Science
AAAS

Supplementary Materials for

Title: An Intrinsic Oscillator Drives the Blood Stage Cycle of the Malaria Parasite, *Plasmodium falciparum*

Authors: Lauren M. Smith, Francis C. Motta, Garima Chopra, J. Kathleen Moch, Robert Riley Nerem, Bree Cummins, Kimberly E. Roche, Christina M. Kelliher, Adam R. Leman, John Harer, Tomas Gedeon, Norman C. Waters, Steven B. Haase
Correspondence to: steve.haase@duke.edu

This PDF file includes:

Other Supplementary Materials for this manuscript include the following:

Data S1 – Transcript levels for strains at each timepoint
Data S2 – Lists of genes represented in heatmaps by figure
Data S3 – Microscopy and culturing metadata for each strain
Data S4 – Microscopy staging values for each strain

Materials and Methods

Plasmodium falciparum strains, culturing, synchronization, and time-series

P. falciparum parasites were synchronized by alanine treatment and temperature cycling as described (*52*). Blood for parasite cultivation was from healthy volunteers under approved protocol WR1868.01. A single blood donation was used to culture the strains FVO-NIH, SA250, and D6, while a different donation was used for culturing the strain 3D7. Briefly, the strains were synchronized by alanine treatment (0.3 M alanine, 10 mM HEPES, pH 7.5) and temperature cycling 1–2 days before the start of the time series sample collection. During the experiment all the strains were grown at 37ºC (no temperature cycling). Growth conditions were: 4% hematocrit, RPMI 1640 supplemented with 10% human serum. Flasks were gassed with 5% CO_2, 5% O_2, 90% N_2 at 37ºC, and suspension cultures were shaken at 44 revolutions per minute. For the time series, samples were collected every three hours for 60–72 hours (1.5 cycles if assuming 48-hour cycle length). At collection, cultures were spun down, media aspirated, and pellet flash frozen and stored at -80ºC until RNA extraction.

For each strain, sampling started at late schizont/segmentor stage (with a few early rings) to capture the invasion/transition to rings time point for each one. The exception was 3D7, where time series collections began at early ring stage (Data S3). Sampling began (0-hour time point) earlier at the schizont stage after synchronization in order to capture the transition to rings at the beginning of the experiment (i.e. our “time 0” sampling protocol was modified after observing variability in invasion times between *P. falciparum* laboratory strains). Each time series experiment was split into two pools of cultures after synchronization to avoid reaching stressful, high parasitemia levels over time: time points 0-33 hours were sampled from culture “A” (labeled as “O” in 3D7), which was seeded at 0.1% parasitemia (0.37% parasitemia in 3D7) and

progressed to ~3–6% between time points 0-33 hours; time points 36–69 hours were sampled from culture "B" (labeled as "N" in 3D7), which was seeded at 0.01% parasitemia (0.023% in 3D7) and progressed to ~3–5%. For SA250, where parasitemia did not increase substantially, samples were used from the "A" culture for the entire time series. Blood smears from each time point were counted and developmentally staged by microscopy (Data S4, Fig. S3). Stage-specific parasitemia levels over time were quantified using SYBR Green 1 and flow cytometric analyses (Data S3).

RNA extraction and sequencing

Total RNA from frozen, packed red blood cells was extracted via a phenol-chloroform protocol (*53*). Citrate-buffered phenol was substituted for acid phenol to compensate for the high buffering capacity of blood. While the volume of red blood cells varied per sample, the ratios of reagent to sample as found in the cited protocol were retained, with the exception of the initial thawing in TES buffer, which was changed to a 2:1 ratio of TES to cells.

Library preparation and sequencing was performed by the Duke Sequencing Core. All samples were prepared using Stranded Total RNA Kapa kits with rRNA Ribo-Zero Globin depletion intro libraries of 50 base pair, single-end reads, except for 3D7 samples, which were prepared in libraries of 125 base pair, paired-end reads. All samples were sequenced on an Illumina HiSeq 4000 instrument, except 3D7 samples, which were sequenced on an Illumina HiSeq 2500 instrument. Samples were multiplexed as follows: seven 3D7 samples were combined in each of three lanes, 12 FVO-NIH samples were combined into each of two lanes, 11 SA250 samples were combined into each of two lanes, and 12 D6 samples were combined into each of two lanes.

Read processing

A summary of read alignment statistics per strain is shown in Table S8. STAR (*54*) was used to create a genome index for read mapping, using the *P. falciparum* 3D7 v3.0 genome as found on www.genedb.org, and the Sanger Institute's 3D7 annotation downloaded on March 4th, 2016. The following commands were used to create the genome build (with generic file names in *italics*):

```
STAR --runThreadN 1 --runMode genomeGenerate --genomeDir
path_to_genome_directory --genomeFastaFiles path_to_fasta_file --
sjdbGTFfile path_to_gff3_annotation_file --sjdbGTFfeatureExon CDS --
sjdbGTFtagExonParentTranscript Parent
```

For read alignment, intron size was set to 10—3000 base pairs in accordance with prior research (*55*). STAR was used for alignment using the following commands (with generic file names in *italics*):

```
STAR --runThreadN 3 --runMode alignReads --genomeDir
path_to_3D7_genome_build --readFilesIn sample.fastq --outFilterType
BySJout --alignIntronMin 10 --alignIntronMax 3000 --outFileNamePrefix
./STAR_output/ --outFilterIntronMotifs RemoveNoncanonical
```

SAMtools was used to created aligned, sorted SAM files for each sample. The Cufflinks (*56*) suite of tools was used for as follows (with generic file names in *italics*):

```
cuffquant --library-type=fr-firststrand path_to_gff3_annotation_file
sample.aligned.sorted.sam
```

Time series samples for each strain were normalized together as follows (with generic file names in *italics*):

```
cuffnorm –library-type=fr-firstrand path_to_gff3_annotation_file *.cxb
```

The "genes.fpkm_table" output, representing normalized FPKM gene expression, were used in all analyses. The FPKM values for all genes and strains are available as tables in Data S1. As values less than 1 were found to interfere with some periodicity detection algorithms (data not shown), 1 was added to each FPKM value.

Periodic gene detection

Genes were filtered for peak FPKM expression of at least 2 (or 3, with the addition of 1) to reduce noisy genes. We used JTK_CYCLE (*21*) to filter for periodic genes with the following initial search parameters: 3D7, 36–39h; FVO-NIH, 45–48h; SA250, 47–50h; D6, 36–39h. After arriving at a more precise determination of period length for each strain (see "Estimation of period length per strain" and Table S2), we re-ran JTK_CYCLE on SA250 with a 51–54h search period, replacing the initial run. These sets of genes were used for all downstream analysis and are the "periodic genes" used in the main body of this paper.

Using data from the initial run of JTK_CYCLE, we sought to find a significance cutoff that favored preserving periodic genes (Table S1). The distribution of JTK_CYCLE's ADJ.P values for all genes was sharply grouped towards 0 (most periodic) in all strains (Fig. S14). We chose 0.25 as a reasonable cutoff based on these distributions. Since SA250 is the strain most weakly skewed towards zero (i.e. less periodic), we used it to visually compare the 0.25 cutoff versus the more traditional 0.05 p-value (Fig. S15). We found that both thresholds yielded a largely periodic transcriptome, with the genes falling in between the two thresholds mostly periodic as well. Therefore, we consider $p < 0.25$ a reasonable cutoff, and use it for all genes termed "periodic" in this study. This corresponds to a false discovery rate (FDR) of 25 – 28%, depending on the strain. (Table S1).

For comparative analyses using mouse tissue transcriptomics, we ran liver, lung, and kidney tissue data through JTK_CYCLE with a 22–24 hr search window and kept genes with BH.Q < 0.05 in all three tissues, yielding 293 periodic genes.

Computing ordering similarities

All datasets were interpolated with a PCHIP (Piecewise Cubic Hermite Interpolating Polynomial) spline to one-hour intervals; the *P. falciparum* timeseries were wrapped as described below and offset to a common starting point (~50% trophozoite-to-schizont transition; Fig. S3), while the mouse datasets were truncated to the first 24-hour circadian cycle. In addition, we down-sampled the *P. falciparum* dataset by removing every odd-number timepoint (1, 3, 5…) to bring sample density in line with the circadian dataset. Expression peak times for genes periodic across mouse tissues or parasite strains were calculated as a percent of cycle length (24 hours for mouse, final estimated period length for *P. falciparum* strains, as shown in Table S2).

We chose 3D7 to be the reference for comparison to the remaining *P. falciparum* strains. For the circadian analysis, we chose liver tissue as the reference to compare lung and kidney tissue to, as these were the tissues with the greatest number of circadian genes (*16*). Kidney and lung peak time values were compared to liver values, and a difference of <5% was considered "in phase" with liver. Any gene in phase with liver in at least one tissue was retained, yielding 126 genes. In *P. falciparum*, FVO-NIH, SA250, and D6 were compared to 3D7. Any genes in phase (<5% peak time difference) with 3D7 in at least two other strains were retained, yielding 122 genes. For computational feasibility (see below), the final gene set was restricted to 107 mouse genes and 119 parasite genes.

From these gene sets, we took subsets of six genes each and computed a partial order representing the relationships between the local minima and maxima of the expression time series over a range of noise levels (plus or minus 6–10%), as presented in (*30*). Similarity at each noise level was computed between the partial order for each strain or tissue using a distance with respect to the reference (3D7 for *P. falciparum* and liver tissue for mouse). Similarity is given as 1–*d*, where *d* is a graph metric varying between 0 and 1 given in (*57*); further details are given below.

A sample size of 5000 subsets of six genes each was chosen from the genes in the in-phase gene list and the computed partial orders were compared to the analogous partial order in the reference to calculate the similarity. The average similarity across these samples approximates the similarity of all the genes in the in-phase gene list. A baseline similarity was computed for comparison. We created a null distribution for each strain or tissue by randomly interchanging gene names and phase shifting each time series by a random amount. Specifically, if a time series is viewed as a list $a_1, a_2, a_3, \ldots, a_n$ of length n, we select a random m such that $1 \leq m \leq n$ and create a new phase shifted time series

$$a_m, a_{m+1}, \ldots, a_n, a_1, a_2, \ldots, a_{m-1}.$$

This operation preserves the characteristics of the data set except for phase shift (ordering of extrema). For each sample, the baseline similarity was computed between the unpermuted and unshifted reference data set and the permuted and shifted data sets. This provides a baseline score for each sample, along with the actual similarity score.

The results are shown in Fig. S16. The mean and plus or minus one standard deviation of the similarity score as a function of noise level (ε) is shown for the liver/kidney, liver/lung, 3D7/SA250, 3D7/FVO, and 3D7/D6 comparisons. The parasite strain comparisons have greater

similarity and greater distance from the baseline comparison than the circadian samples. The mean and standard deviations over all ε for the in-phase and baseline comparisons are shown in Table S3, as well as the mean and standard deviation of the difference. A further quantity is shown

$$\alpha = (mean)_{\epsilon}\ (N_{\epsilon}/5000)$$

where N_{ϵ} is the number of times that the in-phase comparison exceeds the baseline comparison at noise level ε. In other words, α is the proportion of time that the in-phase comparison out-performs the baseline comparison. The range is 61–68% for mouse tissue data and 75–95% for *P. falciparum* strains.

In (*30*), a similarity measure was developed to compare gene expression ordering in time series data. This similarity measure is based on the timing of peaks and valleys for a collection of genes compared across data sets and is independent of amplitude due to time series normalization on the interval [−0.5,0.5]. Each computation of similarity depends on a level of noise. The peaks and valleys below the noise level are disregarded in the similarity computation. As noise increases, the number of peaks and valleys decrease.

