Hunting for Clues of the Circadian Clock in High Throughput and Genomic Data

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften (Dr.rer.nat.)
im Fach Bioinformatik

eingereicht an der
Fakultät für Mathematik und Informatik
der
Freien Universität zu Berlin

von
Robert Lehmann M.Sc.

Juni 2015

Gutachter:
1. Prof. Dr. Martin Vingron
2. Prof. Dr. Hanspeter Herzel

Eingereicht am: 04.06.2015
Tag der mündlichen Prüfung: 08.04.2016

Contents

1 Introduction

> Life is a cyclical chemical process that is regulated in four dimensions.
>
> — Jay C. Dunlap, Molecular Bases for Circadian Clocks

1.1 Life in rhythmic environments

The vast majority of life on earth is subjected to rhythmic influences. Many organisms experience periodical changes of sometimes radical nature in their environment. In adaptation to these external rhythmic changes in the environment, most organisms inhabiting the surface of our planet have developed internal molecular clock mechanisms to predict these changes. Due to the rotation of the earth, plants can only perform photosynthesis and produce energy during the day. Surplus energy is invested into storage metabolites, which they then use for energy production via respiration during night. The amount of available light is similarly important for animals, which are commonly adapted to be active during day (diurnal) or night (nocturnal). This results in alternating phases of high activity and resting, which are synchronised to the light phases. Furthermore, adaptations to seasonal changes are found, such as hibernation. This influence extends even to the popular bacterial model organism *Escherichia coli*, which is found in the intestines of warm-blooded organisms and is thus subject to the rhythmic food intake. In general, such a clock allows the adjustment of physiological parameters for peak performance in anticipation of a recurring challenge, be it the time for hunting of a predator or the time for hiding for its prey [1].

1.2 Three circadian clock oscillators

The three defining properties of the circadian clock A circadian clock must fulfill three basic requirements in order to provide a useful internal reference for the organism. Firstly, a free-running period, *i.e.* without outer force modifying the oscillation period, of approximately 24 h is necessary which approximates the external light rhythm (the term is derived from "circa diem," or "about a day"). This allows mammals, which all feature a circadian clock, to regulate processes such as the sleep-wake cycle and metabolism. Also plants can predict the start of the light phase in order to spread their leaves and thus prepare for starting photosynthesis. Secondly, the clock must be temperature compensated to prevent differing periods between summer and winter.

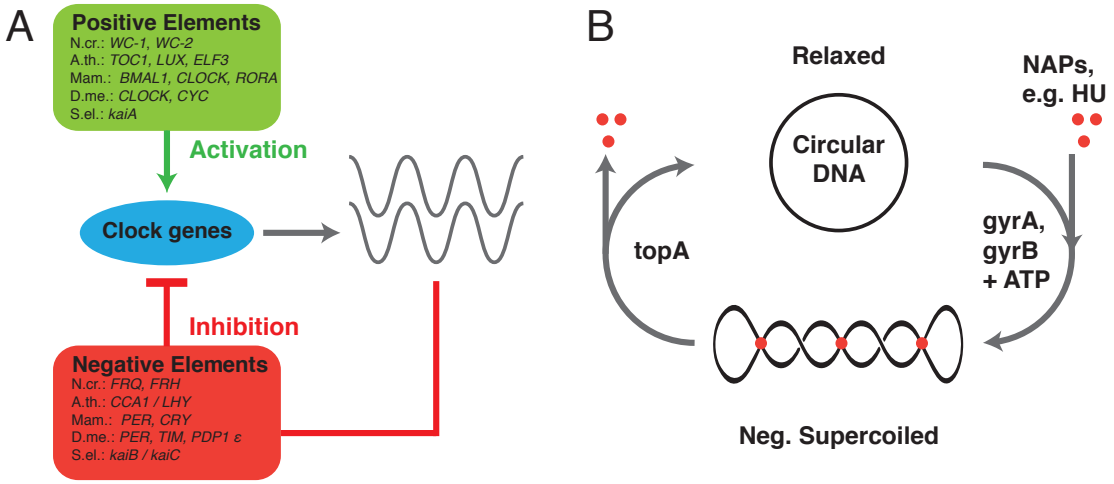


Fig. 1.1: **The core Circadian Clock mechanism based on a Transcription-Translation negative feedback loop oscillator (TTO), and a proposed metabolic oscillator mechanism based on chromosomal structure changes.** A) Constitutively expressed core clock transcription factors of the positive arm (green) bind to other core clock genes (blue), which are then transcribed (grey) and translated into transcription factors, which function as the negative arm (red) to suppress expression of core clock genes. B) The mechanisms of chromosomal structure regulation in *Escherichia coli* via topoisomerase enzymes and nucleoid associated proteins in combination with the observed circadian changes in the supercoiling state of the chromosome in the cyanobacterium *Synechococcus elongatus* provide the necessary building blocks for a metabolic oscillator. The topoisomerase enzymes gyrA and gyrB introduce negative supercoiling and thereby a denser chromosomal structure under energy consumption (ATP). The counteracting enzyme topA relaxes supercoiling without the need of energy. The relaxation inhibits the expression of *topA* and induces the expression of *gyrA* and *gyrB*. The nucleoid associated proteins, such as HU or H-NS, can introduce or stabilise chromosomal structures, such as the depicted plectroneme (common in eubacteria) or toroidal structures, as found for eukaryotic and archaebacterial DNA wound around nucleosomes.

However, the phase of any biochemical oscillator cannot be perfectly robust against intracellular perturbations. Also, organisms might move into zones with shifted rhythms of light and temperature. Therefore, a third feature is required, the ability to reset its phase and to entrain to a rhythmic external zeitgeber, *e.g.* light or temperature. The term zeitgeber was introduced in the 1950s by Prof. J. Aschoff, referring to an external rhythmic parameter used as reference to synchronise the internal clock of an organism. Following the terminology of Prof. J. Dunlap, one can distinguish between the circadian oscillator, the molecular core clock mechanism on the one hand, and the cellular processes regulated by the clock as circadian system on the other [44].

Structure of the mammalian core clock The classical molecular structure of the circadian oscillator is the Transcription-Translation Oscillator feedback loop (TTO) [179],

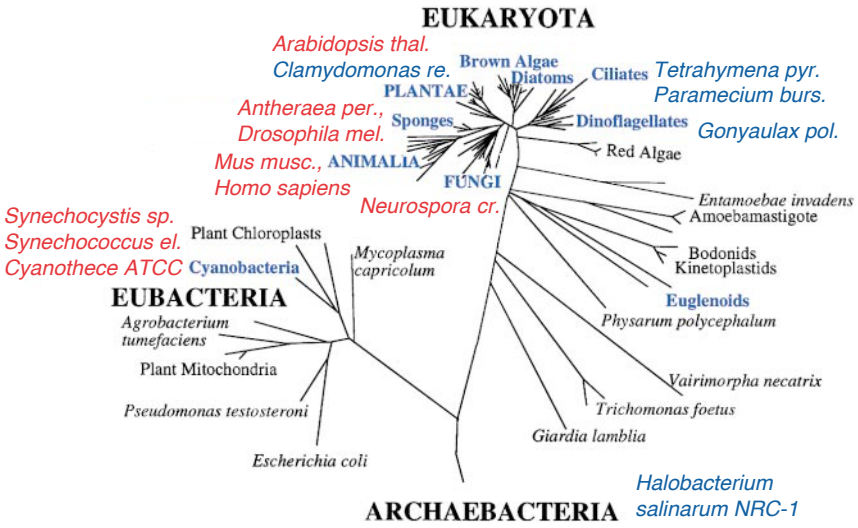


Fig. 1.2: **Circadian rhythms are found across the phylogenetic tree of life.** The early phylogenetic tree with three kingdoms (Eukaryota, Eubacteria, Archaebacteria) based on small subunit rRNA genes is shown (adapted from [44]). Phylogenetic groups where circadian rhythms have been described are indicated in blue, while species with well studied circadian clock mechanisms are shown in red. The archaebacterial kingdom is indicated, featuring one species with described circadian rhythm.

a circular series of steps involving the transcription of core clock genes from the DNA template to RNA, translation of RNA into proteins, and regulation of transcription by DNA-binding proteins. It comprises positive elements, transcriptional activators of the clock genes, as well as negative elements (see Figure 1.1 A). The names of the particular genes are provided for *Neurospora crassa* (N.cr.), *Arabidopsis thaliana* (A.th.), *Mammalia* (Mam.), *Drosophila Melanogaster* (D.me.), and *Synechococcus elongatus* (S.el.). The basic structure of the circadian clock mechanism is shared amongst eukaryotes and will be illustrated in the following by means of the mammalian clock. The basis of the mammalian clockwork are the two transcription factors CLOCK and BMAL1, which contain basic helix–loop–helix PAS (Period–Arnt–Single-minded) domains, which activate transcription and thereby provide the drive to the system. Negative feedback is realised via rhythmic inhibition of the CLOCK–BMAL1 induction signal by negative regulators. Specifically, CLOCK–BMAL1 heterodimers activate the rhythmic transcription of the three period genes (in mouse mPer1–mPer3) and two cryptochrome genes (mCry1, mCry2). The resulting mPER and mCRY proteins translocate back into the nucleus where they act as repressors by directly interacting with CLOCK and/or BMAL1 to inhibit transcription, closing the negative feedback loop. The positive feedback loop is formed by the rhythmic regulation of Bmal1 transcription, whose mRNA levels peak 12 h out of phase relative to mPer and mCry RNAs. The underlying the three-component loop mechanism, involving a negative feedback loop with a delay and a non-linearity, has since been identified in variety of different biochemical oscillators, including somitogenesis, DNA damage

response via p53, cellular stress response (NF-κB), and the synthetic repressilator [3].

Cyanobacteria possess a different clock mechanism Clock systems have been found frequently amongst various members of the Eukaryota and Eubacteria kingdoms (see Figure 1.2), while Archaebacteria were left out. However, the first member of the Archaebacteria kingdom was recently discovered to exhibit circadian rhythms in transcription and metabolism. It was long believed that a TTO is the universal circadian clock mechanism [44]. However, cyanobacteria constitute an exception to the paradigm of the TTO as core clock mechanism. While a Transcription-Translation Oscillator exists with kaiA as positive and kaiC as negative element [71], it was shown only recently that the TTO is a slave oscillator to an underlying master oscillator [176]. This Post-Transcriptional Oscillator (PTO) is based on the phosphorylation state of the proteins kaiA, kaiB, and kaiC. Due to its independence of the cellular machinery, circadian oscillation can even be observed in a test tube merely containing the kai proteins and ATP [145]. Interestingly, the set of *kai* genes and also the composition of the clock is not conserved across cyanobacterial species [45]. While members of the *Prochlorococcus* group appear to have compensated the loss of the *kaiA* gene by genome reduction with an altered "hourglass" clock mechanism which requires periodic input of its zeitgeber [8], *Gloeobacter violaceus* PCC 7421 does not possess any *kai* genes and thus displays no circadian rhythmicity [9]. Various other strains possess multiple copies of *kaiB* and *kaiC*. The transcriptional regulation of *kaiBC* is required for the regulation of the cyanobacterial circadian clockwork [265]. In *Synechococcus elongatus*, the three clock genes, *kaiA*, *kaiB* and *kaiC* are arranged as three adjacent genes. The *kaiB* and *kaiC* genes are arranged in a dicistronic operon, while the *kaiA* gene possesses a separate promoter. The transcript of kaiA is rhythmically abundant while the respective protein is constant [82]. In contrast, both abundances of the *kaiBC* transcripts and the KaiB and KaiC proteins exhibit circadian cycles [81, 84, 256]. Due to this diversity it is currently not clear which strains possess a stable circadian clock and which processes are clock regulated. There are currently five genes *SasA*, *RpaA*, *RpaB*, *LabA*, and *CikA* described as clock output pathways [71].

Metabolic circadian clocks - The original clock? Only recently it was reported that one archaebacterial species, namely *Halobacterium salinarum* NRC-1, does exhibit diurnal expression patterns [250], possibly to anticipate the higher available oxygen levels at night in its highly saline environment. Additional support was provided by the extraordinary observation by Edgar *et al.* that in all considered species with circadian clocks, a second metabolic circadian oscillator (oxidation-reduction cycles of peroxiredoxin proteins) was observed [46], including *Halobacterium salinarum* NRC-1. Presently, the molecular mechanism of this rhythmicity, be it a TTO or PTO or an entirely new mechanism, remains to be elucidated. But already now it is clear that daily timekeeping evolved independently within the different lineages, resulting in the cyanobacterial, fungal, insect, and mammalian clocks. Edgar *et al.* offer a compelling reverse interpretation, according to which the metabolic oscillator constitutes the ancestral underlying clock

which could have given rise to different TTO and PTO oscillator architectures [46].

1.2.1 Circadian rhythms in metabolism, expression, and chromosomal structure

Metabolic oscillations - Not only circadian Further light on the relationship of metabolic oscillators, TTOs, and PTOs was shed by another recent experiment by O'Neill *et al.*, which demonstrated that the cycles in oxidation-reduction state of human peroxiredoxin proteins, as well as haemoglobin tetramer–dimer transitions, and NADH/NADPH ratio persist independently from transcriptional activity, and thus the TTO, in red blood cells [157]. This supports the idea of the metabolic oscillator as the underlying mechanism for the circadian TTO / PTO. They further hypothesise that the binding affinity modulation of the Clock and Bmal1 transcription factors (TF) by the NADH/NADPH ratio can serve as coupling mechanism from metabolic to the TTO oscillator. A similar hypothesis was formulated by Wang *et al.*, hypothesising that the described circadian redox-rhythms in neurons of the suprachiasmatic nucleus of rats [240] could modify the respective neuronal firing rate. The yeast strain *Saccharomyces cerevisiae* provides another interesting example for the connection of metabolic oscillations to transcriptional regulation. Under continuous culture conditions, ultradian oscillations between respirative and fermentative phases can be observed [141, 228], both of which induce a specific transcriptional program [127].

Oscillations in chromosomal structure - Another metabolic oscillator? A third circadian oscillator mechanism was proposed based on structural changes of the cyanobacterial chromosome (see Figure 1.1 B). The main observation was that the chromosome of *Synechococcus elongatus* PCC 7942 exhibits circadian cycles of elevated and reduced supercoiling, which in turn result in rhythmical transcriptional changes [237]. A mechanistic explanation could be a negative feedback loop driven by the elevated energy availability during the photic period, which allows the ATP-dependent type-II topoisomerases gyrA and gyrB to introduce supercoiling into the chromosome and thereby regulate the expression of supercoiling-sensitive genes. The increasing negative supercoiling inhibits the expression of *gyrA* and *gyrB* and induces the supercoiling-removing type-I topoisomerase *topA*, thereby closing a negative feedback loop which ensures that the chromosomal structure returns to the starting condition. A range of proteins with DNA-binding domains, so called nucleoid-associated proteins (NAPs), are able to stabilise certain DNA conformations and even induce DNA structural changes [40].

Supercoiling, nucleoid compaction, and periodic DNA sequence motifs In its relaxed form, the DNA double strand is made up of 10.5 bp long helix subunits which turn for a full 360°. Generally, both eukaryotic and bacterial DNA is found in negatively supercoiled form, *i.e.* slightly undertwisted and thus with subunits of >10.5 bp length [221]. The compaction of the nucleoid, the prokaryotic equivalent of the nucleus, caused by increasing supercoiling has been modelled for *E. coli* [211]. It is well

known that changes in the supercoiling state can alter gene expression levels, *e.g.* for the supercoiling-related topoisomerase enzymes *gyrA*, *gyrB*, and *topA* in *E. coli* [87, 164]. As laid out in the following, there are indications that the preferred supercoiling state of a stretch of DNA can leave marks in the nucleotide sequence.

Sequence motif periodicity, *i.e.* the bias in nucleotide spaces along the DNA, was reported for a range of genome sequences since the 1980s [222]. Eukaryotes and archaea feature signals with a period of 10–10.5 bp which reflect the necessary pitch to wind DNA helices around nucleosomes. However, the causes and effects of ~11 bp period signals in bacterial genomes are less well studied and understood [72, 139, 190, 218, 254]. While dinucleotides usually yield a stronger signal than mononucleotides, the combinations of A and T (WW in IUPAC notation) often yield the strongest signal [139]. This is commonly interpreted with the mechanical properties of the different dinucleotides. Short runs of A and T nucleotides excluding the TpA step called AT-tract motif induce a bend of the DNA backbone into the minor groove of the helix. If these AT-tracts are evenly spaced along the DNA strand and in phase with the ~10.5 bp sections of the DNA double helix, this axial bend can induce an intrinsic curvature of the DNA double helix [183].

However, there is an alternative hypothesis of the connection of sequence periodicity and the chromosomal structure. There are numerous reports of nucleoid-associated proteins (NAP) that can stabilise folding of DNA into specific structures such as plectonemes or tori [131]. For several of these NAPs, binding preferences to AT-rich DNA, AT-tracts, but also curved DNA have been observed [221]. Also, similar to nucleosomes in eukaryotes and archaea, the signal might be related to the wrapping of the DNA around multimeres of the NAP HU [40].

1.3 Quantifying rhythms

Early experiments detecting circadian rhythms Clock controlled biological processes have been found in a wide range of species. In the early days of chronobiology, mainly high level traits have been studied. Rodents or large walking insects like cockroaches, were used to conveniently measure wheel running behaviour under different conditions. Indeed, cockroaches allowed for one compelling experiment proving the location of the circadian pacemaker in the optic lobes. Isolation of the optic lobe from the brain deprives the individual from circadian rhythmicity. It was found by chance, like so many scientific discoveries, that cockroaches can regenerate the injured neuronal connections which then restores rhythmic behaviour. Using two strains with different circadian period (one shorter and one longer than 24 h), it was then shown that the transplantation of the optic lobes is sufficient to change the circadian period of individuals of one strain into the period of the other [158]. But also the activity of songbirds was readily measured by detecting their jumping from one perch to another. Even lizards, while walking around in the cage will tilt the cage from left to right, which is detected by a switch at its bottom.

Plants and fungi have played an important role in understanding the basics of the circadian clock. Fritz Went showed in 1960 that when providing similar total absolute length of dark

and light phases but with varying circadian period, the African violet *Saintpaulina* grew fastest with around 24h period [248]. Plants that exhibit leaf movements (closing at night, opening during the day) were the prime experimental models for a while (*e.g. Kalanchoe*, *Mimosa*, and Tobacco). Contemporarily, the fungus *Neurospora crassa* was shown to exhibit temperature-compensated circadian rhythms with the focus on establishing a genetically tractable clock model organism [169]. However, one of the most prominent examples used humans as subjects. In the bunker experiments of Aschoff and Pohl [6], it was shown that sleep-wake cycles in humans underlie circadian regulation and thus persist independently of external light or temperature rhythms as zeitgeber.

Molecular biology permits finding cyanobacterial circadian clock The observation of rhythmic activity or motion has been supplemented by a rich arsenal of molecular biological methods, leading to new insights. A good example are cyanobacteria, which have long and persistently been assumed to not possess the means for a circadian clock due to their simple makeup [65]. Another assumption was that generation time should be equal or longer than the period of a cellular oscillator and its zeitgeber, thereby disqualifying most procaryotes for circadian rhythmicity [168]. Nevertheless, a circadian rhythm of nitrogen fixation in the cyanobacterium *Synechococcus* RF-1 was observed in Prof. Huang's research group [62]. Only relatively recently, this striking phenotype led to further research and, with the help of molecular biology, to the discovery of the kaiABC clock [102]. Inspired by this discovery, Liu *et al.* [122] attempted to quantify the extent of diurnal expression with the following experiment. A fluorescent reporter gene was introduced at random positions into the chromosome of *Synechococcus elongatus* PCC 7942. Selecting cells with measurable bioluminescence then yielded instances where the reporter is active, *i.e.* the gene was inserted close to an active promotor. Surprisingly, a large fraction of cells with oscillating brightness over time was observed, which was interpreted as extensive diurnal transcription. To extend this single-gene based approach, the expression of the entire transcriptome can be measured in parallel over extended periods of time using microarrays, allowing for more intricate analyses. A successful strategy to elucidate clock regulatory networks was to first detect genes with circadian expression pattern, and then analyse the corresponding promoter regions for overrepresented transcription factor binding sites [26, 103, 178]. While experiments with cyanobacteria yielded large numbers of circadian expressed genes, the definition of distinct motifs within the promoter regions, determining the peak phase, proved to be difficult. Initially, various sequence fragments and promoter regions were found to correlate with circadian expression phases in specific genes [109, 137, 236]. Detailed analyses of these fragments, however, indicated mainly the sequence composition as distinguishing feature but did not reveal specific motifs.

How the clock regulates transcription of its target genes Microarrays have now been replaced by more powerful deep sequencing techniques. Simultaneously, deep sequencing has also supplemented the theoretical sequence-based transcription factor binding site analysis with the ChIP-seq protocol (Chromatin Immuno-Precipitation

Sequencing) to experimentally locate DNA binding sites for a large number of proteins. Studies combining both methods have been very successful in elucidating the circadian clock network and its targets, *e.g.* in mammals [100, 180] and the fungus *Neurospora crassa* [188, 201]. However, even experimental determination of transcription factor binding sites does not allow the straightforward association of transcription factors to target genes. Systematic studies such as that by Sikora-Wohlfeld *et al.* have assessed different methods of assigning ChIP-seq peaks to target genes using microarray expression data for comparison [199]. Specifically, they employ the prediction of differential expression measured by 13 HemoChIP and 8 ESChIP datasets between wildtype and TF-knockout cells as performance measure of the different algorithms, which reflects only the steady states with TF present or absent. The most important concepts of peak-target association are the genomic distance of peak and transcriptional start site (TSS), the number of peaks found in proximity of a gene, and the strength of the peak signal. While the obtained results are encouraging, such methods are not able to completely model expression regulation by TFs. One obvious shortcoming of current methods is, that distant enhancers or repressors are not compatible with the idea to assign peaks to the closest TSS. Another reason is, that gene expression involves subsequent post-transcriptional modification of the nascent RNA. This issue was addressed experimentally in the study by Menet *et al.*, which disentangled the contribution of transcription and post-transcriptional regulation in circadian gene expression [135]. They provide a direct comparison of the amount of nascent RNA with the respective mRNA in mouse liver. Surprisingly, about 70% of genes with circadian oscillating mRNA did not show transcriptional rhythms. This result suggests a much more prominent role of post-transcriptional regulation. This picture was supported by a followup study on *Drosophila* head samples, also presented by the group of Prof. M. Rosbash [182]. A related more extensive work from the group of Prof. F. Naef probed the circadian transcriptional process in mouse liver at three levels, epigenetic modification, RNA polymerase II binding, and resulting mRNA abundance [114]. It was found, that RNA polymerase II binding occurs rhythmically, as opposed to constant binding into an arrested state with rhythmic release into an active state. A similar conclusion of extensive involvement of post-transcriptional regulation in circadian rhythm generation was reached. Another surprising finding was that the epigenetic landscape appeared to be dynamically remodelled during the 24 h cycle, a process previously thought to be non-rhythmic and dynamic only on longer time scales.

1.4 Difficulties in connecting the core clock to cellular processes

Temporal separation of metabolic processes? It is often observed that nitrogen-fixation in unicellular bacteria occurs at night. Since the corresponding nitrogenase complex is inhibited by oxygen, photosynthetic cyanobacteria face the challenge to separate the oxygen produced during the light phase form the nitrogenase. Some cyanobacterial species feature a second cell type, heterocysts, which are specialised in nitrogen fixation to ensure spatial separation, *e.g. Anabaena* sp. PCC 7120 [48]. Mitsui

et al. [138] proposed that oxygen-producing day-active photosynthesis and the night-active nitrogen fixation process are exclusively temporally separated in *Synechococcus* sp. BG 43511 to prevent inhibition. This was also postulated for different strains such as *Oscillatoria* sp. 23 [206] and *Cyanothece* ATCC 51142 [196, 207]. Interestingly, *Trichodesmium* features a complex combination of temporal and spatial separation [17]. Temporal separation predicts that under constant light (LL) cyanobacterial growth would slow down compared to light/dark (LD) regime due to permanent inhibition of nitrogen fixation and eventual nitrogen starvation. However, growth under LL is not slower, but even faster compared to LD cycles, possibly due to the additional photosynthetic energy [154, 253]. This observation is inconsistent with the general prediction of temporal separation. However, it is possible that the predicted reduced growth occurs only under specific experimental conditions (*e.g.* medium, light, carbon source) which have not yet been described. This illustrates the complications of unraveling the exact regulatory functions of the circadian clock. However, the experiment mentioned above could prove the presence of an adaptive advantage of the clock. While grown separately under LL condition, clock-deficient mutants, wild type, or mutants with different period length all grow with similar rate [154, 253]. Only when wild type and mutant are mixed together and grown in direct competition, the strain with a working clock with a free running period closest to the external light period outgrows its competitor.

The Clock does not always tick Another interesting phenomenon, the clock conditionality, emphasises the problem that clock features can be observable only under specific conditions. While the clock period of *Synechococcus elongatus* PCC 7942 is compensated for changes in the environmental temperature, the transcriptional oscillation amplitude of a *kaiBCp::luxAB* reporter gene is dampened and disappears at lower temperatures in the wild type [257]. This temperature dependency is caused by the non-optimal codon usage of the WT multicistronic *kaiBC* gene and disappears when its codon usage is optimised. Crucially, wild type strains featuring clock conditionality exhibit higher growth rates at low temperatures as compared to a mutant strain with optimised *kaiBC* gene, demonstrating the adaptive advantageousness of conditionality.

Varying extent of circadian transcriptional regulation in cyanobacteria High throughput methods have often been employed to elucidate the connection between the core clock and its target processes. The first method was the microarray, for which a sizeable repertoire of circadian time series datasets has accumulated, for eukaryotes as well as for cyanobacteria. While the number of circadian regulated genes in eukaryotes is around $\tilde{1}0\%$ [38, 66, 208, 231] it varies drastically between 1.25 and 79% across the closely-related cyanobacterial phylum [108, 264]. Independent experiments using the same strain similarly yield large differences, such as for *Synechocystis* sp. PCC 6803 with 9% [107] and 37% [110], or *Synechococcus elongatus* PCC 7942 with 29% [83] and 64% [237]. This extensive variation between circadian repertoire might be partly explained by the variation of employed conditions but also by the detection methods. For the detection of periodic expression profiles, the choice of the background model, *i.e.* how

non-periodic profiles are expected to look like, and the microarray preprocessing have a large impact on the periodicity classification [54].

1.5 What is this thesis about?

The questions addressed in this thesis can be summarised as follows:

- Do some cellular processes consistently feature diurnal expression across different cyanobacterial species? **(Chapter 2)**

  Firstly, the diurnal expression program of the cyanobacterium *Synechocystis* sp. PCC 6803 is derived from a microarray time series dataset, following a careful assessment of the applicability of common microarray preprocessing steps. Secondly, an inventory of cyanobacterial diurnal expression is presented. Using a computationally efficient method for oscillation detection, expression oscillation in the *Synechocystis* sp. PCC 6803 dataset is systematically compared with all available microarray time series datasets for other cyanobacterial strains. This analysis addresses the large variation in the extent of circadian transcriptional expression in six cyanobacterial strains and sheds light on the similarities and differences of diurnally expressed cellular processes.

- Do periodically occurring dinucleotides signify a mechanism for diurnal transcriptional regulation facilitated by chromosomal structure changes? **(Chapter 3)** Motivated by the hypothetical metabolic oscillator involving structural changes of the cyanobacterial chromosome (Figure 1.1 B) and the connection between DNA structure and sequence periodicity, the strength and localisation of periodic dinucleotides across the cyanobacterial phylum are systematically investigated.

  Firstly, the 10-12 bp periodicity strength of all possible dinucleotides and two motifs across all available cyanobacterial chromosome sequences is established. A phylogenetic study is performed, hinting a specific function of sequence periodicity. The second part focusses on AT-tract periodicity, investigating whether a concentration in specific chromosomal locations or genes persist. The last analysis studies the correlation of periodicity to circadian and supercoiling-sensitive expression.

- Does *Neurospora crassa* feature ultradian transcribed genes which are regulated by the clock transcription factors WCC and CSP1? **(Chapter 4)**

  Circadian expression in *Neurospora crassa* is analysed in detail using a combined circadian dataset of gene-wise RNA polymerase II occupancy, together with the resulting mRNA abundance. The prediction of peak transcription phases for circadian TFs by ChIP-seq data is tested. The final analysis investigates the validity of a mechanism proposed for the generation of ultradian rhythms by combined circadian transcriptional regulation.

2 A compendium of diurnal gene expression in the cyanobacterial phylum

This chapter contains the collaborative work with Rainer Machné, Manuela Benary, Jens Georg, Ilka M. Axmann, and Ralf Steuer. Since cyanobacteria rely on photosynthesis for energy production, some strains anticipate the extensive metabolic changes due to the natural day-night cycle. This motivated us to measure the transcriptional control of metabolic processes in the cyanobacterium Synechocystis sp. PCC 6803. A first circadian microarray expression time series dataset (12 samples, LD condition) was produced in the group of Ilka M. Axmann, followed by a second one which is more densely sampled (24 samples, LD, LL, and DD condition). The analyses of these datasets were conceived in close collaboration with Rainer Machné. The subsequent experimental validation of the diurnal trend in the total cellular rRNA content was presented in Beck et al. *[11]. However, we found that the expression profiles of metabolic genes peak almost exclusively between dawn and midday [98]. Considering the limited variation in transcriptional rhythmicity of metabolic genes, the subsequent step was to analyse the circadian protein abundance dataset of Synechococcus* elongatus *PCC 7942 which was created in the group of Ilka M. Axmann and of Albert Heck. In contrast to the extensive transcriptional circadian regulation, we could identify 77 proteins with circadian oscillating abundances [63], a fraction of which possess metabolic functions. The results of this work, a systematical comparison of microarray preprocessing methods in the context of our circadian dataset, was published in BMC Bioinformatics [117].*

2.1 Background

2.1.1 The daily transcriptional program of *Synechocystis* sp. PCC 6803

High variation of diurnal transcription among cyanobacterial strains - Biology or technical bias? Photosynthetic organisms such as cyanobacteria employ complex diurnal regulatory patterns in preparation of the light period [7, 207, 252]. The extent, purpose, and mechanism of diurnal and circadian oscillations differ significantly among cyanobacterial species, reflecting the high morphological and genetic diversity [198]. Accordingly, the reported estimates of the number of oscillating transcripts differ drastically between studies of different strains, ranging from 9 to 80% of protein-coding genes in microarray time series [83, 107, 209, 215, 264]. In fact, random insertion of a

luciferase reporter system indicated that up to 100% of genes may be under circadian control [122, 252]. Although the microarray technology is a powerful and widely spread tool, its technical limitations significantly complicate the quantification and interpretation of such extensive transcriptional rearrangements.

Microarray platform-inherent limitations can lead to random or systematic variation added to any biological variation of interest [20]. If differences between the distribution of the measured fluorescence values of individual arrays within an experiment are observed, they are commonly attributed to the variations in the quality of RNA extraction in the respective samples (experimental variation) and to the quality of the individual arrays (technical variation). Normalisation methods attempt to reduce the technical variation between arrays by using assumptions about biologically plausible variations. Hence, the expected extent of change in gene expression is a crucial element in the design of normalisation methods. This can be a hen-egg-problem in less well studied experimental systems, for which little or no information is available on the expected global remodelling of the transcriptional landscape.

Different microarray preprocessing - Different results? Previous studies of the transcriptional landscape in cyanobacteria have employed a range of normalisation methods. A combination of LOESS and quantile normalisation was used by Vijayan *et al.* [235, 237]. Interestingly, this study included spike-in standards, but did not apply this additional information in the normalisation. On the other hand, Kucho *et al.* [107] and Straub *et al.* [209] employed LOWESS normalisation. A modified LOWESS normalisation was used in the work of Stöckel *et al.* [207]. As described in detail by Calza *et al.*[28], the assumption of constant expression for traditional housekeeping genes does not hold under all conditions. Considering the high percentage of diurnally varying genes in cyanobacteria, including central cellular processes, the *a priori* definition of constantly expressed housekeeping genes is not plausible for cyanobacteria. Accordingly, methods employing housekeeping gene-based normalisation are generally not used.

A violation of the assumptions made by global normalisation methods has significant impact on subsequent analyses [28]. The set of differentially expressed genes may change [34, 136], as well as the correlations between genes [57, 121]. Cross-validation of expression measurements from alternative techniques, such as RNA-seq, can be used to discover methodological problems. However, the lack of such diurnal expression datasets impedes verification in the case of cyanobacteria. There is an ongoing debate about how preprocessing affects descriptors for periodic expression, *i.e.* the fraction of oscillating genes and of the peak expression phase [120, 204, 237]. Since the peak phase is generally used to derive the temporal order of cellular processes, prevention of systematical errors is crucial.

Due to the semi-quantitative nature, microarray data require a transformation in addition to the normalisation. The calibration methods permitting a more quantitative interpretation are, however, not widely used [20, 21, 49]. Transcript time series are commonly normalised to the mean value or to the distribution of fluorescence intensities: the fold-change or log_2 mean ratio transformation (in the following: `l2m`) removes the

mean, while standardisation (z-score transformation, in the following: `std`) additionally normalises the standard deviation in order to focus on the pattern of change rather than on its amplitude [261]. On the other hand, the discrete Fourier transformation (DFT) is a less commonly used data transformation due to its applicability specific for periodic signals. The removal of the first DFT component results in a normalisation by the expression mean, whereas an amplitude scaling serves to de-emphasize the amplitude [127]. Only this transformation considers explicitly the time series character of the data.

Clustering is a widespread and versatile tool The biological interpretation of microarray data is possible only after the transformation and normalisation. Due to its high-dimensional nature, a standard step in the interpretation of microarray data is clustering. A variety of clustering algorithms have been proposed, requiring systematic evaluation of the performance on gene expression data [53, 94]. However, due to the diversity of the data domain, a recent work concluded that the choice of a clustering algorithm might depend on the specific experiment [53]. In the case of time series analysis, most clustering algorithms do not consider the pattern of change over time, but treat each sample independently of the temporal order. An increasing number of algorithms propose solutions to this issue [10, 96, 99, 175, 241], however there is yet no accepted standard. An interesting approach, specifically designed to cluster periodic time series, has recently been proposed for the analysis of respiratory oscillations in budding yeast culture [127]. Here, the DFT of the time series was clustered with a model-based algorithm that uses a t-distribution as model (flowClust [123]). Nevertheless, the impact of data transformation and normalisation of time resolved microarray data, the clustering algorithm, and the similarity measure on the corresponding clustering result have not been fully described.

The distinctiveness of cyanobacterial transcription Several studies on various model organisms have reported that accepted standard normalisation methods lead to inaccurate results under certain experimental conditions. The study of a human B cell line by Loven *et al.* [124] verified an increase of the global mRNA abundance per cell. This violation of a common assumption for normalisation challenges the conclusions of a wide range of studies. Due to global oscillatory transcriptional changes of budding yeast under continuous culture conditions, an alternative normalisation scheme was employed to avoid detrimental effects of standard methods [127]. This normalisation method is included in this study, since similar global oscillatory transcriptional changes were also observed in cyanobacteria. However, cyanobacteria are a highly specific model system featuring a small number of genes with a large fraction of diurnally oscillating genes. Therefore, a systematic analysis of normalisation methods is conducted to demonstrate how to circumvent problems while analysing diurnal and circadian expression data.

2.1.2 Integrating the expression program of *Synechocystis* sp. PCC 6803 into the cyanobacterial phylum - A systematic comparison

The study of Beck *et al.*[12] reported a whole genome comparison of 16 phototrophic cyanobacteria. The authors described the genetic diversity, specifically with respect to

the metabolic functionality. Their algorithm employed pair-wise comparisons of protein sequences, which are then used to construct clusters of likely orthologous genes (CLOG). This procedure allows for the classification of genes into three classes. Core genes possess homologs in every considered strain, while shared genes possess homologs in some strains, and unique genes which occur in only one strain. The majority of genes are strain-specific. Metabolic genes are strongly overrepresented within the core genome, indicating that key metabolic processes function similarly across strains. This provides an ideal starting point to test the hypothesis, that these metabolic core processes share a temporal regulatory pattern, since the underlying paradigm of necessary metabolic changes between day and night is universal to phototrophic organisms.

Notably, variation can be expected in temporal organisation between strains, which perform nitrogen fixation during the night and thus are highly active with an altered metabolic task. Accordingly, the microarray study of Stöckel *et al.* reported increased transcriptional activity at night in *Cyanothece* ATCC 51142 with elevated total transcript abundance per cell [207]. The expression of 30% of the 5354 monitored genes oscillated with a period of 24 h, while twice as many genes peaked at the transition between LD and DL, and that ~10 % more genes peaked in the light phase. A much smaller extent of circadian expression was found for *Anabaena* PCC 7120 (78 genes), which is alternatively referred to as *Nostoc* sp. PCC 7120 [108]. Interestingly, the *kai* genes exhibited only low amplitude rhythms or arrhythmicity. The majority of genes peaked at the LD / DL transitions similar to *Cyanothece* ATCC 51142. However, the total transcript abundance per cell was constant.

Non nitrogen-fixing strains on the other hand can be expected to have less metabolic and transcriptional activity during subjective night, *e.g.* focussing on degradation of storage metabolites and basic cell maintenance. Indeed, the total mRNA content in *Synechococcus elongatus* PCC 7942 dropped to ~20 % during the 12 h dark period (CT 12 to 0) as reported by Ito *et al.* [83]. Again, the majority of the 800 circadian expressed genes peaked at the LD and DL transitions. Amongst the ~50 % of dawn-peaking genes were those coding for ATP-synthase subunits, ABC transporters, carboxysomal proteins, and RuBisCO (Ribulose-1,5-bisphosphat-carboxylase/-oxygenase, *rbcLS*). RuBisCO is a crucial enzyme in autotrophic organisms, facilitating the conversion of inorganic carbon into organic compounds within the carboxysome (unique micro-compartments encapsulating the carbon fixation machinery). While the majority of fixed carbon proceeds through the Calvin cycle and is used in the anabolism, a fraction is diverted to the synthesis of the storage metabolite glycogen via *glgA*, *glgB*, and *glgC*. The study of Vijayan *et al.* reported more than twice as many circadian genes in this strain and relates the transcriptional control to structural changes of the chromosome induced by supercoiling.

A gene expression study in *Synechocystis* sp. PCC 6803 under constant light yielded 237 circadian genes with the majority peaking at dusk [107]. In contrast, Labiosa *et al.* described more than five times as many oscillating genes under LD condition, thereby including light-induced genes. Interestingly, they reported peak phases mainly during subjective day and night, not the transitions [110].

Another study of a non-nitrogen-fixing freshwater strain, *Microcystis aeruginosa* PCC 7806, by Straub *et al.* found that the synthesis of the central storage metabolite glycogen

during the day is reflected in the transcriptional program [209]. Accordingly, the degradation of the generated glycogen is reflected in the peak expression of oxidative pentose phosphate pathway genes during the night. The expression of TCA branched pathway genes and ammonium uptake indicated amino acid biosynthesis during the subjective night. An interesting case is the strain *Prochlorococcus marinus* MED4, the smallest oxygenic photoautotroph in physical and genomic size, 0.6 micron in diameter and 1.64 - 2.68 Mbp respectively. In comparison to other cyanobacterial strains, its regulatory capacity is reduced due to the streamlined genome size. Most prominently, members of the *Prochlorococcus* family have lost *kaiA*, the third component of the cyanobacterial KaiA/B/C clock. Nevertheless, there are indications of an alternative circadian oscillator [8, 9]. The transcriptional pattern observed in the study of Zinser *et al.* agrees with that of other strains [264]. Most diurnal genes peaked at the LD and DL transition, while photosynthesis genes peaked at dawn and over the day, ATP-synthase subunits and CO_2 metabolism genes peaked shortly before dawn, and ribosome-coding genes peaked at night. Maximal expression of carbon fixation and glycogen biosynthesis related genes at dawn indicated the storage metabolite accumulation.
A collection of available microarray-based studies of diurnal expression together with key parameters is shown in Table 2.1. The procedures for microarray normalisation and oscillating gene detection employed in the respective studies are highly diverse (see last column), as is the fraction of diurnally oscillating genes. While part of this variation must be attributed to biological differences, technical and experimental sources must be considered as well. As demonstrated here, microarray normalisation procedures affect the observed oscillations. The same holds true for the detection of oscillating expression profiles, as demonstrated by Futschik *et al.* [54]. Another important distinction must be made in the employed experimental conditions, most importantly the light regime. Some studies use constant-light conditions, *i.e.* free-running conditions under which the clock is not driven by light as external zeitgeber to detect genes with clock-driven transcriptional pattern. In contrast, other studies employ light-dark cycles, which do not allow the differentiation between clock-driven and light-induced or -repressed genes. This distinction is important for the interpretation of the results of the following chapter.

2.1.3 Outline of this chapter

Four multi-array normalisation methods and three data transformations are compared with respect to diurnal expression oscillation strength and phase. Furthermore, a variety of clustering algorithms is used to examine the global expression landscape of *Synechocystis* sp. PCC 6803. The results of seven clustering algorithms are then integrated to verify whether and how normalisation shapes the results of downstream analyses. This analysis demonstrates that normalisation methods have significant impact on the estimated number and phases of oscillating transcripts, with major consequences for subsequent analysis and biological interpretation. LOS normalisation is identified as the preferable method, and the respective diurnal expression program of *Synechocystis* sp. PCC 6803 is described. Experimental results confirm the suspected diurnal oscillation of ribosomal RNA.

For a range of 6 cyanobacterial strains, transcriptional studies were performed which probed circadian gene expression patterns. The choice of the oscillatory gene detection method varies amongst the studies and impacts the results. An unbiased comparison of the diurnal expression repertoire between these strains is thus only possible using the same pre-processing. Since raw data were not provided for all nine datasets, the already normalised datasets are subjected to a standardised oscillating gene detection, providing the expression phase, amplitude, and p-value. Diurnal expression patterns of homologous gene clusters across six cyanobacterial strains is systematically compared. The cyanobacterial core gene set as well as the set of metabolic genes is characterised with respect to diurnal expression. The set of genes with at least one oscillating homolog in each dataset is analysed in detail. The materials and methods employed in this chapter are described in detail in Appendix A 5.1.

2.2 Extensive diurnal oscillations complicate microarray normalisation

The microarray expression dataset of *Synechocystis* sp. PCC 6803 features extensive diurnal oscillations. In addition to the large amplitude oscillations in a subset of genes, a second artifactual global oscillation is found in the mean signal of each microarray. In this context, the effects of different microarray normalisation methods and other preprocessing steps are systematically compared. The analysis suggests, that LOS normalisation is best suited to selectively attenuate the oscillatory artefact while preserving the dataset integrity.

A diurnal trend in the chip signal Two independent *Synechocystis* sp. PCC 6803 cultures were synchronised over the course of three days with alternating light and dark phases of 12 h duration (LD). During the fourth day, six samples were taken from two biological replicates. The resulting 12 samples were analysed using micorarrays. The resulting expression profiles of both replicates were concatenated due to their complementing information content, yielding 12 samples over two 12:12 LD cycles of 3,347 protein-coding genes and 2251 non-coding transcripts. In comparison with previous investigations, this microarray technique has the highest transcript resolution and enabled monitoring small noncoding RNAs. The direct RNA labelling strategy guaranteed a very high detection sensitivity and avoided primer-caused sequence bias, as well as reverse transcriptase-introduced artefacts of DNA microarrays. Due to the focus on diurnal periodic expression, time points were provided in hours of circadian time (CT), defining light onset as CT 0.
Periodical transcript abundances were quantified using a Fourier transform based approach (Appendix A Sections 1.2 and 1.3) which uses the peak phase and amplitude parameters in order to describe each gene profile. The non-equidistant sampling time points yield phase values that do not correspond linearly to the time domain, but provide accurate

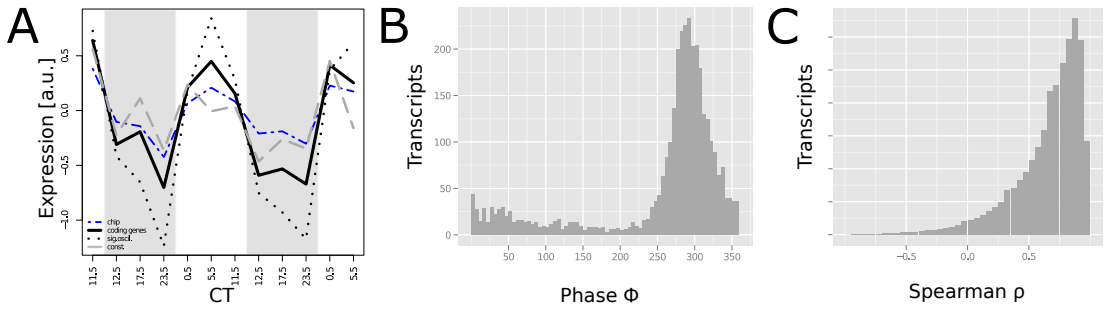


Fig. 2.1: **Diurnal oscillation of the raw total microarray time series of *Synechocystis* sp. PCC 6803.** Mean expression profiles for different transcript groups prior to pre-processing. (**A**) The profiles representing all transcripts on the chip (blue dashed), all 3347 protein-coding genes (black solid), and significantly oscillating genes (black dotted, $p_{osc} < 0.05$) exhibit diurnal oscillations with peaks over the light phase. Genes without significant diurnal profile (grey dashed, $p_{osc} > 0.05$) exhibit a light phase expression increased and a peak at 17.5 CT. (**B**) Unimodal peak expression (ϕ) distribution with the majority of genes in the range of $250 - 350$ °corresponding to expression over the day. (**C**) Correlation distribution (Spearman ρ) for all pairs of the 3447 protein coding genes, showing one subgroup of strongly correlating genes and only few uncorrelated or anti-correlated genes.

account of the temporal sequence of transcript abundance peaks. To account for random occurrence of diurnal transcript profiles, an empirical p-value (p_{osc}) was derived from a permutation-based background model.

Surprisingly, a diurnal oscillatory pattern is found in the average raw microarray signal (Fig. 2.1 A), despite the application of identical amounts of 1.5 μg RNA for each sample and chip. The average profile for significantly oscillating genes ($p_{osc} < 0.05$) resembled the oscillation of the total average profile, while not oscillating genes ($p_{osc} > 0.05$) exhibited increased abundance over the day and a peak at 17.5 CT. Accordingly, the phase distribution attests peak abundance phases 250-350° for the majority of transcripts (Fig. 2.1 B), corresponding to an expression during the light phase.

A diurnal oscillation of the total mRNA content alone, as described for *Synechococcus elongatus* PCC 7942 [83], could not have caused the oscillation found in the total chip signal of the presented dataset due to the hybridisation with a constant amount of total RNA extract. The remaining possibility was the accumulation of an RNA type, which is included in the total RNA extract used for microarray hybridisation, but is not recognised by the probes, *e.g.* ribosomal RNA (rRNA) or tRNA. Accumulation of rRNA during the night would then necessarily reduce the fraction of mRNA independent of the transcription, leading to the appearance of day-induced diurnal profiles. Prior to further analysis, a normalisation was required to distinguish any technical bias from true biological signal. However, these observations indicated that central assumptions of several common normalisation methods might be violated.

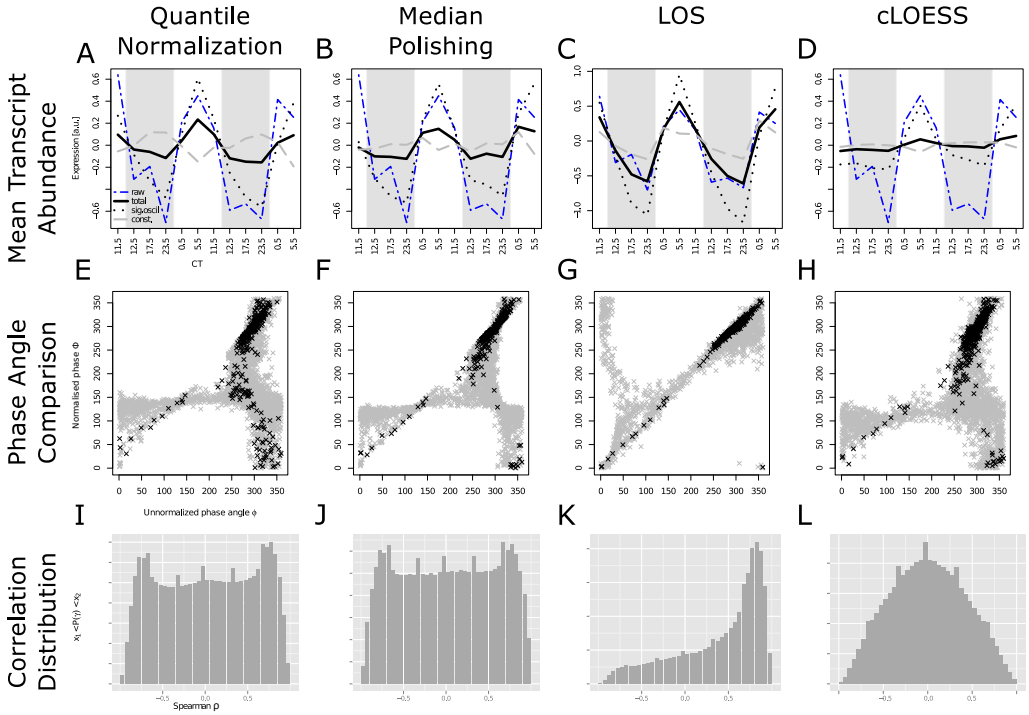


Fig. 2.2: **Varying impact of Between-Array-Normalisation methods on peak phases and pair-wise correlation distribution.** Comparison of time-series specific profile parameters after different Between-Array-Normalisation procedures. Quantile normalisation, median polishing, LOS, and cLOESS are represented in column one to four. Average profiles, peak abundance phase comparisons, and pair-wise correlation distributions are shown in rows one to three. Prominent average profiles of row one are computed similar to Fig. 2.1 A, comparing the normalised data (all transcripts: black solid, significantly oscillating in unnormalized data: black dotted, not oscillating: grey dotted) to the unnormalized mean expression profile (dashed blue). Row two compares the peak phases of the unnormalised (x-axis) versus normalised (y-axis) data, where phases of significantly oscillating profile ($p_{osc} < 0.05$) are shown in black and the remaining transcripts in grey. Pairwise Spearman ρ distributions as proxy of global expression landscape diversity are shown in row three.

Normalisation can strongly alter peak expression times Four prominent between-array-normalisation procedures were selected to quantify and compare their impact on the dataset with focus on diurnally oscillating abundance. While median polishing and quantile normalisation have been employed in previous studies of temporal expression organisation in cyanobacterial species [235, 237], cyclic loess (cLOESS) has become an established standard [260].
The fourth method is a variation of the recently proposed least variant genes method which uses a reference gene set for loess smoothing [28]. In contrast to the original procedure, the set of least diurnally oscillating genes (LOS, genes with largest p_{osc}) was used as reference for the local polynomial regression fitting [127]. Already the fraction of profiles which were classified as significantly oscillating was strongly affected by the normalisation method. A cut-off of $p_{osc} < 0.05$ yielded 25% of all transcripts from raw data, 58% from median polished, 60% from quantiles-normalised, 64% from LOS-normalised and 35% from cLOESS-normalised data. At a very conservative cut-off of $p_{osc} < 0.001$, the number of significant oscillators in cLOESS (1.7%) decreased below the level of raw data (raw: 2.2%; quantiles: 4.4%; median polishing: 4.9%; LOS: 7.8%). Moreover, the pairwise comparison of peak expression phases between the unnormalised and normalised data revealed the effect of some procedures (Fig. 2.2 E–2.2 H). These comparisons reveal a systematic deviation of strong oscillators ($p_{osc} < 0.05$) towards earlier phases, manifested as deviation from the diagonal, after application of all methods other than the LOS-normalisation. Moreover, even not significantly oscillating transcripts exhibit a similar bias. LOS-normalisation leaves significantly oscillating profile phases unchanged while the phases of non-significant oscillators are evenly distributed over the entire period. This even phase distribution indicates, that LOS normalisation successfully removes the global oscillatory trend. This leads to a biologically plausible set of genes with constant transcription and small changes due to noise, for which the resulting phase values are randomly distributed. Assuming technical noise to be independently identically distributed between the samples (microarrays), noise suppression should neither alter the observed phase of periodic profiles nor induce new oscillations. Due to the design of quantile normalisation, median polishing and cLOESS to compensate for the observed global oscillatory trend, anti-phasic oscillations were introduced, resulting in a large number of genes apparently peaking during the night (phases $< \phi$ 125°).
These systematical phase changes underly the alterations in the mean profiles of the strong and weak oscillator sets. The quantile normalisation, median polishing and cLOESS reduce the average variation between subjective night and day in the total average profile (Fig. 2.2A–2.2D, black solid line) and remove the CT 17.5 peak amongst non-oscillating transcripts. In particular, quantile normalisation moves the phase of the mean profile of not oscillating transcripts from subjective day to night. The corresponding profiles for median polishing and cLOESS do not show significant oscillations. LOS normalisation has an opposing effect, effectively reinforcing the global day-peak, but also removing the night peak at CT 17.5 amongst not oscillating transcript profiles. On the other hand, cLOESS normalisation severely dampens the periodicity of oscillating profiles, leading to the decrease in the number of significant oscillators at conservative cut-off thresholds.
Another way of characterising the data is by the pairwise correlation between profiles.

The pairwise Spearman ρ distribution in the raw dataset (Fig. 2.1 C) is unimodal with a pronounced peak around 0.8, attesting one large homogeneous profile group. The absence of uncorrelated pairs could be induced by either a global oscillatory trend present in a majority of transcripts or by array-to-array noise shared between all profiles. Quantile normalisation and median polishing lead to drastic changes of the correlation distribution, featuring bimodal distributions with comparable numbers of correlating and anti-correlating pairs and many uncorrelated gene pairs (Fig. 2.2 I, 2.2 J). This is due to the compensation of the global oscillation leading to anti-phasic oscillations in weakly or non-oscillatory profiles. In contrast, cLOESS yields a unimodal symmetric distribution with a peak at zero (Fig. 2.2 L), indicating extensive correlation reduction. This is consistent with the reduced number of significant oscillators and the dampened global diurnal trend. While only LOS normalisation preserved correlation and phase characteristics of the unnormalised data. The introduction of a small fraction of anti- and non-correlated pairs indicates that LOS normalisation yielded a set of genes with constant transcription as well as subtle night expressed genes. Both groups are expected and biologically reasonable.

The data preprocessing can strongly affect the observed periodicity in a microarray dataset [54]. Normalisation can significantly alter the number of oscillating transcripts as well as the transcript peak phases. In presence of a global oscillatory trend, anti-phasic oscillations can be newly created or phases of oscillating transcripts can be changed. In the context of diurnal expression patterns, day-expressed transcripts may be converted to night-expressed ones and *vice versa*, depending on the choice for a normalisation method. LOS normalisation allowed for the removal of a subtle global oscillatory trend, while best preserving the biologically reasonable extensive diurnal expression oscillations.

Normalisation and transformation methods shape the resulting transcript clusters Clustering of data is often used to identify the temporal or functional organisation of regulatory processes. The following analysis investigated to which extent clustering results are determined by the preprocessing, *i.e.* normalisation and data transformation. A large number of clusterings was generated, associated to all combinations of the described normalisation methods (Appendix A Section 1.4), data transformations (`l2m`, `std`, DFT, see Appendix A Section 1.2), and clustering algorithms (Appendix A Section 1.5). The resulting clusterings were then analysed for similarity. Seven popular diverse clustering approaches were selected (see Appendix A Section 1.5). Two well-established non-hierarchical clustering methods were included (K-means [52], Partitioning Around Medoids (PAM) [91]). The Self-Organising Tree Algorithm (SOTA) [42] and Hclust [47] represented the hierarchical methods. The Self-Organising Maps (SOM) algorithm [70], an approach related to SOTA, was also considered. Mclust [261] and flowClust [123] from the class of model-based methods were included. The Bayesian information criterion curve, an estimate of the optimal number of present clusters obtained from the flowClust results, indicates an optimum between eight to ten clusters (Appendix A Fig. 1). Accordingly, the following analysis employed eight clusters. The Euclidean distance and Spearman ρ similarity measures were used, capturing the absolute difference between

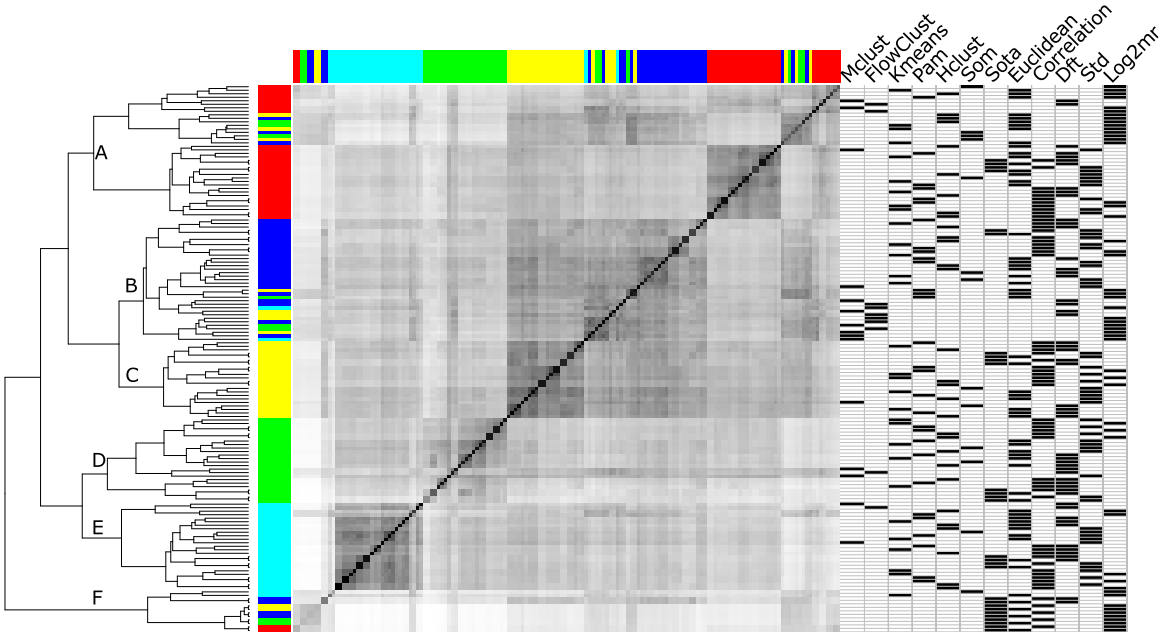


Fig. 2.3: **Normalisation determines groups of clustering results.** Similarity between all clusterings with eight clusters, measured by mutual information. Rows and columns are ordered identically according to hierarchical clustering of the similarities (Hclust, complete linkage, dendrogram on the left). White encodes minimal similarity over grey to black for maximal similarity. Normalisation method colour-code of the underlying data: raw data - blue, median polishing - yellow, LOS - green, cLOESS - cyan, quantile normalisation - red. Subsequent processing steps (transformation, similarity measure, clustering algorithm) are represented as black bars in the corresponding column on the right. The column "correlation" marks the usage of the Spearman ρ as similarity measure except for clusterings obtained from SOTA, which only allows usage of the Pearson correlation.

each value of two time series and the relative differences, respectively. The pairwise similarity between the large number of clusterings obtained from all combinations of the considered processing steps is measured via mutual information (MI, Appendix A Section 1.6). These pairwise similarities are arranged in a symmetrical matrix where each row and column corresponds to one individual clustering (Fig. 2.3). This similarity matrix was again clustered to reveal subgroups. The processing steps which yielded the respective clustering are represented as annotation matrix to the right of each row. Only the normalisation method is colour-coded on the left/top of the similarity matrix. Six large subgroups of clusterings are obtained from visual inspection (Fig. 2.3 A–F). The colour-coded normalisation method, shown on the left, yields the subgroups A to E each dominated by one method. Subgroup A containes mostly clusterings of quantile normalised data, while subgroup B contains mostly clusterings of unnormalised data, but both contain a further sub-branch. Subgroups C, D, and E exclusively contain clusterings of median polished, LOS normalised, and cLOESS normalised data, respectively. Only the

clusterings of subgroup F are organised by clustering algorithm. They are most distant to all other clusterings. Clusterings in this subgroup are derived using all normalisation methods and mostly the SOTA and SOM algorithm. The large branch length and small numbers of leaves in the dendrogram show that clusterings in this subgroup are very diverse. Manual inspection revealed that all clusterings feature at least one small cluster with < 10 genes, pointing to unstable solutions disregarded in the following. Subgroups A and B, quantile-normalised and raw data, contain a sub-branch of clusterings that are based on other normalisation methods. These sub-branches contain mostly clusterings based on `l2m` transformed data (Fig. 2.3, right panel). It can be speculated that the observed dominance of the `l2m` transformation over the normalisation method reflects the design of the `l2m` transformation to retain amplitude information, in contrast to `std` and amplitude-scaling DFT. The presented features were consistently observed for clustering with five to 14 clusters, and when using the normalised Variation of Information and the Rand index as clustering similarity measure (data not shown).
Pairwise comparisons of clusterings showed that the normalisation method determines the resulting clustering more than any other preprocessing step. Furthermore, it is important whether oscillation amplitude is retained or removed by the transformation, such as by standardisation [261] or by DFT with amplitude scaling. Interestingly, the choice of the clustering algorithm has only limited impact on the clustering result.

