

Timing properties of gene expression responses to environmental changes

Gal Chechik ^{*}, Daphne Koller,
Computer Science Department, Stanford University
Stanford, CA, 94305-9010, {gal,koller}@cs.stanford.edu

February 7, 2009

Abstract

Cells respond to environmental perturbations with changes in their gene expression that are coordinated in magnitude and time. Timing information about individual genes, rather than clusters, provides a refined way to view and analyze responses, but is hard to estimate accurately.

To analyze response timing of individual genes, we developed a parametric model that captures the typical temporal responses: an abrupt early response followed by a second transition to a steady state. This *impulse* model explicitly represents natural temporal properties such as the onset and the offset time, and can be estimated robustly, as demonstrated by its superior ability to impute missing values in gene expression data.

Using response time of individual genes, we identify relations between gene function and their response timing, showing, for example, how cytosolic ribosomal genes are only repressed after mitochondrial ribosome is activated. We further demonstrate a strong relation between the binding affinity of a transcription factor and the activation timing of its targets, suggesting that graded binding affinities could be a widely used mechanism for controlling expression timing.

^{*}Current Address for correspondence: Google Inc, 1600 Amphitheater Parkway, Mountain View CA, 94043

Introduction

Over the past few years, significant progress has been made in mapping different components of the cellular architecture: protein complexes, functional modules, and even more complex pathways and cellular networks. However, the static set of components and their interactions tells only part of the story. In reality, cells continuously reconfigure their activity to adapt to their fluctuating environment, and activate different parts of their pathways in a dynamic way. Obtaining insight into the cellular dynamics is a significant challenge, primarily because data measuring aspects of the cell’s activity over different points in time is hard to obtain, especially at a genome-wide scale.

Arguably, the main data so far that have provided a genome-wide view into the cell’s dynamics are measurement of gene expression profiles taken over a time course, following a perturbation to the cell’s environment. Although these measurements probe only a single level of the cellular control hierarchy, the availability of transcription data under multiple conditions could provide significant insights into dynamics of cellular control. With these data, we might hope to study how the transcriptional program changes to cope with an environmental perturbation. We can try to understand the role that expression timing plays in cellular responses, to map those genes and modules that are expressed in a timely manner and to identify molecular mechanisms that control timing.

Unfortunately, gene expression time courses are hard to interpret: they are notoriously noisy, often measured at irregular intervals, and these intervals differ from one experiment to the other. Thus, with the exception of cell cycle data, much of the analysis of gene expression profiles has ignored their temporal aspects, using these data primarily to identify genes that share common responses across experiments, and to associate genes with various cellular processes based on their response profiles.

Some papers do attempt to model the dynamics of expression time courses (see [1] for a recent survey). Several approaches [2, 3, 4, 5] have focused on capturing the dynamics of cell cycle time courses; these methods are tailored to the sinusoidal transcriptional patterns in the cell cycle, and do not generalize to other types of time series. In the more general setting, Bar-Joseph and other researchers [6, 7, 8, 9, 10] showed how splines can be used to encode continuous gene expression profiles, and successfully impute missing values and align “similar” expression profiles that exhibit different temporal properties. Some methods [11, 12, 13] have defined “shape-based” similarity metrics for gene expression time courses, for the purpose of gene clustering, but without attempting to extract or evaluate specific timing properties. Other approaches

[14, 15, 16, 17, 18] use a probabilistic or regression-based time series model to capture the temporal dynamics of gene expression data. These approaches all use generic function representation, capable of capturing a broad family of response profiles, and hence tend to over-fit the data more easily. As a consequence, the parameters of the model are typically estimated using clusters of genes, possibly obscuring finer-grained signal. Most importantly, however, these methods do not easily provide an approach for extracting biologically meaningful timing aspects of the responses in individual time courses, and compare these timing aspects across different conditions.

In this paper we propose a parametric approach that identifies interpretable timing properties of mRNA profiles, and use them to characterize the timing of cellular responses. The idea is to fit any given time course with a function that is parametrized with biologically meaningful and easily interpretable parameters. Specifically, we describe a phenomenological model for encoding a gene’s continuous transcriptional profile over time. The model is designed to capture the typical *impulse*-like response to an environmental perturbation such as changing media or stress condition: transition to a temporary level followed by a second transition to a new steady state. Thus, we define the model in terms of meaningful aspects of the response: its *onset* and *offset* times, the slope of the response, and the short term and long term response magnitudes.

We evaluate the model on a broad compendium of 481 measurements in *Saccharomyces cerevisiae*, comprising 76 different gene expression time courses following diverse environmental perturbations. We show that the impulse model is rich enough to capture a wide variety of expression behaviors and at the same time robust enough to be learned from sparse data. We demonstrate this robustness by providing estimates of missing measurements that are significantly more accurate than other approaches. We then show how we can use the biologically meaningful parameters that we extract from the impulse form to shed light on the cell’s transcriptional response to environmental changes. In a parallel work [19] we use the model presented here to analyze pathway-dependent patterns of activity in metabolism, demonstrating its utility in a notehr context.

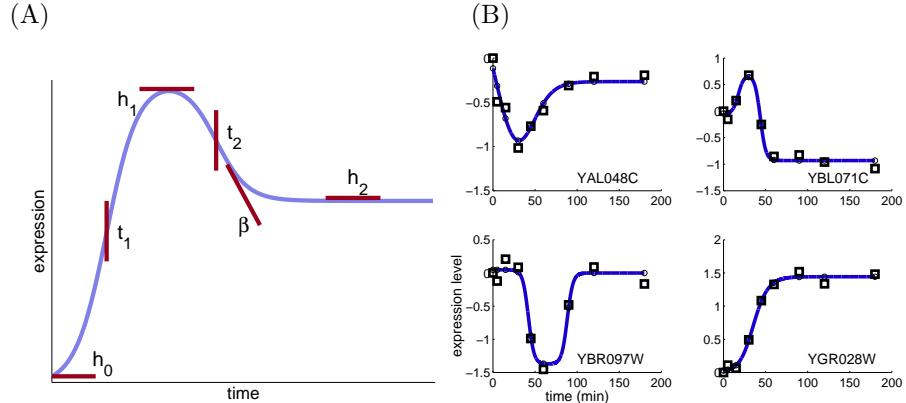


Figure 1: **The impulse model.** (A) The six parameters of the impulse model. (B) Examples of impulse model fit (solid line) to gene expression (squares) in response to 1M sorbitol, as described in [20].

Results

An *impulse* model of responses to changes

When subjected to an abrupt change in the environmental condition, a cell typically responds by increasing the activity level of certain sets of genes and decreasing the activity level of others. For example, when exposed to a heat shock, genes involved in growth related processes are repressed, then shortly followed by repression of ribosomal proteins coding genes. In many cases, the expression level changes temporarily, exhibiting a sharp increase or decrease, and later changes again, reaching a new steady state which is often different from the original “resting” state Fig. 1. This two-step behavior is widely observed in multiple systems, from yeast [21, 13] to human [15]. The reason is that an abrupt environmental change requires two types of adaptive responses. First, the cell actively reconfigures some processes, typically involving both generic emergency responses and specialized processes that the cell recruits. At a second phase, the cell achieves a new homeostasis in its new environment.

We propose an *impulse model* designed to encode a two-transition behavior, allowing us to compactly represent the relevant aspects of expression responses to environmental changes. The impulse model encodes this behavior as a product of two sigmoid functions, one that captures the onset response, and another that models the offset (see Methods). Importantly, this model allows for a sustained expression level different from the resting state. The model function has six free parameters (shown in Fig. 1(A)). Three amplitude (height) parameters

determine the initial amplitude (h_0), the peak amplitude (h_1), and the steady state amplitude (h_2). The onset time t_1 is the time of first transition (where rise or fall is maximal) and the offset t_2 is the time of second transition. Finally, the slope parameter β is the slope of both first and second transitions. Formally, the model has the following parametric form:

$$\begin{aligned}
 f_\theta(x) &= \frac{1}{h_1} \cdot s_1(x) \cdot s_2(x) \\
 s_1(x) &= h_0 + (h_1 - h_0)S(+\beta, t_1) \\
 s_2(x) &= h_2 + (h_1 - h_2)S(-\beta, t_2) \\
 S(\beta, t) &= \frac{1}{1 + e^{-\beta(x-t)}} \\
 \theta &= [h_0, h_1, h_2, t_1, t_2, \beta] \quad .
 \end{aligned} \tag{1}$$

What type of profiles can the impulse model capture? It is designed for modeling temporal profiles that have at most two significant changes in expression levels. Examples of such profiles are depicted in Fig. 1(B), where the impulse model was fit to actual expression measurements of yeast genes. The impulse model is not appropriate for encoding periodic behavior with multiple peaks, such as the characteristic behavior of the cell cycle (like the well-studied data of [22]). Thus, the impulse model is best-suited for capturing a one-time response to some external signal such as an environmental disturbance.

The parameters of the model are learned by minimizing a squared error to fit measured data. Given a set of expression measurements $\{e_1, \dots, e_n\}$ at time points $\{t_1, \dots, t_n\}$, we search for the set of impulse parameters θ that minimize the squared prediction error $\min_\theta \sum_i (f_\theta(t_i) - e_i)^2$. We find the (locally) optimal parameters using a conjugate gradient ascent procedure, repeated 100 times with different starting points (see Methods).

Gene expression measurements are notoriously noisy and hard to model, especially on the level of individual genes. We systematically evaluated the properties of the impulse model using a diverse set of 76 conditions. First, we found the model to be remarkably robust to both timing noise and to expression level noise (see Methods). Furthermore, we estimated the model's coverage — the fraction of genes that can be well-fit with the model — showing that up to 95% of the genes are well described by the impulse model, depending on the condition (see Methods). Finally, we estimated the extent to which genes had a particularly impulse-like response, showing that, on average, 35% of the genes have an impulse-like response profile (see Methods).

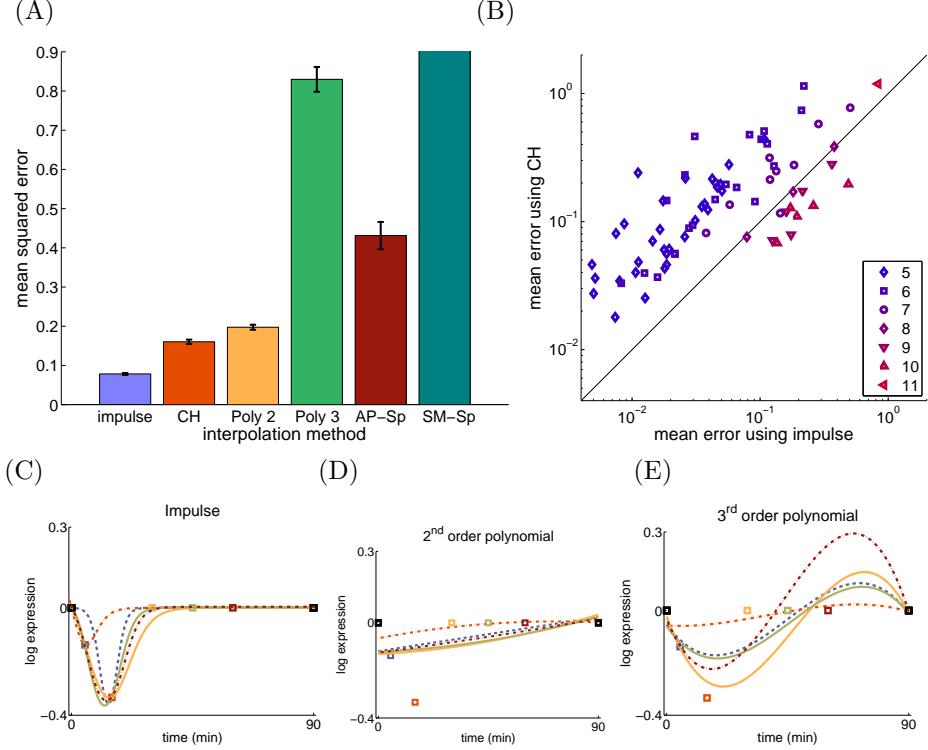


Figure 2: **Imputing missing values.** (A) Mean squared error for imputing missing values with the leave-one-out procedure described in the text. Methods compared are: impulse model, cubic Hermite (CH), 2nd and 3rd order polynomials, approximating splines and smoothing splines. Error is the average over 6209 genes in 76 conditions. Error bars denote the standard error of the mean (s.e.m.) across the 76 conditions. (B) **Scatter plot of the mean error for the impulse model and cubic-Hermite (CH, the second best predictor).** Each point corresponds to a different condition, and its shape shows the number of time point measurements in that condition. The impulse model provides superior fits, especially in conditions with a small number of time points. Note that the figure is in log-log scale, demonstrating that the impulse model is superior across the full range of errors, providing better fit both for easy-to-fit and hard-to-fit profiles. (C)–(E) **Comparison of leave-one-out fits to a gene expression profile.** Squares denote measurements, which are the same for all three panels. For each method, 5 curves are shown, each corresponding to a fit performed with a different single measurement that was left out during the fit. The color of each curve corresponds to the color of the hidden value (square marker). Curves for the rightmost and leftmost measurements are not shown, because polynomials perform very poorly in this extrapolation task. (C) Impulse model. (D) 2nd order polynomial. (E) 3rd order polynomial. A fit for cubic hermit is given as supplemental figure.