The methodology in (*57*) assigns a time interval to every peak and valley (extremum) in each time series interval. The noise level $0 < \varepsilon < 1$ determines which of these intervals are consistent with stochastic behavior and removes these from consideration. The remaining intervals can be partially ordered. If the intervals overlap, then the order of two extrema cannot be distinguished at that ε. If the intervals are disjoint, then the order is known, and we call each such pair an "ordered relation."

This partial ordering has an equivalent graph, where every node in the graph is an interval representing an extremum and every edge in the graph is an ordered relation. There is such a

graph for each data set and each noise level ε. The similarity of these graphs is given by 1 – *d*, where *d* is a specially developed graph distance between partial orders that varies between 0 and 1 (*30, 57*). This similarity measure gives roughly the ratio of shared ordered relations between two data sets to the total possible that could be shared. The proportion of shared ordered relations between data sets is the similarity of the ordering of extrema between data sets.

The similarity score defined here is a combinatorially hard problem to compute as the number of time series increases and as the number of extrema per time series increases. For this reason, we computed similarity using partial orders constructed from six randomly chosen genes at a time. As mentioned above, in the parasite data there were 122 in-phase genes, but we only used 119 that had eight or fewer local extrema in the time series at $\varepsilon = 0.06$. In the mouse tissue data there were 126 in-phase genes, of which we used 107 that had eight or fewer local extrema in the time series at $\varepsilon = 0.06$. This limitation on the number of extrema excluded the noisiest data and increased feasible sample size.

Estimation of period length per strain

To estimate the period length of each strain, a set of periodic genes were used that came from a second preliminary run of JTK_CYCLE with the same parameters as mentioned above, except for SA250 the search range was 50–53h. All genes were filtered to ADJ.P < 0.25.

Microscopy and expression values for each strain were first interpolated via a PCHIP spline from a resolution of 3 hours to 1 hour. A consensus cycle length for each strain was determined on the basis of four measurements: (1) error-minimizing expression cycle length, (2) error-minimizing microscopy cycle length, (3) the distribution of distances in hours of the two largest expression maxima for all genes in each strain, and (4) the distribution of distances of the two

smallest expression minima for all genes in each strain. To determine (1) and (2), microscopy and expression data were independently wrapped so that repeating data in the cycle overlapped and root mean squared error (RMSE) of the overlapping measurements was calculated. To accomplish the wrapping, at least one full cycle was assumed to exist within the time series. Mean error was calculated for all possible wraps of the data that contained a single core cycle within the series, not less than half the length of that series. This restriction allowed us to avoid multiply wrapping the data and was reasonable since, in this dataset, no strain was observed for long enough to have completed a second cycle, as determined by microscopic observation. The procedure described above is analogous to the phase dispersion minimization algorithm of Stellingwerf (*25*). The result was two error-minimizing mappings, by gene expression (Fig. S7) and microscopy (Fig. S8) of observations over more than one cycle to a single cycle. By determining an error-minimizing wrap for each strain, this method also inferred a probable cycle length in hours. Overlapping measurements were averaged in the case of both microscopy and expression data to give a final, single-cycle representation suitable for subsequent analysis. To determine (3) and (4), the distribution of distances between paired maxima was computed on interpolated but unwrapped data. Maxima were identified as critical points in the splined expression data and the distance between the largest two maxima (peak-to-peak distance, in hours) was calculated. The resulting distribution of distances over all genes displayed a peak we interpreted as a consensus cycle length, the assumption being that, most frequently, the paired maxima represented two measurements from the same point in the cycle (Fig. S9). This process was repeated with the two smallest minima in each expression profile to give a distribution over trough-to-trough distances for all genes (Fig. S10).

To arrive at a consensus period length for each strain, we computed a weighted average of the methods by giving more weight to metrics that utilized more information and were likely more accurate. The above metric (transcriptomic wrap, microscopic wrap, peak distance, trough distance) were weighted at 3:2:1:1, respectively, before averaging the values together for a final estimate (Table S2). However, due to the limited cycle length of strain SA250's microscopic and transcriptomic time series, and thus greater unreliability of wrapping estimates, a simple averaging was performed for the final period length estimate.

Comparison of parasite stage lengths between strains

In order to compare the relative lengths of each parasite intraerythrocytic cycle stage (ring, trophozoite, and schizont,) we re-wrapped the microscopy data using the above described wrapping procedure, this time to the consensus period length determined by the method explained above. Since the best-fit microscopic wrap periods of 3D7, SA250, and D6 were identical to the consensus periods, we only needed to re-wrap FVO-NIH. Fig. S11 visualizes this procedure and Table S4 provides RMSE values for the wrap.

Modelling synchrony loss with a population of phase oscillators

The bioluminescence time series traces of the 80 fibroblast cells—which we denote by $s_i(t_k)$, for $i = 1,\ldots,80$ and $k = 1,\ldots,M_i$—were first denoised using a discrete stationary wavelet transform, following the methods employed in the original paper (*36*). In particular, the stationary wavelet transform swt() implemented in MATLAB R2018 was used with an order 12 Daubechies mother wavelet (*58*) to decompose each trace to level 7, after extending continuously each trace to a length equal to the smallest integer divisible by $128 = 2^7$. All but the level 5 and

6 detail coefficients were eliminated, thereby removing the high and low frequency components of each trace, while retaining approximately the spectra associated with period ranges of 11 to 46 hours. Peaks of the denoised traces, $\bar{s}_i(t_k), 1 \leq k \leq M_i$, were readily identified as local maxima (i.e. a point $\bar{s}_i(t_k), 1 < k < T_i$ is a peak if $s_i(t_{k-1}) < s_i(t_k)$ and $s_i(t_{k+1}) < s_i(t_k)$). To minimize the boundary effects caused by extending the traces and by denoising, the first three peaks at the start and end of each denoised trace were removed from consideration. Cycle lengths were taken to be the peak-to-peak times between adjacent peaks by regarding each peak as a phase marker representing the time of initiation of a new cycle.

The denoising and peak identification process yielded a small number of spurious peaks whose amplitude (i.e. the minimal difference between the peak and the preceding and succeeding local minima) was deemed too small to be reliable (Fig. S17). Of the 2831 identified peaks, 18 were considered spurious and were eliminated from consideration when computing peak-to-peak times.

To compare the variability in free-running circadian cycle progression rates observed by single-cell imaging of PER2::LUC bioluminescence in fibroblasts (*36*) with the intraerythrocytic cycle progression rates of the *P. falciparum* strains in this study, and to ensure physically-reasonable models of the distributions of peak-to-peak times, we considered several distributions supported on $(0, \infty)$. Estimation of distribution parameters for the fibroblast circadian cycle lengths was done using the MATLAB algorithm fitdist() and it was found that the data was well modelled by the two-parameter log logistic distribution having the probability density function $LL(x; \mu, \sigma) = \exp(z)/(\sigma x(1 + \exp(z))\text{^}2$), where $z = (\ln(x) - \mu)/\sigma$ (See Fig. S18). Incidentally, the maximum likelihood estimation employed by the fitdist() function found estimates for the scale parameter $\mu = 3.21153 \pm 0.002991$ and shape parameter $\sigma =$

0.0460319 ± 0.001472 which corresponds to an analytic distribution with mean and standard deviation within approximately 0.13% and 0.91% of the sample mean and sample standard deviation respectively.

If X is a random variable representing the progression rate (frequency) of an oscillation, then $Y = 1/X$ represents the cycle length (period) of the oscillation. Let $X \sim LL(\mu, \sigma)$ be a log logistic random variable. A straightforward calculation shows that $Y = 1/X \sim LL(-\mu, \sigma)$. Thus, if we assume the rate of progression through a cycle is log logistically distributed with parameters μ and σ, it follows that the cycle length is likewise distributed log logistically with parameters $-\mu$ and σ. Assume $X \sim LL(\mu, \sigma)$, then the expectation and variance of X are respectively $E[X] = \exp(\mu)\pi\sigma/\sin(\pi\sigma)$ and $Var[X] = \exp(2\mu)(2\pi\sigma/\sin(2\pi\sigma) - (\pi\sigma)^2/(\sin^2(\pi\sigma)))$, if $0 < \sigma < 1/2$.

To identify the variability of *P. falciparum* intraerythrocytic cycle (IEC) period and thereby model the apparent population synchrony loss observed by microscopy, we adopt a simple phase-oscillator model for progression through the IEC. In particular, we parameterize the IEC by phase, $0 \leq \theta \leq 1$, and assume that each parasite progresses according to $d\theta/dt = V$, where $V \sim LL(\ln(\pi\sigma/\sin(\pi\sigma)), \sigma)$ is a constant cycle progression rate (natural frequency), and $P = 1/V \sim LL(\ln(\sin(\pi\sigma)/\pi\sigma), \sigma)$ is the corresponding cycle length. By choosing $\mu = \ln(\sin(\pi\sigma)/\pi\sigma)$, we insist that $E[P] = 1$, and allow the shape parameter σ to control the variability of the cycle length across the population of cells. For small values of σ $E[V] \approx 1$, and so under this model both the average progression rate and the average period length are near 1, provided that σ is small. To compare the model to real data we first rescale the time axis from units of hours to units of periods using the mean period for each strain as estimated from the

transcriptional expression data. To show that the model was insensitive to this choice, we considered these mean periods as well as these mean periods plus or minus one hour.

A calculation shows that the coefficient of variation, $\mathrm{CV}[\mathrm{P}] = \sqrt{(\tan(\pi\sigma)/\pi\sigma - 1)}$, of the periods of oscillation depends only on σ and approaches 0 as σ approaches 0 but is unbounded as σ approaches 1/2.

Let $\theta_i(t)$ be the IEC phase of the i-th parasite in a population of *P. falciparum* at time t. Assume that for a given strain there are well-defined phases at which a parasite transitions between the discrete stages of the IEC. Namely, define $0 = \theta_{SR} < \theta_{RT} < \theta_{TS} < 1$ to be successively the schizont-to-ring, ring-to-trophozoite and trophozoite-to-schizont transition phases. Again, assume that the phase of the i-th parasite evolves according to $d\theta_i/dt = V_i$ so that, by direct integration, $\theta_i(t) = V_i(t) + \theta_i(0)$. We further assume that $\theta_i(0) \sim \mathrm{WN}(\theta_0, \sigma_0)$, where $\mathrm{WN}(\mu, \sigma)$ is the wrapped normal distribution, θ_0 is the unknown mean initial phase of the population, and σ_0 is the unknown scale parameter expressing how well the population is initially phase-synchronized in the IEC. Note that the length of available *P. falciparum* microscopy data is less than two full cycles and our present effort is to compare the variability of cycle lengths across a population. For these reasons we do not explicitly model cycle-over-cycle variability or within-cycle variability of an individual parasite's cycle length due to stochastic variations and instead assume each parasite has a constant natural IEC frequency as doing otherwise would serve only to make a more complicated model whose conclusions would be unchanged.

For a population of n phase oscillators modelling parasite progression through the IEC, in accordance with the five unknown parameters $(\sigma, \theta_{R,T}, \theta_{T,S}, \theta_0, \sigma_0)$ outlined above, we define at each time the fraction of the population in ring, trophozoite, and schizont stages respectively as

$$\tilde{R}(t) = \#\{i \mid 0 \le \theta_i(t) < \theta_{RT}\}/n,$$

$$\tilde{T}(t) = \#\{i \mid \theta_{RT} \le \theta_i(t) < \theta_{TS}\}/n,$$

and

$$\tilde{S}(t) = \#\{i \mid \theta_{TS} \le \theta_i(t) < 1\}/n.$$

Given experimental staging data, $R(t), T(t), S(t)$, collected at times t_k, $1 \le k \le M$, we define the total squared error function over the unknown parameter space $\Omega = (\sigma, \theta_{RT}, \theta_{TS}, \theta_0, \sigma_0)$

$$TSE(\Omega) = \sum_{k=1}^{M} \left(\tilde{R}(t_k) - R(t_k)\right)^2 + \left(\tilde{T}(t_k) - T(t_k)\right)^2 + \left(\tilde{S}(t_k) - S(t_k)\right)^2$$

and search for the parameter choice which minimizes the total error between simulation and experiment, $\Omega^* = \underset{\Omega}{\operatorname{argmin}}\, TSE(\Omega)$. Determination of model parameters Ω^* is done using the constrained genetic algorithm optimization ga() implemented in MATLAB R2018 with parameter constraints $0 < \sigma < 1/2,\ 0 < \theta_{TS}, \theta_{RT}, \theta_0 < 1,\ 0 < \theta_{TS} - \theta_{RT}$, and $0 < \sigma_0 < 1$ necessary to ensure a mathematically-sound and physically-realizable model.

