SI Appendix

SI Materials and Methods

Construction of Promoter/Enhancer Database. Detailed database construction procedures are described here. We performed the following steps to construct the mammalian promoter/enhancer database. We mapped RefSeq (1) (15/11/2004 release 28,712 sequences for human and 8/11/2004 release 26,221 sequences for mouse), UniGene (2) (4,790,589 sequences of build #175 for human and 3,650,800 sequences of build #141 for mouse) and 5′-end sequences (3) (1,394,825 sequences) to human (NCBI Human genome build 35) or mouse (NCBI Mouse genome build 33) genomes respectively using the BLAT program (4). The results of mapping and information from Ensembl genes (19,580 known genes and 2,641 novel genes for human, and 20,718 known genes and 4,665 novel genes for mouse), Ensembl EST genes (27,678 genes for mouse) (Version 26) (5) and Vega genes (6,166 genes for human) (6) were used to identify gene structure and putative transcriptional start site (TSS). We identified 24,749 human genes with 50,373 TSSs and 26,047 mouse genes with 43,863 TSSs. Next, we determined 16,268 human-mouse orthologues (65.7% of human genes and 62.5% of mouse genes) by selecting the reciprocal best match of all genes using the BLAST program.

As a next step, we sought to define orthologous genes across species. Using positional information of adjacent orthologues, we were able to determine 434 human-mouse syntenic regions. These regions were then compared using the unit of the blocks performed by the LAGAN program (7). We determined a total of 750,043 human-mouse genome conserved regions (173 Mb and 5.6% coverage of human genome, and 172 Mb and 6.5% of mouse genome). Information regarding the coding regions was obtained from RefSeq, UniGene, Ensembl Genes and Vega Genes, allowing the non-coding genome conserved regions to be defined. Through this process, we were able to determine 893,798 human and 892,128 mouse human-mouse conserved non-coding regions (145 Mb and 4.7% coverage of human genome and 144 Mb and 5.5% of mouse genome). Next, we sought to define the transcription factor binding sites (TFBSs) by searching the known 862 consensus sequences obtained from TRANSFAC (8) at conserved regions of both genomes. In total, 7,804,559 human-mouse non-coding conserved TFBSs were predicted. Finally, data were integrated as a mammalian genome-wide promoter/enhancer database.

Determination of Genes and TSSs. Here we describe the detailed methods and conditions used in mapping RefSeq, UniGene and 5'-end sequences, determination of gene structures and TSSs, and determination of orthologues. All mRNA, EST and 5'-end sequences were mapped and selected by using blat, pslSort and pslReps programs of BLAT suit with the following parameter: "-q=rna -minIdentity=95" for blat and "-minCover=0.2 –minAli=0.98 –nearTop=0.002" for pslReps, this is determined based on the parameters used for constructing the UCSC genome browser database (9). We used repeat-masked genome sequences obtained from Ensembl (ftp://ftp.ensembl.org/) as the subject of mapping. Redundant mRNA sequences of UniGene, which contained part of the RefSeq sequences, were removed. If UniGene or the 5'-end sequence was mapped at a single continuous region, i.e. not spliced, it would not be used for further analysis in order to decrease genomic sequence contamination, transcripts of repetitive sequences, or pseudo genes, respectively. Determination of gene structures and TSSs was performed by the following five steps. First, we made transcriptional clusters by clustering overlapping chromosomal positions at the same strand of the results of mapping of RefSeq, UniGene and 5'-end sequences, and information of Ensembl genes, Ensembl EST genes and Vega genes. Second, we determined exons from the transcriptional cluster by selecting the regions with at least one or more sets of reliable information (RefSeq, Ensembl known gene or Vega gene). Third, we also determined exon-exon connections—i.e. the splicing junctions—by selecting the connections with at least one reliable set of information or two additional sets of information. Fourth, we determined the TSS or transcription termination site (TTS) containing exons by selecting exons with at least one or more sets of reliable information of 5'-end or 3'-end (5'-EST and mRNA sequence of UniGene and 5'-end sequence for TSS, and 3'-EST and mRNA sequence of UniGene for TTS). Lastly, we determined the position of TSS by selecting the most 5'-end position of reliable information or second most 5'-end position of other information. To determine human TSSs, we used 1,394,825 5'-end sequences from human full-length cDNA libraries generated by the oligo-cap method (3). In an attempt to determine mouse TSSs, we used UniGene sequences which contained 496,856 5'-end sequences from mouse full-length cDNA libraries that were generated by the FANTOM project using the cap-trapper method (10). To determine orthologues, we performed all genes to all genes BLAST homology search between the two species. Due to many of the genes having alternative splicing variants and alternative promoter variants, we used a merged gene sequence, which we were able to generate by joining determined exon sequences of the gene and using this for homology search. Following a nucleotide BLAST search with parameter "-E 1 –G 1", which reduced opening and extension of the alignment gap, and with a cutoff of E-value as 1.0e-4.0, we selected the reciprocal best match as orthologues.

