Supporting Information

Li et al. 10.1073/pnas.0914302107

SI Text

Noise Clustering Based on Mixture Gaussian. To group together genes with similar expression noise level, we used a generative model to learn the inherent noise structure. Briefly, at step (1), we modeled the noise distribution with Gaussian mixtures, whose parameters were automatically learned by the expectation-maximization algorithm. At step (2), the number of Gaussian components, K, was determined by minimizing Bayesian information criterion (or Schwarz criterion, BIC) in our model selection. By varying $K = 1$ to 10, we found BIC reaches the minimal when $K = 7$. At step (3), to unambiguously assign each gene into one of the seven noise clusters, we took a Bayesian approach to calculate the posterior probability for each gene to belong to each of the seven noise clusters; we then assigned the gene to the cluster with the highest posterior probability. We noted that this algorithm is purely based on the intrinsic noise structure and also guarantees that genes with similar expression noise are grouped together.

Calculating Clustering Coefficient *C*. The local *clustering coefficient* of a vertex in a network reflects the possibility of the vertex within a clique formed by its immediate interacting partners. For the ith node with K_i immediate neighbors in an undirected network, its *clustering coefficient* C_i is defined as the following (1): $C_i = \frac{2|\{e_{jk}\}|}{K_i(K_i-1)}$, where $|\{e_{jk}\}|$ represents the total number of edges between its K_i neighbors.

The above clustering coefficients were defined in the undirected networks (1), such as the protein interaction networks; in the gene regulatory network which is a directed network, we followed a generalized definition proposed in ref. 2.

Calculating Modularity Index *Q*. We adopted Newman's approach to calculate network modularity Q (3). Given an undirected network defined by the adjacency matrix A, where $A_{ij} = 1$ if node i and j have an edge while $A_{ij} = 0$ if node i has no connection with node j. For any partition of the network into two modules, the modularity index Q of the partition can be defined as

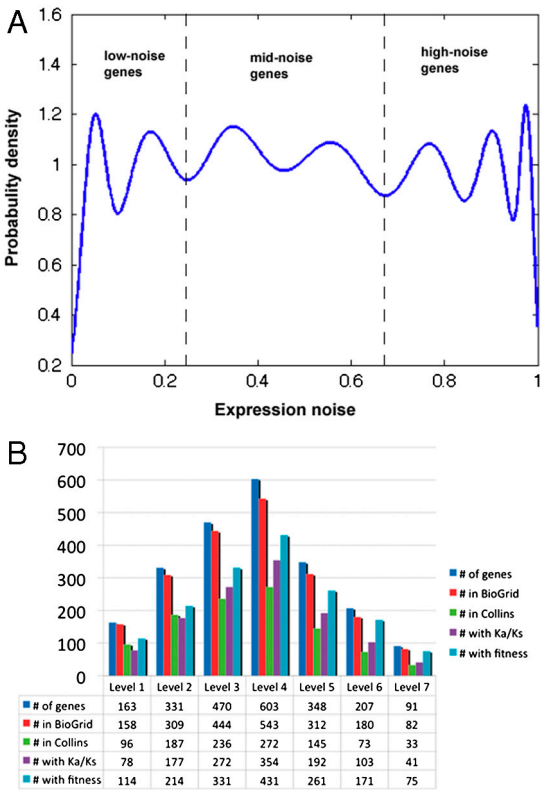
$$Q = Q = \frac{1}{4m} s^T B s, \quad \text{where } B_{i,j} = A_{i,j} - \frac{k_i k_j}{2m}.$$

In the above equation, S is a vector whose element $s_i = 1$ indicates the node i belongs to module 1 while $s_i = -1$ for the node i classified into module 2. k_i,k_j… is the network degree of node i and j; $m = \frac{\Sigma_{ij} A_{ij}}{2}$. In our analysis clustering coefficients in protein interaction networks and the gene regulatory network were calculated using GAIMC toolbox implemented in MATLAB, and betweenness in protein interaction networks were calculated based on the Graph module in Perl.

Automated Image Acquisition. Yeast cells expressing GFP fusion chimeras were grown in YEPD media overnight in 96-well format deep-well blocks and subcultured for 5 hours in prewarmed fresh media to obtain cells in log phase. Cells were then resuspended in low fluorescent media and distributed in 96-well glass bottomed plates (MMI Greiner M plates). An ImageXpress 5000A fluorescence microscopy system from Molecular Devices was used to acquire images. Images were acquired at room temperature for two hours.

