OPEN ACCESS Freely available online PLoS COMPUTATIONAL BIOLOGY

Modeling an Evolutionary Conserved Circadian *Cis*-Element

Eric R. Paquet[1,2], Guillaume Rey[1,2], Felix Naef[1,2*]

1 Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland, **2** Swiss Institute of Experimental Cancer Research (ISREC), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Circadian oscillator networks rely on a transcriptional activator called CLOCK/CYCLE (CLK/CYC) in insects and CLOCK/BMAL1 or NPAS2/BMAL1 in mammals. Identifying the targets of this heterodimeric basic-helix-loop-helix (bHLH) transcription factor poses challenges and it has been difficult to decipher its specific sequence affinity beyond a canonical E-box motif, except perhaps for some flanking bases contributing weakly to the binding energy. Thus, no good computational model presently exists for predicting CLK/CYC, CLOCK/BMAL1, or NPAS2/BMAL1 targets. Here, we use a comparative genomics approach and first study the conservation properties of the best-known circadian enhancer: a 69-bp element upstream of the *Drosophila melanogaster period* gene. This fragment shows a signal involving the presence of two closely spaced E-box–like motifs, a configuration that we can also detect in the other four prominent CLK/CYC target genes in flies: *timeless, vrille, Pdp1,* and *cwo*. This allows for the training of a probabilistic sequence model that we test using functional genomics datasets. We find that the predicted sequences are overrepresented in promoters of genes induced in a recent study by a glucocorticoid receptor-CLK fusion protein. We then scanned the mouse genome with the fly model and found that many known CLOCK/BMAL1 targets harbor sequences matching our consensus. Moreover, the phase of predicted cyclers in liver agreed with known CLOCK/BMAL1 regulation. Taken together, we built a predictive model for CLK/CYC or CLOCK/BMAL1-bound *cis*-enhancers through the integration of comparative and functional genomics data. Finally, a deeper phylogenetic analysis reveals that the link between the CLOCK/BMAL1 complex and the circadian *cis*-element dates back to before insects and vertebrates diverged.

Citation: Paquet ER, Rey G, Naef F (2008) Modeling an evolutionary conserved circadian *cis*-element. PLoS Comput Biol 4(2): e38. doi:10.1371/journal.pcbi.0040038

Introduction

In flies and mammals, circadian timing is controlled via interlocked transcriptional feedback loops that rely on basic helix-loop-helix (bHLH), PAS domain transcription factors [1,2]. In both fly and mammalian systems an evolutionary conserved bHLH heterodimer acts as the central transcriptional activator. The pair is called CLOCK [3] and CYCLE [4] in *Drosophila*, while the mammalian orthologues are CLOCK [5] and BMAL1 [6]. In mammals the CLOCK paralog NPAS2 can substitute for CLOCK function in the suprachiasmatic nucleus [7,8]. Like most transcription regulators of the bHLH family members, DNA binding of the CLK/CYC or CLOCK/BMAL1 pairs has been shown to involve canonical CANNTG E-box sequences [9–11] both in flies and mammals [6,12]. However, the low information content of this motif does not provide a sufficient explanation for the specificity of gene induction by the CLOCK transcription factor, nor does it allow to build a model that can predict clock regulated transcripts on a genome-wide scale.

Both the possibility of informative nucleotides flanking the E-boxes or the possibility that a combination of closely spaced partner signals could contribute cooperatively to the specificity was considered in flies and mammals [13,14]. Either mechanism can in theory significantly increase binding affinity of CLK/CYC to DNA, e.g. an increase in total ΔG_0 of 1 kcal/mol from one additional good hydrogen bond raises binding affinity by a factor of 5.

In *Drosophila*, the best-studied enhancer is that of the *period* (*per*) gene where a 69-bp fragment upstream of the transcription start site (TSS) drives circadian gene expression [9]. This enhancer depends on a canonical E-box, but it was also shown that its immediate 3' flank contributes to drive large amplitudes and tissue specific expression [15]. Interestingly, the fly enhancer can also be activated by the murine CLOCK/BMAL1 complex [6]. The next best studied enhancer is that of the *timeless* (*tim*) gene [11] which harbors closely spaced E and TER boxes, the latter being a variant of the consensus E-box which coincides with the mammalian E'-box [16]. In the mouse, well-studied CLOCK/BMAL1 elements include the *Per1* [6], *Per2* [17], *Avp* [14] and *Dbp* [18] genes. A study of the *Avp* promoter suggested that CLOCK/BMAL1 enhancers use a combination of a canonical E-box and a second more degenerate version thereof [14]. More recently a pyrimidine-rich 22 nucleotides sequence was found to cooperate with the core E-box in the *Avp* promoter [19]. So far, however, it was not possible to compile this information to build a

Editor: Uwe Ohler, Duke University, United States of America

Received September 10, 2007; **Accepted** January 4, 2008; **Published** February 15, 2008

A previous version of this article appeared as an Early Online Release on January 11, 2008 (doi:10.1371/journal.pcbi.0040038.eor).

Abbreviations: bp, base pair(s); bHLH, basic helix-loop-helix; DD, dark-dark; EL, expected log-likelihood; GR-CLK, glucocorticoid receptor-CLOCK fusion; HMM, hidden Markov model; LD, light-dark; LL, log-likelihood; PWM, position weight matrix; TSS, transcription start site

* To whom correspondence should be addressed. E-mail: felix.naef@epfl.ch

Author Summary

Life on earth is subject to daily light/dark and temperature cycles that reflect the earth rotation about its own axis. Under such conditions, organisms ranging from bacteria to human have evolved molecularly geared circadian clocks that resonate with the environmental cycles. These clocks serve as internal timing devices to coordinate physiological and behavioral processes as diverse as detoxification, activity and rest cycles, or blood pressure. In insects and vertebrates, the clock circuitry uses interlocked negative feedback loops which are implemented by transcription factors, among which the heterodimeric activators CLOCK and CYCLE play a key role. The specific DNA elements recognized by this factor are known to involve E-box motifs, but the low information content of this sequence makes it a poor predictor of the targets of CLOCK/CYCLE on a genome-wide scale. Here, we use comparative genomics to build a more specific model for a CLOCK-controlled *cis*-element that extends the canonical E-boxes to a more complex dimeric element. We use functional data from *Drosophila* and mouse circadian experiments to test the validity and assess the performance of the model. Finally, we provide a phylogenetic analysis of the *cis*-elements across insect and vertebrates that emphasizes the ancient link between CLOCK/CYCLE and the modeled enhancer. These results indicate that comparative genomics provides powerful means to decipher the complexity of the circadian *cis*-regulatory code.

predictive algorithm for CLK/CYC or CLOCK/BMAL1-activated enhancers.

Computational strategies for the optimal discovery of *cis*-elements from genomic sequence pose formidable algorithmic challenges [20]. Among the many ways to model transcription factor binding sites, position weight matrices (PWMs) reflect most closely the biophysics of protein-DNA interactions [21–23]. Recent algorithms that exploit phylogeny to infer PWMs apply probabilistic (Gibbs) sampling to evolutionary models [24–26], or implement expectation maximization to optimize scoring schemes that incorporate phylogeny [27–30]. Most of these methods allow for relatively simple model architectures, mostly single block motifs or symmetric structures [31]. Hidden Markov Models (HMMs) [32] and their phylogenetic extensions [33,34] are best suited for more complex model structures like the one we use. The phylogenetic HMMs currently focus on optimizing trees [33] rather than motif identification; the latter would require optimizing the state dependent equilibrium frequencies. However conventional HMMs, for which motif training is well established, can be supplemented with a weighting scheme approximating the phylogenetic dependencies [35,36], which is what we will use here.

Our analysis starts with the five known CLK/CYC targets among the clock genes in *Drosophila*, *per* [9,10,37], *tim* [38], *vrille (vri)* [39], *Par-domain protein 1 (Pdp1)* [40], and *clockwork orange* (*cwo*) (formerly CG17100) [38,41,42]. Starting from the 69-bp enhancer in the *period* gene, we found a *cis*-element that is both common to all five genes and highly conserved among *Drosophila* species. This enhancer, which we validate using functional data, not only refines the core circadian E-box (E1), but also incorporates a flanking partner element (E2) that resembles the more degenerate E-box discussed above, and which is found at a very specific distance of the core E-box with an uncertainty of one nucleotide. While such structures are not implemented in common motif discovery programs, they are conveniently modeled with hidden Markov models (HMMs) [32]. We thus trained such an HMM model from the available fly sequences. Remarkably, the *Drosophila* model was able to predict many known mammalian CLOCK/BMAL1 targets without modification and with high specificity. A deeper phylogenetic analysis revealed the presence of the *cis*-element throughout insects and vertebrates. This shows that despite important differences in the organism's clock architectures, e.g., rhythmic mRNA accumulation of *Clock* in flies versus *Bmal1* in mammals, an ancient element in the circadian *cis*-regulatory code has been maintained since their common ancestor 500 million years ago.

