Supporting Information

Wang et al. 10.1073/pnas.1715225115

SI Materials and Methods

Data Processing for Total RNA-Seq in Mouse Liver Around the Clock. To quantify the temporal accumulation of pre-mRNA and mRNA in mouse liver, we used total RNA-seq data from ref. 8 (GEO accession no. GSE73554), in which wild-type C57BL/6J male mice between 10 and 14 wk of age, under 12-h light/12-h dark and ad libitum feeding conditions were killed every 2 h during 1 d (four biological replicates = 48 time points in total). To assign uniquely mapped paired-end read to exons or introns, we defined exonic regions if these were covered by at least one of mRNA transcript (University of California, Santa Cruz, mm10 annotation), and defined intronic regions if these were not covered by all annotated transcripts. The obtained read counts in exonic and intronic regions are then proportional to the concentrations of mRNAs and pre-mRNAs and the length of the respective features. Further details are in ref. 8.

Modeling the Temporal Profiles of mRNA and Pre-mRNA. Here, we modeled the temporal accumulations of mRNA with the following differential equation:

$$\frac{dm(t)}{dt} = kp(t) - \gamma(t)m(t),$$

in which $m(t)$ and $p(t)$ denote the temporal accumulations (concentrations) of mRNA (m) and pre-mRNA (p), respectively. $\gamma(t)$ describes temporal variation of mRNA degradation rate. The parameter k represents the pre-mRNA processing rate (defined here as the effective rate for processing of pre-mRNA into mRNA, combining several intermediate steps such as splicing, pre-mRNA decay, and nuclear export) from pre-mRNA to mRNA, which is assumed to be fast on the scale of $\gamma(t)$, and can thus be approximated as a gene-specific constant. Furthermore, to distinguish different regulations, we considered the two cases that $p(t) = p_0$ if the pre-mRNA accumulation is constant, and $p(t) = p_{\min} + A_p((1+\cos(\omega t - \varphi_p))/2)^{\beta}$ if it is rhythmic. $p_{\min}$ is the minimum, A_p is the absolute amplitude (different between the maximum and minimum), $\omega = 2\pi/24h^{-1}$ is the angular frequency, and φ_p is the phase of rhythmic pre-mRNA. Depending on the coefficient β, this parametrization allows to capture profiles that are more peaked than standard cosine functions, since temporal profiles of some core clock genes are sharper than simple cosine function (6). Similarly, $\gamma(t) = \gamma_0$ and $\gamma(t) = \gamma_0(1 + \varepsilon_\gamma \cos(\omega t - \varphi_\gamma))$ represent constant and rhythmic mRNA degradation, respectively, in which γ_0 is the mean rate of mRNA degradation, ε_γ is the relative amplitude and φ_γ is the phase of degradation rate. While we could have used more complicated parameterizations, the ones described here provided a good compromise between their number of parameters and ability to accurately describe pre-mRNA and mRNA profiles.

Model Selection and Parameter Estimation. According to the above model, the temporal accumulation of mRNA for each gene results from the balance between constant or rhythmic synthesis and degradation rates. We thus fitted the measured pre-mRNA and mRNA profiles with four models (M1–M4), representing the different combinations of constant (C) or rhythmic (R) synthesis (S) and degradation (D) with the following notation: CS-CD (M1), RS-CD (M2), CS-RD (M3), and RS-RD (M4). In particular, because temporal profiles of pre-mRNA and mRNA were both quantified from the same total–RNA-seq data, the absolute levels were used in the model fitting. Then the optimal model is selected by combining a maximum-likelihood approach with the Bayesian information criterion (BIC), and the estimated kinetic parameters are taken from the optimal model. Details of the model selection and parameter estimation are discussed in the following paragraphs.

Fitting Temporal mRNA and Pre-mRNA Measurements to the Kinetic Model. To take into account the noise structure of RNA-seq count data, we assumed negative binomial distributions RNA-seq read counts (55, 56). This allowed us to formulate a log-likelihood function of the measured counts to follow modeled mRNA [$m(t)$] and pre-mRNA [$p(t)$] levels as follows:

$$\log L = \sum_t \log NB\left(n_m(t)|\mu_m(t), \alpha_m(t)\right) + \log NB\left(n_p(t)|\mu_p(t), \alpha_p(t)\right),$$

