Materials and Methods

Learning an impulse model

The impulse model is a product of two sigmoids

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

Other variants of this model 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:

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

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

$$\frac{\partial E}{\partial \theta} = \sum_i [f(x_i) - y_i] \frac{\partial f(x_i)}{\partial \theta} \tag{3}$$

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.

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 (DeRisi et al., 1997; Gasch et al., 2000, 2001; Causton et al., 2001; Zakrzewska et al., 2004; Lai et al., 2005; Kitagawa et al., 2005; Mercier et al., 2005). We also included 13 new experiments with various media conditions. These experiments are detailed in (Chechik et al., 2008).

For a detailed list of conditions is in supplemental Table 1 online http://robotics.stanford.edu/ gal/Research/
.

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 (Gasch et al., 2000), 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 (Hughes et al., 2000).

Fig. 5(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. 5(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. 5(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. 6(A)). In a typical condition, the impulse model achieved this low error on 75% of the genes (Fig. 6(B)). The distribution of errors across all conditions is given in Fig. 6(C). Those conditions that had larger errors typically had more samples (hence are harder to fit), or were from irradiation experiments (Mercier et al., 2005).

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 ($p < 10^{-6}$, observed 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. 7 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 (Gasch et al., 2000) is shown in Fig. 8(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. 8(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. 8(B) plots the cumulative distribution function (CDF) of the distribution in Fig. 8(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, 2^{nd} order polynomials achieve no excess over the expected random level, and 3^{rd} and 4^{th} order polynomials achieve only a 15% level (data not shown). Under the same definition, Fig. 8(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. **??**) 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.* (Troyanskaya et al., 2001). 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. **??**.

Correction for multiple hypotheses

We used false discovery rate (FDR) as originally described by Benjamini and Hochberg (Benjamini and Hochberg, 1995) 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

Benjamini, Y. and 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**(1):289–300.

Causton, H., Ren, B., Koh, S., Harbison, C., Kanin, E., Jennings, E., Lee, T., True, H., Lander, E., and Young, R., *et al.*, 2001. Remodeling of yeast genome expression in response to environmental changes. *Mol. Biol. Cell*, **12**(2):323–337.

Chechik, G., Oh, E., Rando, O., Weissman, J., Regev, A., and Koller, D., 2008. Submitted.

DeRisi, J., Iyer, V., and Brown, P., 1997. Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science*, **278**(5338):680.

Gasch, A., Huang, M., Metzner, S., Elledge, S., Botstein, D., and Brown, P., 2001. Genomic expression responses to dna-damaging agents and the regulatory role of the yeast atr homolog mec1p. *Mol Biol Cell*, **12**(10):2987–3003.

Gasch, A., Spellman, P., Kao, C., Carmel-Harel, O., Eisen, M., Storz, G., Botstein, D., and Brown, P., 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, **11**:4241–4257.

Hughes, T., Marton, M., Jones, A., Roberts, C., Stoughton, R., Armour, C., Bennett, H., Coffey, E., Dai, H., He, Y., *et al.*, 2000. Functional Discovery via a Compendium of Expression Profiles. *Cell*, **102**(1):109–126.

Kitagawa, E., Akama, K., and Iwahashi, H., 2005. Effects of iodine on global gene expression in saccharomyces cerevisiae. *Biosci Biotechnol Biochem*, **69**(12):2285–2293.

Lai, L., Kosorukoff, A., Burke, P., and 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**(10):4075–91.

Mercier, G., Berthault, N., Touleimat, N., Kepes, F., Fourel, G., Gilson, E., and Dutreix, M., 2005. A haploid-specific transcriptional response to irradiation in Saccharomyces cerevisiae. *Nucleic Acids Res*, **33**(20):6635.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R., 2001. Missing value estimation methods for dna microarrays. *Bioinformatics*, **17**:520–525.

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

supplementary figures

supplementary figure 6

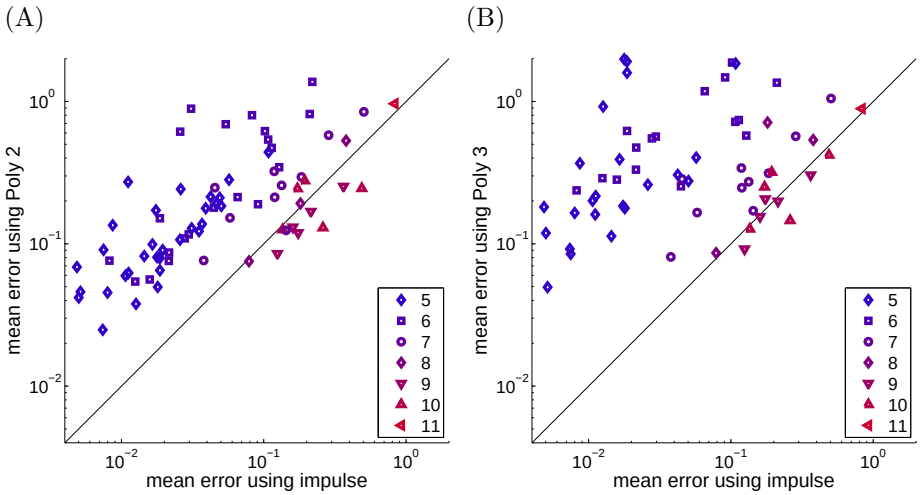

Figure 1: **(A)** Scatter plot of the mean error for each condition using the impulse model and polynomials. **(A)** 2^{nd} . **(B)** 3^{rd} order Polynomial.