Results

Evolutionary Conservation of CLK/CYC Enhancers in *Drosophila*

The 69-bp enhancer upstream of the *per* promoter in *D. melanogaster* was discovered and dissected in great detail [9]. Using genome sequences from 12 *Drosophila* species [43,44], we searched for presence of this enhancer in this clade (Figure 1). Although not immediate to find (in current UCSC alignment the enhancer is absent in half of the species), we identified sequences in all species that show remarkable conservation in a ~25 bp subfragment tightly collocated around the central canonical E-box motif (Figure 1A). The subfragment harbors a half E-box (GTG) located 9 bp to the right of the central E-box in the species close to *D. melanogaster*, and 10 bp for more remote clade members, e.g. *D. grimshawi*. Moreover the subfragment contains the 18 bp E-box [10] and the 3' flanking regions showing the strongest attenuation in activity upon deletion [15]. We then searched for similar flanking signals in the vicinity of other conserved E-boxes near promoters of validated CLK/CYC targets. We noticed that all five known target genes contain such dimeric signals that can be aligned with the *per* enhancer (Figure 1B), and also that this particular signal is conserved in all species considered.

Deriving a Probabilistic Model from Five Known CLK/CYC Targets in *Drosophila*

To make this more systematic we focus on the vicinity of all conserved E-boxes that can be found around the TSSs of the circadian transcripts per-RA, tim-RA, Pdp1-RD, vri-RA, and cwo-RA. We used multiple alignments from the UCSC browser (http://genome.ucsc.edu/) and considered all islands of ± 30 bp around degenerate CANNGT sequences that were present at least in the subclade consisting of *D. melanogaster, D. yakuba, D. simulans, D. sechellia,* and *D. erecta* (in total about 660 nucleotides per gene for each species, available at http://circaclock.epfl.ch/training_seqs.fa). While conservation often extends to all 12 species, sub-optimal alignments required that we apply this milder criterion (cf. alignment of *per*, Figure S3A). A preliminary motif finding analysis of this restricted set of sequences based on the MEME algorithm [45] (using motifs length of 7) confirmed the presence of E-box-like dimers in these sequences (Figure S1). These were spaced with an accuracy of plus or minus one base pair as in the *per* enhancer (Figure 1A). To model this configuration we

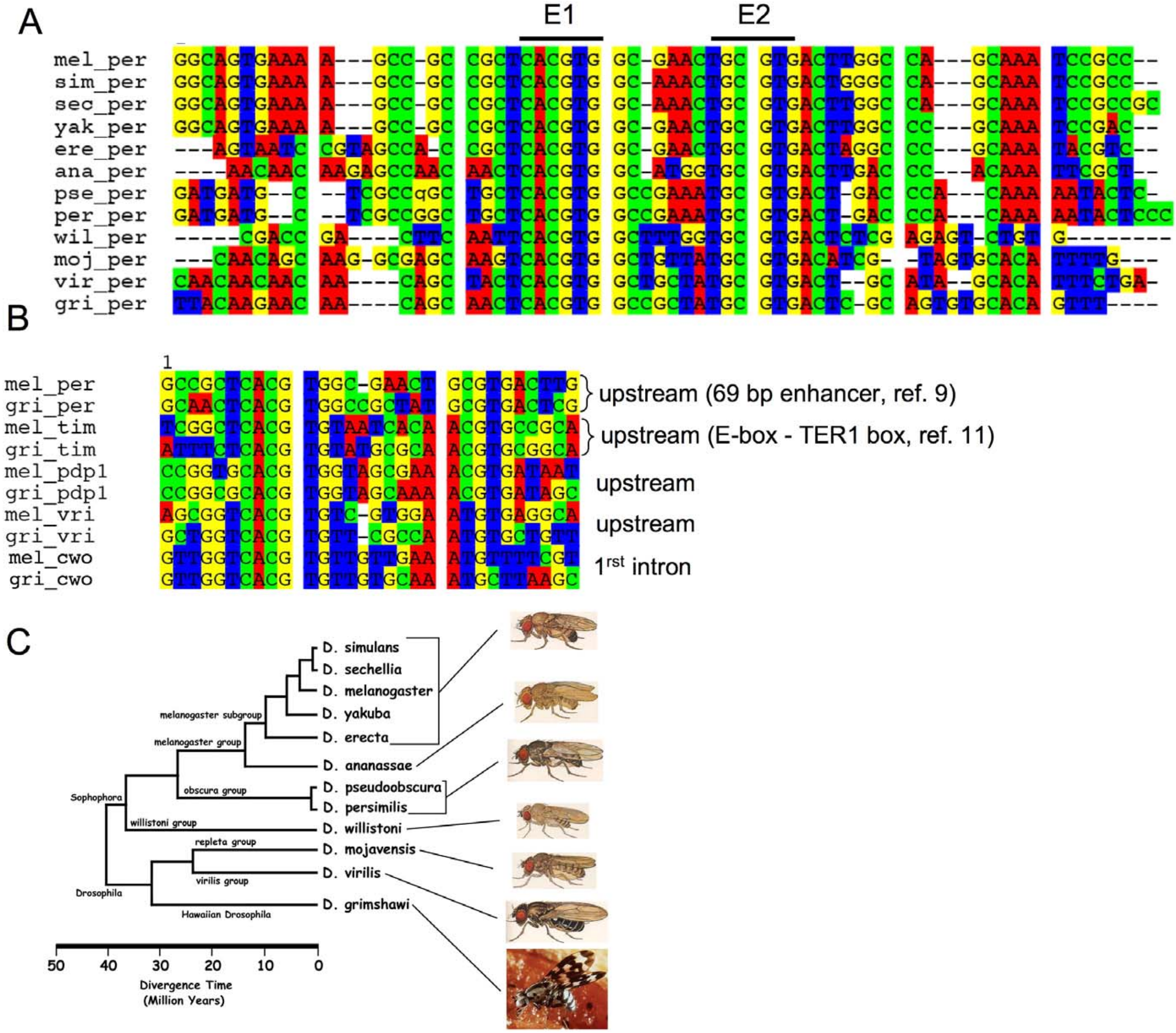

Figure 1. Period Enhancer and Related Sequences in Other CLK/CYC Targets

(A) Alignment of the central part of the 69-bp enhancer in the *Drosophila period* gene in the 12 species. Locations of E1 and E2 boxes are indicated.
(B) Similar sequences were found in promoters of the five known CLOCK/CYCLE targets: *period, timeless, vri, Pdp1,* and *cwo* (*CG17100*). Only the two most distant species are shown (*D. melanogaster* and *D. grimshawi*), but the elements are found across the full species tree.
(C) Species tree of sequenced *Drosophila* (from http://rana.lbl.gov/drosophila).
doi:10.1371/journal.pcbi.0040038.g001

implement a HMM reflecting the dimer structure (Figure 2A), and train the emission probabilities from the example sequences using the Baum-Welsh algorithm [32]. The model is cyclic so that several instances of the motifs can occur per sequence, we also allow to by-pass E2 in the case that it would not be sufficiently supported by the training sequences. We seeded the model only with one E-box (Figure 2B, left) flanked by a weak T nucleotide to break the palindrome symmetry of the bare E-box, while the putative partner site (E2) is initialized with a fully uninformative model. Only the emissions are trained while the transition probabilities p_1 from background to E1, and p_2 form E1 to E2 are held fixed (Methods). These transitions tune the stringency of the E1 and E2 parts, and reflect the chemical potential of the regulators that would bind to the E1 and E2 boxes [23]. We varied p_1 and p_2 over a wide range and retained the combination that maximizes the enrichment of hits among genes that show induction by CLK in functional genomics assays (Figure 3). Importantly, despite the uninformative seed and large search space, converged models do reflect the right flank described above for a wide range of transitions, the combination retained (p_1=2^{-11}, p_2=2^{-4}) show a AACGTG right consensus. Apart from details in the emission probabilities, this model is quite stable for a range of p_1 and p_2 values (Figure S2).