Automated image quantification. Automated image acquisition and analysis were performed with MetaXpress software, v1.63 (Molecular Devices). After images were shade-corrected and background-subtracted, objects were segmented and single cells were defined using background cell fluorescence in the GFP channel. We used a series of MetaXpress modules to segment whole cells. Once cells were identified, dead cells were removed from further analysis by gating average-grayscale as they had high autofluorescence. A minimal set of features (dimension, shape factor, and elliptical form factor) was used to train the software to efficiently classify an unseen image into two categories such as budded and unbudded cells. Each budded or unbudded cell was taken as a region of interest, and the dimension and intensity profiles of each region of interest were quantified individually for each cell.

1. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* 393(6684):440–442.
2. Fagiolo G (2007) Clustering in complex directed networks. *Phys Rev E* 76(2 Pt 2):026107.
3. Newman ME (2006) Modularity and community structure in networks. *Proc Natl Acad Sci USA* 103(23):8577–8582.



	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7
# of genes	163	331	470	603	348	207	91
# in BioGrid	158	309	444	543	312	180	82
# in Collins	96	187	236	272	145	73	33
# with Ka/Ks	78	177	272	354	192	103	41
# with fitness	114	214	331	431	261	171	75

Fig. S1. (*A*) The learned mixture Gaussian model that best fits the distribution of expression noise among the yeast proteins. Yeast genes are assigned to one of the seven Gaussian components based on its maximal Bayesian posterior probability. To provide enough and balanced datasets in the SVM prediction experiments (described in the latter part of the manuscript), the two quietest gene clusters (level 1–2) are grouped as "low-noise," the noisiest clusters (level 5–7) are grouped as "high-noise," and the intermediate levels (level 3–4) are grouped as "mid-noise". (*B*) The number of genes in each of the seven noise clusters. Also shown is the availability of protein–protein interactions, Ka/Ks values, and fitness data for genes in each cluster.

