# The one-message-per-cell-cycle rule: A conserved minimum transcription level for essential genes

Teresa W. Lo,[1] Han Kyou James Choi,[1] Dean Huang,[1] and Paul A. Wiggins[1,2,3,*]

[1]*Department of Physics, University of Washington, Seattle, Washington 98195, USA*
[2]*Department of Bioengineering, University of Washington, Seattle, Washington 98195, USA*
[3]*Department of Microbiology, University of Washington, Seattle, Washington 98195, USA*

The inherent stochasticity of cellular processes leads to significant cell-to-cell variation in protein abundance. Although this noise has already been characterized and modeled, its broader implications and significance remain unclear. In this paper, we revisit the noise model and identify the number of messages transcribed per cell cycle as the critical determinant of noise. In yeast, we demonstrate that this quantity predicts the non-canonical scaling of noise with protein abundance, as well as quantitatively predicting its magnitude. We then hypothesize that growth robustness requires an upper ceiling on noise for the expression of essential genes, corresponding to a lower floor on the transcription level. We show that just such a floor exists: a minimum transcription level of one message per cell cycle is conserved between three model organisms: *Escherichia coli*, yeast, and human. Furthermore, all three organisms transcribe the same number of messages per gene, per cell cycle. This common transcriptional program reveals that robustness to noise plays a central role in determining the expression level of a large fraction of essential genes, and that this fundamental optimal strategy is conserved from *E. coli* to human cells.

## INTRODUCTION

All molecular processes are inherently stochastic on a cellular scale, including the processes of the central dogma, responsible for gene expression [1, 2]. As a result, the expression of every protein is subject to cell-to-cell variation in abundance [1]. Many interesting proposals have been made to describe the potential biological significance of this noise, including bet-hedging strategies, the necessity of feedback in gene regulatory networks, *etc* [1, 3, 4]. However, it is less clear to what extent noise plays a central role in determining the function of the gene expression process more generally. For instance, Hausser *et al.* have described how the tradeoff between economy (*e.g.* minimizing the number of transcripts) and precision (minimizing the noise) explains why genes with high transcription rates and low translation rates are not observed [5]. Although these results suggest that noise may provide some coarse limits on the function of gene expression, this previous work does not directly address a central challenge posed by noise: How does the cell ensure that the lowest expression essential genes, which are subject to the greatest noise, have sufficient abundance in all cells for robust growth?

To investigate this question, we first focus on noise in *Saccharomyces cerevisiae* (yeast), and find that the noise scaling with protein abundance is not canonical. We reanalyze the canonical stochastic kinetic model for gene expression, the telegraph model [6–8], to understand the relationship between the underlying kinetic parameters and the distribution of protein abundance in the cell. As previously reported, we find that the protein abundance for a gene is described by a gamma distribution with two parameters: the *message number*, defined as the total gene message number transcribed per cell cycle, and the translation efficiency, which is the mean protein number translated per message. Protein expression noise is completely determined by the message number [3, 9]. Although these results have been previously reported, the distinction between message number *per cell* versus *per cell cycle* and even between *mean protein number* and *mean message number* is often neglected (*e.g.* [10]).

To explore the distinction between these parameters and provide clear evidence of the importance of the message number, we return to the analysis of noise in yeast. In yeast, the translation efficiency increases with message number [11]. By fitting an empirical model for the translation efficiency, we demonstrate that the noise should scale with a half-power of protein abundance. We demonstrate that this non-canonical scaling is observed and that our translation model makes a parameter-free prediction for the noise. The prediction is in close quantitative agreement with observation [12], confirming that the message number is the key determinant of noise strength.

Finally, we use this result to explore the hypothesis that there is a minimum expression level for essential genes, dictated by noise. The same mean expression level can be achieved by a wide range of different translation and transcription rates with different noise levels. We hypothesize that growth robustness requires that essential genes (but not non-essential genes) are subject to a floor expression level, below which there is too much cell-to-cell variation to ensure growth. To test this prediction, we analyze transcription in three model organisms, *Escherichia coli*, yeast, and *Homo sapiens* (human), with respect to three related gene characteristics: transcription rate, cellular message number, and message number per cell cycle. As predicted by the noise-based mechanism, we observe an organism-independent floor

* pwiggins@uw.edu

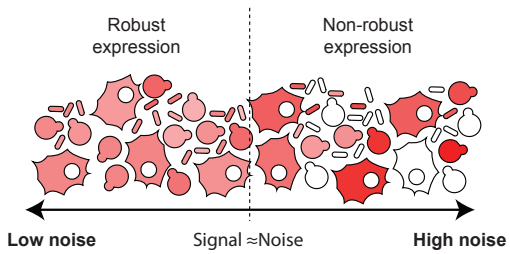


FIG. 1. **Robustness hypothesis:** The stochasticity in gene expression is represented by the red shading. We hypothesize that robust growth requires sufficiently low noise levels for cellular function. We hypothesize that this critical noise level should be below the level where the signal (mean) equals noise (standard deviation).

for the number of messages transcribed per cell cycle for essential genes, but not non-essential genes. We conclude that virtually all essential genes are transcribed at a rate of at least once per cell cycle. This analysis strongly supports the hypothesis that the same biological optimization imperatives, which determine the transcription rates of many low-expression genes, are conserved from *E. coli* to human.

## RESULTS

**Implications of noise on growth robustness.** With the realization of the stochasticity of central dogma processes, a key question is how cells can grow robustly in spite of cell-to-cell variations in protein expression. The noise in protein abundance is defined as the coefficient of variation squared [12–14]:

$$\mathrm{CV}_p^2 \equiv \frac{\sigma_p^2}{\mu_p^2}, \tag{1}$$

where $\sigma_p^2$ is the variance of protein number and $\mu_p \equiv \overline{N}_p$ its mean. It is important to emphasize that protein abundance must double between birth and cell division in symmetrically dividing cells during steady state growth. The protein abundance should therefore be interpreted either as expression per unit volume [15] or the abundance associated with cells of a defined volume [12].

The coefficient of variation is inversely related to protein abundance and therefore low-copy proteins have the highest noise [3, 9, 12–15]. The challenge faced by the cell is that many essential proteins, strictly required for cell growth, are relatively low abundance. How does the cell ensure sufficient protein abundance in spite of cell-to-cell variation in protein number? It would seem that growth robustness demands that, for essential proteins, the mean should be greater than the standard deviation:

$$\mathrm{CV}_p^2 < 1, \tag{2}$$

in order to ensure that protein abundance is sufficiently high enough to avoid growth arrest. To what extent do essential proteins obey this noise threshold?

**What determines the strength of the noise?** Usually, noise is argued to be proportional to inverse protein abundance (*e.g.* [3, 4, 10]):

$$\mathrm{CV}_p^2 \propto \mu_p^{-1}, \tag{3}$$

for low abundance proteins, motivated both by theoretical and experimental results [10, 15] and in some cases obeying a low-translation efficiency limit [15]:

$$\mathrm{CV}_p^2 \approx \mu_p^{-1}. \tag{4}$$

Can this model be used to make quantitative predictions of the noise? *E.g.*, is the scaling of Eq. 3 correct? Can the coefficient of proportionality be predicted? Although Eq. 3 appears to describe *E. coli* quite well [15], the situation in yeast is more complicated [16]. To analyze the statistical significance of the deviation from the canonical noise model in yeast, we can fit an empirical model to the noise [13, 14]:

$$\mathrm{CV}_p^2 = \frac{b}{\mu_p^a} + c. \tag{5}$$

In the null hypothesis, $a = 1$ (canonical scaling), while $b$ and $c$ are unknown parameters. $c$ corresponds to the noise floor. In the alternative hypothesis, all three coefficients are unknown. (A detailed description of the statistical model is given in the Supplemental Material Sec. A 5.)

The canonical model fails to fit the noise data for yeast as reported by Newman *et al.* [12]: The null hypothesis is rejected with p-value $p = 6 \times 10^{-36}$. The model fit to the data is shown in Fig. 2. The estimated scaling exponent for protein abundance in the alternative hypothesis is $a = 0.57 \pm 0.02$, and a detailed description of the statistical model and parameter fits is provided in Supplementary Material Sec. A 5 d. As shown in Fig. 2, even from a qualitative perspective, the scaling of the yeast noise at low copy number is much closer to $\mu_p^{-1/2}$ than to canonical assumption $\mu_p^{-1}$ (Eq. 3). In particular, above the detection threshold, the noise is always larger than the low-translation efficiency limit (Eq. 4).

**Stochastic kinetic model for central dogma.** To understand the failure of the canonical assumptions, we revisit the underlying model. The telegraph model for the central dogma describes multiple steps in the gene expression process: Transcription generates mRNA messages [17]. These messages are then translated to synthesize the protein gene products [17]. Both mRNA and protein are subject to degradation and dilution [18]. (See Fig. 3A.) At the single cell level, each of these processes are stochastic. We will model these processes with the

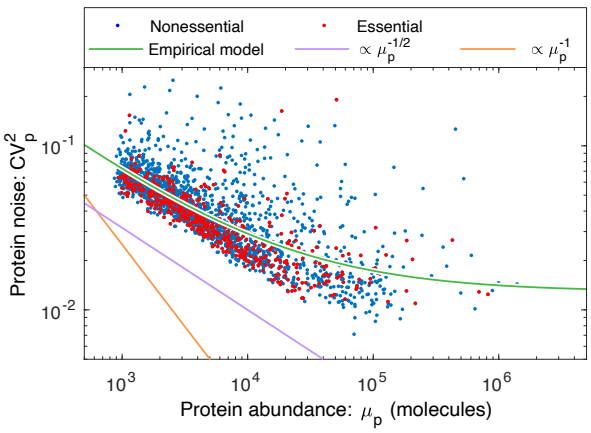


FIG. 2. **A non-canonical scaling is observed for gene-expression noise in yeast.** The protein expression noise ($\mathrm{CV}_p^2$) for yeast scales like $\mu_p^{-1/2}$ (purple) rather than the canonical $\mu_p^{-1}$ (orange) for low-abundance proteins. (Data from Ref. [12].) An empirical noise model (Eq. 5, green) fit to the essential genes gives an estimate of the protein-abundance scaling of $\mu_p^{-0.57}$.

stochastic kinetic scheme [17]:

$$\begin{array}{ccccc} \mathrm{DNA} & \xrightarrow{\beta_m} & \mathrm{mRNA} & \xrightarrow{\beta_p} & \mathrm{Protein} \\ & & \gamma_m \downarrow & & \gamma_p \downarrow \\ & & \emptyset & & \emptyset, \end{array} \tag{6}$$

where $\beta_m$ is the transcription rate ($\mathrm{s}^{-1}$), $\beta_p$ is the translation rate ($\mathrm{s}^{-1}$), $\gamma_m$ is the message degradation rate ($\mathrm{s}^{-1}$), and $\gamma_p$ is the protein effective degradation rate ($\mathrm{s}^{-1}$). The message lifetime is $\tau_m \equiv \gamma_m^{-1}$. For most protein in the context of rapid growth, dilution is the dominant mechanism of protein depletion and therefore $\gamma_p$ is approximately the growth rate [15, 19, 20]: $\gamma_p = T^{-1} \ln 2$, where $T$ is the doubling time. We will discuss a more general scenario below.