Inspection of the converged model indicates that effectively 15 high scoring instances of E1 box were used, and 6 for the E2 box. The latter were from *vri* (2–3 instances), *per* (1–2),

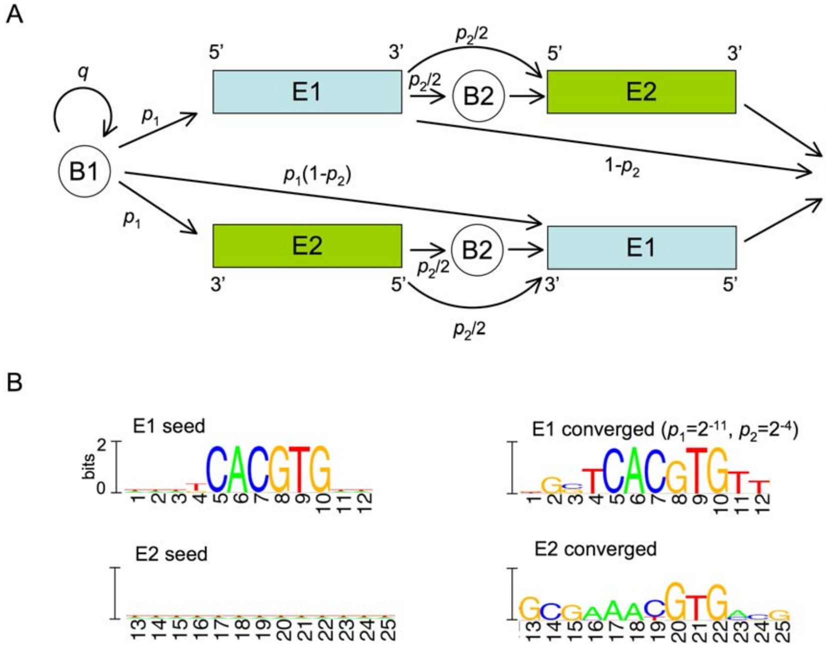

Figure 2. Probabilistic E1-E2 Model and Its Training

(A) Structure of the circular E1-E2 hidden Markov model (HMM). All transition probabilities are indicated except when they are equal to 1, and $q = 1 - 2p_1 - p_1(1 - p_2)$. The background states B1 and B2 have tied emission states; for the E1 (13 positions) and E2 boxes (12 positions) the emissions in the reverse strand model (lower part) are tied with those in the forward direction.

(B) The model is initialized with the matrices in the left and converges to the matrices on the right. Transition probabilities are those used throughout: $p_1 = 2^{-11}$, $p_2 = 2^{-4}$. All models are given at http://circaclock.epfl.ch/Models. Matrices are displayed in information format ($I_a = \max(0, p_a \log_2(p_a/q_a))$, where p_a are the probabilities for letter a in a column and q_a are the (genomic) background frequencies.

doi:10.1371/journal.pcbi.0040038.g002

tim (1), *Pdp1* (1) and *cwo* (1). In these five genes it is noticeable that multiple E1-E2 copies are found, and that E1 also often occurs alone (Figure S3). For instance, the second conserved site in the *per* intron (Figure S3A) could provide an explanation for the promoterless *per* allele found to cycle in a restricted part of the nervous system [46]. Thus, the converged model is consistent with the attenuated CLK/CYC activation in mutated 69-bp enhancers with deletions that are immediately 3' of the right central E-box [15]. Furthermore the model captures the mammalian architecture in which a canonical and a fuzzier E-box are juxtaposed [14].

Model Validation Using Functional Genomics Datasets

Training a model on five genes raises the question about its generalization to further putative CLK/CYC targets. To address this we used several microarray datasets that measure 'CLK targetness' [38] (Methods) and assessed correlation with sequence match from our model. Windows of ± 2500 bp around all annotated TSSs were scanned with our HMM model, in which the five training genes were found among the first 13 highest scores (Figure 3B).

Recently a glucocorticoid receptor-CLK fusion protein (GR-CLK) was used in S2 cells and cultured fly heads to induce CLK targets under cycloheximide treatment [38]. In this assay new protein synthesis is blocked to minimize indirect effects. Even though it is not formally excluded that the fusion protein could interfere with partner complexes, this experiment is best suited to test the specificity of the sequence model. We show that highly induced genes in the GR-CLK experiment are significantly enriched in high scoring hits from the sequence model, so that we can identify a set of ~30 genes among the top 57 induced genes which show highly significant 2- to 6-fold enrichment in sequence specificity (Figure 3A and Table S1). Importantly, the five training genes are excluded from the set of positives in this analysis. When testing how much E2 contributes to the observed enrichment, we found that it contributes only marginally: it reduces specificity for low sensitivities and increases specificity at higher sensitivities (Figure S4A). Nonetheless, several of the highly induced genes in the GR-CLK experiment, e.g., *CG13624*, show presence of E1-E2. Moreover, these sites show highly increased conservation profiles specifically at the predicted locations including the E2 site (Figure S3F and S3G). Below we show that increased specificity from E2 is most important in mammals.

We also considered expression levels in Clk^{Jrk} flies [47,48] since CLK/CYC targets are predicted to be down-regulated in this mutant. Moreover we tested cycling transcripts in light-dark (LD) and dark-dark (DD) conditions with phases that are compatible with known CLK/CYC targets, i.e., peak time accumulations in windows ZT6–20 (Methods). No signature of enriched E1-E2 motifs was detected in either the Clk^{Jrk} or cycler datasets (Figure S4). This can be expected since both differential expression in Clk^{Jrk} mutants, or rhythmic mRNA accumulation, also reflect indirect mechanisms downstream of the CLK/CYC transcription factor. We extensively searched whether other p_1 and p_2 parameters would detect enrichment without success. Consistently, we do not detect enrichment of the motif in mouse transcripts showing differential expres-

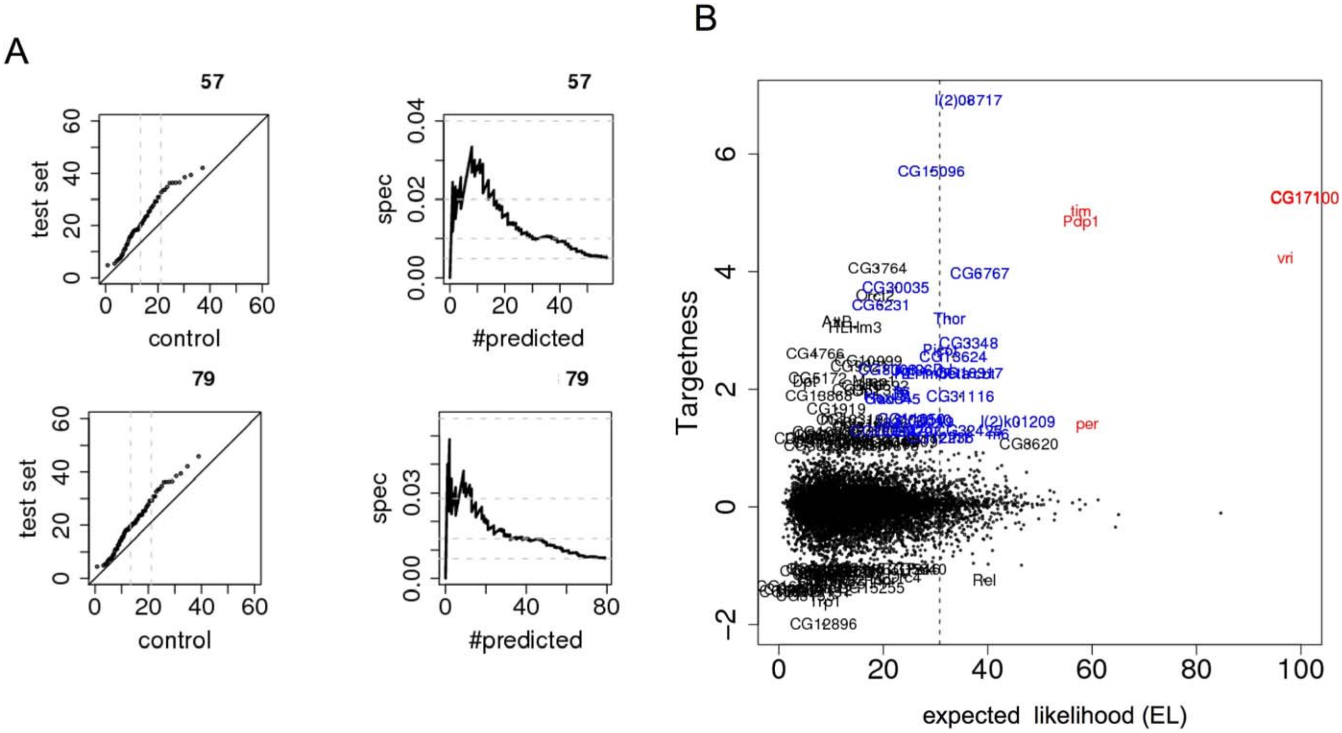

Figure 3. Promoters of Highly Induced Genes in the GR-CLOCK Are Enriched with High Scoring Hits from the E1-E2 Model

(A) Left two diagrams: quantile–quantile plots show that expected likelihood (EL) scores of highly induced genes (positives) are shifted upward with respect to the control (negatives). Positives correspond to the 57 (top 0.5%; upper left diagram) or 79 (top 0.7%; lower left diagram) induced genes (ranked according to fold induction; see Methods), while the negative set consists of all remaining genes. Right two diagrams: Specificity versus number of predicted genes (sensitivity) in the group of 57 (upper right diagram) or 79 (lower right diagram). The horizontal lines represent expected specificity (lowest line), 2-fold, 4-fold, and 8-fold enrichment. Importantly, the five training genes are excluded from the set of positives in all panels. The increased specificities are highly significant: in the top row, $p = 1.4 \times 10^{-7}$ for 10 predicted positives (chi-squared test), $p = 7.5 \times 10^{-6}$ for 20, and $p = 1.3 \times 10^{-3}$ for 30 predicted positives. The top 30 positives are marked in blue in (B).

