Supplementary Materials:

Probing the effect of promoters on noise in gene expression using thousands of designed sequences

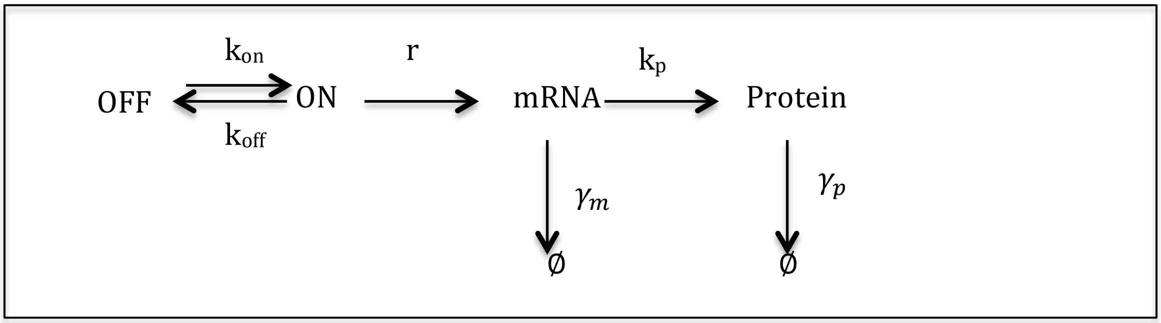
Eilon Sharon[1,2,†], David van Dijk[1,2,†], Yael Kalma[2], Leeat Keren[1,2], Ohad Manor[1], Zohar Yakhini[3,4] and Eran Segal[1,2]

1 Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel
2 Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel
3 Computer Science department, Technion, Haifa, Israel
4 Agilent Laboratories, Santa Clara, California, USA

† These authors contributed equally to this work
Correspondence should be addressed to E.Segal (eran.segal@weizmann.ac.il)

Kinetic model of promoter activation

To investigate, theoretically, how promoter state switching affects protein abundance noise, we assume the following common kinetic model of gene expression(Raser and O'Shea 2004):



Where k_{on} and k_{off} are the promoter on- and off-switching rates respectively, r the transcription rate (of the ON state), γ_m the mRNA degradation rate, and k_p and $\gamma_m$the protein production and degradation rate respectively.

This model assumes that the promoter switches between a transcriptionally active (ON) and inactive (OFF) state, and that this switching is correlated with TF binding and unbinding such that the promoter that has an activator bound to it is transcriptionally active and the promoter that is unbound is inactive.

Following the derivation presented by Sanchez et al. 2011(Sanchez et al. 2011) we assume that the number of proteins produced per mRNA (denoted *b*) follows a geometric distribution, thereby leaving out a specific translation and protein degradation rate. Using the master equation we solve this kinetic model to achieve the steady state mean protein abundance and the noise.

The mean protein abundance can be written as:

$$(1)\ \ \langle p \rangle = \frac{k_{on}}{k_{on} + k_{off}} \cdot \frac{b \cdot r}{\gamma_m}$$

the Noise (CV^2) as:

$$(2)\ \eta^2{}_p = \frac{b+1}{\langle p \rangle} + \frac{k_{off}\gamma_m}{k_{on}(\gamma_m + k_{off} + k_{on})}$$

and the noise strength (Fano factor) as:

$$(3)\ F_p = b + 1 + \frac{k_{off}}{(\gamma_m + k_{off} + k_{on})} \cdot \frac{b \cdot r}{(k_{on} + k_{off})}$$

The first component of equation (2) scales inversely with the mean protein level and is a result of Poissonian transcription and bursty translation. The second component, however, is affected by mRNA degradation rate and promoter activation switching rates. Therefore, an increase in both k_{on} and k_{off} by the same factor (faster switching) does not change mean protein abundance, but decreases the noise. Similarly, a decrease in promoter switching rates (slower switching), does not change mean protein abundance but increases the noise. Finally, both an increase in *r* or k_{on}will increase expression and decrease noise. However, only when increasing *r* will the Fano factor go up; increasing k_{on} decreases the Fano factor.

A kinetic model of gene expression that takes into account transcription factor non-specific DNA binding and 1-dimensional sliding along the DNA

In the above model we investigated promoter state switching as a function of TF binding, however we did not explicitly model binding reactions, nor did we fit the model to measured data. In order to capture the differences between our designed promoters that contain all combinations of 7 possible Gcn4 binding sites, we extend the above model by incorporating specific binding and unbinding reactions for each binding site. To investigate the mechanisms of how binding site configuration can affect noise, we fit our model to the data using two different assumptions on how TFs find their target sites.

In short, we fitted the model parameters using MATLAB using the following procedure:

1) Each of 256 promoters (2^7 combinations of 7 possible binding sites in two DNA sequence background) is mapped to a unique kinetic scheme. Free parameters of transcription rates, translation rate, mRNA and protein degradation rates as and specific binding to each site were shared across the different promoter configurations
2) Each model's free parameters (in a cross-validation) was fitted to the measured mean expression and noise, such that in each iteration:
   a. We generate all 256 kinetic models (one per promoter) and plug in the free parameters or compute rate parameters using a set of equations depending on the specific model assumptions on TF target search
   b. We analytically solve each model to get the predicted mean and noise
   c. We compute the RMSE of both mean and noise across all promoters

Next we will describe our modeling approach in detail.

Mapping a promoter transcription factor binding configuration to a kinetic scheme

The set of designed promoters that we modeled contained 256 promoters– all combinations of 7 predefined Gcn4 binding sites, each in one of two sequence contexts, namely a high (GAL1-10) and a low (HIS3) GC content context. **Table S1** contains a description of these promoters. Each promoter is represented by a binary string of length 7 which is 1 at position *q* if the promoter has a binding site at position *q* and zero otherwise, and by an indicator of the context sequence.

In our model a promoter that has *N* sites has 2^N possible states in which a site is either free or bound by a TF (Gcn4). A transition matrix *K* of size 2^Nx2^N is used to represent the promoter state space, where *Mij* is the rate of transitioning from state *i* to state *j*. Each transition involves either one binding or one unbinding reaction to a single binding site. Therefore, *Mij = 0* when either *i=j* or the difference between promoter state *i* and promoter state *j* is more than one binding/unbinding reaction. The rate transitions that involve binding or unbinding of a single TF are computed according to either the 3D or

3D+1D model as described below. See **Fig. 4A** for two examples of a mapping between a promoter and a kinetic scheme.

Transcription factor binding and unbinding rates

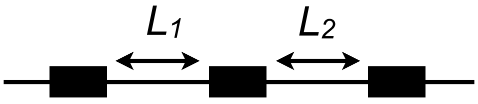
In order to compute the transitions of the above kinetic scheme, that are a function of the binding and unbinding reactions of the TFs, we assume either only 3D diffusion, in which binding and un-binding is independent of any neighboring binding sites, or 3D diffusion followed by 1D diffusion along the DNA (1D-sliding), in which binding and un-binding is influenced by neighboring sites. In both cases we assume the below scheme with a free, non-specific (NS) bound and specific (S) bound state in which specific binding occurs only after non-specific binding, and non-specific un-binding after specific un-binding. This model, and the below equations, are an adaptation of the model presented by Hammar et al.(Hammar et al. 2012).

$$\text{Free} \underset{kd_{ns}}{\overset{ka_{ns}}{\rightleftarrows}} \text{NS} \underset{kd_{s}}{\overset{ka_{s}}{\rightleftarrows}} \text{S}$$

We compute the effective non-specific binding and unbinding as follows:

$$(4)\ \mathrm{ka}_0 = \mathrm{ka}_{\mathrm{ns}}\left(1 + \mathrm{s}\left(\tanh\left(\frac{\mathrm{L}_1}{\mathrm{s}}\right) + \tanh\left(\frac{\mathrm{L}_2}{\mathrm{s}}\right)\right)\right)$$

$$(5)\ \mathrm{kd}_0 = \mathrm{kd}_{\mathrm{ns}}\left(1 + \mathrm{s}\left(\tanh\left(\frac{\mathrm{L}_1}{\mathrm{s}}\right) + \tanh\left(\frac{\mathrm{L}_2}{\mathrm{s}}\right)\right)\right)$$



Where $\mathrm{ka}_{\mathrm{ns}}$ and $\mathrm{kd}_{\mathrm{ns}}$ are the non-specific binding and unbinding rates respectively, s is the sliding distance, L1 and L2 are the distances from the center of the current site to the center of the left and right site respectively, ka_0 and kd_0 are the effective non-specific binding and un-binding rates to the specific site that take into account 1D-sliding along the DNA. When the left or right site does not exist, the values for L1 and L2, respectively, are infinite. When a neighboring site is unbound we take half of the distance to the neighboring site (L/2), and when it is bound we take the whole distance.

We then compute the total effective rates as follows:

$$(6)\ \mathrm{ka_{eff}} = \mathrm{ka_s}\left(\frac{\mathrm{ka_0}}{\mathrm{ka_s} + \mathrm{kd_0}}\right)$$

$$(7)\ \mathrm{kd_{eff}} = \mathrm{kd_s}\left(\frac{\mathrm{kd_0}}{\mathrm{ka_s} + \mathrm{kd_0}}\right)$$

Where $\mathrm{ka_{eff}}$ and $\mathrm{kd_{eff}}$ are the effective binding and un-binding rates respectively to and from the specific site taking into account non-specific binding and 1D-sliding, $\mathrm{ka_s}$ and $\mathrm{kd_s}$ are the specific binding and un-binding rates respectively.

The 3D-diffusion only model is a special case of the 3D-diffusion and 1D-sliding model in which the sliding distance (s) is 0. In this case, neighboring sites do not affect binding or unbinding of each other.

As we describe above, for each promoter configuration we construct a kinetic scheme (transition matrix K) using the equations for effective binding and unbinding. Here are the free parameters that are shared between configurations and that are estimated by the fitting procedure:

***ka_{ns}* - non-specific binding (min^{-1}):**

A single parameter for each of two sequence contexts: GAL1 10 or HIS3 (see Sharon et. al(Sharon et al. 2012) for details). The main difference between these contexts is their GC content, which relates to lower (low GC) or higher (high GC) nucleosome affinity. We therefore assume that the contexts will have different accessibility and therefore different non-specific binding rates.

***kd_{ns}* - non-specific unbinding (affinity, min^{-1}):**

We assume a similar non-specific unbinding rate (i.e. binding affinity) for all promoters since this rate is associated with the affinity of a TF to a non-specific DNA.

ka_s - site specific binding (min^{-1}):

We assume single value for each of the 7 sites, which can be interpreted as each binding site having a specific accessibility/binding affinity.

kd_s - site specific unbinding (affinity, min^{-1}):

We assume a single value for each of the 7 sites, which can be interpreted as each binding site having a specific binding affinity.

s - sliding distance (bp):

This parameter is the average sliding distance of a TF along the DNA before it dissociates and is therefore shared between all promoters. This parameter is only free when 1D sliding is assumed, in the only 3D model it is fixed to zero.

Thus, our set of 256 unique promoters has a combined 18 free parameters in 3D+1D model and 17 in 3D model (since s is fixed to zero) that are used for modeling the promoter state transition matrices of all 256 promoters. The attached MATLAB script '*generate_Ks_mat.m'* implements the computation of the transition matrices.

Modeling transcription and translation

In addition to promoter state switching, as a result of TF binding and unbinding, we model transcription, translation and mRNA and protein degradation. The transcription rate of a promoter with no bound TF is *Roff* (in min^{-1}) and the rate of transcription of any bound state (in min^{-1}) is *Ron* times the number of sites. *Roff* and *Ron* are free parameters that are shared between all configurations in our model. mRNA degradation and translation are captured by the free parameter *b* (in protein/mRNA), which is the average number of proteins produced per mRNA when we assume that this quantity follows the geometric distribution, which enables us to solve the model analytically(Sanchez et al. 2011; Carey et al. 2013). Finally we have the protein degradation rate *delta* (in min^{-1}) as a free parameter.

