Molecular Cell, *Volume 45*

Supplemental Information

Cellular Noise Regulons Underlie Fluctuations in *Saccharomyces cerevisiae*

Jacob Stewart-Ornstein, Jonathan S. Weissman, and Hana El-Samad

Supplementary Experimental Procedures

Construction and Treatment of Strains to Measure Gal1 Promoter Noise

An estradiol inducible system similar to that constructed by Louvion et al. (1993) was integrated at the LEU2 locus in a *w303a* strain background. The construct consists of the Adh1 promoter driving GAL4DBD (1-93AA) fused to the human estrogen receptor hormone binding domain and the activation domain of MSN2(1-303AA)). A GAL1 promoter driving YFP was then integrated at the HIS3 locus, and a second GAL1 promoter driving either YFP or mCherry was integrated at the TRP1 locus. Strains were grown at 30C shaking to saturation, diluted 1:200 into SDC and grown for 12hrs, followed by a dilution of 1:200 into SDC with estradiol ranging from 100nm to 3.7nm in a log1.6 dilution series in 96 well plates (Costar). The cells were allowed to grow for 8hrs before cytometry measurement (OD~=0.05). Intrinsic and extrinsic noise values were computed as described below.

Heat Shock Survival

To induce a range of MSN2/4 target gene expression a constitutively active allele of MSN2 (5A) [] was placed under a GAL1 promoter and induced using an estradiol inducible system in a strain with MSN2/4 deleted (*msn2::NAT, msn4::KAN*). These cells were induced with estradiol concentrations from 1 to 100nm and gown at 30C for 6hrs in 200ul of SDC in 96 well plates, these inductions resulted in 8-log2s of diversity in mean expression of MSN2 target gene HSP12-RFP. 50ul of the induced cells were measured on the flow cytometer, the remaining 150ul were heat shocked to 50C and 5ul sampled into 200ul of SDC at various times. These cells were grown for 12-16hrs at 30C, the number of cells in each well measured by cytometry and the ratio of cells at each temperature and MSN2/4 expression were calculated to give survival rates.

Overexpression of PDE2 and MSN2

To over-express PDE2 or MSN2(5A), the appropriate sequence was amplified from the genome and placed in a vector with a GAL1 promoter, this construct was then integrated at the TRP1 locus in cells expressing an estradiol inducible system as described above. Cells were induced using concentrations of estradiol ranging between 1 and 200nm in SDC for 6hrs before measurement.

Construction and Treatment of One GFP and Two GFP strains for Noise Decomposition

High-expressing genes were selected from the GFP library and checked for growth phenotypes (see table1 for strain list). These strains were mated to an SGA strain with the TRP1 gene deleted by insertion of a URA marker with the promoter of TFS1 driving mKate2 (MATα his3 leu2 met15 ura3 can1::STE3pr-HIS3 lyp1::STE2pr-LEU2 TRP1::TFS1pr-mKate2(Ura3MX)). Diploids were then selected for in SD–Trp/-Ura. The successful diploid strains were sporulated in liquid spo medium at room temperature for 5 days, then transferred to haploid selective medium (SD-Lys-Arg-His-Ura-Leu+SAEC+Canavanine). Two subsequent dilutions into haploid selective medium (total dilutions 1:100000) resulted in pure populations of haploid MATα strains (MATα *his3 leu2 met15 ura3* can1::STE3pr-HIS3 lyp1::STE2pr-LEU2 TRP1::TFS1pr-mKate2(Ura3MX), X::GFP(HIS3MX)), where X is the gene tagged with GFP.

Following sporulation, strains were mated in liquid to either the original GFP strains, or to a BY4741 strain with a his3::TDH3pr-mCherry(HIS3MX). After overnight growth in YPD, selection for diploids was carried out with SD–Trp/-Ura. This resulted in two sets of diploid strains, one with two GFPs and an equivalent diploid strain (1XUra3, 2xHis3) with only one GFP copy. Additionally, the diploid strain harboring one GFP copy is labeled with a bright RFP such that it is distinguishable from the 2XGFP strain in a mixed culture. When co-cultured the two strains could be separated well *in silico* by RFP expression level (<1% of cells were ambiguous in all cases).

The 1-GFP and 2-GFP strains were then mixed (~1:1) in shallow 96 well plates (COSTAR) and grown to saturation overnight in 200ul SD-Trp/Ura. They were then diluted 1:200 and grown for 12hrs, followed by a second dilution of 1:200, which was allowed to grow for 8hrs before measurement (final OD ~0.05-0.1). In all cases, growth was at 30C with orbital shaking.

Dual Color Strain Construction and Covariance Measurements

For crossing to generate dual color strains, individual RFP marked strains were were constructed by PCR based homologous insertion of an RFP protein (mCherry or mKate2) at the C-Terminal end of the open reading frame with a URA3 marker immediately 3'. Reporter strains marked with mCherry (or mKate2) were crossed to strains from the GFP library (Open Biosystems), with a selection step for diploids in SD -Ura/-His. Pure diploid populations were verified as containing no cells lacking either fluorophore using flow cytometry.

Briefly, a RFP marked strain was gown up in 5ml YPD, and GFP strains grown in 96 or 384 well plates (200 or 80ul of YPD). These cells were mixed in liquid on 96 or 384 well plates and incubated at 30C for 24hrs to allow for mating. Diploids were then selected by 1:8000 fold dilution into SD-URA-HIS using a Biomech robotic system. These strains were then diluted 1:32000 using a Biomech robotic system and grown for 20-22 hrs before measurement (final OD ~0.05-0.1), the 96 or 384 plates were measured using an HTS autosampler on the LSRII cytometer (BD). Plate growth was at 30C with orbital shaking on Elim heater shakers.