(B) Scatter plot representing the targetness score (fold induction in $\log_2$ units; see Methods) in function of the expected log-likelihood score of the E1-E2 model in windows of $\pm$2,500 bp around the TSSs. Genes in blue are the 30 genes (from the group of 57) with highest match to the sequence model.

doi:10.1371/journal.pcbi.0040038.g003

sion in a recent mRNA profiling of *Clock* mutants [49] (Figure S5). Similarly, in an early study of rhythmic transcript profiles in fly heads, we did not detect enrichment of consensus E-boxes in the vicinity of periodic transcripts [50].

Further annotating the list of 57 GR-CLK induced genes with the sequence score from the E1-E2 model, the 24-hour periodicity and phase of the transcripts in LD and DD, or with the differential regulation in Clk^{Jrk} flies show that some genes qualify as CLK/CYC regulated genes according to several independent criteria (Table S1). Among those, the C2H2 zinc finger transcription factor *cbt, CG3348, CG11050, CG8008* are the most noticeable. From the purely genomic side, conserved E1-E2 sites are enriched in *D. melanogaster* when compared to permuted E1-E2 matrices (Figure S6A and S6B). From the likelihood scores of known examples, we estimate about one hundred genes to be potentially controlled by medium to high affinity E1-E2 sites (Figure S6C, gene lists in at http://circaclock.epfl.ch/fly_conserved_16.txt).

The *Drosophila* E1-E2 Model Also Predicts Circadian Genes in the Mouse

Even though the model was derived from fly sequences, the core E-box shows similarities to the brain-specific in vitro measured NPAS2/BMAL1 binding consensus GGGTCACGTGT[TC]C[AC] (underlined bases are consistent with our model) [51]. Scanning the mouse genome with the full E1-E2 model taken straight from the flies revealed that many common circadian transcripts show instances of this signal that are highly conserved in mammals (Figure S7). Several of these genes also contain multiple instances of the motif, as in the flies. With few exceptions, sites are found in the vicinity of the core promoter (e.g., *Per2*, *Tef*) or in the introns (*Dbp*, *Cry2*, *RevErbα*). Given the much greater complexity of mammalian genomes as compared to insects, it comes as a great surprise that the fly model predicts known circadian genes in mouse with highly enriched specificity (Figure 4). Among the 13 common circadian genes used as a test set, we find 7 among the top 1% of predictions when we would expect none ($p < 10^{-12}$, binomial distribution). In addition the restriction to sites that are highly conserved in mammals (measured using PhastCons [52]) increases the specificity (compare Figures 4 and S8). From the scores of known examples, we thus estimate in the order of hundred CLOCK/BMAL1 binding sites in mouse (Figure S6D). Finally, the two spacer lengths were about equally represented among the conserved hits with scores above 15 bits (given at http://circaclock.epfl.ch/bedFiles).

Importantly, while the E2 sequence played a marginal role in the specificity analysis of the GR-CLK data in flies, it plays

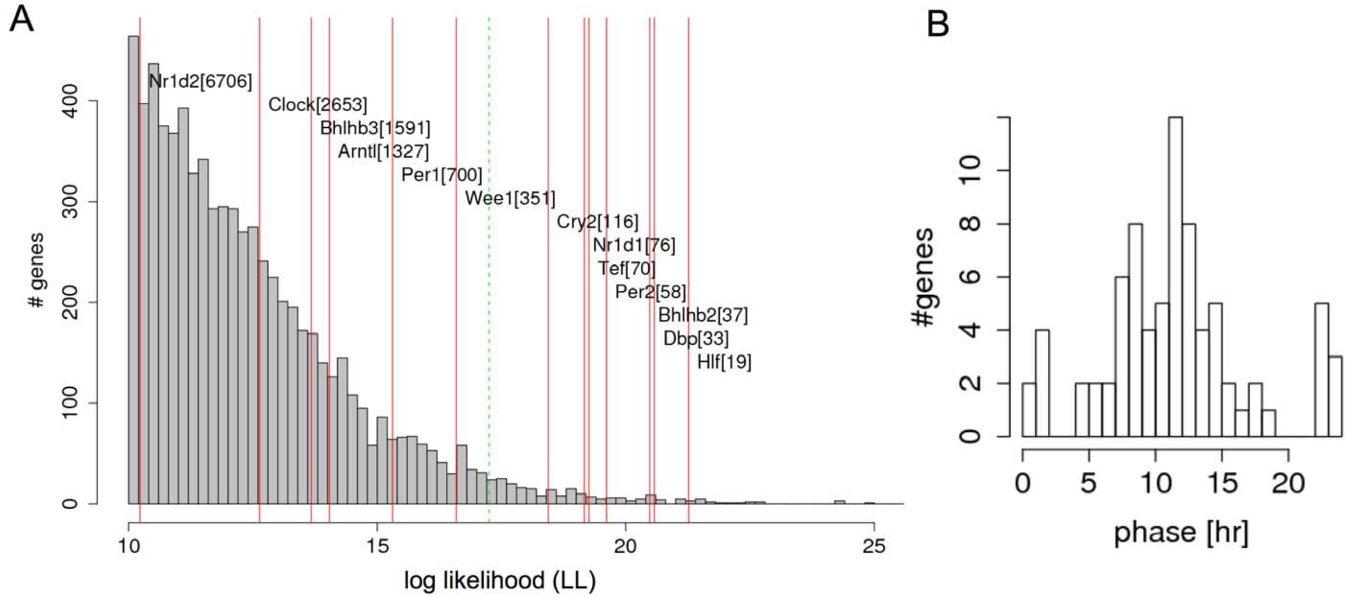

Figure 4. Genome-Wide Scan of the *Drosophila* E1-E2 Model onto the Mouse Genome

(A) Genome-wide rankings of circadian mouse genes in function of their sequence likelihood score.

The E1-E2 chained matrices are scanned along all mouse genes including ± 2 kb flanking sequence (Methods). Only hits with average PhastCons scores above 0.5 are counted (in PhastCons 0 implies no conservation and 1 maximal conservation). The unconstrained result is shown in Figure S8. The name of the gene is always aligned to the right of its score line (in red), and its rank is indicated in brackets. A total of 7 of 13 expected circadian genes (Methods) are found above the 1% line (green dashed line), while we expected zero at this cutoff ($p < 10^{-12}$). Notice Wee1 is just below the 1% line; E1-E2 is in the 3′ region for this gene (Figure S7). Known circadian genes represented are *Cry1, Cry2, Per1, Per2, Per3, Dbp, Tef, Hlf, Wee1, Bhlhb2 (Dec1), Bhlhb3 (Dec2), Nr1d1 (RevErbα), Nr1d2 (RevErbβ), Arntl (Bmal1),* and *Clock*. The latter two are expressed in anti-phase with respect to known target and are mainly controlled by *Ror* orphan receptors and their repressors *RevErbα,β* [70]. Thus, *Clock* and *Bmal1* are not included in the test set.

(B) Phase distribution of all conserved hits (as in [A]) with scores above 15 bits, and which show cycling in the liver in [54] (Fourier component F24 > 0.1; Methods).

doi:10.1371/journal.pcbi.0040038.g004

a much more prominent role in mouse. For example, the *Dbp* site ranks only at position 804 and that of *Per2* at position 3021 when E2 is not used in the prediction (Figure S8, right); overall the 13 test genes are clearly shifted to the bulk of scores. The conservation pattern of many of these hits shows tight increase around the E1-E2 sequences (Figure S7), which further supports the functional role of the predicted loci. Moreover, several of these predictions coincide with known CLOCK/BMAL1 functional circadian enhancers, e.g., those in the *Per1* [6], *Per2* [17] or *Dbp* [18] genes. As with the Drosophila Clk^{Jrk} data, putative CLOCK/BMAL1-induced genes identified from a *Clock* mutant array experiments in mice [49] did not show enriched E1-E2 boxes presumably due to indirect effects, except perhaps for a weak tendency in the liver (Figure S5). Consistent with our model, recent circadian band shift assays with mouse liver extracts indicate that a sequence closely related to the E1-E2 site is able to shift the CLOCK/BMAL1 complex more specifically than single E-boxes [53]. Finally, the phase distribution among the conserved hits that cycle in liver [54] shows a clear phase preference around ZT12, as expected for CLOCK/BMAL1 targets (Figure 4B).