with

$$\mu_m(t) = m(t)S_m(t)L_m,$$

$$\mu_p(t) = p(t)S_p(t)L_p.$$

Here, $n_m(t)$ [or $n_p(t)$] denotes the read count for mRNA (or pre-mRNA) at time t. $\alpha_m(t)$ [or $\alpha_p(t)$] is the time-dependent and gene-specific dispersion parameter for mRNA (or pre-mRNA) in the negative binomial distribution. Before maximizing the log-likelihood, these dispersion parameters were estimated with DESeq2 (55) using four biological replicates at each time point. $\mu_m(t)$ [or $\mu_p(t)$] is the expected count, given by the product of the concentration of mRNA $m(t)$ [or $p(t)$ for pre-mRNA], the exon length L_m (or L_p for intron length), and the scaling factor of each sample (which depends on the sampling depth of each library) $S_m(t)$ [or $S_p(t)$], which was also estimated with DESeq2. Moreover, the concentrations of mRNA and pre-mRNA, $m(t)$ and $p(t)$, are numerically calculated (using numerical integration) from the kinetic model in the previous section for given kinetic parameters. Some of the parameters needed to be constrained, for example, the relative amplitude, can only take values between 0 and 1. Since we experience that relative amplitude near 1 renders the optimization very sensitive, the constraint at 1 has been implemented using a sigmoid function, which penalizes relative amplitude beyond 0.8 in our case. The log-likelihoods for each model, M1, M2, M3, and M4, were maximized using the function "optim" with L-BFGS-B method in R. In addition, to avoid local minima, 10 initial parameter values for M2 and 12 for M3 and M4 were sampled to start the optimization, and the solution with the best log-likelihood was selected. After the parameters were estimated, the corresponding Hessian matrix was calculated to obtain SEs for the estimated parameters.

Model Selection with BIC. To select the optimal model for each gene, given the measured read counts of mRNA and pre-mRNA, the BIC was used: $\mathrm{BIC} = -\log(L) + K\log(N)$ in which $\log(L)$ is the log-likelihood, K is the number of parameters, and N is the number of data points. Thereafter, the probability of each model can be approximated using Schwarz weight: $w_j = e^{1/2\ \Delta BIC_j} / \sum e^{1/2\ \Delta BIC_j}$, where $\Delta\mathrm{BIC}_j = \mathrm{BIC}_j - \mathrm{BIC}_{\min}$ with $\mathrm{BIC}_{\min}$ the minimum value. In our analysis, only genes with weight $w_j > 0.5$ for the optimal model j were used.

Outlier Detection. Because the data at the different time points were collected from individual mice, variations across biological replicates could be quite large. Consequently, some data points appear as outliers when we fit the data to the model. These outliers may contribute dominantly to the BIC score and weaken the penalty term. To alleviate the effects from outliers, we followed an established method (57) to be appropriate for significant sample sizes (we used 48 samples for both mRNA and pre-mRNA), which allowed us to identify outliers iteratively. Specifically, for each iteration, after fitting the data to models M2–M4, we examined the contribution of –log-likelihood of each time point and identified outliers if these were out of range [Q1 − 1.5 × IQR, Q3 + 1.5 × IQR], where Q1 and Q3 are the first and third quartiles, respectively, and IQR is the interquartile range. Fixing the maximum number of outliers to be six (out of 48 measurements), we removed shared outliers across models M2–M4 and fitted again the resulting reduced data to all models until shared outliers were no longer found.

Identifiability Analysis for Kinetic Parameters. Depending on the data for each gene, some kinetic parameters, notably k, γ_0, ε_γ, and φ_γ in models M3 and M4, were practically nonidentifiable (37), because the estimation of parameters for rhythmic degradation relied on the phase shift and amplitude ratio between mRNA and pre-mRNA and also the shape of temporal profiles. In particular, the mRNA degradation parameters solely depend on the mRNA profiles in M3 (because the pre-mRNA was flat) and thus easily became practically nonidentifiable. It is also clear that k is underdetermined if γ_0 is, because only the ratio of both is constrained by the data. To mitigate the parameter non-identifiability in the fitting, we first chose combinations of parameters that can be identified easily from the data: $a = k/\gamma_0$, $\varepsilon_\gamma' = \varepsilon_\gamma \gamma_0 / \sqrt{\gamma_0^2 + \omega^2}$, $\varphi_\gamma' = \varphi_\gamma + atan(\omega/\gamma_0)$ with $\omega = 2\pi/24h^{-1}$. The good identifiability of those combinations can be understood from the approximations described in ref. 9. Second, we implemented the profile likelihood (PL) (37), which allowed us to assess, gene by gene, if the mean degradation rate γ_0 is identifiable. Specifically, we followed ref. 37: the parameter γ_0 was sampled in its whole range of values and the likelihood is remaximized for each value of γ_0. γ_0 is structurally non-identifiable if the likelihood does not change; and γ_0 is practically nonidentifiable if the log-likelihood varies within certain threshold; otherwise, γ_0 is well estimated. Here, we found γ_0 was either practically nonidentifiable or well estimated. In addition, the SEs of parameter combinations a, ε_γ', and φ_γ' were obtained using the Hessian matrix; then the associated SEs for the parameters k, ε_γ, and φ_γ were computed using error propagations. Due to the dependence of φ_γ on γ_0, SEs for φ_γ became very large when the degradation rate γ_0 was nonidentifiable (in Dataset S1, the SEs were cut at a maximum value of 12 h).

For the analyses of degradation parameters (Figs. 3–5), we focused on genes with high-confidence estimates. We filtered for coefficient of variation (CV) of estimated half-life <0.4 for Fig. 3; CV of relative amplitude of degradation rate <0.4 and SE of estimated degradation phase <1 h for Fig. 4. In addition to such filters, only genes with sufficiently rhythmic mRNAs and pre-mRNAs (FDR < 0.05 rhythmicity test, and relative amplitude > 0.1) were considered in Fig. 5.