2.3 The diurnal expression program of *Synechocystis* sp. PCC 6803

A clustering method, specifically focussing on periodic signals, is applied to the appropriately preprocessed expression dataset. The resulting groups of genes with diurnal expression patterns are presented and analysed by functional enrichment analysis. The results confirm that the rRNA content of *Synechocystis* sp. PCC 6803 varies diurnally, justifying the suitability of the presented normalisation strategy.

In order to demonstrate the described normalisation effects, two significantly oscillating day-expressed genes were selected (Appendix A Fig. 2). LOS normalised profiles resemble the raw profiles but exhibit dampened expression spikes at CT 17.5. Quantile normalisation lead to phase shifts of $\approx 130 - 160°$ from subjective day to night, and dampening of the oscillation amplitude of photosystem I gene *ycf37* and complete absence of oscillations for transposon *ISY120b* (Appendix A Fig. 2 A and B, photosystem I gene *ycf37* and transposon *ISY120b*). Median polishing and cLOESS preserve phases well but also to severely dampen oscillation amplitudes. Since peak expression phase and amplitude preservation were of crucial for the characterisation of the diurnal expression organisation, LOS normalisation was selected for further analyses. The final clustering was performed using LOS normalised and DFT transformed data, the flowClust clustering method, and ten clusters to ensure a finer resolution of the data for the following biological

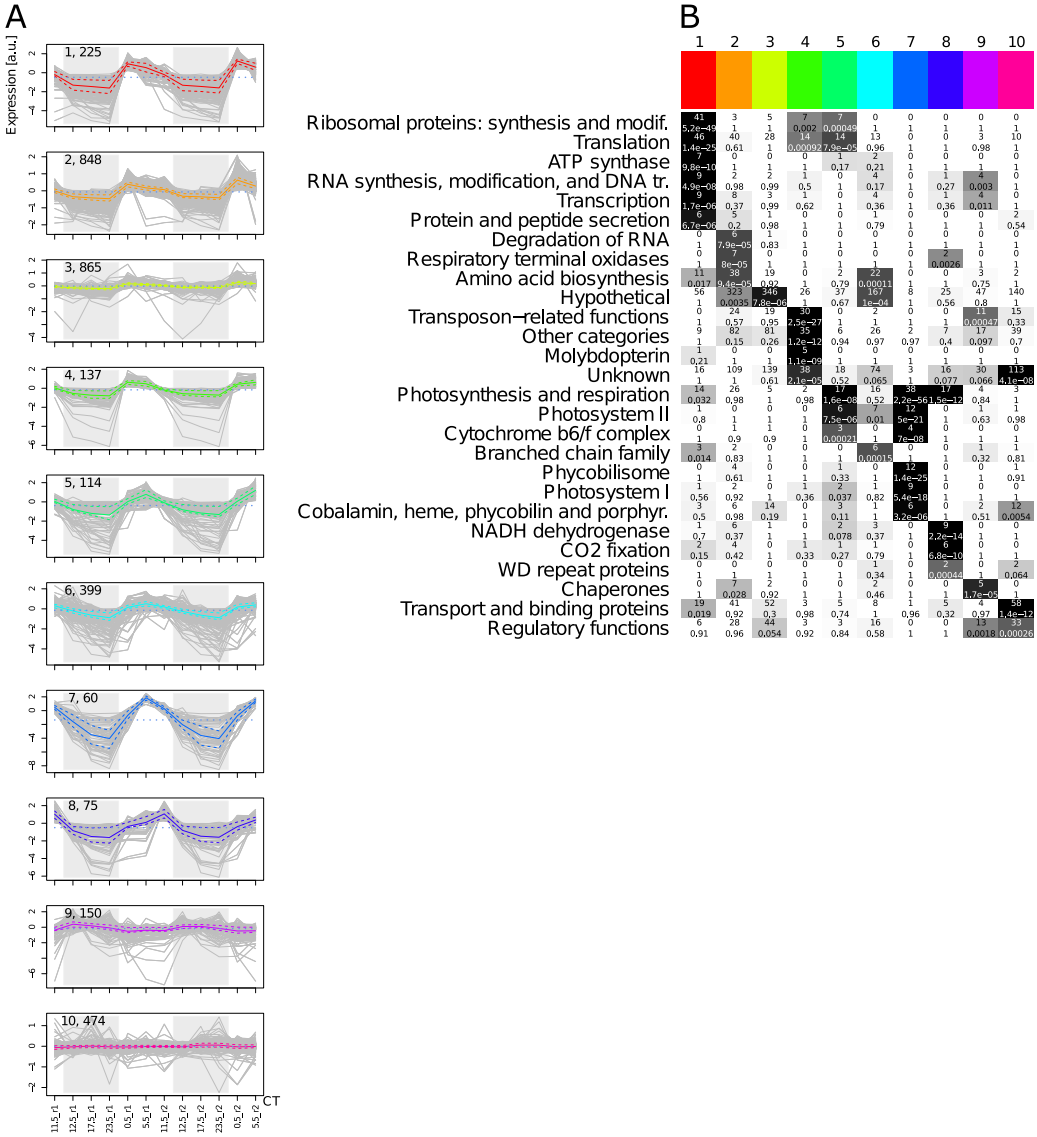


Fig. 2.4: **Clustering yields diurnal expression organisation in *Synechocystis* sp. PCC 6803.** (**A**) Clustering of expression profiles for all protein-coding genes in *Synechocystis* sp. PCC 6803. The data were LOS normalised, DFT transformed and clustered using flowClust with ten clusters. Individual profiles of are represented by grey lines using `12m`transformed data, solid coloured line mark the cluster mean, whereas the 5% and 95% quantiles are marked by dashed lines of corresponding colour. Cluster index and number of comprised genes are shown in the upper left corners. The clusters are sorted by the mean peak phase ϕ. Subjective dark periods are grey shaded. (**B**) Cluster-wise functional enrichment analysis. Rows correspond to biological functions, while the columns correspond to the clusters on the left with matching indices and colours. Each cell provides the number of genes annotated with the corresponding function and below the corresponding enrichment p-value. Simultaneously, the cell background colour represents the enrichment p-value (black - highly enriched, white - not enriched). The rows/functions were rearranged according to temporal ordering.

interpretation (still in plateau of optimal BIC values, see Appendix A Fig. 1). The resulting clustering reveals large homogeneous gene groups with distinct diurnal expression profiles (Fig. 2.4 A). The cluster-wise functional enrichment analysis yields the coarse biological program that *Synechocystis* sp. PCC 6803 follows in a diel LD rhythm (Fig. 2.4 B).

As expected for a photoautotrophic organism, the three photosynthesis-related clusters 5, 7, and 8 peaked in the morning, midday, and evening, respectively. The expression of components of the transcriptional and translational machinery in cluster 1 increased sharply during the DL transition. This could mark the extensive metabolic changes associated with the transition from respiration during dark phase to photosynthesis during light phase, together with the induction of various processes utilising the readily available photosynthetic energy. Expression of amino acid biosynthesis-related genes increased shortly after, possibly to provide the building blocks for the previously induced protein synthesis machinery. CO_2 fixation related genes showed an increased expression only in the second half of the day (cluster 8). This might reflect a separation between protein synthesis and cellular maintenance during the first half of the day from storage metabolite accumulation during the second half as preparation for the night. A comparable organisation of transcriptional activity over the day was found in *Cyanothece* sp. ATCC 51142 [30]. On the other hand, the enrichment of regulatory function-related genes in the non-oscillating cluster 10 indicates that the transcriptional activity of many stress-protection mechanisms, *e.g.* against UV radiation, dehydration, or osmotic stress, remained unchanged during the diel cycle.

Diurnal oscillation of long RNAs Despite the consistent application of $1.5\mu g$ RNA extract on each individual microarray chip, there is a diurnal trend in the mean chip signal (Fig. 2.1). One of the suspected causes for this trend is a variation of the ratio between probed (mRNA) to non-probed (rRNA, tRNA) transcripts in the total RNA extract. Indeed, the subsequent measurement of the total cellular RNA abundance revealed significant diurnal rhythmicity [11]. The total RNA abundance per cell increases from $1.76x10^{-14}\frac{gRNA}{cell}$ during light period to $3.76x10^{-14}\frac{gRNA}{cell}$ during dark (Fig. 2.5). Interestingly, this RNA accumulation seemed to be size dependent, with selective accumulation of longer transcripts (>500 bp), such as the 16S and 23S RNA, and constant short RNAs like the 5S RNA. While this accumulation was also found among long mRNAs, their lower relative abundance (18-40% of total RNA) compared to rRNA limits the impact of mRNA accumulation on the total RNA content, similar to *E. coli* [39, 170]. While 16S and 23S rRNA accumulate during night, the ribosomal mRNAs show induction during the light phase, similar to the ribosomal protein abundance. This selective anti-phasic accumulation might therefore hint at a functional relevance.

The accumulation of ribosomal RNA contradicts previous knowledge about RNAs in bacteria.

For *Synechococcus elongatus* PCC 7942, rRNA content remains constant while only the total mRNA content drops to ~20 % during the course of a 12 h dark period (CT 12 to 0) [83]. In another strain, *Synechococcus* sp. PCC 6301, the accumulation of total RNA and

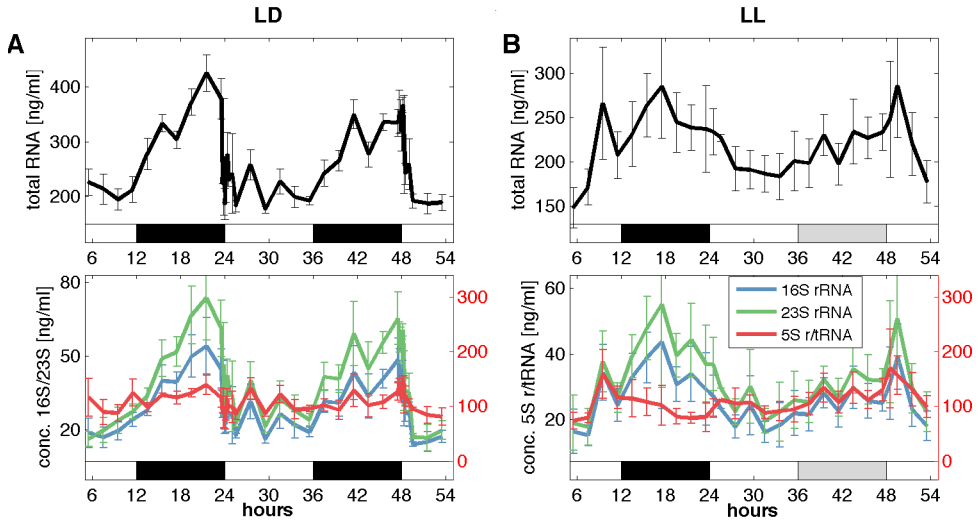


Fig. 2.5: **Diurnal variation in the total RNA amount and composition in *Synechocystis* sp. PCC 6803.** (**A**) Total RNA and rRNA abundance in LD. The RNA profile is the average of four biological replicates of Agilent 2100 Bioanalyzer and a NanoDrop spectrometer measurement. The rRNA profile comprises four biological replicates measured via Bioanalyzer. (**B**) Concentrations of total RNA (four replicates) and rRNAs (six replicates) in LL. All measurements indicate RNA abundance in 1ml cell culture with OD750 of 1. The standard errors of the sample-wise average is indicated as error bars. 16S and 23S RNA concentrations relate to the black axis on the left of both lower panels, that of 5S RNA to the red axis on the right. Adapted from Beck *et al.* [11].

16S rRNA during the subjective night phase was observed [118]. However, the authors explained this finding with the confinement of cell division to subjective day together with constant *de-novo* RNA biosynthesis. This interpretation is not applicable for the presented dataset since the conditions were chosen to prevent cell division during the course of the experiment. Instead, rRNA content is known to vary in correlation with the growth rate [93, 97, 160, 185], and thus simultaneously with the number of ribosomes [59].

2.4 Diurnal expression in *Synechocystis* sp. PCC 6803 compared to other cyanobacteria

The previous analysis yielded a coarse picture of the diurnal adjustments to gene expression in the cyanobacterium *Synechocystis* sp. PCC 6803. Currently, there are several microarray-based diurnal expression studies available for a range of cyanobacterial strains. The following analysis compares the presented diurnal program with all available experiments, focussing on diurnal oscillatory gene expression and addressing the following questions. I) Was temporal expression orchestration reproduced in different experiments

using the same strain, despite the differences in, *e.g.* growth condition? II) To what extent can the variation in the fraction of diurnally oscillating genes be explained by different detection methods? III) Is temporal expression orchestration similar between cyanobacterial strains? IV) Is diurnal regulation more frequent in the cyanobacterial core genome or amongst metabolic genes?

Diurnal transcriptional regulation in cyanobacteria For a set of six cyanobacterial strains, genome-wide datasets probing the diurnal expression program were available (Table 2.1). Unfortunately, the data for the two reported *Synechocystis* sp. PCC 6803 experiments by Labiosa *et al.* [110] and Kucho *et al.* [107] could not be obtained.Due to the employed ultradian light cycles with 6 h of light followed by 6 h of darkness, one study by Toepel *et al.* had to be discarded [216]. Throughout the following section, the cyanobacterial species contained in Table 2.1 will be referred to by the corresponding genus while other species will be fully named.
The *Synechocystis* datasets, the *Synechococcus* dataset of Ito *et al.*, and the *Anabaena* dataset were `l2m` transformed, while the *Cyanothece* datasets were already in transformed form. On the other hand, the dataset provided by Vijayan *et al.* appears to be `l2m` transformed due to the occurrence of negative expression values, but is not annotated as such. The two biological replicates of the *Anabaena* dataset were concatenated for the following analyses [108], similar to the *Synechocystis* dataset [117]. Expression profiles were smoothed using a Savitzky-Golay lowpass filter, as proposed by Yang *et al.* [259], in order to remove pseudo peaks prior to the detection of periodic genes. Diurnally oscillating expression profiles were detected using harmonic regression analysis, as described in Appendix A Section 2.1. This methodical changed offered several advantages. As described earlier, the Fourier-based procedure does not provide exact phase values for non-equidistant sample intervals, which otherwise complicates the comparison between datasets. Furthermore, harmonic regression is computationally more efficient, allowing for fast processing of the dataset collection. The p-value is based on the assumption of a linear background profile as compared to the sinoidal foreground model. After multiple testing correction according to Benjamini-Hochberg, all datasets yielded significantly oscillating genes ($q <= 0.05$) except for *Microcystis*, with 7 samples the shortest dataset. Accordingly, *Microcystis* data are shown for comparison reasons but were not utilised in the comparative analyses.

Anti-phasic gene clusters as common feature The comparison of peak expression phase ϕ distributions across all considered datasets reveals two gene groups, mostly of comparable size and with anti-phasic expression, *i.e.* $\approx$ 12 h delay (see Fig. 2.6). Nevertheless, some deviations from this pattern could be observed. Dominant midday expression with a narrow phase distribution was found in the *Synechocystis* experiments, and mostly late-night expression in *Anabaena*. Anti-phasic groups were not found in *Prochlorococcus* which featured a mid-night group peaking around CT 19 and an evening group around CT 11. Most experiments yielded few genes with intermediate phases,

Strain	Strain Abbrev.	N_2 Fixation	Habitat	Total Genes	Ref.	Dataset Abbrev.	Diurnal Genes	Light Conditions	Culture Conditions	Normalization / Oscillation Detection
Prochlorococcus marinus MED4	ProMED4	-	M	1766	**[264]**	zinser09	1403 (79%)	LD (14:10, T_{samp} 2h)	Pro99 Medium, stirred, 24 °C, Batch	RMA, Fourier Analysis
Synechocystis sp. PCC 6803	Syn6803	-	F	3628	**[117]**	leh13	1133 (31%)	LD (12:12, T_{samp} irreg.)	BG11 Medium, air bubbling, 30°C, Batch	LOS, Fourier Analysis
					[11]	beck14	27%	LD (12:12, T_{samp} 2h)	BG11 Medium, air bubbling, 30°C, Batch	LOS, Fourier Analysis
					[110]	-	1349 (37%)	LD (14:10, T_{samp} 2h)	BG11 Medium, stirred, 3% CO_2 air bubbling, 27°C, Turbidostat	Stanford Microarray Database standard, ANOVA and correlation with light
					[107]	-	237 (9%)	LL (T_{samp} 4h)	BG11 Medium, air bubbling, 30°C, stirred, Batch with manual dilution	LOWESS, modified Cosiner
Synechococcus elongatus PCC 7942	Syc7942	-	F	2719	**[237]**	vijayan09	1748 (64%)	LL (T_{samp} 4h)	BG11 Medium, 1% air CO_2 bubbling, 30°C, Continuous Culture	Loess and Quantile, Fourier Analysis
					[83]	ito09	800 (29%)	LL (T_{samp} 4h)	BG11 Medium, 30°C, Continuous Culture	Replicate Mean Polishing, Correlation to Sine
Microcystis aeruginosa PCC 7806	Mic7806	-	F	6360	**[209]**	straub11	1344 (21%)	LD (12:12, T_{samp} irreg.)	BG11 Medium, 1% air CO_2 bubbling, 22°C, Batch	LOWESS, significant difference to CT0
Anabaena sp. PCC 7120	Ana7120	•	F	6222	**[108]**	kushige13	78 (1.25%)	LL (T_{samp} 4h)	BG11 + N Medium, 30°C, Continuous Culture	Replicate Mean Polishing, Correlation to Sine
Cyanothece ATCC 51142	Cyn51142	•	M	5354	**[207]**	stoeckel08	1445 ($\approx$ 30%)	LD (12:12, T_{samp} 4h)	ASP2 Medium, 30°C, air bubbling, Batch	LOWESS, Correlation Network
					[215]	toepel08	1424 ($\approx$ 20%)	LD (12:12)/ 24h LL (T_{samp} 4h)	ASP2 Medium, 30°C, Airlift Bioreactor	LOWESS, Differential Expression
					[216]	-	1400 (27%)	LD (6:6, T_{samp} 2h)	ASP2 Medium, 30°C, Airlift Bioreactor	LOWESS, Correlation Network

Table 2.1: **Cyanobacterial strains used to determine the core oscillatory genome.** Information is provided about the ability to fix nitrogen, the habitat (freshwater F, saltwater S), the total number of genes, the publication reference, the absolute and relative number of diurnally expressed genes reported in the original publications, the applied light and sampling schema, the experimental culture conditions, and the methods for microarray normalisation and oscillating gene detection. References to datasets employed in the following comparison are shown bold.

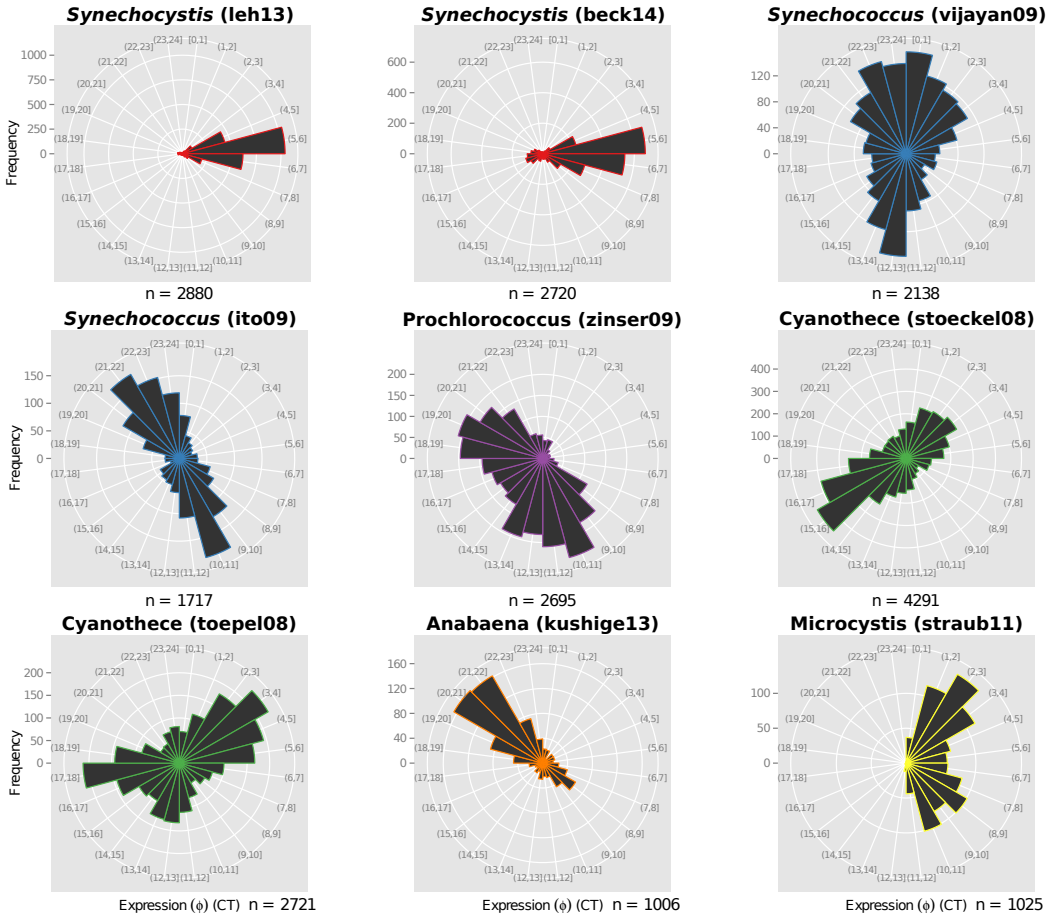


Fig. 2.6: **Expression phases ϕ of diurnally expressed genes in all available cyanobacterial circadian microarray datasets.** Circular phase histograms of significantly oscillating genes across 6 cyanobacterial strains. Peak expression phases are shown on the circular axis in circadian time [CT], the number of transcripts in each bin is indicated on the y-axis, note the varying scale between the datasets. The total number of significantly oscillating genes n is provided below. The significance threshold $q <= 0.05$ is used for false discovery rate corrected p-values according to Benjamini-Hochberg. Since no profiles of the *Microcystis* dataset remained significant after FDR correction, uncorrected values were used ($p <= 0.05$) for display purposes only. The histogram colours assigned to the cyanobacterial strains are used consistently throughout this chapter.

i.e. outside of the dominant phase groups. There were, however, two exceptions: the datasets of *vijayan09* (comparable number of intermediate day and night phases), and the *Prochlorococcus* dataset (mostly night-intermediate phased).
Interestingly, the group of night-expressed genes in the *toepel08* dataset showed a division into two subgroups, one at the LD transition and one before midnight (CT 17). This subdivision was not captured in the biological replicate experiment of *stoeckel08*. Minor differences were also observed between the biological replicates of *Synechococcus*. The width of the morning cluster in the *vijayan09* dataset was significantly larger than in the *ito09* dataset, which also exhibited a lower number of genes with intermediate phases. Furthermore, a delay of peak phases of about 2 h was observed between both datasets. Since the *Microcystis* dataset did not yield significantly circadian genes after multiple testing correction, expression phases for significant oscillators before correction are shown for comparison. The resulting phase distribution yields two groups of day-expressed genes and a complete absence of night-expressed genes. This suggests that reliable oscillation detection with harmonic regression was not possible in combination with the sampling scheme, involving 7 non-equidistant samples.

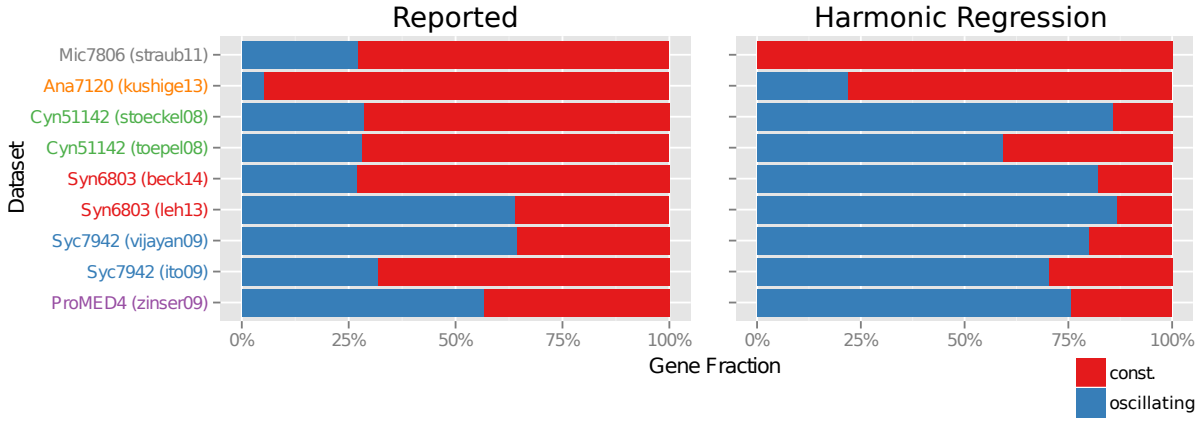


Fig. 2.7: **Reported number of diurnally expressed genes between available datasets (left panel) compared to harmonic regression-derived numbers (right panel).** For each of the considered expression datasets (y-axis, datasets marked in strain-specific colour), the number of genes classified as diurnally expressed (dark blue) and with other expression patterns (red) are shown, relative to the number of genes covered in the respective microarray platform. Underlaying data are provided in Appendix A Table 1.

The reported fractions of diurnally oscillating genes varied drastically among biological replicates and across strains. The presented data allow to determine the influence of the employed oscillation detection methods on the reported numbers of diurnally expressed genes, and to estimate how much variation must be attributed to biological and other technical sources. The reported numbers of diurnal genes (Fig. 2.7 left, underlaying data in Appendix A Table 1) fall into two groups, one with $\approx 60\%$ diurnal genes, one with $\approx 30\%$. *Anabaena* constitutes an outlier with only $\approx 6\%$, which might reflect its biological distinctness. Importantly, the redundant experiments for *Synechococcus* and *Synechocystis* disagreed, one belonging to the 30% group, and one to 60% group,

respectively. Only the redundant *Cyanothece* experiments yield remarkably similar results. Application of harmonic regression resolved these discrepancies. The redundant experiments for *Synechococcus* and *Synechocystis* achieve similarly high numbers of diurnal genes (Fig. 2.7 right). In contrast, the variation between the two *Cyanothece* experiments is increased to $\approx 30\%$. This is biologically reasonable, since *toepel08* switched from LD rhythms during the first day to constant light during the second day, thereby suppressing oscillations of purely light-driven genes in the second period. In the *stoeckel08* dataset, light-driven genes were detected in addition to clock driven genes due to consistent LD illumination. Again, the *Anabaena* dataset yielded a relatively small number of diurnal genes, confirming its biological distinctiveness. Since no significant oscillators could be detected in the *Microcystis* dataset, it was excluded from the analysis.

In order to estimate how much variation in the fraction of oscillating genes α is due to varying oscillation detection methods, the intrinsic variance was compared between the reported values and the harmonic regression derived values. The two datasets of *Microcystis* and *Anabaena* had to be excluded due to technical and biological reasons, respectively. While the *Microcystis* dataset did not yield significantly oscillating genes, the *Anabaena* dataset yielded significantly fewer diurnal genes consistently for both oscillation detection methods. This can be attributed to its biological distinctiveness as a heterocyst forming strain, which was examined under nitrogen-fixing condition and is also reflected in its unique expression phase distribution with a majority of dusk genes (Fig. 2.6). A 2.5-fold reduction was found between the inter-experiment variance for the reported values ($\sigma^2 = 314$) compared to the harmonic regression-obtained values ($\sigma^2 = 127$), which shows that only about 40% of the variation is due to biological differences between the considered strains and technical differences other than the oscillator detection method. In summary, the application of a standardised oscillation detection method reduced the variation significantly, thereby improving comparability and confirming expected biological differences between the datasets.

Reproducible peak expression time between biological replicates The comparison of peak expression phases ϕ between the redundant datasets revealed good agreement for genes with diurnal profiles in both datasets. The *Synechococcus* datasets yield the highest circular correlation ($\rho_{ccc} = 0.61$) between the 1490 common diurnal genes (Fig. 2.8 A) and also high correlation of the observed amplitudes ($\rho = 0.78$) with a tendency of higher values in *ito09*. The systematic delay of $\approx 2h$ is similar in the dusk and dawn clusters. This delay, together with the large dawn cluster, leads to the relatively small non-circular correlation of 0.29. In contrast, the comparison of the 2566 diurnal genes common to the *Synechocystis* datasets reveals a group of drastically delayed genes (up to 12 h) in the *beck14* data, leading to a circular correlation of $\rho_{ccc} = 0.31$, but with the highest amplitude correlation of 0.93. Interestingly, the circular correlation between the two *Cyanothece* datasets was negative with $\rho_{ccc} = -0.51$. The non-circular phase correlation of 0.43 was similar to that between the *Synechocystis* datasets. The observed amplitude correlation of 0.77 is similar to that of the *Synechococcus* datasets, with the *stoeckel08* dataset tending to higher amplitudes.

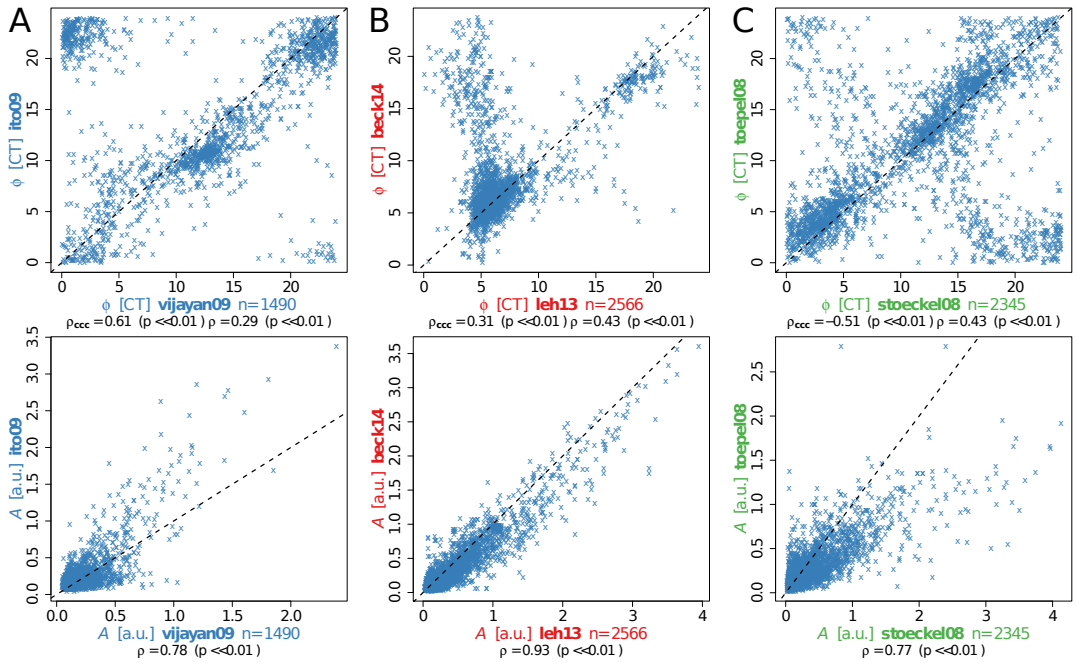


Fig. 2.8: **Comparison of expression phase ϕ (top, [CT]) and amplitude A (bottom, [a.u.]) of diurnal genes shared between independent datasets of the same cyanobacterial strain.** (**A**) *Synechococcus* datasets of Vijayan *et al.* [237] (x-axis) and Ito *et al.* [83] (y-axis), (**B**) *Synechocystis* in the datasets of Lehmann *et al.* [117] (x-axis) and Beck *et al.* [11] (y-axis), and (**C**) *Cyanothece* in Stöckel *et al.* [207] (x-axis) and Toepel *et al.* [215] (y-axis). The number of genes n found to oscillate significantly and the corresponding Pearson correlation coefficient ρ between ϕ and A is provided below each panel, followed by the respective p-value for ρ differing from 0. For ϕ, the circular correlation coefficient ρ_{ccc} is also provided. Axis labels are shown in the strain-specific colour.

Substantial variation in diurnal expression phase between cyanobacterial strains The study of 16 cyanobacterial genomes presented by Beck *et al.* [12] provided a grouping of genes into families across the strains with a shared ancestor. A clustering of reciprocal protein blast scores was used to measure protein similarity. The resulting clusters of orthologous genes (CLOG) occur in at least one cyanobacterial strain, and contain one or more homologous genes. The following analysis employed this family grouping to detect diurnally expressed orthologous gene clusters across the diurnal microarray datasets. The hypothesis to be tested was whether the diurnal expression phase is measurably preserved amongst genes with common origin, as could be expected for, *e.g.*photosynthesis related genes. The strain *Microcystis* was excluded due to lacking significantly diurnal genes, as described above. Since each CLOG can contain multiple genes in one strain, all possible combinations of diurnal gene pairs between different strains were considered. This could mitigate a phase correlation signal, *e.g.* in cases where homologous isoenzymes follow drastically different diurnal activity patterns. The impact

is likely limited since many examples of such differentially regulated isoenzymes were not classified as homologous in the employed dataset, such as the prominent glyceraldehyde-3-phosphate dehydrogenase isoenzymes Gap1 (*slr0884*) and Gap2 (*sll1342*). While the Gap1 selectively catalyses a conversion step in the glycolysis as part of storage metabolite breakdown (dusk-peaking diurnal expression in all expression dataset, data not shown), the Gap2 catalyses the reaction in the opposite direction as part of the day-active Calvin cycle (dawn-peaking diurnal expression). Similarly, the differentially regulated glycolytic isoenzymes studied by Tabei *et al.* were not classified as homologous [212].

Visual inspection of pairwise phase scatterplots of homologous diurnally expressed genes and the circular phase correlations ρ_{ccc} reveals substantial variation between cyanobacterial strains (Appendix A Fig. 4). The highest phase correlations with a $\rho_{ccc} = 0.23$ were observed between the *stoeckel08* (*Cyanothece*) and *vijayan09* (*Synechococcus*) datasets, as well as between *zinser09* (*Prochlorococcus*) and *kushige13* (*Anabaena*). The highest phase anti-correlation was found between the *zinser09* and *toepel08* (*Cyanothece*) datasets with $\rho_{ccc} = -0.2$. A clear diagonal patterns indicated phase correlation for the biological replicate datasets. However, such patterns were not found in any across-strain comparison. Instead, all comparisons of *Synechocystis* datasets reflected the strong bias towards expression during the subjective mid-day (CT ≈ 6), independent of the corresponding phase in other strains (strong vertical cluster in the *leh13* vs. *vijayan09* comparison). Another prototypical pattern emerges due to the strongly bimodal phase distributions. Dusk genes in the *ito09* dataset (CT 10) were found to be separated into two groups, namely, dusk (CT 10) and dawn (CT 20) in the *zinser09* dataset. A similar separation of dawn genes resulted in four gene clusters, indicating maximal disagreement between dusk and dawn-assignments between the two strains with accordingly low circular phase correlation ($\rho_{ccc} = 0.027$) for the 709 considered genes. General disagreement was found between the *Synechocystis* datasets and the *kushige13 Anabaena* datasets, resulting in small circular phase correlations of $\rho_{ccc} = -0.076$ and $\rho_{ccc} = -0.085$, respectively.

Diurnal expression is not more prevalent in core genes and metabolic genes

Using their orthology prediction, Beck and coworkers defined a set of 660 core CLOGs common to all 16 considered cyanobacterial strains [12], as opposed to 6668 CLOGs shared by multiple strains and 13910 strain-specific CLOGs. The authors noted that 55% of the core CLOGs included genes with enzymatic functions, demonstrating a central position of the metabolism in the core genome. However, no overrepresentation of genes coding for a specific metabolic process was detected, but the prominent Calvin Benson cycle, the oxidative pentose phosphate pathway, nucleotide synthesis, and amino acids synthesis were found to be highly conserved.

The following analysis investigated whether core CLOG genes exhibit a higher propensity to diurnal expression as compared to strain-specific genes. The results could hint towards a shared cyanobacterial diurnal expression program, reflecting an adaptation to alternating photic and aphotic phases common to all cyanobacteria. Due to the prominent role of enzymatic genes in the core genome, a similar comparison of diurnal expression was performed between genes with and without enzymatic function. The

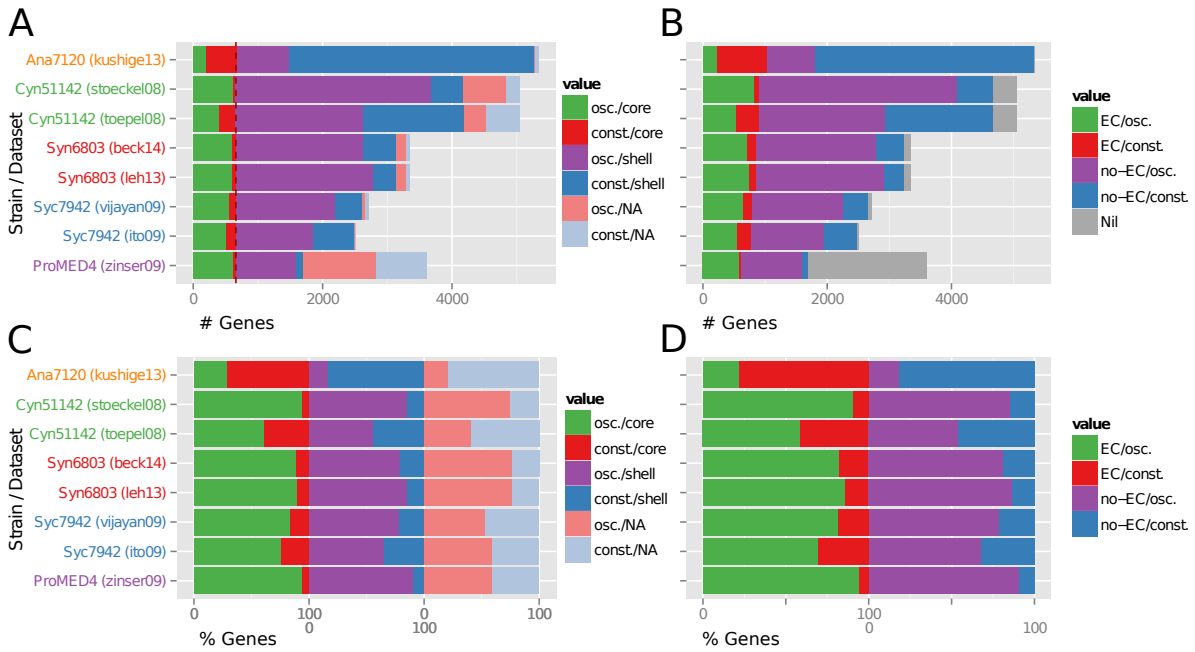


Fig. 2.9: **Overlap of diurnally oscillating genes with cyanobacterial core genes and metabolic genes.** (**A**) Distribution of diurnally expressed genes across the core and shell genome. The total number of measured transcripts for each dataset is divided into core (green/red) and shell (purple/blue) genes according to Beck *et al.* [12], as well as into diurnally oscillating and constant. The size of the core genome across 16 species is marked as red dashed vertical line. The cyanobacteria strains and expression dataset names are provided on the left. The numbers of genes not included in the CLOG dataset are shown in coral and light blue. (**B**) Similar analysis of diurnally oscillating genes versus metabolic genes (defined as possessing EC number annotation). The number of transcripts, which were measured but which could not be mapped to any CLOG is indicated in grey. (**C**) Relative contributions of diurnally expressed genes to core and shell genome, compare to absolute contributions shown in A. Each class (core, shell, not included) is considered as 100% as indicated on the x-axis, and relative fractions of diurnal and constant genes are shown. (**D**) Relative contributions of diurnally expressed genes to metabolic genes, compare to absolute contributions shown in B.

Enzyme Commission (EC) numbers provided in Beck *et al.* were used. Again, the strain *Microcystis* was excluded. While the number of shell genes varied widely (1296-4622), the number of individual genes associated to the 660 core CLOGs varied only slightly across the considered strains (661-686). However, the absolute and relative fractions of diurnally expressed genes in core, shell, and also unclassified genes (Fig. 2.9 A and C) are highly similar. The Pearson correlation between the relative fractions of diurnal genes between core and shell genes was nearly perfect with $\rho = 0.99$, which points to equal distribution of diurnally expressed genes over all three gene classes and thus the absence of enrichment. The two Cyanothece datasets are particularly informative as the the diurnal gene fraction in the *ito08* is reduced equally across all gene classes with respect to the *stoeckel08* dataset. This reduction is likely triggered by the changed light regimen.
Similarly, the comparison of of metabolic genes (Fig. 2.9 B and D) indicates absence of enrichment ($\rho = 0.997$).

2.5 The core diurnal genome

The set of orthologous genes with diurnally oscillating expression in all considered cyanobacterial strains and datasets is analysed. Expression phase distributions and individual profiles are shown for groups of genes with prominent function, as well as for important individual genes.

A total of 95 genes exhibit diurnally oscillating expression in all eight datasets of five cyanobacterial strains (Appendix A Table 2). Genes were excluded from this list if they do not feature homologs in all considered strains but oscillate in the remaining ones. Importantly, the whole genome datasets allow for the extrapolation of the number of shared diurnal genes beyond the strains in this analysis. The size of the core diurnal genome does not change drastically with 5 or more datasets (Fig. 2.10), suggesting that a core diurnal genome is shared amongst all cyanobacteria.
In this set, genes coding for proteins in central metabolic processes are found almost exclusively. The small subunit of the enzyme RuBisCO, crucial for the carbon fixation part of photosynthesis, exhibits high amplitude diurnal oscillations in transcript abundance with a peak phase at dawn (Appendix A Fig. 5). Only in *Synechocystis* it's phase is delayed to midday. The amplitude in *Cyanothece* is significantly larger in the experiment of Stöckel (~3.5 fold) compared to that of Toepel (~2 fold). Seven out of 27 genes annotated with "photosystem II" fall into the core diurnal set, together with 18 out of 66 ribosomal protein genes. Another large group is comprised of 15 genes marked "hypothetical" in *Synechocystis*. Figure 2.11 shows the phase distributions for each of these gene groups, while Figure 2.12 presents the underlying expression profiles. Particularly the ribosomal genes exhibit high synchronicity in phases as well as the underlying expression profiles (Fig. 2.12 A). The peak expression phase of the ribosomal gene group is consistent between same strain experiments, but varies between the strains. The strains

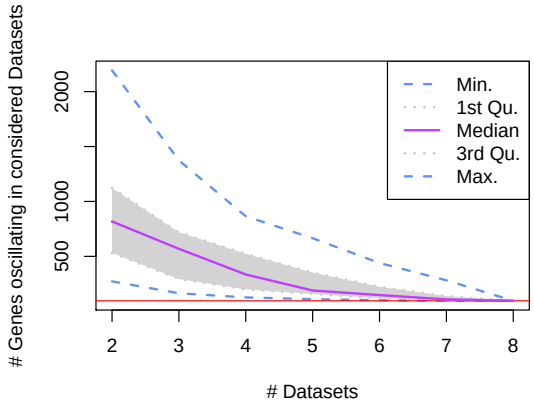


Fig. 2.10: **The estimated size of the cyanobacterial diurnal core gene set.** Estimated number of genes that invariably exhibit diurnal expression regulation with rising number of considered datasets. With increasing number of included datasets (x-axis) the number of genes in the diurnal core decreases (y-axis). When using less that eight datasets, all possible dataset combinations are tested. The obtained distribution is illustrated as median (purple), first and third quartile (grey shaded area), minimum, and maximum (blue dashed). The size of the final diurnal core gene set (95 genes) is marked with a red line.

can be ordered by ribosomal peak phase, starting with *Microcystis* after-dawn and shortly after *Synechocystis*, then *Synechococcus* shortly after dawn, followed by *Cyanothece* and *Prochlorococcus* around midnight. Notable spread of phases from dark to light phase is is observed only in the *Cyanothece* experiment of Toepel. The phase spread of the photosystem II genes is surprisingly large for *Synechococcus* and *Prochlorococcus*, whereas expression peaks during early day to midday for the remaining strains. The corresponding expression profiles show less synchronicity compared to the ribosomal genes (Fig. 2.12 B). The highest variability in phase distributions and expression profile shape is observed for hypothetical genes (Fig. 2.12 and 2.12 C). The two RNA polymerase subunits (α and γ) represent the core transcriptional machinery. Furthermore, the analysis reveals two transcription-regulation genes: the group 2 sigma factor sigB (*sll0306* in *Syn6803*) and the anti-sigma factor F antagonist (*rsbV* in ProMED4).

Already in 1996, it was reported that sigma factor RpoD2, whose homolog in Syc7942 is sigB, is involved in the circadian clock mechanism of Syc7942 [227]. It was shown that RpoD2 expression peaks at subjective dusk [144], similar to a large fraction of the diurnal genome [83]. Moreover, translation is represented by 18 out of 64 ribosome subunit-coding genes. The three photosystem I-related gene encoding NADH dehydrogenase subunits 1 and 4 (*sll0519*, *slr0331*) and rubredoxin (*slr2033*) appear in the list (Appendix A Table 2). While the NADH dehydrogenase subunits show consistent high-amplitude oscillations with dusk peaks in *Prochlorococcus* and *Cyanothece*, rubredoxin expression peaks at dawn. All three genes show midday peaks in *Synechocystis* and variable phases

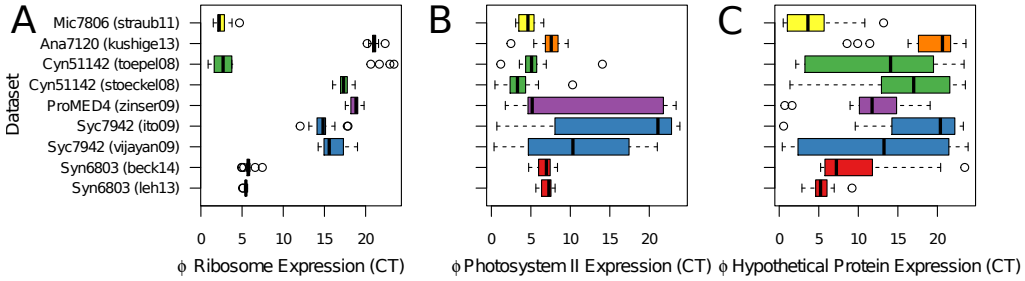


Fig. 2.11: **Expression phase distributions of prominent core diurnal gene classes across all datasets.** (**A**) Ribosomes (**B**) Photosystem (**C**) Hypothetical Protein genes.

with low amplitudes in *Anabaena* and *Synechococcus*. The ATP synthase subunit b gene (*sll1323*) peaks consistently at dawn, however, with a strongly varying amplitude. With the light-dependent NADPH-protochlorophyllide oxidoreductase (PCR, *slr0506*) and the light-independent protochlorophyllide reductase subunit ChlB (*slr0772*), two key enzymes in the chlorophyll synthesis are diurnally regulated. Expression of the light-dependent PCR peaks around dawn, except for *Anabaena* in which it peaks at dusk. Interestingly, the light-independent ChlB gene is also consistently diurnally expressed with varying phases. In *Synechocystis*, it shows strong induction during day and in the remaining strains around dusk. The only exception is the *Cyanothece* experiment of Stöckel where the phase is significantly delayed towards dawn. Fructose-bisphosphate aldolase (*sll0018*) is consistently diurnally regulated with expression peaks around dawn in all datasets except for *Synechocystis*, in which it is delayed towards midday. A diurnal regulation pattern of this enzyme is not surprising since this enzyme is used by opposing reactions of glycolysis in forward direction, and of the Gluconeogenesis and the Calvin-Benson-Bassham cycle in backward direction. Carbamoyl-phosphate synthase catalyses the first committed step in pyrimidine and arginine biosynthesis, making it an ideal gene for temporal regulation. While *Synechocystis* and *Synechococcus* feature dawn- to day- peaking expression with low amplitudes (<0.8 fold), higher-amplitude oscillations (1-3 fold) peaking around dusk are found for the remaining strains. The glyceraldehyde-3-phosphate dehydrogenase genes *gap1* and *gap2*, described above, are consistently diurnally expressed in anti-phase. Gap1 exhibits peak expression around dusk with fold changes up to 6, whereas Gap2 is dawn specific with smaller fold changes up to 3. However, they are not amongst the core diurnal genome since no *gap1* homolog is found in *Prochlorococcus* and no *gap2* in *Cyanothece*. For *Microcystis*, high amplitude dawn-specific diurnal expression changes were described for the glycogen synthesising genes *glgA1*, *glgA2*, *glgB*, and *glgC*. However, these genes are not part of the core diurnal genome due to the high variability found between strains and also between same-strain experiments. Only *Prochlorococcus* shows similarly high dawn-phased oscillations, whereas *Cyanothece* features similar phases but mostly with small amplitudes. In *Synechocystis*, all genes peak during midday but only *glgC* exhibits notable amplitudes of >1 fold change.

The most prominent non-metabolic diurnally expressed gene is the *kaiB* homolog *kaiB1*, one of the core components of the cyanobacterial circadian clock. Due to the integration

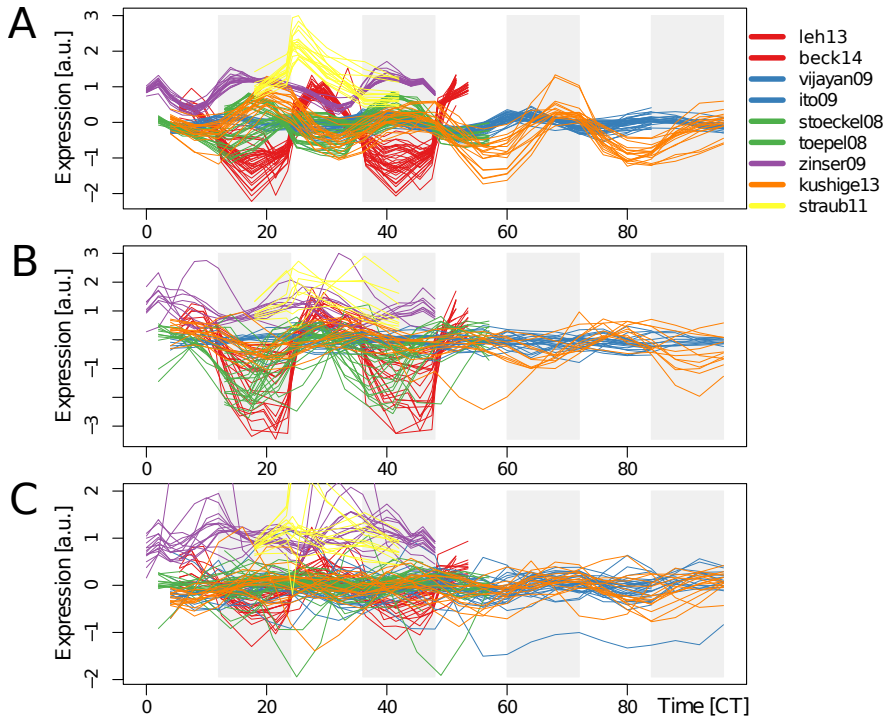


Fig. 2.12: **Expression profiles of prominent core diurnal gene classes across all datasets.** (**A**) Ribosome-coding genes. (**B**) Photosystem-II subunit coding genes. (**C**) Genes annotated as hypothetical in *Synechocystis*.

of the transcriptional regulation, *e.g.* the *kaiBC* operon in *Synechococcus*, similarities between the expression patterns of individual genes can be expected. To evaluate this, the expression patterns of the most common homologs *kaiA*, *kaiB1*, and *kaiC1* are shown in Figure 2.13. Interestingly, the *kaiA* gene features only very low amplitude expression oscillations and is arrhythmic in the vijayan09 *Synechococcus* dataset. The expression phases vary from dawn (*Microcystis*) to morning (*Synechocystis*), over midday (*stoeckel08* dataset of *Cyanothece*), and dusk (*ito09* dataset of *Synechococcus*), into night (*Anabaena*). The observed expression phases of *kaiB1* are comparable to those of *kaiA*, but with significantly larger amplitude in *Synechococcus* datasets. The phase of the *kaiB1* homolog in *Prochlorococcus* peaks before dawn, comparable to *Anabaena*. *KaiC1* expression phases and amplitudes match those of *kaiB1*, with the notable exception of *Cyanothece* for which antiphasic late-night peaks are observed. In *Prochlorococcus*, *kaiC1* peaks during the early night in contrast to the late night phase of *kaiB1*.

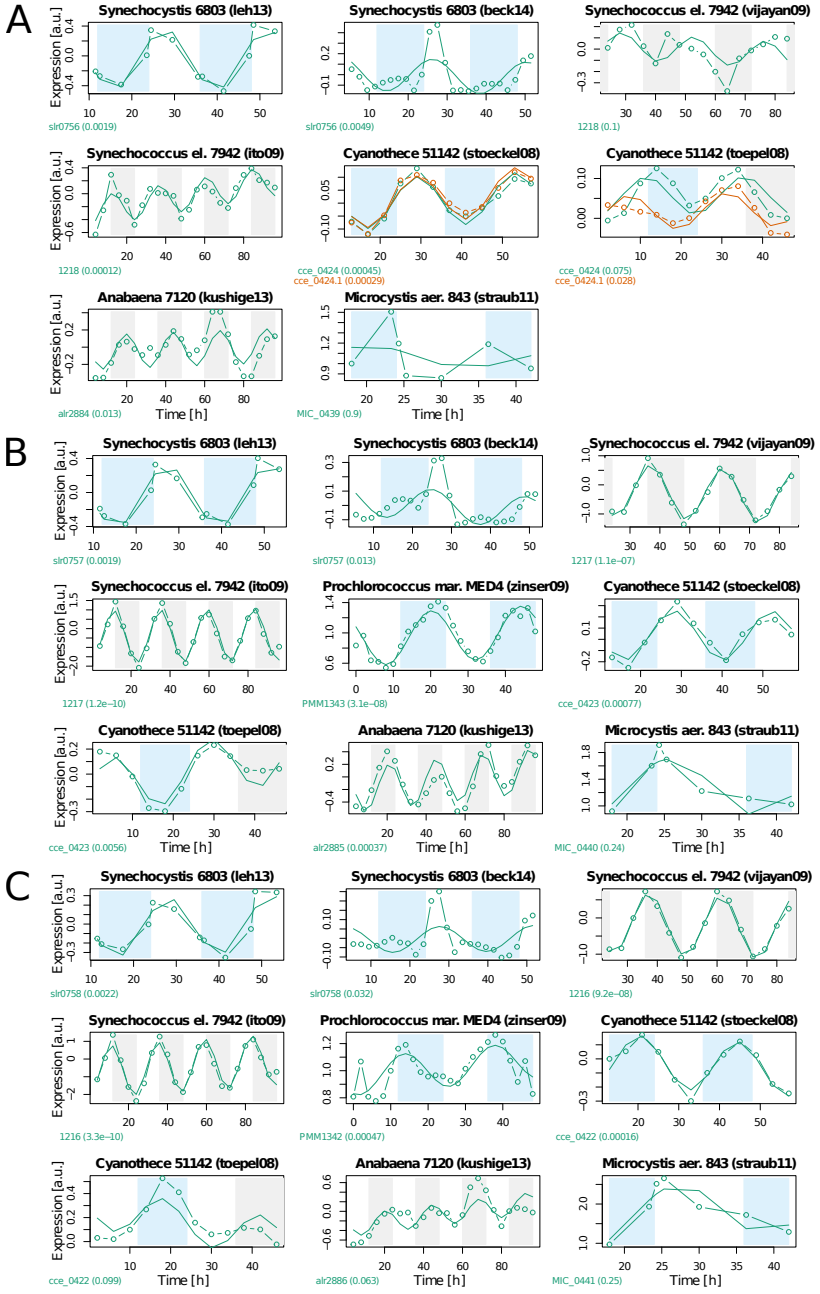


Fig. 2.13: **Expression profiles of *kaiA*, *kaiB1*, and *kaiC1* across all datasets.** Expression profiles of all homologs of the core clock genes *kaiA* (**A**), *kaiB1* (**B**), and *kaiC1* (**C**). The expression profile and the corresponding harmonic regression function are shown as circles and solid lines, respectively. Different homologs are distinguished by colour. Subjective dark periods in LL conditions are marked as shaded grey, actual dark periods as blue shaded. The homolog identifiers and p-value are shown below in matching colour. Only *kaiB1* is significantly oscillating in all considered datasets. The *kai* homologs other than *kaiB1* and *kaiC1* were omitted due to their sparse occurrence across the considered strains.

2.6 Discussion and conclusions

2.6.1 Diurnal expression in *Synechocystis* sp. PCC 6803

Microarray data indicate extensive diurnal expression change in *Synechocystis* and a distinctive biological feature Extensive diurnal expression is a phenomenon that has been observed in various cyanobacterial strains. In this work, the presented microarray-based dataset of the cyanobacterium *Synechocystis* sp. PCC 6803 revealed a large number of diurnally expressed genes. However, a global diurnal background trend in the mRNA signal is not explained by mRNA oscillations. Previously unknown circadian oscillations in the cellular rRNA content were experimentally confirmed. These changes in rRNA content introduced oscillations in the subsequently analysed mRNA fraction of the total cellular RNA extract. Common multi-chip normalisation methods were not applicable to selectively remove these undesirable effects of rRNA oscillation while preserving genuine mRNA oscillations. Specifically, several time series specific descriptors (phase, oscillatory p-value) and clustering analyses were employed to systematically assess the impact of four normalisation methods on the presented dataset.

Common microarray normalisation methods introduce extensive phase shifts due to diurnal background trend The common normalisation methods median polishing, quantile normalisation and cyclic LOESS (cLOESS) systematically changed the estimated expression phases of oscillating genes in comparison to the raw data. Expression phase information was best preserved by the reference gene-based least oscillating set (LOS) normalisation method. The underlaying assumption is that technical noise is reflected best by the change pattern which is common to the set of genes with least diurnal oscillation. This method selectively attenuated the diurnal background trend in reference gene set which was introduced by the circadian rRNA oscillation. Accordingly, LOS normalisation preserved the pairwise correlation distribution amongst all expression profiles. In contrast, quantile normalisation and median polishing substantially changed the original correlation structure by introducing anti-phasic oscillations. On the other hand, cLOESS suppressed oscillations without introducing anti-phasic ones. The estimated number of oscillating genes differed drastically with respect to the different normalisation methods due to the varying treatment of the oscillation in the mean transcript abundance. Overall, only LOS normalisation avoided the removal of the low-amplitude global trend, anti-phasic oscillations, or severe dampening of oscillations. However, it remains not possible to distinguish low amplitude oscillatory expression of individual genes which are in phase with the global trend from the genes of unchanged expression following a superimposed trend.