Determination of Non-coding Genome Conserved Regions. Here we describe the detailed methods and conditions for determining genome conserved regions and coding regions. After the determination of syntenic regions by use of positional information of the adjacent orthologues, we used the following two steps to determine conserved regions in the genome was due to the size of the syntenic region being too long for comparing genome sequences between the two species. First, we divided the syntenic region into blocks by pairs of conserved exons between two species, which we determined by comparing the nucleotide sequence of exon regions of orthologous gene pairs using nucleotide BLAST. We then determined the conserved regions by comparing the unit of the blocks of synteny region of the two species. Genome comparisons were performed using the LAGAN program, and conserved regions were selected with 75% identity over 100 bp, as almost all of the evolutionally conserved elements previously determined (11) were detected using this cutoff value. These alignments were used for further analysis. Genome positions of the coding regions were determined by the following procedures. First, we compared the coding regions of RefSeq sequences with the mapped position of genome sequences. Second, we then compared the coding regions of mRNA and HTC sequences of UniGene with these mapped positions. Thirdly, we used the information from the coding regions of Ensembl genes and Vega genes. Finally, the genome position of all coding regions were merged and used for further analysis. In an attempt to determine the non-coding genome conserved regions, overlapping regions were removed.

Prediction of Evolutionally Conserved putative TFBSs. In this section we describe the prediction of evolutionally conserved putative TFBSs. We used 1,002 known consensus sequences of TFBS that were obtained from TRANSFAC. We searched the consensus sequences at non-coding genome conserved regions of the two species, then selected evolutionally conserved putative TFBSs that exist in the corresponding positions of the two species using alignments of sequence of genome conserved regions determined above. The search results of 140 consensus sequences of human-mouse conserved putative were not used as no conserved putative TFBS was detected at non-coding genome conserved regions or excess number (> 250,000) of conserved TFBSs were detected.

Annotation and Integration. Here we describe the annotation, integration, and interface to the database. All determined data, including genes, TSSs, non-coding genome conserved regions, and putative TFBSs, were related to gene annotations obtained from the Entrez Gene (2). Additionally, we performed a rpsblast search of the conserved domain database (CDD) (12) to estimate the molecular function of these genes. All data were integrated in a MySQL database. Finally, we constructed a web user interface for use with the mammalian promoter/enhancer database. This meant that a large amount of data, including the alignment of genome conserved regions, sequence and conservation of putative TFBSs, was made viewable by using the interface. Visualized genome conservation of genome conserved regions, which were outputted from the VISTA program (13), were also viewable. Additionally, to browse genome information annotated with determined genes, TSSs, non-coding genome conserved regions and putative TFBSs, we introduced the Ensembl Genome Browser (5) and the distributed annotation system (DAS) (14).

Determination of the distributions of distance from TSSs for natural and randomly positioned elements. To determine the distributions of distance from TSSs for clock-controlled elements, we used predicted 1,108 E-boxes, 2,314 D-boxes and 3,288 RREs (available on the circadian section of the mammalian promoter/enhancer database: http://promoter.cdb.riken.jp/circadian.html) and calculate the distance from nearest TSS. The distributions of distance from TSSs for random positioned clock-control elements were determined by randomly distribute the same number elements (1,108 for E-box, 2,314 for D-box and 3,288 for RRE) within conserved non-coding regions. The distributions of randomly positioned elements were generated 100 times and the averaged distributions were used.

Determination of FDR of putative elements. The false discovery rate (FDR) was determined for predicted elements. We used the result of predictions of a randomized genome as a background distribution of false positives. Profile HMM searches were performed in three conditions, (I) searching for conserved elements within conserved non-coding regions ("Conserved element"), (II) searching for mouse elements within the conserved non-coding regions by relaxing the requirement of element conservation ("Non-coding region") and (III) searching for mouse elements within the entire mouse genome by relaxing both element conservation and search space ("Whole genome"). The mouse genome sequence was randomized for these three conditions preserving the frequency of sequential nucleotide pair ("dinucleotide"), because a relatively small number of "CG" dinucleotides are existed in mammalian genomes. Otherwise (i.e. without preservation of dinucleotide), it will affect on the FDR for E-box, which contains "CG" dinucleotide. For the "Whole genome", a random mouse genome sequence was generated to keep the frequency of dinucleotide of the entire original mouse genome. For the "Non-coding region", a random genome sequence was generated to keep the frequency of each dinucleotide of the non-coding genome conserved regions of the original mouse genome. After randomization, genome sequence was masked except in the corresponding regions of the conserved non-coding genome regions of the original mouse genome. To reflect conservation between mouse and human among the "Conserved element", non-coding genome conserved regions of mouse and human genomes were generated to maintain the dinucleotide frequency at the parallel positions of the conserved genome region between mouse and human, including the case of a mismatch and or gap. For each condition, 200 randomized genome sequences were generated, and then HMM searches were performed. The averaged hit counts of 200 searches of these models on randomized genome sequence were used to obtain a background distribution and estimate the false discovery rate for each model taking into consideration the total number of base pairs searched. If the simulated value of FDR accidentally exceeds 1.0, the value is set to 1.0.

Animals. Male Balb/c mice (JAPS, Osaka, Japan) were purchased 5 weeks after birth. Mice were on a 12 hr light (400 lux): 12h dark (LD12:12) cycle for at least 2 weeks and were given food and water *ad lib*. Then animals were transferred to constant darkness conditions (DD) and, during the second DD cycle starting at CT0, animals were sacrificed every four hours under deep anesthesia and tissue samples were removed and frozen in liquid nitrogen for RNA extraction. This study was performed in compliance with the Rules and Regulations of the Animal Care and Use Committee, Kinki University School of Medicine, and followed the Guide for the Care and Use of Laboratory Animals, Kinki University School of Medicine.