**Statistical model for protein abundance.** To study the stochastic dynamics of gene expression, we used a stochastic Gillespie simulation [21, 22]. (See Supplemental Material Sec. A 1.) In particular, we were interested in the explicit relation between the kinetic parameters $(\beta_m, \gamma_m, \beta_p, \gamma_p)$ and experimental observables.

Consistent with previous reports [3, 9], we find that the distribution of protein number per cell (at cell birth) was described by a gamma distribution: $N_p \sim \Gamma(\theta_\Gamma, k_\Gamma)$, where $N_p$ is the protein number at cell birth and $\Gamma$ is the gamma distribution which is parameterized by a scale parameter $\theta_\Gamma$ and a shape parameter $k_\Gamma$. (See Supplementary Material Sec. A 1.) The relation between the four kinetic parameters and these two statistical parameters has already been reported, and have clear biological interpretations [9]: The scale parameter:

$$\theta_\Gamma = \varepsilon \ln 2, \tag{7}$$

is proportional to the translation efficiency:

$$\varepsilon \equiv \frac{\beta_p}{\gamma_m}, \tag{8}$$

where $\beta_p$ is the translation rate and $\gamma_m$ is the message degradation rate. $\varepsilon$ is understood as the mean number of proteins translated from each message transcribed. The shape parameter $k_\Gamma$ can also be expressed in terms of the kinetic parameters [9]:

$$k_\Gamma = \frac{\beta_m}{\gamma_p}; \tag{9}$$

however, we will find it more convenient to express the scale parameter in terms of the cell-cycle message number:

$$\mu_m \equiv \beta_m T = k_\Gamma \ln 2, \tag{10}$$

which can be interpreted as the mean number of messages transcribed per cell cycle. Forthwith, we will abbreviate this quantity *message number* in the interest of brevity.

In terms of two gamma parameters, the mean and the squared coefficient of variation are:

$$\mu_p = k_\Gamma \theta_\Gamma = \mu_m \varepsilon \tag{11}$$

$$\mathrm{CV}_p^2 = \frac{1}{k_\Gamma} = \frac{\ln 2}{\mu_m}, \tag{12}$$

where the noise depends on the message number ($\mu_m$), not the mean protein number ($\mu_p$). (Eq. 12 only applies when $\varepsilon \gg 0$ [3, 9].) Are these theoretical results consistent with the canonical model (Eq. 3)? We can rewrite the noise in terms of the protein abundance and translation efficiency:

$$\mathrm{CV}_p^2 = \frac{\varepsilon \ln 2}{\mu_p}, \tag{13}$$

which implies that the canonical model only applies when the translation efficiency ($\varepsilon$) is independent of expression ($\mu_p$).

**Measuring the message number.** The prediction for the noise (Eq. 12) depends on the message number ($\mu_m$). However, mRNA abundance is typically characterized by a closely related, but distinct quantity: Quantitative RNA-Seq and methods that visualize fluorescently-labeled mRNA molecules typically measure the number of messages per cell [6]. We will call the mean of this number the *cellular message number* $\mu_{m/c}$. In the kinetic model, these different message abundances are related:

$$\mu_m = \frac{T}{\tau_m} \mu_{m/c}, \tag{14}$$

by the message recycling ratio, $T/\tau_m$, which can be interpreted as the average number of times messages are recycled during the cell cycle. To estimate the message number, we will scale the observed cellular message number $\mu_{m/c}$ by the message recycling ratio, using the mean message lifetime. Fig. 3C illustrates the difference

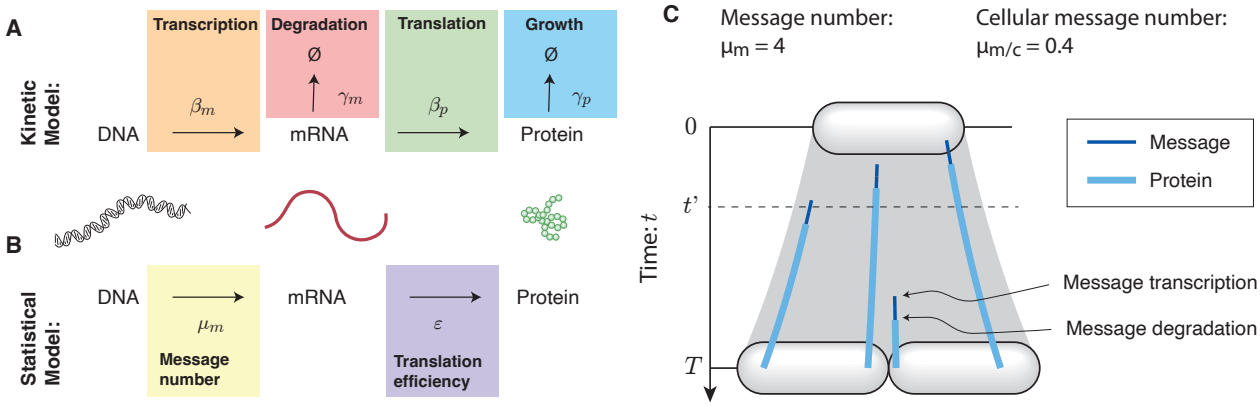


FIG. 3. **Panel A: Kinetic model for the central dogma.** The telegraph model is a stochastic kinetic model for protein synthesis, described by four gene-specific rate constants: the transcription rate ($\beta_m$), the message degradation rate ($\gamma_m$), the translation rate ($\beta_p$), and the dilution rate ($\gamma_p$). **Panel B: Statistical model for the central dogma.** The predicted distribution in protein abundance is described by a gamma distribution, which is parameterized by two unitless constants: the shape parameter $\mu_m$, the mean number of messages transcribed per cell cycle, and the scale parameter $\varepsilon$, the mean number of proteins translated per message. **Panel C: Message number.** The *message number* ($\mu_m$) is defined as the mean total number of messages (dark blue) transcribed per cell cycle. Here, four total messages are transcribed and translated to protein (light blue); however, due to message degradation, at time $t'$, only one message is present in the cell. Cellular message number ($\mu_{m/c}$) is defined as the mean number of messages per cell at time $t$.

| Model organism | Growth condition | Doubling time: $T$ | Message lifetime: $\tau_m$ | Message recycling ratio: $T/\tau_m$ | Total number of messages /cell: $N^{\text{tot}}_{m/c}$ | Total number of messages /cell-cycle: $N^{\text{tot}}_{m}$ | Total number of proteins: $N^{\text{tot}}_{p}$ | Average translation efficiency: $\varepsilon$ | Average translation rate: $\beta_p$ (h$^{-1}$) |
|---|---|---|---|---|---|---|---|---|---|
| *Escherichia coli* (*E. coli*) | LB | 30 min | 2.5 min | 12 | $7.8 \times 10^3$ | $9.4 \times 10^4$ | $3 \times 10^6$ | 32 | 770 |
| | M9 | 90 min | 2.5 min | 36 | $2.4 \times 10^3$ | $8.6 \times 10^4$ | $3 \times 10^6$ | 35 | 833 |
| *Sacchromyces cerevisiae (Yeast–haploid)* | YEPD | 90 min | 22 min | 4 | $2.9 \times 10^4$ | $1.2 \times 10^5$ | $5 \times 10^7$ | 420 | 1100 |
| *Homo sapiens (Human)* | Tissue | 24 h | 14 h | 1.7 | $3.6 \times 10^5$ | $6.2 \times 10^5$ | $2 \times 10^9$ | $3.2 \times 10^3$ | 230 |

TABLE I. **Central dogma parameters for three model organisms.** Columns three through seven hold representative values for measured central-dogma parameters for the model organisms described in the paper. The sources of the numbers and estimates are described in the Supplemental Material Sec. A 2.

between the message number and the cellular message number. The mean lifetimes, message recycling ratios, as well as the total message number for three model organisms are shown in Tab. I.

**Construction of an empirical model for protein number.** To model the noise as a function of protein abundance ($\mu_p$), we will determine the empirical relation between mean protein levels and message abundance by fitting to Eq. 11. Note that the objective here is only to estimate $\mu_m$ from $\mu_p$, not to model the process mechanistically (*e.g.* [23].) The message numbers are estimated from RNA-Seq measurements, scaled as described above (Eq. 14). The protein abundance numbers come from fluorescence and mass-spectrometry based assays [12, 24], with overall normalization chosen to match reported total cellular protein content. (See Supplemental Material Sec. A 3 d.) The resulting fit generates our empirical translation model for yeast:

$$\mu_p \;=\; 8.0\,\mu_m^{2.1}, \tag{15}$$

where both means are in units of molecules. (An error analysis for both model parameters is described in Supplementary Material Sec. A 4 b.) The data and model are shown in Fig. 5A.

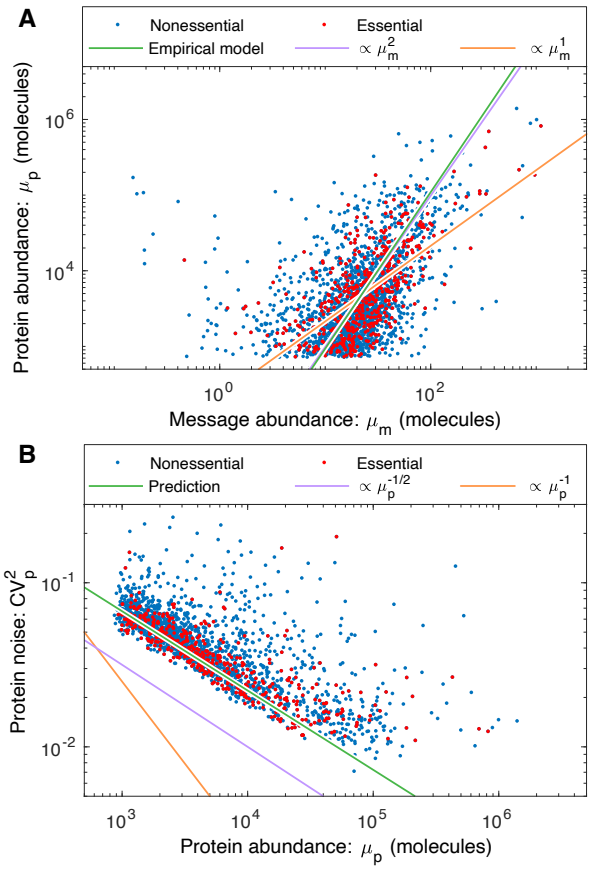


FIG. 4. **Panel A: An empirical model for protein number $\mu_p$ in yeast.** The canonical noise model assumes constant translation efficiency, which would imply that protein number is proportional to the message number (orange); however, the empirical fit (green) shows that protein number scales close to the square of message number (purple): $\mu_p \propto \mu_m^2$. The protein abundance has a cutoff near $10^1$ due to the autofluorescence cutoff [12]. **Panel B: The statistical noise model predicts the observed noise.** The statistical noise model (Eq. 12) and empirical model for protein number (Eq. 15) make a parameter-free prediction of the noise (green). This prediction both closely matches the observed scaling ($\propto \mu_p^{-1/2}$, purple) relative to the canonical scaling ($\propto \mu_p^{-1}$, orange) and quantitatively estimates magnitude (vertical offset). This prediction does not include the contribution of noise floor, relevant for describing high-expression proteins.

