Supplementary Materials for

Promoter Architecture Dictates Cell-to-Cell Variability in Gene Expression

Daniel L. Jones, Robert C. Brewster, Rob Phillips*

*correspondence to: phillips@pboc.caltech.edu

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S11
Tables S1 to S3
Full Reference List

Contents

1 Materials and Methods 3

1.1 Strains 3

1.2 Growth 4

1.3 mRNA FISH 4

1.3.1 Fixation and labeling 4

1.3.2 FISH data acquisition 5

1.3.3 FISH analysis 6

1.4 Miller LacZ assay 9

2 Supplementary Text 10

2.1 Calibration of mRNA FISH data versus Miller assay 10

2.2 Estimation of Additional Noise Sources 11

2.2.1 Quantification error in image analysis 11

2.2.2 Gene copy number variation 17

2.2.3 Extrinsic noise due to repressor copy number fluctuations 20

2.2.4 Extrinsic noise due to RNAP copy number fluctuations 21

2.2.5 Extrinsic sources of noise: concluding remarks 26

2.3 mRNA histograms for constitutive expression 27

2.4 Error bars in Fano factor measurements 28

2.5 Copy Number Variation: Uncorrected Figures 29

2.6 Testing gene copy number noise by cell size segregation 29

2.7 Determination of Rate Parameter Values 31

3 Supplementary Figures 33

4 Supplementary Tables 44

1 Materials and Methods

1.1 Strains

The genetic modifications used to create the strains used in the various sections of the main paper are listed below with a bold heading characterizing which parameter is tuned within that class of strains.

Constitutive expression: tuning r. As described in [19], promoter constructs consist of an RNAP binding site with a LacI O2 binding site immediately downstream of the transcription start site, as shown in Fig. 1A. (The O2 binding site does not affect expression because *lacI* is deleted from the host strain used in gene expression measurements.) Expression of a LacZ reporter is tuned over a factor of ≈ 500 by changing the DNA sequence of the RNAP binding site. Promoter + reporter constructs are chromosomally integrated at the *gal* locus in an *E. coli* strain (HG105) in which *lacI* and *lacZYA* are deleted.

Simple repression: tuning $k_{\mathbf{on}}^{\mathbf{R}}$. In two of the constitutive promoter constructs (*lac*UV5 and 5DL1) the O2 LacI binding site is replaced with an Oid LacI binding site. This construct is integrated into an *E. coli* strain (RCB110) in which wild-type *lacIZYA* is deleted as before, but with the addition of a genetic circuit allowing inducible control of LacI expression. This circuit consists of two components: first, the *tet* repressor TetR is chromosomally integrated at the *gspI* locus under the control of the strong constitutive promoter $\mathrm{P}_{\mathrm{N25}}$; and second, LacI is integrated at the *ybcN* locus under the control of the $\mathrm{P}_{\mathrm{LtetO\text{-}1}}$ promoter [24], which is repressed by TetR. Expression of LacI can thus be induced by the small molecule anhydrotetracycline (aTc), which interacts with TetR and prevents it from repressing transcription of LacI. The ribosomal binding sequence of the LacI is the “1147” version from reference [25]: at full induction this produces roughly 40 LacI per cell. By varying the aTc concentration, we tune mean gene expression by tuning the intracellular concentration of LacI, yielding the curves shown in Fig. 3A.

Simple repression: tuning $k_{\mathbf{off}}^{\mathbf{R}}$. These measurements exploit the same strain (RCB110) from the k_R^{on} tuning; however in this case we use only the *lac*UV5 promoter strain and create constructs in which the O2 LacI binding site is replaced by O1, O3 or Oid. These constructs are integrated into the *galK* locus as before, yielding four constructs total each with a different LacI binding site. These strains are measured at constant inducer concentration (to achieve equal repressor copy numbers across all samples) for three distinct induction conditions. This was done for each of 3 different LacI concentrations as shown in Fig. 3B.

1.2 Growth

Cultures are grown overnight to saturation (at least 8 hours) in LB and diluted 1:4000 into 30 mL of M9 minimal media supplemented with 0.5% glucose in a 125mL baffled flask. Growth in minimal media continues approximately 8 hours and cells are harvested in exponential phase when OD600$= 0.3 - 0.5$ is reached.

1.3 mRNA FISH

1.3.1 Fixation and labeling

Our assay is based on that used in Ref. [5]. Once a culture reaches OD600$= 0.3-0.5$, it is immersed in ice for 15 minutes before being harvested in a large centrifuge chilled to $4°C$ for 5 minutes at 4500 g. The cells are then fixed by resuspending in 1 mL of 3.7% formaldehyde in 1x PBS which is then mixed gently at room temperature for 30 minutes. Next, they are centrifuged (8 minutes at 400 g) and washed twice in 1 mL of 1x PBS twice. The cells are permeabilized by resuspension in 70% Ethanol which proceeds, with mixing, for 1 hour at room temperature. The cells are then pelleted (centrifuge at 600 g for 7 minutes) and resuspended in 1 mL of 20% wash solution (200 μL formamide, 100 μL 20x SSC, 700 μL water). This mixture is allowed to sit several minutes before centrifugation (7 minutes at 600g) and resuspended in 50 μL of DNA probes (consisting of a mix of 72 unique DNA probes; individual oligo sequences listed in Table S1) labelled with ATTO532 dye (Atto-tec) in hybridization solution (0.1 g dextran sulfate, 0.2 mL formamide, 1 mg *E.coli* tRNA, 0.1 mL 20x SSC, 0.2 mg BSA, 10 μL of 200 mM Ribonucleoside vanadyl complex). This hybridization reaction is allowed to proceed overnight. The hybridized product is then washed four times in 20% wash solution before imaging in 2x SSC.

1.3.2 FISH data acquisition

Samples are imaged on a 1.5% agarose pad made from PBS buffer. Each field of view is imaged with phase contrast at the focal plane and with 532 nm epifluorescence (Verdi V2 laser, Coherent Inc.) both at the focal plane and in 8 z-slices spaced 200 nm above and below the focal plane (for a total of 17 slices), sufficient to cover the entire depth of the *E. coli.* The images are taken with an EMCCD camera (Andor Ixon2) under 150× magnification. The phase image is used for cell segmentation and the fluorescence images are used in mRNA detection. A total of 100 unique fields of view are imaged in each sample and a typical field of view has between 5 and 15 viable cells (cells which are touching and cells that have visibly begun to divide are ignored) resulting in roughly 900 individual cells per sample on average. However, due to differences in plating density and position quality, the actual number can vary. A histogram of the sample size for all samples in this study is shown in fig. S1.

1.3.3 FISH analysis

The FISH data is analyzed in a series of Matlab (The Mathworks) routines. The overview of the workflow is as follows: identifying individual cells, segmenting the fluorescence to identify possible mRNA, and quantifying the mRNA which are found (because of the small size of *E. coli*, at high copy number mRNA can be difficult to distinguish and count by eye).

Cell identification and segmentation: In phase contrast imaging, *E. coli* are easily distinguishable from the background and automated programs can identify, segment and label cells with high fidelity. The results of the phase segmentation are manually checked for accuracy: Cells which are touching or overlapping other cells, misidentification of cells or their boundaries or cells which have visibly begun to undergo division are all discarded manually.

Fluorescence segmentation: First we perform several steps to process the raw intensity images. The images are flattened, a process to correct for any uneven elements in the illumination profile, using a flattening image. The flattening image is an average over $10 - 15$ images of an agarose pad coated with a small drop of fluorescein (such that the drop spreads evenly across most of the pad); the resulting image is a map of illumination intensity at any given pixel I_{flat}. Each pixel of every fluorescence image is scaled such that the corresponding pixel in the flattening image would be a uniform brightness (typically each pixel is scaled up to the level of the brightest pixel). This can be achieved by renormalizing each pixel in the data images and dividing by the ratio of the intensity of the corresponding pixel in the flattening image to the intensity of the brightest pixel. In other words the raw images, I, are renormalized such that for pixel i, j with raw intensity $I^{(i,j)}$,

$$I^{(i,j)}_{\text{corrected}} = \left(I^{(i,j)} - I^{(i,j)}_{\text{dark}}\right) \times \left(\frac{\max_{i,j}(I^{(i,j)}_{\text{flat}} - I^{(i,j)}_{\text{dark}})}{I^{(i,j)}_{\text{flat}} - I^{(i,j)}_{\text{dark}}}\right), \tag{S1}$$

where I_{dark} corresponds to an image taken with no illumination (mostly these counts are from camera offset). In the preceding equation, the first term in parentheses is the signal from the i, jth pixel, while the second term in parentheses corrects the signal for nonuniform illumination using the flattening image. We then subtract from every pixel the contribution to our signal associated with autofluorescence. The value for the autofluorescence is obtained by averaging over the fluorescence of every pixel in a control sample (one which underwent the entire FISH protocol but did not possess the *lacZ* gene). Finally, all local 3D maxima (where $x - y$ is the image plane) in fluorescence are identified. We require that the maxima be above a threshold in fluorescence (typically $300 - 400\%$ above the background autofluorescence signal). This threshold eliminates all fluorescence maxima in the control sample, which does not contain the *lacZ* gene.

mRNA quantification: Each identified maximum pixel is dilated in the image plane to a 5×5 box of surrounding pixels. These 5×5 boxes are referred to as “spots”. If multiple spots overlap, the pixels which make up each overlapping spot are merged into one larger spot to avoid double counting the signal from any one pixel. Since, due to the small size of the *E. coli* we cannot guarantee that every spot corresponds to exactly one mRNA, we must have a way to quantify the relation between signal and mRNA copy number. An example histogram of the intensity of identified spots is shown in fig. S2A. The histogram has two clear peaks in probability: one corresponding to background noise, at approximately 0 intensity, and the other corresponding to the intensity of a single mRNA. The low intensity peak, corresponding to background noise, is removed by thresholding the spots and rejecting spots that are less bright than the threshold. The threshold is selected to eliminate spots in a control sample that does not contain the target mRNA. However, we find choice of this threshold does not alter our results significantly since these spots are already significantly dimmer than an mRNA. To determine the calibration between signal intensity and mRNA copy number we take an average over all remaining isolated spots (meaning no merge events with other, nearby spots) in very low expression samples (where the mean $\ll 1$ and mRNA are statistically unlikely to overlap), see fig. S3A and B for examples of these thresholded histograms. Once this single mRNA intensity value is identified, when possible we also verify in other low expression strains that as we increase the mean expression it simply increases the frequency of spots with the single mRNA intensity but does not increase the mean intensity of each spot. An example of this is shown in figs. S3A where the single spot intensity histograms of seven unique strains are shown with each histogram normalized by the number of cells in each sample. The growing peak at 1 mRNA shows that as we transition from low expression towards having upwards of 1 mRNA per cell, the increased signal is primarily due to an increased number of identified single mRNA, although some brighter spots begin to appear corresponding to multiple mRNA per spot. Normalizing these same histograms by the total number of identified spots, as shown in fig S3B, demonstrates that the identified spots have the same character in each sample regardless of mean. The dashed black line shows the result of a Gaussian fit to the combined data from all seven samples in the figure. Finally, the day to day variability in these histograms is shown in fig. S3C for five different acquisitions on two distinct constitutive expression strains.