LOS normalisation differentiates between diurnal mRNA expression and diurnal background trend The mechanism leading to diurnal oscillations in the mean transcript abundance, despite the consistent application of $1.5\mu g$ RNA on each individual microarray chip, may have several additive components. The employed microarrays probed only a subset of genes excluding ribosomal RNAs. The suspected circadian

oscillatory pattern of ribosomal RNAs was subsequently experimentally confirmed for the 16S and 23S rRNA [11]. The diurnal variations of the fraction of probed to non-probed transcripts in the total RNA extract led to an anti-phasic trend, with a peak phase around CT 0. A constant level over the subjective day and a linear decrease during the night followed an instantaneous increase at dawn (inverse to Fig. 2.5). However, none of the gene clusters displayed a corresponding profile (Fig. 2.4). Furthermore, the majority of genes peak between CT 4 and 7 rather than CT 0 (Fig. 2.6). Accordingly, it is deduced that the extensive diurnal expression oscillation is not exclusively due to the rRNA oscillations. Instead, the influence of the rRNA trend must be small compared to the differential expression signal in individual genes. However, the amplitudes of day- and night-peaking genes might be enhanced and reduced, respectively, and the observed phase distribution might have been skewed. The functional relevance of the diurnal rRNA oscillation is yet unknown. In addition to being a segment of the ribosome, long rRNAs could serve as storage metabolite during the night which is degraded at the DL transition to facilitate phototrophic growth or jump-start the Calvin cycle early in the morning. Such a mechanism was recently reported for *Saccharomyces cerevisiae* which reacts with RNA degradation upon carbon starvation, producing ribulose-5-phosphate [258]. Furthermore, microarrays are known to be susceptible to biases in the GC content of the probes. Substantial overrepresentation of GC-rich sequences amongst day- or night-expressed genes might thus lead to oscillations. However, the analysis of the different gene clusters did not yield sufficiently strong signals. While the normalisation of the microarray signal to the cell number via spike-ins proved useful to examine a global expression induction [124], this approach does not allow the detection of sequence biases or rRNA fraction changes.

Overall, only LOS normalisation removes a low-amplitude diurnal background trend caused by oscillating cellular rRNA content, while maintaining the extensive diurnal oscillations amongst the mRNA.

A clustering approach to characterise diurnal organisation The presented analysis demonstrated that the clustering analysis result is governed by the choice of the normalisation method and that it is widely independent of the data transformation, similarity measure, or clustering algorithm. The only exception are transformations like the log_2 mean ratio, which emphasises amplitude information more than the standardisation and DFT transformation. However, amplitude information is less informative in order to define the temporal order of expression. It was thus attenuated by standardisation and DFT transformation, which allowed for exclusive clustering by the pattern of change. The comparison of individual expression profiles with existing biological knowledge showed that the combination of LOS normalisation, clustering via flowClust and after DFT transformation yielded the best results. Functional enrichment analysis of the resulting clusters were used to outline the basic diurnal biological program of *Synechocystis*. Other normalisation methods caused large phase shifts or the attenuation of diurnal oscillations, which are inconsistent with biological knowledge. For cyanobacteria, standardised more robust multi-chip normalisation methods must be used to study temporal expression

organisation. Furthermore, the biological function of the diurnal rRNA oscillation should be investigated.

2.6.2 Diurnal expression patterns are strain-specific

Characteristic phase distributions in cyanobacteria In cyanobacteria, all peak expression phase distributions reveal a bimodal distribution (Fig. 2.6), which were separated by ≈12 h in most datasets. Such a bimodality has been observed in many circadian clock models such as mammals and fungi [80, 262]. In most organs of the mouse, circadian genes show a preference for peak expression before subjective dusk or dawn while genes with intermediate phases are less common [262]. Interestingly, the genes peaking around subjective dusk or dawn tend to have higher amplitudes compared to genes with other phases. These observations led to the description of dawn and dusk as "transcriptional rush hours". Reassuringly, biological replicates reproduced the strain-specific phase distributions. However, the main times of diurnal transcriptional activity, *i.e.* the phase distribution modes, varied between the cyanobacterial strains. In contrast to mammals, cyanobacteria feature transcriptional rush hours with strain-specifically adjusted timings. Here, it is important to emphasise that the impact of data pre-processing, such as multi-array-normalisation (Section 2.2) or linear detrending in short time series (Chapter 4), on the observed diurnal expression parameters can be profound. Therefore, the application of a standardised pre-processing could further improve the fidelity of this analysis.
So far, the reports on the diurnal transcriptional regulation of *Synechocystis* sp. PCC 6803 have been conflicting. For instance, Kucho *et al.* described 237 circadian expressed genes under LL conditions [107]. Circadian gene phases clustered around dawn and dusk, with more genes featured in the latter. In contrast, Labiosa *et al.* described more than five times as many diurnally transcribed genes when using LD condition [110]. Importantly, peak phases cluster mainly during subjective day and night. The *Synechocystis* datasets in this analysis featured a large group of day-expressed genes of similar phase and only few night-expressed genes under LD conditions. Both datasets featured extensive midday-expressed (CT 6) gene clusters and only few night-expressed (CT 18) genes. This indicates a large number of DL transition-induced genes being detected in addition to clock controlled genes. Unfortunately, the Kucho and Labiosa datasets were not included in this comparative study. On the other hand, additional experiments revealed no detectable circadian clock-driven gene expression under LL and DD conditions [11]. These results indicate that the circadian clock of *Synechocystis* sp. PCC 6803 does not work under certain experimental conditions, as recently described by Xu *et al.* for *Synechococcus elongatus* PCC 7942 [257].

Standardised oscillating gene detection resolves variation in diurnal gene fractions The variation in the fraction of diurnally oscillating genes in cyanobacteria is renowned. However, diverse methods for microarray preprocessing and oscillation detection were employed to yield these results (Table 2.1). The application of a standardised oscillating gene detection procedure (harmonic regression) reduced the observed variance

between four different strains in seven datasets 2.5-fold. Two strains had to be excluded: firstly the *Microcystis* dataset since the covered time span was too short and sparsely sampled to reliably detect diurnal expression via harmonic regression. Secondly, the *Anabaena* dataset was not used due to its biological distinctiveness (heterocyst under nitrogen-fixing condition). Accordingly, this dataset yielded significantly fewer diurnal genes consistently for both oscillation detection methods, and a unique expression phase distribution with a majority of dusk genes (CT 20-22). The different extents of diurnal expression between the redundant *Synechococcus* and *Synechocystis* datasets disappeared when using the same procedure (Fig. 2.7 right), consistent with the highly similar phase distributions (Fig. 2.6). Only a $\approx 2h$ delay between the *Synechococcus* datasets was found, together with a wider dusk gene cluster in the *vijayan09* dataset. The latter points to a finer temporal regulation as preparation for the increased photosynthetic activity and growth rate, perhaps permitted by the CO_2 addition in this experiment. Otherwise, the addition of CO_2 had no influence on the extent of diurnal expression. Only for *Cyanothece* the variation of the diurnal expression extent decreased in the *toepel08* dataset compared to the *stoeckel08* dataset. While the oscillation amplitude correlates (Fig. 2.8 C), some differences are found in the phase distributions. While the *stoeckel08* dataset features one large dusk gene cluster at CT 15, the *toepel08* dataset exhibits a smaller dusk gene group which is subdivided into a CT 12 and CT 17 group. These differences were also found in the expression phases of the prominent ribosomal genes, which were drastically delayed from CT 17 in *stoeckel08* up into the subjective day $>$ CT 0 in *toepel08* (Fig. 2.11 and 2.12 A), and in the very low amplitude expression pattern of *kaiA* (Fig. 2.13). These differences likely reflect the different light regimes. While the second subjective night phase in the *toepel08* experiment was replaced by constant illumination, the *stoeckel08* dataset reflects continuous LD cycles. Apparently, the *stoeckel08* dataset features a significant fraction of genes which are induced by the LD transition. A differential analysis of both datasets could thus be used to distinguish clock-driven from light-induced genes, allowing a more accurate analysis of regulatory sequence features in both groups.

The pairwise comparison of peak expression phases between biological replicate datasets revealed consistent phases for diurnally expressed genes. In contrast, this consistency of peak phases does not hold for orthologous genes across different strains, which revealed extensive peak time variations. This suggests, that diurnal expression patterns have adapted in a strain-specific way to the universal challenge of alternating photic and aphotic phases, potentially driven by the differences in the lifestyle and environment. Interestingly, diurnal expression patterns were equally common amongst the cyanobacterial core genome, the genes common to 16 cyanobacterial species as defined by Beck *et al.* [12], as amongst strain specific genes. Similarly, metabolic genes did not show increased propensity for diurnal expression. These observations suggest that diurnal and circadian transcriptional regulation is highly adaptable in cyanobacteria and does not constitute a conserved paradigm. Instead, cyanobacteria implement transcriptional regulation at different locations in the metabolic network with different phases, which still might lead to similar metabolic fluxes through the corresponding pathways.

2.6.3 The core diurnal genome comprises central metabolic functions

The comparison of oscillatory expression properties between eight datasets of five strains yielded several insights. The five considered strains share a core diurnal genome, *i.e.* genes that consistently exhibit diurnal expression patterns in all available experiments. While the asymptotic size appeared to be constant, its actual size remains unknown. With 95 genes as found here, the core diurnal genome is considerably compared to the number of strain-specific diurnal genes. The expression phases of individual core diurnal genes agree well with previous knowledge. In *Synechococcus*, the core clock genes *kaiB* and *kaiC* are arranged in the *kaiBC* operon resulting similar expression patterns [81, 84, 256], while *Cyanothece* features the *kaiAB1C1* operon [133]. Interestingly, *Cyanothece* featured consistent anti-phasic expression of *kaiB1* and *kaiC1* whereas the remaining strains showed co-expression, hinting at *Cyanothece*-specific post-transcriptional regulation of *kaiB1* or *kaiC1*. Oscillations in the *kaiA* gene expression, as reported by Ishiura *et al.*, featured small expression amplitudes compared to *kaiB* and *kaiC* [82]. In fact,*kaiA* consistently fell below the threshold of 2-fold expression change for the classification as circadian oscillator, which is commonly employed in microarray studies. This demonstrates again that the amplitude information must be considered secondary to the expression pattern, and that comparative studies are a valuable tool to emphasise consistently occurring low-amplitude oscillations which are otherwise disregarded.

In contrast, the crucial carbon fixation enzyme RuBisCO exhibited high amplitude diurnal expression oscillations (Appendix A Fig. 5) under LD and also LL conditions, indicating at least indirect circadian clock input. This contrasts experimental findings, in which RuBisCO expression is largely constitutive with only limited induction under CO_2 limitation [4, 172]. Interestingly, the amplitude of diurnal oscillations of RuBisCO expression in *Synechococcus* was more doubled when providing substantially more CO_2 but almost half as much light (*vijayan09*, 1% CO_2, 25 photons $m^{-2}s^{-1}$), as compared to the normal CO_2 supply via air (*ito09*, atmospheric 0.04% CO_2, 40 photons $m^{-2}s^{-1}$). Despite its central metabolic function, the transcriptional controls required for upregulation of carboxysome genes and specifically RuBisCO genes are presently not known. It can therefore only be speculated, that higher CO_2 availability in the *vijayan09* experiment permited higher carbon fixation rates, which in turn required more facilitating enzyme.

The peak expression of ATP-synthase at dawn, as pointed out for *Synechococcus*, is a consistent feature of the considered strains [83] and illustrates the increased metabolic activity due to photosynthesis as compared to respiratory activity during aphotic phases. The glyceraldehyde-3-phosphate dehydrogenase genes *gap1* and *gap2* represent anti-phasic expressed isoenzymes, a pattern that is shared amongst the considered strains. *gap1* facilitates glycolytic glucose breakdown whereas *gap2* is used in the photosynthetic Calvin cycle and the non-photosynthetic gluconeogenesis [101]. The observed peak expression of *gap2* during dawn before the photic phase and of *gap1* during dusk are compatible with this notion, indicating the synthesis of glucose during the day and subsequent breakdown during the night. In contrast, the expression patterns of genes coding for the enzymes necessary to convert glucose to the major storage metabolite glycogen were diverse and did not indicate circadian regulation. On the one hand, this might indicate

the application of alternative storage metabolites in some strains. On the other hand, the employed experimental conditions might not permit extensive storage generation as they are not optimised for fast growth. Instead, slow growing conditions were deliberately selected for technical reasons in case of *Synechocystis*. However, expression regulation studies must be interpreted carefully with respect to biological processes facilitated by the corresponding proteins due to the various intermediate processes modifying the temporal relationship between mRNA and protein. This was demonstrated for the ribosomes of the *Shewanella* bacterium, for which the expression changes under dark-stress to ensure a constant concentration of the corresponding protein [213].
This chapter considers the hypothesis that cyanobacteria feature a general diurnal expression program due to the similar requirements of the oxygenic phototrophic lifestyle and despite their varying lifestyles. The transitions between photic and aphotic phases play an important role due to the necessary vast changes in metabolic processes. The analysis was complicated by the fact, that not all available experiments were conducted under constant light conditions. As a result, some datasets reflect the combination of circadian-clock regulated and light-induced genes, or possibly only light-induced genes due to clock conditionality. Accordingly, the purpose of this analysis is a) to delineate the set of genes which is critical for the adjustment to photic and aphotic phases, b) to suggest candidates for new clock-driven genes in strains with a known working clock, and c) to suggest marker genes that can be used to test for a working circadian clock. The estimate of the core diurnal genome based on the presented data collection spans 95 genes which are mostly involved in central metabolic processes (*e.g.* ribosomal proteins, photosynthesis). The analysed whole genome expression datasets allow for the extrapolation of the number of genes in the core diurnal genome beyond the strains explicitly considered in the analysis. The size of the core diurnal genome remains constant with 5 or more datasets (Fig. 2.10) which hints at a small core diurnal genome which is shared amongst all cyanobacteria. While observed phases in the core diurnal genome agree widely with previous knowledge, the peak expression phases of other diurnal genes did not demonstrate substantial correlation between different strains. This indicates a high degree of adaptation of diurnal and circadian regulation to the strain-specific environment. With the exception of ribosomal proteins, diurnally expressed genes are found interspersed across the metabolic network, but are not specific for certain pathways. Additionally, 15 genes annotated as "hypothetical" constitute a candidate set for further study. The core diurnal genome is a good starting point to study sequence determinants of clock input in a focussed fashion, which could potentially improve the weak signals observed, *e.g.* in *Synechocystis* [236].

3 Periodic nucleotide sequences in cyanobacterial chromosomes

Cyanobacteria feature several remarkable circadian regulatory properties, not least the circadian rhythm in the chromosomal supercoiling state reported by Vijayan et al.*. The long-standing hypothesis that sequence periodicity with* ≈ 10.5 *bp period indicates the preferred supercoiling state (Trifonov and Sussmann, 1980), together with the remarkably strong genome-wide periodicity in cyanobacteria were the starting point for this work. Windowed AT-tract periodicity analyses of promoter regions, which yielded subtle differences for supercoiling-induced and -repressed genes in E. coli, did not provide conclusive results for the corresponding gene sets in Synechocystis and Synechococcus. This raised the questions whether AT-tracts are indeed the crucial motifs and whether these strains are relevant. Both questions are answered here. Inspired by the cyanobacteria-specific short intergenic sequences which complicate the detection of periodic patterns in contrast to the E. coli study, we improved upon described techniques to achieve highly localised periodic pattern detection independent from genomic features. Interestingly, we found that genes contribute more to the genome-wide periodic signal rather than the intergenic regions. However, genes with shared periodic patterns do not exhibit coherent transcriptional regulation with respect to the supercoiling state and diurnal profile. These findings do not support the hypothesis connecting circadian supercoiling to circadian gene expression via sequence periodicity. The subsequent phylogenetic analysis revealed the association of stronger genome-wide periodicity with increased chromosomal copy numbers. This led to the modified hypothesis that the genome-wide periodicity could aid chromosomal packaging in the presence of many chromosome copies, but is not involved in transcriptional regulation of individual genes. The comparison of spectral gene clusters with supercoiling-sensitive genes, as well as the phylogenetic analysis and a vast set of biological knowledge were contributed by Rainer Machné, with whom I generated the figures and wrote the manuscript. The fruitful collaboration with Hanspeter Herzel has led to this chapter, which has been published as "Lehmann"* et al. *in Nucleic Acids Research [116].*

3.1 Periodic dinucleotide sequences are implicated in DNA structure

Regular periodic occurrences of nucleotides along the DNA sequence are well known in many species for a long time (Trifonov *et al.* 1980 [222], Satchwell *et al.* 1986 [189]). In archaea and eukaryotes, sequences which are wrapped around the nucleosomes often

feature periodic nucleotide occurrences with 10-10.5 bp period [27, 148], reflecting the helical pitch (length of one full DNA double-helix subunit), the nucleosome positioning code. In contrast, the signal in bacterial genomes with ≈ 11 bp period is less well described and understood [72, 139, 190, 218, 254]. Generally, dinucleotides yield stronger periodic signals than mononucleotides, but counting dinucleotide motifs consisting of A and T (WW in IUPAC notation) often exhibit the strongest signal [139]. This has been interpreted in reference to the different structural properties of the different possible dinucleotides. While short runs of A and T nucleotides without the TpA step (A-tract or AT-tract), induce a bend of the DNA backbone into the minor groove of the helix, other dinucleotides do so much less. Inserting such bends, *i.e.* AT-tracts, along the DNA helix in phase ("phased AT-tract") with the local pitch, *e.g.* 10.5 bp, the bends along the helical axis accumulate and lead to persistent intrinsic backbone curvature [183]. Notably, the DNA pitch is not constant and can be modified by over- or under-twisting the helix, commonly referred to as positive and negative supercoiling [72, 73]. DNA with modified pitch assumes preferential structures, so called plectonemes (Fig. 3.1 A left, period >10.5 bp) or solenoids (torus-shaped loops, Fig. 3.1 A right) [220]. In contrast, DNA wound around nucleosomes or other proteins assumes solenoidal loops but is still over-twisted (pitch <10.5 bp).

DNA plectronemes exhibit high axial curvature in loops, in which atomic force and electron microscopy experiments preferentially locate phased AT-tracts [112, 162, 193, 226]. The overlap of these curved plectroneme loops with promoter regions could facilitate transcriptional regulation [69, 90, 156, 165]. In such a mechanism, the torsion of negatively supercoiled DNA could aid double-helix unwinding during transcriptional initiation [142]. Following this idea, different dinucleotide periods in promoter sequences have been suggested to explain differential transcription (10.3 and 11 bp in *E. coli*) in response to changing (negative) DNA supercoiling in bacteria [106, 142, 151]. The supercoiling state of DNA in bacteria is altered by topoisomerase enzymes, which can introduce negative supercoiling via double strand breaks under energy consumption (*gyrA*, *gyrB*), or relax supercoiling (topA) without the need of energy.

While cyanobacteria do not possess proteins homologous to nucleosome subunits in eucaryotes and archaea, nucleoid associated proteins such as HU wrap DNA in solenoids [131]. Accordingly, a signal similar to the nucleosome positioning code could exist for NAPs.

Periodic signals in DNA sequence have different sources. Most prominently, the bias in codons employed to encode the amino acid sequence introduces a strong signal with 3 bp period [195, 205, 225]. The secondary structure of the encoded RNA contributes to this 3 bp signal [194, 224]. A well known source for 10-11 bp periodic patterns are amphipathic *alpha*-helices [247, 263]. Since such helices span only 30 bp, they can be distinguished from longer-spanning signals with similar period, such as the 100 bp spanning 10-11 bp periodicity in archaeal and bacterial genomes [37, 72, 139, 217]. In *E. coli*, phased AT-tracts with the period of 11 bp and of length 100 bp were found to occur in clusters along the genome covering intergenic and also coding regions. Tolstorukov *et al.* suggested for this AT-tract distribution to reflect a "structural code for DNA condensation into a nucleoid" [217].

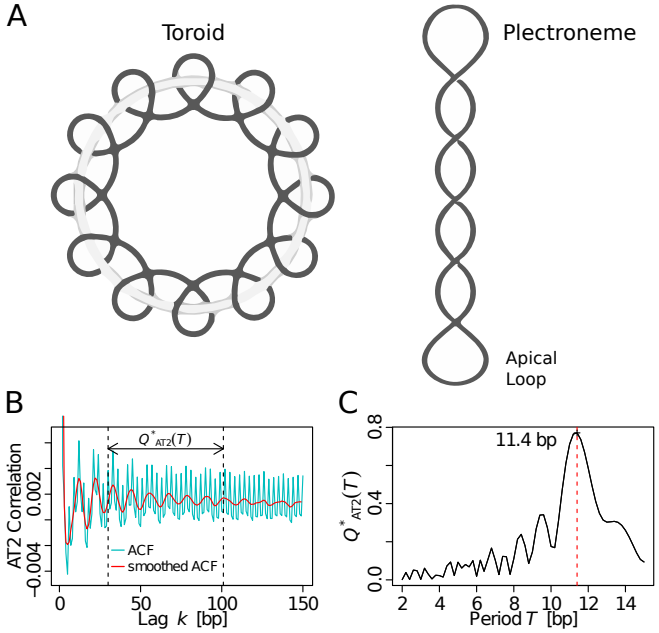


Fig. 3.1: **Plectronemic and toroidal DNA structure and dinucleotide periodicity in cyanobacterial genomes.** (**A**) DNA can assume plectronemic (left) or toroidal (right) geometries, depending on supercoiling-status. DNA is toroidal if wound around nucleosomes. Adapted from Travers and Muskhelishvili [221]. (**B**) The autocorrelation function (ACF) of chromosomal AT2 motif positions in *Synechocystis* sp. PCC 6803 before ($C_{NN_NN}(k)$, blue) and after 3 bp window smoothing ($\bar{C}_{NN_NN}(k)$, red). The dotted vertical lines indicate the range used to quantify periodicity. (**C**) Power spectrum $Q^*_{AT2}(T)$ of the interval $k = [30, 101]$ bp of the smoothed ACF. The dominant AT2 period at 11.4 bp is marked (red dashed vertical line).

In previous analyses, cyanobacterial genomes exhibited exceptionally strong 11 bp periodic signals [139]. Another remarkable feature is the chromosomal supercoiling state, which oscillates in a circadian fashion in light/dark cycles [252] and also under constant light [237]. Moreover, this oscillation correlates with the genome-wide remodelling of the transcriptome [237]. These observations predispose the cyanobacterial clade to test the suggested relations of the periodicity signal to negative DNA supercoiling [72], and thereby to in supercoiling-dependent mRNA transcription [106, 151] or DNA packaging [217].

The cyanobacterial clade is morphologically and genetically very diverse. Cyanobacteria are traditionally classified into five morphological subsections [181], which differ in the mode of cell division, and also include multicellular and differentiated lifestyles. Several independent transitions in morphology are found throughout the cyanobacterial phylogeny [22, 191, 230]. Comparative genomics analyses are not able to identify signature genes for complex morphologies, presumably due to frequent horizontal gene transfer. However, filamentous species tend to have a higher number of signalling and regulatory proteins [198].

3.1.1 Outline of this chapter

This chapter presents the first systematical assessment of sequence periodicity in the cyanobacterial clade. The comparison between 54 strains shows AT-tracts as strongest periodicity source. The loss of strong genome-wide periodicity is associated with transitions in cell morphology or lifestyle. A windowed scan allows to localise the signal along the genomes. Coding sequence is identified as main periodicity source, and particularly the first and third codon position within the coding sequence. Two distinct transposons show a strong ≈ 11 bp signal. Potential functions in the formation of an active transpososome or regulation of transposase transcription are discussed in this context. Importantly, no strong connection between periodicity signal and supercoiling-dependent mRNA transcription is found. This could indicate a role of periodicity in DNA packaging, instead of gene-specific transcriptional regulation.

3.2 Materials and methods

The following paragraphs provide a summary of the materials and methods, which is elaborated upon in Appendix B.

Cyanobacterial genomic sequences, lifestyle, and phylogeny The genomic sequences of 54 cyanobacterial strains constituted the basis for this systematical periodicity study, including the strains which have recently been sequenced by Shih *et al.* [198]. Since different aspects of sequence periodicity have been studied in detail in non-cyanobacterial species, these were considered as reference species in this study. Most notably, the two related enterobacteriaceaeal species *E. coli* K-12 and *Dickeya dadantii* 3937 were included, due to extensive periodicity studies in the promoter regions of the former [105], and to the reports of transcriptional regulation of genes by supercoiling in both [153, 164]. Furthermore, the archaeum *Methanococcus maripaludis* S2 exhibited strong genome-wide periodicity [139]. On the other hand, the unicellular eukaryote *Saccharomyces cerevisiae* (chromosome IV) was included because of the ≈ 10 bp periodicity reported in connection with the nucleosome positioning code [89]. The genome selection was limited to sequences ≥ 1 Mbp excluding plasmids and genomes marked as unfinished. The genomic sequences, protein-coding sequences (CDS), and intergenic annotations were obtained from the GenBank [16] or from the Joint Genome Institute database [61]. Phylogenetic trees were obtained from Shih *et al.* [198] (based on 31 conserved proteins) and the Integrated Microbial Genomes (IMG) database [129] (based on 16S rRNA alignments of the SILVA database [174]). Information to the lifestyle of cyanobacterial species were obtained from the IMG database and Shih *et al.*. Functional annotations and InterProScan protein domain hits for CDS for *Synechocystis* sp. PCC 6803 and *Cyanothece* sp. 8801 were obtained from the CyanoBase database [146]. The curated annotations along with the results reported here are listed in a table in Appendix B File 1.

Transcriptional data The diurnal time-series data of *Synechocystis* sp. PCC 6803 was used in this study as presented in Chapter 2. The work of Prakash *et al.* [171] reports 4 groups of *Synechocystis* sp. PCC 6803 genes according their transcription with respect to an increase in gyrase-induced negative supercoiling. Besides consistently up- (group 1) and down-regulated (group 3) genes, some genes showed a mixed response (group 2), while a large number was non-responsive (nr). A table of CDS with the additional curated data is provided as Appendix B File 2, additional data are provided in Appendix B.

Dinucleotide periodicity measures and statistics Periodicity of genome-wide and locally occurring motifs was quantified using the autocorrelation function (ACF) following the approach presented by Schieg and Herzel [190] (Fig. 3.1 B blue). The strong signal with 3 bp period induced by the codon usage bias was removed via smoothing with a 3 bp sliding window across the ACF (Fig. 3.1 B red). The Fourier transform was used to derive the power spectrum of the ACF in the range from $k_{min} = 30$ bp to $k_{max} \approx 101$ bp in order to exclude short-range signals induced by amphipathic α-helices (< 30 bp) [72]. To compensate for varying sequence composition across strains, the normalised power spectrum $Q^*_{NN}(T)$ was obtained (Fig. 3.1 C) as described by Mrázek [139]. For more details on the procedure, see Appendix B Section 3.1.

Extensions to localise periodicity The ACF based periodicity quantification was extended to detect local dinucleotide periodicity occurrences. Each sequence was subdivided into non-overlapping windows of 200 bp width, to which the ACF procedure was applied. The empirical periodicity p-value $P_{T,i}$ for each window i and period T was derived. Therefore, the background distribution for the periodicity statistic was approximated with random permutations of the window sequence ($n = 5000$). The sequence permutations were obtained using `uShuffle` [88] with preservation of dinucleotide content, to account for local variation of sequence composition. Optionally, the spectra of consecutive window i and $i + 1$ were averaged (termed 200Avg4) to achieve higher sensitivity and accepting lower spacial resolution. More details are provided in Appendix B Section 3.2.

Spatial overlap of periodic windows with genes To test whether periodic sequence windows exhibit any preference for coding or intergenic regions, the overlap was statistically tested. First, periodic windows ($P_{T,i} < 0.01$) were sorted into four groups of similar period (< 9 bp,$9 - 10.5$ bp,$10.5 - 12$ bp,> 12 bp). Within each group, adjacent windows were then concatenated to yield sets of non-overlapping periodic segments. The overlap between these four sets and CDS or intergenic sequence was then statistically tested using the Jaccard test with interval permutation as provided by the R package `GenometriCorr` [50]. For additional information see Appendix B Section 3.3.

Localising periodicity in coding sequence via permutation To quantify the contribution of codon order, synonymous codon preference (codon usage), and individual

codon position (I-III) to the observed genome-wide periodicity, different random permutation methods were applied to the CDS sequences to remove potential regularities from the original sequence. Comparison of periodicity strength before and after permutation ($Q^{CDS}_{AT2}(T)$ of the concatenated CDS) then indicates the contribution to the overall observed signal. The permutations were performed using a customised version of the R package `seqinr`. Replacement of synonymous codons was performed under preservation of the overall codon usage bias, similarly was the position-wise codon permutation performed to preserve the overall nucleotide composition. To quantify the feature-wise contribution to the overall-signal, the ratio of AT2 periodicity strength $Q^{CDS}_{AT2}(T)$ at $T = 11.8$ bp before to after permutation was introduced (Appendix B Section 3.4).

Spectral CDS clustering and enrichments After calculating the normalised AT2 periodicity spectra $Q^{*}_{AT2}(T, i)$ for individual CDS within a cyanobacterial strain, these spectra were hierarchically clustered (Ward's method, euclidean distance, $k = 14$). These clusters of CDS of similar dominant period were then compared to a range of other gene properties using a hypergeometric test. For gene features with observed significant overlap to the spectral clusters, multiple testing-corrected p-values via Benjamini-Hochberg are reported (see Appendix B Table 14 and 15)

3.3 Genome-wide sequence periodicity across the cyanobacterial clade

The genome-wide periodic occurrence is compared between a range of cyanobacterial genomes across all possible dinucleotides and the two motifs WW and AT2. The strongest periodicity and a pattern across the strains is observed for AT2 motifs, indicating functional relevance and motivating the focus of following analyses on this motif.

Comparing dinucleotides by genome-wide periodicity strength This study focussed on dinucleotide motifs (NN in Iupac notation), which are considered as the minimal unit in the context of DNA structure [24, 113]. Previous studies also explored WW motifs (all possible A and T permutations) and AT2 (WW excluding TpA, a minimal AT-tract) [72, 139, 190]. Here, the genome-wide periodic occurrence of these motifs was quantified as spectral signal-to-noise ratio $Q^{*}_{SNR}(NN)$ for the interesting period range 10-12 bp against the remaining evaluated spectrum (Fig. 3.1 and Appendix B Section 3.1). Hierarchical clustering of the 54 considered cyanobacterial strains and four reference species by the periodicity strength profile across the dinucleotide motifs revealed that WW induces the strongest grouping (Fig. 3.2 A). Particularly the AT2 motif was the strongest periodic motif in 35 species, followed by WW in eight species. Accordingly, the species clustering was dominated by the strain groups with varying AT2 periodicity strength (clusters A-D). In contrast, there were strongly periodic strain specific dinucleotides. The small group of three highly periodic *Cyanothece* strains exhibited

periodic CpA and TpG dinucleotides. Another example is the archaeum *Methanococcus marsipaludis* S2, which featured periodic ApC and GpT dinucleotides in addition to the strong AT2 periodicity. In contrast, only *Synechococcus* sp. PCC 7335 exhibited very low periodicity in all other dinucleotides but the strongly periodic WW/AT2 motifs. This analysis confirmed the observation of wide-spread strong AT2 periodicity amongst cyanobacteria, with a typical period of 11-11.6 bp (data not shown, similar to Fig. 3.1). Interestingly, the two enterobacteriaceaeal reference species (genome-wide period 11 bp) and budding yeast (9.8 bp) exhibited only weak AT2 periodicity despite its reported functional relevance in these species . In contrast, the archaeum harboured a very strong signal but with significantly different dominant period (9.8 bp).

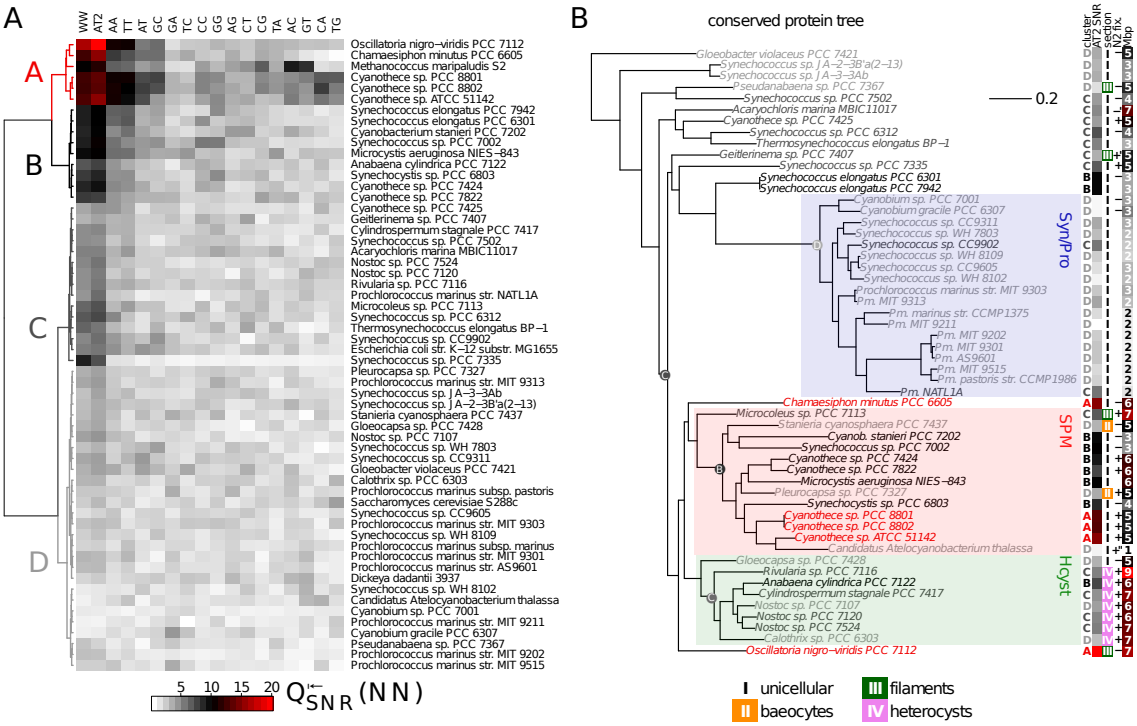


Fig. 3.2: **Genome-wide dinucleotide periodicity in different groups of the cyanobacterial clade.** (**A**) Strength of the 10-12 bp periodicity across 54 cyanobacterial and 4 "control" species genomes (rows) and all dinucleotides NN, WW and the AT-tract motif AT2 (columns). Signal strength is measured by the power spectrum signal-to-noise ratio Q^*_{SNR} (black low, red high). The species were hierarchically clustered (Ward's method, complete linkage, $k = 4$, tree on the left) to obtain clusters A-D. (**B**) Phylogenetic similarity tree for the corresponding strains (alignments of 31 conserved proteins, [198]), coloured shading indicates major subclades ($\geq 70\%$ bootstrap-support). Annotations provided on the right, corresponding to the row's strain: cluster index, colour-coded AT2 periodicity strength ($Q^*_{SNR}(AT2)$), morphological section, ability for nitrogen fixation, and genome size excl. plasmids [Mbp].

Correlation of genome-wide periodicity with lifestyle Cyanobacteria can be grouped according to morphological properties into five section [181]. Phylogenetic studies have shown that morphological sections do not reflect true relationships amongst the taxa. Possibly, none of the morphological sections originates in only a single ancestor, implying that morphologies must have evolved repeatedly. Attempts to reconstruct ancestral cyanobacteria point to unicellular, freshwater-inhabiting cells with small diameters, which lack the traits to form thick microbial mats [22]. The number of cyanobacterial clades varies between different studies with the choice of the underlying sequences, the number of included strains, and parameters. However, the core cyanobacterial tree described by Blank and Sánchez-Baracaldo [22] features three clades (SPM, PNT, SynPro) and is consistently found in the works of others, such as Schirrmeister *et al.* [191] (three clades), Shih *et al.* [198] (seven clades), and Turner *et al.* [230] (ten clades). With SPM (*Synechocystis*, *Pleurocapsas*, *Microcystis*), only one of the core clades was found to feature strong AT2 periodicity (Fig. 3.2 B and Appendix Fig. 7 A). This clade is dominated by members of morphological section I, featuring unicellular lifestyle with binary fission or budding. Importantly, only the four strains in this clade, which exhibit weak periodic signal, underwent drastic transitions in cellular morphology or lifestyle. Two strains changed morphologically to section II, reproducing by multiple fission into several small (greek baeo-) daughter cells and, thus called baeocytes [245] (*Stanieria cyanosphaera* PCC 7437, *Pleurocapsa* sp. PCC 7327). The strain *Candidatus Atelocyanobacterium thalassa* (alternative name cyanobacterium UCYN-A) has adapted to an unique symbiotic lifestyle with extensively reduced metabolism [214, 223]. Only *Microcoleus* sp. PCC 7113, located at the base of the SPM clade, is part of section III featuring undifferentiated filamentous lifestyle. Its intermediate level of AT2 periodicity is consistent with the majority of filamentous species, mainly found in the Hcyst clade. The filamentous Hcyst strains are grouped into section IV due to their ability to differentiate into cells specialised in nitrogen-fixation (heterocysts), thereby marking filamentous differentiated lifestyle compared to the filamentous un-differentiated lifestyle of section III. Two strains with very high periodic signal are found at the base of the two clades Hcyst/SPM, *Chamaesiphon minutus* PCC 6605 and *Oscillatoria nigro-viridis* PCC 7112. However, their placement in the tree achieved only small bootstrap support, and phylogenetic tree construction using 16S rRNA placed them within the Hcyst clade [198] (Appendix Fig. 7 A).
The second main clade Syn/Pro is characterised by weak genome-wide periodicity. It is comprised exclusively by picocyanobacteria of section I, which feature small cell size, a marine habitat, and streamlined genomes with reduced regulatory capacity for a minimal oxyphototroph lifestyle [264]. While *Prochlorococcus* strains exhibit a strict coupling of S-phase to cell division, *i.e.* a monoploid lifestyle [234], some marine *Synechococcus* strains were found to harbour multiple genome copies [19]. Several cellular properties correlate with the absence of genome-wide AT2 periodicity, such as cell size, genome length, and the number of metabolic genes relative to the number of genes (Appendix Fig. 7 B). Interestingly, the placement of *Synechococcus* species of cluster B with strong genome-wide periodicity has only low bootstrap support in the study of Shih *et al.*[198] and also in that of Blank *et al.*, both strains could instead branch basal to the SPM and Hcyst clade. This uncertainty reflects the general problems of current phlyogenetic

methods to place several cyanobacterial strains such as *Thermosynecho-coccus elongatus*, *Acaryochloris marina* and *Synechococcus elongatus* [22]. The remaining, more distantly related, species which are not included in the three main clades exhibit intermediate or weak periodicity (cluster C or D). Genome periodicity appears to have evolved in the SPM/Hcyst sister clades. Unfortunately, the two strongly periodic basal placed strains *Chamaesiphon minutus* and *Oscillatoria nigro-viridis* hamper the inference of the ancestral periodicity strength. Accordingly, there is no evidence on whether genome-wide AT2 periodicity evolved in SPM or was abolished in Hcyst.
In summary, AT-tracts represented by the AT2 motif reveal the strongest genome-wide periodicity. This is consistent with the proposed role of phased AT-tracts for the induction of intrinsic curvature in the DNA backbone facilitating DNA packaging [217]. Importantly, strong genome-wide periodicity is limited to the unicellular SPM clade, and periodicity loss accompanies lifestyle and thereby morphological transitions.

3.4 Genomic localisation of periodic dinucleotide sequences

The previous step has demonstrated the importance of AT2 as periodic motif on a genome-wide level in a subgroup of cyanobacterial strains, the SPM clade. However, all mechanistic interpretations predict genomic localisation of phased AT-tracts in, *e.g.* promoter regions [106], or in clusters of AT-tracts forming a genome-wide structural code for DNA packaging [217]. The genome-wide analysis is therefore extended to test for spatial occurrences of sequence periodicity, the results of which are presented in the following.

AT2 periodicity originates in coding sequence To obtain a localised periodicity measure, ten representative cyanobacterial strains from weak to strong genome-wide AT2 periodicity were selected. The genomic sequences were divided into 200 bp wide non-overlapping basic windows and the normalised Fourier spectrum $Q^*_{AT2}(T)$ was calculated. An empirical p-value was obtained for every period T indicating the probability to observe periodicity with equal or higher strength by chance (see Appendix B Section 3.2 for details). To each window, the period T_{sig} resulting in the lowest p-value was assigned. Inspection of the T_{sig} distributions for significantly periodic windows ($p < 0.01$) revealed that four *Cyanothece* strains and *Synechocystis* sp. PCC 6803 feature a group of periodic windows with ≈ 10 bp period in addition to the expected group with ≈ 11 bp period, as indicated by the genome-wide analysis (Appendix Fig. 8 A-J) and as described in [139]. To increase the sensitivity of the periodicity detection and uncover possible ≈ 10 bp window groups in other strains, an additional averaging of AT2 spectra $Q^*_{AT2}(T)$ over four adjacent 200 bp windows (200Avg4) was introduced. However, the resulting T_{sig} distributions reconfirmed the result: ≈ 11 bp periodic windows were found in weakly periodic species, *e.g.* in *E. coli*, without indicating ≈ 10 bp periodicity in additional

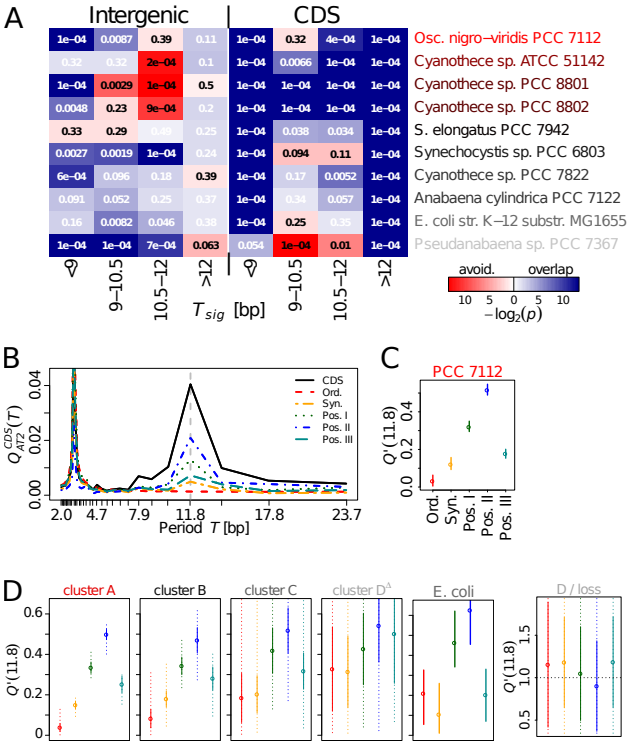


Fig. 3.3: **AT2 periodicity in protein coding sequence.** (**A**) Statistical overlap test of periodic genome segments with protein coding sequence (CDS) and alternatively intergenic regions using the Jaccard test [50]. Periodic 200 bp windows are grouped into four period ranges (columns), and adjacent windows in the same group are are concatenated. Species name (rows) colours indicate AT2 periodicity strength ($Q^{*}_{SNR}(AT2)$, see Fig. 3.2). (**B**) The AT2 periodicity spectrum ($Q^{CDS}_{AT2}(T)$) before and after permutation with various methods in the concatenated CDS of *Oscillatoria nigro-viridis* PCC 7112 (average of 50 random permutations). Ord.: codon order permutation, Syn.: synonymous codon replacement, Pos. I-III: codon position-wise permutations. (**C**) AT2 periodicity loss by permutation ratio ($Q'(T)$). The power at $T = 11.8$ bp (vertical dashed grey line in B) of the CDS spectrum is divided by the spectral power obtained after each permutation method is applied (mean and range of 50 permutations). (**D**) $Q'(T)$ as in C, applied to all members of AT2 periodicity clusters A-D (Fig. 3.2). Non-cyanobacterial reference species are not included, *E. coli* K-12 MG1655 is shown separately. Four strains with exceptionally low AT signal are removed from cluster D (D^{Δ}) and are separately shown ("D/loss"). Mean, standard deviation, and range of 50 permutations are represented as circle, solid line, and dashed line respectively.

strains. Due to the reduced spatial resolution in the averaging approach (200Avg4), the distinction between the ≈ 10 and ≈ 11 bp groups is hampered. This leads to an attenuation of the ≈ 10 peak and indicates spatial proximity between windows of both period.

With the localisation of periodicity in place, the presence of periodicity in specific genomic regions such as coding sequence or conversely intergenic regions was investigated. Periodic windows were pooled into four groups of similar period, as indicated by the previous analysis, and tested for statistically significant overlap with CDS and intergenic regions (Fig. 3.3 A, method details in Section 3.2 and Appendix B Section 3.2). This analysis yielded a differential localisation pattern for different periods. Most strikingly, ≈ 11 bp periodic windows significantly overlapped with CDS in most species, and avoided intergenic regions in *Cyanothece*. Conversely, the second group with periods of ≈ 10 bp tended to overlap with intergenic regions specifically in strains with lower genome-wide measured periodicity and *Oscillatoria nigro-viridis*. Intergenic enrichments are consistent with a range of previous observations and functional interpretations of bacterial promotor curvature [69, 90, 106, 151, 156, 165], as well as transcriptional termination regions [104]. However, the strong genome-wide periodicity signal can not be attributed to the promoter regions, which are remarkably short in cyanobacterial genomes, but is due to significant contributions of the CDS. The remaining groups with < 9 and > 12 bp period were considered as statistical background. Accordingly, no interpretable pattern was observed in the overlap study (both groups overlapped significantly with CDS, but did not yield significant results for intergenic regions).

Periodicity is encoded in specific codon positions Previous studies demonstrated that sequence periodicity is encoded in the protein coding sequence across all three kingdoms [37, 73], in particular in the third codon position. This was established using a range of permutation methods removing different aspects of the original sequence and comparing the resulting periodicity signals. However, these studies did not consider codon position-wise permutations and instead used synonymous codon permutation to test implicitly for the influence of the third codon position.

The concatenation of all CDS sequences for each respective cyanobacterial genome was subjected to the described ACF spectra $Q_{AT2}^{CDS}(T)$. The impact of five different permutation techniques on the sequence-wise ≈ 11 bp periodicity was then quantified by the ratio of signal strength in the original sequence to the permuted sequence ($Q'(T)$) at period $T = 11.8$ bp (Fig. 3.3 B-D). Pertubation of the amino acid sequence by shuffling the codon order (Ord.) under preservation of the overall codon usage bias, the most invasive permutation method, attenuated the ≈ 11 bp period completely. This yielded a Q' of 0-10% in the highly periodic species of cluster A and 0-30% in cluster B (Fig. 3.3 D). With weaker genome-wide periodicity, the impact of codon order permutation also weakened. Synonymous codon permutation (Syn.) constituted the next less invasive method, preserving the amino acid sequence of the gene but not its codon usage bias. This permutation method was used to test for the contribution of amino acid patterns such as amphipathic α-helices, which are known to induce periodicities [263]. The periodicity

signal in cluster A was also reduced to 10-20%, which confirmed that a significant fraction of the ≈ 11 bp periodicity is encoded in the most flexible third codon position. In contrast, contributions of the amino acid sequence were effectively attenuated by the employed method. Permutations of the codon positions I and III (Pos.I and Pos.III) under preservation of the sequence composition drastically decreased the ≈ 11 bp signal to 20-40%, with a stronger effect of position III. Position II (Pos.II) consistently yielded the highest Q' with ≈ 50%, *i.e.* the lowest periodicity reduction. Similar trends were found for genomes of cluster C and D including the reference species (*E. coli* shown in Fig. 3.3 D), but permutation impact was reduced due to the overall weaker periodicity. A group of four cyanobacterial strains stood out in this analysis, in which a CDS-wide ≈ 11 bp periodicity was completely absent in the original CDS sequences, but was introduced upon permutation (D/loss in Fig. 3.3 D, *Candidatus Atelocyanobacterium thalassa*, *Cyanobium* sp. PCC 7001, *Cyanobium gracile* PCC 6307, *Prochlorococcus marinus str.* MIT 9211.

3.4.1 Diurnal and supercoiling-sensitive transcripts

The previous analyses (Section 3.4) have demonstrated, that I) the main source of periodicity in cyanobacteria are CDS, and II) that strains can have periodic regions featuring periods different from the genome-wide dominant period. This motivated the question whether coding sequences feature characteristic periodicities. To address this point, the AT2 spectra $Q^*_{AT2}(T)$ of all *Synechocystis* sp. PCC 6803 and *Cyanothece* sp. PCC 8801 CDS were calculated as before. For each strain, the 1000 CDS with the highest maximal periodicity signal were selected and clustered using a hierarchical approach. This yielded clear spectral clusters of genes, where most clusters featured a single dominant period (Fig. 3.4 A, Appendix Fig. 9-10, and Table 14-15). A surprising range of dominant periods was observed amongst the spectral clusters. However, clusters with dominant periods T_{max} at 10-12.5 bp were the largest in accordance with the genome-wide period.

Since curved DNA and phased AT-tracts downstream of the promoter were shown to affect transcription by plectonemic DNA looping [58, 149, 156], the spectral CDS clusters were compared with two expression data sets. The first experiment by Prakash *et al.* quantified supercoiling-sensitive transcriptional regulation in *Synechocystis* sp. PCC 6803 [171]. In addition, the clustering of the diurnal expression dataset presented in Chapter 2 was consulted.

Due to the focus on ≈ 11 bp periodicity, CDS clusters with a wide range of dominant periods T_{max} were divided into four groups similarly to previous analyses. Only two statistically weak overlaps were detected. Amongst the supercoiling-sensitive gene groups from Prakash *et al.* (Fig. 3.4 B), supercoiling unresponsive genes were enriched in aperiodic genes (not in CDS clustering) ($p < 0.012$), and supercoiling-inhibited genes were found slightly enriched amongst genes with $T_{max} \leq 10$ bp ($p < 0.017$).

On the other hand, no general correlation was found between spectral CDS clusters and clusters of diurnally co-transcribed genes (Fig. 3.4 D). Two exceptions were expression clusters 5 and 11 (genes not on the microarray). Both clusters contained several transposons and unknown proteins (Appendix Fig. 11 B and 13 A) and featured unusually

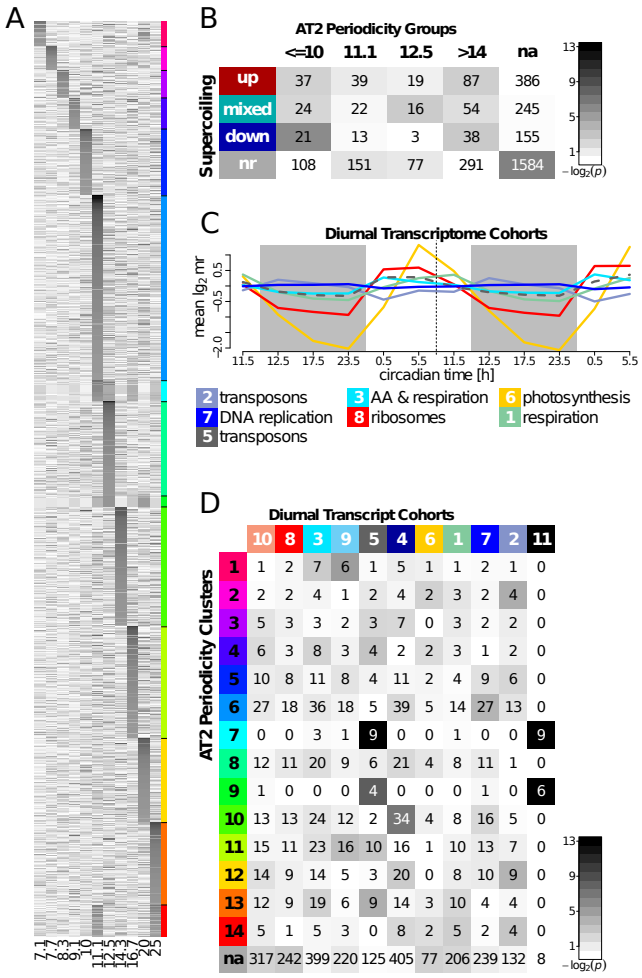


Fig. 3.4: **Periodicity in *Synechocystis*PCC 6803 coding regions.** (**A**) AT2 periodicity spectrum ($Q^*_{AT2}(T,i)$) clustering of the 1000 strongest periodic CDS of *Synechocystis* sp. PCC 6803 (rows - Genes, columns - Period T [bp], black indicates higher spectral power). Each cluster is dominated by one main spectral component, cluster membership is colour-indicated on the right (1-14 from top to bottom). (**B**) Overlap of the spectral CDS cluster groups with supercoiling-sensitive gene groups. The CDS clusters from A are condensed into four groups with similar dominant period T_{max} (columns, see Appendix Fig. 10 A). These groups are tested for overlap with gene groups found to be "up"-regulated, "down"-regulated, and genes that showed a "mixed" or no response ("nr") to changes in DNA supercoiling (rows, from [171]). The number of shared genes is shown, the significance p-value obtained from a hypergeometric test is colour-coded (legend on the right). (**C**) Mean transcript abundance time-series of diurnally co-transcribed cohorts. Only clusters with typical features of supercoiling-sensitivity are shown (function, strong bias in GC-content, Appendix B Fig. 12 and 13). (**D**) Overlap between clusters of co-transcribed genes (columns) in *Synechocystis* (Section 2.3, Appendix Fig. 12) with CDS periodicity clusters from A (rows). Both clusterings are provided in Appendix B File 2.

high AT contents (Appendix Fig. 12 B). Elevated AT-content is typical of genes of foreign origin [58, 149]. Cluster 5 exhibited diurnal transcriptional oscillations of low amplitude in-phase with the supercoiling-induced and GC-rich growth genes (clusters 3 and 8, Fig. 3.4 C). Striking members of this expression cluster 5 were the distinct $5'$ and $3'$ parts of the *ISY100* transposase, which caused the "Transposase" annotation enrichment in the overlapping spectral clusters 7 and 9 (T_{max} 11.1 and 12.5 bp, see also Appendix Tables 14 and 15). Both spectral clusters were comparably small and featured highly similar spectra. Accordingly, previous work suggested that arrangement of curved DNA into plectonemes is important for transpososome formation [29, 64] and transposon silencing [150]. Specifically, the 3' sections reflected the reported curvature of the *Mos1* transposon in *Drosophila mauritiana*[29] which also belongs to the *Tc1 / mariner / IS630* transposon superfamily. However, the spectral clustering as well as the diurnal expression were diverse across the entirety of the 120 known transposons in *Synechocystis* sp. PCC 6803 (Appendix Fig. 13). This might be due to the dependency of transposon expression upon its (randomly) acquired genomic context, rather than transcriptionally regulating AT2 periodicity.
Interestingly, significant overlaps were found between the diurnal expression clusters and supercoiling-sensitive gene groups (Fig. 3.4 C and Appendix Fig. 12 A). The typical bias in GC-content, previously reported for supercoiling-sensitive genes in *E. coli*[164] and *Synechococcus elongatus* PCC 7942 [237], was also found in *Synechocystis* sp. PCC 6803 (Appendix Fig. 12 A). GC-rich genes encoding amino-acid synthesis and ribosomes (diurnal expression clusters 3 and 8) were maximally expressed in the morning and were induced by negative DNA supercoiling, while night-expressed genes were supercoiling-repressed (cluster 2) or AT-rich (cluster 7).
Nucleoid associated proteins, such as HU, were found to wrap DNA in solenoids [131] and preferentially bind to AT-rich sequence. As outlined in the introduction (see Fig. 1.1), a signal similar to the nucleosome positioning code could exist for NAPs in order to stabilise the diurnal structural changes. Compellingly, HU expression profiles exhibited significant diurnal oscillations in all datasets, except *Anabaena*, with phases around dusk (see Appendix A, Fig. 6). In the *Anabaena* dataset, irregular but visible oscillations in the first replicate prevent the classification as significantly oscillating. Interestingly, the two HU homologs in both *Cyanothece*datasets exhibit highly significant circadian out of phase oscillations, one at dawn and one at dusk.

3.5 Discussion and conclusions

3.5.1 AT2 periodicity and transcriptional regulation

The systematic study revealed that combinations of W (A or T) dinucleotides induce the strongest genome-wide 10-12 bp periodicity in 74% of the 58 tested genomes, while the minimal AT-tract motif AT2 was the strongest in 60% of the genomes. Only a small group of picoyanobacteria and the symbiotically living cyanobacterium *Candidatus Atelocyanobacterium thalassa* were characterised by absence of genome-wide 10-12 bp periodicity. Analysing the localisation of periodicity indicated CDS as major source

in cyanobacteria. It was reported previously [37, 73] that the periodic dinucleotides are mainly encoded in codon positions III. Application of various permutation methods showed, that position I also contributes to this signal. The contribution of position II was the weakest but not absent. The protein's amino acid order and the strain-specific codon usage bias were partially adapted to encode this putative DNA structural information. These observations are consistent with the idea that phased AT-tracts constitute a structural code which can induce nucleoid condensation in a pre-arranged manner, as proposed for *E. coli* by Tolstorukov *et al.* [217]. In this context, theoretical [193] and experimental studies of negatively supercoiled DNA [112, 162, 226] reported DNA curvature in the apical loops of plectonemes due to phased AT-tracts (Fig. 3.1 A right).

Periodicity and transcriptional regulation Periodic signals as well as DNA curvature have been reported in a range of species and studies. It was found that the binding of RNA polymerase to a plectronemic *E. coli* promoter region leads to changes in the DNA structure, placing the binding site in the apical loop [69]. Furthermore, the binding motif for RNA polymerase in *Campylobacter jejuni* and other species was characterised as strongly curved AT-tract containing sequence [165]. Detailed studies of promoter periodicity in *E. coli* were able to correlate differential promoter periodicity signals with the supercoiling-sensitive expression of the corresponding genes [106, 151]. The analysis of 170 prokaryotic species also indicated a role of DNA curvature in transcriptional termination [104]. While the cyanobacterial intergenic regions featured significant enrichments of 200 bp windows of pronounced AT2 periodicity, the major contribution of the genome-wide ≈ 11 bp signal was located in coding sequence. This can be partly explained by the shorter cyanobacterial promoter lengths compared to eucaryotes. However, gene expression can also be modulated by DNA curvature downstream of the promoter region, as found for the H-NS mediated expression silencing mechanism targeting foreign genes [58, 150, 156]. Notably, the presented analysis does not reveal strong correlations between the AT2 periodicity in coding regions and supercoiling sensitive or diurnal transcription in *Synechocystis* sp. PCC 6803 (Fig. 3.4, Appendix Fig. 11 and 12). Several factors might reduce the power of this analysis. Firstly, transcriptional regulation may involve long-range interactions implemented by more extensive chromosomal domain architecture changes [163, 203]. Secondly, operons complicate the association between transcriptional pattern and periodicity. However, *Synechocystis* sp. PCC 6803 belongs to a cyanobacterial clade with smaller and less densely packed operons [133]. Thirdly, the expression dataset of *Synechocystis* sp. PCC 6803 is likely to reflect extensive light-driven diurnal rather than clock-driven expression (see Chapter 2). A similar periodicity to circadian expression correlation analysis using *Synechococcus elongatus* data could extend the presented insights. RNAseq-based transcriptome data under free-running conditions may refine this result in future studies.

Since no strong correlation between the AT2 periodicity and diurnal transcription was found, the involvement of the AT-rich sequence binding nucleoid associated protein HU in the transcriptional regulation could not be tested. However, the consistent dusk-peaking

diurnal expression oscillations, observed in the microarray data collection presented in the first part, imply functional relevance. Particularly interesting are the two HU homologs of *Cyanothece*, one of which peaks at dawn and the other one at dusk. This predestines *Cyanothece*for future studies of the function of HU in cyanobacteria.

Periodicity mosaic - multiple DNA-structural codes The localised periodicity analysis revealed additional regions with AT2 periods ≈ 10 bp in highly periodic species with genome-wide ≈ 11 period, and enrichment in intergenic regions of both ≈ 10 and ≈ 11 bp period ranges in weakly periodic genomes (Fig. 3.3 A and Appendix Fig. 8 K). These observations are consistent with previous reports [106, 139, 151, 217], and might be attributed to multiple DNA structural features encoded in bacterial genomes, *e.g.* plectonemes (> 10.5 bp) and solenoids (< 10.5 bp) [220]. Here AT-tracts are the dominant source of sequence periodicity, indicating functional relevance. This is supported by the prominent role of AT-tracts in DNA curvature formation [183]. However, other dinucleotides were also shown to achieve notable DNA curvature [24]. In general, YR stretches (alternating pyrimidine/purine) support transition to Z-DNA [233]. Interestingly, very general lifestyle features, such as pathogenicity, growth temperature, and oxygen requirement were found to correlate with genomic features like curved DNA and Z-DNA forming patterns [23, 76]. It can thus be expected that the high diversity of lifestyles in the cyanobacterial clade, covering marine, freshwater, thermophilic, soil and even desert habitats, is reflected in sequence periodicity and DNA structural patterns. Efforts to systematically study sequence pattern abundances have only recently started [76] and more work on the relative genomic localisation is necessary to unravel the different DNA structural codes.