Having found an optimal choice of parameters for each strain, we subsequently sample a small number of individual oscillators (100) from the entire population at each time point and use these samples to estimate the fraction of cells in each developmental stage. By subsampling the population many times (100,000), we estimate the expected staging curves and their expected variability caused by finite sampling. This process is consistent with the experiment in which small samples of the parasites are drawn at each time point to estimate the fraction of the population in each stage (Fig. S3).

To better understand the impact of parasite replication on stage percentage curve dynamics, and the interplay between replication rates and period-length variance over a

population, we extended the model described above to accommodate parasite replication. In accordance with the current understanding of the *P. falciparum* IEC and the experimental process used here to count the fraction of parasites in each stage, we model replication as occurring instantaneously at the schizont-ring-transition by instantiating some specified, fixed-number, N, new "daughter" oscillators whenever any "mother" oscillator's phase crosses an integer value. This model reflects the fact that a parasite in late schizont, which has already undergone multiple rounds of division, is experimentally counted as a single parasite in schizont stage, and only after the "daughter" merozoites emerge from their current red blood cell and invade a new red blood cell will they be counted as individuals (in ring stage). In this way, we model the population as having a replication rate of N by having each parasite "divide" into N ring-stage parasites after crossing the schizont-ring phase transition. The instantaneous "birth" of N parasites into ring stage as any oscillator crosses the schizont-to-ring transition is in agreement with a previously published discrete-time model (*37*), which our model generalizes. The period lengths of the new oscillators are drawn at random according to the same distribution as the initial population: $\mathrm{LL}(\ln(\sin(\pi\sigma)/\pi\sigma), \sigma)$.

Figs S1 – S18

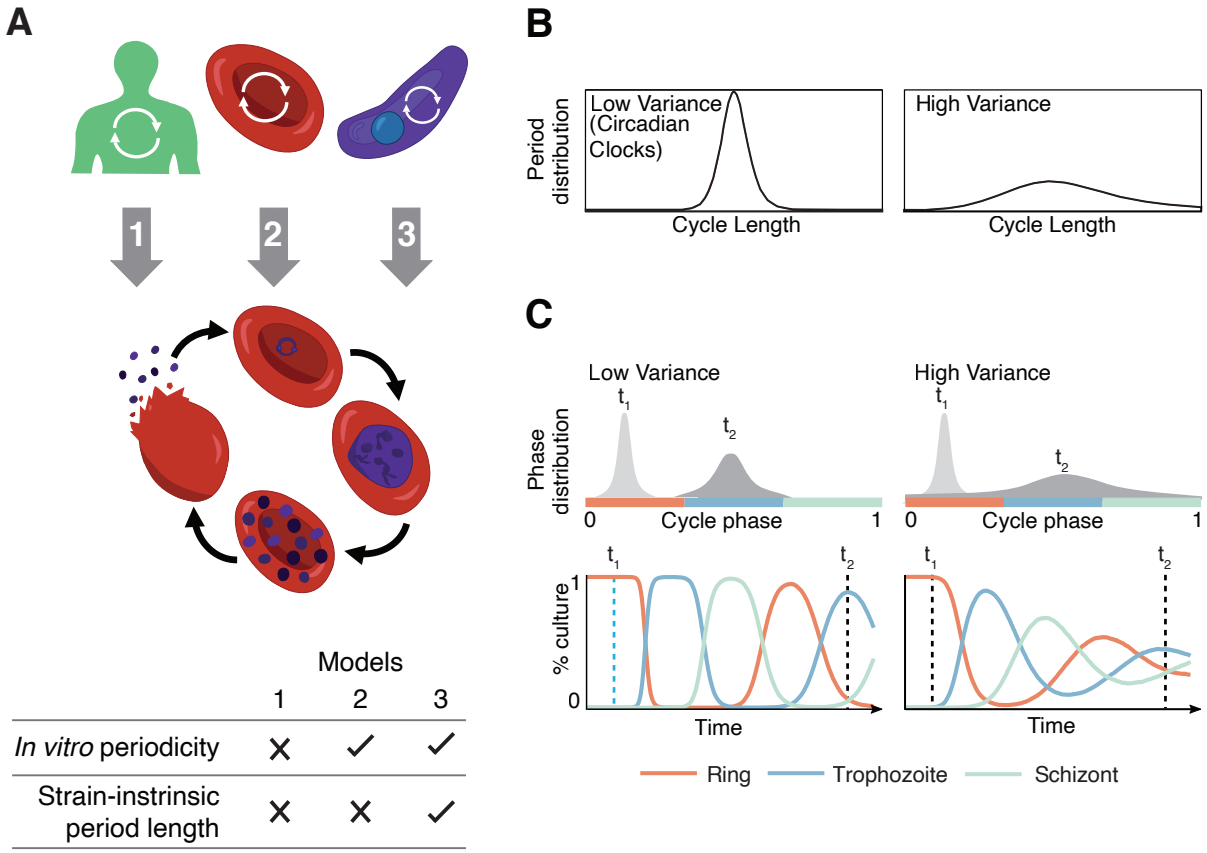

Fig. S1. (A) Three models for the source of oscillation in the malaria parasite *P. falciparum*. In Model 1, the human circadian rhythm or a circadian-controlled process drives oscillation of the parasite. In Model 2, the proposed 24-hour peroxiredoxin cycle of red blood cells is responsible. In Model 3, the parasite itself has an independent innate oscillator capable of driving its own intraerythrocytic cycle. Experimental predictions from each of these models are shown. (B) Models of relatively low and high variance distributions of the period lengths of members of an oscillating population. (C) Models of the continuous-phase distributions of a population of *P. falciparum* parasites whose phases are initially well-aligned, measured at an early time (t_1) and a

later time (t_2) (top), together with the corresponding discrete-stage percentage curves as a function of time (bottom).

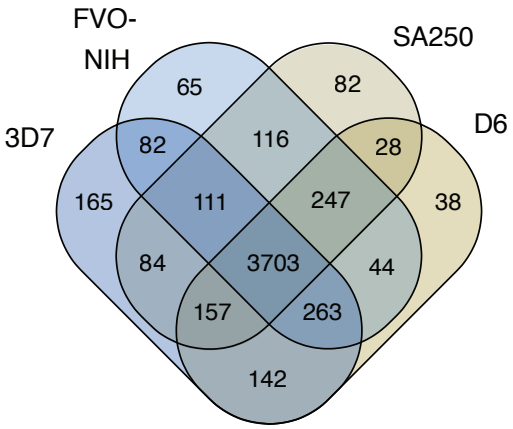

Fig. S2. Overlap between periodic genes (ADJ.P < 0.25) among the four strains. 3,703 genes (between 79–82% of each strain) are part of a shared set of periodic genes.

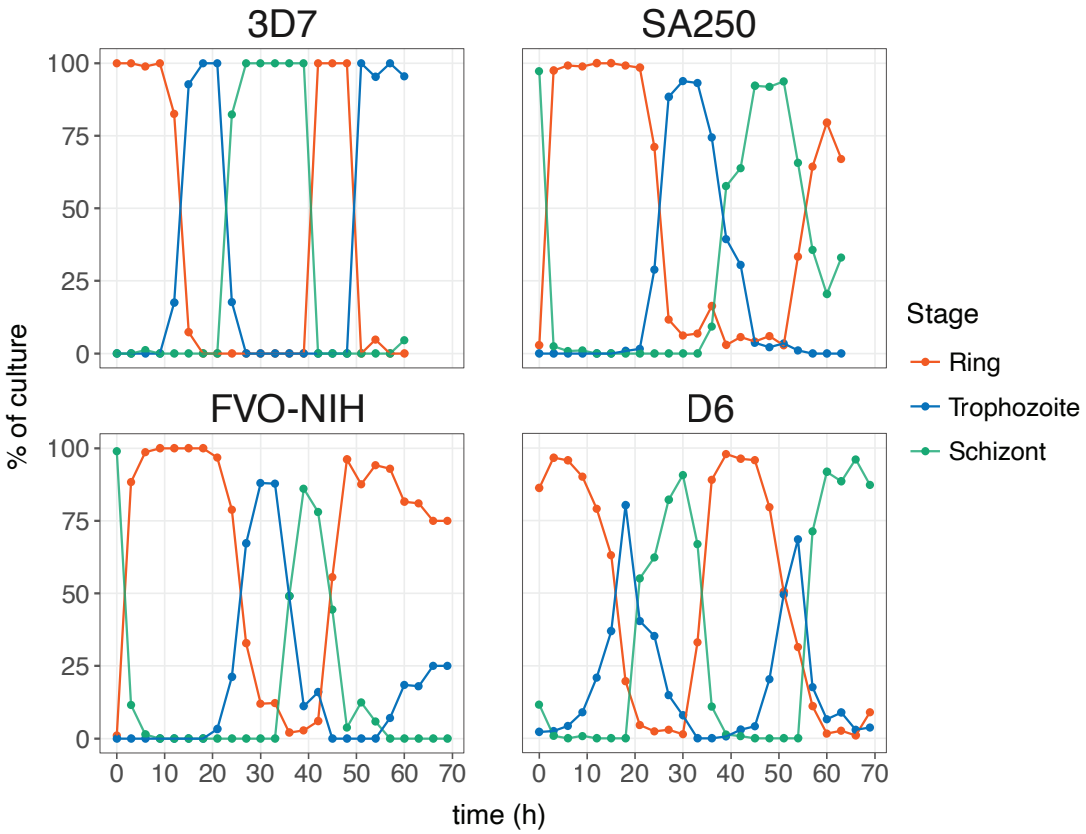

Fig. S3. Parasite culture staging data by microscopic examination for the four strains used in this study (see Data S4). Giemsa-stained counts were taken every three hours at the time of culture sampling for RNA-seq. Parasites were grouped by into ring, trophozoite, or schizont stages, and each stage is presented as percent of total parasites per time point.

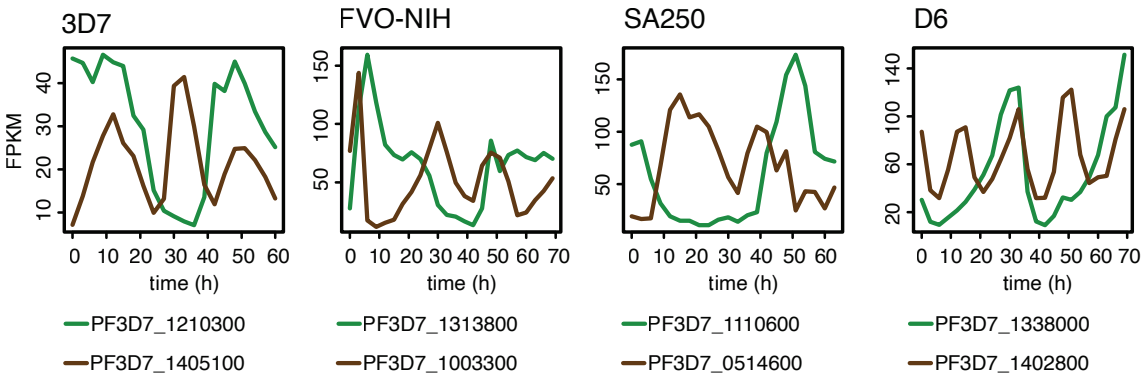

Fig. S4. Examples of harmonic (half-period; brown lines) gene expression compared to full-period expression (green lines).