**Prediction of the noise scaling with abundance.** Now that we have fit an empirical model that relates $\mu_p$ and $\mu_m$, we return to the problem of predicting the yeast noise. We apply the relation (Eq. 15) to Eq. 12 to make a parameter-free prediction of the noise as a function of protein abundance:

$$\mathrm{CV}_p^2 = 1.9\ \mu_p^{-0.48}. \tag{16}$$

An error analysis for both model parameters is described in Supplementary Material Sec. A 4 e. Our noise model (Eq. 16) makes both a qualitative and quantitative prediction: (i) From a qualitative perspective, the model suggests that the $\mu_p$ exponent should be roughly $\frac{1}{2}$ for yeast, rather than the canonically assumed scaling exponent of 1. (ii) From a quantitative perspective, the model also predicts the coefficient of proportionality if the empirical relation between protein and message abundances is known (Eq. 15).

**Observed noise in yeast matches the predictions of the empirical model.** Newman *et al.* have characterized protein noise by flow cytometry of strains expressing fluorescent fusions expressed from their endogenous promoters [12]. The comparison of this data to the prediction of the statistical expression model (Eq. 16) are shown in Fig. 5. From a qualitative perspective, the predicted scaling exponent of $-0.48$ comes very close to capturing the scaling of the noise, as determined by the direct fitting of the empirical noise model (Eq. 5 and Fig. 2). From a quantitative perspective, the predicted coefficient of Eq. 16 also fits the observed noise.

From both the statistical analysis (Eq. 5) and visual inspection (Fig. 5C), it is clear that the noise in yeast does not obey the canonical model (Eq. 3). However, the noise in *E. coli* does obey the canonical model for low copy messages [15]. (See Fig. 5C.) Why does the noise scale differently in the two organisms? The key difference is that the empirical relation between the protein and message numbers are different. In *E. coli*, $\mu_p \propto \mu_m^1$ [25]. Our analysis therefore predicts the canonical model (Eq. 3) should hold for *E. coli*, but not for yeast, as illustrated schematically in Fig. 5. (Additional discussion can be found in the Supplementary Material Sec. A 5 f.).

**Implications of growth robustness for translation.** Before continuing with the noise analysis, we to focus on the significance of the empirical relationship between the protein and message numbers (Eq. 15). How can the cell counteract noise-induced reductions in robustness? Eq. 11 implies that gene expression can be thought of as a two-stage amplifier [17]: The first stage corresponds to transcription with a gain of message number $\mu_m$, and the second stage corresponds to translation with a gain in translation efficiency $\varepsilon$. (See Fig. 5AB.) The noise is completely determined by the first stage of amplification, provided that $\varepsilon \gg 0$ [3, 9]. Genes with low transcription levels are the noisiest. For these genes, the cell can achieve the same mean gene expression ($\mu_p$) with lower noise by increasing the gain of the first stage (increasing message number) and decreasing the gain of the second stage (the translation efficiency) by the same factor. This is most clearly understood by reducing $\varepsilon$ at fixed $\mu_p$ in Eq. 13. Highly transcribed genes have low noise and can therefore tolerate higher translation efficiency in the interest of economy (decreasing the total number of messages) [5]. Growth robustness therefore predicts that the translation efficiency should grow with transcription level.

**Translation efficiency increases with expression level in yeast.** The translation efficiency (Eq. 8) can be deter-

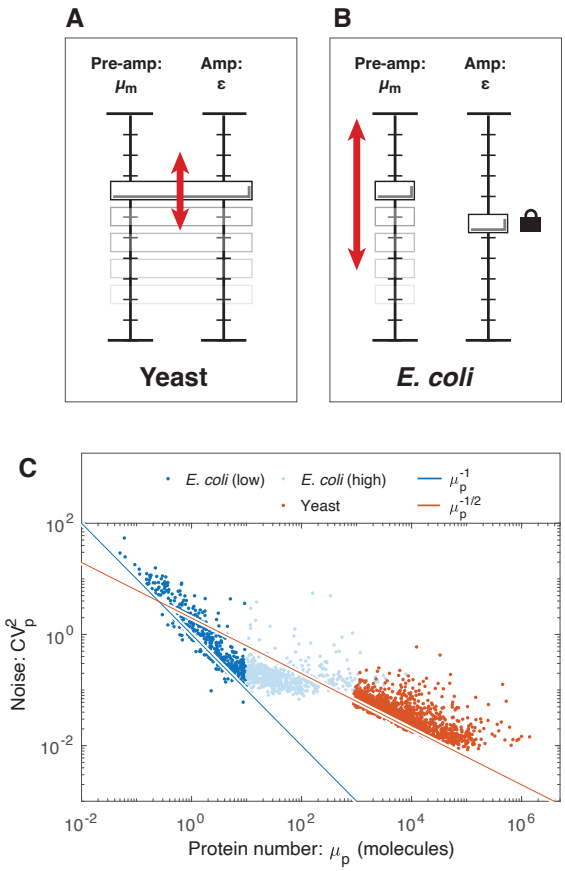


FIG. 5. **Understanding the distinct central dogma strategies using the amplifier analogy. Panel A: Yeast.** High expression ($\mu_p$) is typically achieved by coordinated small increases in both transcription ($\mu_m$) and translation ($\varepsilon$), relative to low-expression genes. **Panel B: *E. coli*.** High expression ($\mu_p$) is typically achieved by a large increase in transcription ($\mu_m$) only, relative to low-expression genes. Translation ($\varepsilon$) is uncorrelated. **Panel C: Distinct noise scaling with gene expression.** Due to the coordinated changes in both transcription and translation in yeast, noise scaling is weaker than in *E. coli*, where only transcription changes. The noise of high-expression *E. coli* genes is determined by the noise floor.

mined from the empirical translation model (Eq. 15):

$$\varepsilon \;=\; 8.0\,\mu_m^{1.1}, \tag{17}$$

as a function of message number. (An error analysis for both model parameters is described in Supplementary Material Sec. A 4 d.) In yeast, the translation efficiency clearly has a strong dependence on message number $\mu_m$, and grows with the expression level, exactly as predicted by robustness arguments. We note the contrast to the translation efficiency in *E. coli*, which is roughly constant [25]. (See Supplementary Material Sec. A 5 f.) We will speculate about the rationale for these differences in the discussion below.

**Implications of growth robustness for transcription.** In addition to the prediction of translation efficiency depending on transcription, a second qualitative prediction of growth robustness is that essential gene expression should have a noise ceiling, or maximum noise level (Eq. 2), where noise above this level would be too great for robust growth. The fit between the statistical model and the observed noise has an important implication beyond confirming the predictions of the telegraph and statistical models for noise: The identification of the message number, $\mu_m$, as the key determinant of noise allows us to use this quantity as a proxy for noise in quantitative transcriptome analysis.

To identify a putative transcriptional floor, we now broaden our consideration beyond yeast to characterize the central dogma in two other model organisms: the bacterium *Escherichia coli* and *Homo sapiens* (human). We will also analyze three different transcriptional statistics for each gene: transcription rate ($\beta_m$), cellular message number ($\mu_{m/c}$), and message number ($\mu_m$). Analysis of these organisms explores orders-of-magnitude differences in characteristics of the central dogma, including total message number, protein number, doubling time, message lifetime, and number of essential genes. (See Tab. I.) In particular, as a consequence of these differences, the three statistics describing transcription: transcription rate, cellular message number and message number are all distinct. Genes with matching message numbers in two different organisms will not have matching transcription rates or cellular message numbers. We hypothesize that cells must express essential genes above some threshold message number for robust growth; however, we expect to see that non-essential genes can be expressed at much lower levels since growth is not strictly dependent on their expression. The signature of a noise-robustness mechanism would be the absence of essential genes for low message numbers.

**No organism-independent threshold is observed for transcription rate or cellular message number.** Histograms of the per-gene transcription rate and cellular message number are shown in Fig. 6 for *E. coli*, yeast, and human. Consistent with existing reports, essential genes have higher expression than non-essential genes on average; however, there does not appear to be any consistent threshold in *E. coli* (even between growth conditions), yeast, or human transcription, either as characterized by the transcription rate ($\beta_m$) or the cellular message number ($\mu_{m/c}$). For instance, the per gene rate of transcription is much lower in human cells than *E. coli* under rapid growth conditions, with yeast falling in between.

**An organism-independent threshold is observed for message number for essential genes.** In contrast to the other two transcriptional statistics, there is a consistent lower limit, or floor, on message number ($\mu_m$) of somewhere between 1 and 10 messages per cell cycle for essential genes. (See Fig. 7.) Non-essential genes can be expressed at a much lower level. This floor is consistent

| Model organism | Estimated minimum essential gene | | | | |
|---|---|---|---|---|---|
| | Maximum essential gene noise: $\max \mathrm{CV}_p^2$ | messages /cell-cycle: $\mu_m^{\min}$ | messages /cell: $\mu_{m/c}^{\min}$ | transcription rate: $\beta_m^{\min}$ ($\mathrm{h}^{-1}$) | proteins: $\mu_p^{\min}$ |
| *E. coli* (LB) | 0.7 | 1 | 0.08 | 2 | 30 |
| (M9) | 0.7 | 1 | 0.03 | 0.7 | 30 |
| Yeast | 0.7 | 1 | 0.2 | 0.7 | 400 |
| Human | 0.7 | 1 | 0.6 | 0.04 | 3000 |

TABLE II. **Estimates of threshold levels for the central dogma in three model organisms.** Estimates for the lower thresholds of transcription statistics as inferred from our analysis based on the *one-message-per-cell-cycle rule*.