Imputing missing values

The impulse model is a continuous function that provides an estimate for a gene's expression measurement at each point in time. We show that the impulse model can accurately predict the value of missing expression measurements. Imputing missing values is an important problem in gene expression data, hence the success of our impulse model at this task is both a validation of the model, and one of its applications.

We applied the impulse model to a compendium of 76 gene expression time courses in *Saccharomyces cerevisiae*, which measure the response of yeast to different environment stress conditions and changing media [23, 20, 24, 25, 26, 27, 28, 29]. Time courses had between 5 and 10 measurements (see a full list of data sets and time courses in Supplemental Table 1).

We evaluated the performance of the model on the imputation task in two ways: using information only at the level of individual genes; and incorporating information from other, similar genes.

Using individual genes

First, we considered the ability of an impulse model to estimate the value of an unmeasured expression value for a gene, given the other expression measurements for that gene alone. For a given gene, we held out one of the expression measurements, fit an impulse model to the remaining measurements, and used the resulting function to estimate the expression value at the held out time point. We compared this value to the measured held-out value, and computed the error. We repeated this experiment for all 6209 genes in our compendium and all measurements, and computed the mean prediction error. For comparison, we applied the same procedure using other methods for function estimation, including both interpolation methods such as interpolating splines and cubic-Hermite polynomials, and fitting methods using polynomials of degrees two to five, and smoothing splines. All of these methods used information at the level of single genes only, using measurements taken at all available time point to predict the value in a single hidden time point. The results of this comparison are shown in Fig. 2(A). The prediction of the impulse model are significantly superior to all the other methods.

Fig. 2(B) shows a scatter plot of average prediction error for each of the 76 conditions, as obtained with the impulse model and the cubic-Hermite (CH, the second best predictor). It shows that the impulse model is particularly better at fitting time courses with a small number of points, suggesting that it avoids over-fitting more effectively.

Interestingly, a comparison to a third order polynomial yields similar results. This similarity suggests that even though the impulse model has 6 free parameters, it avoids over-fitting better than a model with 4 free parameters. The reason is that polynomials are generic function approximators, capable of fitting any function, hence could predict fits that are highly unlikely for gene expression timecourses. In comparison, the impulse model focuses on a restricted set of behaviors, and hence uses the domain-specific knowledge to avoid large mistakes.

This effect can be understood by comparing the actual functions learned by the different fitting procedures. Fig. 2(C)-(E) compares the fits to a particular gene expression profile for three methods: polynomials of degree 2 and 3, and the impulse model. The descriptive power of the 2^{nd} order polynomial is too limited, leading to a “flat” curve that changes little in time. On the other hand, the 3^{rd} degree polynomial is too expressive, and over-fits for several time-points. Conversely, the impulse model, despite having a larger number of free parameters, successfully avoids over-fitting the measurements.

Figure 3

Using whole genome information

When imputing missing values, a valuable source of information is the similarity in expression profiles between different genes. Two approaches are commonly used for taking this information into account. First, missing values can be inferred from neighboring genes, where the neighborhood is based on the observed measurements. Second, genes can be clustered and the cluster profiles are then used for imputing the missing values. We compare the performance of the impulse model with two standard methods that take these two approaches.

For the first evaluation, we follow the approach of Troyanskaya *et al.* [30] and use profiles of similar genes to complete missing measurements. Troyanskaya *et al.*, in their KNN-impute procedure, propose a k -nearest neighbor procedure, estimating the value of a time t measurement for gene g as the average of the time t expression values measured for the k genes most similar to g . KNN-impute uses a Euclidean distance over the vector of expression measurements to find the nearest neighbors. To evaluate the gain in using the impulse model we applied the same procedure, but using the values predicted by the impulse model fit, rather than the raw original measurements.

Specifically, we hid a randomly selected single time point in the expression profile of each gene, and used the remaining measurements to estimate the left-out values (see Methods); overall, this process resulted in a level of about

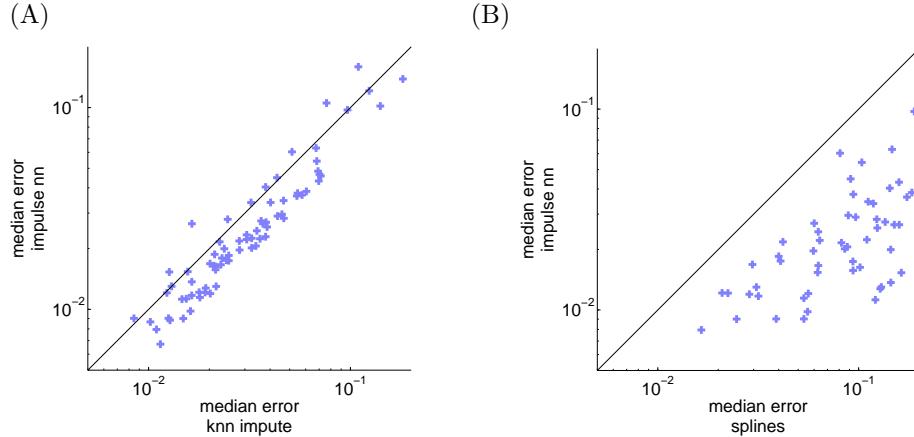


Figure 3: Whole genome comparisons. Median squared prediction error obtained with *Impulse-KNN* compared with two other methods. Both figures are in log-log scale, demonstrating that impulse NN is superior across a large range of error values. **(A) Comparison with KNN-impute**, where missing values are filled using nearest neighbors (according to a Euclidean distance between experimental measures). Impulse-model errors are on average 20% lower than those of KNN-impute with a Euclidean distance. **(B) Comparison with spline clustering** [6], where expression profiles are simultaneously clustered and modeled as a time course. *Impulse-KNN* errors are on average 35% lower than with spline clustering.

10–20% missing values, depending on the number of measurements in each time course. For each gene, we estimated the curve fit to the remaining measurements of that gene. We then estimated the value of a missing time t measurement for gene g by selecting the k genes nearest to g , using Euclidean distance over the predicted values, and averaging the predicted expression values at time t . Note that the predicted values were used both for selecting the neighbors and as a basis for estimating the time t value. For comparison, we also applied the standard KNN-impute procedure to the same data.

Fig. 3(A) compares the median error obtained with the two distance measures across 76 conditions. Using the impulse model reduced the error in 64 out of 76 conditions, yielding an average error reduction of 20% of the KNN-impute. This difference was highly significant (paired t -test: $p < 3 \times 10^{-6}$). The analysis was repeated for $k = 10$ and $k = 20$, with almost identical results (data not shown).

Bar-Joseph *et al.* [6] used another approach for utilizing similarity of expression profiles across genes. They cluster genes and train a model based on approximating splines for cluster profiles. We compared this method with the *Impulse-KNN* method described above, over the same data set described above,

We used code supplied to us by Bar-Joseph, and selected values of the parameters that performed well in the experiments described by Bar-Joseph *et al.* We used 10 clusters, since we found that this number of clusters captures well most of the structure in the data. The results, shown in Fig. 3(B) show that the *Impulse-KNN* model outperformed spline-based clustering by 35% on average.

Temporal patterns of response to changes

The impulse form directly provides meaningful parameters that characterize the shape of the response profile, including the response onset, offset and profile peak. We chose to focus on the onset response time, since it directly captures the timing at which the cell initiates the production of a gene's mRNA, and this timing could be critical to the survival of the organism upon an environmental change. We therefore extracted the onset of every response profile, and used these timing data to explore the relationship between response onset and gene function.

To illustrate the insights arising from this type of analysis, we can consider the timing patterns arising when the cell is exposed to diamide [20]. Here, we can see that genes involved in *gene expression* respond at a wide range of delays (Fig. 4(A)). Looking at three main subsets of this group, we find that genes that are involved in *RNA processing* typically respond earlier than the other genes; transcription genes also respond early, and translation is last. Interestingly, translation occurs in two peaks, one observed early (~7 minutes) and a second occurring much later (~18) minutes.

To understand this phenomenon better, we look into the distribution of onset times and peak responses of all ribosomal genes under diamide exposure. A finer breakdown of the set of ribosomal genes reveals that the vast majority of the early onset events correspond to induction of the mitochondrial ribosome, whereas the later events represent the repression of the cytosolic ribosome (see Fig. 5). We note that previous studies of these data [20, 8] have noted the differential expression of the ribosomal genes: while most cytosolic translation is repressed, the mitochondrial ribosome is induced in order to handle the oxidative stress caused by diamide. However, our onset timing analysis provides an additional dimension to this standard result, demonstrating that there is also a difference in the timing of these two events. We hypothesize that the reason for this delay is that upregulation and translation of mitochondrial genes is required to deal with the stress. Hence, cytosolic ribosomal genes can only be repressed after translation of mitochondrial genes is completed.

The data in Fig. 5 also reveals a fairly large group of cytosolic ribosomal genes that are repressed considerably earlier than the bulk of the genes in this

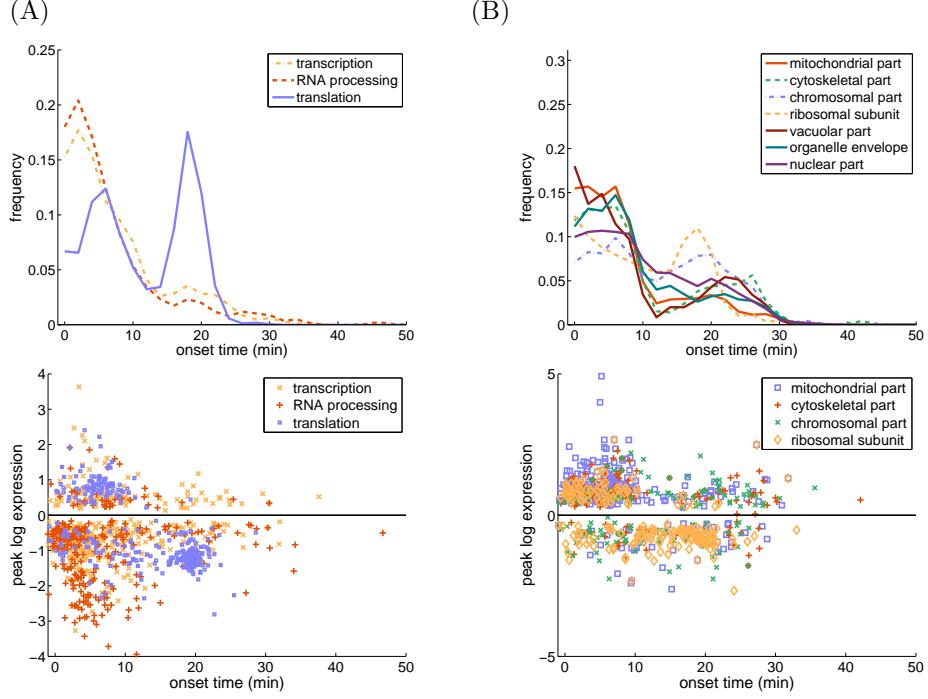


Figure 4: **Distribution of onset time and peak responses.** Upper panels: Distribution of onset time in sibling GO categories. Bottom panels: Distribution of onset time and peak response level per gene. **(A)** Subclasses of the *gene expression* GO category, under exposure to diamide [20]. **(B)** Subclasses of *intracellular organelle part*; Exposure to acid [25]. To reduce clutter, Not all subclasses are shown.

category (see Supplemental Table B). An in-depth investigation of these two groups of genes shows two interesting trends. First, in the early group, many of the genes (10 out of 33) are not ribosomal components but are more likely required for creation of ribosomes and for RNA processing or translational fidelity; by comparison, such genes are a small proportion of the late group (3 out of 115, $p < 10^{-6}$). One hypothesis is that the cell first represses accessory proteins, whereas the structural components are only shut off at the end, giving enough time for translation of the mitochondrial ribosome, as well as any other proteins necessary for the cell’s immediate response. As a second trend, for the large ribosomal subunit, we see nine genes in the early group that code for the same component as a gene in the later group (for example, *RPL13A* shuts down early, whereas *RPL13B* shuts down later). The only case where both copies are shut down early is *RPL41A* and *RPL41B*, which code for a non-essential component of the ribosome. An interesting hypothesis is that, to conserve resources,

the cell begins by shutting off one copy of each component, and only then shuts down the other. The situation is a little less clear with the small subunit, where three components have both copies shut down early; however, these are not in the central part of the ribosome. It would be interesting to understand whether and why these components are not required during the transition phase.