Analytical solution

The resulting kinetic scheme (matrix *K*) and above rate parameters for transcription, translation and degradation (using in total 21 or 22 parameters) form a system of equations that can be solved analytically to obtain the mean and the noise of the steady state protein abundance distribution. (Implemented in MATLAB script: *solveMasterEquation.m*)

Converting predicted protein abundance to measured arbitrary fluorescence units

Our experimental system measures the gene expression distribution indirectly by using sequencing reads counts to identify the single-cell distribution across bins of increasing promoter driven fluorescence (see below section "Extracting expression mean and variance from pooled promoter expression distribution measurements"). In addition, the expression is a normalized one, namely it is the ratio between variable promoter driven YFP and constant TEF2 promoter driven mCherry. Therefore, the expression value that our computational pipeline produces is in arbitrary units and only has meaning when comparing between promoters. However, this is only the case for the mean expression - noise (defined as the coefficient of variation squared) is unit less and therefore can be interpreted in absolute terms. Our model predicts gene expression levels in terms of protein abundance. So, when comparing the predicted protein abundance mean to the measured mean expression (but not noise), we convert it to arbitrary units. Therefore, we introduce an additional free parameter to our modeling procedure, namely a scaling parameter *S* (in units of proteins per fluorescence value) whose value is fitted together with the other free parameters.

Fitting procedure

We fitted the model's parameters to the measured data in a cross-validation scheme, where we used 30 bootstrapped cross-validations, each with 20% of the data held out as test set and the rest used as train set. The fitting procedure is as follows: We used a non-linear constrained optimization function to minimize an objective function that computes for each of the 2x128 promoters the root mean square error (RMSE) of the predicted mean expression and noise compared to the measured mean expression and noise. In each iteration of the fitting process, 256 kinetic schemes are generated and solved as we describe above. For the optimization we used MATLAB's fmincon function (using the interior-point algorithm), where we minimize the following error function:

$$(7)\ f(x) = \sqrt{\left(\log(\eta_p) - \log(\eta_m)\right)^2} + \sqrt{\left(\log(\mu_p) - \log(\mu_m)\right)^2}$$

where μ_m and μ_p are vectors of measured and predicted mean expression respectively, and η_m and η_p are vectors of measured and predicted noise values respectively.

See **Table S1 for** the input data for the model (promoter configurations and contexts) and the measured and fitted mean and noise values for each promoter. We note that simulating our model numerically, using the Gillespie algorithm(Gillespie 1977) gives the same result as the analytical solution (R^2=0.99, **Fig. S15**). See **Fig. S11** and **S12** for the performance of the model in a 30x bootstrapped 5-fold (20% test data) cross-validation and see **Table S2** for its parameter values.

Sensitivity analysis

To test whether our model contains any redundant parameters we investigate the sensitivity of the parameters of our model by performing the rigorous sensitivity analysis procedure, LHS-PRCC, described by Marino et al. (Marino et al. 2008). First, we sample 10,000 instances of the model, each with a unique parameter setting. This sampling is done using a Latin Hypercube, to ensure uniform sampling in the multidimensional parameter space of our model. Next we perform a rank transformation on the sampled instances using the distance of the model instances to the measured data (for both mean expression and noise) and quantify the correlation that each parameter has with the goodness of fit measure. See **Fig. S16 and Fig. S17** for the scatter plots and correlation coefficient for each parameter of the kinetic model. We note that one of the most sensitive parameters in the model is the scaling parameter (S). This parameter converts measured fluorescence units to number of proteins and therefore we would expect this parameter to influence model outcome hugely. Other (some even more) sensitive parameters were *b* (protein burst size), r_{on}/r_{off} (transcription rates), and non-specific binding. It is not surprising that the model is sensitive to these parameters as they directly affect the mean expression but also the noise (such as *b* which directly affects the mean and noise via bursting). Reassuringly, we find that the model is sensitive to almost all parameters (with exception to kdsp3 and kdsp5), suggesting that separation between non-specific and specific binding, modeling binding at individual sites, and including sliding are non-redundant components of our model. In addition, the significant sensitivity to D (diffusion coefficient) suggests that interaction between sites is an important mechanism in this model. However, the sensitivity analysis also tells us that the model is more sensitive to general non-specific binding and unbinding parameters, such as *kans* and *kdns*, than to the specific (un)binding parameters of the individual sites, suggesting that the model is dominated by general binding and unbinding to the promoter, rather than by binding events at individual sites.

Extracting expression mean and variance from pooled promoter expression distribution measurements

Removing experimental noise from expression distribution measurements using peak detection

To remove experimental noise from measurements of promoter distribution across expression bins, the distribution peak that contained the largest fraction of cells of each promoter was detected and any cells outside of the peak were considered as technical noise. Here is a description of the procedure applied to each promoter expression distribution:

1. Expression bins that contained a fraction of cells smaller than a threshold were set to zero. The threshold used in this work was 1 / (#bins * 10) =0.3125%.
2. The main peak of the distribution across the expression bins was detected using the following procedure:

a. Bins values (fraction of cells in each bin) were smoothen using ‘rlowess’ procedure with a span of three bins (MATLAB, Curve fitting toolbox, smooth function).
b. The segment of non-empty neighboring bins with the largest sum of cells was selected.
c. Within this segment the bin with the largest fraction of cells was detected as the center of the peak.
d. The peak was defined as bins with decreasing cell fractions relative to bins closer to the peak center. In case that the first increase in cell fraction was less than 30% the bin in which the fraction increased was also included in the peak. This heuristic allows small improvements in the detection of the peak edges as estimated by eye examination.
e. Fractions of cells in bins outside the main peak were set to zero.
f. Finally, the filtered distribution of cells across expression bins was calculated by normalizing the sum of non-smoothen fraction of cells in the main peak bins to one. The value of bins outside the main peak was set to zero.

This procedure is implemented in the MATLAB script: “SynLibConvertBinFractionVec2SinglePeak.m”, which is publically available in our web site. **Table S3-4** contain the un-processed fractions of cell containing each promoter in each expression bin.

Extracting expression mean and variance by fitting a gamma distribution to the expression distribution of each promoter

Following the removal of experimental noise from the data through expression distribution peak detection the mean and the variance of each promoter expression distribution were extracted by fitting a Gamma distribution to the data. The below procedure (implemented in MATLAB script “FitGammaDistributionToValRangesProbabilities.m”) was used for this purpose:

First, the mean and the variance of the expression distribution were estimated directly from the data. The mean expression of each promoter was calculated as a weighted average of the mean expression of all bins, where the weight of each bin is the fraction of cell containing the promoter in that bin. We used the following formula: $\overline{Exp} = \sum_{i=1}^{32} \overline{Exp_i} * P_i$ where $\overline{Exp_i}$ is the mean expression of cells in bin I and P_i is the fraction of cells containing the promoter in bin i. Similarly, we computed the standard deviation of each promoter using the standard deviation of each bin and the distribution of the promoter across the bins. We used the following formula: $var(Exp) = \sum_{i=1}^{32}(var(Exp_i) + \overline{\left(Exp - \overline{Exp_\imath}\right)}^2 \, (\overline{Exp} - \overline{Exp_i})^2) * P_i$ where: $\overline{Exp}$ is promoter expression mean, $\overline{Exp_i}$ is the mean expression of cells in bin i and $var(Exp_i)$ is the variance of the expression of cells in bin i (implemented in the MATLAB script: ComputeStdFromBinStdsAndMeans.m). Expression levels mean, standard deviation and minimal and maximal expression level of cells in all bins are attached (**Sup Table S5-6**).

Next a Gamma distribution was fitted to the distribution of cells using MATLAB optimization and statistics toolboxes. The mean of and variance calculated above were used as a starting point for the learning algorithm. The algorithm (implemented in the script: FitGammaDistributionToValRangesProbabilities.m and functions called within) apply optimization algorithm that minimizes three criteria: kolmogorov-smirnov distance, Kuiper's test distance and L2 distance between the fraction of cells in each bin and the predicted fraction according to the Gamma function. The formula used the L2 distance is: $D = \sum_{i=1}^{32}(P_i - \int_{x=min_i}^{max_i} P(x)_{\sim Gamma(a,b)} dx)^2$ where P_i is the fraction of cells in bin i , min_i and $max_i$are the minimal and maximal expression levels of cells in bin i. The integral calculates the fraction of cells predicted by the Gamma function to be in the bin. The results of the optimization algorithm that preform best on the three criteria were selected as the output. **Table S7** contains the processed values and this filter indicator.

Assessing the quality of pool promoter expression mean and variance measurements

The quality of our pooled measurements of the noise ($\frac{variance}{mean^2}$) and mean of expression level was assessed by two criteria – reproducibility and accuracy. These criteria were used to assess three alternative data processing methods: 1) extracting mean and variance directly from the data ("Raw data") 2) Extracting mean and variance directly from data after removing experimental noise by detecting the peak of the distribution across expression bins as described above ("Expression distribution peak") 3) Extracting mean and variance by fitting a Gamma function to the expression distribution peak ("Gamma fitted", this is the data used for the analysis in this manuscript). As shown in **Fig. S1** and **Fig. S18** the reproducibility of expression noise level measurements decreased following the expression peak detection step probably due to the technical noise added by the procedure (Pearson's R^2 = 0.74, 0.47, 0.53 for methods 1-3 correspondingly). However, after filtering low quality measurements (as described in **Methods**), the reproducibility of our noise estimates was similar to using the raw data (Pearson's R^2=0.78).

To assess the accuracy of our pooled mean and noise measurements we compared them to measurements of 54 strains that were isolated from the library and measured using FACS separately. As shown in **Fig. S19,** while the expression mean can be predicted accurately by all three methods (Pearson's R^2 = 0.976, 0.974, 0.939 for methods 1-3 correspondingly) the accuracy of expression noise level measurements increased considerably by using our data processing scheme (Pearson's R^2 = 0.20,0.71,0.80 for methods 1-3 correspondingly).

We concluded that, while extracting expression mean directly from the data gives highly reproducibility and accurate results, extracting the noise (which requires to extract the variance) requires filtering of low quality measurements and data processing by a combination of peak detection and Gamma function fitting.

Supplementary Figures

Fig. S1:

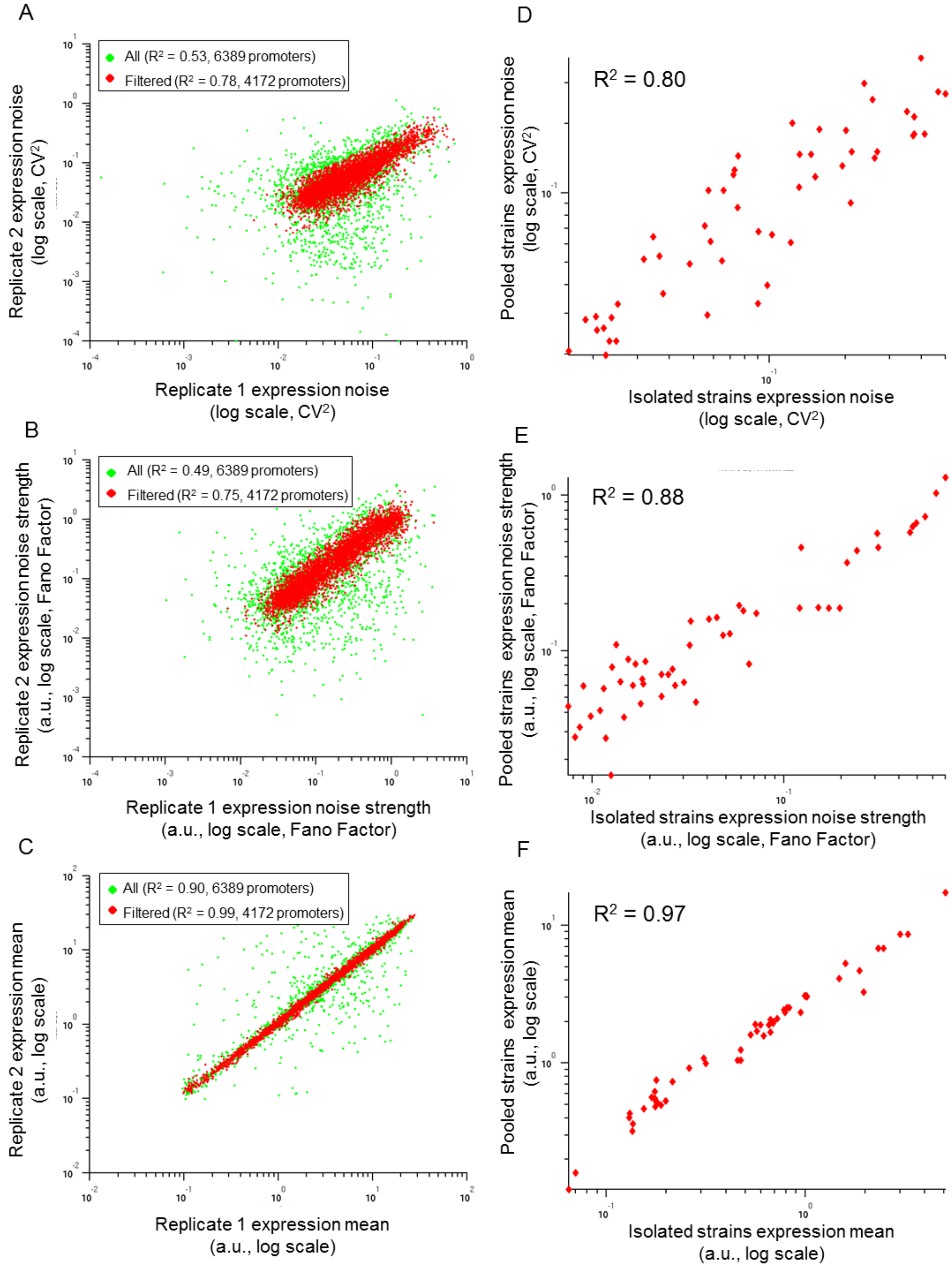


Figure S1. Obtaining reproducible and accurate expression and noise measurements for thousands of designed promoter sequences. (A-C) A comparison of expression noise **(A)**, noise strength **(B)** and mean expression **(C)** measurements obtained for two independent replicates (x-axis: replicate 1, y-axis: replicate 2) for all 6500 constructs in the library (green points) or the subset of the promoters that passed our quality control filter (red dots, see **Methods**). **(D-F)** A comparison of expression noise **(D)**, noise strength **(E)** and mean expression **(F)** for 54 individual strains isolated and sequenced from the pool of transformed yeast cells (x-axis). Each strain was measured in isolation using a flow cytometer (x-axis) or within a single experiment using our method (y-axis).

Fig. S2:

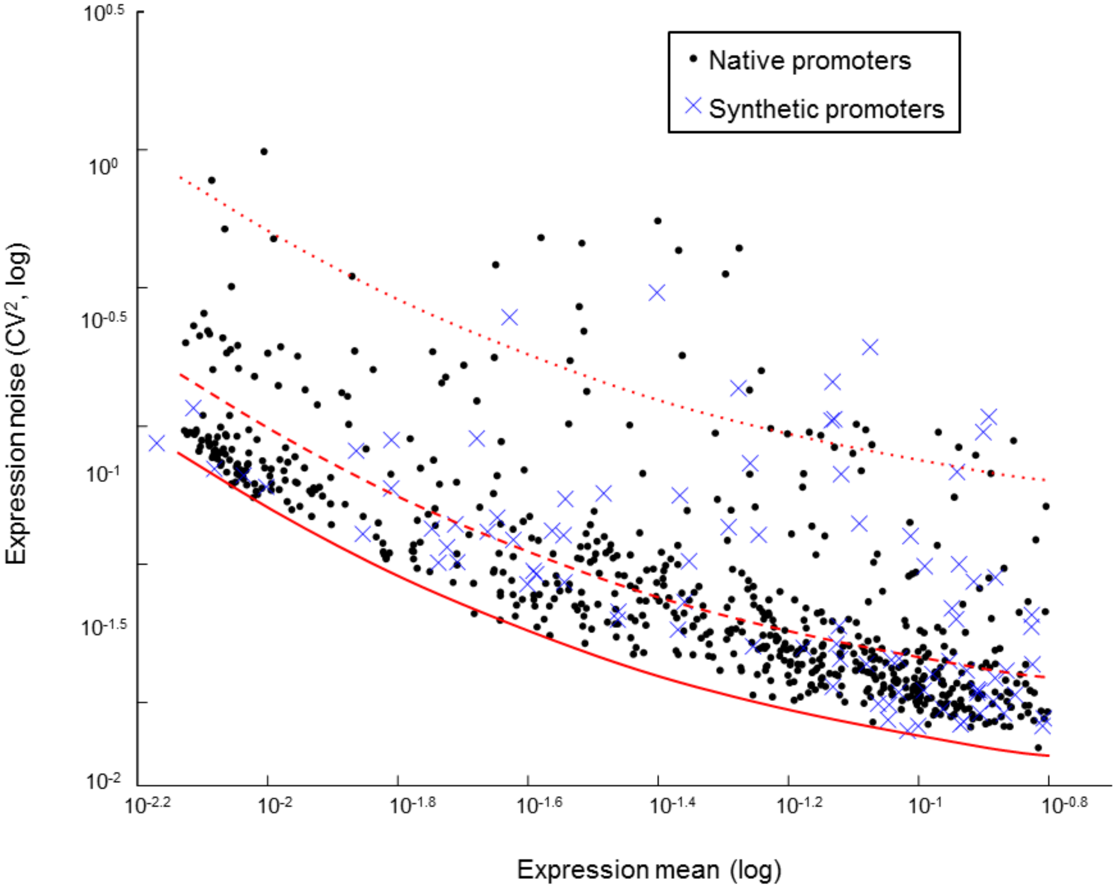


Figure S2. Synthetically designed promoters span similar ranges of mean and noise expression levels as native promoters. Shown are mean (x-axis) and noise (y-axis) measurements of individual, non-pooled strains from the synthetic promoter library (blue) and from a native promoter library (black)(Lubliner et al. 2013). The solid red line marks the lower noise limit, the dashed red line marks two-fold above the lower noise limit, and the dotted red line marks 10-fold above the lower noise limit.

Fig. S3:

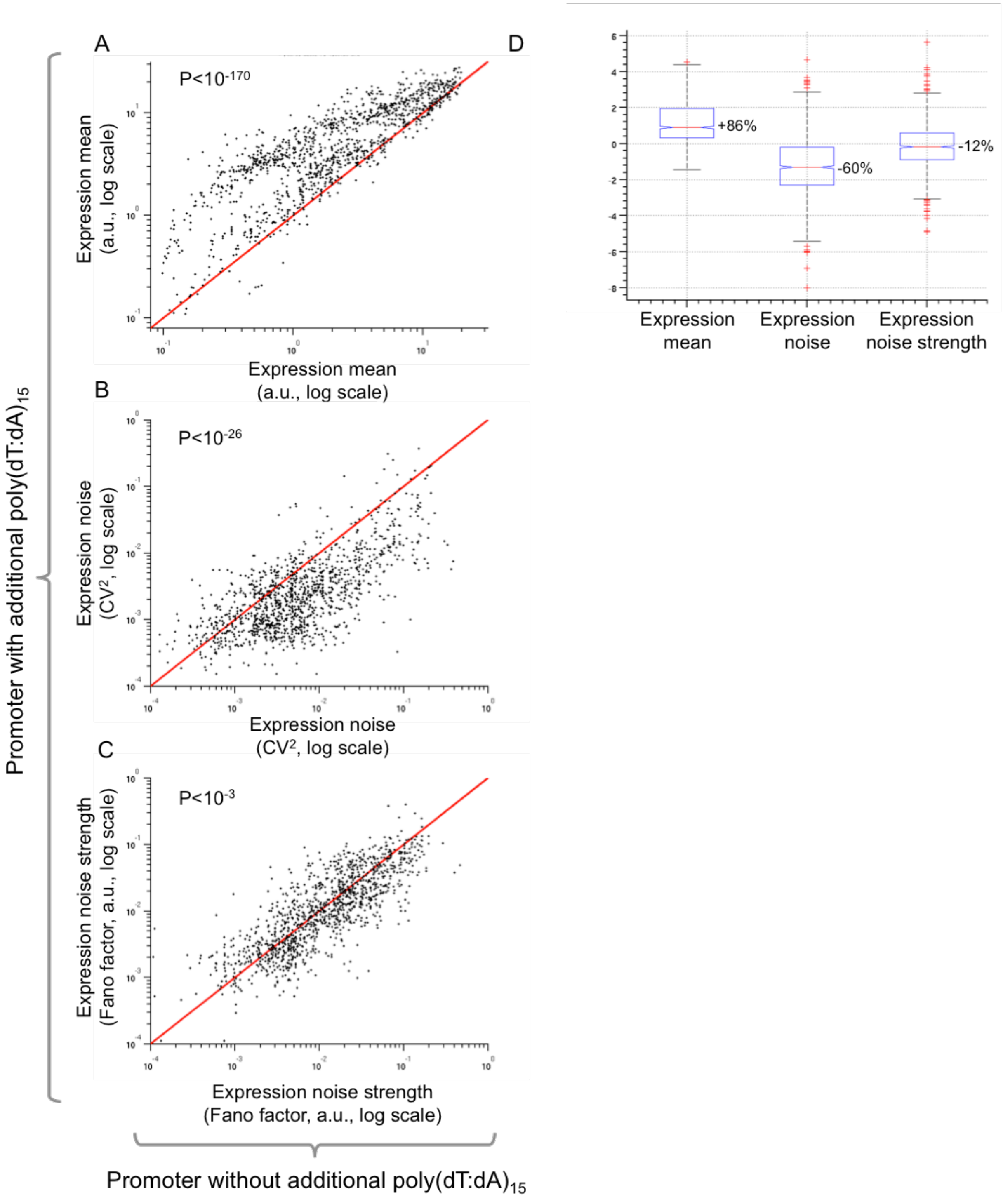

A
P<10^-170
Expression mean
(a.u., log scale)
Expression mean
(a.u., log scale)
B
P<10^-26
Expression noise
(CV², log scale)
Expression noise
(CV², log scale)
C
P<10^-3
Expression noise strength
(Fano factor, a.u., log scale)
Expression noise strength
(Fano factor, a.u., log scale)
Promoter with additional poly(dT:dA)15
Promoter without additional poly(dT:dA)15
D
+86%
-60%
-12%
6
4
2
0
-2
-4
-6
-8
Expression mean
Expression noise
Expression noise strength

Figure S3. Nucleosome disfavoring sequences increase mean expression, decrease expression noise and have little effects on noise strength. Shown is a comparison of expression mean **(A)**, noise **(B)** and noise strength **(C)** for 1268 promoter pairs with (y-axis) or without (x-axis) a nucleosome disfavoring sequence (15bp long poly(dT:dA) tract). Note that in most pairs, adding a nucleosome disfavoring sequence significantly increases the mean expression and significantly decreases the noise, but has little effect on noise strength. *P* values were computed using Student's *t*-test. **(D)** Boxplot of $\log_2$ ratio of expression mean, noise and noise strength of the 1268 promoter pairs presented in (A)-(C). Ratio values median and median 95% confidence intervals (CI, computed using 100,000 bootstrapping iterations) are 1.86 (CI 1.76-1.93), 0.40 (CI 0.37-0.43) and 0.88 (0.84-0.93) for expression mean, noise and noise strength respectively. The median of each pair of the three distributions is significantly different (Wilcoxon rank sum test *P* values < 10^{-80})

Fig. S4:

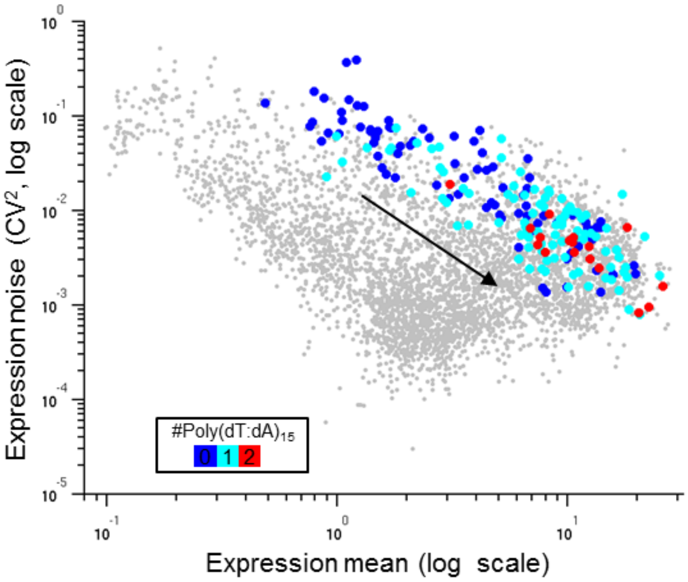


Figure S4. Promoters with more nucleosome disfavoring sequence have higher mean expression and lower noise. Shown is the mean expression and noise of 205 promoters with 0 (blue points), 1 (light blue points), or 2 (red points) poly(dT:dA) sequences of length 15bp and two Gcn4 binding sites. Promoters with more poly(dT:dA) tracts tend to have higher expression and lower noise.