Genome sequences. Genome sequences of NCBI Human genome build 35 and NCBI Mouse genome build 33 were downloaded from FTP site of Ensembl Project (5) (ftp://ftp.ensembl.org/) and used in this study.

Oligonucleotide sequences.

Primer sequence for quantitative PCR of mRNA.

Tbp-forward: 5'-GTTGTGCAGAAGTTGGGCTTC-3'

Tbp-reverse: 5'-TCACAGCTCCCCACCATGTT-3'

Per2-forward: 5'-TGTGCGATGATGATTCGTGA-3'

Per2-reverse: 5'-GGTGAAGGTACGTTTGGTTTGC-3'

Cry1-forward: 5'-TGAGGCAAGCAGACTGAATATTG-3'

Cry1-reverse: 5'-CCTCTGTACCGGGAAAGCTG-3'

Arntl-forward: 5'-CCACCTCAGAGCCATTGATACA-3'

Arntl-reverse: 5'-GAGCAGGTTTAGTTCCACTTTGTCT-3'

MGI:1926224-forward: 5'-GACCTGGCGGTGGATGG-3'

MGI:1926224-reverse: 5'-AACACATTTGCGTCCTGCC-3'

Lrrc35-forward: 5'-TGCTGCAGGCCTAATCCTTT-3'

Lrrc35-reverse: 5'-CGGTTGGGTTGGATGAGACT-3'

1300001I01Rik-forward: 5'-AAGATGGTTTTGCACTGGTTCA-3'

1300001I01Rik-reverse: 5'-TTTCGTGTCTTCAATTAGGCCTC-3'

Cpne7-forward: 5'-ATGGAAAGGGTGGTGAAGGG-3'

Cpne7-reverse: 5'-TCTCCACACGATCAAATGGC-3'

Gria2-forward: 5'-CTGAGTGCCTTACACAATGGTTTC-3'

Gria2-reverse: 5'-CGGATGCCTCTCACCACTTT-3'

Etv4-forward: 5'-GCCTCTGCCTAGGTCTTGCTC-3'

Etv4-reverse: 5'-ACACTGGATCTCTGTGGTGGG-3'

Pogz-forward: 5'-TTTATGCCACCACTCCCAGC-3'

Pogz-reverse: 5'-CGGCGTTCCTAATAACCCAC-3'

Plcb1-forward: 5'-AGTGCACGCCTTGCAACTC-3'

Plcb1-reverse: 5'-CTTCTTGAGGCTGTCGGACAC-3'

Irf2bp1-forward: 5'-TCGTGGCTTGCCTTTTCC-3'

Irf2bp1-reverse: 5'-CTTCCCCGCCCCCTG-3'

Trim8-forward: 5'-ACCCTCTTTCTAGCCGGAAGTT-3'

Trim8-reverse: 5'-GGTTTGAAGATGCCAAAGGC-3'

Tspan7-forward: 5'-GTATGGCATCGAGGAGAATGG-3'

Tspan7-reverse: 5'-ATGAGGAGGGTTTTGAGACAGG-3'

Atp1b2-forward: 5'-CTCGAATTTTGGAGCCGTCT-3'

Atp1b2-reverse: 5'-CACACACCGCCTAGAAGCAA-3'

Spsb3-forward: 5'-CAGGGACATCTCTGGTTCATTCA-3'

Spsb3-reverse: 5'-GGCTGAGCGCCGTATAAGAA-3'

Mgea5-forward: 5'-CCTTAATAGCAGATCCGCATGTG-3'

Mgea5-reverse: 5'-CAGTCCCCTTACCCTTACTTAACAAT-3'

Adam12-forward: 5'-CACTGTCCAGCCAATGTGTACC-3'

Adam12-reverse: 5'-AGTAACCATCCACGCCCTGA-3'

Myo1b-forward: 5'-ACGAGTGTTTGTCTCTCTCTCCCT-3'

Myo1b-reverse: 5'-CAGACTTCAGCAGCCCTTTAGC-3'

D13Ertd150e-forward: 5'-CGCTTTTGCAACCAGGTGTT-3'

D13Ertd150e-reverse: 5'-GGTGGGAGCGAACGTGG-3'

The oligonucleotide sequence for competitive binding assay.

Canonical consensus sequences for E-box (CACGTG), D-box (TTATGTAA) and RRE ([A/T]A[A/T]NT[A/G]GGTCA) are indicated in bold while mutated core sequences for each element are indicated in bold and italics.