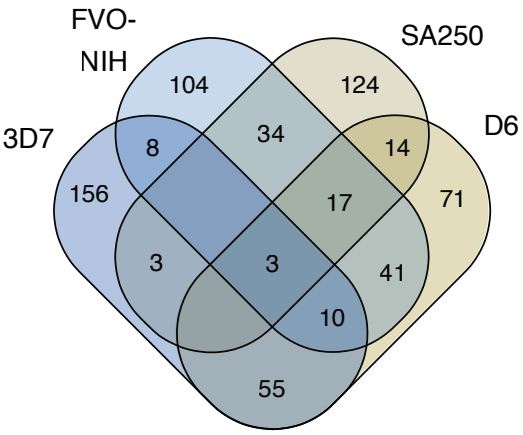

Fig. S5. Overlap between genes identified as oscillating at the first harmonic (half-period.)

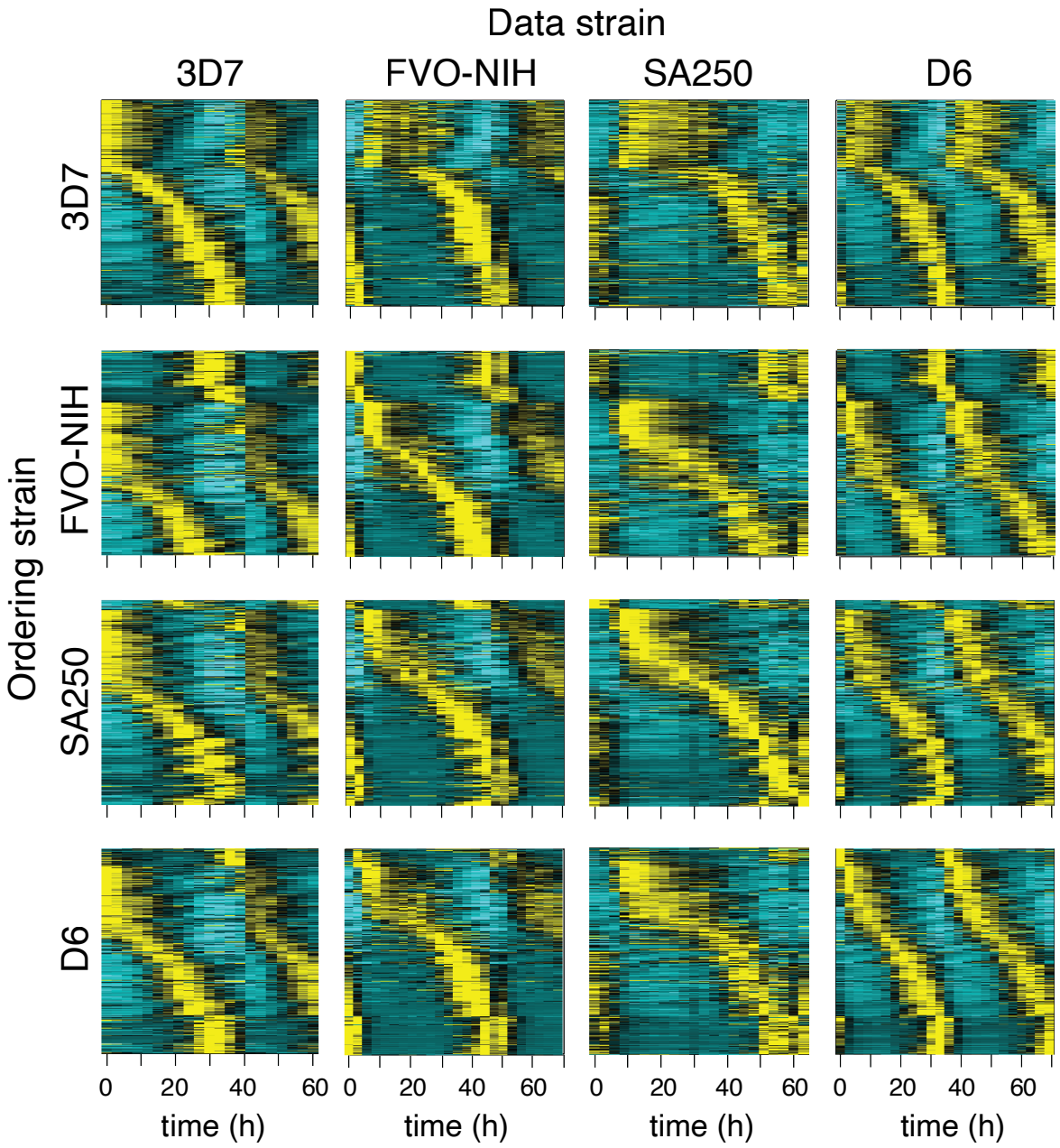

Fig. S6. Gene ordering is largely conserved between strains, regardless of which strain is used for an ordering standard. The ordering of the set of 3,703 shared periodic genes is determined per strain by peak expression time (columns) and applied to all strains (rows).

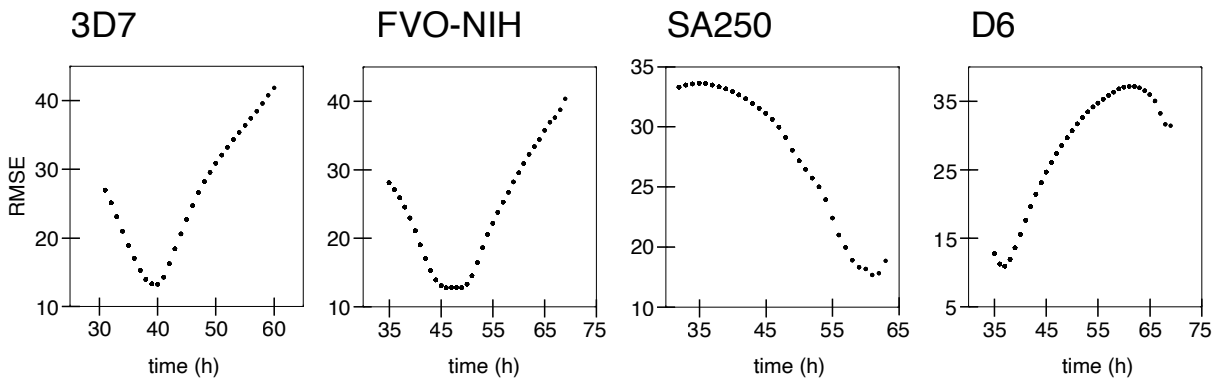

Fig. S7. Wrapping procedure used to find the period length of each strain according to the periodic transcriptome in bulk. For each strain, expression data was splined to 1 hour using PCHIP. Root Mean Square Error (RMSE) of the overlapping measurements for a succession of period lengths. See Table S4 for RMSE values of best-fit period.

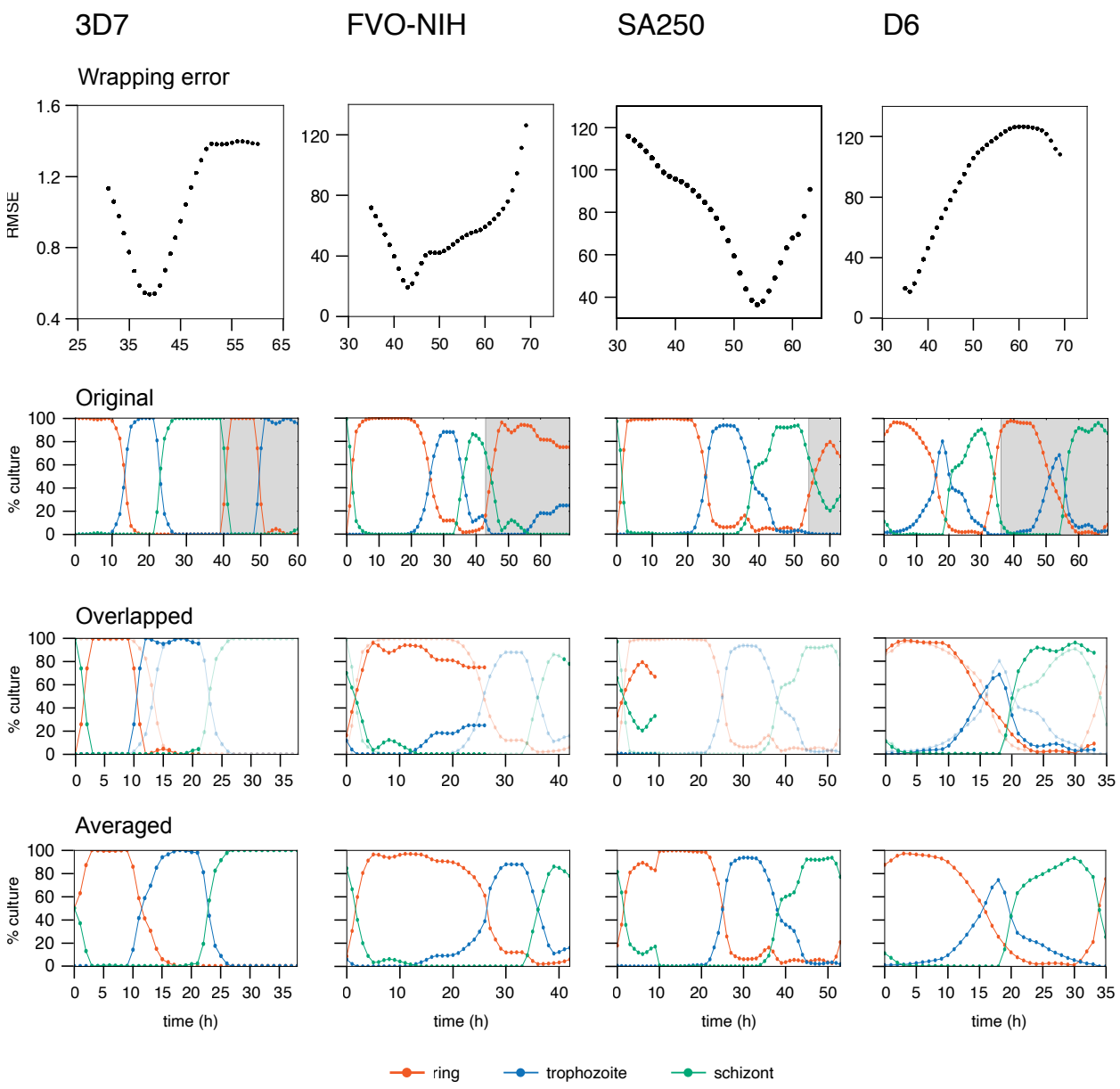

Fig. S8. Wrapping procedure used to find the period length of each strain according to recorded microscopic data. For each strain, data was splined to 1 hour using PCHIP. A succession of period lengths for wrapping were measured for RMSE of the overlapping measurements, here pictured per strain (see also Table S4 for RMSE values of best-fit values.) For each strain's best-fit wrap, we show the original microscopy data (with grey shading representing the data wrapped over the cycle), a visualization of the overlap, and the final wrapped data averaged together.