To generalize this type of analysis and identify other functions whose RNA levels are carefully timed, we looked at the distribution of onsets across genes grouped by their GO associations. In each condition, we then searched for GO categories whose onsets are significantly different from a baseline distribution of onsets. A relevant baseline should contain genes of similar (but not identical) functions. We therefore defined a separate baseline for each category using all genes from sibling categories in the GO hierarchy (other children of its parent category). For each GO category and each condition, we calculated a Wilcoxon score to quantify how significantly its gene onsets appear earlier or later than the baseline onsets. This comparison provides a tool for identifying sub-functions that are controlled in time. We found 151 sub-categories that exhibited highly significant (Wilcoxon test, $p < 10^{-5}$, Bonferroni corrected) onset differences at least in one condition (see Supplemental Table IV for a full list).

Fig. 4(B) shows another example, for the main sub categories of *intracellular organelle part*, under exposure to Acid [25]. Mitochondrial genes are again regulated significantly earlier, and so are cytoskeletal genes, while a larger fraction of chromosomal respond late. Ribosomal genes again have two peaks, and these correspond again to mitochondrial and cytosolic ribosome; indeed, as we discuss below, this distinction is found across a variety of conditions. Here, vacuolar genes also appear to have two distinct peaks, with 53 genes responding before $t = 12$ minutes and 20 genes responding after. Relative to the late vacuolar genes, we find that the early vacuolar genes are enriched for *vacuolar membrane* (hypergeometric $p < 10^{-15}$).

We can also utilize our timing analysis to construct a system-level “response timeline”, by looking at how multiple functional categories are ordered in time. Under each condition, we calculated the ordering score for every pair of GO categories, and used these ordering scores to identify sets of categories that are regulated in a timing-distinct manner (see Methods). As one example, we consider the onset timing extracted from the responses to DNA-damaging gamma irradiation [24]. Fig. 6 plots the median peak and median onset time for each of the top four timed categories in the *cellular-component* hierarchy. First, genes of the nucleolus (a sub-organelle of the cell nucleus) are repressed, followed by repression of ribonucleoproteins, then cytoplasmic proteins. Finally, membrane proteins are activated. A similar analysis on annotations in the *molecular func-*

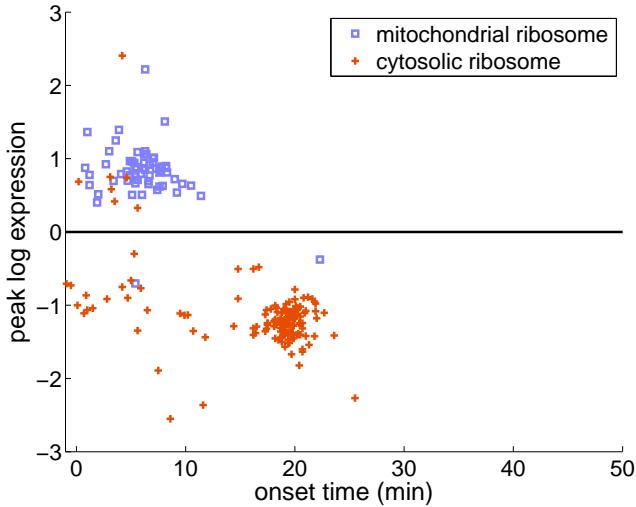


Figure 5: **The timing of of ribosomal gene responses under exposure to diamide [20].** The figure shows the timing and peak level for two subclasses of ribosomal genes: mitochondrial (blue squares) and cytosolic (red crosses). These two groups exhibit response profiles that are distinct both in their expression level and their timing, demonstrating that the timing of the ribosome to this condition is finely controlled.

tion and biological processes hierarchies in the same condition (Supplementary Figures 10 and 11), is consistent with this view: The biological processes of *ribosome biogenesis and assembly* (which takes place at the nucleolus) are repressed first, followed by the activation of the *localization* and *transport* genes (processes that take place at cytoplasm and membranes). Similarly, the molecular function *structural constituent of the ribosome* are repressed first, while multiple functions related to transport are activated later.

Another interesting perspective on this finding is the observation that the stronger the repression of the genes in these timed categories, the earlier the onset of the repression. This phenomenon holds not only for the medians of the groups in Fig. 6, but in fact the onset time is correlated with the peak response across all genes in these categories (Pearson correlation, $p < 10^{-10}$); this phenomenon holds only for genes in timed categories (the background correlation across all genes in this condition is p -value = 0.04). As one hypothesis, if a group of genes is highly detrimental to the cell (leading to a strong repression), it may be desirable to shut them off as soon as possible. In particular, if mRNA degradation mechanisms are used to decrease mRNA abundance in this condition [31], this finding may also suggest a sequential targeting of the RNA

degradation machinery, ordered by the cell’s current priorities.

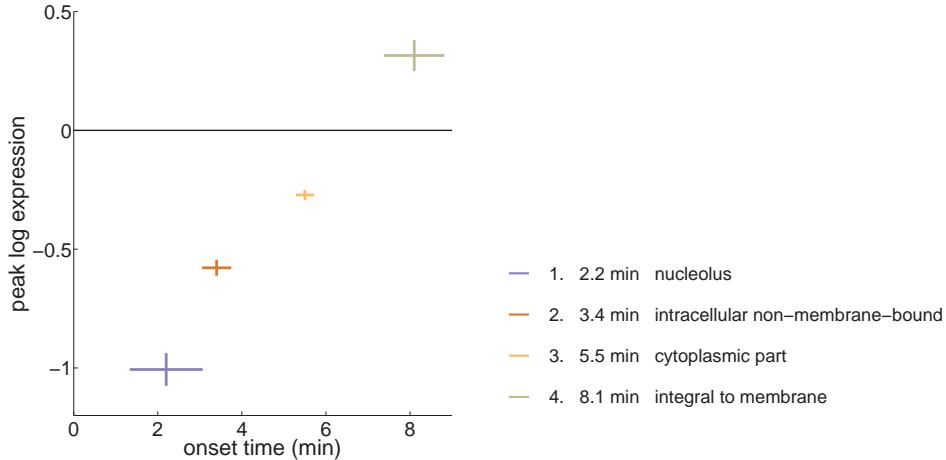


Figure 6: **A timeline of functional responses following gamma irradiation.** Each cross denotes the median peak and median onset time of all genes in the corresponding GO category [24]. Length of bars denote the standard error of the mean (s.e.m.) of each group (see Methods).

Finally, we looked at functional differences in timing across multiple conditions. We counted the number of conditions in which each pair of categories is significantly timed (p -value < 0.001 , Wilcoxon test, Bonferroni corrected). In general, nuclear and mitochondrial components respond earlier than cytosolic and ribosomal components. For instance, for the *cellular component* hierarchy, the mitochondrion, shown above to be activated early under exposure to diamide (Fig. 5), and acid (Fig. 4, responds significantly earlier (with $p < 10^{-3}$) than the cytosolic ribosome in 16 out of the 76 conditions tested (yielding an overall $p < 10^{-40}$, Binomial distribution with $p = 10^{-3}, N = 76$). These conditions were mostly stress conditions (rather than media changes), including exposure to diamide, dtt, KCL and heat shock. Many of these stress conditions create oxidative stress which elicit differential mitochondrial response. These results show that these mitochondrial genes typically respond early. Mitochondrial responses in the remaining conditions were more scattered in time and never significantly late.

For the *biological processes* hierarchy, *translation* often responds significantly later than other various metabolic and transcription processes. For instance, in 12 out 76 conditions *translation* response occurs significantly after *transcription*. Also, *biosynthesis* processes tend to follow metabolic processes. For instance, in 11 conditions, *biosynthetic process* responds significantly after *biopolymer*

metabolic process. For the *molecular function* hierarchy, almost all significant pairs were due to late response of *structural constituent of ribosome* and *structural molecule activity*. Supplemental Table IV lists all category pairs and conditions that exhibited significant timing relationships.

Graded binding affinity: A mechanism for controlling transcription timing

The above findings suggest that cells control the timing of transcription activation to shape their responses to environmental changes. What mechanisms could achieve fine timing control?

One possible mechanism is that sequential activation of genes is achieved by cooperative binding by several transcription factors (TFs), each activated in its turn. This hypothesis requires that TF's are themselves sequentially activated by some mechanism. A different (albeit not exclusive) mechanism is that a single transcription factor binds to multiple target genes, but with different binding affinities. Indeed, the recent work of Tanay [32] shows that binding affinities, as measured in ChIP-chip data [33], have functional consequences even in weak affinities that were previously considered insignificant. This work demonstrates that transcription binding is not an all-or-none phenomenon, and graded binding is achieved through graded sequence affinity. The reason and purpose for having a wide range of binding affinities is still unknown, but it was recently shown that gene expression in the phosphate response (PHO) pathway is tuned to different environmental phosphate levels using both binding-site affinities and chromatin structure [34]. In the specific context of metabolism, we demonstrate in concurrent work that some linear chains of metabolic reactions, whose transcription timing is ordered, also have ordered binding affinities [19]. However, it is unclear if the relation between binding affinity and expression timing holds across other cellular functions, or if it is specific to the ordering constraints of linear metabolic chains; and what range of binding affinities are relevant for controlling transcription timing.

If graded binding affinities are used for regulating the timing of gene expression, we expect the shape of a gene expression profile to depend on the strength of binding to its regulating TFs. Since binding operates as a stochastic equilibrium, the stronger the binding affinity of an activating TF to a binding site, the higher the probability of the TF to remain bound to the corresponding promoter and recruit the transcriptional machinery, and hence the earlier the gene would be expressed on average.

To test this single-TF hypothesis, we measured how binding affinities are related to the onset time of transcription activation. Specifically, we combined

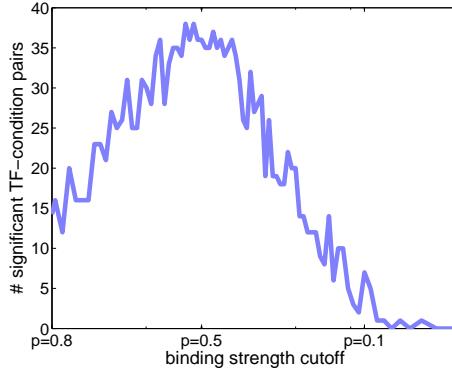


Figure 7: **Number of significant TF+condition pairs as a function of binding strength considered.** The peak is achieved at p -value = 0.52, with 38 significant pairs out of 48. A p -value of 0.5 corresponds to by-chance binding.

whole genome binding affinity measurements [33] with gene expression measurements as described above. We selected a subset of affinity and expression measurements that were taken in matching conditions. We collected a total of 48 affinity-expression experiment pairs (see Supplemental Table B2), including amino acid starvation (34 TFs), exposure to acid (2 TFs), and to heat shock (12 TFs).