Validation of the Method with Simulations. To validate our model selection and parameter estimation, we tested the method with simulated data for models (M1–M4), taking kinetic parameters from realistic distributions. For M1 to M4, read counts of mRNA and pre-mRNA for 200 synthetic genes were simulated. To mimic the real data, the same number of samples with the same time resolution was simulated. The read counts of pre-mRNA and mRNA were sampled from negative binomial (NB) distribution using the dispersion $\alpha_m(t)$ and $\alpha_p(t)$, exon lengths L_m, and intron lengths L_p, p_0, or p_{min} of randomly chosen genes in real data as well as sample scaling factors $S_m(t)$ and $S_p(t)$. The absolute amplitudes A_p and the sharpness parameter β for pre-mRNAs were sampled from the distributions obtained from rhythmic pre-mRNAs in the real data. In particular, the relative amplitudes of mRNA degradation (ε_γ) were sampled from the normal distribution $N(0.2, 0.2)$ with the boundary of [0.05, 0.8]; and the mRNA half-life $[\log(2)/\gamma_0]$ is sampled from a log-normal distribution $\log N(3, 1)$ that resembles the distribution of mRNA half-lives measured in mouse NIH 3T3 cells (38). Comparisons between the identified and true models, as well as the comparison between the estimated and true values of parameters, are shown in Fig. S2.

Inference of mRNA Binding Proteins Involved in Rhythmic mRNA Degradation Using a Linear Model with Elastic-Net Regularization. To infer mRNA binding proteins (mRBPs) with rhythmic activities from the identified rhythmic mRNA degradations, we used genes from group 1 (569 mRNAs rhythmic in both AL and RF WT and RF *Bmal1*$^{-/-}$) and group 2 (292 mRNAs rhythmic in AL and RF WT), and mRBP motif library in ref. 47. 3′-UTR of mRNAs (RefSeq) for those genes were scanned with FIMO (58) to find hits to mRBP motifs, which are potentially responsible for rhythmic mRNA degradation. The rhythmic variation of mRNA degradation (in log-scale) was assumed as a linear combination of diurnal activities of mRBP motifs [analogous to linear models for transcription factor activities (35, 48)]:

$$\log 2\left(\gamma_g(t)\right) = N_{gm} A_m(t) + \text{noise},$$

where $\gamma_g(t)$ is the rhythmic degradation rate for gene g at time t; N_{gm} is the number of occurrence for mRBP motif m, and $A_m(t)$ represents the temporal activity of motif m at time t. Here, we performed the linear regression in the subspace of 24-h periodic functions, which uses directly our estimated parameters γ_0, ε_γ, and φ_γ. To control for overfitting and also redundancy of motifs, we employed elastic-net penalty [implemented in R package glmnet (59)]. Splicing factors were excluded from the motif occurrence matrix N_{gm}. In addition, the inference was done separately for genes in group 1 and 2 because of their distinct phase distribution of rhythmic degradation; and the glmnet mixing parameters $\alpha = 0.15$ was chosen for both groups.

Rhythmicity Assessment of Transcripts in WT and *Bmal1*$^{-/-}$ Mice. To decipher the role of systemic cues and the role of the circadian clock in diurnal mRNA degradation, RNA-seq of WT and *Bmal1*$^{-/-}$ mice under night-restricted feeding regiment was retrieved from ref. 8. Rhythmicity in mRNA abundances for transcripts classified in M3 (Dataset S1) in different conditions was assessed using a model selection approach as described in ref. 8, in which rhythmicity parameters (amplitudes and phases) could be specific to an experiments or shared between two or three conditions. Briefly, profiles from the three datasets (WT ad libitum, WT restricted feeding, and *Bmal1*$^{-/-}$ restricted feeding) were fitted with harmonic regression and BIC was used as a criterion for model selection.

Gene Ontology Analysis. Gene ontology (GO) analysis was performed using the TopGO R package (60). Enrichment analysis for GO terms derived from "Biological Process" ontology was done in the different models and significance assessed using the Fisher exact test. GO terms with P value < 0.05, a minimum number of three genes, and less than 500 annotated genes were considered. Genes defined as rhythmic (from all models) were used as background. All GO terms derived from the GO analysis can be found in Dataset S4.