supplementary figure 7

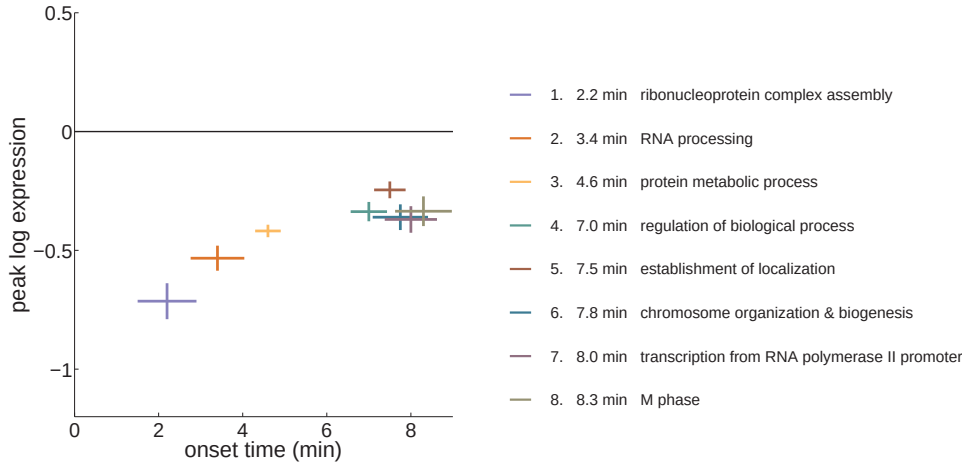

Figure 2: **A timeline of responses to gamma irradiation, *biological processes*.** Each cross denotes the median peak and onset time of all genes in the relevant GO category (Gasch, 2001). Length of bars denote the standard error of the mean (s.e.m.) across genes associated with the category.

supplementary figure 8

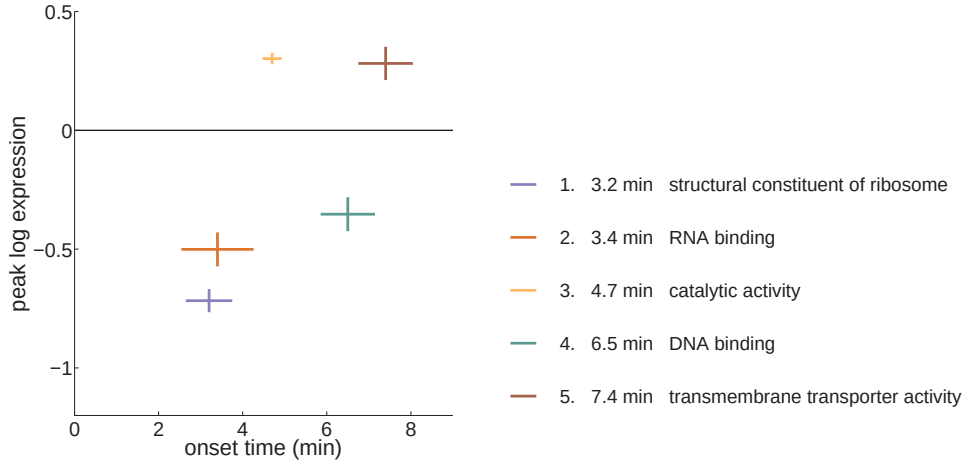

Figure 3: **A timeline of responses to gamma irradiation, *molecular function*.** Each cross denotes the median peak and onset time of all genes in the relevant GO category (Gasch, 2001). Length of bars denote the standard error of the mean (s.e.m.) .

supplementary figure 9

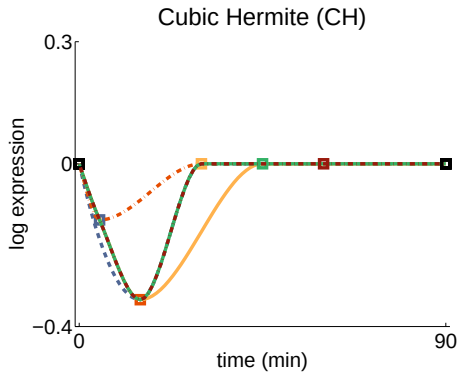

Figure 4: **Comparison of leave-one-out fits to a gene expression profile**. Same analysis as in Fig 2 but for cubic Hermite interpolation (CH). CH interpolation is often heavily local: in this example, it is close to a piece-wise linear interpolation. Fig 2(A) shows that, on average, this type of fit 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 10

