics/btu027 PMID: 24451627

37. Kryuchkova-Mostacci N, Robinson-Rechavi M. A benchmark of gene expression tissue-specificity metrics. Briefings in Bioinformatics. 2016; 18(2):205–214.

38. Hutchison AL, Allada R, Dinner AR. Bootstrapping and Empirical Bayes Methods Improve Rhythm Detection in Sparsely Sampled Data. Journal of Biological Rhythms. 2018; 33(4):339–349. https://doi.org/10.1177/0748730418789536 PMID: 30101659

39. Schibler U. The daily rhythms of genes, cells and organs. EMBO reports. 2005; 6(S1):S9–S13. https://doi.org/10.1038/sj.embor.7400424 PMID: 15995671

40. Iuchi H, Sugimoto M, Tomita M. MICOP: Maximal information coefficient-based oscillation prediction to detect biological rhythms in proteomics data. BMC Bioinformatics. 2018; 19(1):249. https://doi.org/10.1186/s12859-018-2257-4 PMID: 29954316

41. Altenhoff AM, Gonnet GH, Train CM, Dylus D, Glover NM, de Farias TM, et al. The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. Nucleic Acids Research. 2017; 46(D1):D477–D485. https://doi.org/10.1093/nar/gkx1019

42. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. UpSet: Visualization of Intersecting Sets. IEEE Transactions on Visualization and Computer Graphics. 2014; 20(12):1983–1992. https://doi.org/10.1109/TVCG.2014.2346248 PMID: 26356912

43. Pizarro A, Hayer K, Lahens NF, Hogenesch J. CircaDB: A database of mammalian circadian gene expression profiles. Nucleic acids research. 2012; 41. https://doi.org/10.1093/nar/gks1161 PMID: 23180795

44. Sukumaran S, Jusko WJ, DuBois DC, Almon RR. Light-dark oscillations in the lung transcriptome: implications for lung homeostasis, repair, metabolism, disease, and drug action. Journal of Applied Physiology. 2011; 110(6):1732–1747. https://doi.org/10.1152/japplphysiol.00079.2011 PMID: 21436464

45. Christopher B, Gill S, Melkani G, Panda S. type; 2015. Available from: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64108. GSE64108.

46. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Research. 2017; 46(D1):D754–D761. https://doi.org/10.1093/nar/gkx1098

47. Giraldo-Calderón GI, Emrich SJ, MacCallum RM, Maslen G, Dialynas E, Topalis P, et al. VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. Nucleic Acids Research. 2014; 43(D1):D707–D713. https://doi.org/10.1093/nar/gku1117 PMID: 25510499

48. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research. 1999; 27(1):29–34. https://doi.org/10.1093/nar/27.1.29 PMID: 9847135

49. the FlyBase Consortium, Thurmond J, Goodman JL, Kaufman TC, Strelets VB, Calvi BR, et al. FlyBase 2.0: the next generation. Nucleic Acids Research. 2018; 47(D1):D759–D765. https://doi.org/10.1093/nar/gky1003

2.2 Supporting information

2.2.1 S1 Table

Species		 *Papio anubis*	 *Mus musculus*		 *Rattus norvegicus*	 *Danio rerio*	 *Drosophila melanogaster*	 *Aedes aegypti*	 *Anopheles gambiae*
Transcriptome profiling technique		RNA-seq	microarray	RNA-seq	microarray	microarray	RNA-seq	microarray	microarray
Tissue	Aorta	★	★	★					
	Brain stem	★	★	★					
	Cerebellum	★	★	★					
	Heart	★	★	★			★		
	Liver	★	★	★		★			
	Lung	★	★	★	★				
	Muscle	★	★	★					
	SCN	★	★						
	White adipose	★	★	★					
	Head						★	★	★
	Body						★	★	
Experimental conditions	Total period length	24h	48h	48h	72h	48h	24h	48h	48h
	Exp. conditions	LD	DD	DD	LD	LD	LD	LD	LD
	Regular sampling each	2h	2h	6h	2h	4h	3h	4h	4h
	Number of biological replicates per time-point	x1	x3	x3	x1	x3	x1	x1	x2
References		Mure, 2018	Zhang, 2014		Sukumaran, 2011	Boyle, 2017	Gill, 2017	Leming, 2014	Rund, 2011

Images credit: Anthony Caravaggi (mouse), Ian Quigley (zebrafish) both with license CC BY-NC-SA 3.0, wikipedia GNU GPL Muhammad Mahdi Karim (baboon), and Public Domain for other images (from http://phylopic.org/)

Table S1: Gene expression time-series datasets. Gene expression time-series datasets that come from circadian experiments. We kept data from healthy, wild-type individuals for these seven species, allowing comparisons among vertebrates and among insects. We preferred data from light-dark (LD) and ad-libitum experimental conditions whenever possible as providing a better representation of wild conditions. LD for regular alternation of light and darkness each 24h; and DD for continuous darkness usually after an entrainment to a 12h:12h light:dark

2.2.2 S1 File

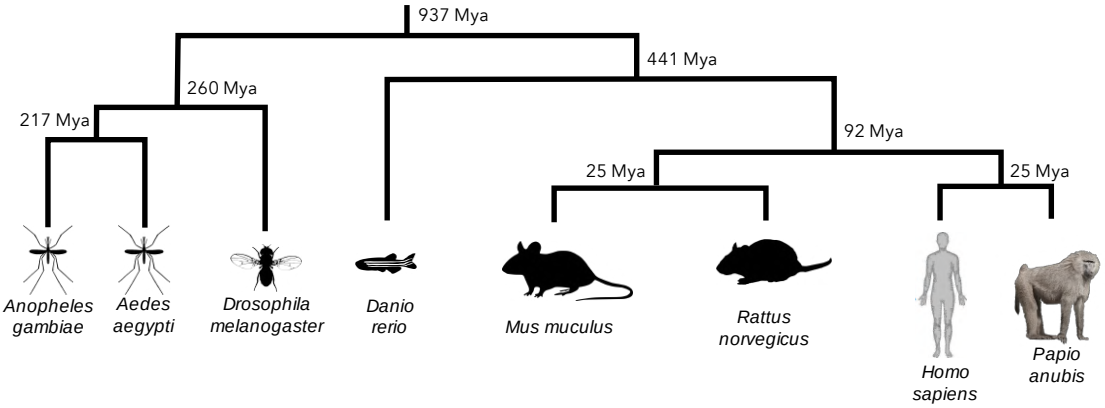


Fig S1. Phylogenetic tree of species studied. (images credits at the end of the file)

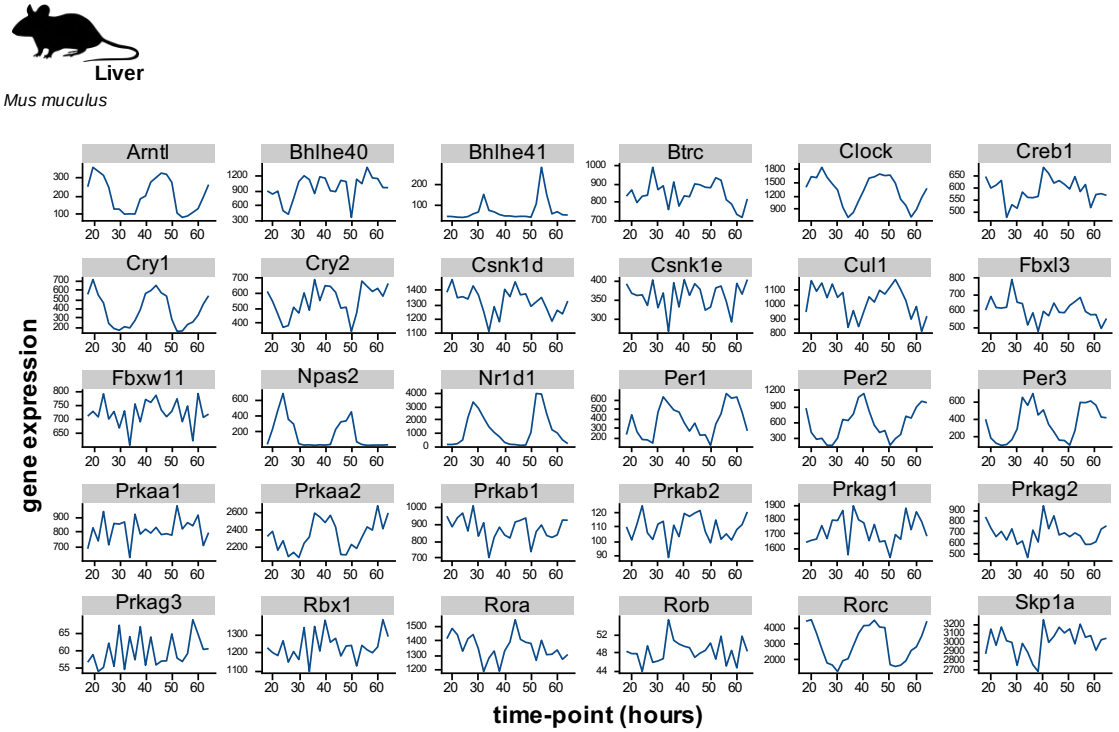


Fig S2. Expression over time of some known circadian genes in liver from mouse experiment data [2] (microarray).

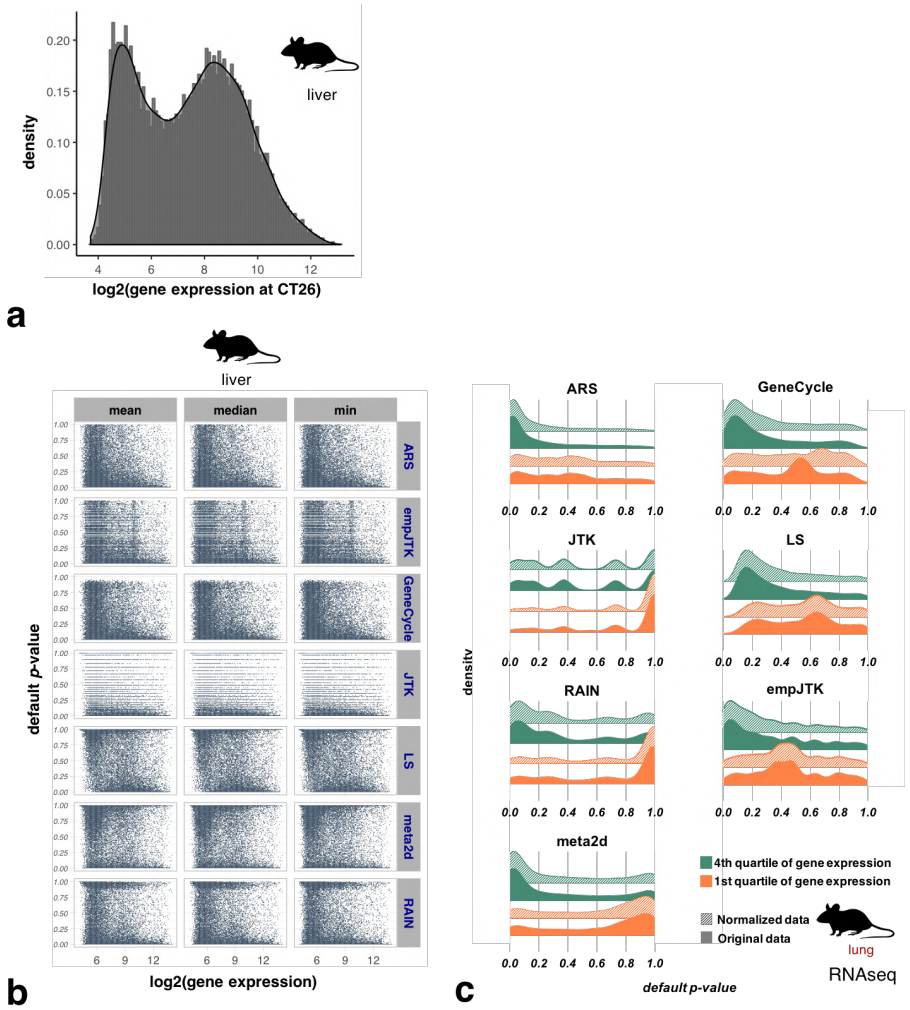


Fig S3. a) Typical bimodal density distribution of gene expression at time-point CT26 from mouse liver data (microarray). **b)** Gene expression per time-point, calculated with the median, the mean, or the minimum of time-points, as a function of default p-values obtained for the seven methods applied to mouse liver data.
c) Methods applied to original vs normalized gene expression values of mouse lung data (RNAseq) produce the same distributions of p-values within highly expressed genes, or within lowly expressed genes.

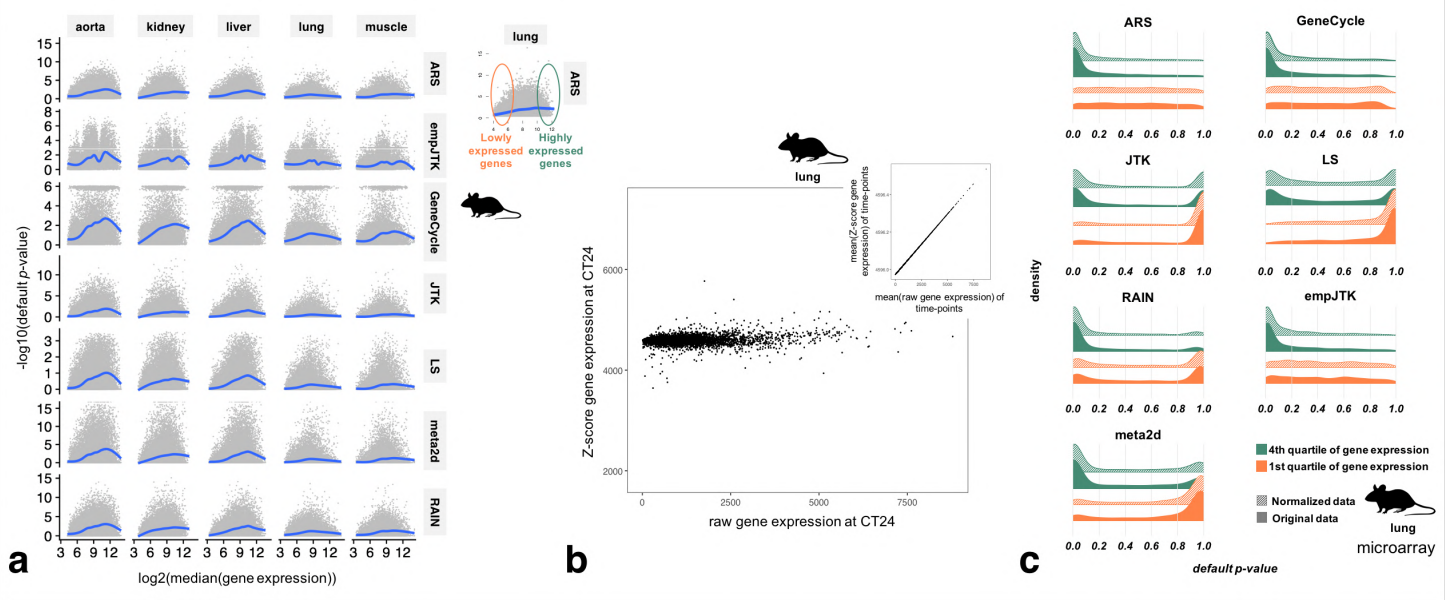


Fig S4. Detected rhythmic genes are enriched in highly expressed genes.
a) Scatter plot showing the relation between the expression level of genes (median of time-points) and their default p-values. The blue line is the smoothed conditional mean. Higher expression levels imply a higher power to detect rhythmic patterns.
b) Control of the normalization by Z-score of gene expression values at a given time-point (CT24).
c) Methods applied to original vs normalized gene expression values produce the same distributions of p-values within highly expressed genes, or within lowly expressed genes. Particularly, the normalization of gene expression values does not allow to recover rhythmicity within lowly expressed genes.

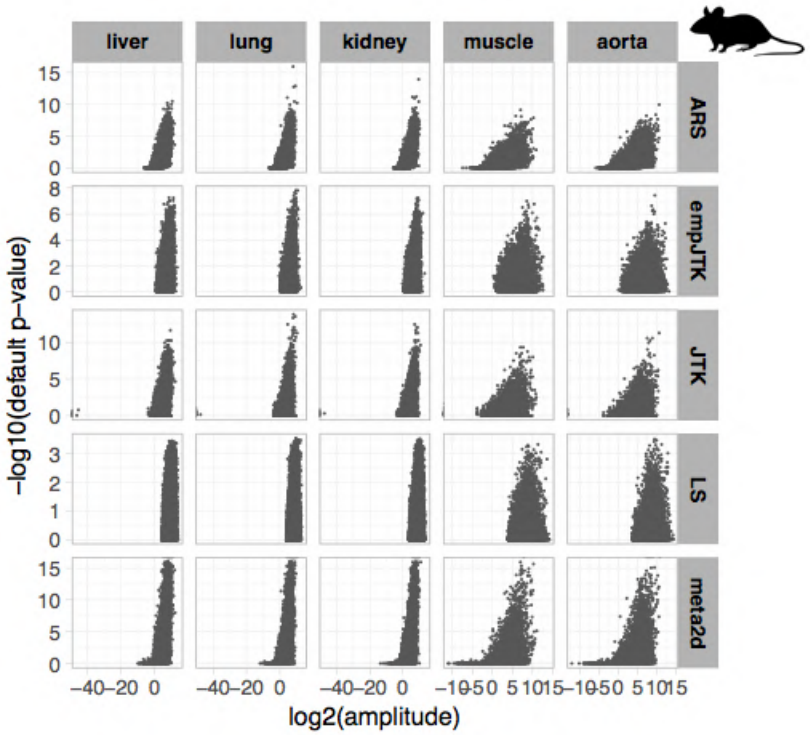


Fig S5. Scatter-plots showing the amplitude of gene expression as a function their default p-values obtained for the methods applied to five mouse tissues (microarray). Only five methods are shown since they are the only ones giving estimations of amplitudes.

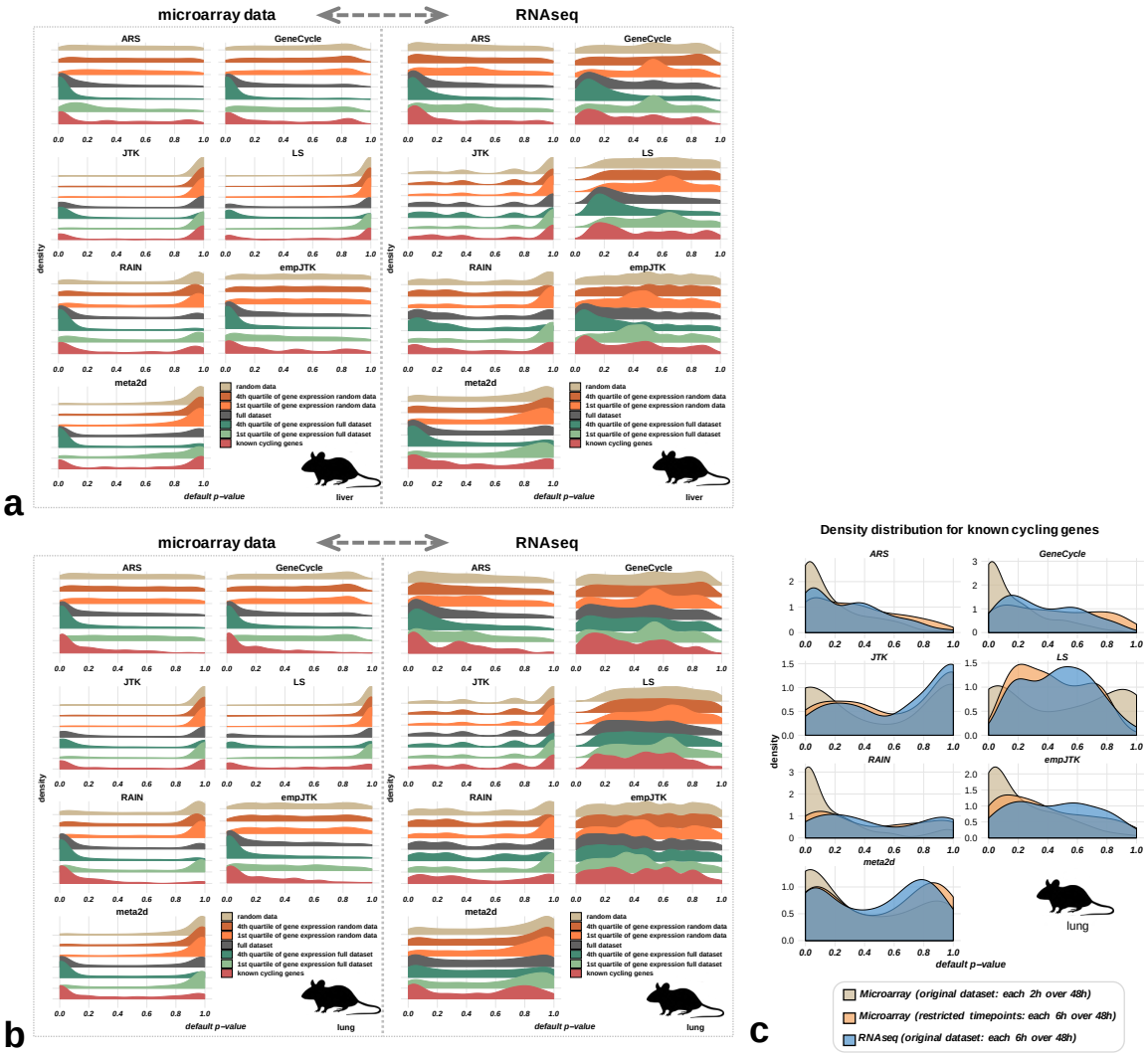


Fig S6. a)b) p-values distributions obtained from microarray vs RNAseq in mouse liver (**a**) and lung (**b**).
c) The restriction of microarray time-series to the same time-points as in the RNAseq series applied to known cycling genes produces similar p-value distributions to those obtained with RNAseq.

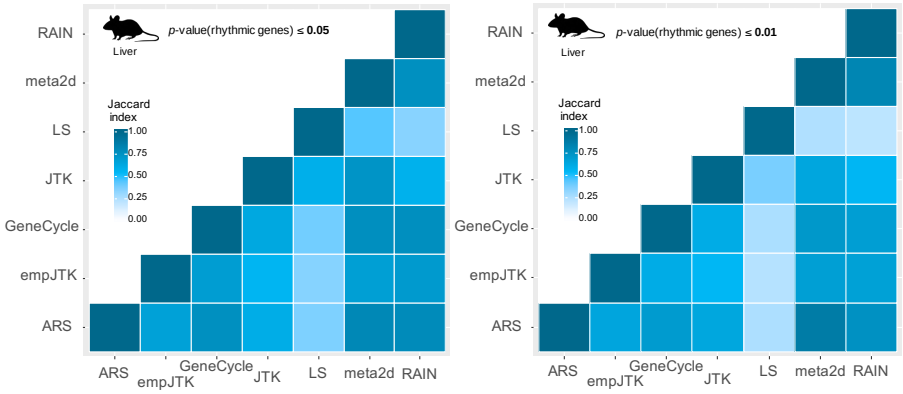


Fig S7. Heatmaps of Jaccard-indices comparing the similarity of genes called rhythmic (default p-value ≤ 0.05 or 0.01) between the methods applied to mouse liver dataset (microarray).

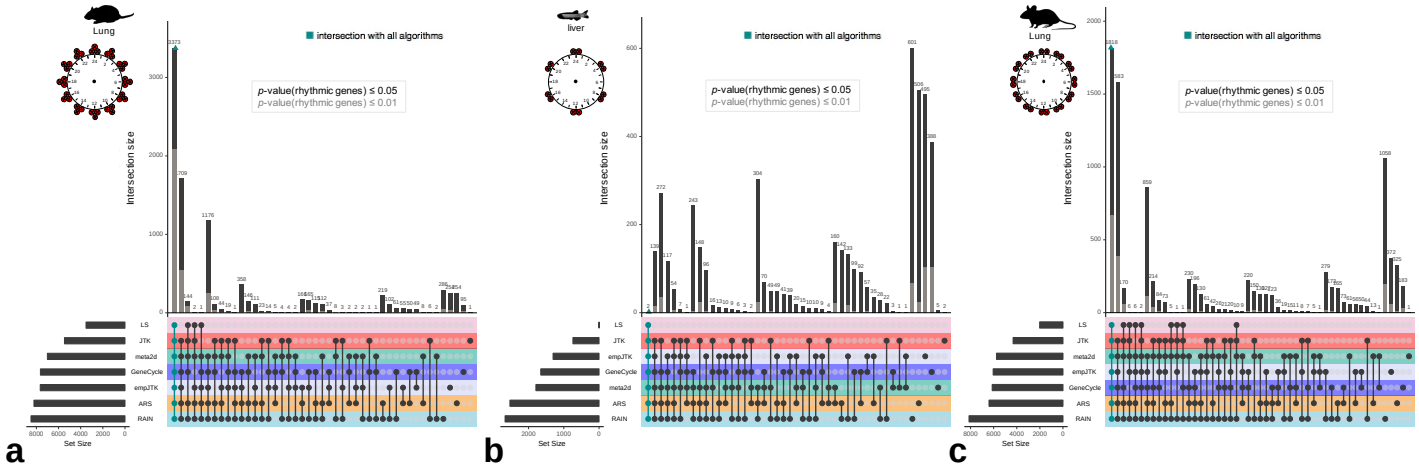


Fig S8. Upset diagram for mouse lung dataset (microarray) (**a**), zebrafish liver dataset (**b**), and mouse lung dataset (**c**) for the p-value thresholds of 0.05 (black) or 0.01 (grey) for calling genes rhythmic.

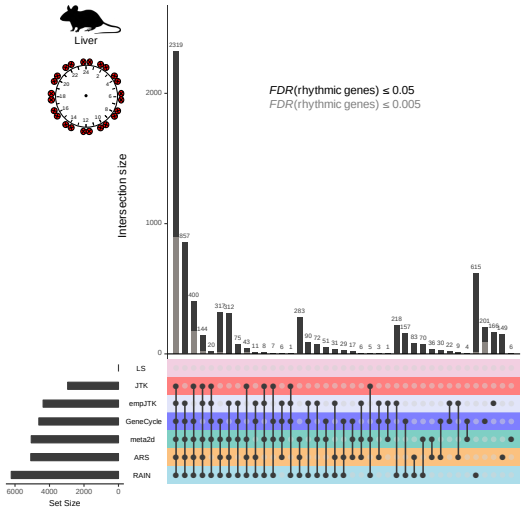


Fig S9. Using a very low false positive tolerance, all methods except LS overlap largely.
Upset diagram for mouse liver dataset (microarray) for the FDR thresholds of 0.05 (black) or 0.005 (grey) for calling genes rhythmic. FDR correspond to the false discovery rate adjustment of default p-values using `p.adjust` R function.

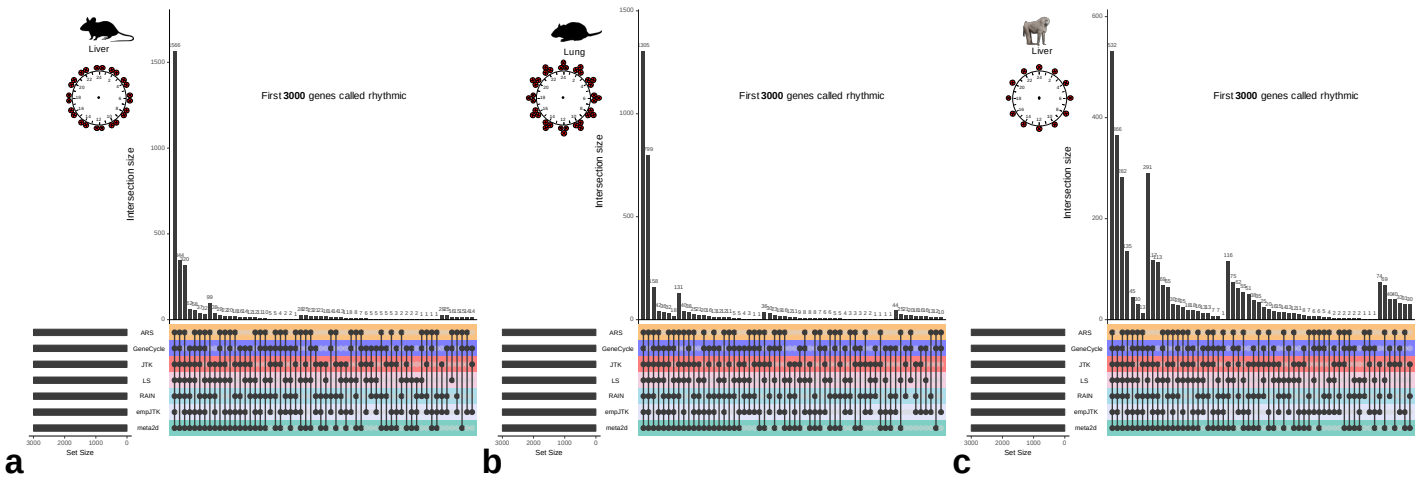


Fig S10. Upset diagram for mouse liver dataset (microarray) (**a**), rat lung dataset (**b**), and baboon liver dataset (**c**) for the first 3000 genes detected rhythmic for each method.

Fig S11. Default p-values density distribution of rhythmic orthologs vs non-rhythmic orthologs obtained for the seven methods applied to mouse liver data. Rhythmic orthologs are mouse-zebrafish 1:1 orthologs detected rhythmic in zebrafish liver each rhythm detection method (in red) and a p-value threshold of 0.01 or 0.05.

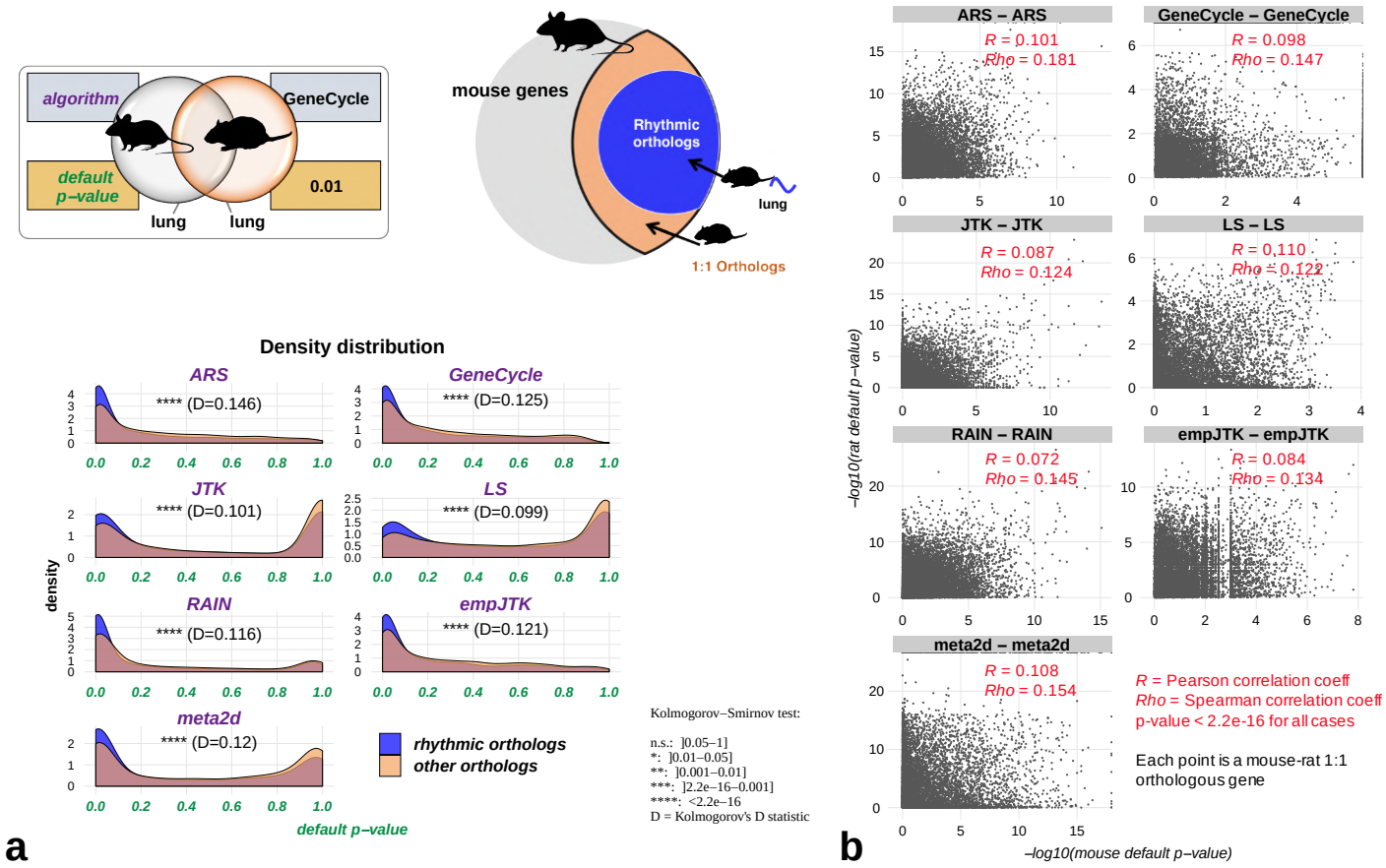


Fig S12. a) Distribution of default p-values of rhythmic and non-rhythmic mouse-rat orthologs obtained for the seven methods applied to mouse lung data. Rhythmic genes of rat lung are detected using the GeneCycle method and a p-value threshold of 0.01.
b) Scatter-plot comparing p-values for each mouse-rat one-to-one orthologous gene obtained for the seven methods (same method used for each comparison). Pearson R and Spearman Rho coefficients have been obtained with a significance p-value $< 2.2e{-}16$.

