A statistical modeling approach is proposed for use in searching large microarray data sets for genes that have a transcriptional response to a stimulus. The approach is unrestricted with respect to the timing, magnitude or duration of the response, or the overall abundance of the transcript. The statistical model makes an accommodation for systematic heterogeneity in expression levels. Corresponding data analyses provide gene-specific information, and the approach provides a means for evaluating the statistical significance of such information. To illustrate this strategy we have derived a model to depict the profile expected for a periodically transcribed gene and used it to look for budding yeast transcripts that adhere to this profile. Using objective criteria, this method identifies 81% of the known periodic transcripts and 1,088 genes, which show significant periodicity in at least one of the three data sets analyzed. However, only one-quarter of these genes show significant oscillations in at least two data sets and can be classified as periodic with high confidence. The method provides estimates of the mean activation and deactivation times, induced and basal expression levels, and statistical measures of the precision of these estimates for each periodic transcript.

Advances in microarray technologies (1–5) have enabled investigators to explore the dynamics of transcription on a genomewide scale. The current challenge is to extract useful and reliable information out of these large data sets. A common, first approach is cluster analysis. The primary objective of cluster analysis is to group genes that have comparable patterns of variation. This method is valuable for reducing the complexity of large data sets and for identifying predominant patterns within the data (6). However, additional methods are needed to extract information about individual genes from these large data sets.

 Toward these goals we consider a statistical method to identify genes whose transcript profiles respond to a stimulus. In general terms, this approach involves modeling the association of a generic response with a specific experimental variable, for example, timing, cell type, temperature, or drug dosage, using a set of interpretable parameters. One objective is to estimate pertinent parameters for individual transcripts, with the goal of testing specific hypotheses concerning transcription response to the stimulus. If the statistical model provides an adequate representation of the expression data for a specific gene, then the corresponding model parameter estimates can provide certain response characteristics for that gene. For example, model parameters may describe the magnitude, duration, or timing of the response. This modeling strategy can be used for two group comparisons, where the objective may be to identify genes that are differentially expressed between normal and abnormal tissues, or in drug discovery studies, where the objective may be to identify transcripts affected by drug dosage.

To demonstrate the utility of this approach, we formulated a model for identifying periodically transcribed genes of the budding yeast Saccharomyces cerevisiae. In this case, the stimulus is synchronous resumption of the cell cycle by releasing the cells from a fixed arrest point. The response is a pulse of transcription, and the key experimental variable is cell cycle timing (7–9).

Four synchronized cell cycle data sets have been generated and made available for general exploration (8, 9). These large data sets have been analyzed by visual inspection (8), Fourier transform and hierarchical clustering (9), k-means (10) and QT clustering (11), self-organizing maps (12), and singular value decomposition (13, 14). Fourier transform analysis of three data sets, where the threshold for periodicity was based on the behavior of known periodic genes, led to a report that there are 800 periodically transcribed genes (9). Later, k-means clustering was applied to one data set, and five periodic clusters with 524 members were identified (10). However, only 330 genes were identified by both approaches. For comparison, we have used statistical modeling to look for regularly oscillating profiles within these large data sets. This approach complements clustering methods in that rather than seeking to group together genes having similar expression patterns it aims to directly identify transcripts affected by a given stimulus and to provide specific information regarding individual response patterns. As amplified below, the method also allows for heterogeneity in response patterns among samples, with expected robustness of inferences on response parameters to certain types of experimental variations.