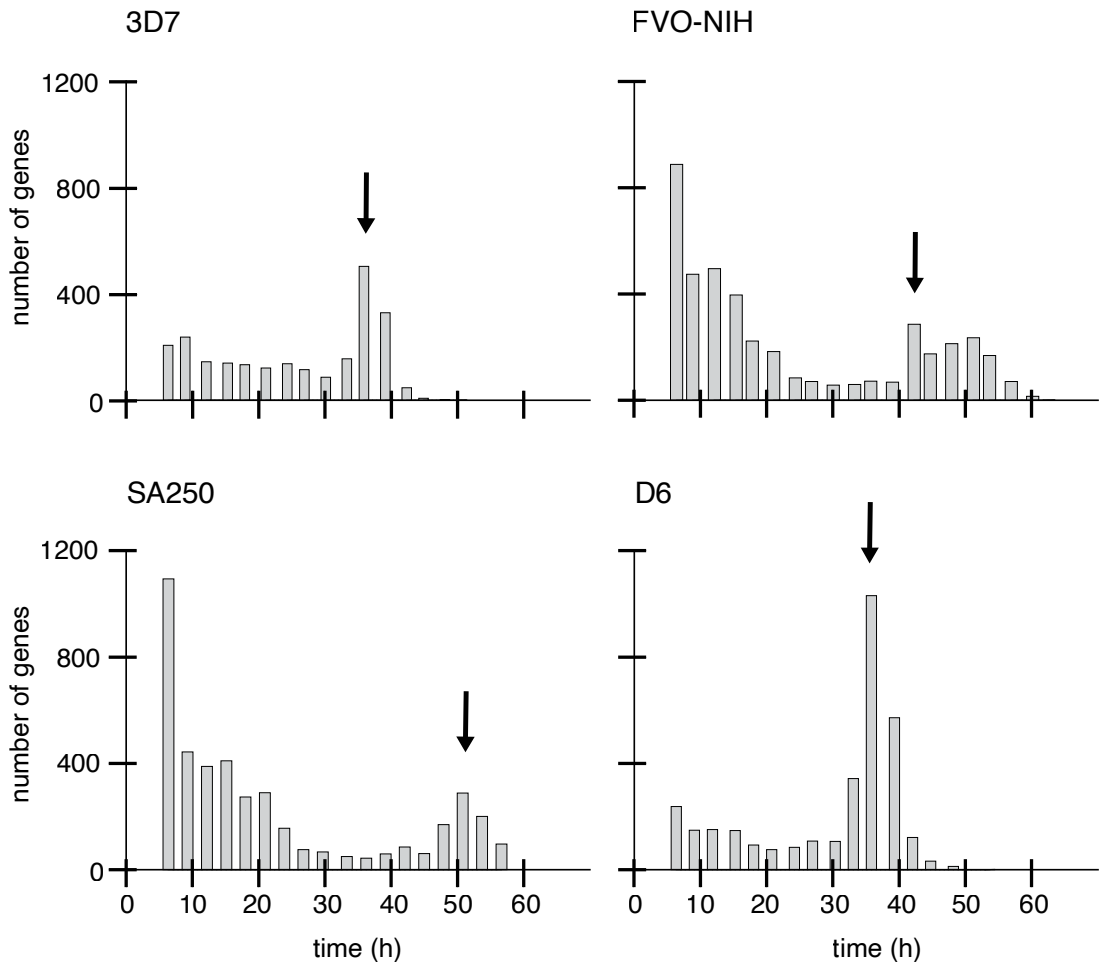

Fig. S9. Distribution of distance in hours between maxima of genes for each strain (peak-peak distance.) Using expression data PCHIP splined to one-hour intervals, the distances between the two largest maxima (gene expression peaks) were calculated. The most frequent (locally modal) distance is shown by arrows. Values are found in Table S2.

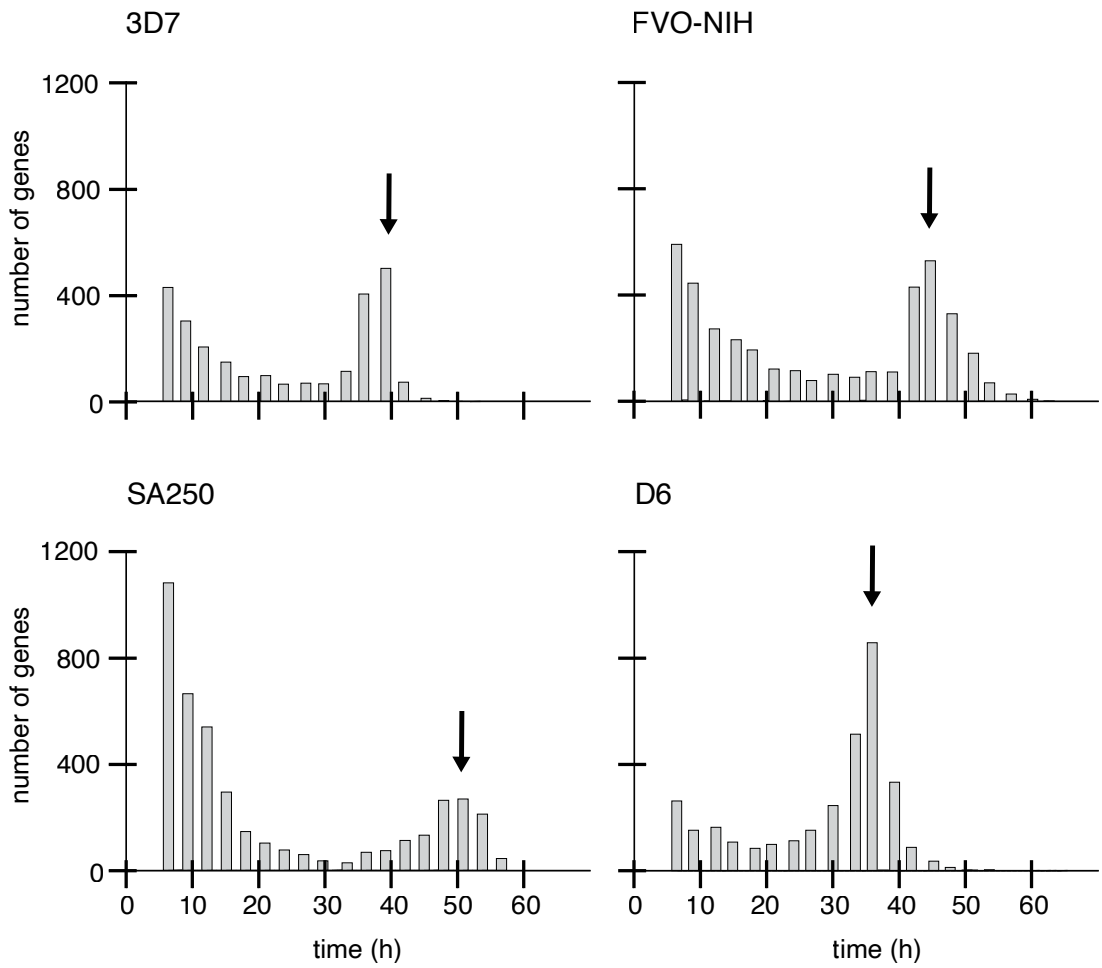

Fig. S10. Distribution of distance in hours between minima of genes for each strain (trough-trough distance.) Using expression data PCHIP splined to one-hour intervals, the distances between the two lowest minima (gene expression peaks) were calculated. The most frequent (locally modal) distance is shown by arrows. Values are found in Table S2.

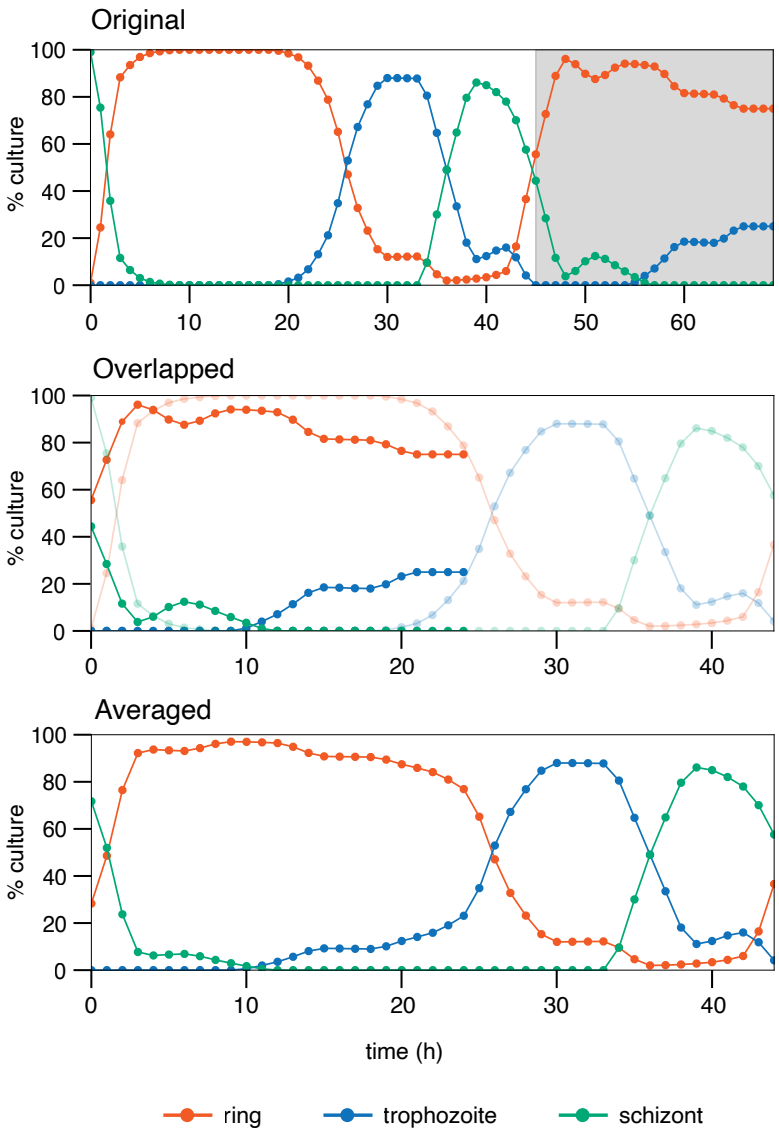

Fig. S11. Wrapping procedure used to re-wrap microscopic data for FVO-NIH to the consensus period length (Table S2), after a PCHIP spline to 1-hour sampling. Shown here are the original microscopy data (with grey shading representing the data wrapped over the cycle), a visualization of the overlap, and the final wrapped data averaged together. See Table S4 for RMSE values.