For each affinity-expression experiment pair, we restricted attention to genes that were differentially expressed (absolute peak response $> R$), and measured the Spearman correlation between their onset time and the binding affinity of the measured TF, using the p -value as the quantitative measure of affinity. Of course, not all genes are bound by a particular TF; we therefore wanted to restrict attention only to those genes where TF binding plausibly occurs. As discussed above, Tanay [32] showed that measured binding affinity p -values are correlated with binding prediction based on sequence models, even for very weak binding, suggesting that measured weak binding may reflect actual binding rather than noise. We therefore considered the whole range of possible p -value thresholds for treating a binding event as valid (where the chance level is p -value = 0.5). Specifically, for a range of different affinity thresholds C , we computed the Spearman correlation between onset time and binding affinity, restricting the analysis to all genes that are both differentially expressed (crossing a threshold R) and have a binding affinity stronger than a cutoff value C . Fig. 7 shows the number of pairs that obtained significant Spearman correlation as a function of the affinity cutoff value C ; here, we used a gene expression response threshold $R = 0.7$, chosen to maximize the number of significant pairs. The number of

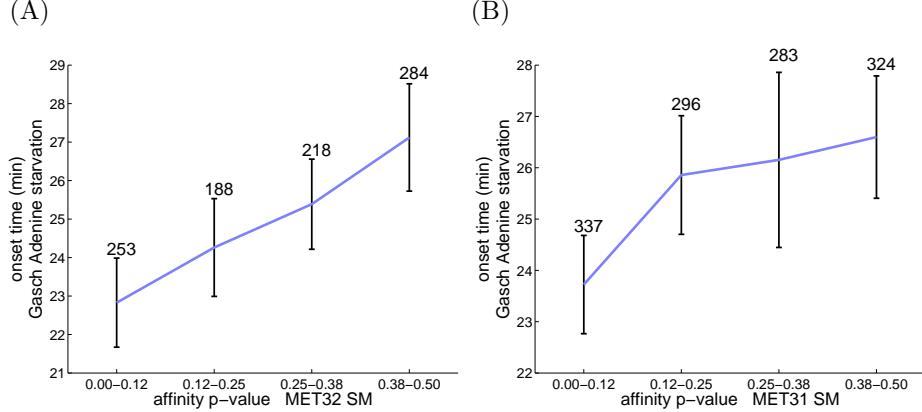


Figure 8: **Mean onset time across genes grouped by their binding affinity.** (A) Binding of MET31 measured under amino acid starvation and expression measured under adenine starvation, Spearman correlation $r = 0.14$ $p < 2.5 \times 10^{-10}$. (B) MET32 measured under amino acid starvation and expression measured under adenine starvation. Spearman correlation $r = 0.13$, $p < 1.9 \times 10^{-9}$. Error bars denote standard deviations, numbers above error bars denote group sizes.

significant pairs peaks near p -value = 0.50, where 38 of 48 TF-condition pairs have a significant correlation (FDR q -value $\leq 10^{-3}$) (the optimum is actually obtained at 0.52, which is larger than the chance level 0.5, but this is likely to be due to noise, . Typically, the correlations became even stronger when limiting the analysis to more strongly expressed genes (larger values of R), but the p -values decrease due to the smaller sample size.

Fig. 8 visualizes the relation between binding affinity and expression onset; here, to more clearly illustrate the pattern, we used an expression cutoff of $R = 1$. We aggregated the genes in our set into four groups according to their binding affinities, and calculated the mean onset time of each group. The left panel shows the results of this analysis for the targets of MET32, a transcription factor involved in methionine biosynthesis; here, the binding affinities were measured under amino acid starvation, and the transcription onset extracted from a time course following adenine starvation [20]. A clear trend can be observed in the mean onset time as a function of MET32 binding affinity, across the whole range of relevant affinity strengths. This effect is highly significant (Spearman correlation $r = 0.14$ across 943 samples, $p < 1.9 \times 10^{-7}$, Bonferroni corrected for 48 hypotheses). Other such trends were observed under amino acid starvation, including MET31 (Fig. 8b) (Spearman $r = 0.14$, Bonferroni $p < 2.5 \times 10^{-8}$), CBF1 ($p < 9.7 \times 10^{-8}$) and SFP1 ($p < 6 \times 10^{-9}$).

We also found pairs that exhibited significant negative correlations (for in-

stance YAP1 and HSF1 under a heat shock), where higher binding affinity was associated with delayed onset time. The mechanism for such associations is unclear at this point, and could be related to competition between TFs.

This finding has two implications. First, it shows that graded binding affinities are very commonly correlated with expression timing, and could be a commonly used mechanism for controlling the timing of response onsets. Second, it suggests that even (very) weak binding affinities have a functional effect on the concerted profile of cellular expression responses.

Discussion

Environmental changes may threaten the survival of cells and force them to respond quickly and reconfigure their gene expression profiles. To respond efficiently to changing conditions, cells have to control not only the magnitude of their responses, but also their timing. Indeed, it was shown that expression timing in *E. Coli* is tightly controlled, even to the level where sequences of individual proteins are expressed in an ordered manner [35, 36]. It is unknown, however, if such controlled timing is to be found across multiple biological processes, and if responses are similarly timed in Eukaryotes, which have more complex hierarchy of pre- and post-transcriptional control mechanisms. Our work suggests that fine-grained control of transcriptional timing exists also in Eukaryotes.

The time course of gene expression responses often follows a typical *impulse* curve: starting with an initial abrupt response that saturates and is then followed by a relaxation to a new steady state. In this paper, we used this common behavior to build a parametric model that can be robustly fit to a single expression profile, while capturing the essential timing aspects of the response: its onset time, peak response and offset time.

Since the impulse model is tuned to typical cellular responses, it provides robust estimates of response characteristics, even when given very few samples per time course. We found that it provides superior prediction for imputing missing or corrupted measurements, both using single gene and using whole genome information. We believe that this model has other valuable uses, such as the alignment and comparison of time courses taken at different time points, or as the basis for determining a set of differentially expressed genes [9].

Perhaps most important, the impulse model allows us to study response timings directly. Using the distribution of onsets across functional categories, we found multiple functions that are timed differently from closely related functions. We also observed a global response pattern, roughly moving outwards

from the nucleus towards the cytoplasm and membranes. Finally, we found strong correlations between the onset of responses and the binding affinity to specific transcription factors. This last finding suggests a hypothesis in which gradual binding affinities are widely used by cells to tune the timing of expression responses, extending on recent findings in the context of specific pathways [34].

Transcriptional regulation is one mechanism in a series of hierarchical controls including regulation of mRNA, translation, and protein activation. Importantly, we note that our finding relates to overall mRNA levels, which encompass effects from both transcriptional changes and mRNA degradation. Our analysis is done purely on the timing in the change in mRNA levels, and we make no attempt to identify the cause(s) for the change. Indeed, it seems quite likely that many of the timing changes we observe result from a combination of these two regulatory mechanisms. Regardless of the mechanism, our findings suggest that the timing in fluctuations of gene expression levels is regulated in a way that optimizes for the role of the resulting protein product. For instance, the distribution of timing in Fig. 5 suggests a bifurcated response in the cytosolic ribosome: those components that are not required for translation of other protein products are repressed early, whereas the necessary components are repressed later, after fulfilling their role. Therefore, even though several regulatory phases separate mRNA levels from active protein levels, our findings support a model in which response onsets of mRNA are tuned with respect to the corresponding protein function.

The impulse model has its limitations. Since it is designed to capture typical two-transition responses, it could provide a poor fit to expression profiles that have more than two transitions, which we observe both in certain environmental response profiles (see Methods) and in the cyclic behavior of cell cycle. In addition, the impulse model is currently fit to noisy expression measurements without taking into account the physical mechanisms that lead to the observed expression levels. It would be interesting to explore ways in which prior knowledge regarding transcriptional or degradation dynamics can be integrated into the impulse model.

The impulse model captures one kind of typical response profiles, but other typical behavior may exist, such as the periodic behavior observed due to cell cycle. Such typical behaviors can be identified by unsupervised clustering of time courses, as in [13]. As in our analysis, one can then construct a specialized model that utilizes biologically relevant parameters that characterize that type of response, allowing these parameters to be extracted and used in further analysis.

Impulse-shaped responses are not limited to mRNA responses to stress. Sim-

ilar patterns are observed in gene expression profiles along early development [37] or in protein profiles. The modeling and visualizations techniques discussed in this paper could be usefully applied in these cases as well. Analysis of gene expression data can be used to analyze the dynamics of cellular networks, seeing how they adapt in response to changes in the cell condition. Recent work uses these data to obtain important insights into the dynamics of complex formation during the cell cycle [38]. We hope that the fine-grained timing information provided by our work will allow us to understand the reconfiguration of cellular complexes and pathways in response to environmental perturbations.

Materials and Methods

Learning an impulse model

The impulse model is a product of two sigmoids

$$\begin{aligned}
 f_\theta(x) &= \frac{1}{h_1} \cdot s_1(x) \cdot s_2(x) \\
 s_1(x) &= h_0 + (h_1 - h_0)S(+\beta, t_1) \\
 s_2(x) &= h_2 + (h_1 - h_2)S(-\beta, t_2) \\
 S(\beta, t) &= \frac{1}{1 + e^{-\beta(x-t)}} \\
 \theta &= [h_0, h_1, h_2, t_1, t_2, \beta] \quad .
 \end{aligned} \tag{2}$$

Clearly, other variants can be defined, such as using different slopes for the two sigmoids. With the data discussed in this paper, we found that such a model did not improve overall fit to data.

Fitting a single gene profile

We first consider the task of estimating the impulse model for the response profile of an individual gene. We assume that a gene's expression profile is given as a set of measurement (x_i, y_i) , where x_i is a particular time point, and y_i the expression value observed at that point. To estimate the parameters that best fit the gene's observed expression measurements, we search for the maximum likelihood parameter values, under an assumption of additive and independent Gaussian noise. Equivalently, we define an error function that we aim to minimize, which equals the negative of the log likelihood:

$$E = -\log P(D | \theta) = \frac{1}{2} \sum_i [f_\theta(x_i) - y_i]^2 + \text{const.} \tag{3}$$

This impulse function is differentiable with respect to all of the parameters of the model, and its derivatives $\frac{\partial f(x_i)}{\partial \theta}$ are given below. We therefore have the gradient of the error function

$$\frac{\partial E}{\partial \theta} = \sum_i [f(x_i) - y_i] \frac{\partial f(x_i)}{\partial \theta} \quad (4)$$

which we use with a conjugate gradient procedure to search for the optimal parameter set that minimizes the error function. Due to the form of the sigmoid and impulse function, the error function may have multiple local minima; 100 random restarts were used to find a good local minimum. Typically, many of the restarts converged to the same minimum which was also the best one found, suggesting that the error function tends to have a strong basin of attraction, which is likely the global optimum.

The gradient

We use the fact that the derivative of the sigmoid function $S(\beta, x) = 1/(1 + \exp(-\beta x))$, is $\frac{\partial S(\beta, x)}{\partial x} = (-\beta)S(\beta, x)[1 - S(\beta, x)]$. Using the auxiliary functions $s_1(x), s_2(x)$ as defined above, we obtain the gradient of the impulse function $f_\theta(x)$ with respect to θ at x :

$$\begin{aligned} \frac{\partial f}{\partial h_0} &= -\frac{1}{h_1} [1 - S(+\beta, t_1)] s_2 & (5) \\ \frac{\partial f}{\partial h_1} &= -\frac{1}{h_1^2} s_1 s_2 + \frac{1}{h_1} [S(+\beta, t_1) s_2 + s_1 S(-\beta, t_2)] \\ \frac{\partial f}{\partial h_2} &= -\frac{1}{h_1} [1 - S(-\beta, t_2)] s_1 \\ \frac{\partial f}{\partial t_1} &= \frac{1}{h_1} \left[-\beta(h_1 - h_0)S(+\beta, t_1)(1 - S(+\beta, t_1)) \right] s_2 \\ \frac{\partial f}{\partial t_2} &= \frac{1}{h_1} \left[-\beta(h_1 - h_2)S(-\beta, t_2)(1 - S(-\beta, t_2)) \right] s_1 \\ \frac{\partial f}{\partial \beta} &= +\frac{s_2}{h_1}(h_1 - h_0)(t_1 - x)S(+\beta, t_1)[1 - S(+\beta, t_1)] \\ &\quad + \frac{s_1}{h_1}(h_1 - h_2)(t_2 - x)S(-\beta, t_2)[1 - S(-\beta, t_2)]. \end{aligned}$$

Extracting response onset

We defined the onset of the response as the *time-to-half-peak* — the time at which the cell first reached half of its peak response, according to the fitted impulse model). More precisely, we first compute the peak of the profile (the maximum or the minimum, depending if the gene was activated or repressed); then we find the first time where the profile reaches half of this peak level. This

measure is widely used in analysis of sigmoidal functions and was found in our case to be robust to noise, as discussed in detail below.