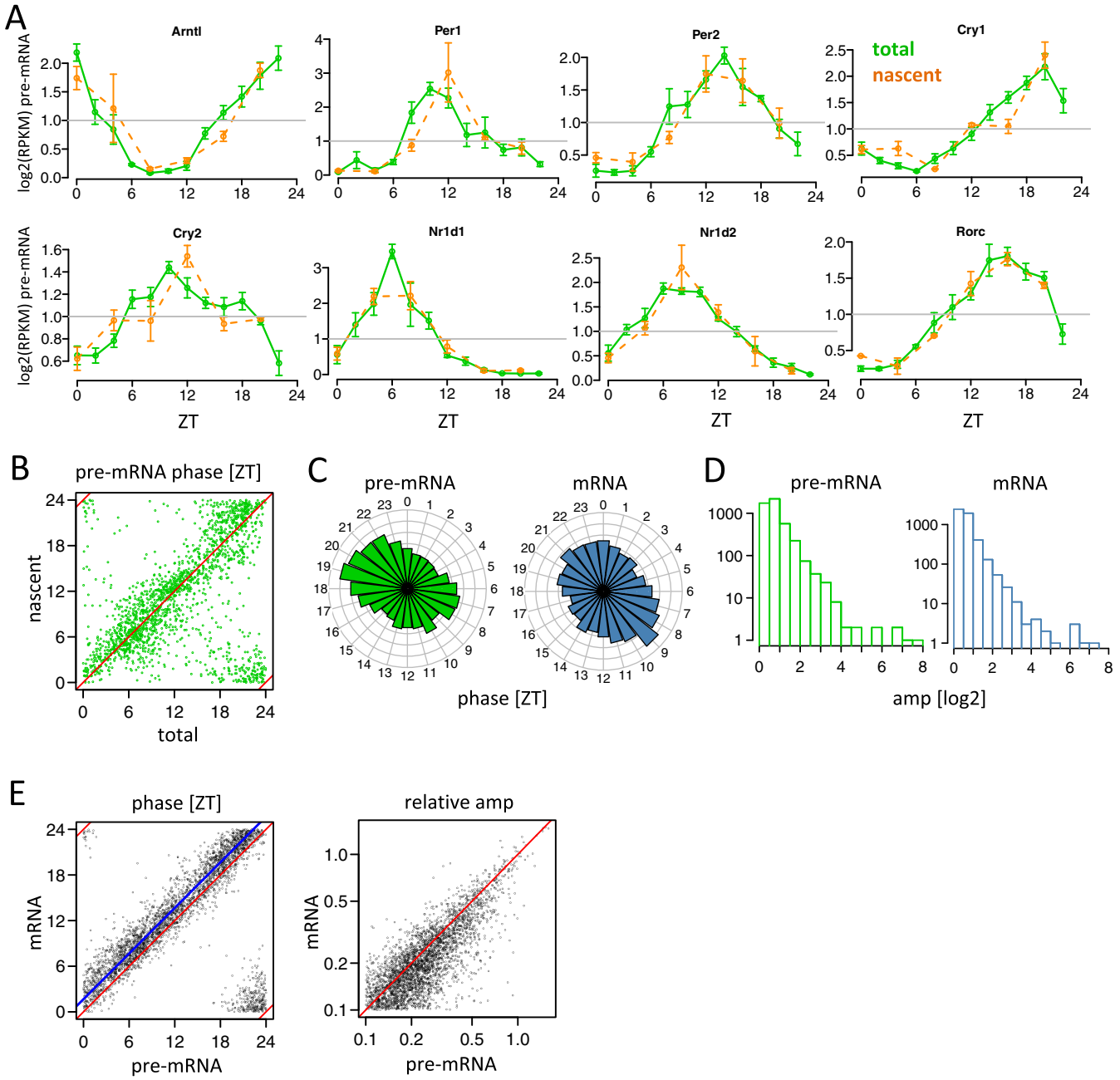

Fig. S1. Rhythmic temporal accumulations of pre-mRNA and mRNA quantified by total RNA-seq in mouse liver. (*A*) Temporal accumulations pre-mRNA for core-clock genes quantified by total RNA-seq compared with nascent RNA-seq. (*B*) Scatterplot of phases of rhythmic pre-mRNAs from total RNA-seq vs. those from nascent RNA-seq. Circular correlation: pval < 10e-4. (*C* and *D*) Distribution of phase (*C*) and peak-to-trough amplitude (*D*) for rhythmic pre-mRNAs (*Left*) and mRNAs (*Right*) (FDR < 0.05 harmonic regression). (*E*) Scatterplot of peak times (*Left*) and relative amplitude (*Right*), defined as the difference between the maximum and minimum levels divided by twice the average expression, between pre-mRNA and mRNA accumulations. Red lines (*Left*) show circular regression predicting a phase delay of 1.7 h.