Fig. S5:

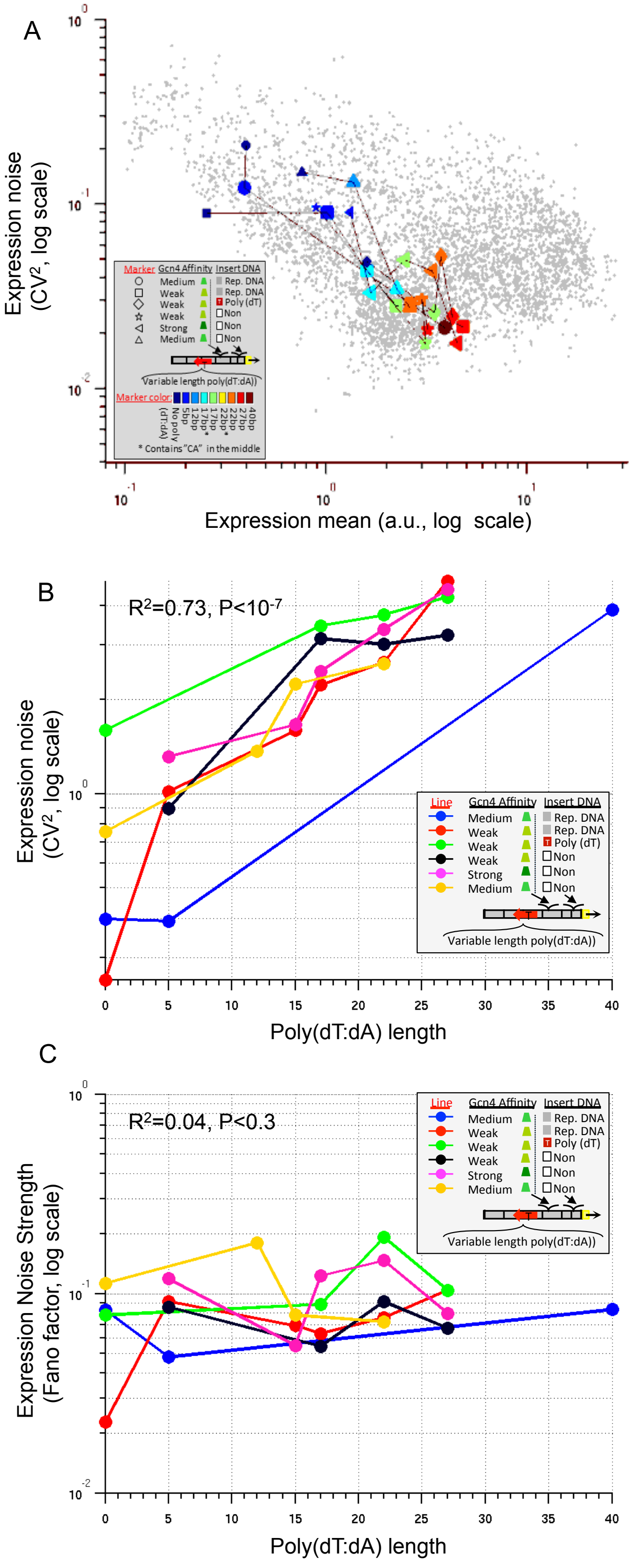

A
Expression noise
(CV^2, log scale)
10^0
10^{-1}
10^{-2}
Expression mean (a.u., log scale)
10^{-1}
10^0
10^1
Marker
Gcn4 Affinity
Insert DNA
○ Medium Rep. DNA
□ Weak Rep. DNA
◇ Weak Poly (dT)
☆ Weak Non
◁ Strong Non
△ Medium Non
Variable length poly(dT:dA))
Marker color:
No poly (dT:dA)
5bp
12bp
17bp*
17bp
22bp*
22bp
27bp
40bp
* Contains "CA" in the middle
B
R^2=0.73, P<10^{-7}
Expression noise
(CV^2, log scale)
10^0
Poly(dT:dA) length
0
5
10
15
20
25
30
35
40
Line
Gcn4 Affinity
Insert DNA
Medium Rep. DNA
Weak Rep. DNA
Weak Poly (dT)
Weak Non
Strong Non
Medium Non
Variable length poly(dT:dA))
C
R^2=0.04, P<0.3
Expression Noise Strength
(Fano factor, log scale)
10^0
10^{-1}
10^{-2}
Poly(dT:dA) length
0
5
10
15
20
25
30
35
40
Line
Gcn4 Affinity
Insert DNA
Medium Rep. DNA
Weak Rep. DNA
Weak Poly (dT)
Weak Non
Strong Non
Medium Non
Variable length poly(dT:dA))

Figure S5 Promoters with longer poly(dT:dA) tracts have higher mean expression and lower noise. (**A)** Shown is the effect of gradually increasing the length of a poly(dT:dA) tract from 0bp (none, dark blue) to 40bp (dark red) in six different promoter configurations (each marked with a different symbol, for a detailed description see Sharon et al.(Sharon et al. 2012)). Also shown is the effect of the length of the inserted poly(dT:dA) tract (x-axis) on expression noise (**B,** y-axis) and noise strength (**C**, y-axis) in the six promoter configurations. While increasing the length of the poly(dT:dA) tract increases noise (R^2=0.73) it does not have a consistent effect on noise strength (R^2=0.04).

Fig. S6:

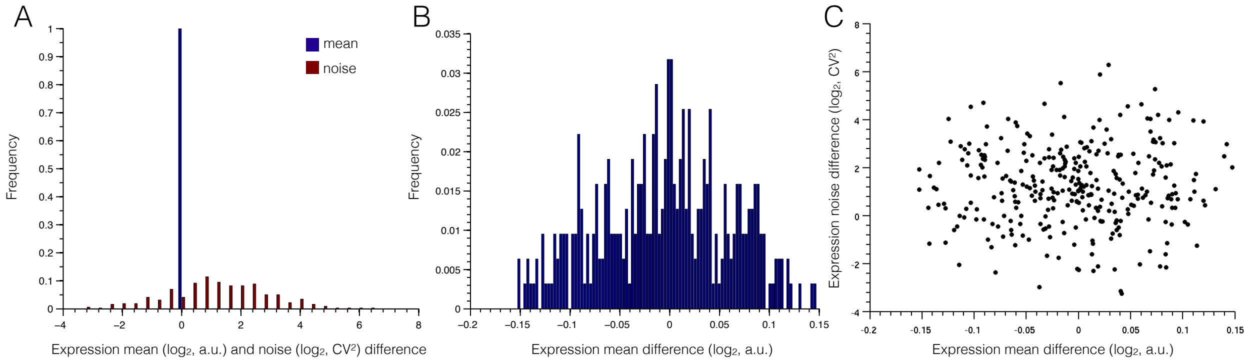


Figure S6 Noise changes for pairs of similar mean expression. (**A**) Shown are the changes in mean expression and noise (log2 ratio between each pair promoter with TF binding site and promoter with nucleosome disfavoring sequence) for 417 promoter pairs in which adding a TF binding site or a poly(dT:dA) tract resulted in similar increase of the mean expression, as is shown in **Fig. 2B**. The blue bar (**A**) shows that each pair has very small difference in mean expression. The red bars show the distribution of changes in noise. (**B**) Shown is the distribution of the changes in mean expression (similar to the blue bar in (**A**) but zoomed in). (**C**) Shown is the data of (**A)** and (**B**), but in a dot plot where the X-axis shows the change, for each pair (dot), in mean expression and the Y-axis the difference in noise (both in log2). While the pairs of promoters were selected such that they induce highly similar expression levels (mean of expression mean changes ratio distribution is almost zero, Student's *t*-test $P>0.13$) the noise of promoters with TF binding site is significantly larger than promoters with nucleosome sequence (expression noise changes ration distribution Student's *t*-test $P<10^{-30}$)

Fig. S7:

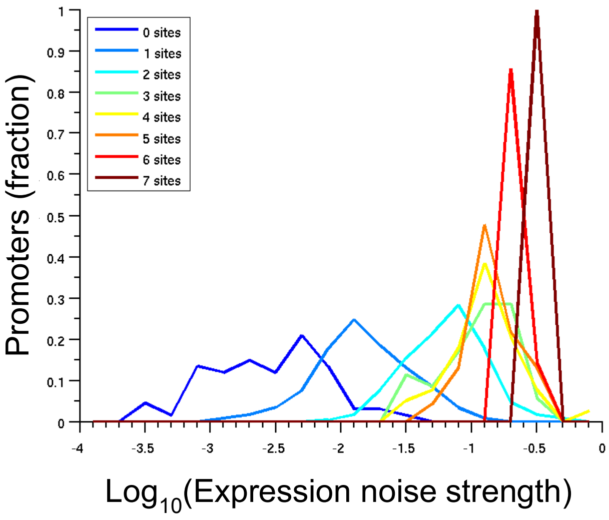


Figure S7 Promoters with more Gcn4 binding sites have higher expression noise strength. Shown is a histogram of noise strength for the promoter sets of **Fig. 2**. The distributions have significantly different means (ANOVA test $P<10^{-19}$).

Fig. S8:

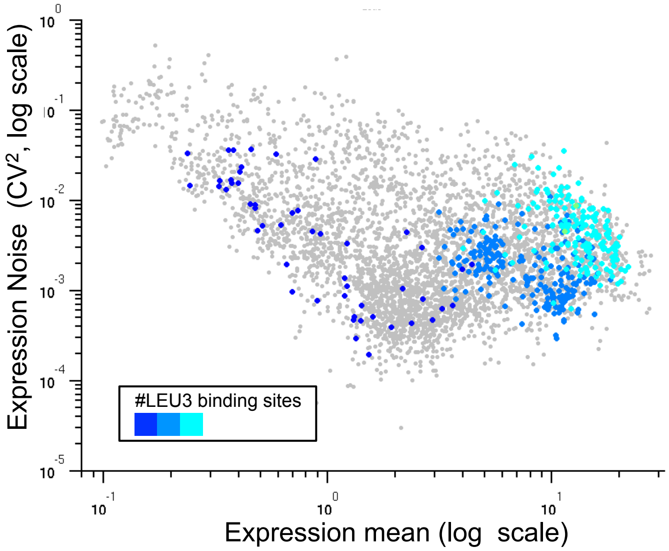


Figure S8. Promoters with more Leu3 binding sites have higher mean expression and higher noise. Shown is expression mean and noise for 442 promoters with 0 (dark blue points) to 2 (cyan points) Leu3 binding sites. For a given expression level, promoters that contain more Leu3 binding sites have higher noise.