Hetrozygous Deletion Strain Construction and Analysis

Strains for HSP12 and PGM2 covariance measurements were constructed by deletion of the stated genes (IRA2, GPB, PDE2, RAS2) by homologous recombination with a NAT marker from a PGM2-mCherry MATα strain (MATα *his3 leu2 met15 ura3* can1::STE3pr-HIS3 lyp1::STE2pr-LEU2 PGM2-mCherry (URA3) YFG::NAT). This strain was then crossed to (MATa *his3 leu2 met15* HSP12-GFP(HIS3)) to construct heterozygous strains expressing PGM2-mCherry and HSP12-GFP and selected for in SD-HIS-URA.

To construct diploids hetrozygous strains, a MATα strain with tagged copies of ARG4 and CIT1 (MATα *his3 leu2 met15 ura3* can1::STE3pr-HIS3 lyp1::STE2pr-LEU2 CIT1-mCherry (URA3) ARG4-GFP (HIS3)) was crossed to deletion strains from the Yeast deletion collection (MATa *his3 leu2 met15 YFG::*KAN). Diploids were selected for in two steps with an initial selection in YPD+G418(200ug/ml) followed by selection in SD-URA+G418(500ug/ml). A complete list of these strains is in table S4.

To study the effects of heterozygous deletions of key genes on the covariance between ARG4 and many different GFP tagged genes, we crossed an ARG4-mCherry strain with a specific gene deletion (MATα *his3 leu2 met15 ura3* can1::STE3pr-HIS3 lyp1::STE2pr-LEU2 ARG4-mCherry (URA3) YFG::NAT) to the 182 strains used in the covariance matrix analysis. Selection for diploids was in SD–URA-HIS.

These strains were inoculated into shallow 384 well plates (Fisher) and grown to saturation overnight in 80ul SD-URA/HIS. They were then diluted 1:32000 using a Biomech robotic system and grown for 22-24 hrs before measurement (final OD ~0.05-0.1), the 384 plates were measured using an HTS autosampler on the LSRII cytometer (BD). In all cases, growth was at 30C with orbital shaking. Flow cytometery data for these strains was collected, the data analyzed as described below.

Data Processing and Analysis of Flow Cytometry Data

Flow cytometry data was first processed to remove outliers by filtering the data using a minimum covariance determinant method [Rousseeuw and Van Driessen, 1999] (~80-90% of original cells are kept). Other approaches to filtering data or calculating 'robust' measures of variance and mean (e.g. Mean Absolute Deviation or S-estimators) were explored and gave similar, though less reproducible, outcomes. In practice these approaches proved similar to gating as both remove the most extreme outliers.

To test that filtering was not artificially inflating or reducing measured correlations or variance, we generated random samples of a given size drawn from a large set of cells (~1 million) and calculated medians, variance, and covariance. A sample was either MCD filtered or gated by forward and side scatter such that an equivalent percentage of the population is included. All methods used converged to similar medians and variances as the MCD filtering method.

Second, to correct for variations in cell size (and cell cycle), we utilized forward and side scatter information (FSC and SSC). Traditionally this is done by selecting only a subset of cells with similar size/shape (gating), and calculating means and covariances from this subset of data [Newman et al., 2005]. As the gate size becomes infinitely small (i.e. all cells within the gate have identical forward and side scatter properties), any covariance measured will be independent of these parameters. On the other hand, as the gate becomes smaller and the number of measured cells decreases errors magnify. Further gating typically discards the vast majority of the data (80-99% in some reports).

We addressed this problem in two ways. First, we viewed the measurements from a particular experiment as a population from which to draw samples. We selected a single cell randomly and found the N cells that are most similar to it in forward and side scatter space (based on a distance metric with a determined threshold). We then computed variance and covariance from this sample of cells. Note that by taking a fixed number of nearby cells, the error in computing the covariance for each group is constant, but the degree of similarity between cells within a group is not. By repeating this procedure many times (500 was sufficient to produce sampling errors in measured covariance of <1%), robust estimates of variance and covariance were computed. The size of N was empirically set at 100, as further reductions result in negligible changes in covariance. We will refer to this procedure as multi-gating.

The second approach is to view the flow-cytometry (GFP, RFP, FSC, SSC) measures as a multi-dimensional dataset and determine using partial correlations the covariance between GFP and RFP (or

variance within GFP or RFP) measurements given the correlation due to cell size as measured by side scatter. This is similar in spirit to a recent analysis of cell-to-cell variation in flow cytometry data [Rinott et al., 2011]. This assumes relatively linear and normal relationship between side scatter and fluorescence. In practice, this approach results in estimates for mean and noise that are similar to a simple linear transformation of the data by dividing by side scatter or other measures of cell size which other groups have adopted [Murphy et al., 2010;, Raser and O'Shea, 2005].

Multi-gating and partial covariance based approaches for estimating noise and covariance in practice are nearly equivalent (R^2 in a given plate is typically ~0.97). Partial covariance is however much less data intensive and typically requires no more than 1000-2000 cells to produce good estimates. A comparison between partial correlation and mutual information also found very good agreement (R^2 of 0.92, N=192), suggesting that this data is well explained by linear relationships.

To evaluate our ability to measure noise and covariance reproducibly, we examined the error across biological replicates (for example, duplicate correlation measurements of a large set of genes with ribosomal reporter RPL17B (S1)). These measurements suggest a median error of ~0.025 (R^2 of replicates of 0.92). The variation in correlation could result from intrinsic measurement errors or variation in the cellular state of biological replicates. Our metric of sensitivity showed somewhat greater variation (R^2 of 0.88, N=362), probably as it incorporates multiple potentially noise measures (two covariance measures, two means). Similar errors were observed for stress reporter replicates (correlation R^2 of 0.90, sensitivity R^2 0.87, N=365).