We also experimented with other measures, and found the time-to-half-peak to be numerically more stable than other onset measures including: (1) *half-time-to-peak* — half the time it takes the curve to reach its peak; (2) the parameter t_1 of the impulse model; (3) *fastest-change-time* — the time with steepest curve (zero second derivative). The superior stability of the *time-to-half-peak* estimate is largely because it is numerically more stable to estimate the level of the peak than its time.

Importantly, the *time-to-half-peak* definition of the onset time is independent of the measurement scale, since rescaling a curve does not change the time it reaches its peak. As a result, the onset provides **orthogonal** information to the peak level.

Gene Expression Data

We collected 63 gene expression time courses from multiple published experiments, including responses to changing media and various types of environmental stress [23, 20, 24, 25, 26, 27, 28, 29]. We also included 13 new experiments with various media conditions. These experiments are detailed in another paper that is currently under review [19]. For completeness, we describe the experimental procedure below, and will remove this section in the final version of this manuscript.

We generated a set of 13 time courses by measuring gene expression following a metabolic change. Yeast strain KCN118 (MATalpha ade2) was grown at 28 C in 400 mL of synthetic complete media with 2% dextrose (SCD) to an OD600 of 0.6. Synthetic complete was prepared using the standard recipe, except 75 uM inositol was included. At OD600 of 0.6, 100 mL of cells were collected by centrifugation and frozen as a reference sample, and remaining cells were rapidly collected by filtration, washed with distilled water, and resuspended in 300 mL of one of the following media: SCE (SC+ 2% ethanol), SCG (SC + 2% galactose), SM1 (SCD lacking amino acids A, R, N, C, Q, G, K, P, S, F, and T), SM2 (SCD lacking amino acids L, I, V, W, H, and M), 14 S0 (SCD lacking all amino acids), S0G (no amino acids, 2% galactose), or S0E (no amino acids, 2% ethanol). To measure response profiles, 50 mL aliquots of resuspended yeast were added to 500 mL flasks shaking in a 28 degree water bath for 15, 30, 60, 120, or 240 minutes. At the indicated times, cells were collected by centrifugation for 2 minutes at 3700 rpm, and were flash frozen in liquid nitrogen. Poly-A RNA extraction, mRNA labeling, and cDNA microarray hybridization were performed as previously described 30.

For a detailed list of conditions see supplemental Table 1.

Model properties: Robustness, coverage, impulse-ness

The impulse-shape of expression responses is prevalent, suggesting that it could be used as a meaningful characterization of response profiles. But could an impulse function be accurately estimated from sparse and noisy samples?

To answer this question, we evaluate three aspects of the impulse model: How robustly it can be estimated from scarce data; what fraction of cellular responses it fits well; and how “impulse”-like are cellular responses to environmental changes.

Robustness

Microarray measurements of mRNA levels are notoriously noisy, often causing individual expression measurements to be unreliable. Importantly however, the variability in the estimate of the response onset from a time course is considerably lower than the variability of each individual measurement. This robustness results from two complementing effects: robustness to expression level noise, and robustness to timing noise.

Recall that we defined the onset-time as the time it takes to reach half-peak level. This definition of the onset is invariant to linear transformations of the data like rescaling and shift. Furthermore, since the onset is largely determined by the lowest and highest measurements, additive noise often has small effect on the estimate of onset time. To demonstrate this effect, we calculated the correlation between two sets of onset times: one extracted from timecourses measured in response to Peroxide exposure [20], and another extracted from a corrupted version of the same timecourses, achieved by adding Gaussian noise with zero mean and standard deviation of 0.1. The magnitude of the noise was chosen to reflect experimental noise observed between replicates [39].

Fig. 13(A) shows that the two estimates of the onsets are strongly repeatable (correlation coefficient is $\rho = 0.89$). Adding simulated noise to the measured expression levels can also be used as a procedure for estimating reliability of onset estimates from an individual expression profile: if adding noise results in large variation of the onset estimate, the estimate can be viewed as unreliable.

Second, onset time estimates are robust to timing noise, a crucial property for analyzing dynamical responses. Timing noise has multiple sources, both biological and experimental. For the current discussion, we consider timing noise that originates from variability in experimental conditions, and study the sensitivity of our onset estimates in face of such timing noise. In particular, we

tested the effects of timing noise on onset estimates, using a simple noise model. We consider the (unobserved) impulse curve that underlies the mRNA measurements, and added noise to it, in the form of convolution with a Gaussian curve that had a 2-minute standard deviation and a magnitude that was 20% of the impulse peak. When a sigmoid function is convolved with a Gaussian (or any symmetric function), the onset time of the original sigmoid and the convolved ones are the same. This fact is a simple consequence of the properties of a convolution of two symmetric functions. The result is that noise added to the timing of the mRNA transcription has little effect on the estimate of the onset from the mRNA time course. Fig. 13(B) illustrates this point, showing the convolution of an impulse curve (blue) with a 2-minute standard deviation Gaussian (red), and demonstrating that their resulting convolution Fig. 13(C) preserves the onset time.

Coverage

The impulse model is designed to capture a restricted set of expression response types. What is the fraction of the genome that is adequately described by the model?

To address this question we looked into the distribution of the normalized errors across genes. The normalized error is the L_2 prediction error, normalized by the standard deviation of the expression measurements. This measure yields a measure of error that is invariant to the scale of the expression levels.

We found that the impulse model was able to fit up to 95% of the genes in some conditions with an error as low as half a standard deviation of the profile variability (adenine starvation Fig. 14(A)). In a typical condition, the impulse model achieved this low error on 75% of the genes (Fig. 14(B)). The distribution of errors across all conditions is given in Fig. 14(C). Those conditions that had larger errors typically had more samples (hence are harder to fit), or were from irradiation experiments [29].

We complete the study of coverage by looking at the functional annotation of genes that are well described by an impulse behavior. We defined a set of K *impulse genes* to be the top K genes with lowest relative error, and tested for functional enrichment of this group using GO. We chose $K = 750$ since the number of significant categories peaked at this value.

In some conditions, impulse genes were enriched for GO categories relevant to the condition. For instance, under nitrogen depletion, the impulse genes were enriched for *amine metabolic process* $p < 5 \times 10^{-8}$ and *nitrogen compound biosynthetic process* $p < 2.5 \times 10^{-5}$. In other cases, environmental changes elicited generic responses, most notably the ribosome and its subunits ($p < 10^{-6}$, ob-

served in multiple stress conditions including DTT, diamide, and hypo-osmotic stress). Another category that repeated significantly was non-membrane-bound organelle ($p < 10^{-15}$, DTT, $p < 10^{-6}$ heat shock), and genes whose product are located in the cytosol ($p < 10^{-15}$ in DTT, heat shock, and gamma irradiation). A similar GO enrichment analysis for non-impulse genes (genes with bad impulse fits), did not reveal significantly enriched GO categories.

As could be expected, some genes do not follow a two-transition impulse response, and fitting their profiles with an impulse model may miss important components of the response. One family of such responses was observed in responses to KCl. Fig. 15 shows three examples, where response starts with an activation (repression in panel (C)), and later rebounds with a stronger repression (activation in panel (C)). The impulse model can be generalized to capture such three-transition profiles, in cases were the number of samples is sufficiently large to allow fitting a model with additional parameters.

Impulse-ness

The above experiments estimate the fraction of the genome that can be described by the model with a low error, but some of these genes may have profiles that are easy to fit by any model. We therefore used a Monte Carlo approach to estimate the fraction of the genes that are characteristically impulse-shaped, that is, they can be described considerably better with an impulse profile.

Our intuition is that some temporal profiles are easy to fit with small error. For example, genes that remain non-responsive to the induced stress, exhibit near constant expression profile, which is very easy to fit, but also easy to fit when measurements are shuffled in time. We therefore estimated the *Impulse-ness* of a gene, by measuring the extent to which it is easier to fit the original profile with an impulse model, as compared to time-shuffled timecourses with the same measurements.

In particular, we first fit an impulse model to each gene profile and calculated its error. We then randomly shuffled the expression measurements in time, fit an impulse model to the shuffled data, and calculated the fit error. We repeated the shuffling 100 times, yielding an estimate of the error distribution under the null hypothesis. We finally used this distribution to calculate a p -value for each gene. For the case of a non-responsive gene, both the original profile and its shuffled versions are easy to fit, hence in this case, the p -value assigned to this profile will be non-significant. On the other hand, in genes where the stress induces a pronounced impulse response (like those observed in Fig. 1), the impulse model can achieve low error, but many of the shuffled version will have multi-peak profiles, which cannot be fitted well with a single impulse. These genes will

therefore achieve a significant p -value.

The distribution of p -values across all 6209 genes under exposure to diamide [20] is shown in Fig. 16(A). We use this distribution to estimate model coverage, by calculating the excess in the fraction of genes observed for each p -value as compared to the expected random level. Under the null hypothesis of random errors, the expected distribution of p -values is flat (Fig. 16(A), black horizontal line), simply following the definition of a p -value as the probability of observing result under the null hypothesis. Many more genes in our model show smaller p -values than expected (red zone above the black line). Fig. 16(B) plots the cumulative distribution function (CDF) of the distribution in Fig. 16(A). The area of that zone provides the fraction of genes that are well described by the impulse model beyond what is expected at random, yielding that 54% of the genes exhibit behavior well captured by the model in diamide. Comparing this result to a similar analysis for other parametric families (estimated using the same procedure), we find that the impulse model provides a significantly better fit. In particular, 2nd order polynomials achieve no excess over the expected random level, and 3rd and 4th order polynomials achieve only a 15% level (data not shown). Under the same definition, Fig. 16(C) shows the distribution of excess coverage across the 76 conditions in our data, showing that on average 35% of the genes are above the baseline.

Comparisons with other modeling methods

Single gene profiles were fit (Fig. 2) using the impulse model, and compared with the following methods. (1) Polynomials fit of degree 2,3 and 4. The fitting procedure finds a polynomial of degree d that fits the data best in a least-squares sense. (2) Piecewise cubic Hermite interpolation, as calculated by the matlab function *interp1*. (3) Piecewise cubic spline interpolation, as calculated by the matlab function *interp1*. (4) Approximating splines calculated using code supplied by Bar-Joseph.

K Nearest Neighbors procedure

For K-nearest neighbor imputation (KNN-impute), we followed the approach of Troyanskaya *et al.* [30]. The known measurements are used to calculate distances between gene profiles, and the k nearest neighbors of each gene are identified. The missing measurement at a time t for gene g is estimated as the average of the time t expression values measured for the k genes most similar to g . KNN-impute uses a Euclidean distance over the vector of expression measurements to find the nearest neighbors.

To evaluate the impulse model in this context, we hid a randomly selected single time point in the expression profile of each gene, and used the remaining measurements to estimate the left-out values. Overall, this process resulted in a level of about 10–20% missing values, depending on the number of measurements in each time course. For each gene, we estimated the curve fit to the remaining measurements of that gene. We then estimated the value of a missing time t measurement for gene g by selecting the k genes nearest to g , using Euclidean distance over the predicted values, and averaging the predicted expression values at time t . Note that the predicted values were used both for selecting the neighbors and as a basis for estimating the time t value.

For comparison, we also applied the standard KNN-impute procedure to the same data. We used the on-line version of KNN-impute, available for download at <http://smi-web.stanford.edu/projects/helix/pubs/impute/>. We used $k = 15$, which is in the middle of the range of optimal values for k in the analysis of Troyanskaya *et al.*.