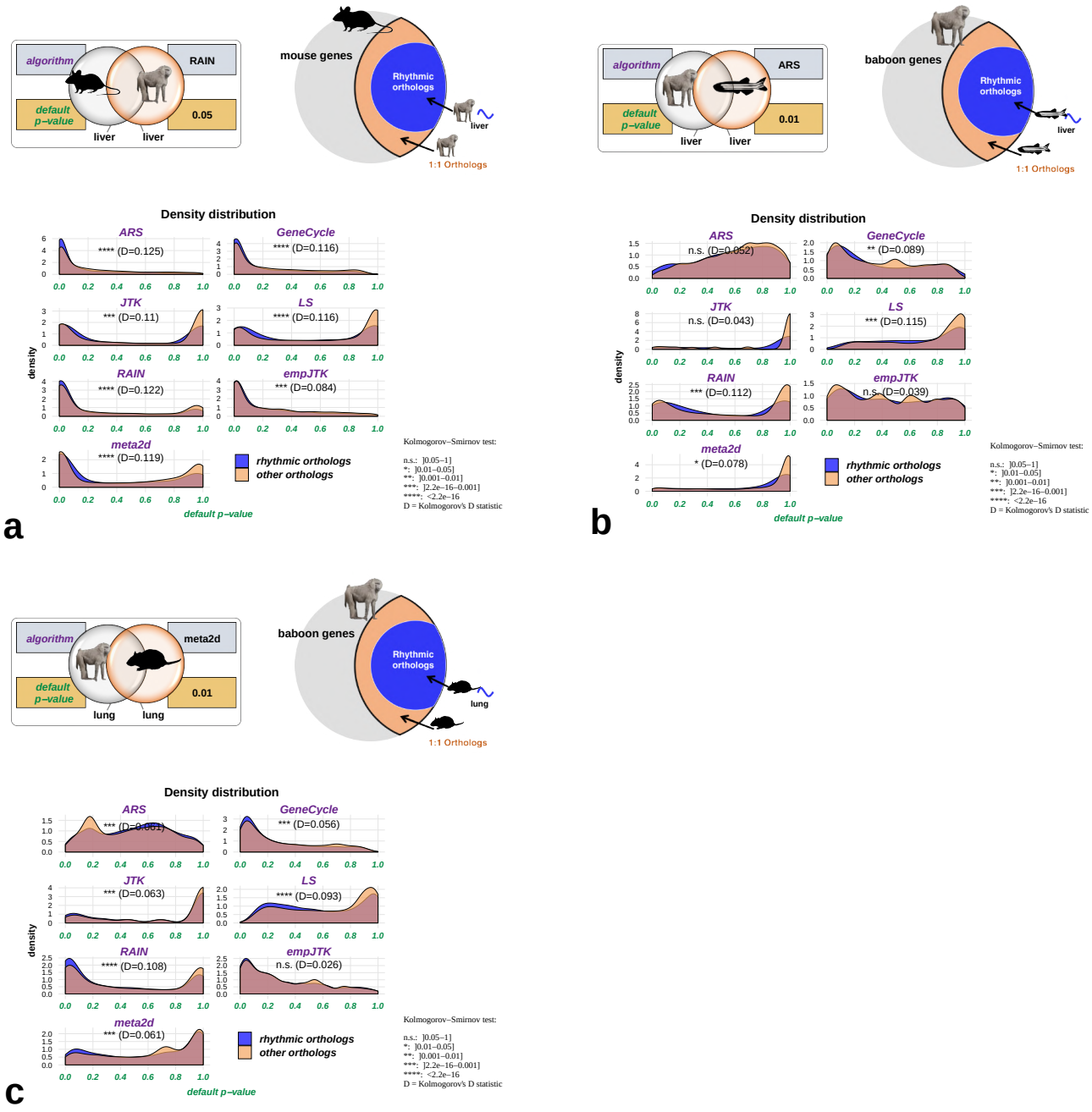


Fig S13. Distribution of default p-values of rhythmic and non-rhythmic mouse-baboon (**a**), baboon-zebrafish (**b**), and baboon-rat (**c**) orthologs obtained for the seven methods applied to the homologous tissue of species_1.

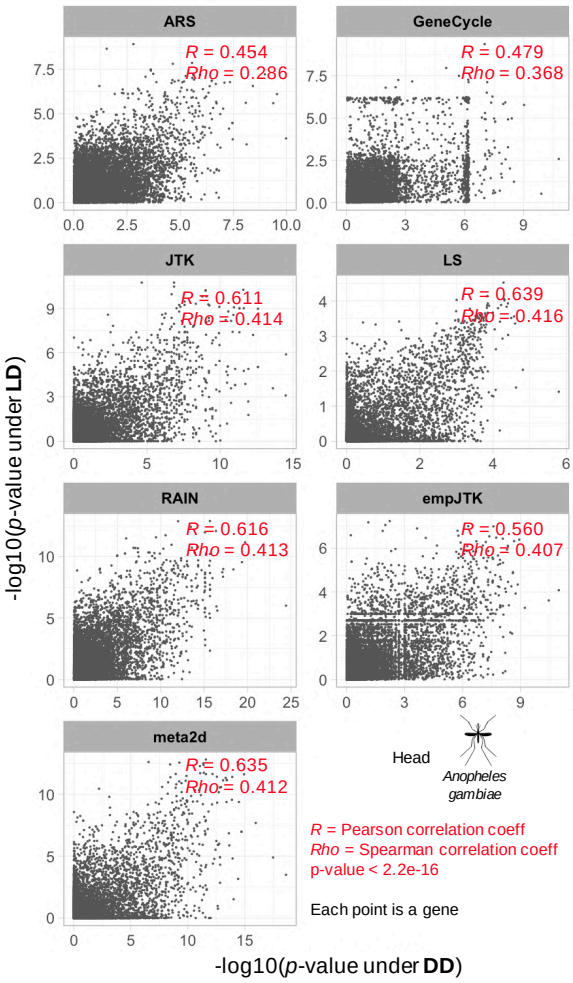


Fig S14. Scatter-plot comparing p-values obtained for the seven methods applied to the head dataset of *Anopheles gambiae* under light-dark (LD) or dark-dark (DD) conditions [1]. Pearson R and Spearman *Rho* coefficients have been obtained with a significance p-value $< 2.2e{-}16$.

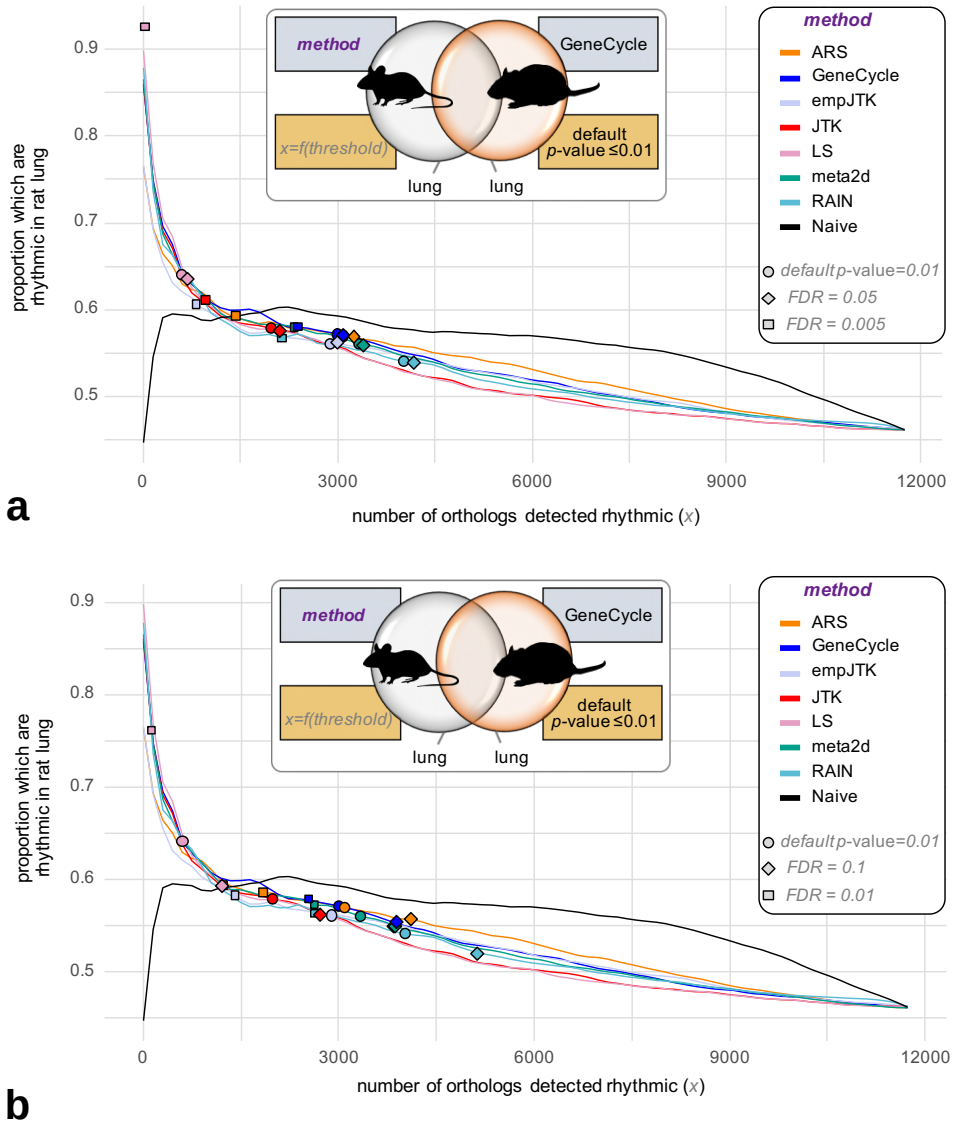


Fig S15.
Variation of the proportion rhythmic orthologs/all orthologs in mouse lung as a function of the number of mouse orthologs detected rhythmic, for each method applied to the mouse lung dataset. The benchmark gene set is composed of mouse-rat orthologs, detected rhythmic in rat lung by the GeneCycle method with default p-value ≤ 0.01. The black line is the Naive method which orders genes according to their median expression levels (median of time-points), from highest expressed to lowest expressed gene, then, for each gene, calculates the proportion of rhythmic orthologs among those with higher expression. Rings correspond to a p-value threshold of 0.01, diamonds correspond to a FDR threshold of 0.05 (**a**) or 0.1 (**b**), and squares correspond to a FDR threshold of 0.005 (**a**) or 0.01 (**b**). FDR correspond to the false discovery rate adjustment of default p-values using `p.adjust` R function.

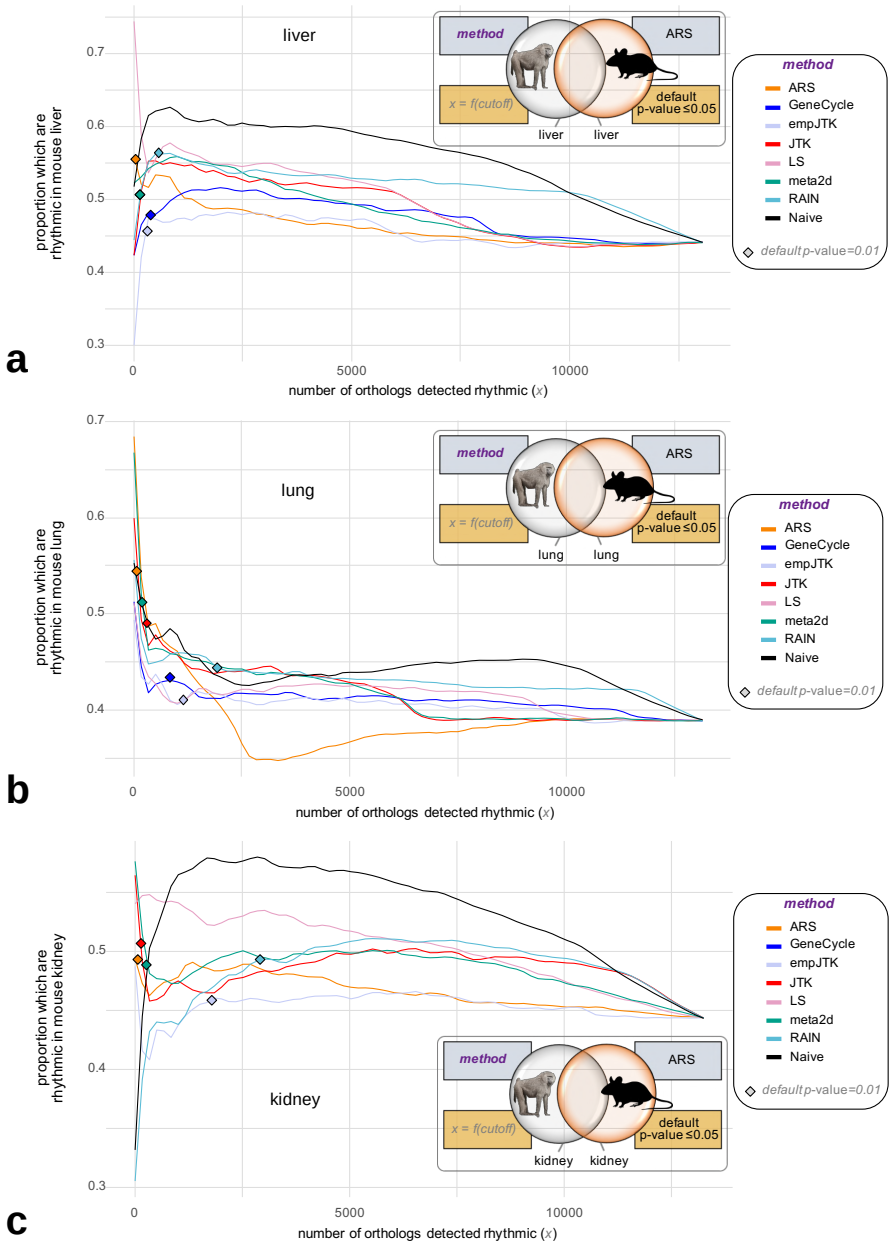


Fig S16. Variation of the proportion rhythmic orthologs/all orthologs in baboon as a function of the number of mouse orthologs detected rhythmic, for each method applied to the baboon lung (**a**), lung (**b**), and kidney (**c**) dataset. The benchmark gene set is composed of baboon-mouse orthologs, detected rhythmic in the homologous tissue of mouse by the ARS method with default p-value ≤ 0.05.

References

1. S. S. C. Rund, T. Y. Hou, S. M. Ward, F. H. Collins, and G. E. Duffield. Genome-wide profiling of diel and circadian gene expression in the malaria vector anopheles gambiae. *Proceedings of the National Academy of Sciences*, 108(32):E421–E430, 2011.

2. R. Zhang, N. F. Lahens, H. I. Ballance, M. E. Hughes, and J. B. Hogenesch. A circadian gene expression atlas in mammals: Implications for biology and medicine. *Proceedings of the National Academy of Sciences*, 111(45):16219–16224, 2014.

2.2.3 S2, S3, S4, S5, and S6 Files are available from the paper online

RESULTS 2

3 RESULTS 2: ARTICLE EN PREPARATION, *Energetic costs and expression noise of rhythmically expressed genes*

Energetic costs and expression noise of rhythmically expressed genes

David Laloum[1,2], Marc Robinson-Rechavi[1,2,*]

[1] Department of Ecology and Evolution, Batiment Biophore, Quartier UNIL-Sorge, Université de Lausanne, 1015 Lausanne, Switzerland
[2] Swiss Institute of Bioinformatics, Batiment Génopode, Quartier UNIL-Sorge, Université de Lausanne, 1015 Lausanne, Switzerland

* marc.robinson-rechavi@unil.ch

3.1 Introduction

Living organisms have to adapt to complex and changing environments. In a given environment, the conditions are not uniform but vary over time, generating patterns of variation. Physiological systems able to accommodate themselves to changing circumstances are expected to have a higher stability of survival and reproduction [47]. Circadian rhythms can be considered as a physiological system adapted to the cyclic environment imposed to all living organisms on Earth. “Circadian rhythms” denotes an entity characterized by an endogenous and entrainable oscillator clock able to persist in constant conditions (such as in constant darkness) whose phases can be altered (reset or entrained). However, many physiological systems display non-autonomous rhythms, directly or indirectly controlled by the local clock or mainly by the environment itself, or by both [5][6][7][8][9]. Such rhythms are found at all levels: molecular, cellular, organs, and behavioural level, and several regulatory networks appear to play roles in the synchronization of these levels [54]. The ubiquity of such rhythms can be seen as an adaptation to the most robust external patterns in nature, mostly characterized by the light/dark cycle and by the nycthemeral variations of the temperature, which govern the energetic cycles on Earth. Here, we studied the evolutionary trade-offs that shape the rhythmic nature of gene expression. For this, we analysed characteristics we presume to be part of the trade-off.
We call ”rhythmic genes” all genes displaying a 24-hours periodic variation of their mRNA or protein level, or of both, constituting the nycthemeral transcriptome or proteome. The rhythmic expression of these genes can be entrained directly by the internal clock but also indirectly by external inputs, such as the light-dark cycle or food-intake [5][6][7][8][9]. This is why we use the term ”nycthemeral” to avoid confusion with the specific features of ”circadian” rhythms, although these are included in the nycthemeral rhythms. Because the alternation of light and dark can be considered as a permanent signal for a long time scale for most living organisms [55], we consider that the entirety of nycthemeral biological rhythms are relevant.

Why develop cyclic systems, costly and complex?

a. Rhythmicity as an adaptation

The endogenous nature of circadian rhythms is an anticipation strategy, providing a clear advantage to the organisms able to anticipate their environmental changes. The evolutionary origin of maintaining large cyclic biological systems, in term of adaptability, can be seen as a trade-off between the disadvantages (cost and noise induced by the added complexity) and the advantages (economy over a daily time-scale, temporal

organization, adaptability). Indeed, gene expression is costly for the cell in terms of energy and cellular materials. Wang et al. [56] provided first results showing that in the liver of mice, abundant proteins that are required at one time are down-regulated at other times, apparently to economize on overall production. Wang et al. also showed that at each time-point, the total metabolic cost was ∼4 fold higher for the set of cycling genes compared to the non-cycling genes set at both transcriptional and translational levels [56] – although the proteomic data used from mouse fibroblasts appear to have been underestimated and have since been corrected [57]. Furthermore, cycling genes have been shown to be over-represented among highly expressed genes [56][58]. Here we present results which support the hypothesis that cycling gene expression allows minimizing the overall cellular energy usage by repressing periodically highly expressed genes. Thus, a first evolutionary advantage given by rhythmic biological processes is an optimization of the overall cost (over a 24-hours period), relative to constant expression at a high level, when that high level is needed for fitness at least at some point of time.

b. Noise and cost optimization

The nycthemeral transcriptome is known to be tissue-specific [32][33][34], i.e. a given gene can be rhythmically expressed in some tissues, and constantly expressed in others. The regulatory mechanism behind this tissue-specificity seems to be precise chromatin loops recruiting clock- and tissue-specific transcription factors (TFs) [35]. In addition, the rhythmic expression concerns a poorly defined set of genes whose features can be explored to better understand the putative adaptive role of their rhythmic regulation. It should be noted that rhythmic expression has costs, and thus it is not always obvious that it should be adaptive. For instance, regulatory dynamics can cause substantial changes in noise levels, e.g. the noise strength immediately following gene induction is almost twice the final steady-state value [59]. Here, the noise is the stochastic variation in proteins numbers within one cell (intrinsic noise), and is opposed to the gene expression precision. This noise is larger when there are few mRNAs per protein unit, i.e. many proteins are translated per mRNA unit.

These considerations lead to simple predictions which we test here: i) decreased stochasticity during a specific period of time for constant proteins translated from rhythmically accumulated mRNAs; ii) an economizing strategy for genes whose protein expression is rhythmic from constant mRNAs abundances (these genes are presumed to be noise-tolerant genes); and iii) a combined strategy for genes rhythmic at both layers.

3.2 Results

3.2.1 Cyclicality of highly expressed but normally costly proteins

The expression cost per gene has been shown to be dominated by the costs generated at the translational step [60][61]. Thus, for each gene, the expression cost per time-unit can be simplified by the formula (*1*) which take into account the averaged amino-acid (AA) synthesis cost, the protein length, and the protein abundance; we have neglected here the costs of protein decay. It gives an estimation of the cost that the cell requested to produce the number of proteins available at time t. To obtain a comparable estimation of expression costs between rhythmic and non-rhythmic proteins, we calculated the average and an approximation of the maximum protein expression level (Methods formula (2) and (3)) over time-points (Figure 3.1d).

$$C_p = N_p L_p (\bar{c}_{AA} - 1) \quad (1)$$

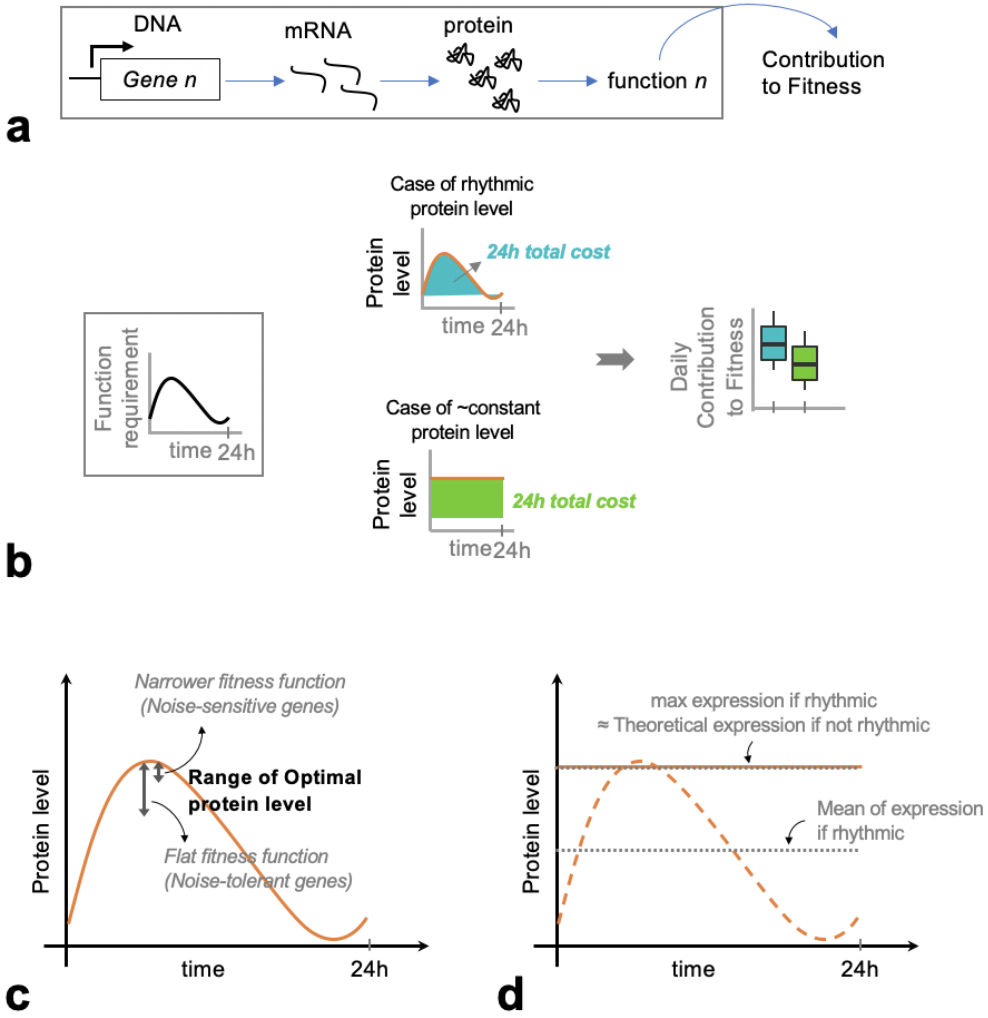


Figure 3.1: a) Gene expression contributes to the organismal fitness. **b)** Rhythmic protein regulation is a way of maintaining or increasing the contribution to fitness by regulating the expression costs. The integration into the rhythmic regulation might come from a trade-off between the costs saved, the costs generated by its integration into the rhythmic system, and the advantages provided. Assuming that a given protein level is only needed at some times, the integration into the rhythmic network minimize the overall costs while maintaining benefits. **c)** The range of optimal protein levels depends on the sensitivity of the function to deviations from the level which maximizes the contribution of the function to fitness. "Narrower" is the term used by Hausser et al. [62], but we could also use "steeper". Noise sensitive genes have steeper fitness function, i.e. a small deviation from the optimum rapidly decreases the contribution to fitness. Precision is less critical for genes with flat fitness functions. **d)** Mean or maximum expression level calculated from time-series datasets. We assume that the maximal expression level gives an estimation of the theoretical level that would have constantly been maintained in case of no rhythmic regulation.

In the case of a biological function periodically requested, the integration of its gene expression into the rhythmic regulation can be see as a balance between the costs saved by not producing proteins when they are not needed, and the costs involved in making it rhythmic (Figure 3.1b). This is why we might expect to retrieve costlier genes as rhythmic genes. First, we confirm that cycling genes are enriched in highly

expressed genes [58] [56] and thus constitute the most costly group of proteins to transcribe and translate in the genome [56]. However, per unit of protein, they were not more expensive to produce, except for the mouse liver dataset (Figure 3.2b). To compare with an acceptable group size, we considered as rhythmic the first 15% of proteins from *p*-values ranking obtained from the rhythm detection algorithms (see Methods). The AA biosynthesis costs estimated in *E. coli* (Supplementary Table S2) were used as representative for all species since biosynthetic pathways are nearly universal conserved [61]. Thus, the higher cost observed for cycling genes (Figure 3.2a) was especially due to their higher expression levels, observed both at the proteome level (in average and at their maximum levels, (Figure 3.2c and 3.2d), and at the transcriptional level, except in Ostreococcus (S1 File: Figure S1). Furthermore, we found that rhythmic proteins can be significantly longer than non-rhythmic ones (for mouse and cyanobacteria) (Figure 3.2b). Thus, rhythmic proteins are longer and therefore more expensive to produce per unit of protein. However, the immediate causality implying that a protein is rhythmic because its costs are higher per unit of protein (because longer) is probably wrong. Rather, it can be proposed that conserved genes (known to be longer [63][64]) are often more expressed [65][66], therefore more expensive to produce, explaining their rhythmic expression.

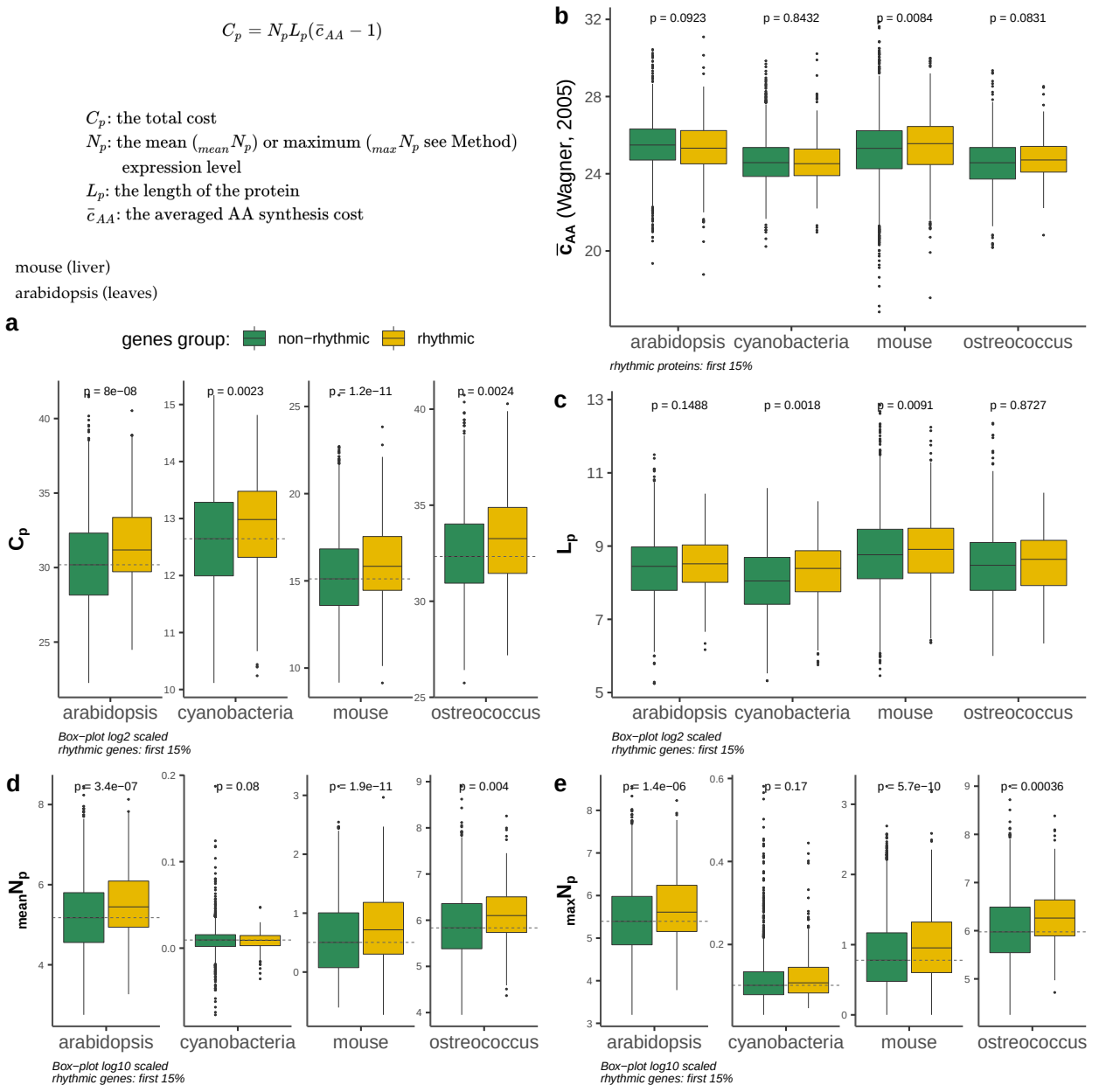


Figure 3.2: Rhythmic proteins are costly proteins due to their high level of expression. **a)** The total cost of rhythmic proteins is higher than those of other proteins. **b)** With the exception of mouse liver, rhythmic proteins do not contain more expensive amino-acids than other proteins. **c)** Rhythmic proteins can be longer in some species. **d-e)** Mean or maximum expression level calculated from time-series datasets: rhythmic proteins are highly expressed proteins.

These results support the hypothesis according to which costlier genes are preferentially under rhythmic regulation. If this is true globally, then the tissue-specificity of the rhythmic regulation can be expected to be due to the tissue-specific expression level requested. This leads us to propose that each gene should be found rhythmic specifically in tissues where a high expression of that gene is requested, i.e. where it is costliest.

3.2.2 For a given tissue, rhythmically expressed proteins are proteins whose function requires a higher level of expression

To test this hypothesis we used time-series datasets which included circadian sampling for several tissues from the same species (See Methods). For each gene we separated the tissues in two groups, ones for which the gene was rhythmic (p-value$\leq$cutoff.1) and those for which it was not rhythmic (p-value$>$cutoff.2). Because of the difficulty of setting reliable thresholds for rhythmicity [58], we ignored intermediate values (cutoff.1$<p$-value$\leq$cutoff.2). For each gene, we estimated the difference δ of expression levels between these two groups of tissues (expression levels were Z-score normalized, see Methods). We tested the hypothesis that the δ distribution mean is equal to 0 using the Student's test. As predicted, genes tend to have higher expression in tissues where they are rhythmic than in those where they are not rhythmic (Table 3.1). These results support the hypothesis that in a given tissue, genes whose function requires a high expression level in this tissue are rhythmically regulated. However, δ has a bimodal distributions for proteomic data (Figure 3.3), meaning that rhythmic proteins are separated in two groups: i) one group consists of lowly expressed proteins in tissues in which they are rhythmic, and ii) another consists of highly expressed proteins in tissues in which they are rhythmic.

species	**omics**	**technique**	**nb tissues**	**nb genes**	**parameter**	rhythmic	non-rhythmic
	Mean of Delta tissues	**t**	**df**	**lower**	**upper**	**Signif**	
mouse	**transcript**	microarray	**11**	19262	max expression level	p<0.01	p>0.5
	0.0146	28.29	11683.00	0.0136	0.0156	< 2.2e-16	
drosophila	**transcript**	RNAseq	**3**	8286	max expression level	p<0.01	p>0.5
	0.0024	2.16	115.00	0.0046	0.0046	0.0327	
anopheles	**transcript**	microarray	**2**	11269	max expression level	p<0.01	p>0.5
	0.0517	10.60	879.00	0.0422	0.0613	< 2.2e-16	
mouse	**protein**	SILAC	**3**	737	max expression level	p<0.05	p>0.1
	0.1036	4.66	255.00	0.0598	0.1474	5.2e-06	
mouse	**protein**	SILAC	**4**	388	max expression level	p<0.05	p>0.1
	0.0365	1.50	198.00	-0.0116	0.0846	0.1362	

Table 3.1: Student's test testing the hypothesis that the δ distribution mean is equal to 0. The number of genes is the number of genes for which there were data for all tissues. We used the maximum expression level to calculate δ (See Methods). The second row gives the results of the Student's test. Rhythmic and non-rhythmic are the thresholds used to make the groups. Because the mouse cartilage dataset limited drastically the number of common genes, 3 or 4 tissues have been included in the calculation (liver, forebrain, tendon, +/- cartilage).