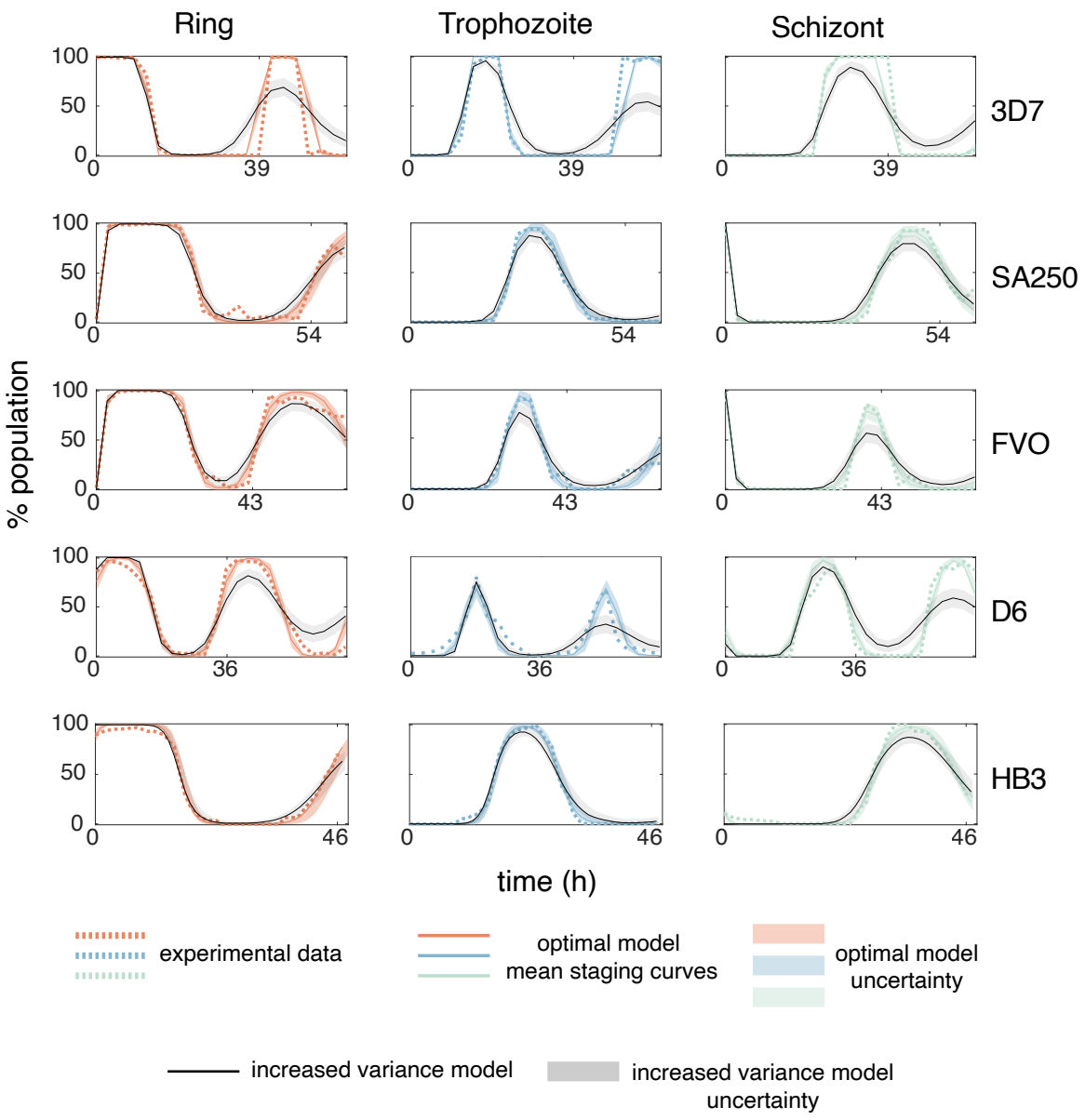

Fig. S12. Microscopic data are well captured within two standard deviations of the mean staging curves in a simple phase-oscillator model; standard deviations were estimated by repeated sampling from a simulated large parasite population with optimal parameters (Table S6). Also shown are the mean and two standard deviations of repeated samples from simulated parasite populations modelled with best fit parameters under the assumption of a relatively large coefficient of variation (CoV) in period lengths of 14% of the mean period length (Table S7).

The experimental staging curves fall notably outside the expected ranges of the model with this larger CoV in period lengths, especially after the first cycle when the loss of synchrony due to period length variation becomes more apparent. All simulations were conducted assuming the best estimate of strain-specific mean periods via microscopy: 39, 54, 43, 36 and 46 hours for 3D7, SA250, FVO, D6, and HB3, respectively.

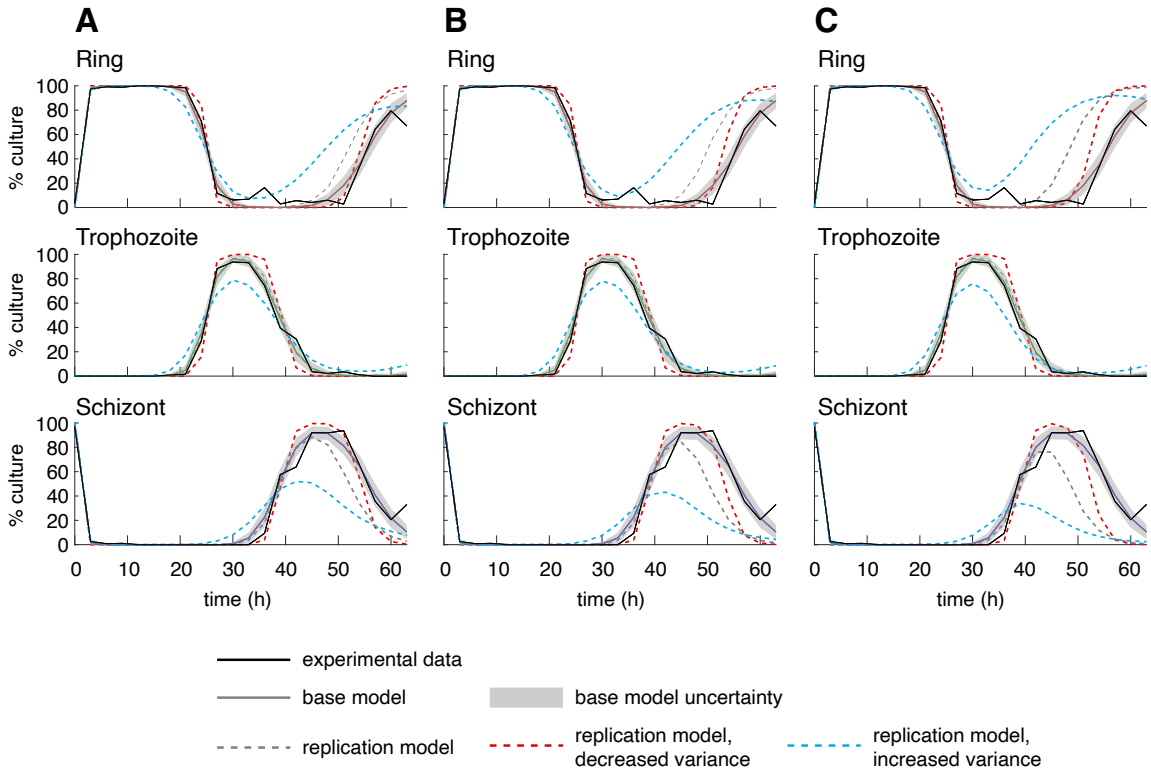

Fig. S13. Simulated staging curves produced by a phase-oscillator model, with and without replication, fit to the SA250 strain experimental data. For the non-replicative model, the estimated uncertainties are given by two standard deviations around the mean staging curves and were estimated by repeated sampling of a small number of parasites from a large parasite population. The replicative model incorporates (A) 4x (B) 8x (C) 16x replication occurring at the schizont-ring transition phase for different choices of population progression variability. All simulations were carried out with the same best-fit model parameters (Table S6) except, for the simulations with replication, parasite progression variability was chosen to be either one half, equal to, or twice the best-fit choice of parasite progression variability. Under the replication model, an increase in period variability reduces the model's ability to fit the experimental data. The data is better fit assuming smaller variance in the population periods. The same behavior is observed in the models of all other strains.

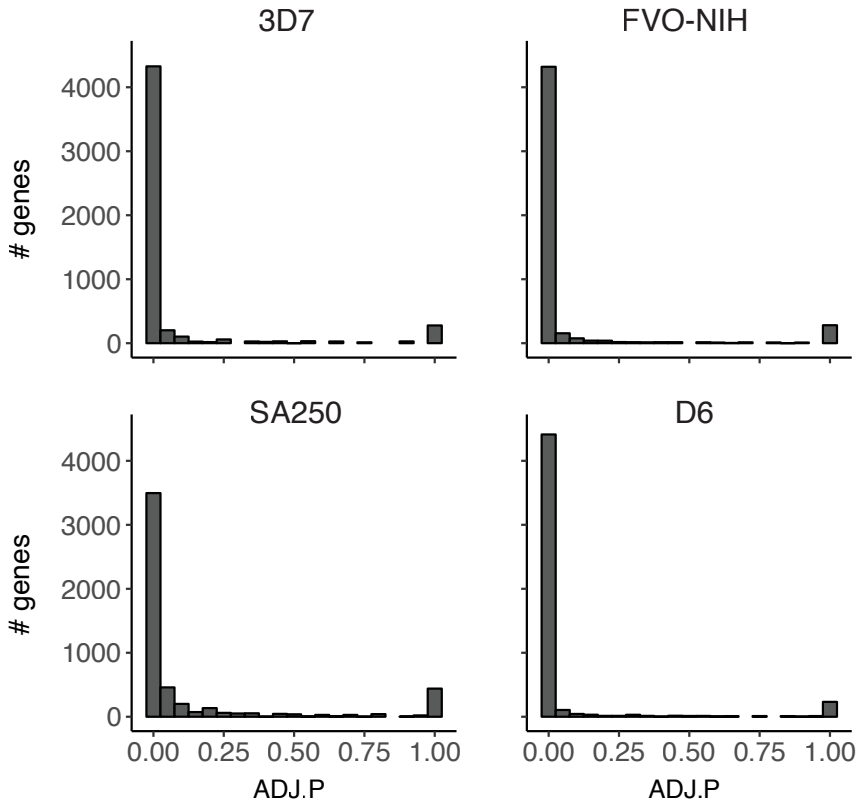

Fig. S14. The distribution of ADJ.P values from JTK_CYCLE analysis of each strain. All strains are dramatically distributed towards zero, indicating that the majority of genes are considered periodic. Data are taken from the initial runs of JTK_CYCLE on each strain, which included a 47–50h search parameter for SA250.

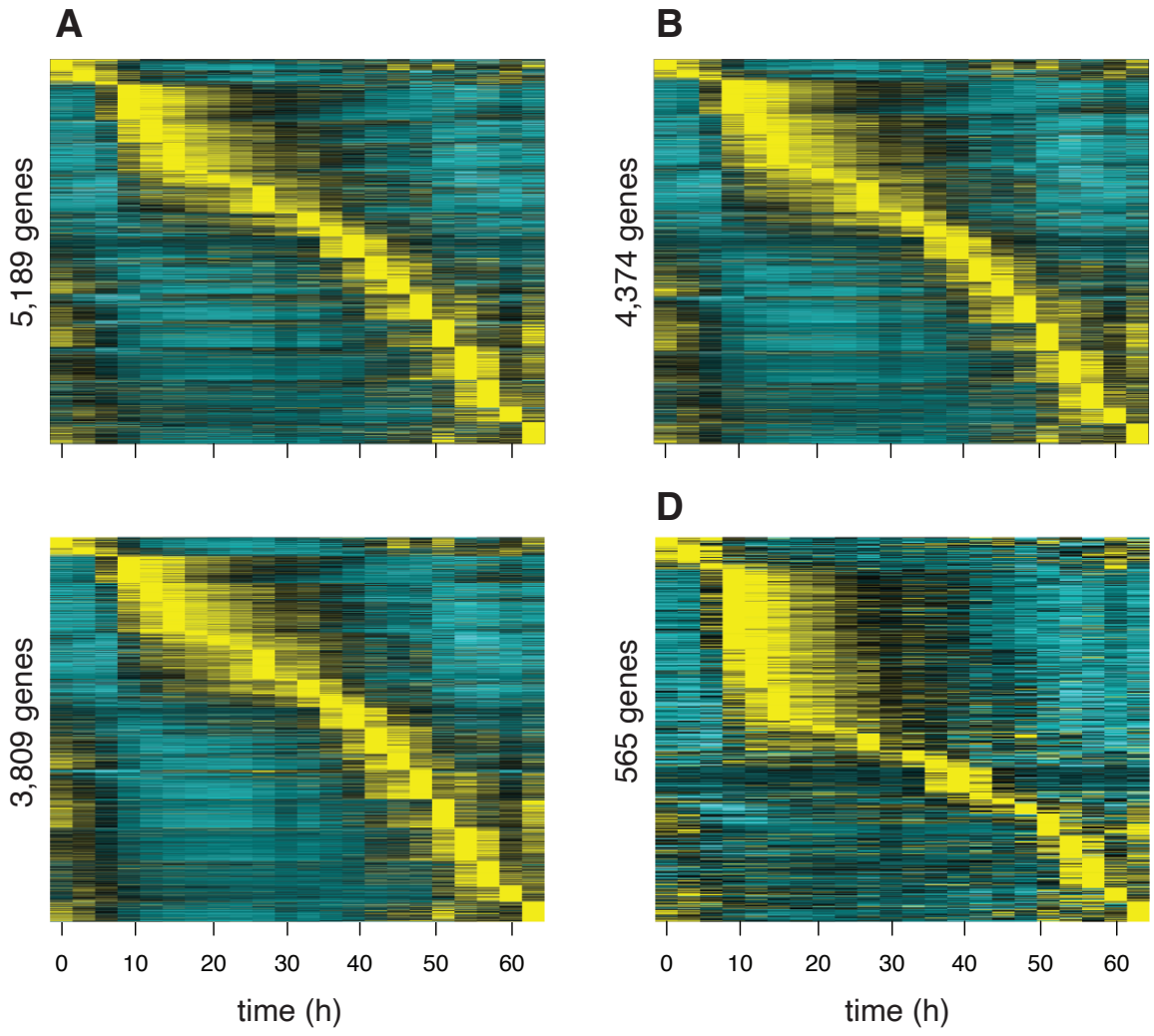

Fig. S15. A p-value cutoff of <0.25 (using JTK_CYCLE's ADJ.P value) retains periodic genes while not introducing excessive non-periodic genes. We used SA250 at the initial JTK_CYCLE search parameters of 47–50h for this visualization to determine the threshold, as SA250 skews the least periodic of the strains (Fig. S14). All gene counts are shown on the y-axis. (A) The set of all mapped genes, with no periodicity filter applied. (B) All genes p <0.25. (C) All genes p <0.05. (D) The genes falling between 0.05> p <0.25. They are predominantly periodic in nature and we judged them worthy of inclusion. This informed our decision to use p <0.25 as the study-wide periodicity p-value threshold.