Methods

A Modeling Framework. Let $Y_{jk}$ denote the expression level for the $j$th gene in the $k$th sample in a stimulus experiment. The number, $J$, of genes studied often will be of high dimension, typically in the thousands, while the number of samples, $K$, may be comparatively few. A standard statistical approach would relate the mean of the vector response, $Y_k = (Y_{1k}, \ldots, Y_{Jk})$ for the $k$th sample to a corresponding vector of $p$ covariates $x_k = (x_{1k}, \ldots, x_{pk})$ that codes the stimulus categories and possibly other characteristics of the $k$th sample using a regression function, say $\Delta(x_k, \theta)' = (\Delta_1(x_k, \theta), \ldots, \Delta_k(x_k, \theta))$, where $\theta = (\theta_1, \ldots, \theta_J)$ may include gene-specific and other parameters and is to be estimated. Under such a regression model the elements of the vector of differences $Y_k - \Delta(x_k, \theta)$ have mean zero, but can be expected to be correlated due, for example, to variations in mRNA extraction, amplification, and assessment among samples. Such variations can be accommodated by introducing additional parameters, which we refer to as heterogeneity parameters into the model for the mean of $Y_k$. In fact, for sample $k$, one can introduce both an additive heterogeneity parameter $\delta_k$ and a multiplicative heterogeneity parameter $\lambda_k$ giving a model, $\delta_k + \lambda_k \Delta(x_k, \theta)$ for the expectation of $Y_k$. The average of the $\delta_k$ and $\lambda_k$ are restricted to be zero and one, respectively, to avoid possible identifiability problems relative to the regression parameters $\theta$ of primary interest. The high dimension of $Y_k$ will allow those heterogeneity parameters to be precisely estimated in many applications. The inclusion of these parameters may

Abbreviation: SPM, single-pulse model.

To whom reprint requests should be addressed. E-mail: lbreeden@fhcrc.org or lzhao@fhcrc.org.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.
make plausible an assumption that \( Y_k \) given \( x_k \) are nearly independent, especially for in vitro experiments. Under such an assumption, one can simplify the modeling and numerical procedure for the estimation of \( \theta \).

Following the approach described in the seminal statistical paper by Liang and Zeger (15), estimation of the mean parameter vector \( \eta = (\delta_1, \ldots, \delta_K, \lambda_1, \ldots, \lambda_K, \theta) \) can proceed by specifying a “working” covariance matrix for \( Y_k \), which under the above independence assumption will be approximated by a diagonal matrix, written as \( V_k = \text{diag}(v^2_1, \ldots, v^2_J) \), so that the expression level for each of the \( J \) genes is allowed to have a distinct variance.

Estimates of the vector of mean parameters \( \eta \) can now be estimated as \( \hat{\eta} = (\hat{\delta}_1, \ldots, \hat{\delta}_K, \hat{\lambda}_1, \ldots, \hat{\lambda}_K, \hat{\theta}) \), a solution to the estimating equation,

\[
\sum_{k=1}^K D_k \hat{V}_k^{-1} [Y_k - \delta_k 1 - \lambda_k \Delta_k(x_k, \theta)] = 0,
\]

where \( D_k \) is the matrix of partial derivatives of the mean of \( Y_k \) with respect to the parameter \( \eta \). \( \hat{V}_k \) denotes \( V_k \) with each \( v^2_j \) replaced by a consistent estimate \( \hat{v}^2_j \), and \( 1 \) denotes a column vector of ones of length \( J \). Under the above modeling assumptions, \( \hat{\eta} \) will be approximately jointly normally distributed provided both \( J \) and \( K \) are large, and the variance of \( \hat{\eta} \) can be consistently estimated (as \( J \) and \( K \) become large) by a standard sandwich formula (15, 16).

The mean parameter estimation procedure just outlined is expected to be useful in various types of microarray data sets. It will allow the estimation of meaningful gene-specific parameters to characterize expression levels in response to a stimulus, and, in that sense, is complementary to cluster analysis that seeks to group genes having similar expression patterns, with less emphasis on pattern characteristics. For example, in the context of comparing the expression patterns between diseased and nondiseased tissues, one may define a binary indicator \( x_k \) that takes value zero for nondiseased tissue samples and one for diseased tissue samples, and specify a regression function, \( \Delta_k(x_k, \theta) = \theta_0 + \theta_1 x_k \), under which the \( j \)th gene would be differentially expressed between normal and abnormal tissues, whenever \( \theta_1 \neq 0 \). The regression variable \( x_k \) also could be expanded to allow the regression function to depend on other measured characteristics of the \( k \)th sample (or the \( k \)th study subject). Similarly, in a study of variations of expression over time one would define \( x_k = t_k \), the timing of the \( k \)th sample to be gathered, and one can choose a linear or other functional form to model the regression function \( \Delta_k(x_k, \theta) \).