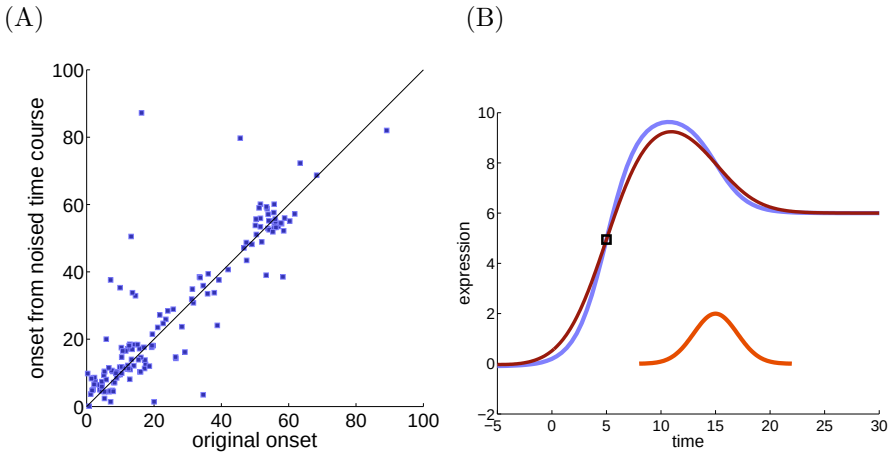

Figure 5: **(A) Robustness to expression level noise:** Onset extracted from corrupted time courses is highly correlated with the original onset (correlation coefficient is $\rho = 0.89$). Time courses were corrupted by additive Gaussian noise $N(0, 0.1)$. Each point corresponds to a gene; shown are differentially expressed genes (log expression > 1) under exposure to peroxide (Gasch, 2000).**(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 11

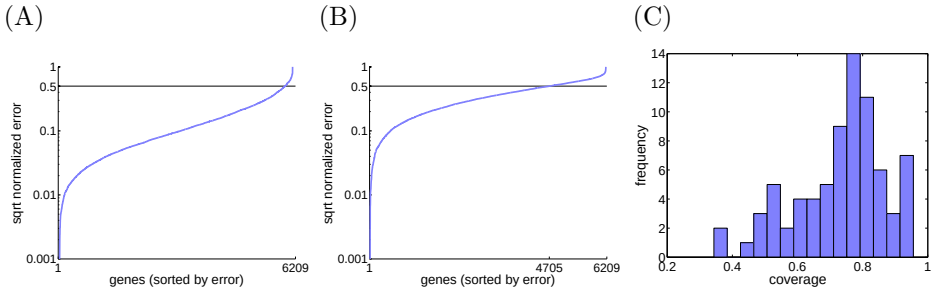

Figure 6: **Distribution of normalized error across all genes. (A)** The condition with largest number of genes with low error, adenine starvation (Gasch, 2000); more than 95% of the genes (5934) have a normalized error below half a standard deviation of the expression. **(B)** A condition with typical 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 12

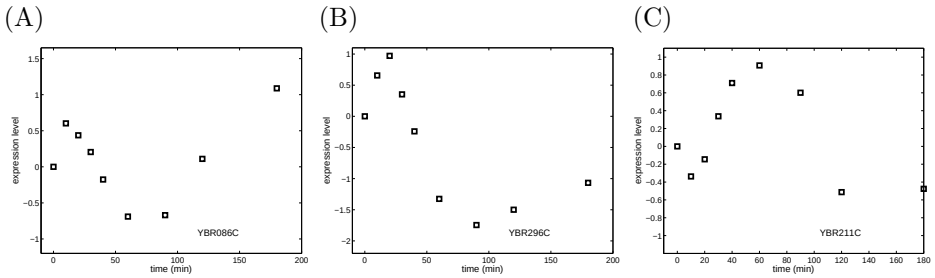

Figure 7: **Non-impulse responses to exposure to 1M KCl (ORourke, 2002).** These genes follow at least three transitions.

supplementary figure 13

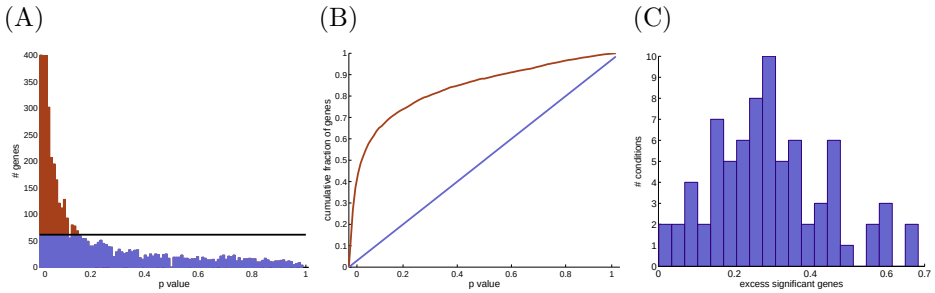

Figure 8: **Impulseness. (A)** Distribution of p-values for impulse model fitting for responses of yeast genes to diamide. The expected distribution of p-values (flat) is plotted as a black line, and the excess of significant genes over random is colored in red. **(B)** Cumulative distribution of the 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.