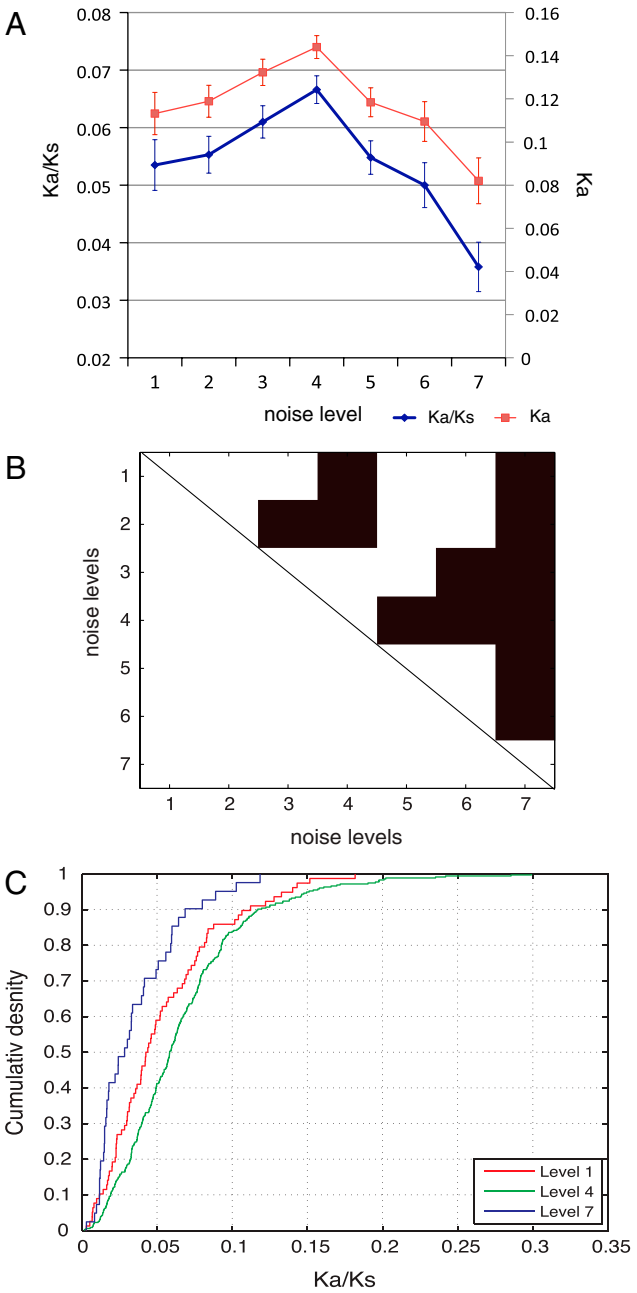


Fig. S2. Yeast proteins with the highest expression noise (level 7) are under the strongest selective pressure on their coding sequences. (*A*) The mean evolutionary rate of genes at each noise level. Error bars represent one standard error. (*B*) The statistical significance of pairwise comparison of selective pressure (Ka/Ks) between any two noise levels. Clearly genes at noise level 7 have the lowest Ka and Ka/Ks. Statistically significant pairwise comparisons between two noise levels are highlighted in black; we used the Wilcoxson rank sum test. (*C*) The cumulative density functions for the Ka/Ks distribution for genes in the noise levels 1, 4, and 7. Clearly the extremely noisy genes (level 7, the highest curve, blue color) have the lowest Ka/Ks whereas genes with the intermediate noise level (level 4, the lowest curve, green color) have the highest Ka/Ks.

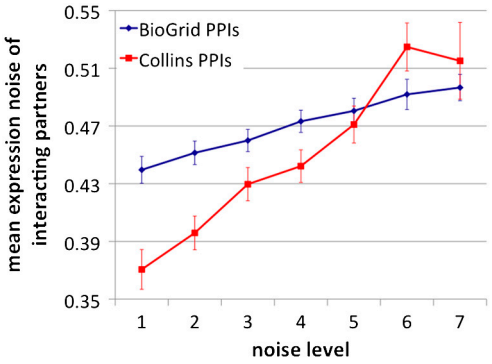


Fig. S3. Interaction partners of quiet proteins tend to be quiet and interaction partners of noisy proteins tend to be noisy too. Error bars represent one standard error.

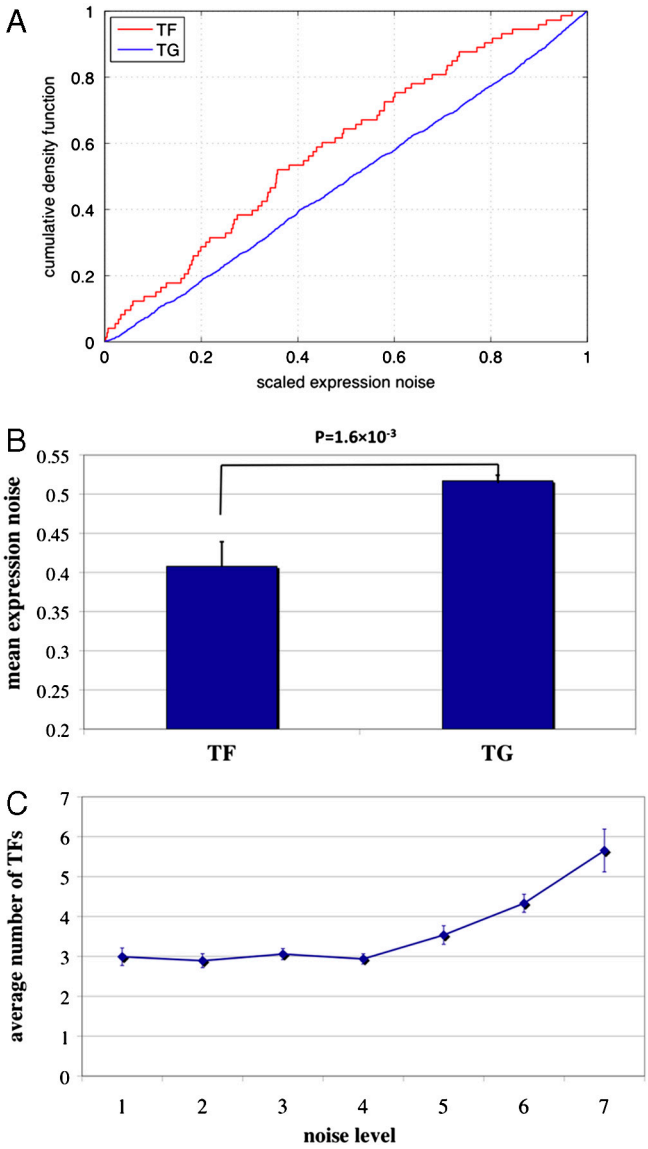


Fig. S4. (*A*) Comparison of cumulative density functions of expression noise between transcription factors (TFs) and target genes (TGs). Clearly TFs have reduced noise than TGs. (*B*) TFs have lower mean expression noise than TGs ($P = 1.6 \times 10^{-3}$, Wilcoxon rank sum test). Error bars represent one standard error. (*C*) TGs with increased expression noise tend to be regulated by more TFs ($R = 0.18$, $P = 1.14 \times 10^{-12}$, Pearson's correlation, and $R = 0.17$, $P = 1.54 \times 10^{-11}$, Spearman's correlation). Error bars represent one standard error.