Identifying timed functions

To study the timeline of cellular responses, we identified GO categories that are timed distinctly earlier or later than other categories, using the following procedure. First, we defined a list of gene-sets pairs. The first set in each pair consisted of all genes in a GO category. We only considered medium size categories, and therefore ignored categories whose size was not between 1%–20% of the genome (60–1200 assigned genes). The second set in a pair, consisted of all genes in sibling categories (other children of the parent category). This set of genes provide a baseline to which the GO category can be compared.

Second, We collected the set of onset times for each gene set, based on all genes with relevant functional annotation. Finally, we used a Wilcoxon test to the timing difference between every pair of categories.

To produce our functional time lines, we needed to identify a subset of k categories that are strongly ordered. We chose to select the subset whose sum of pairwise scores is maximal. However, finding such an optimal set is computationally very costly, since it requires to go over all subset of size k (in fact, this is a case of the NP-Hard problem *max weighted clique*, that is believed to be impossible to solve efficiently for large instances). Instead, we followed a greedy procedure. We initialized the set with the two most distant categories (highest pairwise score), and repeatedly added a category whose sum of scores with the current set was maximal. We collected N categories with this procedures, and then manually pruned away categories that had high overlap (50%) with other categories in term of the number of genes, removing the category with

lower score. N was chosen to show many categories while avoiding clutter. This procedure yields interpretable results, as demonstrated in Fig. 6.

Correction for multiple hypotheses

We used false discovery rate (FDR) as originally described by Benjamini and Hochberg [40] to correct for multiple hypotheses. In some cases noted in the text, we used the more conservative Bonferroni correction for simplicity. In these cases, the reported upper bounds on the p -values were simply multiplied by the number of hypotheses.

Acknowledgments

This work was supported by the National Science Foundation under grant BDI-0345474. We are very grateful to Maya Schuldiner for useful comments and discussions. We thank Z. Bar-Joseph for sharing his splines clustering software, A. Regev and E. Segal for fruitful discussions of earlier versions of this work.

References

- [1] Androulakis I, Yang E, Almon R (2007) Analysis of time-series gene expression data: Methods, challenges, and opportunities. *Annual Review of Biomedical Engineering* 9:205–228.
- [2] Zhao L, Prentice R, Breeden L (2001) Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc Natl Acad Sci U S A* 98:5631–5636.
- [3] Alter O, Brown P, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A* 97:10101–10106.
- [4] Shedden K, Cooper S (2002) Analysis of cell-cycle gene expression in human cells as determined by microarrays and double-thymidine block synchronization. *Proc Natl Acad Sci U S A* 99:4379–4384.
- [5] Wichert S, Fokianos K, Strimmer K (2004) Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics* 20:5–20.
- [6] Bar-Joseph Z, Gerber G, Gifford D, Jaakkola T, Simon I (2003) Continuous representations of time series gene expression data. *J Comput Biol* 10:241–256.

[7] Luan Y, Li H (2003). Clustering of time-course gene expression data using a mixed-effects model with B-splines.

[8] Simon I, Siegfried Z, Ernst J, Bar-Joseph Z (2005) Combined static and dynamic analysis for determining the quality of time-series expression profiles. *Nature Biotechnology* 23:1503–1508.

[9] Storey J, Xiao W, Leek J, Tompkins R, Davis R (2005) Significance analysis of time course microarray experiments. *Proc Natl Acad Sci U S A* 102:12837.

[10] Ma P, Castillo-Davis C, Zhong W, Liu J (2006) A data-driven clustering method for time course gene expression data. *Nucleic Acids Res* 34:1261.

[11] Qian J, Dolled-Filhart M, Lin Y, Yu H, Gerstein M (2001) Beyond syn-expression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *J Mol Biol* 314:1053–66.

[12] Balasubramaniyan R, Hullermeier E, Weskamp N, Kamper J (2005) Clustering of gene expression data using a local shape-based similarity measure. *Bioinformatics* 21:1069–77.

[13] Ernst J, Nau G, Bar-Joseph Z (2005) Clustering short time series gene expression data. *Bioinformatics* 21:i159–i168.

[14] Holter N, Maritan A, Cieplak M, Fedoroff N, Banavar JR (2001) Dynamic modelling of gene expression data. *Proc Natl Acad Sci U S A* 98:1693–1698.

[15] Ramoni M, Sebastiani P, Kohane I (2002) Cluster analysis of gene expression dynamics. *Proc Natl Acad Sci U S A* 99:9121–9126.

[16] Schliep A, Schonhuth A, Steinhoff C (2003) Using hidden Markov models to analyze gene expression time course data. *Bioinformatics* 19:1264–72.

[17] Perrin B, Ralaivola L, Mazurie A, Bottani S, Mallet J, et al. (2003) Gene networks inference using dynamic Bayesian networks. *Bioinformatics* 19:138–148.

[18] Zou M, Conzen S (2005) A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* 21:71–79.

[19] Chechik G, Oh E, Rando O, Weissman J, Regev A, et al. (2008) Activity motifs reveal principles of timing in transcriptional control of the yeast metabolic network. *Nature Biotechnol* Accepted for publication.

[20] Gasch A, Spellman P, Kao C, Carmel-Harel O, Eisen M, et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* 11:4241–4257.

[21] Holter N, Mitra M, Maritan A, M Cieplak JB, Fedoroff N (2000) Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proc Natl Acad Sci U S A* 97:8409–8414.

[22] Spellman P, Sherlock G, Zhang M, Iyer V, Anders K, et al. (1998) Comprehensive identification of cell cycle regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* 9:3273–3297.

[23] DeRisi J, Iyer V, Brown P (1997) Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science* 278:680.

[24] Gasch A, Huang M, Metzner S, Elledge S, Botstein D, et al. (2001) Genomic expression responses to dna-damaging agents and the regulatory role of the yeast atr homolog meclp. *Mol Biol Cell* 12:2987–3003.

[25] Causton H, Ren B, Koh S, Harbison C, Kanin E, et al. (2001) Remodeling of yeast genome expression in response to environmental changes. *Mol Biol Cell* 12:323–337.

[26] Zakrzewska A, Boorsma A, Brul S, Hellingwerf K, Klis F (2004) Transcriptional Response of *Saccharomyces cerevisiae* to the Plasma Membrane-Perturbing Compound Chitosan. *Eukaryotic Cell* 4:703–715.

[27] Lai L, Kosorukoff A, Burke P, Kwast K (2005) Dynamical remodeling of the transcriptome during short-term anaerobiosis in *saccharomyces cerevisiae*: differential response and role of msn2 and/or msn4 and other factors in galactose and glucose media. *Mol Cell Biol* 25:4075–91.

[28] Kitagawa E, Akama K, Iwahashi H (2005) Effects of iodine on global gene expression in *saccharomyces cerevisiae*. *Biosci Biotechnol Biochem* 69:2285–2293.

[29] Mercier G, Berthault N, Touleimat N, Kepes F, Fourel G, et al. (2005) A haploid-specific transcriptional response to irradiation in *Saccharomyces cerevisiae*. *Nucleic Acids Res* 33:6635.

[30] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, et al. (2001) Missing value estimation methods for dna microarrays. *Bioinformatics* 17:520–525.

[31] Keene J (2007) RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet* 8:533–543.

[32] Tanay A (2006) Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Research* 16:962.

[33] Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* 431:99 – 104. URL <http://dx.doi.org/10.1038/nature02800>.

[34] Lam F, Steger D, O’Shea E (2008) Chromatin decouples promoter threshold from dynamic range. *Nature* 453:246.

[35] Zaslaver A, Mayo A, Rosenberg R, Bashkin P, Sberro H, et al. (2004) Just-in-time transcription program in metabolic pathways. *Nat Genet* 36:486–491.

[36] Kalir S, McClure J, Pabbaraju K, Southward C, Ronen M, et al. (2001). Ordering Genes in a Flagella Pathway by Analysis of Expression Kinetics from Living Bacteria.

[37] Wen X, Fuhrman S, Michaels G, Carr D, Smith S, et al. (1998) Large-scale temporal gene expression mapping of central nervous system development. *Proc Natl Acad Sci U S A* 95:334.

[38] de Lichtenberg U, Jensen L, Brunak S, Bork P (2005). Dynamic Complex Formation During the Yeast Cell Cycle.

[39] Hughes T, Marton M, Jones A, Roberts C, Stoughton R, et al. (2000) Functional Discovery via a Compendium of Expression Profiles. *Cell* 102:109–126.

[40] Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)* 57:289–300.

[41] O’Rourke S, Herskowitz I (2002) A Third Osmosensing Branch in *Saccharomyces cerevisiae* Requires the Msb2 Protein and Functions in Parallel with the Sho1 Branch. *Mol Cell Biol* 22:4739–4749.

A Supplemental Figures

supplementary figure 9

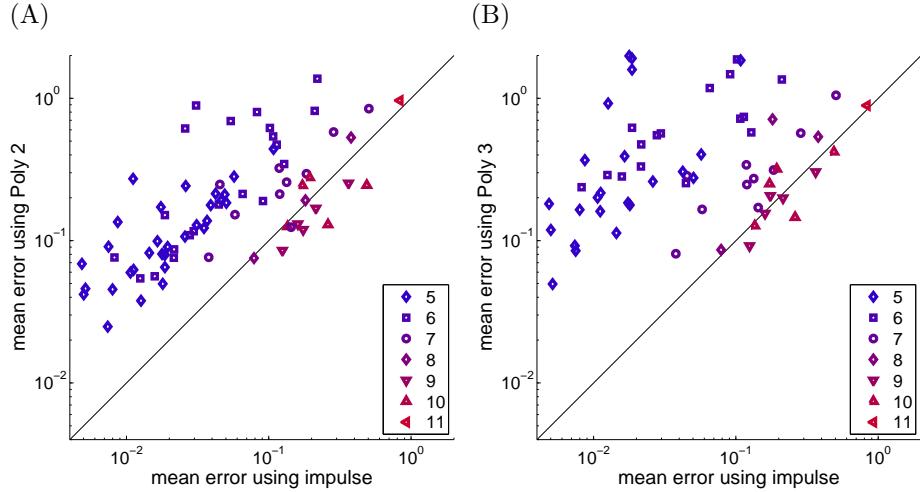


Figure 9: (A) Scatter plot of the mean error for each condition using the impulse model and polynomials. (A) 2^{nd} . (B) 3^{rd} order Polynomial.

supplementary figure 10

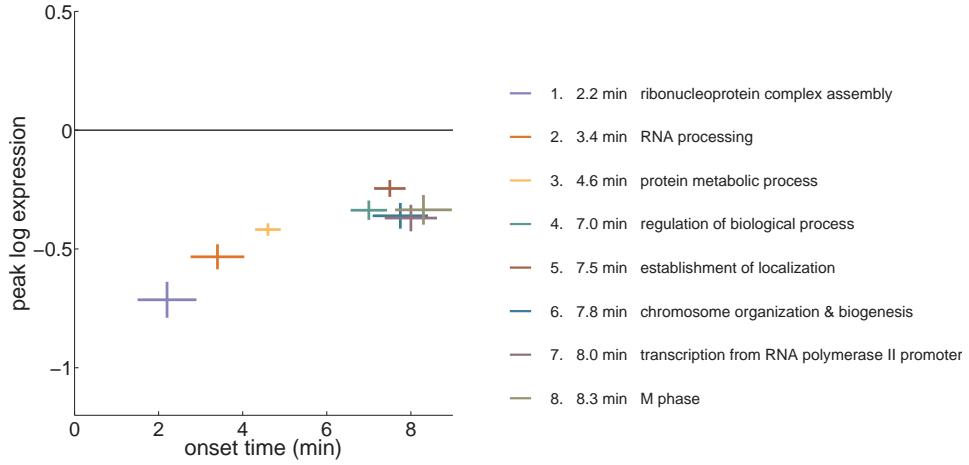


Figure 10: **A timeline of responses to gamma irradiation, biological processes.** Each cross denotes the median peak and onset time of all genes in the relevant GO category [24]. Length of bars denote the standard error of the mean (s.e.m.) across genes associated with the category. (see Methods).