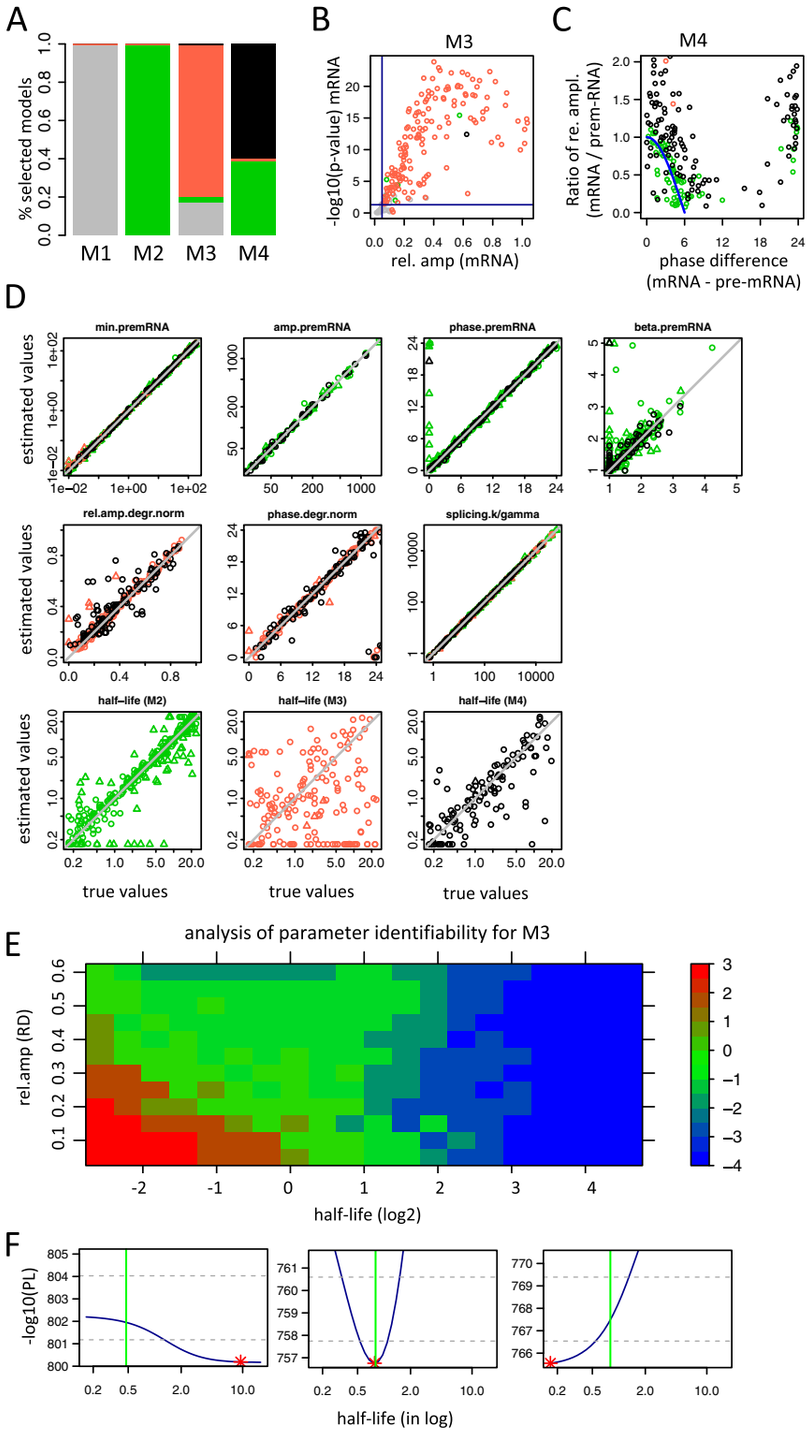

Fig. S2. Validation of model selection and parameter estimation by simulated data and analysis of parameter identifiability. (*A*) Fractions of simulated genes generated from known models (*x* axis) identified as M1 (gray), M2 (green), M3 (red), or M4 (black) by the model selection. (*B*) Rhythmicity (*P* values, harmonic regression) and relative amplitudes of mRNA for simulated genes generated from M3. Colors represent the model chosen by the model selection algorithm (M1, gray; M2, green; M3, red; M4, black). The blue horizontal and vertical lines indicate thresholds of *P* value of 0.05 and relative amplitude of 0.05. Most incorrectly classified mRNAs had low relative amplitudes and blurred rhythm. (*C*) Ratios of relative amplitude ($\varepsilon_m/\varepsilon_s$) and phase delays ($\varphi_m - \varphi_s$) between mRNA (*m*) and pre-mRNA (*s*) for simulated genes generated from M4 (same color coding as *B*). The theoretical relationship for genes generated from M2 was indicated by blue curve. For incorrectly classified mRNAs, the rhythmic degradation did not affect mRNA amplitudes or phases. For these mRNA, which are very close to the blue curve, the model selection between M2 and M4 relies solely on the shape of temporal profiles of pre-mRNA and mRNA and does not allow for a reliable discrimination between them. (*D*) Comparisons between true values of parameters and parameter combinations (e.g., rel.amp.degr.norm, phase.degr.norm, *Materials and Methods*) for the simulated data and estimated by our method. (*E*) For each pair of half-life and relative amplitude of rhythmic degradation displayed as a cell on the grid, 50 realizations of pre-mRNA and mRNA temporal read counts in M3 were simulated by sampling the negative-binomial distributed noise ($\alpha = 10^{-4}$) and keeping other parameters constant. The comparison (mean of $\log_2$ratio) between the estimated and true half-life is shown by the color scale. High values of the mean $\log_2$ratio (red) indicate an overestimation of half-lives, which happens for short half-lives and small relative amplitude of rhythmic degradation, while low values (blue) indicate an underestimation of half-lives. (*F*) Profile likelihood (PL) as a function of half-life sampled in its parameter range for three scenarios: half-lives nonidentifiable and overestimated (*Left*, associated to the red area in *E*), identifiable (*Middle*, associated to the green area in *E*), and nonidentifiable and underestimated (*Right*, associated to the blue area in *E*). The red asterisk indicates the estimated values of half-life, and the green lines, the true values. Two gray dashed horizontal lines represent pointwise and simultaneous confident intervals.