With this calibration in hand, we sum the signal from all identified spots in a given cell and determine how many mRNA are in that cell by dividing by the single mRNA intensity calibration found previously. A different technique, used in some studies, is to first quantify every individual identified spot then determine the copy number by summing the total number of identified spots in each cell. Fig. S4 shows a direct comparison of these methods on simulated data (crosses) and real data (circles) with red corresponding to the "whole cell" method used in our study, where signal is summed over the whole cell and black corresponding to the single spot analysis where each spot is quantized and rounded to the nearest whole number of mRNA. The methods gives roughly identical values for the mean; however, the corresponding Fano factor is very slightly systematically higher using the single spot analysis, probably due to the rounding performed on each spot.

1.4 Miller LacZ assay

Concurrent with the mRNA FISH protocol, each sample also has LacZ activity measured by Miller assay. The protocol is identical to that described previously [19, 25]. Once cultures are ready for measurement, the OD600 of each sample is recorded. Next, a volume of cells between 5 μL and 200 μL is added to Z-buffer (60mM Na_2HPO_4, 40 mM NaH_2PO_4, 10 mM KCl, 1 mM $MgSO_4$, 50 mM β-mercaptoethanol, pH 7.0) to reach a total of 1 mL in a 1.5 mL Eppendorf tube. The time of the enzymatic reaction is inversely proportional to the volume of cells, and thus low expression samples require more volume and high expression samples require less to ensure that the time of reaction is reasonable ($\sim 1 - 10$'s of hours) to avoid measurement uncertainty, and to ensure that the yellow color is easily distinguishable from a blank sample of 1 mL of Z-buffer. The cells are lysed with 25 μL of 0.1% SDS and 50 μL of chloroform, mixed by a 10s vortex. To begin the reaction, 200 μL of 4mg/mL 2-nitrophenyl β-D-galactopyranoside (ONPG) in Z-buffer is added to the Eppendorf tube. The tube is monitored for the development of yellow color and once sufficient yellow has developed in a sample (sufficient absorbance at 420nm, without saturating the reading), the reaction is stopped through the addition of 200 μL of 2.5 M Na_2CO_3. Once all samples have been stopped, cell debris is removed from the supernatant by centrifugation at $> 13,000$ g for 3 min. 200 μL of each sample, including the blank which contains no cells, are loaded into a 96 well plate and absorbance at 420nm and 550nm is measured for each well with a Tecan Safire2. The LacZ activity in Miller units is then,

$$MU = 1000\frac{OD420 - 1.75 \times OD550}{t \times v \times OD600}0.826, \tag{S2}$$

where t is the reaction time in minutes, v is the volume of cells used in mL and OD refers to the optical density measurements obtained from the plate reader. The factor of 0.826 accounts for the use of 200 μL Na_2CO_3 as opposed to 500 μL which changes the concentration of ONPG in the final solution. While some alternative protocols involve time resolved measurements of LacZ activity over a range of cell densities, we believe the protocol used here is a simple and accurate method for providing a consistent relative calibration for our mRNA measurements that has been shown to be equivalent in terms of accuracy and reproducibility to more complicated and time consuming measurement protocols [25].

2 Supplementary Text

2.1 Calibration of mRNA FISH data versus Miller assay

As a test of our ability to accurately measure mean mRNA copy number with mRNA FISH, we directly compare our results to simultaneously acquired measurements of mean LacZ enzymatic activity (the protein produced by the mRNA targeted in our

FISH labelling). In fig. S5, we show the mean mRNA expression vs mean LacZ activity in Miller units for every measurement of every strain in this study. These two measurements of expression give consistent results as demonstrated by direct proportionality between these two measurement techniques over several orders of magnitude of expression. Error bars represent the standard deviation from repeated measurements.

2.2 Estimation of Additional Noise Sources

2.2.1 Quantification error in image analysis

As described in the main text, the intensity of a single mRNA molecule is determined using a low mRNA expression sample such that detected spots are most likely to contain either 0 (*i.e.* the fluorescence maxima is due to background noise) or 1 mRNA. A representative histogram of detected spots is shown in fig. S2A. However, this identification and quantification process is not without uncertainty of its own. While ideally each spot should have the same clear value for its integrated intensity, the intensity of a given identified mRNA varies significantly, due to factors such as fluctuations in probe hybridization efficiency and non-specifically bound probes. We wish to estimate the effect that variability in the single mRNA intensity has on the overall observed variability. In order to do so, we will make the following assumptions:

- mRNA copy numbers are distributed with mean μ_m and variance σ_m^2. Aside from this, we make no assumptions about the specific form of the mRNA distribution.

- Integrated intensities for a single mRNA are distributed with mean $\langle I_1 \rangle$ and variance σ_I^2.

- Intensities for more than one mRNA are independent and additive. For cells containing k mRNA, integrated intensities have mean $k\langle I_1\rangle$ and variance $k\sigma_I^2$ (since variances add for independent random variables).

In this work we sought to measure the Fano factor for various mRNA copy number distributions. However, what we actually measure experimentally (as described above) is the following:

$$\mathrm{Fano}_{\mathrm{exp}} = \mathrm{Fano}\left(\frac{I}{\langle I_1\rangle}\right) = \frac{1}{\langle I_1\rangle}\frac{\mathrm{var}(I)}{\langle I\rangle} \tag{S3}$$

where I is a random variable denoting the integrated intensity of a cell, and $\langle I_1\rangle$ is the mean intensity of a single mRNA. That is, we measure the Fano factor of a set of observed integrated intensities divided by the single mRNA intensity. We will now use the assumptions listed above to further investigate eq. S3, and proceed by computing the mean and variance of I.

The random variable I is distributed according to:

$$P(I) = \sum_{k=0}^{\infty} p_m(k) p_I(I|k) \tag{S4}$$

where $p_m(k)$ is the probability that a cell contains k mRNA, and $p_I(I|k)$ is the probability of obtaining intensity I given that a cell contains k mRNA. We will denote the conditional expectation of the nth moment of I given k as $\langle I^n|k\rangle$: *i.e.*,

$\langle I^n|k\rangle = \int I^n p_I(I|k)dI$. Then the expected value of I is given by

$$\langle I\rangle = \int_{-\infty}^{\infty} IP(I)dI \tag{S5}$$

$$\langle I\rangle = \int_{-\infty}^{\infty} I \sum_{k=0}^{\infty} p_m(k)p_I(I|k)dI \tag{S6}$$

$$\langle I\rangle = \sum_{k=0}^{\infty} p_m(k) \int Ip_I(I|k)dI \tag{S7}$$

$$\langle I\rangle = \sum_{k=0}^{\infty} p_m(k)\langle I|k\rangle \tag{S8}$$

$$\langle I\rangle = \sum_{k=0}^{\infty} p_m(k)k\langle I_1\rangle = \langle k\rangle\langle I_1\rangle \tag{S9}$$

$$\langle I\rangle = \mu_m\langle I_1\rangle \tag{S10}$$

which is exactly what one would naively expect (the mean integrated intensity equals the mean number of mRNA times the mean single mRNA intensity). To compute the variance of I, we next need to compute the expected value of I^2:

$$\langle I^2\rangle = \int_{-\infty}^{\infty} I^2P(I)dI \tag{S11}$$

$$\langle I^2\rangle = \int_{-\infty}^{\infty} I^2 \sum_{k=0}^{\infty} p_m(k)p_I(I|k)dI \tag{S12}$$

$$\langle I^2\rangle = \sum_{k=0}^{\infty} p_m(k) \int_{-\infty}^{\infty} I^2p_I(I|k)dI \tag{S13}$$

$$\langle I^2\rangle = \sum_{k=0}^{\infty} p_m(k)\langle I^2|k\rangle. \tag{S14}$$

According to the assumptions above,

$$\mathrm{var}(I|k) = \langle I^2|k\rangle - \langle I|k\rangle^2 = k\sigma_I^2 \tag{S15}$$

and hence

$$\langle I^2|k\rangle = k^2\langle I_1\rangle^2 + k\sigma_I^2. \tag{S16}$$

Plugging this back into eq. S14, we obtain:

$$\langle I^2\rangle = \langle I_1\rangle^2 \sum_{k=0}^{\infty} k^2 p_m(k) + \sigma_I^2 \sum_{k=0}^{\infty} k p_m(k) \tag{S17}$$

$$\langle I^2\rangle = \langle k^2\rangle\langle I_1\rangle^2 + \langle k\rangle\sigma_I^2 \tag{S18}$$

$$\langle I^2\rangle = (\mu_m^2 + \sigma_m^2)\langle I_1\rangle^2 + \mu_m\sigma_I^2 \tag{S19}$$

where we have used the fact that $\langle k^2\rangle = \sigma_m^2 + \mu_m^2$. Hence

$$\mathrm{var}(I) = (\mu_m^2 + \sigma_m^2)\langle I_1\rangle^2 + \mu_m\sigma_I^2 - \mu_m^2\langle I_1\rangle^2 \tag{S20}$$

$$\mathrm{var}(I) = \sigma_m^2\langle I_1\rangle^2 + \mu_m\sigma_I^2 \tag{S21}$$

and

$$\mathrm{Fano}_{\mathrm{exp}} = \frac{1}{\langle I_1\rangle}\frac{\sigma_m^2\langle I_1\rangle^2 + \mu_m\sigma_I^2}{\mu_m\langle I_1\rangle} \tag{S22}$$

$$\mathrm{Fano}_{\mathrm{exp}} = \frac{\sigma_m^2}{\mu_m} + \frac{\sigma_I^2}{\langle I_1\rangle^2}. \tag{S23}$$

The two terms in eq. S23 have simple interpretations. The first term is the Fano factor of the actual underlying mRNA distribution. The second term reflects uncertainty in the intensity of a single mRNA spot and is essentially the squared coefficient of variation of the intensity of a single mRNA spot. This value depends slightly on the conditions of the specific acquisition. For instance, the single mRNA peaks from one experiment are shown in fig. S3B. For this acquisition, one can fit a Gaussian to the observed single spot mRNA intensity distribution (dashed black line) and make a measurement of both the mean and standard deviation of this distribution to calculate the expected contribution to the Fano factor from quantization error, as in eq. S23. For this acquisition, we see that $\sigma_I^2/\langle I_1\rangle^2 = 0.16$. This result is typical for our experiments (as demonstrated in fig. S3C) and therefore the green shaded region in Fig. 2B has a height equal to 0.16. Of course, this value is not static and depends on the particular acquisition conditions and could be calculated for each separate acquisition independently. However, these values are small enough relative to the range of Fano factors (≈ 1 to 8) observed in our experiments that this effect will not change the qualitative conclusions reached in this work.

As a complementary test of the performance of our image analysis routines, we created simulated FISH data sets at a variety of mRNA expression levels. Our goal was as much as possible to faithfully reproduce the images coming off of our microscope. We acquire raw microscopy data by spotting FISHed *E. coli* cells on agarose pads, mounting the cells on the microscope, and running an automated acquisition script. The script generates a grid of ≈ 100 positions on each pad; at each position, a phase contrast image is taken for segmentation purposes, followed by a fluorescence z stack (separated by 0.2 μm) to image mRNAs. Our simulated data thus also consisted of sets of ≈ 100 "positions", with each position consisting of a simulated phase contrast image and a simulated fluorescence z stack. The data generation algorithm at each position can be roughly described as follows:

1. Generate a phase contrast image. 25 "*E. coli*" cells are placed at random in a field of view. Cells are modeled as ellipsoids 22 pixels in length, 10 pixels wide, and 4 pixels tall.