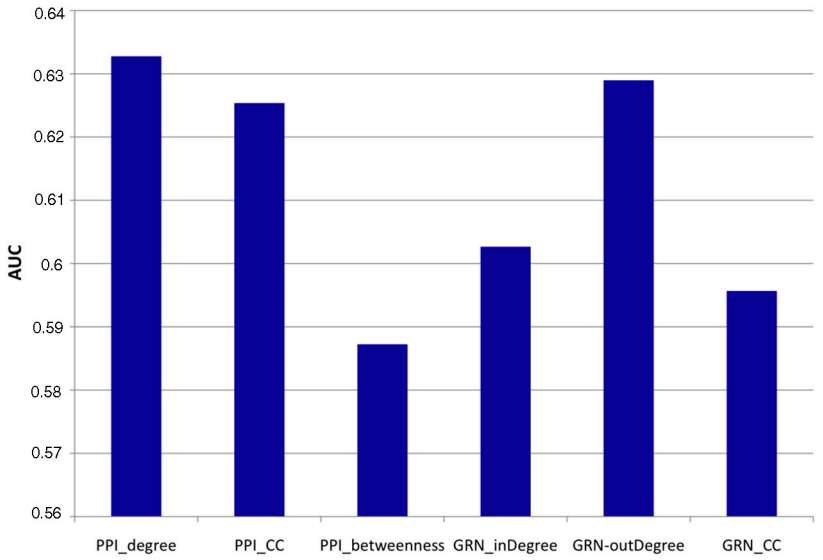


Fig. S5. Prediction accuracy of the individual topological features in the protein–protein interaction networks and the gene regulatory network. Because only TFs in GRN have outgoing edges, the prediction accuracy by out-degree in the GRN only applies to TFs. CC: clustering coefficient.

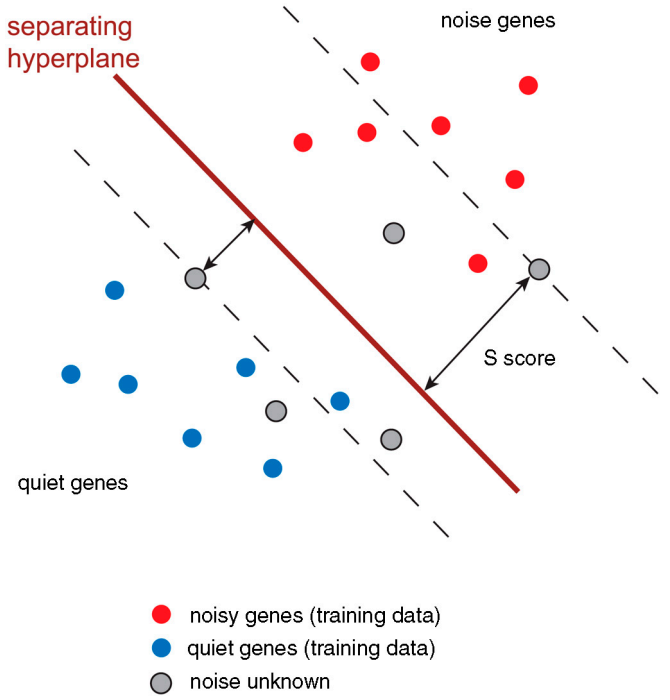


Fig. S6. Schematic diagram of the support vector machine (SVM) algorithm. The SVM is trained on the *noisy genes* (red circles) and *quiet genes* (blue circles); a hyperplane is determined that can best separate these two groups of training genes in the induced kernel space of high dimension. For a gene of unknown noise, its noise can be quantified by the ***S*** score, reflecting the distance of the gene to the hyperplane.