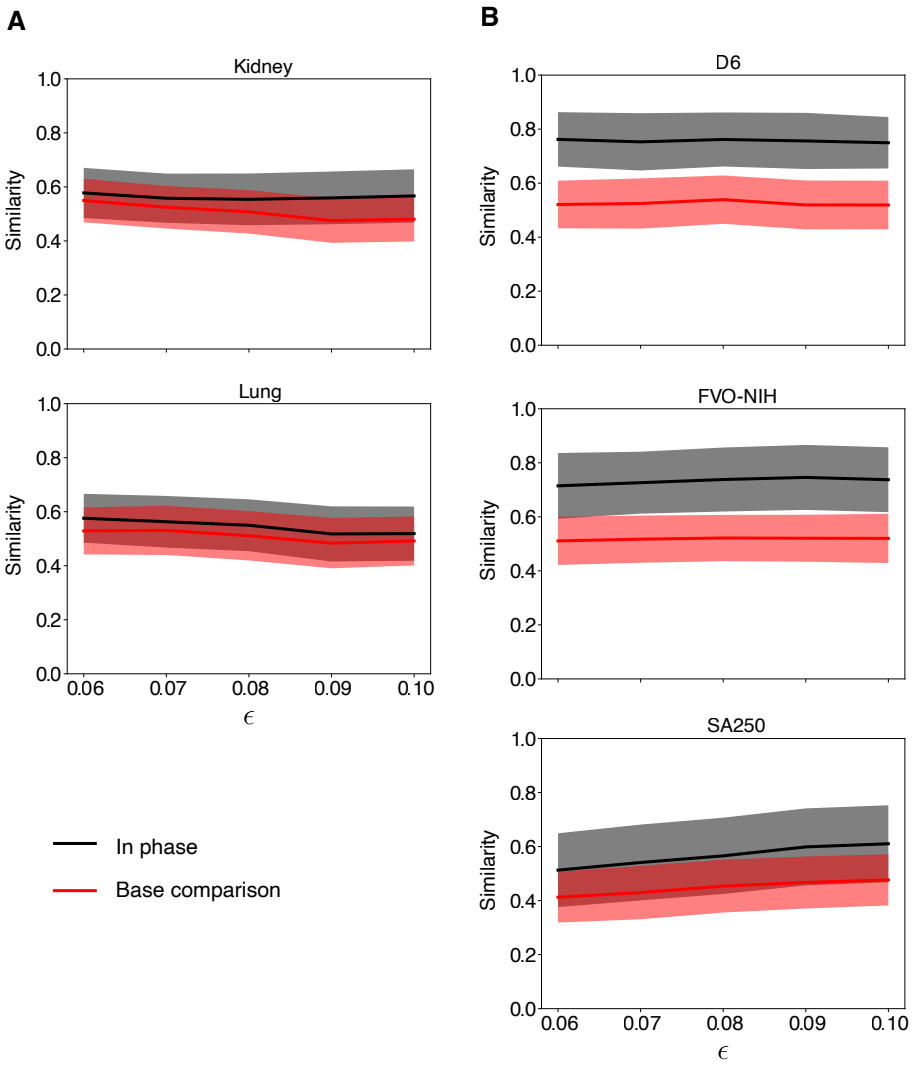

Fig. S16. Mean similarities between in phase and baseline comparisons between (A) kidney or lung gene expression data and the reference tissue liver, and (B) D6, FVO, SA250, and the reference strain 3D7, as a function of noise levels ε. Base comparisons are between (A) liver or (B) 3D7 and a permuted, phase-shifted time series of the same. Grey and red areas indicate ± 1 standard deviation. Summary statistics are found in Table S3.

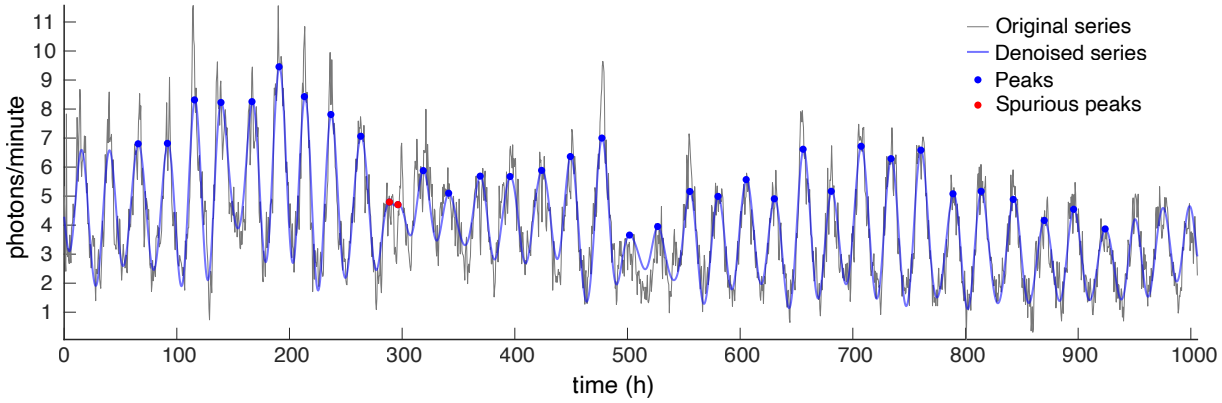

Fig. S17. Original time series of PER2::LUC bioluminescence in fibroblast cell number 34 (*36*), together with the denoised time series, identified peaks, and spurious peaks that were omitted from consideration when calculating peak-to-peak intervals.

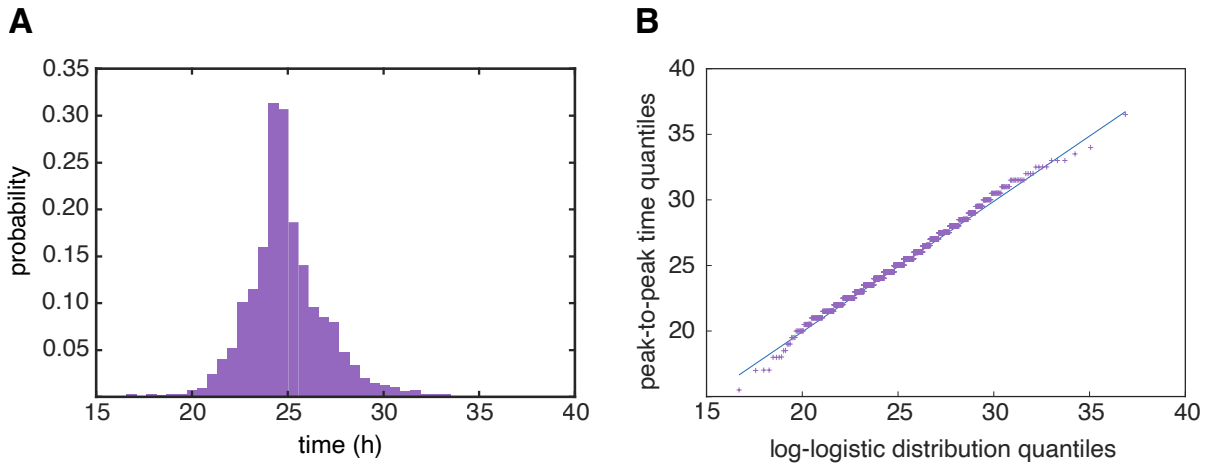

Fig. S18. (A) Histogram of circadian cycle periods determined from approximately 2300 peak-to-peak intervals in PER2::LUC bioluminescence data (*36*). (B) Q-Q plot of the quantiles of the circadian cycle periods against the quantiles of a log-logistic distribution with shape and scale parameters $\boldsymbol{\sigma}$=0.04603 and $\boldsymbol{\mu}$=3.2115, determined by maximum likelihood estimation.

Tables S1 – S8

	all genes	p < 0.01	p < 0.05	p < 0.1	p < 0.25
3D7	5188	81% (1.1%)	86% (5.4%)	88% (9.3%)	91% (25%)
FVO-NIH	5051	82% (1.2%)	88% (5.6%)	90% (10%)	92% (26%)
SA250*	5189*	58%*	73%*	79%*	84%*
SA250	5189	69% (1.3%)	79% (6.1%)	83% (11%)	87% (28%)
D6	4989	86% (1.1%)	90% (5.3%)	91% (9.8%)	93% (26%)
union	5457	5084	5228	5269	5327
intersection	86%	54%	62%	66%	70%

Table S1. Results using differing JTK_CYCLE ADJ.P value filters per strain. The initial set of unfiltered genes ("all genes") is shown, and for each cutoff the percent of the total unfiltered set is given, with the false discovery rate in parentheses. The number of genes in the union and intersection of all strains at each cutoff is also shown. A p-value of <0.25 was chosen for its high retention of genes while maintaining a periodic characteristic (Fig. S15) instead of the more conventional <0.05. *Initial JTK_CYCLE run using a 47–50h search parameter, used in the initial establishment of the significance threshold, and not included in calculation of the total union or intersection.

Strain	Transcriptomic wrap (h)	Microscopic wrap (h)	Peak-peak distance (h)	Trough-trough distance (h)	Final estimate (h)
3D7	40	39	36	39	39
FVO-NIH	47	43	42	45	45
SA250	61	54	51	51	54
D6	37	36	36	36	36

Table S2. *P. falciparum* shows strain-intrinsic period lengths. We used several approaches to estimate period length: an error-minimizing wrap on the bulk periodic transcriptome data (Fig. S7), an error-minimizing wrap of microscopic staging data (Fig. S8), and the mode of the distribution of distances between maxima (peak-peak distance; Fig. S9) and minima (trough-trough distance; Fig. S10) of each individual periodic gene. These values were weighted 3:2:1:1, respectively, before averaging to arrive at the final estimate. The exception was SA250; given the limited cycle length of the time series, all values were weighted equally for averaging.

	Mean (std)	Mean (std) baseline	Mean (std) of difference	α
Lung	0.545 (0.100)	0.509 (0.093)	0.036 (0.126)	0.610
Kidney	0.563 (0.096)	0.508 (0.086)	0.056 (0.118)	0.678
D6	0.756 (0.101)	0.525 (0.091)	0.232 (0.136)	0.948
FVO	0.732 (0.119)	0.518 (0.088)	0.214 (0.150)	0.917
SA250	0.566 (0.145)	0.448 (0.099)	0.118 (0.172)	0.747

Table S3. Mean and standard deviations of similarity across all samples and all ε for in-phase genes (first column) and baseline (second column). The mean and standard deviation of the difference between in-phase similarity and baseline similarity across all samples and all ε is shown in the third column. Also shown is the proportion of samples for which the in-phase similarity score was higher than the corresponding baseline calculation, averaged across ε (α).

