Methods

Data pre-processing

The bulk RNA-seq time course raw count data were normalized using MedianNorm function in EBSeq, where each time point was assigned its own condition. Lowly expressed genes were then filtered out. Specifically, 3881 genes whose maximum expression among all time points less than 10 were removed. This gives a normalized and filtered expression matrix of 15218 genes and 96 time points. A bandpass smoother was then applied to each gene to remove redundant variations. It uses sliding windows such that the expression of each time point is subtracted by the average within the window. The two parameters for the window sizes were set as 10 and 9 time points.

Shifted-correlation analysis

We used 4 target genes (HES7, NanoLuc, tdTomato and LFNG) as gold-standard oscillation genes. Since the RNA-seq data is noisier in the beginning and end of the time course epxeriment, we decided to only use time points from 100min to 800min for the gold standard genes. We then applied shifted-correlation analysis to identify novel genes oscillating together with gold-standard genes but with phase shift. In particular, we considered all possible phase shifts from -100min to +100min. Since the estimated oscillation period is around 400 minutes from prior study, and by identifying both positive correlated and negative correlated genes, it is enough to capture all possible shifts within the 400 minutes period.

For a given gene g and a shift level k (k=-100, -87.5, -75, ..., 100), we calculate the Pearson correlation between the expression of g in time interval [100+k, 800+k] and the expression of four gold-standard genes in time interval [100, 800]. The k giving the largest absolute value of correlation is considered as the best phase shift for gene g, and the corresponding correlation is considered as the best correlation between gene g and gold standard genes. We selected top correlated genes bla bla...