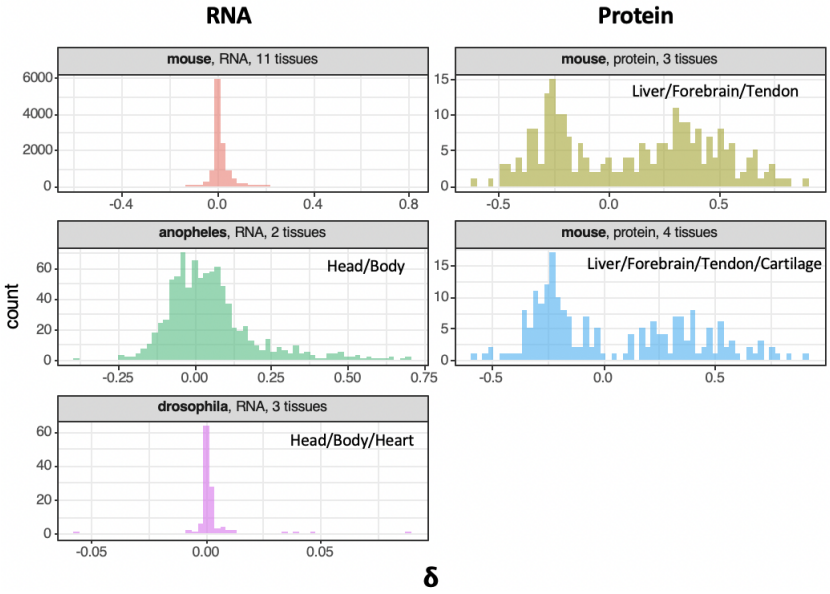


Figure 3.3: Distribution of the difference of expression levels between rhythmic tissues group and non-rhythmic tissues group (δ) obtained for every gene (see Methods). δ shows a bimodal distribution at protein level.

3.2.3 Rhythmic genes are tissue-specific

To clarify whether rhythmic genes tend to be tissue-specific highly expressed genes, we first analysed the relation between the number of tissues in which a gene is rhythmic and its tissue-specificity τ [67][68]. Partial correlations show that tissue-specific rhythmic genes are tissue-specific expressed genes (Table 3.2, Pearson correlation are given in Supp Table S7). This partial correlation seems to be stronger at the transcriptional level, although data was available in much less tissues at the protein level (only mouse forebrain, cartilage, and liver). For each tissue, we also found significant partial correlations between the *p*-values obtained from rhythm detection algorithm and the tissue-specificity τ obtained at the transcriptional level (for mouse, Supplementary Table S4 and S5). For each gene, its rhythmic RNA abundance in a given tissue is a function of its tissue-specificity level (Supplementary Table S4 and S5).

species	omics	nb.tissues	nb.genes	parameters
rhythmic	non.rhythmic	Pearson_cor	Pearson_t	Pearson_Signif
mouse	**transcript**	**11**	17736	tau VS nb rhythmic tissues
p<0.01	p>0.5	-0.37	-5.3e+01	< 2.2e-16
baboon	**transcript**	**9**	16816	tau VS nb rhythmic tissues
p<0.01	p>0.5	-0.13	-1.7e+01	< 2.2e-16
drosophila	**transcript**	**3**	8286	tau VS nb rhythmic tissues
p<0.01	p>0.5	-0.03	-2.3e+00	0.0229
mouse	**protein**	**3**	0	tau VS nb rhythmic tissues
p<0.01	p>0.5	0.01	2.7e-01	0.7902

Table 3.2: Spearman correlation test: τ vs. the number of tissues in which the gene is rhythmic. The number of genes is the number of genes for which there were data for all tissues. Rhythmic is the threshold used to consider the gene as rhythmic.

3.2.4 Lower cell-to-cell variability for genes with rhythmic transcripts

Because increasing translation for a fixed amount of mRNA can increase noise in final protein levels, we expect that genes with rhythmic proteins but constant mRNA levels be genes with noise-tolerant functions. To test this, we compared the noise distribution between rhythmic (p-value$\leq$cutoff.1) and non-rhythmic genes (p-value>cutoff.2). Because the protein coefficient of variation (= cell-to-cell variation in protein abundance) decreases with increasing transcription and decreasing translation for a given protein level [62], we used this as an estimation of the noise. Because the noise is negatively correlated to mean expression level [62][1], we preferentially used the method ($F*$) of Barroso et al. [1] which control the biases associated with the correlation between the expression mean (μ) and the variance (σ^2). It was the most efficient compared with other methods, see Supporting information. We applied it to different single-cell RNA data: *Arabidopsis thaliana* roots [69], and *Mus musculus* liver, lung, limb muscle, heart, and aorta [70] (Supplementary Table S1). Then, we assessed rhythmicity based on time-series datasets: RNAs and proteins in leaves of Arabidopsis and liver of Mouse from data used above; and RNAs in lung, kidney, muscle, heart, and aorta from the transcriptome time-series data of Zhang et al. [32] (Supplementary Table S1). For almost all these cases, we found trends of a lower intercellular variability for genes with rhythmic mRNA levels (Table 3.3), as well as for genes with rhythmic protein levels (Table 3.3). These results support the hypothesis that for genes which are rhythmic and that we predict to be noise-sensitive, there is rhythmic regulation at the transcriptional level. However, this noise has been estimated at the RNA level, whereas the functional impact of noise is expected to be at the protein level. Moreover, single-cells data used for the estimation of noise in Arabidopsis are from the root, while transcriptomic time-series data used to detect rhythmicity are from the leaves. Thus our power to test the hypothesis is limited by the data available at present.

3.2.5 Genes with rhythmic transcripts are more under selective constraint than non-rhythmic ones.

We tested next the hypothesis that protein evolutionary conservation (estimated by the dN/dS ratio) was equal between rhythmic (p-value$\leq$cutoff.1) versus non-rhythmic genes (p-value>cutoff.2). In all cases, in plants, vertebrates, and insects, we found that genes with rhythmic abundance of their mRNAs were significantly more conserved (Supplementary Table S3). We obtained similar results after controlling for gene expression bias (residuals in Supplementary Table S3). Interestingly, we didn't obtain such clear results for genes with

Noise comparison between rhythmic versus non-rhythmic genes

		Arabidopsis leaves	Mouse Liver	Mouse Lung	Mouse Muscle	Mouse Heart	Mouse Kidney	Mouse Aorta
RNA	**rhythmic cutoff | F***	$p \leq 0.001$ (148/5746 genes) 1.06	$p \leq 0.01$ (1847/3980 genes) 1.32	$p \leq 0.001$ (2899/13232 genes) 1.25	$p \leq 0.01$ (1420/12945 genes) 1.34	$p \leq 0.01$ (2089/14174 genes) 1.22	$p \leq 0.01$ (4067/11997 genes) 1.16	$p \leq 0.01$ (1459/13022 genes) 1.12
	non-rhythmic cutoff | F*	$p > 0.8$ (693/5746 genes) 1.05	$p > 0.2$ (721/3980 genes) 1.51	$p > 0.5$ (2618/13232 genes) 1.28	$p > 0.5$ (3942/12945 genes) 1.24	$p > 0.5$ (4643/14274 genes) 1.3	$p > 0.5$ (2029/11997 genes) 1.22	$p > 0.5$ (4684/13022 genes) 1.19
	t.test	t = 0.41765, df = 222.34, p-value = 0.6766 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.03823673 0.05880229	*** t = -2.0697, df = 1121.2, p-value = 0.03871 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.357314811 -0.009537716	t = -0.79918, df = 5245, p-value = 0.4242 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.09870270 0.04153398	t = 1.7194, df = 1873.3, p-value = 0.08571 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.01302986 0.19829610	*** t = -1.9735, df = 3621.4, p-value = 0.04852 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.1448638364 -0.0004736541	*** t = -2.7214, df = 4636.7, p-value = 0.006525 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.10446009 -0.01697686	*** t = -2.8496, df = 2298, p-value = 0.004417 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.10270758 -0.01897147
Protein	**rhythmic cutoff | F***	$p \leq 0.05$ (96/636 genes) 1.07	$p \leq 0.05$ (428/3980 genes) 1.31					
	non-rhythmic cutoff | F*	$p > 0.5$ (198/636 genes) 1.1	$p > 0.8$ (638/3980 genes) 1.32					
	t.test	t = -0.89787, df = 155.18, p-value = 0.3706 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.11957490 0.04484205	t = -0.035863, df = 827.98, p-value = 0.9714 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.2084844 0.2010026					
rhythmic Proteins ($p \leq 0.05$): rRNA VS nrRNA	**rhythmic cutoff | F***	$p \leq 0.05$ (44/91 genes) 1.01	$p \leq 0.05$ (301/428 genes) 1.28					
	non-rhythmic cutoff | F*	$p > 0.2$ (34/91 genes) 1.13	$p > 0.1$ (104/428 genes) 1.37					
	t.test	t = -1.6119, df = 75.376, p-value = 0.1112 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.28271235 0.02980919	t = -0.38301, df = 145.5, p-value = 0.7023 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.5434054 0.3669798					

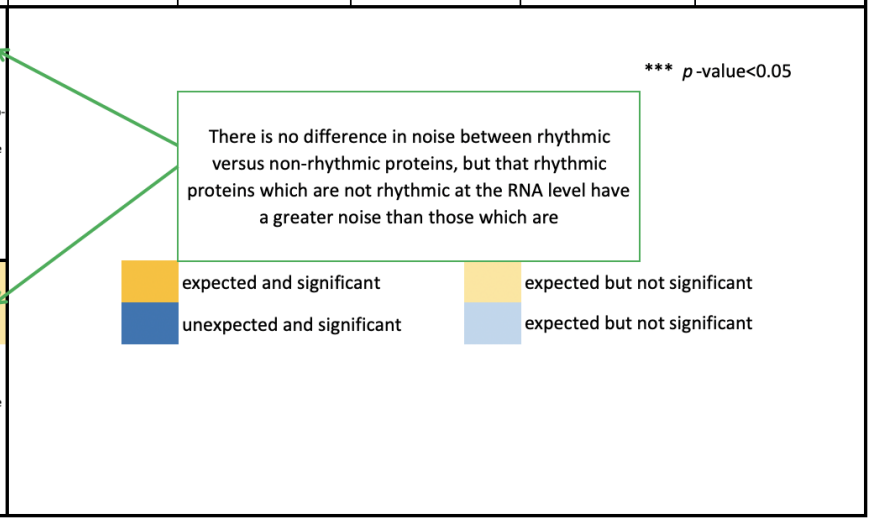


Table 3.3: Student's test testing the hypothesis that the noise is equal between rhythmic versus non-rhythmic transcripts, proteins, and between rhythmic versus non-rhythmic transcripts among rhythmic proteins. $F*$ is an estimation of the noise based on Barroso et al. method [1].

rhythmic proteins (Supplementary Table S3). In Arabidopsis, genes rhythmic at both steps were significantly less conserved than genes rhythmic at only one step. Results were unclear for the mouse. This suggests that rhythmicity at the transcriptional level might play a relevant role for the expression of important genes (defined as being under strong purifying selection).

3.2.6 Is there a subtle cost adaptation based on AA composition of proteins?

Among expression costs, the protein synthesis costs represent the constraint which is the most sensitive to natural selection [61][60], while costs at the genome and the transcriptome level are often below the threshold for efficient selection in multicellular species [60]. In addition, the degree of selective constraint caused by an increase in expression will be a function of effective population size (N_e) [61][60]. Indeed, the smaller the population size, the more important the effect of genetic drift .

The average cost for the cell of each protein depends partly on its composition in AA, whose biosynthesis costs (c_{AA}) can vary from 9.5∼P to 75.5∼P (according to data from [61]) (see Material and Methods). We compared the averaged AA synthesis costs per protein unit ($\bar{c}_{AA}$) between species with different population sizes. To do this, we took advantage of the estimatation of synonymous nucleotide diversity (π_S) in 76 non-

model animal species by [71]. This diversity of genetic polymorphism within a population is correlated with N_e and can be used as a proxy to compare N_e between species. Among the species with π_S estimates, whole genome protein sequence data was available for 13; 7 Chordata, 4 Arthropoda, 1 Mollusca, and 1 Nematoda (Supplementary Table S6). We calculated the averaged AA synthesis cost for each protein ($\bar{c}_{AA}$, Methods Formula (5)). Comparing these two parameters between species, π_S versus $\bar{c}_{AA}$, two clusters of animals seem to appear (Figure 3.4). We have not found a satisfactory explanation for these two clusters. Two range of orders of π_S are found between them (π_S<0.01 for all chordata species except for *Ciona intestinalis*). In both clusters, the $\bar{c}_{AA}$ appears to decrease with the synonymous nucleotide diversity (not shown). Within each cluster, the higher the effective number of individuals, the greater the selective constraint on cheap AA content of proteins.

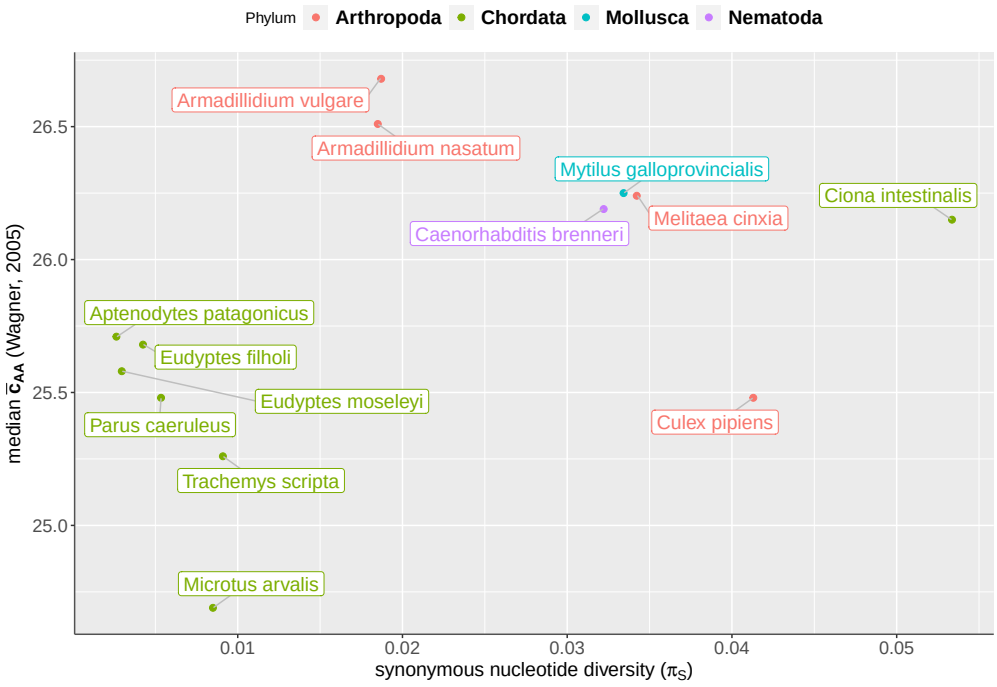


Figure 3.4: a) Placement of species as a function of their synonymous nucleotide diversity, π_S, and the median averaged AA synthesis cost of their proteins, median $\bar{c}_{AA}$.
Species π_S come from Source Romiguier et al. [71] (Supplementary Table S6)

3.3 Discussion

The endogenous generation of circadian rhythms is an anticipation strategy, which is optimized if the internal clock resonates with the external cycle [43][44]. Indeed, the autonomous nature of such mechanisms provides a clear advantage to the organism able to anticipate its environmental changes before they take place, allowing it to be "ready" before organisms who would not be endowed with such capacity. If we consider nycthemeral rhythms without considering their endogenous or exogenous nature, one can ask the question of the evolutionary origin of maintaining large cyclic biological systems, in term of adaptability in a given environment. The question to know whether cyclic systems are more complex than uniform ones is still complicated in biology (discussed in Discussion section 4.3.1). Indeed, they are very widely

found among living organisms, ubiquitous at every levels, highlighting their obvious necessity for adapted phenotypes. Thus, one can note that in all living organisms there are periodic regulatory systems as well as permanent regulatory systems.

During evolution, living systems have made trade-offs between energy efficiency and noise reduction [62][59] whose degree of sensitivity depends on the functions and on the selective pressure of the environment. Indeed, each gene has its own sensitivity to intrinsic noise (expression sensitivity), i.e. the optimality of its function (sometimes called fitness-function) is more or less sensitive to the protein level deviations away from its optimum expression level (= expression level which maximizes the organismal fitness) [72]. Schmiedel et al. constructed fitness landscapes for each gene representing the theoretical effect of the protein level variations and cell-to-cell variability on the fitness (assessed by the growth rate relative to wild-type in yeast) [72]. However, these fitness landscapes are assessed for presumed steady-state protein levels measured at a given time-point, which should be representative only for non-rhythmic proteins. The case of rhythmically regulated genes is more complex and the fitness effects of expression variability is more problematic to understand inside the concept of optimality in gene expression, which presupposes constant expressions (See Discussion section 4.2.1). Indeed, the cyclic nature of biological systems can be seen as a robust oscillatory system giving a framework for complex structures based on temporal organisation of regulatory circuits (See Discussion section 4.2.3). It's this component of temporal dynamics which provide robustness faced with environmental changes, optimizing the adaptation of the individual to its environment, thus its survival and fitness.
Thus, the dynamics of expression of a gene must depend on a trade-off between the degree of requirement of the function, its sensitivity to protein level variation from the optimum level, and the cost. Note that the parameters of this trade-off could have different effects over time (at daylong time-scale) (See Discussion 4.2.1). For instance, the oscillation of negative feedback loop systems has been shown to be maintained thanks to noise oscillations whose source seem to be low-frequency fluctuations of cell-to-cell variability of protein production rates - rather than other parameters such as degradation rates - as shown for the p53-Mdm2 system (ultradian rhythm) which amplifies the frequency component of the noise in the vicinity of its natural frequency [73]. Thus, the stochasticity would also be beneficial by maintaining oscillations of an oscillator system that would be damped otherwise (from deterministic simulations).

Based on proteomic and transcriptomic data in several species, we show that the average expression of rhythmic proteins is higher than others and explains their higher expression costs observed. This supports the hypothesis according to which costlier genes are preferentially under rhythmic regulation. Support is even stronger if we assume that in the absence of rhythmicity they would be constantly at their maximum of expression, which is assumed to be the optimal expression level. Thus, their rhythmic regulation allows minimizing the overall costs, over at least day-night time-scales. Furthermore, rhythmically expressed genes are genes whose function requires a higher expression level in the tissue in which they are rhythmic. Cyclic systems might play a role in the optimization of expression precision for a given period of time.

3.4 Materials Methods

3.4.1 Datasets

Datasets details are available in Supplementary Table S1.

Mus musculus

Mouse liver transcriptomic and proteomic time-series datasets come from Mauvoisin et al. [26]. Original protein counts dataset was downloaded from ProteomeXchange (PXD001211) file Combined_WT, and cleaned from data with multiple Uniprot.IDs affectations between raw peptides data. Time-serie transcriptomic data was downloaded from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) accession (GSE33726) [24] Multi-tissues time-serie transcriptomic data used for tissue-specificity expression comparisons are microarray data from Zhang et al. [32] downloaded from the NCBI GEO accession GSE54652 (see Materials of Laloum and Robinson-Rechavi [58] for more details). Tissues analysed are: adrenal gland, aorta, brain stem, brown adipose, cerebellum, heart, kidney, liver, lung, muscle, and white adipose. Multi-tissues time-serie proteomic data come from: Mauvoisin et al. [26] for the liver, Noya et al. [74] for the forebrain, Chang et al. [75] for the tendon, and Dudek et al. [76] for the cartilage. Finally, single-cell data of thousand of cells in organs for which we had time-series datasets, i.e. liver, lung, kidney, muscle, aorta, and heart, were downloaded from figshare using R objects from FACS single-cell datasets [70].

Arabidopsis thaliana

Leaves time-series proteomic data are the Dataset I from Krahmer et al. [77], cleaned from data with multiple protein identifications. Leaves time-series transcriptomic data were downloaded from the NCBI GEO accession (GSE3416) [78]. Single-cell data of twenty root cells were downloaded from the NCBI GEO accession (GSE46226) [69].

Ostreococcus tauri

Unicellular alga proteomics time-series dataset (normalized abundances) come from Noordally et al. [79]. Transcriptomic time-series dataset come from Monnier et al. [80], was downloaded from the NCBI GEO accession (GSE16422) and was cleaned for genes with too much missing values (more than seven).

Synechococcus elongatus (PCC 7942)

Unicellular cyanobacterium proteomics time-series dataset come from Guerreiro et al. [81]. Transcriptomic time-series dataset come from Ito et al. [82] and was recovered from Guerreiro et al. [81].

Drosophila melanogaster

Transcriptomic time-series datasets for the body, the head, and the heart, come from Gill et al. [83] and were downloaded from the NCBI GEO accession (GSE64108) (see Materials of Laloum and Robinson-Rechavi [58] for more details).

Papio anubis [Olive baboon]

Multi-tissues time-serie transcriptomic data used for tissue-specificity expression comparisons are RNA-seq data from Mure et al. [84] (see Materials of Laloum and Robinson-Rechavi [58] for more details).

3.4.2 Pre-processing

For each time-series dataset, only protein coding genes were kept. ProbIDs assigned to several proteins were removed. Probesets were cross-referenced to best-matching gene symbols by using either Ensembl BioMart software [85], or UniProt [86].

3.4.3 Rhythm detection

To increase power of rhythm detection [58], we considered biological replicates as new cycles when it was possible. We used `GeneCycle` R package (version 1.1.4) [87] available from CRAN and used the robust.spectrum function developped by [88] - with parameters *periodicity.time=24* and *algorithm="regression"* - that computes a robust rank-based estimate of the periodogram/correlogram and that we improved with `try-catch` function to avoid error of dimension with MM-estimation method. When the *p*-values distribution obtained did not correspond to the expected distribution - skewed towards low *p*-values because of the presence of rhythmic genes - we used the results obtained by the rhythm detection method used by the original paper from where the data came after checking they presented a classic skewed *p*-values distribution. Finally, for each gene or protein having several data (ProbIDs or transcripts), we combined *p*-values by Brown's method using the `EmpiricalBrownsMethod` R package (See Supporting information). Thus, for each dataset, we obtained a unique rhythm *p*-value per gene or per protein. Low-amplitude or maybe less-accurate measurements implied that it was statistically more challenging to identify rhythms in proteomics data. That is why, in general, we used lower stringency for proteins.

3.4.4 Consistent gene expression levels

Tissue-specific mRNA or protein abundances were the average or the maximum level of the n time-points j such as for gene i:

$$ {}_{max}N_i = \frac{{}_{max1}N_{i,j} + {}_{max2}N_{i,j}}{2} \quad \text{with} \quad {}_{max1}N_{i,j} \neq {}_{max2}N_{i,j} \tag{2} $$

$$ {}_{mean}N_i = \frac{\sum_{j=1}^{n} N_{i,j}}{n} \quad \text{with} \quad n\text{: the number of time-points} \tag{3} $$

$$ \tag{4} $$

N_{pi}: for the abundance of the protein i

N_{RNAi}: for the abundance of the transcript i

3.4.5 Expression costs

Energetic costs of each AA (unit: high-energy phosphate bonds per molecule) come from Akashi and Gojobori [89], or Wagner [61] which are linearly correlated (Supplementary File S1 Figure S2). The averaged AA synthesis cost of one protein of lenght *Lp* is:

$$ \bar{c}_{AA} = \frac{\sum_{j=1}^{L_p} c_{AAj}}{L_p} \tag{5} $$

Protein sequences come from FASTA files downloaded from EnsemblPlants [85] or UniProt [86] for: Proteome UP000002717 for *Synechococcus elongatus* PCC 7942, Uniprot Reviewed [Swiss-Prot] for *Mus musculus*, Proteome UP000009170 for *Ostreococcus tauri*. For species used in Romiguier et al. [71] paper, we downloaded protein sequences FASTA files from EnsemblMetazoa for *Caenorhabditis brenneri* and from NCBI GEO database [90] for *Aptenodytes patagonicus, Armadillidium nasatum, Armadillidium vulgare, Ciona intestinalis, Culex pipiens, Eudyptes filholi, Eudyptes moseleyi, Melitaea cinxia, Microtus arvalis, Mytilus galloprovincialis, Parus caeruleus*, and *Trachemys scripta*.
Main results use Wagner [61] AA costs data and we provide supplementary results using Akashi and Gojobori [89] AA costs data (Supplementary File S1 Figure S3).

3.4.6 Multi-tissues analysis

To obtain comparable expressions levels between different tissues or datasets, we normalized expression values by Z-score transformation such as in the dataset of n genes, the mean expression of the gene i becomes:

$$ {}_{mean}Z_i = \frac{{}_{mean}N_i - \bar{N}}{{}_{max}Z.\sigma} \quad \text{with} \quad {}_{mean}N_i\text{: the average expression level of gene } i \text{ (formula (3))} \tag{6} $$

$$ \bar{N} = \frac{\sum_{i=1}^{n} N_i}{n} $$

σ: the standard deviation of ${}_{mean}N_i$

${}_{max}Z$: the maximal value of the ${}_{mean}Z$-scores

n: the number of genes

and the maximal expression of the gene i becomes:

$$ {}_{max}Z_i = \frac{{}_{max}N_i - \bar{N}}{{}_{max}Z.\sigma} \quad \text{with} \quad {}_{max}N_i\text{: the maximal expression level of gene } i \text{ (formula (2))} \tag{7} $$

${}_{max}Z$: the maximal value of the ${}_{max}Z$-scores

σ: the standard deviation of ${}_{max}N_i$

$$ Z_i \in [0,1] $$

To compare the expression levels between the set of n_r rhythmic tissues and the set of $n_{\bar{r}}$ non-rhythmic tissues, we estimated the difference (δ) of expression levels between these two groups for each gene i such as:

$$ \delta_i = \frac{\sum_{j_r=1}^{n_r} ({}_{max}Z_{i,j_r})}{n_r} - \frac{\sum_{j_{\bar{r}}=1}^{n_{\bar{r}}} ({}_{max}Z_{i,j_{\bar{r}}})}{n_{\bar{r}}} \tag{8} $$

with n_r: the number of tissues in which gene i is rhythmic ($p \leq 0.01$ or 0.05)
$n_{\bar{r}}$: the number of tissues in which gene i is not rhythmic ($p>0.1$ or 0.5)
${}_{max}Z_{i,j_r}$: the maximum expression level Z-score normalized of gene i in the tissue j_r in which it is rhythmic
${}_{max}Z_{i,j_{\bar{r}}}$: idem for non-rhythmic tissue
${}_{max}Z_i \in [0,1]$, defined as formula (7)

Finally, we analysed the distribution of δ_i and generated a Student's t-test to compare with an expected theoretical mean of 0.

3.4.7 Tissue-specificity of gene expression

To calculate a tissue-specificity τ for each gene i, we log-transformed the averaged gene expression and followed Kryuchkova-Mostacci and Robinson-Rechavi instructions [68] to make expression values manageable. Thus, among the n tissues, the tissue-specificity of gene i is:

$$\tau_i = \frac{\sum_{j=1}^{n}(1-\hat{N}_{i,j})}{n-1} \quad \text{with} \quad n\text{: the number of tissues} \tag{9}$$

$$\hat{N}_{i,j} = \frac{log({}_{mean}N_{i,j})}{\max_{1\leq j\leq n}(log({}_{mean}N_{i,j}))}$$

$\max\limits_{1\leq j\leq n}(log({}_{mean}N_{i,j}))$ is the maximal expression level of gene i among the n tissues

$\hat{N}_{i,j} \in [0,1]$

The tissue-specificity formula was described by Yanai et al. [67]. Finally, we performed linear regressions to analyse the respective and the interaction influences of gene expression level and tissue-specificity into the rhythmicity. For mouse [32], we used RNA-seq data to estimate the mean expression used for the calculation of τ, and used rhythm p-values obtained from microarray dataset since there was more time-points in the microarray time-series (as discussed in benchmark paper). Figure S4 (File S1) shows the distributions of τ obtained.

3.4.8 Gene expression noise quantification

We estimated a unique expression noise per gene by taking advantages of *Arabidopsis thaliana* roots [69] and Mus musculus liver, lung, kidney, muscle, aorta, and heart [70] (from figshare using R objects of FACS single-cell datasets) single-cell RNAseq data (see Materials). For Arabidopsis, we obtained gene expression for twenty root quiescent centre (QC) cells for which a total of 14,084 genes had non-zero expression in at least one of the 20 single cells. For mouse, we obtained gene expression for a number of cells ranging from 180 to 3772 cells, for which a total of genes with non-zero expression was around 19,000 genes. As in Barroso et al. [1], we kept only genes whose expression level satisfied log[FPKM + 1] >1.5 in at least one single cell. We calculated the stochastic gene expression ($F*$) defined by Barroso et al. [1] as a measure for

gene expression noise (cell-to-cell variability) designed to control the biases associated with the correlation between the expression mean (μ) and the variance (σ^2) by using the second lowest degree polynomial regression which decorrelate them. We tested different degrees of the polynomial regression to estimate $\log(\sigma^2)$ in the calculation of $F*$ and measure the correlation based on Kendall's rank. For Arabidopsis, a polynomial regression of the first degree was enough (Kendall's rank correlation test between $F*$ and μ was =0.0016, p-value=0.7805), fourth or fifth degree for the mouse tissues.

3.4.9 dN/dS analysis

dN/dS data have been downloaded from Ensembl BioMart Archive 99 [85]. The homologous species used are: *Mus musculus - Rattus norvegicus*; *Arabidopsis thaliana - Arabidopsis lyrata*; and *Anopheles gambiae - Aedes aegyti*. We first tested the hypothesis that rhythmic (p-value$\leq$cutoff.1) versus non-rhythmic genes (p-value$>$cutoff.2) have equal ratio of non-synonymous to synonymous substitutions. Then, since rhythmic genes are now known to be enriched in highly expressed genes and because highly expressed genes are under purifying selection, we controlled for the effect of gene expression on dN/dS ratio by testing the hypothesis that rhythmic (p-value$\leq$cutoff.1) versus non-rhythmic (p-value$>$cutoff.2) genes among residuals of the linear regression fitting gene expression level and dN/dS ratio.
In R code it gives:

```
> lmTest = lm(log(dNdS)~log(gene.expr.level), data)
> rhythmic.residuals = lmTest$residuals[data$rhythm.pvalue≤cutoff.1]
> nonrhythmic.residuals = lmTest$residuals[data$rhythm.pvalue>cutoff.2]
> t.test(rhythmic.residuals, nonrhythmic.residuals)
```

Plots have been generated using ggplot2 package (version 3.3.2) run in R version 4.0.2.