Formulas for Noise Decomposition Using the One GFP and Two GFP Strains

To extract values for extrinsic and intrinsic noise from fluorescent measurements in the 1-GFP and 2-GFP strains, we used the following formula:

$$Var(GFP_1 + GFP_2) = 2 * Var(GFP_1) + 2 * Cov(GFP_1, GFP_2)$$

Here, GFP_1 and GFP_2 are the two (identical) copies in the 2-GFP diploid strains.
Therefore, $\text{var}(GFP_1) = \text{var}(GFP_2)$ and $Cov(GFP_1, GFP_2)$ is given by

$$Cov(GFP_1, GFP_2) = [Var(GFP_1 + GFP_2)] / 2 - Var(GFP_1))$$

Also, measurements in these diploid strains show that $mean(GFP_1 + GFP_2) = 2 * mean(GFP_1)$. As a result,

$$CV(GFP_1) = \frac{\sqrt{Var(GFP_1)}}{Mean(GFP_1)}$$

$$CV(GFP_1 + GFP_2) = \frac{\sqrt{Var(GFP_1 + GFP_2)}}{2 * Mean(GFP_1)}$$

We define extrinsic noise as: $CV_{ext} = \frac{\sqrt{Cov(GFP, GFP)}}{Mean(GFP)}$

Therefore,

$$CV_{ext}^2 = \frac{Cov(GFP_1, GFP_2)}{Mean(GFP_1)^2} = \frac{Var(GFP_1 + GFP_2)}{2 * Mean(GFP_1)^2} - \frac{Var(GFP_1)}{Mean(GFP_1)^2} = 2 * CV^2(GFP_1 + GFP_2) - CV^2(GFP_1)$$

where we have again used the fact that that $mean(GFP_1 + GFP_2) = 2 * mean(GFP_1)$. This finally gives the expression for extrinsic noise as:

$$CV_{ext} = \sqrt{2 * CV^2(GFP_1 + GFP_2) - CV^2(GFP_1)}$$

And therefore, intrinsic noise is given by:

$$CV_{\text{int}} = \sqrt{CV(GFP_1)^2 - CV_{ext}^2}$$

The quantities above were computed after subtraction of background fluorescence, which was obtained by measurement of isogenic diploid strains without any GFP tagged proteins.

Error in measurement of noise components was computed using replicate measurements of the dataset. The R^2 between replicate measurements of total noise (CV) was 0.96, while that for extrinsic noise was 0.85. Specific wells were discarded if they did not meet criteria for number of cells, absence of non-florescent cells, and clear separation of the two populations; remaining replicates were averaged (table 1).

Extrinsic Noise Error Models and Number of Noisy Genes

The CV for extrinsic noise showed similar reproducibility to the covariance measurements (R^2 of 0.85 for extrinsic noise, N=468, S1). Residuals from a least square fit to replicates of the extrinsic noise measurements showed a Gaussian distribution centered on zero with a standard deviation of 0.022.

To estimate the number of genes with measurably elevated (or reduced) noise, we noted that the histogram of extrinsic noise values [S1B] could be represented as the convolution of two Gaussian distributions: the first is centered around 0.1 and the second contains the subset of genes with greater extrinsic noise. This would be consistent with a global noise 'floor' experienced by all genes, with a fraction of genes experiencing noise greater than this floor value.

To verify that the data is well represented by this model, we pursued two strategies to estimate the mean and standard deviation of this noise 'floor'. First, we estimated the mean by finely binning the data and selecting the largest bin (the mode) as the mean. The standard deviation was estimated by subtracting this value from the dataset, data points with less than zero extrinsic noise after this operation likely represent the 'noise' in measurement so we reflected them across the zero line and computed the standard deviation of this transform. Alternatively, we used an estimate of measurement noise computed from

replicate data. The two approaches gave indistinguishable results and produced an excellent Gaussian fit to the first peak in the extrinsic noise histogram. This suggests that points within this distribution are consistent with measurement noise, whereas data falling outside represents genuine elevation of extrinsic noise. Using this approach, we found that 84 genes representing ~1/5 of the dataset had elevated extrinsic noise (>2*std + median, Table S1).

Consistent with this characterization, extrinsic noise for replicates of the 25% (N=108) of genes with lowest extrinsic noise agreed with an R^2 of 0.0701, whereas the replicates of the 25% highest extrinsic noise genes agreed with an R^2 of 0.9145. We do not observe the same differential for intrinsic noise—genes with the highest intrinsic noise agreed with R^2 of 0.90 and the lowest with 0.73—suggesting that only extrinsic noise exhibits a floor in expression noise and that intrinsic noise has measureable signal across its dynamic range.

Metrics of Relatedness among Genes from Single-Cell Measurements (S-score)

We assume that in an exponentially growing population, gene expression can be approximated as a linear process. The expression of a given gene in a single cell (b_i) is then equal to the sum of a basal expression (β_i), the activity of upstream signaling pathways in that cell (x_i) multiplied by the gene's susceptibility to that pathway (α), and a noise term (ε_i). The following derivation will show that it is not possible to get an absolute value for α from single cell noise data, but that it is possible to measure the strength of gene *b*'s response to pathway *x* relative to a reporter gene (*a*) which is simultaneously measured in the same cell (i.e. you can find α_b/α_a). We define:

$$a_i = \beta_1 + \alpha_a x_i + \varepsilon_{1i}$$
$$b_i = \beta_2 + \alpha_b x_i + \varepsilon_{2i}$$
$$x_i = \bar{x} + \theta_i$$
$$\varepsilon_i = Norm(0, \sigma_i), i = 1,2$$
$$\theta = Norm(0, \sigma_\theta)$$
$$Corr(\varepsilon_1, \varepsilon_2) = Corr(\varepsilon_2, \theta) = Corr(\varepsilon_1, \theta) = 0$$

a_i = expression of gene *a* in cell *i*

b_i = expression of gene *b* in cell *i*

β = basal expression of a gene

x_i =activity of pathway *X* in cell *i*

α =change in a gene in response to pathway X

Under the assumption that noise terms are uncorrelated, we obtain:

$$\bar{a} = \beta_1 + \alpha_a \bar{x}$$
$$\bar{b} = \beta_2 + \alpha_b \bar{x}$$