Fig. S9:

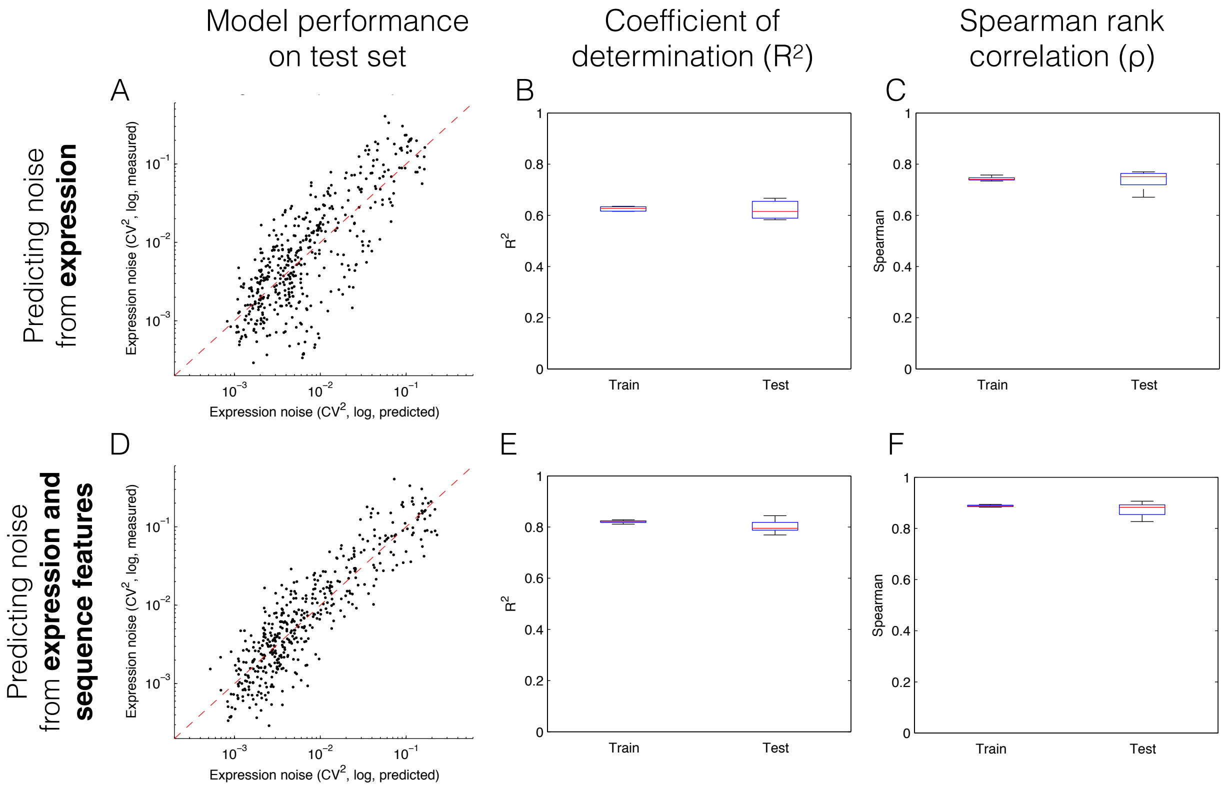


Figure S9. Five-fold cross-validation results of the linear model on the single site set. (A,B,C) Noise predicted from expression. **(D,E,F)** Noise predicted from sequence features and expression. (**B,E**) Coefficient of determination (R^2) of 5-fold cross-validation. (**C,F**) Spearman's rank correlation coefficient (ρ) of 5-fold cross-validation.

Fig. S10:

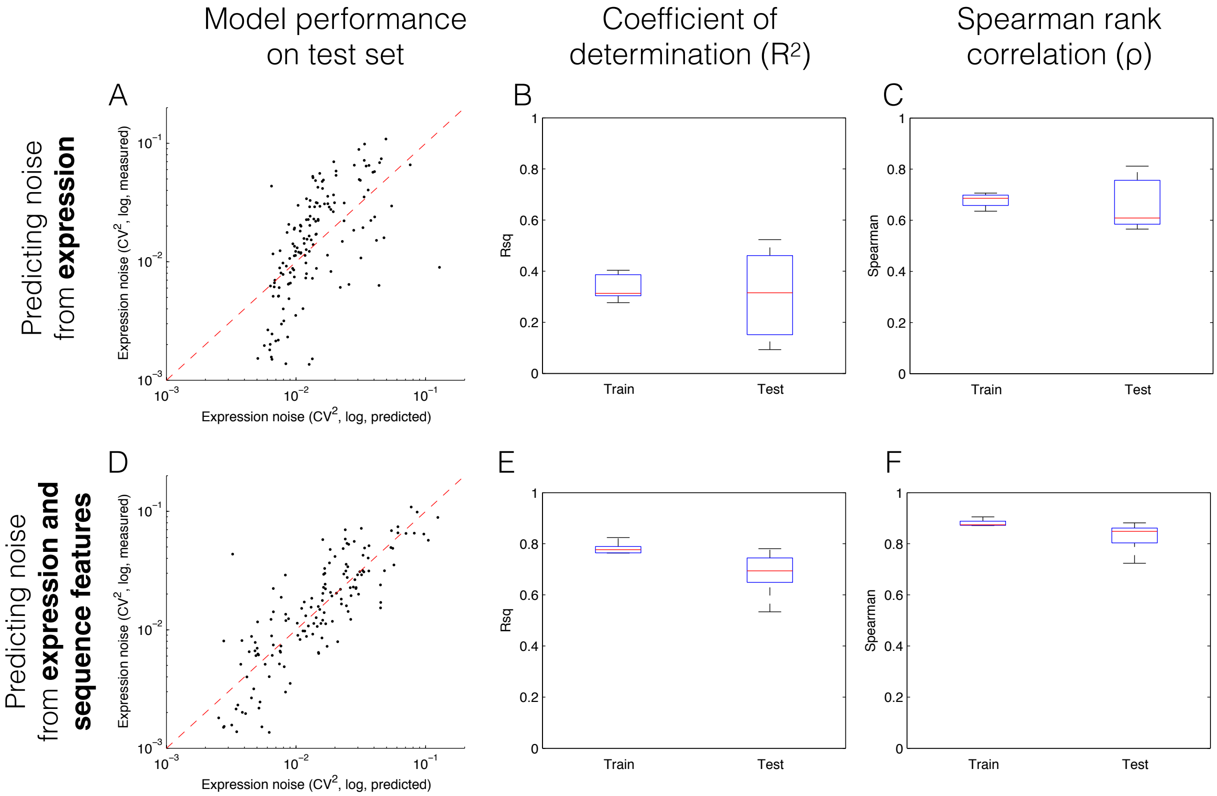


Figure S10. Five-fold cross-validation results of the linear model on the multiple site set. (A,B,C) Noise predicted from expression. **(D,E,F)** Noise predicted from sequence features and expression. (**B,E**) Coefficient of determination (R^2) of 5-fold cross-validation. (**C,F**) Spearman's rank correlation coefficient (ρ) of 5-fold cross-validation.

Fig. S11:

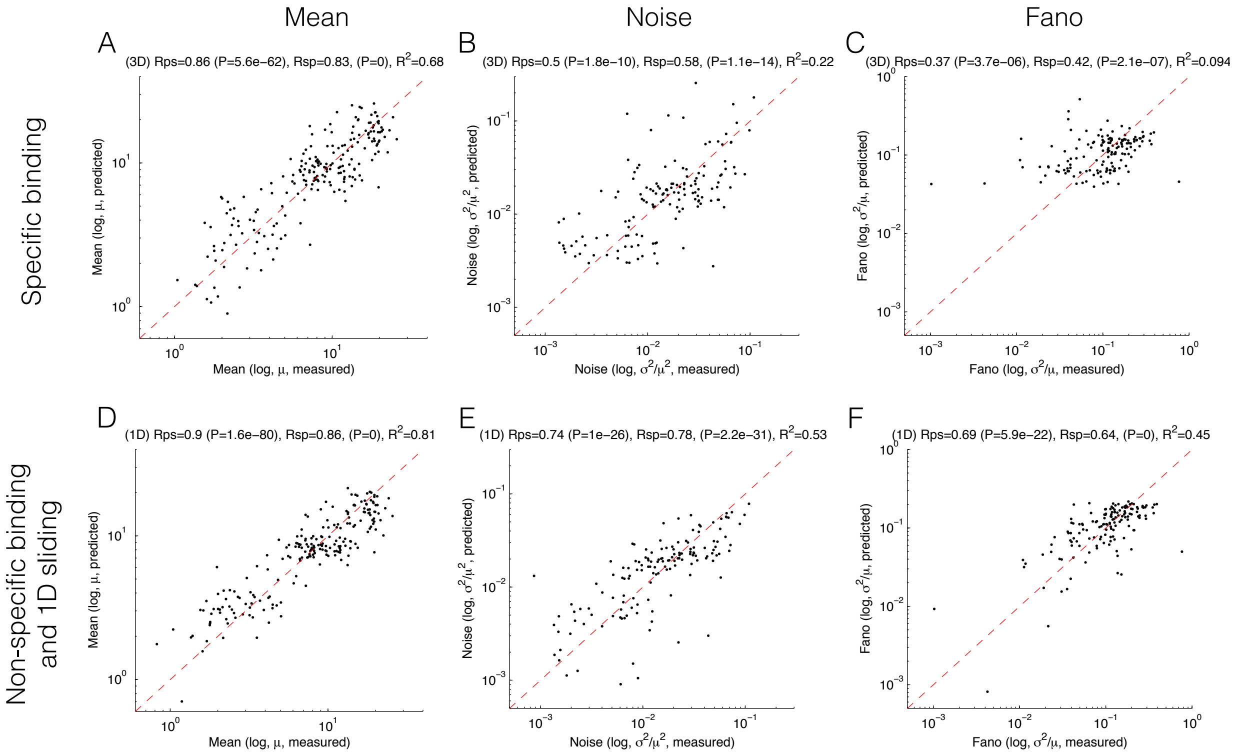


Figure S11. Prediction of the kinetic model of transcriptional regulation on held-out data. Shown are measured versus predicted data for the mean (A,D) noise (B,E) and Fano factor (C,F). (A,B,C) Show the model in which binding is assumed to be only specific. (D,E,F) show the model in which binding is non-specific and TFs slide to their target sites.

Fig. S12:

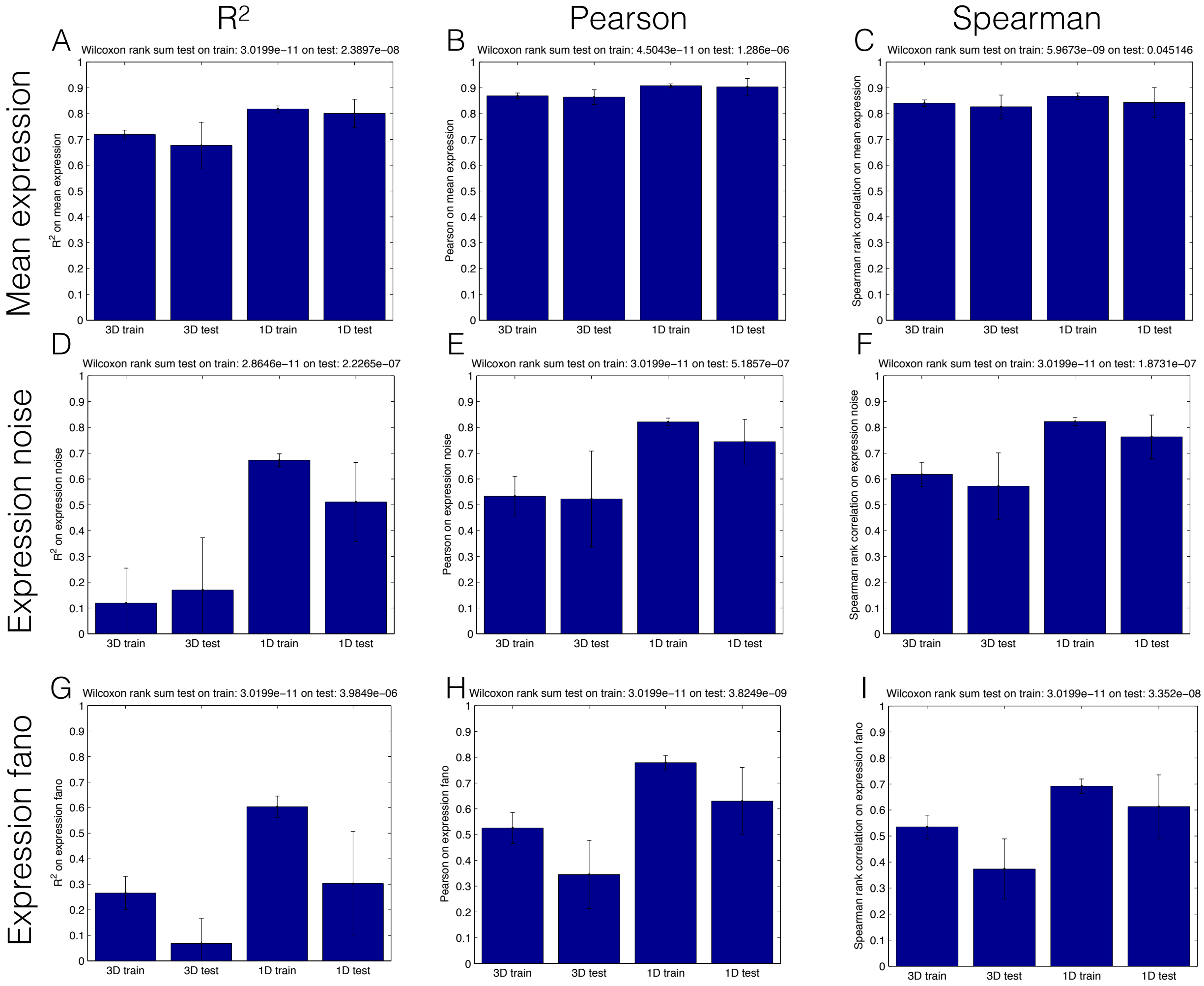


Figure S12. Performance of the kinetic model of transcriptional regulation on train and test sets. Shown are the model performances on mean expression (A,B,C), noise (D,E,F) and fano factor (G,H,I) in terms of R^2 (A,D,G), Pearson's correlation (B,E,H) and Spearman's rank correlation (C,F,I) for both models (3D only and 3D+1D sliding) on train and test data (from a 5-fold cross-validation in 30 bootstraps, see Methods). Plot titles show the P-values for Wilcoxon rank sum tests on the train and test data to compare the 3D model performance with the 1D model performance. The significant P-values indicate that the 1D model performs significantly better.