3.5 Additional files

3.5.1 Supplementary Tables

	ANIMALS									PLANTS					BACTERIA	
	VERTEBRATES							INSECTS		multicellular			unicellular alga		unicellular cyanobacterium	
Species	*Mus musculus* (*Ne* ~ 2.5–12 x10^4)						*Rattus norvegicus* (*Ne* ~ 12.4 x10^4)	*Anopheles gambiae* (*Ne* ~ 62 x10^6)	*Aedes aegypti*	*Arabidopsis thaliana* (*Ne* ~ 127 x10^6)			*Ostreococcus tauri* (*Ne* ~ 6-29 x10^6)		*Synechococcus elongatus* (*Ne* ~ 142 x10^9)	
tissue	liver		forebrain	tendon	cartilage	lung, kidney, muscle, aorta, adrenal gland, brain stem, brown adipose, cerebellum, heart, white adipose	liver, lung, kidney, muscle, aorta, heart	lung	head, body	head	Leaves (Rosettes without roots and shoots)	Leaves (Full rosettes without roots)	root	*unicellular alga*		*unicellular cyanobacterium*	
profile	**proteome**	**transcriptome**	**proteome**	**proteome**	**proteome**	**transcriptome**	**single-cell RNAseq**	**transcriptome**	**transcriptome**	**transcriptome**	**proteome**	**transcriptome**	**single-cell RNAseq**	**proteome**	**transcriptome**	**proteome**	**transcriptome**
profiling technique	SILAC	array (totRNA)	LC-MS/MS	LC-MS/MS	LC-MS/MS	microarray	RNAseq FACS	microarray	microarray	microarray	IGEPAL-TCA	microarray	RNAseq	Bradford Assay	microarray	TMT labeling	microarray
Experimental conditions – Total period	24h	48h	24h	48h	48h	48h	_	72h	48h	48h	24h	24h	_	24h	24h	48h	48h
Experimental conditions – Experimental conditions	ZT*	ZT*	LD	LD	LD	CT	_	LD (12:12 LD cycles)	LD (12:12 LD cycles)	LD (12:12 LD cycles)	LL (entrained with 22 LD cycles and then transferred to LL)	LD (entrained with 5 weaks of 12:12 LD cycles) (flowering had not commenced)	_	LD (entrained under 12:12 LD cycles)	LD (entrained under 12:12 LD cycles)	LD (12:12 LD cycles)	LL (entrained with 2 LD cycles and then transferred to LL)
Experimental conditions – Sampling	3h	2h	4h	3h	4h	2h	_	2h	4h	4h	4h	4h	_	4h	3h	from 1h to 3h	4h
Experimental conditions – Number of biological replicates	x2	x2	x2-4	x4	x6	x3	~3000 cells per organ	x1	x2	x1	x5 (each containing 8 rosettes)	x3 (3 independant biological experiments performed at 2-month intervals) (5 samples of 3 rosettes where harvested at each time-point)	20 cells from quiescent center (4 technical replicates)	x5 (cells pellet obtained from 135ml of culture per sample replicate, each from a separate culture vessel)	x3 (cells pellet obtained from one-milliliter cell samples of culture, each from three independant experiments)	x1 (cyanobacteria pellet obtained from 40ml of culture)	x2 (2 independent experiments)
considered as (1 cycle = 24 hrs)	x1 every 3h over **2 cycles**	x1 every 2h over **2 cycles**	x1 every 4h over ~**3 cycles**	x1 every 3h over 2 **cycles**	x1 every 4h over 1 **cycle**	x1 every 2h over **2 cycles**	_	x1 every 2h over 3 **cycles** (ZT00 is the average expression value of ZT23.75 and ZT00.25)	x1 every 4h over 4 **cycles**	x1 every 4h over 2 **cycles**	x1 every 4h over **5 cycles**	x1 every 4h over **3 cycles**	_	x1 every 4h over **5 cycles**	x1 every 3h over **3 cycles**	x1 every 1h-3h over **2 cycles**	x1 every 4h over **4 cycles**
References	Mauvoisin, 2013	Jouffe, 2013	Noya, 2019	Chang, 2020	Dudek, 2019	Zhang, 2014	Tabula Muris Consortium, 2018	Sukumaran, 2011	Rund, 2011	Leming, 2014	Krahmer, 2019	Bläsing, 2005	Efroni, 2015	Noordally, 2018	Monnier 2010	Guerreiro, 2014	Ito, 2009
Dataset	Dataset S1 / PXD001211 (ProteomeXcha	GSE33726	Table S7	PXD013450	Supplementary Table 1	GSE54652	figshare: facs R_objects	GSE25612	GSE22585	GSE60496	Dataset I	GSE3416	GSE46226	Supplementary Table 1	GSE16422	Supp Table .xlsx Guerreiro 2014	Supp Table .xlsx Guerreiro 2014 / GSE14225

ZT: Zeitgeber
LD: light/dark
LL: constant light

proteomic
transcriptomic
single-cell

Mouse image credit: Anthony Caravaggi ,license CC BY-NC-SA 3.0

*maintained under standard animal housing conditions with free access to food and water in 12-h light–dark(LD) cycles.
Then, animals were fed only at night starting 4 days before the experiment to control for geno-type-dependent feeding rhythms.

Table S1 Time-series and Single-Cell datasets.

Energetic costs per amino acid (unit: high-energy phosphate bonds per molecule ~P)			
Amino Acid	cost_Akashi	cost_Wagner	
A	11.7	14.5	Akashi and Gojobori: "Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis", Hiroshi Akashi and Takashi Gojobori, 2002, in E. coli
R	27.3	20.5	
N	14.7	18.5	
D	12.7	15.5	
C	24.7	26.5	
Q	16.3	10.5	
E	15.3	9.5	
G	11.7	14.5	
H	38.3	29	Akashi: also see: "The bioenergetic costs of a gene", Michael Lynch and Georgi K. Marinov, 2015
I	32.3	38	
L	27.3	37	
K	30.3	36	
M	34.3	36.5	Wagner: "Energy Constraints on the Evolution of Gene Expression", Andreas Wagner, 2005, from E. coli
F	52	61	
P	20.3	14.5	
S	11.7	14.5	
T	18.7	21.5	
W	74.3	75.5	
Y	50	59	
V	23.3	29	

Table S2 Energetic costs estimated for each amino acid.

REFs:

Hiroshi Akashi and Takashi Gojobori. Metabolic efficiency and amino acid composition in the proteomes of escherichia coli and bacillus subtilis. Proceedings of the National Academy of Sciences, 99(6):3695–3700, 2002. ISSN 0027-8424. doi: 10.1073/pnas.062526999

Michael Lynch and Georgi K. Marinov. The bioenergetic costs of a gene. Proceedings of the National Academy of Sciences, 112(51):15690–15695, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1514974112

Andreas Wagner. Energy Constraints on the Evolution of Gene Expression. Molecular Biology and Evolution, 22(6):1365–1374, 03 2005. ISSN 0737-4038. doi: 10.1093/molbev/msi126

		PLANT			VERTEBRATES																	
		Arabidopsis leaves			**Mouse Liver**			**Mouse Lung**			**Mouse Muscle**			**Mouse Heart**			**Mouse Kidney**			**Mouse Aorta**		
RNA	**rhythmic cutoff | dN/dS**	$p \leq 0.01$ (1854/21092 genes)	0.213	residuals (~expression level)	$p \leq 0.001$ (1145/3777 genes)	0.1459	residuals (~expression level)	$p \leq 0.001$	0.1731	residuals (~expression level)	$p \leq 0.01$	0.2111	residuals (~expression level)	$p \leq 0.01$	0.1636	residuals (~expression level)	$p \leq 0.01$	0.1768	residuals (~expression level)	$p \leq 0.01$	0.1631	residuals (~expression level)
	non-rhythmic cutoff | dN/dS	$p > 0.8$ (3102/21092 genes)	0.307		$p > 0.2$ (872/3777 genes)	0.2417		$p > 0.8$	0.3035		$p > 0.4$	0.2458		$p > 0.4$	0.2614		$p > 0.4$	0.2831		$p > 0.4$	0.2624	
	t.test	*** t = -14.486, df = 4953.9, p-value < 2.2e-16 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.10646727 -0.08108515		t = -1.6658, df = 3899.4, p-value = 0.09582 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.086870133 0.007060047 sample estimates: mean of x mean of y -0.01666033 0.02324471	*** t = -9.6855, df = 1392.2, p-value < 2.2e-16 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.11513071 -0.07634907		t = -8.8751, df = 1872.1, p-value < 2.2e-16 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.5545160 -0.3537954 sample estimates: mean of x mean of y -0.2281692 0.2259865	*** t = -17, df = 2483.7, p-value < 2.2e-16 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.1454142 -0.1153368		t = -1.0376, df = 5998.7, p-value = 0.2995 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.06718169 0.02067787 sample estimates: mean of x mean of y 0.01639971 0.03965162	*** t = -5.7077, df = 2795.4, p-value = 1.265e-08 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.04669015 -0.02281323		t = 1.7159, df = 2536.7, p-value = 0.0863 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.006871158 0.103128405	*** t = -21.064, df = 4987.5, p-value < 2.2e-16 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.10683632 -0.08864291		t = -2.9348, df = 3470.6, p-value = 0.00336 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.12150958 -0.02417922 sample estimates: mean of x mean of	*** t = -23.681, df = 10387, p-value < 2.2e-16 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.11512955 -0.09752694		t = -4.3601, df = 9770.1, p-value = 1.314e-05 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.12831352 -0.04872158 sample estimates: mean of x mean of	*** t = -19.197, df = 3010.4, p-value < 2.2e-16 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.10938070 -0.08910761		t = -2.7058, df = 2266, p-value = 0.006866 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.13044960 -0.02081796 sample estimates: mean of x mean of y -0.05130777 0.02432601
Protein	**rhythmic cutoff | dN/dS**	$p \leq 0.01$ (107/1610 genes)	0.172	residuals (~expression level)	$p \leq 0.01$ (142/3777 genes)	0.1737	residuals (~expression level)															
	non-rhythmic cutoff | dN/dS	$p > 0.8$ (148/1610 genes)	0.193		$p > 0.7$ (888/3777 genes)	0.1659																
	t.test	t = -0.84799, df = 240.5, p-value = 0.3973 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.07152045 0.02847479		t = 0.032465, df = 240.2, p-value = 0.9741 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.2218513 0.2292862 sample estimates: mean of x mean of y -0.01124218 -0.01495965	t = 0.53426, df = 203.3, p-value = 0.5937 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.02109180 0.03677047		t = 1.2088, df = 113.61, p-value = 0.2292 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.09190226 0.37963468 sample estimates: mean of x mean of y 0.10218469 -0.04168152														
rhythmic Protein ($p \leq 0.05$): rRNA VS nrRNA	**rhythmic cutoff | dN/dS**	$p \leq 0.01$ (51/263 genes)	0.271		$p \leq 0.01$ (241/451 genes)	0.1633																
	non-rhythmic cutoff | dN/dS	$p > 0.3$ (95/263 genes)	0.148		$p > 0.2$ (241/451 genes)	0.2045																
	t.test	** t = 2.656, df = 59.303, p-value = 0.01014 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: 0.03039419 0.21601111			t = -1.5517, df = 78.34, p-value = 0.1248 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.09405767 0.01165610																
rhythmic RNA ($p \leq 0.01$): rProtein VS nrProtein	**rhythmic cutoff | dN/dS**	$p \leq 0.05$ (51/215 genes)	0.271		$p \leq 0.01$ (241/1989 genes)	0.1633																
	non-rhythmic cutoff | dN/dS	$p > 0.3$ (101/215 genes)	0.18		$p > 0.2$ (1145/1989 genes)	0.157																
	t.test	t = 1.9368, df = 62.078, p-value = 0.05732 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.002914993 0.184779631			t = 0.59515, df = 418.6, p-value = 0.5521 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.01455001 0.02718707																

expected and significant — expected but not significant — *** p-value<0.01

unexpected and significant — expected but not significant — ** p-value<0.05

Table S3$_1$

VERTEBRATES							INSECTS											
Mouse Tendon			**Mouse Forebrain**			**Mouse Cartilage**			**Rat Lung**			**Anopheles Head**			**Anopheles Body**			**Aedes Head**		
									$p \leq 0.001$	0.1626	residuals (~expression level)	$p \leq 0.01$	0.0259	residuals (~expression level)	$p \leq 0.01$	0.0277	residuals (~expression level)	$p \leq 0.01$	0.02423	residuals (~expression level)
									$p > 0.4$	0.2085		$p > 0.4$	0.0369		$p > 0.4$	0.036		$p > 0.4$	0.031	
									*** t = -8.7227, df = 3264.5, p-value < 2.2e-16 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.05623448 - 0.03559330		t = -0.86331, df = 3607.7, p-value = 0.388 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.08101977 0.03148251 sample estimates: mean of x mean of y -0.027167131 -0.002398498	*** t = -2.2041, df = 4286.9, p-value = 0.02757 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.020797081 - 0.001216195		t = -3.983, df = 3832.1, p-value = 6.93e-05 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.21328370 - 0.07257498 sample estimates: mean of x mean of	t = -1.361, df = 3383.4, p-value = 0.1736 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.020256164 0.003656927		t = -0.82302, df = 3902.7, p-value = 0.4105 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.10433354 0.04263738 sample estimates:	*** t = -2.6578, df = 368.27, p-value = 0.008209 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.011692698 - 0.001748094		t = -1.2772, df = 335.34, p-value = 0.2024 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.24262617 0.0515926 sample estimates: mean of x mean of y -0.08360014 0.01191664
$p \leq 0.05$	0.1157	residuals (~expression level)	$p \leq 0.05$	0.117	residuals (~expression level)	$p \leq 0.05$	0.0987	residuals (~expression level)												
$p > 0.4$	0.1426		$p > 0.4$	0.112		$p > 0.4$	0.1342													
t = -1.68, df = 178.8, p-value = 0.09471 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.058442554 0.004692579		t = 0.76802, df = 113.05, p-value = 0.4441 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.1738606 0.3939943 sample estimates: mean of x mean of y 0.09712011 -0.01294672	t = 0.60366, df = 1066, p-value = 0.5462 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.009845866 0.018595764		t = -0.95786, df = 503.51, p-value = 0.3386 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.22510631 0.07754943 sample estimates: mean of x mean of	*** t = -3.2737, df = 517.85, p-value = 0.001132 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.05682665 - 0.01420198		t = -2.7655, df = 374.29, p-value = 0.005965 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.48988684 - 0.08273798 sample estimates: mean of x mean of											

Table S3$_2$ Student's test testing the hypothesis that the dN/dS distributions are equal among rhythmic versus non-rhythmic genes. Residuals correspond to

Student's test testing the hypothesis that the dN/dS are equal among rhythmic versus non-rhythmic genes in residuals controlled for the gene expression effect (See Methods).

species	omics	technique	tissue	nb.tissues	nb.genes	parameters	terms	Estimate	Std. Error	t value	Signif	Adj R-squared	F-statistic
mouse	transcripts	microarray	cerebellum	11	17736	-log10(rhythm.p-value) ~ log(mean.RNA.level) + tau	(Intercept)	0.043	0.032	1.356	0.1751	0.074	658.134
							log(mean.RNA.level)	0.109	0.003	32.551	< 2.2e-16	0.074	658.134
							tau.mean	0.422	0.043	9.912	< 2.2e-16	0.074	658.134
			brown_adipose	11	17736	-log10(rhythm.p-value) ~ log(mean.RNA.level) + tau	(Intercept)	-0.696	0.058	-11.978	< 2.2e-16	0.136	1251.539
							log(mean.RNA.level)	0.238	0.006	39.608	< 2.2e-16	0.136	1251.539
							tau.mean	1.353	0.08	16.842	< 2.2e-16	0.136	1251.539
			muscle	11	17736	-log10(rhythm.p-value) ~ log(mean.RNA.level) + tau	(Intercept)	0.282	0.04	6.968	3.30E-12	0.051	426.648
							log(mean.RNA.level)	0.091	0.004	21.648	< 2.2e-16	0.051	426.648
							tau.mean	0.264	0.056	4.7	2.60E-06	0.051	426.648
			adrenal_gland	11	17736	-log10(rhythm.p-value) ~ log(mean.RNA.level) + tau	(Intercept)	-0.348	0.049	-7.172	7.70E-13	0.077	687.443
							log(mean.RNA.level)	0.158	0.005	32.978	< 2.2e-16	0.077	687.443
							tau.mean	0.875	0.058	15.105	< 2.2e-16	0.077	687.443
			brain_stem	11	17736	-log10(rhythm.p-value) ~ log(mean.RNA.level) + tau	(Intercept)	0.156	0.033	4.71	2.50E-06	0.047	413.796
							log(mean.RNA.level)	0.082	0.003	24.974	< 2.2e-16	0.047	413.796
							tau.mean	0.221	0.039	5.669	1.50E-08	0.047	413.796
			liver	11	17736	-log10(rhythm.p-value) ~ log(mean.RNA.level) + tau	(Intercept)	-0.028	0.051	-0.547	0.5842	0.169	1544.447
							log(mean.RNA.level)	0.241	0.005	44.99	< 2.2e-16	0.169	1544.447
							tau.mean	0.422	0.08	5.299	1.20E-07	0.169	1544.447
			lung	11	17736	-log10(rhythm.p-value) ~ log(mean.RNA.level) + tau	(Intercept)	-0.902	0.063	-14.319	< 2.2e-16	0.128	1221.208
							log(mean.RNA.level)	0.265	0.006	42.921	< 2.2e-16	0.128	1221.208
							tau.mean	1.378	0.074	18.551	< 2.2e-16	0.128	1221.208
			kidney	11	17736	-log10(rhythm.p-value) ~ log(mean.RNA.level) + tau	(Intercept)	-0.165	0.055	-2.977	2.90E-03	0.132	1237.393
							log(mean.RNA.level)	0.234	0.006	40.267	< 2.2e-16	0.132	1237.393
							tau.mean	0.745	0.074	10.116	< 2.2e-16	0.132	1237.393
			heart	11	17736	-log10(rhythm.p-value) ~ log(mean.RNA.level) + tau	(Intercept)	-0.299	0.053	-5.621	1.90E-08	0.108	970.371
							log(mean.RNA.level)	0.186	0.006	33.764	< 2.2e-16	0.108	970.371
							tau.mean	0.819	0.074	11.13	< 2.2e-16	0.108	970.371
			aorta	11	17736	-log10(rhythm.p-value) ~ log(mean.RNA.level) + tau	(Intercept)	-0.047	0.05	-0.944	0.3454	0.106	966.891
							log(mean.RNA.level)	0.144	0.005	27.507	< 2.2e-16	0.106	966.891
							tau.mean	0.231	0.065	3.585	3.40E-04	0.106	966.891
			white_adipose	11	17736	-log10(rhythm.p-value) ~ log(mean.RNA.level) + tau	(Intercept)	0.596	0.039	15.194	< 2.2e-16	0.042	371.716
							log(mean.RNA.level)	0.049	0.004	11.753	< 2.2e-16	0.042	371.716
							tau.mean	-0.317	0.046	-6.87	6.60E-12	0.042	371.716

Table S4: T-tests testing the independence between rhythm p-values obtained in each tissue and the tissue-specificity τ

species	omics	technique	tissue	nb.tissues	nb.genes	parameters	terms	Estimate	Std. Error	t value	Signif	Adj R-squared	F-statistic
mouse	transcripts	microarray	cerebellum	11	17736	-log10(rhythm.p-value) ~ log(mean.RNA.level) * tau	(Intercept)	0.214	0.044	4.882	1.10E-06	0.076	450.244
							log(mean.RNA.level)	0.084	0.006	14.982	< 2.2e-16	0.076	450.244
							tau.mean	0.082	0.074	1.111	0.2665	0.076	450.244
							log(mean.RNA.level):tau.mean	0.063	0.011	5.655	1.60E-08	0.076	450.244
			brown_adipose	11	17736	-log10(rhythm.p-value) ~ log(mean.RNA.level) * tau	(Intercept)	-0.546	0.065	-8.371	< 2.2e-16	0.138	843.959
							log(mean.RNA.level)	0.213	0.008	27.101	< 2.2e-16	0.138	843.959
							tau.mean	1.002	0.107	9.411	< 2.2e-16	0.138	843.959
							log(mean.RNA.level):tau.mean	0.085	0.017	5	5.80E-07	0.138	843.959
			muscle	11	17736	-log10(rhythm.p-value) ~ log(mean.RNA.level) * tau	(Intercept)	0.519	0.049	10.594	< 2.2e-16	0.055	310.046
							log(mean.RNA.level)	0.054	0.006	8.899	< 2.2e-16	0.055	310.046
							tau.mean	-0.267	0.084	-3.194	1.40E-03	0.055	310.046
							log(mean.RNA.level):tau.mean	0.113	0.013	8.543	< 2.2e-16	0.055	310.046
			adrenal_gland	11	17736	-log10(rhythm.p-value) ~ log(mean.RNA.level) * tau	(Intercept)	-0.209	0.063	-3.312	9.30E-04	0.078	462.514
							log(mean.RNA.level)	0.139	0.007	19.202	< 2.2e-16	0.078	462.514
							tau.mean	0.604	0.098	6.164	7.30E-10	0.078	462.514
							log(mean.RNA.level):tau.mean	0.045	0.013	3.429	6.10E-04	0.078	462.514
			brain_stem	11	17736	-log10(rhythm.p-value) ~ log(mean.RNA.level) * tau	(Intercept)	0.285	0.047	6.027	1.70E-09	0.048	280.958
							log(mean.RNA.level)	0.065	0.006	11.748	< 2.2e-16	0.048	280.958
							tau.mean	-0.03	0.076	-0.397	0.6915	0.048	280.958
							log(mean.RNA.level):tau.mean	0.04	0.01	3.822	1.30E-04	0.048	280.958
			liver	11	17736	-log10(rhythm.p-value) ~ log(mean.RNA.level) * tau	(Intercept)	-0.426	0.073	-5.87	4.50E-09	0.172	1053.763
							log(mean.RNA.level)	0.298	0.009	32.719	< 2.2e-16	0.172	1053.763
							tau.mean	1.29	0.137	9.407	< 2.2e-16	0.172	1053.763
							log(mean.RNA.level):tau.mean	-0.152	0.02	-7.768	8.50E-15	0.172	1053.763
			lung	11	17736	-log10(rhythm.p-value) ~ log(mean.RNA.level) * tau	(Intercept)	-0.695	0.081	-8.611	< 2.2e-16	0.129	820.506
							log(mean.RNA.level)	0.237	0.009	25.567	< 2.2e-16	0.129	820.506
							tau.mean	0.981	0.122	8.029	1.00E-15	0.129	820.506
							log(mean.RNA.level):tau.mean	0.069	0.017	4.096	4.20E-05	0.129	820.506
			kidney	11	17736	-log10(rhythm.p-value) ~ log(mean.RNA.level) * tau	(Intercept)	-0.359	0.075	-4.769	1.90E-06	0.133	830.439
							log(mean.RNA.level)	0.262	0.009	27.917	< 2.2e-16	0.133	830.439
							tau.mean	1.131	0.125	9.024	< 2.2e-16	0.133	830.439
							log(mean.RNA.level):tau.mean	-0.069	0.018	-3.804	1.40E-04	0.133	830.439
			heart	11	17736	-log10(rhythm.p-value) ~ log(mean.RNA.level) * tau	(Intercept)	-0.247	0.063	-3.954	7.70E-05	0.108	647.805
							log(mean.RNA.level)	0.178	0.008	23.18	< 2.2e-16	0.108	647.805
							tau.mean	0.704	0.103	6.809	1.00E-11	0.108	647.805
							log(mean.RNA.level):tau.mean	0.026	0.017	1.578	0.1145	0.108	647.805
			aorta	11	17736	-log10(rhythm.p-value) ~ log(mean.RNA.level) * tau	(Intercept)	-0.234	0.058	-4.018	5.90E-05	0.108	659.009
							log(mean.RNA.level)	0.174	0.007	24.457	< 2.2e-16	0.108	659.009
							tau.mean	0.633	0.091	6.942	4.00E-12	0.108	659.009
							log(mean.RNA.level):tau.mean	-0.089	0.014	-6.226	4.90E-10	0.108	659.009
			white_adipose	11	17736	-log10(rhythm.p-value) ~ log(mean.RNA.level) * tau	(Intercept)	0.263	0.05	5.225	1.80E-07	0.049	286.603
							log(mean.RNA.level)	0.098	0.006	15.742	< 2.2e-16	0.049	286.603
							tau.mean	0.332	0.077	4.329	1.50E-05	0.049	286.603
							log(mean.RNA.level):tau.mean	-0.12	0.011	-10.559	< 2.2e-16	0.049	286.603

Table S5: T-tests testing the independence between rhythm p-values obtained in each tissue and the tissue-specificity τ by controlling the effect of gene expression.

Supp Table 2 from *Romiguier et al. 2014*

"Species"	"Common name"	"Family"	"Order"	"Class"	"Phylum"	"Number_of_Individuals"	"SNP_number"	"piS"	"piN"	"piNpiS"	"Fit"	"Propagule_Size"	"Fecundity"	"Body_Mass"	"Adult_Size"	"Longevity"	"Marine_Continental"
"Caenorhabditis_brenneri"	"soil roundworm"	"Rhabditidae"	"Rhabditida"	"Chromadorea	"Nematoda"	10	10574	0.032215	0.0012752	0.039584	0.159726	0.0052	77.8	5.12E-06	0.15	0.16	Continental
"Armadillidium_vulgare"	"pill woodlouse"	"Armadillidiidae	"Isopoda"	"Malacostraca	"Arthropoda'	10	119157	0.0187064	0.00193442	0.103409	0.247406	0.12	0.959	0.04	1.8	3.42	Continental
"Armadillidium_nasatum"	"pill woodlouse"	"Armadillidiidae	"Isopoda"	"Malacostraca	"Arthropoda'	2	33749	0.0185015	0.00233782	0.126358	-0.0534897	0.12	0.959	0.04	1.8	3.42	Continental
"Culex_pipiens"	"common house mosquito"	"Culicidae"	"Diptera"	"Insecta"	"Arthropoda'	10	84967	0.0412972	0.00112016	0.0271243	0.418858	0.08	100	0.0025	0.5	0.205	Continental
"Mytilus_galloprovincialis"	"Mediterranean mussel"	"Mytilidae"	"Mytiloida"	"Bivalvia"	"Mollusca"	6	74366	0.0334238	0.00225418	0.0674425	0.0723712	0.01	110000	37.5	7.5	25	Marine
"Ciona_intestinalis"	"vase tunicate"	"Cionidae"	"Enterogona"	"Ascidiacea"	"Chordata"	10	42351	0.0533579	0.00297222	0.0557034	0.109717	0.016	1000	29.7	15.5	3	Marine
"Microtus_arvalis"	"common vole"	"Cricetidae"	"Rodentia"	"Mammalia"	"Chordata"	7	35650	0.00849582	0.00096359	0.113419	0.346856	7.8	0.0768	27.5	11.1	4.8	Continental
"Melitaea_cinxia"	"Glanville fritillary"	"Nymphalidae"	"Lepidoptera"	"Insecta"	"Arthropoda'	10	43631	0.0342226	0.00320756	0.0937264	0.516966	0.0556	95.2	0.165	3.96	1	Continental
"Trachemys_scripta"	"pond slider"	"Emydidae"	"Testudines"	"NA"	"Chordata"	2	4171	0.00909918	0.00154465	0.169758	-0.117794	3.39	0.0822	240	20.7	41.3	Semiaquatic
"Parus_caeruleus"	"blue tit"	"Paridae"	"Passeriformes"	"Aves"	"Chordata"	10	4893	0.00534488	0.00073719	0.137925	0.0287238	11.5	0.0247	10.3	11.5	14.6	Continental
"Aptenodytes_patagonicus"	"king penguin"	"Spheniscidae"	"Sphenisciformes"	"Aves"	"Chordata"	10	5840	0.00263499	0.00045691	0.173402	-0.0153924	90	0.0027	11800	90	41	Semiaquatic
"Eudyptes_moseleyi"	"northern rockhopper penguin"	"Spheniscidae"	"Sphenisciformes"	"Aves"	"Chordata"	4	2331	0.00296217	0.00045379	0.153196	-0.0208097	55	0.0055	2500	55	29	Semiaquatic
"Eudyptes_filholi"	"eastern rockhopper penguin"	"Spheniscidae"	"Sphenisciformes"	"Aves"	"Chordata"	4	2749	0.00424895	0.00053738	0.126476	-0.0291077	55	0.0055	2500	55	29	Semiaquatic

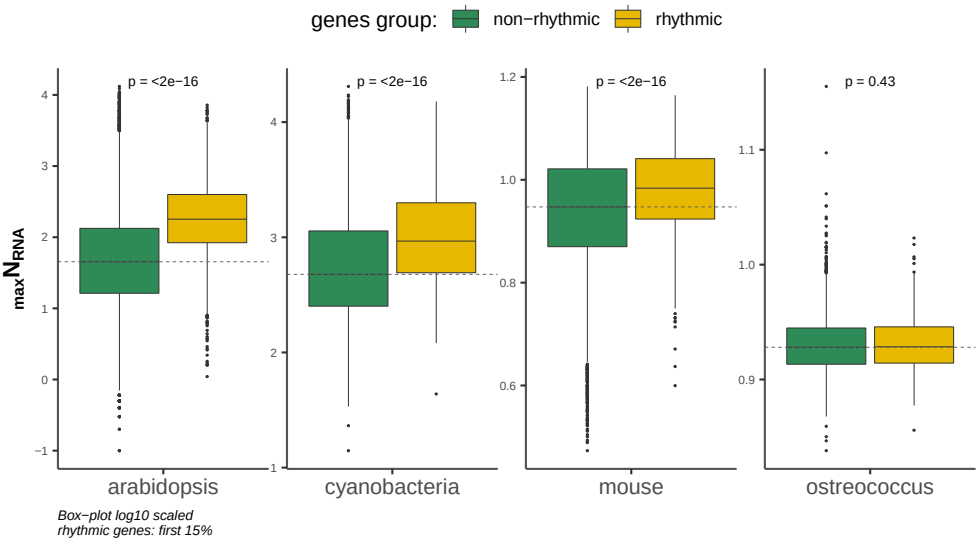
Table S6 Estimation of synonymous nucleotide diversity, π_S, in some non-model animal species from Romiguier et al., for which we found protein sequences FASTA files.

species	omics	nb.tissues	nb.genes	parameters	rhythmic	non.rhythmic	Pearson_cor	Pearson_t	Pearson_Signif
mouse	transcript	11	17736	tau VS nb rhythmic tissues	p<0.01	p>0.5	-0.37	-5.3e+01	< 2.2e-16
baboon	transcript	9	16816	tau VS nb rhythmic tissues	p<0.01	p>0.5	-0.13	-1.7e+01	< 2.2e-16
drosophila	transcript	3	8286	tau VS nb rhythmic tissues	p<0.01	p>0.5	-0.03	-2.3e+00	0.0229
mouse	protein	3	0	tau VS nb rhythmic tissues	p<0.01	p>0.5	0.01	2.7e-01	0.7902

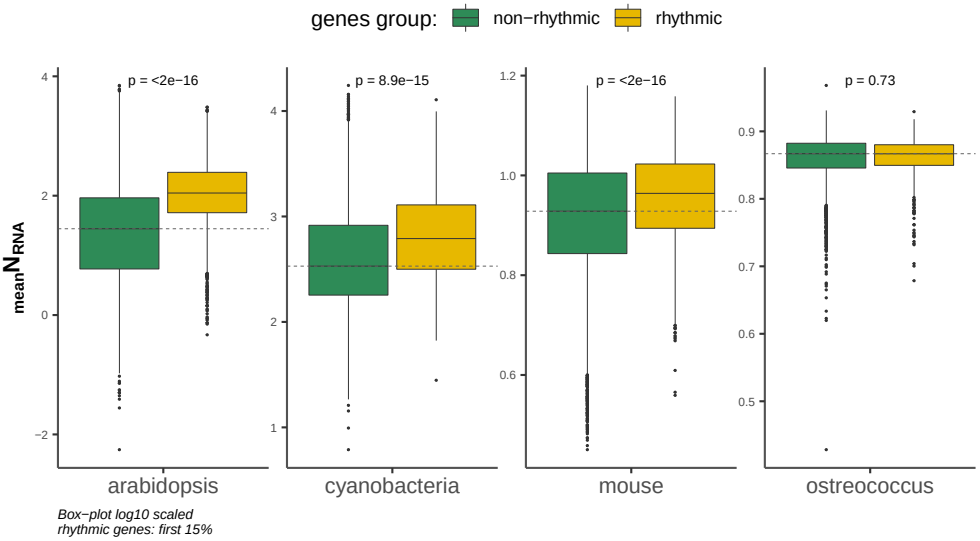
Table S7 Pearson correlation test: τ vs. the number of tissues in which the gene is rhythmic. The number of genes is the number of genes for which there were data for all tissues. Rhythmic is the threshold used to consider the gene as rhythmic..

3.5.2 S1 File

S1 File



(a) Maximum RNA level over time-points (See Methods)



(b) Mean RNA level over time-points

Figure **S1**: Mean or maximum mRNA expression level calculated from time-series datasets. Rhythmic transcripts are highly expressed transcripts.

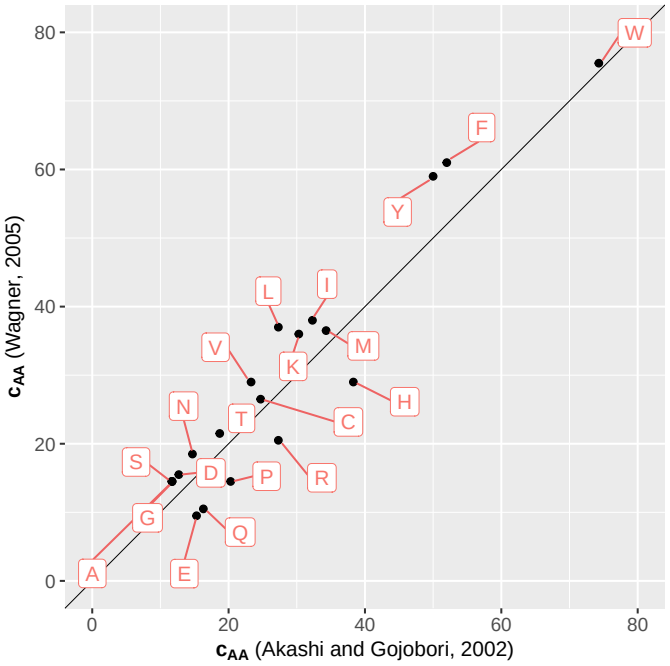


Figure **S2**: Linear relationships of amino acid (AA) biosynthesis costs estimated by Akashi and Gojobori [1], and Wagner [2].