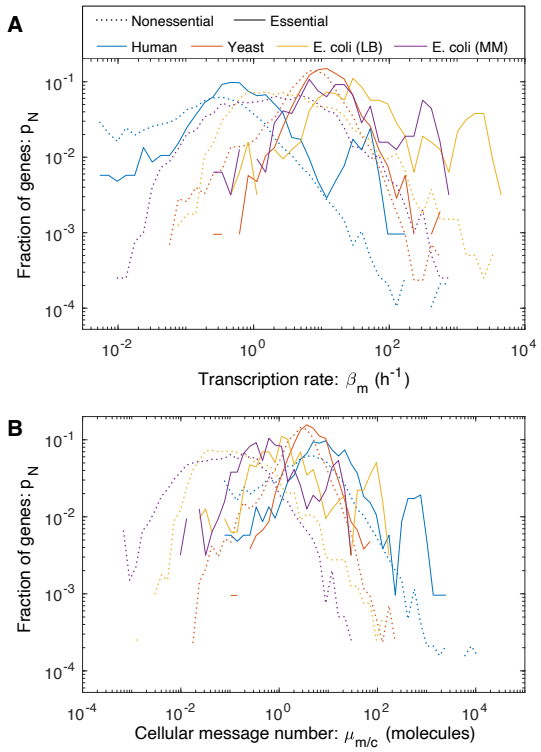


FIG. 6. **Transcription in three model organisms.** We characterized different gene transcriptional statistics in three model organisms. In *E. coli*, two growth conditions were analyzed. **Panel A: The distribution of gene transcription rate.** The transcription rate varies by two orders-of-magnitude between organisms. **Panel B: The distribution of gene cellular message number.** There is also a two-order-of-magnitude variation between cellular message numbers.

not only between *E. coli*, growing under two different conditions, but also between the three highly-divergent organisms: *E. coli*, yeast and human. We will conservatively define the minimum message number as

$$\mu_m^{\min} \equiv 1, \tag{18}$$

and summarize this observation as the *one-message-per-cell-cycle rule* for essential gene expression.

In addition to the common floor for essential genes, there is a common gene expression distribution shape shared between organisms dependent on the message numbers, especially for low-expression essential genes. This is observed in spite of the significantly larger number of essential genes in human relative to *E. coli*. (See Fig. 7.) Interestingly, there is also a similarity between the non-essential gene distributions for *E. coli* and human, but not for yeast, which appears to have a much lower fraction of genes expressed at the lowest message numbers.

**What genes fall below-threshold?** We have hypothesized that essential genes should be expressed above a threshold value for robustness. It is therefore interesting to consider the function of genes that fall below this proposed threshold. Do functions of these genes give us any insight into essential processes that do not require robust gene expression?

Since our own preferred model system is *E. coli*, we focus here. Our essential gene classification was based on the construction of the Keio knockout library [26]. By this classification, 10 essential genes were below threshold. (See Supplementary Material Tab. IV.) Our first step was to determine what fraction of these genes were also classified as essential using transposon-based mutagenesis [27, 28]. Of the 10 initial candidates, only one gene, *ymfK*, was consistently classified as an essential gene in all three studies, and we estimate that its message number is just below the threshold ($\mu_m = 0.4$). *ymfK* is located in the lambdoid prophage element e14 and is annotated as a CI-like repressor which regulates lysis-lysogeny decision [29]. In $\lambda$ phase, the CI repressor represses lytic genes to maintain the lysogenic state. A conserved function for *ymfK* is consistent with it being classified as essential, since its regulation would prevent cell lysis. However, since *ymfK* is a prophage gene, not a host gene, it is not clear that its expression should optimize host fitness, potentially at the expense of phage fitness. In summary, closer inspection of below-threshold essential genes supports the threshold hypothesis.

**Maximum noise for essential genes.** The motivation for hypothesizing a minimum threshold for message number was noise-robustness, or the existence of a hypothesized noise ceiling above which essential gene expression is too noisy to allow robust cellular proliferation.

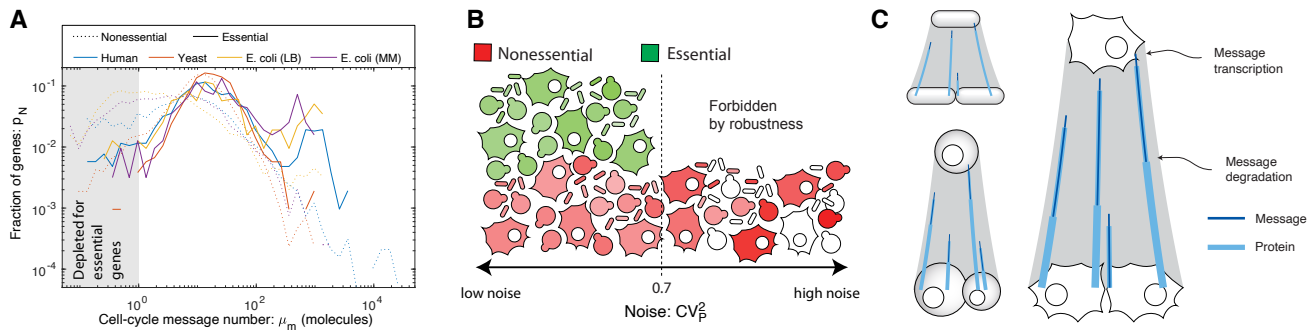


FIG. 7. **Panel A: Transcription in three model organisms. The distribution of gene message number.** All organisms have roughly similar distributions of message number for essential genes, which are not observed for message numbers below a couple per cell cycle. However, non-essential genes can be expressed at much lower levels. **Panel B: Nonessential genes tolerate higher noise levels than essential genes.** The floor of message number is consistent with a noise ceiling of $\mathrm{CV}_p^2 = 0.7$ for essential genes (green). Nonessential genes (red) are observed with lower transcription levels. **Panel C: Conserved transcriptional program for essential genes.** The message number per gene (number of messages transcribed per cell cycle) is roughly identical in *E. coli*, yeast, and human. We show this schematically.

With the *one-message-per-cell-cycle rule*, $\mu_m^{min} \equiv 1$, we can estimate the essential gene noise ceiling using Eq. 12:

$$\mathrm{CV}_p^2 \leq 0.7, \tag{19}$$

for essential genes. Since noise depends only on the message number, we expect to observe the same limit in all organisms if the message number floor is conserved.

**Estimating the floor on central-dogma parameters.** If message number floor is conserved, a limit can be estimated for the floor value on other transcriptional parameters. Using Eq. 14, we can estimate the floor on the cellular message number (as measured in RNA-Seq measurements):

$$\mu_{m/c}^{\min} = \tfrac{\tau_m}{T}, \tag{20}$$

for essential genes. Similarly, we can use Eq. 9 to estimate the minimum transcription rate:

$$\beta_m^{\min} = \tfrac{1}{T}, \tag{21}$$

for essential genes. Again, this result has an intuitive interpretation as the one-message-per-cell-cycle rule. Finally, we can estimate a floor on essential protein abundance, assuming a constant translation efficiency using Eq. 11:

$$\mu_p^{\min} = \varepsilon, \tag{22}$$

for essential genes, where $\varepsilon$ is the translation efficiency (which we will assume is well approximated by the mean in the context of the estimate). All four floor estimates for each model organism are shown in Tab. II.

## DISCUSSION

**Noise by the numbers.** Although there has already been significant discussion of the scaling of biological noise with protein abundance [3, 9, 10, 12, 15], our study is arguably the first to test the predictions of the telegraph and statistical noise models against absolute measurements of protein and message abundances. This approach is particularly important for the message number ($\mu_m$), which determines the magnitude of the noise in protein expression, and facilitates direct comparisons of noise between organisms as well as identifying the common distributions of message number for genes, that are conserved from bacteria to human.

**Noise scaling in *E. coli* versus yeast.** A key piece of evidence for the significance of the message number was the observation of the non-canonical scaling of the yeast noise with protein abundance (Fig. 5); however, the canonical model (Eq. 3) does accurately describe the noise in *E. coli* (see Fig. 11). Why does the noise scale differently? In *E. coli*, the translation efficiency is only weakly correlated with the gene expression [25], and therefore the canonical model is a reasonable approximation (Supplementary Material Sec. A 5 d). However, we also argued that translation efficiency should grow with expression level. Why is this not observed in *E. coli*? Due to the high noise floor in *E. coli*, nearly all essential genes are expressed at a sufficiently high expression level such that the noise is dominated by the noise floor [15]. As a consequence, increasing the message number, while decreasing translation efficiency, does not decrease the noise even as it increases the metabolic load as a result of increased transcription. (A closely related point has recently been made in *Bacillus subtilis* [30], where Deloupy *et al.* report that the noise cannot be tuned by adjusting the message number due to the noise floor.) Our expectation is therefore that other bacterial cells will look similar to *E. coli*: They will have a higher noise floor and a similar scaling of noise with protein abundance.

In contrast, due to the lower noise floor, we expect eukaryotic cells to optimize the central dogma processes like yeast and as a result will have a similar non-canonical scaling of noise with protein abundance. Although this non-canonical scaling is clear from the abundance data (Fig. 5B), there is an important qualification to emphasize: the mechanism that gives rise to the non-canonical scaling is due to the correlation between translation efficiency and transcription. Regulatory changes that effect only transcription (*i.e.* increase $\mu_m$) and not translation ($\varepsilon$) should obey the canonical noise model (Eq. 3). This scenario may help explain why Bar-Even *et al.* claim to observe canonical noise scaling in yeast [10], studying a subset of genes under a range of conditions resulting in differential expression levels. The failure of the canonical noise model (Eq. 3) at the proteome level in yeast (Eq. 16) is a consequence of genome-wide optimization of the relative transcription and translation rates.

**Essential versus non-essential genes.** What genes are defined as *essential* is highly context specific [31]. It is therefore important to consider whether the comparison between these two classes of genes is informative in the context of our analysis. We believe the example of *lac* operon in *E. coli* is particularly informative in this respect. The genes *lacZYA* are conditionally essential: they are required when lactose is the carbon source; however, these genes are repressed when glucose is the carbon source. Our expectation is that these conditionally essential genes will obey the one-message-per-cell-cycle rule when these genes are required; however, they need not obey this rule when the genes are repressed. By analyzing essential genes, we are limiting the analysis to transcriptionally-active genes, whereas the non-essential category contains both transcriptionally-active and silenced genes.