2. Determine the number of mRNA copies in each cell. For each cell, the number of mRNA it will contain is drawn from the appropriate probability distribution

(for instance, from a Poisson distribution with a given mean).

3. Determine the spatial distribution of mRNAs within a cell. For each cell, mRNA are distributed uniformly at random within the cell. For instance, if a cell has 4 mRNA assigned to it, then 4 pixels within the cell are chosen at random, with each one corresponding to the center of an mRNA.

4. Determine the intensity of each mRNA. As seen in fig. S2, the integrated fluorescence intensity of a single mRNA can vary substantially. We choose the intensity of each mRNA from a Gaussian distribution with mean 0.4 and standard deviation 0.16 (thus $\sigma_I^2/\langle I_1 \rangle^2 = 0.16$ as in fig. S3B). These values were chosen as reasonable representations of our physical data sets. Each mRNA pixel (as determined in the previous step) is assigned a fluorescence intensity drawn from this distribution.

5. Convolve with point-spread-function. In reality mRNA do not show up as single bright pixels but rather as diffraction-limited spots. To simulate this we convolve the fluorescence stack with a Gaussian point-spread function with a standard deviation of 0.875 pixels. This value was chosen as a reasonable representation of the point spread functions observed in our actual data.

6. Generate background and noise. In addition to the signal from actual mRNA molecules, our images are subject to background from cellular autofluorescence and the agarose pad, as well as noise from unbound or non-specifically bound fluorescent probes. To simulate this, a random fluorescence background is generated for each cell by drawing pixel values from a geometric distribution with mean 466 (reflecting typical mean background fluorescences encountered in experimental data), convolving with a Gaussian with mean 1.0 pixels (to reflect

spatially correlated noise from e.g. unbound probes), then adding these values to the "signal" as determined in the previous step. This background is added to a constant offset of 1080 counts to mimic a typical camera offset.

In fig. S4 we show the measured Fano factor for simulated data for a population of cells with Poisson distributed mRNA copy number (circles) using two distinct mRNA quantification schemes as described in the methods. As expected, even at low mRNA expression, the measured Fano factor is greater than the correct value of 1, due to the variability in intensity measured for a single mRNA. The single mRNA intensity distribution is a Gaussian with $\sigma_I^2/\langle I_1\rangle^2 = 0.16$ and is designed to mimic our experimental data (see fig. S2B). Eq. S23 thus predicts that the measured Fano factor will be 1.16, this prediction is shown as the green bar in fig. S4 for comparison to the Fano factor in the simulated data. For reference, our Fano factor measurements (with gene copy number noise subtracted) for the constitutive expression strains are shown as crosses. While the quantification noise matches at low means, at higher means RNAP fluctuations in the real data are likely also contributing to the measured noise and pushing the experimental noise above the simulations' noise.

2.2.2 Gene copy number variation

As cells grow and divide, their chromosomes are replicated, causing the copy number of a given gene to change over the course of the cell cycle. This effect can potentially obscure our measurements of transcriptional noise. To that end, we wish to calculate the effect of changes in gene copy number on variability in gene expression. Under the growth conditions of our experiments, *E. coli* cells contain 1 or 2 chromosomes, and hence 1 or 2 copies of the gene of interest. Let $1 - f$ denote the fraction of the cell cycle for which 1 copy of the gene of interest is present. Then f is the fraction for which 2 gene copies are present. The probability that m mRNA are present in a cell is given by

$$p(m) = (1 - f)\, p_1(m) + f\, p_2(m) \tag{S24}$$

where $p_1(m)$ is the probability of m mRNA given 1 gene copy, and $p_2(m)$ is the probability of m mRNA given 2 gene copies. We will assume that, when two gene copies are present, expression from the two copies is statistically independent. Thus, we can use well-known properties of sums of independent random variables to calculate properties of $p_2(m)$.

We will proceed by computing the mean and variance of $p(m)$ given in equation S24.

$$\langle m \rangle = \sum_{m=0}^{\infty} m p(m) = (1 - f) \sum_{m=0}^{\infty} m p_1(m) + f \sum_{m=0}^{\infty} m p_2(m) \tag{S25}$$

$$= (1 - f)\langle m \rangle_1 + f \langle m \rangle_2 \tag{S26}$$

It can easily be shown that $\langle m \rangle_2 = 2\langle m \rangle_1$ and hence

$$\langle m \rangle = (1 + f)\langle m \rangle_1. \tag{S27}$$

Similarly for $\langle m^2 \rangle$, we have:

$$\langle m^2 \rangle = (1 - f)\langle m^2 \rangle_1 + f \langle m^2 \rangle_2. \tag{S28}$$

It can be shown that $\langle m^2 \rangle_2 = 2\langle m^2 \rangle_1 + 2\langle m \rangle_1^2$ (this follows from the fact that the variance of a sum of independent random variables is equal to the sum of the variances). Thus we obtain

$$\langle m^2 \rangle = (1 - f)\langle m^2 \rangle_1 + f \left[2\langle m^2 \rangle_1 + 2\langle m \rangle_1^2 \right]. \tag{S29}$$

Putting these expressions together, we find that

$$\begin{aligned} \mathrm{var}(m) &= \langle m^2 \rangle - \langle m \rangle^2 && \text{(S30)} \\ &= (1+f)\langle m^2 \rangle_1 + 2f\langle m \rangle_1^2 - (1+f)^2 \langle m \rangle_1^2 && \text{(S31)} \\ &= (1+f)\langle m^2 \rangle_1 - (1+f^2)\langle m \rangle_1^2 && \text{(S32)} \end{aligned}$$

The Fano factor is then

$$\begin{aligned} F &= \mathrm{var}(m)/\langle m \rangle \\ &= \frac{(1+f)\langle m^2 \rangle_1 - (1+f^2)\langle m \rangle_1^2}{(1+f)\langle m \rangle_1} \\ &= \frac{\langle m^2 \rangle_1}{\langle m \rangle_1} - \frac{(1+f)\langle m \rangle_1^2 - f(1-f)\langle m \rangle_1^2}{(1+f)\langle m \rangle_1} \\ F &= \underbrace{\frac{\langle m^2 \rangle_1 - \langle m \rangle_1^2}{\langle m \rangle_1}}_{\text{Transcription}} + \underbrace{\frac{f(1-f)}{1+f}\langle m \rangle_1}_{\text{Gene copy number}}, && \text{(S33)} \end{aligned}$$

reproducing equation 1 from the main text. The two terms of this expression each have straightforward interpretations. The first term is simply the (architecture-dependent) Fano factor of a single copy of a gene. The second term is the contribution from copy number variation. We can make two observations. First, the contributions to overall noise from promoter architecture and from gene copy number change are independent and additive. This is unsurprising since the two processes are (by assumption) independent and uncorrelated. Second, the contribution due to gene copy number increases linearly with expression. The predicted contribution to the Fano factor from copy number variation, the second term in Eq. S33, is shown in fig. S6 as a function of the average gene copy number ($= 2 - f$). As expected, if the copy number has a defined, static value ($f = 0$ or $f = 1$) there is no contribution to the Fano factor from variation in copy number. However, between these two minima, the variance reaches a maximum at $f = 0.5$, when the cell spends equal time with 1 or 2 copies, and thus the contribution to Fano factor has a maximum shifted towards slightly lower means (which appears in the denominator of Fano factor). For our experiments, our growth rates and locus of integration gives $f = 2/3$ [26], meaning that 2/3 of our cells have two copies of the reporter gene and the rest have one copy. In a section to follow and in fig. S11, we show that, as predicted, if the cells are binned into a population expected (based on physical size) to have only one or only two copies of the measured gene, the Fano factor deceases by roughly the same magnitude as that expected from copy number variations.

2.2.3 Extrinsic noise due to repressor copy number fluctuations

In addition to the "intrinsic" variability characterized by our modeling efforts, extrinsic sources of variability including (*e.g.*) changes in TF copy number can also contribute to overall variability in gene expression. To investigate the potential effects of fluctuations in repressor copy number, we performed numerical studies in which the repressor copy number was allowed to vary across a population of cells. Let $P_{\mathrm{arc}}(m|R = k)$ denote the promoter architecture dependent probability that a cell contains m mRNA given k copies of a repressor TF. Let $P_{TF}(R = k)$ denote the probability that a cell contains k repressor copies. (Our analysis of "intrinsic" cell-to-cell variability implicitly assumes that all cells have the same repressor copy number - that is, $P_{TF}(R = k) = \delta_{kk'}$ for some k'.) Then the overall probability of observing m mRNA in a cell is given by

$$P(m) = \sum_{k=0}^{\infty} P_{\mathrm{arc}}(m|R = k) \cdot P_{TF}(R = k). \tag{S34}$$

The quantity $P_{\text{arc}}(m|R = k)$ can be computed numerically as described in [2], and thus we can compute $P(m)$ numerically for any repressor copy number distribution $P_{TF}(R = k)$.

Here, we analyzed a population of cells in which the repressor copy numbers of individual cells are distributed according to a negative binomial distribution. The negative binomial distribution

$$P_{TF}(k; n, p) = \binom{n+k-1}{n-1} p^n (1-p)^k \tag{S35}$$

gives the probability that the nth success occurs on the $(k + n)$th trial, where the probability of success on any single trial is p. It is often used to model a more dispersed or long-tailed distribution than the Poisson distribution, and has been shown to correspond to constitutive mRNA production with a geometrically distributed number of proteins translated from each mRNA [1]. The degree of dispersal can be tuned via the parameter n as shown in fig. S7A. For a range of different n values, we tuned mean repressor expression via the parameter p while holding n constant. The resulting Fano vs mean curves for the target gene are shown in fig. S7B. We observe that, despite substantial variability in repressor copy number, the overall variability is predominantly contributed from intrinsic sources. This conclusion is robust across both relatively narrow and relatively dispersed repressor copy number distributions.

2.2.4 Extrinsic noise due to RNAP copy number fluctuations

In addition to repressor copy number fluctuations, RNAP copy number fluctuations also have the potential to contribute to the overall observed variability. We will follow an approach similar to the one outlined in the previous section. The distribution of RNAP copy numbers will be estimated from sources in the literature. The average

RNAP copy number is reported as $\approx 10,000$ per cell [27]. In [3], the authors report that for a typical protein with 10,000 copies per cell, the standard deviation in protein copy number is approximately 3200. We will thus model the RNAP copy number distribution as a negative binomial distribution with mean equal to 10,000 and standard deviation equal to 3200, show in fig. S8A. We assume that the transcription rate r is proportional to the RNAP copy number in the cell. The resulting Fano vs mean curves are plotted in fig. S8B for both the constitutive expression and simple repression architectures. In both cases, we see that extrinsic variability due to RNAP fluctuations increases with increasing mean expression. In the case of constitutive expression, this increasing trend in the Fano vs mean curve is markedly similar to the increasing trend we observe in our constitutive expression data. In the case of simple repression, the addition of extrinsic variability does not change the overall qualitative features of the predicted curve. However, it does lead to the prediction that the overall observed Fano factor will not fall all the way back down to 1 in the absence of repressor, which is consistent with our experimental observations.

In the case of constitutive expression, it is possible to derive an informative analytical expression for the extrinsic noise contributed by variation in r. While this is a general approach to variations in r, later we will relate this to the specific circumstance of RNAP fluctuations expected in our constitutive expression measurements.