In the remainder of this paper we focus on the specific setting of transcript response patterns when cells are released from cell cycle arrest. Our particular interest will be the identification of genes having periodic expression level changes over multiple cell cycles. A key data analysis challenge relates to the possibility of a single nonoscillating pulse of transcription for some genes, after the resumption of cell cycling.

**Modeling Periodic Transcription.** Consider a synchronized experiment to identify mRNAs that are transcribed once per cell cycle. Suppose that when activated, the \( j \)th mRNA reaches an elevated value \( (\alpha_j + \beta_j) \) and when deactivated it falls to a basal expression level \( (\alpha_j) \) (Fig. 1). Naturally, \( \beta_j \) is interpreted as the difference between averaged peak and trough expression levels. Considering multiple copies of the \( j \)th mRNA, transcribed and dissipated at consecutive times in multiple cells with imperfect synchronization, the mean expression level of the \( j \)th transcript at the time \( t_k \) may be modeled as:

\[
\mu_j(t_k) = \delta_k + \lambda_k \left\{ \alpha_j + \beta_j \left[ \sum_{c=0}^\infty \phi_c \left( \frac{e^{\theta + \xi_j - t_k^*}}{\alpha_k} \right) - \phi_c \left( \frac{e^{\theta + \xi_j - t_k^*}}{\alpha_k} \right) \right] \right\},
\]

in which \( j = 1, 2, \ldots, J \), and \( k = 1, 2, \ldots, K \) for all \( J \) transcripts at all \( K \) time points, where \( \xi_k, \xi_j \) are the activation and deactivation times for the \( j \)th gene, respectively, \( t_k^* = t_k + \tau \), where \( \tau \) denotes the difference of actual cell cycle timing and observed timing and is typically known as phase, \( \theta \) is the cell cycle span, and the summation is over multiple cell cycles, \( c = 0, 1, 2, \ldots \). The standard deviation, \( \sigma_k \), describes the variation of “true” cell-specific timings around \( t_k^* \), which we assume to follow a normal distribution with mean \( t_k \), resulting in the cumulative normal distribution function \( \Phi(.) \) in the mean model. Also, \( \delta_k, \lambda_k \) are the additive and multiplicative heterogeneity parameters for the \( k \)th sample as described above, and here \( x_k = t_k \). The above single-pulse model (SPM) specifies a model for the mean expression of each gene as the cell cycle proceeds. Gene-specific
activation and deactivation times as well as the background and elevated expression levels are estimated for each gene. SPM also allows for variation between samples, for the fact that the synchrony is imperfect and, as described below, for the synchrony to deteriorate over time (Fig. 1). Further detail on the development of the SPM is given in the Appendix, which is published as supplemental material on the PNAS web site, www.pnas.org. The resulting mean expression model has been shown visually to reproduce the profiles observed for periodic transcripts measured by conventional means (17).

The SPM described above can be applied by using the mean model estimation procedure outlined above. To simplify numerical aspects we have used a multistage procedure: (i) heterogeneity parameters, \( (\delta_k, \lambda_k), k = 1, \ldots, K \), are estimated by using all genes when the pulse heights are set to zero, (ii) the cell cycle span, \( \Theta \), is estimated by using a group of known cell cycle genes under a pulse model, (iii) the synchronization variability, \( \sigma_k, k = 1, \ldots, K \), is estimated by using the same group of known genes, and (iv) gene-specific parameters \( (\alpha_j, \beta_j, \epsilon_j, \delta_j), j = 1, \ldots, J \) are estimated while other estimated parameters are treated as fixed at their estimated values. Although a simultaneous estimation approach using the estimating Eq. 1 would be preferable, the impact on estimation of the gene-specific parameters of their variance estimates is likely to be minimal as gene-specific parameters are weakly correlated with other parameters.