supplementary figure 11

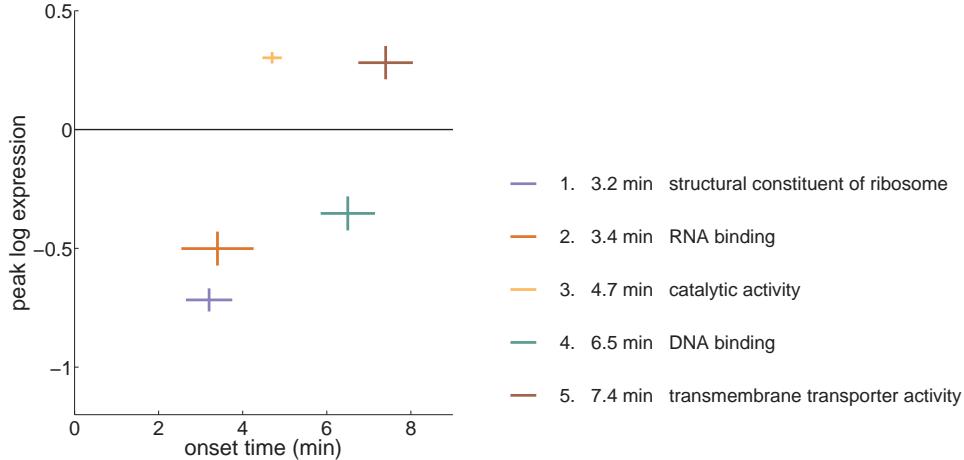


Figure 11: **A timeline of responses to gamma irradiation, molecular function.** Each cross denotes the median peak and onset time of all genes in the relevant GO category [24]. Length of bars denote the standard error of the mean (s.e.m.) (see Methods).

supplementary figure 12

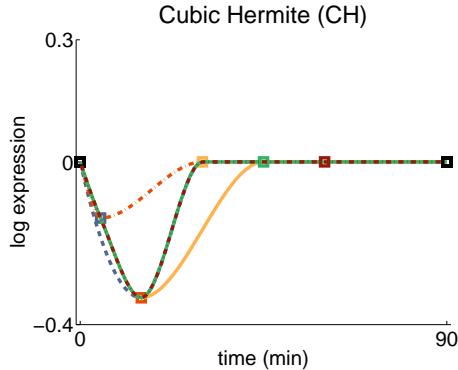


Figure 12: **Comparison of leave-one-out fits to a gene expression profile.** Same analysis as in Fig. 2 but for cubic Hermite interpolation (CH). CH interpolation is often heavily local: in this example, it is close to a piece-wise linear interpolation. Fig. 2(A) shows that, on average, this type of fit yields poor approximations in comparison with the impulse model. Each curve corresponds to a fit performed with a different single measurement that was left out during the fit. The color of each curve corresponds to the color of the hidden value (square marker).

supplementary figure 13

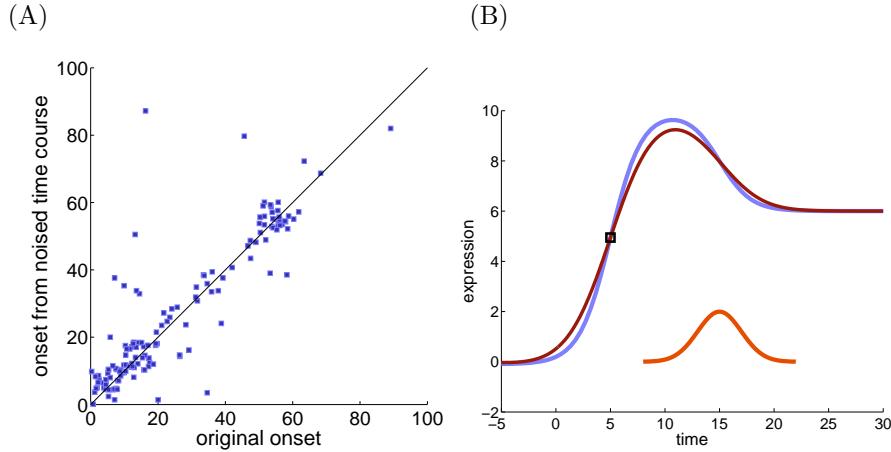


Figure 13: **(A) Robustness to expression level noise:** Onset extracted from corrupted time courses is highly correlated with the original onset (correlation coefficient is $\rho = 0.89$). Time courses were corrupted by additive Gaussian noise $N(0, 0.1)$. Each point corresponds to a gene; shown are differentially expressed genes (log expression > 1) under exposure to peroxide [20]. **(B) Robustness to timing noise:** Convolution of an impulse model (blue curve) that has an onset at 5 minutes (black square), with a Gaussian (red curve, 2 minutes standard deviation). The resulting convolution (purple curve), has essentially the same onset as the original impulse blue curve.

supplementary figure 14

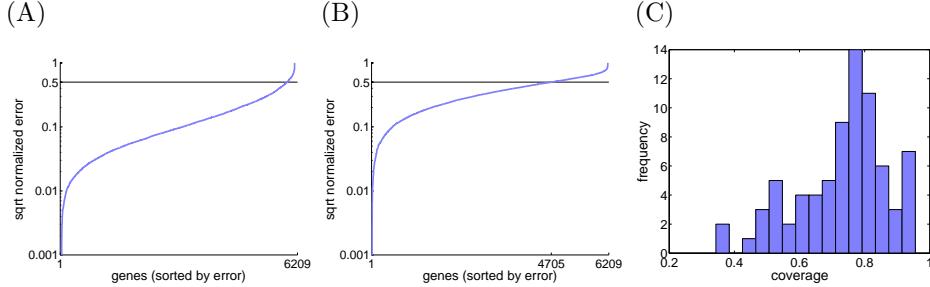


Figure 14: **Distribution of normalized error across all genes.** **(A)** The condition with largest number of genes with low error, adenine starvation [20]; more than 95% of the genes (5934) have a normalized error below half a standard deviation of the expression. **(B)** A condition with typical error profile (synthetic complete media with Ethanol and inositol). Condition was chosen as the median across conditions; 4705 genes (78%) had normalized variance below half a standard deviation. **(C)** Distribution of coverage (fraction of genes with error below cutoff) across conditions.

supplementary figure 15

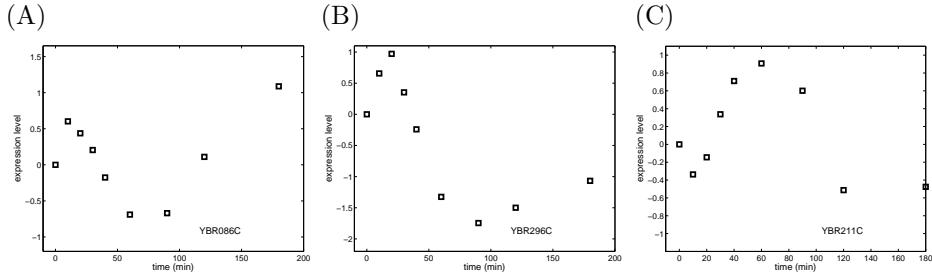


Figure 15: **Non-impulse responses to exposure to 1M KCl [41].** These genes follow at least three transitions.

supplementary figure 16

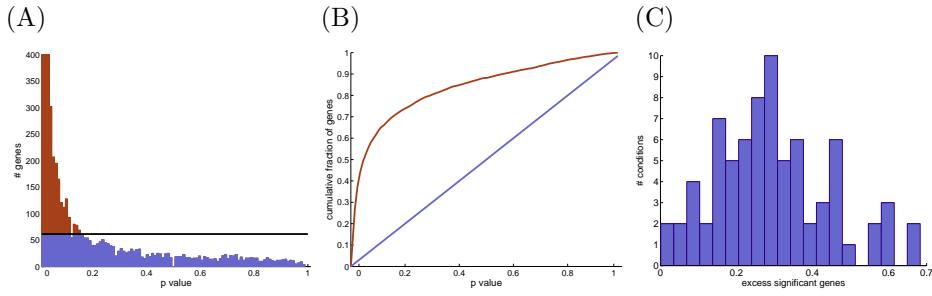


Figure 16: **Impulseness.** (A) Distribution of p -values for impulse model fitting for responses of yeast genes to diamide. The expected distribution of p -values (flat) is plotted as a black line, and the excess of significant genes over random is colored in red. (B) Cumulative distribution of the p -values (red) vs the expected in random (blue). (C) Distribution of the excess of significant genes across 76 stress conditions. It shows that the impulse model fits on average about 35% of the yeast genome above the baseline level.

B Supplemental Tables

Time course	samples	Time course	samples
[20]		[41]	
sorbitol 1 <i>M</i>	7	WT 0.125M KCl	5
diamide 1.5 <i>mM</i>	9	WT 0.25M KCl	5
menadione 1 <i>mM</i>	10	WT 0.5M KCl	10
25 °C	6	WT 1M KCl	9
DTT 2.5 <i>mM</i>	9	WT 1M Sorbitol	10
adenine starvation	6	WT alpha	5
hypo osmotic shock	6	hog1 0.125M KCl	5
nitrogen depletion	10	hog1 0.5M KCl	10
H2O2 0.32 <i>mM</i>	11	pbs2 0.5M KCl	5
DTT	8	sho1 0.5M KCl	5
heat shock	8	ssk1 0.0625M KCl	5
[24]		ste11 0.0625M KCl	5
MMS DES459 <i>mec1</i> 0.02	8	ssk1 0.125M KCl	5
MMS DES460 0.02	7	ssk1 0.5M KCl	5
<i>mec1</i> plus gamma	9	ssk1sho1 0.5M KCl	5
wt plus gamma	9	ssk1ste11 0.0625M KCl	5
[23]		ste11 0.5M KCl	5
diauxic shift	6	ssk1ste11 0.125M KCl	5
New Data		ste11 0.125M KCl	5
SD	6	ssk1ste11 0.25M KCl	5
SD+aa	6	ssk1ste11 0.5M KCl	5
SD+aaAR+ino	6	ssk1ste11 1M KCl	5
SD+aaLI+ino	6	[27]	
SD+ino	6	WT Gal N2 rep1	5
SEtOH	6	WT Gal N2 rep2	5
SEtOH+aa	6	WT Gal N2 rep3	5
SEtOH+aa+ino	6	WT Glu N2 rep1	5
SEtOH+ino	6	WT Glu N2 rep2	5
Sgal	6	WT Glu N2 rep3	5
Sgal+aa	6	msn2/4 Galactose rep1	5
Sgal+aa+ino	6	msn2/4 Galactose rep2	5
Sgal+ino	6	msn2/4 Galactose rep3	5
[25]		WT Gal N2 mean	5
Acid	7	WT Glu N2 mean	5
Alkali	7	msn2/4 Galactose mean	5
Heat	6	[29]	
NaCl	6	18733 200GY	7
Peroxide	6	18734 200Gy	7
Sorbitol	6	18735 200Gy	7
[26]		6053 200Gy	7
Chitosan	6	hdf1 LM79 200Gy	7

Table I: List of time courses that were analyzed. conditions are grouped by publication.

Adenine starvation (Gasch 2000)	ADR1 SM, ARG80 SM, ARG81 SM, ARO80 SM, BAS1 SM, CAD1 SM, CBF1 SM, CHA4 SM, DAL81 SM, DAL82 SM, FHL1 SM, GAT1 SM, GCN4 SM, GCR2 SM, GLN3 SM, HAP4 SM, HAP5 SM, LEU3 SM, MET28 SM, MET31 SM, MET32 SM, UGA3 SM, MET4 SM, MOT3 SM, PHO2 SM, PUT3 SM, RAP1 SM, STP1 SM, RCS1 SM, RPH1 SM, RTG1 SM, RTG3 SM, SFP1 SM, SIP4 SM,
Heat shock (Gasch 2000)	ADR1 HEAT, GAT1 HEAT, HSF1 HEAT, MSN2 HEAT, SKN7 HEAT, YAP1 HEAT,
Heat (Causton 2001)	ADR1 HEAT, GAT1 HEAT, HSF1 HEAT, MSN2 HEAT, SKN7 HEAT, YAP1 HEAT
Acid (Causton 2001)	MSN2 Acid, MSN4 Acid

Table II: List of matching TF-condition pairs. The left column gives the gene expression condition. The right column gives the name of the name of the TF the conditions at which it was tested.