Per1 E-box-forward: 5'-CAAGTC**CACGTG**CAGGGACAAGTC**CACGTG**CAGGGACAAGTC**CACGTG**CAGGGA-3'

Per1 E-box-reverse: 5'-TCCCTG**CACGTG**GACTTGTCCCTG**CACGTG**GACTTGTCCCTG**CACGTG**GACTTG-3'

Mutated *Per1* E-box-forward: 5'-CAAGTC***ACCGGT***CAGGGACAAGTC***ACCGGT***CAGGGACAAGTC***ACCGGT***CAGGGA-3'

Mutated *Per1* E-box-reverse: 5'-TCCCTG***ACCGGT***GACTTGTCCCTG***ACCGGT***GACTTGTCCCTG***ACCGGT***GACTTG-3'

High-scoring E-box-forward: 5'-CGGGGC**CACGTG**CAGGCGCGGGGC**CACGTG**CAGGCGCGGGGC**CACGTG**CAGGCG-3'

High-scoring E-box-reverse: 5'-CGCCTG**CACGTG**GCCCCGCGCCTG**CACGTG**GCCCCGCGCCTG**CACGTG**GCCCCG-3'

Low-scoring E-box-forward: 5'-GTTAAA**CACGTG**TTTTACGTTAAA**CACGTG**TTTTACGTTAAA**CACGTG**TTTTAC-3'

Low-scoring E-box-reverse: 5'-GTAAAA**CACGTG**TTTAACGTAAAA**CACGTG**TTTAACGTAAAA**CACGTG**TTTAAC-3'

Per3 D-box-forward: 5'-

CCCGCGCG**TTATGTAA**GGTACTCGCCCGCGCG**TTATGTAA**GGTACTCGCCCGCGCG**TTATGTAA**GGTACTCG-3'

Per3 D-box-reverse: 5'-

CGAGTACC**TTACATAA**CGCGCGGGCGAGTACC**TTACATAA**CGCGCGGGCGAGTACC**TTACATAA**CGCGCGGG-3'

Mutated *Per3* D-box-forward: 5'-

CCCGCGCG***CACCCGGC***GGTACTCGCCCGCGCG***CACCCGGC***GGTACTCGCCCGCGCG***CACCCGGC***GGTACTCG-3'

Mutated *Per3* D-box-reverse: 5'-

CGAGTACC***GCCGGGTG***CGCGCGGGCGAGTACC***GCCGGGTG***CGCGCGGGCGAGTACC***GCCGGGTG***CGCGCGGG-3'

High-scoring D-box-forward: 5'-

CCCGCGCG**TTATGTAA**CGAGCCCGCCCGCGCG**TTATGTAA**CGAGCCCGCCCGCGCG**TTATGTAA**CGAGCCCG-3'

High-scoring D-box-reverse: 5'-

CGGGCTCG**TTACATAA**CGCGCGGGCGGGCTCG**TTACATAA**CGCGCGGGCGGGCTCG**TTACATAA**CGCGCGGG-3'

Low-scoring D-box-forward: 5'-

ATGAAAAT**TTATGTAA**GTTTAAACATGAAAAT**TTATGTAA**GTTTAAACATGAAAAT**TTATGTAA**GTTTAAAC-3'

Low-scoring D-box-reverse: 5'-

GTTTAAAC**TTACATAA**ATTTTCATGTTTAAAC**TTACATAA**ATTTTCATGTTTAAAC**TTACATAA**ATTTTCAT-3'

Arntl RRE-forward: 5'-

AGGCAG**AAAGTAGGTCA**GGGACGAGGCAG**AAAGTAGGTCA**GGGACGAGGCAG**AAAGTAGGTCA**GGGACG-3'

Arntl RRE-reverse: 5'-

CGTCCC**TGACCTACTTT**CTGCCTCGTCCC**TGACCTACTTT**CTGCCTCGTCCC**TGACCTACTTT**CTGCCT-3'

Mutated *Arntl* RRE-forward: 5'-

AGGCAG***AAAGTCCTAGC***GGGACGAGGCAG***AAAGTCCTAGC***GGGACGAGGCAG***AAAGTCCTAGC***GGGACG-3'

Mutated *Arntl* RRE-reverse: 5'-

CGTCCC***GCTAGGACTTT***CTGCCTCGTCCC***GCTAGGACTTT***CTGCCTCGTCCC***GCTAGGACTTT***CTGCCT-3'

High-scoring RRE-forward: 5'-

AGAAAG**AAAGTAGGTCA**GTGCGGAGAAAG**AAAGTAGGTCA**GTGCGGAGAAAG**AAAGTAGGTCA**GTGCGG-3'

High-scoring RRE-reverse: 5'-

CCGCAC**TGACCTACTTT**CTTTCTCCGCAC**TGACCTACTTT**CTTTCTCCGCAC**TGACCTACTTT**CTTTCT-3'

Low-scoring RRE-forward: 5'-

CTCCCC**TATTTGGGTCA**ACCGACCTCCCC**TATTTGGGTCA**ACCGACCTCCCC**TATTTGGGTCA**ACCGAC-3'

Low-scoring RRE-reverse: 5'-

GTCGGT**TGACCCAAATA**GGGGAGGTCGGT**TGACCCAAATA**GGGGAGGTCGGT**TGACCCAAATA**GGGGAG-3'

The oligonucleotide sequence for construction of expression plasmids.

Underlines indicate linker sequence, which incorporated the recognition sequence for a restriction enzyme.