The covariance of *a* and *b* can also be computed:

$$Cov(a,b) = \overline{(a_i - \bar{a})(b_i - \bar{b})}$$
$$Cov(a,b) = \overline{(\alpha_a \theta_i + \varepsilon_{1i})(\alpha_b \theta_i + \varepsilon_{2i})}$$
$$Cov(a,b) = \overline{\alpha_a \theta_i \varepsilon_{2i} + \alpha_b \theta_i \varepsilon_{1i} + \alpha_a \alpha_b \theta_i^{\,2} + \varepsilon_{1i} \varepsilon_{2i}}$$
$$Cov(a,b) = \alpha_a \alpha_b \sigma_x^{\,2}$$

Now, if we assume that variance in gene expression is proportional to mean expression level of the gene, then the total variance in *a* and *b* is given by

$$\sigma_a^{\,2} = \alpha_a^{\,2} \sigma_x^{\,2} + \phi \bar{a} + \gamma_a$$
$$\sigma_b^{\,2} = \alpha_b^{\,2} \sigma_x^{\,2} + \phi \bar{b} + \gamma_b$$

Where the first term represent the pathway noise from *x*, and subsequent terms represent intrinsic noise (whose variance is proportional to mean expression – See Fig1), and extrinsic noise from other processes unrelated to *x*. The correlation between *a* and *b* can be written as:

$$Corr(a,b) = \frac{\alpha_a \alpha_{b2} \sigma_x^{\,2}}{\sqrt{(\alpha_a^{\,2} \sigma_x^{\,2} + \phi \bar{a} + \gamma_a)} \sqrt{(\alpha_b^{\,2} \sigma_x^{\,2} + \phi \bar{b} + \gamma_b)}}$$

Note that the correlation depends on the mean expression level of *a* and *b*, indicating that as the mean expression level of the genes increase their correlation will also decrease. Correlation has a fundamental bias towards weakly expressed genes, and is therefore problematic as a measure of relatedness. Correlation does however describe in an unbiased way the proportion of variance shared by a given gene pair and is useful to compare across different genes. To achieve a measure of relatedness independent of mean expression we turn to the covariance.

$$Cov(a,a) = \alpha_a^{\,2} \sigma_x^{\,2}$$
$$\frac{Cov(a,b)}{Cov(a,a)} = \frac{\alpha_a \alpha_b \sigma_x^{\,2}}{\alpha_a^{\,2} \sigma_x^{\,2}} = \frac{\alpha_b}{\alpha_a} = G$$
$$S = G * \frac{\bar{a}}{\bar{b}}$$

From the covariance we define two metrics of relatedness, the gain (G) of a particular gene relative to a given reporter, and the sensitivity (S) of a gene to fluctuations in the reporter as a proportion of its mean.

To derive the expressions above, we made two main assumptions. First, we assumed that gene expression could be modeled as a linear process. This assumption seems to be warranted for the Msn2/4 regulated genes which respond linearly to changes in the activity of Msn2/4. Indeed when a strain containing two distinct Msn2/4 reporters is tested under many different conditions (deletion strains) the two reporters across experiments were linearly related (R^2 =0.59, data not shown).

The second assumption is that the noise terms (ε) are independent. In our experiments, these terms represent noise in measurement and also factors not explicitly accounted for in the model of gene expression such as global correlators. For example, cell size (or transcriptional/translation capacity) can uniformly affect the expression of all genes in a given cell and therefore induce interdependence. We attempted to control for this explicitly in our flow cytometry data analysis by only processing cells with relatively equivalent dimensions in forward and side scatter.

Supporting this assumption, we find that on average the correlation of the ~700 genes we measured to the stress responsive gene Pgm2 was 0.0471 (+/-0.0022), looking at ribosomal genes only this correlation drops further to 0.0113 (+/-0.0042). These results suggest that most of the noise from global fluctuations has been accounted for with our data analysis approach, validating that our normalized covariance reports on pathway interactions between genes.

A General Linear Framework for the Relationship between Genes in Different Transcriptional Modules

We can extend and generalize the above analysis to deal with genes which share upstream regulatory pathways, but not specific transcription factors. As our example we focus on the MSN2 stress responsive pathway as it is a noisy system amenable to experimental manipulation.

Let's consider a simple model of gene expression for PGM2 (target gene for Msn2) and some gene G in a single cell. This model is similar to the model we use to derive the 'S-score', but contains some additional terms to account for experimental manipulations to the system. We write:

$$PGM2_i = C_i * (\alpha_p x_i + \varepsilon_{1i})$$
$$G_i = C_i * (\alpha_g y_i + \varepsilon_{2i})$$

Here, x_i and y_i are active transcription factors (TF) upstream of these genes in cell *i,* α is the sensitivity of PGM2 or ARG4 to their TFs, C_i is a measure of transcriptional capacity in that particular cell (often referred to as cell-to-cell global variation), and ε terms represent variability not specific to any process. We express active TF as an average value perturbed in every cell by two sources of variability: one originating from the regulatory pathway upsteam of the TF and the other originating from fluctuations in TF expression. Therefore, we write:

$$x_i = MSN2_a + \theta_{xi} + \kappa_i$$
$$y_i = A + \theta_{yi} + \eta_i$$

As a result:

$$PGM2_i = C_i * [\alpha_p (MSN2_a + \theta_{xi} + \kappa_i) + \varepsilon_{1i}]$$
$$G_i = C_i * [\alpha_g (A + \theta_{yi} + \eta_i) + \varepsilon_{2i}]$$

We assume that the fluctuations are uncorrelated. More precisely,

$C_i = \bar{c} + \theta_{ci}$

$\text{cov}(C, x) = \text{cov}(C, y) = \text{cov}(C, \varepsilon) = \text{cov}(\varepsilon, \theta) = 0$

$\bar{\theta} = \bar{\kappa} = \bar{\varepsilon} = 0$

Case 1:

If $PGM2$ and G have independent TFs and uncorrelated signaling pathways upstream of the TFs, then

$$Cov(PGM2, G) = \alpha_a * \alpha_g * A * MSN2_a * \sigma_c^{\ 2}$$

This correctly recapitulates that these two genes should only be correlated through global fluctuations.