Fig. S13:

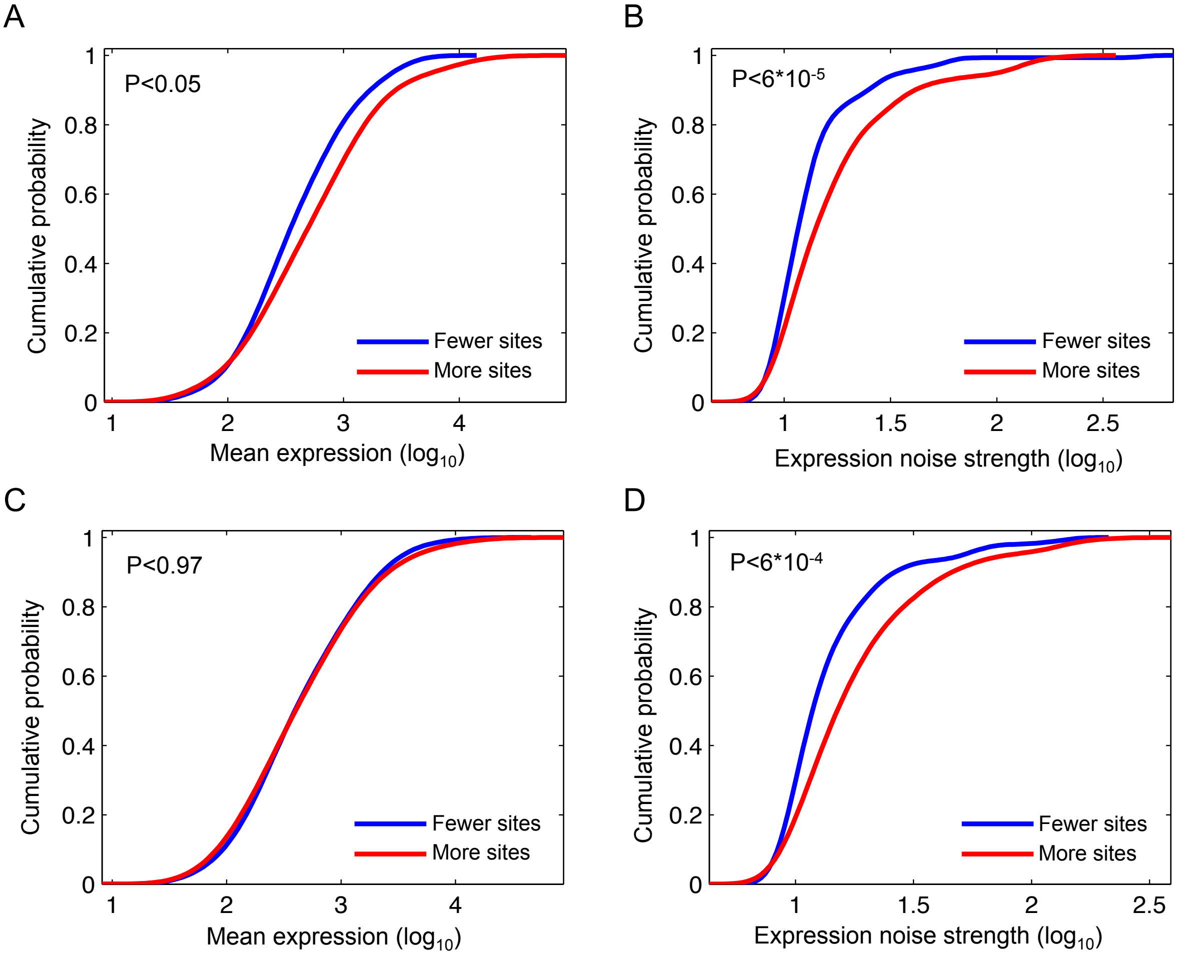


Figure S13. Mean expression and noise as a function of the predicted number of binding sites and measured TF binding in native yeast genes. Shown is the cumulative distribution of measured expression mean (**A,C**) and noise strength (**B,D**) of a set of 466 native yeast genes(Stewart-Ornstein et al. 2012). Genes are split by their high (red line) and low (blue) TF binding as measured by ChIP-seq(Venters et al. 2011) (**A,B**) or average of all TF PSSM scores(Basehoar et al. 2004; MacIsaac et al. 2006; Portales-Casamar et al. 2010; Pachkov et al. 2013)(**C,D**). P-values are computed using Kolmogorov-Smirnov test. Notice that while native genes with more binding sites may or may not have higher expression (**A,C**), they have, on average, significantly higher noise (**B,D**) for both methods of estimating TF binding.

Fig. S14:

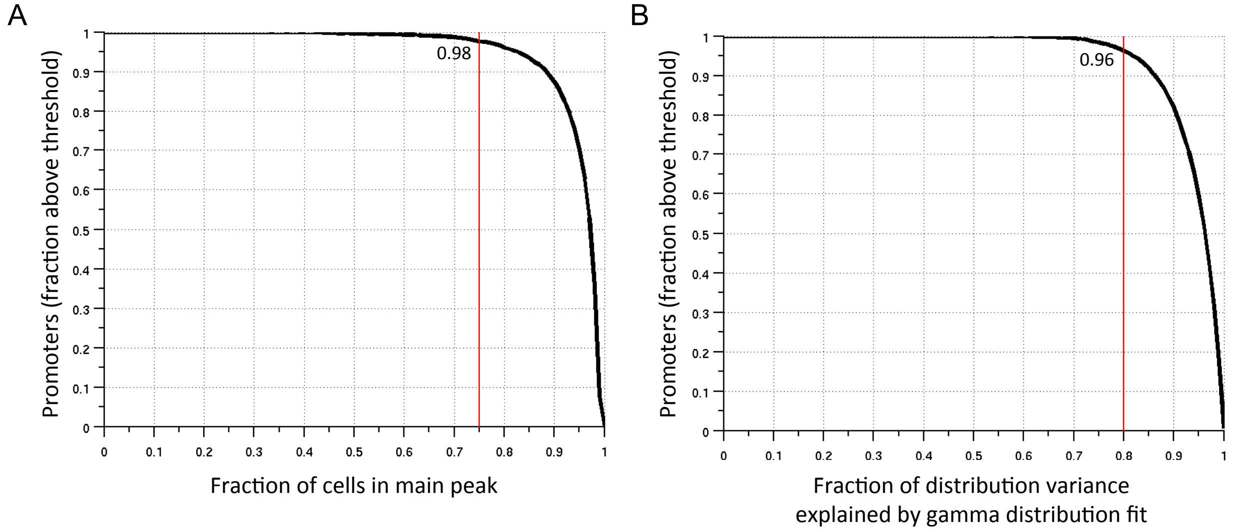


Figure S14. Filtering low quality expression measurements. (A) Shown is the cumulative distribution, for all promoters, of the fraction of cells in the main detected peak of the distribution. The main peak contains 75% of the cells for 98% of the promoters. (**B)** Shown is the cumulative distribution, for all promoters, of the fraction of the distribution across the bins that is explained by fitting a gamma distribution. This is an estimate of how well the derived expression distributions are matching the expected gamma distribution shape. A gamma distribution can explain 80% of the distribution across the bins for 96% of the promoters.

Fig. S15:

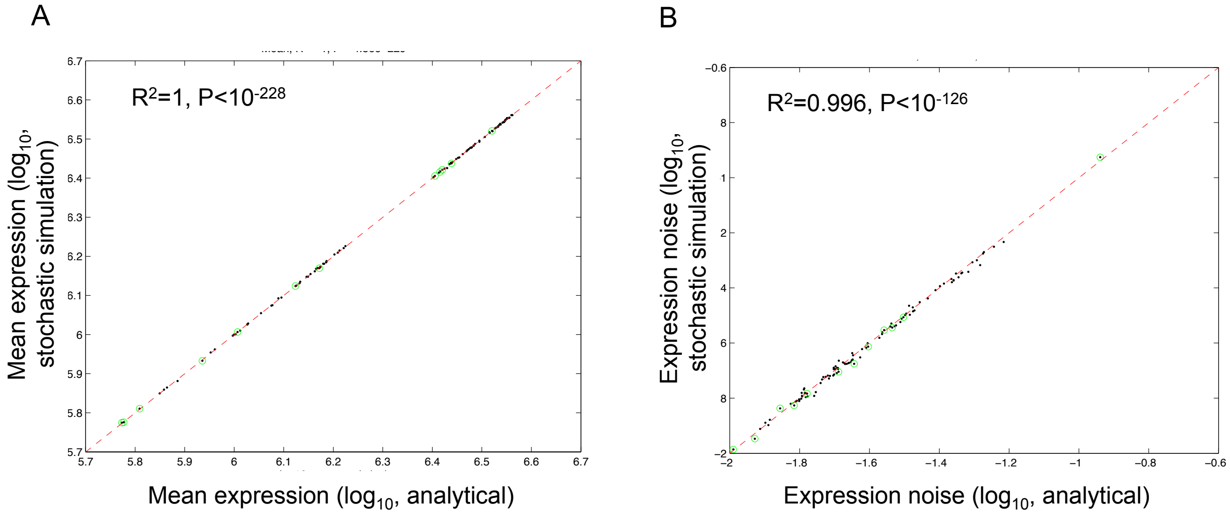


Figure S15 Analytical solution versus stochastic simulations. (**A**) Mean expression predicted per construct using an analytical solution (x-axis) versus mean expression predicted using stochastic simulations using the Gillespie algorithm(Gillespie 1977) (y-axis). (**B**) Noise predicted per construct using an analytical solution (x-axis) versus noise predicted using stochastic simulations using the Gillespie algorithm (y-axis).