3.5.2 A role for AT2 periodicity in transposon function

The *ISY100* transposon, member of the *Tc1 / mariner / IS630* superfamily, was found to feature strong AT2 ≈ 11 bp periodicity. Importantly, it is the only active transposon in *Synechocystis* sp. PCC 6803 [232], and is able to function without additional co-factors. However, its efficiency is modulated by the degree of negative supercoiling in its host DNA[51], similar to several other transposons [36, 95, 200]. It has been suggested previously that *Tc1 / mariner* transposons feature intrinsically curved DNA sequence to aid the transposase in simultaneously binding both inverted terminal repeat (ITR) ends [29, 64]. It can thus be speculated that the phased AT-tracts in the transposon sequence induce a plectronemic structure at a specific location [193], so that ITR sequences are placed in spatial proximity in preparation for the transposase binding as proposed for eukaryotic transposons [36]. Interestingly, *ISY100* copies also feature a diurnal transcription profile (Fig. 3.4 C and Appendix Fig. 11). This might indicate synchronisation between the transposon lifecycle to the host's physiological rhythm with negative DNA supercoiling as a coupling mechanism. Another interesting observation was made in *Arabidopsis thaliana* which features sequence periodicity in centromere-proximal transposons [140]. Similar periodicity in strong nucleosome binding sites, also located near the centromeres [186], allows for the possibility that periodicity was originally only

involved in one mechanism, *e.g.* transposons, but was subsequently reutilised.

3.5.3 AT2 periodicity for DNA packaging in cyanobacteria

Comparison of the genome-wide AT2 $\approx$ 11 bp periodicity revealed groups with significantly varying signal strengths. An analysis of the corresponding cyanobacterial phylogeny showed high periodicity particularly in the SPM clade (morphological section I, unicellular, Fig. 3.2 and Appendix B Fig. 7). Intermediate signal strength denoted the Hcyst clade (section IV, heterocyst-forming), and weak periodicity was found in the Syn/Pro clade (section I, picocyanobacteria). While both enterobacteriaceaeal reference species *E. coli* and *Dickeya dadantii* use supercoiling for transcriptional regulation, and a relationship between promoter periodicity and transcription has been described in the former, only relatively weak genome-wide periodicity was found here. This pattern, together with the weak connection found between periodicity and the transcriptional regulation of individual genes, points to a functional interpretation. The correlation between genome-wide AT2 periodicity and the cell division mode, as well as events in which the loss or gain of high periodicity is accompanied by lifestyle transitions, indicate a function in DNA packaging.

Strains with high AT2 periodicity Several cyanobacterial species contain multiple complete chromosomal copies, a feature which sets them apart from more traditionally studied bacteria, *e.g. E. coli*, which typically contains only one or two copies [197]. In particular, several members of the SPM/Hcyst sister clades as well as *Synechococcus* feature more than one chromosome copy (oligoploidy) up to several hundred (polyploid). Interestingly, the ploidy level was found to vary nearly fourfold between exponential growth and the stationary phase in *Synechococcus* [60], rather than being coupled to cell division. *Synechocystis* sp. PCC 6803 has shown significant variance in DNA content between daughter cells [192], pointing at unequal chromosome partitioning upon cell division. Nucleoid separation takes place only very late in the cell cycle, immediately before the completion of cell septum formation, suggesting the constriction as partitioning mechanism. It was thus proposed that chromosome segregation is random and passive without the help of a dedicated partitioning machinery, similar to *E. coli* or *Bacillus subtilis*. The genome replication of *Synechococcus elongatus* PCC 7942 is similarly independent from the cell division, but chromosome segregation was found to be less random [31, 244], with chromosomal copy alignment along the long axis of the cell prior to septum constriction [31, 85]. Monoploid model bacteria on the other hand, such as *E. coli* or *Bacillus subtilis*, exhibit a coupling between DNA replication, chromosome segregation, and cell division [41, 255].

Strains with low AT2 periodicity More alike to *E. coli* and *Bacillus subtilis*, picocyanobacteria from the *Prochlorococcus* and *Synechococcus* genera showed a strict coupling of S-phase to cell division and stable ploidy [19, 161, 234]. Lacking the cellular volume and resources for high numbers of chromosomal copies, picocyanobacteria might have been forced to evolve more precise chromosomal replication and cell division mechanisms necessary for a monoploid lifestyle.

Interestingly, three strains in the highly periodic SPM clade were found to exhibit anomalously low periodicity, and also deviated in their morphological section assignment. Two strains are assigned to section II due to cell division via multiple fission, in which the vegetative cell is divided into at least four to over 100 spherical and small daughter cells called baeocytes [245]. While the exact mechanism of this multiple fission is still unclear, it is clearly non-random as it involves a combination of multiple chromosomal replication with the ordered partitioning between the daughter cells [167]. The third genome belongs to the unclassified strain *Candidatus Atelocyanobacterium thalassa*. While it is closely related to highly periodic *Cyanothece* strains, it has undergone genomic reduction due to its symbiotic lifestyle with an unicellular alga [214] which has might have impacted its cell division mechanisms.
While homologs of the histon-like HU protein and the SMC chromosome condensation complex can be found in cyanobacteria, many other nucleoid organisation and segregation proteins known from eukaryotes are absent [192, 219]. Furthermore, only very weak association of AT2 periodicity and supercoiling sensitive expression was found to support the hypothesis of sequence-determined NAP binding to change DNA conformation and thereby transcription. Interestingly, the expression profiles of all cyanobacterial HU homologs indicated diurnal rhythmicity suggesting functional relevance. The experimental determination of genome-wide binding sites of HU via ChIP-seq would be valuable to hint at possible target genes or its functional principle.
It can be speculated, that sequence periodicity constitutes a structural code which aids chromosomal packaging in highly polyploid cyanobacterial strains under fast growing conditions. Such high growth rates were found to be marked by elevated negative DNA supercoiling levels [164, 203], which would lead to extensive plectroneme formation and thus more efficient nucleoid compaction, allowing for higher numbers of chromosome copies. According to this interpretation, the schema in Figure 1.1 depicts the transition from slow growth (top) to fast growth (bottom) and back. In contrast, no compelling evidence was found for the utilisation of localised periodicity-aided supercoiling for transcriptional regulation of individual genes, as described by the hypothetical metabolic circadian oscillator.

4 Circadian and ultradian transcriptional rhythms in *Neurospora crassa*

This chapter presents a combined time-resolved dataset of RNA polymerase II binding and mRNA abundance, both measured every 2 h over a total of 22 h in constant darkness. The dataset is integrated with genomic binding locations for the clock TFs WCC and CSP1, both measured via ChIP-seq. Observed target gene phases are compared with theoretical predictions, with focus on genes with combined input from both TFs. Since frequency doubling is predicted for combined targets of circadian transcription factors, ultradian transcriptional rhythms are analysed in more detail. The presented work represents the collaborative efforts with Bharath Ananthasubramaniam, Gencer Sancar, Michael Brunner, and Hanspeter Herzel. The experimental work and the assembly of the RNAPII and mRNA abundance profiles have been conducted in the group of Michael Brunner. The bioinformatic analysis was performed in collaboration with Bharath Ananthasubramaniam. The dataset is published as Sancar et al. *in BMC Biology [187].*

4.1 Two major transcription factors WCC and CSP1 involved in the circadian clock of *Neurospora crassa*

4.1.1 The *Neurospora crassa* core clock

The filamentous fungus *Neurospora crassa* is a well established model organism for the molecular mechanism of the circadian clock. Its molecular core clock consists of a Transcription-Translation negative feedback loop oscillator (TTO) similar to the mammalian clock (see Section 1.1). It was found that the circadian clock can control the expression of between 10%, the assumed extent of circadian expression in mammals [79], up to 40% of the *Neurospora crassa* transcriptome, which is rather typical for cyanobacteria (see Chapter 2). The primary output of the circadian core clock in *Neurospora crassa* is facilitated by the transcription factor *White Collar Complex* (WCC) which binds to the promoters of a subset of clock controlled genes (ccg). The positive element WCC, a heterodimer of the two GATA family proteins White Collar 1 and 2 (WC-1, WC-2), induces the expression of its target gene *Frequency* (*frq*). *frq*, like most of its targets, is maximally expressed around dawn. After dimerisation with the Frequency Interaction RNA Helicase (FRH), the FRQ/FRH complex is progressively phosphorylated and, when completely phosphorylated, closes the negative feedback loop on WCC (see Fig. 4.1). Interestingly, WCC, and specifically WC-1, functions also as the major blue

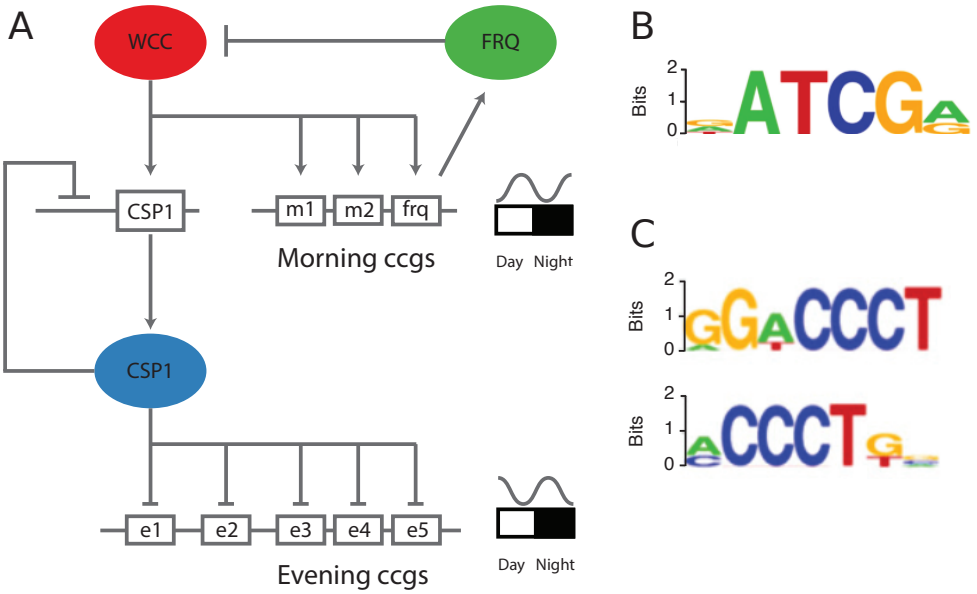


Fig. 4.1: **The circadian core clock of *Neurospora crassa*.** (**A**) Hierarchical organisation of clock controlled genes (ccg). The WCC is active in the subjective morning directly activating morning-specific ccgs (m1, m2, *frq*). The transcription repressor *csp1* is amongst these ccgs. Newly synthesised CSP1 is rapidly inactivated by progressive phosphorylation and degradation (deg). The target genes of CSP1 (e1-5) are repressed in the subjective morning. Accordingly, CSP1 regulated second tier ccgs display evening-specific expression rhythms. CSP1 inhibits its own gene in a negative feedback loop. DNA binding motifs of the core clock transcription factors White-Collar-Complex (**B**) and CSP1 (**C**) derived from ChIP-seq experiments [188, 201].

light receptor, thereby directly allowing for light entrainment of the circadian clock.

4.1.2 Major transcription factors WCC and CSP1 regulate morning- and evening-specific expressed genes

There are about 24 transcription factors amongst the WCC regulated ccgs [201], potentially relaying the circadian clock signal towards their target genes. The addition of this intermediate step could introduce various delays, allowing for different phases in the respective target genes. The most prominent example is the second tier transcriptional repressor conidial separation-1 (CSP1), which is induced by WCC. The morning-specific expression of this transcriptional repressor was found to result in evening-peaking expression in its target genes [188]. Similar to FRQ, the stability of CSP1 is regulated by successive phosphorylation steps which ultimately lead to its degradation, also referred to as phospho-timer.

4.1.3 Combination of regulatory input permits phase- and period-variation

While it has been experimentally shown that WCC is responsible for the activation of morning-genes [201] and CSP1 for evening-genes [188], it is not clear which mechanism is used to achieve the observed intermediate phases. As described by Korenčič *et al.*[103], the combination of circadian regulatory signals is sufficient to achieve phase changes and even frequency doubling in the respective target gene, depending on the phase relationship, amplitude, and regulatory strength of the input. The work by Westermark and Herzel elaborates further on this model and presents additional biological evidence [249]. The importance of the described second tier TFs in *Neurospora crassa*, particularly the transcriptional repressor CSP1, predispose *Neurospora crassa* to test predictions of this model towards the generation of intermediate phases and ultradian expression patterns.

4.1.4 Genome-wide assessment of circadian expression

Various methods have been employed to monitor diurnal transcriptional activity, microarrays being the most prominent. Previous microarray-based circadian expression analyses in *Neurospora crassa* by Nowrousian *et al.* [152] and Correa *et al.* [38] probed only a fraction of genes. Accordingly, Nowrousian *et al.* reported 27 clock-controlled genes out of a total of $\approx$ 10.000 protein coding genes ($\approx$ 1000 genes measured), whereas Correa *et al.* found 145 out of 1343 measured genes. The work of Chen *et al.* presents time resolved expression measurements up to 240 min after the transition from dark to light (DL) condition [32]. The first RNA-seq based study of the circadian *Neurospora crassa* transcriptome by Hurley *et al.* suggests a significantly higher fraction of circadian regulated genes (10-40%) [80]. Functional enrichment analysis of dawn- and dusk-peaking genes indicates a temporal separation between catabolic processes over the day and anabolic processes during the night. Interestingly, their luciferase reporter-based promoter analysis of 296 genes with circadian oscillating mRNA abundance levels yields constant promoter activity in 66 genes. The this result was interpreted as indication for the major role of post-transcriptional processes in circadian regulation.
A similar conclusion was drawn in an earlier work comparing circadian RNAPII binding and mRNA abundance [114], and abundances of nascent RNA with mRNA [126, 135] in mouse liver cells. It was found that diurnal transcription is achieved mainly via rhythmic RNAPII binding, not by the concerted RNAPII release from the promotor. However, many genes exhibited constant RNAPII binding profiles with oscillating mRNA levels or significant phase delays between both peaks, indicating post-transcriptional regulation. Furthermore, some of the genes exhibited oscillating RNAPII binding but constant RNA abundance profiles. This was interpreted as indication of long mRNA lifetimes which can dampen and effectively avoid the oscillations in mRNA synthesis.
RNA-Seq has various advantages compared to the standard microarray technology. The background signal, caused by falsely mapped reads which match multiple locations in the genome, is generally low. Furthermore, the upper limit for transcript quantification

is given by the sequencing dept which can be chosen freely. Accordingly, the achievable dynamic range in RNA-Seq is much greater (~9000-fold [143]) compared to 100-fold for micro arrays [242]. However, this method comes with its own complications. Depending on whether cDNA or RNA fragmentation was used during library generation, a bias towards 3' end sequences or a depletion of sequences at the transcript ends is found. Furthermore, alternative splicing complicates the process of read mapping. However, the fraction of alternatively spliced genes in fungi appears to be significantly lower with ~4% compared to estimates of >40% in humans [55].

4.1.5 Outline of this chapter

In the following chapter, genome-wide simultaneous measurements of RNA polymerase II (RNAPII) binding and the resulting mRNA abundance are compared to assess the extent of circadian expression in *Neurospora crassa*. Genome-wide binding studies of the core clock TFs WCC and CSP1 are used to predict corresponding target genes. Phase distributions for these target gene sets are evaluated with focus on genes with combined WCC and CSP1 input. The observed phase distributions are compared with theoretical predictions. Mathematical modelling predicts frequency doubling, *i.e.* ultradian rhythmicity, for genes with combined input from two TFs with according phase relationship. Genes with ultradian transcription pattern in the presented dataset are examined with respect to input from WCC or CSP1, and other TFs via discriminative motif overrepresentation analysis.

4.2 Materials and methods

The methods used in this chapter are described in the following, while supplementary analyses are provided in Appendix C.

***Neurospora crassa* Strains and culture conditions** The *Neurospora crassa* wt strain (FGSC#2489) was acquired from FGSC. As standard growth medium, 1× Vogel's medium supplemented with 2 % glucose, 0.5 % L-arginine, and 10 ng / ml biotin was used. The cultures were entrained in 11h / 11h LD cycles and then released at CT 12 into constant dark for 22 h. Cultures were transferred in a staggered manner to allow culture harvesting within 10 h.

RNA analysis RNA was extracted with peqGOLD TriFAST (peqLab, Erlangen, Germany) according to the manufacturer's protocol. After dissolving RNA in 70μl nuclease free water with 80u Ribolock RNAse inhibitor (ThermoScientific), cDNA was prepared using the Maxima First Strand cDNA Synthesis Kit (ThermoScientific). Transcript levels were analysed by quantitative real-time PCR in 96-well plates with the StepOnePlus Real-Time PCR System (Applied Biosystems) using TaqMan Gene Expression Master

Mix (Applied Biosystems). Primers and probes are listed in [187], Table S11. rRNA was used for normalisation.

Chromatin Immunoprecipitation (ChIP) ChIP was performed by using an antibody specific for the Serine-2 phosphorylated C-terminal tail of RNAPII as described previously [201]. Polyclonal anti-rabbit RNAPII Ser2-P was raised against the peptide (pS)PTSPSY(pS)PTSPSC.

RNA sequencing and ChIP sequencing cDNA was prepared by using NEBNext® Ultra RNA Prep kit with NEBNext®Multiplex oligos according to the manufacturer's instructions. ChIP DNA libraries were prepared with NEBNext®ChIP-Seq Library Prep Reagent Set for Illumina®with NEBNext®Multiplex oligos. A 2100 Bioanalyzer was used to check the size and the quality of the libraries. Un-paired sequencing with 50 bp reads was performed with a HiSeq 2000 at GeneCore EMBL Heidelberg for RNA-seq and by the BGI, Hong Kong for ChIP-seq. Individual sequence reads for each run are available in SRA database under the study name PRJNA248256. Accession numbers for experiments and number of sequence reads are listed in [187], Table S12.

High-throughput data analysis The raw sequence reads were mapped to the *Neurospora crassa* genome (NC10) using Bowtie [111], where parameters were set to allow a maximum of 3 mismatches and suppress alignments which mapped to more than one location. Gene expression was quantified by the number of reads falling into the annotated exons. For analysis of the Ser2P-RNAPII ChIP-seq data, the reads falling into a 500 bp window upstream of each gene until its end position were integrated. Normalisation was carried out using the size factor as described by Anders *et al.* [5].

Oscillation detection 1499 Genes with median RNA-seq read counts <16 were considered insufficiently high expressed and were discarded from further analysis, ten of which also exhibited median ChIP-seq counts below this threshold (see Appendix C Fig. 16). The profiles were log2 transformed prior to harmonic regression to detect oscillating profiles. When analysing microarray data, expression values are commonly log2 transformed to achieve approximately normal distributed data. However, it is not obvious whether the underlying assumption of harmonic regression, *i.e.* sinusoidal shape of the expression profiles, should be applied in normal or log-space. A comparison of the detection results revealed extensive accordance. log_2 mean ratio transformation (Appendix A section 1.2) yielded a slightly higher number of oscillating genes and was therefore used for further analyses. Detrending prior to harmonic regression is advisable since liquid cultures were used, which can lead to correlations between successive samples due to culture conditions [54]. In this case, linear detrending introduces a bias into phase and amplitudes obtained from subsequent harmonic regression. As demonstrated in Appendix C 4.2, amplitude values of profiles with phases other than about 0 CT and 12 CT are suppressed in datasets which cover only one period. The procedure employed here is similar to that describe in Appendix A 2.1 with exception of the p-value derivation.

Due to the extensive low amplitude oscillations observed in the RNAPII and mRNA abundance data (see Fig. 4.4), an empirical p-value was employed. Instead of using the F-distribution as background model to assign a p-value to each genes F statistic, the F statistic was repeatedly computed for randomly permuted versions of the individual profile. The oscillatory p-value was then approximated as probability to observe the original profiles F statistic amongst those of the permuted background. A number of genes are strongly induced by light. They still lead to high expression values in the first sample of the time series followed by a sharp decrease in the second sample, *e.g. frq* (see Fig. 4.3). As this drastic decay hampers the oscillation detection via harmonic regression, the procedure was repeated while omitting this first sample. The results were combined preferring the lower p-value. The oscillation detection procedure was repeated separately with the period parameter 22 h and 11 h, to assess ultradian oscillations. For the following analyses, a cutoff of $p < 0.025$ was chosen. The estimated circadian and ultradian amplitude is independent of the level of RNAPII occupancy and mRNA abundance (Appendix C Fig. 20), indicating sufficient sequencing depth to resolve the majority of oscillations.

Functional enrichment analysis Gene enrichment analysis for Gene ontology terms was performed using the TopGO R package, specifically the "elim" algorithm with Fisher statistic. Current GO annotations were obtained from the Broad Institute (`www.broadinstitute.org/ annotation/genome/neurospora/MultiHome.htm`). Enrichment of functional categories (FunCat) were performed via FungiFun [173]. In both cases, all genes probed in the presented dataset were provided as background set.

ChIP-seq target gene prediction The White Collar Complex (WCC) has been shown to transcriptionally regulate a large number of target genes [201]. Of the 456 reported WCC binding sites found via ChIP-seq ([201], Suppl. Table 1), 307 sites are located in the promoter region of a known protein-coding gene, 86 in intergenic regions, and 64 in the coding region of genes (one peak is in the promoter of one gene and intergenic for another). 48.7% of the 40.5 Mb genome assembly NC10 of *Neurospora crassa* is occupied by the coding sequence of the 9732 genes, which are on average 2025 bp long. Each genes contains in average 1.7 introns with an average intron size of 134 nucleotides, *i.e.* $\approx 3\%$ of the genome [55]. Approximately 10% of the genome are taken up by repetitive sequence. Assuming a uniform peak distribution across the entire genome as background model would yield an estimated 208 peaks in the coding region. Comparison with the observed 64 coding sequence peaks indicates significant 3.25 fold underrepresentation and thus specific peak localisation in intergenic and promoter sequences. 351 genes are found with WCC binding sites in their promoter, 145 in their intergenic region, and 60 in their coding region, yielding 352 net predicted target genes. However, this prediction does not include established WCC target genes such as the light-responsive GATA family transcriptional activator *sub-1* (*NCU01154*, essential for late light induced genes) [32]. The transcriptional repressor CSP1, one of WCCs target genes, has been shown to cause circadian expression anti-phasic to WCC [188].

It was furthermore shown that CSP1 forms a transient complex with its co-repressor RCM1/RCO1 and ubiquitin ligase UBR1. ChIP-seq experiments of DD cultures, 30 min after a short light pulse, yielded 695 CSP1 binding peaks (>4-fold enrichment), including 185 peaks with a RCO1 peak in its vicinity, as well as 323 individual RCO1 peaks (>2.5-fold enrichment). In their analysis, the authors consider all genes located next to a peak (up- and downstream) as potential target gene, yielding 1310 genes for CSP1 alone or with RCO1 as well as 618 RCO1 target genes.
Here, the association of genomic peak location and target gene (genome build NC10 for the CSP1 dataset, NC7 for the WCC dataset) was predicted by the transcription factor association score (TFAS) [155]. It models the association strength as exponentially decaying with growing distance between gene and peak while integrating scores over multiple peaks for each TF and gene. Specifically, the association $a_{i,j}$ of TF j on gene i is defined as sum over all peaks of TF j as:

$$a_{i,j} = \sum_k g_k e^{-d_k/d_0} \quad ,$$

where g_k is the number of aligned reads for binding peak k of the TF j. d_k is the distance in bp between the TSS of gene i and binding peak k in the reference genome, and d_0 is a constant scaling factor to determine the decay of the influence of a peak with growing distance to the TSS. Here, the scaling factor was set to $d_0 = 5000$ whereas the score threshold was set to 200. The TFAS implementation of the R package *TFTargetCaller* was used for the calculation. This procedure yielded 1590 predicted WCC targets, 374 targets bound by CSP1 and its co-repressor RCO1, 1250 targets bound only by CSP1, and 826 RCO1 targets. This procedure predicted the known association between WCC and *csp-1*, *sub-1*, *vvd*, *cry*, and *bli-4*. Similarly, known CSP1 targets were found, such as *stl1*, *sty1*, *gld1*, *NCU07161*, and *NCU01924*. Only the TFAS based prediction included the negative feedback loop of CSP1 on itself [188]. The published target gene predictions, as well as the one presented here, do not include all literature knowledge. The grainy-head gene *ghh* (*NCU06095*) is involved with the circadian clock [159], but its CSP1/RCO1 binding sites are not located at the beginning of the gene. The same holds true for the CSP1 target genes, *NCU07161* and *NCU01924*, and the WCC targets *al-3* and *fl*. Accordingly, the observed association scores are low and fall below the threshold. Another interesting case is the gene osmo-regulation gene *os-2* (*NCU07024*) which has been shown to be an output pathway of the circadian clock [238]. The total amount of OS-2 and the amount of phosphorylated OS-2 was observed to oscillate with a peak phase of CT 0. In absence of the circadian clock, *os-2* is regulated by the response regulator RRG-1. In the presented target prediction, *os-2* was found to exhibit strong CSP1 and RCO1 association and only slightly below threshold association with WCC, indicating a complex regulatory mechanism.

Motif overrepresentation A differential motif overrepresentation was performed to test whether potential transcription factor binding site motifs occur in the promoters of

genes with 11 h ultradian expression profiles. 1 kb wide DNA sequence windows symmetrically around the transcriptional start site were excised according to the *Neurospora crassa* NC12 genome annotation (`www.broadinstitute.org/annotation/genome/neurospora/MultiHome.htm`). The set of 357 11h ultradian transcribed genes was separated into four groups according to the dominant phase, group I ($0 \geq \phi \geq 3.7$, n = 48), II ($3.7 \geq \phi \geq 5.7$, n = 111), III ($5.7 \geq \phi \geq 7.3$, n = 36), and IV ($7.3 \geq \phi \geq 10.9$, n = 165). The general background set was defined as all genes with oscillatory p-value > 0.2 in RNAPII and mRNA profiles with 22 and 11 h period and an estimated amplitude below 0.06, resulting in 423 genes. For the analysis of each foreground group, a matching number of background sequences was randomly sampled from the general background set. The GC content vs. the CpG content scatterplot revealed one homogenous group, *i.e.* no CpG islands as observed in vertebrates. Therefore, a distinction of high-GC and low-GC promoters as recommended by Roider *et al.* [184] is not necessary. The DECOD tool was applied for the differential motif overrepresentation analysis between foreground and sampled background set [77]. The six most overrepresented motifs were reported for each group.

4.3 Pervasive circadian rhythms in transcription and mRNA abundance

The extensive circadian rhythmicity in RNAPII and mRNA abundance profiles is descriptively analysed. Genes are grouped by oscillatory pattern and phase for successive analyses. Functional annotation enrichment analysis reveals a coarse temporal separation between cellular functions. The transcriptional patterns of prominent genes are compared to literature knowledge with a good agreement of the presented data with previous experiments. A mathematical framework is used to derive a prediction for target genes with combined regulatory input from WCC and CSP1.

Liquid cultures of the fungus *Neurospora crassa* were entrained to 11h/11h LD cycles and then released to constant dark. mRNA abundance was measured via RNA-seq and in parallel RNA polymerase II (RNAPII) occupancy was measured via Ser2P-RNAPII ChIPseq. Samples were taken every two hours, starting from CT 12 until CT 12 of the following subjective day. At the beginning of the experiment, a light flash was given to synchronise the culture. For RNAPII occupancy, the ChIP-seq signal was integrated within a window from 500 bp upstream of the gene start to the end. 1499 genes out of the available total 9732 genes were discarded due to low detected mRNA expression (see Appendix C 4.1). Identical statistical analyses and thresholds were used to determine rhythmicity in both datasets (see Materials and Methods 4.2).

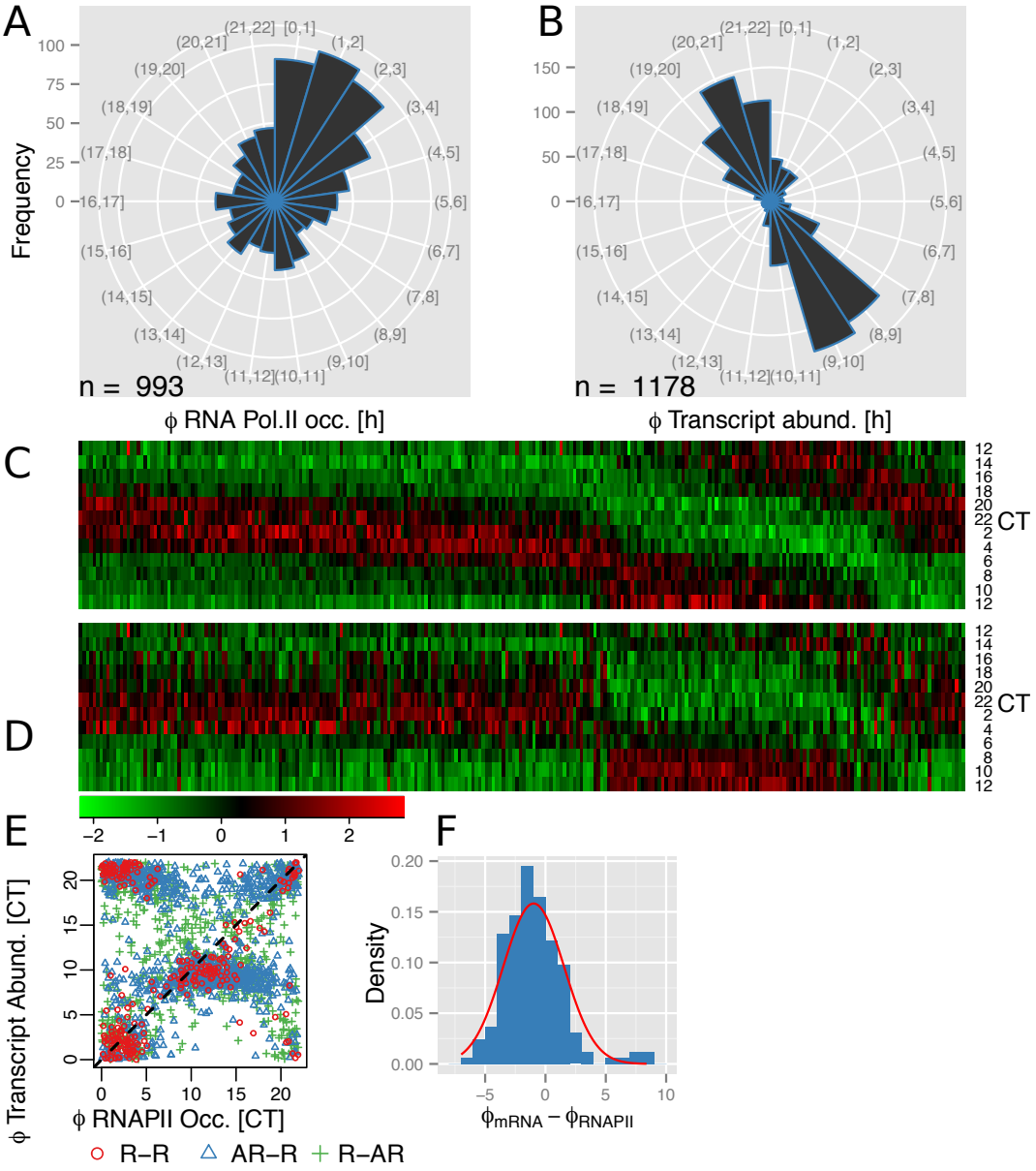


Fig. 4.2: **Comparison of oscillatory parameters of *Neurospora crassa* RNAPII binding and transcript abundance.** (**A**) Phase histogram for genes with significantly oscillating RNAPII profiles ($p < 0.025$). (**B**) Similar to (A) for significantly oscillating mRNA abundance. (**C**) RNAPII occupancy profiles for R-R genes, and (**D**) the corresponding mRNA abundance profiles for R-R genes (n = 262, similar ordering and colour mapping). The gene profiles are z-transformed and ordered by phase ϕ, green marks low and red high values. (**E**) Oscillation phase ϕ (CT) of RNA polymerase II binding (x-axis) versus transcript abundance (y-axis). Transcripts with significantly oscillating ($p < 0.025$) RNAPII and mRNA profiles (R-R set) are marked with red circles, only significant RNAPII profiles (R-AR set, n = 682) with green '+', and significant mRNA profiles (AR-R set, n = 916) with blue triangles. The consistently arhythmic AR-AR set (n = 6372) is omitted for clarity. (**F**) Distribution of phase differences between RNAPII and mRNA abundance for the R-R gene set.

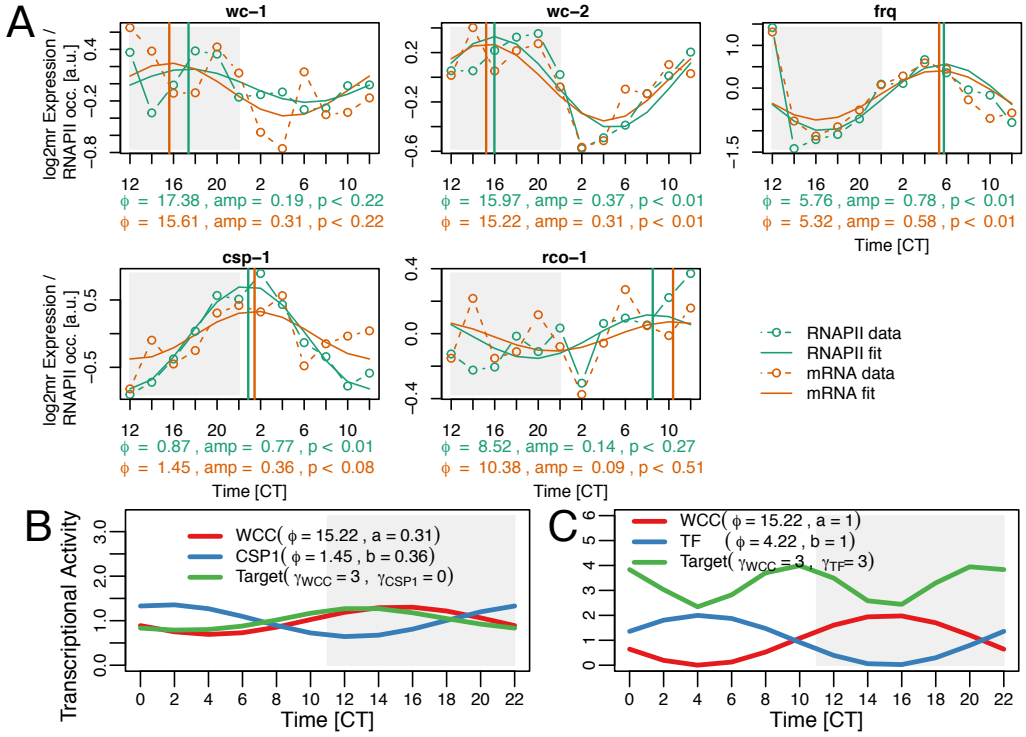


Fig. 4.3: **RNAPII occupancy profiles and Expression profiles of the *Neurospora crassa* core clock transcription factors together with a target gene simulation.** (**A**) The RNAPII occupancy (green dash-dotted) and mRNA abundance (orange dash-dotted) core clock gene profiles for white collar (*wc-1/2*), frequency (*frq*), *csp-1*, and its co-repressor *rco-1* are shown, together with the harmonic regression models (same colour, solid lines). The harmonic regression model parameters phase (ϕ), amplitude (amp), and the resulting p-value (p) are provided below each panel in corresponding colour. (**B**) Simulated transcriptional activity of a combined WCC / CSP1 target gene. (**C**) Simulated ultradian transcriptional activity of a gene regulated by WCC and a second hypothetical activating TF in anti-phase to WCC.

4.3.1 RNAPII and mRNA rhythms vary extensively

In the following, the relationship of rhythmic transcriptional activity and mRNA abundance was analysed. Two phase clusters emerge in the phase distributions of all genes with significantly oscillating RNAPII or mRNA abundance, represented by circular histograms in Figure 4.2 A and B, respectively. RNAPII phases feature one dominant dawn cluster at CT 0-3 and a weak dusk cluster at CT 11. In contrast, the size of the dawn phase cluster (CT 19-22) in mRNA abundances is comparable to the dusk phase cluster (CT 8-10). The dominant role of dawn- and dusk-specific expression agrees well with previous reports [38, 80, 152]. The fractions of rhythmic profiles with a period of 22 h in RNAPII 10.2% (993) and mRNA 12.1% (1178) are comparable. The gene set featuring rhythmic RNAPII and mRNA is comprised of 3.2% of all genes (R-R set, 262 genes). 682 genes (8.3%) exhibit rhythmic RNAPII occupancy and arhythmic mRNA abundance (R-AR

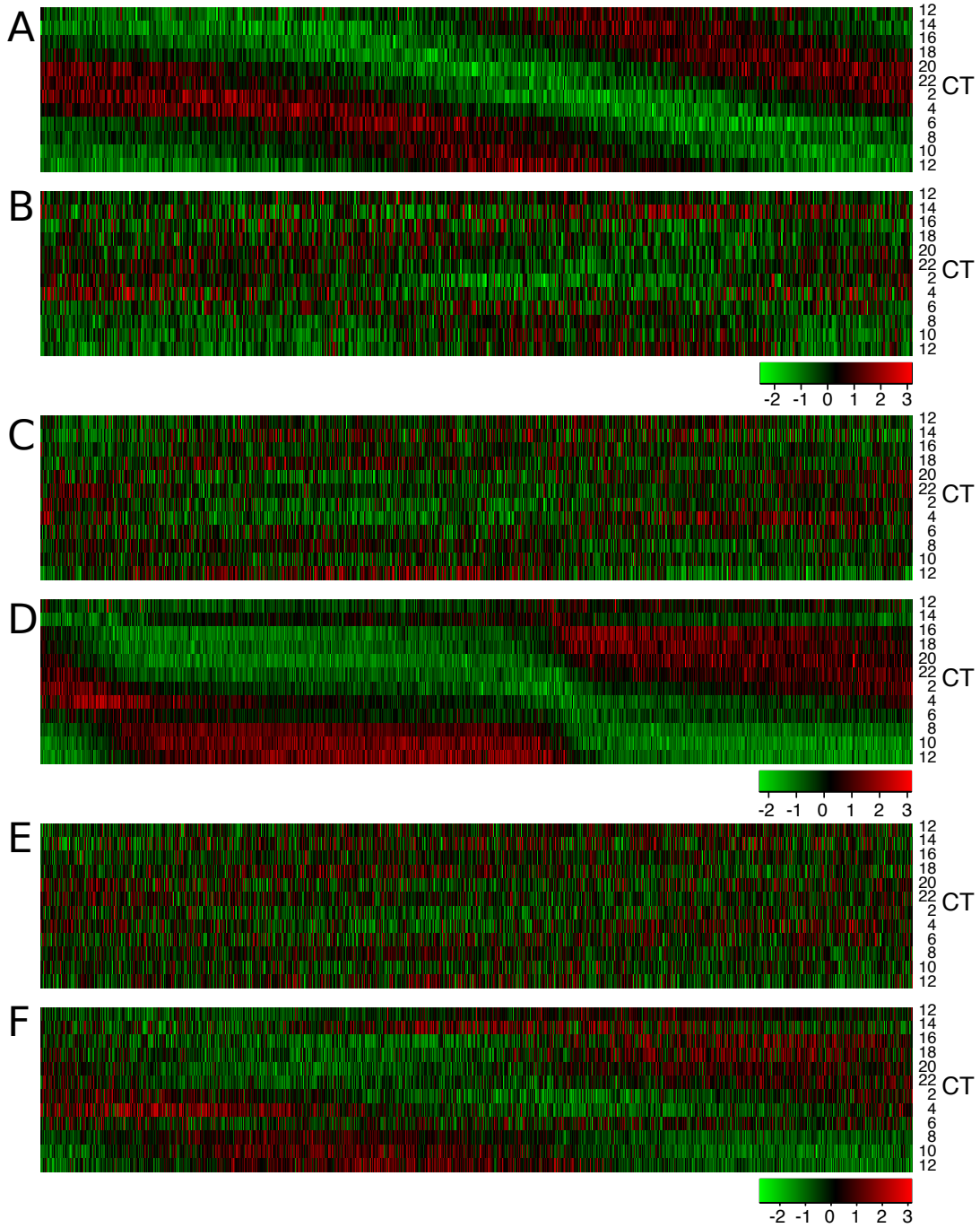


Fig. 4.4: **RNAPII occupancy and mRNA abundance of the R-AR, AR-R and AR-AR gene sets.** RNAPII occupancy (**A**) and mRNA abundance profiles (**B**) for the R-AR gene set (n = 682), the AR-R gene set (n = 916, **C** and **D**), and the AR-AR gene set (n = 6372, **E** and **F**) are shown, where green marks low and red high values similar to Fig. 4.2 C and D. The genes are ordered by peak phase ϕ of the significantly oscillating group, and the values are gene-wise z-transformed. Each pair of RNAPII and mRNA heatmaps are scaled identically according to the colour-bar in the respective right bottom corner, to allow direct visual comparison of oscillatory amplitudes.

set), 916 genes (11.1%) vice versa (AR-R), and 6372 genes (82.3%) are classified as completely arhythmic. Only genes with sufficiently high expression were considered in this grouping, discarding 42 genes with rhythmic RNAPII but low mRNA abundance. Comparison of the corresponding phases of the R-R genes shows high correspondence (circular correlation coefficient $\rho_{ccc} = 0.44, p < 3.4 \times 10^{-10}$) of the peak times (Figure 4.2 E, red) with the characteristic dawn and dusk clusters. The variability of RNAPII phases in the R-R set is larger than for the mRNA phases, resulting horizontally protruding elliptical clusters. No significant circular phase correlation was found for the AR-R set ($\rho_{ccc} = 0.03, p < 0.28$) and the R-AR set ($\rho_{ccc} = -0.009, p < 0.82$).
The mean phase difference between mRNA abundance and RNAPII binding in the R-R set is -1.1 h, indicating that the mRNA phase generally precedes the RNAPII phase. The mean phase difference amongst all expressed genes is -0.7 h, accordingly. The phase difference distribution (Fig. 4.2 F) reaches from -7 h until 8.3 h with a small overrepresentation of large positive differences >5 h.
Only 8 genes feature a delay $> 5h$ (see Appendix C Fig. 22), the small nucleolar ribonucleoprotein Lcp5 (*NCU07331*), the H/ACA ribonucleoprotein complex subunit 2 (*NCU08951*), a translocase of outer mitochondrial membrane 6 (*NCU05772*), the cell cycle control protein cwf16 (*NCU01390*), tRNA-ser *NCU11837*, and the hypothetical protein genes *NCU00733*, *NCU03451*, *NCU06513*. Comparison of the delays between nascent RNA and mRNA in mouse liver reported by Menet *et al.*[135] with the observed distribution yields similar shape, but a markedly smaller width of the latter.
With six out of 12, about half of the known clock controlled genes (ccgs 1, 2, 4, 6, 7, 8, 9, 12, 13, 14, 15, 16) exhibited significant circadian rhythms on RNAPII or mRNA level (see Appendix C Fig. 23). The peak phases agreed surprisingly well with values previously reported by Bell-Pedersen *et al.* [13]. Specifically, ccg-1 peaks at CT ≈ 3, ccg-2 at CT $21 - 22$ and ccg-9 at CT $19 - 21.5$. The largest difference was observed for ccg-6, peaking at CT 10.6 in the presented dataset, whereas CT 19 was reported. Moreover, ccg-6 (*NCU01418*) exhibited a significant RNAPII rhythm but no corresponding mRNA rhythm. This hints towards a long mRNA lifetime, or post-transcriptional modification mechanisms, both of which could suppress mRNA rhythms. The observed diversity of rhythms amongst ccgs agrees well with recent reports [80].

Temporal separation of cellular functions Functional enrichment analysis of dawn and dusk phase genes was performed to test the compatibility of the timing of central biological processes in this dataset with literature knowledge. As described in detail in Appendix C, Section 4.4, the temporal organisation in this dataset is in agreement with recent reports of Hurley *et al.* [80].

The R-AR, AR-R, and AR-AR gene sets The R-AR set contained 682 genes (8.3%) which exhibit rhythmic RNAPII occupancy and arhythmic mRNA abundance (Fig. 4.4 A and B, ordered by ϕ_{RNAPII}). Direct comparison of the RNAPII phase distributions between the R-R (Fig. 4.2 A) and the R-AR set (Appendix C Fig. 24 A) revealed several distinctions. The dawn-cluster in the R-AR set was smaller and a

dusk-specific cluster was not discernible. Instead, intermediate phased genes were found across the entire day, yielding a more dispersed phase distribution. Visual inspection of the mRNA abundance heatmap indidacted low amplitude oscillations often appearing in phase with the RNAPII profile (Fig 4.4 B). The corresponding mRNA phase distribution however displayed only rudimentary dawn and dusk clusters emerging above the uniform background (data not shown). This confirms that these mRNA oscillations were below the sensitivity of the employed procedure, and that the p-value threshold was chosen appropriately to exclude these cases.

Genes with arhythmic RNAPII occupancy and rhythmic mRNA abundance (Fig. 4.4 C and D, respectively, ordered by ϕ_{mRNA}) were collected in the AR-R set (916 genes, 11.1%). The phase distribution of mRNA profiles resembled that of the R-R set with two distinct dawn and dusk clusters (Appendix C Fig. 24 B for direct comparison). An interesting pattern was observed for the sample at CT 14. Dusk-peaking genes exhibit low expression at the first and third sample with an intermediate peak at CT 14 (Fig. 4.4 D, left). On the other hand, dawn-peaking genes show the opposite pattern, with high expression at CT 12 and 16 with a transient repression at CT 14 (Fig. 4.4 D, right). Considering that WCC is the major blue light receptor in *Neurospora crassa*, these fluctuations could reflect light-induction or repression by a light flash which was applied to the culture as synchronising signal immediately before the beginning of the experiment. Again, visual indications of rhythmicity in the arhythmic RNAPII profiles were not reflected in the uniform phase distribution (data not shown).

As anticipated, the visual representation of RNAPII profiles of the 6372 genes in the AR-AR set does not indicate significant rhythmicity (Fig. 4.4 E, ordered by ϕ_{RNAPII}) and thus yields a uniform phase distribution. Surprisingly, this is not the case for the mRNA profiles (Fig. 4.4 F, ordered by ϕ_{mRNA}). They exhibit extensive low-amplitude oscillations and a similar transient abundance change at CT 14 as found in the AR-R set. Furthermore, the phase distribution is strongly bimodal with peaks at CT 8-9 and CT 19-20 (not shown).

In order to further explore the connection between only visually observed RNAPII rhythmicity and significantly detected mRNA rhythmicity, the variability of both signals was compared. The hypothesis was that increased non-rhythmic variability in the RNAPII signal still promotes circadian rhythmicity in the mRNA abundance. Variability was defined as gene-wise standard deviation across all samples of the corresponding dataset. Including all genes, the mean mRNA variability ($\bar{\sigma} = 0.36$) is generally larger than the mean RNAPII ($\bar{\sigma} = 0.26$) variability (single sided Mann-Whitney-U-Test, $p < 8 \times -243$). The only exception is the R-R set. Circadian oscillatory pattern increased variability in RNAPII leading to similar values in both data sets ($\bar{\sigma} = 0.32$, $p < 0.18$).

As expected, the R-R set exhibits the highest and the AR-AR set the lowest standard deviation in RNAPII and mRNA (Appendix C Fig. 21 A and B). Surprisingly, RNAPII variability in the AR-R set (*i.e.* no RNAPII rhythmicity) is elevated from the base level and similar to the R-AR set (*i.e.* rhythmic RNAPII). This variability might represent the low-amplitude circadian rhythmicity revealed by visual inspection. This interpretation is supported by the comparison of mRNA variability. Only rhythmic mRNA gene sets (R-R and AR-R) exhibited elevated standard deviations relative to the AR-AR

base line suggesting directionality, *i.e.* RNAPII variability leads to mRNA variability but not vice versa. This posed the question whether elevated RNAPII variability, which was not classified as circadian oscillation, is correlated with circadian oscillations in the corresponding mRNA. Indeed, the analysis revealed a Pearson correlation of $\rho = 0.63$ (Appendix C Fig. 21 C). In contrast, the amplitude of circadian mRNA profiles is independent from the corresponding RNAPII variability (Appendix C Fig. 21 D, $\rho = 0.21$).

Overall, visual inspection of gene sets not classified as rhythmic revealed sublime circadian rhythmicity. Elevated variability in the arhythmic RNAPII profile was correlated with higher variability and circadian oscillations in the corresponding mRNA profile, whereas a reverse relationship was not found. This suggests that a second group of genes receives only weak circadian transcriptional regulation which is translated or amplified into regular mRNA rhythmicity.

Core clock gene transcriptional patterns As described before, the core circadian clock in *Neurospora crassa* is composed of the transcription factors white collar 1 and 2 (WC-1, WC-2), frequency (FRQ), and CSP1. The three genes *wc-2*, *frq*, and *csp-1* exhibit significantly oscillating RNAPII profiles (Fig. 4.3 A) which persist in the mRNA profiles of *wc-2* and *frq*. In case of *csp-1*, a premature increase of mRNA abundance from noon on towards the evening is observed reducing its significance. Importantly, the phase relationships of these key elements agree with previous knowledge. *wcc* expression peaks around midnight (CT 15.61) so that the protein dimer can be present at dawn to act as transcriptional activator. Accordingly, transcription of the *csp-1*, a known WCC target gene, peaks shortly after dawn (CT 1.45). *frq*, another known WCC target gene, exhibits a much slower induction resulting in a peak phase at CT 5.32. This anti-phasic relationship between *frq* and *wcc* is particularly plausible due to the negative feedback loop formed by both TFs. The observed arhythmicity of *wc-1* and *rco-1* is of no consequence, since both TFs require rhythmically transcribed complex partners, thereby producing circadian activity patterns.

4.3.2 Minor phase shift predicted for shared WCC and CSP1 targets

Morning-specific circadian expression is facilitated by WCC and evening-specific expression by CSP1 [188, 201]. As shown by Westermark and Herzel [249], the combination of circadian transcription factors can lead to different target gene transcription dynamics, including doubling of the oscillation frequency or phase shifts. The combination of WCC and CSP1 can be modelled as circadian AND funnel with activator and repressor. As necessary prerequisite, the DNA binding motifs differ sufficiently to exclude cross-hybridisation of the two transcription factors and thereby competition for binding sites (DNA binding motifs shown in Figure 4.1 B and C). Therefore, the transcriptional activity of target genes can be modelled as:

$$
\begin{aligned}
x_{WCC} &= 1 + a\cos(\omega t - \phi_{WCC}) \\
x_{CSP1} &= 1 + b\cos(\omega t - \phi_{CSP1}) \\
x_t &\approx \frac{1+\gamma_{WCC}(1+a\cos(\omega t-\phi_{WCC}))}{1+(1+a\cos(\omega t-\phi_{WCC}))} * \frac{1+\gamma_{CSP1}(1+b\cos(\omega t-\phi_{CSP1}))}{1+(1+b\cos(\omega t-\phi_{CSP1}))}
\end{aligned}
$$

Here, ϕ denotes the transcription factor activity phase, x_{WCC}, x_{CSP1} are the corresponding TF activity profiles, x_t is the transcriptional activity of the target gene, $a_{WCC} = 0.31$ and $b = 0.36$ are the estimated amplitudes of the respective TF's, $\omega = 2\pi/22h^{-1}$ is the oscillation frequency, and γ is the fold change factor specifying the TF's activation or repression efficiency. In this case, parameters were set to $\phi_{WCC} = 2\pi/22 * 22$ due to the estimated RNAPII phase of WC-1, $\phi_{CSP1} = 2\pi/22 * 1.45$, $\gamma_{WCC} = 3$ to model transcriptional induction, and $\gamma_{CSP1} = 0$ for transcriptional repression (see Figure 4.3 A and B). The predicted target gene does not exhibit harmonic oscillation, but an amplitude reduction and a minor phase shift away from WCC towards CSP1. It must be noted that the fold change factors of γ_{WCC} and γ_{CSP1} can vary amongst target genes, thereby widening the expected target gene phase distribution. To demonstrate the generation of an ultradian target gene, a second hypothetical inducing TF was assumed which features an anti-phasic transcriptional profile to WCC. The transcriptional pattern of the hypothetical ultradian target gene together with both TFs is shown in Figure 4.3 C. In the following, the observed phase distributions for target genes sharing input from WCC and CSP1 will be compared to this prediction. Furthermore, genes with ultradian transcription will be tested for TF binding preferences.

4.4 Transcriptional patterns of core clock transcription factor target genes

The phase distributions of target genes for WCC and CSP1, as predicted from published ChIP-seq data, are compared with literature knowledge. Phases for combined targets of both TFs are compared with the theoretical prediction.

4.4.1 Target gene sets for WCC and CSP1/RCO1

Genomic locations of transcription factor binding for WCC, CSP1, and RCO1 were taken from the original publications [188, 201]. Transcriptionally regulated target gene sets were predicted using the transcription factor association score (TFAS) as defined by Ouyang *et al.* [155] with a score threshold of 200. This resulted in 1590 WCC target genes, 374 genes targeted by CSP1 and RCO1, 1250 CSP1 target genes, and 826 RCO1 targets.

CSP1 association predicts circadian transcription better than WCC association Fig. 4.5 represents the number of predicted target genes (x-axis), distinguished by oscillating and non-oscillating genes. The different combinations of datasets and TFs constitute the y-axis. While RCO1 is not a circadian TF by itself, as shown above, it

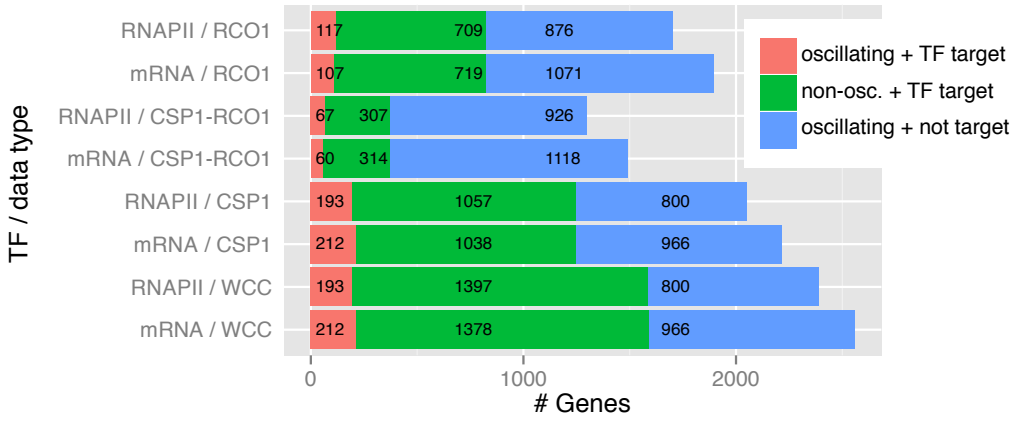


Fig. 4.5: **Number of predicted target genes of WCC, CSP1, and RCO1, distinguished by oscillating and non-osillating** The absolute number of genes (x-axis) in each of the of the predicted target gene sets (WCC, CSP1, CSP1 and RCO1, RCO1) found oscillating (red) or non-oscillating (green) compared with the number of oscillating non-target genes (blue). The combinations of data set (RNAPII or mRNA) and TF target gene sets are shown along the y-axis.

acts as co-repressor for CSP1 [188]. Accordingly, target genes alone should not exhibit circadian transcription due to exclusive RCO1 input (*i.e.* background), whereas targets of CSP1 and RCO1 should. While the absolute number of oscillating target genes varied between the TFs, the relative fraction of oscillators amongst all predicted targets was generally similar. This result is in contradiction with the expectation of fewer oscillations amongst RCO1 targets, pointing to either the effect of another circadian TF or extensive circadian background signal. However, CSP1 targets and combined CSP1 / RCO1 targets consistently showed higher fractions of oscillating genes ($\approx 20\%$) as compared to the background of RCO1 targets ($\approx 15\%$). On the other hand, WCC targets did not show a higher fraction oscillating genes compared to the background. This indicates higher complexity in the transcriptional regulatory mechanism of WCC compared to CSP1. Alternatively, a lower quality of the corresponding dataset might lead to such an outcome. However, inferring transcriptional regulatory function from genomic TF binding sites remains challenging and the fractions of oscillating target gene sets were accordingly small.

A detailed analysis of the TFAS scores substantiated the difference between WCC and CSP1 targets. TFAS scores for oscillating genes were compared to those of non-oscillating genes, independently for RNAPII and mRNA. CSP1's association scores to oscillating genes were significantly higher, whereas WCC did not exhibit a significant difference. Specifically, the mean CSP1 score for genes with RNAPII oscillations was greater with 104 compared to 72 for non-oscillating ones (Mann-Whitney-U-Test, single sided, $p < 1 \times 10^{-5}$). Similarly, the mean for genes with oscillating mRNA was greater with 98.5 compared to 72 ($p < 1 \times 10^{-7}$). WCC score differences were small for RNAPII (106.5 vs. 98, $p < 0.04$) and mRNA (103 vs. 98, $p < 0.12$). The mean association scores

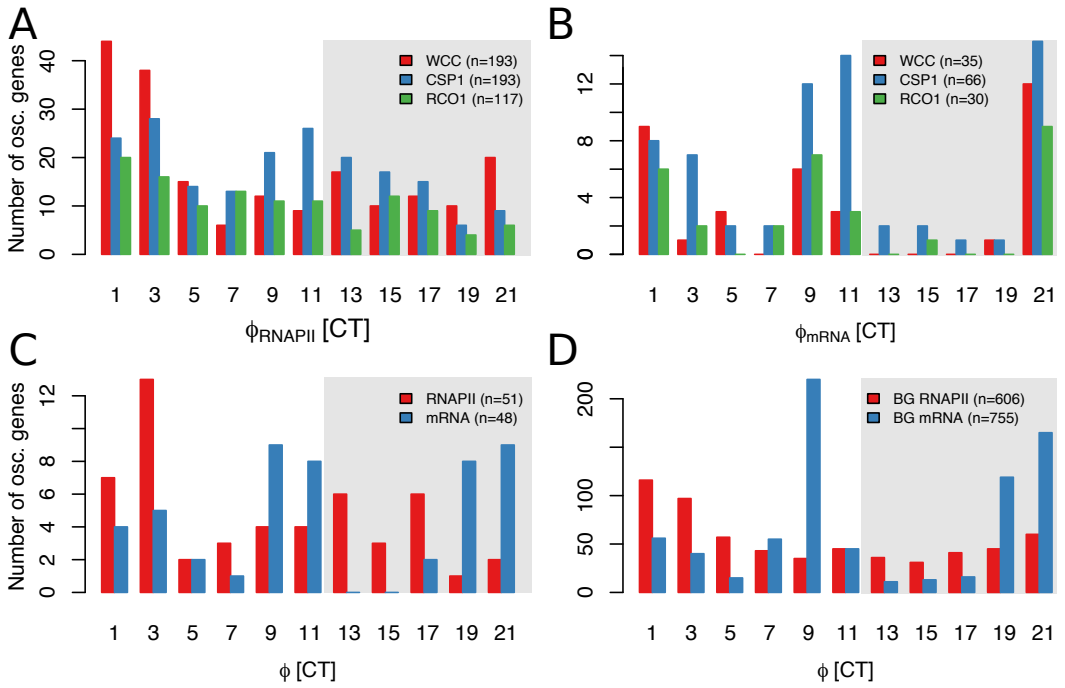


Fig. 4.6: **Phase distributions of predicted target genes of WCC, CSP1, and RCO1.** Phase angle (ϕ) comparison of RNAPII binding and mRNA abundance for oscillating target genes of clock transcription factors. (**A**) RNAPII peak phases of predicted target genes for WCC, CSP1, and RCO1. The numbers of genes in the corresponding class is given in the legend. (**B**) mRNA peak phases for predicted target gene groups, including only genes from the R-R set. (**C**) RNAPII and mRNA peak phases for genes predicted to be target of WCC and CSP1. (**D**) Background phase distributions of oscillating genes without predicted core clock transcription factor association. Subjective night shown as grey shaded area.

increased for both TFs when comparing only the R-R gene set versus the remaining genes (WCC: 116.5 vs. 98, $p < 0.04$, and CSP1: 134 vs. 74, $p < 1.7 \times 10^{-5}$), but remained non-significant for WCC.