Deep Evolutionary Conservation of the E1-E2 Motif

We first provide a phylogenic analysis of the activators CLOCK/BMAL1 binding E1 in mammals, birds, frogs, fishes, flies, mosquito and honey bee. Beyond these species, notably in the nematodes, no orthologues can be found. Both CLOCK and BMAL1 harbor two conserved PAS domains, in addition to the preserved DNA binding bHLH domain (Figure 5A; full-length protein alignments are given at http://circaclock.epfl.ch/jarFiles), whose conservation exceeds by far the bHLH consensus motif [55,56]. As the complex is expected to bind the E1 site, its conservation is consistent with the high information content (11.0 bits) of the E1 motif.

To track the presence of the E1-E2 motif in a broader set of species, we consider two gene families among the best conserved circadian CLK/CYC or CLOCK/BMAL1 targets. First, the *Period* genes are primary targets of CLOCK/BMAL1 whose genes products function as repressors of CLOCK/BMAL1, hence closing a negative feedback loop at the core of the circadian oscillator. While flies have a single *period* gene, vertebrates have multiple copies, e.g., three in mammals. The presence of E1-E2 signals near promoters of *period* genes generalizes beyond flies and mammals to a broad set of species including birds, frogs, fishes, flies, mosquito and honey bee (Figure 5B). While the mammalian site is at the TSS and that of fly is around −500 bp, the fish promoter is unannotated and the site is at 2.6 kbp upstream of the annotated PER3 protein. Interestingly the mammalian E2 motif shares many similar bases with the fish. Even though nematodes have a putative *period* homologue (*lin-42*), we could not detect presence a proximal E1-E2 in *C. elegans* and *C. briggsae*, which is both consistent with the absence of CLOCK/BMAL1 and the still uncertain existence of circadian rhythms in nematodes [57].

Second, the PAR-domain basic leucine zipper (PAR bZip)

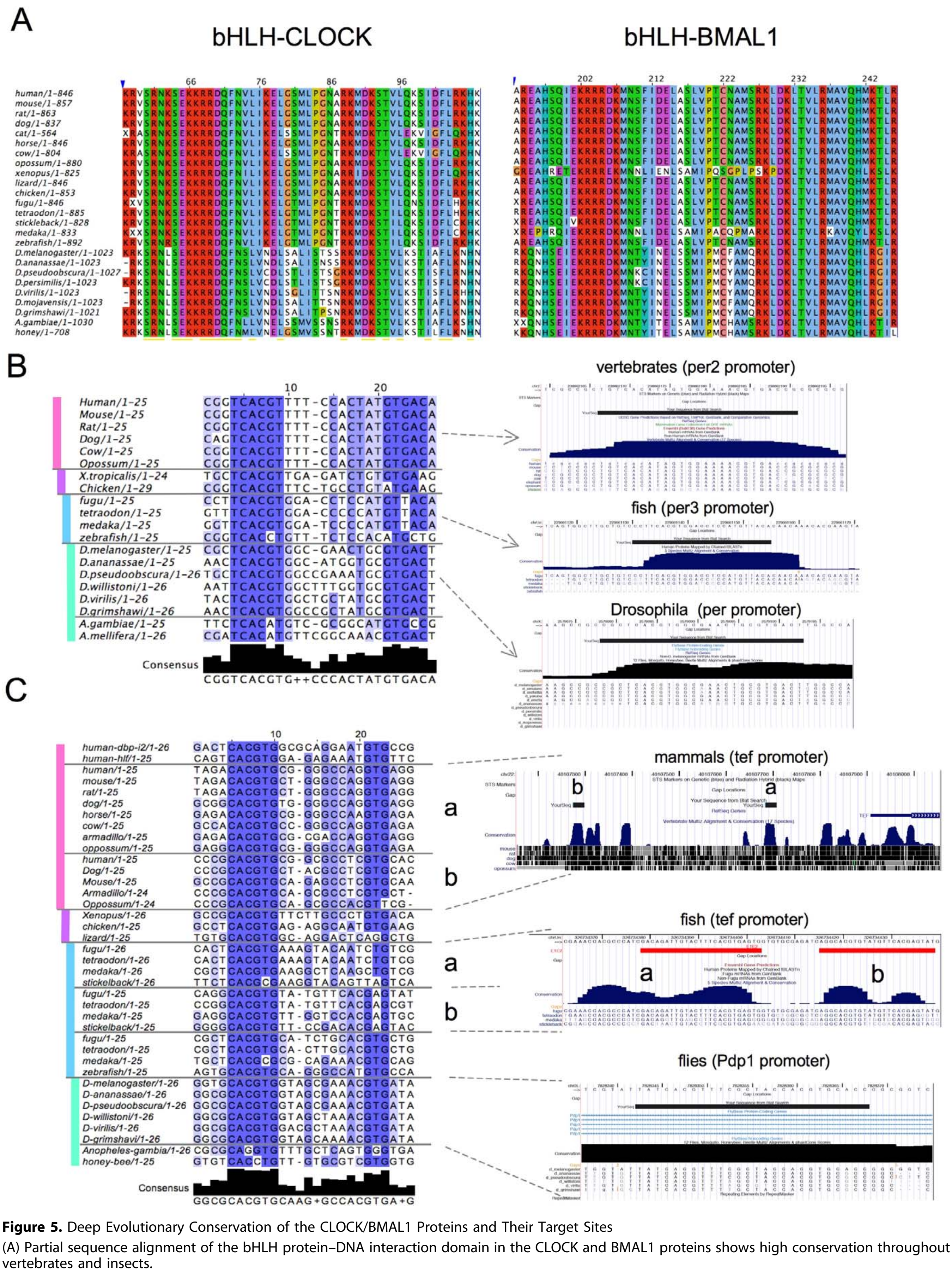

Figure 5. Deep Evolutionary Conservation of the CLOCK/BMAL1 Proteins and Their Target Sites

(A) Partial sequence alignment of the bHLH protein–DNA interaction domain in the CLOCK and BMAL1 proteins shows high conservation throughout vertebrates and insects.

(B) Left: E1-E2 motifs in Period genes; for space reasons too redundant species (e.g., the apes) are not shown. Notice the similarity in E2 between fish and mammals. Right (top): the murine *Per2* enhancer located near the TSS [17] can be aligned up to chicken. Right (middle): conserved E1-E2 enhancer upstream of *Per3* gene (2.6 kbp upstream of the start codon, the TSS is unannotated). Right (bottom): the *period* enhancer in flies (idem Figure S3A), positioned at –530 in *D. melanogaster*. As explained in the text, the more distant flies are missing from the MultiZ alignment, even though the enhancer is present (Figure 1A).
(C) Left: E1-E2 motifs in *Tef* genes. Sites in the human paralogues *Dbp* and *Hlf* are shown in the first two lines. Two sites are shown for the mammals and three for the fish. Right (top): the 5′ end of the *Tef* mRNA is shown with two conserved E1-E2 sites (see Figure S7G) located at –200 (a) and –600 (b) bp. Right (middle): two closely spaced E1-E2 sites in the *Tef* promoter at –500 bp of the putative start site. Right (bottom): the enhancer in the *Pdp1* gene (Figures 1B and S3D) at –2.2 kbp of the Pdp1-RD transcript.
doi:10.1371/journal.pcbi.0040038.g005

transcription factors *Tef/Hlf/Dbp* (mouse) are homologues of the fly circadian gene *Pdp1* (PAR domain protein 1) and are prominent clock output genes directly regulated by E-box motifs [12,18]. Their function is to mediate rhythmic physiology in organs such as the liver and kidney, where they induce, e.g., the cytochrome P450 enzymes [58]. Among the three murine paralogues, *Tef* is the most ancient representative with putative orthologues in most vertebrates and insects. In few species, e.g., in zebrafish and *Xenopus tropicalis*, full-length mRNA are available for *Tef*, elsewhere we relied on annotations inferred from a combination of ESTs and proteins (from other species) to genome alignments provided in the UCSC web browser. We could find E1-E2 elements in the vicinity of the *Tef* promoter in most of the vertebrates and insects, some harboring several copies (Figure 5C). Interestingly, the locations of the instances of the E1-E2 motif shows a typical conservation structure (in the PhastCons scores) in subgroups where non-coding sequences can be multiply aligned, i.e., the mammals, the fishes, and the flies.

Even if the exact position of the TSS is poorly documented in many of these species, we find that more than 85% of the shown sequences for both the *Period* and *Tef* genes occur within 1.5 kbp of an annotated start. Furthermore, 75% (respectively 25%) of the likelihood scores are above 15.1 bits (respectively 19.5 bits) and the median score is at 17.1 bits. Using background statistics for the E1-E2 likelihood score computed as in [23] (Figure S9), we estimate that the probability per position to find a motif having a likelihood score greater than 17 bits is 5×10^{-7}, or 2×10^{-6} for scores of 15 bits. Assuming independent positions, we estimate that the probability p to find conserved hits (PhastCons > 0.5) in regions of 1.5 kbp around the mammalian, fish and insect promoters is $p = 2 \times 10^{-9}$ for 17 bits hits and $p = 10^{-7}$ for 15 bits. Here we used that the genomic fraction of conserved sites (PhastCons > 0.5) is 10% in mammals (UCSC mm8 assembly, PhastCons score based on 18 species), 23% in fish (fr2 assembly, 4 species), and 40% flies (dm3, 15 species). This simple calculation thus suggests that the conserved configurations found for the *Period* and *Tef* genes are highly unlikely due to chance.