Arntl-forward: 5'-ATTACCCTGTTATCCCTAATGCGGACCAGAGAATGGAC -3'

Arntl-reverse: 5'-ACCCATAATACCCATAATAGCTGTTTGCCACTACAGCGGCCATGGCAAGTC-3'

Clock-forward: 5'-ATTACCCTGTTATCCCTAATGTGTTTACCGTAAGCTGTAG -3'

Clock-reverse: 5'-ACCCATAATACCCATAATAGCTGTTTGCCACTGTGGCTGGACCTTGG -3'

Bhlhb2-forward: 5'-ATTACCCTGTTATCCCTAATGAACGGATCCCCAGCGC -3'

Bhlhb2-reverse: 5'-ACCCATAATACCCATAATAGCTGTTTGCCAGTCTTTGGTTTCTAAG -3'

Dbp-forward: 5'-ATTACCCTGTTATCCCTAATGCGCGGCCTCTGAGCGAC -3'

Dbp-reverse: 5'-ACCCATAATACCCATAATAGCTGTTTGCCACAGTGTCCCATGCTGG -3'

Nfil3-forward: 5'-ATTACCCTGTTATCCCTAATCAGCTGAGAAAAATGCAG -3'

Nfil3-reverse: 5'-ACCCATAATACCCATAATAGCTGTTTGCCACCTGGAGTCCGAAGCCG -3'

Nr1d1-forward: 5'-ATTACCCTGTTATCCCTAATACGACCCTGGACTCCAATAAC -3'

Nr1d1-reverse: 5'-ACCCATAATACCCATAATAGCTGTTTGCCACTGGGCGTCCACCCGG -3'

Rora-forward: 5'-ATTACCCTGTTATCCCTAATTATTTGTGATCGCAGCG -3'

Rora-reverse: 5'-ACCCATAATACCCATAATAGCTGTTTGCCACCCATCGATTTGCATGGCTG-3'

Plasmid constructions. The SV40-dLuc (15) and *Per2*-dLuc (16) reporters are described elsewhere. The sequences containing three tandem repeats of putative *cis*-acting elements were inserted into *Mlu*I/*Bgl*II site of the SV40-dLuc vector. Detailed information on putative *cis*-acting elements is available in **Tables S1 and S3**. The DNA sequences of all constructs generated in this study were verified by standard methods. For construction of expression plasmids, we amplified the full length coding sequence of mouse *Arntl*, *Clock*, *Bhlhb2*, *Dbp*, *Nfil3*, *Nr1d1* and *Rora* from pCI-*Arntl* (17), pCI-*Clock* (17), pSPORT6-*Bhlhb2* (MGC clone # 3707474, Invitrogen, Carlsbad, CA), pSPORT6-*Rora* (MGC clone # 3592667, Invitrogen) or NIH3T3 cDNA library by PCR with forward primers containing *I-Sce*I recognition sequence and reverse primers containing *PI-Psp*I recognition sequence (Hokkaido System Science, Hokkaido, Japan). PCR product was digested with *I-Sce*I (NEW ENGLAND BioLabs, Ipswich, MA) and *PI-Psp*I (NEW ENGLAND BioLabs), and cloned into pMU2 vector (18) and termed as pMU2-*Arntl*, pMU2-*Clock*, pMU2-*Bhlhb2*, pMU2-*Dbp*, pMU2-*Nfil3*, pMU2-*Nr1d1* and pMU2-*Rora*. Those genes were fused in-frame with 1 × Flag Tag at N-terminal by *I-Sce*I recognition site, and regulated by T7 promoter in pMU2.

Quantitative PCR. Quantitative PCR was performed with the ABI PRISM 7900HT and SYBR Green Reagents (Applied Biosystems, Foster City, CA). cDNAs were synthesized from 0.25 μg of total RNA using SuperScript II Reverse Transcriptase (Invitrogen) and Random Primers (Promega). Samples contained 1 × SYBR Green PCR Master mix, 0.8 μM primers and 1/50 synthesized cDNA in a 10 μl volume. The PCR conditions were as follows: 10min at 95°C, then 45 cycles of 15 sec at 95°C and 1min at 59°C. Absolute cDNA abundance was calculated using the standard curve obtained from mouse genomic DNAs. *Tbp* expression levels were quantified and used as the internal control. Detailed data of quantitative PCR is available at the circadian section of the mammalian promoter/enhancer database.

Rhythmicity analysis of real-time bioluminescence data. Bioluminescence time-series data beginning 21 h after forskolin stimulation were used for analysis to distinguish endogenous circadian oscillations from acute effects of stimulation. Bioluminescence data were detrended by using the trend curve calculated by the smoothing spline method, and statistical significance and the period of circadian oscillation in the detrended data was evaluated as previously described ($p < 0.01$) (16). To visualize the normalized bioluminescence data (i.e. the oscillatory component of the bioluminescence data) shown in **Fig. 2*A***, the moving average of the absolute value of the detrended bioluminescence data was calculated first. The window size of the moving average was set to half of the period calculated above. Then, the oscillatory component of the detrended data was calculated by dividing the data by the moving average of the data at each time point.

Amplitude analysis of real-time bioluminescence data. Bioluminescence time-series data of 21–96 h after forskolin stimulation were normalized so that the average bioluminescence is 1.0, and then detrended as described above. Detrended and normalized time-series data were used in further analysis. To determine the amplitude of each sample, 151 "reference" time-series data were generated for E-box, D-box and RRE, respectively, by multiplying time-series data of positive controls (*Per1* E-box, *Per3* D-box and *Arntl* RRE) (11) with the value of "relative amplitude" from 0 to 1.5 with 0.01 step. Time-series data for each sample were compared with reference data using the least-squares method to determine the best fit for reference data, the "relative amplitude" of which was used as the control amplitude and used for comparisons.