Let the probability distribution for values of r be denoted by $P_{\text{ext}}(r)$. The steady-state probability distribution for mRNA copy number, m, given a particular value of r is a Poisson distribution with mean r/γ, such that

$$P_{\text{arc}}(m|r) = \frac{(r/\gamma)^m}{m!} e^{-r/\gamma}. \tag{S36}$$

We assume that r changes on a timescale sufficiently long compared to $1/\gamma$ that we can use the steady-state probability distribution. Then the overall probability distribution for mRNA copy number, integrated over all possible values of r, is given by

$$P(m) = \int P_{\text{arc}}(m|r)P_{\text{ext}}(r)dr \tag{S37}$$

To compute the Fano factor for this overall distribution, we will as usual proceed by computing $\langle m \rangle$ and $\langle m^2 \rangle$.

$$\langle m \rangle = \sum_{m=0}^{\infty} m \int P_{\text{arc}}(m|r)P_{\text{ext}}(r)dr \tag{S38}$$

$$\langle m \rangle = \int P_{\text{ext}}(r) \sum_{m=0}^{\infty} m P_{\text{arc}}(m|r)dr \tag{S39}$$

$$\langle m \rangle = \int P_{\text{ext}}(r)\frac{r}{\gamma}dr \tag{S40}$$

$$\langle m \rangle = \frac{\langle r \rangle}{\gamma} \tag{S41}$$

where $\langle r \rangle$ is the mean value of r. Similarly, for $\langle m^2 \rangle$,

$$\langle m^2 \rangle = \sum_{m=0}^{\infty} m^2 \int P_{\text{arc}}(m|r)P_{\text{ext}}(r)dr \tag{S42}$$

$$\langle m^2 \rangle = \int P_{\text{ext}}(r) \sum_{m=0}^{\infty} m^2 P_{\text{arc}}(m|r)dr \tag{S43}$$

$$\langle m^2 \rangle = \int P_{\text{ext}}(r)\left(\frac{r^2}{\gamma^2} + \frac{r}{\gamma}\right)dr \tag{S44}$$

$$\langle m^2 \rangle = \frac{\langle r \rangle}{\gamma} + \frac{\langle r^2 \rangle}{\gamma^2}. \tag{S45}$$

Hence, the Fano factor is given by

$$\text{Fano} = \frac{\langle m^2 \rangle - \langle m \rangle^2}{\langle m \rangle} \tag{S46}$$

$$\text{Fano} = \frac{\frac{\langle r \rangle}{\gamma} + \frac{\langle r^2 \rangle}{\gamma^2} - \frac{\langle r \rangle^2}{\gamma^2}}{\frac{\langle r \rangle}{\gamma}} \tag{S47}$$

$$\text{Fano} = 1 + \frac{1}{\gamma} \frac{\langle r^2 \rangle - \langle r \rangle^2}{\langle r \rangle} \tag{S48}$$

$$\text{Fano} = 1 + \frac{1}{\gamma} \, \text{Fano}(r). \tag{S49}$$

Thus far, we have assumed nothing about the specific mechanism causing fluctuations in r. Let's now explore the case in which fluctuations in r are caused by fluctuations in RNAP copy number. We will assume that the transcription rate r is proportional to the RNAP copy number, so that $r = r_0 E$ where E is the RNAP polymerase copy number and r_0 is a constant of proportionality that can be thought of as roughly the transcription rate per RNAP molecule. The constant of proportionality r_0 is assumed to depend on the strength of the promoter, so that when we tune mean expression by tuning promoter strength, it is really (by assumption) the parameter r_0 that we are changing. Under this assumption, eq. S49 becomes:

$$\text{Fano} = 1 + \frac{1}{\gamma} \, \text{Fano}(r_0 E) \tag{S50}$$

$$\text{Fano} = 1 + \frac{r_0}{\gamma} \, \text{Fano}(E) \tag{S51}$$

$$\text{Fano} = 1 + \frac{\langle m \rangle}{\langle E \rangle} \, \text{Fano}(E) \tag{S52}$$

$$\text{Fano} = 1 + \langle m \rangle \frac{\sigma_E^2}{\langle E \rangle^2} \tag{S53}$$

$$\text{Fano} = 1 + \frac{1}{10} \langle m \rangle \tag{S54}$$

where we have used the fact that by assumption $\langle m\rangle = \langle r\rangle/\gamma = r_0\langle E\rangle/\gamma$, and used the result of [3] that the squared coefficient of variation $\sigma_E^2/\langle E\rangle^2 \approx 10^{-1}$ for a protein with $\approx 10^4$ copies per cell. Eq. S54 tells us simply that fluctuations in RNAP copy number contribute an additional term to the Fano factor that increases linearly with mean gene expression, with a slope equal to the squared coefficient of variation of the RNAP copy number.

However, it is worth noting that RNAP copy number fluctuations are by no means the only extrinsic mechanism capable of generating this linear relationship between Fano factor and mean expression. For instance, one could imagine that DNA supercoiling renders the promoter inaccessible and thus silences transcription some fraction s of the time. We could model this scenario by saying that the effective RNAP copy number is 0 for a fraction s of the time, and E_0 for the remainder (*i.e.*, $1 - s$) of the time, where E_0 is some number of order 10^4. As before, we assume that the transcription rate is proportional to the RNAP copy number. We can thus proceed from eq. S51. It can easily be shown that $\text{Fano}(E) = sE_0$ in this case, and hence eq. S51 becomes

$$\text{Fano} = 1 + \frac{r_0}{\gamma} s E_0. \tag{S55}$$

If we use the fact that $\langle m\rangle = r_0\langle E\rangle/\gamma = r_0(1-s)E_0/\gamma$, we obtain finally

$$\text{Fano} = 1 + \frac{s}{1-s}\langle m\rangle. \tag{S56}$$

So again, we have an extrinsic noise term that increases linearly with mean expression, here with a slope that depends on the fraction of time s for which expression is silenced. The implications of this result will be discussed in the following section.

2.2.5 Extrinsic sources of noise: concluding remarks

To conclude this exploration of sources of extrinsic noise, we will offer a few observations concerning model selection and the interpretation of experimental evidence. In a 2011 paper, Huh and Paulsson pointed out that protein partitioning at cell division yields the same $1/\langle x \rangle$ overall scaling in the cell-to-cell variability in protein levels as does Poissonian transcription [28]. Thus, experimental observation of $1/\langle x \rangle$ noise scaling does not in itself provide a basis for distinguishing between these two mechanisms. Although this specific point is not relevant in the case of our experiments, since mRNA lifetimes are sufficiently short compared with division times that partitioning effects are negligible, the overall spirit of Huh and Paulsson's argument is relevant.

In particular, we found that the effect of gene copy number variation on the Fano factor increases linearly with mean gene expression. However, in the case of constitutive expression, we also find that the effect of RNAP fluctuations on the Fano factor increases linearly with mean expression. Furthermore, the same would be true (in the case of constitutive expression) if one postulated that a mechanism such as DNA supercoiling causes transcriptional silencing *e.g.* 25% of the time: one would find a linear relationship between Fano factor and mean expression. So how can we have any confidence in the breakdown of noise sources in Fig. 2B? In our view, this discussion highlights the importance of independent corroboration of each of the pieces shown in Fig. 2B. The quantization error is corroborated through both theoretical calculation and analysis of simulated data (above). The gene copy number variation effect is corroborated below by using cell size as a proxy for gene copy number (older, larger cells will have two chromosome copies, while younger, smaller cells will have one). The RNAP fluctuation effect is the most speculative. Although we believe it is defensible both in terms of the underlying assumption that expression is proportional to RNAP copy number [29], and in the magnitude of RNAP fluctuations taken from literature sources [3,30], our data does not provide us with a means to independently corroborate this effect. (To do so, one might perform an experiment in which fluorescently tagged RNAP molecules are used to quantify RNAP fluctuations.) Thus, it is possible that the RNAP fluctuation effect is instead something else entirely such as transcriptional silencing by DNA supercoiling. This is the reasoning behind our statement in the Discussion of the main text that "our data does not rule out these [alternative] hypotheses."

2.3 mRNA histograms for constitutive expression

Representative mRNA copy number histograms are shown in Fig. S9 for each strain, with the strain names above each histogram for reference. For each strain, we plot the predicted mRNA copy number distributions both with (black lines) and without (blue dashed lines) accounting for gene copy number variation. Specifically, the blue curves are given by a Poisson distribution parameterized by the observed mean $\langle m\rangle$ of each strain:

$$P(m) = (\langle m\rangle^m/\langle m\rangle!)\, e^{-\langle m\rangle}. \tag{S57}$$

The black curves are given by combining the preceding equation with Eq. S24, namely,

$$P(m) = (1-f)(\langle m\rangle_1^m/\langle m\rangle_1!)\, e^{-\langle m\rangle_1} + (f)((2\langle m\rangle_1)^m/(2\langle m\rangle_1)!)\, e^{-2\langle m\rangle_1} \tag{S58}$$

where $\langle m\rangle_1 = \langle m\rangle/(1+f)$, $f = 2/3$, and $\langle m\rangle$ is the observed mean for each strain. It can readily be seen that accounting for gene copy number variation improves the agreement between theory and experiment without requiring additional free parameters. To quantify this improvement, we report the log-likelihood ratio (LLR) for each strain. This quantity is the logarithm of the ratio of the likelihood of the data given the variable copy number probability distribution (black circles, Eq. S58) to the likelihood of the data given by the simple Poisson prediction (blue dashed line, Eq. S57). LLR = 0 implies that the observed data is equally likely given either theoretical distribution, whereas LLR > 0 implies that the data is more likely to have been observed given the variable gene copy distribution of Eq. S58. We obtain positive LLR values for every strain, with the most positive values tending to occur at high mRNA expression, where the difference between Eq. S57 and Eq. S58 is most pronounced.

2.4 Error bars in Fano factor measurements

In the main text Figs. 2 and 3 as well as SI figs. S10 and S11B contain experimental measurements for the Fano factor. Typically there are at least three repeats of any given condition and each individual data point represents an individual experiment; all available data points are plotted on every figure. Error bars are determined by bootstrapping the single cell copy number distribution 1000 times and calculating the standard deviation in the Fano factor for those 1000 independent bootstrapped data sets. In other words, the single cell mRNA copy number distribution, which typically contains roughly 900 entries of the number of mRNA in a given cell, is randomly resampled with replacement (the same cell may appear multiple times in a given bootstrapped data set) to create a new data set with the equivalent number of entries. This is repeated 1000 times and the standard deviation of the Fano factor in these measurements is used as an error bar for the measurement.

2.5 Copy Number Variation: Uncorrected Figures

In the main text, our focus was on the promoter architecture dependent component of gene expression variability, and thus we subtracted the gene copy number dependent term (as defined by the second term in equation S33) from the measured Fano factor in Figs. 2 and 3 of the main text. However, doing so does not change the qualitative conclusion that variability is promoter architecture dependent. In fig. S10, we plot the data from Figs. 2 and 3 without subtracting the gene copy number dependent term.

