Supplementary material for Zhao *et al.* (2001) *Proc. Natl. Acad. Sci. USA* **98** (10), 5631–5636. (10.1073/pnas.101013198)

Appendix: Single Pulse Model (SPM) and Estimation

The SPM can be developed in several steps. The first step models a single transcript in a single cell across cell cycles as a binary process:

$$Y(t) = \begin{cases} 1 & t \in [\varsigma + c\Theta, \xi + c\Theta), \text{ some c } \in \{0,1,2,...\} \\ 0 & \text{otherwise} \end{cases},$$

where $Y(t)$ denotes expression level at time t, (ς, ξ) with $(0 \le \varsigma < \xi \le \Theta)$ denote activation and deactivation times, Θ is the cell cycle span, c = 0,1,2,… denotes 1st, 2nd, 3rd, … , cell cycle. Alternatively, the above display may be written as

$$Y(t) = \sum_{c \ge 0} I\{\varsigma + c\Theta < t \le \xi + c\Theta\}$$

with the summation over 1st, 2nd, 3rd, … , cycle, and $I\{\bullet\}$ is an identity function.

The second step considers multiple transcripts within a single cell, giving an expression pulse for the cell having background and elevated expression levels $(\tilde{\alpha}, \tilde{\alpha} + \tilde{\beta})$ and activation and deactivation times (ς, ξ) (Fig.1). The model for the expected expression level for the cell may be written as

$$\tilde{\alpha} + \tilde{\beta} \sum_{c \ge 0} I\{\varsigma + c\Theta < t \le \xi + c\Theta\}.$$

A third step acknowledges the fact that multiple cells are pooled and are synchronized, but that the synchronization is not perfect. Let t_k denote the targeted timing. The actual timing of single cells, T_k, is randomly distributed around t_k, and is assumed to have a normal

distribution with mean t_k and standard deviation σ. Notationally, let $Y(t) = \sum_{i=1}^{N} Y_i^*(t + T_i)$, where N is the number of cells in the synchrony, $(t + T_i)$ is the age (timing) of the ith cell, and Y_i^* is the expression level of a particular gene in the ith cell. Modeling the mean expression level Y_i by SPM gives the expected values of $Y_i^*(t + T_i)$ as

$$\tilde{\alpha} + \tilde{\beta} \sum_{c \geq 0} I\{\varsigma + c\Theta < t + T_i \leq \xi + c\Theta\}.$$

The mean expression for the synchrony then arises from summation over the N cells and taking the expectation over the random timing (T_i). Following some simple algebra, we can show that the mean expression level at time t_k can then be written as:

$$\tilde{\alpha} + \tilde{\beta}\left[\sum_{c \geq 0} \phi\left(\frac{\xi + c\Theta - t_k}{\sigma}\right) - \phi\left(\frac{\varsigma + c\Theta - t_k}{\sigma}\right)\right]$$

where $\phi(x)$ is the Gaussian cumulative distribution function and $\alpha = N\tilde{\alpha}$ and $\beta = N\tilde{\beta}$.

A fourth step acknowledges that the synchronization deteriorates over time, an inherent limitation with all synchronization protocols. We model this deterioration by allowing σ to monotonically increase with time t. Specifically, we assume that the standard deviation for the timing of cells in sample k follows an exponential form model:

$$\sigma_k = \exp(\gamma_0 + \gamma_1 t_k),$$

where (γ_0, γ_1) are parameters to be estimated.

A fifth step incorporates multiplicative (λ_k) and additive (δ_k) heterogeneity factors between samples. Variations in mRNA extraction, amplification and assessment may result in

heterogeneity between samples. As mentioned previously, the desire to accommodate such heterogeneities leads to the following model for the mean expression level,

$$\mu(t_k) = \delta_k + \lambda_k\{\alpha + \beta\left[\sum_{c\geq 0} \Phi(\frac{\xi + c\Theta - t_k}{\sigma_k}) - \Phi(\frac{\varsigma + c\Theta - t_k}{\sigma_k})\right]\}$$

in which δ_k and λ_k are specific to the *k*th sample and the δ_k's and λ_k's average to zero and *1*, respectively, over the *K* samples. As written, the model is applicable to measurements of the abundance of transcripts directly. To analyze ratios of transcript levels we choose to eliminate the multiplicative heterogeneity factors ($\lambda_k \equiv 1$).

Each gene is allowed to have its own activation and deactivation time and its own background and elevated expression level, giving the SPM model for the mean expression for the *j*th gene as

$$\mu_j(t_k) = \delta_k + \lambda_k\{\alpha_j + \beta_j\left[\sum_{c\geq 0} \Phi(\frac{\xi_j + c\Theta - t_k}{\sigma_k}) - \Phi(\frac{\varsigma_j + c\Theta - t_k}{\sigma_k})\right]\},$$

where $j = 1, 2, \ldots, J$ and $k = 1, 2, \ldots, K$ denote all *J* genes in all *K* samples.