The following analysis addressed whether the described TFAS differences translate into significant overrepresentation or depletion of predicted target genes in the gene sets with oscillating RNAPII mRNA, and both (R-R). For this purpose, hypergeometric tests were performed, interpreting the predicted target genes as success and the oscillating genes as sample from the entire RNAPII or mRNA dataset. No significant depletion of WCC or CSP1 targets was found in the gene sets with oscillating RNAPII, mRNA, and the R-R set. In contrast, predicted CSP1 target genes were significantly enriched in each oscillating set. Eenrichment was strongest in the oscillating RNAPII set (4.7×10^{-13}), intermediate in the mRNA set ($p < 9.5 \times 10^{-11}$), and weakest in the R-R set ($p < 4 \times 10^{-9}$). On the other hand, enrichments for WCC targets were less clear, where the RNAPII set was again most significantly enriched ($p < 8.7 \times 10^{-4}$), followed by the R-R set ($p < 1.5 \times 10^{-3}$)

and the mRNA set ($p < 0.027$).
In summary, TFAS statistics and hypergeometric tests suggest that WCC target genes predicted from the ChIP-seq experiment by Smith *et al.* [201] are not more often than expected by chance circadian oscillators on RNAPII and mRNA level. In contrast, CSP1 target genes predicted from the experiment by Sancar *et al.* [188] are more often circadian oscillators than expected by chance.

WCC and CSP1 target gene RNAPII phases match previous knowledge Circadian oscillating WCC target genes exhibited a clear predisposition for peak phases around dawn in their RNAPII occupancy profiles (Fig. 4.6 A, red). Out of the 24 transcription factors with associated WCC binding sites, described by Smith *et al.* [201], the three genes *sub-1/NCU01154*, *NCU04295*, *NCU06095* were part of the R-R set with peak phases around dawn. Four genes exhibited only significant RNAPII oscillations (*csp-1/NCU02713*, *bek-1/NCU00097*, *NCU06536*, *NCU07846*) with dawn phase, except *NCU07846* ($\phi_{RNAPII} = 14.5$). The genes *NCU01871* and *adv-1/NCU07392*) showed only mRNA oscillations, with dawn- and dusk-specific phases, respectively. The circadian CSP1 target genes exhibited a bimodal phase distribution (blue), with one peak at dusk as pointed out by Sancar *et al.* [188]. The second peak, however, matches that of the WCC target genes at dawn. The target gene phase distribution for RCO1, the non-circadian expressed co-repressor of CSP1, is comparably uniform with a small peak at dawn. The background phase distributions of all circadian RNAPII and mRNA profiles are depicted in Fig. 4.6 D. The dawn peak of WCC targets is much more distinct compared to the background. While the dusk peak is specific for CSP1 target genes and matches previous descriptions, the presence of the background-typical dawn peak indicates a significant number of false-positives in the target prediction. Finally, the phase distribution of RCO1 (green), which is not a circadian regulator by itself, matches that of the background and, therefore, successfully serves as negative control.
Only genes of the R-R set were considered for the comparison of mRNA phases. The observed differences between the mRNA phase distributions for the three TFs (Fig. 4.6 B) are considerably smaller amongst each other, and also compared to the background distribution (Fig. 4.6 D, blue). All phase distributions exhibit two peaks of comparable height shortly before dawn and dusk. Again, WCC targets showed the strongest preference for dawn phases. As expected, the negative control RCO1 yielded the smallest phase peaks.
A total of 221 genes feature predicted simultaneous regulatory input from WCC and CSP1. These genes were employed as test set for the predictions derived from the phase-based model. Compellingly, the RNAPII phase distribution of WCC and CSP1 target genes (Fig. 4.6 C, red) match that of purely WCC targeted genes. This result is in agreement with the prediction. As described in section 4.3.2, the transcriptional repressor CSP1 is not sufficiently in phase with the activator WCC such that the resulting circadian transcriptional phase is only marginally shifted (Fig. 4.3 B, green) instead of leading to frequency doubling. Likewise, the mRNA phase distribution matches that of purely WCC or CSP1 targeted genes with its bimodality, featuring dawn and dusk peaks.

4.5 Ultradian expression rhythms not linked to WCC or CSP1

The set of genes with ultradian transcriptional rhythms is analysed regarding functional annotations, transcriptional regulation by WCC or CSP1, and the overrepresented motifs indicating putative TF binding sites.

As described in section 4.3.2, mathematical modelling predicts that combined input of two circadian transcriptional regulators can result in frequency doubling, *i.e.* ultradian rhythmicity if the right phase relationship is given between both regulators. *wcc* and *csp-1* are, however, not sufficiently in phase to predict ultradian rhythmicity in combined target genes. Therefore, the harmonic regression procedure with permutation background was repeated, while changing the period parameter from 22 h to 11 h. This analysis resulted in a total of 357 genes with ultradian RNAPII profiles and 185 genes with ultradian mRNA profiles, which are depicted in Figure 4.7 A and B respectively. The corresponding phase histograms are shown in panels C and D. Surprisingly, the set of genes with ultradian rhythmicity in RNAPII profile and mRNA profile was very small (n = 6) when compared to the total number of ultradian oscillating RNAPII genes (n = 357) and mRNA genes (n = 185). The phase distributions are bimodal with two nearly anti-phasic clusters, resembling those of circadian genes. Furthermore, the RNAPII phase clusters lagged $\approx$ 1 h behind the mRNA phase clusters. In order to perform functional enrichment analysis and the discriminative motif overrepresentation analysis of gene promoters, the set of genes with ultradian RNAPII profiles was divided into four subgroups (I-IV) with comparable peak phase, as indicated in Fig. 4.7 A. This was necessary due to the underlaying assumption that genes with similar transcriptional patterns are more likely functionally related. According to the mathematical model, each subgroup features a specific combination of two circadian transcriptional regulators in order to achieve the ultradian rhythmicity with the particular phase. Gene ontology annotation enrichment analysis revealed overrepresentation of translation related genes amongst all ultradian genes (*GO:0006412*, 6 genes, $q < 0.009$) caused by an enrichment in group III (5 genes, $q < 4 \times 10^{-5}$). Specifically, the 60S ribosomal proteins L18 (*NCU03988*), L19 (*NCU05804*), L20 (*NCU08389*), L36 (*NCU03302*), L39 (*NCU08990*) were found in group III, whereas the mitochondrial ribosomal protein subunit L32 (*NCU03516*) was sorted into group II due to a slightly earlier transcriptional onset. Group II was enriched for the annotation "coenzyme metabolic process" (3 genes, $q < 0.03$) whereas group IV was significantly enriched for the term "response to light stimulus" (4 genes, $q < 0.03$). FunCat functional enrichment analysis of the total set of ultradian RNAPII genes reflected these results, yielding 9 genes annotated with "cell cycle and DNA processing" ($q < 0.004$, also slightly enriched in group IV, $q < 0.04$), 21 with "protein synthesis" ($q < 0.009$, 15 of which were enriched in group III, $q < 2 \times 10^{-11}$), and 57 with "protein with binding function or cofactor requirement" ($q < 0.04$, also highly enriched in groups II and III). There were also group-specifically enriched functional annotations, "Energy" for group

II ($q < 0.03$) and "Metabolism" for group III ($q < 0.001$). Group IV featured a range of specific enriched functional annotations, namely "Cell rescue" ($q < 2 \times 10^{-4}$), "interaction with environment" ($q < 2 \times 10^{-3}$), "cellular transport" ($q < 0.015$), and "transcription" ($q < 0.035$).

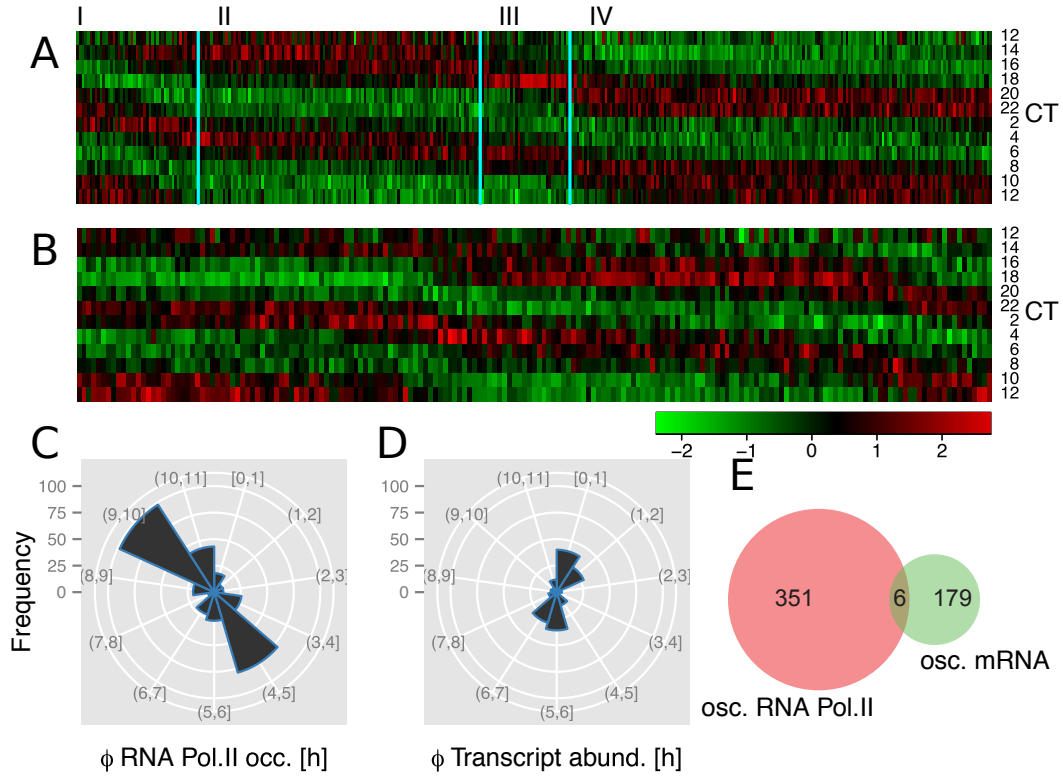


Fig. 4.7: **Ultradian transcriptional rhythmicity in *Neurospora crassa*.** (**A**) Heatmap representation of all genes with ultradian RNAPII profiles (11 h period), divided into four groups according to the dominant phase (I-IV). (**B**) Heatmap of genes with ultradian mRNA profiles, similar to A. (**C**) The peak phase distributions of all genes shown in A, and (**D**) similarly for all genes in B. (**E**) Overlap of genes with ultradian RNAPII and mRNA profiles in the R-R set (see Fig. 4.8 for profiles of R-R set genes).

Ultradian R-R genes The profile of the six prominent genes in the ultradian R-R set is depicted in Figure 4.8. The nuclear protein localisation protein 4 (NPL4, *NCU02680*) is involved in the import of nuclear-targeted proteins into the nucleus and in the export of poly(A) RNA out of the nucleus, as well as in the endoplasmic reticulum-associated degradation (ERAD) pathway. The peak phase at dawn and dusk is consistent with the results of the functional enrichment analysis which indicated elevated transcriptional activity at dusk and protein processing which could result in the transport to the nucleus. Two other transcription-related genes are amongst the ultradian R-R set: the pre-mRNA splicing factor (*NCU11251*) and the SNF2 family helicase / ATP-ase (*NCU03652*). The latter is particularly interesting as members of the SWI2/SNF2 family are involved in

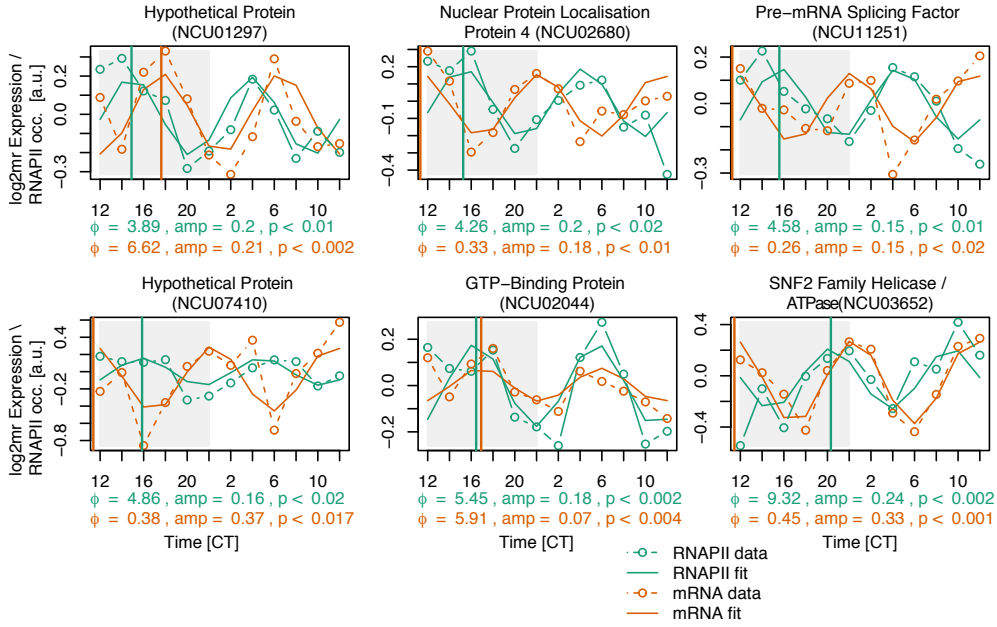


Fig. 4.8: **Six genes exhibited ultradian RNAPII and mRNA rhythms.** Ultradian RNAPII profiles (green) and mRNA abundance profiles (orange) of all six genes in the R-R set ($p < .025$). Peak phases are represented as vertical solid lines, the harmonic regression model as solid sinoidal line, and the underlying data as circle-dotted line in corresponding colour. The model parameters are shown below each panel similarly coloured.

transcription regulation, chromatin remodelling, and DNA repair including the disruption of protein-protein and protein-DNA interactions. Its peak expression during dawn and dusk is biologically reasonable since necessary LD/DL transition specific adjustments of gene transcription are likely to involve also chromatin remodelling. Adding further complexity to its transcriptional pattern, recent research has shown that the SWI–SNF complex is recruited by WCC which then modulates rhythmic opening of chromatin at the *frq* locus [239]. It was suggested that this might be a general principle of WCC acting as a pioneer TF in order to prepare the chromatin structure around it for the binding of further TFs. Another interesting ultradian gene is the GTP-binding protein (*NCU02044*). Previous work demonstrated down-regulated expression in a *Neurospora crassa mak-1* knockout strain [15]. Compellingly, the *mak-1* map kinase (MAPK) pathway is regulated by the circadian clock. In particular, the circadian activation of *mak-1* suggests that growth and development in *Neurospora crassa* might be clock controlled partly by the *mak-1* MAPK pathway. This would fit to the temporal expression regulation of the GTP-binding protein which peaks at mid-night and day, presumably after LD/DL transition adjustments are accomplished and the metabolism is working efficiently. Furthermore, the expression of the hypothetical protein gene *NCU01297* is regulated by the map kinase *mak-2* in *Neurospora crassa* [119] as well as its homolog in *Podospora anserina* [18]. The peak phase was close to that of the *mak-1* regulated GTP-binding protein *NCU02044*.

Unfortunately, no details were available to the hypothetical protein gene *NCU07410*.

WCC and CSP1 target genes not overrepresented amongst ultradian genes The ultradian oscillating RNAPII gene set is composed of 255 genes not targeted by WCC or CSP1. 50 of these genes are predicted WCC targets, 45 are predicted CSP1 targets, and seven are both. Similarly, the ultradian oscillating mRNA gene set contains 136 genes which are not predicted WCC or CSP1 target. 24 genes are predicted target genes for each TF, and one gene targeted by both. Assessment of the enrichment of predicted target genes via hypergeometric test did yield neither significant enrichment nor depletion of WCC or CSP1 amongst genes with ultradian oscillating RNAPII or mRNA. Out of the six genes in the ultradian R-R set, three genes possess predicted WCC binding sites (NPL4, *NCU01297*, *NCU07410*) whereas only SNF2 (*NCU03652*) exhibits a CSP1 binding site.

Overrepresented motifs vary between ultradian gene groups Since WCC and CSP1 did not did not exhibit notable binding preferences in proximity of ultradian expressed genes, the next step was to analyse ultradian gene promoters for overrepresented motifs indicating binding of other potentially circadian transcriptional regulators. The promoter was defined as 1 kb sequence window, that is centred around the transcriptional start site annotated in the Broad Institute NC12 genome build. Discriminative motif overrepresentation analysis allows for the efficient detection of short motifs, typical core transcription factor binding sites, which occur more often in the foreground sequence than randomly expected. The power of this analysis relies heavily on background sequence as the definition of random chance. A common approach is to randomly shuffle the foreground sequence, which ensures preservation of the base composition but removes any correlations between neighbouring bases. This has the drawback of not considering other sequence properties such as the CpG content. A better approach to the background definition is the selection of a gene set which does not exhibit oscillatory transcription but matches the basic sequence composition. All genes exceeding a harmonic regression p-value of 0.2 and with an estimated amplitude below 0.06 were selected, encompassing 423 genes in total. Since *Neurospora crassa* possesses only few low-GC / high-CpG promoter sequences (Fig. 4.9 bottom right), stratification is not necessary in this case. As different phases are presumably caused by different transcription factors or combinations, the ultradian gene set was subdivided into four groups (I-IV) as indicated in Figure 4.7 A. The respective number of available gene sequences in group I-IV was 45, 105, 35, and 155. For each group, the matching number of background sequences was randomly selected from the full background set. The resulting sequence sets were then tested for differential overrepresentation of short sequence motifs of 8 bp length using DECOD [77]. This yielded sets of the six most significantly overrepresented motifs for each group and the total set of ultradian genes (Fig. 4.9). The discovered motifs for each group were compared to uncover possible similarities in the regulation. The pairwise motif similarity scores were calculated as described in Pietrokovski *et al.* [166] (Fig. 4.9, bottom right "Motif Similarity"). The similarity scores were generally low between all discovered motifs.

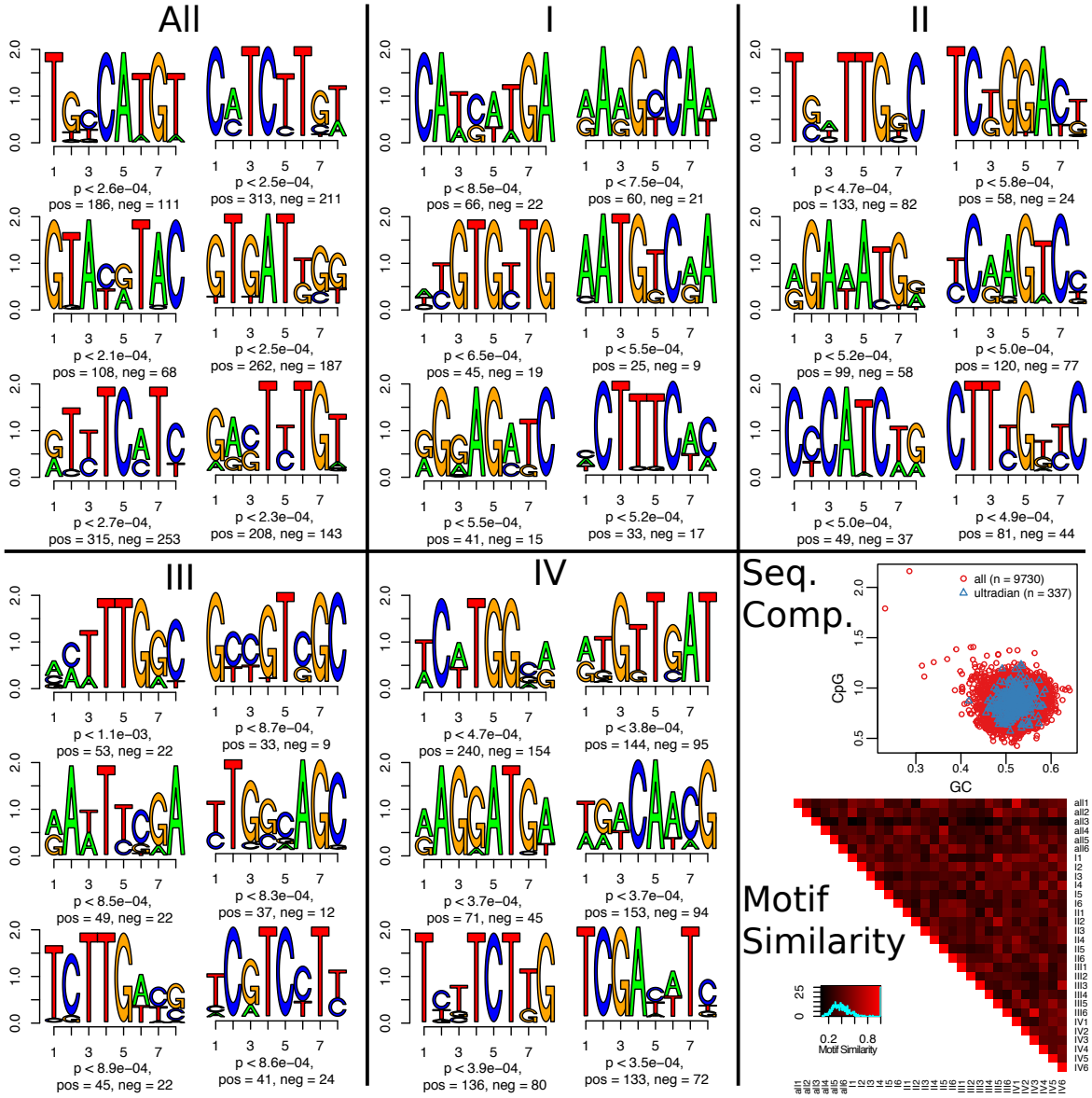


Fig. 4.9: **Overrepresented motifs in ultradian RNAPII gene promoters and sequence composition.** (**All**) Overrepresented motif sequence logos obtained from a discriminative motif overrepresentation analysis of all 337 ultradian RNAPII gene promoter sequences (1 kb centered around the transcriptional start site) compared to a background set of genes with least-oscillating RNAPII profiles. The six most significant motifs are shown. The overrepresentation p-value is given below each panel, together with the number of occurrences in the foreground ('pos') and background ('neg'). (**I-IV**) Results of discriminative motif overrepresentation analyses of the ultradian gene subgroups. Motifs are row-wise numbered 1-6. (**Composition**) Comparison of GC (x-axis) and CpG content (y-axis) of promoter sequences between genes with ultradian RNAPII oscillation (blue) against all genes in *Neurospora crassa* with available promoter sequence (red). (**Motif Similarity**) Pairwise similarity scores shown as symmetrical heatmap. Red represents high scores (max. 1) and similarity whereas black represents dissimilarity (min. 0) as indicated in the colour map. The distribution of similarity scores is shown in the colour map (cyan).

Only two motif pairs reached scores >0.8, the first motifs in group all and group IV, as well as motif 6 of group III and motif 3 of group IV. This dissimilarity suggests group specific and, thus, phase specific TFs regulating the ultradian transcriptional patterns. Presently, no collection of known binding motifs is available for *Neurospora crassa*. The commonly used JASPAR database features a list of 177 TF binding motifs for the fungus *Saccharomyces cerevisiae* [130]. Theses motifs were compared with the presented motifs. Prior to the comparison, edges below one bit information content were trimmed. Pearson correlation was employed as similarity measure after motif alignment via the Smith-Waterman algorithm, as used in the STAMP algorithm [128]. The two most similar *Saccharomyces cerevisiae* TFs for each discovered motif are shown in Appendix C, Fig. 25 for the groups "all", "I-III" and for the group "IV" in Fig. 26. The empirical similarity p-value, based on simulated PSSMs, was not significant ($p > 0.001$) for 31 of all 60 comparisons, as can be expected for a comparison between different species. Only motif 1 in group II showed highly significant similarity to the CCAAT motifs bound by HAP3 ($p < 6.5 \times 10^{-8}$) and HAP5 ($p < 4 \times 10^{-7}$). Multiple other motifs also showed significant similarity to HAP TFs. Indeed, two HAP genes (HAP2, HAP3) occur in the *Neurospora crassa* genome. Another motif, RIM101, occurred most often in this similarity search. The RIM101 gene in *Neurospora crassa* is referred to as *pacC* and encodes a pH-response transcription factor. Recent work suggests that the *pacC* signal transduction pathway in *Neurospora crassa* is similar to the PacC/Rim101 of *Saccharomyces cerevisiae* [33]. This transcription factor mediates the regulation of both acid- and alkaline-expressed genes in response to ambient pH. At alkaline ambient pH, it activates transcription of alkaline-expressed genes and represses transcription of acid-expressed genes. Two motifs were found to be significantly similar to the STE12 binding site. The STE like transcription factor *pp-1* of *Neurospora crassa* exhibited significant circadian RNAPII rhythmicity in this dataset. Compellingly, it is a downstream target of the *mak-2* map kinase [115], which also exhibited circadian rhythmicity in this dataset. While motif 5 in group II showed significant similarity to MCM1, only the DNA replication licensing factor MCM3 exhibited significant circadian oscillations. Importantly, none of the TFs showed ultradian rhythmicity.
Overall, the discriminative motif overrepresentation analysis yielded a set of biologically plausible candidate transcription factors constituting a good starting point to study ultradian transcriptional regulation in *Neurospora crassa*.

4.6 Discussion and conclusions

4.6.1 Temporal relationships of RNA polymerase II occupancy and mRNA accumulation

This chapter involved a genome-wide combined Ser2P-RNAPII ChIP-seq measurement of RNAPII bound to genes, and RNA-seq measurements of the corresponding mRNA abundance in a *Neurospora crassa* liquid culture. Constant darkness and a sampling frequency of 2 h over a total of 22 h provided a detailed picture of the temporal organisation of circadian transcription. Classification of circadian oscillators by harmonic regression

yielded 993 genes with RNAPII rhythms and 1178 genes with mRNA rhythms. According to this classification, 262 genes exhibited circadian rhythms in both, RNAPII and mRNA, termed R-R set. Significant rhythms in only RNAPII were found in 682 cases (R-AR set), and 916 genes featured only significant mRNA rhythms (AR-R set), leaving 6372 of the captured genes as fully arhythmic. However, visual inspection of heatmap representations of the AR-AR set, the mRNA profiles of the R-AR set, and the RNAPII profiles of the AR-R set suggested the widespread subtle circadian rhythmicity, particularly in the mRNA dataset. Detection of these oscillations was frequently prevented by the low amplitude. In the RNAPII dataset, dampening of oscillatory amplitudes might be related to the integration of the read signal over the proximal promoter (500 bp) together with the entire gene. This yields a combined representation of rhythmic transcription initiation, elongation and termination. A second signal integrating only the polymerase signal at the transcriptional start site might prove more sensitive to rhythmicity. Importantly, the increased variability, here defined as the standard deviation of each gene profile, in arhythmic RNAPII profiles correlated with more frequent circadian mRNA rhythmicity. In a related study Menet *et al.*[135] encountered similarly widespread sublime circadian oscillations in nascent RNA and mRNA levels. Furthermore, they found that fluctuations in non-rhythmic nascent RNA levels contribute to the generation of rhythmic mRNA profiles. These weak transcriptional patterns might signify a specific group of genes which employ a combination of weak transcriptional regulation with a second mode of regulation to achieve a stable circadian mRNA abundance. In this context, a range of post-transcriptional processes and degradation have been recently highlighted by Lück *et al.* as a means to achieve circadian rhythmicity [126]. Due to the low-amplitude oscillations in the presented dataset, a precise prediction of genes featuring such rhythmic post-transcriptional processing was not attempted. The case of rhythmic RNAPII with arhythmic mRNA can be explained by long mRNA half-life times dampening changes in transcription as studied in detail for the mouse liver model by Le Martelot *et al.* [114]. However, it can not be excluded that the widespread low-amplitude oscillation is due to experimental normalisation to the rRNA content, which is commonly assumed to be constant but was observed to oscillate in circadian fashion in *Synechocystis* sp. PCC 6803 (see section 2.3).

A range of genes was found to be highly over-expressed in the first time point compared to the remaining time series. A critical example is the *frequency* gene (see Fig. 4.3) with a peak phase of CT 0-6 which showed a two-fold larger RNAPII and mRNA signal in the first sample (CT 12) compared to the expected peak transcription at CT 5.3 [13]. This strong induction is likely due to a light flash which was applied prior to the sampling start to synchronise the culture. This feature is shared by other light-induced immediate early genes such as *al-1* and *vvd* [67]. As a consequence, the oscillation detection procedure was repeated without the first sample, and significantly oscillating genes were added to the results of the full dataset. A second interesting observation was the transient induction or repression specifically at the second sample (CT 14) amongst the genes of the AR-R and the AR-AR set (Fig. 4.4, C-F). This behaviour might represent a second delayed reaction to the synchronising light flash. However, additional steps to moderate the effect of this transient to the oscillation detection were not taken with respect of the

integrity of the presented dataset.
Comparison of the peak phases in the R-R gene set showed that RNAPII profiles lagged the corresponding mRNA phases by ≈ 1 h. Even when including phase estimates of not significantly oscillating genes, mRNA profiles preceeded RNAPII by ≈ 0.7 h. This is contradictory to the expectation, since RNA polymerase binding to the gene promoter precedes nascent RNA synthesis and processing into mRNA. Accordingly, Le Martelot *et al.* reported average delays of ≈ 1 h between RNAPII binding and mRNA abundance in mouse liver cells [114]. The larger number of steps in the Ser2P-RNAPII ChIP-seq experiment might introduce a delay in the resulting profiles. Measurements of transcriptional activity and mRNA abundance of a small number of known circadian genes could serve as reference to correct for this delay. Expression peak phases clustered around dawn and dusk, which concurs with a recent report of circadian mRNA rhythmicity in *Neurospora crassa* [80]. According observations of circadian transcription in various mouse tissues led to the description of dawn and dusk as "transcriptional rush hours" [262].
Neurospora crassa exhibited fundamentally different transcriptional programs during day and night. Functional enrichment analysis revealed transcription of catabolism related genes in the morning, which are separated from anabolism related genes in the evening. As expected, many light response, salt stress, and cell defence related genes peaked during the morning. These observations agree very well with previous results reported by Hurley *et al.* [80].

4.6.2 Target genes of the central circadian transcription factors WCC and CSP1

Genome-wide binding studies based on ChIP-seq were presented for the two circadian clock output transcription factors WCC and CSP1. Target gene sets were predicted using the published genomic peak locations and calculating the transcription factor association score (TFAS) according to Ouyang *et al.* [155]. Comparing the TFAS statistics and applying hypergeometric tests between circadian and non-rhythmic genes in the presented dataset suggested that WCC target genes (from [201]) are not more often than expected by chance circadian oscillators on RNAPII and mRNA level. In contrast, circadian oscillators are overrepresented amongst CSP1 target genes predicted from the experiment by Sancar *et al.* [188].
The observed RNAPII phase distributions for the predicted target gene sets of WCC and CSP1 agreed widely with previous reports (Fig. 4.6 A). WCC targets showed a clear preference for peak phases shortly after dawn. Additionally, numerous genes were found at all possible phase bins forming an even background. This agrees with previous observations of diverse transcriptional phases of WCC target genes [80]. BMAL1-CLOCK, the mammalian WCC counterpart, was shown to open its surrounding chromatin to allow other TFs to bind subsequently [134]. Indices for this "pioneer-like" operating principle were recently found for WCC [239]. Importantly, the background distribution of all genes with circadian RNAPII profiles also showed a small peak after dawn which presumably contributed to the dawn phase cluster amongst WCC targets. In contrast, CSP1 target

phases were separated mainly into two groups, one centred around dusk as demonstrated earlier [188]. The second group of genes exhibited phases around dawn. While genes often receive regulatory input from more than one TF, circadian anti-phasic profiles can not be explained by combined circadian TF input according to the employed model. Furthermore, CSP1 is characterised as transcriptional repressor, so that its dawn-specific induction can not directly relayed to its targets. The dawn phase peak thus likely represents false positive CSP1 target classifications, which follow the background phase distribution. The observed dawn phase preference of the ≈ 50 WCC and CSP1 target genes agrees well with the prediction of the theoretical model, justifying the subsequent analysis of genes with ultradian oscillations.
To focus on transcriptionally regulated genes, only members of the R-R gene set were considered in the comparison of mRNA phases between target sets (Fig. 4.6 B). As before, CSP1 and RCO1 target sets exhibited pronounced phase clusters before dawn and dusk, similar to those of the background distribution (Fig. 4.6 D, blue). However, the extent of dawn-specific phases in the WCC target set was drastically reduced. Genes with combined WCC and CSP1 input werere divided equally between dawn and dusk phase. It can be concluded that the phases of RNAPII occupancy correlate better to the activity patterns of the circadian TFs. In contrast, the corresponding mRNA rhythms exhibited only weak phase specificity. It must be expected that including the additional steps of transcription and RNA processing complicates the correlation to TF binding, as compared to RNAPII binding as primary readout. Some mRNA rhythms might remain hidden in the undetected low-amplitude oscillations. Furthermore, post-transcriptional modification can modify the resulting mRNA phases. Overall, the presented RNAPII phase dataset is more informative to study the effects of circadian transcriptional regulation.

4.6.3 Ultradian transcription in *Neurospora crassa*

Mathematical modelling predicts doubling of transcription frequency, *i.e.*ultradian rhythms, for genes which are regulated by two circadian TFs with appropriate phase relationship. Specifically, two circadian activators with distinct regulatory sites in the target gene promoter must exhibit approximately anti-phasic abundances to achieve ultradian rhythmicity. A circadian activator and a circadian repressor occurring approximately in phase achieve the same result. While the results, described above, did not support the hypothesis that ultradian rhythms can be generated by WCC and CSP1, 24 additional genes have been proposed to be secondary circadian transcription factors [201]. In the presented datasets, nine of these genes exhibited circadian rhythms in either RNAPII or mRNA profiles. Subsequent analyses therefore focussed on genes with ultradian transcriptional rhythms.
Genes with ultradian rhythmicity were detected via harmonic regression, choosing the period parameter to be 11 h. This yielded 357 genes with significant RNAPII oscillations and 185 genes with mRNA oscillations. The phase distributions were bimodal with two anti-phasic clusters of comparable size. Surprisingly, only six genes showed oscillations in RNAPII and mRNA abundance. Compellingly, two of these genes (*NCU02044*, *NCU01297*) receive input from the circadian clock via the map kinase pathway and thus

might mediate circadian control of growth and development in *Neurospora crassa*. Three genes in this set are important for different steps in gene expression, namely protein import and RNA export from the nucleus (*NCU02680*), chromatin remodelling (*NCU03652*), and splicing (*NCU11251*). Several factors might explain the small overlap between the observed ultradian RNAPII and mRNA rhythmicity. The detrimental influence of long mRNA half-life relative to the transcriptional oscillation frequency, already discussed for circadian genes, is greater for fast ultradian rhythms as compared to circadian rhythms. Similarly, the integration of the ChIP-seq signal over the promoter and the gene should dampen faster transcriptional rhythms more compared to circadian rhythms. Additionally, post-transcriptional regulatory mechanisms might be also responsible for ultradian rhythms, either in a combinatorial fashion of two circadian processes with appropriate phase relationship, or by a single inherently ultradian process.

The genes with significant ultradian RNAPII rhythms were investigated in detail. These were separated into four subgroups according to phase and expression profile shape, as shown in Figure 4.7 A. Functional enrichment analysis highlighted several genes encoding 60S ribosomal subunits in group III, peaking at $CT \approx 7$ and $CT \approx 18$. This agrees well with the observed overrepresentation of genes for protein synthesis and processing during dawn and also dusk amongst circadian genes (see Table 3). It appears that ultradian transcriptional regulation in *Neurospora crassa* is more frequently found amongst protein synthesis-related genes.

The overlap between ultradian rhythms in transcription and mRNA abundances was limited with only six genes. One particularly interesting ultradian gene is the SWI2/SNF2 family member *NCU03652* which is involved in transcription regulation, chromatin remodelling, and DNA repair, including the disruption of protein-protein and protein-DNA interactions. This gene peaks during dawn and dusk. Furthermore, its binding to WCC allows for targeted chromatin remodelling in the surrounding of the corresponding binding sites. Remodelling then prepares the binding of other transcription factors, as found for the *frq* locus [239]. This suggests distinct chromatin remodelling patterns via directed placement by WCC and undirected binding or directed binding by another TF in the evening. It must be noted that *NCU03652* is not part of the SWI/SNF complex as described in Wang *et al.* [239].

The number of ultradian RNAPII profiles was considerable. As expected in the light of the previous results, no evidence of overrepresentation of WCC or CSP1 target genes was found. Overrepresented sequence motifs in the promoter sequences were then determined for all ultradian genes and separately for each subgroup. Discriminative motif overrepresentation analysis yielded disjunct sets of motifs for each subgroup. A comparison of these motifs with the set of 117 known TF binding models in the fungus *Saccharomyces cerevisiae* highlighted a range of TFs with homologs in *Neurospora crassa*. Most notably, two motifs were similar to the binding site of STE. Its homologous STE like transcription factor *pp-1* exhibited circadian RNAPII rhythms in this dataset. Furthermore, one motif was found to be similar to MCM1, a component of the mini chromosome maintenance complex, for which another component, MCM3, also exhibited circadian rhythms.

Considering that nine of the 24 putative secondary circadian TFs in *Neurospora crassa*

exhibited circadian rhythmicity in the presented dataset, it is possible that each of the ultradian gene subgroups are regulated by two specific TFs. However, further research is required to study the properties of ultradian gene promoters in more detail with respect to sequence motif overrepresentation, motif localisation, sequence composition and conservation, occurrences and localisation of known regulatory motifs such as the proximal and distal light responsive elements or the clock box, and expression patterns.

5 Summary and outlook

The circadian clock has sparked interest amongst researchers for a long time. The concept of an internal rhythm was first described as early as 1729 by Jean-Jacques d'Ortous de Mairan, who observed continued plant leaf movements under constant illumination. According to current knowledge, gene expression is the interface between the circadian core clock mechanism and other biological processes. Hence, a comprehensive assessment of the circadian expression program is crucial to understand the physiological impact of the clock on the corresponding organism. A wide range of circadian clock model organisms have been studied with respect to the circadian transcriptome, such as mammals, cyanobacteria, plants, and fungi. A central hypothesis is the circadian control of metabolic gene expression. In plants and cyanobacteria, the beginning and ceasing of photosynthetic activity are accompanied with extensive metabolic changes and its anticipation should allow for better preparation. A range of high throughput methods have been used to identify clock controlled genes. While microarrays are the most commonly applied technique, they are increasingly supplemented by current techniques like RNA-Seq and ChIP-seq.

Circadian transcriptional program in cyanobacteria probed by microarrays
The first part of this thesis (Chapter 2) provides a detailed analysis of a diurnal microarray expression dataset of *Synechocystis* sp. PCC 6803, which is then embedded in a collection of diurnal transcription datasets of six cyanobacterial strains.
Previously undescribed diurnal oscillations in the 16S and 23S rRNA content led to a diurnally varying fraction of mRNA in the total RNA content of *Synechocystis* sp. PCC 6803. Application of a constant amount of RNA extract to the microarrays of the presented time series led to fundamental low-amplitude oscillations. Standard multi-array normalisation techniques drastically altered the observed transcriptional phases and were thus unsuited. Instead, multi-array normalisation based on the set of least oscillating genes yielded a description of the diurnal transcriptional landscape.
Cyanobacteria are a very diverse clade featuring strains with drastically different lifestyles. However, photoautotrophic organisms share the fundamental challenge to adjust their metabolism to photic and aphotic phases according to the natural succession of day and night. A comparative study of a range of microarray datasets describing the circadian and diurnal transcription in the most popular cyanobacterial strains was performed. The presence of a fundamental transcriptional program, which is shared amongst the cyanobacterial clade, was investigated. Comparing the number of genes with diurnal expression highlighted the importance of standardised methods for oscillation detection, allowing for a better comparability.
A comprehensive homology prediction was used to detect consistently diurnally expressed

genes across six different cyanobacterial strains in a total of nine experiments. This core diurnal genome contained 95 genes. As expected, this set was dominated by genes encoding for proteins with metabolic functions, such as the photosystem or ribosomal proteins. Importantly, pairwise comparison of the phase relationships between homologous diurnal genes suggested that diurnal gene expression is strain specific. Cyanobacterial strains have thus adapted their diurnal transcriptional repertoire to the individual lifestyle, while central metabolic genes were the only exceptions to this differentiation.

Sequence periodicity patterns and circadian expression in cyanobacteria

Cyanobacterial chromosomes have been observed to undergo rhythmical structural changes under LD conditions [252]. The work of Vijayan *et al.* demonstrated that the chromosomal compaction via supercoiling follows circadian rhythmicity under constant light and regulates gene transcription [237]. Several genes in *E. coli* have been found to be expressed in a supercoiling-dependent fashion [87, 164]. Furthermore, there is evidence suggesting a connection between supercoiling-sensitive expression and sequence periodicity patterns in the corresponding genes [69, 90, 104, 105, 151, 156]. Specifically, periodic AT-dinucleotides in phase with the local DNA pitch can induce curvature in the DNA backbone. Muskhelishvili and Travers [142, 221] have combined these observations in a general model, in which the gene expression landscape is regulated by the metabolic state of the cell via modification of the chromosomal supercoiling state. Integral part of this model are nucleoid associated proteins like HU, which stabilise DNA structures after binding to curved and AT-rich DNA [40]. Indeed, HU homologous genes can be found in a range of cyanobacteria. Compellingly, these HU homologs exhibited circadian or diurnal expression patterns in almost all of the available datasets of six strains. As an extension of this model, a metabolic oscillator can be envisioned. In this oscillator, ATP-dependent topoisomerase enzymes facilitate the energy produced via photosynthesis over the day to introduce supercoiling. Energy-independent topoisomerase could then relax the supercoiling during the night. The second part of this thesis (Chapter 3) examined whether the localisation of dinucleotide periodicity patterns in cyanobacterial genomes supports this model.

Dinucleotides consisting of combinations of A and T exhibited the strongest genome-wide periodicity in the majority of the considered genomes. AT2 periodicity was found to be encoded mainly in the first and third codon position of the coding sequence. The comparison of AT2 periodicity strength and diurnal gene expression in *Synechocystis* sp. PCC 6803 did not reveal notable correlation. Individual genes exhibited AT2 periodicity in a range of different periods. Interestingly, transposons exhibited strong 11 bp periodicity in significantly bent sections of the gene in agreement with the idea of a structural role of periodic AT2 dinucleotides. However, systematic comparison of genome-wide AT2 periodicity between a wide range of cyanobacterial strains revealed a correlation to the number of chromosomal copies. These results suggest that high AT2 periodicity facilitates chromosomal compaction, specifically in cyanobacterial strains with large chromosome copy numbers.

Ultradian expression patterns through circadian transcription factors in *Neurospora crassa* In fungi and animals alike, the core clock ("circadian oscillator") consists of a rhythmically present heterodimeric TF which induces the expression of clock controlled genes ("circadian system"). In *Neurospora crassa*, this is the morning-specific transcriptional activator WCC. Conceptually, this circadian TF can give rise to two groups of target genes: one with similar (morning) phase acting as transcriptional activator, and one with the opposite (evening) phase acting as repressor. Instead of a direct repressor function of WCC, *Neurospora crassa* realises evening phases via the secondary circadian transcriptional repressor CSP1, which is induced by WCC. However, genome-wide measurements of circadian transcriptional activity routinely yield a wide range of intermediate phases which can not directly be explained. Moreover, many species also exhibit ultradian transcriptional rhythms. The combination of two circadian transcriptional regulators with appropriate phase relationship can explain the phase modifications and ultradian rhythms in a mathematical model [249]. The third part of this thesis (Chapter 4) presents a combined dataset of RNAPII binding and corresponding mRNA abundance in *Neurospora crassa* liquid cultures. The extent and biological relevance of circadian and ultradian oscillations was determined. Furthermore, the impact of combined WCC and CSP1 binding on the resulting transcriptional phase and the generation of ultradian transcription patterns was investigated.

The RNAPII binding phases featured a clear morning phase cluster and a range of intermediate phases. In contrast, mRNA abundance phases clustered around morning and evening with few intermediate phases. Functional annotation analysis of genes with similar phases indicated a clear separation of cellular processes between day and night. As reported previously, *Neurospora crassa* gene expression focussed on catabolic activity during the day in order to generate metabolic intermediaries [80]. These resources appear to be invested into the biosynthesis of complex cellular components and growth during the following night.

In addition to fully rhythmic genes, a significant number of genes exhibited circadian oscillations only in RNAPII or mRNA profiles. The analysis suggested that elevated transcriptional variability in arhythmic RNAPII profiles, including only visually detected low-amplitude circadian oscillations, systematically promotes rhythmicity in mRNA abundance. While such a low-amplitude oscillatory trend has been observed previously in mouse liver when comparing nascent RNA with mRNA, it was interpreted as transcriptional noise [135]. The study of Hurley *et al.* tested a subset of genes with rhythmic mRNA abundance and found widespread constant transcription, which was interpreted as hint towards extensive post-transcriptional regulation in *Neurospora crassa* [80]. Considering the diverse potential complications of circadian expression measurements (as described in Chapter 2 for microarrays), all assumptions on the biology of *Neurospora crassa* should be carefully re-examined to exclude an artifactual nature of this low-amplitude oscillatory trend. Alternatively, these results would suggest the involvement of a second wide-spread post-transcriptional process which relies on the amplification of the weak transcriptional oscillations to achieve stable mRNA rhythms.

Numerous genes exhibited ultradian transcriptional and mRNA abundance profiles. However, analyses of ChIP-seq experiments suggested that ultradian transcriptional

rhythms are not directly generated by the circadian output TFs WCC and CSP1. This points towards alternative secondary circadian transcription factors. The study of Smith *et al.* [201] described 24 TFs, which were predicted to receive transcriptional regulatory input from WCC. De novo motif discovery was performed in the promoter of genes with ultradian transcriptional profiles. Subsequent comparison of the overrepresented motifs to known TF binding motifs of the fungus *Saccharomyces cerevisiae* revealed two TFs with circadian transcription, *pp-1* and *mcm3*, which might be involved in the transcriptional regulation of ultradian genes in *Neurospora crassa*. Due to its various roles in clock driven transcription in *Neurospora crassa*, WCC might not be optimally suited to implement a precise combinatorial regulation. It was shown recently that BMAL1-CLOCK, the mammalian counterpart of WCC, can act as pioneer TF facilitating rhythmic opening of chromatin and thereby allowing other TFs to bind adjacent to BMAL1-CLOCK [134]. The observed heterogeneity of WCC target gene phases, also reported by Hurley *et al.* [80], suggests a similar role for WCC not only as phase-determining activator but as pioneer TF. Indeed, the SWI/SNF complex is recruited by WC-1 to remodel and loop chromatin at the *frq* gene promoter, thereby activating *frq* expression to initiate the circadian cycle [239]. This chromatin remodelling and potential subsequent binding of additional TFs might complicate the achievement of a precise activation phase, which in combination with a second TF is a prerequisite for ultradian rhythmicity.
While a significant number of genes displayed ultradian rhythms in transcription or in mRNA abundance, only a small set of six genes featured both. Most notably, the gene *NCU03652*, containing SNF2 family helicase / ATP-ase domains, peaked during dusk and dawn. As mentioned above, the SWI/SNF complex is recruited by WC-1 to remodel chromatin at target gene promoters. However, the comparative study of *Saccharomyces cerevisiae* did not recover *NCU03652* as member of the *Neurospora crassa* SWI/SNF complex [239]. It would be an attractive hypothesis, that WC-1 mediates the pioneer TF functionality of WCC in the morning by recruiting the SWI/SNF complex to achieve chromatin remodelling around its targets. In absence of WCC in the evening, the chromatin remodelling by the SWI/SNF complex could then focus on alternative genomic regions. Interestingly, a comparative study of *Saccharomyces cerevisiae* and *Neurospora crassa* presents *NCU03652* as possible member of the error-free post-replication repair pathway [92]. In this context, ultradian transcription is surprising since in *Neurospora crassa* the cell cycle is coupled to the circadian clock [75].

5.1 Outlook

This thesis comprises a comprehensive view on diurnal and circadian transcriptional patterns in cyanobacteria and *Neurospora crassa*. A representative collection of nine time series microarray dataset of six popular cyanobacterial strains was analysed, revealing a set of consistently diurnally expressed genes or core diurnal genome. While this is not a final definition of the cyanobacterial core diurnal genome, it suggests a short list of metabolic processes that rely on diurnal and probably circadian transcriptional control. A detailed analysis of the localisation of these genes within the metabolic network

and pathways might provide further insights into how the metabolism is regulated by the circadian clock and whether there is a feedback from the metabolic state back to the clock. Specifically, it is conceivable that core diurnal genes predominantly serve housekeeping functions and critical metabolic functions, whereas strain specific diurnal genes relate to environmental features of the corresponding strain. Further work is required to elucidate the functions of genes amongst the core diurnal genome which are annotated as "hypothetical protein". Most importantly, homology information should be used to assemble existing knowledge from other cyanobacterial strains. Classical knock-out experiments targeting these hypothetical protein genes could yield further information towards their function.

While experimental work demonstrated the existence of diurnal oscillations of 16S and 23S rRNA in *Synechocystis* sp. PCC 6803, its functional relevance is still unclear. The hypothesis that these rRNAs might be degraded in the morning to jump-start the Calvin cycle could be tested experimentally using metabolomics techniques. Application of RNA-seq for the measurement of the cyanobacterial circadian transcriptomes would be advantageous due to the higher dynamic range of this method. Crucially, a standardised culture system should be established in the cyanobacteria community in order to ensure better comparability of the obtained results. This would facilitate the systematical examination of the circadian oscillator and the circadian system under different culture conditions, *e.g.* light regime, temperature, and medium. It is conceivable that other factors besides the temperature, which inactivates the circadian oscillator in *Synechococcus elongatus* [257], can trigger the conditionality of the clock.

The questions whether the cyanobacterial HU homolog is functional and whether it takes a role in the circadian nucleoid compaction could be addressed with a time resolved ChIP-seq experiment. This could be complemented by an assessment of the chromosomal structure by chromosomal conformation capture [43], which would allow to study the co-localisation of AT2-periodic regions with apical DNA loops and HU binding sites.

Reprocessing of the presented *Neurospora crassa* RNAPII dataset by separately integrating over the gene promoter and the gene body would enhance our understanding of circadian transcription. The comparison of a promoter-specific with a gene body-specific signal enables a quantitative answer to whether circadian transcription is also regulated by RNAPII binding, as found in mouse liver [114], and not by the transition from paused to productive elongation. Furthermore, a promoter-specific signal would provide a temporally better resolved picture of the transcriptional regulation, independent from gene-specific properties such as transcriptional efficiency. Future work should focus on the properties of ultradian gene promoters with respect to sequence motifs, localisation of putative and known motifs such as the proximal and distal light responsive element or the clock box, sequence composition and conservation, and on the expression patterns explored in previous studies.

Appendix A

1 Methods used in microarray time series

1.1 The *Synechocystis* sp. PCC 6803 time series expression dataset

Synechocystis sp. strain PCC 6803 was grown in BG11-medium [181] at 30 °C under continuous illumination with white light of 120 µmol of photons $m^{-2}s^{-1}$ and a continuous stream of air. The optical density of the culture was monitored by measuring the absorbance at 750 nm. Cultures were synchronised with three cycles of light/dark 12h:12h prior sampling. Aliquots were taken at OD750 0.5. Over a 24 h time course, 6 samples for RNA isolation were taken at the following time points: 30 minutes before and after light is switched off, (sample 1 - CT 11.5 and sample 2 - CT 12.5), 30 minutes before midnight (sample 3 - CT 17.5), 30 minutes before and after light onset (sample 4 - CT 23.5 and sample 5 - CT 0.5) and 30 minutes before noon (sample 6 - CT 5.5). Cells were filtered rapidly through Supor 0.45µm membrane filters (PALL), immediately stowed with TRIzol reagent (Invitrogen) and frozen in liquid nitrogen. Total RNA samples stored at −20 °C were transferred directly to a 65 °C waterbath for 5 minutes, mixed with 0.2 ml chloroform per ml of TRIzol and incubated for 15 minutes. The dissolving of the membrane and lyses of the cells were supported by vortexing. Centrifugation at maximum speed for 10 min at 4 °C separated the phases. The RNA in the supernatant was precipitated by adding 0.5 ml of isopropanol per ml TRIzol used in the initial homogenisation. Two replicates were prepared from two synchronously growing cultures. The microarray design and hybridisation procedure have been described previously [56]. The custom made Agilent single channel expression microarray holds probe sets for all annotated genes from the chromosome (NC_000911) as well as the seven plasmids. The detailed description of the employed microarray is deposited at gene expression omnibus (GEO) under the series identifiers GSE16162 and GSE14410. The extracted RNA was labeled directly for microarray hybridisation to avoid labelling artifacts from reverse transcription and second strand synthesis during cDNA synthesis. The same amount of 1.5μg RNA was applied for every array, *i.e.* time point. The spot intensities were extracted with the 'Agilent Feature Extraction Software 10.5.1.1' using the Protocol GE1_105_Dec08. No background correction was performed. Probe summarisation yields expression values for 8907 mRNAs, of which 3242 can be mapped onto protein coding genes located on the chromosome and 105 located on plasmid pSYSA. Only those genes were selected for further analysis.

1.2 Data transformation

The brightness of spots in a microarray experiment, from which the expression strength is derived, depends not only on the number of mRNAs in the sample, which is applied to the array chip. Large differences in hybridisation energy and experimental effects like cross hybridisation lead to expression values, which span several orders of magnitude and of which only relative changes for one probe set between the conditions can be interpreted. By the use of different transformations, it is common to bring raw expression data into the same order of magnitude. To allow for comparability, raw data were included in every analysis step.

Log2 mean ratio

The `l2m` mean ratio is defined as

$$x' = log_2 \frac{x}{\bar{x}} \quad ,$$

where x, $\bar{x}$, and x' denote the original time series, the average expression over the genes entire expression profile, and the transformed time series, respectively.

Standardisation (Z transformation)

The standardisation is defined as

$$x' = \frac{(x - \bar{x})}{\sigma_x} \quad ,$$

where σ_x denotes the standard deviation of the genes expression profile from its average, which is calculated as

$$\sigma_x = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2}$$

for an expression profile x of length N.

Discrete Fourier transformation

A series of measurements $x = \{x_0, ..., x_{N-1}\}$, acquired at times $\{t_0, ..., t_{N-1}\}$, can be approximated as a set of sine-functions with different frequency and amplitude. This transformation into frequency-space is done by applying the Discrete Fourier Transform (DFT) to each gene's time series

$$X_k = \sum_{n=0}^{N-1} x_n e^{-2\pi i \frac{kn}{N}} \quad , \quad k = \{0, \ldots, N-1\} \quad ,$$

where X is a vector of complex numbers representing the decomposition. Each component X_k represents a sine with period $P_k = (t_{N-1} - t_0)/k$ where X_0 represents the non-oscillating component or an offset from 0 of the time series. For each component X_k the amplitude A_k and the phase angle ϕ_k can be calculated as $A_k = |X_k|/N$ and $\phi_k = tan^{-1}(Im(X_k)/Re(X_k))$. Since the obtained spectrum is symmetrical relative to $k = N/2$, it can be restricted to $0 < k < N/2$ (in this case 0 to 6) without loss of any information. It must be noted that the computed phase angles ϕ_k provide a distorted measure of the diurnal expression time due to the non-equidistant sampling. However, the phase angles provide an excellent means to obtain a temporal order of oscillating expression patterns.
To be able to cluster frequency spectra, the uninformative non-oscillating component X_0 and the highest frequency component X_6 were discarded creating a series of values out of the 5 real and imaginary parts of the remaining frequency spectrum for every gene. This component omission can be interpreted as subtracting the mean for each gene's time series. For the remaining components X_k, the amplitude is scaled to emphasise the shape of the expression pattern instead of the absolute amplitude, which is less informative for microarray data. Therefore, the scaled amplitude a_k is the amplitude at component k divided by the mean of amplitudes at all other non-zero components, $a_k = A_k/\bar{A}_{i\neq\{0,k\}}$.

1.3 Fourier transform-based detection of periodic expression profiles

As proposed previously, a permutation-based method is used to detect diurnal periodic expression profiles [120]. As diurnal periodicity is reflected in a large magnitude of the corresponding Fourier component X_k, its significance can be assessed by the probability p_{osc} to observe X_k in a random permutation of the original time series. Therefore, the the Fourier spectra of 100000 random permutations of each time series were calculated followed by the empiric relative probability for each X_k to observe a Fourier coefficient equal or larger in a random permutation.
It must be emphasised that the Fourier transform uses a sine function as underlying model which in case of a sinusoidal expression profiles leads to a distinct peak in X at the corresponding frequency k. For periodic signals with non-sinusoidal shape, *e.g.* spike-shaped, the magnitude of the corresponding frequency component is distributed across the harmonic and neighbouring frequency components. This hampers the detection of low-amplitude periodic non-sinusoidal profiles in comparison with sinusoidal profiles, since the lower magnitude of X_k receives a higher probability in the permutation background model.

1.4 Data normalisation

Strategies for the compensation of experimental variations in multi-chip experiments are generally considered necessary. Basis for such approaches are assumptions of similarity between different arrays in the same experiment.
The quantile-normalisation approach by Bolstad *et al.* [25] assumes that the real distribution between the arrays is identical and only a small number of genes show differential

expression due to the experiment. The array-wide normalisation was performed using the R-implementation in package `limma` [202] (`normalizeBetweenArrays` with method quantile).
Median polishing [229] is a classical method in exploratory data analysis. It is used within the RMA and GCRMA preprocessing protocols to summarise the probe sets. In this study, it is used to remove differences in the total median between individual arrays. We, thereby, illustrate the relaxation of the assumption of similar distribution shape, which is made in quantile normalisation, while maintaining the assumption that the majority of genes are not differentially expressed.
With the LOESS normalisation [260], another non-microarray specific normalisation method finds wide acceptance. In this method, the observation of an expression amplitude-dependent non-linear relationship between multiple microarrays is accounted for using a polynomial correction function instead of a linear one for the equalisation of two arrays. For the extension of this pairwise normalisation, the gene-wise mean expression over all samples can be used as reference array for each individual sample array. In the work of Bolstad *et al.* [25], the cyclical application of the LOESS normalisation was included, in the following referred to as cLOESS. The implementation in the R-package `limma` within the method `normalizeCyclicLoess` was employed with default settings.
The least oscillatory set (LOS) normalisation is proposed. The method is related to the least variant set normalisation (LVS) [28], selecting a subset of expression profiles for the fitting of a LOESS polynomial.
While LVS attempts to define a set of housekeeping genes by finding profiles with minimal array-to-array variation (after partitioning the observed variation into array-to-array variation, within-probeset variation and residual variation), LOS follows a more intuitive approach. Here, housekeeping genes are defined as the set, which exhibits the least pronounced diurnal oscillations (measured by oscillatory p-value p_{osc}). Defining the lower cutoff $p_{osc} > 0.7$ and considering all transcripts on the chip yields a LOS set of 1173 expression profiles. The mean expression for each of these LOS profiles is used to fit a LOESS normalisation curve to each individual array, which is then used to perform the normalisation. For the presented dataset, LOS normalisation leads to the dampening of the spike at the first CT 17.5.

1.5 Clustering algorithms

From the plethora of clustering algorithms, which have been proposed for the clustering of expression data, seven diverse methods were selected covering different clustering principles.

K-means

The non-hierarchical K-means clustering algorithm is implemented in the R-function `Kmeans` (package: `amap`). In this function, 100 random starting sets of k cluster centers are used to run 1000 iterations of the Lloyd-Forgy algorithm [52] each. From the set of available distances measures, Euclidean distance and Spearman correlation coefficient

ρ were selected. In this case as in every following correlation coefficients have been transformed into a distance measure by:

$$\hat{\rho} = 1 - \rho$$

taking 1 minus the correlation coefficient.

Partitioning around medoids (PAM)

Similar to K-means, PAM is a non-hierarchical clustering algorithm that partitions the data by attempting to minimise the squared error of a distance measure [91]. In contrast to K-means PAM takes data points as cluster centers, which are then called exemplars or medoids. The R-implementation `pam` (package: `cluster`) was used with Euclidean and Spearman correlation distance.

Hclust

The bottom-up hierarchical cluster [47] analysis included in this study is implemented in the R-function `hclust` (package: `stats`). The clustering is based on a set of dissimilarities between the samples. Here, dissimilarities based on the Euclidean distance and the Spearman correlation coefficient were used with Ward's method [243].