Rhythmicity analysis of quantitative PCR data. Two statistical tests, cosine fitting and analysis of variance (ANOVA), were combined to identify circadian expression profiles with high-amplitude. To evaluate the wave form of expression profiles, statistical analysis (cosine fitting test) was performed in parallel on two independent expression profiles (two days each). 2-cycle cosine waves of 24-h period with a different phase were generated by shifting phase with 0.4-h interval (total 60 waves). A two day expression profile was compared with 60 cosine test waves by calculating the correlation coefficient to determine the most correlated cosine wave, the correlation coefficient of which was used as "maximum correlation coefficient" for the expression profile. Statistical significance of the maximum correlation coefficient for the expression profile was evaluated by calculating the maximum correlation coefficient for 100,000 random expression profiles. Two *P*-values calculated on two independent expression profiles (two days each) were combined by Fisher's probability combination method (19) to calculate the synthetic *P*-value for cosine fitting test. To evaluate the amplitude of expression profile, statistical analysis (one-way ANOVA test) was also performed on experimental duplicates. *P*-values for cosine fitting and ANOVA tests were combined by Fisher's probability combination method to calculate the synthetic *P*-value, which was used to identify circadian expression profiles with high-amplitude ($p < 0.03$). The peak time of an expression profile was estimated by the peak time of the best correlated cosine wave.

Over representation analysis of clock-controlled genes. To determine the significance of the enrichment of clock-controlled genes within the genes having predicted clock-controlled elements, we first selected the 100 most significant predicted sequences for each clock-controlled elements (E-box, D-box and RRE) that were mapped to 98 genes on U74 microarray after removing the 21 clock-controlled genes used for HMM generation and training. As the dataset of clock-controlled genes, we used previously identified clock-controlled genes in the SCN and liver by using mouse U74 microarray (15, 20). After removing the 21 clock-controlled genes in the above, we found additional 19 putative clock-controlled out of the 6,195 genes common in our mammalian promoter/enhancer database and U74 mouse microarray, which is significantly higher than the expected 10.67 genes that would have arisen from chance ($p = 0.009$).

Estimation of the number of high-amplitude E-boxes. To estimate the number of clock-controlled conserved E-boxes that likely confer circadian rhythmicity, we used 1,108 E-boxes (minimum HMM score = 11.56; available on http://promoter.cdb.riken.jp/circadian.html) except E-boxes used for HMM generation and training. Linear interpolation from the true positive rate (40%) of high-scoring conserved E-boxes (mean HMM score = 16.15) and the false negative rate (7.1%) of low- scoring conserved E-boxes (mean HMM score = 2.5143) was used to estimate the probability of rhythmicity for each E-box. The sum of the probability of 1,108 E-boxes was calculated to estimate the number of high-amplitude E-boxes.

Microarray expression data analysis of genes with predicted clock-controlled elements. To investigate the averaged expression of genes with predicted clock-controlled elements, we used time-series microarray expression data of mouse liver in our previous study (15). Averaged expression value under LD and DD conditions were used for this analysis. We selected 100 most significant sequences for each clock-controlled elements (E-box, D-box and RRE). 21 genes, which were used for training data set of HMMs, were excluded in the selection. We then related them to microarray probe set data (36 genes for E-box, 29 genes for D-box and 34 genes for RRE respectively). In case that more than one probe sets exists for a single gene, averaged expression data was used. Then, averaged expression value for each time points were determined for each clock-controlled element. Data were then normalized so that the average expression value over 2-day 12-point time courses is 1.0. To statistical tests, cosine fitting was performed. 2-cycle cosine waves of 24-h period with a different phase were generated by shifting phase with 0.4-h interval (total 60 waves). A 2-day expression profile was compared with 60 cosine test waves by calculating the correlation coefficient to determine the most correlated cosine wave, the correlation coefficient of which was used as "maximum correlation coefficient" for the expression profile. Statistical significance of the maximum correlation coefficient for the expression profile was evaluated by calculating the maximum correlation coefficient for 100,000 random expression profiles. *P*-values for cosine fitting were used to evaluate circadian oscillation of averaged expression profiles. The peak times of the averaged expression profile were estimated by the peak time of the best correlated cosine wave.

Sequence logos. Sequence logos shown in **Fig. S4** and **Table S2** were created by WebLogo (http://weblogo.berkeley.edu/) (21).

Calculation of relative affinity from competitive DNA binding assay data. A relative binding affinity between a regulator (R) and an unlabeled competitive DNA element (D_u) in comparison with that to the labeled control DNA element (D_l) was determined by competitive DNA binding assays. In each competitive DNA binding assay, the concentration of bound regulator to the labeled element ($[RD_l]$) can be described as follows;

$$[RD_l] = ([R]_{all} - [R]) \frac{[D_l]}{[D_l] + \frac{A_u}{A_l}[D_u]} \qquad \text{(Eq. 1)}$$

, where $[R]_{all} \equiv [RD_l] + [RD_u] + [R]$ is a total concentration of a regulator,

$[RD_l]$ is a concentration of the regulator-labeled element complex,

$[RD_u]$ is a concentration of the regulator-unlabeled element complex,

$[R]$ is a concentration of a free regulator,

$[D_l]$ is a concentration of a free labeled element,

$[D_u]$ is a concentration of a free unlabeled element,

$A_l \equiv \frac{[RD_l]}{[R][D_l]}$ is an affinity constant between a regulator and a labeled element, and

$A_u \equiv \frac{[RD_u]}{[R][D_u]}$ is an affinity constant between a regulator and an unlabeled element.