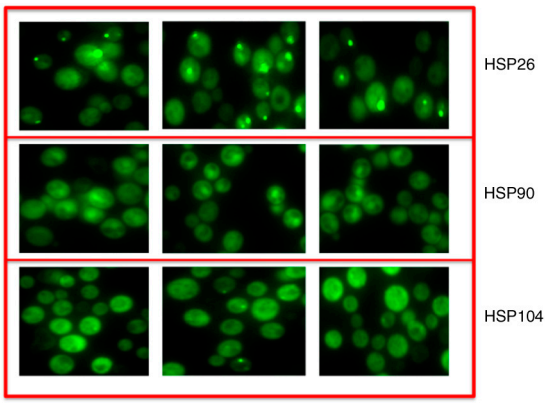


Fig. S7. Fluctuation of fluorescence intensities for three heat-shock proteins in the microscopic assay.

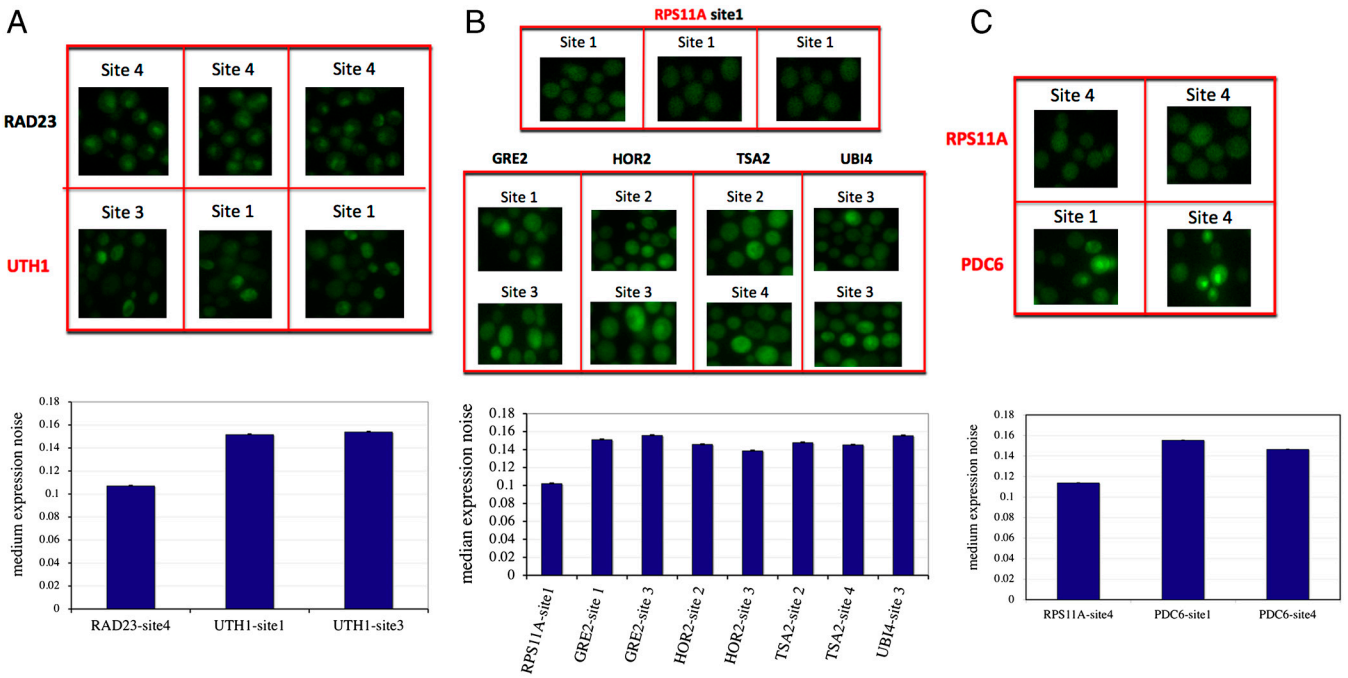


Fig. S8. Yeast cells with GFP fused proteins have reduced noise are shown on top, whereas yeast cells with elevated noise are shown at the bottom. Cells in the same image and between images all have comparable extrinsic characteristics. Expression noise of the gene in red had not been previously measured. The bar chart indicates the boot-strapped noise median and its associated standard error estimated from the entire cell population of the corresponding sites for each gene. (*A*) Comparison between RAD23 and UTH1. (*B*) Comparison between RPS11A, GRE2, HOR2, TSA2, and UBI4. (*C*) Comparison between RPS11A and PDC6.

Other Supporting Information Files

Table S1 (XLS)
Table S2 (XLS)
Table S3 (XLS)
Table S4 (XLS)
Table S5 (XLS)