Discussion

Even though novel post-transcriptional mechanisms regulating the circadian clockworks are regularly uncovered [59], transcriptional control remains an essential ingredient of molecular clocks that is particularly relevant for relaying circadian output functions [2]. Output genes can be induced by the transcription factors of the core oscillator, or via tissue specific effectors such as *Dbp*, *Hlf* and *Tef* in mouse, which are themselves direct CLOCK/BMAL1 targets [58]. This layered design complicates the interpretation of experiments such as mRNA steady state time courses, particularly if one is interested in deciphering new direct targets of the core regulators. This task can be greatly facilitated using functional experiments like the glucocorticoid-CLK fusion experiments, which have improved specificity compared with the profiling of mutants, and accurate models for the *cis*-regulatory sequences bound by the regulators. Presently the mechanisms that facilitate the recruitment to DNA and subsequent trans-activating activity of the main circadian regulator CLK/CYC or CLOCK/BMAL1 are not fully understood. Likely though, this situation will evolve rapidly, helped by approaches such as large-scale chromatin immunoprecipitation analyses or comparative genomics. We used the latter to derive a probabilistic model for CLK/CYC-regulated circadian enhancers consisting of two partner signals, E1 and E2, linked by a spacer that can tolerate a variability of one nucleotide. E1 has an E-box core flanked by informative T's (or A on the reverse strand), while the second half is more degenerate and resembles previously reported TER boxes [11] or E' boxes [16]. The close proximity of the two sites suggests a cooperative binding of two partner complexes, one of which is the CLK/CYC heterodimer, while the second possibly identical factor needs to be identified.

To validate the predictive power of the model in *Drosophila,* we analyzed a recent study in which a GR-CLK fusion was used to induce CLK/CYC targets in S2 cells. We found an unusual number of high sequence scores among the highest induced genes, even though the E2 part did not contribute a large improvement in this case. This could reflect two scenarios: either the fusion protein interferes with a putative E2 binding complex, or it could simply be that the list of highest affinity CLK/CYC targets does not extend much beyond the list of known five, even though we identified several strong candidates that harbor the expected *cis*-element (Figure S3 and Table S1). Consistent with the first functional study of the period enhancer [9] we find no preferential orientation of the E1-E2 elements. Anecdotally, it is interesting that the double E1-E2 site around –2.5 kb in the *vrille* promoter (Figure S3C) is located on a fragment that is inverted in *D. grimshawi* only (Figure S10).

Having built the model from *Drosophila* sequences only, it was quite remarkable that the unchanged E1-E2 model identified high scoring hits in the majority of known CLOCK/BMAL1 targets in mouse. Among genes with putative E1-E2 elements, many instances of the motif are highly conserved, and the conservation patterns are often concentrated just on top of the identified elements while rapidly decreasing outside of it. Unlike in flies, the E2 element appears to be a determinant for specificity in mouse. Given that tissue-specific expression analyses [60,61] revealed largely non-overlapping circadian regulation programs, it is not excluded that future analyses will reveal enhancer elements permitting tissue specific predictions. We showed that our model predicted peak expression phases in mouse liver that were preferentially centered around ZT12 (Figure 4B), which is consistent with an induction by CLOCK/BMAL1. It might be possible to find subclasses in the E1-E2 model that drive expression with more specific phases, e.g., by modifying the binding affinity of the E2 element. There should nevertheless be limits to this undertaking as mRNA accumulation is also influenced by processes downstream of transcription. Noticeably, many of our predicted CLOCK/BMAL1 targets show non-cycling steady state mRNA abundances, at least when assessed in liver [54]. It is likely that some will cycle in other tissues, however, long mRNA half-lives can easily mask rhythmic transcription rates as has been reported for the albumin gene [62].

In conclusion we built a probabilistic sequence model, termed E1-E2, that predicts enhancers driven by the bHLH proteins CLK/CYC in insects and CLOCK/BMAL1 in mammals. This model not only refines the circadian E-box beyond its core nucleotides but also emphasizes the role of a flanking partner motif that may involve binding of a novel co-regulator complex. A deeper phylogenetic analysis showed that conserved instances of E1-E2 are found both in promoters of core circadian clock genes, and in genes mediating circadian output. E1-E2 seems to occur in vertebrates and insects but not in nematodes. This is perhaps not surprising as the existence of circadian behavior in nematodes is still controversial [57]. Absence of E1-E2 could also reflect the Coelomata hypothesis that groups arthropodes with chordates in a monophyletic clade [63]. In this perspective our findings would suggest that the CLOCK/BMAL1 based oscillator evolved after the nematodes separated from a common ancestor. Alternatively, the nematodes could have lost some oscillator components as a result of their live style in the soil, which largely shields them from daily light cues. Our report is not the first example of an ancient linkage between bHLH regulators and companion *cis*-elements. An even deeper conservation of a *cis*-regulatory element has been reported in proneural genes controlled by bHLH factors of the Hes family [64]. Several reasons, e.g., the necessity to maintain highly stable key developmental programs, were proposed to explain such unusually high conservation. Here, it is interesting that the BMAL1 protein, unlike genes in the *Period* or *Crytochromes* families, stands out as the only circadian component in the murine clock with no functionally redundant paralogues. The high degree of conservation in its target sites is thus consistent with the unique function of BMAL1 (CYC) as the master activator in the circadian network. We surely expect that comparative genomics combined with functional datasets will allow further dissecting the circadian and other *cis*-regulatory codes.

Methods

***Drosophila* sequence data.** MultiZ [65] Multiple alignments were downloaded from the UCSC table browser (Multiple alignments of 14 insects with *D. melanogaster*, dm3, April 2006, but we restricted these to *Drosophila* species). We used the *Drosophila melanogaster* genome and annotations version r5.1 to analyze windows of ±2,500 bases around all annotated transcripts. These sequences were used to identify flanking sequences around conserved CANNGT motifs in the five training genes; for the *period* gene we added the 69-bp enhancer from the species missed in the multiple alignment (Figure 1A and Figure S3A).

Model training. The sequences used for the model training are given at http://circaclock.epfl.ch/training_seqs.fa. We implemented a standard Baum-Welsh optimization in which each sequence is independent (no explicit use of the multiple alignments is made). We took into account phylogenetic relationships by attributing a geometric weighting reminiscent of [35] reflecting the *Drosophila* species tree (Figure 1C): droGri2: weight = 1/8, dp4: 1/8, droYak2: 1/16, droEre2: 1/16, droPer1: 1/8, droWil1: 1/4, droSim1: 1/32, dm3: 1/16, droAna3: 1/8, droSec1: 1/32, droMoj3: 1/16, droVir3: 1/16. Thus each gene is counted as one and we used fixed pseudo-count of 0.3 for each nucleotide. Species identifiers are those used in the UCSC alignments. Training is done on both strands simultaneously with tied (reverse complemented) emission probabilities using a custom HMM implementation following [32].

Genome scans in *Drosophila*. We scanned (decoded) windows of ±2,500 bp for all annotated transcripts (r5.1) with the cyclic E1-E2 model. The converged HMM model is provided at http://circaclock.epfl.ch/Models/M_11_4_0.3_3_2_13_0_1.mod, while the seed model is http://circaclock.epfl.ch/Models/seed.M_11_4_3_2_13_0_1.mod.

We used posterior decoding to compute the posterior state probabilities P_{si} for state s at position i (Figure S3), and the expected likelihood (EL) for a sequence is computed as $\sum_{si} P_{si}\log_2(e_s\ (O_i))$ minus the likelihood of the background (Figure 3). Here, $e_s(O_i)$ is the (emission) probability to observe nucleotide O_i at position i in the state s. In the case of multiple transcripts, the highest score was used as the gene score. Correspondence between Affymetrix oligos and genes was done with the Annotations provided at NetAffx.com for the DrosGenome1 and Drosophila_2 arrays (July 2007 versions).