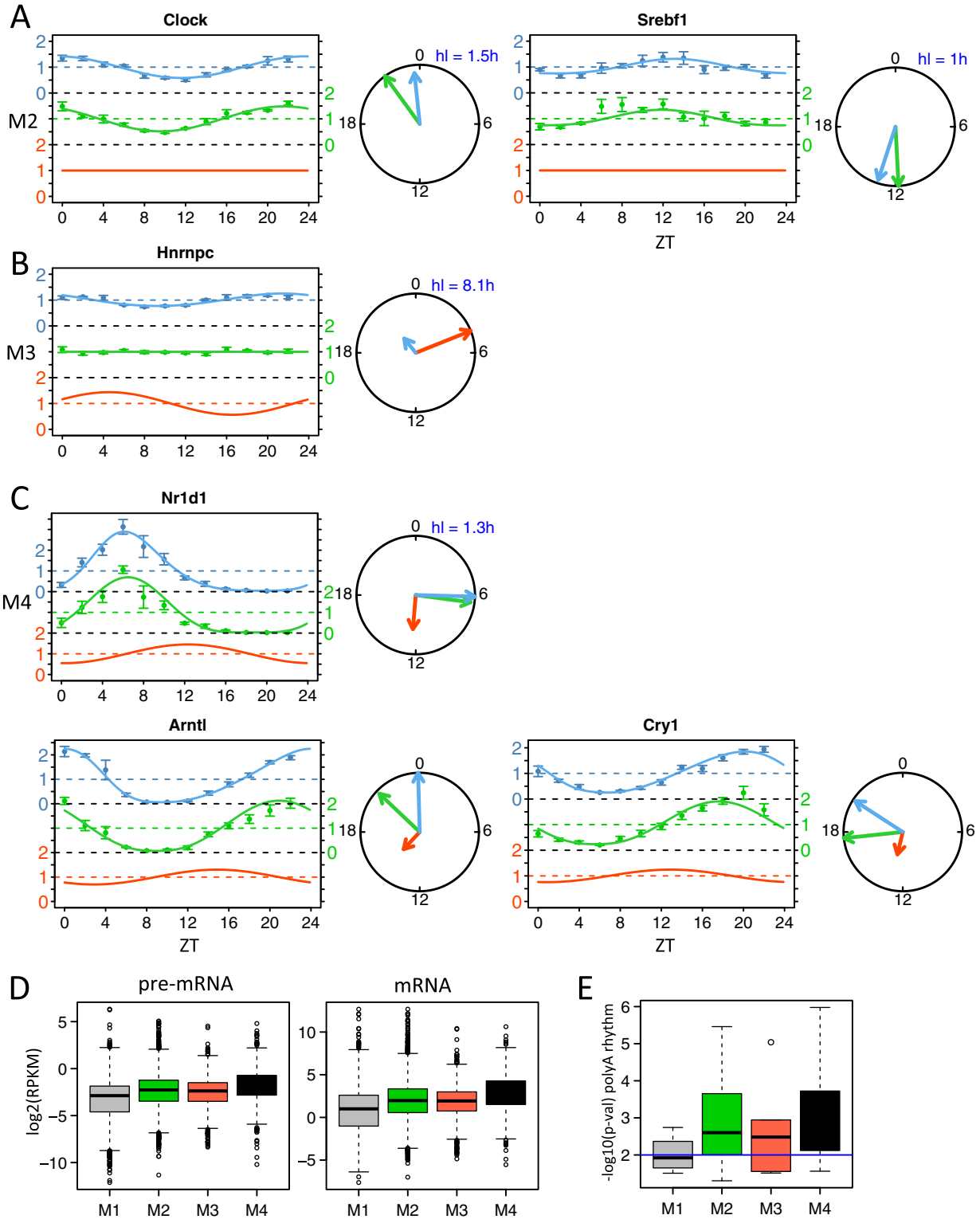

Fig. S3. Temporal profiles of mRNA and pre-mRNA of genes assigned to models M2–M4. (*A*) Two examples of genes identified as M2 (RS-CD). Data for mRNA (blue) and pre-mRNA (green) with error bars (SEM over four biological replicates) are shown in as relative expression, that is, the total normalized counts divided by the average value over time. Solids curves are the fitting for the optimal model and estimated parameters (blue, mRNA; green, pre-mRNA; red, degradation). Peak times (phase) and amplitudes are summarized in circle plots. Absolute half-lives [$\log(2)/\gamma_0$, in hours] are labeled. (*B*) Idem for M3. Here, the half-life was identifiable and is indicated on the plot. (*C*) Idem for M4. Half-life of *Nr1d1* was estimated to be 1.3 h, whereas half-lives were nonidentifiable for *Arntl* and *Cry1*. (*D*) Mean expression levels [$\log_2$(RPKM)] for pre-mRNA and mRNA in M1–M4. (*E*) *P* values of rhythmicity (Harmonic regression) of poly(A) tail length for genes classified in M1–M4. Blue horizontal line indicates a *P* value threshold of 0.01.