2.6 Testing gene copy number noise by cell size segregation

In the main text we claim that one significant contribution to the measured cell-to-cell mRNA copy number variability stems from the fact that our measurements contain a mix of cells with one or two copies of the gene of interest. To test this claim, we take the data from each measurement of our constitutive strains and divide the data into two subsets based on their physical size. The idea is to use our knowledge of the cell cycle, based on growth rate and gene position in the chromosome, to divide each data set into one set with cells likely to have a single copy of the target gene (referred to as "small cells") and cells likely to have two copies of the target gene (referred to as "large cells"). As mentioned in the main text, at 60 minute growth rate we expect that the galK locus has a copy number of 1.66 [26], which implies that we should set our division line at roughly 1/3 of the way through the cell cycle. We determine this point by plotting the cumulative probability distribution of the cell area of every cell in every sample and identifying the area value where 1/3 of the cells are smaller and 2/3 are larger. For our cells this is at an area of roughly 3.75 μm^2. To help ensure that this division is "clean" we discard cells 1/8 above and below the division line so that our small cells contain the smallest 21% of cells and the large cells are the 54% largest cells.

In fig. S11A, we show the result of plotting the mean mRNA copy number of the small cells bin versus the mean mRNA copy number of the large cell bin for every measurement of every constitutive strain (black points). As stated previously while deriving gene copy number noise, we expect that the large cells, binned to have two copies of the reporter gene, should have twice the transcriptional activity of the small cells. This is precisely what is observed: the red line is a line of slope 2 and intercept 0. While we do not expect this method to achieve a perfect division of the total population, this test indicates that these subsets of data contain primarily cells with one and two copies of the reporter gene.

In fig. S11B, we show the Fano factor of each of these two data subsets (black squares for small cells, black circles for large cells) as well as the Fano factor of the full data sets as red diamonds. Once again, the relevant sources of intrinsic and quantization noise are shaded in green (quantization error), red (RNAP fluctuation error) and blue (gene copy number noise). First, when the mean is small (<1 per cell per gene copy) the expected contribution from copy number variations is small (the blue shaded region is small) and thus the two subsets and the full data set give similar results for the Fano factor as expected. Above this threshold we begin to see that the Fano factor of either subset of data (both large cells and small cells) falls below the corresponding Fano factor of the full data set. Furthermore, we see that the reduction in Fano factor causes the subsets to fall approximately on the interface between the shaded blue and red region; the subset data is now consistent with a Poisson process with quantization error and RNAP fluctuations but without gene copy number fluctuations.

2.7 Determination of Rate Parameter Values

The values of $k_{\rm off}^{\rm R}$ used in this work are taken from [31] and [2], and are shown in table S2. Specifically, ref. [31] used a single molecule *in vitro* assay to measure the dissociation rate from the LacI Oid operator. Ref. [2] used this rate, along with knowledge of the dissociation constants of the Oid, O1, O2, and O3 operators (reported in [13]), to estimate the dissociation rates for the three additional operators, using the assumption that the ratio of the dissociation rates for two particular operators is equal to the ratio of their dissociation constants. In order to determine the three different values of $k_{\rm on}^{\rm R}$ used in Fig. 3B of the main text, slightly more work was required. Recall that we are assuming a diffusion-limited on rate for which $k_{\rm on}^{\rm R} = k_0[R]$. Ref. [32] reports that $k_0 = 2.7 \times 10^{-3}(\text{s nM})^{-1}$. To determine $k_{\rm on}^{\rm R}$ for each of the three aTc concentrations, we must determine the repressor concentration $[R]$ at each aTc concentration. Unfortunately we do not independently possess the exact input-output relation between aTc concentration and repressor copy number, but we can estimate the repressor concentration at each aTc concentration by looking at how strongly gene expression is repressed at each aTc concentration.

More specifically, in a recent work [25], the authors defined the "repression" as the ratio between gene expression in the absence of repressor transcription factors (TFs) and gene expression in the presence of repressor TFs. They showed that, for the type of "simple repression" promoter architecture used in this paper, the repression is given by

$$\text{Repression} = 1 + \frac{2R}{N_{\rm NS}}\, e^{-\Delta\epsilon_{\rm rd}/k_{\rm B}T} \tag{S59}$$

where R is the repressor copy number, $N_{\rm NS}$ is the number of non-specific binding sites (taken to be equal to the size of the *E. coli* genome, or $2 \times 5 \times 10^6$), and $\Delta\epsilon_{rd}$ is the repressor-DNA binding energy ($-17.3\, k_B T$) for the Oid LacI binding site. This expression can be solved to determine the repressor copy number R as a function of the repression

$$R = (\text{Repression} - 1) \times \frac{N_{NS}}{2}\, e^{\Delta\epsilon_{rd}/k_BT}. \tag{S60}$$

Garcia and Phillips [25] used this equation to determine the effective repressor copy number R, and verified their results using quantitative Western blot analysis. We used a similar approach by computing the repression at each of the aTc concentrations for the Oid operator construct, then using equation S60 coupled with the fact that the volume of an *E. coli* cell is approximately 1 fL to determine the repressor concentration at each aTc concentration. Finally, we multiplied these concentrations by k_0 to determine the appropriate value of k_{on}^{R}. The results are summarized in table S3.

3 Supplementary Figures

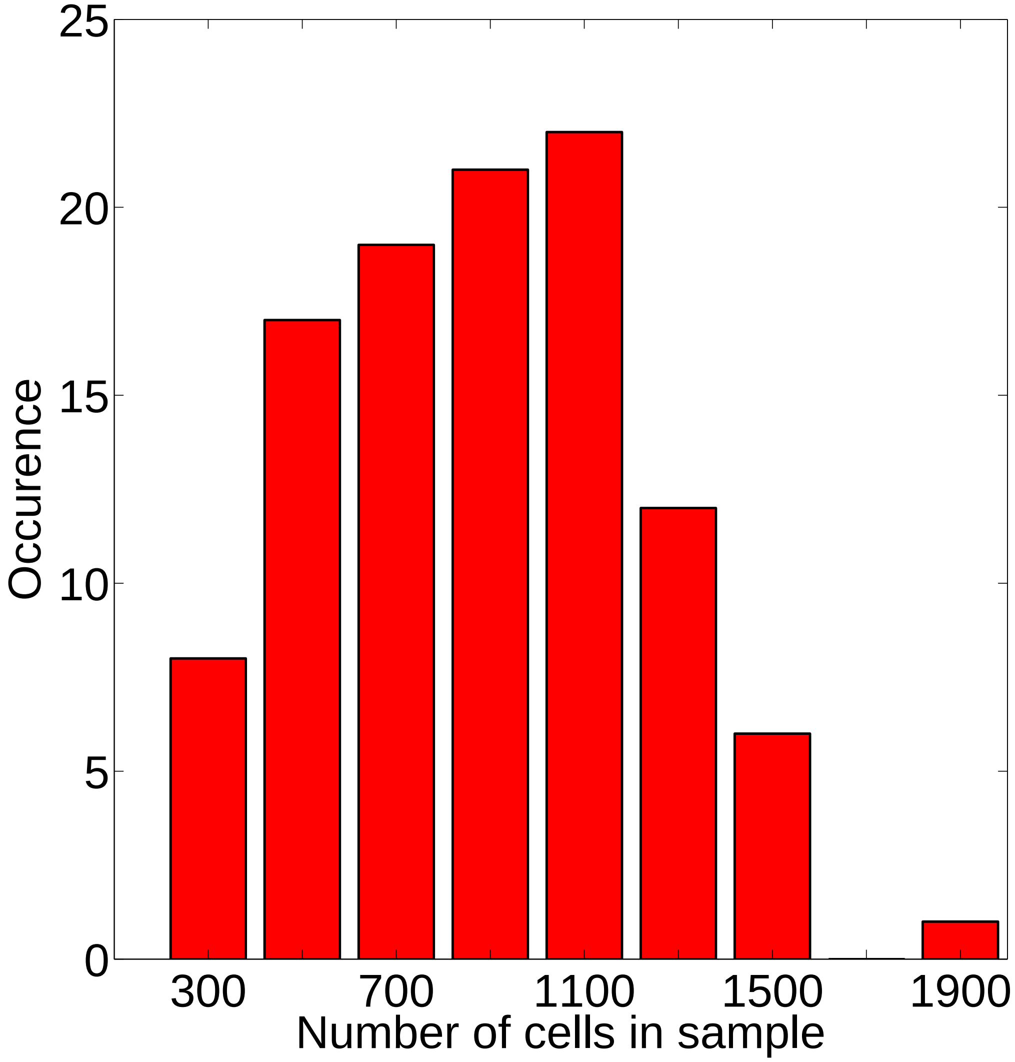

Figure S1: **Histogram of the number of cells per FISH sample.**
Each sample has 100 unique positions imaged. Due to differences in cell density and position quality (positions are chosen in an automated process), samples range in size and have roughly 900 cells on average per sample.

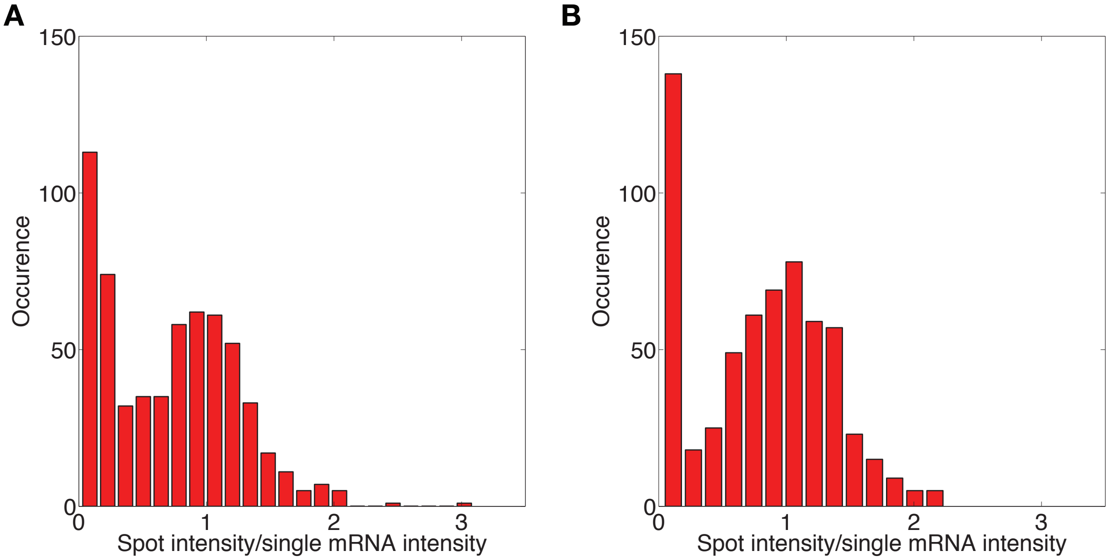

Figure S2: **Histograms of detected spot intensities for low expression FISH data.**

Detected spots (local maxima in fluorescence signal), in principle, correspond to either 0 or 1 mRNA. Part (**A**) shows a representative histogram from FISH experiments, while part (**B**) shows a histogram from simulated data. The signal of each spot has been normalized by the "single mRNA intensity". For both histograms, we identify a "noise" or background peak at fluorescence intensity ≈ 0. This peak corresponds to unbound or nonspecifically bound probes. In both cases, although we can discern distinct peaks, distinguishing between 0 and 1 mRNA is not completely unambiguous.