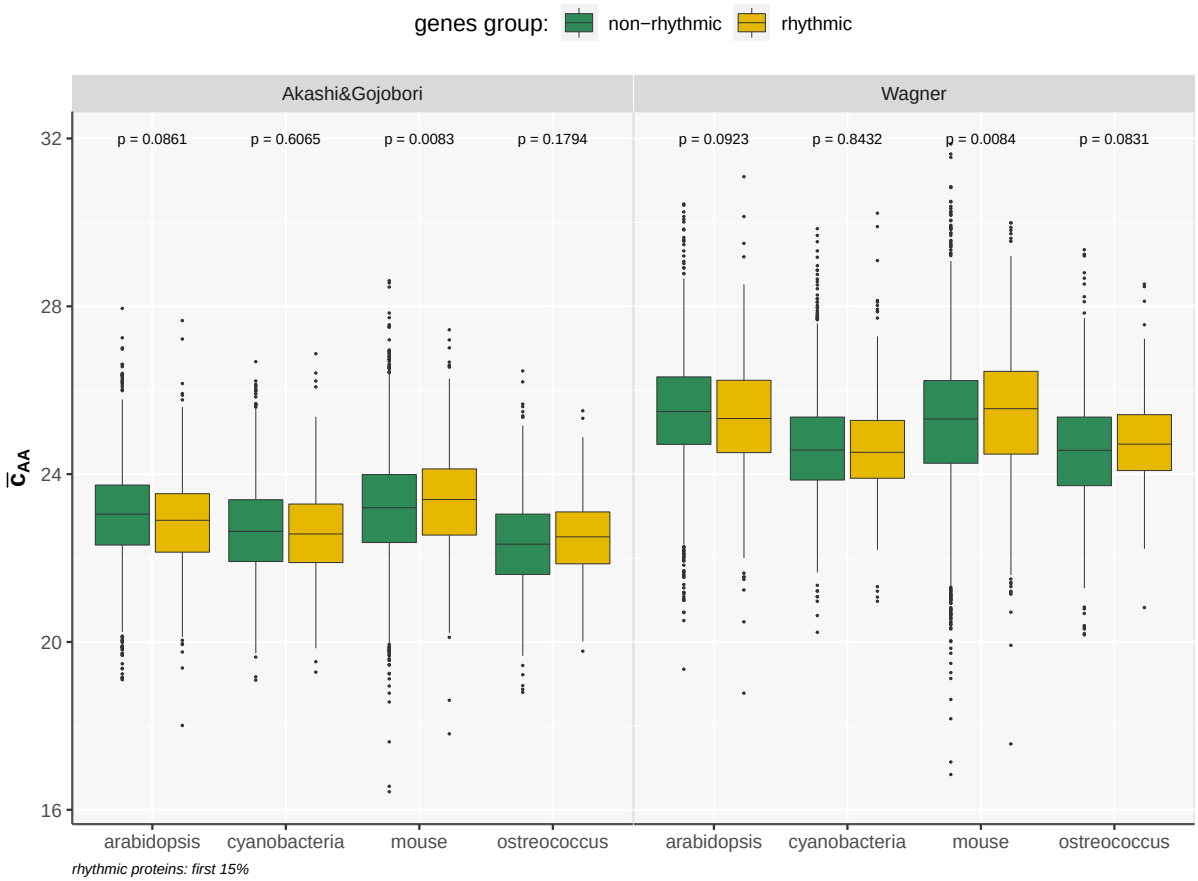


Figure **S3**: Comparison of the averaged AA synthesis costs calculated in both groups: rhythmic VS random genes group, using Akashi and Gojobori versus Wagner AA costs data.

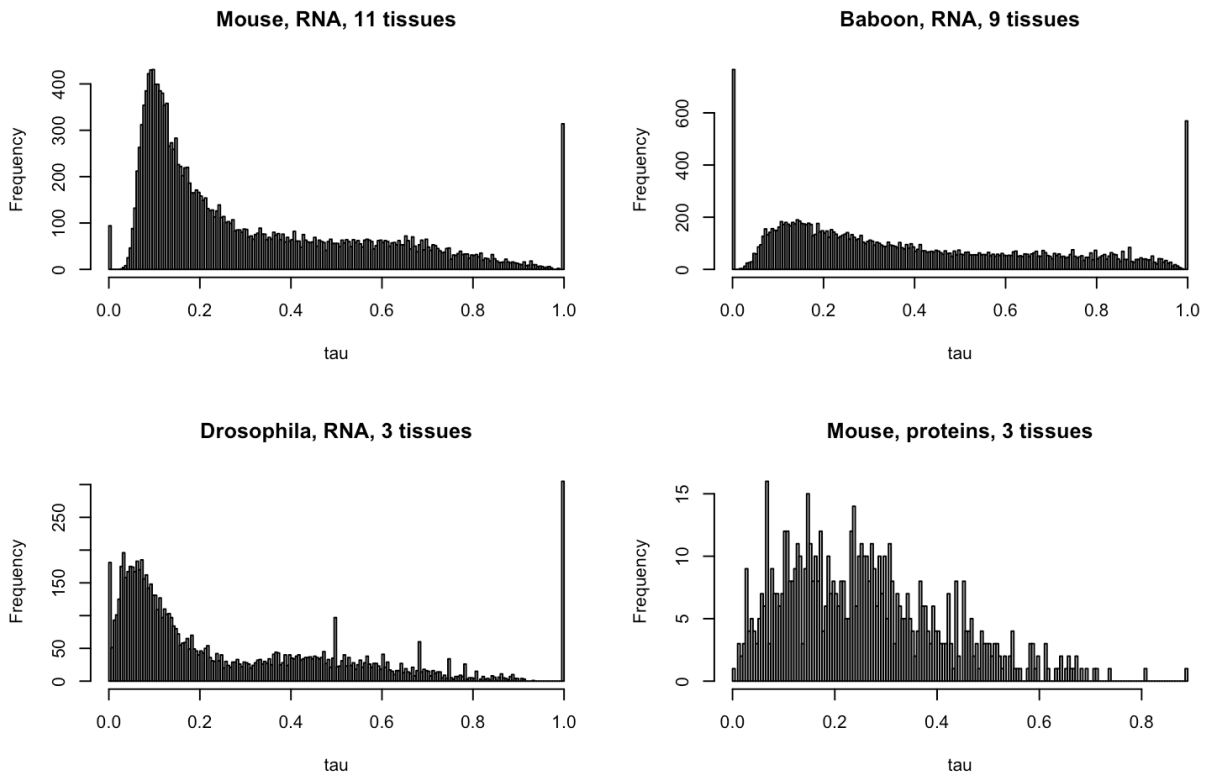


Figure **S4**: Histograms of tissue-specificity τ.

References

[1] H. Akashi and T. Gojobori. Metabolic efficiency and amino acid composition in the proteomes of escherichia coli and bacillus subtilis. *Proceedings of the National Academy of Sciences*, 99(6):3695–3700, 2002.

[2] A. Wagner. Energy Constraints on the Evolution of Gene Expression. *Molecular Biology and Evolution*, 22(6):1365–1374, 03 2005.

3.5.3 Supporting information

Supporting information

1 Detection of rhythmic gene expression

We mainly used the GeneCycle algorithm to detect rhythmic patterns in gene expression time-series (See Methods for more details). We checked that density distributions of p-values obtained from rhythm detection methods used in this paper was producing expected left-skewed distributions (Supporting information Fig S1 and S3). For each gene or protein having several data (ProbIDs or transcripts), we combined p-values by Brown's method using the `EmpiricalBrownsMethod` R package. Fig S2 shows the density distributions of p-values obtained after this Brown's normalization for the transcriptome time-series of Ostreococcus.

2 Gene expression level

Figures S4 and S5 show that distributions of the mean (Fig S4) and the maximum (Fig S5) expression level calculated over time-points (See Methods) can be seen as normal, so relevant for our statistical analysis such as the Student's test.

3 Averaged AA synthesis cost and protein length

Before applying the t-tests, we checked that the averaged AA synthesis cost obtained for each protein and their length were normally distributed in both groups (rhythmic, first 15%, and non-rhythmic) (Supporting information Fig S6 and S8). We also checked that quantile-quantile-plots showed comparable distributions between the theoretical distribution and the empirical distribution for both groups (Supporting information Fig S7 and S9).

Our results support the hypothesis also claimed by Wang et al. [3] that cycling expression of the more expensive genes is a conserved strategy for minimizing overall cellular energy usage. In this study, we provide new results base on relevant data. Indeed, data used by Wang et al. [3] for the calculation of costs seem to be biased mainly due to two points: i. translation rates come from fibroblasts cells[2] and there was errors in the estimation of protein levels resulting in a systematic underestimation of protein levels and derived translation rate constants (Cf Corrigendum [2]).

4 Noise estimation

We compared several noise estimation methods and found that the $F*$ polynomial degree of Barroso et al. [1] method was the best method across all datasets (Supporting information Fig S10) and especially the most efficient method in controlling for the effect of mean expression (Supporting information Table S1). To do this, we calculated the slope of the linear model which best fit the correlation between the noise and mean expression, and the R^2, applied to normally distributed noise estimations. The linear model that explained the variance the least well (small R^2) was considered to be approximately the best expected value. Indeed, we expect a good noise estimation model to be independent of the mean gene expression (slope=0) and with a large range of noise for every mean expression (small R^2).

5 PCA on replicates

PCA on biological replicate or replicates from a different day show how difficult can be the interpretation of results obtained from time-series datasets.

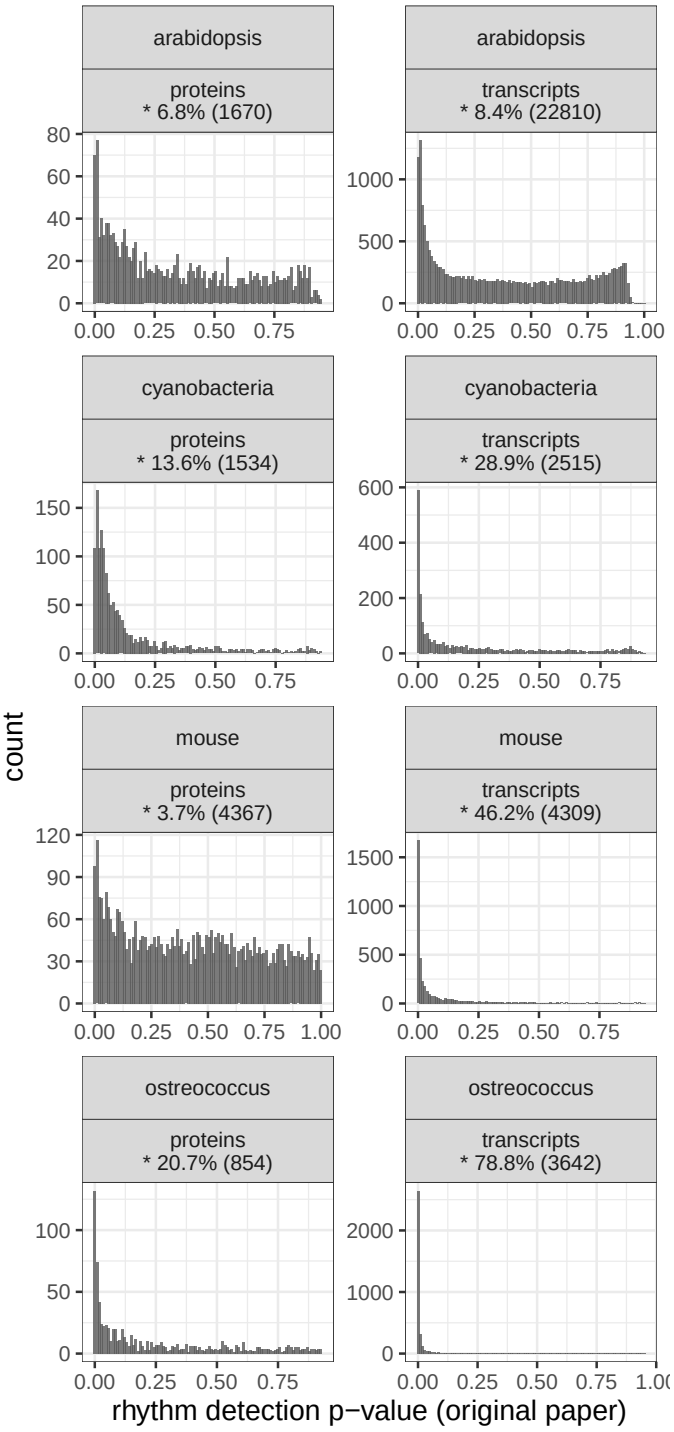


Figure **S1**: **Density distribution of *p*-values obtained from rhythm detection**. All by GeneCycle except for the mouse liver proteome dataset for which no classic rhythm detection methods worked (we used the Harmonic regression method used in the original article). Percentage of genes detected rhythmic with p-value$\leq$0.01 (among total of genes).

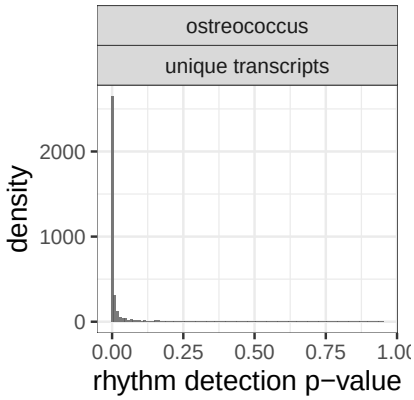


Figure **S2**: Distribution of p-values obtained after the Brown normalization to score unique genes in *Ostreococcus tauri* transcripts time-serie dataset.

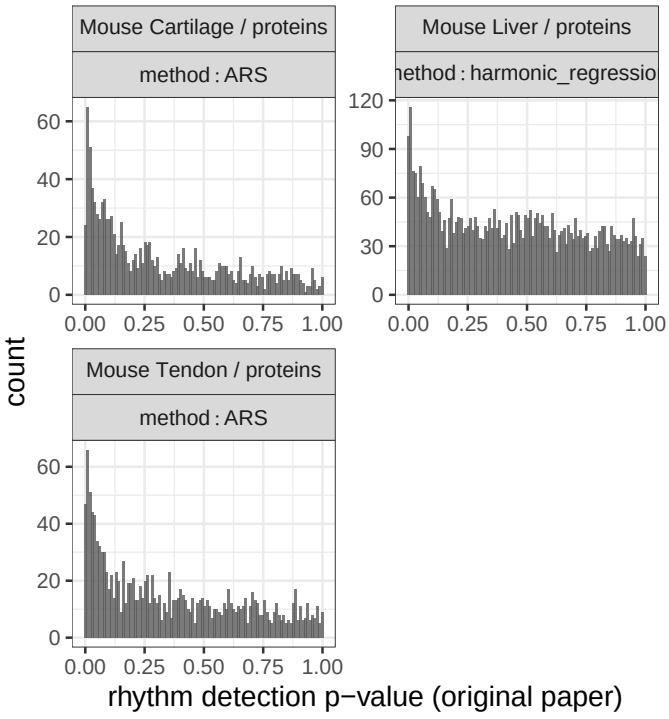


Figure **S3**: Density distribution of p-values obtained from rhythm detection using the method used in this paper.

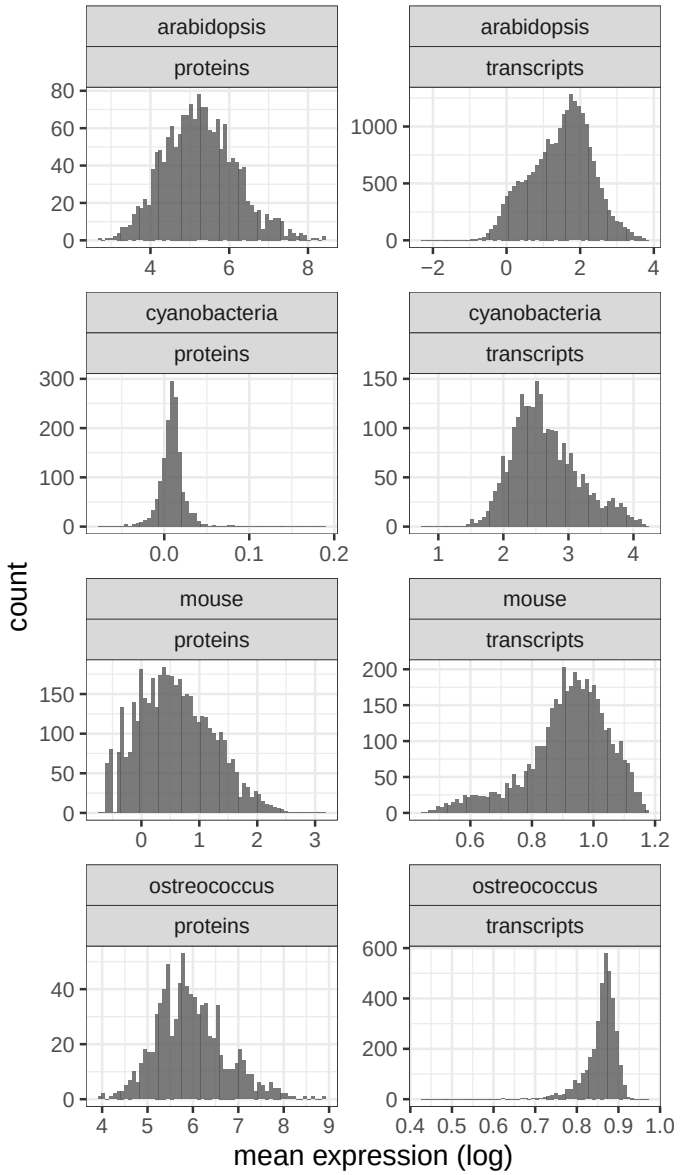


Figure **S4**: Histograms of the mean expression values calculated over time-points.

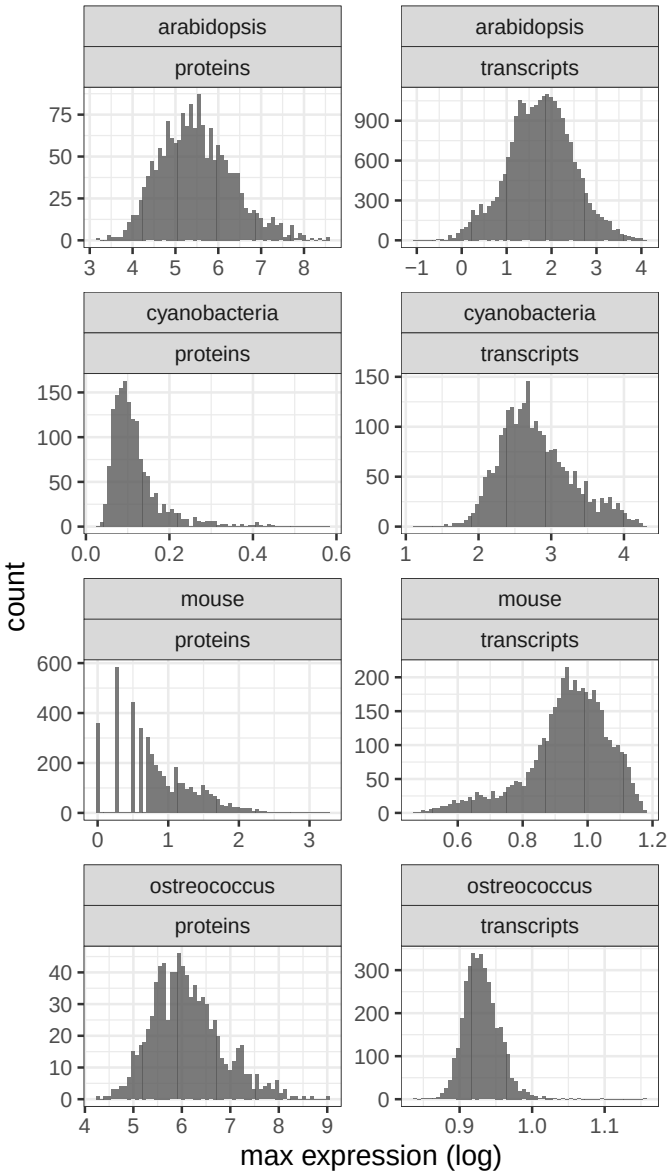


Figure **S5**: Histograms of the maximum expression values calculated (average of the two maximum expression values among time-points).

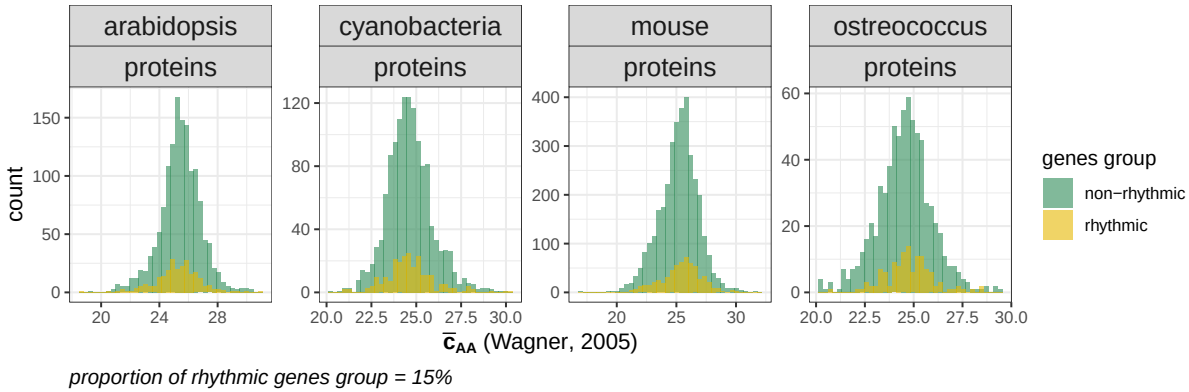


Figure **S6**: Histograms of the averaged AA synthesis costs calculated in both groups.

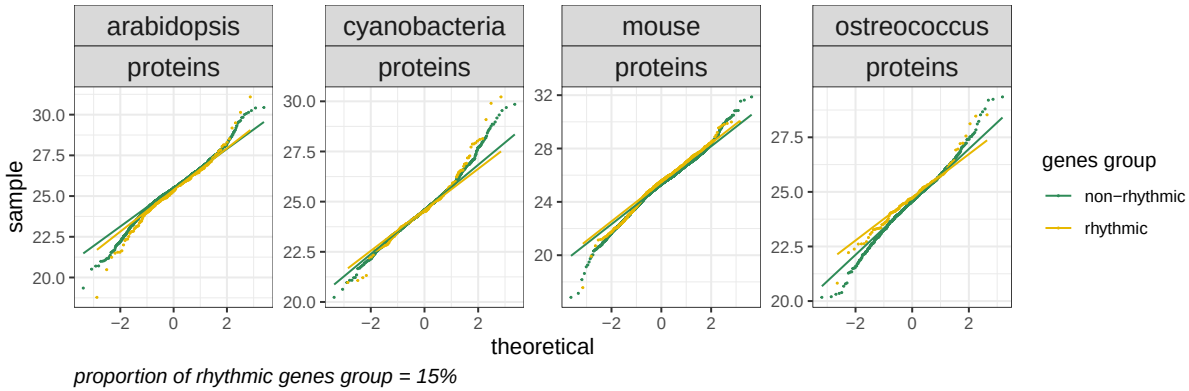


Figure **S7**: QQ-plots of the averaged AA synthesis costs calculated in both groups.

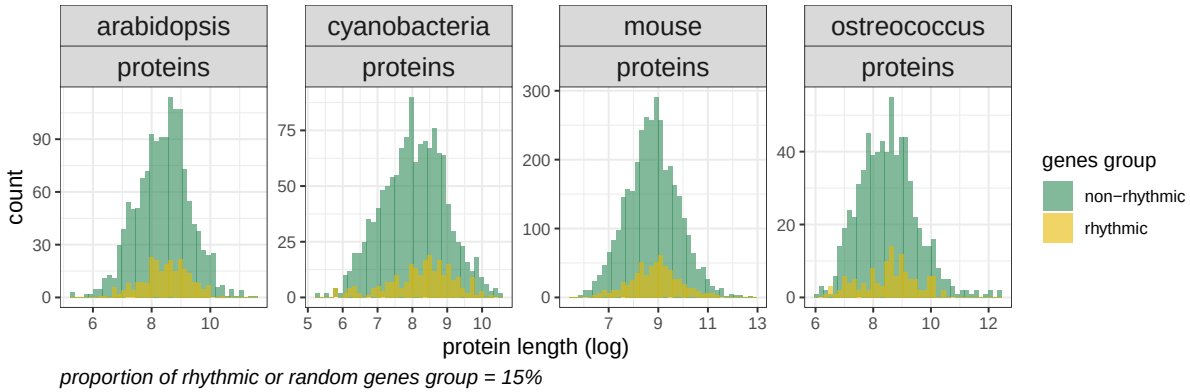


Figure **S8**: Histograms of the protein lengths in both groups.

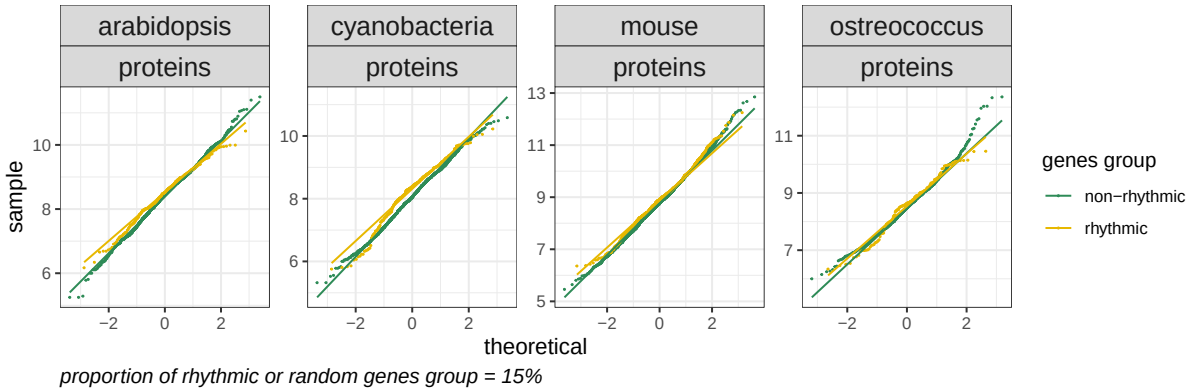
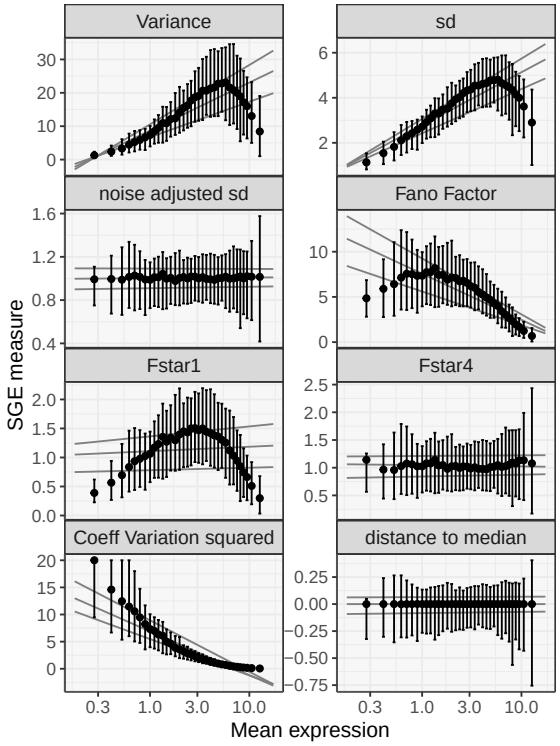
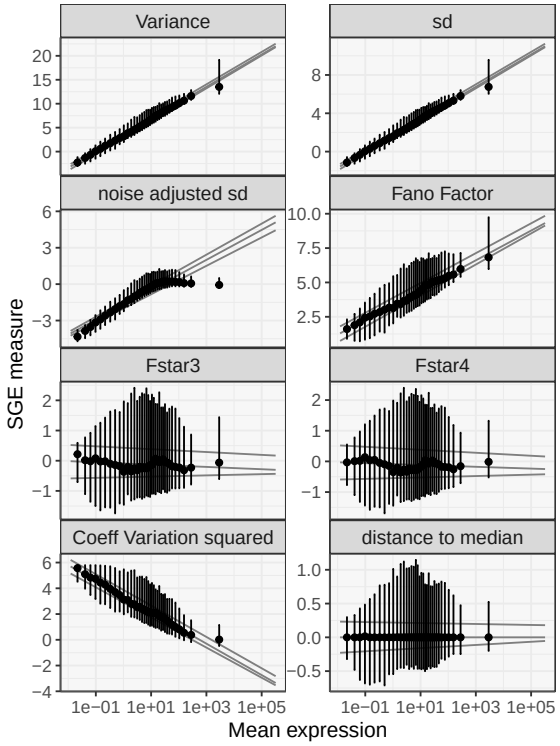


Figure **S9**: QQ-plots of the protein lengths in both groups.

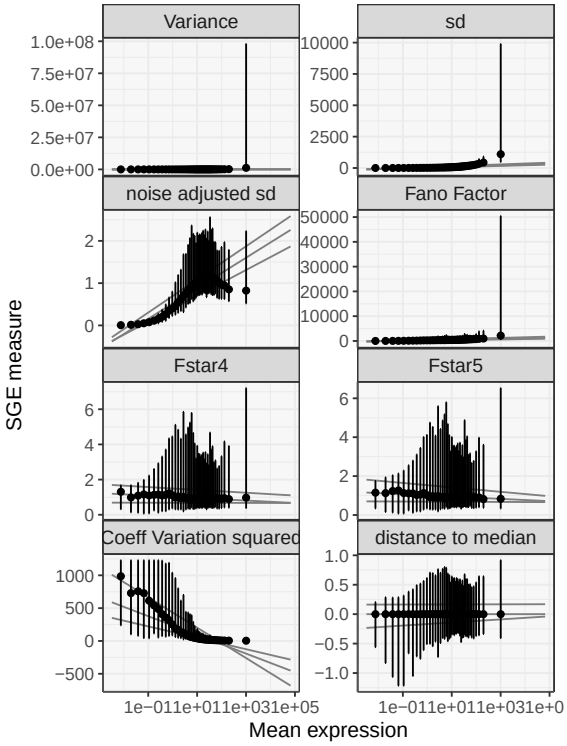
(a) ***Arabidopsis thaliana*** **(root)**



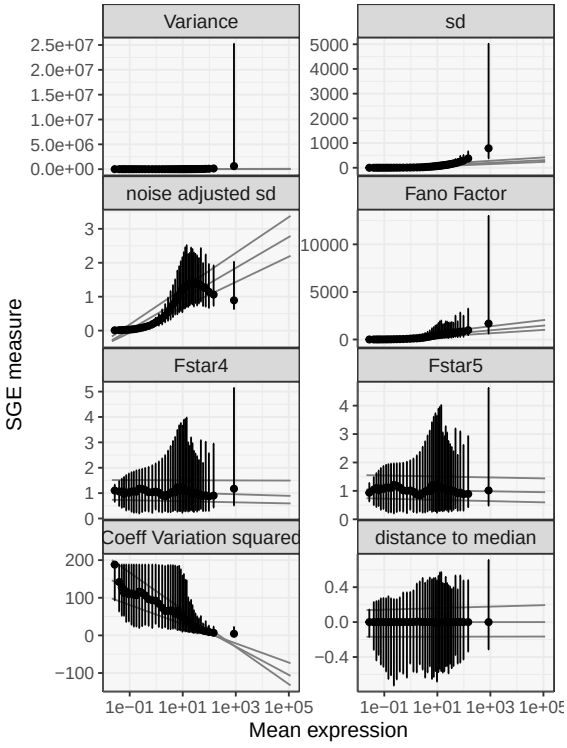
(b) ***Mus musculus*** **(Liver)**



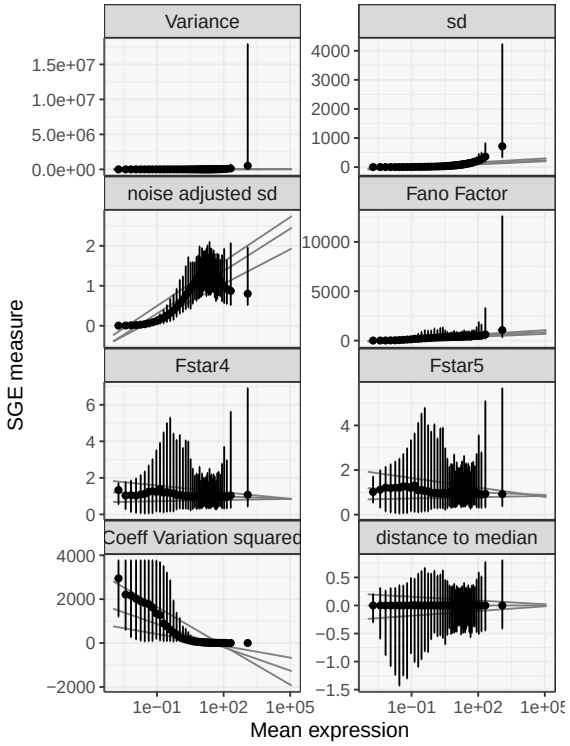
(c) ***Mus musculus*** **(Lung)**



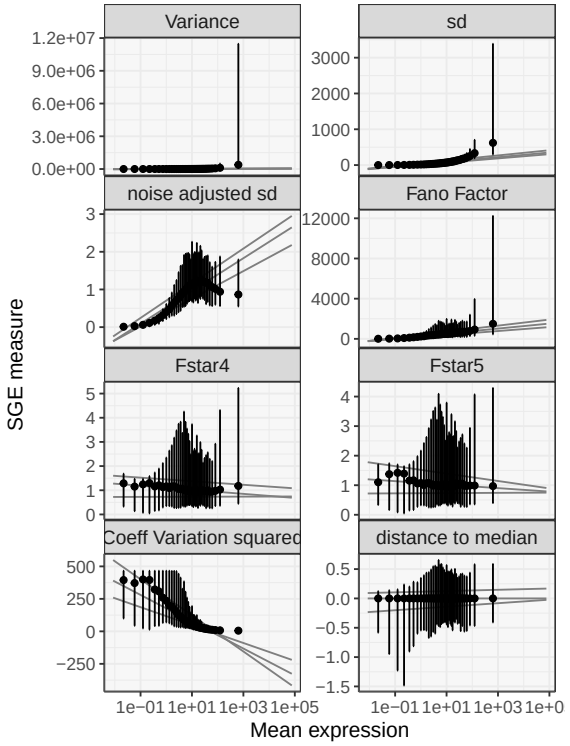
(d) ***Mus musculus*** **(Aorta)**



(e) *Mus musculus* (Heart)



(f) *Mus musculus* (Kidney)



(g) *Mus musculus* (Muscle)

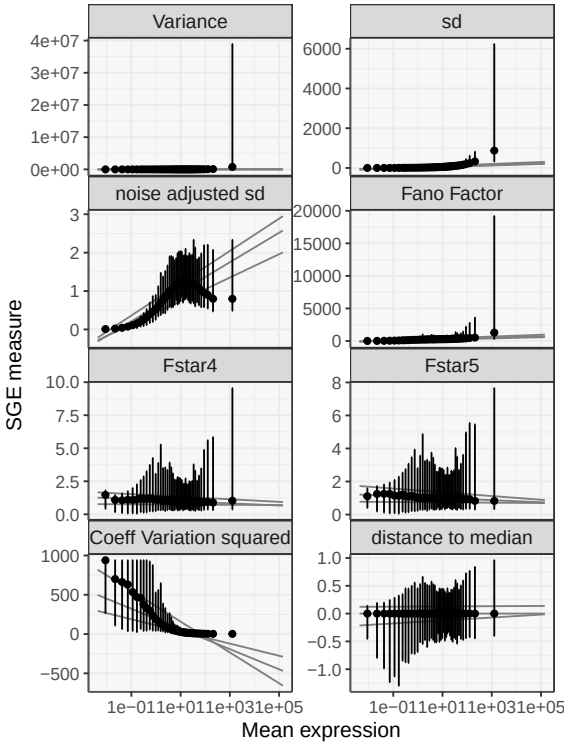


Figure **S10**: **a-f)** Relationships between different stochastic gene expression (SGE) estimations and the mean gene expression. $F*$ is the noise estimated by the method of Barroso et al. [1]. $F*$_min is the first polynomial degree which break the correlation between the noise with mean expression, and $F*$_max is the next one.