Early repressed ribosomal gene in Fig. 5 (onset before 13 min)
ARD1, CAM1, EGD1, FUN12, GCN1, GCN20, HEF3, MAP1, NAT1, NIP7, NMD3, PAT1, RLI1, RPL10, RPL13B, RPL31B, RPL35A, RPL37B, RPL39, RPL40A, RPL41A, RPP1B, RPS19B, RPS20, RPS25A, RPS25B, RPS28A, RPS28B, RPS30B, RPS31, SIS1, SQT1, YGR054W
Late repressed ribosomal gene in Figure Fig. 5 (onset after 13 min)
BTT1, GCN2, RPL11A, RPL11B, RPL12A, RPL12B, RPL13A, RPL14A, RPL14B, RPL15A, RPL15B, RPL16A, RPL16B, RPL17A, RPL17B, RPL18A, RPL18B, RPL19B, RPL1A, RPL1B, RPL20A, RPL20B, RPL21A, RPL21B, RPL22A, RPL23A, RPL23B, RPL24A, RPL24B, RPL25, RPL26A, RPL26B, RPL27A, RPL27B, RPL28, RPL2B, RPL3, RPL30, RPL31A, RPL32, RPL33A, RPL33B, RPL34B, RPL35B, RPL36A, RPL37A, RPL38, RPL40B, RPL42A, RPL42B, RPL43A, RPL4A, RPL4B, RPL5, RPL6A, RPL6B, RPL7A, RPL7B, RPL8A, RPL8B, RPL9A, RPL9B, RPP0, RPP1A, RPP2A, RPP2B, RPS0A, RPS0B, RPS10A, RPS10B, RPS11A, RPS11B, RPS12, RPS13, RPS14A, RPS14B, RPS15, RPS16A, RPS16B, RPS17A, RPS17B, RPS18A, RPS18B, RPS19A, RPS1A, RPS1B, RPS2, RPS21A, RPS21B, RPS22A, RPS22B, RPS23A, RPS23B, RPS24A, RPS24B, RPS26A, RPS26B, RPS27A, RPS27B, RPS29A, RPS29B, RPS3, RPS4A, RPS4B, RPS5, RPS6A, RPS6B, RPS7A, RPS7B, RPS8A, RPS8B, RPS9A, RPS9B, STM1, TIF5

Table III: Early repressed genes (onset before 13 minutes) and late ones in figure Fig. 5.

B SUPPLEMENTAL TABLES

Condition	GO category (median time)	parent (median time)
Gasch sorbitol	translation (10.00)	gene expression (9.40)
Gasch diamide	translation (10.75) rib. (16.30) translation (10.75) intracellular mb. org. (6.10) mb. org. (6.10) biopolymer bios. p. (5.80) cytosolic rib. (19.10) organellar rib. (6.15) translation (10.75) cellular bios. p. (8.90) rib. (16.30) ribosomal subunit (17.65) structural constituent of rib. (18.10) translation (10.75)	gene expression (6.50) intracellular non-mb. org. (6.60) macromolecule bios. p. (7.40) intracellular org. (6.30) org. (6.30) macromolecule bios. p. (7.40) rib. (16.30) rib. (16.30) cellular protein metabolic p. (7.50) bios. p. (7.45) rib.nuc.prot complex (8.90) rib.nuc.prot complex (8.90) structural molecule activity (8.25) cellular bios. p. (8.90)
Gasch dtt	intracellular mb. org. (31.35) mb. org. (31.35) translation (54.55) translation (54.55) rib. (59.90) translation (54.55) rib.nuc.prot complex (51.10) ribosomal subunit (61.30) translation (54.55) biopolymer bios. p. (28.40) cytosolic part (61.60) rib. biog. and assembly (53.50) biopolymer metabolic p. (31.60) rib. (59.90) cytosolic rib. (62.00) cellular bios. p. (45.00)	intracellular org. (32.85) org. (32.85) cellular protein metabolic p. (36.65) macromolecule bios. p. (36.50) intracellular non-mb. org. (44.30) gene expression (37.60) macromolecular complex (33.50) rib.nuc.prot complex (51.10) cellular bios. p. (45.00) macromolecule bios. p. (36.50) cytosol (55.60) org. organization and biog. (32.70) macromolecule metabolic p. (33.70) rib.nuc.prot complex (51.10) rib. (59.90) bios. p. (36.65)
Gasch heat shock	intracellular mb. org. (7.10) mb. org. (7.10) translation (10.00) cytosolic rib. (11.15)	intracellular org. (7.40) org. (7.40) cellular bios. p. (8.50) rib. (10.20)
Gasch MMS	macromolecule metabolic p. (7.60)	metabolic p. (10.00)
Gasch wt+gamma	rib. biog. and assembly (2.00) rRNA metabolic p. (1.80)	org. organization and biog. (5.70) RNA metabolic p. (6.20)
Rando SD	ribosomal subunit (11.25)	rib.nuc.prot complex (10.00)
Rando SD+AR+I	intracellular mb. org. (9.60) mb. org. (9.60) biopolymer metabolic p. (9.20) rib. (12.60) nucleolus (7.65) ribosomal subunit (13.10) rib. (12.60) translation (12.15) translation (12.15)	intracellular org. (9.90) org. (9.90) macromolecule metabolic p. (10.00) intracellular non-mb. org. (10.70) intracellular non-mb. org. (10.70) rib.nuc.prot complex (11.80) rib.nuc.prot complex (11.80) macromolecule bios. p. (11.10) cellular protein metabolic p. (11.20)
Rando SD+I	nucleolus (7.10) translation (11.60) rib. (11.70) intracellular mb. org. (9.00) ³⁹ mb. org. (9.00) ribosomal subunit (12.40) translation (11.60)	intracellular non-mb. org. (10.05) gene expression (9.60) intracellular non-mb. org. (10.05) intracellular org. (9.20) org. (9.20) rib.nuc.prot complex (10.70) cellular protein metabolic p. (9.95)
Causton Heat	nuclear lumen (15.00)	org. lumen (14.40)

B SUPPLEMENTAL TABLES

Condition	GO category (median time)	parent (median time)
Gasch sorbitol	translation (10.00)	gene expression (9.40)
Causton Peroxide	intracellular mb. org. (9.60) mb. org. (9.60) ribosomal subunit (18.95) structural constituent of rib. (19.10) rib. (17.45) cytosolic part (19.10) nucleolus (12.05) cytosolic rib. (19.90) translation (15.60) nucleoplasm (9.35)	intracellular org. (9.80) org. (9.80) rib.nuc.prot complex (14.10) structural molecule activity (14.80) rib.nuc.prot complex (14.10) cytosol (15.70) nuclear lumen (10.20) rib. (17.45) cellular bios. p. (12.90) nuclear lumen (10.20)
ORourke 0.0625MKCl	nucleolus (1.60) nucleolus (1.60)	intracellular non-mb. org. (2.10) nuclear part (1.70)
ORourke 0.125MKCL	rib. (4.75) translation (4.50) biopolymer metabolic p. (2.60) intracellular mb. org. (2.80) mb. org. (2.80) nucleolus (2.45) RNA p.ing (2.50) nuclear part (2.10) protein metabolic p. (3.40)	intracellular non-mb. org. (2.90) gene expression (3.10) macromolecule metabolic p. (2.90) intracellular org. (2.90) org. (2.90) intracellular non-mb. org. (2.90) gene expression (3.10) intracellular org. part (2.80) macromolecule metabolic p. (2.90)
ORourke 0.5MKCl	organellar rib. (4.30) cytosolic rib. (10.45) translation (9.65) intracellular mb. org. (6.10) mb. org. (6.10) structural constituent of rib. (10.20)	rib. (9.90) rib. (9.90) macromolecule bios. p. (5.90) intracellular org. (6.45) org. (6.45) structural molecule activity (9.00)
ORourke 1Msorb	cytosolic rib. (12.10) organellar rib. (6.75) intracellular mb. org. (8.70) mb. org. (8.70) translation (10.75) cellular bios. p. (10.00) rib. (11.40)	rib. (11.40) rib. (11.40) intracellular org. (8.90) org. (8.90) macromolecule bios. p. (8.80) bios. p. (9.10) intracellular non-mb. org. (9.70)
ORourke hog1 0.5MKCl	intracellular mb. org. (6.80) mb. org. (6.80) translation (11.10) rib. (13.50) organellar rib. (3.00) cytosolic rib. (14.60) cellular bios. p. (9.70) translation (11.10) rib. (13.50)	intracellular org. (7.10) org. (7.10) macromolecule bios. p. (7.30) intracellular non-mb. org. (8.50) rib. (13.50) rib. (13.50) bios. p. (7.80) cellular protein metabolic p. (8.40) rib.nuc.prot complex (10.30)
ORourke pbs2 0.5MKCl	intracellular mb. org. (7.20) mb. org. (7.20) rib. (10.20) translation (9.85) translation (9.85) translation (9.85) rib. (10.20) ribosomal subunit (10. ⁴⁰ 30) cellular bios. p. (9.10) translation (9.85) biopolymer bios. p. (6.20) biopolymer metabolic p. (7.10) protein modification p. (6.80)	intracellular org. (7.40) org. (7.40) intracellular non-mb. org. (8.40) gene expression (7.80) macromolecule bios. p. (7.85) cellular protein metabolic p. (7.80) rib.nuc.prot complex (9.30) rib.nuc.prot complex (9.30) bios. p. (7.90) cellular bios. p. (9.10) macromolecule bios. p. (7.85) macromolecule metabolic p. (7.40) cellular protein metabolic p. (7.80)

Condition	GO category (median time)	parent (median time)
ORourke sho1 0.5MKCl	translation (9.70) translation (9.70) rib. (9.90) intracellular mb. org. (7.10) mb. org. (7.10) organellar rib. (5.75) cellular bios. p. (9.20) cytosolic rib. (10.10) rib. (9.90)	cellular protein metabolic p. (8.00) macromolecule bios. p. (7.40) intracellular non-mb. org. (8.35) intracellular org. (7.40) org. (7.40) rib. (9.90) bios. p. (7.70) rib. (9.90) rib.nuc.prot complex (9.10)
ORourke ssk1sho1 0.5MKCl	translation (9.80) translation (9.80) cytosolic rib. (10.45) rib. (10.00) intracellular mb. org. (7.10) mb. org. (7.10) cellular bios. p. (9.00) translation (9.80) biopolymer bios. p. (6.55)	macromolecule bios. p. (7.90) gene expression (8.10) rib. (10.00) intracellular non-mb. org. (8.15) intracellular org. (7.40) org. (7.40) bios. p. (8.10) cellular protein metabolic p. (8.20) macromolecule bios. p. (7.90)
ORourke ssk1ste11 0.5MKCl	rib. (9.15) translation (9.00) intracellular mb. org. (6.10) mb. org. (6.10) nucleolus (4.55) translation (9.00) translation (9.00) RNA p.ing (5.40)	intracellular non-mb. org. (6.85) gene expression (6.60) intracellular org. (6.40) org. (6.40) intracellular non-mb. org. (6.85) macromolecule bios. p. (7.20) cellular protein metabolic p. (7.60) gene expression (6.60)
ORourke ssk1ste11 1MKCL	rib. (39.20) biopolymer bios. p. (27.70) cellular bios. p. (35.00) translation (35.85)	intracellular non-mb. org. (31.10) macromolecule bios. p. (30.40) bios. p. (31.60) macromolecule bios. p. (30.40)
ORourke ste11 0.5MKCl	cytosolic rib. (10.40) organellar rib. (5.85) intracellular mb. org. (7.80) mb. org. (7.80)	rib. (10.05) rib. (10.05) intracellular org. (8.00) org. (8.00)
Lai Gal R2	organellar rib. (4.50) cytosolic rib. (10.50)	rib. (9.90) rib. (9.90)
Lai Gal R3	organellar rib. (4.65)	rib. (9.50)
Lai Mean	organellar rib. (5.15) cytosolic rib. (9.75)	rib. (9.15) rib. (9.15)

Table IV: **Differentially-timed sibling GO categories.** List of GO categories whose timing is significantly different than sibling categories (Wilcoxon test, $p < 10^{-5}$, Bonferroni corrected). Only categories with 30–300 genes were considered. Notations: ‘p.’ for ‘process; ‘bios’ for ‘biosynthetic’, ‘biog’ for ‘biogenesis’, ‘org.’ for ‘organelle’, ‘i.c.’ for ‘intracellular’, ‘rib’ for ‘ribosome’, ‘m.b.’ for ‘membrane-bound’.