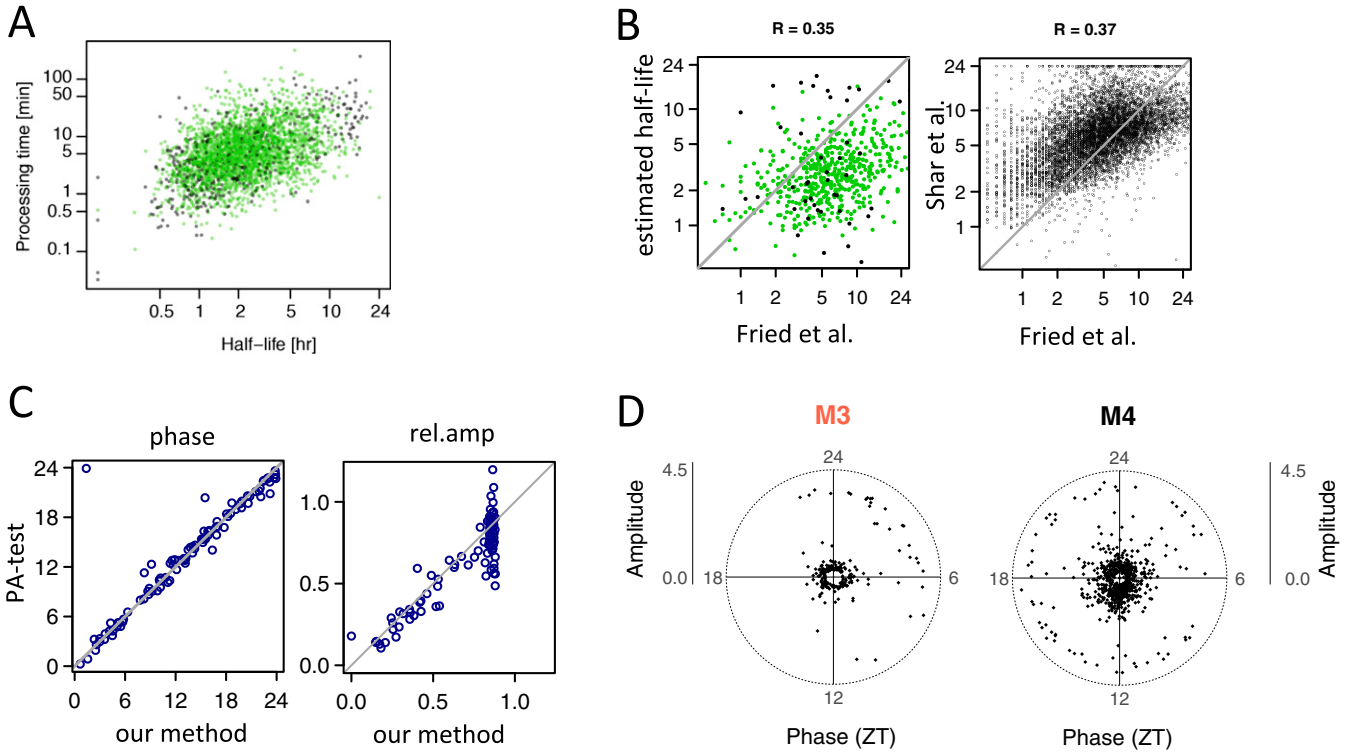

Fig. S4. Estimated kinetic parameters for mRNA degradation. (*A*) Scatterplot of half-lives (in hours) vs. processing times (in minutes) for transcripts with identifiable parameters ($R = 0.78$, $P < 10^{-16}$). Colors indicate the model that best described the synthesis–degradation dynamics (green for M2, red for M3, and black for M4). (*B*, *Left*) Scatterplot of half-lives (in hours) measured in NIH 3T3 cells in ref. 38 vs. ones estimated by our method ($R = 0.35$, $P < 10^{-17}$). (*Right*) Scatterplot of half-lives (in hours) measured in NIH 3T3 cells (21) vs. ones in mES cells in ref. 41 shows similar correlation ($R = 0.37$). (*C*) Comparisons of phases and relative amplitudes between our method and those obtained from the PA test. The relative amplitudes estimated by our method show a soft cutoff around 0.9 due to a penalizing sigmoid function used in the optimization algorithm. (*D*) Polar plot of rhythmic degradation rate ($\log_2$ peak-to-trough amplitude, radial axis) vs. phase (angular axis) of transcripts classified in M3 (*Left*) or M4 (*Right*) with identifiable degradation parameters (*Materials and Methods*).