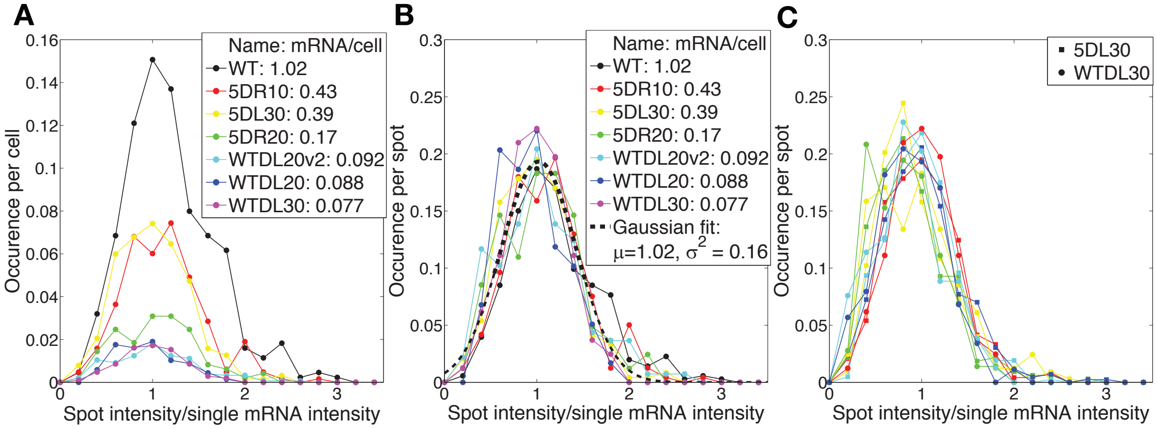

Figure S3: **Comparison of identified single spot intensities across different samples and experimental acquisitions.**

(**A**) Histograms of single spot intensities for seven samples, normalized by the number of cells in each sample. The value on the y-axis corresponds to the probability of finding a spot with a given intensity in any one cell in the sample. Mean expression in the samples ranges between 0.077 mRNA per cell and 1.02 mRNA per cell. When expression is low, increases in the mean expression level increase the probability of finding a spot with intensity equal to a single mRNA, rather than, for instance, increasing the intensity of identified spots. (**B**) Histograms of single spot intensity values for the same seven samples, normalized by the total number of identified spots in each sample. In this case, the value on the y-axis corresponds to the probability that a given identified spot has a particular intensity. The spots have roughly the same properties in each of these samples, although in the highest expression samples, we begin to see increased probability to have spots with intensity corresponding to more than 1 mRNA. The day-to-day reproducibility in this identification process is shown in part (**C**) where two different strains (5DL30 and WTDL30) are shown measured across five different acquisitions.

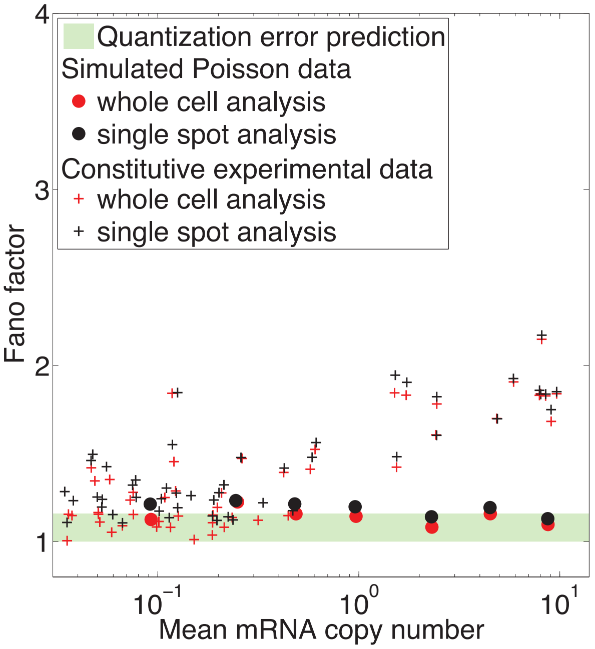

Figure S4: **Fano factor vs mean plot for simulated Poisson distributed data.** Simulated mRNA FISH data with Poisson distributed mRNA copy numbers (circles) is analyzed over a range of mean mRNA levels to evaluate our analysis code. Since the simulated data is Poisson distributed, the true value of the Fano factor is 1. However, we see here that the measured Fano factor is always slightly greater than one. This persistent noise represents the “quantization error” discussed in the text; the expected contribution to the Fano factor from quantization error is indicated by the height of the green bar. For comparison, the crosses are the Fano factors (corrected for gene copy number noise) from our constitutive expression strains (data from Fig. 2B). The different colors (black and red) represent two distinct methods for quantifying the resulting mRNA signal. For the black symbols, individual mRNA spots are identified and quantified (divided by the intensity of a single mRNA) and rounded to the nearest whole number of mRNA and the copy number in a cell is the sum of the number of mRNA in each identified spot for a given cell. The red symbols correspond to summing the signal of all identified spots in a cell and determining a cell’s copy number by dividing the summed signal by the single mRNA intensity (and, in this case, not rounded). This second method (red symbols) is used in this work, but this choice does not significantly influence the outcome.

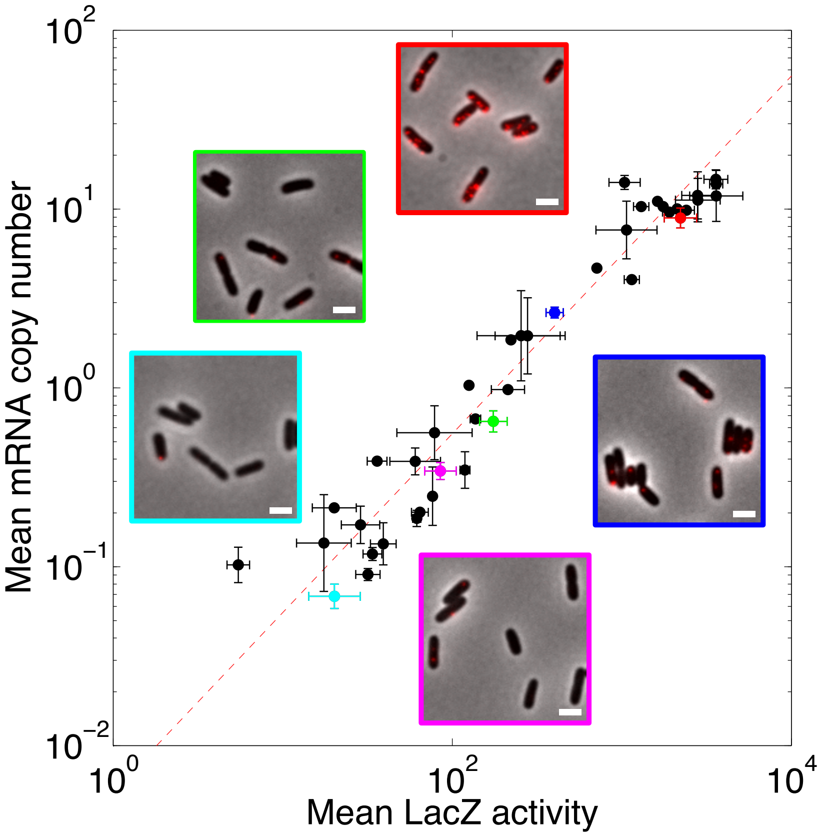

Figure S5: **Experimental comparison of mean mRNA FISH measurements to enzymatic assay.**
Direct comparison of the average mRNA copy number to the average enzymatic activity of the encoded protein for every data strain and condition used in the text. The red line is a linear fit to the data. Error bars are standard deviation from multiple measurements.

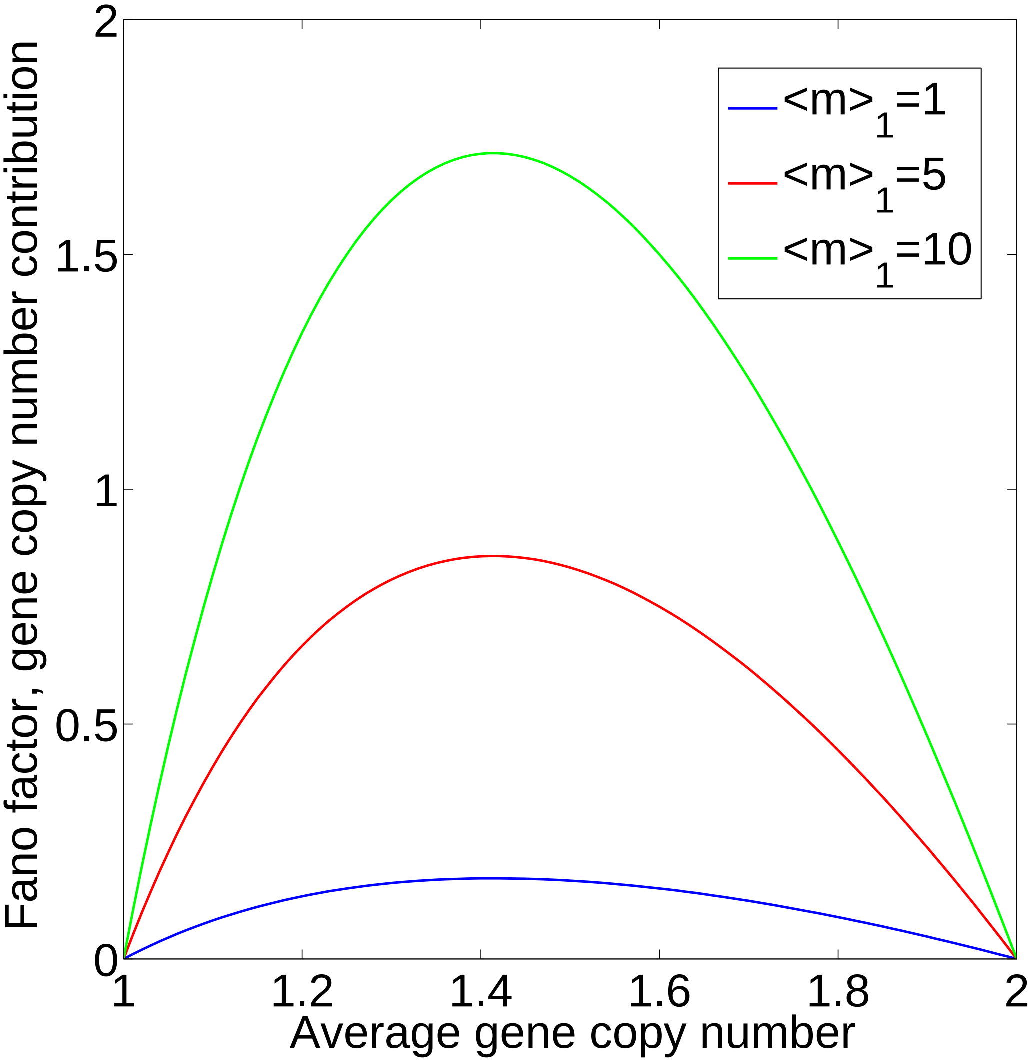

Figure S6: **Fano factor contribution from gene copy number variation.** Predicted contribution to the Fano factor from gene copy number variation for three distinct mean expression levels, 1 (blue curve), 5 (red curve) and 10 (green curve) mRNA copies per cell per gene copy. The effect increases with transcription rate and is largest when the gene spends approximately half the cell cycle with 1 copy and the other half with 2 copies.