To test the fit of the SPM we introduced additional polynomial functions of time in the mean model and tested the hypothesis that the polynomial coefficients were identical zero. Specifically, the SPM is augmented and written as

\[
\tilde{\mu}_j(t_k) = \mu_j(t_k) + \gamma_1 z_k^1 + \gamma_2 z_k^2 + \gamma_3 z_k^3,
\]

allowing a departure from the SPM. A score-type test statistic for \( (\gamma_1, \gamma_2, \gamma_3) = (0, 0, 0) \) is then constructed using the asymptotic normal theory described above. This score statistic, \( X^2_{n-3} \), will have an approximate chi-square distribution with three degrees of freedom under the SPM model, for sufficiently large \( J \) and \( K \). We choose 11.3, the 1% upper percentage of this chi-square distribution to identify genes with patterns that depart significantly from the SPM. For the cdc28 data set, for example, only 262 genes give test statistics that exceeded the critical value. Note that other deviations than those polynomial terms could be specified but are not pursued here.

For those genes for which the expression pattern does not depart significantly from SPM, we estimate activation time \( (\xi_j) \), deactivation time \( (\eta_j) \), basal expression level \( (\alpha_j) \), and elevation in expression level during the interval \( (\delta_j) \), along with their estimated standard deviations. Under the SPM, expression levels are cell cycle regulated and if and only if \( \beta_j \neq 0 \). We choose a critical value of 5 for the absolute value of each \( Z_j \), the ratio of the estimate of \( \beta_j \) to its estimated standard deviation, to reject the null hypothesis. This value, far in the tail of normal distribution, is expected to preserve a genomewide significance level of about 0.3% (two-sided) even with as many as 6,000 genes under study. Some of genes that showed evidence of departure from SPM also may have expression patterns that vary with the cell cycle. One could test \( \beta_j = 0 \) also for those genes in the context of the augmented mean model, \( \tilde{\mu}_j(t) \), described above, though the interpretation of such a test would be conditional on the adequacy of the augmented model.

Nature of the Data. Three data sets were used in this analysis. The cdc28 data set was generated by Cho et al. (8), and synchrony was established by using a temperature-sensitive cdc28 mutation to reversibly arrest cells in G1. Oligonucleotide arrays (Affymetrix, Santa Clara, CA) were hybridized to fluorescently labeled cDNAs made from each sample, and the absolute fluorescence intensity values are assumed to be proportional to the amounts of each transcript in each target sample (7). These data were downloaded from http://genome-www.stanford.edu. The two other sets of data (alpha factor and cdc15) were generated by Spellman et al. (9) using an alpha factor-mediated G1 arrest and a temperature-sensitive cdc15 mutation to induce a reversible M-phase arrest, respectively. In this case, fluorescently labeled cDNAs were made from RNA from each time point and a second fluorescent dye was used to label cDNA from an asynchronous control culture. Control and test cDNAs were mixed and hybridized to arrays of PCR-amplified yeast ORFs. Fluorescence intensity values of both dyes were measured, and logarithmic ratios of test versus control values were generated. Obtained ratios are assumed to approximate the corresponding true ratios of test versus control mRNA levels (9). These data and the cdc28 data, resealed to mimic the ratio data, were accessed from the public domain site http://cellcycle-www.stanford.edu. Our results are based on the analysis of these data sets and as such will be influenced by all sources of variation involved in the preparation and processing of these arrayed samples.

Results and Discussion

The primary assumptions of SPM are that cell cycle-regulated transcripts will peak only once per cycle and that these pulses occur at invariant times in consecutive cycles. SPM includes terms that enable additive and multiplicative heterogeneity across samples to be accommodated. Fig. 2 shows additive heterogeneity estimates for each data set. Additive heterogeneity is minimal when logarithmic ratios are used. When the absolute intensity is considered for the cdc28 data set, the additive heterogeneity is most evident at the 90-min time point. This confirms the concern over this particular time point (8) and provides a means of correcting for its heterogeneity.