Fig. S16:

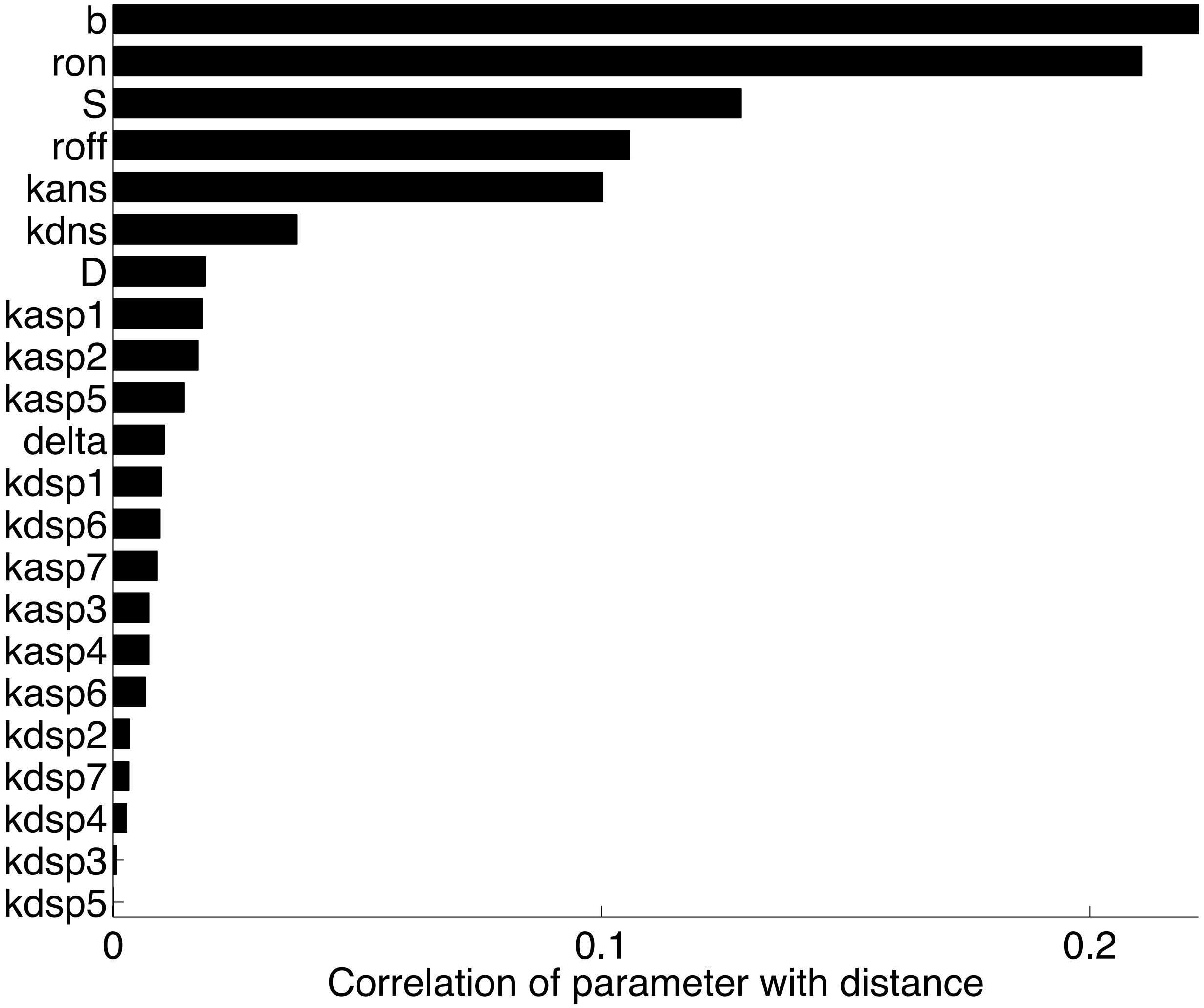
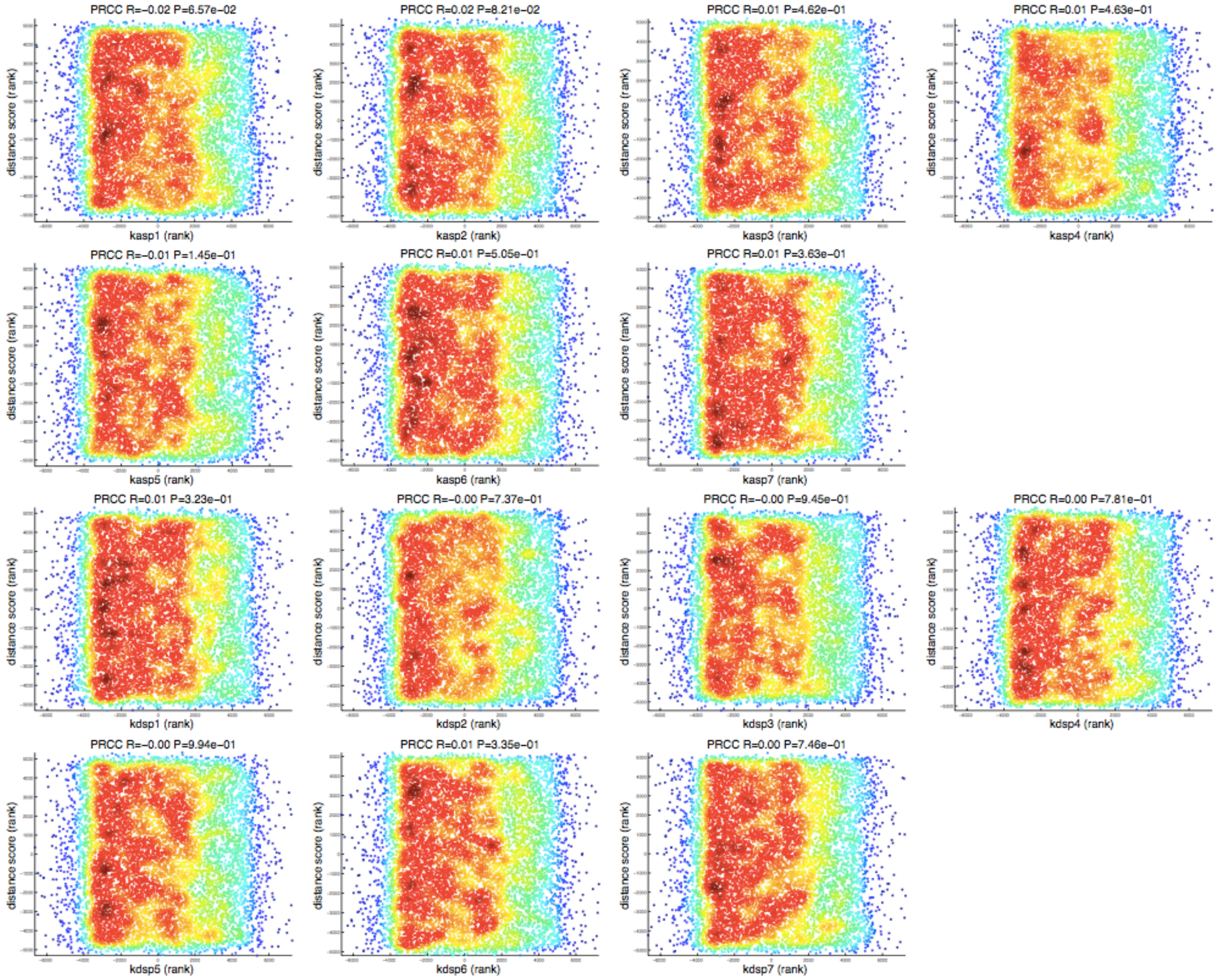


Figure S16 Sensitivity analysis. Shown are the parameter sensitivities obtained from the LHS-PRCC sensitivity analysis. See supplemental methods for details of the procedure.

Fig. S17:


PRCC R=−0.02 P=6.57e−02
distance score (rank)
kasp1 (rank)
PRCC R=0.02 P=8.21e−02
distance score (rank)
kasp2 (rank)
PRCC R=0.01 P=4.62e−01
distance score (rank)
kasp3 (rank)
PRCC R=0.01 P=4.63e−01
distance score (rank)
kasp4 (rank)
PRCC R=−0.01 P=1.45e−01
distance score (rank)
kasp5 (rank)
PRCC R=0.01 P=5.05e−01
distance score (rank)
kasp6 (rank)
PRCC R=0.01 P=3.63e−01
distance score (rank)
kasp7 (rank)
PRCC R=0.01 P=3.23e−01
distance score (rank)
kdsp1 (rank)
PRCC R=−0.00 P=7.37e−01
distance score (rank)
kdsp2 (rank)
PRCC R=−0.00 P=9.45e−01
distance score (rank)
kdsp3 (rank)
PRCC R=0.00 P=7.81e−01
distance score (rank)
kdsp4 (rank)
PRCC R=−0.00 P=9.94e−01
distance score (rank)
kdsp5 (rank)
PRCC R=0.01 P=3.35e−01
distance score (rank)
kdsp6 (rank)
PRCC R=0.00 P=7.46e−01
distance score (rank)
kdsp7 (rank)

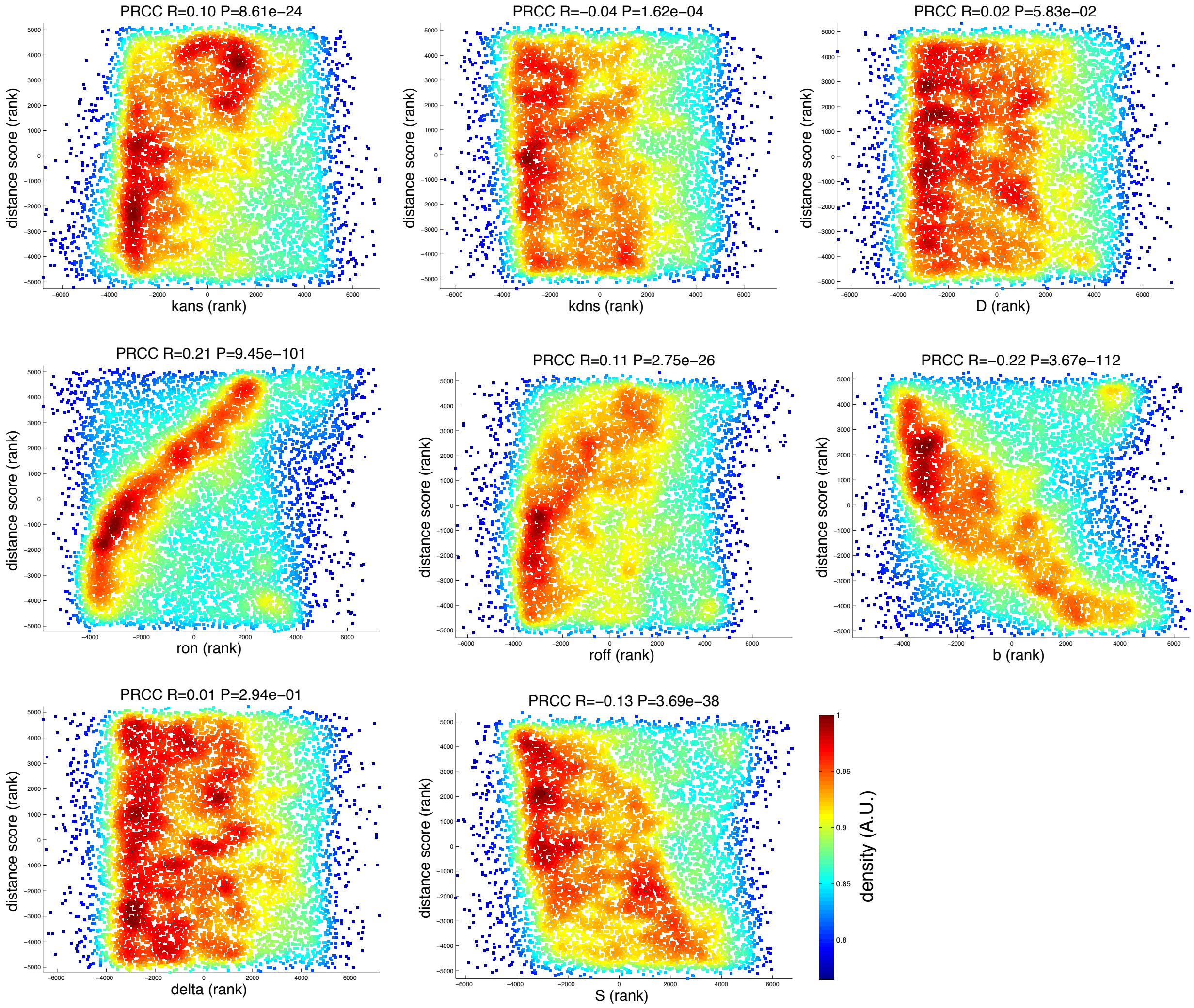


Figure S17 Sensitivity analysis. Shown are the sensitivity plots for each of the model's parameters. X-axes show the parameter values that were sampled randomly using latin hypercube sampling. Y-axes show the rank transformed distance scores of the model to the measured data. Plot titles show the PRCC (ranked correlation) coefficient and P-value. The colors represent density of the scatter plot (in arbitrary units). See supplemental methods for details.