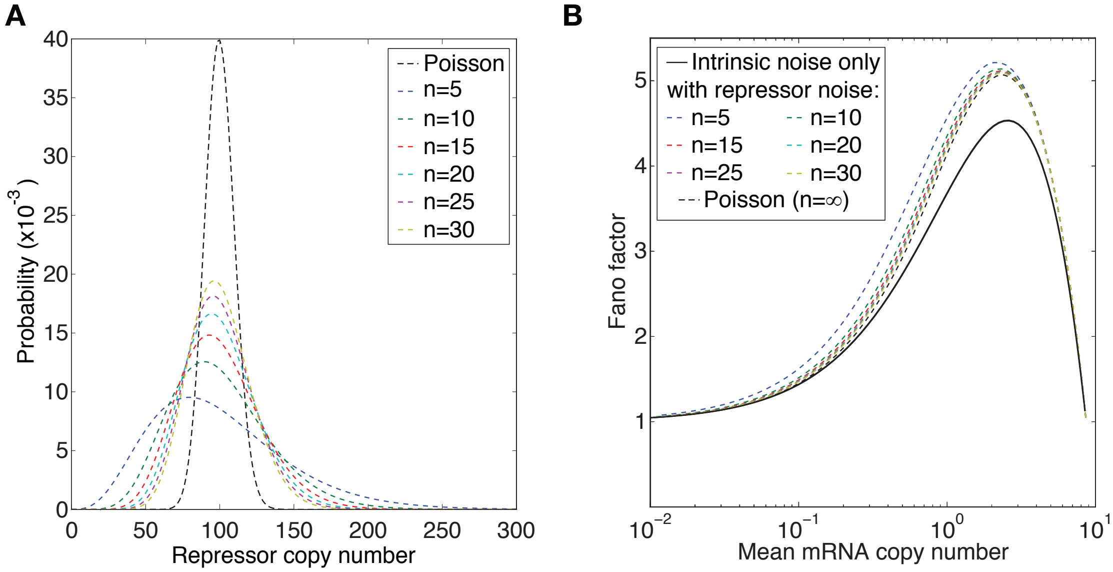

Figure S7: **Quantifying the extrinsic noise contribution of repressor copy number variation.**

(**A**) Single-cell repressor distribution for the negative binomial distribution with various choices for the parameter n and for a Poisson distribution. (**B**) Predicted Fano factor for simple repression with a static value for the repressor copy number without distribution (solid black line) along with the Fano factor for the distributions shown in (A) of this figure. Even when the distribution is quite wide, the added noise above the intrinsic piece is relatively small.

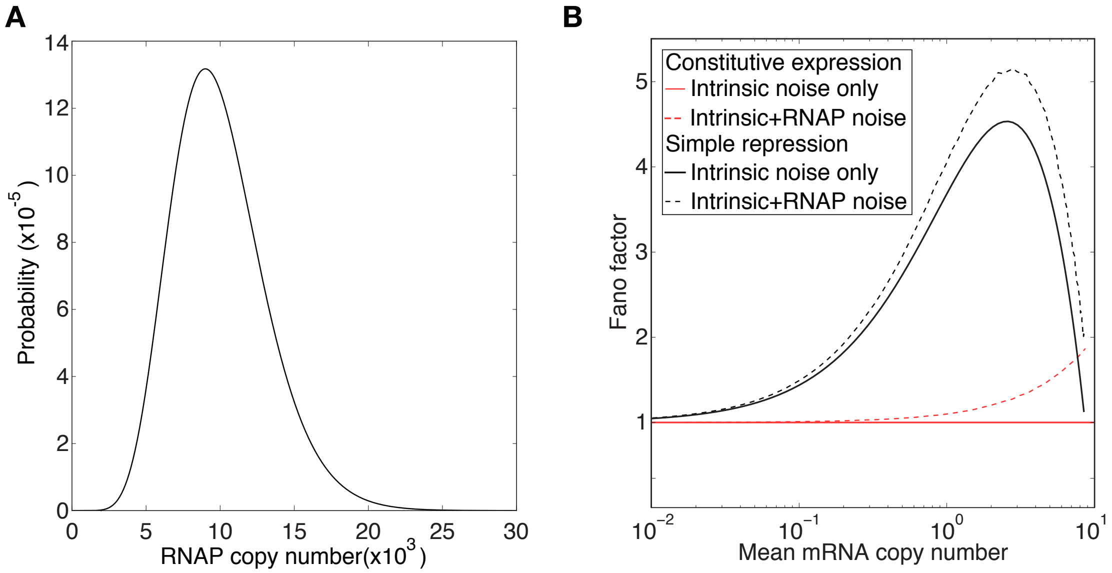

Figure S8: **Quantifying the extrinsic noise contribution of RNAP copy number variation.**
(**A**) Negative binomial model of RNAP copy number distribution with width chosen to coincide with reported literature values [3]. (**B**) The resulting contribution to the Fano factor from extrinsic noise in RNAP copy number. The solid lines are the theoretical predictions without any source of extrinsic noise for constitutive (red solid line) and simple repression (black solid line) and with RNAP fluctuation noise (corresponding dashed lines).

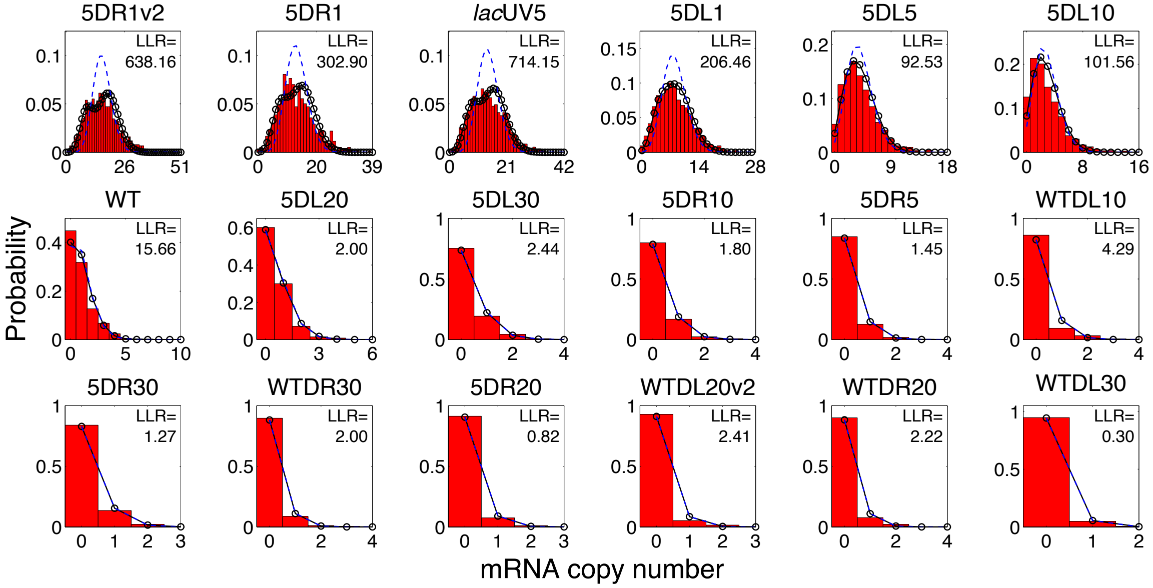

Figure S9: **mRNA copy number histograms for constitutive expression.** Observed mRNA copy number distributions for library of 18 constitutive promoters. For each promoter, we plot the predicted mRNA copy number distribution assuming Poissonian production and degradation, both with (black circles) and without (dashed blue lines) accounting for gene copy number variation. The log-likelihood ratio (LLR) of the observed data with and without accounting for copy number variation is shown on each histogram. Accounting for gene copy number variation substantially improves agreement between theory and data, as indicated by positive LLRs. The mean number of cells included in a histogram is 1238 ± 267 cells for each sample.

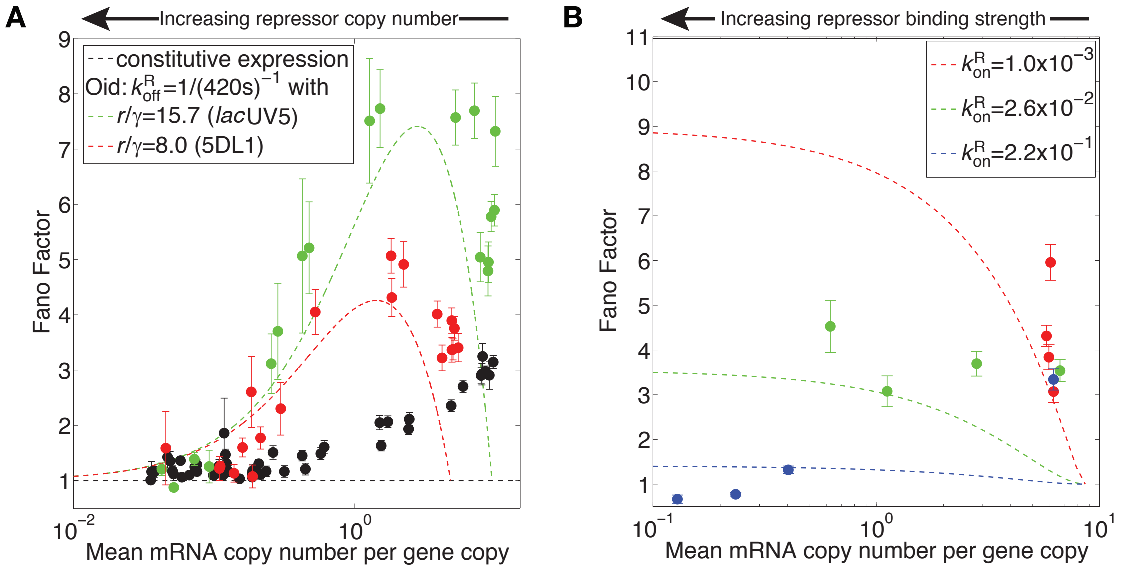

Figure S10: **Fano factor vs. mean mRNA copy number.** The data from Fig. 3 of the main text are plotted without subtracting the effect of gene copy number variation. (**A**) Fano factor vs. mean mRNA copy number for two promoters (choices of r/γ): 5DL1 (red points) and *lac*UV5 (green points) while tuning k_{on}^{R} by inducing LacI to varying levels. The parameter-free predictions from the kinetic theory of transcription are shown as dashed lines in the corresponding color holding promoter (r/γ) and repressor binding strength ($k_{\text{off}}^{\text{R}}$) constant. For reference, the black data is the constitutive data from figure 2. (**B**) Fano factor vs. mean mRNA copy number for *lac*UV5 while tuning $k_{\text{off}}^{\text{R}}$ by changing repressor binding site identity at fixed repressor copy number, each color is a different induction condition from red (lowest LacI induction) to blue (highest LacI induction). Again, the predictions from the kinetic theory of transcription are shown as dashed lines in the corresponding color. For both panels, not subtracting gene copy number variation slightly worsens the fit between theory and data, but the overall conclusion that variability is promoter architecture dependent is not affected. Error bars are the result of bootstrap sampling of the expression measurements in each sample