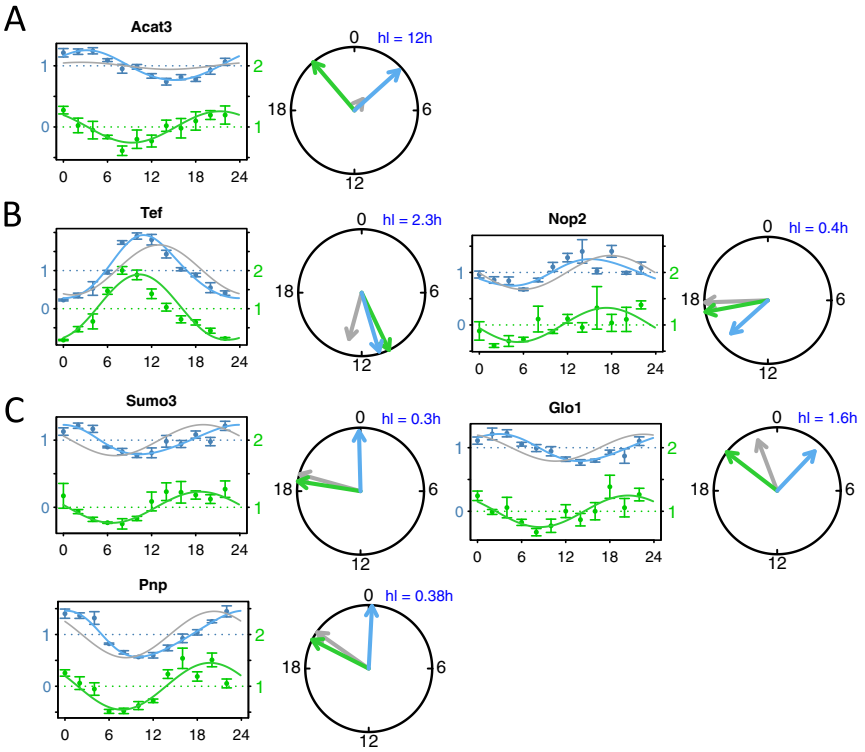

Fig. S5. Individual examples illustrating the distinct roles of rhythmic mRNA degradation in transmitting diurnal rhythms from transcription to mRNA. (*A*) Temporal relative expression profiles (absolute levels divided by the average expression over time) of pre-mRNA (green) and mRNA (blue) of C1 genes *Cbs* and *Acat3* (error bars show SEM over four biological replicates). The gray line shows predicted mRNA profile under assumption of constant degradation (relative amplitude of mRNA degradation ε_γ set to zero). (*B*) Idem as *A* for C2 genes *Tef* and *Nop2*. (*C*) Idem as *A* for C3 genes *Sumo3*, *Glo1*, and *Pnp*.

Dataset S1. Selected models and estimated kinetic parameters for genome-wide mRNAs by combining a maximum-likelihood approach with the Bayesian information criterion (BIC)

Dataset S1

(Sheet 1) The lengths and read counts of pre-mRNAs and mRNAs identified from the time-resolved total RNA-seq data (8) as well as the rhythmicity analysis using RPKM transformed read counts in $\log_2$ scale. (Sheet 2) Time-dependent dispersion parameters for pre-mRNAs and mRNAs estimated with DEseq2 (58); the identified outliers and parameter pre-mRNAs and mRNAs were also included. The optimal model selected with BIC, associated probabilities, estimated parameters, and SEs/mean (CV) are provided. In addition, identifiability analysis with profile likelihood (PL) for parameter γ_0 are also indicated. (Sheet 3) Explanation of column headers.

Dataset S2. Functional enrichment analysis for genes identified in M2–M4 performed using the DAVID Bioinformatics Resources 6.7

Dataset S2

Dataset S3. *P* values of PA test, estimated amplitudes, and phase of rhythmic mRNA degradation

Dataset S3

Results obtained by using mRNA half-lives measured from cell lines as described in ref. 9.

Dataset S4. GO analyses

Dataset S4

(*A*) Significant GO terms (*P* value < 0.001, hypergeometric test) of genes with rhythmic degradation in M4 model. GO term enrichment was performed in a 6-h sliding window to group genes with similar phases of mRNA degradation. (*B* and *C*) Functional enrichment analysis for genes with rhythmic mRNA accumulation in both WT and *Bmal1*$^{-/-}$ (*B*) and for genes with rhythmic mRNA accumulation in WT only (*C*). (*D*) Annotation of genes mentioned in the main text.

Dataset S5. Comparison of mRNA expression in ad libitum (AD) vs. restricted feeding (RF) conditions and in wild-type (WT) vs. *Bmal1*$^{-/-}$ (KO) mice

Dataset S5

(*A*) mRNA expression (mRNA-seq, $\log_2$ of counts per million reads), average level, phase, and amplitude of M3 transcripts in AL-WT, RF-WT, and RF-KO. Classification of each transcript into models describing their rhythmic pattern in each condition (rhythmic or constant, with similar phase and amplitude or not). (*B*) *A* column names and model ID descriptions.

Dataset S6. Prediction of mRBP activities

Dataset S6

(*A*) Phase and amplitude of predicted activities of mRBP motifs in both group (BMAL1-dependent and independent). (*B*) Estimation of phase and amplitude and SE on the estimates of mRNA binding proteins abundance profiles from ref. 50. FDR values for the assessment of their rhythmicity.