Case 2:

If $PGM2$ and G have independent TFs and partially correlated signaling pathways upstream of the TFs (that is they are influenced by fluctuations in signaling pathways upstream of Msn2, but also fluctuations coming from some other uncorrelated pathway z), then $\theta_{yi} = \theta_{xi} + \theta_{zi}$. In this case,

$$Cov(PGM2, G) = \alpha_a * \alpha_g * A * MSN2_a * \sigma_c^{\ 2} + \alpha_a * \alpha_g * (\sigma_c^{\ 2} + \bar{c}^2) * \sigma_x^{\ 2}$$

Here, $\sigma_x^{\ 2} = \text{var}(\theta_x)$. In this case, which corresponds to $ARG4$, an experimental assay which titrates increasing amounts of $MSN2_a$ is expected to show a **linear increase** in $Cov(PGM2, G)$ (first term of expression above) as a function of $PGM2$ mean expression ($\overline{PGM2}$). Since signaling upstream is unaffected, the second term is an offsetting constant.

In contrast, upon titration of a component of signaling upstream in the PKA pathway (e.g. PDE2), one one would expect $\sigma_x^{\ 2}$ to change. As a result, a **nonlinear** change in $Cov(PGM2, G)$ is expected as a function of $\overline{PGM2}$.

Case 3:

If $PGM2$ and G are both targets of Msn2 (and by definition have the same signaling pathway upstream), then

$$Cov(PGM2, G) = \alpha_a * \alpha_g * MSN2_a^{\ 2} * \sigma_c^{\ 2} + \alpha_a * \alpha_g * (\sigma_c^{\ 2} + \bar{c}^2) * \sigma_x^{\ 2} + \alpha_a * \alpha_g * (\sigma_c^{\ 2} + \bar{c}^2) * \sigma_\kappa^{\ 2}$$

This case corresponds to $HSP12$. Here, direct manipulation of $MSN2_a$ should increase $Cov(PGM2, HSP12)$ quadratically as a function of $\overline{PGM2}$. However, since such a such a

manipulation would be expected to change σ_κ^2 (last term in equation above), then a linear relationship between $Cov(PGM2, HSP12)$ and $\overline{PGM2}$ would **NOT** be expected as $MSN2_a$ changes.

On the other hand, titration of a component of signaling upstream in the PKA pathway (e.g. PDE2), should change σ_x^2 but not σ_κ^2. Given the structure of these equations, this case is not differentiable from titration of $MSN2_a$.

These predictions can be verified by placing MSN2 (constitutively active) under the control of the GAL1 promoter and titrating its expression with an Estradiol inducible system . In this assay, the DNA binding domain of the GAL1 activator GAL4 is fused to the ligand binding domain of the human estrogen receptor. The nuclear localization and activity of this fusion transcription factor is modulated by estradiol (see supp. Methods). Using this assay, we indeed determined that overexpression of MSN2 results in linear (r^2=0.93) increase of the COV(PGM2, ARG4) as predicted by the model. In contrast, $Cov(PGM2, HSP12)$ increased nonlinearly and non-loglinearly as a function of $\overline{PGM2}$ (fig. S4). These results argue that perturbations that produce non-linear changes in covariance are perturbations which affect *both* genes and arise from pathways common to these genes, in this case the PKA pathway and more generally experimentally validate our simple linear approach to covariance analysis.

Covariance Matrix Analysis

To obtain the covariance matrix described in Figure 3 of the main text, we normalized every covariance by dividing by GFP and RFP mean expression– (the RFP is nearly constant across any given set of measurements we obtain very similar results if we normalize by GFP expression alone). Data was generally collected in duplicates. Replicates disagreeing by more than 30% were discarded, and all remaining replicates were averaged. From this analysis we obtained a 182 by 45 matrix of interaction scores (Table S3). This matrix was hierarchically clustered using a Pearson correlation metric for distance.

To analyze this data with principal component analysis, a row-wise correlation of this matrix was computed to give a 182x182 correlation matrix (using the Matlab `corrcoef` function set to ignore NAN values) on which hierarchical clustering (using a correlation distance metric and the Matlab `linkage` function with *'complete'* settings) and PCA analysis (`PCACOV` function) was performed using standard Matlab functions.

Query genes tagged with mCherry were as follows: ADE6, ARG4, ATP1, CAR2, ENO1, ENO2, GAP1, GOR1, HSP104, HSP12, HSP26, INO1, MET6, PGM2, SER3, SOD2, TDH2, URA1, ADE4, AGP1, ARG8, ATP14, ATP5, CIT1, FAS1, HOM2, HOR2, HSP42, HSP82, HTB2, LYS21, MET14, MUP1, PDR16, POR1, PRX1, SAM1, SAM2, SER33, SOL4, SSB2, SUR7, TSA1, URA4, ZWF1.

The Amino Acid group consisted of: MET6, SER1, BNA3, MET10, ARG4, GAP1, AGP1, YOL098C, TIF4632, HIS4, PFK2, HTB1, SSA4, ARG5,6, BAT2, ACO2, LYS21, SAM3, ARG8, ARG3, MET2,

LYS20, SER33, MET22. The Mitochondrial group: FAA3, COX15, DDR48, LSC2, NDI1, COR1, SOD2, HEM15, CIT1, MDH1, ATP14, ATP4, ATP3, PUT2, LPD1, PAM17, ALO1, ISU1, RUD3, STR3, GCV1, MSS51, RIB4, MRPL13, ISU2, ABF2. And the stress response group of: GDB1, GPH1, TPS3, HSP104, SSA1, ENO1, GPD1, CWP2, PIL1, PBI2, YDL124W, GSY2, PGM2, GLK1, HOR2, TPS1, RTC3, HOR7, NOP8, HSP12, TDH1.

Supplementary References

Louvion JF, Havaux-Copf B, Picard D. (1993) Fusion of Gal4-VP16 to a steroid-binding domain provides a tool for gratuitous induction of galactose responsive genes in yeast. Gene **131**:129-134.