**Protein degradation and transcriptional bursting.** Two important mechanisms can act to significantly increase the noise above the levels we predict: protein degradation and transcriptional bursting. Although the dominant mechanism of protein depletion is dilution in *E. coli*, protein degradation plays an important role in many organisms, especially in eukaryotic cells [32, 33]. If protein degradation depletes proteins faster than dilution, the shape parameter decreases below our estimate (Eq. 9), increasing the noise. Likewise, the existence of transcriptional bursting, in which the chromatin switches between transcriptionally active and quiescent periods, can also act to increase the noise [1, 7, 34]. Since the presence of both these mechanisms increases the noise beyond what is predicted by the message number, they do not affect our estimate of the minimum threshold for $\mu_m$.

**The biological implications of noise.** What are the biological implications of gene expression noise? Many important proposals have been made, including bet-hedging strategies, the necessity of feedback in gene regulatory networks, *etc* [1]. Our analysis suggests that noise influences the optimal function of the central dogma process generically. Hausser *et al.* have already discussed some aspects of this problem and use this approach to place coarse limits on transcription versus translation rates [5]. The transcriptional floor for essential genes that we have proposed places much stronger limits on the function of the central dogma.

Although we describe our observations as a floor, a more nuanced description of the phenomenon is a common distribution of gene message numbers, peaked at roughly 15 messages per cell cycle and cutting off close to one message per cell cycle. Does this correspond to a hard limit? We expect that this does not since there are a small fraction of genes, classified as essential, just below this limit; however, it does appear that virtually all essential genes have optimal expression levels above this threshold. The common distribution of message number clearly suggests that noise considerations shape the function of the central dogma for virtually all genes. Exploring this hypothesis will require quantitative models that explicitly realize the high cost of noise-induced low essential-protein abundance. We will present such an analysis elsewhere.

**Adapting the central dogma to increased cell size and complexity.** Although core components of the central dogma machinery are highly-conserved, there has been significant complexification of both the transcriptional and translational processes in eukaryotic cells [35]. Given this increased regulatory complexity, it is unclear how the central dogma processes should be adapted in larger and more complex cells. An important clue to this adaptation comes from *E. coli* proliferating with different growth rates. Although there are very significant differences between the cellular message number as well as the overall transcription rate under the two growth conditions, there is very little difference in message number. In short, roughly the same number of messages are made during the cell cycle, but they are made more slowly under slow growth conditions.

How does this picture generalize in eukaryotic cells? Although both the total number of messages and the number of essential and non-essential genes are larger in both yeast and human cells, the distribution of the message number per gene is essentially the same as *E. coli* (Fig. 7). The conservation of the message number between organisms is consistent with all of these organisms being optimized with respect to the same trade-off between economy and robustness to noise.

**Data availability.** We include a source data file which includes the estimated message numbers as well as essential/nonessential classifications for each organism.

**Acknowledgments.** The authors would like to thank B. Traxler, A. Nourmohammad, J. Mougous, K. Cutler, M. Cosentino-Lagomarsino, S. van Teeffelen, and S. Murray. This work was supported by NIH grant R01-GM128191.

**Author contributions:** T.W.L., H.K.J.C., D.H. and P.A.W. conceived the research. T.W.L. and P.A.W. performed the analysis. H.K.J.C. and D.H. performed experiments and analysis. T.W.L., H.K.J.C., D.H. and P.A.W. wrote the paper.

**Competing interests:** The authors declare no competing interests.

[1] Raser, J. M. & O'Shea, E. K. Noise in gene expression: origins, consequences, and control. *Science* **309**, 2010–3 (2005).

[2] Phillips, R., Kondev, J., Theriot, J. & Orme, N. *Physical Biology of the Cell* (Garland Science, 2013).

[3] Paulsson, J. & Ehrenberg, M. Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks. *Phys Rev Lett* **84**, 5447–50 (2000).

[4] Paulsson, J. Summing up the noise in gene networks. *Nature* **427**, 415–8 (2004).

[5] Hausser, J., Mayo, A., Keren, L. & Alon, U. Central dogma rates and the trade-off between precision and economy in gene expression. *Nat Commun* **10**, 68 (2019).

[6] Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* **4**, e309 (2006).

[7] Iyer-Biswas, S., Hayot, F. & Jayaprakash, C. Stochasticity of gene products from transcriptional pulsing. *Phys Rev E Stat Nonlin Soft Matter Phys* **79**, 031911 (2009).

[8] J, P. & B, Y. Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.* **48**, 222–234 (1995).

[9] Friedman, N., Cai, L. & Xie, X. S. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys Rev Lett* **97**, 168302 (2006).

[10] Bar-Even, A. *et al.* Noise in protein expression scales with natural protein abundance. *Nat Genet* **38**, 636–43 (2006).

[11] Weinberg, D. E. *et al.* Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. *Cell Rep* **14**, 1787–1799 (2016).

[12] Newman, J. R. S. *et al.* Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* **441**, 840–6 (2006).

[13] Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–6 (2002).

[14] Swain, P. S., Elowitz, M. B. & Siggia, E. D. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Natl Acad Sci U S A* **99**, 12795–800 (2002).

[15] Taniguchi, Y. *et al.* Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–8 (2010).

[16] Although there have been claims that Eq. 3 is consistent with the data [10], these authors did not fit competing models, nor did they perform a proteome-wide analysis of protein abundance and noise.

[17] Crick, F. Central dogma of molecular biology. *Nature* **227**, 561–3 (1970).

[18] Hargrove, J. L. & Schmidt, F. H. The role of mRNA and protein stability in gene expression. *FASEB J* **3**, 2360–70 (1989).

[19] Koch, A. L. & Levy, H. R. Protein turnover in growing cultures of *Escherichia coli*. *J Biol Chem* **217**, 947–57 (1955).

[20] Martin-Perez, M. & Villén, J. Determinants and regulation of protein turnover in yeast. *Cell Syst* **5**, 283–294.e5 (2017).

[21] Gillespie, D. A rigorous derivation of the chemical master equation. *Physica A* **188**, 404–425 (1992).

[22] Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry* **81**, 2340–2361 (1977).

[23] Shah, P., Ding, Y., Niemczyk, M., Kudla, G. & Plotkin, J. B. Rate-limiting steps in yeast protein translation. *Cell* **153**, 1589–601 (2013).

[24] de Godoy, L. M. F. *et al.* Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251–4 (2008).

[25] Balakrishnan, R. *et al.* Principles of gene regulation quantitatively connect DNA to RNA and proteins in bacteria. *Science* **378**, eabk2066 (2022).

[26] Baba, T., Huan, H.-C., Datsenko, K., Wanner, B. L. & Mori, H. The applications of systematic in-frame, single-gene knockout mutant collection of *Escherichia coli* K-12. *Methods Mol Biol* **416**, 183–94 (2008).

[27] Gerdes, S. Y. *et al.* Experimental determination and system level analysis of essential genes in *Escherichia coli* mg1655. *J Bacteriol* **185**, 5673–84 (2003).

[28] Goodall, E. C. A. *et al.* The essential genome of *Escherichia* coli k-12. *mBio* **9** (2018).

[29] Mehta, P., Casjens, S. & Krishnaswamy, S. Analysis of the lambdoid prophage element e14 in the *E. coli* k-12 genome. *BMC Microbiol* **4**, 4 (2004).

[30] Deloupy, A. *et al.* Extrinsic noise prevents the independent tuning of gene expression noise and protein mean abundance in bacteria. *Sci Adv* **6** (2020).

[31] Chin, B. L., Ryan, O., Lewitter, F., Boone, C. & Fink, G. R. Genetic variation in *Saccharomyces cerevisiae*: circuit diversification in a signal transduction network. *Genetics* **192**, 1523–32 (2012).

[32] Ciechanover, A. Intracellular protein degradation: from a vague idea thru the lysosome and the ubiquitin-proteasome system and onto human diseases and drug targeting. *Cell Death Differ* **12**, 1178–90 (2005).

[33] Eden, E. *et al.* Proteome half-life dynamics in living human cells. *Science* **331**, 764–8 (2011).

[34] Golding, I., Paulsson, J., Zawilski, S. M. & Cox, E. C. Real-time kinetics of gene activity in individual bacteria. *Cell* **123**, 1025–36 (2005).

[35] Cooper, G. M. *The Cell: A Molecular Approach. 2nd edition* (Sinauer Associates 2000, 2000).

[36] Bernstein, J. A., Khodursky, A. B., Lin, P.-H., Lin-Chao, S. & Cohen, S. N. Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *Proc Natl Acad Sci U S A* **99**, 9697–702 (2002).

[37] Chen, H., Shiroguchi, K., Ge, H. & Xie, X. S. Genome-wide study of mRNA degradation and transcript elongation in *Escherichia coli*. *Mol Syst Biol* **11**, 808 (2015).

[38] Bartholomäus, A. *et al.* Bacteria differently regulate mRNA abundance to specifically respond to various stresses. *Philos Trans A Math Phys Eng Sci* **374** (2016).

[39] Milo, R. What is the total number of protein molecules per cell volume? A call to rethink some published values. *Bioessays* **35**, 1050–5 (2013).

[40] Alberts, B. *et al.* *Molecular Biology of the Cell* (Garland, 2002), 4th edn.

[41] Chia, L. L. & McLaughlin, C. The half-life of mRNA in *Saccharomyces cerevisiae*. *Mol Gen Genet* **170**, 137–44 (1979).

[42] Pelechano, V., Chávez, S. & Pérez-Ortín, J. E. A complete set of nascent transcription rates for yeast genes. *PLoS One* **5**, e15442 (2010).

[43] Futcher, B., Latter, G. I., Monardo, P., McLaughlin, C. S. & Garrels, J. I. A sampling of the yeast proteome. *Mol Cell Biol* **19**, 7357–68 (1999).

[44] Yang, E. *et al.* Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res* **13**, 1863–72 (2003).

[45] Milo, R., Jorgensen, P., Moran, U., Weber, G. & Springer, M. Bionumbers–the database of key numbers in molecular and cell biology. *Nucleic Acids Res* **38**, D750–3 (2010).

[46] Blevins, W. R. *et al.* Extensive post-transcriptional buffering of gene expression in the response to severe oxidative stress in baker's yeast. *Sci Rep* **9**, 11005 (2019).

[47] Hereford, L. M. & Rosbash, M. Number and distribution of polyadenylated RNA sequences in yeast. *Cell* **10**, 453–62 (1977).

[48] von der Haar, T. A quantitative estimation of the global translational activity in logarithmically growing yeast cells. *BMC Syst Biol* **2**, 87 (2008).

[49] Zenklusen, D., Larson, D. R. & Singer, R. H. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat Struct Mol Biol* **15**, 1263–71 (2008).

[50] Miura, F. *et al.* Absolute quantification of the budding yeast transcriptome by means of competitive PCR between genomic and complementary DNAs. *BMC Genomics* **9**, 574 (2008).

[51] van Leeuwen, J. *et al.* Systematic analysis of bypass suppression of essential genes. *Mol Syst Biol* **16**, e9828 (2020).