Since an excessive amount of a labeled or unlabeled element is usually applied in the competitive DNA binding assay, the concentration of a labeled ($[D_l]$) or unlabeled ($[D_u]$) element is much greater than the concentration of bound regulator to the labeled ($[RD_l]$) or unlabeled ($[RD_u]$) element. Thus, the total concentration of the labeled ($[D_l]_{all} \equiv [D_l] + [RD_l]$) or unlabeled ($[D_u]_{all} \equiv [D_u] + [RD_u]$) element can be approximated as follows;

$$[D_l]_{all} \cong [D_l]$$

$$[D_u]_{all} \cong [D_u]$$

Since an amount of a labeled or unlabeled element vastly exceeds the amount of regulator, the amount of a free regulator ($[R]$) is much less than the total amount of a regulator ($[R]_{all}$) in the competitive DNA binding assay. Thus, $[R]_{all} - [R]$ can be also approximated as follows;

$$[R]_{all} - [R] \cong [R]_{all}$$

In addition, since $[RD_l]$ is proportional to the measured value of competitive DNA binding assay (M_{450}), Equation 1 can be rewritten as follows;

$$M_{450} \propto [RD_l] \cong [R]_{all} \frac{[D_l]_{all}}{[D_l]_{all} + \frac{A_u}{A_l}[D_u]_{all}}$$

Furthermore, as $[R]_{all}$ is constant in the assay, we can derive the following equation;

$$M_{450} = C\frac{[D_l]_{all}}{[D_l]_{all} + A*[D_u]_{all}} \qquad \text{(Eq. 2)}$$

, where C is a proportional constant, and $A \equiv \frac{A_u}{A_l}$ is an relative affinity constant of the unlabeled element in comparison with that of the labeled element.

To determine the relative affinity constant from the measured value of competitive DNA binding assay, we first determined the value of $[D_l]_{all}$ and C from series of data for unlabeled known clock-controlled element (positive control) where A was defined as 1.0. In details, by changing the value of $[D_l]_{all}$ and C, the most fitting $[D_l]_{all}$ and C values were determined using the least-square method applied to series of measured values and model data calculated from Equation 2. We then determined the values of A for unlabeled elements including the "high-scoring", "low-scoring" and "negative control" by changing the value of A.

***In silico* analysis of affinity to amplitude mechanism.** By modifying the previous described formula (11), we formulated transcriptional activity $T(t)$ at time t regulated by competition between a clock-controlled activator and a repressor as follows:

$$T(t) \equiv \frac{(\frac{A(t)}{K_a})^n}{1 + (\frac{A(t)}{K_a})^n + (\frac{R(t)}{K_b})^n} + \alpha$$

where $1/K_a$ and $1/K_b$ represent the affinity of an activator and a repressor, respectively. α represents transcriptional activity that does not depend on the circadian clock. n represents the Hill coefficient at competitive regulation. $A(t)$ and $R(t)$ represent expression of a clock-controlled activator and repressor, which are defined as follows:

$$A(t) \equiv \beta_a (1 + Cos(2\pi \frac{t-a}{24})) + \gamma_a$$

and

$$R(t) \equiv \beta_b (1 + Cos(2\pi \frac{t-b}{24})) + \gamma_b$$

where β_a and β_b represent half amplitude of expression of an activator and a repressor. γ_a and γ_b represent expression of an activator and a repressor that does not depend on the circadian clock. $a(-12 \le a \le 12)$ and $b(-12 \le b \le 12)$ represent phases of expression of an activator and a repressor. Then, we formulated output $P(t)$ at time t which depended on transcriptional activity $T(t)$ as follows:

$$\frac{d}{dt} P(t) = T(t) - \lambda P(t)$$

where λ represent decay constant and defined as follows:

$\lambda \equiv \log 2 / T_{1/2}$

where $T_{1/2}$ represent half life. For simplicity, we used $a = 0$, $b = 12$, $\alpha = 0.2$, $\beta_a = \beta_b = 1$, $\gamma_a = \gamma_b = 0.2$, $T_{1/2} = 3$, $P(0) = 0$ and $48 \le t \le 144$ in the analysis. We used $n = 1$ and $K_b = 1$ for the analysis of activator affinity to amplitude mechanism (**Fig. 4 *B* and *C***). Output time-series data were normalized so that the center value of maximum and minimum value of time-series data is 100%. The difference between maximum and minimum values of normalized output time-series data was used to define amplitude.