Genome scans in mouse. To scan the full mm8 mouse genome (from the UCSC genome browser) we extracted the two weight matrices from the *Drosophila* HMM (given at http://circaclock.epfl.ch/Models/M_11_4_0.3_3_2_13_0_1.p1.mat and http://circaclock.epfl.ch/Models/M_11_4_0.3_3_2_13_0_1.p2.mat), and computed the standard likelihood (LL) $\sum_i \log_2(w_i(O_i)/b(O_i))$ for the chained matrices at each genomic position. Here $w_i(O_i)$ is the probability to observe nucleotide O_i at position i and $b(O_i)$ is the background probability for nucleotide O_i. As in flies we allow for a zero or one nucleotide spacer and consider the maximum of the two scores. We used a single nucleotide background (0-th order) with 29% of A and T's, and 21% of C or G's. To filter for conservation (Figures 4 and S8), we average PhastCons scores [52] (from alignments with 17 vertebrates, UCSC genome browser) at the positions of the hit (25 or 26 bases depending on spacer). Hits are mapped to genes when they occur in windows of ±2 kb of the transcription units from the affyMOE430 table at UCSC. The latter was used for easy comparison with expression data. A set of 15 known circadian genes was used to test the specificity of prediction in mouse: *Cry1, Cry2, Per1, Per2, Per3, Dbp, Tef, Hlf, Wee1, Bhlhb2 (Dec1), Bhlhb3 (Dec2), Nr1d1 (RevErbα), Nr1d2 (RevErbβ), Bmal1 (Arntl),* and *Clock*, of which the latter two are not expected to be self-induced.

Array datasets. Two Clk^{Jrk} mutant time series of 12 time points each [47,48] were used to quantify differential regulation induced by the mutation, we applied a one-sample t-test to the 24 merged $\log_2$-expression ratios at each time point. GR-CLK induction data was from [38]; replicated conditions were averaged and the fold induction between stimulated and un-stimulated cells was computed separately for the S2 cells and the cultured fly heads. The two were then summed to make a single score for each gene. The obtained rankings correlate tightly with the original analysis. DD and LD cycling scores (amplitude and phases) were compiled from a comprehensive collection of previously described time courses [66] using established methods [67]. Information regarding genes induced or repressed in *Clock* mutant mice is taken from [49]. Mouse liver data used for Figure 3B is from [54], rhythmicity was assessed via the 24 hour Fourier component (F24) as in [67].

Phylogenetic analysis. The protein sequences for the CLOCK and BMAL1 homologues of insects and vertebrates, was taken from NCBI when available, and if not, we used the tBlastn table and the predicted protein from the UCSC database. The protein alignments were produced using ClustalW [68] and visualized using Jalview [69]. To identify instances of the E1-E2 motifs in the *Period* and *Tef* promoters, we used the UCSC browser to find the genomic regions around *bona fide* (when supported by full length mRNA in the specie) or putative (inferred from aligning mRNA, ESTs or protein from other species) homologues of these genes. We then scanned these sequences for instances of E1-E2 and the highest scoring instances near putative promoters were retained.

Website. Additional data and model files are given at http://circaclock.epfl.ch. Genes used for Figure 4B are listed in the file http://circaclock.epfl.ch/cyclers__mouse__fig4B.txt. Predictions (.bed file format) for flies and mouse can be uploaded to the UCSC Genome browser as custom tracks.

Supporting Information

Figure S1. MEME Analysis of Sequences Surrounding Conserved CACGTG in Five Circadian Genes in Flies

Found at doi:10.1371/journal.pcbi.0040038.sg001 (1.7 MB PDF).

Figure S2. Models Obtained with Different Transition Probabilities *p1* and *p2*

Found at doi:10.1371/journal.pcbi.0040038.sg002 (391 KB PDF).

Figure S3. Positions of E1 and E2 sites Around TSSs of Circadian Genes in Flies

Found at doi:10.1371/journal.pcbi.0040038.sg003 (4.5 MB PDF).

Figure S4. In *Drosophila,* E1-E2 Signal Is Not Enriched in Cyclic Genes or Genes That Are Downregulated in Clk^{Jrk} Flies

Found at doi:10.1371/journal.pcbi.0040038.sg004 (1.0 MB PDF).

Figure S5. E1-E2 Enhancers in Δ19 *Clock* Mutants

Found at doi:10.1371/journal.pcbi.0040038.sg005 (262 KB PDF).

Figure S6. Estimated Number of E1-E2 Sites in Flies and Mice

Found at doi:10.1371/journal.pcbi.0040038.sg006 (527 KB PDF).

Figure S7. Position of E1-E2 Enhancers in Circadian Mouse Genes

Found at doi:10.1371/journal.pcbi.0040038.sg007 (1.3 MB PDF).

Figure S8. Sequence Conservation and E2 Improve the Specificity of E1-E2 Sites in Mouse

Found at doi:10.1371/journal.pcbi.0040038.sg008 (388 KB PDF).

Figure S9. Background Distribution for the Likelihood Score

Found at doi:10.1371/journal.pcbi.0040038.sg009 (163 KB PDF).

Figure S10. Sequence Inversion in the *D. Grimshavi vrille* Promoter

Found at doi:10.1371/journal.pcbi.0040038.sg010 (387 KB PDF).

Table S1. Top 57 GR-CLK–Induced Genes

Found at doi:10.1371/journal.pcbi.0040038.st001 (77 KB PDF).

Acknowledgments

We thank Hans Reinke and Juergen Ripperger for insightful discussions, Ueli Schibler and Herman Wijnen for their suggestions on the manuscript, and Bernhard Sonderegger with help on the computation of the background model.

Author contributions. FN conceived and designed the experiments. ERP and GR performed the experiments. ERP, GR, and FN analyzed the data. ERP and FN wrote the paper.

Funding. ERP acknowledges support from a National Institutes of Health (NIH) administrative supplement to parent grant GM54339. FN receives support from the National Competence Center in Research (NCCR) program in Molecular Oncology and the Swiss National Science Foundation.

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

References

1. Young MW, Kay SA (2001) Time zones: A comparative genetics of circadian clocks. Nat Rev Genet 2: 702–715.
2. Schibler U, Naef F (2005) Cellular oscillators: Rhythmic gene expression and metabolism. Curr Opin Cell Biol 17: 223–229.
3. Allada R, White NE, So WV, Hall JC, Rosbash M (1998) A mutant *Drosophila* homolog of mammalian Clock disrupts circadian rhythms and transcription of period and timeless. Cell 93: 791–804.
4. Rutila JE, Suri V, Le M, So WV, Rosbash M, et al. (1998) CYCLE is a second bHLH-PAS clock protein essential for circadian rhythmicity and transcription of *Drosophila* period and timeless. Cell 93: 805–814.
5. King DP, Zhao Y, Sangoram AM, Wilsbacher LD, Tanaka M, et al. (1997) Positional cloning of the mouse circadian clock gene. Cell 89: 641–653.
6. Gekakis N, Staknis D, Nguyen HB, Davis FC, Wilsbacher LD, et al. (1998) Role of the CLOCK protein in the mammalian circadian mechanism. Science 280: 1564–1569.
7. Reick M, Garcia JA, Dudley C, McKnight SL (2001) NPAS2: An analog of clock operative in the mammalian forebrain. Science 293: 506–509.
8. DeBruyne JP, Weaver DR, Reppert SM (2007) CLOCK and NPAS2 have overlapping roles in the suprachiasmatic circadian clock. Nat Neurosci 10: 543–545.
9. Hao H, Allen DL, Hardin PE (1997) A circadian enhancer mediates PER-dependent mRNA cycling in *Drosophila melanogaster*. Mol Cell Biol 17: 3687–3693.
10. Darlington TK, Lyons LC, Hardin PE, Kay SA (2000) The period E-box is sufficient to drive circadian oscillation of transcription in vivo. J Biol Rhythms 15: 462–471.
11. McDonald MJ, Rosbash M, Emery P (2001) Wild-type circadian rhythmicity is dependent on closely spaced E boxes in the *Drosophila* timeless promoter. Mol Cell Biol 21: 1207–1217.
12. Ripperger JA, Shearman LP, Reppert SM, Schibler U (2000) CLOCK, an essential pacemaker component, controls expression of the circadian transcription factor DBP. Genes Dev 14: 679–689.
13. Kyriacou CP, Rosato E (2000) Squaring up the E-box. J Biol Rhythms 15: 483–490.
14. Munoz E, Brewer M, Baler R (2002) Circadian transcription. Thinking outside the E-box. J Biol Chem 277: 36009–36017.
15. Lyons LC, Darlington TK, Hao H, Houl J, Kay SA, et al. (2000) Specific sequences outside the E-box are required for proper per expression and behavioral rescue. J Biol Rhythms 15: 472–482.
16. Ueda HR, Hayashi S, Chen W, Sano M, Machida M, et al. (2005) System-level identification of transcriptional circuits underlying mammalian circadian clocks. Nat Genet 37: 187–192.
17. Yoo SH, Ko CH, Lowrey PL, Buhr ED, Song EJ, et al. (2005) A noncanonical E-box enhancer drives mouse Period2 circadian oscillations in vivo. Proc Natl Acad Sci U S A 102: 2608–2613.
18. Ripperger JA, Schibler U (2006) Rhythmic CLOCK-BMAL1 binding to multiple E-box motifs drives circadian Dbp transcription and chromatin transitions. Nat Genet 38: 369–374.
19. Munoz E, Brewer M, Baler R (2006) Modulation of BMAL/CLOCK/E-Box complex activity by a CT-rich *cis*-acting element. Mol Cell Endocrinol 252: 74–81.
20. Tompa M, Li N, Bailey TL, Church GM, De Moor B, et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23: 137–144.
21. Berg OG, von Hippel PH (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. J Mol Biol 193: 723–750.
22. Stormo GD, Fields DS (1998) Specificity, free energy and information content in protein-DNA interactions. Trends Biochem Sci 23: 109–113.
23. Djordjevic M, Sengupta AM, Shraiman BI (2003) A biophysical approach to transcription factor binding site discovery. Genome Res 13: 2381–2390.
24. Li X, Wong WH (2005) Sampling motifs on phylogenetic trees. Proc Natl Acad Sci U S A 102: 9481–9486.
25. Siddharthan R, Siggia ED, van Nimwegen E (2005) PhyloGibbs: A Gibbs sampling motif finder that incorporates phylogeny. PLoS Comput Biol 1: e67.
26. Newberg LA, Thompson WA, Conlan S, Smith TM, McCue LA, et al. (2007) A phylogenetic Gibbs sampler that yields centroid solutions for *cis*-regulatory site prediction. Bioinformatics 23: 1718–1727.
27. Wang T, Stormo GD (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. Bioinformatics 19: 2369–2380.
28. Moses AM, Chiang DY, Eisen MB (2004) Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. Pac Symp Biocomput: 324–335.
29. Prakash A, Blanchette M, Sinha S, Tompa M (2004) Motif discovery in heterogeneous sequence data. Pac Symp Biocomput: 348–359.
30. Sinha S, Blanchette M, Tompa M (2004) PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences. BMC Bioinformatics 5: 170.
31. Favorov AV, Gelfand MS, Gerasimova AV, Ravcheev DA, Mironov AA, et al. (2005) A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. Bioinformatics 21: 2240–2245.
32. Durbin R, Eddy S, Krogh A, Mitchison G (1999) Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge (United Kingdom): Cambridge University Press. 356 p.
33. Felsenstein J, Churchill GA (1996) A Hidden Markov Model approach to variation among sites in rate of evolution. Mol Biol Evol 13: 93–104.
34. Siepel A, Haussler D (2004) Combining phylogenetic and hidden Markov models in biosequence analysis. J Comput Biol 11: 413–428.
35. Thompson JD, Higgins DG, Gibson TJ (1994) Improved sensitivity of profile