Murphy KF, Adams RM., Wang X., Balazsi G., Collins J.J. (2010). Tuning and controlling gene expression noise in synthetic gene networks. Nucleic Acids Research, 1–15

Newman JR, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, Weissman JS. (2005) Single-cell proteomic analysis of S. cerevisiae reveal the architecture of biological noise. Nature **441**, 840-846.

Raser JM, O'Shea EK. (2004) Control of Stochasticity in Eukaryotic Gene Expression. Science **304**: 1811-1814.

Rinott R, Jaimovich A, and Friedman N. (2011) Exploring transcriptional regulation through cell-to-cell variability. PNAS **108**(15): 6329-6334.

Rousseeuw, P.J. and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," Technometrics, 41, pp. 212-223.

S1

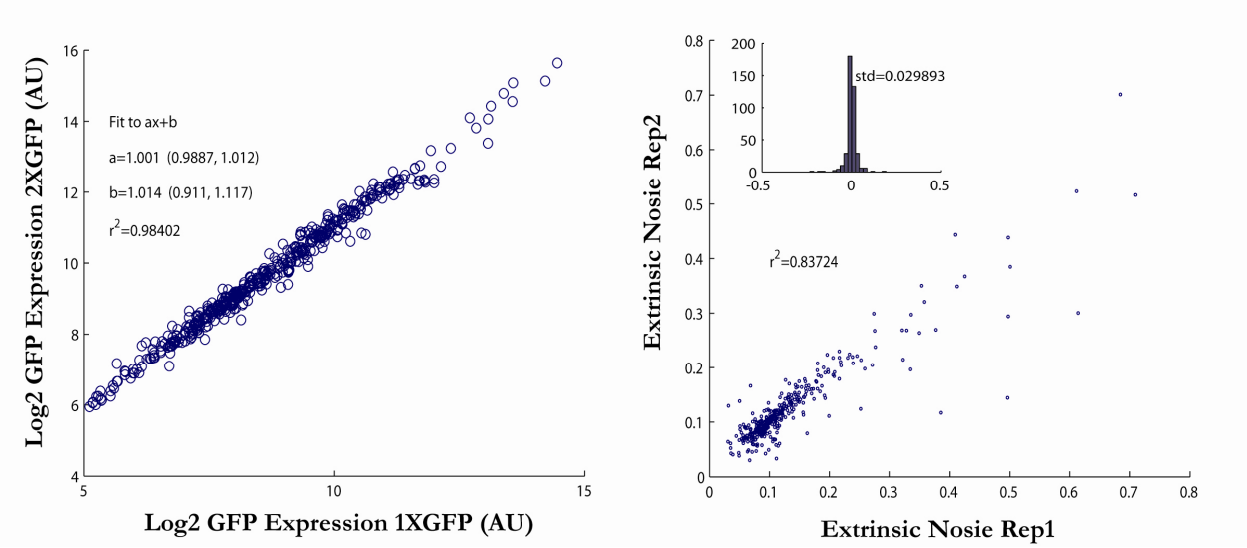

Figure S1 Related to Figure 1

Measurements of extrinsic noise using the one FP strategy are reproducible and reveal a subset of genes which are extrinsically noisy. (a) A scatter plot of the log2 expression of strains with one and two copies of YFG-GFP. (b) Replicates of noise measurements show strong reproducibility and a Gaussian error structure (inset).

S2

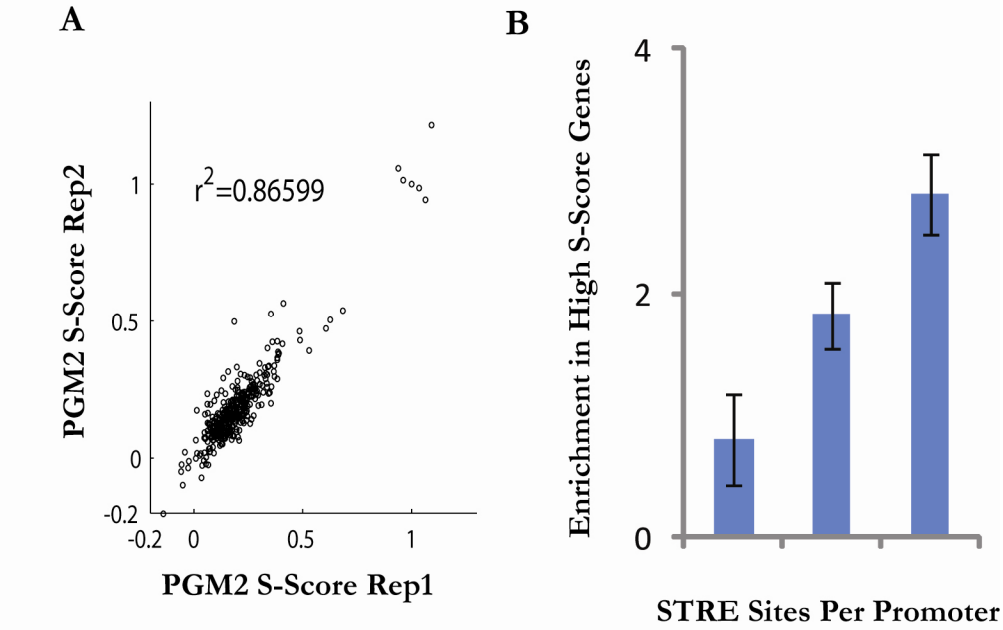

Figure S2. Related to Figure 2

S-score measurements are reproducible and genes which show high PGM2 S-score are enriched in MSN2/4 binding sites (STRE). (a) Biological replicates of S-Score measurements (N=362). (b) Enrichment of genes with S-scores greater than 0.3 (N=53) in genes with one, two, or three STRE binding sites in their promoters, error bars represent standard error.