Strain	Best fit microscopic wrap		Best fit expression wrap		Consensus period	Microscopic re-wrap to consensus	Expression re-wrap to consensus
	Period (h)	RMSE	Period (h)	RMSE		RMSE	RMSE
3D7	39	0.53	40	13.25	39	-	13.26
FVO-NIH	43	19.36	47*	12.88	45	28.22	13.10
SA250	54	36.40	61	17.67	54	-	23.95
D6	36	17.53	37	10.96	36	-	11.24

Table S4. Microscopic and expression data was interpolated using PCHIP to 1-hour increments, and the best-fit microscopic and expression wraps were found using the procedure described (Methods and Materials.) Root Mean Square errors (RMSE) scores for both these wraps are displayed. Once a consensus period length was decided upon (Methods and Materials, Table S2), microscopic and expression data were re-wrapped for all other analyses. RMSE scores for these re-wraps are displayed. *Since 45–49 hour RMSE scores were locally plateaued (Fig. S7), we used 47 hours as it is the center value.

Strain	Period length	Ring		Trophozoite		Schizont	
D6	36	18h	50.0%	4h	11.1%	14h	38.9%
3D7	39	12h	30.8%	11h	28.2%	16h	41.0%
FVO-NIH	45	24h	53.3%	11h	24.4%	10h	22.2%
SA250	54	12h	44.4%	13h	24.1%	17h	31.5%

Table S5. Variation in cycle length in *P. falciparum* strains is not due to change in any individual intraerythrocytic cycle stage. Interpolated, wrapped staging data (Figs. S3, S8, S11) was used to determine the percent of cycle spent in each phase. Strains are listed by increasing period length. There does not appear to be any consistent pattern of stage lengthening or shortening to achieve the range of period lengths.

			Period Variability		Initial Synchrony Variability		Stage Transition Phases		
Strain	**Mean Period**	**CoV**	$\boldsymbol{\sigma}^*$	$\boldsymbol{\mu}^*$	$\boldsymbol{\sigma_0}^*$	$\boldsymbol{\mu_0}^*$	$\boldsymbol{\theta}^*_{R,T}$	$\boldsymbol{\theta}^*_{T,S}$	**Model Error**
circadian	24.93	8.45E-02	4.64E-02	-	-	-	-	-	-
3D7	38	3.65E-02	2.01E-02	6.63E-04	7.69E-07	7.03E-06	3.23E-01	6.19E-01	1.59E+00
	39	2.76E-02	1.52E-02	3.80E-04	8.37E-05	1.95E-04	3.12E-01	6.02E-01	1.16E+00
	40	4.20E-02	2.31E-02	8.81E-04	1.96E-06	5.89E-05	3.05E-01	5.81E-01	1.30E+00
SA250	53	9.95E-02	5.45E-02	4.89E-03	7.52E-04	9.51E-01	4.24E-01	6.91E-01	2.81E-01
	54	1.02E-01	5.58E-02	5.12E-03	2.76E-05	9.54E-01	4.19E-01	6.80E-01	2.41E-01
	55	1.01E-01	5.54E-02	5.06E-03	3.19E-04	9.71E-01	4.27E-01	6.83E-01	2.43E-01
FVO	42	8.53E-02	4.68E-02	3.61E-03	5.63E-03	9.38E-01	5.73E-01	7.99E-01	3.68E-01
	43	8.17E-02	4.49E-02	3.32E-03	2.03E-02	9.56E-01	5.71E-01	7.97E-01	3.07E-01
	44	8.53E-02	4.68E-02	3.61E-03	1.89E-03	9.79E-01	5.76E-01	8.02E-01	2.98E-01
D6	35	1.16E-03	6.41E-04	6.76E-07	7.53E-02	2.16E-02	4.71E-01	6.30E-01	8.68E-01
	36	2.27E-03	1.25E-03	2.59E-06	7.50E-02	5.33E-02	4.78E-01	6.29E-01	6.47E-01
	37	1.18E-02	6.53E-03	7.02E-05	7.83E-02	7.94E-02	4.77E-01	6.26E-01	6.96E-01
HB3	45	9.13E-02	5.01E-02	4.13E-03	1.94E-02	8.05E-05	3.83E-01	6.57E-01	1.94E-01
	46	8.27E-02	4.54E-02	3.39E-03	3.18E-02	1.57E-02	3.90E-01	6.57E-01	1.68E-01
	47	7.89E-02	4.33E-02	3.09E-03	3.59E-02	3.63E-02	4.02E-01	6.64E-01	1.78E-01

Table S6. Model-determined period CoVs, optimal parameters, and model errors for each of five *P. falciparum* strains and three choices of mean IDC period lengths. HB3 data was extracted from Figure 1 of Bozdech et al. (*23*). For each strain, the best estimate of the mean IDC period length (Table S2) is used along with +/- one hour. The values $\boldsymbol{\sigma}^*$and $\boldsymbol{\mu}^*$are respectively the optimal choices of shape and scale parameters for the log-logistic distribution of periods of phase oscillators, $\boldsymbol{\sigma_0}^*$ and $\boldsymbol{\mu_0}^*$ are respectively the optimal choices of shape and scale parameters of the wrapped-normal distribution of initial oscillator phases, and $\boldsymbol{\theta}^*_{R,T}$ and $\boldsymbol{\theta}^*_{T,S}$ are the optimal chocies of ring-to-trophozoite and trophozoite-to-schizont transition phases. For comparison, the first row provides the relevant values for circadian cycles: the empirical CoV and mean period determined by peak-to-peak intervals of fibroblast cells (*36*), and the maximum likelihood

estimation of the shape and scale parameters for a log-logistic distribution that was fit to the period-normalized peak-to-peak interval data.

			Period Variability		Initial Synchrony Variability		Stage Transition Phases		
Strain (Mean Period)	Min. σ	σ^*	$\Delta\sigma^*$	$\Delta\mu^*$	$\Delta\sigma_0{}^*$	$\Delta\mu_0{}^*$	$\Delta\theta^*_{R,T}$	$\Delta\theta^*_{T,S}$	ΔModel Error
3D7 (39)	5.80E-02	5.80E-02	281%	1356%	-76%	-70%	3%	1%	93%
	6.96E-02	6.96E-02	358%	1998%	-100%	-57%	3%	2%	134%
	8.12E-02	8.12E-02	434%	2758%	-94%	127%	4%	3%	178%
SA250 (54)	5.80E-02	5.80E-02	4%	8%	215%	0%	0%	0%	1%
	6.96E-02	6.96E-02	25%	56%	819%	0%	0%	0%	24%
	8.12E-02	8.12E-02	46%	112%	2942%	0%	1%	0%	75%
FVO (43)	5.80E-02	5.80E-02	29%	67%	-21%	0%	-1%	0%	16%
	6.96E-02	6.96E-02	55%	141%	-36%	-1%	-1%	0%	56%
	8.12E-02	8.12E-02	81%	228%	-67%	-1%	-2%	-1%	117%
D6 (36)	5.80E-02	5.80E-02	4526%	214135%	-65%	-39%	-3%	-2%	90%
	6.96E-02	6.96E-02	5451%	308479%	-70%	-54%	-5%	-2%	138%
	8.12E-02	8.12E-02	6375%	420125%	-82%	-71%	-6%	-3%	193%
HB3 (46)	5.80E-02	5.80E-02	28%	63%	-39%	-39%	-1%	-1%	24%
	6.96E-02	6.96E-02	53%	135%	-100%	-69%	-2%	-1%	94%
	8.12E-02	8.12E-02	79%	220%	-100%	-98%	-3%	-2%	231%

Table S7. Changes in model-determined optimal parameters, and model errors for each of five *P. falciparum* strains across three choices of minimum allowable shape parameter (Min. $\boldsymbol{\sigma}$) for the log-logistic distribution of the phase oscillator periods compared to the model which allowed an unrestricted range of values for $\boldsymbol{\sigma}$ (Table S6). The lower bounds on $\boldsymbol{\sigma}$ were chosen to be 1.25, 1.50, and 1.75 times the optimal shape parameter of a log-logistic distribution determined by maximum liklihood estimation of the peak-to-peak intervals of fibroblast cells *(36)*. The values

$\boldsymbol{\Delta\sigma}^*$ and $\boldsymbol{\Delta\mu}^*$ are respectively the percent change in the optimal choices of shape and scale parameters for the log-logistic distribution of periods of the phase oscillators, $\boldsymbol{\Delta\sigma_0}^*$ and $\boldsymbol{\Delta\mu_0}^*$ are respectively the percent change in the optimal choices of shape and scale parameters of the wrapped-normal distribution of initial oscillator phases, and $\boldsymbol{\Delta\theta^*_{R,T}}$ and $\boldsymbol{\Delta\theta^*_{T,S}}$ are the percent change in the optimal choices of ring-to-trophozoite and trophozoite-to-schizont transition phases.

Strain	Avg. input reads	Avg. uniquely mapped	Avg. multiply mapped	Avg. unmapped (too short)
3D7	37,781,603	68.2%	19.5%	12.3%
FVO-NIH	32,564,491	48.4%	34.6%	16.9%
SA250	36,030,416	48.7%	35.2%	16.0%
D6	33,837,883	56.1%	40.6%	3.3%

Table S8. Alignment statistics from STAR output. All strains were mapped to the 3D7 reference genome. Each statistic represents the average across all time point samples.

Data S1 – Gene expression values per strains. Excel spreadsheet with RNA-seq expression time series for *P. falciparum* strains 3D7, FVO-NIH, SA250, and D6. Each strain is in a separate tab. Time points are in hours.

Data S2 – Lists of genes used in each figure. Excel spreadsheet with a separate tab for each main text figure.

Data S3 – Microscopy metadata for each strain. Excel spreadsheet with a separate tab for each strain. Contains parasitemia data, dates of experiments, and qualitative observations.

Data S4 – Microscopy quantitative values for each strain. Single-tabbed Excel spreadsheet with microscopy data that was used for analyses in this paper. Early and multiply invaded ring and trophozoites were grouped with ring and trophozoite counts; segmentor and multiply-invaded schizonts were grouped with schizont counts. Counts are displayed as percent of culture at each time point. Time points are in hours. Sample IDs match those in Data S3.

References

52. J. D. Haynes, J. K. Moch, Automated synchronization of Plasmodium falciparum parasites by culture in a temperature-cycling incubator. *Methods Mol. Med.* **72**, 489–497 (2002). doi:10.1385/1-59259-271-6:489 Medline

53. A. R. Leman, S. L. Bristow, S. B. Haase, Analyzing transcription dynamics during the budding yeast cell cycle. *Methods Mol. Biol.* **1170**, 295–312 (2014). doi:10.1007/978-1-4939-0888-2_14 Medline

54. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013). doi:10.1093/bioinformatics/bts635 Medline

55. K. Sorber, M. T. Dimon, J. L. DeRisi, RNA-Seq analysis of splicing in Plasmodium falciparum uncovers new splice junctions, alternative splicing and splicing of antisense transcripts. *Nucleic Acids Res.* **39**, 3820–3835 (2011). doi:10.1093/nar/gkq1223 Medline

56. C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, L. Pachter, Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31**, 46–53 (2013). doi:10.1038/nbt.2450 Medline

57. R. Nerem, P. Crawford-Kahrl, B. Cummins, T. Gedeon, A poset metric from the directed maximum common edge subgraph. arXiv:1910.14638 [cs.DS] (31 October 2019).

58. I. Daubechies, *Ten Lectures on Wavelets* (CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, 1992).