We estimated cell cycle span for each data set by using a set of 108 known cell cycle-regulated genes and profiling over a range of possible cell cycle spans (see Appendix). As expected, the cycle span differs for each synchrony method. Cycle spans for the alpha factor and cdc15 data sets show bimodal distributions (Fig. 2). These may be due to recovery artifacts that differentially affect the first cycle and alter the timing of a subset of the transcripts. We have used the estimated cell cycle span that minimizes a certain weighted sum of squares, giving a value of 58 min for the alpha factor synchrony, 115 for the cdc28 cells, and 85 for the cdc28 culture. Fig. 2 also shows the estimated standard deviations associated with loss of synchrony over time. Once these values have been obtained, the \( x_i^2 \) values are calculated for the \( j \)th gene for \( j = 1, \ldots, J \), and gene-specific parameters are estimated for all genes having transcription patterns consistent with the SPM (i.e., \( x_i^2 < 11.3 \)). Gene-specific parameters include the mean activation and deactivation times and the basal and elevated levels.

Fig. 3 shows the microarray data (solid lines) for five periodic genes and the fitted SPM to these profiles (dotted lines). Clearly, the model closely approximates the profile of the data and provides mean activation and deactivation times (in brackets) that are consistent with the patterns observed. The \( Z \) values for these oscillations vary from about 18 for RF41 in the cdc15 data set to about 3.5 for MCM3 in the alpha factor data set. The fact that the periodic behavior of MCM3 is still evident gives us confidence that we have set a sufficiently conservative threshold for each \( Z_j \). The top three transcripts have been classified as G1-specific, MCB-regulated genes (9). However, the PDS1 pulse is delayed compared to the other two. RF41 and CLB6 are activated at about the same time, but the pulse of CLB6 mRNA
is shorter lived. These differences are reflected in the activation and deactivation times calculated for each gene by SPM and can be used to identify coordinately regulated transcripts.

**Identification of Cell Cycle-Regulated Transcripts Using SPM on the cdc28 Data Set.** A total of 607 genes met SPM thresholds for periodicity (i.e., $|Z|$ value of 5 or greater) using absolute fluorescence intensity measurements directly from the cdc28 data (8). We obtained about the same number of genes by using either the logarithm of the intensity or the logarithmic ratios of intensities as generated by Spellman et al. (1, 2, 9). However, only about 500 genes were identified in all three analyses. Thus, any single data transformation may miss about 20% of the potential positives, due to $Z$ values that are close to our threshold. In all subsequent analysis, we have used logarithmic ratios of the cdc28 data to be consistent with the alpha factor and cdc15 data.

Lists of cell cycle-regulated genes in the cdc28 data set have been compiled by visual inspection (8) and k-means clustering (10). SPM analysis confirms a majority of these assignments and identifies many more candidate oscillating transcripts. The application of the k-means approach provided by Tavazoie et al. (10) used an initial filtering strategy to select the 3,000 yeast genes, which showed the highest coefficient of variation over the time course. Then, the iterative k-means procedure was used to partition all 3,000 profiles into 30 clusters. The requirement that all 3,000 profiles fit into one of 30 clusters necessitated the assembly of large clusters with loosely correlated patterns of expression. Five of these clusters had mean temporal profiles that were clearly periodic over two cell cycles. However, only about half of the profiles of the 524 cluster members exceeded the thresholds for periodicity in SPM.