searches through the use of sequence weights and gap excision. Comput Appl Biosci 10: 19–29.
36. Gerstein M, Sonnhammer EL, Chothia C (1994) Volume changes in protein evolution. J Mol Biol 236: 1067–1078.
37. Lee C, Bae K, Edery I (1999) PER and TIM inhibit the DNA binding activity of a *Drosophila* CLOCK-CYC/dBMAL1 heterodimer without disrupting formation of the heterodimer: A basis for circadian transcription. Mol Cell Biol 19: 5316–5325.
38. Kadener S, Stoleru D, McDonald M, Nawathean P, Rosbash M (2007) Clockwork Orange is a transcriptional repressor and a new *Drosophila* circadian pacemaker component. Genes Dev 21: 1675–1686.
39. Blau J, Young MW (1999) Cycling vrille expression is required for a functional *Drosophila* clock. Cell 99: 661–671.
40. Cyran SA, Buchsbaum AM, Reddy KL, Lin MC, Glossop NR, et al. (2003) vrille, Pdp1, and dClock form a second feedback loop in the *Drosophila* circadian clock. Cell 112: 329–341.
41. Lim C, Chung BY, Pitman JL, McGill JJ, Pradhan S, et al. (2007) Clockwork orange encodes a transcriptional repressor important for circadian-clock amplitude in *Drosophila*. Curr Biol 17: 1082–1089.
42. Matsumoto A, Ukai-Tadenuma M, Yamada RG, Houl J, Uno KD, et al. (2007) A functional genomics strategy reveals clockwork orange as a transcriptional regulator in the *Drosophila* circadian clock. Genes Dev 21: 1687–1700.
43. Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, et al. (2007) Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. Nature 450: 219–232.
44. Consortium DG (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450: 203–218.
45. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2: 28–36.
46. Frisch B, Hardin PE, Hamblen-Coyle MJ, Rosbash M, Hall JC (1994) A promoterless period gene mediates behavioral rhythmicity and cyclical per expression in a restricted subset of the *Drosophila* nervous system. Neuron 12: 555–570.
47. Ceriani MF, Hogenesch JB, Yanovsky M, Panda S, Straume M, et al. (2002) Genome-wide expression analysis in *Drosophila* reveals genes controlling circadian behavior. J Neurosci 22: 9305–9319.
48. Ueda HR, Matsumoto A, Kawamura M, Iino M, Tanimura T, et al. (2002) Genome-wide transcriptional orchestration of circadian rhythms in *Drosophila*. J Biol Chem 277: 14048–14052.
49. Miller BH, McDearmon EL, Panda S, Hayes KR, Zhang J, et al. (2007) Circadian and CLOCK-controlled regulation of the mouse transcriptome and cell proliferation. Proc Natl Acad Sci U S A 104: 3342–3347.
50. Claridge-Chang A, Wijnen H, Naef F, Boothroyd C, Rajewsky N, et al. (2001) Circadian regulation of gene expression systems in the *Drosophila* head. Neuron 32: 657–671.
51. Hogenesch JB, Gu YZ, Jain S, Bradfield CA (1998) The basic-helix-loop-helix-PAS orphan MOP3 forms transcriptionally active complexes with circadian and hypoxia factors. Proc Natl Acad Sci U S A 95: 5474–5479.
52. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034–1050.
53. Reinke H, Fleury-Olea F, Dibner C, Benjamin IJ, Schibler U (2008) Differential display of DNA-binding proteins reveals heat shock factor 1 as a circadian transcription factor. Genes Dev 22: 331–345.
54. Kornmann B, Schaad O, Bujard H, Takahashi JS, Schibler U (2007) System-driven and oscillator-dependent circadian transcription in mice with a conditionally active liver clock. PLoS Biol 5: e34.
55. Littlewood TD, Evan GI (1995) Transcription factors 2: Helix-loop-helix. Protein Profile 2: 621–702.
56. Ledent V, Vervoort M (2001) The basic helix-loop-helix protein family: Comparative genomics and phylogenetic analysis. Genome Res 11: 754–770.
57. Simonetta SH, Golombek DA (2007) An automated tracking system for *Caenorhabditis elegans* locomotor behavior and circadian studies application. J Neurosci Methods 161: 273–280.
58. Gachon F, Olela FF, Schaad O, Descombes P, Schibler U (2006) The circadian PAR-domain basic leucine zipper transcription factors DBP, TEF, and HLF modulate basal and inducible xenobiotic detoxification. Cell Metab 4: 25–36.
59. Merrow M, Mazzotta G, Chen Z, Roenneberg T (2006) The right place at the right time: Regulation of daily timing by phosphorylation. Genes Dev 20: 2629–2623.
60. Panda S, Antoch MP, Miller BH, Su AI, Schook AB, et al. (2002) Coordinated transcription of key pathways in the mouse by the circadian clock. Cell 109: 307–320.
61. Storch KF, Lipan O, Leykin I, Viswanathan N, Davis FC, et al. (2002) Extensive and divergent circadian gene expression in liver and heart. Nature 417: 78–83.
62. Wuarin J, Schibler U (1990) Expression of the liver-enriched transcriptional activator protein DBP follows a stringent circadian rhythm. Cell 63: 1257–1266.
63. Philip GK, Creevey CJ, McInerney JO (2005) The Opisthokonta and the Ecdysozoa may not be clades: Stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. Mol Biol Evol 22: 1175–1184.
64. Rebeiz M, Stone T, Posakony JW (2005) An ancient transcriptional regulatory linkage. Dev Biol 281: 299–308.
65. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, et al. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 14: 708–715.
66. Wijnen H, Naef F, Boothroyd C, Claridge-Chang A, Young MW (2006) Control of daily transcript oscillations in *Drosophila* by light and the circadian clock. PLoS Genet 2: e39.
67. Wijnen H, Naef F, Young MW (2005) Molecular and statistical tools for circadian transcript profiling. Methods Enzymol 393: 341–365.
68. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.
69. Clamp M, Cuff J, Searle SM, Barton GJ (2004) The Jalview Java alignment editor. Bioinformatics 20: 426–427.
70. Preitner N, Damiola F, Lopez-Molina L, Zakany J, Duboule D, et al. (2002) The orphan nuclear receptor REV-ERBalpha controls circadian transcription within the positive limb of the mammalian circadian oscillator. Cell 110: 251–260.