[52] Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

[53] Velculescu, V. E. *et al.* Analysis of human transcriptomes. *Nat Genet* **23**, 387–8 (1999).

[54] Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350**, 1096–101 (2015).

[55] Lahtvee, P.-J. *et al.* Absolute quantification of protein and mRNA abundances demonstrate variability in gene-specific translation efficiency in yeast. *Cell Syst* **4**, 495–504.e5 (2017).

[56] Hellton, K. H. & Thoresen, M. The impact of measurement error on principal component analysis. *Scandinavian Journal of Statistics* **41**, 1051–1063 (2014). URL https://doi.org/10.1111/sjos.12083.

[57] Gerdes, S. Y. *et al.* Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J Bacteriol* **185**, 5673–84 (2003).

[58] Mori, M. *et al.* From coarse to fine: the absolute *Escherichia coli* proteome under diverse growth conditions. *Mol Syst Biol* **17**, e9536 (2021).

CONTENTS


Introduction 1

Results 2

Discussion 8

References 10

A. Supplemental analysis 12
1. Gillespie Simulation of the telegraph model 12
2. Selection of central dogma parameter estimates 12
a. *E. coli* data 13
b. Yeast data 13
c. Human data 14
3. Quantitative estimates of central dogma parameters 14
a. Estimating the *cellular message number*: $\mu_{m/c}$ 14
b. Estimating the *transcription rate*: $\beta_m$ 14
c. Estimating the *message number*: $\mu_m$ 14
d. Estimating the *protein number*: $\mu_p$ 14
4. Empirical models for yeast gene expression 14
a. The meaning of the error estimates 15
b. Empirical model for protein number 15
c. Empirical model for message number 15
d. Empirical model for translation efficiency 15
e. Empirical model for the coefficient of variation 16
5. Supplemental analysis of gene expression noise 16
a. Inclusion of noise floor in the yeast analysis 16
b. Hypothesis test I 16
c. Hypothesis test II 17
d. Maximum likelihood estimates of the parameters 17
e. Statistical details MLE estimate of the variance 17
f. Detailed discussion of noise in *E. coli* 17


## Appendix A: Supplemental analysis

### 1. Gillespie Simulation of the telegraph model

Protein distributions of the telegraph model for *E. coli* were simulated with a Gillespie algorithm. Assuming the lifetime of the cell cycle ($T_{cc} = 30$ min) [36], mRNA lifetime ($\tau_m = 2.5$ min) [37], and translation rate ($\beta_p \approx 500$ hr$^{-1}$), the protein distributions for several mean expression levels were numerically generated for exponential growth with 100,000 stochastic cell divisions, with protein partitioned at division following the binomial distribution.

The gamma distributions for each mean message number with scale and shape parameters determined by the corresponding translation efficiency and message number ($\theta = \varepsilon \ln 2$, $k = \frac{\mu_m}{\ln 2}$) as used for the Gillespie simulation were also plotted with the protein distributions.

$$p(n|\theta, k) = \frac{1}{\Gamma(k)\theta^k} n^{k-1} e^{-\frac{n}{\theta}} \tag{A1}$$

### 2. Selection of central dogma parameter estimates

The estimates for central dogma model parameters come from two types of data: (i) quantitative measurement of cellular-scale parameters for each organism (total number of messages in the cell, cell cycle duration, *etc*) and (ii) genome-wide studies quantitative of mRNA and protein abundance.

For the cellular-scale central dogma parameters, we relied heavily on an online compilation of biological numbers: BioNumbers [45]. This resource provides a collection of curated quantitative estimates for biological numbers, as well as their original source. In the interest of conciseness, we have cited only the original source in the Tab. I, although we are extremely grateful and supportive of the creators of the BioNumbers website for helping us very efficiently identify consensus estimates for the parameters of the central dogma parameters.

For the selection of genome-wide studies on abundance, we used many of the same resources cited in BioNumbers as well as studies selected by a previous study of a quantitative analysis of the central dogma: Hausser *et al.* [5].

| Model organism | Growth condition | Doubling time: $T$ | Message lifetime: $\tau_m = \gamma_m^{-1}$ | Message recycling ratio: $m = T/\tau_m$ | Total number of messages /cell: $N_{m/c}^{\text{tot}}$ | Total number of messages /cell-cycle: $N_m^{\text{tot}}$ | Total number of proteins: $N_p^{\text{tot}}$ | Average translation efficiency: $\varepsilon$ | Average translation rate: $\beta_p$ ($\text{h}^{-1}$) |
|---|---|---|---|---|---|---|---|---|---|
| *Escherichia coli* (*E. coli*) | LB | 30 min [36] | 2.5 min [37] | 12 | $7.8 \times 10^3$ [38] | $9.4 \times 10^4$ | $3 \times 10^6$ [39] | 22 | 530 |
| | M9 | 90 min [36] | 2.5 min [37] | 36 | $2.4 \times 10^3$ [38] | $8.6 \times 10^4$ | $3 \times 10^6$ [39] | 24 | 580 |
| *Sacchromyces cerevisiae (Yeast–haploid)* | YEPD | 90 min [40] | 22 min [41] | 4 | $2.9 \times 10^4$ [42] | $1 \times 10^5$ | $5 \times 10^7$ [43] | $4 \times 10^2$ | 410 |
| *Homo sapiens (Human)* | Tissue | 24 h [35] | 14 h [44] | 1.7 | $3.6 \times 10^5$ [40] | $5 \times 10^5$ | $2 \times 10^9$ [39] | $4 \times 10^3$ | 120 |

TABLE III. **Central dogma parameters for three model organisms with detailed references.** Columns three through seven hold representative values for measured central-dogma parameters for the model organisms described in the paper. Each value is followed by a reference for its source. The last three columns hold estimates for the lower threshold on transcription inferred from our analysis. The sources of the numbers and estimates are described in the Supplemental Material Sec. A 2.

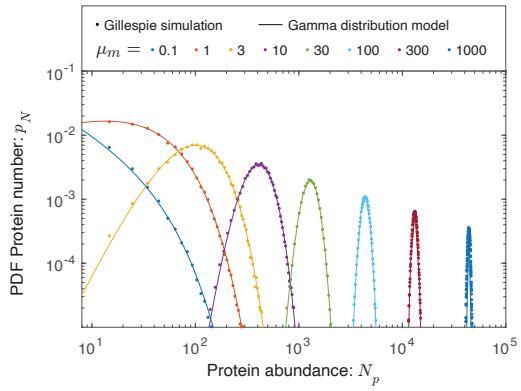


FIG. 8. **The protein abundance is approximately gamma distributed.** Protein abundance was modeled for eight different transcription rates using a Gillespie simulation, including the stochastic partitioning of the proteins between daughter cells at cell division. The range in abundance matches the observed range of expression levels in the cell. We observed that the simulated protein abundances were well fit by gamma distributions.

### *a. E. coli data*

**Message lifetimes:** The message lifetimes (and median lifetime) were taken from a recent transcriptome-wide study by Chen *et al.* [37]. These investigators measured the lifetime in both rapid (LB) and slow growth (M9).

**Noise:** Taniguchi *et al.* have performed a beautiful simultaneous study of the proteome and transcriptome with single-molecule sensitivity [15]. Although we use the noise analysis data from this study for our supplemental analysis of *E. coli* noise, it is not the source for our *E. coli* transcriptome data due to the extremely slow growth of the cells in this study (150 minute doubling time), which is not consistent with the growth conditions for the other sources of data.

**mRNA abundance:** Instead, we used data from the more recent Bartholomaus *et al.* study [38], which characterizes the transcriptome in both rapid (LB) and slow growth (M9).

**Total cellular message number.** This study was chosen since it was the source of the BioNumbers estimates of cellular message number in *E. coli* (BNID 112795 [45]).

**Doubling time:** The source of the doubling times for rapid (LB) and slow (M9) growth of *E. coli* comes from Bernstein [36].

**Essential gene classification.** The classification of essential genes in yeast comes from the construction of the Keio knockout collection from Baba *et al.* [26].

**Protein number.** The total protein number in *E. coli* came from Milo's recent review of this subject [39].

### *b. Yeast data*

**Message lifetimes:** The message lifetimes (and median lifetime) were taken from Chia *et al.* [41].

**Noise:** The noise data was taken from the Newman *et al.* study, which used flow cytometry of a library of fluorescent fusions to characterize protein abundance with single-cell resolution [12].

**mRNA abundance:** The transcriptome data comes from the very recent Blevins *et al.* study [46].

**Total cellular message number.** There are a wide-range of estimates for the total cellular message number in yeast: $1.5 \times 10^4$ [47] (BNID 104312 [45]), $1.2 \times 10^4$ [48] (BNID 102988 [45]), $6.0 \times 10^4$ [49] (BNID 103023 [45]), $2.6 \times 10^4$ [42] (BNID 106763 [45]) and $3.0 \times 10^4$ [50]. We used the compromise value of $2.9 \times 10^4$.

**Doubling time:** The doubling time was taken from [40].

**Protein number.** The total protein number in yeast comes from Futcher *et al.* [43].

**Essential gene classification.** The classification of essential genes in yeast comes from van Leeuwen *et al.* [51].

**Proteome abundance data:** The proteome abundance data came from two sources: flow cytometry of fluorescent fusions from Newman *et al.* [12] as well as mass-spec data from de Godoy *et al.* [24].

#### c. Human data

**Message lifetimes:** The message lifetimes (and median lifetime) were taken from Yang *et al.* [44] who reported a median half life of 10 h which corresponds to a lifetime of 14 h.

**mRNA abundance:** The transcriptome data comes from the data compiled by the Human Protein Atlas [52], which we averaged over tissue types.

**Total cellular message number.** The total cellular message number in human comes from Velculescu *et al.* [53] (BNID 104330 [45]).

**Doubling time:** The doubling time was taken from [35].

**Protein number.** The total protein number in human came from Milo's recent review of this subject [39].

**Essential gene classification.** The classification of essential genes in human comes from Wang *et al.* [54].

### 3. Quantitative estimates of central dogma parameters

#### a. Estimating the cellular message number: $\mu_{m/c}$

For each model organism (and condition), we found a consensus estimate from the literature for the total number of mRNA messages per cell $N^{\text{tot}}_{m/c}$. This number and its source are provided in Tab. I. To estimate the number of messages corresponding to gene $i$, we re-scaled the un-normalized abundance level $r_i$:

$$N_{m/c,i} = N^{\text{tot}}_{m/c} \frac{r_i}{\sum_j r_j}, \tag{A2}$$

where the sum over gene index $j$ runs over all genes.