Self-organizing maps (SOM)

The non-hierarchical Self Organising Map (SOM) approach represents multidimensional data in a low-dimensional topological map. The grid used here is one-dimensional and the number of grid points equals the number of clusters [70]. The implementation of SOM in the R-function `som` (package: `kohonen`) [246] is used. During the training phase the data are presented for 3000 times to the network.The learning rate alpha is set to start from 0.5 and decreases linearly to 0.05 over the 3000 repetitions. A rectangular network topology with 1 by k nodes was chosen.

Self-organising tree algorithm (SOTA)

The top-down approach called self-organising tree algorithm or SOTA was proposed as strategy for phylogenetic reconstruction [42]. It has also been used to cluster microarray gene expression data [70]. In a top-down fashion, SOTA produces a hierarchical binary tree structure by repeatedly training a neural network and splitting the most diverse neuron into two neurons of the new network. The R-implementation `clValid` (package: `clValid`) with default parameters [42] was used.

Mclust

A non-hierarchical model-based clustering approach was included using expectation maximisation initialised by hierarchical clustering for parametrised Gaussian mixture

models [261]. Each mixture component represents a cluster. The full set of 10 possible models is calculated for each number of clusters k and the model yielding the highest Bayesian information criterion (BIC) is selected. The R-implementation `Mclust` (package: `Mclust`) is employed with default parameters.

flowClust

`flowClust` [123] was chosen as second member of the family of model-based clustering methods. The main difference to Mclust is the usage of a multivariate t distribution as model for each cluster instead of a Gaussian distribution. The R-implementation `flowClust` (package: `flowClust`) was used with default parameters. The application of `flowClust` to standardised and unnormalised data prevented the convergence of the algorithm or lead to clusterings that include clusters of less than 10 genes. This suggests incompatibility of the algorithm to these transformations and justified the exclusion of these combinations from further analysis.

1.6 Clustering comparison

Adjusted Rand index

The Rand index [177] between two clusterings counts for all pairs in the dataset how often both are in the same cluster (a) or in different clusters (b) within both clusterings (agreement of clusterings). Also the number of disagreements in between all pairs is counted, *i.e.* for how many pairs both are in the same cluster in clustering 1, but not in clustering 2 (c) and vice versa (d). The counts are then combined to a score:

$$R = \frac{a+b}{a+b+c+d}$$

The adjusted Rand index furthermore accounts for similarities in the clusterings which are expected by chance. The adjusted Rand index values are of interval [0,1] where 1 is reached by maximally similar and 0 by maximally dissimilar clusterings. The R-implementation of the adjusted Rand index in function cluster.stats (package: `fpc`) was used.

Mutual information

The mutual information is defined as

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log \left(\frac{p(x,y)}{p_1(x) p_2(y)} \right) \quad ,$$

where $p(x,y)$ is the joint probability function for elements of the two clusterings $x \in X, y \in Y$ and $p_1(x), p_2(y)$ are the marginal probabilities for elements in the individual clusterings. The joint probability function is estimated by a contingency table whereas the marginal distributions are estimated by a histogram with each cluster being one bin. The

mutual information values range from 0 for maximally dissimilar clustering to a maximum of the entropy of one clustering when both are identical. Therefore, the maximum mutual information increases with the cluster number enabling for a larger entropy value in a clustering. The R-implementation of the mutual information in function mi.empirical (package: `entropy`) was used.

Normalised variation of information

The variation of information was proposed by Meila [132] is defined as follows:

$$VI(X,Y) = H(X) + H(Y) - 2I(X,Y)$$
$$nVI(X,Y) = \frac{VI(X,Y)}{H(X,Y)}$$

where $H(X), H(Y)$ are the entropies of the individual clusterings, $I(X,Y)$ is the already introduced mutual information. Instead of the variation of information $VI(X,Y)$, the normalised variation of information was employed in order to facilitate comparability between *e.g.* clusterings with different k. Values of the normalised Variation of Information are of interval [0,1] where 0 is reached by maximally similar and 1 by maximally dissimilar clusterings. The R-implementation of the VI in function cluster.stats (package: `fpc`) with subsequent normalisation was used.

1.7 Functional enrichment analysis

The functional enrichment analysis was performed using the gene annotations as provided by the Cyanobase database [147]. The overrepresentation of genes with a certain functional annotation was then computed with the R-library topGO [2], using the *classic* algorithm and the *Fisher* test statistic.

2 Diurnal expression comparison across cyanobacterial strains

2.1 Regression-based detection of periodic signals in microarray data

The described microarray experiments provide time series data *i.e.*gene-expression values in a well-defined order. To detect periodic signals within the large datasets, several different approaches have been put forward ranging from simple visual inspection [35], over fitting of sinusoidal functions [210], Fourier transform based methods [54, 251] and as described in Section 1.3 above, to statistical techniques such as Fourier-based periodic-normal mixture models [125] and the Jonckheere-Terpstra-Kendall (JTK) algorithm [78]. Many algorithms are based on the assumption that the diurnal expression profiles follow a sinusoidal shape, which can be described by the parameters phase and amplitude. Harmonic regression provides an efficient method to fit a sinusoidal function to a given expression profile, using the following formulation:

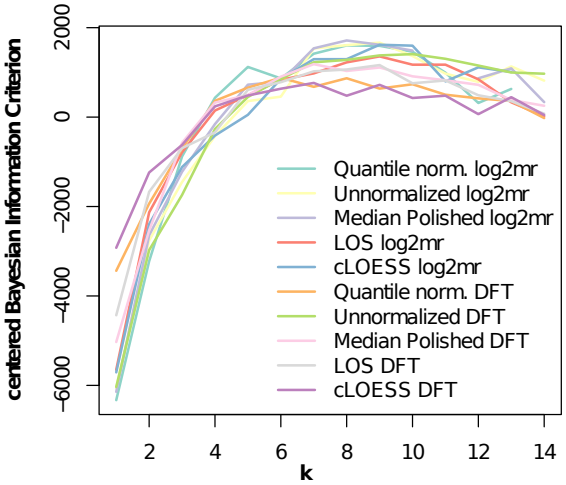


Fig. 1: **Bayesian information criterion versus number of clusters in the *Synechocystis*expression dataset.** Bayesian information criterion for clusterings of data with different normalisation and transformation obtained via flowClust.

$$x_t = \beta_1 \cos(2\pi\omega t) + \beta_2 \sin(2\pi\omega t) + t + \epsilon_t$$
$$\beta_1 = A\cos(\phi)$$
$$\beta_2 = -A\sin(\phi)$$

where A is the amplitude, ω the frequency, ϕ is the phase of the signal, and ϵ_t is the noise component. The addition of the sample time t allows for a linear trend in the data and is used optionally. The phase and amplitude are then obtained from the regression parameters:

$$A = \sqrt{\beta_1^2 + \beta_2^2}$$

$$\phi = \begin{cases} \arctan(\frac{-\beta_2}{\beta_1}) > 0 & \frac{1}{\omega} - \frac{\arctan(-\beta_2/\beta_1)}{2\pi\frac{1}{\omega}} \\ \arctan(\frac{-\beta_2}{\beta_1}) \leq 0 & \frac{|\arctan(-\beta_2/\beta_1)|}{2\pi\frac{1}{\omega}} \end{cases}$$

Here, negative values for ϕ correspond to earlier peak times and positive values for later peak times. Since microarray data provide simultaneously measured transcript levels for many thousand genes, spurious periodicity can be expected. To assess the significance of periodic signals, it is necessary to define what distribution of descriptive parameters is expected if the corresponding profile exhibits no true periodicity. This is equivalent to the definition of a null hypothesis of non-periodic expression. A straightforward approach to model non-periodic expression is based on the randomisation of the observed profile, as applied in the LOS normalisation procedure employed earlier in this chapter. By repeating the random reordering of the samples of a time series and obtaining the model parameters, an empirical background distribution can be obtained. Due to the potentially

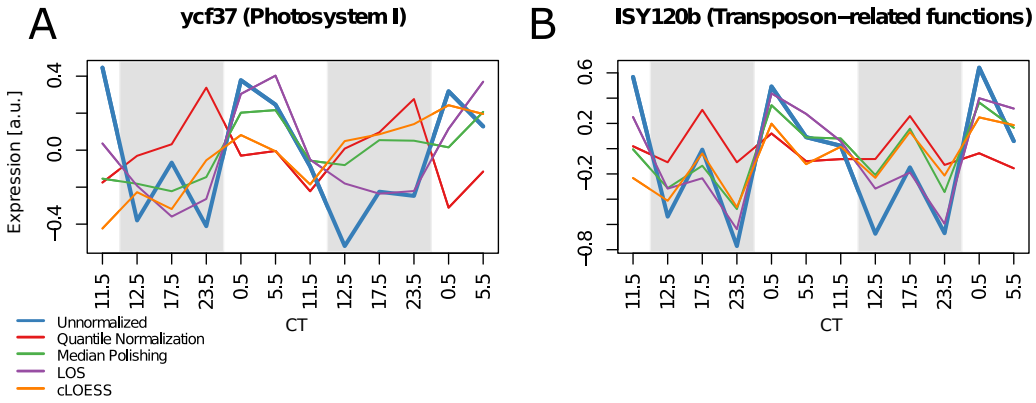


Fig. 2: **Phase alterations of diurnal expression profiles due to normalisation.** The expression profiles of four diurnally oscillating genes before and after normalisation with several methods (see legend for colour-method mapping) using `l2m`transformed data. The expression phase ϕ of the photosynthesis-related gene ycf37 (**A**) was shifted by $\approx 130°$ by quantile normalisation. The phase of transposon ISY120b (**B**) was phase shifted by $\approx 160°$ after quantile normalisation. The subjective night is shaded grey.

high number of required samples of the background distribution, this procedure can be prohibitively computationally expensive. Alternatively, non-periodic expression can be derived using a statistical model. In the case of harmonic regression, a simple F-test can be employed to test whether an expression profile fits better the alternative hypothesis of a sinusoidal function or the null hypothesis of a linear or constant function.

$$F = \frac{\frac{RSS_1 - RSS_2}{p_2 - p_1}}{\frac{RSS_2}{n - p_2}}$$

where n is the number of samples in the expression time series, RSS_1 and RSS_2 are the residual sum of squares of the linear model and the periodic model, respectively. p refers to the number of parameters in each model, in this case three for the harmonic model, two for the linear and one for the constant null model. The gene-wise p-values obtained from the F-statistic are then corrected for multiple testing according to Benjamini and Hochberg [14] to obtain the false discovery rate.

As demonstrated in Appendix C (4.2), linear detrending is detrimental for oscillatory time series which cover only one period (one apex and one valley) or less, due to alterations in the observed amplitude or phase. Accordingly, linear detrending is applied in all cases, except the *Microcystis aeruginosa* dataset of Straub *et al.*[209] which contains only 7 samples covering one day.

The residuals obtained form the harmonic regression are shown in Figure 3.

To establish similarity between the peak expression phases observed in two datasets, the circular correlation coefficient ρ_{ccc} was computed [86].

Strain	Dataset Abbrev./ Publication	Accession ID	Criterion	Measured Genes	Oscillating Genes (Pub.)	Oscillating (Harmonic Reg. $q < 0.05$)
Prochlorococcus marinus MED4	zinser09 [264]	Supplement	q < 0.1	3610	2043	2695
Synechocystis sp. PCC 6803	leh13 [117]	GSE45667	Fourier $p < 0.05$	3347	2142	2880
	beck14 [11]	GSE47482	Fourier $E < 0.01$	3347	904	2720
Synechococcus elongatus PCC 7942	vijayan09 [237]	GSE18902	Fourier Analysis	2716	1746	2138
	ito09 [83]	GSE14225	Correlation to Sine $p < 0.1, A > 0.1$	2514	800	1717
Microcystis aeruginosa PCC 7806	straub11 [209]	Supplement	Differential Expression to CT 0	5085	1344	0
Anabaena sp. PCC 7120	kushige13 [108]	Supplement	nitrogen-fixing condition, Correlation to Sine, $p < 0.05, A > 10^{-0.7}$	5336	283	1006
Cyanothece ATCC 51142	stoeckel08 [207]	E-TABM-337	Correlation Network	5048	1445	4291
	toepel08 [215]	E-TABM-386	Differential Expression	5048	1424	2721

Table 1: **Comparison of the number of genes with diurnal expression profiles in different cyanobacterial strains and datasets.** Comparison of the number of diurnally expressed genes reported ('Oscillating Genes (Pub.)') in the corresponding publication ('Dataset Abbrev./ Publication') amongst all genes covered by the microarray platform ('Measured Genes') using various methods of oscillation detection ('Criterion'). The last column provides the corresponding number of oscillating genes using harmonic regression.

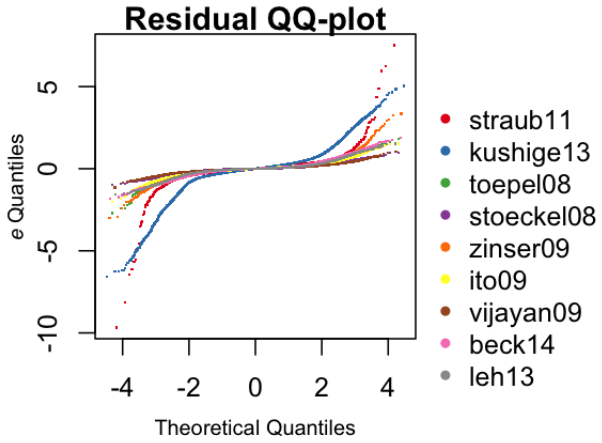


Fig. 3: **Harmonic regression residual QQ-plot for all cyanobacterial expression datasets**

$$\rho_{ccc} = \frac{\sum_{i=1}^{n} \sin(\alpha_i - \mu_c(\alpha)) \sin(\beta_i - \mu_c(\beta))}{\sqrt{\sum_{i=1}^{n} \sin(\alpha_i - \mu_c(\alpha))^2 \sum_{i=1}^{n} \sin(\beta_i - \mu_c(\beta))^2}}$$

$$\mu_c(x) = arctan(\sum_{i=1}^{n} \sin(x_i), \sum_{i=1}^{n} \cos(x_i))$$

Here, α and β refer to the peak expression phases in radians for i genes, and $\mu_c(x)$ denotes the circularised mean. The ρ_{ccc} is provided for all phase comparison plots shown in Fig. 4 and Fig. 2.8 in the main text.

2.2 Core diurnal gene set

Table 2: Diurnal core CLOGs across cyanobacterial strains excluding *Microcystis aeruginosa* PCC 7806. In every CLOG (row) at least one gene of each of the considered datasets (columns) exhibited diurnal expression.

Syn6803	Syc7942	ProMED4	Cyn51142	Ana7120	Function
sll0158	1085	PMM0584	cce_2248	all0713	1,4-alpha-glucan branching enzyme
sll1817	2210	PMM1536	cce_4038	all4192	30S ribosomal protein S11
sll1096	0887	PMM1511	cce_4091	all4340	30S ribosomal protein S12
ssl3437	2223	PMM1549	cce_4023	asl4206	30S ribosomal protein S17
ssl3432	2228	PMM1554	cce_4018	asl4211	30S ribosomal protein S19
sll1260	2530	PMM0753	cce_0705	all4792	30S ribosomal protein S2
sll1804	2226	PMM1552	cce_4020	all4209	30S ribosomal protein S3
sll1812	2216	PMM1542	cce_4031	all4199	30S ribosomal protein S5
sll1097	0886	PMM1510	cce_4090	all4339	30S ribosomal protein S7
sll1809	2219	PMM1545	cce_4028	all4202	30S ribosomal protein S8
sll1822	2205	PMM1531	cce_4043	all4187	30S ribosomal protein S9

...continues on next page

(...continued)

Syn6803	Syc7942	ProMED4	Cyn51142	Ana7120	Function
sll1821	2206	PMM1532	cce_4042	all4188	50S ribosomal protein L13
sll1802	2229	PMM1555	cce_4017	all4212	50S ribosomal protein L2
slr1678	1219	PMM1344	cce_1391	all0147	50S ribosomal protein L21
sll1803	2227	PMM1553	cce_4019	all4210	50S ribosomal protein L22
sll1807	2221	PMM1547	cce_4025	asl4204	50S ribosomal protein L24
sll1799	2232	PMM1558	cce_4013	all4215	50S ribosomal protein L3
sll1800	2231	PMM1557	cce_4015	all4214	50S ribosomal protein L4
sll1810	2218	PMM1544	cce_4029	all4201	50S ribosomal protein L6
sll0329	0039	PMM0770	cce_3746	alr5275	6-phosphogluconate dehydrogenase
sll1323	0333	PMM1454	cce_4485	all0008	ATP synthase subunit b' of CF(0)
sll1908	1501	PMM1354	cce_2134	alr1890	D-3-phosphoglycerate dehydrogenase
sll0519	1343	PMM0160	cce_2224	alr0223	NADH dehydrogenase subunit 1
slr0331	1976	PMM0150	cce_4299	alr3957	NADH dehydrogenase subunit 4 (involved in photosystem-1 cyclic electron flow)
sll0689	2359	PMM1600	cce_1688, cce_2468	alr0091, all1303	Na+/H+ antiporter
sll1818	2209	PMM1535	cce_4039	all4191	RNA polymerase alpha subunit
slr1265	1523	PMM1484	cce_3838	alr1595	RNA polymerase gamma-subunit
sll0306, sll2012	1746, 0672	PMM1629	cce_3594, cce_0644	alr3810, alr3800	RNA polymerase group 2 sigma factor
slr1424	1740	PMM0021	cce_2372	alr5066	UDP-N-acetylenolpyruvoylglucosamine reductase
sll0108	0442	PMM0263	cce_3261	alr0991	ammonium/methylammonium permease
ssl2667	0450	PMM0418	cce_1017	asr1309	an assembly factor for iron-sulfur culsters
slr0585	0009	PMM1707	cce_4370	alr4798	argininosuccinate synthetase
slr0549	1848	PMM1654	cce_0293	all3680	aspartate beta-semialdehyde dehydrogenese
slr1720	1313	PMM1688	cce_4152	all2436	aspartyl-tRNA synthetase
sll1498	2122	PMM0951	cce_0902	alr1155	carbamoyl-phosphate synthase small chain
sll1028, sll1029	1421	PMM0549	cce_4283, cce_4282	alr0867, all0868	carbon dioxide concentrating mechanism protein CcmK
slr1390	0942	PMM0743	cce_1270	all3642	cell division protein FtsH
sll0109	1915	PMM1181	cce_3372	all4012	chorismate mutase
slr0757	1217	PMM1343	cce_0423	alr2885	circadian clock protein KaiB homolog
slr1138	2604	PMM0444	cce_1975	alr0952	cytochrome c oxidase subunit III
slr0550	1847	PMM1653	cce_0294	all3679	dihydrodipicolinate synthase
slr1626	0040	PMM0591	cce_4420	alr3512	dihydroneopterin aldolase
slr2026	2303	PMM0830	cce_4305	alr4386	dihydropteroate synthase
slr1051	0126	PMM0282	cce_0460	all4391	enoyl-[acyl-carrier-protein] reductase
sll0018	1443	PMM0781	cce_1357	all4563	fructose-bisphosphate aldolase, class II
slr1843	2334	PMM1074	cce_2536	all4019	glucose 6-phosphate dehydrogenase
slr0638	2457	PMM1165	cce_3990	all1985	glycyl-tRNA synthetase alpha chain
slr1718	2589	PMM0614	cce_1018	all2568	hypothetical protein
sll0098	1758	PMM1481	cce_2727	all0355	hypothetical protein
sll0996	2374	PMM1305	cce_2810	alr1312	hypothetical protein
slr1847	0464	PMM0020	cce_4379	alr5067	hypothetical protein
slr1471	1617	PMM1186	cce_1364	alr3415	hypothetical protein

...continues on next page

(...continued)

Syn6803	Syc7942	ProMED4	Cyn51142	Ana7120	Function
slr1577	0426	PMM0909	cce_0429	all5166	hypothetical protein
ssr0332	0658	PMM1563	cce_4619	asr4076	hypothetical protein
slr0742	2023	PMM1491	cce_1842	alr3828	hypothetical protein
ssl0353	2017	PMM0061	cce_0650	asl0940	hypothetical protein
slr0362	2396	PMM0789	cce_0268	alr3102	hypothetical protein
sll1898	2601	PMM0447	cce_4599	all0949	hypothetical protein
slr1896	0286	PMM0390	cce_2055	all0876	hypothetical protein
sll0372	0362	PMM0239	cce_2501	all2849	hypothetical protein
sll1866	1864	PMM0319	cce_4203	alr0116	hypothetical protein
sll0854	0826	PMM1517	cce_3387	all3378	hypothetical protein
slr0506	2503	PMM0542	cce_0320	all1743	light-dependent NADPH-protochlorophyllide oxidoreductase
slr0772	1838	PMM0544	cce_1954	alr3441	light-independent protochlorophyllide reductase subunit ChlB
slr1540	1826	PMM0797	cce_3779	alr4831	mRNA-binding protein
slr1055	2137	PMM0831	cce_4358	all4365	magnesium protoporphyrin IX chelatase subunit H
slr2033	1179	PMM0295	cce_1309	alr3843	membrane-associated rubredoxin, essential for photosystem I assembly
sll0902	2514	PMM1263	cce_3251	alr4907	ornithine carbamoyltransferase
sll1553	1293	PMM0871	cce_1321	alr4958	phenylalanyl-tRNA synthetase
sll0684	2441	PMM0725	cce_0883	all4572	phosphate transport ATP-binding protein PstB homolog
slr0597	0396	PMM0266	cce_4354	all3093	phosphoribosyl aminoimidazole carboxy formyl formyltransferase/inosinemonophosphate cyclohydrolase (PUR-H(J))
slr1645	0343	PMM0507	cce_3633	all1258	photosystem II 11 kD protein
sll0851	0656	PMM1158	cce_0659	alr4291	photosystem II CP43 protein
slr1311, sll1867, slr1181	0893, 1389, 0424	PMM0223	cce_3411, cce_0267, cce_3501, cce_0636	alr3742, alr4866, alr4592, alr3727, all3572	photosystem II D1 protein
slr0906	0697	PMM0315	cce_1837	all0138	photosystem II core light harvesting protein
sll0427	0294	PMM0228	cce_2572	all3854	photosystem II manganese-stabilizing polypeptide
sll0849, slr0927	1637, 0655	PMM1157	cce_2485, cce_0660	alr4290, alr4548	photosystem II reaction center D2 protein
sll0171	2308	PMM1687	cce_4346	all4609	probable aminomethyltransferase
slr1673	1199	PMM1299	cce_0402	alr0175	probable tRNA/rRNA methyltransferase
slr0774	0142	PMM0929	cce_4646	all0121	protein-export membrane protein SecD
sll1786	1521	PMM1486	cce_3489	alr1593	putative deoxyribonuclease, tatD homolog
slr2034	1178	PMM0296	cce_1308	alr3844	putative homolog of plant HCF136, which is essential for stability or assembly of photosystem II
sll1841	1068	PMM0405	cce_2750	alr3606	pyruvate dehydrogenase dihydrolipoamide acetyltransferase component (E2)
sll1282	2244	PMM1643	cce_4679	alr3993	riboflavin synthase beta subunit
slr0194	0584	PMM1489	cce_0103	all0888	ribose 5-phosphate isomerase
slr0012	1427	PMM0551	cce_3164	alr1526	ribulose bisphosphate carboxylase small subunit

...continues on next page

(...continued)

Syn6803	Syc7942	ProMED4	Cyn51142	Ana7120	Function
slr0743	2022	PMM1492	cce_1841	alr3829	similar to N utilization substance protein
ssr1600	2121	PMM0950	cce_1801	asr1156	similar to anti-sigma f factor antagonist
sll1820	2207	PMM1533	cce_4041	all4189	tRNA pseudouridine synthase 1
sll1980	2128	PMM0242	cce_0972	alr0570	thiol:disulfide interchange protein TrxA
sll1615	1582	PMM0189	cce_4596	all4677	thiophen and furan oxidation protein
sll0755	2309	PMM0856	cce_2409	alr4641	thioredoxin peroxidase
slr1793	2297	PMM0519	cce_4687	all2563	transaldolase
sll1957	0688	PMM0714	cce_3556	alr2766	transcriptional regulator
slr1884	1308	PMM0598	cce_3534	all1269	tryptophanyl-tRNA synthetase

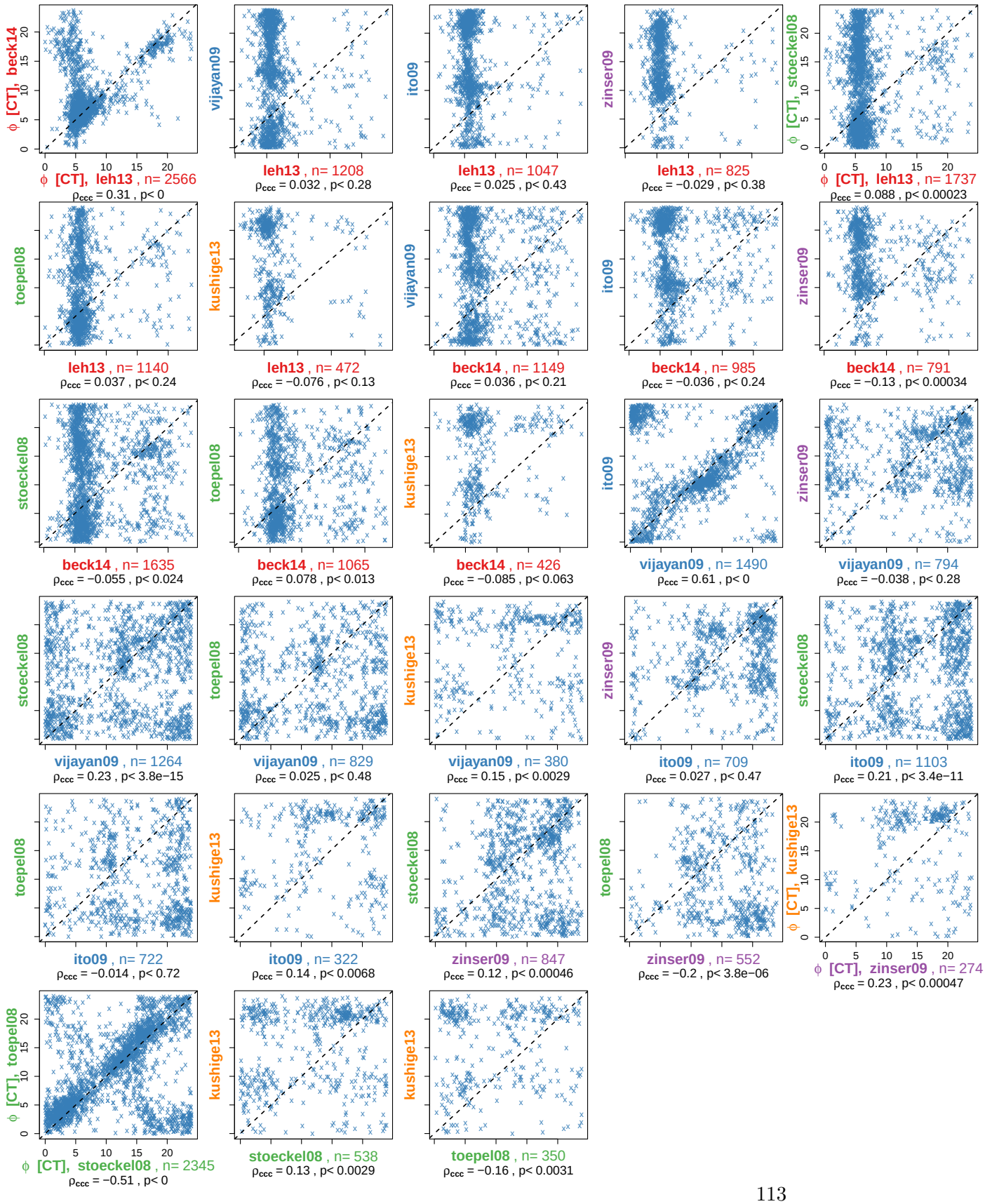


Fig. 4: **Pairwise comparison of expression phase between all available datasets across cyanobacterial strains.** The peak expression phase ϕ [CT] is compared for all possible pairwise combinations of available circadian datasets, considering only gene pairs which oscillate significantly ($fdr < 0.05$) in both datasets. For same-strain combinations phases can be compared directly, whereas for different strains gene pairs are derived via homology prediction. The dataset name is shown on each axis using the respective specie's colour, the number of homologous significantly oscillating genes is provided with the x-axis label, together with the circular correlation coefficient ρ_{ccc} and its corresponding p-value.

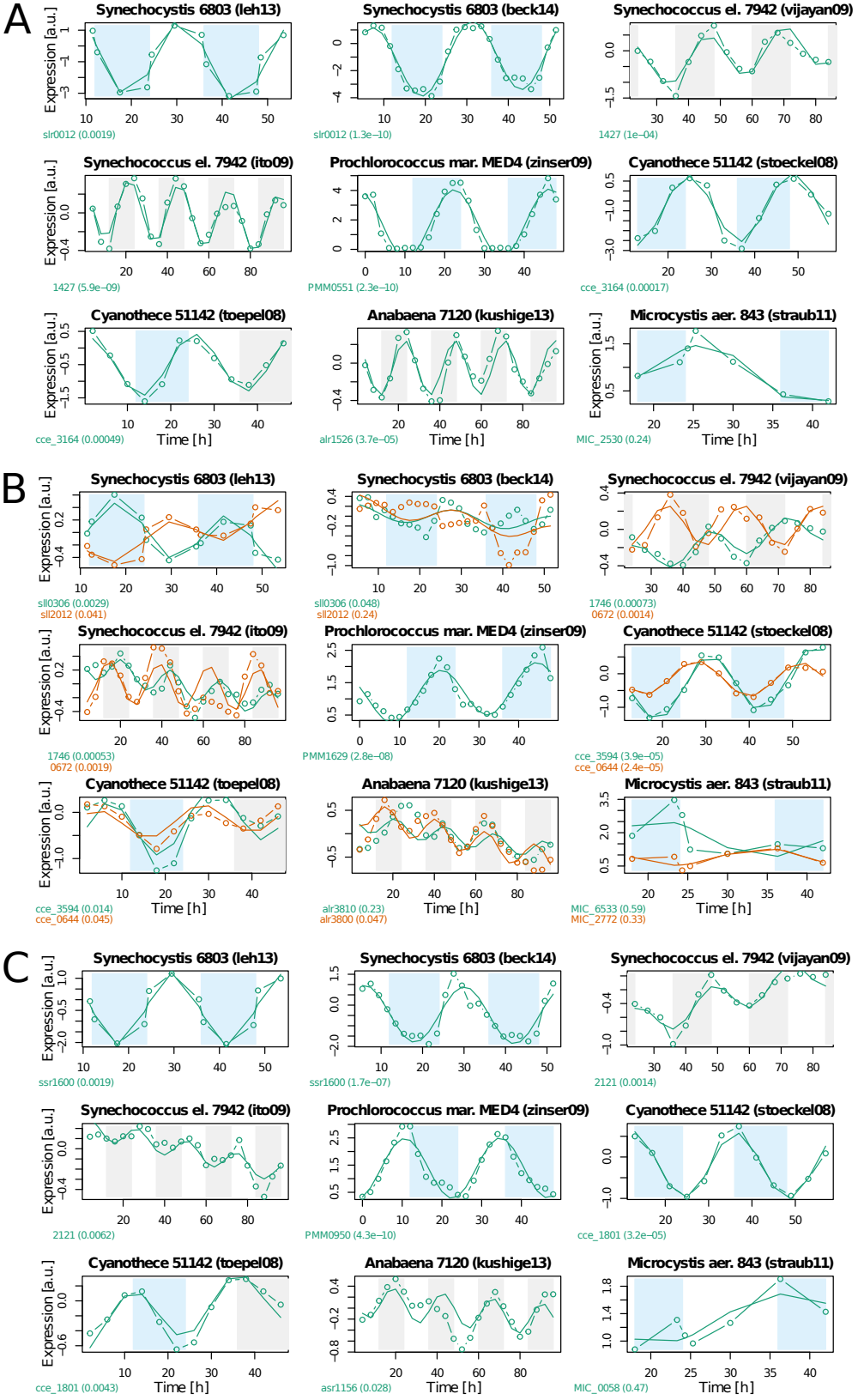


Fig. 5: **Expression profiles of RuBisCO, sigma factor B (sigB), and anti sigma factor F across all datasets.** Expression profiles and harmonic regression functions in all datasets for (**A**) RuBisCO (Ribulose-1,5-bisphosphat-carboxylase/-oxygenase), (**B**) sigma factor B, and (**C**) anti sigma factor F. Expression data are shown circles and solid line, whereas the harmonic regression function is a solid line, different homologs are distinguished by colour. The homolog identifiers and p-value are shown below in matching colour.

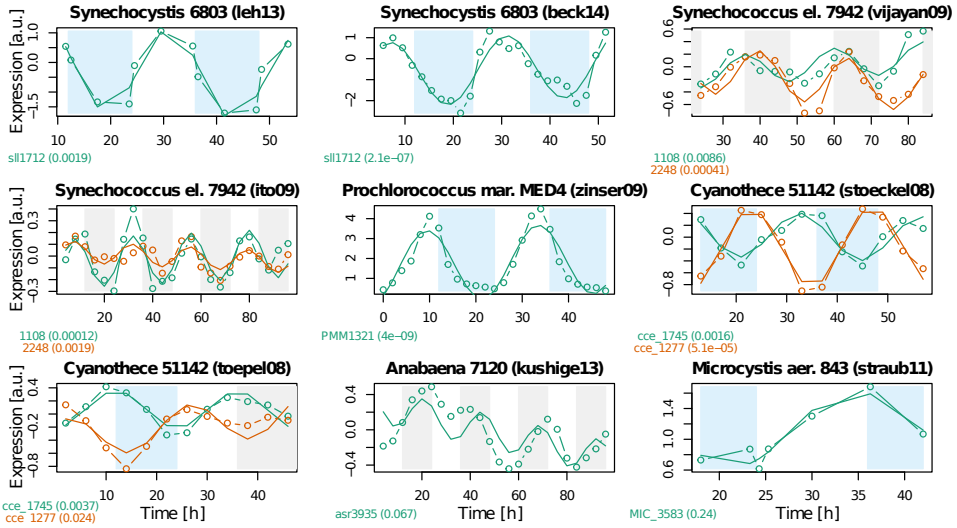


Fig. 6: **Expression profiles of cyanobacterial homologs of the nucleoid associated protein HU across all datasets.** Expression profiles and harmonic regression functions in all datasets for homologs of the NAP HU. Expression data are shown circles and solid line, whereas the harmonic regression function is a solid line, different homologs are distinguished by colour. The homolog identifiers and p-value are shown below in matching colour.

Appendix B

3 Dinucleotide periodicity detection

3.1 Dinucleotide periodicity measurement

Autocorrelation

The autocorrelation function (ACF) was used as described in Schieg *et al.*[190]. The number of dinucleotides, *e.g.* NN, in distance k ($N_{NN-NN}(k)$) was counted and normalised it by the number of possible pairs

$$p_{NN-NN}(k) = \frac{N_{NN-NN}(k)}{N-k-1} \quad ,$$

to obtain the pair probability in a sequence of length N. The background probability ${p_{NN}}^2$ to observe a NN-pair due to the sequence composition is obtained from the observed NN-dinucleotide probability p_{NN}. The employed correlation measure is then the difference between the observed and the background pair probability

$$C_{NN-NN}(k) = p_{NN-NN}(k) - {p_{NN}}^2 \quad .$$

As coding sequences introduce a strong 3 bp periodicity, a much better understood observation [74, 205, 225], a window smoothing of width 3 is applied to suppress this signal.

$$\bar{C}_{NN-NN}(k) = \frac{\sum_{i=k-1}^{k+1} C_{NN-NN}(i)}{3}$$

Spectral analysis of whole genome autocorrelation function

To analyse the dinucleotide autocorrelation function (ACF) $\tilde{C}_{NN-NN}$ obtained from the entire chromosomal sequence, the power spectrum was calculated via Fourier transform of the window $k = 30 - 131$ base pairs, as indicated in Figure 3.1 A, according to

$$Q_{NN}(T) = \left| \sum_{k=k_{min}}^{k_{max}} \bar{C}_{NN-NN}(k) exp\left(-ik\frac{2\pi}{T}\right) \right| \quad ,$$

where $k_{min} = 30$ and $k_{max} = 131$ specify the interesting range of the ACF. The spectra were evaluated for the period range $T = [4, 50]$bp to exclude the 3 bp periodic signal of the coding sequence.
The obtained spectra are normalised according to ensure comparability between genomes with varying sequence composition.

$$Q^*_{NN}(T) = \frac{(T_{max} - T_{min} + 1)Q_{NN}(T)}{\sum_{i=T_{min}}^{T_{max}} Q_{NN}(T)} \quad .$$

The goal of this analysis is to determine if species exhibit preferences in the choice of periodic dinucleotides and if such preferences are consistent across species. To facilitate this comparison, dinucleotide spectra were condensed to a scalar reflecting the information whether the $\sim$ 10 bpperiodicity is smaller, similar or stronger than its periodicity in the remaining spectrum. The four components in the interval $T = [9, 13]$bp (9.2, 10.1, 11.2, 12.6) were defined as foreground spectrum $Q^*_F(NN)$ of dinucleotide NN, while the remaining 18 spectral components ($T = [4, 9] \cup [13, 34]$bp) constitute its background spectrum $Q^*_B(NN)$. The signal-to-noise ratio (maximum of the foreground spectrum $Q^*_F(NN)$ divided by the median of the background spectral power $Q^*_B(NN)$) was employed as indicator of the periodicity strength of each dinucleotide:

$$Q^*_{SNR}(NN) = \frac{max(Q^*_F(NN))}{\tilde{Q^*}_B(NN)} \quad .$$

Due to the prominent role of W nucleotides, the WW and AT2 motifs were included in hierarchical cluster analysis, using Euclidean distances, and 'complete linkage' for dinucleotides and Ward's minimum variance method for genomes. This clustering, cut at level k = 4 (Fig. 3.2 A), is provided with the strain list in Appendix B File 1. Ancestral states of $Q^*_{SNR}(NN)$ and species cluster assignments (Fig. 3.2 B) were inferred on the phylogenetic tree obtained from Shih *et al.*using maximum-likelihood methods via the R package `ape`. Only direct transitions $A \leftrightarrow B \leftrightarrow C \leftrightarrow D$ were allowed for discrete state modelling of ancestral cluster assignments.

3.2 Periodicity localisation analysis

The dinucleotide periodicity analysis of whole genome can be extended to determine whether $\sim$ 10 bpperiodicity is found in specific genomic localisations. Therefore, the chromosomal sequence was divided into non-overlapping windows, calculated the ACF for each window, and performed the spectral analysis on the ACF with $k_{min} = 30$ and $k_{max} = 101$. This yields a matrix of $Q^*(T, i)$ values, where T denotes the period in bp of the corresponding spectral component, and i denotes the index of the sequence window. While the accuracy of the spectral analysis improves with increasing window size, application of large genomic windows is problematic since periodic regions are generally only 100 bp in length [37, 72, 139, 217]. When computing the ACF across a sequence

that contains several periodic stretches which are out of phase, the obtained ACF will show a severely reduced signal. To alleviate this problem, the ACF was calculated on small non-overlapping sequence windows (200 bp) and averaged the spectra, to effectively increase the window-size to the necessary width. A related spectrum averaging approach has been used previously by Kravatskaya *et al.*[105, 106] to analyse promoter sequences in *E. coli*. As already described, the signal strength of the ~ 10 bpperiodicity strongly varies amongst different species [139, 190]. Therefore, only cyanobacterial species with high signal strength were selected (*Cyanothece* sp. 8801, *Cyanothece* ATTC 51142, *Synechocystis* sp. 6803, *Synechococcus* sp. 7942), and the number of averaged spectra was varied between 2 to 4, effectively achieving a window size of 400 to 800 bp.

The statistical significance of the spectral components $Q^*(T,i)$ for each window i was assessed to exclude effects purely due to sequence composition from the subsequent analysis. Therefore, the sequence of each window i was randomly permuted and its Fourier spectrum $Q^*_{perm}(T,i)$ was recalculated to numerically obtain a p-value for each spectral component T as:

$$P_{T,i} = Pr(Q^*_{perm}(T,i) > Q^*(T,i)) \quad .$$

The composition of the sequence window permutations was kept constant with respect to dinucleotides in the original sequence, since dinucleotides are the subject of this analysis. `ushuffle` as implemented by Jiang *et al.*[88] was used to calculate 5000 permutations per window. For subsequent analyses, a set of periodic sequence windows i_{per} for each species was defined, which exceeds the significance threshold $\min_T (P_{T,i}) < 0.01$ in the interesting period range $T \in [9, 13]$bp.

To test whether the window periodicity is reflecting the global ~ 10 bpperiodicity, the distribution of the periods with smallest p-value $T_{sig} = \operatorname{argmin}_T P_{T,i},\ i \in i_{per}$ was compared across periodic windows with the period measured in the whole chromosome.

3.3 Testing significance of overlap between periodic windows and coding-sequence

After selection of significantly periodic sequence windows, adjacent windows were combined to sets of non-overlapping periodic genomic intervals. The significance of the overlap between these intervals and coding, or alternatively intergenic, sequences was tested, using R package `GenometriCorr` [50]. All regions annotated with CDS, tRNA, rRNA, and CRISPR in the gene annotations available from GenBank [16] were defined as coding. The *p*-values derived from the Jaccard measure and the projection test were used. Both tests focus on the overlap of two groups of intervals. This is particularly important, since the relative distance tests are considered problematic in small genomes with close distances between reference points, which is the case for cyanobacteria.

3.4 Locating periodicity in codon positions

As described previously, a significant contribution to the genome-wide ~ 10 bpperiodicity is located in the third codon position within coding sequence [37]. This is explained with the degenerateness of the genetic code, which allows to encode a specific amino acid with various codons with the greatest variability in the third codon position. Cohanim *et al.*have used two different background models to perturb properties of the original coding sequence and thereby locate the periodicity-related ones. Firstly, they permuted the dinucleotide order while keeping the overall composition constant, which affects all three codon positions and destroys the amino acid sequence. Secondly, a synonymous codon permutation was used which keeps the amino acid sequence and codon usage intact and mainly affects the third codon position. For the selected set of 12 prokaryotic species, they conclude that the majority of the ~ 10 bpperiodicity is encoded in the third codon position, but also find contributions in the first and second position.

This analysis was repeated and extended to cyanobaterial species which exhibit particularly strong ~ 10 bpperiodicity. In addition to the dinucleotide, codon order, and synonymous codon permutations, permutations were performed exclusively in either the first, second, or third codon position with preservation of the original base composition, to further characterise the localisation of the periodicity signal. A customised version of the R package `seqinr` was used to perform codon position permutations. All genes were excluded for which the annotated length is not a multiple of three. After the different treatments of the individual sequences and subsequent concatenation, the periodicity was quantified as described before via analysis of the ACF spectra $Q^*(T)$. The parameters for the spectral analysis were set to $k_{min} = 30$, $k_{max} = 101$, and $T = [2, 30]$bp. No 3 bp smoothing was performed in this case to preserve the codon usage signal. In Figure S3, a spectra comparison between the entire original chromosome, only coding sequence, synonymous codon-permuted, and codon position-specifically permuted sequences is shown for a) *Synechocystis* sp. 6803 and b) *Cyanothece* sp. 8801.

3.5 Clustering genes by spectrum

The windowed periodicity measurement approach as described above (spectral analysis of ACF with permutation-based background for significance assessment) was modified to accept coding region annotations instead. Genes with less than 300 bp in length or without correct open reading frame annotation (length not multiple of three) were excluded from this analysis. A total of 306 coding sequences for *Synechocystis* sp. PCC 6803 and 612 for *Cyanothece* sp. 8801 were excluded, the majority due to the length criterium. The ACF was calculated, and the spectral analysis was performed using the range of 30-130 bp. Due to the high computational cost and good correlation between p-values and amplitudes in the window based analysis, sequence permutations were not employed as background model. The resulting matrix $Q^*_{AT2}(T, i)$ of normalised spectral components with gene index i and period T was row-wise hierarchically clustered across genes using Ward's method, euclidean distance, and 14 clusters.

Properties of spectral gene clusters

The spectral pattern based gene clusters were tested for systematic variation in different properties, such as gene length, GC content, CpG content, gene functional enrichment analysis using Gene Ontology terms, and enrichment analysis of protein domains annotated to the genes. For the CpG measurement the expected-to-observed ratio was used:

$$P_{CpG} = \frac{P(CpG)}{P(C) * P(G)} \quad .$$

For the cluster-wise gene ontology enrichment analysis the gene function category annotations were obtained from the CyanoBase database [146]. A hypergeometrical test was performed for each category in each cluster, while categories one and two were treated as separate. The obtained p-values were adjusted due to multiple testing using the Benjamini-Hochberg correction. The outcome of this procedure is described in the results section.
A similar approach was used to test the gene clusters for enrichment of protein domains. The Interpro protein domain annotations from the prediction algorithms HMMPfam, HMMTigr, Gene3D, HMMPanther, HMMSmart, BlastProDom, ProfileScan, FPrintScan, and HMMPIR were obtained from CyanoBase database. The identical hypergeometrical test with Benjamini-Hochberg correction for multiple testing was performed. The results of this procedure are shown in Table S1 for *Cyanothece* sp. 8801 and in Table S2 for *Synechocystis* sp. PCC 6803.

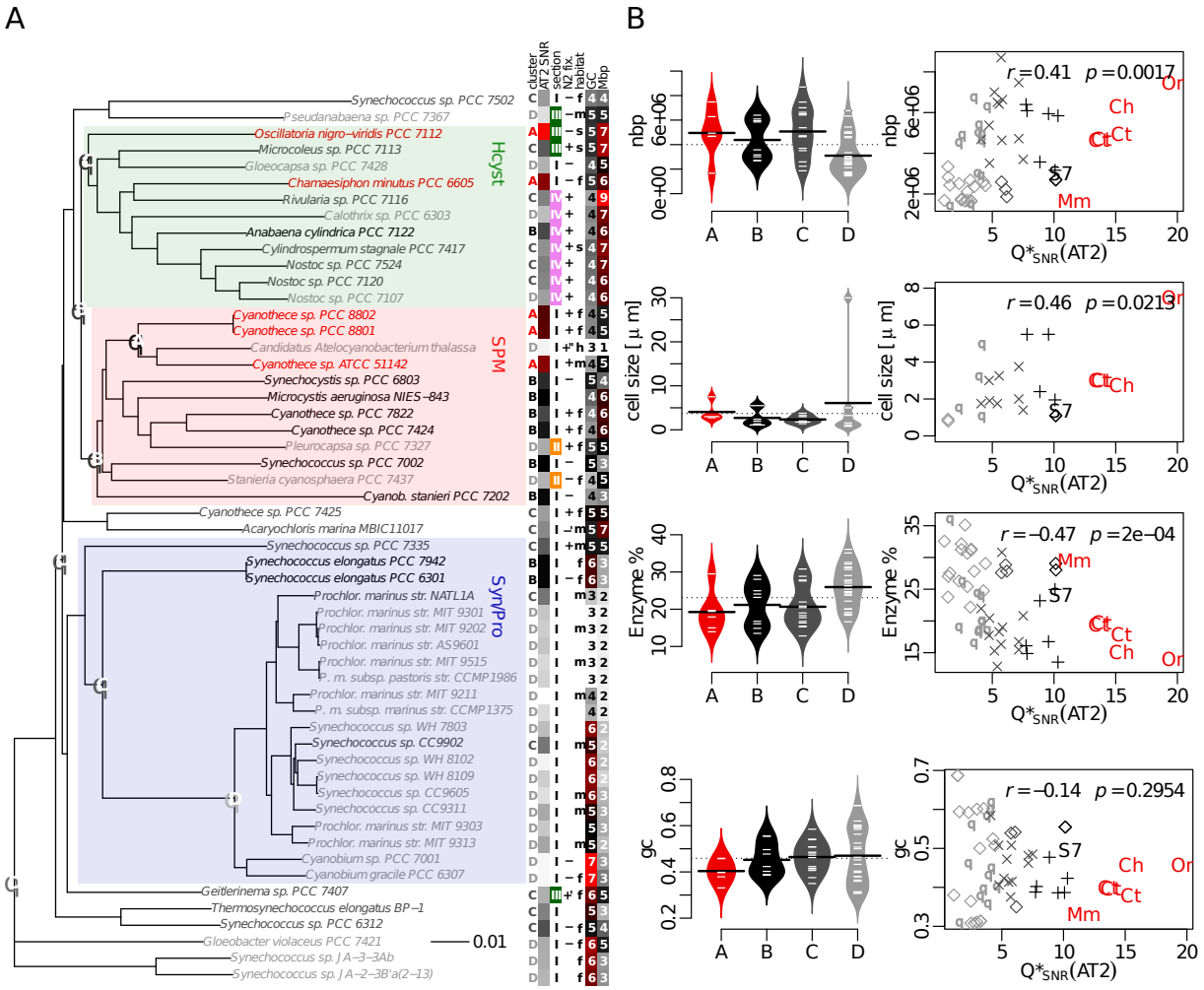
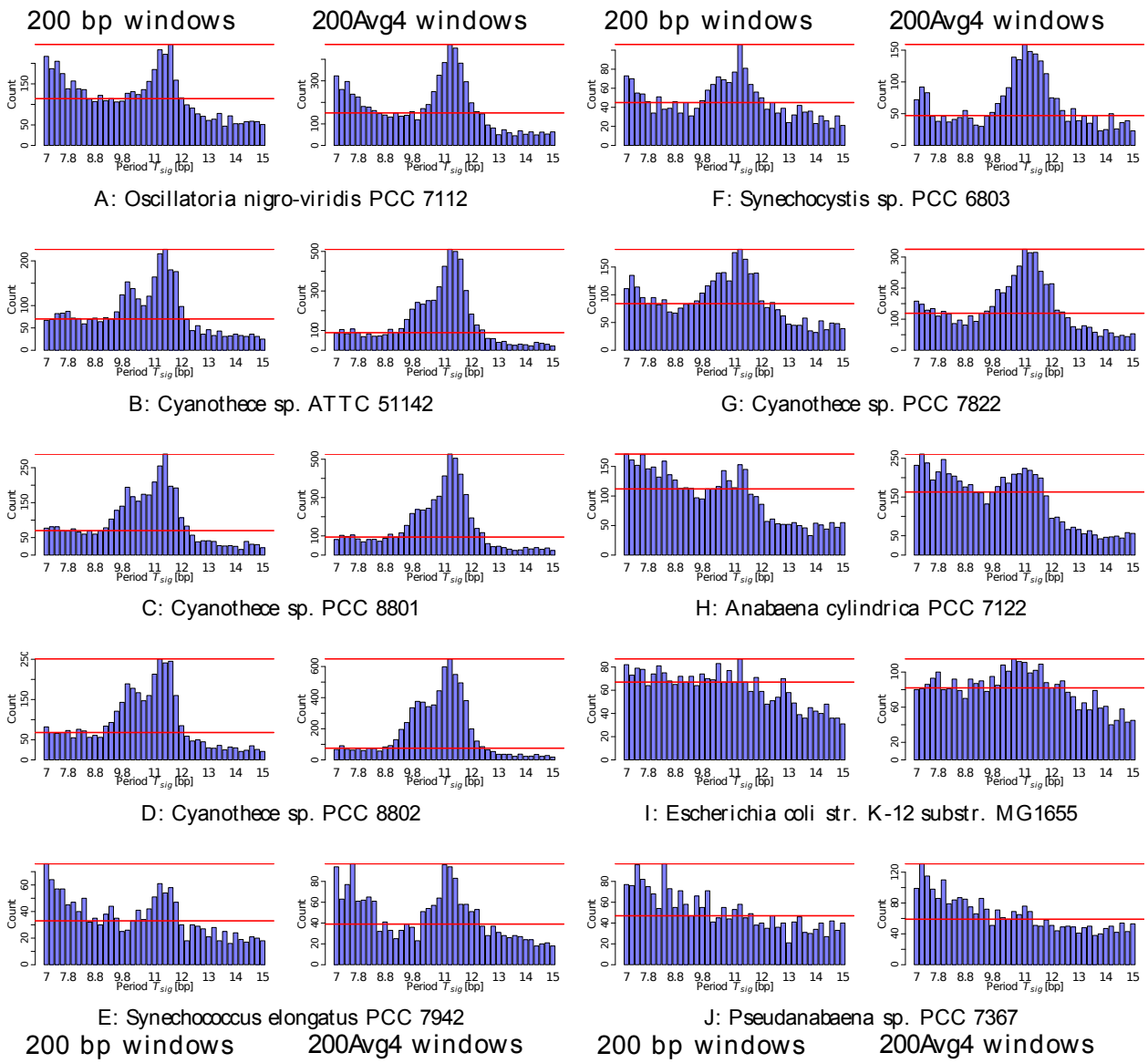
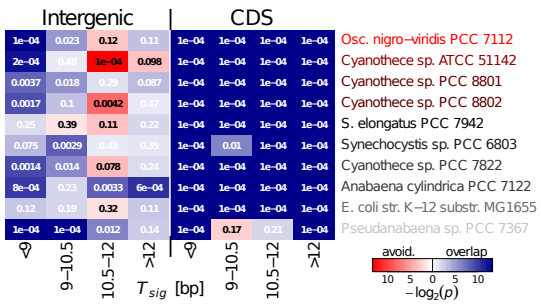


Fig. 7: **Cyanobacterial 16S rRNA phylogeny vs. genomic and physiological properties.** (**A**) The 16S rRNA based phylogeny is shown (SILVA [174], obtained from IMG database [129]), similar to Figure 3.2 B. In comparison, the genomic GC content (%/10) and the habitat (from IMG and [198]; f: fresh water, m: marine, s: soil) are shown. (**B**) Genomic properties, the number of base pairs 'nbp', GC-content 'gc', the mean of cell diameter ranges from Shih *et al.*[198] 'cell size', the fraction of enzyme genes as provided by IMG 'Enzyme %'. (Left) The distributions for the species clustering in Fig. 3.2 A. (Right) The properties are plotted against the species $Q^*_{SNR}(AT2)$. The Pearson correlation (r, p-value) were calculated only for eubacterial species and the outlier species in cell size (*Stanieria cyanosphaera* PCC 7437 with 30 µm) was further excluded from this test. Cluster memberships are marked as follows. Diamonds: picocyanobacteria from the Syn/Pro clade in several clusters; light grey circles: cluster D; dark grey 'x': cluster C; black '+': cluster B (S7: *Synechococcus* sp. PCC 7002); red text: cluster A, where species names are indicated by abbreviations (Mm: the Archaeum *Methanococcus marsipaludis* S2; On: *Oscillatoria nigro-viridis* PCC 7112; Ch: *Chamaesiphon minutus* PCC 6605; Ct: *Cyanothece* sp. PCC 8801, PCC 8802 and ATCC 51142).





K: overlap tests for 200Avg4 windows

Fig. 8: **Distributions of local dominant periods in different cyanobacteria.** (**A**-**J**): Histogram of the dominant period T_{sig} in all genomic windows i with $min(P_{T,i}) < 0.01$ for primary windows of 200 bp length (left plot in each panel) and with averaging over 4 adjacent spectra (200Avg4, right). The red lines indicate maximum and median of the T_{sig} distribution. (**K**) identical to Fig. 3.3 A, using 200Avg4 windows.

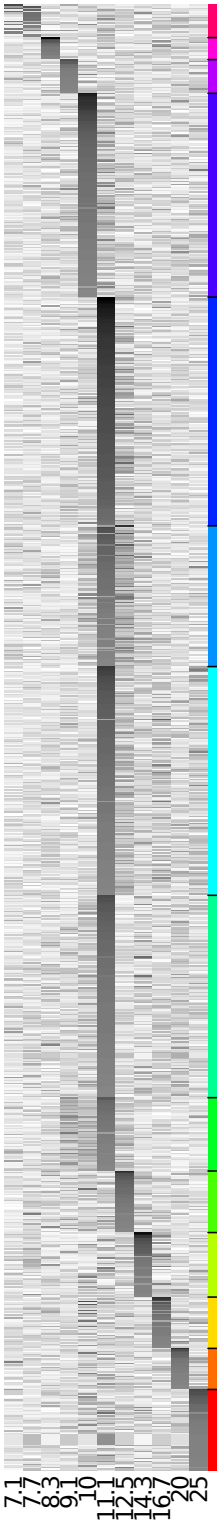


Fig. 9: **CDS clustering by AT2 spectrum in *Cyanothece* sp. PCC 8801.** Clustering of the 1000 *Cyanothece* sp. PCC 8801 CDS with the highest AT2 spectral amplitude, similar to Fig. 3.4 A. Refer to Fig. 10 for details on the clusters.

Synechocystis sp. PCC 6803

cluster	1	2	3	4	5	6	7
T_{max}	7.1	7.7	8.3	9.1	10	11.1	11.1
count	27	26	30	34	73	202	23
group	< = 10					11.1	

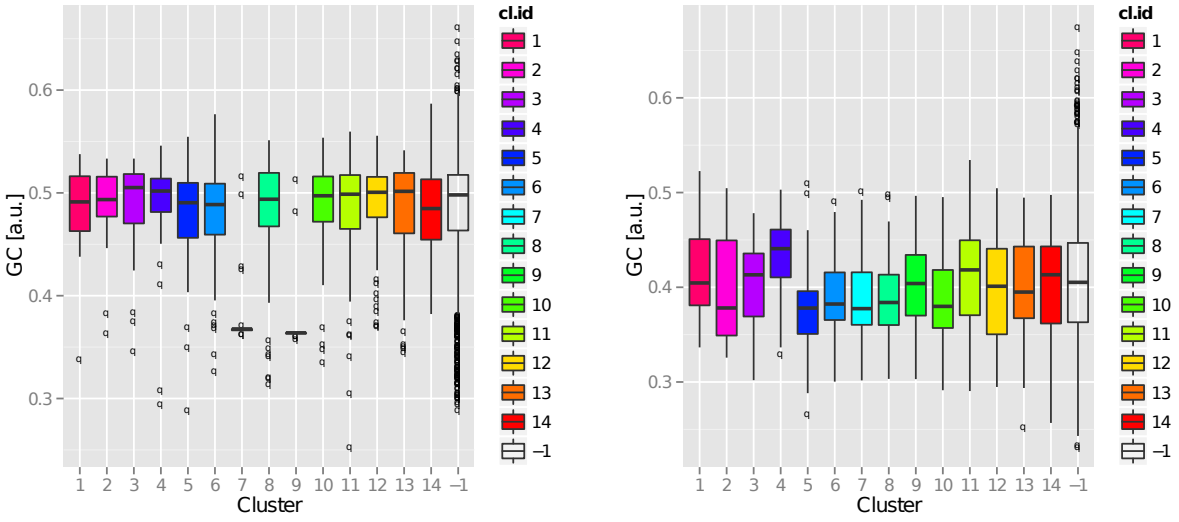
cluster	8	9	10	11	12	13	14
T_{max}	12.5	12.5	14.3	16.7	20	25	25
count	103	12	131	122	92	90	35
group	12.5		> 14				

Cyanothece sp. PCC 8801

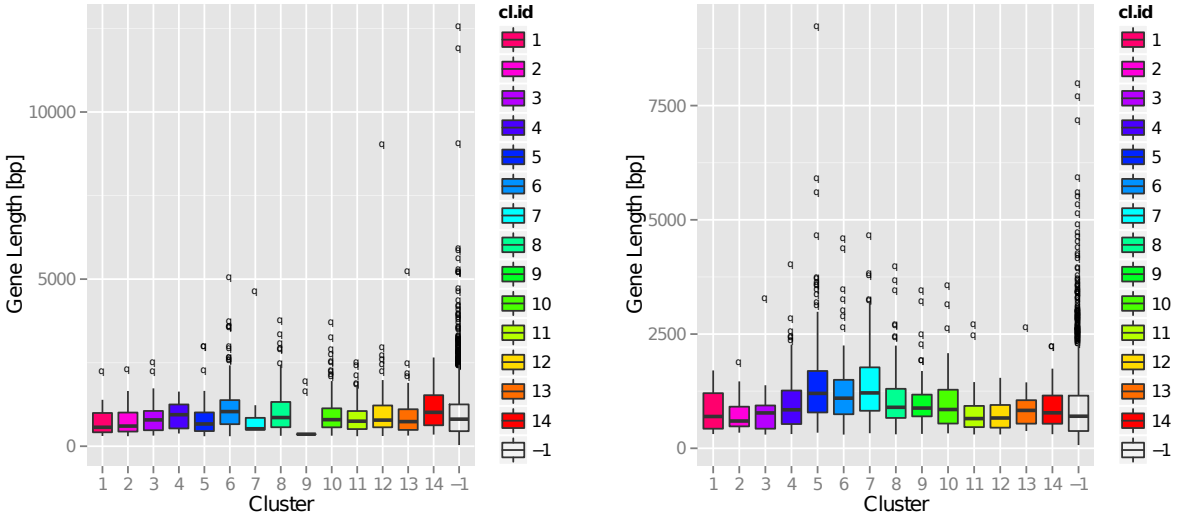
cluster	1	2	3	4	5	6	7
T_{max}	7.7	8.3	9.1	10	11.1	11.1	11.1
count	22	15	23	139	156	96	156

cluster	8	9	10	11	12	13	14
T_{max}	11.1	11.1	12.5	14.3	16.7	20	25
count	138	50	42	44	35	28	56

A: CDS Cluster Counts and Periods



B: CDS Cluster GC Content



C: CDS Cluster Length

Fig. 10: **Sequence properties of CDS clusters by dominant period in *Synechocystis* sp. PCC 6803 and *Cyanothece* sp. PCC 8801.** (**A**) Dominant period and number of members for spectral CDS clusters, period grouping only for *Synechocystis* sp. PCC 6803, as presented in Fig. 3.4 A and 9. (**B**) Cluster-wise GC-content and (**C**) gene length. Group '-1' represents CDS not included in the clustering due to insufficient amplitude in CDS spectrum.