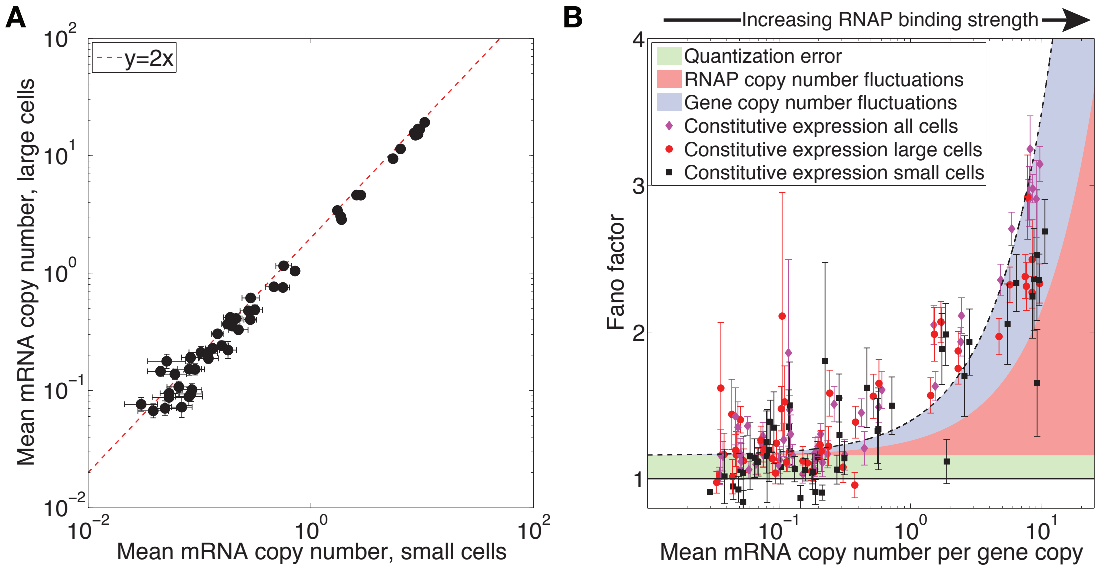

Figure S11: **Analysis of small cell and large cell data subsets in constitutive expression.**

Each individual constitutive expression sample measurement (multiple measurements of all 18 strains) is divided into two subsets of "large" and "small" cells based on cell area. The division line between these sets is chosen such that small cells are expected to have one copy of the reporter gene while large cells are expected to contain two copies. (**A**) Mean mRNA copy number of large cells vs mean mRNA copy number of small cells within the same sample. The mean copy number of the large cells is double the mean copy number of the small cells, supporting the assertion that the data sets are correctly divided based on gene copy number. (**B**) Fano factor vs mean mRNA copy number for the full data sets (purple diamonds, as from Fig. 2B), large cells (red circles) and small cells (black squares). The Fano factor for the full data samples agrees with the noise prediction including quantization noise, RNAP fluctuation noise and gene copy number noise. The subsets, divided to remove gene copy number variation in a sample, are described best without the gene copy number noise term. All error bars are the result of bootstrap sampling of the expression measurements in each sample.

4 Supplementary Tables

Name	Sequence	Name	Sequence
lacZ1	gtgaatccgtaatcatggtc	lacZ37	gatcgacagatttgatccag
lacZ2	tcacgacgttgtaaaacgac	lacZ38	aaataatatcggtggccgtg
lacZ3	attaagttgggtaacgccag	lacZ39	tttgatggaccatttcggca
lacZ4	tattacgccagctggcgaaa	lacZ40	tattcgcaaaggatcagcgg
lacZ5	attcaggctgcgcaactgtt	lacZ41	aagactgttacccatcgcgt
lacZ6	aaaccaggcaaagcgccatt	lacZ42	tgccagtatttagcgaaacc
lacZ7	agtatcggcctcaggaagat	lacZ43	aaacggggatactgacgaaa
lacZ8	aaccgtgcatctgccagttt	lacZ44	taatcagcgactgatccacc
lacZ9	taggtcacgttggtgtagat	lacZ45	gggttgccgttttcatcata
lacZ10	aatgtgagcgagtaacaacc	lacZ46	tcggcgtatcgccaaaatca
lacZ11	gtagccagctttcatcaaca	lacZ47	ttcatacagaactggcgatc
lacZ12	aataattcgcgtctggcctt	lacZ48	tggtgttttgcttccgtcag
lacZ13	agatgaaacgccgagttaac	lacZ49	acggaactggaaaaactgct
lacZ14	aattcagacggcaaacgact	lacZ50	tattcgctggtcacttcgat
lacZ15	tttctccggcgcgtaaaaat	lacZ51	gttatcgctatgacggaaca
lacZ16	atcttccagataactgccgt	lacZ52	tttaccttgtggagcgacat
lacZ17	aacgagacgtcacggaaaat	lacZ53	gttcaggcagttcaatcaac
lacZ18	gctgatttgtgtagtcggtt	lacZ54	ttgcactacgcgtactgtga
lacZ19	ttaaagcgagtggcaacatg	lacZ55	agcgtcacactgaggttttc
lacZ20	aactgttacccgtaggtagt	lacZ56	atttcgctggtggtcagatg
lacZ21	ataatttcaccgccgaaagg	lacZ57	acccagctcgatgcaaaaat
lacZ22	tttcgacgttcagacgtagt	lacZ58	cggttaaattgccaacgctt
lacZ23	atagagattcgggatttcgg	lacZ59	ctgtgaaagaaagcctgact

lacZ24	ttctgcttcaatcagcgtgc	lacZ60	ggcgtcagcagttgtttttt
lacZ25	accattttcaatccgcacct	lacZ61	tacgccaatgtcgttatcca
lacZ26	ttaacgcctcgaatcagcaa	lacZ62	taaggttttcccctgatgct
lacZ27	atgcagaggatgatgctcgt	lacZ63	atcaatccggtaggttttcc
lacZ28	tctgctcatccatgacctga	lacZ64	gtaatcgccatttgaccact
lacZ29	ttcatcagcaggatatcctg	lacZ65	agttttcttgcggccctaat
lacZ30	cacggcgttaaagttgttct	lacZ66	atgtctgacaatggcagatc
lacZ31	tggttcggataatgcgaaca	lacZ67	ataattcaattcgcgcgtcc
lacZ32	ttcatccaccacatacaggc	lacZ68	tgatgttgaactggaagtcg
lacZ33	tgccgtgggtttcaatattg	lacZ69	tcagttgctgttgactgtag
lacZ34	atcggtcagacgattcattg	lacZ70	attcagccatgtgccttctt
lacZ35	tgatcacactcgggtgatta	lacZ71	aatccccatatggaaaccgt
lacZ36	atacagcgcgtcgtgattag	lacZ72	agaccaactggtaatggtag

Table S1: Names and sequences of LacZ mRNA probes.

Operator	$k_{\text{off}}^{\text{R}}(s^{-1})$
Oid	0.0023
O1	0.0069
O2	0.091
O3	2.1

Table S2: **Repressor dissociation rates**. These rates are taken directly from refs. [2] and [31]. The dissociation rate of the Oid operator was directly measured *in vitro*, while the O1, O2, and O3 dissociation rates were computed using the ratios of these binding sites' equilibrium occupancies to that of Oid.

aTc concentration	R (copy number)	[R] (nM)	$k_{\text{on}}^{\text{R}}(s^{-1})$
0.5 ng/mL	0.21	0.35	0.0010
2 ng/mL	5.9	9.8	0.026
10 ng/mL	50	83	0.22

Table S3: **Repressor binding rates**. As described in the supplementary text, these rates were computed by combining the association rate per repressor reported in ref. [32] with an estimate of the repressor copy number at each aTc concentration. The overall association rate is then the product of the estimated repressor copy number with the association rate per repressor molecule.

References

[1] J. Paulsson, M. Ehrenberg, *Phys. Rev. Lett.* **84**, 5447 (2000).

[2] A. Sanchez, H. G. Garcia, D. Jones, R. Phillips, J. Kondev, *PLoS Comput. Biol.* **7**, e1001100 (2011).

[3] Y. Taniguchi, *et al.*, *Science.* **329**, 533 (2010).

[4] W. J. Blake, *et al.*, *Mol. Cell* **24**, 853 (2006).

[5] L. H. So, *et al.*, *Nature Genet.* **43**, 554 (2011).

[6] H. Salman, *et al.*, *Phys. Rev. Lett.* **108**, 238105 (2012).

[7] H. Maamar, A. Raj, D. Dubnau, *Science.* **317**, 526 (2007).

[8] A. Eldar, M. B. Elowitz, *Nature.* **467**, 167 (2010).

[9] G. M. Süel, R. P. Kulkarni, J. Dworkin, J. Garcia-Ojalvo, M. B. Elowitz, *Science.* **315**, 1716 (2007).

[10] M. Thattai, A. van Oudenaarden, *Genet.* **167**, 523 (2004).

[11] E. Kussell, S. Leibler, *Science.* **309**, 2075 (2005).

[12] H. Salgado, *et al.*, *Nucleic Acids Res.* **41**, D203 (2013).

[13] L. Bintu, *et al.*, *Curr. Opin. Genet. Dev.* **15**, 125 (2005).

[14] T. Kuhlman, Z. Zhang, J. Saier, M. H., T. Hwa, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 6043 (2007).

[15] J. M. Vilar, S. Leibler, *J. Mol. Biol.* **331**, 981 (2003).

[16] Materials and methods are available as supplementary material on *Science* Online.

[17] A. Sanchez, J. Kondev, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 5081 (2008).

[18] P. Swain, M. Elowitz, E. Siggia, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12795. (2002).

[19] R. Brewster, D. Jones, R. Phillips, *PLoS Comput. Biol.* **8**, e1002811 (2012).

[20] P. Hammar, *et al.*, *Science.* **336**, 1595 (2012).

[21] I. Golding, J. Paulsson, S. M. Zawilski, E. C. Cox, *Cell* **123**, 1025 (2005).

[22] A. Sanchez, S. Choubey, J. Kondev, *Annu. Rev. Biophys.* **42**, 469 (2013).

[23] N. Mitarai, I. B. Dodd, M. T. Crooks, K. Sneppen, *PLoS Comput. Biol.* **4**, e1000109 (2008).

[24] R. Lutz, H. Bujard, *Nucleic Acids Res.* **25**, 1203 (1997).

[25] H. G. Garcia, R. Phillips, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 12174 (2011).

[26] H. Bremer, P. P. Dennis, *Escherichia coli and Salmonella: Cellular and Molecular Biology*, F. C. Neidhardt, *et al.*, eds. (ASM Press, Washington DC, 1996), pp. 1553–1569.

[27] I. L. Grigorova, N. J. Phleger, V. K. Mutalik, C. A. Gross, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 5332 (2006).

[28] D. Huh, J. Paulsson, *Nature Genet.* **43**, 95 (2011).

[29] S. Klumpp, Z. Zhang, T. Hwa, *Cell* **139**, 1366 (2009).

[30] S. Bakshi, A. Siryaporn, M. Goulian, J. C. Weisshaar, *Mol. Microbiol.* **85**, 21 (2012).

[31] O. K. Wong, M. Guthold, D. A. Erie, J. Gelles, *PLoS Biol.* **6**, e232 (2008).

[32] J. Elf, G. W. Li, X. S. Xie, *Science.* **316**, 1191 (2007).

[33] The authors wish to thank H. J. Lee, C. Wiggins, Y. Lin, X. Zhu, F. Weinert, M. Rydenfelt, R. Milo, H. Garcia, N. Belliveau and J. Sheung for useful discussions. We are grateful for support from the NIH through award numbers DP1 OD000217 (Directors Pioneer Award), R01 GM085286, and 1 U54 CA143869 (Northwestern PSOC Center); from La Fondation Pierre Gilles de Gennes (RP); and from the Donna and Benjamin M. Rosen Center for Bioengineering at Caltech (DLJ). Raw microscopy image data is archived in the Phillips laboratory at Caltech, and is available upon request.