#### b. Estimating the transcription rate: $\beta_m$

To estimate the transcription rate for gene $i$, we start from the estimated cellular message number $N_{m/c,i}$ and use the telegraph model prediction for the cellular message number:

$$N_{m/c,i} = \beta_{m,i}/\gamma_{m,i}, \tag{A3}$$

where $\gamma_{m,i}$ is the message decay rate. Since gene-to-gene variation in message number is dominated by the transcription rate (*e.g* [37]), we estimate the decay rate as the inverse gene-median message life time:

$$\gamma_{m,i} = \tau_m^{-1}, \tag{A4}$$

for which a consensus value was found from the literature. This number and its source are provided in Tab. I. We then estimate the gene-specific transcription rate:

$$\beta_{m,i} = N_{m/c,i}/\tau_m. \tag{A5}$$

#### c. Estimating the message number: $\mu_m$

To estimate the message number of gene $i$, we use the predicted value from the telegraph model:

$$N_{m,i} = T\beta_{m,i} = \tfrac{T}{\tau_m} N_{m/c,i}, \tag{A6}$$

where $T$ is the doubling time and $N_{m/c,i}$ is the cellular message number (Eq. A2).

#### d. Estimating the protein number: $\mu_p$

The protein abundance data for yeast grown in YEPD media and measured with flow cytometry fluorescence [12] were given in arbitrary units (AU). In order to convert from AU to protein number, the fluorescence values were rescaled by comparing with mass-spectrometry protein abundance data for yeast grown in YNB media [24]. Since the protein abundance from mass-spectrometry was given in terms of Intensity, the Intensity values were first rescaled by the total number of proteins in yeast, $5 \times 10^7$. The mass-spectrometry protein data was thresholded at 10 proteins, based on the assumption that the noise of the data for 10 and fewer proteins makes the data unreliable. Next, the log of the fluorescence protein abundance in AU as a function of the log of thresholded mass-spectrometry protein abundance was fit as a linear function with an assumed slope of 1 to find the offset, 3.9, (Fig. 9) which corresponds to a multiplicative scaling factor (Eqn. A8). We then used that offset value to rescale the fluorescence data from AU to protein number. We also compared to yeast grown in SD media [12] and found a similar offset result.

$$\log \mu_P^{\text{F}} = m \log \mu_P^{\text{MS}} + b \tag{A7}$$

$$\mu_P^{\text{F}} = b(\mu_P^{\text{MS}})^m \tag{A8}$$

### 4. Empirical models for yeast gene expression

To generate the empirical model for protein number as a function of message number, we used protein abundance data from Newman *et al.* [12], re-scaled to estimate protein number (Sec. A 3 d) and transcriptome

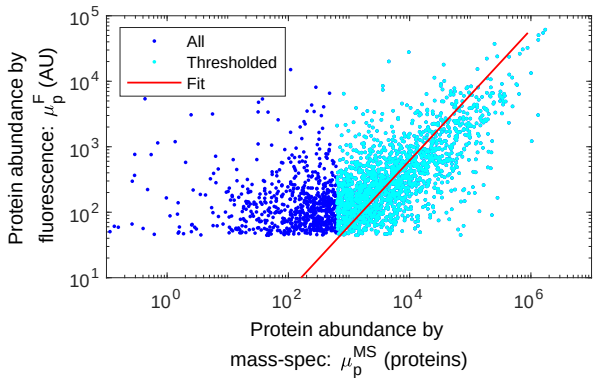


FIG. 9. **Fit to rescale fluorescence intensity to protein number.** Protein abundance from flow cytometry fluorescence [12] as a function of mass-spectrometry scaled abundance [24]. The mass-spectrometry data was thresholded at 10 proteins, and then a linear fit was performed to find the offset of 3.9, which was used to convert protein fluorescence AU to number.

data from Lahtvee *et al.* [55], re-scaled to estimate message number (Sec. A 3 c).

### a. *The meaning of the error estimates*

Before providing a detailed error analysis, it is important to place our error estimates in perspective. The error that we will be estimating is the statistical error associated with the finite sample size; however, *this is not the dominant source of error*. A far more important consideration are systematic problems with our analysis and the underlying experiments. For instance, since we do not have a detailed model for the error of the experiments analyzed, there are multiple distinct analyses (*i.e.* assumptions about the error model) that could be implemented for the data fitting, each giving slightly different model parameters. These model to model differences still give rise to predictions consistent with our qualitative conclusions; however, they are likely larger than the statistical uncertainty we compute (while assuming a particular model).

### b. *Empirical model for protein number*

We initially fit the empirical model for protein number,

$$\mu_p = C_0 \mu_m^{\alpha_0}, \tag{A9}$$

to the data using a standard least-squares approach; however, the algorithm led to a very poor fit since it does not account for uncertainty in both independent and dependent variables. We therefore used an alternative approach [56], which assumes comparable error in both variables. The model parameters are:

$$\alpha_0 = 2.1 \pm 0.04, \tag{A10}$$
$$C_0 = 8.0 \pm 1.0, \tag{A11}$$

where the uncertainties are the estimated standard errors.

### c. *Empirical model for message number*

For the prediction of the coefficient of variation, it is useful to invert Eq. A9 to generate a model for message number as a function of protein number:

$$\mu_m = C_0^{-1/\alpha_0} \mu_p^{1/\alpha 0}, \tag{A12}$$
$$= C_1 \mu_p^{\alpha_1}, \tag{A13}$$

where the last line defines two new parameters: a coefficient $C_1$ and an exponent $\alpha_1$. The resulting parameters and uncertainties are:

$$\alpha_1 \equiv 1/\alpha_0, \tag{A14}$$
$$= 0.48 \pm 0.01, \tag{A15}$$
$$C_1 \equiv C_0^{-1/\alpha_0}, \tag{A16}$$
$$= 0.37 \pm 0.02, \tag{A17}$$

where the uncertainties are the estimated standard errors.

### d. *Empirical model for translation efficiency*

To generate an empirical model for translation efficiency, we started from the empirical model for protein number (Eq. A9), and then use Eq. 11 to relate protein number, message number, and translation efficiency:

$$\varepsilon = \frac{\mu_p}{\mu_m}, \tag{A18}$$
$$= C_0 \mu_m^{\alpha_0 - 1}, \tag{A19}$$
$$= C_2 \mu_m^{\alpha_2}, \tag{A20}$$

where the last line defines two new parameters: a coefficient $C_2$ and an exponent $\alpha_2$. The resulting parameters and uncertainties are:

$$\alpha_2 = \alpha_0 - 1, \tag{A21}$$
$$= 1.07 \pm 0.04, \tag{A22}$$
$$C_2 = C_0, \tag{A23}$$
$$= 8.0 \pm 1.0, \tag{A24}$$

where the uncertainties are the estimated standard errors.

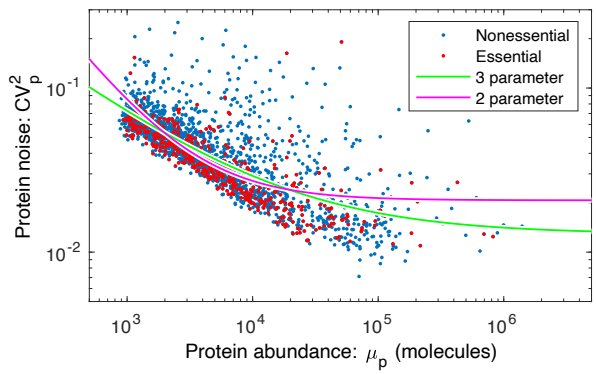


FIG. 10. **Yeast noise fit against canonical noise model, with a noise floor.** Yeast noise data fit with the 2- (null hypothesis with $\mu_p^{-1}$ dependence) and 3- parameter ($\mu_p^a$) models.

#### e. *Empirical model for the coefficient of variation*

To generate an empirical model for the coefficient of variation, we started from the empirical model for message number (Eq. A13), and then substitute this into the statistical model prediction for $\mathrm{CV}_p^2$ (Eq. 12):

$$\begin{aligned} \mathrm{CV}_p^2 &= \tfrac{\log 2}{\mu_m}, && \text{(A25)} \\ &= C_0^{1/\alpha_0} \log 2 \cdot \mu_p^{-1/\alpha_0}, && \text{(A26)} \\ &= C_3 \mu_p^{\alpha_3}, && \text{(A27)} \end{aligned}$$

where the last line defines two new parameters: a coefficient $C_3$ and an exponent $\alpha_3$. The resulting parameters and uncertainties are:

$$\begin{aligned} \alpha_3 &\equiv -1/\alpha_0, && \text{(A28)} \\ &= -0.48 \pm 0.01, && \text{(A29)} \\ C_3 &\equiv C_0^{1/\alpha_0} \log 2, && \text{(A30)} \\ &= 1.9 \pm 0.1, && \text{(A31)} \end{aligned}$$

where the uncertainties are the estimated standard errors.

### 5. Supplemental analysis of gene expression noise

The quantitative model for gene expression noise includes multiple contributions:

$$\mathrm{CV}_p^2 \approx \tfrac{1}{\mu_p} + \tfrac{\log 2}{\mu_m} + c_0, \qquad \text{(A32)}$$

where the first term can be understood to represent the Poisson noise from translation, the second term the Poisson noise from transcription, and the last term, $c_0$, is called the *noise floor* and is believed to be caused by the cell-to-cell variation in metabolites, ribosomes, and polymerases *etc* [13, 14].

#### a. *Inclusion of noise floor in the yeast analysis*

In the main text of the paper, we have ignored the role of the noise floor in the analysis of noise in yeast. Unlike *E. coli*, where the noise floor is high ($\mathrm{CV}_p^2 = 0.1$) and is determinative of the noise associated with almost all essential genes [13–15], in yeast the noise floor is much lower ($\mathrm{CV}_p^2 = 0.01$) and therefore affects only genes with the highest expression.

In this section, we will consider models that include the noise floor, since its presence can make the noise scaling more difficult to interpret. To determine if the scaling of the noise is consistent with the canonical assumption that the noise is proportional to $\mu_p^{-1}$ for low expression, we will consider two competing empirical models for the noise (Fig. 10). In the null hypothesis, we will consider a model:

$$\eta_0(\mu_p; b, c) = \tfrac{b}{\mu_p} + c, \qquad \text{(A33)}$$

and an alternative hypothesis with an extra exponent parameter $a$:

$$\eta_1(\mu_p; a, b, c) = \tfrac{b}{\mu_p^a} + c. \qquad \text{(A34)}$$

We will assume that $\mathrm{CV}_p^2$ is normally distributed about $\eta$ with unknown variance $\sigma_\eta^2$.