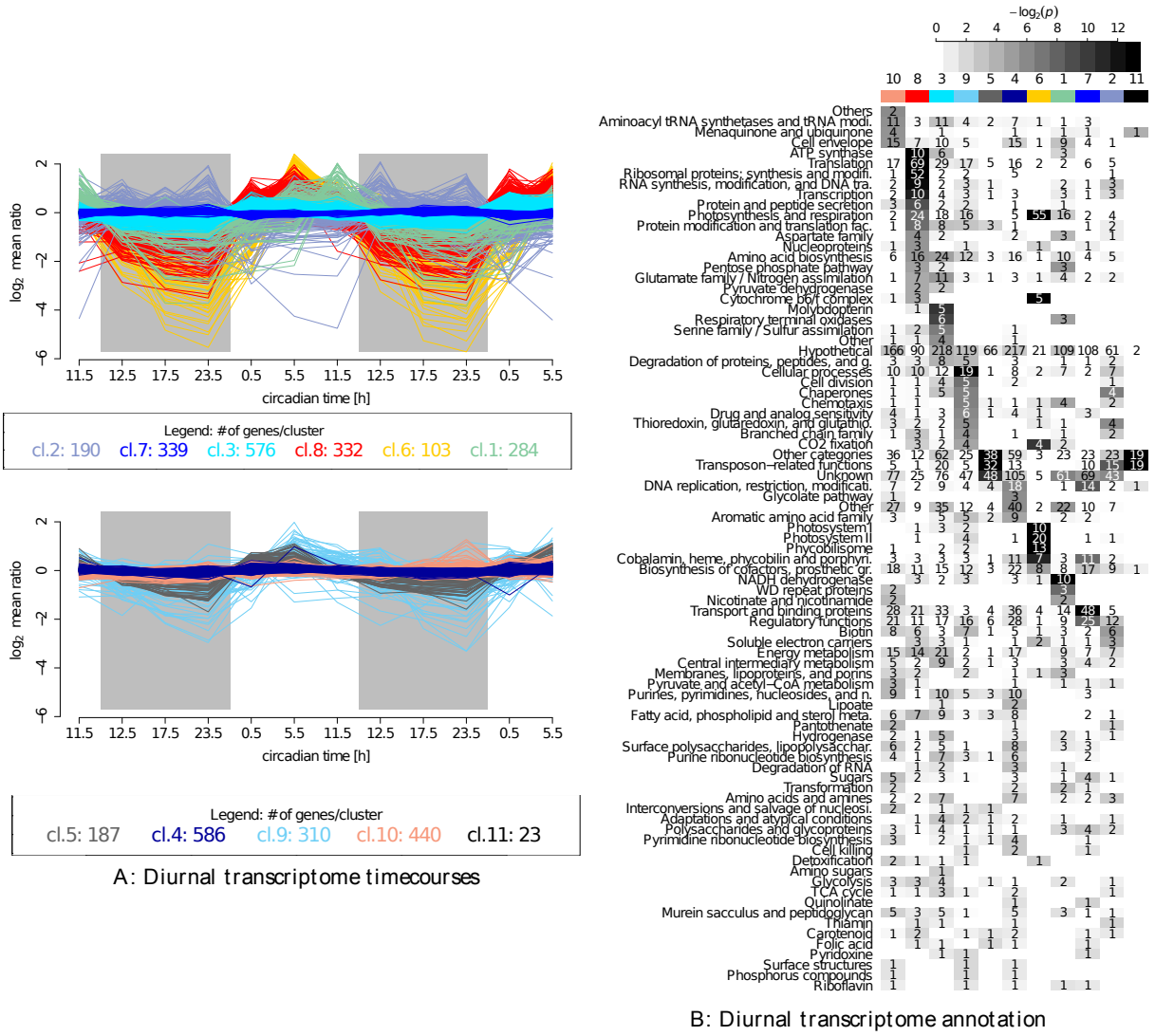
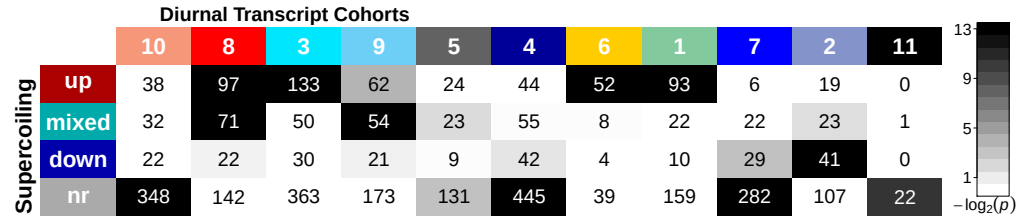
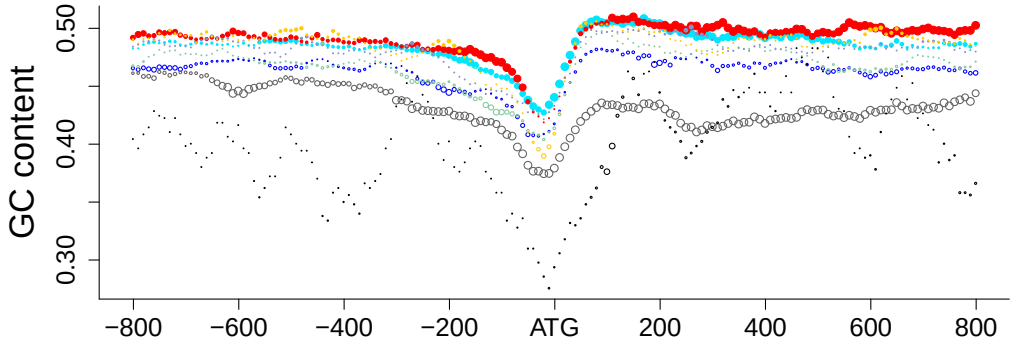


Fig. 11: **Diurnal co-expression clustering in *Synechocystis* sp. PCC 6803** (**A**) Clustering of the 3370 protein-coding transcripts of the diurnal expression dataset in *Synechocystis* sp. PCC 6803 presented in Chapter 2, using the described method. Cluster 11 represents 23 annotated genes not included in the microarray. The clusters are sorted by peak phase and divided into high and low amplitude groups. (**B**) Overlap between co-expressed clusters and functional annotations obtained from CyanoBase (cumulative hypergeometric test, not controlled for multiple testing).



A: Diurnal vs. Supercoiling-sensitive transcripts



B: GC content of supercoiling-sensitive genes

Fig. 12: **Supercoiling-sensitive expression vs. diurnal expression pattern in *Synechocystis* sp. PCC 6803.** (**A**) Numbers of genes found in diurnal co-expression clusters (columns) and simultaneously in groups of genes with supercoiling-sensitive expression (rows) (up: expression induced by negative supercoiling, down: repressed, mixed: mixed response, nr: no response, from [171]). The p-value of a hypergeometric test in each matrix cell is colour-coded, with black showing high significance. (**B**) Average GC-content in co-expression clusters with supercoiling-sensitivity, sliding-window of 71 bp width around start-codon (statistical DNA profile, [127]). Circle size inversely correlates with the two-sided t-test p-value (window at corresp. position for each gene in cluster versus window for all other genes).

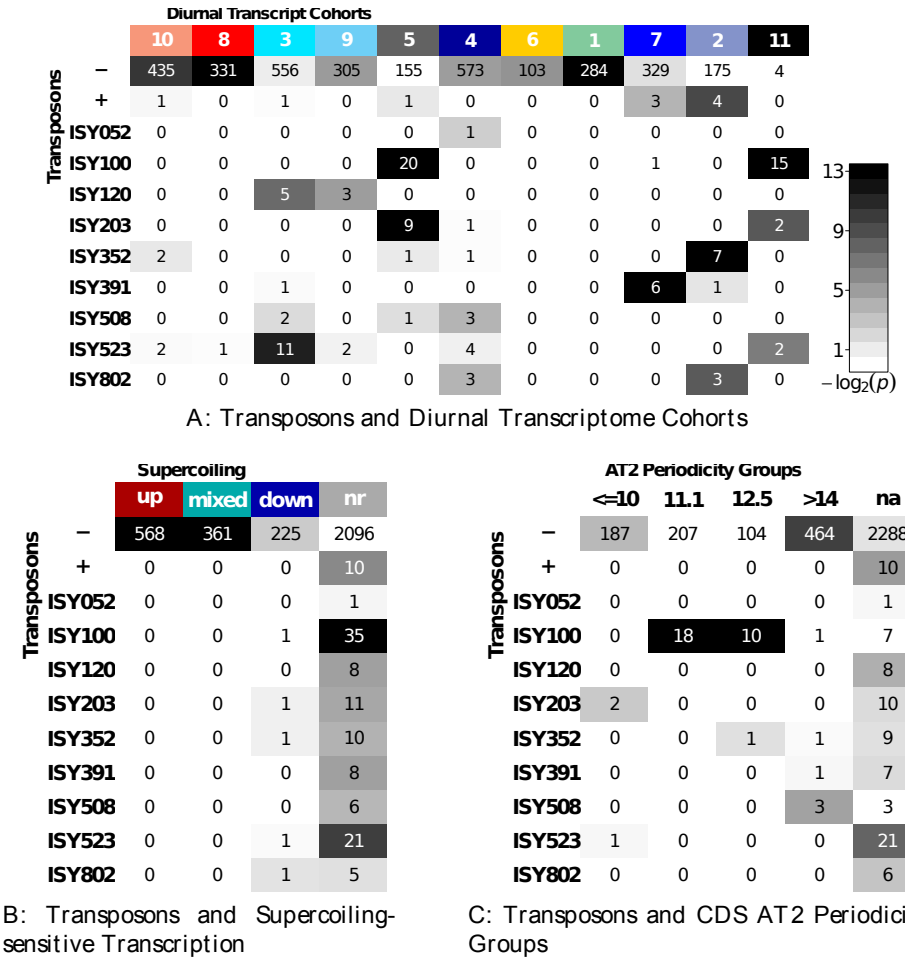


Fig. 13: **Properties of transposons in *Synechocystis* sp. PCC 6803.** Comparison of diurnal expression profile (**A**), supercoiling-sensitive expression (**B**), and AT2 periodicity (**C**) of transposon genes in *Synechocystis*. In each matrix, rows refer to the different transposon types ('ISY', each with multiple copies), not separately named transposons ('+'), and all remaining CDS ('-'), while the columns refer to the different property groups. The co-expression clusters in (A) refer to the clustering shown in Chapter 2 Fig. 2.4 and Chapter 3 Fig. 3.4 C. Supercoiling-sensitive gene clusters in (**B**) are obtained from Prakash *et al.*[171], and CDS clusters by dominant periodicity period (**C**) refer to Chapter 3 Fig. 3.4 A. Importantly, ISY transposons are frequently found in two separate adjacent genes coding only half of the mRNA, respectively. ISY100 is the most frequent transposon in *Synechocystis*, for a detailed analysis see Urasaki *et al.*[232].

Cluster-wise Domain Annotation Enrichment 6803			
Domain-ID	p-value/ BH	in Cluster	total
Cluster 3 (30 genes, T_{max} 8.3 bp)			
PF01850	0.0074/ 1	2	4
PF00395	0.0038/ 1	2	3
Cluster 4 (34 genes, T_{max} 9.1 bp)			
PF01609	0.0048/ 1	2	3
Cluster 6 (202 genes, T_{max} 11.1 bp)			
PF04055	0.0041/ 1	5	6
Cluster 7 (23 genes, T_{max} 11.1 bp)			
PF01710	3.1e-07/ 1.6e-4	7	20
Cluster 9 (12 genes, T_{max} 12.5 bp)			
PF01710	3.2e-16/ 1.7e-13	10	20
Cluster 10 (131 genes, T_{max} 14.3 bp)			
PF00805	9e-05/ 0.048	6	7
Cluster 14 (35 genes, T_{max} 25 bp)			
PF00070	0.01/ 1	2	4

Cluster-wise Domain Annotation Enrichment 8801			
Domain-ID	p-value/ BH	in Cluster	total
Cluster 1 (22 genes, T_{max} 7.7 bp)			
PF01385	1.1e-4/ 0.054	4	11
Cluster 3 (23 genes, T_{max} 9.1 bp)			
PF01590	0.0072/ 1	2	5
Cluster 7 (156 genes, T_{max} 11.1 bp)			
PF00924	0.0068/ 1	3	3
Cluster 8 (138 genes, T_{max} 11.1 bp)			
PF01527	7.7e-4/ 0.37	4	4
Cluster 11 (44 genes, T_{max} 14.3 bp)			
PF00805	4.8e-14/ 2.3e-11	12	15
Cluster 14 (56 genes, T_{max} 25 bp)			
PF01609	3.6e-08/ 1.7e-05	7	8

Pfam Domain Index:

ID	Family	Description	T_{max}
PF01385	OrfB_IS605	Probable transposase	7.7 bp
PF01850	PIN	PIN domain	8.3 bp
PF00395	SLH	S-layer homology domain	8.3 bp
PF01590	GAF	GAF domain	9.1 bp
PF04055	Radical_SAM	Radical SAM superfamily	11.1 bp
PF00924	MS_channel	Mechanosensitive ion channel	11.1 bp
PF01527	HTH_Tnp_1	helix-turn-helix motif, transposase	11.1 bp
PF01710	HTH_Tnp_IS630	helix-turn-helix motif, transposase	11.1 & 12.5 bp
PF00805	Pentapeptide	Pentapeptide repeat	14.3 bp
PF00070	Pyr_redox	Pyridine nucleotide-disulphide oxidoreductase	14.3 bp
PF01609	DDE_Tnp_1	Transposase DDE domain	9.1 & 25 bp

Fig. 14: **Protein domain annotation enrichment of *Synechocystis* sp. PCC 6803 (left) and *Cyanothece* sp. PCC 8801 (right) CDS periodicity clusters in Figure 4A of the main article and Figure S9.** The p-values were calculated by cumulative hypergeometric distribution tests and p-values before ('p-value') and after adjustment for multiple testing by the Benjamini-Hochberg method ('BH') are shown. Individual domains from the prediction algorithms HMMPfam, HMMTigr, Gene3D, HMMPanther, HMMSmart, BlastProDom, ProfileScan, FPrintScan, and HMMPIR were tested for PCC 6803 (total: 6813 domains) and PCC 8801 (total: 6599 domains). All overlaps that were significant ($p \leq 0.01$) before BH control are shown. For both species only HMMPfam annotations yielded significant results. The domain names and descriptions of enriched domains are provided together with the main periods T_{max} of the respective CDS periodicity clusters. Summary: the pentapeptide repeat domain (PF00805) which coincidentally may mimic DNA structure and inhibit DNA gyrase [68], is found in clusters with $T_{max} = 14{:}3$ bp in both species and likely reflects secondary effects of the amino acid code (pentapeptide repeat: 5 amino acids = 15 bp). None of the enrichments could comprehensively explain the CDS periodicity profiles by such secondary effects.

Cluster-wise Category Annotation Enrichment 6803			
Category	p-value/ BH	in Cluster	total
Cluster 2 (26 genes, T_{max} 7.7 bp)			
Chaperones	0.007/ 0.48	2	16
Cluster 5 (73 genes, T_{max} 10 bp)			
Fatty acid, phospholipid and sterol metabolism	0.0076/ 0.14	3	39
Cluster 7 (23 genes, T_{max} 11.1 bp)			
Other categories	2.6e-18/ 4.6e-17	20	305
Transposon-related functions	1.6e-23/ 1.1e-21	18	107
Cluster 9 (12 genes, T_{max} 12.5 bp)			
Other categories	5.1e-11/ 9.2e-10	11	305
Transposon-related functions	6.7e-14/ 4.6e-12	10	107
Cluster 13 (90 genes, T_{max} 25 bp)			
Soluble electron carriers	6.1e-4/ 0.042	4	15
Cluster 14 (35 genes, T_{max} 25 bp)			
Cellular processes	0.0086/ 0.15	4	76
NADH dehydrogenase	0.0016/ 0.11	3	22

Cluster-wise Category Annotation Enrichment 8801			
Category	p-value/ BH	in Cluster	total
Cluster 1 (22 genes, T_{max} 7.7 bp)			
Replication, recombination and repair	2.8e-4/ 0.0066	5	120
Cluster 4 (139 genes, T_{max} 10 bp)			
Amino acid transport and metabolism	0.0056/ 0.079	11	142
Translation, ribosomal structure and biogenesis	0.0066/ 0.079	11	145
Cluster 5 (156 genes, T_{max} 11.1 bp)			
Replication, recombination and repair	0.0038/ 0.077	11	120
Coenzyme transport and metabolism	0.0064/ 0.077	10	111
Cluster 7 (156 genes, T_{max} 11.1 bp)			
Secondary metabolites biosynthesis, transport and catabolism	0.0071/ 0.085	6	48
Defense mechanisms	0.0036/ 0.085	6	42
Cluster 8 (138 genes, T_{max} 11.1 bp)			
Multiple Categories	0.009/ 0.22	17	286
Cluster 11 (44 genes, T_{max} 14.3 bp)			
Function unknown	5.4e-07/ 1.3e-05	15	338

Fig. 15: **Function enrichments in CDS clusters.** As Table S1 but for CyanoBase Function categories of *Synechocystis* sp. PCC 6803 (top, total: 84 categories) and *Cyanothece* sp. PCC 8801 (bottom, total: 23 categories).

Appendix C

4 *Neurospora crassa* RNA polymerase II binding and mRNA abundance

4.1 Defining expressed genes

The logarithmic gene-wise median RNAPII occupancy and mRNA abundance distributions are shown in Figure 16 A and B, respectively. The mRNA abundance distribution exhibited a peak at 9 with an extensive shoulder towards the lower end with a small peak around 0, indicating genes with expression levels below detection limit. To exclude transcripts with too low signals in both datasets, all transcripts with a logarithmic median signal below 3 were discarded from further analyses, which removes 2 transcripts from the RNAPII occupancy and 1218 from the mRNA abundance dataset.

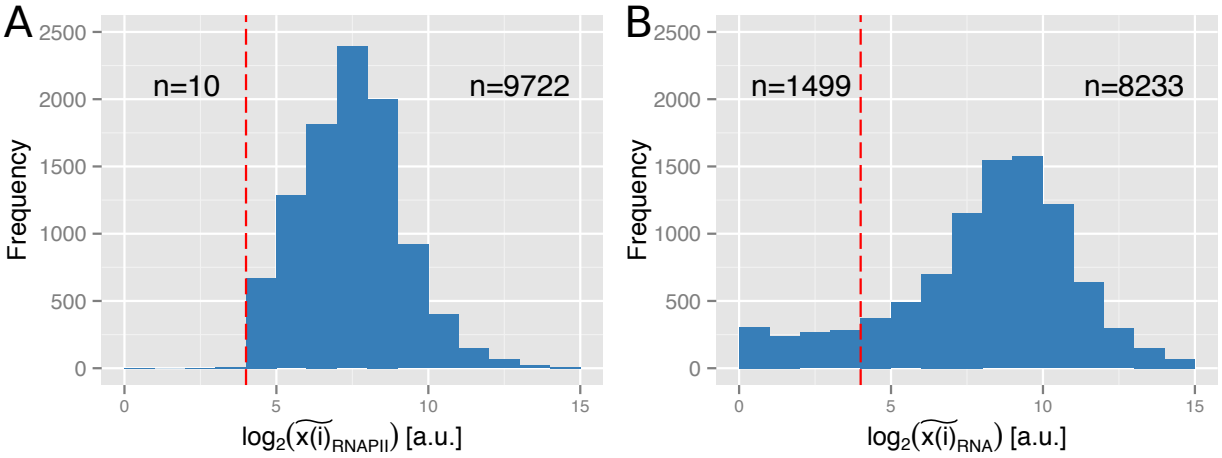


Fig. 16: **Median gene-wise signals for (A) RNAPII occupancy levels $x(i)_{RNAPII}$ and (B) transcript abundance $x(i)_{RNA}$.** The minimal threshold for admission for further analyses is marked as red dashed line, with the number of genes below and above shown left and right to it.

4.2 Linear detrending in short diurnal time series

The Fourier transform assumes that the signal is stationary and that the signals in the sample continue infinitely. If these assumptions are not met, the Fourier transform performs poorly. Hence, the removal of linear trends is a common preprocessing step

prior to a classical Fourier analysis or also harmonic regression [259]. However, linear detrending can have significant impact on the observed phase distributions. The phase distributions obtained from raw data using harmonic regression without and with a linear trend in the the model (internal detrending), as well as from linear detrended data (external detrending) are shown in Figure 17. Observed mRNA phases fell mostly into two large clusters (Figure 17 right column) around CT 8 and CT 20 in the raw data (B). Linear detrending (internal and external) consistently yielded peak phases delayed by 3 hours (D), where internal detrending resulted in a more dispersed phase cluster around CT 0 (F). In case of RNAPII occupancy, a wider range of phases was obtained from the raw data with one distrinct peak around CT 2 (A) and a minor cluster around CT 14. Consistent with the mRNA abundance dataset, internal detrending flattened the phase distribution and yielded more significant oscillators compared to the raw data (C). More importantly, external detrending led to a marked condensation of two phase clusters around CT 2 and 14, resembling the mRNA phase distribution, and less significant oscillators (E). This posed the question, if external detrending selectively suppresses profiles with phases other than ~0 and ~12 or changes the observed phases.
The direct comparison of phases obtained without detrending and with internal detrending (Fig. 18 A) showed that internal detrending alters the observed phases around 0 and 12 CT but preserves the significant oscillation (grey = significant with and without detrending). In contrast, significant oscillators with phases around 12 and 18 CT in the raw data are not changed but fell below the significance threshold in the internally detrended data (blue). External detrending showed a similar effect, but additionally introduced a bias of the preserved oscillators towards 0 and 12 CT phases, which then resulted the described narrow phase clusters (see Figure 18 B). Comparison of internal and external detrending-derived phases demonstrated that both have a similar effect but the latter introduced a stronger bias (see Figure 18 C). The observed effects of linear detrending on individual gene profiles are demonstrated in Figure 19.

4.3 Oscillation amplitude is independent from sequencing depth

To test whether the sequencing depth in the presented datasets was high enough to resolve circadian oscillations the median RNAPII occupancy and mRNA abundance signals for each gene were compared with the corresponding estimated amplitudes (see Fig. 20 A and B). No dependency was found for either dataset.

4.4 Functional enrichment analysis

The oscillatory RNAPII gene set was separated into dawn (CT $1 > \phi_{RNAPII} >$ CT 18) and dusk-phased (CT $8 > \phi_{RNAPII} >$ CT 10), before subjecting each set to functional enrichment analysis. This revealed a fundamental separation of cellular tasks between day and night (see Table 3). Enriched GO annotations of dawn peaking genes reflect stress response to light and salt. Different aspects of anabolic processes appeared enriched, such as the GO term for to thiamine, amino acid, and cellular nitrogen compound biosynthesis. FunCat enrichment confirmed these results, most importantly yielding

	Dawn (464)				
Id	Term	An.	Sig.	Exp.	q-value
0009416	response to light stimulus	136	16	6.92	3e-04
0009228	thiamine biosynthetic process	2	2	0.1	0.003
0009651	response to salt stress	4	2	0.2	0.01
0008652	cellular amino acid biosynthetic process	24	4	1.22	0.03
0044271	cellular nitrogen compound biosynthetic process	18	5	0.92	0.04
FunCat Id	Term		Sig.		p-value
01	metabolism		123		1.2e-10
10	cell cycle and DNA processing		5		2.3e-08
14	protein fate (folding, modification, destination)		15		9.1e-07
12	protein synthesis		2		5e-06
16	protein with binding function or cofactor requirement		37		1.8e-05
11	transcription		16		0.00071
30	cellular communication/signal transduction mechanism		4		0.005
42	biogenesis of cellular components		12		0.005
18	regulation of metabolism and protein function		4		0.04
32	cell rescue, defense and virulence		38		0.04
	Dusk (336)				
Id	Term	An.	Sig.	Exp.	q-value
0006364	rRNA processing	12	6	0.73	2e-05
0042273	ribosomal large subunit biogenesis	8	3	0.49	0.01
0044260	cellular macromolecule metabolic process	164	20	10.01	0.02
0034622	cellular macromolecular complex assembly	13	3	0.79	0.04
0006974	response to DNA damage stimulus	13	3	0.79	0.04
0090304	nucleic acid metabolic process	62	12	3.78	0.04
0050794	regulation of cellular process	23	4	1.4	0.05
0000054	ribosomal subunit export from nucleus	6	2	0.37	0.05
FunCat Id	Term		Sig.		p-value
11	transcription		63		2.1e-12
12	protein synthesis		31		9.1e-06
10	cell cycle and DNA processing		42		6.8e-05
20	cellular transport, facilitation and routes		19		0.0003
16	protein with binding function or cofactor requirement		71		0.002
01	metabolism		49		0.006
40	cell fate		16		0.035
02	energy		7		0.04

Table 3: **GO enrichment analysis of genes with circadian RNAPII occupancy profiles separated into dawn and dusk-specific.** Gene sets are defined by peak RNAPII binding where dawn corresponds to phases between CT $18-4$ and dusk to CT $5-14$. The number of genes in the dawn and dusk set is given in parentheses. The column "Id" lists the gene ontology id of the term specified in the second column "Term". Column "An." provides the number of term occurrences in the entire dataset (background and foreground), followed by the number of occurrences in the foreground "Sig.", the number of expected occurrences "Exp." and the q-value.

a large number of genes with the FunCat annotation "Metabolism", but also "cell rescue, defense, and virulence" and "biogenesis of cellular compounds". In contrast, dusk specific genes are enriched for catabolic processes. Protein synthesis related GO annotations such as rRNA processing, ribosome biogenesis and nuclear export, cellular macromolecule metabolism and complex formation were found to be enriched. FunCat results complemented the picture, yielding numerous "transcription", "protein synthesis", and "cell cycle and DNA processing" related genes. Interestingly, protein synthesis-related genes were expressed during dusk while genes involved in crucial protein processing steps like "folding, modification, destination" are only expressed 11 h later during dawn.
In summary, *Neurospora crassa* exhibited fundamentally different transcriptional programs during day and night. Transcription of catabolism related genes in the morning was found to be separated from anabolism related genes in the evening. As expected, many light response, salt stress, and cell defense related genes peak during the morning. More importantly, a large number of metabolic genes for the synthesis of amino acids, nitrogen compounds, and specifically thiamine are also transcribed during the morning. A large number of genes involved in transcription and translation, as well as the synthesis of macromolecules were found amongst the evening peaking group. These results suggest that *Neurospora crassa* generates metabolic precursors during the day, while handling higher cellular stress, to synthesise cellular components, grow, and divide during the night, when cellular stress is lower. These observations agree very well with previous results reported by Hurley *et al.* [80].

4.5 Variability in RNAPII and mRNA profiles

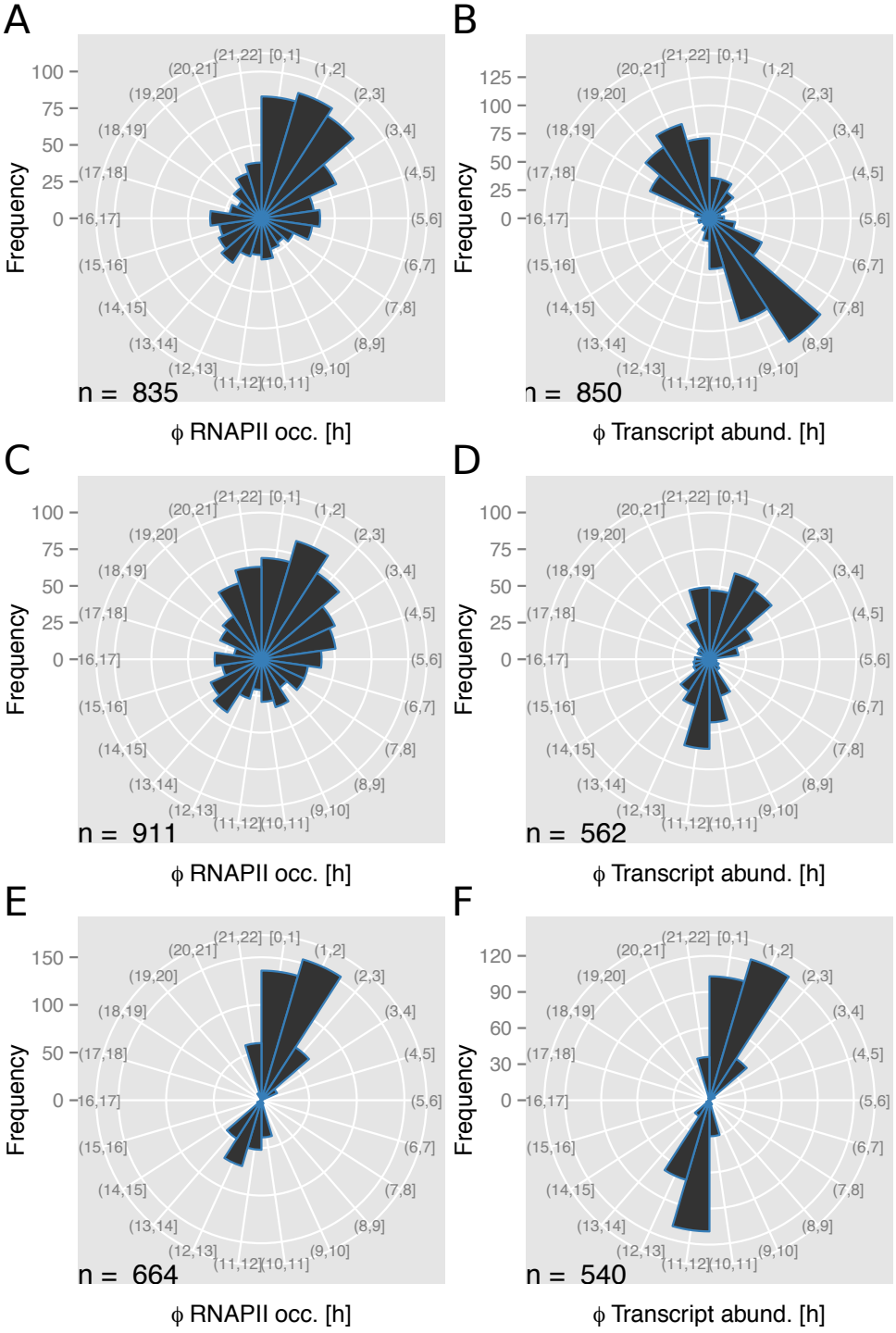


Fig. 17: **Linear detrending can change observed circadian phases.** Observed peak phase distributions of the RNAPII occupancy (left column) and the mRNA abundance datasets (right column), without any detrending (**A**,**B**), with a linear trend included in the harmonic regression model (**C**,**D**), and with linear detrending performed separately before harmonic regression (E,F). The raw datasets were `l2mtransformed` prior to any processing. Only transcripts with a harmonic regression $p < 0.025$ were considered in this comparison.

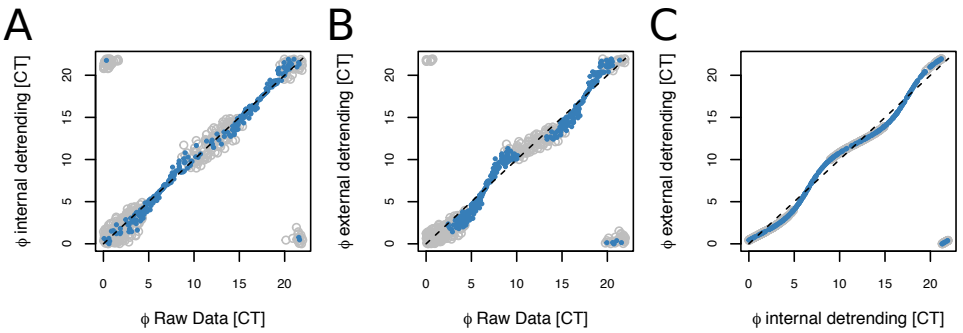


Fig. 18: **Linear detrending changes some phases and suppresses oscillation for others.** (**A**) Comparison of phases obtained from the raw data (x-axis) with those from internally detrended data (y-axis). Genes which oscillate significantly in both dataset sets are shown in grey, whereas genes which oscillate significantly only in the raw data are shown in blue. (**B**) Comparison of phases from obtained from raw data (x-axis) and externally detrended data (y-axis), and (**C**) phases from internally (x-axis) and externally detrended data.

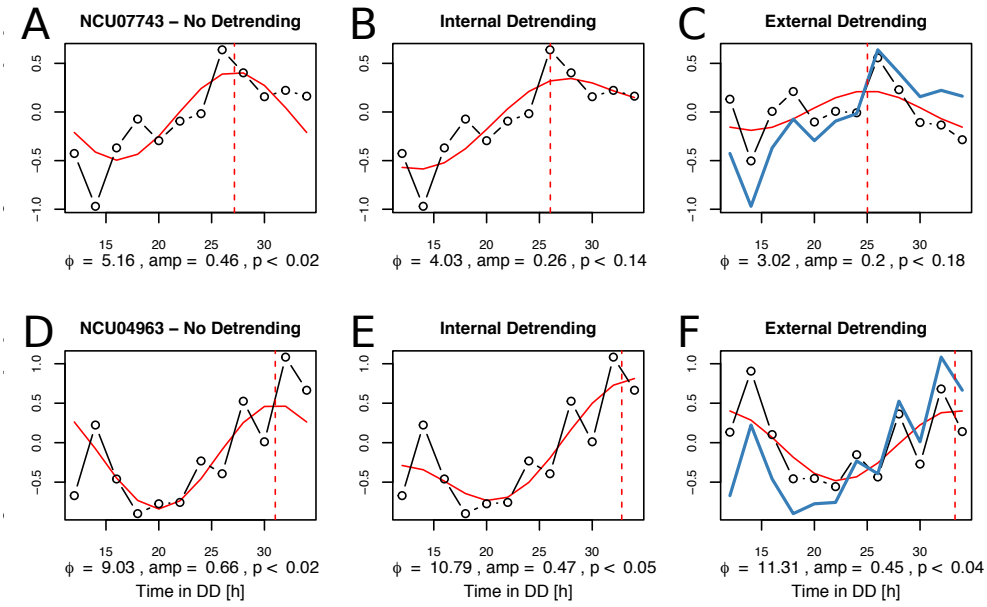


Fig. 19: **The two genes NCU05855 and NCU09823 exemplify how detrending can change the observed phase or suppress oscillation.** The harmonic regression results without detrending (**A**, **D**), internal detrending (**B**, **E**), and external detrending (**C**, **F**) are shown for the RNAPII occupancy profiles of the genes *NCU05855* (top) and *NCU09823* (bottom). The `l2m`transformed profiles are shown in black, the harmonic regression curve in red, the original data in blue in panels (C, F), and the obtained phase is marked with a vertical red dashed line. The oscillatory parameters are shown below each plot.

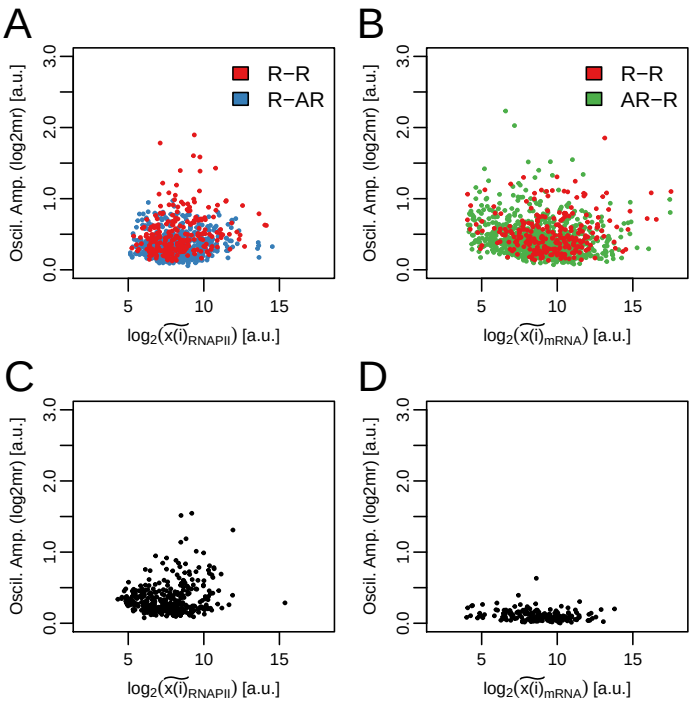


Fig. 20: **Oscillatory amplitude does not depend on expression level.** The median logarithmic RNAPII occupancy (**A**, x-axis) and mRNA abundance (**B**, x-axis) of circadian oscillating genes is compared with the corresponding amplitude (y-axis), distinguishing the R-R set and the sets of only RNAPII or mRNA oscillating genes. The same comparison for all genes with ultradian RNAPII or mRNA profile is shown in **C** and **D**, respectively.

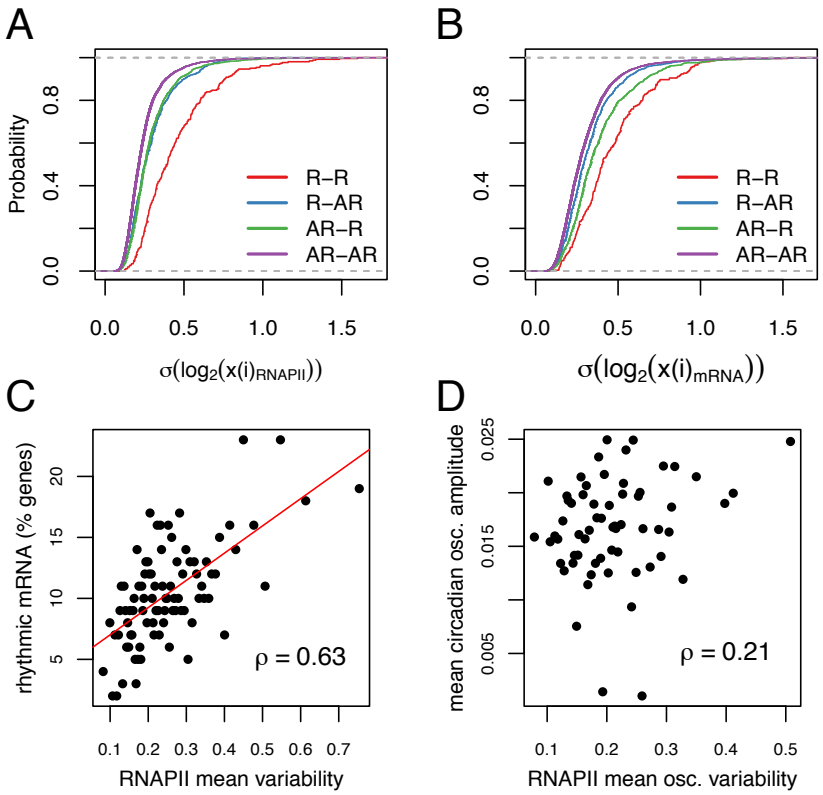


Fig. 21: **Standard deviation of RNAPII occupancy signal and mRNA abundance, distinguishing the four groups of circadian and non-circadian profiles.** The gene-wise standard deviations σ of the `l2m`transformed RNAPII occupancy (**A**) and the mRNA abundance dataset (**B**), shown as empirical cumulative density distribution and distinguishing the oscillatory gene sets. (**C**) Higher RNAPII σ is weakly associated with more rhythmic mRNA expression. All genes without significant RNAPII rhythm were binned (n = 100), and the mean RNAPII standard deviation (x-axis) was compared with the percentage of circadian mRNA occurrences. (**D**) RNAPII standard deviation of arhythmic genes does not influence the observed amplitude of the corresponding circadian mRNA profiles.

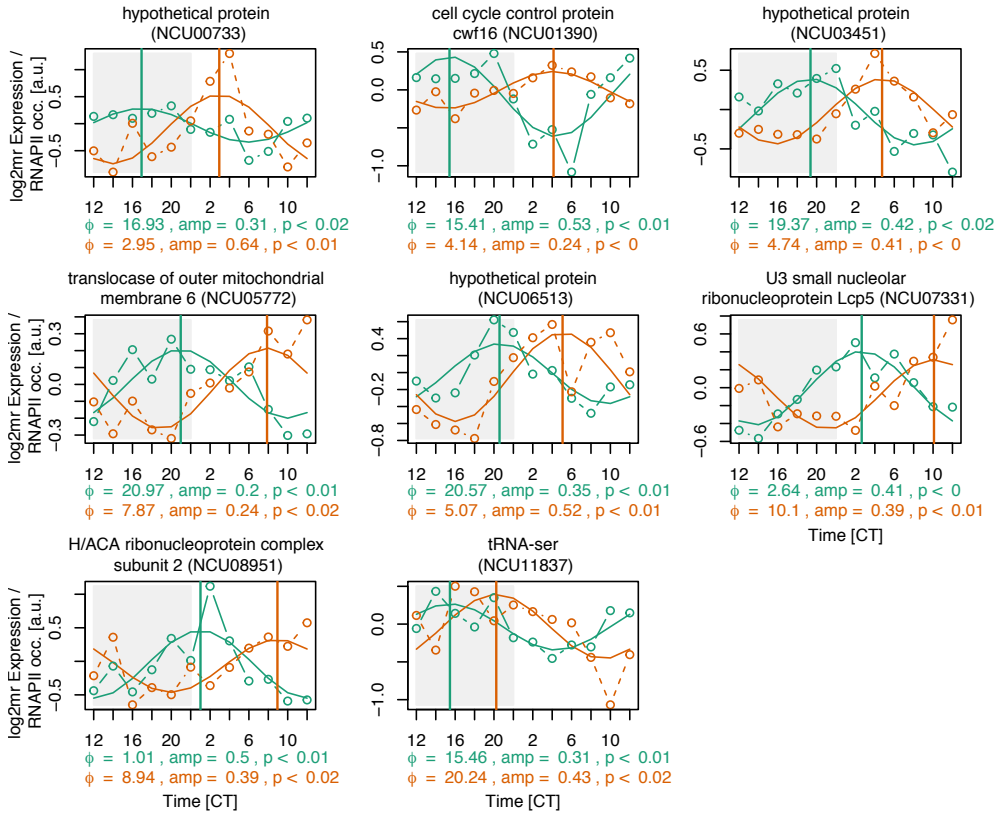


Fig. 22: **Profiles of genes with largest phase delay between RNAPII and mRNA signals.**

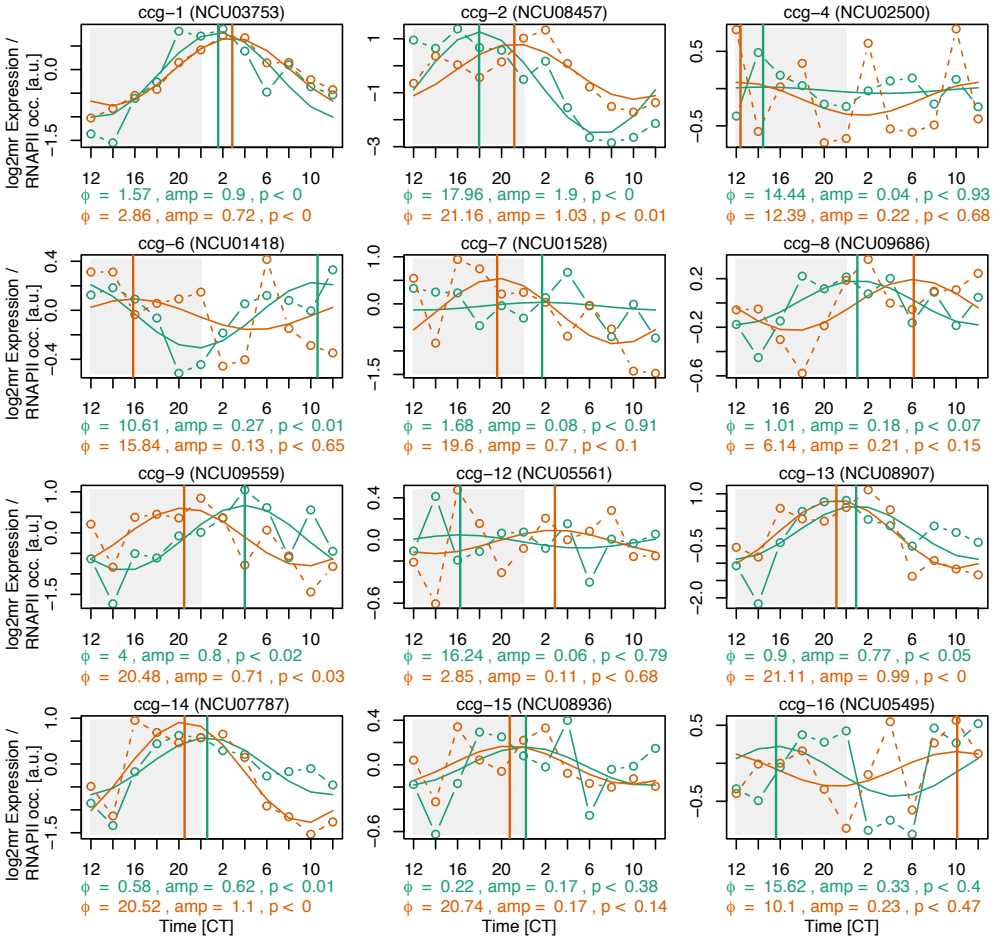


Fig. 23: **RNAPII occupancy and mRNA profiles of known clock controlled genes in *Neurospora crassa***

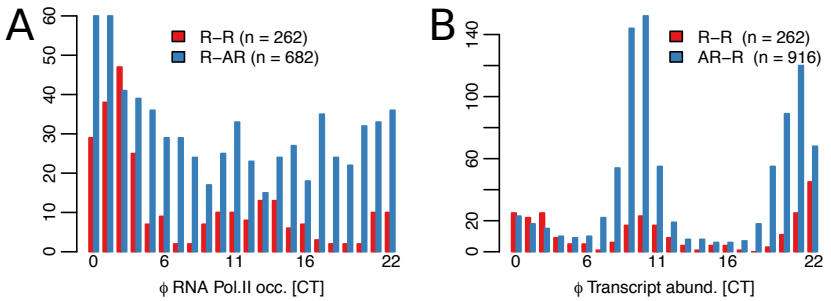


Fig. 24: **Comparison of phase distributions between R-R, R-AR, and AR-R gene sets.** (**A**) RNAPII peak phase distributions for the R-AR gene set compared to the R-R set. (**B**) mRNA abundance peak phase distributions for the AR-R gene set compared to the R-R set.

Fig. 25: **Similarity between known TF binding site models of the fungus *S. cerevisiae* and the overrepresented motifs of group all and I-III.** For each group, the six most overrepresented motifs (Motif 1-6, Logo on top) are shown (order as in Fig. 4.9) together with the two most similar *S. cerevisiae* TF binding site models (logo left, name and similarity p-value right). Only group all and I-III are shown here, for group IV see Fig. 26. Similarity between the set of 177 *S. cerevisiae* TFs and each motif was measured using STAMP, as implemented in the MotIV R package.

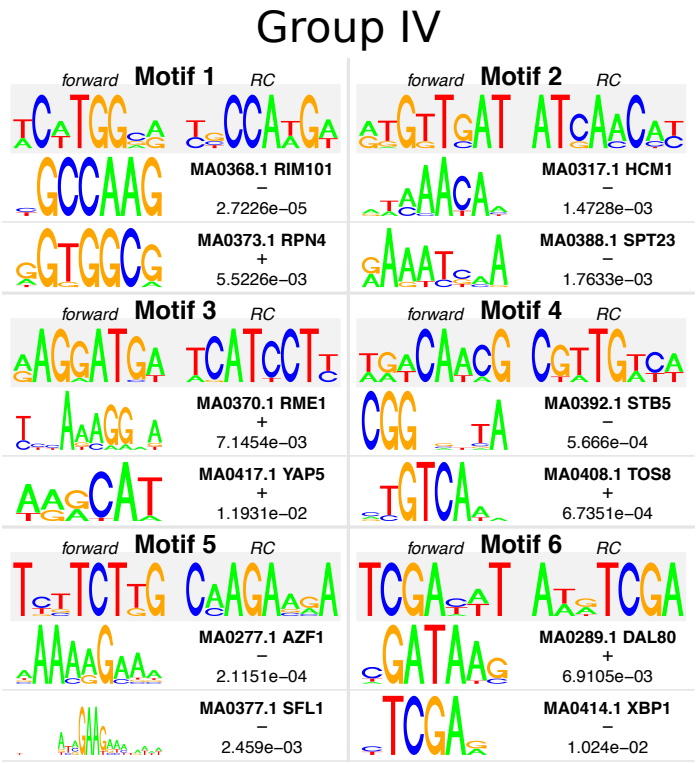


Fig. 26: **Similarity between known TF binding site models of the fungus *S. cerevisiae* and the overrepresented motifs of group IV.** Structure similar to Fig. 25.

Glossary

CDS Coding sequence (CDS) is the sequence region of a gene between the start and stop codon(s).

Circadian Oscillator The core mechanism of a the circadian clock.

Circadian Rhythm Endogenously generated rhythmic patterns with a period of 24 h, similar to the rotation of the Earth.

Circadian System The circadian system includes the circadian oscillator, *i.e.* the molecular core mechanism, and its input and output processes that connect the oscillator to the cell.

CLOG In the phylogenetic context, CLOGSs refer to Clusters of Likely Orthologous Genes, i.e. genes with similar genetic origin and thus highly similar sequence and often function.

Diurnal Process The term diurnal refers to rhythmic processes with a period of 24 h which is synchronised with the day/night cycle, which does not persist without zeitgeber input.

Infradian Process The term diurnal refers to rhythmic processes with a period significantly longer than 24 h.

LD LD in the circadian field refers to the experimental condition with light on during the subjective day and light off during the night.

LL LL in the circadian field refers to the experimental condition with constant light on during the subjective day and night.

Nucleoid The nucleoid is an irregularly-shaped region in the prokaryotic cell which holds all or most of the genetic material.

Operon Operons are neighbouring genes that are transcribed together resulting in the formation of polycistronic mRNA.

PTO Post-transcriptional Oscillator.

TF Transcription Factor, a protein which induces or represses gene transcription by binding to the proximal or distal promoter DNA of its target genes.

TFBS Transcription factor binding site on the DNA strand.

TTO Transcription-Translation Oscillator.

Ultradian Process The term diurnal refers to rhythmic processes with a period significantly shorter than 24 h.

Zeitgeber External rhythmic parameter which is used as reference to synchronise the internal clock of an organism.

Bibliography

[1] U. Albrecht and G. Eichele. "The mammalian circadian clock." In: *Current opinion in genetics & development* 13.3 (June 2003), pp. 271–7.

[2] A. Alexa, J. Rahnenführer, and T. Lengauer. "Improved scoring of functional groups from gene expression data by decorrelating GO graph structure." In: *Bioinformatics (Oxford, England)* 22.13 (2006), pp. 1600–1607.

[3] B. Ananthasubramaniam and H. Herzel. "Positive feedback promotes oscillations in negative feedback loops." In: *PLoS ONE* 9.8 (Jan. 2014), e104761.

[4] A. K. And and L. Reinhold. "CO2 concentrating mechanisms in photosynthetic microorganisms". In: *Annual Review of Plant Physiology and Plant Molecular Biology* 50 (Nov. 2003), pp. 539–570.

[5] S. Anders and W. Huber. "Differential expression analysis for sequence count data." In: *Genome Biology* 11.10 (Jan. 2010), R106.

[6] J Aschoff and H Pohl. "Phase relations between a circadian rhythm and its zeitgeber within the range of entrainment." In: *Die Naturwissenschaften* 65.2 (Feb. 1978), pp. 80–4.

[7] R. Aurora et al. "A network of genes regulated by light in cyanobacteria." In: *Omics : a journal of integrative biology* 11.2 (Jan. 2007), pp. 166–85.

[8] I. M. Axmann et al. "Biochemical evidence for a timing mechanism in prochlorococcus." In: *Journal of Bacteriology* 191.17 (Sept. 2009), pp. 5342–7.

[9] I. M. Axmann et al. "Diversity of KaiC-based timing systems in marine Cyanobacteria." In: *Marine genomics* 14 (Apr. 2014), pp. 3–16.

[10] Z. Bar-Joseph et al. "Continuous representations of time-series gene expression data." In: *Journal of Computational Biology* 10.3-4 (Jan. 2003), pp. 341–56.

[11] C. Beck et al. "A daily expression pattern of protein-coding genes and small non-coding RNAs in Synechocystis sp. PCC 6803." In: *Applied and environmental microbiology* 80.17 (June 2014), pp. 5195–5206.

[12] C. Beck et al. "The diversity of cyanobacterial metabolism: genome analysis of multiple phototrophic microorganisms." In: *BMC genomics* 13.1 (Feb. 2012), p. 56.

[13] D Bell-Pedersen et al. "Circadian clock-controlled genes isolated from Neurospora crassa are late night- to early morning-specific." In: *Proceedings of the National Academy of Sciences* 93.23 (Nov. 1996), pp. 13096–101.

[14] Y. Benjamini and Y. Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *Journal of the Royal Statistical Society* 57.1 (1995), pp. 289 –300.

[15] L. D. Bennett et al. “Circadian activation of the mitogen-activated protein kinase MAK-1 facilitates rhythms in clock-controlled genes in Neurospora crassa.” In: *Eukaryotic cell* 12.1 (Jan. 2013), pp. 59–69.

[16] D. A. Benson et al. “GenBank.” In: *Nucleic Acids Research* 41.Database issue (Jan. 2013), pp. D36–42.

[17] I. Berman-Frank, P. Lundgren, and P. Falkowski. “Nitrogen fixation and photosynthetic oxygen evolution in cyanobacteria.” In: *Research in Microbiology* 154.3 (Apr. 2003), pp. 157–64.

[18] F. Bidard, E. Coppin, and P. Silar. “The transcriptional response to the inactivation of the PaMpk1 and PaMpk2 MAP kinase pathways in Podospora anserina.” In: *Fungal genetics and biology* 49.8 (Aug. 2012), pp. 643–52.

[19] B. Binder and S. Chisholm. “Cell Cycle Regulation in Marine Synechococcus sp. Strains”. In: *Applied and environmental microbiology* 61.2 (Feb. 1995), pp. 708–717.

[20] H. Binder, K. Krohn, and S. Preibisch. “"Hook"-calibration of GeneChip-microarrays: Chip characteristics and expression measures”. In: *Algorithms for Molecular Biology* 3.1 (2008), p. 11.

[21] H. Binder and S. Preibisch. “GeneChip microarrays—signal intensities, RNA concentrations and probe sequences”. In: *Journal of Physics: Condensed Matter* 18.18 (May 2006), S537–S566.

[22] C. E. Blank and P Sánchez-Baracaldo. “Timing of morphological and ecological innovations in the cyanobacteria–a key to understanding the rise in atmospheric oxygen.” In: *Geobiology* 8.1 (Jan. 2010), pp. 1–23.

[23] J. Bohlin, S. P. Hardy, and D. W. Ussery. “Stretches of alternating pyrimidine/purines and purines are respectively linked with pathogenicity and growth temperature in prokaryotes.” In: *BMC genomics* 10 (Jan. 2009), p. 346.

[24] A. Bolshoy et al. “Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles.” In: *Proceedings of the National Academy of Sciences* 88.6 (Mar. 1991), pp. 2312–2316.

[25] B. Bolstad et al. “A comparison of normalization methods for high density oligonucleotide array data based on variance and bias”. In: *Bioinformatics (Oxford, England)* 19.2 (2003), p. 185.

[26] K. Bozek et al. “Regulation of clock-controlled genes in mammals.” In: *PLoS ONE* 4.3 (Jan. 2009), e4882.

[27] K. Brogaard et al. “A map of nucleosome positions in yeast at base-pair resolution.” In: *Nature* 486.7404 (June 2012), pp. 496–501.

[28] S. Calza, D. Valentini, and Y. Pawitan. "Normalization of oligonucleotide arrays based on the least-variant set of genes". In: *BMC Bioinformatics* 9.140 (2008), p. 140.

[29] S. Casteret et al. "Physical properties of DNA components affecting the transposition efficiency of the mariner Mos1 element." In: *Molecular genetics and genomics* 282.5 (Nov. 2009), pp. 531–46.

[30] J. Cervený and L. Nedbal. "Metabolic rhythms of the cyanobacterium Cyanothece sp. ATCC 51142 correlate with modeled dynamics of circadian clock." In: *Journal of Biological Rhythms* 24.4 (Aug. 2009), pp. 295–303.

[31] A. H. Chen et al. "Spatial and temporal organization of chromosome duplication and segregation in the cyanobacterium Synechococcus elongatus PCC 7942." In: *PLoS ONE* 7.10 (Jan. 2012), e47837.

[32] C.-H. Chen et al. "Genome-wide analysis of light-inducible responses reveals hierarchical light signalling in Neurospora." In: *The EMBO journal* 28.8 (Apr. 2009), pp. 1029–42.

[33] J. L. Chinnici et al. "Neurospora crassa female development requires the PACC and other signal transduction pathways, transcription factors, chromatin remodeling, cell-to-cell fusion, and autophagy." In: *PloS ONE* 9.10 (Jan. 2014), e110603.

[34] M. Chiogna et al. "A comparison on effects of normalisations in the detection of differentially expressed genes." In: *BMC Bioinformatics* 10 (Jan. 2009), p. 61.

[35] R. J. Cho et al. "A genome-wide transcriptional analysis of the mitotic cell cycle." In: *Molecular Cell* 2.1 (July 1998), pp. 65–73.

[36] C. Claeys Bouuaert, D. Liu, and R. Chalmers. "A simple topological filter in a eukaryotic transposon as a mechanism to suppress genome instability." In: *Molecular and cellular biology* 31.2 (Jan. 2011), pp. 317–27.

[37] A. B. Cohanim, E. N. Trifonov, and Y. Kashi. "Specific selection pressure at the third codon positions: contribution to 10- to 11-base periodicity in prokaryotic genomes." In: *Journal of Molecular Evolution* 63.3 (Sept. 2006), pp. 393–400.

[38] A. Correa et al. "Multiple oscillators regulate circadian gene expression in Neurospora." In: *Proceedings of the National Academy of Sciences* 100.23 (Nov. 2003), pp. 13597–602.

[39] E. DeLong, G. Wickham, and N. Pace. "Phylogenetic stains: ribosomal RNA-based probes for the identification of single cells". In: *Science (New York, N.Y.)* 243.4896 (Mar. 1989), pp. 1360–1363.

[40] S. C. Dillon and C. J. Dorman. "Bacterial nucleoid-associated proteins, nucleoid structure and gene expression." In: *Nature reviews. Microbiology* 8.3 (Mar. 2010), pp. 185–95.

[41] W. D. Donachie. "Relationship between Cell Size and Time of Initiation of DNA Replication". In: *Nature* 219.5158 (Sept. 1968), pp. 1077–1079.

[42] J. Dopazo and J. Carazo. "Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree". In: *Journal of Molecular Evolution* 44.2 (1997), pp. 226–233.