SI Discussion

Utility of the Promoter/Enhancer Database. As a tool for understanding systems-level transcriptional regulation in mammals, we constructed the mammalian promoter/enhancer database (http://promoter.cdb.riken.jp) by integrating information of conserved non-coding regions, TSSs, and TFBSs. Users can input a gene name or symbol, or UniGene or RefSeq identifiers, and get back a page that summarizes promoter information with additional links to outside databases such as SymAtlas and NCBI. The promoter sequences are available in a default view 1000 bp 5' of the TSS; this default view can be changed arbitrarily by the user 5' or 3' of the TSS, and the DNA sequence information can be downloaded in FASTA format. Users can also highlight conserved sequence regions between humans, mice, and rats, as well as TSSs, exons and TFBSs. Where known, links to alternative promoters are available. Although we applied this database to the understanding of circadian transcriptional regulation, it is generically useful and can be applied to any aspect of mammalian transcriptional regulation.

In silico* modeling of affinity amplitude mechanism.** From competitive DNA binding assays, the "high-scoring" D-box and RRE show approximately the same affinity as positive control whereas the "low-scoring" D-box and RRE show relatively weak affinity (**Fig. 4*A and **Fig. S5*A***). These results can be reasonably explained by using *in silico* model that has been extended from our previous model for gene expression of transcriptional activators and repressors to generate high-amplitude circadian output (11) by introducing expression dynamics of an activator and a repressor, affinity of a DNA element to an activator and a repressor, and half life of output (see also ***SI Materials and Methods***). This *in silico* analysis shows that, if a DNA element has a weak affinity to both an activator and a repressor, then the transcription system exhibits low-amplitude of output oscillations, which can be intuitively elucidated (**Fig. S5*B***). In this *in silico* analysis also shows that the amplitude of output oscillations depend not only on the strength of affinity but also on a hill coefficient (i.e. a parameter for nonlinearity in transcriptional response) (**Fig. S5*B***).

On the other hand, the "high-scoring" E-box shows approximately the same affinity as positive control, whereas the "low-scoring" E-box shows, surprisingly, 4.8-times higher affinity only to *Arntl/Clock* activator (**Fig. 4*A* and Fig. S5*A***). Since "low-scoring" E-box shows approximately the same affinity to *Bhlhb2* repressor as positive control, this result suggests that the "low-scoring" E-box has an unbalanced affinity stronger for *Arntl/Clock* activator than *Bhlhb2* repressor. In order to interpret this seemingly complicated result, we also performed *in silico* analysis of affinity-amplitude relationship especially in the case that a DNA element has an unbalanced affinity between an activator and a repressor. This *in silico* analysis shows that, if a DNA element has a 5-times higher affinity to an activator than a repressor, then the transcription system exhibits less than half amplitude of output oscillations (**Fig. 4*B***). Further *in silico* analysis also shows that, if a DNA element has a lower affinity to an activator than a repressor, then the transcription system exhibits the reduced amplitude of output oscillations (**Fig. 4*C***). Collectively, these results suggest that not only the strength of affinity to regulators but also the balance of affinity between an activator and a repressor are important in generating high-amplitude outputs. In this study, we supposed the same level of expression for a clock-controlled activator and repressor for simplicity in the modeling, and drew the conclusion that appropriate affinity balance between activators and repressors is important. We can easily generalize it into different levels of expression between activator and repressor. In such a case, the conclusion also holds for the product of affinity and concentration instead (i.e. appropriate balance in the product of concentration and affinity is important).

SI References

1. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35:D61-65.

2. Wheeler DL*, et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 36:D13-D21.

3. Kimura K*, et al.* (2006) Diversification of transcriptional modulation: Large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res* 16:55-65.

4. Kent WJ (2002) BLAT--the BLAST-like alignment tool. *Genome Res* 12:656-664.

5. Flicek P*, et al.* (2008) Ensembl 2008. *Nucleic Acids Res* 36:D707-D714.

6. Wilming LG*, et al.* (2008) The vertebrate genome annotation (Vega) database. *Nucleic Acids Res* 36:D753-D760.

7. Brudno M*, et al.* (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* 13:721-731.

8. Matys V, *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 34:D108-110.

9. Karolchik D, *et al.* (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* 36.

10. Okazaki Y, *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420:563-573.

11. Ueda HR, *et al.* (2005) System-level identification of transcriptional circuits underlying mammalian circadian clocks. *Nat Genet* 37:187-192.

12. Marchler-Bauer A, *et al.* (2007) CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res* 35:D237-240.

13. Mayor C, *et al.* (2000) VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* 16:1046-1047.

14. Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L (2001) The distributed annotation system. *BMC Bioinformatics* 2:7.

15. Ueda HR, *et al.* (2002) A transcription factor response element for gene expression during circadian night. *Nature* 418:534-539.

16. Sato TK, *et al.* (2006) Feedback repression is required for mammalian circadian clock function. *Nat Genet* 38:312-319.

17. Hida A, *et al.* (2000) The human and mouse Period1 genes: five well-conserved E-boxes additively contribute to the enhancement of mPer1 transcription. *Genomics* 65:224-233.

18. Ukai H, *et al.* (2007) Melanopsin-dependent photo-perturbation reveals desynchronization underlying the singularity of mammalian circadian clocks. *Nat Cell Biol* 9:1327-1334.

19. Fisher RA (1970) Statistical methods for research workers.

20. Panda S, *et al.* (2002) Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell* 109:307-320.

21. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Res* 14:1188-1190.