Slope and R^2 of the relationships between gene expression and the variance, standard deviation, and different methods of noise estimation.

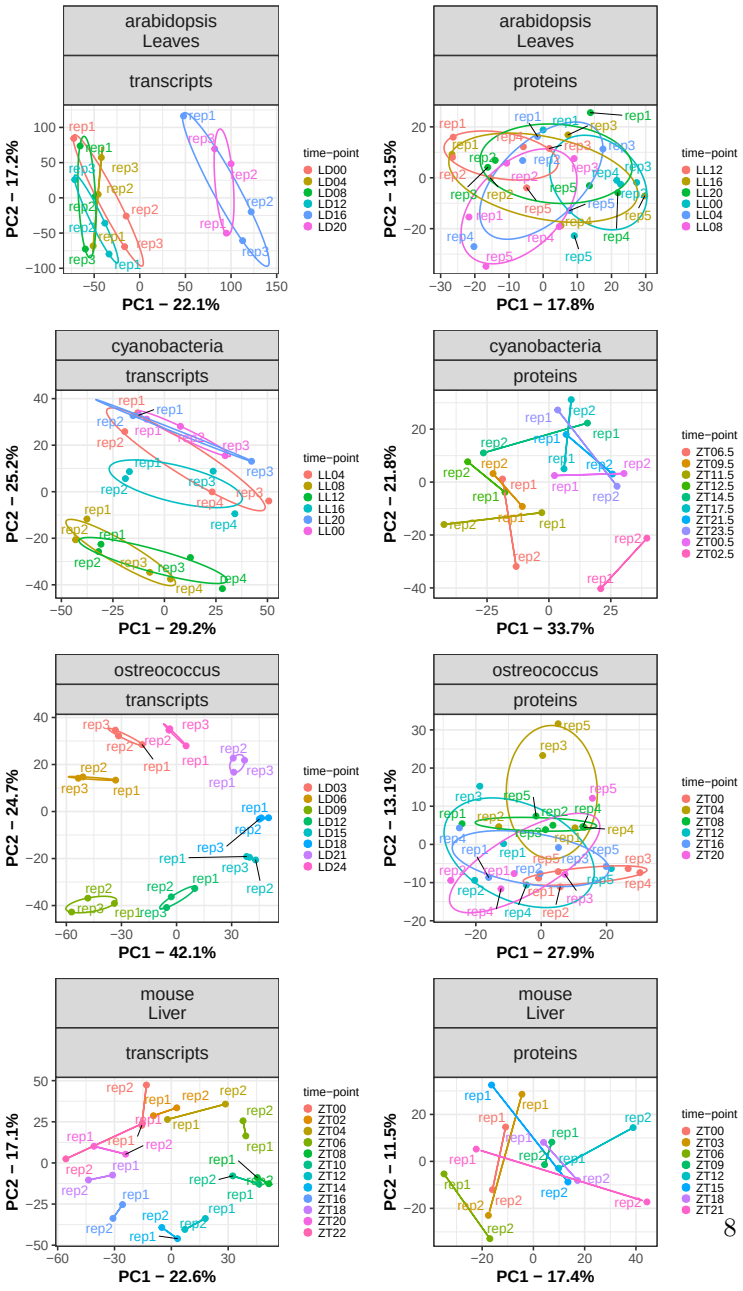
	Arabidopsis root		Mouse Liver		Mouse Lung		Mouse Muscle		Mouse Heart		Mouse Kidney		Mouse Aorta	
	slope	R^2	slope	R^2	slope	R^2	slope	R^2	slope	R^2	slope	R^2	slope	R^2
Variance	0.176	1.79E-01	77.8	0.0114	21.3	1.96E-02	31.3	0.00947	20.3	8.06E-03	19.2	0.00917	16.3	5.94E-03
sd	1.25	1.89E-01	156	0.0114	42.6	1.96E-02	62.5	0.00947	40.6	8.06E-03	38.3	0.00917	32.6	5.94E-03
noise adjusted sd	0.0511	5.07E-06	49.2	0.000582	16.2	1.44E-03	17.8	0.000358	14.7	6.24E-04	12.2	0.000476	15.1	7.48E-04
Fano Factor	-0.958	5.96E-01	211	0.012	54.9	1.66E-02	89.7	0.00874	60.6	7.05E-03	40.2	0.00606	33.3	4.60E-03
F*_min	-2.04	7.79E-02	44.8	0.000147	-57.7	4.34E-03	-156	0.00728	-124	6.77E-03	-98.4	0.00813	-105	4.42E-03
F*_sup	0.638	3.33E-03	-110	0.000885	-2.32	6.92E-06	12.4	0.000045	-1.21	6.26E-07	-6.26	0.000032	-33.4	4.43E-04
Coeff Variation squared	-0.438	4.58E-01	-116	0.00384	-40	1.27E-02	-46.8	0.00505	-32.4	5.34E-03	-45.1	0.00796	-70	6.19E-03
distance to median	-0.631	6.61E-04	-70.7	0.000068	-17.1	7.15E-05	45.1	0.000114	-5.93	2.86E-06	-60.9	0.00058	21.8	3.49E-05

Approximatively the best value Approximatively the second best value

Table S1: Slope and R^2 of the relationships between gene expression and the variance, standard deviation, and different methods of noise estimation of Figures S4. $F*$ is the noise estimated by the method of Barroso et al. [1]. $F*$_min is the first polynomial degree which break the correlation between the noise with mean expression, and $F*$_max is the next one. (See Section *Noise estimation*)

PC1-PC2: RNA and Protein time-series
Arabidopsis; Cyanobacteria; Ostreococcus; Mouse

(a)



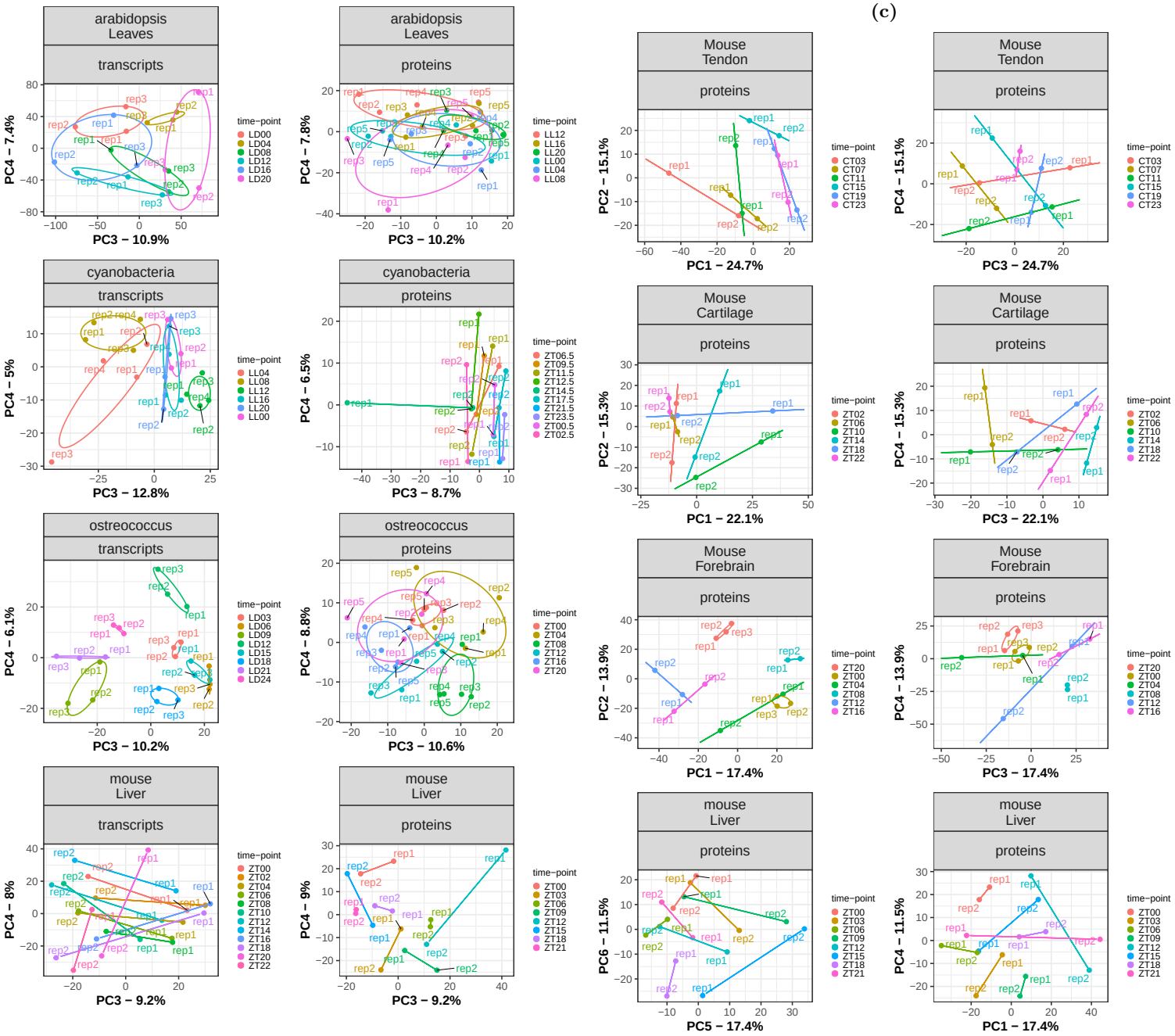


Figure **S11**: **a-c)** PCA of biological replicates or replicates of the same time-point (but from a different day)

References

[1] G. V. Barroso, N. Puzovic, and J. Y. Dutheil. The evolution of gene-specific transcriptional noise is driven by selection at the pathway level. *Genetics*, 208(1):173–189, 2018.

[2] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. Correction: Corrigendum: Global quantification of mammalian gene expression control. *Nature*, 495(7439):126–127, 2013.

[3] G.-Z. Wang, S. L. Hickey, L. Shi, H.-C. Huang, P. Nakashe, N. Koike, B. P. Tu, J. S. Takahashi, and G. Konopka. Cycling transcriptional networks optimize energy utilization on a genome scale. *Cell Reports*, 13(9):1868–1880, 2019/09/30 2015.

Availability of data and scripts

The data and scripts are available at https://github.com/laloumdav/cost_noise_conservation_rhythmicity

Acknowledgements

We thank Johanna Krahmer (CIG, UNIL, Lausanne Switzerland) for her useful advice about the datasets in plants.

DISCUSSION

«Les limites de mon langage signifient les limites de mon propre monde»
Tractatus logico-philosophicus, Ludwig Wittgenstein

4 DISCUSSION

At the beginning of this thesis, I was faced with the problem of detecting rhythmic genes. With a stringent cutoff, there was very little overlap between different algorithms showing that each algorithm has its own criteria for rhythmic gene expression. Each algorithm has its own standard of what a rhythmic gene should be. We have brought a biological approach to this by considering that a good algorithm should preferentially detect rhythmic genes whose rhythmicity is conserved between species. The evolutionary conservation permits to highlight important aspects of rhythmicity in gene expression.
The broader question I worked on was trying to understand why some genes have rhythmic RNAs but not proteins, and vice versa. Apart from mechanistic causes that explain `how` as discussed in the Introduction, it was not clear `why` in some cases the rhythmicity is lost over the processes (from transcription initiation to protein decay), while in others it is initiated during later processes. In some cases it might be a by-product of the evolution of regulatory processes, but there are so many nycthemeral genes that all of these rhythmic patterns have little chance to only be due to genetic drift.

4.1 Evolutionary trade-offs in rhythmic gene expression

4.1.1 Expression level: a barometer for tissue-specific rhythmicity

Relative gene expression levels are an important part in function, under selection. For instance, the fact that transcriptional activity is strongly correlated in sister cells and is transmitted from mother to daughter cells [91], show how important expression levels are for functions. Wild-type expression levels in some conditions are not optimal for growth, and genes whose fitness is greatly affected by small changes in expression level tend to exhibit lower cell-to-cell variability in expression. [92].

Our results suggest that in a given tissue, genes whose function requests a high expression level in this tissue and are specifically expressed in this tissue seem to be rhythmically regulated mainly at the RNA level. This high expression level required might be environmental-dependant.

4.1.2 Expression costs

According to the Schwanhausser et al. study [57], proteins are on average about 2800 times more abundant than their corresponding transcripts in mammalian cells. They estimated the median translation rate at 140 proteins per mRNA molecule per hour (in mouse fibroblasts) [57] which can be up to 1000 proteins per mRNA molecule per hour for highly expressed proteins. Abundant proteins are translated around 100 times more efficiently than those of low abundance [57]. Thus, the energy used for the production of proteins has been estimated to be from 28 times [57] to 125 times [60] larger than the one for transcripts. To produce a significant change in protein versus transcript levels, it requires considerably more catabolic or anabolic activity. That is why the global expression cost per gene is largely represented by the cost at the protein level (data from *Mus musculus* fibroblasts [57], *Escherichia coli* [60][61], *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Arabidopsis thaliana* [60]). Thus, the protein synthesis rates are energetically substantially more constraining than mRNA synthesis rates [61] (Figure 4.1). Especially for genes that are moderately to highly translated. They can impose a high enough energetic burden to be opposed by selection if they do not confer sufficient added benefits [60] (Figure 4.2). The major contribution that pushes the protein expression cost to past the drift barrier in eukaryotes is the cost of translation [60] (Figure 4.2).

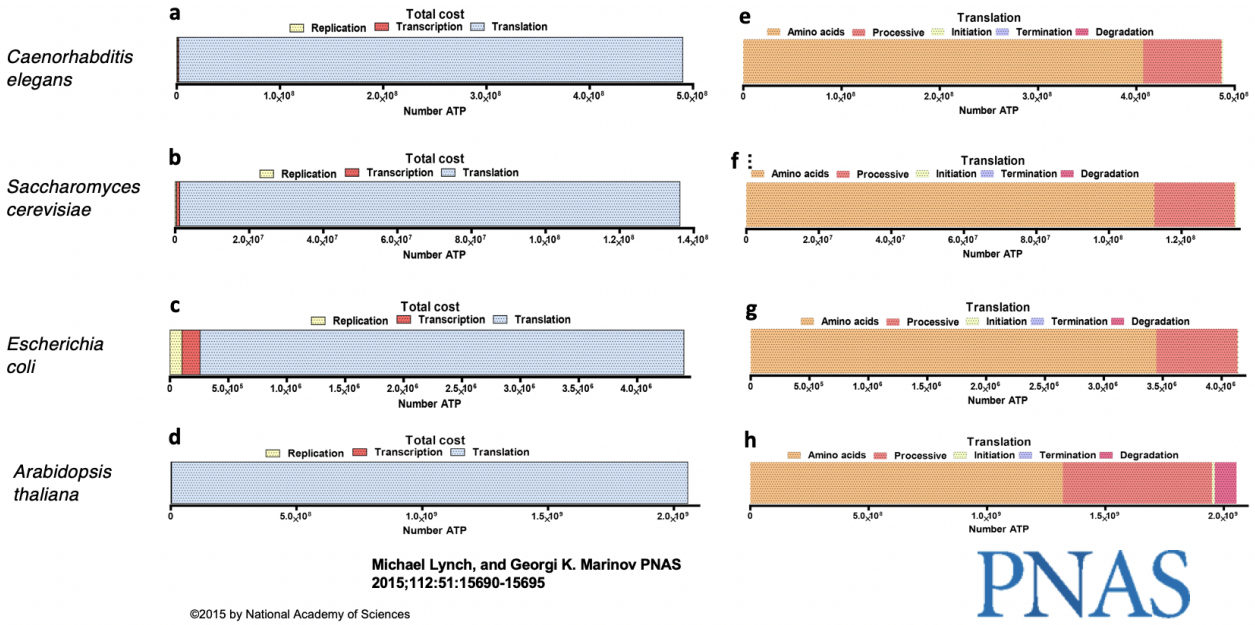


Figure 4.1: a-c) Relative contributions of replication, transcription and translation, into the total expression cost of a gene assessed in *Caenorhabditis elegans, Saccharomyces cerevisiae, Escherichia coli*, and *Arabidopsis thaliana*. **d-e)** Relative contributions of AA synthesis, the "processive" cost of peptide bond formation, translation initiation, translation termination, and protein degradation, into protein expression costs. Source [60].

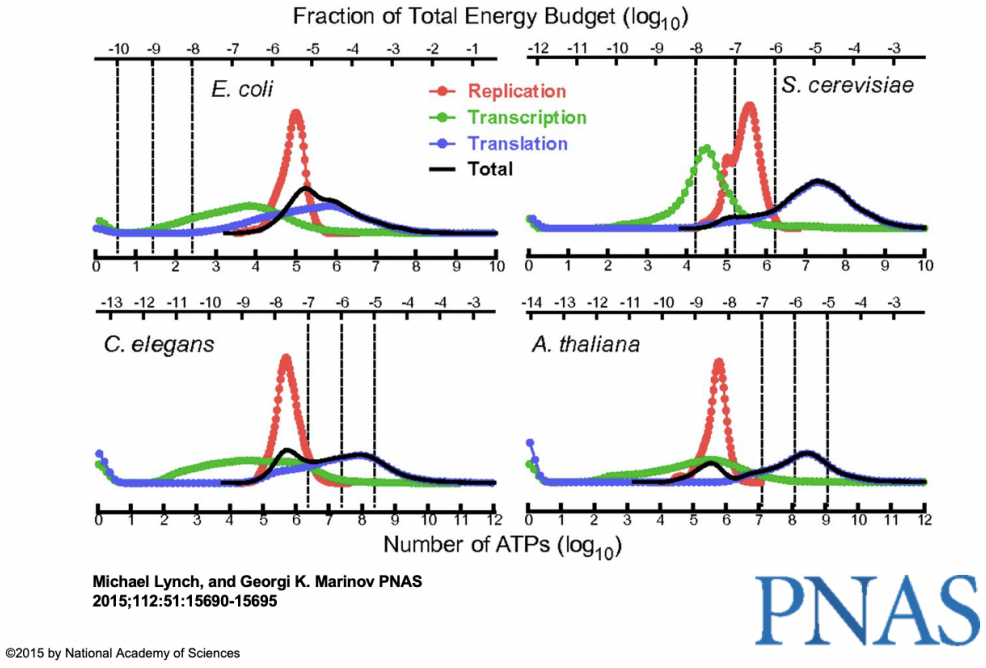


Figure 4.2: Distribution of energy costs for the full sets of annotated genes in one bacterium (E. coli) and four eukaryotic species (Saccharomyces cerevisiae, C. elegans, and A. thaliana). The bottom axis shows the absolute costs in ATP units, and the upper axis shows the corresponding costs as the fraction of the cell's lifetime energy budget. The dashed vertical lines denote key positions below which the energy cost is expected to be too low to be opposed by selection (in the absence of any additional advantages for the gene); for genes to the left of a particular vertical bar (with logarithmic value x on the upper axis), the energetic cost would be effectively neutral if the effective population size (N_e) were $>10^x$. They are equal to the inverse of the effective population sizes for these species, and the surrounding lines are arbitrary +/- orders of magnitude to give the reader an idea of the likely upper / lower bounds, providing the approximate range in which Ne is likely to reside for species in the same broad taxonomic categories as the characterized species. Source [60]

Although these considerations might be overestimated, other work suggests that energetic costs are largely embodied by protein production. Andreas Wagner shows that among expression costs, the protein synthesis costs are the most sensitive constraints to natural selection [61]. An increase of the expression cost was much greater (in yeast) than the selection coefficient (selective disadvantage) above which the natural selection takes over the genetic drift in the genotype [61].

In the absence of rhythmicity, cells should have maintained a constant abundance of gene products at an optimal level (= expression level which maximizes the organismal fitness). At least, at the minimal level at which the function is functional. We show that rhythmic proteins are on average more expressed than others. This suggests that biological systems took advantage of the circadian network to periodically decrease the expression and its main costs (proteins) at times when functions are not needed. In a given issue, genes whose function requires a high expression level in this tissue should be rhythmically regulated due to a "low-cost keeper" strategy. Although it intuitively seems to be obvious, it has never been really highlighted. Rhythmic biological processes seem to provide a first evolutionary advantage by saving energetic costs over a 24-hours time-scale period.

4.1.3 Expression noise

Even from the same experiment, many transcripts show nycthemeral fluctuations without rhythmicity of the abundance of their proteins; in mouse liver for instance [26] and in plants [93]. As we just discussed above, costs at RNA level are probably too negligible to be a satisfactory evolutionary explanation of this rhythmicity at RNA level.
Thattai et al. and Hausser et al. propose that living systems have made tradeoffs between energy efficiency and noise reduction [59][62]. Indeed, the control of noise (stability against fluctuations) plays a key role in the functionality of biological systems, i.e. in the robustness of gene expression, mostly during key periods such as in some developmental stages [94][59].

What is noise? Two types of noise must be defined here:

1. The noise entrained by fluctuations caused by the stochastic nature of biochemical reactions (sometimes called intrinsic noise), such as transcriptional noise. (Figure 4.3B)

2. Cell-to-cell variability (extrinsic noise). (Figure 4.3A)

Inherent stochasticity, or intrinsic noise, is the remaining part of the total noise arising from the discrete nature of the biochemical process of gene expression [95]. To be more precise, if we could measure any phenotype (a protein level for instance) from a same single cell several times in the same conditions and at the same time, we would not obtain a unique value, but a range of values. The higher the stochasticity, the broader the distribution of values. Elowitz et al. [95] defined it as the correlation fail of the activities of two identical copies of one gene in the same cell (Figure 4.3B). Intrinsic noise fundamentally limits the precision of gene regulation.
Extrinsic noise is due to a non-perfect synchronous between cells. It is often considered to be induced by fluctuations of the environment. Although environmental fluctuations seem to affect each cell in the population equally [95], they will not change the heterogeneity of the cell population.

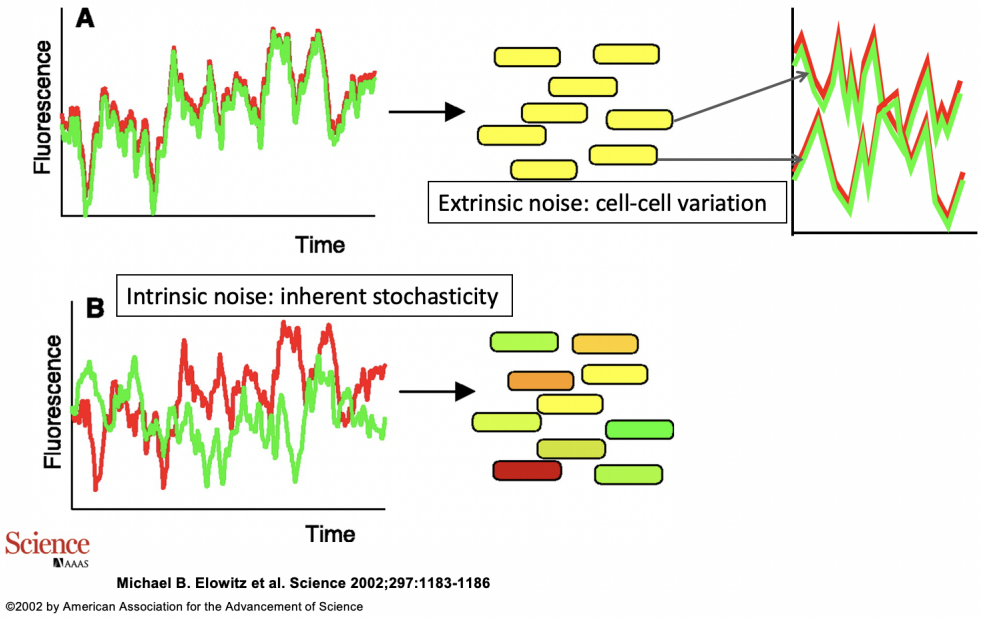


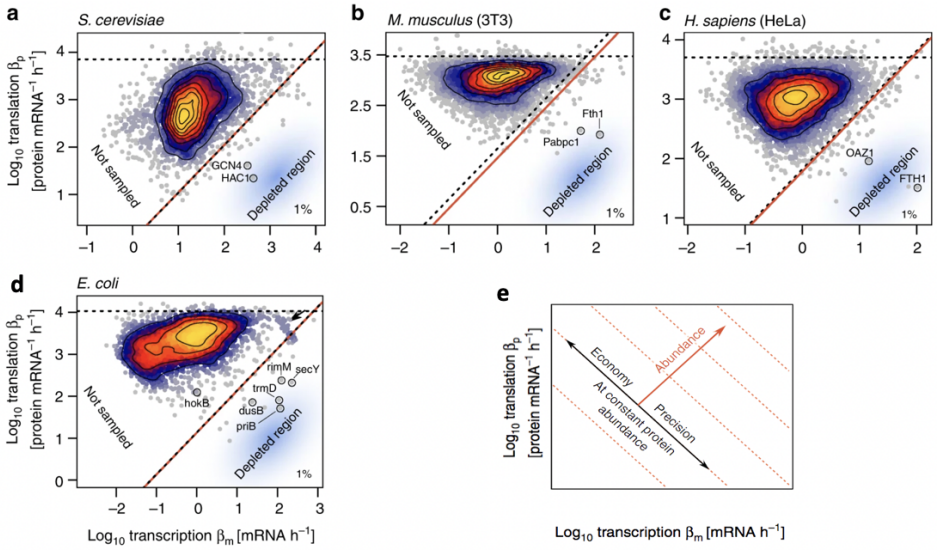
Figure 4.3: Intrinsic and extrinsic noise can be measured and distinguished with two genes (cfp, shown in green; yfp, shown in red) controlled by identical regulatory sequences. Cells with the same amount of each protein appear yellow, whereas cells expressing more of one fluorescent protein than the other appear red or green. (A) In the absence of intrinsic noise, the two fluorescent proteins fluctuate in a correlated fashion over time in a single cell (left). Thus, in a population, each cell will have the same amount of both proteins, although that amount will differ from cell to cell because of extrinsic noise (right). (B) Expression of the two genes may become uncorrelated in individual cells because of intrinsic noise (left), giving rise to a population in which some cells express more of one fluorescent protein than the other. Source [95]

Both types of noise contribute substantially to the overall variation [95]. Noise is negatively correlated to mean expression level [62][1]. Hausser et al. have shown that the noise (intrinsic) is larger when there are few mRNAs per protein unit [62], i.e. many proteins are translated per mRNA unit. An analysis of the literature about noise tend to support the hypothesis of a strong regulation of noise at transcriptional level [62][59].
Two major observations must be noted:

- At constant protein abundance, increasing transcription increases precision [62] (the authors mention an increase of costs of expression but since it should not be relevant for selection, as discussed above, I do not include it into the current discussion)
- Combinations with high transcription and low translation have been eliminated by natural selection [62] (they call it the depleted region, Figure 4.4).

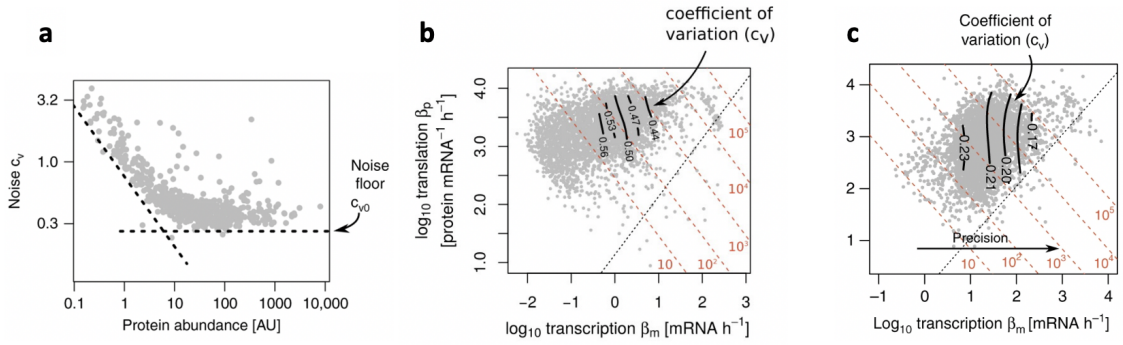
The last point is presumed to be due to the noise floor, at least in part. For highly expressed genes, the variability tends to a constant level called the noise floor [96][62] (Figure 4.5a). Cells can not decrease the noise below a minimum threshold for highly expressed genes. The depleted region corresponds to transcription and translation rates whose functions are probably too sensitive to noise [62], i.e. the cost of fine regulation dominates the benefit of the resulting precise expression. Low translation rates (per mRNA per hour) might be sustained in the genome through the resulting beneficial reduction of the noise it causes [59], until

a certain limit: the noise floor. Another explanation we discuss below might be that noise floor would be the minimum variability under which the noise level would be too low to generate a phenotypic diversity.



Hausser et al. 2019 [62]

Figure 4.4: Genes combining high transcription and low translation are depleted. Transcription and translation rates were estimated from ribosome profiling and mRNA sequencing data. The top percentile of translation rates (β_p^{max}) is represented as a horizontal dashed line. The observed boundary of the depleted region (diagonal dashed line) has slope 1 and is such that 99% of the genes have a larger translation/transcription ratio. β_p and β_m are translation and transcription rates. Source [62]



Hausser et al. 2019 [62]

Figure 4.5: a) The precision of gene expression is limited by the noise floor cv0. Protein abundance and CV (= coefficient of variation) data re-plotted from [97]. The noise floor is also found in the *E. coli* measurements of [98]. **b-c) Increasing transcription at constant protein abundance decreases stochastic fluctuations in protein abundance. b)** In *E. coli*, coefficients of variation (CV, black lines) scale with transcription rates. Transcription and translation rates inferred from [99]. CVs from [97]. Diagonal dotted lines are lines of constant protein abundance from 10 to 105 proteins per cell. **c)** In *S. cerevisiae*, CVs from [100], ribosome profiling and mRNAseq data from [101]. Source [62]

In our work, we used the cell-to-cell variability as an estimation of the overall noise. Indeed, Hausser et al. showed that in *S. cerevisiae* and *E. coli*, the cell-to-cell variation in protein abundance (= protein coefficient of variation, CV) decreases with increasing transcription and decreasing translation on equi-protein level [62] (Figure 4.5b-c). In our case, we used this estimation of noise based on single-cell RNA data instead of proteins, limiting our interpretation.

Rhythmicity at the transcriptional level might be a way to regulate the noise.
We have seen that cost generated at the transcriptional level might not be significant enough to be under selection. However, the noise regulated by the transcriptional layer could be under selection. Indeed, small fluctuations of expression away from optimal wild-type expression has been shown to impact organismal fitness in yeast [72], where noise was nearly as detrimental as sustained (mean) deviation [72]. Noise-increasing mutations in its endogenous promoter have been found to be under purifying selection [102]. Highly expressed genes are under stronger selection and, as proposed by Horvath et al. [103], genes experiencing strong selection could also be less tolerant to high expression noise levels. Our results show a lower mean noise trend among rhythmic transcripts compared to non-rhythmic ones (Results 2 Table 3.3). Thus, rhythmic expression of RNAs could be a way to reduce expression noise of highly expressed genes (Results 2: Figure 3.2 and File S1 Figure S1). These rhythmic and highly expressed transcripts are under stronger selection (dN/dS in Results 2 Supplementary Table S4).

One last point we must mention here is that for some genes, mRNAs oscillating at specific times during the nycthemeral cycle may be required not to regulate the noise, but to adjust the energy consumption of the cell at certain growth stages or in response to the environment. This could also explain the differences in growth stage- and experiment-dependent diurnal transcript oscillations observed [93].

4.2 Fitness, Optimization, and Innovations through Rhythmicity

To what extent is a function is affected by changes in protein levels or by deviation away from the optimal expression level? This raises the question of what an optimal expression is? The first answer is: it depends on the genetic background and the environment, and it's relative to other individuals. For instance, wild-type expression levels in *Saccharomyces cerevisiae* can be non-optimal for growth in some conditions [92].

4.2.1 Optimization

We generally see a function as a biological entity functioning optimally when the level of expression and the noise are optimal given the genetic background and the environment. These optimal levels are those which maximize the organism's fitness in its environment. For complex organisms with a lifespan of at least a few days and living in a cyclic environment, the notion of an optimal expression is blurred. A permanent optimization of protein levels in changing environment could be considered as optimal since, contrary to "optimal", "optimized" integrates a time-scale at the individual level. Optimization implies a corrective tendency rather than a reached state. The optimality would therefore be what the sum of the optimizations tends towards.

Optimization of the precision over a given period of time.
For many cellular processes, timing is essential. DNA replication, cellular divisions, cellular response to environmental changes, hormonal secretion, etc., many biological processes are time-organized. Gene

expression is one of the first step of this regulation. Reach a precise protein level at a given time t can be a challenge for cells. Thus, the timing can be seen as a constraint that cells might need to control. Based on simulations, Co et al. [104] show that for an induced gene, timing variability (uncertainty due to not perfect timing) is minimal if the level of expression at the time when it's induced is approximately half of the level to reach. A certain precision level might be needed for some functions at a given period of time and might be allowed by rhythmic RNA expressions. In this context, an optimal expression would be the fact that the precision in optimized over a given period of time.

Possible mechanistic explanations of the temporal regulation of precision.
In mammals [105][16], and it seems in prokaryotes as well [106] (but unclear in plants [107]), the transcription occurs during active and inactive periods (ON-OFF). Active periods are short bursts characterized by their frequency (= number of bursts per time-unit) and size (= average number of transcripts produced during one burst episode). Bursting is correlated with the noise. For instance, increasing burst frequency increases mRNA abundances while decreasing mRNA cell-to-cell variability for equi-burst-size [108]. According to the literature, it seems that mechanisms which transform discontinuous transcription (bursting) at individual nuclei into precise macro-scale expression patterns are still unknown. Nevertheless, what regulates bursts? It seems that histone acetylation might play a role in burst frequency [16] (we have no idea what regulates burst sizes). Particularly, in the case of the mammal circadian gene Bmal1 (Introduction Figure 1.1), the variation of bursting frequency of its transcription has been shown to be correlated with a rhythmic acetylation of its promoter [16]. Furthermore, a recent study appears to show the role of histone acetylation into the behavioral rhythmicity in social insects (ants) [15].
These results lead to suggest that periodic regulation of bursts could allow high-frequency bursts to be concentrated over a specific period of time, optimizing the precision of expression of the genes required at that time. This type of regulation based on the size and frequency of transcriptional bursts could be a way to timely control noise and to being in tune with the environment through epigenetic regulations. Indeed, histone modifications, such as acethylation or methylation of the DNA, play important roles in the epigenetic regulation of gene expression. Epigenetic status of promoters appears to affect transcriptional bursts [109][16] and, for instance, gene body methylation is negatively associated with transcriptional noise [109]. Thus, histone modifications, known to be affected by environmental factors, might in some way adjust the innate regulation of expression noise with the environment, allowing permanent optimizations.

Advantages given by stochasticity of gene expression.
On the other hand, in some situations, it is possible that the intrinsic noise of a regulator can actually increase the sensitivity with which its signal is transmitted [59][110]. The hypothesis suggested by Paulsson et al. [110] twenty years ago was to say that in non-linear systems, such as transcription, the signal noise (which I understand to be stochastic signals due to random fluctuations of regulatory molecules which are present in low copy numbers per cell) can reduce the uncertainty in regulated processes. The noise would facilitate the signal detection in nonlinear systems thanks to the stochastic focusing (SF). Although this hypothesis does not seem to have been really investigated, it has the advantage of suggesting a mechanism that would explain some erratic patterns of gene expression.
Another point to raise is that, in changing environments, stochastic gene expression creates phenotypic diversity that can potentially be beneficial for rapid evolution (shown in unicellular organisms: yeast and *E. coli*) [111][112]. The heterogeneity in genetically identical cells ensure that some cells are always prepared for changes of the environment [111] (they call it "blind anticipation"). Importantly, it seems that de-novo promoters in *E. coli* exhibit low noise by default and that the main role of transcription factors would be to

increase the noise of its targets [112]. Thus, expression noise would allow phenotypic diversity enhancing the population fitness in fluctuating environments [113].

Thus, the evolutionary trade-off takes into account the advantages provided by the function, its expression costs and precision required, as well as the expression variability which improves the organism's adaptation in a fluctuating environment.

Repressive system.
To conclude this part about optimization, another interesting point is that the major structure of regulation for rhythmicity appear to be a repression, i.e. it is mainly based on repressive actions [114]. For instance, the promoter sequences of clock gene Bmal1 has receptor response elements (ROREs, Introduction: Figure 1.1) which seem to be required for the drop in burst frequency at the expression trough [16]. This repressive nature presupposes the initial existence of a default gene expression. Indeed, the emergence of repressive systems among networks of initially unexpressed genes would not make sense.

4.2.2 Rhythmicity and Fitness

We have seen to which extent the rhythmicity of gene expression (i.e. the entirety of genes with nycthemeral variations) can improve organism's fitness. By:

- Anticipation (circadian rhythms)
- Lower energetic costs (Proteins)
- Periodic noise adjustment (RNA)
- Temporal compartmentalization of biological processes
- More hypothetical:
  - Optimization of timing, precision, and transmitted signals
  - Periodic noise control (RNA) for noise sensitive functions in fluctuating environments.

The fitness is conceivable in the context of rhythmic gene expression since organisms with physiological systems able to accommodate themselves to changing circumstances are expected to have a higher stability of survival and reproduction [47]. To my knowledge, there are no multicellular organisms (reproducing by generations) with a lifespan shorter than 24 hours. They all live at least a few days. Since fitness is the ability to survive, find a mate, reproduce and transmit its genes to a fecund generation, an averaged daily probability to survive, day after day, might be a time-scale parameter to combine with the fitness concept. One more day in an individual's life increases its chances of transmitting its genes.

4.2.3 Innovations through rhythmicity

Such periodic mechanisms also allowed a temporal compartmentalization of biological processes that might be incompatible. For example, nitrogen fixation versus photosynthesis in plants appears to be temporally separated [115]. At the same time, temporal compartmentalization can also impose a high enough constraint to become pathological in some case. For instance, a dysregulation of the temporal organization of

four pathways (protein expression associated with ribosomes, ATP synthesis, glucose metabolism, and the cytoskeleton) in articulations of mice has been observed in osteoarthritis or with aging[76].
For complex organisms in variable environments mainly govern by nycthemeral changes, the rhythmic network might have been a fertile ground for better adaptation and for evolutionary innovations. Here, by innovation I mean novel gene expression state or new function. A new property (or new state of a property) could be advantageous for the individual only inside the rhythmic regulation network whereas the same new property would not be advantageous outside the rhythmic network. Even whether the rhythmic network only provides to the new property with an ability of responsiveness to nycthemeral changes in the environment.
First, the rhythmic network provides a temporal frame-work allowing to temporally arrange the pathways. Secondly, the circadian network could have made it possible to keep complex and costly new properties (or new state of a property) that would have been eliminated otherwise (i.e., without regulation of costs or noise allowed by the rhythmic network for instance). The new property can have increased the fitness of individuals simply because the trade-off between the advantages brought by the new property, and the new costs relaxed by rhythmicity, has increased their survival or reproduction.

4.2.4 Rhythmicity and Essentiality

Are rhythmic genes evolutionarily conserved due to their rhythmicity or is there a selection for the rhythmicity of expression of important genes in specific conditions?
Organisms living in extreme environments, fairly constant environments, such as Mexican cavefishes *Astyanax mexicanus* (Figure 4.6), have lost their rhythmicity (behavior and circadian clock) [116]. Nevertheless, contrary to other cavefish species such as *Phreatichthys andruzzii*, the circadian clock of *Astyanax mexicanus* is still functional (if we put them under light-dark cycles). Furthermore, it seems that in these caves there are some external nycthemeral signals, such as the bats population which displays clear nycthemeral rhythmicity of their activity, defecation, etc. These external nycthemeral signals could have been used by cavefishes as external cues to synchronize and entrain their circadian clocks. Which does not seem to have been the case. Fishes have lost their rhythmicity in these habitats with little nycthemeral changes in less than 30,000 years [117]. This shows that rhythmicity provides an advantage to individuals only if they live in environments with nycthemeral changes large enough to be a selective constraint. Moreover, if rhythmicity had become an essential property for the functionality of many biological processes, evolution would have maintained it, especially in this cave environment where some external nycthemeral signals would have made it possible to maintain the entrainment of circadian systems.

Figure 4.6: Maxican blind Cavefish *Astyanax mexicanus* and direct descendants of ancestral surface fish. They diverge less than 30,000 years ago. (Image: © Richard Borowsky)

Essential genes are genes indispensable for the viability of an organism. The evolutionary conservation of a gene is associated with its essentiality (shown in bacteria and mammals) [118][119][120]. These essential genes play such a fundamental role that their dysfunction cannot be compensated by other genes (duplicated

genes [121]). As we just discussed, rhythmicity does not appear to be an essential property for viability in weakly changing environments where the genetic drift, operating faster than selection, has rapidly changed the functions of mexican cavefishes. Nevertheless, we have shown that rhythmicity at RNA level appear to concern of genes under stronger selection. We also found a lower conservation for genes rhythmic at the protein level and for genes rhythmic at both levels (Results 2: Supplementary Table S3). Thus, genes rhythmic at protein level are highly expressed (Results 2: Figure 3.2) and less conserved genes (Results 2: Supplementary Table S3). This is all the more true when we know that conserved genes tend to be highly expressed. One explanation would be that the rhythmicity at the protein level confers an advantage for new functions, i.e. not yet essential, that are more costly than usual (usually, the gene expression of recent genes is weaker relatively to others). This hypothesis fits with the hypothesis discussed above proposing that the circadian network could have made it possible to keep complex and costly new properties that would have been eliminated otherwise.

4.3 Complex or simple system?

It is now recognized that oscillations occur in a broad spectrum of chemical and biological systems spanning the most primitive to the most complex. We have used the term "complex" several times to refer to the rhythmic network, but here it is not clear what the complexity refers to. Some parts of a given system can be complex in appearance, appearing more chaotic than other parts, although they are all governed by the same rules. It can sometimes be a non-sense to try to find theoretical rules that very specific observations would follow. The Game of Life [122] would be an interesting example for me to illustrate this. The study of complexity often leads to try to find complex laws to which they seem to be subject while these are by-products of the complete system which responds entirely to simple laws.

4.4 Implications for future medicine

4.4.1 Evolutionary medicine

Evolutionary medicine is a growing field which apply the evolutionary concepts to our understanding of diseases [123]. It is based on the fact that we (*Homo sapiens*) are both at the same time adapted and adapting, i.e. that past evolutionary events can help to understand the functionalities of the human body, investigates human disease vulnerability and disease etiologies. Another point which is not clearly specified in papers talking about evolutionary medicine is that comparative transcriptomic and proteomic can help to make the difference between the normal (healthy) and the pathological.
Such as for this work trying to understand `why` genes are rhythmic and not `how`, evolutionary medicine try to understand `why` do we have susceptibility to certain diseases by studying the evolution of parasites, human, their coevolution, and their evolutionary trade-offs.

4.4.2 Chrono-Medicine, the future of the Medicine of precision

Many recent work show the relationships between rhythmicity and diseases. Severe metabolic disturbance have been observed both in circadian mutant mice and in humans subject to shift work [124]. Inappropriate meal-times may be a major cause of coronary heart disease in shift workers [124]. The efficacy of chemotherapy as well differs greatly depending on the time of the day when it is administrated [125]. Thousands of cycling genes in human encode for proteins that either transport or metabolize drugs or are themselves drug targets [126]. Even in surgery: Montaigne et al. [127] have shown that the peri-operative

myocardial ischaemia-reperfusion tolerance in patients undergoing aortic valve replacement surgery depends of the time-of-day of the surgery (which was concomitant with transcriptional alterations in expression of the circadian receptor gene Rev-Erb).
These results support the idea that in order to obtain the best drug or therapeutic response, biological time must be taken into consideration. Further characterization of the rhythmic transcriptome and proteome will thus highlight potential therapeutic targets and could facilitate more effective chrono-therapeutic strategies.
The future of the personalized medicine is a chrono-personalized medicine: For each patient, there is a right drug administrated at the right time.

REFERENCES

[1] Gustavo Valadares Barroso, Natasa Puzovic, and Julien Y. Dutheil. The evolution of gene-specific transcriptional noise is driven by selection at the pathway level. *Genetics*, 208(1):173–189, 2018. ISSN 0016-6731. doi: 10.1534/genetics.117.300467. URL https://www.genetics.org/content/208/1/173.

[2] Rodrigo Sigala, Sebastian Haufe, Dipanjan Roy, Hubert Dinse, and Petra Ritter. The role of alpha-rhythm states in perceptual learning: insights from experiments and computational models. *Frontiers in Computational Neuroscience*, 8:36, 2014. ISSN 1662-5188. doi: 10.3389/fncom.2014.00036. URL https://www.frontiersin.org/article/10.3389/fncom.2014.00036.

[3] Teresa L. Iglesias, Jean G. Boal, Marcos G. Frank, Jochen Zeil, and Roger T. Hanlon. Cyclic nature of the rem sleep-like state in the cuttlefish sepia officinalis. *Journal of Experimental Biology*, 222(1), 2019. ISSN 0022-0949. doi: 10.1242/jeb.174862. URL https://jeb.biologists.org/content/222/1/jeb174862.

[4] Fabio Piccolin, Lavinia Suberg, Robert King, So Kawaguchi, Bettina Meyer, and Mathias Teschke. The seasonal metabolic activity cycle of antarctic krill (euphausia superba): Evidence for a role of photoperiod in the regulation of endogenous rhythmicity. *Frontiers in Physiology*, 9:1715, 2018.

[5] Jeffrey Hubbard, Mio Kobayashi Frisk, Elisabeth Ruppert, Jessica W. Tsai, Fanny Fuchs, Ludivine Robin-Choteau, Jana Husse, Laurent Calvel, Gregor Eichele, Paul Franken, and Patrice Bourgin. Melanopsin-dependent direct photic effects are equal to clock-driven effects in shaping the nychthemeral sleep-wake cycle. *bioRxiv*, 2020. doi: 10.1101/2020.02.21.952077. URL https://www.biorxiv.org/content/early/2020/02/26/2020.02.21.952077.

[6] Seung-Hee Yoo, Shin Yamazaki, Phillip L. Lowrey, Kazuhiro Shimomura, Caroline H. Ko, Ethan D. Buhr, Sandra M. Siepka, Hee-Kyung Hong, Won Jun Oh, Ook Joon Yoo, Michael Menaker, and Joseph S. Takahashi. Period2::luciferase real-time reporting of circadian dynamics reveals persistent circadian oscillations in mouse peripheral tissues. *Proceedings of the National Academy of Sciences*, 101(15):5339–5346, 2004. ISSN 0027-8424. doi: 10.1073/pnas.0308709101. URL https://www.pnas.org/content/101/15/5339.

[7] Catharine E Boothroyd, Herman Wijnen, Felix Naef, Lino Saez, and Michael W Young. Integration of light and temperature in the regulation of circadian gene expression in drosophila. *PLOS Genetics*, 3:1–16, 04 2007. doi: 10.1371/journal.pgen.0030054. URL https://doi.org/10.1371/journal.pgen.0030054.

[8] Emi Nagoshi, Camille Saini, Christoph Bauer, Thierry Laroche, Felix Naef, and Ueli Schibler. Circadian gene expression in individual fibroblasts: Cell-autonomous and self-sustained oscillators pass time to daughter cells. *Cell*, 119(5):693 – 705, 2004. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2004.11.015. URL http://www.sciencedirect.com/science/article/pii/S0092867404010542.

[9] A. Gerber, C. Saini, T. Curie, Y. Emmenegger, G. Rando, P. Gosselin, I. Gotic, P. Gos, P. Franken, and U. Schibler. The systemic control of circadian gene expression. *Diabetes, Obesity and Metabolism*, 17(S1):23–32, 2015. doi: 10.1111/dom.12512. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/dom.12512.

[10] Philip B. Kidd, Michael W. Young, and Eric D. Siggia. Temperature compensation and temperature sensation in the circadian clock. *Proceedings of the National Academy of Sciences*, 112(46):E6284–E6292, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1511215112. URL https://www.pnas.org/content/112/46/E6284.

[11] Lihong Chen and Guangrui Yang. Recent advances in circadian rhythms in cardiovascular system. *Frontiers in pharmacology*, 6:71–71, 04 2015. doi: 10.3389/fphar.2015.00071. URL https://pubmed.ncbi.nlm.nih.gov/25883568.

[12] Shuqun Shi, Akiko Hida, Owen P. McGuinness, David H. Wasserman, Shin Yamazaki, and Carl Hirschie Johnson. Circadian clock gene bmal1 is not essential; functional replacement with its paralog, bmal2. *Current Biology*, 20(4):316 – 321, 2010. ISSN 0960-9822. doi: https://doi.org/10.1016/j.cub.2009.12.034. URL http://www.sciencedirect.com/science/article/pii/S0960982209021587.

[13] Jason P. DeBruyne, Elizabeth Noton, Christopher M. Lambert, Elizabeth S. Maywood, David R. Weaver, and Steven M. Reppert. A clock shock: Mouse clock is not required for circadian oscillator function. *Neuron*, 50(3):465 – 477, 2006. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2006.03.041. URL http://www.sciencedirect.com/science/article/pii/S0896627306002674.

[14] Charna Dibner, Ueli Schibler, and Urs Albrecht. The mammalian circadian timing system: Organization and coordination of central and peripheral clocks. *Annual Review of Physiology*, 72(1):517–549, 2010. doi: 10.1146/annurev-physiol-021909-135821. URL https://doi.org/10.1146/annurev-physiol-021909-135821. PMID: 20148687.

[15] Romain Libbrecht, Dennis Nadrau, and Susanne Foitzik. A role of histone acetylation in the regulation of circadian rhythm in ants. *iScience*, 23(2):100846–100846, 02 2020. doi: 10.1016/j.isci.2020.100846. URL https://pubmed.ncbi.nlm.nih.gov/32004990.

[16] Damien Nicolas, Benjamin Zoller, David Suter, and Felix Naef. Modulation of transcriptional burst frequency by histone acetylation. *Proceedings of the National Academy of Sciences*, 115:201722330, 06 2018. doi: 10.1073/pnas.1722330115.

[17] Jerome S Menet, Joseph Rodriguez, Katharine C Abruzzi, and Michael Rosbash. Nascent-seq reveals novel features of mouse circadian transcriptional regulation. *eLife*, 1:e00011, nov 2012. ISSN 2050-084X. doi: 10.7554/eLife.00011. URL https://doi.org/10.7554/eLife.00011.

[18] Nobuya Koike, Seung-Hee Yoo, Hung-Chung Huang, Vivek Kumar, Choogon Lee, Tae-Kyung Kim, and Joseph S. Takahashi. Transcriptional architecture and chromatin landscape of the core circadian clock in mammals. *Science*, 338(6105):349–354, 2012. ISSN 0036-8075. doi: 10.1126/science.1226339. URL https://science.sciencemag.org/content/338/6105/349.

[19] Rukeia El-Athman, Dora Knezevic, Luise Fuhr, and Angela Relógio. A computational analysis of alternative splicing across mammalian tissues reveals circadian and ultradian rhythms in splicing events. *International journal of molecular sciences*, 20(16):3977, 08 2019. doi: 10.3390/ijms20163977. URL https://pubmed.ncbi.nlm.nih.gov/31443305.

[20] Esteban J Beckwith, Carlos E Hernando, Sofía Polcowñuk, Agustina P Bertolin, Estefania Mancini, M Fernanda Ceriani, and Marcelo J Yanovsky. Rhythmic behavior is controlled by the srm160 splicing factor in drosophila melanogaster. *Genetics*, 207(2):593–607, 10 2017. doi: 10.1534/genetics.117.300139. URL https://pubmed.ncbi.nlm.nih.gov/28801530.

[21] Shihoko Kojima, Elaine L Sher-Chen, and Carla B Green. Circadian control of mrna polyadenylation dynamics regulates rhythmic protein expression. *Genes & development*, 26(24):2724–2736, 12 2012. doi: 10.1101/gad.208306.112. URL https://pubmed.ncbi.nlm.nih.gov/23249735.

[22] Peggy Janich, Alaaddin Bulak Arpat, Violeta Castelo-Szekely, Maykel Lopes, and David Gatfield. Ribosome profiling reveals the rhythmic liver translatome and circadian clock regulation by upstream open reading frames. *Genome research*, 25(12):1848–1859, 12 2015. doi: 10.1101/gr.195404.115. URL https://pubmed.ncbi.nlm.nih.gov/26486724.

[23] Nicholas T. Ingolia, Liana F. Lareau, and Jonathan S. Weissman. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, 147(4):789–802, 2020/08/16 2011. doi: 10.1016/j.cell.2011.10.002. URL https://doi.org/10.1016/j.cell.2011.10.002.

[24] Céline Jouffe, Gaspard Cretenet, Laura Symul, Eva Martin, Florian Atger, Felix Naef, and Frédéric Gachon. The circadian clock coordinates ribosome biogenesis. *PLOS Biology*, 11(1):1–17, 01 2013. doi: 10.1371/journal.pbio.1001455. URL https://doi.org/10.1371/journal.pbio.1001455.

[25] Sarah Lück, Kevin Thurley, Paul F. Thaben, and Pål O. Westermark. Rhythmic degradation explains and unifies circadian transcriptome and proteome data. *Cell Reports*, 9(2):741 – 751, 2014. ISSN 2211-1247. doi: https://doi.org/10.1016/j.celrep.2014.09.021. URL http://www.sciencedirect.com/science/article/pii/S221112471400789X.

[26] Daniel Mauvoisin, Jingkui Wang, Céline Jouffe, Eva Martin, Florian Atger, Patrice Waridel, Manfredo Quadroni, Frédéric Gachon, and Felix Naef. Circadian clock-dependent and -independent rhythmic proteomes implement distinct diurnal functions in mouse liver. *Proceedings of the National Academy of Sciences*, 111(1):167–172, 2014. ISSN 0027-8424. doi: 10.1073/pnas.1314066111. URL https://www.pnas.org/content/111/1/167.

[27] Arisa Hirano, Ying-Hui Fu, and Louis J Ptáček. The intricate dance of post-translational modifications in the rhythm of life. *Nature Structural & Molecular Biology*, 23(12):1053–1060, 2016. doi: 10.1038/nsmb.3326. URL https://doi.org/10.1038/nsmb.3326.

[28] Daniel Mauvoisin. Circadian rhythms and proteomics: It's all about posttranslational modifications! *WIREs Systems Biology and Medicine*, 11(5):e1450, 2019. doi: 10.1002/wsbm.1450. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wsbm.1450.

[29] Maria S. Robles, Jürgen Cox, and Matthias Mann. In-vivo quantitative proteomics reveals a key contribution of post-transcriptional mechanisms to the circadian regulation of liver metabolism. *PLOS Genetics*, 10(1):1–15, 01 2014. doi: 10.1371/journal.pgen.1004047. URL https://doi.org/10.1371/journal.pgen.1004047.

[30] Christopher Jang, Nicholas F Lahens, John B Hogenesch, and Amita Sehgal. Ribosome profiling reveals an important role for translational control in circadian gene expression. *Genome research*, 25(12):1836–1847, 12 2015. doi: 10.1101/gr.191296.115. URL https://pubmed.ncbi.nlm.nih.gov/26338483.

[31] Jérôme Mermet, Jake Yeung, and Felix Naef. Systems chronobiology: Global analysis of gene regulation in a 24-hour periodic world. *Cold Spring Harbor perspectives in biology*, 9(3):a028720, 03 2017. doi: 10.1101/cshperspect.a028720. URL https://pubmed.ncbi.nlm.nih.gov/27920039.

[32] Ray Zhang, Nicholas F. Lahens, Heather I. Ballance, Michael E. Hughes, and John B. Hogenesch. A circadian gene expression atlas in mammals: Implications for biology and medicine. *Proceedings of the National Academy of Sciences*, 111(45):16219–16224, 2014. ISSN 0027-8424. doi: 10.1073/pnas.1408886111. URL https://www.pnas.org/content/111/45/16219.

[33] Greg Boyle, Kerstin Richter, Henry D. Priest, David Traver, Todd C. Mockler, Jeffrey T. Chang, Steve A. Kay, and Ghislain Breton. Comparative analysis of vertebrate diurnal/circadian transcriptomes. *PLOS ONE*, 12(1):1–18, 01 2017. doi: 10.1371/journal.pone.0169923. URL https://doi.org/10.1371/journal.pone.0169923.

[34] Anja Korenčič, Rok Košir, Grigory Bordyugov, Robert Lehmann, Damjana Rozman, and Hanspeter Herzel. Timing of circadian genes in mammalian tissues. *Scientific Reports*, 4(1):5782, 2014. doi: 10.1038/srep05782. URL https://doi.org/10.1038/srep05782.

[35] Jake Yeung, Jérôme Mermet, Céline Jouffe, Julien Marquis, Aline Charpagne, Frédéric Gachon, and Felix Naef. Transcription factor activity rhythms and tissue-specific chromatin interactions explain circadian gene expression across organs. *Genome Research*, 2017. doi: 10.1101/gr.222430.117. URL http://genome.cshlp.org/content/early/2017/12/15/gr.222430.117.abstract.

[36] Joshua R Beytebiere, Alexandra J Trott, Ben J Greenwell, Collin A Osborne, Helene Vitet, Jessica Spence, Seung-Hee Yoo, Zheng Chen, Joseph S Takahashi, Noushin Ghaffari, and Jerome S Menet. Tissue-specific bmal1 cistromes reveal that rhythmic transcription is associated with rhythmic enhancer-enhancer interactions. *Genes & development*, 33(5-6):294–309, 03 2019. doi: 10.1101/gad.322198.118. URL https://pubmed.ncbi.nlm.nih.gov/30804225.

[37] Violeta Castelo-Szekely, Alaaddin Bulak Arpat, Peggy Janich, and David Gatfield. Translational contributions to tissue specificity in rhythmic and constitutive gene expression. *Genome Biology*, 18(1):116, 2017. doi: 10.1186/s13059-017-1222-2. URL https://doi.org/10.1186/s13059-017-1222-2.

[38] Violeta Castelo-Szekely and David Gatfield. Emerging roles of translational control in circadian timekeeping. *Journal of Molecular Biology*, 432(12):3483 – 3497, 2020. ISSN 0022-2836. doi: https://doi.org/10.1016/j.jmb.2020.03.023. URL http://www.sciencedirect.com/science/article/pii/S0022283620302564. Circadian Regulation: from Molecules to Physiology.

[39] Stephen Smith and Ramon Grima. Single-cell variability in multicellular organisms. *Nature Communications*, 9(1):345, 2018. doi: 10.1038/s41467-017-02710-x. URL https://doi.org/10.1038/s41467-017-02710-x.

[40] Maria Loza-Correa, Laura Gomez-Valero, and Carmen Buchrieser. Circadian clock proteins in prokaryotes: Hidden rhythms? *Frontiers in Microbiology*, 1:130, 2010. ISSN 1664-302X. doi: 10.3389/fmicb.2010.00130. URL https://www.frontiersin.org/article/10.3389/fmicb.2010.00130.

[41] Florian Atger, Cédric Gobet, Julien Marquis, Eva Martin, Jingkui Wang, Benjamin Weger, Grégory Lefebvre, Patrick Descombes, Felix Naef, and Frédéric Gachon. Circadian and feeding rhythms differentially affect rhythmic mrna transcription and translation in mouse liver. *Proceedings of the National Academy of Sciences*, 112(47):E6579–E6588, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1515308112. URL https://www.pnas.org/content/112/47/E6579.

[42] Jana Husse, Gregor Eichele, and Henrik Oster. Synchronization of the mammalian circadian timing system: Light can control peripheral clocks independently of the scn clock. *BioEssays*, 37(10):1119–1128, 2015. doi: 10.1002/bies.201500026. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.201500026.

[43] Yan Ouyang, Carol R. Andersson, Takao Kondo, Susan S. Golden, and Carl Hirschie Johnson. Resonating circadian clocks enhance fitness in cyanobacteria. *Proceedings of the National Academy of Sciences*, 95(15):8660–8664, 1998. ISSN 0027-8424. doi: 10.1073/pnas.95.15.8660. URL https://www.pnas.org/content/95/15/8660.

[44] Antony N. Dodd, Neeraj Salathia, Anthony Hall, Eva Kévei, Réka Tóth, Ferenc Nagy, Julian M. Hibberd, Andrew J. Millar, and Alex A. R. Webb. Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage. *Science*, 309(5734):630–633, 2005. ISSN 0036-8075. doi: 10.1126/science.1115581. URL https://science.sciencemag.org/content/309/5734/630.

[45] Matthew J. Rubin, Marcus T. Brock, Amanda M. Davis, Zachary M. German, Mary Knapp, Stephen M. Welch, Stacey L. Harmer, Julin N. Maloof, Seth J. Davis, and Cynthia Weinig. Circadian rhythms vary over the growing season and correlate with fitness components. *Molecular Ecology*, 26(20):5528–5540, 2017. doi: 10.1111/mec.14287. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.14287.

[46] Charles Darwin and Alfred Wallace. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. *Journal of the Proceedings of the Linnean Society of London. Zoology*, 3(9):45–62, 1858. doi: 10.1111/j.1096-3642.1858.tb02500.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1096-3642.1858.tb02500.x.

[47] Bell Graham. The basics of selection. *Chapman & Hall*, 1997.

[48] T. Ryan Gregory. Understanding natural selection: Essential concepts and common misconceptions. *Evolution: Education and Outreach*, 2(2):156–175, 2009. doi: 10.1007/s12052-009-0128-1. URL https://doi.org/10.1007/s12052-009-0128-1.

[49] Zachary Gerhart-Hines and Mitchell A. Lazar. Circadian Metabolism in the Light of Evolution. *Endocrine Reviews*, 36(3):289–304, 06 2015. ISSN 0163-769X. doi: 10.1210/er.2015-1007. URL https://doi.org/10.1210/er.2015-1007.

[50] Romain A. Studer and Marc Robinson-Rechavi. How confident can we be that orthologs are similar, but paralogs differ? *Trends in Genetics*, 25(5):210–216, 2009. doi: https://doi.org/10.1016/j.tig.2009.03.004. URL http://www.sciencedirect.com/science/article/pii/S0168952509000559.

[51] Toni Gabaldón and Eugene V. Koonin. Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics*, 14:360 EP –, 04 2013. URL https://doi.org/10.1038/nrg3456.

[52] Stefano Castellana, Tommaso Mazza, Daniele Capocefalo, Nikolai Genov, Tommaso Biagini, Caterina Fusilli, Felix Scholkmann, Angela Relógio, John B Hogenesch, and Gianluigi Mazzoccoli. Systematic analysis of mouse genome reveals distinct evolutionary and functional properties among circadian and ultradian genes. *Frontiers in Physiology*, 9:1178, 2018. ISSN 1664-042X. doi: 10.3389/fphys.2018.01178. URL https://doi.org/10.5167/uzh-157688.

[53] Ohad Nachtomy, Ayelet Shavit, and Zohar Yakhini. Gene expression and the concept of the phenotype. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 38(1):238 – 254, 2007. ISSN 1369-8486. doi: https://doi.org/10.1016/j.shpsc.2006.12.014. URL http://www.sciencedirect.com/science/article/pii/S1369848606000999.

[54] C. Saini, D.M. Suter, A. Liani, P. Gos, and U. Schibler. The mammalian circadian timing system: Synchronization of peripheral clocks. *Cold Spring Harbor Symposia on Quantitative Biology*, 76:39–47, 2011. doi: 10.1101/sqb.2011.76.010918. URL http://symposium.cshlp.org/content/76/39.abstract.