To find parameter estimates that solve the estimating Eq. **1** we can minimize the weighted sum of squares,

$$\sum_{j=1}^{J} \hat{v}_j^{-2}\left(\sum_{k=1}^{K} [Y_{jk} - \mu_j(t_k)]^2\right). \qquad \textbf{[A1]}$$

Because the mean activation and deactivation times represent changing points and are restricted ($\varsigma_j \geq 0, \xi_j \geq 0$ and $\xi_j > \varsigma_j$), we minimize the above sum of squares (Eq. **A1**) with respect to the other parameters at each point on fine grid values for (ς_j, ξ_j), and select the set of parameters estimates giving an overall minimum for Eq. **A1**. We restricted the profiling

procedure to points such that (ς_j, ξ_j) included at least two t_k values. The weight function in the calculation was defined as

$$\hat{v}_j^2 = \tfrac{1}{K} \sum_{k=1}^{K} [Y_{jk} - \hat{\mu}_j^0(t_k)]^2,$$

where $\hat{\mu}_j^0(t_k) = \hat{\delta}_k + \hat{\lambda}_k \hat{\alpha}_j$ denotes the estimated value of $\mu_j(t_k)$ upon requiring $\beta_j = 0$. Note also that upon estimating all model parameters,

$$R_j^2 = 1 - \tfrac{1}{K} \hat{v}_j^{-2} \sum_{k=1}^{K} [Y_{jk} - \hat{\mu}_j(t_k)]^2,$$

is simply the percentage of the variation in expression levels for gene *j*, after heterogeneity parameter adjustment, that is explained by the cycle aspects of the SPM model. Hence an R_j^2 value close to unity implies that the SPM is providing a good representation of the observed expression profile for the *j*th gene.

As mentioned in *Methods*, we carried out the parameter estimation in multiple stages to simplify calculations. The first stage led to estimates of $(\hat{\delta}_k, \hat{\lambda}_k)$, *k = 1,…,K*, by minimizing Eq. **A1** with all β_j values restricted to be zero. Under this restriction we also have

$$\hat{\alpha}_j = K^{-1} \sum_{k=1}^{K} (Y_{jk} - \hat{\delta}_k) / \hat{\lambda}_k,$$

so that $\hat{\mu}_j^0(t_k)$ values and weights $\hat{v}_j^2$ can be calculated. Next the cell cycle span estimate, $\hat{\Theta}$, was calculated by minimizing Eq. **A1** under a SPM. Because most of the transcripts are not cell cycle regulated, we used only a set of 104 known periodic transcripts to ensure an appropriate estimate of the cell cycle span. The calculation involves profiling over the cell

cycle span Θ, for example, for the *cdc28* data set, from 40 to 80 min in units of 1 min. On the same set of genes, we estimated the synchronization variability, σ_k, by minimizing Eq. **A1**.

Upon fixing these parameters the minimization of Eq. **A1** with respect to the parameters $(\varsigma_j, \xi_j, \alpha_j, \beta_j)$ for the *j*th gene simply requires the minimization of

$$\sum_{k=1}^{K} [Y_{jk} - \mu_j(t_j)]^2,$$

separately for $j = 1, \ldots, J$, much simplifying the calculation. Also estimated standard deviations for these parameter estimates arise from applying the sandwich formula (1) to data for the *j*th gene alone under the model assumption and the independence assumption of Y_k given x_k. These calculations give statistics Z_j, the ratio of $\hat{\beta}_j$ to its standard deviation, that will have an approximate standard normal distribution if $\beta_j = 0$, for each $j = 1, \ldots, J$. Under such a standard normal distribution the probability that Z_j exceeds 5 in absolute values is about 5.7×10^{-7}, so that the probability that any one of the $\hat{\beta}_j$ values, say 6,000 genes, exceeds 5 if all β_j values equal zero is conservatively estimated, using a Bonferroni approximation, as $6{,}000 \times 5.7 \times 10^{-7} = 0.003$. This suggests that our threshold of 5 may be too extreme, especially because the Bonferroni correction is conservative, but the standard normal distribution approximation for Z_j may be rather liberal, especially if the number of samples (K) is fairly small. Hence we have chosen to retain the rather extreme threshold of 5.

The numerical procedure outlined above ensures that parameter estimates of all model parameters can be obtained under minimal constraints on the data [e.g., heterogeneity corrected values, $(Y_{jk} - \hat{\delta}_k) / \hat{\lambda}_k$ must exhibit some variation across samples]. It would be desirable to have further statistical development to ensure that the multistage estimation procedure has minimal effect on *Z* statistics compared to a procedure that simultaneously estimates all model parameters, and to examine the conservatism associated with asymptotic normal approximation for the distribution of model parameter estimates. In the context of the two group comparison problems and the time-course analyses mentioned in *Methods*, we find that each Z_j value does not depend much on whether heterogeneity and regression parameters were estimated in multistages as in this paper, or jointly (L. P. Z., unpublished observations). Asymptotic normal approximations, however, appeared to be more liberal in the extreme tails than did certain empirical approximations to the Z_j distribution that arise by comparing Z_j values under various permutations of the regression variable among samples.

Reference:

1. Liang, K. Y. & Zeger, S. L. (1986) *Biometrika* **73**, 13-22.