In this context, a maximum likelihood analysis is equivalent to least-squares analysis. Let the sum of the squares be defined:

$$S_I(\boldsymbol{\theta}) \equiv \sum_i [\mathrm{CV}_{p,i}^2 - \eta_I(\mu_{p,i}; \boldsymbol{\theta})]^2 \qquad \text{(A35)}$$

for model $I$ where $\boldsymbol{\theta}$ represents the parameter vector. The maximum likelihood parameters are

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} S_I(\boldsymbol{\theta}), \qquad \text{(A36)}$$

with residual norm:

$$\hat{S}_I = S_I(\hat{\boldsymbol{\theta}}). \qquad \text{(A37)}$$

To test the null hypothesis, we will use the canonical likelihood ratio test with the test statistic:

$$\Lambda \equiv 2 \log \frac{q_1}{q_0}, \qquad \text{(A38)}$$

where $q_0$ and $q_1$ are the likelihoods of the null and alternative hypotheses, respectively. Wilks' theorem states that $\Lambda$ has a chi-squared distribution of dimension equal to the difference of the dimension of the alternative and null hypotheses ($3 - 2 = 1$).

#### b. *Hypothesis test I*

In our first analysis, we will estimate the variance directly. We computed the mean-squared difference for

successive $\mathrm{CV}_p^2$ values, sorted by mean protein number $\mu_p$. The variance estimator is

$$\hat{\sigma}_\eta^2 = \tfrac{1}{2}\left\langle (\mathrm{CV}_{p,i}^2 - \mathrm{CV}_{p,i+1}^2)^2 \right\rangle_i = 6.3 \times 10^{-4}, \quad \text{(A39)}$$

where the brackets represent a standard empirical average over gene $i$ for the $\mu_p$-ordered gene $\mathrm{CV}_p^2$ values. The test statistic can now be expressed in terms of the residual norms:

$$\begin{aligned} \Lambda &= (\hat{S}_1 - \hat{S}_2)/\hat{\sigma}_\eta^2, & \text{(A40)} \\ &= 3.3 \times 10^4, & \text{(A41)} \end{aligned}$$

which corresponds to a p-value far below machine precision. We can therefore reject the null hypothesis.

### c. Hypothesis test II

In a more conservative approach, we can use maximum likelihood estimation to estimate the variance of each model independently as a model parameter. In this case, the test statistic can again be expressed in terms of the residual norms:

$$\begin{aligned} \Lambda &= N \log \tfrac{\hat{S}_1}{\hat{S}_2}, & \text{(A42)} \\ &= 1.6 \times 10^2, & \text{(A43)} \end{aligned}$$

where $N$ is the number of data points. In this case, the p-value can be computed assuming the Wilks' theorem (*i.e.* the chi-squared test):

$$p = 6 \times 10^{-36}, \quad \text{(A44)}$$

again, strongly rejecting the null hypothesis.

### d. Maximum likelihood estimates of the parameters

In the alternative hypothesis, the maximum likelihood estimate (MLE) of the empirical noise model (Eq. 5) parameters are (Fig. 10):

$$\begin{aligned} a &= 0.57 \pm 0.02, & \text{(A45)} \\ b &= 3.0 \pm 0.5, & \text{(A46)} \\ c &= 0.013 \pm 0.001, & \text{(A47)} \end{aligned}$$

where the parameter uncertainty has been estimated using the Fisher Information in the usual way using the MLE estimate of the variance.

The noise model parameters were also determined for *E. coli*:

$$\begin{aligned} a &= 1.22 \pm 0.01, & \text{(A48)} \\ b &= 1.27 \pm 0.02, & \text{(A49)} \\ c &= 0.154 \pm 0.002, & \text{(A50)} \end{aligned}$$

with the corresponding fit shown in Fig. 11. Since $a$ is close to 1, the canonical model with $a = 1$ (Eqn. 3) is a somewhat reasonable approximation for the noise in *E. coli*.

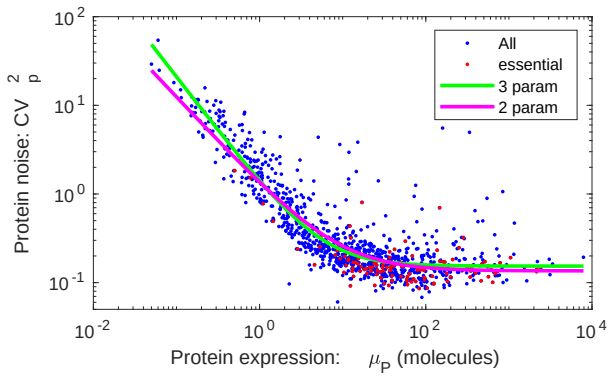


FIG. 11. **Three-parameter fit to *E. coli* noise.** The noise as a function of protein abundance from Taniguichi *et al.* was fit to the 3 parameter noise model (Eqn. 5). From the fit, protein noise scales proportionally with $\mu_p^{-1.22}$, which is a close result to the canonical model with $\mu_p^{-1}$.

### e. Statistical details MLE estimate of the variance

The minus-log-likelihood for the normal model $I$ is:

$$h_I(\hat{\boldsymbol{\theta}}, \sigma^2) = \tfrac{N}{2} \log 2\pi\sigma^2 + \tfrac{1}{2\sigma^2}\hat{S}_I, \quad \text{(A51)}$$

where $\hat{S}_I$ is the least-square residual. We then minimize $h_I$ with respect to the variance $\sigma^2$:

$$\partial_{\sigma^2} h|_{\hat{\sigma}^2} = 0, \quad \text{(A52)}$$

to solve for the MLE $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \tfrac{1}{N}\hat{S}_I. \quad \text{(A53)}$$

Next we evaluate $h$ at the variance estimator:

$$h_I(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2) = \tfrac{N}{2}\left[\log 2\pi \tfrac{\hat{S}_I}{N} + 1\right]. \quad \text{(A54)}$$

The test statistics can be written in terms of the $h$'s:

$$\begin{aligned} \Lambda &= 2h_0(h_1(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2) - 2h_2(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2), & \text{(A55)} \\ &= N \log \tfrac{\hat{S}_1}{\hat{S}_1}, & \text{(A56)} \end{aligned}$$

which can be evaluated directly in terms of the residual norms for the null and alternative hypotheses.

### f. Detailed discussion of noise in E. coli

In general, the Telegraph model predicts that the noise will have a coefficient of variation [3, 9]:

$$\mathrm{CV}_p^2 \approx \frac{1}{\mu_p} + \frac{\varepsilon \ln 2}{\mu_p}, \quad \text{(A57)}$$

where the first term is significant whenever the translation efficiency isn't $\varepsilon \gg 1$. In both *E. coli* ($\varepsilon \approx 30$) and

| Gene name | Message number: $\mu_m$ | Annotated function from Ecocyc | Essential (E)/ Nonessential (N) Ref. [26], [57], [28] |
|---|---|---|---|
| *alsK* | 0.3 | The alsK gene encodes a D-allose kinase. Its role in the degradation of D-allose is unclear; AlsK is not required for utilization of a D-allose carbon source; this effect may be due to the presence of other ambiguous sugar kinases within *E. coli* K-12. | E, N, N |
| *bcsB* | 0.4 | BcsB is encoded in a predicted operon together with *bcsA*, *bcsZ* and *bcsC*. In other organisms, these genes are involved in cellulose biosynthesis, a characteristic of the rdar (red, dry and rough) morphotype. However, the K-12 laboratory strain of *E. coli* does not show a rdar morphotype and does not produce cellulose. | E, N, N |
| *entD* | 0.4 | AcpS is the founding member of a 4'-phosphopantetheinyl (P-pant) transferase protein family that includes *E. coli* EntD, *E. coli* o195 protein, and *Bacillus subtilis* Sfp; family members share two conserved motifs but relatively low sequence identity overall. | E, N, N |
| *yafF* | 0.4 | No information about this protein was found by a literature search conducted on April 19, 2017. | E,-, N |
| *yagG* | 0.6 | *yagGH* is predicted to be a member of the XylR regulon; its products may mediate transport (YagG) and hydrolysis (YagH) of xylooligosaccharides; putative XylR and CRP binding sites are identified upstream of *yagGH*. | E,-, N |
| *yceQ* | 0.2 | No information about this protein was found by a literature search conducted on July 12, 2017. | E, E, N |
| *ydiL* | 0.2 | No information about this protein was found by a literature search conducted on April 7, 2017. | E, N, N |
| *yhhQ* | 0.4 | YhhQ is an inner membrane protein implicated in the uptake of queuosine (Q) precursors - 7-cyano-7-deazaguanine (preQ0) and 7-aminomethyl-7-deazaguanine (*preQ1*) - for Q salvage. Q-modified tRNA is absent in Δ*queD* and Δ*queD* Δ*yhhQ* strains grown in minimal media with glycerol; Q-modified tRNA is detected when a Δ*queD* strain is grown in minimal media plus 10 nM 0*preQ0* or *preQ1* but is absent when a Δ*queD* Δ*yhhQ* strain is grown under these conditions. *yhhQ* expressed from a plasmid restores the presence of Q-modified tRNA in a Δ*queD* ΔextityhhQ strain. | E,-, N |
| *yibJ* | 0.3 | No information about this protein was found by a literature search conducted on July 9, 2018. | E, N, N |
| *ymfK* | 0.4 | YmfK is a component of the relic lambdoid prophage e14 and is likely the SOS-sensitive repressor. It is similar to the P34 gene of the Shigella flexneri bacteriophage SfV and belongs to the LexA group of SOS-response transcriptional repressors. | E, E, E |

TABLE IV. **Below-threshold essential genes identified in *E. coli*.** This table describes the message numbers and annotations for essential genes that we estimated to have expression below the threshold of one message per cell cycle. However, in the final column, we show classifications from three different studies. Only one of the identified genes, *ymfK*, was consistently defined as essential.

yeast ($\varepsilon \approx 420$), this would seem naively to be the case. However, since translation efficiency in yeast is not uniform, we must consider its variation for low-expression proteins. We estimate that the detection efficiency in yeast is roughly $10^3$ molecules. Using Eq. 15, we estimate that $\varepsilon \approx 100$ and our approximation holds at the low-expression detection limit.

In *E. coli*, the situation is somewhat more complicated. Unlike yeast, the translation efficiency is roughly constant (at high to intermediate expression levels) with respect to expression level [25], and therefore both terms in Eq. A57 are expected to scale like the canonical model ($\propto \mu_p^{-1}$). However, it is clear that the translation efficiency must significantly decrease for the lowest abundance proteins. This is visible even in Ref. [25] Fig. 1B, where the data falls below the predicted protein abundance at low message number. Note that these mass-spec measurements are not as sensitive as fluorescence-based measurements (*e.g.* only 64% proteome could be detected [58]). Furthermore, fits to the *E. coli* noise (Eq. A49) are consistent only with low values of $\varepsilon$. At sufficiently high expression levels such that we are confident about the translation efficiency, the noise is already very close to the noise floor.