S3

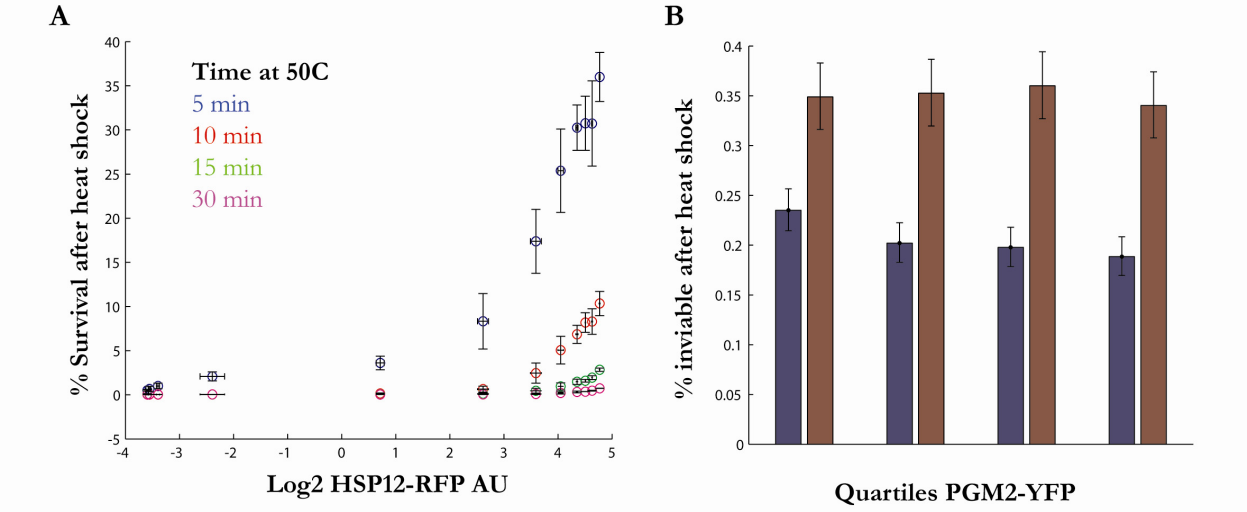

Figure S3. Related to Figure 3

Variation in MSN2/4 activity has phenotypic consequences. (a) Survival of cells after a 50C heat shock as a function of their basal Hsp12 expression. Cells with the endogenous copies of MSN2/4 deleted were transformed with an estradiol inducible constitutively active copy of MSN2 (constitutively active 5A allele). The cells were induced by addition of estradiol to a range of HSP12-RFP levels, and then exposed to heat shock at 50C for the indicated times. Survival was accessed by number of viable cells in each population after recovery at 30C, error bars represent std. error of triplicate measurements. (b) Survival of cells as a function of their basal PGM2 levels. PGM2-YFP expression was determined for individual cells in early (OD=0.05 (red)) mid-exponential phase (OD=0.5 (blue)). Cells were then heat shocked (50C, 20 min), and stained with propidium iodide to detect dead cells. For the mid-exponential phase population, the probability of cell death was lower (18.84% (95% CI = 16.96-20.84)) in the top 25% of PGM2 expressing cells compared to the bottom quarter (23.51% (95% CI = 21.45-25.66%)). No statistically significant difference was detected for the early exponential phase population. Statistics were computed from using a binomial test (N=1608 in each quartile, error bars reflect 95% confidence interval).

S4

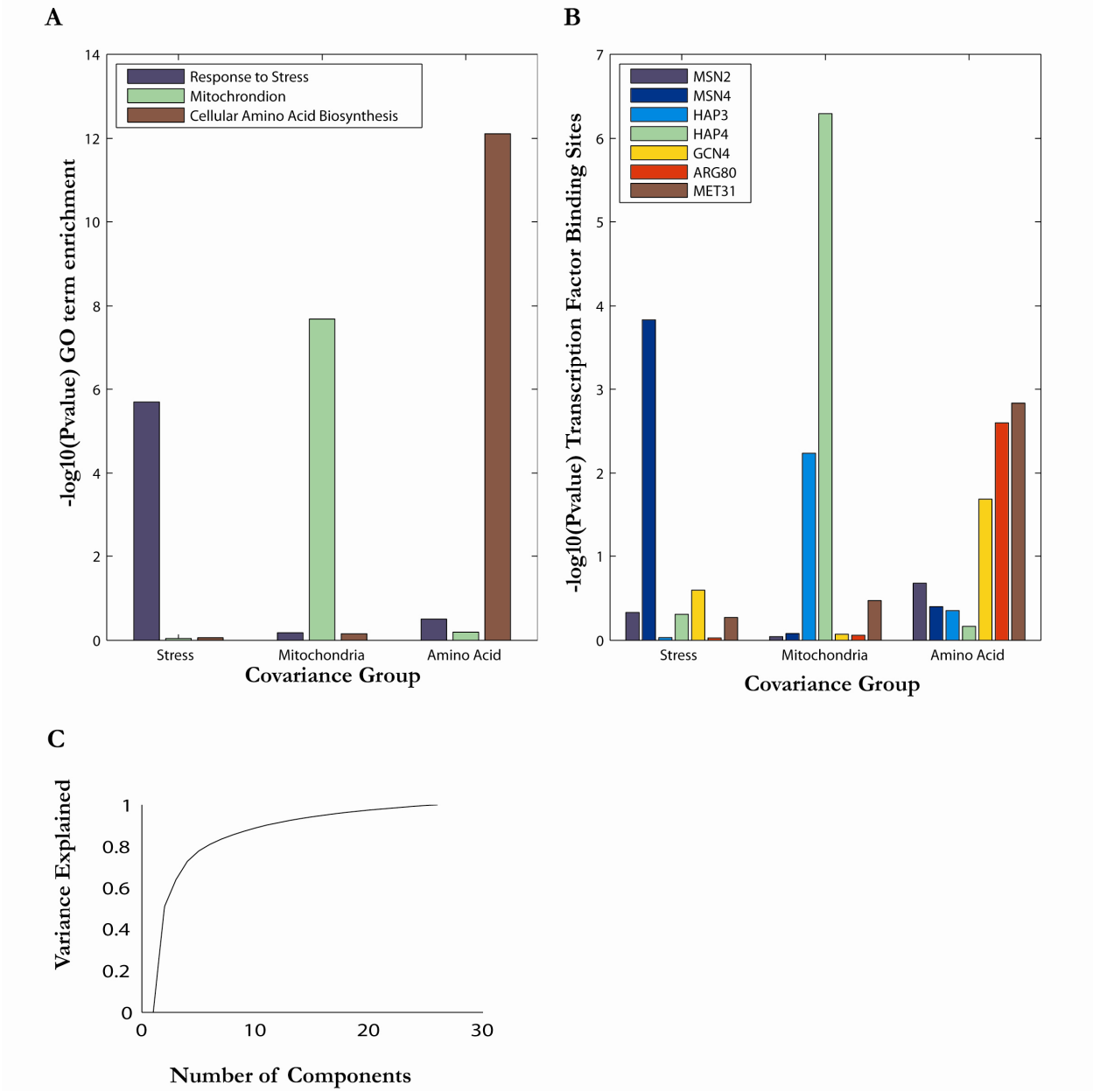
A
Response to Stress
Mitochrondion
Cellular Amino Acid Biosynthesis
-log10(Pvalue) GO term enrichment
Stress
Mitochondria
Amino Acid
Covariance Group
B
MSN2
MSN4
HAP3
HAP4
GCN4
ARG80
MET31
-log10(Pvalue) Transcription Factor Binding Sites
Stress
Mitochondria
Amino Acid
Covariance Group
C
Variance Explained
Number of Components