To see whether SPM could identify a tight cluster of periodic genes, we computed the $\chi^2$ and $Z$ values for a cluster of G1-specific transcripts which was assembled at three different thresholds by Heyer et al. (11), using the QT Clust algorithm. In this case, we find that all the tightest cluster members either exceed or come very close to the threshold for periodicity set in SPM (Fig. 4 Top). Inspection of the borderline cases indicates that they are likely to be periodic and thus our Z value threshold is conservative. When the cluster threshold is set lower, membership doubles and again nearly all profiles are at the SPM threshold or well above it (Fig. 4 Middle). However, as noted by Heyer et al. (11) further relaxation of the cluster threshold to include 272 profiles leads to the inclusion of many poorly matching patterns that also have low $Z$ values by SPM (Fig. 4 Bottom). This finding indicates the efficiency of both approaches in identifying the most periodic transcripts. It also
Periodic transcripts that peak in the G1 phase of the cell cycle were identified by Heyer et al. (11) using the QT Clust algorithm and varying the cluster diameter threshold from <0.3 (top 41 genes), to <0.5 (83 genes), to <1.2 (272 genes). The transcript profiles for members of these successively larger G1 clusters were analyzed by SPM, and their Z and \( x^2 \) values are plotted (Left). The Z score and \( x^2 \) thresholds of SPM are superimposed on these plots to show that the proportion of these profiles that would be classified as periodic (lower right quadrant of each plot). (Right) the distribution of mean activation and deactivation times is plotted for each group. These parameter estimates were calculated by SPM only for those profiles that exceed SPM thresholds.

Fig. 4. Periodic transcripts identified by SPM with thresholds of \( |Z| > 5 \) and \( x^2 < 11.3 \) are depicted to show the extent of agreement between the three data sets. Logarithmic ratio data provided by Spellman et al. (9) for each of the three data sets was analyzed by SPM. The total number of periodic genes identified in each data set is shown and is represented by a circle. Agreement between data sets is indicated by the intersections of the circles. A total of 71 genes meet the SPM threshold for periodicity in all three data sets; 254 score as periodic in at least two databases; 834 appear periodic in only one data set; and 1,088 meet the SPM threshold in only one database. If we use an additional criterion of \( R^2 > 0.6 \) to identify the profiles among these 1,088 for which the model provides an explanation for 60% or more of the expression data variation, we find 473 profiles. This list is available at our website (www.fhcrc.org/labs/breeden/SPM).

Using All the Data Sets to Estimate the Number of Periodic Genes.

One limitation of these cell cycle data sets is the small number of samples and the lack of multiple measurements at any time point, which makes the identification of false positives and false negatives problematic. To mitigate this problem, we have used SPM to identify periodic transcripts from the cdc28, cdc15, and alpha factor data sets separately and then compared the results. SPM identifies about twice as many periodic genes in the cdc28 data set as in either of the other two synchronies (Fig. 5), and overall there are 1,088 genes that show significant oscillations in at least one data set. Included among the 1,088 candidate periodic genes identified by SPM are 81% of the 104 known periodic genes. A total of 254 genes oscillate significantly in at least two databases, which represents 4% of all genes, but includes 46% of the known periodic genes. Thus, SPM identifies the known periodic transcripts well above the level expected by chance. Only one-quarter of the known periodic genes are among the 71 genes that score as periodic in all three data sets. A total of 834 genes appear periodic in only one data set and as such further data collection will be required before this large group of genes can be unambiguously classified. Complete lists of the periodic transcripts identified by SPM are available at our web site (www.fhcrc.org/labs/breeden/SPM).

Spellman et al. (9) used Fourier analysis of the combined data from the same three data sets to identify periodic transcripts. Using the known periodic genes as a guide for setting their threshold, they estimated that 799 genes are periodic. Only 65% of these genes also are picked up by SPM as being periodic in at least one data set. This difference may be explained, in part, by our conservative threshold for Z because reducing the threshold value for Z to 4.0 enables 79% of these genes to be classified as periodic in at least one data set.