[43] J. Dostie et al. "Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements." In: *Genome Research* 16.10 (Oct. 2006), pp. 1299–309.

[44] J. C. Dunlap. "Molecular bases for circadian clocks." In: *Cell* 96.2 (Jan. 1999), pp. 271–90.

[45] V. Dvornyk, O. Vinogradova, and E. Nevo. "Origin and evolution of circadian clock genes in prokaryotes." In: *Proceedings of the National Academy of Sciences* 100.5 (Mar. 2003), pp. 2495–500.

[46] R. S. Edgar et al. "Peroxiredoxins are conserved markers of circadian rhythms." In: *Nature* 485.7399 (May 2012), pp. 459–64.

[47] M. B. Eisen et al. "Cluster analysis and display of genome-wide expression patterns." In: *Proceedings of the National Academy of Sciences* 95.25 (Dec. 1998), pp. 14863–8.

[48] C. Elmerich and W. E. Newton, eds. *Associative and Endophytic Nitrogen-fixing Bacteria and Cyanobacterial Associations*. Vol. 5. Nitrogen Fixation: Origins, Applications, and Research Progress. Dordrecht: Springer Netherlands, 2007.

[49] M. Fasold, P. F. Stadler, and H. Binder. "G-stack modulated probe intensities on expression arrays - sequence corrections and signal calibration". In: *BMC Bioinformatics* 11.Mm (2010), p. 207.

[50] A. Favorov et al. "Exploring massive, genome scale datasets with the GenometriCorr package." In: *PLoS Computational Biology* 8.5 (May 2012), e1002529.

[51] X. Feng and S. D. Colloms. "In vitro transposition of ISY100, a bacterial insertion sequence belonging to the Tc1/mariner family." In: *Molecular Microbiology* 65.6 (Sept. 2007), pp. 1432–43.

[52] E. Forgy. "Cluster analysis of multivariate data: efficiency versus interpretability of classifications". In: *Biometrics* 21 (1965), pp. 768–769.

[53] E. Freyhult et al. "Challenges in microarray class discovery: a comprehensive examination of normalization, gene selection and clustering." In: *BMC Bioinformatics* 11.1 (Jan. 2010), p. 503.

[54] M. E. Futschik and H. Herzel. "Are we overestimating the number of cell-cycling genes? The impact of background models on time-series analysis." In: *Bioinformatics (Oxford, England)* 24.8 (Apr. 2008), pp. 1063–9.

[55] J. E. Galagan et al. "Genomics of the fungal kingdom: insights into eukaryotic biology." In: *Genome Research* 15.12 (Dec. 2005), pp. 1620–31.

[56] J. Georg et al. "Evidence for a major role of antisense RNAs in cyanobacterial gene regulation." In: *Molecular Systems Biology* 5.305 (Jan. 2009), p. 305.

[57] F. M. Giorgi et al. “Algorithm-driven artifacts in median Polish summarization of microarray data.” In: *BMC Bioinformatics* 11.1 (Jan. 2010), p. 553.

[58] B. R. G. Gordon et al. “Structural basis for recognition of AT-rich DNA by unrelated xenogeneic silencing proteins.” In: *Proceedings of the National Academy of Sciences* 108.26 (June 2011), pp. 10690–5.

[59] R. L. Gourse et al. “rRNA transcription and growth rate-dependent regulation of ribosome synthesis in Escherichia coli.” In: *Annual review of microbiology* 50 (Jan. 1996), pp. 645–77.

[60] M. Griese, C. Lange, and J. Soppa. “Ploidy in cyanobacteria.” In: *FEMS Microbiology Letters* 323.2 (Oct. 2011), pp. 124–31.

[61] I. V. Grigoriev et al. “The genome portal of the Department of Energy Joint Genome Institute.” In: *Nucleic Acids Research* 40.Database issue (Jan. 2012), pp. D26–32.

[62] N. Grobbelaar et al. “Dinitrogen-fixing endogenous rhythm in Synechococcus RF-1”. In: *FEMS Microbiology Letters* 37.2 (1986), pp. 177–173.

[63] A. C. L. Guerreiro et al. “Daily rhythms in the cyanobacterium Synechococcus elongatus probed by high-resolution mass spectrometry based proteomics reveals a small-defined set of cyclic proteins.” In: *Molecular & cellular proteomics* 13 (Mar. 2014), pp. 2042–2055.

[64] N Halaimia-Toumi et al. “The GC-rich transposon Bytmar1 from the deep-sea hydrothermal crab, Bythograea thermydron, may encode three transposase isoforms from a single ORF.” In: *Journal of Molecular Evolution* 59.6 (Dec. 2004), pp. 747–60.

[65] F Halberg et al. “Prokaryotic and eukaryotic unicellular chronomics.” In: *Biomedicine & pharmacotherapy* 59 Suppl 1 (Oct. 2005), S192–202.

[66] S. L. Harmer. “Orchestrated Transcription of Key Pathways in Arabidopsis by the Circadian Clock”. In: *Science (New York, N.Y.)* 290.5499 (Dec. 2000), pp. 2110–2113.

[67] Q. He and Y. Liu. “Molecular mechanism of light responses in Neurospora: from light-induced transcription to photoadaptation.” In: *Genes & Development* 19.23 (Dec. 2005), pp. 2888–99.

[68] S. S. Hegde et al. “A fluoroquinolone resistance protein from Mycobacterium tuberculosis that mimics DNA.” In: *Science (New York, N.Y.)* 308.5727 (June 2005), pp. 1480–3.

[69] B ten Heggeler-Bordier et al. “The apical localization of transcribing RNA polymerases on supercoiled DNA prevents their rotation around the template.” In: *The EMBO journal* 11.2 (Feb. 1992), pp. 667–72.

[70] J Herrero, A Valencia, and J Dopazo. “A hierarchical unsupervised growing neural network for clustering gene expression patterns.” In: *Bioinformatics (Oxford, England)* 17.2 (2001), pp. 126–136.

[71] S. Hertel, C. Brettschneider, and I. M. Axmann. "Revealing a two-loop transcriptional feedback mechanism in the cyanobacterial circadian clock." In: *PLoS Computational Biology* 9.3 (Jan. 2013), e1002966.

[72] H. Herzel, O. Weiss, and E. N. Trifonov. "10-11 bp periodicities in complete genomes reflect protein structure and DNA folding." In: *Bioinformatics (Oxford, England)* 15.3 (Mar. 1999), pp. 187–93.

[73] H. Herzel, O. Weiss, and E. N. Trifonov. "Sequence periodicity in complete genomes of archaea suggests positive supercoiling." In: *Journal of biomolecular structure & dynamics* 16.2 (Oct. 1998), pp. 341–5.

[74] D Holste, I Grosse, and H Herzel. "Statistical analysis of the DNA sequence of human chromosome 22." In: *Physical review. E, Statistical, nonlinear, and soft matter physics* 64.4 Pt 1 (Oct. 2001), p. 041917.

[75] C. I. Hong et al. "Circadian rhythms synchronize mitosis in Neurospora crassa". In: *Proceedings of the National Academy of Sciences* 111.4 (2014), pp. 1397–1402.

[76] Y. Huang and J. Mrázek. "Assessing Diversity of DNA Structure-Related Sequence Features in Prokaryotic Genomes." In: *DNA research* (Jan. 2014), dst057.

[77] P. Huggins et al. "DECOD: fast and accurate discriminative DNA motif finding." In: *Bioinformatics (Oxford, England)* 27.17 (Sept. 2011), pp. 2361–7.

[78] M. E. Hughes, J. B. Hogenesch, and K. Kornacker. "JTK_CYCLE: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets." In: *Journal of Biological Rhythms* 25.5 (Oct. 2010), pp. 372–80.

[79] M. E. Hughes et al. "Harmonics of circadian gene transcription in mammals." In: *PLoS Genetics* 5.4 (Apr. 2009), e1000442.

[80] J. M. Hurley et al. "Analysis of clock-regulated genes in Neurospora reveals widespread posttranscriptional control of metabolic potential." In: *Proceedings of the National Academy of Sciences* 111.48 (Oct. 2014), pp. 16995–17002.

[81] K. Imai et al. "Circadian rhythms in the synthesis and degradation of a master clock protein KaiC in cyanobacteria." In: *The Journal of biological chemistry* 279.35 (Aug. 2004), pp. 36534–9.

[82] M Ishiura et al. "Expression of a gene cluster kaiABC as a circadian feedback process in cyanobacteria." In: *Science (New York, N.Y.)* 281.5382 (Sept. 1998), pp. 1519–23.

[83] H. Ito et al. "Cyanobacterial daily life with Kai-based circadian and diurnal genome-wide transcriptional control in Synechococcus elongatus." In: *Proceedings of the National Academy of Sciences* 106.33 (Aug. 2009), pp. 14168–73.

[84] H. Iwasaki et al. "KaiA-stimulated KaiC phosphorylation in circadian timing loops in cyanobacteria." In: *Proceedings of the National Academy of Sciences* 99.24 (Nov. 2002), pp. 15788–93.

[85] I. H. Jain, V. Vijayan, and E. K. O'Shea. "Spatial ordering of chromosomes enhances the fidelity of chromosome partitioning in cyanobacteria." In: *Proceedings of the National Academy of Sciences* 109.34 (Aug. 2012), pp. 13638–43.

[86] S. R. Jammalamadaka and A. SenGupta. *Topics in Circular Statistics*. World Scientific, 2001.

[87] K. S. Jeong, J. Ahn, and A. B. Khodursky. "Spatial patterns of transcriptional activity in the chromosome of Escherichia coli." In: *Genome Biology* 5.11 (Jan. 2004), R86.

[88] M. Jiang et al. "uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts." In: *BMC bioinformatics* 9 (Jan. 2008), p. 192.

[89] N. Kaplan et al. "The DNA-encoded nucleosome organization of a eukaryotic genome." In: *Nature* 458.7236 (Mar. 2009), pp. 362–6.

[90] S. Katayama et al. "Mode of binding of RNA polymerase α subunit to the phased A-tracts upstream of the phospholipase C gene promoter of Clostridium perfringens." In: *Anaerobe* 23 (Oct. 2013), pp. 62–9.

[91] L Kaufman and P Rousseeuw. "Clustering by means of medoids". In: *Statistical Data Analysis Based on the L1Norm and Related Methods*. Ed. by Y Dodge. Vol. 1. Elsevier/North-Holland, 1987, pp. 405–416.

[92] T. Kawabata et al. "Neurosprora crassa RAD5 homologue, mus-41, inactivation results in higher sensitivity to mutagens but has little effect on PCNA-ubiquitylation in response to UV-irradiation". In: *Current genetics* 52.3-4 (Sept. 2007), pp. 125–135.

[93] L. Kerkhof and P. Kemp. "Small ribosomal RNA content in marine Proteobacteria during non-steady-state growth". In: *FEMS Microbiology Ecology* 30.3 (Nov. 1999), pp. 253–260.

[94] G Kerr et al. "Techniques for clustering gene expression data." In: *Computers in Biology and Medicine* 38.3 (Mar. 2008), pp. 283–93.

[95] Y. V. Kil and W. S. Reznikoff. "DNA length, bending, and twisting constraints on IS50 transposition." In: *Proceedings of the National Academy of Sciences* 91.23 (Nov. 1994), pp. 10834–10838.

[96] J. Kim and H. Kim. "Clustering of change patterns using Fourier coefficients." In: *Bioinformatics (Oxford, England)* 24.2 (Jan. 2008), pp. 184–91.

[97] N. Kjeldgaard and C. Kurland. "The distribution of soluble and ribosomal RNA as a function of growth rate". In: *Journal of Molecular Biology* 6.4 (Apr. 1963), pp. 341–348.

[98] H. Knoop et al. "Flux Balance Analysis of Cyanobacterial Metabolism: The Metabolic Network of Synechocystis sp. PCC 6803". In: *PLoS Computational Biology* 9.6 (June 2013), e1003081.

[99] L. Koenig and E. Youn. "Hierarchical signature clustering for time series microarray data." In: *Advances in Experimental Medicine and Biology* 696 (Jan. 2011), pp. 57–65.

[100] N. Koike et al. "Transcriptional Architecture and Chromatin Landscape of the Core Circadian Clock in Mammals." In: *Science (New York, N.Y.)* 338.6105 (Aug. 2012), pp. 349–354.

[101] O Koksharova et al. "Genetic and biochemical evidence for distinct key functions of two highly divergent GAPDH genes in catabolic and anabolic carbon flow of the cyanobacterium Synechocystis sp. PCC 6803." In: *Plant molecular biology* 36.1 (Jan. 1998), pp. 183–94.

[102] T Kondo et al. "Circadian rhythms in prokaryotes: luciferase as a reporter of circadian gene expression in cyanobacteria." In: *Proceedings of the National Academy of Sciences* 90.12 (June 1993), pp. 5672–6.

[103] A. Korenčič et al. "The interplay of cis-regulatory elements rules circadian rhythms in mouse liver." In: *PLoS ONE* 7.11 (Jan. 2012), e46835.

[104] L. Kozobay-Avraham, S. Hosid, and A. Bolshoy. "Involvement of DNA curvature in intergenic regions of prokaryotes." In: *Nucleic Acids Research* 34.8 (Jan. 2006), pp. 2316–27.

[105] G. I. Kravatskaya et al. "Coexistence of different base periodicities in prokaryotic genomes as related to DNA curvature, supercoiling, and transcription." In: *Genomics* 98.3 (Sept. 2011), pp. 223–31.

[106] G. I. Kravatskaya et al. "Structural attributes of nucleotide sequences in promoter regions of supercoiling-sensitive genes: how to relate microarray expression data with genomic sequences." In: *Genomics* 101.1 (Jan. 2013), pp. 1–11.

[107] K. Kucho, K. Okamoto, and Y. Tsuchiya. "Global analysis of circadian expression in the cyanobacterium Synechocystis sp. strain PCC 6803". In: *Journal of Bacteriology* 187.6 (2005), p. 2190.

[108] H. Kushige et al. "Genome-wide and heterocyst-specific circadian gene expression in the filamentous Cyanobacterium Anabaena sp. strain PCC 7120." In: *Journal of Bacteriology* 195.6 (Mar. 2013), pp. 1276–84.

[109] S. Kutsuna et al. "Transcriptional regulation of the circadian clock operon kaiBC by upstream regions in cyanobacteria." In: *Molecular Microbiology* 57.5 (Sept. 2005), pp. 1474–84.

[110] R. G. Labiosa et al. "Examination of diel changes in global transcript accumulation in synechocystis (cyanobacteria)". In: *Journal of Phycology* 42.3 (June 2006), pp. 622–636.

[111] B. Langmead et al. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." In: *Genome Biology* 10.3 (Jan. 2009), R25.

[112] C. H. Laundon and J. D. Griffith. "Curved helix segments can uniquely orient the topology of supertwisted DNA". In: *Cell* 52.4 (Feb. 1988), pp. 545–549.

[113] R. Lavery et al. "A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA." In: *Nucleic Acids Research* 38.1 (Jan. 2010), pp. 299–313.

[114] G. Le Martelot et al. "Genome-wide RNA polymerase II profiles and RNA accumulation reveal kinetics of transcription and associated epigenetic changes during diurnal cycles." In: *PLoS biology* 10.11 (Jan. 2012), e1001442.

[115] A. C. Leeder et al. "Early colony establishment in Neurospora crassa requires a MAP kinase regulatory network." In: *Genetics* 195.3 (Nov. 2013), pp. 883–98.

[116] R. Lehmann, R. Machne, and H. Herzel. "The structural code of cyanobacterial genomes". In: *Nucleic Acids Research* (July 2014), gku641.

[117] R. Lehmann et al. "How cyanobacteria pose new problems to old methods: challenges in microarray time series analysis." In: *BMC bioinformatics* 14 (Jan. 2013), p. 133.

[118] P. W. Lepp and T. M. Schmidt. "Nucleic acid content of Synechococcus spp. during growth in continuous light and light/dark cycles". In: *Archives of Microbiology* 170.3 (Aug. 1998), pp. 201–207.

[119] D. Li et al. "A mitogen-activated protein kinase pathway essential for mating and contributing to vegetative growth in Neurospora crassa." In: *Genetics* 170.3 (July 2005), pp. 1091–104.

[120] U. de Lichtenberg et al. "Comparison of computational methods for the identification of cell cycle-regulated genes." In: *Bioinformatics (Oxford, England)* 21.7 (Apr. 2005), pp. 1164–71.

[121] W. K. Lim et al. "Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks." In: *Bioinformatics (Oxford, England)* 23.13 (July 2007), pp. i282–8.

[122] Y Liu et al. "Circadian orchestration of gene expression in cyanobacteria." In: *Genes & Development* 9.12 (1995), pp. 1469–1478.

[123] K. Lo et al. "flowClust: a Bioconductor package for automated gating of flow cytometry data." In: *BMC Bioinformatics* 10 (Jan. 2009), p. 145.

[124] J. Lovén et al. "Revisiting Global Gene Expression Analysis". In: *Cell* 151.3 (Oct. 2012), pp. 476–482.

[125] X. Lu et al. "Statistical resynchronization and Bayesian detection of periodically expressed genes." In: *Nucleic Acids Research* 32.2 (Jan. 2004), pp. 447–55.

[126] S. Lück et al. "Rhythmic degradation explains and unifies circadian transcriptome and proteome data." In: *Cell reports* 9.2 (Oct. 2014), pp. 741–51.

[127] R. Machné and D. B. Murray. "The yin and yang of yeast transcription: elements of a global feedback system between metabolism and chromatin." In: *PLoS ONE* 7.6 (Jan. 2012), e37906.

[128] S. Mahony and P. V. Benos. "STAMP: a web tool for exploring DNA-binding motif similarities." In: *Nucleic Acids Research* 35.Web Server issue (July 2007), W253–8.

[129] V. M. Markowitz et al. "IMG: the Integrated Microbial Genomes database and comparative analysis system." In: *Nucleic Acids Research* 40.Database issue (Jan. 2012), pp. D115–22.

[130] A. Mathelier et al. "JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles." In: *Nucleic Acids Research* 42.Database issue (Jan. 2014), pp. D142–7.

[131] S. Maurer, J. Fritz, and G. Muskhelishvili. "A systematic in vitro study of nucleoprotein complexes formed by bacterial nucleoid-associated proteins revealing novel types of DNA organization." In: *Journal of Molecular Biology* 387.5 (Apr. 2009), pp. 1261–76.

[132] M Meila. "Comparing clusterings—an information based distance". In: *Journal of Multivariate Analysis* 98.5 (May 2007), pp. 873–895.

[133] D. Memon et al. "A global analysis of adaptive evolution of operons in cyanobacteria." In: *Antonie van Leeuwenhoek* 103.2 (Feb. 2013), pp. 331–46.

[134] J. S. Menet, S. Pescatore, and M. Rosbash. "CLOCK:BMAL1 is a pioneer-like transcription factor." In: *Genes & Development* 28.1 (Jan. 2014), pp. 8–13.

[135] J. S. Menet et al. "Nascent-Seq reveals novel features of mouse circadian transcriptional regulation." In: *eLife* 1 (Jan. 2012), e00011.

[136] F. F. Millenaar et al. "How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results." In: *BMC Bioinformatics* 7 (Jan. 2006), p. 137.

[137] H. Min et al. "Phase determination of circadian gene expression in Synechococcus elongatus PCC 7942." In: *Journal of Biological Rhythms* 19.2 (Apr. 2004), pp. 103–12.

[138] A. Mitsui et al. "Strategy by which nitrogen-fixing unicellular cyanobacteria grow photoautotrophically". In: *Nature* 323.6090 (Oct. 1986), pp. 720–722.

[139] J. Mrázek. "Comparative analysis of sequence periodicity among prokaryotic genomes points to differences in nucleoid structure and a relationship to gene expression." In: *Journal of Bacteriology* 192.14 (July 2010), pp. 3763–72.

[140] J. Mrázek, T. Chaudhari, and A. Basu. "PerPlot & PerScan: tools for analysis of DNA curvature-related periodicity in genomic nucleotide sequences." In: *Microbial informatics and experimentation* 1.1 (Jan. 2011), p. 13.

[141] D. B. Murray, M. Beckmann, and H. Kitano. "Regulation of yeast oscillatory dynamics." In: *Proceedings of the National Academy of Sciences* 104.7 (Feb. 2007), pp. 2241–6.

[142] G. Muskhelishvili and A. Travers. "Integration of syntactic and semantic properties of the DNA code reveals chromosomes as thermodynamic machines converting energy into information". In: *Cellular and molecular life sciences* 70.23 (June 2013), pp. 4555–4567.

[143] U. Nagalakshmi et al. "The transcriptional landscape of the yeast genome defined by RNA sequencing." In: *Science (New York, N.Y.)* 320.5881 (June 2008), pp. 1344–9.

[144] U. Nair et al. "Roles for sigma factors in global circadian regulation of the cyanobacterial genome." In: *Journal of Bacteriology* 184.13 (July 2002), pp. 3530–8.

[145] M. Nakajima et al. "Reconstitution of circadian oscillation of cyanobacterial KaiC phosphorylation in vitro." In: *Science (New York, N.Y.)* 308.5720 (Apr. 2005), pp. 414–5.

[146] M. Nakao et al. "CyanoBase: the cyanobacteria genome database update 2010." In: *Nucleic Acids Research* 38.Database issue (Jan. 2010), pp. D379–81.

[147] M. Nakao et al. "CyanoBase: the cyanobacteria genome database update 2010." In: *Nucleic Acids Research* 38.Database issue (Jan. 2010), pp. D379–81.

[148] N. Nalabothula et al. "Archaeal nucleosome positioning in vivo and in vitro is directed by primary sequence motifs." In: *BMC genomics* 14 (Jan. 2013), p. 391.

[149] W. W. Navarre et al. "Selective silencing of foreign DNA with low GC content by the H-NS protein in Salmonella." In: *Science (New York, N.Y.)* 313.5784 (July 2006), pp. 236–8.

[150] W. W. Navarre et al. "Silencing of xenogeneic DNA by H-NS-facilitation of lateral gene transfer in bacteria by a defense system that recognizes foreign DNA." In: *Genes & Development* 21.12 (June 2007), pp. 1456–71.

[151] T. Nov Klaiman, S. Hosid, and A. Bolshoy. "Upstream curved sequences in E. coli are related to the regulation of transcription initiation." In: *Computational biology and chemistry* 33.4 (Aug. 2009), pp. 275–82.

[152] M. Nowrousian et al. "The frequency Gene Is Required for Temperature-Dependent Regulation of Many Clock-Controlled Genes in Neurospora crassa". In: *Genetics* 164.3 (July 2003), pp. 923–933.

[153] Z.-A. Ouafa et al. "The nucleoid-associated proteins H-NS and FIS modulate the DNA supercoiling response of the pel genes, the major virulence factors in the plant pathogen bacterium Dickeya dadantii." In: *Nucleic Acids Research* 40.10 (May 2012), pp. 4306–19.

[154] Y Ouyang et al. "Resonating circadian clocks enhance fitness in cyanobacteria." In: *Proceedings of the National Academy of Sciences* 95.15 (July 1998), pp. 8660–4.

[155] Z. Ouyang, Q. Zhou, and W. H. Wong. "ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells." In: *Proceedings of the National Academy of Sciences* 106.51 (Dec. 2009), pp. 21521–6.

[156] T. A. Owen-Hughes et al. “The chromatin-associated protein H-NS interacts with curved DNA to influence DNA topology and gene expression.” In: *Cell* 71.2 (Oct. 1992), pp. 255–65.

[157] J. S. O’Neill and A. B. Reddy. “Circadian clocks in human red blood cells”. In: *Nature* 469.7331 (Jan. 2011), pp. 498–503.

[158] T. L. Page. “Effects of optic-tract regeneration on internal coupling in the circadian system of the cockroach”. In: *Journal of Comparative Physiology* 153.3 (1983), pp. 353–363.

[159] A. Paré et al. “The Functions of Grainy Head-Like Proteins in Animals and Fungi and the Evolution of Apical Extracellular Barriers”. In: *PLoS ONE* 7.5 (May 2012), e36254.

[160] L. M. Parrott and J. H. Slater. “The DNA, RNA and protein composition of the cyanobacterium Anacystis nidulans grown in light- and carbon dioxide-limited chemostats”. In: *Archives of Microbiology* 127.1 (Aug. 1980), pp. 53–58.

[161] F. Partensky, W. R. Hess, and D. Vaulot. “Prochlorococcus, a Marine Photosynthetic Prokaryote of Global Significance”. In: *Microbiology and Molecular Biology Reviews* 63.1 (Mar. 1999), pp. 106–127.

[162] J. W. Pavlicek et al. “Supercoiling-induced DNA bending.” In: *Biochemistry* 43.33 (Aug. 2004), pp. 10664–8.

[163] A. G. Pedersen et al. “A DNA structural atlas for Escherichia coli.” In: *Journal of Molecular Biology* 299.4 (June 2000), pp. 907–30.

[164] B. J. Peter et al. “Genomic transcriptional response to loss of chromosomal supercoiling in Escherichia coli.” In: *Genome Biology* 5.11 (Jan. 2004), R87.

[165] L. Petersen et al. “RpoD promoters in Campylobacter jejuni exhibit a strong periodic signal instead of a -35 box.” In: *Journal of Molecular Biology* 326.5 (Mar. 2003), pp. 1361–72.

[166] S Pietrokovski. “Searching databases of conserved sequence regions by aligning protein multiple-alignments.” In: *Nucleic Acids Research* 24.19 (Oct. 1996), pp. 3836–45.

[167] A. V. Pinevich, O. V. Gavrilova, and S. G. Averina. “Chromatin morphology and cytokinesis in pleurocapsalean cyanobacteria”. In: *Cell and Tissue Biology* 2.1 (June 2008), pp. 53–56.

[168] C. S. Pittendrigh. “Temporal organization: reflections of a Darwinian clock-watcher.” In: *Annual review of physiology* 55 (Jan. 1993), pp. 16–54.

[169] C. S. Pittendrigh et al. “Growth Patterns in Neurospora: A Biological Clock in Neurospora”. In: *Nature* 184.4681 (July 1959), pp. 169–170.

[170] L. K. Poulsen, G Ballard, and D. A. Stahl. “Use of rRNA fluorescence in situ hybridization for measuring the activity of single cells in young and established biofilms.” In: *Applied and environmental microbiology* 59.5 (May 1993), pp. 1354–60.

[171] J. S. S. Prakash et al. “DNA supercoiling regulates the stress-inducible expression of genes in the cyanobacterium Synechocystis.” In: *Molecular bioSystems* 5.12 (Dec. 2009), pp. 1904–12.

[172] G. D. Price et al. “The functioning of the CO 2 concentrating mechanism in several cyanobacterial strains: a review of general physiological characteristics, genes, proteins, and recent advances”. In: *Canadian Journal of Botany* 76.6 (June 1998), pp. 973–1002.

[173] S. Priebe et al. “FungiFun: a web-based application for functional categorization of fungal genes and proteins.” In: *Fungal genetics and biology* 48.4 (Apr. 2011), pp. 353–8.

[174] E. Pruesse et al. “SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB.” In: *Nucleic Acids Research* 35.21 (Jan. 2007), pp. 7188–96.

[175] J. Qian et al. “Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions1”. In: *Journal of Molecular Biology* 314.5 (2001), pp. 1053–1066.

[176] X. Qin et al. “Coupling of a core post-translational pacemaker to a slave transcription/translation feedback loop in a circadian system.” In: *PLoS biology* 8.6 (Jan. 2010), e1000394.

[177] W. Rand. “Objective criteria for the evaluation of clustering methods”. In: *Journal of the American Statistical Association* 66.336 (1971), pp. 846–850.

[178] A. Relógio et al. “Ras-mediated deregulation of the circadian clock in cancer.” In: *PLoS Genetics* 10.5 (May 2014), e1004338.

[179] S. M. Reppert and D. R. Weaver. “Coordination of circadian timing in mammals.” In: *Nature* 418.6901 (Aug. 2002), pp. 935–41.

[180] G. Rey et al. “Genome-wide and phase-specific DNA-binding rhythms of BMAL1 control circadian output functions in mouse liver.” In: *PLoS biology* 9.2 (Feb. 2011), e1000595.

[181] R Rippka et al. “Generic Assignments, Strain Histories and Properties of Pure Cultures of Cyanobacteria”. In: *Microbiology* 111.1 (1979), pp. 1–61.

[182] J. Rodriguez et al. “Nascent-Seq analysis of Drosophila cycling gene expression.” In: *Proceedings of the National Academy of Sciences* 110.4 (2013), E275–84.

[183] R. Rohs et al. “The role of DNA shape in protein-DNA recognition.” In: *Nature* 461.7268 (Oct. 2009), pp. 1248–53.

[184] H. G. Roider et al. “CpG-depleted promoters harbor tissue-specific transcription factor binding signals—implications for motif overrepresentation analyses”. In: *Nucleic Acids Research* 37.19 (2009), pp. 6305–6315.

[185] R. Rosset, J. Julien, and R. Monier. “Ribonucleic acid composition of bacteria as a function of growth rate”. In: *Journal of Molecular Biology* 18.2 (July 1966), pp. 308–320.

[186] B. Salih and E. N. Trifonov. "Strong nucleosomes of A. thaliana concentrate in centromere regions." In: *Journal of biomolecular structure & dynamics* 33.1 (Nov. 2013), pp. 10–13.

[187] C. Sancar et al. "Dawn- and dusk-phased circadian transcription rhythms coordinate anabolic and catabolic functions in Neurospora." In: *BMC biology* 13.1 (Feb. 2015), p. 17.

[188] G. Sancar et al. "A Global Circadian Repressor Controls Antiphasic Expression of Metabolic Genes in Neurospora". In: *Molecular Cell* 44.5 (2011), pp. 687–697.

[189] S. C. Satchwell, H. R. Drew, and A. A. Travers. "Sequence periodicities in chicken nucleosome core DNA." In: *Journal of Molecular Biology* 191.4 (Oct. 1986), pp. 659–75.

[190] P. Schieg and H Herzel. "Periodicities of 10-11bp as indicators of the supercoiled state of genomic DNA." In: *Journal of Molecular Biology* 343.4 (Oct. 2004), pp. 891–901.

[191] B. E. Schirrmeister, A. Antonelli, and H. C. Bagheri. "The origin of multicellularity in cyanobacteria." In: *BMC evolutionary biology* 11.1 (Jan. 2011), p. 45.

[192] D. Schneider et al. "Fluorescence staining of live cyanobacterial cells suggest non-stringent chromosome segregation and absence of a connection between cytoplasmic and thylakoid membranes." In: *BMC cell biology* 8.1 (Jan. 2007), p. 39.

[193] R. Schöpflin et al. "Probing the elasticity of DNA on short length scales by modeling supercoiling under tension." In: *Biophysical journal* 103.2 (July 2012), pp. 323–30.

[194] S. A. Shabalina, A. Y. Ogurtsov, and N. A. Spiridonov. "A periodic pattern of mRNA secondary structure created by the genetic code." In: *Nucleic Acids Research* 34.8 (Jan. 2006), pp. 2428–37.

[195] P. M. Sharp and W. H. Li. "The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications." In: *Nucleic Acids Research* 15.3 (Feb. 1987), pp. 1281–95.

[196] L. A. Sherman, P. Meunier, and M. S. Colón-López. "Diurnal rhythms in metabolism: A day in the life of a unicellular, diazotrophic cyanobacterium". In: *Photosynthesis Research* 58.1 (Oct. 1998), pp. 25–42.

[197] D. J. Sherratt. "Bacterial chromosome dynamics." In: *Science (New York, N.Y.)* 301.5634 (Aug. 2003), pp. 780–5.

[198] P. M. Shih et al. "Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing." In: *Proceedings of the National Academy of Sciences* 110.3 (Jan. 2013), pp. 1053–8.

[199] W. Sikora-Wohlfeld et al. "Assessing Computational Methods for Transcription Factor Target Gene Identification Based on ChIP-seq Data". In: *PLoS Computational Biology* 9.11 (Nov. 2013), e1003342.

[200] L. Sinzelle et al. “Factors acting on Mos1 transposition efficiency.” In: *BMC molecular biology* 9.1 (Jan. 2008), p. 106.

[201] K. M. Smith et al. “Transcription factors in light and circadian clock signaling networks revealed by genomewide mapping of direct targets for neurospora white collar complex.” In: *Eukaryotic cell* 9.10 (Oct. 2010), pp. 1549–56.

[202] G. K. Smyth. “Limma: linear models for microarray data”. In: *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. New York: Springer, 2005. Chap. 23, pp. 397–420.

[203] P. Sobetzko, A. Travers, and G. Muskhelishvili. “Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle.” In: *Proceedings of the National Academy of Sciences* 109.2 (Jan. 2012), E42–50.

[204] P. T. Spellman et al. “Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.” In: *Molecular Biology of the Cell* 9.12 (Dec. 1998), pp. 3273–97.

[205] R Staden and A. D. McLachlan. “Codon preference and its use in identifying protein coding regions in long DNA sequences.” In: *Nucleic Acids Research* 10.1 (Jan. 1982), pp. 141–56.

[206] L. J. Stal and W. E. Krumbein. “Temporal separation of nitrogen fixation and photosynthesis in the filamentous, non-heterocystous cyanobacterium Oscillatoria sp.” In: *Archives of Microbiology* 149.1 (Nov. 1987), pp. 76–80.

[207] J. Stöckel et al. “Global transcriptomic analysis of Cyanothece 51142 reveals robust diurnal oscillation of central metabolic processes.” In: *Proceedings of the National Academy of Sciences* 105.16 (Apr. 2008), pp. 6156–61.

[208] K.-F. Storch et al. “Extensive and divergent circadian gene expression in liver and heart.” In: *Nature* 417.6884 (May 2002), pp. 78–83.

[209] C. Straub et al. “A day in the life of microcystis aeruginosa strain PCC 7806 as revealed by a transcriptomic analysis.” In: *PLoS ONE* 6.1 (Jan. 2011), e16208.

[210] M. Straume. “DNA microarray time series analysis: automated statistical assessment of circadian rhythms in gene expression patterning.” In: *Methods in enzymology* 383 (Jan. 2004), pp. 149–66.

[211] R. Stuger et al. “DNA supercoiling by gyrase is linked to nucleoid compaction.” In: *Molecular biology reports* 29.1-2 (Jan. 2002), pp. 79–82.

[212] Y. Tabei, K. Okada, and M. Tsuzuki. “Sll1330 controls the expression of glycolytic genes in Synechocystis sp. PCC 6803.” In: *Biochemical and biophysical research communications* 355.4 (Apr. 2007), pp. 1045–50.

[213] R. C. Taylor et al. “Changes in translational efficiency is a dominant regulatory mechanism in the environmental response of bacteria.” en. In: *Integrative biology* 5.11 (Nov. 2013), pp. 1393–406.

[214] A. W. Thompson et al. "Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga." In: *Science (New York, N.Y.)* 337.6101 (Sept. 2012), pp. 1546–50.

[215] J. Toepel et al. "Differential transcriptional analysis of the cyanobacterium Cyanothece sp. strain ATCC 51142 during light-dark and continuous-light growth." In: *Journal of Bacteriology* 190.11 (June 2008), pp. 3904–13.

[216] J. Toepel et al. "Transcriptional Analysis of the Unicellular, Diazotrophic Cyanobacterium Cyanothece Sp. Atcc 51142 Grown Under Short Day/Night Cycles". In: *Journal of Phycology* 45.3 (June 2009), pp. 610–620.

[217] M. Y. Tolstorukov et al. "A-tract clusters may facilitate DNA packaging in bacterial nucleoid." In: *Nucleic Acids Research* 33.12 (Jan. 2005), pp. 3907–18.

[218] M Tomita, M Wada, and Y Kawashima. "ApA dinucleotide periodicity in prokaryote, eukaryote, and organelle genomes." In: *Journal of Molecular Evolution* 49.2 (Aug. 1999), pp. 182–92.

[219] H. Tong and J. Mrázek. "Investigating the interplay between nucleoid-associated proteins, DNA curvature, and CRISPR elements using comparative genomics." In: *PLoS ONE* 9.3 (Jan. 2014). Ed. by M. Helmer-Citterich, e90940.

[220] A. A. Travers, G. Muskhelishvili, and J. M. T. Thompson. "DNA information: from digital code to analogue structure." In: *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 370.1969 (June 2012), pp. 2960–86.

[221] A. Travers and G. Muskhelishvili. "A common topology for bacterial and eukaryotic transcription initiation?" In: *EMBO reports* 8.2 (Feb. 2007), pp. 147–51.

[222] E. N. Trifonov and J. L. Sussman. "The pitch of chromatin DNA is reflected in its nucleotide sequence." In: *Proceedings of the National Academy of Sciences* 77.7 (July 1980), pp. 3816–20.

[223] H. J. Tripp et al. "Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium." In: *Nature* 464.7285 (Mar. 2010), pp. 90–4.

[224] E. Trotta. "Selection on codon bias in yeast: a transcriptional hypothesis." In: *Nucleic Acids Research* 41.20 (Nov. 2013), pp. 9382–95.

[225] E. Trotta. "The 3-base periodicity and codon usage of coding sequences are correlated with gene expression at the level of transcription elongation." In: *PLoS ONE* 6.6 (Jan. 2011), e21590.

[226] H. Tsen and S. D. Levene. "Supercoiling-dependent flexibility of adenosine-tract-containing DNA detected by a topological method". In: *Proceedings of the National Academy of Sciences* 94.7 (Apr. 1997), pp. 2817–2822.

[227] N. F. Tsinoremas et al. "A sigma factor that modifies the circadian expression of a subset of genes in cyanobacteria." In: *The EMBO journal* 15.10 (May 1996), pp. 2488–95.

[228] B. P. Tu et al. “Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes.” In: *Science (New York, N.Y.)* 310.5751 (Nov. 2005), pp. 1152–8.

[229] J. W. Tukey. *Exploratory Data Analysis*. Ed. by N Wrigley and R. J. Bennet. Vol. 2. Quantitative applications in the social sciences 1. Addison-Wesley, 1977. Chap. 2, p. 688.

[230] S. Turner et al. “Investigating Deep Phylogenetic Relationships among Cyanobacteria and Plastids by Small Subunit rRNA Sequence Analysis”. In: *The Journal of Eukaryotic Microbiology* 46.4 (July 1999), pp. 327–338.

[231] H. R. Ueda et al. “Genome-wide transcriptional orchestration of circadian rhythms in Drosophila.” In: *The Journal of biological chemistry* 277.16 (Apr. 2002), pp. 14048–52.

[232] A. Urasaki, Y. Sekine, and E. Ohtsubo. “Transposition of Cyanobacterium Insertion Element ISY100 in Escherichia coli”. In: *Journal of Bacteriology* 184.18 (Sept. 2002), pp. 5104–5112.

[233] D. Ussery et al. “Bias of purine stretches in sequenced chromosomes.” In: *Computers & chemistry* 26.5 (July 2002), pp. 531–41.

[234] D Vaulot et al. “Growth of prochlorococcus, a photosynthetic prokaryote, in the equatorial pacific ocean.” In: *Science (New York, N.Y.)* 268.5216 (June 1995), pp. 1480–2.

[235] V. Vijayan, I. H. Jain, and E. K. O’Shea. “A high resolution map of a cyanobacterial transcriptome.” In: *Genome Biology* 12.5 (May 2011), R47.

[236] V. Vijayan and E. K. O’Shea. “Sequence determinants of circadian gene expression phase in cyanobacteria.” In: *Journal of Bacteriology* 195.4 (Feb. 2013), pp. 665–71.

[237] V. Vijayan, R. Zuzow, and E. OShea. “Oscillations in supercoiling drive circadian gene expression in cyanobacteria”. In: *Proceedings of the National Academy of Sciences* 106.52 (Dec. 2009), pp. 22564–8.

[238] M. W. Vitalini et al. “Circadian rhythmicity mediated by temporal regulation of the activity of p38 MAPK.” In: *Proceedings of the National Academy of Sciences* 104.46 (Nov. 2007), pp. 18223–8.

[239] B. Wang et al. “Neurospora WC-1 recruits SWI/SNF to remodel frequency and initiate a circadian cycle.” In: *PLoS Genetics* 10.9 (Sept. 2014), e1004599.

[240] T. A. Wang et al. “Circadian rhythm of redox state regulates excitability in suprachiasmatic nucleus neurons.” In: *Science (New York, N.Y.)* 337.6096 (Aug. 2012), pp. 839–42.

[241] X. Wang et al. “Short time-series microarray analysis: methods and challenges.” In: *BMC Systems Biology* 2 (Jan. 2008), p. 58.

[242] Z. Wang, M. Gerstein, and M. Snyder. “RNA-Seq: a revolutionary tool for transcriptomics.” In: *Nature reviews. Genetics* 10.1 (Jan. 2009), pp. 57–63.

[243] J. Ward. "Hierachical grouping to optimize an objective function." In: *Journal of the American Statistical Association* 58.301 (1963), pp. 236–244.

[244] S. Watanabe et al. "Light-dependent and asynchronous replication of cyanobacterial multi-copy chromosomes". In: *Molecular Microbiology* 83.4 (Feb. 2012), pp. 856–865.

[245] J. B. Waterbury and R. Y. Stanier. "Patterns of growth and development in pleurocapsalean cyanobacteria." In: *Microbiological reviews* 42.1 (Mar. 1978), pp. 2–44.

[246] R. Wehrens and L. Buydens. "Self-and super-organizing maps in R: the Kohonen package". In: *Journal of Statistical Software* 21.5 (2007), p. 19.

[247] O. Weiss and H. Herzel. "Correlations in protein sequences and property codes." In: *Journal of Theoretical Biology* 190.4 (Feb. 1998), pp. 341–53.

[248] F. W. Went. "Photo- and Thermoperiodic Effects in Plant Growth". In: *Cold Spring Harbor Symposia on Quantitative Biology* 25.0 (Jan. 1960), pp. 221–230.

[249] P. l. O. Westermark and H. Herzel. "Mechanism for 12 hr rhythm generation by the circadian clock." In: *Cell reports* 3.4 (Apr. 2013), pp. 1228–38.

[250] K. Whitehead et al. "Diurnally entrained anticipatory behavior in archaea." In: *PLoS ONE* 4.5 (Jan. 2009). Ed. by F. Rodriguez-Valera, e5485.

[251] S. Wichert, K. Fokianos, and K. Strimmer. "Identifying periodically expressed transcripts in microarray time series data." In: *Bioinformatics (Oxford, England)* 20.1 (Jan. 2004), pp. 5–20.

[252] M. A. Woelfle and C. H. Johnson. "No promoter left behind: global circadian gene expression in cyanobacteria." In: *Journal of Biological Rhythms* 21.6 (Dec. 2006), pp. 419–31.

[253] M. A. Woelfle et al. "The adaptive value of circadian clocks: an experimental assessment in cyanobacteria." In: *Current biology* 14.16 (Aug. 2004), pp. 1481–6.

[254] P. Worning et al. "Structural analysis of DNA sequence: evidence for lateral gene transfer in Thermotoga maritima." In: *Nucleic Acids Research* 28.3 (Feb. 2000), pp. 706–9.

[255] L. J. Wu. "Structure and segregation of the bacterial nucleoid." In: *Current opinion in genetics & development* 14.2 (Apr. 2004), pp. 126–32.

[256] Y Xu, T Mori, and C. H. Johnson. "Circadian clock-protein expression in cyanobacteria: rhythms and phase setting." In: *The EMBO journal* 19.13 (July 2000), pp. 3349–57.

[257] Y. Xu et al. "Non-optimal codon usage is a mechanism to achieve circadian clock conditionality." In: *Nature* 495.7439 (Mar. 2013), pp. 116–20.

[258] Y.-F. Xu et al. "Nucleotide degradation and ribose salvage in yeast." In: *Molecular Systems Biology* 9 (Jan. 2013), p. 665.

[259] R. Yang and Z. Su. "Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation." In: *Bioinformatics (Oxford, England)* 26.12 (June 2010), pp. i168–74.

[260] Y. H. Yang et al. "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation." In: *Nucleic Acids Research* 30.4 (Feb. 2002), e15.

[261] K. Yeung et al. "Model-based clustering and data transformations for gene expression data". In: *Bioinformatics (Oxford, England)* 17.10 (2001), pp. 977–987.

[262] R. Zhang et al. "A circadian gene expression atlas in mammals: Implications for biology and medicine". In: *Proceedings of the National Academy of Sciences* 111.45 (Oct. 2014), pp. 16219–16224.

[263] V. B. Zhurkin. "Periodicity in DNA primary structure is defined by secondary structure of the coded protein." In: *Nucleic Acids Research* 9.8 (Apr. 1981), pp. 1963–71.

[264] E. R. Zinser et al. "Choreography of the transcriptome, photophysiology, and cell cycle of a minimal photoautotroph, prochlorococcus." In: *PLoS ONE* 4.4 (Jan. 2009), e5135.

[265] D. Zwicker, D. K. Lubensky, and P. R. ten Wolde. "Robust circadian clocks from coupled protein-modification and transcription-translation cycles." In: *Proceedings of the National Academy of Sciences* 107.52 (Dec. 2010), pp. 22540–5.

List of Figures

List of Tables

Abstract

An internal clock mechanism enables organisms to predict and adapt to the daily change between day and night. Comprehensive assessments of circadian expression programs are crucial to elucidate the physiological relevance of the clock in an organism. In photoautotrophic organisms, the adjustments of metabolic processes at the beginning and end of the photic phases play a central role. A leading hypothesis is that gene expression facilitates the circadian control of the cellular metabolism. However, circadian regulation patterns are also found in other cellular processes, such as the chromatid structure in cyanobacteria. This thesis is divided into three main sections which consider different aspects of the transcriptional control exerted by the circadian clock.

Chapter 2 presents a microarray time series dataset of diurnal gene expression in *Synechocystis sp.* PCC 6803. The analysis of this data set indicates diurnal oscillations in the 16S and 23S rRNA content, which cause a systematic bias in the observed diurnal transcriptional patterns when using common multi-array normalisation methods. In order to address this issue, a normalisation procedure is proposed which resolves these issues by using the least oscillating gene set. In addition to this dataset, a collection of microarray datasets, which quantifies diurnal or circadian expression in six popular cyanobacterial strains, is systematically compared. Standardised oscillation detection improved the comparability between individual datasets. Strain-specific peak expression phases in independent experiments demonstrated reliable reproducibility. In contrast, peak expression phases of homologous genes vary significantly between cyanobacterial strains. Moreover, the set of 95 genes with consistent diurnal expression in all considered datasets, the core diurnal genome, is analysed in detail. Many of the contained genes code for proteins with metabolic functions. A pairwise comparison of diurnal expression phases between the datasets and phases in the core diurnal genome reveal that cyanobacterial strains have adapted their diurnal expression programs to their individual environment and do not follow a general expression program.

Chapter 3 focusses on the periodic occurrence of AT dinucleotides in a number of cyanobacterial genomes. It is known that AT dinucleotides induce bending of the DNA backbone which accumulates when the dinucleotides occur in phase with the helical period of the DNA. Current hypotheses connect such a bending signal either with a global structural compaction of the chromatid or alternatively with the transcriptional regulation of target genes. The latter hypothesis is supported by the observation that the chromatid of the cyanobacterium *Synechococcus elongatus* exhibits circadian rhythms of compaction and relaxation due to changes in the supercoiling accompanied with transcriptional changes in a range of genes. The described analysis compares evidence for both hypotheses. In a wide range of cyanobacterial genomes, combinations of A and T (AT2) induce the strongest genome-wide ≈ 11 bp periodic signal among all dinucleotides. However, only transposons are found to feature particularly strong AT2 periodicity in gene sections, which are thought to be curved to aid transposase binding. Instead, genome-wide AT2 periodicity correlates with the number of chromosomal copies present in individual cells of the particular strain. These results suggest that high AT2 periodicity facilitates chromosomal compaction and is thus specifically favoured in cyanobacterial strains with polyploid lifestyle.

Finally, chapter 4 is dedicated to the question of how phases of circadian target genes are tuned. A mathematical model predicts that the combined regulatory input of two circadian TFs in appropriate phase relationship yields a modification of the target gene phase or frequency doubling, *i.e.*, ultradian rhythmicity. This prediction addressed in an in-depth study of circadian transcriptional rhythms in *Neurospora crassa*. The first analysis step quantifies to which extent rhythmic binding of RNA II polymerase (RNAPII) induces circadian rhythmicity in the abundance of the corresponding mRNA. Accordingly, the binding of RNAPII to gene promoters and bodies is compared to the corresponding mRNA abundance. Three classes of genes are identified: one showing rhythmicity both in transcriptional and mRNA accumulation, a second class with rhythmic transcription but non-rhythmic mRNA levels, and a third class with non-rhythmic transcription but rhythmic mRNAs. The third group featured elevated transcriptional variability, which might facilitate circadian rhythmicity in the corresponding mRNA abundance. Due to the regulation of morning- and evening-expressed genes by WCC and CSP1, respectively, both circadian TFs are analysed to verify the prediction of frequency doubling of combined target genes. However, the phase relationship of these TFs in the presented dataset is inadequate for the generation of ultradian rhythms. Accordingly, experimental data do not show preferential binding in proximity of ultradian genes for either of the TFs. Instead, proximal promoter sequence analyses suggest transcriptional regulation by different pairs of transcription factors specific to the ultradian phase.

Zusammenfassung

Ein interner Uhr-Mechanismus erlaubt es diversen Spezies, den täglichen Wechsel zwischen Tag und Nacht vorher zu sagen und sich entsprechend anzupassen. Umfassende Messungen zirkadianer Gen-Expressionsmuster sind notwendig, um den Einfluss der Uhr auf die Physiologie abbilden zu können. In photoautotrophen Spezies spielt die Anpassung metabolischer Prozesse zum Anfang und Ende der photischen Phase eine zentrale Rolle. Eine häufig bestätigt gefundene Annahme ist daher die zirkadiane Expression von Genen metabolisch wichtiger Proteine. Neben der Gen-Expression können zirkadiane Regulationsmuster auch in anderen zellulären Prozessen wie z.B. der Chromatid-Struktur in Cyanobakterien festgestellt werden. Diese Dissertation ist in drei Teile geteilt, welche verschiedene Aspekte der transkriptionellen Regulation durch die zirkadiane Uhr beleuchten.

Kapitel 2 stellt einen Microarray Zeitserien-Datensatz zur Beschreibung diurnaler Expressionsrhythmik im Cyanobakterium *Synechocystis sp.* PCC 6803 vor. Für diesen Stamm bisher nicht beschriebene diurnale Oszillationen im zellulären 16S und 23S rRNA-Gehalt haben zu Komplikationen bei der Verwendung von verbreiteten Microarray-Normalisierungsmethoden geführt. Mittels Normalisierung hinsichtlich gering oszillierender Gene kann dies verhindert werden. Das diurnale Expressionsprogramm wurde mit weiteren Microarray-Datensätzen verglichen, die zirkadiane oder diurnale Expressionsprogramme in sechs verschiedenen Cyanobakterien-Stämmen abbilden. Die standardisierte Detektion oszillierender Gene verbesserte die Vergleichbarkeit der Datensätze. Es wurde gute Reproduzierbarkeit diurnaler Expressionmuster in unabhängigen Experimenten festgestellt. Dagegen variiert die Phase maximaler Expression stark zwischen oszillierenden Homologen in versch. Stämmen. Das 95 Mitglieder umfassende Set von Genen, deren Homologe in allen Datensätzen oszillierende Expressionsmuster aufwiesen (core diurnal genome), wurde im Detail analysiert. Wie erwartet sind metabolische Gene stark vertreten. Paarweise Vergleiche der Expressionsphasen zwischen den einzelnen Datensätzen und im core diurnal genome legen nahe, dass Cyanobakterien-Stämme ihre diurnalen Expressionsmuster an die jeweiligen Umweltbedingungen angepasst haben und kein generelles diurnales Expressionmuster aufweisen.

Kapitel 3 analysiert das periodische Vorkommen von AT Dinukleotiden in einer Reihe von Cyanobakterien-Genomen. Da AT Dinukleotide eine Krümmung im DNA-Rückgrat hervorrufen, kann deren Vorkommen in Phase mit der helikalen Periode der DNA eine Biegung induzieren. Es wurde bereits vermutet, dass eine solche Biegung der DNA entweder eine globale strukturelle Verdichtung des Nukleotids erleichtert oder alternativ der transkriptionellen Regulation bestimmter Zielgene dienen könnte. Letzteres wird durch die Beobachtung gestützt, dass das chromosomale Supercoiling im Cyanobakterium *Synechococcus elongatus* eine zirkadiane Rhythmik aufweist, welche sich in der Expression einer Reihe von Genen widerspiegelt. Die hier beschriebene Analyse vergleicht eine Reihe von Hinweisen für beide Hypothesen. Ein Vergleich zeigte, dass Kombinationen von A und T (AT2) in einem Abstand von ≈ 11 bp die stärkste genomische Periodizität in einer Reihe von Cyanobakterien aufweisen. Jedoch kann nur in Transposons besonders starke AT2 Periodizität in Bereichen festgestellt werden, welche aufgrund ihrer Funktionsweise starke Krümmung aufweisen. Stattdessen korreliert die Stärke Genom-weit gemessener AT2 Periodizität gut mit der Anzahl der typischerweise in der Zelle vorliegenden Chromosomen-Kopien. Diese Ergebnisse weisen darauf hin, dass AT2 Periodizität die Verdichtung

des Nukleoids vereinfacht und daher verstärkt in Stämmen mit besonders vielen Chromosomen-Kopien auftritt.

Kapitel 4 ist der Frage gewidmet, wie Expressionsphasen in Zielgenen der zirkadianen Uhr modifiziert werden können. Mathematische Modellierung sagt vorher, dass die kombinierte Regulation zweier zirkadianer Transkriptionsfaktoren (TF) in entsprechendem Phasenverhältnis zu modifizierten Phasen oder Frequenzverdopplung und damit zu ultradianer Rhythmik des Zielgens führt. Es wird eine detaillierte Studie zirkadianer Transkriptionsrhythmen in *Neurospora crassa* präsentiert. In einem ersten Schritt wird beleuchtet, zu welchem Grad die rhythmische Bindung von RNA Polymerase II (RNAPII) für die zirkadiane Rhythmik in der mRNA Menge verantwortlich ist. Hierzu wird die Bindung von RNAPII an Promotoren und Gene mit der gleichzeitig gemessenen entsprechenden mRNA Konzentration verglichen. Es wurden drei Genklassen identifiziert, eine erste mit rhythmischer Transkription und mRNA Menge, eine zweite mit rhythmischer Transkription aber nicht-rhythmischer mRNA Menge, und eine dritte mit nicht-rhythmischer Transkription und rhythmischer mRNA Menge. Letztere weist erhöhte transkriptionelle Variabilität auf, welche die zirkadiane Rhythmizität der mRNA befördern könnte. Da die beiden Transkriptionsfaktoren WCC und CSP1 für die Regulation von jeweils am Morgen und Abend exprimierten Genen verantwortlich sind, ist dieses Paar ideal um die vorhergesagte Frequenzverdopplung zu beobachten. Deren Phasenbeziehung im vorgestellten Datensatz ist jedoch ungeeignet für die Erzeugung von ultradianen Rhythmen, was die beobachtete Phasenverteilung der Zielgene bestätigt. Dementsprechend deutet die Auswertung experimenteller Daten nicht auf eine Transkriptionsregulation der ultradianen Gene durch WCC oder CSP1 hin. Stattdessen weist die Analyse der Promotorsequenzen auf alternative phasen-spezifische TFs als transkriptionelle Regulatoren hin.

Publications and Software

- Lehmann R., Childs L., Thomas P., Abreu M., Fuhr L., Herzel H., Leser U. & Relógio A. (2015). Assembly of a Comprehensive Regulatory Network for the Mammalian Circadian Clock: A Bioinformatics Approach. *PLOS One*

- Sharon S., Salomon E., Kranzler C., Lis H., Lehmann R., Georg J., Zer H., Hess W.R., & Keren N. (2014). The hierarchy of transition metal homeostasis: Iron controls manganese accumulation in a unicellular cyanobacterium. *BBA Bioenergetics*, pp. 1990-1997

- Korenčič A., Košir R., Bordyugov G., Lehmann R., Rozman D., & Herzel H. (2014) Timing of circadian genes in mammalian tissues. *Scientific Reports*, p. 5782

- Lehmann R., Machné R., & Herzel H. (2014). The structural code of cyanobacterial genomes. *Nucleic Acids Research*, p. gku641

- Guerreiro A. C. L., Benevento M., Lehmann R., van Breukelen B., Post H., Giansanti P., Maarten Altelaar A. F., Axmann I. M. & Heck A. J. R. (2014). 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*, pp. 2042-2055

- Beck C., Hertel S., Rediger A., Lehmann R., Wiegard A., Kölsch A., Heilmann B., Georg J., Hess W. R. & Axmann I. M. (2014). A daily expression pattern of protein-coding genes and small non-coding RNAs in Synechocystis sp. PCC 6803. *Applied and Environmental Microbiology*, pp. 5195–5206

- Lehmann, R., Machné R., Georg J., Benary M., Axmann I. M., & Steuer R. (2013). How cyanobacteria pose new problems to old methods: Challenges in microarray time series analysis. *BMC Bioinformatics*, p. 133

- Knoop H., Gründel M., Zilliges Y., Lehmann R., Hoffmann S., Lockau W., & Steuer R. (2013). Flux balance analysis of cyanobacterial metabolism: The metabolic network of Synechocystis sp. PCC 6803. *PLoS Computational Biology*, p. e1003081

- Adams R., Worth CL., Guenther S., Dunkel M., Lehmann R. & Preissner R. (2012). Binding sites in membrane proteins–diversity, druggability and prospects. *European Journal of Cell Biology*, pp. 326-339

- Hohberg M., Knöchel J., Hoffmann C., Chlench S., Wunderlich W., Alter A, Maroski J., Vorderwülbecke B, JSilva-Azevedo L, Knudsen R., Lehmann R., Fiedorowicz K.,

Bongrazio M., Nitsche B.,Hoepfner M., Styp-Rekowska B., Pries A. & Zakrzewicz A., (2010) Expression of ADAMTS1 in endothelial cells is induced by shear stress and suppressed in sprouting capillaries. *Journal of Cellular Physiology*, pp. 350-361

- Lehmann R., & Steuer R. (2014) CAPTOR - a Cyanobacterial Arduino-based PhotobioreacTOR. *https://github.com/roblehmann/captor*
- Benary M., Kroeger S., Lee Y., & Lehmann R. (2013). cobindR: Finding Co-occuring motifs of transcription factor binding sites. *http://bioconductor.org/packages/release/bioc/html/cobindR.html*

Acknowledgements

First and foremost, I want to thank ***Hanspeter Herzel*** *for supervising my thesis. His guidance, patience, encouragement, and excellent support were invaluable for the success of this work. Many thanks go to* ***Martin Vingron*** *for supporting and monitoring this thesis as member of my PhD committee and graduate school. I want to thank* ***Ralf Steuer***, ***Ilka Axmann***, *and* ***Rainer Machne*** *for their support, guidance, expertise, and fruitful collaborations. Special thanks go to* ***Angela Relogio***, *it was a pleasure working with you. I thank* ***Pal Westermark*** *for stimulating discussions. I thank* ***Nir Keren*** *for hosting me at Hebrew University where I have learned a lot about cyanobacteria in theory and practice. I thank* ***Satoru Miyano*** *and* ***Seiya Imoto*** *for hosting me at the Tokyo university.*

I want to thank my colleagues ***Manuela Benary***, ***Christian Beck***, ***Henning Knoop***, ***Markus Rauer***, *and* ***Christian Brettschneider*** *for the great collaboration and the stimulating discussions about science and everything else. It was a delight to build the CAPTOR with* ***Adrian Kölsch***, *who introduced me to experimental work with cyanobacteria. I thank* ***Andreas Hantschmann*** *and* ***Tiziano Zito*** *for the great and instant IT support. I want to thank all former and current members of the Institute for Theoretical Biology for creating an outstanding working atmosphere, for scientific discussions, and for long evenings of table tennis.*

I thank ***Raphael Bauer*** *for his guidance and inexhaustible and contagious enthusiasm for science. Many thanks go to* ***Kim Joppen*** *and* ***Inka Löck***, *who always provided me with the right perspective. I thank my parents* ***Hans-Werner*** *and* ***Renate*** *and my sisters* ***Nadja***, ***Dörte***, ***Vivian*** *and* ***Silka*** *for their love and support.*

Lastly, I am grateful for the support and love of my wife ***Sahika***.

Selbständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Berlin, den 1.6.2015 Robert Lehmann