Figure S4. Related to Figure 4

Covariance measurements reveal the modular regulatory structure of the cell. (A) Go term enrichment analysis identifies three major functional clusters: Response to Stress, Mitochondrial regulation, and Cellular Amino Acid Biosynthesis (b) Enrichment of transcription factor binding sites in the promoters of genes featured across the three groups (c) Variance explained by principal components of covariance data plotted against the number of components. Much of the structure of the dataset can be explained by 5 principal components. All p-values are calculated using a hypergeometic test (N=182).

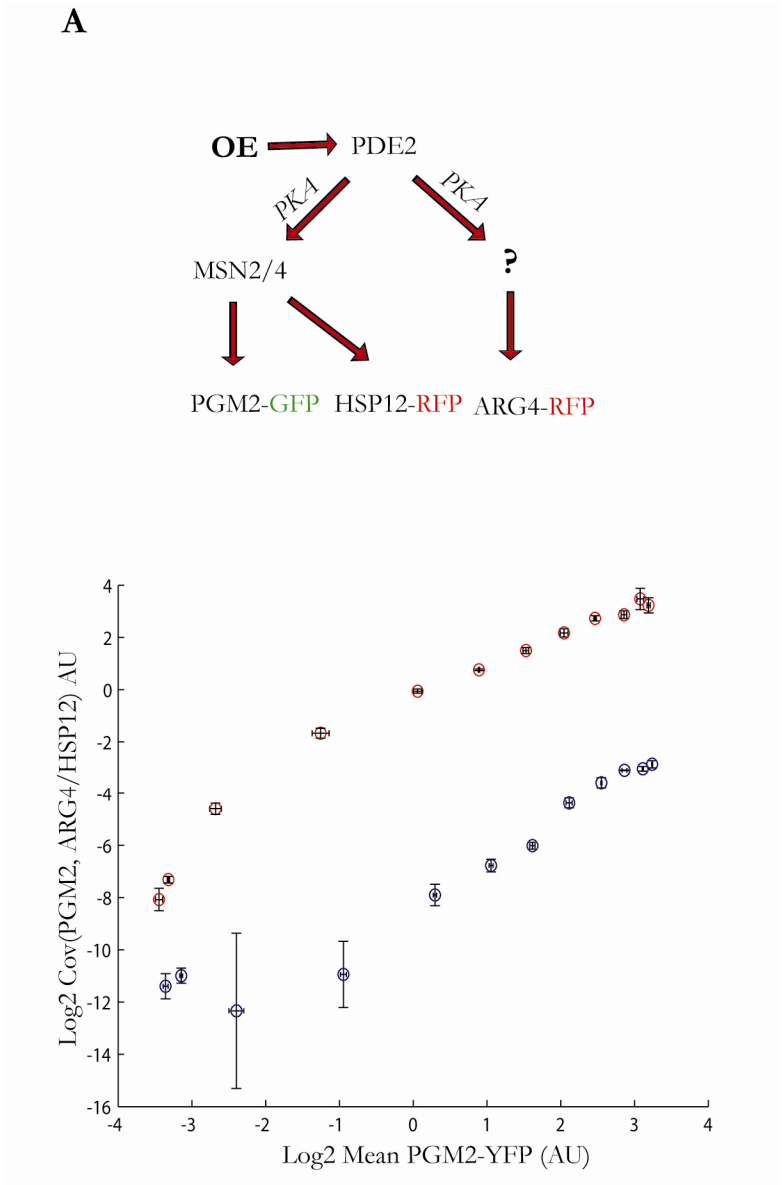
A
OE
PDE2
PKA
PKA
MSN2/4
?
PGM2-GFP
HSP12-RFP
ARG4-RFP
Log2 Cov(PGM2, ARG4/HSP12) AU
Log2 Mean PGM2-YFP (AU)
4
2
0
-2
-4
-6
-8
-10
-12
-14
-16
-4
-3
-2
-1
0
1
2
3
4

Figure S5. Related to Figure 5

Overexpression of PDE2 results in distinct diagnostic patterns of covariance between the MSN2 sensitive gene PGM2 and amino acid biosynthesis gene ARG4. (A) Graded inhibition of PKA by overexpression of PDE2 results in non-linear increases in covariance between ARG4 and PGM2, and PGM2 and HSP12. Error bars represent standard Error of triplicate measurements.