Fig. S18:

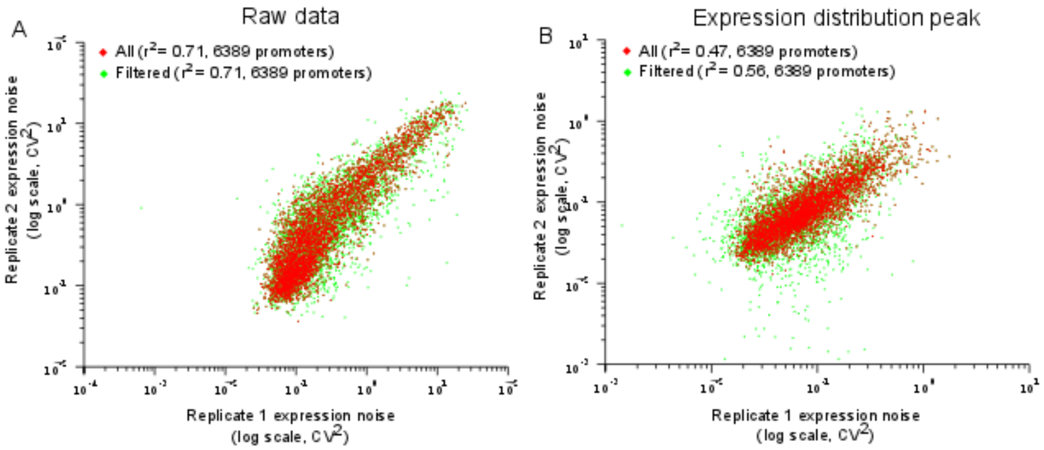


Figure S18 Reproducibility of noise level estimates form raw data and from expression distribution peak. A comparison of expression noise measurements estimated directly from raw data **(A)** or from expression distribution peaks **(B)**. The measurements were obtained for two independent replicates (x-axis: replicate 1, y-axis: replicate 2) for all 6500 constructs in the library (green points) or the subset of the promoters that passed our quality control filter (red dots). Compare this result with similar measurements extracted from a gamma fitted data in **Fig. S1**.

Fig. S19:

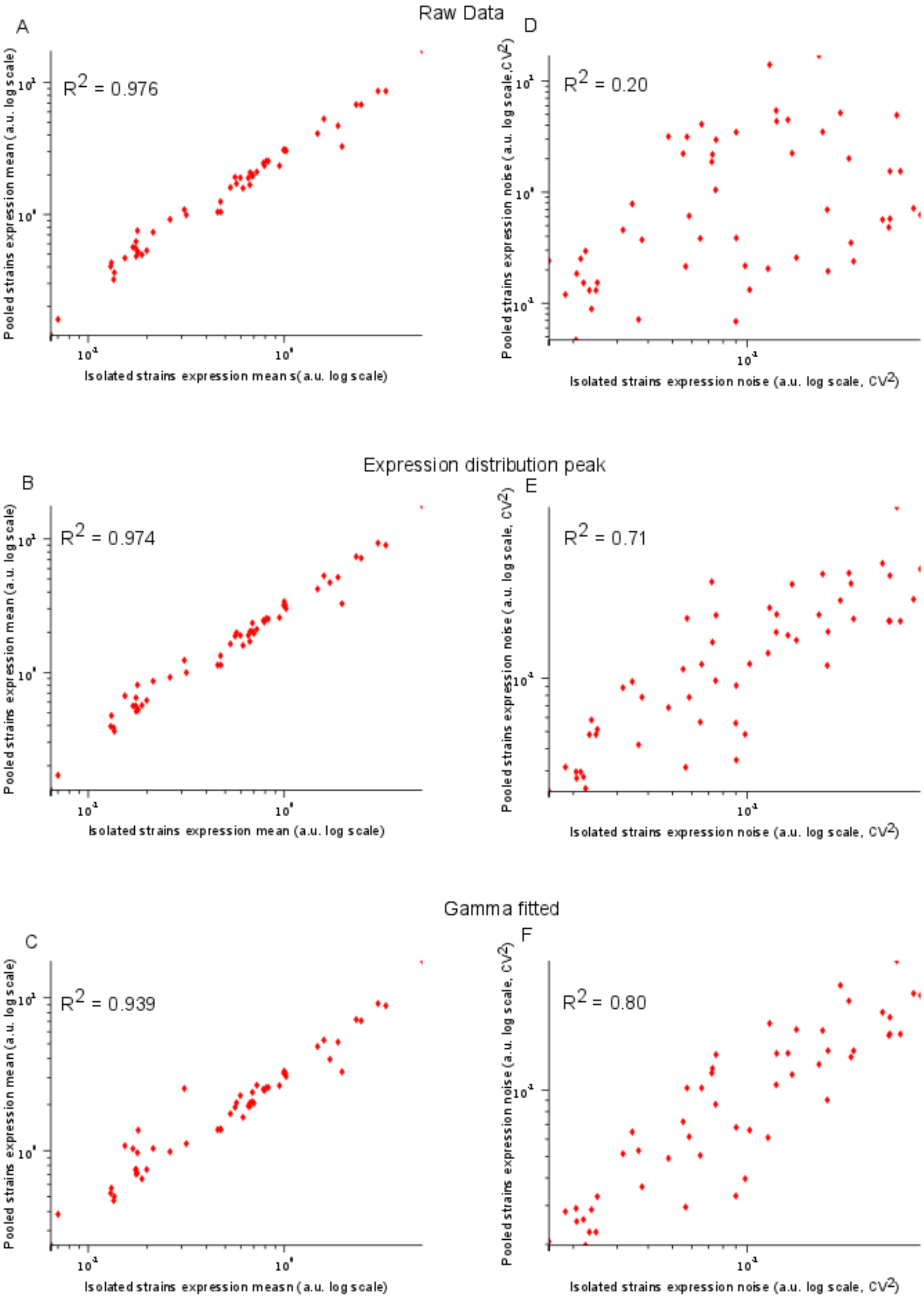

Raw Data
A
R^2 = 0.976
Pooled strains expression mean (a.u. log scale)
10^1
10^0
10^{-1}
10^0
Isolated strains expression mean s(a.u. log scale)
D
R^2 = 0.20
Pooled strains expression noise (a.u. log scale,CV^2)
10^1
10^0
10^{-1}
10^{-1}
Isolated strains expression noise (a.u. log scale, CV^2)
Expression distribution peak
B
R^2 = 0.974
Pooled strains expression mean (a.u. log scale)
10^1
10^0
10^{-1}
10^0
Isolated strains expression mean (a.u. log scale)
E
R^2 = 0.71
Pooled strains expression noise (a.u. log scale, CV^2)
10^{-1}
10^{-1}
Isolated strains expression noise (a.u. log scale, CV^2)
Gamma fitted
C
R^2 = 0.939
Pooled strains expression mean (a.u. log scale)
10^1
10^0
10^{-1}
10^0
Isolated strains expression measn (a.u. log scale)
F
R^2 = 0.80
Pooled strains expression noise (a.u. log scale, CV^2)
10^{-1}
10^{-1}
Isolated strains expression noise (a.u. log scale, CV^2)

Figure S19 Comparison of three methods of estimating expression mean and noise levels from our data. A comparison of expression mean **(A-C)** and noise **(D-F)** levels estimated directly from the raw data **(A,D)**, estimated from the data expression distribution peak **(B,E)** and estimated from expression distribution peak by fitting a gamma distribution to is **(C,F)**. Pooled measurements (y-axis) were compared to measurements of 54 individual strains of transformed yeast cells isolated from the pool, sequenced and measured in isolation using a flow cytometer (x-axis).

Table S1: Promoter configurations and expression noise values
For each promoter configuration (shown as a binary vector of length 7 representing the presence of a binding site in each of the 7 predefined positions) and sequence context (Gal1 10 or His3) the measured and predicted mean and noise values (in log10) are given. NaN values mean that the value didn't pass our quality control filters.

Table S2: Fitted parameter values of the kinetic model. Shown are the lower and upper bounds for each parameter that were used in the fitting procedure, and the final fitted values. Are values are in log10 and in min^{-1}. Kasp1 to kasp7 are the specific binding rates. Kdsp1 to kdsp7 are the specific unbinding rates. Kans (gal110 and his3 for each context) and kdns are the non-specific binding and unbinding rates respectively. D is the 1D diffusion coefficient, from which we compute the sliding distance (s) through s = $(D/kd_{ns})^{1/2}$. Ron is the contribution to transcription per bound site. Roff is the transcription rate of the unbound state. B is the average burst size (proteins per mRNA). Delta is the protein degradation rate, mostly due to dilution from cell division. S is a scaling factor to convert protein abundance to fluorescence units.

Table S3: Replicate 1 mapping of strains to expression bins. Fraction of cells containing each promoter any of the 32 expression bins.

Table S4: Replicate 2 mapping of strains to expression bins. Fraction of cells containing each promoter any of the 32 expression bins.

Table S5: Replicate 1 expression bin values. The standard deviation, mean, minimal value (left edge) and maximal value (right edge) of expression level of cells sorted to each expression bin.

Table S6: Replicate 2 expression bin values. The standard deviation, mean, minimal value (left edge) and maximal value (right edge) of expression level of cells sorted to each expression bin.

Table S7: Promoters expression values. The table contains for each promoter the number of sequencing reads mapped to the promoter, expression mean and expression noise in the two replicates. Library id is similar to Sharon et al.(Sharon et al. 2012).

Supplementary References and Notes

Basehoar AD, Zanton SJ, Pugh BF. 2004. Identification and distinct regulation of yeast TATA box-containing genes. *Cell* **116**: 699–709. http://www.sciencedirect.com/science/article/pii/S0092867404002053 (Accessed July 12, 2013).

Carey LB, van Dijk D, Sloot PMA, Kaandorp JA, Segal E. 2013. Promoter sequence determines the relationship between expression level and noise. *PLoS Biol* **11**: e1001528. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3614515&tool=pmcentrez&rendertype=abstract (Accessed June 3, 2013).

Gillespie DT. 1977. Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* **81**: 2340–2361. http://pubs.acs.org/doi/abs/10.1021/j100540a008 (Accessed May 29, 2013).

Hammar P, Leroy P, Mahmutovic A, Marklund EG, Berg OG, Elf J. 2012. The lac repressor displays facilitated diffusion in living cells. *Science* **336**: 1595–8. http://www.ncbi.nlm.nih.gov/pubmed/22723426 (Accessed June 26, 2013).

Lubliner S, Keren L, Segal E. 2013. Sequence features of yeast and human core promoters that are predictive of maximal promoter activity. *Nucleic Acids Res*. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3675475&tool=pmcentrez&rendertype=abstract (Accessed June 11, 2013).

MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E. 2006. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics* **7**: 113. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1435934&tool=pmcentrez&rendertype=abstract (Accessed May 23, 2013).

Marino S, Hogue IB, Ray CJ, Kirschner DE. 2008. A methodology for performing global uncertainty and sensitivity analysis in systems biology. *J Theor Biol* **254**: 178–96. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2570191&tool=pmcentrez&rendertype=abstract (Accessed March 30, 2014).

Pachkov M, Balwierz PJ, Arnold P, Ozonov E, van Nimwegen E. 2013. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Res* **41**: D214–20. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531101&tool=pmcentrez&rendertype=abstract (Accessed June 20, 2013).

Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A. 2010. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res* **38**: D105–10. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2808906&tool=pmcentrez&rendertype=abstract (Accessed May 21, 2013).

Raser JM, O'Shea EK. 2004. Control of stochasticity in eukaryotic gene expression. *Science* **304**: 1811–4. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1410811&tool=pmcentrez&rendertype=abstract (Accessed June 26, 2013).

Sanchez A, Garcia HG, Jones D, Phillips R, Kondev J. 2011. Effect of promoter architecture on the cell-to-cell variability in gene expression. *PLoS Comput Biol* **7**: e1001100. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3048382&tool=pmcentrez&rendertype=abstract (Accessed June 26, 2013).

Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A, Segal E. 2012. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* **30**: 521–30. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3374032&tool=pmcentrez&rendertype=abstract (Accessed June 6, 2013).

Stewart-Ornstein J, Weissman JS, El-Samad H. 2012. Cellular noise regulons underlie fluctuations in Saccharomyces cerevisiae. *Mol Cell* **45**: 483–93. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3327736&tool=pmcentrez&rendertype=abstract (Accessed June 26, 2013).

Venters BJ, Wachi S, Mavrich TN, Andersen BE, Jena P, Sinnamon AJ, Jain P, Rolleri NS, Jiang C, Hemeryck-Walsh C, et al. 2011. A comprehensive genomic binding map of gene and chromatin regulatory proteins in Saccharomyces. *Mol Cell* **41**: 480–92. http://www.sciencedirect.com/science/article/pii/S1097276511000426 (Accessed May 28, 2013).