[55] Maxime Policarpo, Julien Fumey, Philippe Lafargeas, Delphine Naquin, Claude Thermes, Magali Naville, Corentin Dechaud, Jean-Nicolas Volff, Cedric Cabau, Christophe Klopp, Peter Rask Møller, Louis Bernatchez, Erik García-Machado, Sylvie Rétaux, and Didier Casane. Contrasting gene decay in subterranean vertebrates: insights from cavefishes and fossorial mammals. *Molecular Biology and Evolution*, 09 2020. ISSN 0737-4038. doi: 10.1093/molbev/msaa249. URL https://doi.org/10.1093/molbev/msaa249. msaa249.

[56] Guang-Zhong Wang, Stephanie L. Hickey, Lei Shi, Hung-Chung Huang, Prachi Nakashe, Nobuya Koike, Benjamin P. Tu, Joseph S. Takahashi, and Genevieve Konopka. Cycling transcriptional networks optimize energy utilization on a genome scale. *Cell Reports*, 13(9):1868–1880, 2019/09/30 2015. doi: 10.1016/j.celrep.2015.10.043. URL https://doi.org/10.1016/j.celrep.2015.10.043.

[57] Björn Schwanhäusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. Correction: Corrigendum: Global quantification of mammalian gene expression control. *Nature*, 495(7439):126–127, 2013.

[58] David Laloum and Marc Robinson-Rechavi. Methods detecting rhythmic gene expression are biologically relevant only for strong signal. *PLOS Computational Biology*, 16(3):1–23, 03 2020. doi: 10.1371/journal.pcbi.1007666. URL https://doi.org/10.1371/journal.pcbi.1007666.

[59] Mukund Thattai and Alexander van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 98(15):8614–8619, 2001. ISSN 0027-8424. doi: 10.1073/pnas.151588598. URL https://www.pnas.org/content/98/15/8614.

[60] Michael Lynch and Georgi K. Marinov. The bioenergetic costs of a gene. *Proceedings of the National Academy of Sciences*, 112(51):15690–15695, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1514974112. URL https://www.pnas.org/content/112/51/15690.

[61] Andreas Wagner. Energy Constraints on the Evolution of Gene Expression. *Molecular Biology and Evolution*, 22(6):1365–1374, 03 2005. ISSN 0737-4038. doi: 10.1093/molbev/msi126. URL https://doi.org/10.1093/molbev/msi126.

[62] Jean Hausser, Avi Mayo, Leeat Keren, and Uri Alon. Central dogma rates and the trade-off between precision and economy in gene expression. *Nature Communications*, 10(1):68, 2019. doi: 10.1038/s41467-018-07391-8. URL https://doi.org/10.1038/s41467-018-07391-8.

[63] Stephen Branden Van Oss and Anne-Ruxandra Carvunis. De novo gene birth. *PLOS Genetics*, 15(5):1–23, 05 2019. doi: 10.1371/journal.pgen.1008160. URL https://doi.org/10.1371/journal.pgen.1008160.

[64] Rafik Neme and Diethard Tautz. Phylogenetic patterns of emergence of new genes support a model of frequent de novoevolution. *BMC Genomics*, 14(1):117, 2013. doi: 10.1186/1471-2164-14-117. URL https://doi.org/10.1186/1471-2164-14-117.

[65] Dennis P. Wall, Aaron E. Hirsh, Hunter B. Fraser, Jochen Kumm, Guri Giaever, Michael B. Eisen, and Marcus W. Feldman. Functional genomic analysis of the rates of protein evolution. *Proceedings of the National Academy of Sciences*, 102(15):5483–5488, 2005. ISSN 0027-8424. doi: 10.1073/pnas.0501761102. URL https://www.pnas.org/content/102/15/5483.

[66] D. Allan Drummond, Jesse D. Bloom, Christoph Adami, Claus O. Wilke, and Frances H. Arnold. Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences*, 102(40):14338–14343, 2005. ISSN 0027-8424. doi: 10.1073/pnas.0504070102. URL https://www.pnas.org/content/102/40/14338.

[67] Itai Yanai, Hila Benjamin, Michael Shmoish, Vered Chalifa-Caspi, Maxim Shklar, Ron Ophir, Arren Bar-Even, Shirley Horn-Saban, Marilyn Safran, Eytan Domany, Doron Lancet, and Orit Shmueli. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, 21(5):650–659, 09 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti042. URL https://doi.org/10.1093/bioinformatics/bti042.

[68] Nadezda Kryuchkova-Mostacci and Marc Robinson-Rechavi. A benchmark of gene expression tissue-specificity metrics. *Briefings in Bioinformatics*, 18(2):205–214, 02 2016. ISSN 1467-5463. doi: 10.1093/bib/bbw008. URL https://doi.org/10.1093/bib/bbw008.

[69] Idan Efroni, Pui-Leng Ip, Tal Nawy, Alison Mello, and Kenneth D. Birnbaum. Quantification of cell identity from single-cell gene expression profiles. *Genome Biology*, 16(1):9, 2015. doi: 10.1186/s13059-015-0580-x. URL https://doi.org/10.1186/s13059-015-0580-x.

[70] Tabula Muris Consortium. Robject files for tissues processed by Seurat. 11 2018. doi: 10.6084/m9.figshare.5821263.v3. URL https://figshare.com/articles/dataset/Robject_files_for_tissues_processed_by_Seurat/5821263.

[71] J. Romiguier, P. Gayral, M. Ballenghien, A. Bernard, V. Cahais, A. Chenuil, Y. Chiari, R. Dernat, L. Duret, N. Faivre, E. Loire, J. M. Lourenco, B. Nabholz, C. Roux, G. Tsagkogeorga, A. A. T. Weber, L. A. Weinert, K. Belkhir, N. Bierne, S. Glémin, and N. Galtier. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, 515(7526):261–263, 2014. doi: 10.1038/nature13685. URL https://doi.org/10.1038/nature13685.

[72] Jörn M. Schmiedel, Lucas B. Carey, and Ben Lehner. Empirical mean-noise fitness landscapes reveal the fitness impact of gene expression noise. *Nature Communications*, 10(1):3180, 2019. doi: 10.1038/s41467-019-11116-w. URL https://doi.org/10.1038/s41467-019-11116-w.

[73] Naama Geva-Zatorsky, Nitzan Rosenfeld, Shalev Itzkovitz, Ron Milo, Alex Sigal, Erez Dekel, Talia Yarnitzky, Yuvalal Liron, Paz Polak, Galit Lahav, and Uri Alon. Oscillations and variability in the p53 system. *Molecular Systems Biology*, 2(1):2006.0033, 2006. doi: 10.1038/msb4100068. URL https://www.embopress.org/doi/abs/10.1038/msb4100068.

[74] Sara B. Noya, David Colameo, Franziska Brüning, Andrea Spinnler, Dennis Mircsof, Lennart Opitz, Matthias Mann, Shiva K. Tyagarajan, Maria S. Robles, and Steven A. Brown. The forebrain synaptic transcriptome is organized by clocks but its proteome is driven by sleep. *Science*, 366(6462), 2019. ISSN 0036-8075. doi: 10.1126/science.aav2642. URL https://science.sciencemag.org/content/366/6462/eaav2642.

[75] Joan Chang, Richa Garva, Adam Pickard, Ching-Yan Chloé Yeung, Venkatesh Mallikarjun, Joe Swift, David F. Holmes, Ben Calverley, Yinhui Lu, Antony Adamson, Helena Raymond-Hayling, Oliver Jensen, Tom Shearer, Qing Jun Meng, and Karl E. Kadler. Circadian control of the secretory pathway maintains collagen homeostasis. *Nature Cell Biology*, 22(1):74–86, 2020. doi: 10.1038/s41556-019-0441-z. URL https://doi.org/10.1038/s41556-019-0441-z.

[76] Michal Dudek, Constanza Angelucci, Jayalath P.D. Ruckshanthi, Ping Wang, Venkatesh Mallikarjun, Craig Lawless, Joe Swift, Karl E. Kadler, Judith A. Hoyland, Shireen R. Lamande, John F. Bateman, and Qing-Jun Meng. Circadian time series proteomics reveals daily dynamics in cartilage physiology. *bioRxiv*, 2019. doi: 10.1101/654855. URL https://www.biorxiv.org/content/early/2019/05/31/654855.

[77] Johanna Krahmer, Matthew Hindle, Laura K Perby, Tom H Nielsen, Gerben VanOoijen, Karen J Halliday, Thierry Le Bihan, and Andrew J Millar. Circadian protein regulation in the green lineage ii. the clock gene circuit controls a phospho-dawn in arabidopsis thaliana. *bioRxiv*, 2019. doi: 10.1101/760892. URL https://www.biorxiv.org/content/early/2019/09/08/760892.

[78] Oliver E. Bläsing, Yves Gibon, Manuela Günther, Melanie Höhne, Rosa Morcuende, Daniel Osuna, Oliver Thimm, Björn Usadel, Wolf-Rüdiger Scheible, and Mark Stitt. Sugars and circadian regulation make major contributions to the global regulation of diurnal gene expression in arabidopsis. *The Plant Cell*, 17(12):3257–3281, 2005. ISSN 1040-4651. doi: 10.1105/tpc.105.035261. URL http://www.plantcell.org/content/17/12/3257.

[79] Zeenat B. Noordally, Matthew M. Hindle, Sarah F. Martin, Daniel D. Seaton, T. Ian Simpson, Thierry Le Bihan, and Andrew J. Millar. Circadian protein regulation in the green lineage i. a phospho-dawn anticipates light onset before proteins peak in daytime. *bioRxiv*, 2018. doi: 10.1101/287862. URL https://www.biorxiv.org/content/early/2018/04/04/287862.

[80] Mickael Moulager, Annabelle Monnier, Béline Jesson, Régis Bouvet, Jean Mosser, Christian Schwartz, Lionel Garnier, Florence Corellou, and François-Yves Bouget. Light-dependent regulation of cell division in ostreococcus: Evidence for a major transcriptional input. *Plant Physiology*, 144(3):1360–1369, 2007. ISSN 0032-0889. doi: 10.1104/pp.107.096149. URL http://www.plantphysiol.org/content/144/3/1360.

[81] Ana C L Guerreiro, Marco Benevento, Robert Lehmann, Bas van Breukelen, Harm Post, Piero Giansanti, AF Maarten Altelaar, Ilka M Axmann, and Albert J R Heck. Daily rhythms in the cyanobacterium synechococcus elongatus probed by high-resolution mass spectrometry-based proteomics reveals a small defined set of cyclic proteins. *Molecular cellular proteomics : MCP*, 13 (8):2042—2055, August 2014. ISSN 1535-9476. doi: 10.1074/mcp.m113.035840. URL https://europepmc.org/articles/PMC4125736.

[82] Hiroshi Ito, Michinori Mutsuda, Yoriko Murayama, Jun Tomita, Norimune Hosokawa, Kazuki Terauchi, Chieko Sugita, Mamoru Sugita, Takao Kondo, and Hideo Iwasaki. Cyanobacterial daily life with kai-based circadian and diurnal genome-wide transcriptional control in synechococcus elongatus. *Proceedings of the National Academy of Sciences*, 106(33):14168–14173, 2009. ISSN 0027-8424. doi: 10.1073/pnas.0902587106. URL https://www.pnas.org/content/106/33/14168.

[83] Benner Christopher, S Gill, GC Melkani, and S Panda. Transcriptomic changes in drosophila tissues under time-restricted feeding, 2015. URL https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64108. GSE64108.

[84] Ludovic S. Mure, Hiep D. Le, Giorgia Benegiamo, Max W. Chang, Luis Rios, Ngalla Jillani, Maina Ngotho, Thomas Kariuki, Ouria Dkhissi-Benyahya, Howard M. Cooper, and Satchidananda Panda. Diurnal transcriptome atlas of a primate across major neural and peripheral tissues. *Science*, 359(6381), 2018. ISSN 0036-8075. doi: 10.1126/science.aao0318. URL https://science.sciencemag.org/content/359/6381/eaao0318.

[85] Daniel R Zerbino, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, Laurent Gil, Leo Gordon, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G Izuogu, Sophie H Janacek, Thomas Juettemann, Jimmy Kiang To, Matthew R Laird, Ilias Lavidas, Zhicheng Liu, Jane E Loveland, Thomas Maurel, William McLaren, Benjamin Moore, Jonathan Mudge, Daniel N Murphy, Victoria Newman, Michael Nuhn, Denye Ogeh, Chuang Kee Ong, Anne Parker, Mateus Patricio, Harpreet Singh Riat, Helen Schuilenburg, Dan Sheppard, Helen Sparrow, Kieron Taylor, Anja Thormann, Alessandro Vullo, Brandon Walts, Amonida Zadissa, Adam Frankish, Sarah E Hunt, Myrto Kostadima, Nicholas Langridge, Fergal J Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Dan M Staines, Stephen J Trevanion, Bronwen L Aken, Fiona Cunningham, Andrew Yates, and Paul Flicek. Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761, 11 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx1098. URL https://doi.org/10.1093/nar/gkx1098.

[86] The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, 11 2018. ISSN 0305-1048. doi: 10.1093/nar/gky1049. URL https://doi.org/10.1093/nar/gky1049.

[87] Miika Ahdesmaki, Konstantinos Fokianos, and Korbinian Strimmer. *GeneCycle: Identification of Periodically Expressed Genes*, 2012. URL https://CRAN.R-project.org/package=GeneCycle. R package version 1.1.4.

[88] Miika Ahdesmäki, Harri Lähdesmäki, Ron Pearson, Heikki Huttunen, and Olli Yli-Harja. Robust detection of periodic time series measured from biological systems. *BMC Bioinformatics*, 6(1):117, May 2005. ISSN 1471-2105. doi: 10.1186/1471-2105-6-117. URL https://doi.org/10.1186/1471-2105-6-117.

[89] Hiroshi Akashi and Takashi Gojobori. Metabolic efficiency and amino acid composition in the proteomes of escherichia coli and bacillus subtilis. *Proceedings of the National Academy of Sciences*, 99(6):3695–3700, 2002. ISSN 0027-8424. doi: 10.1073/pnas.062526999. URL https://www.pnas.org/content/99/6/3695.

[90] Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995, 11 2012. ISSN 0305-1048. doi: 10.1093/nar/gks1193. URL https://doi.org/10.1093/nar/gks1193.

[91] Nicholas E. Phillips, Aleksandra Mandic, Saeed Omidi, Felix Naef, and David M. Suter. Memory and relatedness of transcriptional activity in mammalian cell lineages. *Nature Communications*, 10(1):1208, 2019. doi: 10.1038/s41467-019-09189-8. URL https://doi.org/10.1038/s41467-019-09189-8.

[92] Leeat Keren, Jean Hausser, Maya Lotan-Pompan, Ilya Vainberg Slutskin, Hadas Alisar, Sivan Kaminski, Adina Weinberger, Uri Alon, Ron Milo, and Eran Segal. Massively parallel interrogation of the effects of gene expression levels on fitness. *Cell*, 166(5):1282–1294.e18, 2016. doi: https://doi.org/10.1016/j.cell.2016.07.024. URL http://www.sciencedirect.com/science/article/pii/S009286741630931X.

[93] Katja Baerenfaller, Catherine Massonnet, Sean Walsh, Sacha Baginsky, Peter Bühlmann, Lars Hennig, Matthias Hirsch-Hoffmann, Katharine A Howell, Sabine Kahlau, Amandine Radziejwoski, Doris Russenberger, Dorothea Rutishauser, Ian Small, Daniel Stekhoven, Ronan Sulpice, Julia Svozil, Nathalie Wuyts, Mark Stitt, Pierre Hilson, Christine Granier, and Wilhelm Gruissem. Systems-based analysis of arabidopsis leaf growth reveals adaptation to water deficit. *Molecular systems biology*, 8: 606–606, 2012. doi: 10.1038/msb.2012.39. URL https://pubmed.ncbi.nlm.nih.gov/22929616.

[94] Jialin Liu, Michael Frochaux, Vincent Gardeux, Bart Deplancke, and Marc Robinson-Rechavi. Inter-embryo gene expression variability recapitulates the hourglass pattern of evo-devo. *BMC Biology*, 18(1):129, 2020. doi: 10.1186/s12915-020-00842-z. URL https://doi.org/10.1186/s12915-020-00842-z.

[95] Michael B. Elowitz, Arnold J. Levine, Eric D. Siggia, and Peter S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002. ISSN 0036-8075. doi: 10.1126/science.1070919. URL https://science.sciencemag.org/content/297/5584/1183.

[96] Jakub Jedrak and Anna Ochab-Marcinek. Contributions to the 'noise floor' in gene expression in a population of dividing cells. *Scientific reports*, 10(1):13533–13533, 08 2020. doi: 10.1038/s41598-020-69217-2. URL https://pubmed.ncbi.nlm.nih.gov/32782314.

[97] Yuichi Taniguchi, Paul J. Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X. Sunney Xie. Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–538, 2010. ISSN 0036-8075. doi: 10.1126/science.1188308. URL https://science.sciencemag.org/content/329/5991/533.

[98] Olin K. Silander, Nela Nikolic, Alon Zaslaver, Anat Bren, Ilya Kikoin, Uri Alon, and Martin Ackermann. A genome-wide analysis of promoter-mediated phenotypic noise in escherichia coli. *PLOS Genetics*, 8(1):e1002443–, 01 2012. URL https://doi.org/10.1371/journal.pgen.1002443.

[99] Gene-Wei Li, David Burkhardt, Carol Gross, and Jonathan S. Weissman. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, 157(3):624–635, 2020/09/23 2014. doi: 10.1016/j.cell.2014.02.033. URL https://doi.org/10.1016/j.cell.2014.02.033.

[100] John R. S. Newman, Sina Ghaemmaghami, Jan Ihmels, David K. Breslow, Matthew Noble, Joseph L. DeRisi, and Jonathan S. Weissman. Single-cell proteomic analysis of s. cerevisiae reveals the architecture of biological noise. *Nature*, 441(7095):840–846, 2006. doi: 10.1038/nature04785. URL https://doi.org/10.1038/nature04785.

[101] David E. Weinberg, Premal Shah, Stephen W. Eichhorn, Jeffrey A. Hussmann, Joshua B. Plotkin, and David P. Bartel. Improved ribosome-footprint and mrna measurements provide insights into dynamics and regulation of yeast translation. *Cell Reports*, 14(7):1787–1799, 2016. doi: https://doi.org/10.1016/j.celrep.2016.01.043. URL http://www.sciencedirect.com/science/article/pii/S2211124716300213.

[102] Brian P. H. Metzger, David C. Yuan, Jonathan D. Gruber, Fabien Duveau, and Patricia J. Wittkopp. Selection on noise constrains variation in a eukaryotic promoter. *Nature*, 521(7552):344–347, 2015. doi: 10.1038/nature14244. URL https://doi.org/10.1038/nature14244.

[103] Robert Horvath, Benjamin Laenen, Shohei Takuno, and Tanja Slotte. Single-cell expression noise and gene-body methylation in arabidopsis thaliana. *Heredity*, 123(2):81–91, 2019. doi: 10.1038/s41437-018-0181-z. URL https://doi.org/10.1038/s41437-018-0181-z.

[104] Alma Dal Co, Marco Cosentino Lagomarsino, Michele Caselle, and Matteo Osella. Stochastic timing in gene expression for simple regulatory strategies. *Nucleic acids research*, 45(3):1069–1078, 02 2017. doi: 10.1093/nar/gkw1235. URL https://pubmed.ncbi.nlm.nih.gov/28180313.

[105] Keren Bahar Halpern, Sivan Tanami, Shanie Landen, Michal Chapal, Liran Szlak, Anat Hutzler, Anna Nizhberg, and Shalev Itzkovitz. Bursty gene expression in the intact mammalian liver. *Molecular Cell*, 58(1):147–156, 2020/09/24 2015. doi: 10.1016/j.molcel.2015.01.027. URL https://doi.org/10.1016/j.molcel.2015.01.027.

[106] Shasha Chong, Chongyi Chen, Hao Ge, and X. Sunney Xie. Mechanism of transcriptional bursting in bacteria. *Cell*, 158(2):314–326, 2020/09/24 2014. doi: 10.1016/j.cell.2014.05.038. URL https://doi.org/10.1016/j.cell.2014.05.038.

[107] Susan Duncan and Stefanie Rosa. Gaining insight into plant gene transcription using smfish. *Transcription*, 9(3):166–170, 2018. doi: 10.1080/21541264.2017.1372043. URL https://doi.org/10.1080/21541264.2017.1372043. PMID: 28990856.

[108] Roy D. Dar, Sydney M. Shaffer, Abhyudai Singh, Brandon S. Razooky, Michael L. Simpson, Arjun Raj, and Leor S. Weinberger. Transcriptional bursting explains the noise–versus–mean relationship in mrna and protein levels. *PLOS ONE*, 11(7):e0158298–, 07 2016. URL https://doi.org/10.1371/journal.pone.0158298.

[109] Iksoo Huh, Jia Zeng, Taesung Park, and Soojin V. Yi. Dna methylation and transcriptional noise. *Epigenetics & Chromatin*, 6(1):9, 2013. doi: 10.1186/1756-8935-6-9. URL https://doi.org/10.1186/1756-8935-6-9.

[110] Johan Paulsson, Otto G. Berg, and Måns Ehrenberg. Stochastic focusing: Fluctuation-enhanced sensitivity of intracellular regulation. *Proceedings of the National Academy of Sciences*, 97(13):7148–7153, 2000. ISSN 0027-8424. doi: 10.1073/pnas.110057697. URL https://www.pnas.org/content/97/13/7148.

[111] Murat Acar, Jerome T Mettetal, and Alexander van Oudenaarden. Stochastic switching as a survival strategy in fluctuating environments. *Nature Genetics*, 40(4):471–475, 2008. doi: 10.1038/ng.110. URL https://doi.org/10.1038/ng.110.

[112] Luise Wolf, Olin K Silander, Erik van Nimwegen, and Ido Golding. Expression noise facilitates the evolution of gene regulation. *eLife*, 4:e05856, 2015. doi: 10.7554/eLife.05856. URL https://doi.org/10.7554/eLife.05856.

[113] Arantxa Urchueguía, Luca Galbusera, Gwendoline Bellement, Thomas Julou, and Erik van Nimwegen. Noise propagation shapes condition-dependent gene expression noise in escherichia coli. *bioRxiv*, page 795369, 01 2019. doi: 10.1101/795369. URL http://biorxiv.org/content/early/2019/10/07/795369.abstract.

[114] J. Patrick Pett, Anja Korenčič, Felix Wesener, Achim Kramer, and Hanspeter Herzel. Feedback loops of the mammalian circadian clock constitute repressilator. *PLOS Computational Biology*, 12(12):1–15, 12 2016. doi: 10.1371/journal.pcbi.1005266. URL https://doi.org/10.1371/journal.pcbi.1005266.

[115] Jana Stöckel, Eric A. Welsh, Michelle Liberton, Rangesh Kunnvakkam, Rajeev Aurora, and Himadri B. Pakrasi. Global transcriptomic analysis of cyanothece 51142 reveals robust diurnal oscillation of central metabolic processes. *Proceedings of the National Academy of Sciences*, 105(16):6156–6161, 2008. ISSN 0027-8424. doi: 10.1073/pnas.0711068105. URL https://www.pnas.org/content/105/16/6156.

[116] Andrew Beale, Christophe Guibal, T. Katherine Tamai, Linda Klotz, Sophie Cowen, Elodie Peyric, Víctor H. Reynoso, Yoshiyuki Yamamoto, and David Whitmore. Circadian rhythms in mexican blind cavefish astyanax mexicanus in the lab and in the field. *Nature Communications*, 4(1):2769, 2013. doi: 10.1038/ncomms3769. URL https://doi.org/10.1038/ncomms3769.

[117] Julien Fumey, Hélène Hinaux, Céline Noirot, Claude Thermes, Sylvie Rétaux, and Didier Casane. Evidence for late pleistocene origin of astyanax mexicanus cavefish. *BMC Evolutionary Biology*, 18(1):43, 2018. doi: 10.1186/s12862-018-1156-7. URL https://doi.org/10.1186/s12862-018-1156-7.

[118] I. King Jordan, Igor B. Rogozin, Yuri I. Wolf, and Eugene V. Koonin. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Research*, 12(6):962–968, 2002. doi: 10.1101/gr.87702. URL http://genome.cshlp.org/content/12/6/962.abstract.

[119] Han Liang and Wen-Hsiung Li. Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends in Genetics*, 23(8):375–378, 2007. doi: https://doi.org/10.1016/j.tig.2007.04.005. URL http://www.sciencedirect.com/science/article/pii/S0168952507001540.

[120] Tim Wang, Kıvanç Birsoy, Nicholas W. Hughes, Kevin M. Krupczak, Yorick Post, Jenny J. Wei, Eric S. Lander, and David M. Sabatini. Identification and characterization of essential genes in the human genome. *Science*, 350(6264):1096–1101, 2015. ISSN 0036-8075. doi: 10.1126/science.aac7041. URL https://science.sciencemag.org/content/350/6264/1096.

[121] Takashi Makino, Karsten Hokamp, and Aoife McLysaght. The complex relationship of gene duplication and essentiality. *Trends in Genetics*, 25(4):152–155, 2009. doi: https://doi.org/10.1016/j.tig.2009.03.001. URL http://www.sciencedirect.com/science/article/pii/S0168952509000389.

[122] Martin Gardner. Mathematical games - the fantastic combinations of john conway's new solitaire game 'life. *Scientific American*, 223(4):120–123, 1970. doi: 10.1038/scientificamerican1070-120.

[123] Frank Jakobus Rühli and Maciej Henneberg. New perspectives on evolutionary medicine: the relevance of microevolution for human health and disease. *BMC Medicine*, 11(1):115, 2013. doi: 10.1186/1741-7015-11-115. URL https://doi.org/10.1186/1741-7015-11-115.

[124] Akhilesh B. Reddy, Natasha A. Karp, Elizabeth S. Maywood, Elizabeth A. Sage, Michael Deery, John S. O'Neill, Gabriel K. Y. Wong, Jo Chesham, Mark Odell, Kathryn S. Lilley, Charalambos P. Kyriacou, and Michael H. Hastings. Circadian orchestration of the hepatic proteome. *Current Biology*, 16(11):1107–1115, 2006. doi: https://doi.org/10.1016/j.cub.2006.04.026. URL http://www.sciencedirect.com/science/article/pii/S0960982206014874.

[125] Victoria Y. Gorbacheva, Roman V. Kondratov, Renliang Zhang, Srujana Cherukuri, Andrei V. Gudkov, Joseph S. Takahashi, and Marina P. Antoch. Circadian sensitivity to the chemotherapeutic agent cyclophosphamide depends on the functional status of the clock/bmal1 transactivation complex. *Proceedings of the National Academy of Sciences of the United States of America*, 102(9):3407, 03 2005. doi: 10.1073/pnas.0409897102. URL http://www.pnas.org/content/102/9/3407.abstract.

[126] Marc D. Ruben, Gang Wu, David F. Smith, Robert E. Schmidt, Lauren J. Francey, Yin Yeng Lee, Ron C. Anafi, and John B. Hogenesch. A database of tissue-specific rhythmically expressed human genes has potential applications in circadian medicine. *Science Translational Medicine*, 10(458), 2018. ISSN 1946-6234. doi: 10.1126/scitranslmed.aat8806. URL https://stm.sciencemag.org/content/10/458/eaat8806.

[127] David Montaigne, Xavier Marechal, Thomas Modine, Augustin Coisne, Stéphanie Mouton, Georges Fayad, Sandro Ninni, Cédric Klein, Staniel Ortmans, Claire Seunes, Charlotte Potelle, Alexandre Berthier, Celine Gheeraert, Catherine Piveteau, Rebecca Deprez, Jérome Eeckhoute, Hélène Duez, Dominique Lacroix, Benoit Deprez, Bruno Jegou, Mohamed Koussa, Jean-Louis Edme, Philippe Lefebvre, and Bart Staels. Daytime variation of perioperative myocardial injury in cardiac surgery and its prevention by rev-erbα antagonism: a single-centre propensity-matched cohort study and a randomised study. *The Lancet*, 391(10115):59–69, 2020/09/26 2018. doi: 10.1016/S0140-6736(17)32132-3. URL https://doi.org/10.1016/S0140-6736(17)32132-3.

REFERENCES

Etude de l'expression des gènes nycthéméraux à la lumière de l'évolution

Résumé grand public

Depuis des millions d'années, le vivant s'est progressivement synchronisé sur le rythme de l'alternance du jour et de la nuit. La terre met environ 24 heures pour réaliser une rotation complète, donnant une alternance quotidienne d'exposition à la lumière du soleil. Ainsi, la nature est gouvernée par des cycles énergétiques dus aux cycles lumière-obscurité.

L'expression des gènes, c'est-à-dire l'ensemble des processus qui transforment l'information génétique qui se trouve dans l'ADN en protéines biologiquement fonctionnelles, est une première étape de régulation des activités biologiques. Dans nos cellules, on constate que de très nombreux gènes sont concernés par une utilisation périodique. Au cours de cette thèse, je me suis particulièrement intéressé à ces gènes, c'est-à-dire à ceux qui sont exprimés de façon périodique, toutes les 24 heures, période appelée nycthémérale. En effet, certains gènes sont plus utilisés en début de journée, d'autres plus tard dans la journée, d'autres en début de nuit, d'autres encore à la fin de la nuit. Ainsi, la production de nombreuses protéines dans les cellules oscille avec une périodicité de 24h. Ces processus périodiques permettent à chaque cellule de produire en grande quantité, à certaines heures du jour ou de la nuit, des protéines qui assurent la même fonction ou des fonctions compatibles.

Le rythme circadien est un système autorégulé, capable de s'auto-générer, fournissant aux organismes la capacité d'anticiper les changements de leur environnement sur une échelle de temps de 24 heures. Un gène rythmique est un gène qui présente une variation quotidienne de l'abondance de sa molécule intermédiaire transcrite (l'ARN) ou de sa protéine traduite. Ces variations quotidiennes peuvent également être entraînées directement ou indirectement par des facteurs environnementaux tels que l'alternance des périodes de lumière-obscurité, de l'alimentation, des variations de température ou des activités sociales. C'est pourquoi il y a de nombreux gènes dont l'expression rythmique n'est pas autonome, mais directement ou indirectement entraînée par l'environnement lui-même. Une telle expression périodique des gènes est retrouvée presque partout dans la Nature : chez les animaux, les plantes, les bactéries et les champignons.

20% à 50% des protéines sont produites périodiquement (toutes les 24 heures) à partir d'ARN présent en quantité constante. Et inversement, il existe de nombreux ARN dont l'abondance est périodique alors que ce n'est pas le cas pour leurs protéines. Mais alors, pourquoi ?

Mes résultats suggèrent que les variations quotidiennes concernent des protéines produites en quantité relativement abondante. La production en grande quantité de ces protéines est coûteuse pour la cellule. En effet, la fabrication de protéines nécessite une certaine somme d'énergie et de matériaux moléculaires que la cellule n'a pas forcément en quantité infinie. De surcroit, ces protéines périodiques seraient, en effet, encore plus coûteuses à produire si la cellule avait dû maintenir en permanence (i.e. de manière constante) un niveau suffisamment élevé de protéines pour assurer leur fonction biologique. En effet, les coûts de production des protéines sont suffisamment conséquents pour être soumis à la sélection naturelle. A contrario, les coûts de production des ARN sont probablement trop négligeables pour soutenir l'hypothèse que leur variation serait due à une stratégie d'économie pour la cellule. Par contre, pour un gène donné, la quantité de son ARN joue un rôle dans la variabilité du nombre de ses protéines entre les cellules. Ceci s'explique car la quantité de protéines n'est pas exactement la même dans chaque cellule. Plus la production de protéines par unité d'ARN est grande, plus la variabilité entre cellules est faible. Il semblerait que les gènes qui ont des ARN qui varient quotidiennement ont en moyenne une variabilité entre cellules plus faible que celle des autres gènes.

Au cours des millions d'années d'évolution du vivant, des individus ont présenté, par hasard, des modifications qui, dans leur environnement, leur ont procuré un avantage par rapport aux autres individus. C'est le principe de sélection naturelle. Pour certains gènes, leur fonction requiert un niveau élevé de protéines. Les gènes qui produisent périodiquement des protéines sont des gènes dont la quantité requise en protéines s'est avérée être coûteuse pour la cellule, et de fait, ce coût a donc été soumis à la sélection naturelle. Il est possible que la variation périodique des ARN dont les fonctions biologiques sont sensibles à de grandes différences de quantité de protéines entre les cellules (variabilité), ait également apporté un jour, un avantage suffisamment important à l'individu dans son environnement cyclique pour être sélectionné au cours de l'évolution.

Comprendre les systèmes périodiques au sein du monde vivant nous aide à mieux appréhender les relations étroites que nous entretenons avec notre environnement. Cela passe par une compréhension des dynamiques temporelles qui s'opèrent au sein de nos cellules. La chrono-médecine, et peut-être la chrono-chirurgie, prendront un jour en compte le « tic-tac » qui règne dans l'expression de nos gènes ainsi que dans l'organisation temporelle des événements biologiques au cœur de nos cellules.