Nearly all of the genes that exceed the threshold for periodicity by SPM in at least two data sets also are recognized by the method of Spellman et al. (9). Here again, as with clustering, the most robust periodic patterns are identified by both methods. However, there are 571 genes that appear to be periodic by SPM criteria in at least one data set but are not classified as such by Spellman et al. (9). As noted above, these cannot be unambiguously classified as periodic without further corroborating data. They are either false negatives in two data sets or false positives in one. Experimental variation is much more likely to result in a nonperiodic pattern than it is to produce a smoothly oscillating profile. With SPM, the peaks also must occur at the same time in consecutive cell cycles and peaks and troughs are not recognized if they are represented by a single point in the profile (see Appendix). These restrictions should reduce the impact of noise and result in a lower false positive error rate. However, we cannot eliminate the impact of noise in the data and, with so few data points to base these assignments upon, many remain
characterize single pulses of transcription that occur at invariant times in consecutive cell cycles. SPM is a specific application of statistical modeling, but the basic strategy can be applied to any large data set to identify genes undergoing a transcriptional response to a stimulus. Due to its relative simplicity, statistical modeling can be used to interrogate large data sets without using additional filters to reduce the number of genes to be analyzed. It also includes heterogeneity parameters that will tend to reduce the impact of noise in the data sets. SPM identifies regularly oscillating transcripts without regard to the abundance of the transcript or the height or timing of the peak and provides estimates of the mean time of activation and deactivation. These values are only estimates, but they are unbiased under the assumed SPM and can be considered defining characteristics of individual genes. SPM also provides statistical measures of the precision of the parameter estimates so that optimal groupings can be made and subjected to further analysis. These features of statistical modeling complement and augment the other methods used to analyze microarray data.

We thank Drs. M. Campbell, J. Cooper, and J. Sidorova for helpful discussions about the data. We also thank the investigators who contributed their yeast expression data to the public domain, and L. Heyer, S. Kruglyak and G. Church for providing cluster information for comparison. This research is supported by grants from the National Institutes of Health (GM41073 to L.B. and PO1 CA53996 to L.P.Z. and R.P.).

Conclusion
In this report we use a statistical model (SPM) to identify and characterize single pulses of transcription that occur in variant

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:

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

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

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

with the summation over 1st, 2nd, 3rd, \ldots, cycle, and \( I(\cdot) \) 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 \((\varsigma, \xi)\) (Fig.1). The model for the expected expression level for the cell may be written as

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

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 $\sigma$. 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 $i$th cell, and $Y_i^*$ is the expression level of a particular gene in the $i$th cell. Modeling the mean expression level $Y_i$ by SPM gives the expected values of $Y_i^*(t + T_i)$ as

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

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:

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

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 $\sigma$ 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 $(\gamma_0, \gamma_1)$ are parameters to be estimated.

A fifth step incorporates multiplicative ($\lambda_k$) and additive ($\delta_k$) 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,

$$
\mu(t_k) = \delta_k + \lambda_k \{ \alpha + \beta \left[ \sum_{c=0}^{\infty} \Phi\left( \frac{\xi_c + c\Theta - t_k}{\sigma_k} \right) - \Phi\left( \frac{\xi_j + c\Theta - t_k}{\sigma_k} \right) \right] \}
$$

in which $\delta_k$ and $\lambda_k$ are specific to the $k$th sample and the $\delta_k$’s and $\lambda_k$’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

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

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,

$$
\sum_{j=1}^{J} \sum_{k=1}^{K} \left[ Y_{jk} - \mu_j(t_k) \right]^2
$$

Because the mean activation and deactivation times represent changing points and are restricted ($\zeta_j \geq 0, \xi_j \geq 0$ and $\xi_j > \xi_j$), we minimize the above sum of squares (Eq. A1) with respect to the other parameters at each point on fine grid values for ($\zeta_j, \xi_j$), and select the set of parameters estimates giving an overall minimum for Eq. A1. We restricted the profiling
procedure to points such that \((\xi_j, \xi_j)\) included at least two \(t_k\) values. The weight function in the calculation was defined as

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

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,

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

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 \(R^2_j\) 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 \((\hat{\delta}_k, \hat{\lambda}_k)\), \(k = 1, \ldots, K\), by minimizing Eq. A1 with all \(\beta_j\) 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 $\Theta$, 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, $\sigma_k$, 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 $5.7 \times 10^{-7}$, so that the probability that any one of the $\hat{\beta}_j$ values, say 6,000 genes, exceeds 5 if all $\beta_j$ 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: