Title: Noise robustness and metabolic load determine the principles of central dogma regulation

Authors: Teresa W. Lo,[1,*] Han 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*

Abstract: The processes of gene expression are inherently stochastic, even for essential genes required for growth. How does the cell maximize fitness in light of noise? To answer this question, we build a mathematical model to explore the trade-off between metabolic load and growth robustness. The model predicts novel principles of central dogma regulation: Optimal protein expression levels for many genes are in vast overabundance. Essential genes are transcribed above a lower limit of one message per cell cycle. Gene expression is achieved by load balancing between transcription and translation. We present evidence that each of these novel regulatory principles is observed. These results reveal that robustness and metabolic load determine the global regulatory principles that govern central dogma processes, and these principles have broad implications for cellular function.

One-sentence summary: Fitness maximization predicts protein overabundance, a transcriptional floor, and the balancing of transcription and translation.

INTRODUCTION

What rationale determines the optimal transcription and translation level of a gene in the cell? Protein expression levels optimize cell fitness [1, 2]: Too low of an expression level of essential proteins slows growth by compromising the function of essential processes [3, 4], whereas the overexpression of proteins slows growth by increasing the metabolic load [5]. This trade-off naïvely predicts that the cell maximizes its fitness by a Goldilocks principle in which cells express just enough protein for function [6]; however, achieving growth robustness is nontrivial, since all processes at the cellular scale are stochastic, including gene expression [7]. This biological noise leads to significant cell-to-cell variation in protein numbers, even for essential proteins that are required for growth [8, 9]. The optimal expression program must therefore ensure robust expression of hundreds of distinct essential gene products. In this paper, we explore the consequences of growth robustness on the central dogma regulatory program.

RESULTS

Defining the RLTO Model. To study the consequences of growth robustness on the central dogma quantitatively, we propose and analyze a minimal model: the Robustness-Load Trade-Off (RLTO) Model. The model includes three critical components: (i) Protein levels are stochastic and the single-cell growth rate depends upon them, (ii) gene transcription and translation generate a metabolic load, and (iii) cell growth is dependent on a large number of essential genes. These model characteristics result in a highly-asymmetric fitness landscape. The optimization of expression on this asymmetric landscape predicts new phenomenology absent from previous models (*e.g.* [10]).

The protein number N_p expressed from gene i is the product of two sequential stochastic processes: transcription and translation [11], leading to cell-to-cell variation in protein number, which we will refer to as *noise*. In our analysis, we will model gene expression using the canonical equilibrium Paulsson model [12]. (See Fig. 1A.) In this model, the numbers of proteins N_p for gene i are predicted to be gamma-distributed [13], in close agreement with observation [9]. The distribution is described by two gene-specific statistical parameters: *message number* (μ_m), defined as the mean number of messages transcribed per cell cycle for gene i, and the *translation efficiency* (ε), the mean number of proteins translated from each message transcribed for gene i. The mean protein abundance is their product: $\mu_p = \mu_m \varepsilon$. These parameters can be expressed in terms of ratios of the rates of the underlying central dogma processes, as described in Supplementary Material Sec. A 1.

How should the effect of essential protein expression on growth rate be modeled in the context of the RLTO model? Much recent work has focused on cellular resource allocation to functional sectors (*e.g.* [14]). In this approach, an optimization is performed by coherently modulating the abundance of all protein in a particular sector, leading to a trade-off between functional capacities of the cell. However, in the RLTO model, the optimization is fundamentally different: We consider the incoherent variation in the abundance of protein species i due to noise. For these incoherent changes, we generically expect proteins to exhibit rate-limited kinetics: Increases in the protein number N_p above a threshold level n_p has minimal effect on the rate since other chemi-

* These two authors contributed equally
† pwiggins@uw.edu

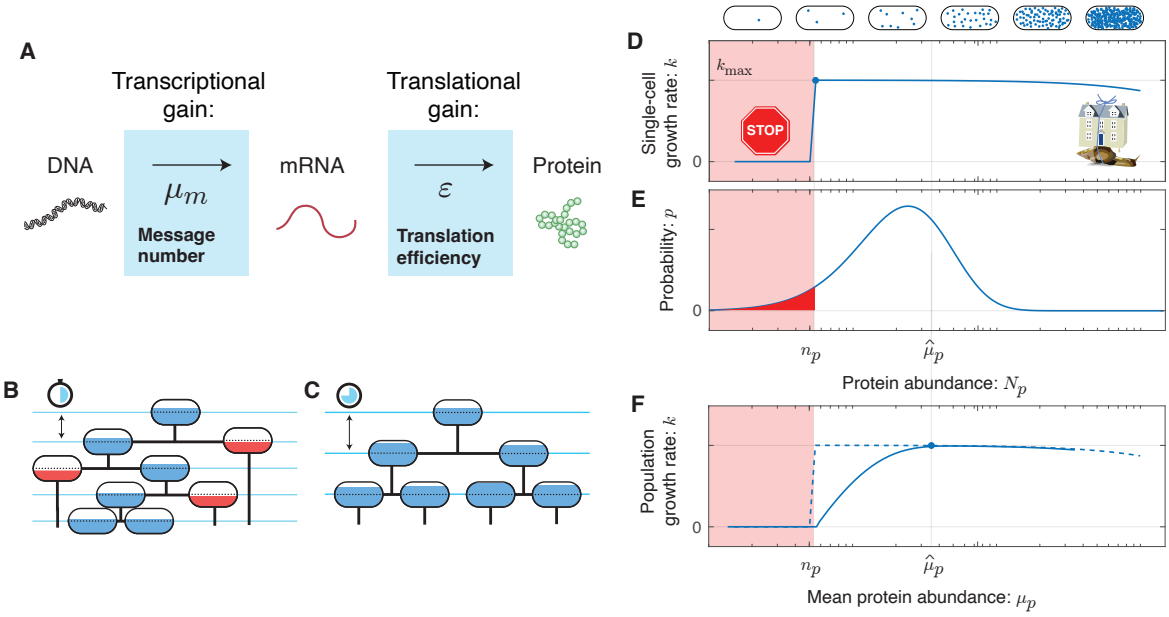

FIG. 1. **The RLTO Model. Panel A: Gene expression is stochastic.** The central dogma describes a two-stage stochastic process in which genes are first transcribed and then translated. The transcription process transcribes an average of μ_m messages per cell cycle. The translation process translates an average of ε proteins per message. **Panel B & C:** A schematic cell lineage tree is shown during exponential growth. For a specific protein i, the cell fill represents the protein number N_p relative to its threshold number n_p required for cell growth. **Panel B:** Reducing the mean expression level reduces doubling time; however, stochasticity in expression results in below-threshold cells (red fill) which grow slowly. **Panel C:** Increasing protein expression increases the doubling time; however, all cells are above threshold (blue fill). **Panel D: The fitness landscape as a function of protein number.** Growth arrests for protein number N_p smaller than the threshold level n_p (red) due to the failure of essential processes. High expression levels are penalized due to the metabolic cost of protein expression. This trade-off leads to a highly asymmetric fitness landscape: The relative metabolic cost of overabundance is small relative to the cost of growth arrest due to the large size of the total metabolic load N_0. **Panel E: The gene expression process is stochastic.** There is significant cell-to-cell variation in protein abundance (N_p) around the mean level (μ_p). Even for mean expression levels significantly above the threshold level n_p, some cells fall below threshold (red). The distribution in protein number is modeled using a gamma distribution [9], parameterized by message number μ_m and translation efficiency ε. **Panel F: The robustness load trade-off determine the optimal expression level.** The population growth rate depends on the distribution of the protein number. The asymmetry of the fitness landscape drives the optimal expression level far above the threshold level due to the high fitness cost of low protein abundance.

cal species (proteins, metabolites, *etc.*) are rate limiting [15]. However, if the protein number N_p falls below the threshold n_p, then protein species i becomes rate limiting and leads to a significant slowdown of the growth rate. In the RLTO model, we coarse-grain the details of this growth slowdown as growth arrest. (See Fig. 1BCD.) There is already some precedent for the use of this type of threshold (*e.g.* [16]), but we will demonstrate that the detailed form of the fitness landscape is not important. (See Supplementary Material Sec. B.) Although sufficiently detailed knowledge of the relevant molecular and cellular biology could be used to predict the protein thresholds n_p, we will treat these as gene-specific unknown parameters.

As shown in Materials and Methods, the relative cellular fitness with respect to the expression of gene i can be computed by combining the fitness losses associated with robustness (Eq. 6) and metabolic load (Eq. 8):

$$\frac{\Delta k}{k_0} = -(\Lambda + \tfrac{\varepsilon}{N_0})\mu_m - \tfrac{1}{\ln 2}\gamma(\tfrac{\mu_m}{\ln 2}, \tfrac{n_p}{\varepsilon \ln 2}), \qquad (1)$$

where the first term represents fitness loss due to metabolic load of transcription and translation while the second term represents loss due to the arrest of essential processes. In summary, the model depends only on a single global parameter: the relative metabolic load Λ and three gene-specific parameters: the threshold number n_p, the message number μ_m and the relative translation efficiency ε/N_0. We propose that the cell is regulated to maximize the growth rate with respect to transcription (message number) and translation (translation efficiency). The fitness landscape predicted by the RLTO model for representative parameters is shown in Fig. 2A.

RLTO predicts protein overabundance. The optimal regulatory program (μ_m and ε values) can be predicted analytically. They depend on only a single global parameter, the relative load Λ, and the gene-specific threshold number n_p. Since the threshold number is not directly observable experimentally, we will instead pre-

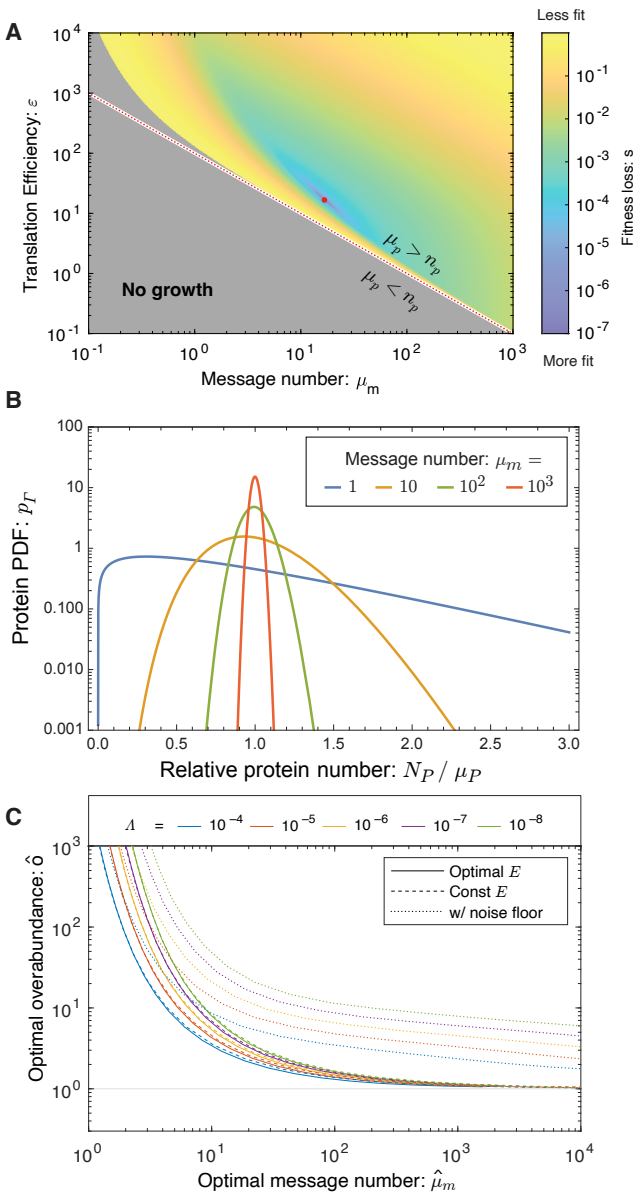

FIG. 2. **The RLTO model predicts overabundance is optimal. Panel A: Fitness landscape determines optimal message number and translation efficiency.** The fitness loss ($s \equiv \ln k_{\max}/k$) is shown as a function of message number (μ_m) and translation efficiency (ε). The red dotted curve represents programs where the mean protein number is equal to the threshold ($\mu_p = n_p$) and the red dot represents the optimal regulatory program ($\hat{\mu}_m$, $\hat{\varepsilon}$). **Panel B: Gene-expression noise.** Due to the stochasticity of the central dogma processes at equilibrium, the protein number N_p is gamma-distributed [12]. For high-expression genes, expression has low noise and the protein number is tightly distributed about its mean; however, for low-expression genes, expression is noisy and the distribution is extremely wide. **Panel C: Overabundance is optimal for all genes.** For high-expression genes, low overabundance is optimal ($\mu_p \approx n_p$); however, for low-expression genes, vast overabundance is optimal ($\mu_p \gg n_p$). From a quantitative perspective, overabundance depends on the relative load Λ; however, the qualitative dependence is invariant to over an orders-of-magnitude range of values.

dict the optimal overabundance o, defined as the ratio of the mean protein number to the threshold number:

$$o \equiv \mu_p / n_p. \tag{2}$$

As shown in Fig. 2C, The RLTO model generically predicts that the optimal protein fraction is overabundant ($o > 1$); however, the overabundance is not uniform for all proteins. For highly-transcribed genes ($\mu_m \gg 1$) like ribosomal genes, the overabundance is predicted to be quite small ($o \approx 1$); however, for message numbers approaching unity, the overabundance is predicted to be extremely high ($o \gg 1$). At a quantitative level, the relation between optimal overabundance and message number depends on the relative load (Λ), but its phenomenology is qualitatively unchanged over orders of magnitude variation in Λ.

Understanding the rationale for overabundance. To explore both the robustness of the protein overabundance prediction and to understand its mathematical rationale, we explored a collection of more complex models numerically. (Supplementary Material Sec. B.) The key mathematical feature that drives overabundance is not the assumption of growth arrest, but rather the strong asymmetry of the fitness landscape: the high cost of protein *underabundance* and the low cost of protein *overabundance*. (See Fig. 1EF.) In the RLTO model, this asymmetry is parameterized by the relative load (Λ), defined as the relative metabolic cost of transcribing an additional message. Since we estimate that $\Lambda < 10^{-5}$, this cost is very low relative to the total metabolic cost of the cell, therefore we expect this asymmetry, and the prediction of the RLTO model, to be robust.

Overabundance is observed in a range of experiments. The RLTO Model predicts that all essential proteins are overabundant. In general, the RLTO model predicts that protein numbers have very significant robustness (*i.e.* buffering) to protein depletion. Although this result is potentially surprising, it is in fact consistent with many studies. For instance, Belliveau *et al.* have recently analyzed the abundance of a wide range of metabolic and other essential biological processes, and conclude that protein abundance appears to be in significant excess of what is required for function [6]. Likewise, CRISPRi approaches have facilitated the characterization of essential protein depletion. The qualitative results from these experiments are consistent with overabundance: Large-magnitude protein depletion is typically required to generate strong phenotypes [3, 17, 18]. In particular, Peters *et al.* engineered a complete collection of CRISPRi essential-gene depletion constructs in *Bacillus subtilis*. Importantly, when *dcas9* is constitutively expressed, these constructs deplete essential proteins about three-fold below their endogenous expression levels [3]; however, roughly 80% grew without measurable fitness loss in log-phase growth despite the depletion. When grouped by functional category, only

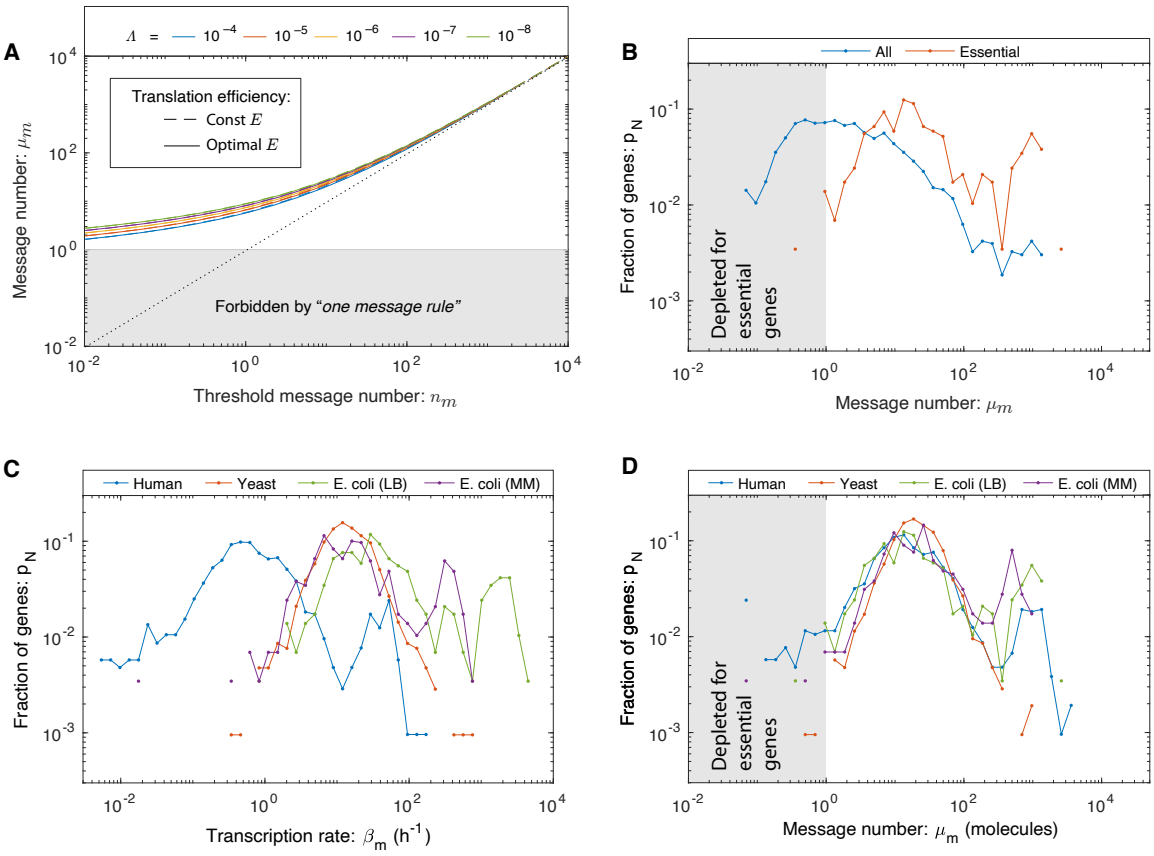

FIG. 3. **A lower threshold for transcription: The one message rule. Panel A: RLTO predicts the one-message rule.** For high-expression genes, overabundance is low and the message number μ_m is predicted to be comparable to the threshold level n_m (dotted line); however, for low-expression genes there is a lower threshold ($\mu_m \geq 1$) below which expression is too noisy for robust growth. The threshold is weakly dependent on relative load Λ. **Panel B: A one-message threshold is observed in *E. coli* for essential genes.** A histogram shows the distribution of gene message numbers for all genes (blue) versus essential genes (orange). As predicted by the RLTO model, virtually all essential genes are expressed above the one-message-per-cell-cycle threshold. **Panel C: The distribution of transcription rates for essential genes.** No alignment is observed between the distributions of transcription rates in three evolutionarily-divergent organisms. For instance, the per gene transcription rate is significantly lower in human cells relative to *E. coli*. **Panel D: The distribution of message numbers for essential genes in three evolutionarily-divergent organisms.** The alignment of distributions of message number per gene between human, yeast, and *E. coli* (under two distinct growth conditions) reveals a nontrivial commonality between central dogma regulatory programs. We propose that the rationale for this alignment is the one-message rule that predicts that all essential genes must be expressed above one message per cell cycle. Both yeast and *E. coli* come very close to satisfying this proposed threshold; however, a greater proportion of genes in human break the one-message threshold. We speculate that this is due in part to the *ad hoc* nature of the essential-gene classification in the context of complex multicellular organisms.

ribosomal proteins were found to have statistically significant reductions in fitness [3]. As shown in Fig. 2C, the RLTO model predicts that all but the highest expression proteins are expected to show minimal fitness reductions in response to a three-fold depletion of essential enzymes. The optimality of protein overabundance explains the paradox of protein expression levels being simultaneously optimal [1] and in excess of what is required for function [3, 4, 6, 18]. Although this qualitative picture of essential protein overabundance is clear, there has yet to be a quantitative and detailed measurement of protein overabundance, and in particular, an analysis of the relationship between protein overabundance and message number.

RLTO predicts a one-message transcription threshold. The RLTO model predicts protein overabundance, but is there a clear transcriptional signature? To analyze this question, we define the message threshold $n_m \equiv \mu_m/o$. (This parameterization is convenient since it is independent of the translation efficiency.) We can then analyze the relation between optimal message number and threshold message number, as shown in Fig. 3A. The model predicts that even for genes that have extremely small threshold message numbers (*e.g.* $n_m = 10^{-2}$), the optimal message number stays above one message tran-

scribed per cell cycle. Qualitatively, expressing messages below this level is simply too noisy even for proteins needed at the lowest expression levels. (See the blue curve in Fig. 2B corresponding to the protein number distribution of $\mu_m = 1$.) The model therefore predicts a lower floor on transcription for essential genes of one message per cell cycle.

A lower threshold is observed for message number. To identify a putative transcriptional floor, we first analyzed the transcriptome in *Escherichia coli*. We hypothesize that cells must express essential genes above the one-message threshold for robust growth. The distinction between essential and nonessential genes is critical in this context, since nonessential genes can be inducibly expressed. For instance, in *E. coli*, the *lac* operon is repressed in the absence of lactose and therefore need not satisfy the one-message threshold. The transcriptional threshold is only hypothesized to apply to genes whose products are required to maintain cell fitness under the measured conditions.

We generated histograms for *E. coli* growing rapidly on rich media for these two classes of genes. (See Supplementary Material Sec. C.) The message numbers for *nonessential* genes are widely distributed, with a significant fraction of genes falling below the one-message threshold; however, only one *essential* gene is expressed below the one-message threshold (0.3% of essential genes). (See Fig. 3B.) The threshold is not sharp, but rather a smooth depletion relative to a median of 18 messages per cell cycle. This observation is consistent with the predictions of the RLTO model.

To further test this prediction, we then analyzed *E. coli* transcription under slow-growth conditions. Since these cells are less transcriptionally active, we hypothesized that this analysis would constitute a more stringent test of the one-message rule. (See Supplementary Material Sec. C.) To our surprise, although the transcription rate is indeed reduced in slow growth, the essential gene message numbers still satisfy the one-message rule (with a two gene exception, 0.7%), again consistent with the predictions of the RLTO model. (See Fig. 3D.)

Next, we analyzed eukaryotic transcriptomes in *Saccharomyces cerevisiae* (yeast) and *Homo sapiens* (human). (See Supplementary Material Sec. C.) For yeast, there is a well-defined notion of essential genes [19]. As predicted, yeast essential genes obey the one-message threshold (with two exceptions, 0.2%). (See Fig. 3D.) The interpretation is less clear-cut in human cells: An essential gene classification has been generated in the context of proliferation in cell culture [20]. In order to try to capture a generic picture, we average the human transcriptome of cell types. We find that the vast majority of essential genes obey the one-message rule; however, there are significantly more genes that break the rule (81 genes, 8%) than in the other organisms.

Message number distribution is conserved. To what extent is this human data consistent with the RLTO model? For human cells, our test of the one-message rule is too simplistic in two respects: (i) We ignore the significant transcriptional differences associated with distinct cell types and (ii) the essential gene classification itself is defined by the ability of mutants or knockdowns to proliferate in cell culture; in marked contrast to the *in vivo* context where cell proliferation is tightly regulated [20]. Due to these subtleties, we decided to take a complementary approach: We considered the distribution of three different transcriptional statistics for each gene: transcription rate, cellular message number, defined as the average number of messages instantaneously, and message number (μ_m), defined as the number of messages transcribed in a cell cycle. (See Supplementary Material Sec. C.) The RLTO model predicts a one-message threshold with respect to message number, but not the other two statistics. We therefore predict that the message number distributions in each organism (*E. coli*, yeast, and human) should align for low expression genes with respect to message number, but not for the other two transcriptional statistics. Consistent with the predictions of the RLTO model, there is a striking alignment of message number for essential genes between all three model organisms and growth conditions for message number. (See Fig. 3D.) This alignment is non-trivial: It is not observed with respect to other transcriptional statistics (Fig. 3C and Supplementary Material Fig. 3). The observed alignment is consistent with a conserved transcriptional regulatory program from *E. coli* to human, and we hypothesize that the rationale for this alignment is the abutment of the lowest essential-gene transcription levels in each organism against the one-message threshold.

Translation efficiency is predicted to increase with transcription. What does the RLTO model predict about how the cell should balance the gene expression process between transcription and translation? Minimizing transcription (at fixed protein abundance) reduces the metabolic load; however, it decreases robustness. Growth rate maximization balances these two costs. Quantitatively, the maximization of the growth rate (Eq. 1) with respect to the translation efficiency can be performed analytically, predicting the optimal translation efficiency, shown in Fig. 4A. We provide an exact expression in the Supplementary Material Sec. A 4; however, an approximate expression for the translation efficiency is more clearly interpretable:

$$\hat{\varepsilon} \approx 0.1\lambda\hat{\mu}_m. \tag{3}$$

The optimal translation efficiency has two important qualitative features for central dogma regulation. The first prediction is that as the message cost (λ) rises, the optimal translation efficiency ($\hat{\varepsilon}$) increases in proportion while the message number decreases. We present evidence for this prediction in the Supplementary Material Sec. A 11.

The second prediction is that the optimal translation efficiency is also approximately proportional to message

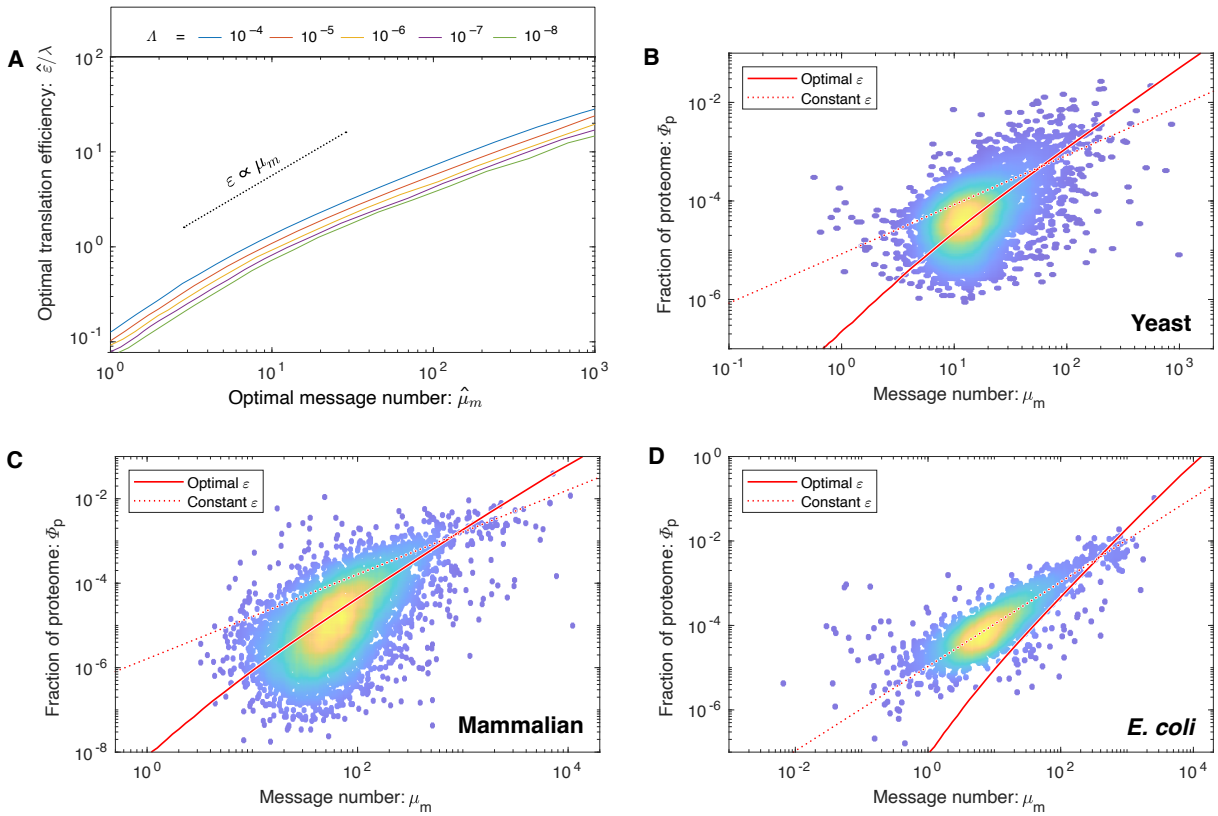

FIG. 4. **How are transcription and translation balanced? Panel A: The RLTO model predicts load balancing.** The ratio of the optimal translation efficiency ($\hat{\varepsilon}$) to the message cost (λ) is roughly independent of the relative load (Λ). The translation efficiency ε is predicted to be roughly proportional message number μ_m. **Panel B: RLTO predicts the protein-message-abundance relation in yeast.** The observed proteome fraction is compared to two models: the RLTO optimal model (solid red line) and constant-translation-efficiency model (dotted red line). Both models make parameter-free predictions. The RLTO optimum predicts the global trend. (Data from Ref. [21].) **Panel C: Mammalian proteome fraction.** The RLTO prediction (solid) is superior to the constant-translation-efficiency prediction (dashed). **Panel D: *E. coli* proteome fraction.** In contrast, the constant-translation-efficiency prediction (dashed) is superior to RLTO prediction (solid).

number ($\hat{\varepsilon} \propto \mu_m$). Therefore, the RLTO model predicts that low expression levels should be achieved with low levels of transcription and translation, whereas high-expression genes are achieved with high levels of both. We call this relation between optimal transcription and translation the *load balancing principle*. The most direct test of load balancing is measuring the protein-message abundance relation. Due to load balancing, the RLTO model predicts protein number (and proteome fraction) to scale like:

$$\hat{\mu}_p \propto \hat{\mu}_m^2, \tag{4}$$

whereas a constant-translation-efficiency model has linear scaling ($\mu_p \propto \mu_m$). Computing proteome fraction, rather than protein number, results in a parameter-free prediction. (See Supplemental Sec. A 12.)

Load balancing is observed in eukaryotic cells. To test the RLTO predictions, we compare observed proteome measurements in three evolutionarily divergent species, *E. coli* [22], yeast [21] and mammalian cells [23], to two models: the RLTO and the constant-translation-efficiency models. The results of the parameter-free predictions are shown in Fig. 4BCD for each organism. The RLTO model clearly captures the global trend in the proteome-fraction message-number relation in eukaryotic cells and a direct fit to a power law with an unknown exponent is consistent with Eq. 4 (Supplementary Material Sec. D 4 c).

In *E. coli*, the constant-translation-efficiency model better describes the data. Why does this organism appear not to load balance? In the supplementary material, we demonstrate that the observed translation efficiency is consistent with the RLTO model, augmented by a ribosome-per-message limit. Hausser *et al.* have proposed just such a limit, based on the ribosome footprint on mRNA molecules [10]. (See Supplementary Material Sec. A 17.) Although this augmented model is consistent with central dogma regulation in *E. coli*, it is not a complete rationale. This proposed translation-rate limit could be circumvented by increasing the lifetime

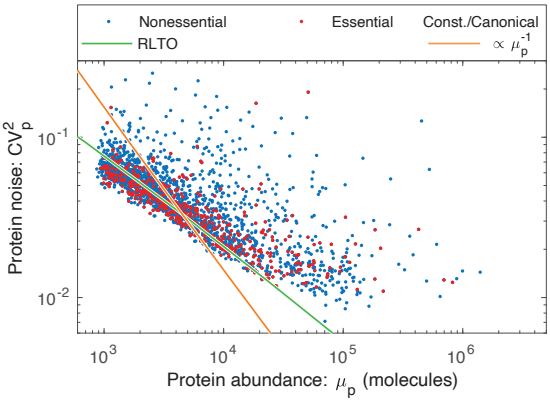

FIG. 5. **RLTO predicts the magnitude of noise in yeast.** The observed gene expression noise is yeast is shown for essential and nonessential genes. Two protein-message abundance models are compared to the data: The RLTO model (green) versus the constant-translation-efficiency (canonical model, orange). The RLTO model predicts both the magnitude of the noise, as well as its scaling with protein abundance. The reduced slope of the RLTO model is the consequence of load balancing, which reduces the noise for the noisiest, low-expression genes. (Data from Ref. [8].)

of *E. coli* messages, which would increase the translation efficiency. A more in-depth analysis specific to *E. coli* is needed to understand why the observed message lifetime is so short.

RLTO model predicts observed noise in yeast. Although the protein fraction measurements support the RLTO predictions for the translation efficiency in eukaryotic cells, these measurements do not provide a compelling rationale for why load balancing maximizes the growth rate. To understand its rationale, we explore its implications for noise.

In a typical biological context, $\mu_m \ll \varepsilon$ and as a result, noise production is dominated by the transcription step of the gene expression process [12, 13]. (A table of central dogma parameters for each model organism appears in the Supplementary Material Tab. S2.) Quantitatively, the Paulsson model predicts that the noise should be inversely related to the message number [12, 13]:

$$\mathrm{CV}_p^2 = \frac{\ln 2}{\mu_m}, \tag{5}$$

however, it is the relation between mean protein abundance μ_p and noise (CV_p^2) which is typically reported [8, 9]. Based on the scaling of the optimal translation efficiency with the message number in eukaryotic cells (Eq. 3), we find the protein number to scale with message number (Eq. 4), which predicts that noise should scale with protein abundance $\mathrm{CV}_p^2 \propto \mu_p^{-1/2}$ in yeast (see Supplementary Material Sec. D 1); however, due to the observed absence of translation-efficiency scaling in bacteria, the noise should scale as $\mathrm{CV}_p^2 \propto \mu_p^{-1}$ in bacteria, as observed [9]. Does the yeast noise show the predicted scaling? The parameter-free RLTO noise prediction closely matches the observed noise in both magnitude and scaling, as shown in Fig. 5.

Reducing noise is the rationale for load balancing. This noise analysis also provides a conceptual insight into the rationale for load balancing. The load balanced (RLTO-green) and constant-translation-efficiency (orange) predictions for the noise are shown in Fig. 5. Load balancing results in decreased noise for low-expression, noisy genes over what is achieved with constant translation efficiency. This decreased noise is predicted to increase growth robustness. In principle, the noise could be reduced further by tipping the balance even more towards transcription; however, the RLTO model predicts that this approach is too metabolically costly, and the optimal strategy is that observed for noise scaling in yeast.

DISCUSSION

What are the biological implications of noise? Many important proposals have been made, including bet-hedging strategies, the necessity of feedback in gene regulatory networks, *etc.* [7]. Our model suggests that overcoming cell-to-cell variation may fundamentally reshape the metabolic budget: Typically, proteins constitute 50-60% of the dry mass of the cell [5] and therefore overabundance could increase the overall protein budget by a significant factor. Why does the cell tolerate this significant increase in metabolic load above what would be predicted by a resource allocation analysis (*e.g.* [24])? This program dramatically reduces the consequence of stochastic expression of proteins on the rate of single-cell proliferation.

A second source of stochasticity, environmental fluctuations, has been proposed as a rationale for overabundance [25], especially in the context of metabolic genes [26]. In short, cells express protein to hedge against starvation [25] or changes in the carbon source, *etc.* [26]. How does this hypothesis compare to our growth robustness hypothesis? There are some similarities between these environmental-fluctuation models and the RLTO model: In both models, it is a fluctuations-based mechanism that drives overabundance; however, there are important distinctions between the model predictions. In the environmental fluctuation model, there is a trade-off between log-phase fitness and the rapidity of adaptation [25]; whereas in the RLTO model, overabundance corresponds to the log-phase optimum. Organisms experiencing prolonged periods of balanced growth would therefore be expected to reduce overabundance. Furthermore, the environmental fluctuation model most naturally explains overabundance for proteins related to metabolic processes, whereas the RLTO model predicts overabundance generically, dependent only on message number, which appears to be much

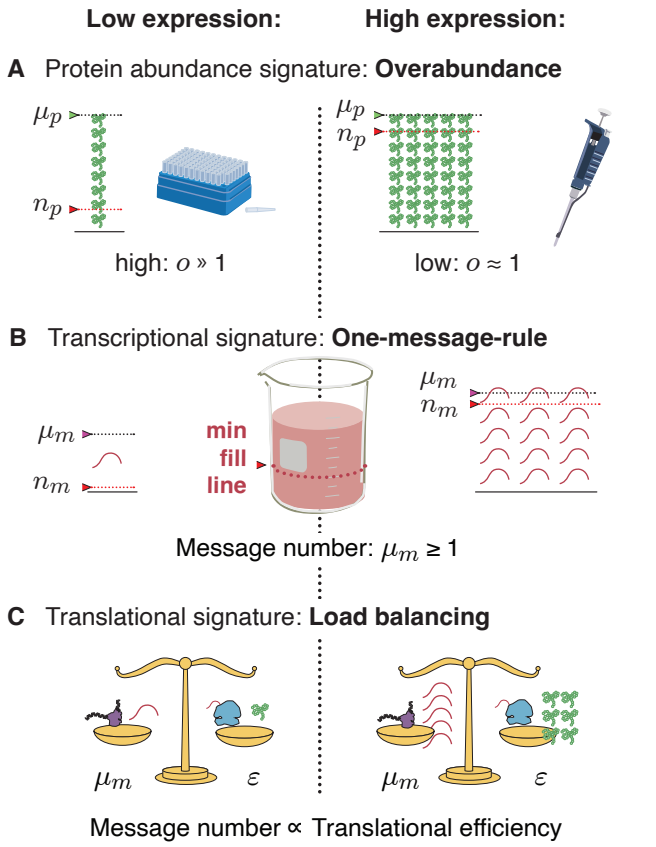

FIG. 6. **Central dogma regulatory principles. Panel A: Overabundance.** Low-expression essential genes are expressed with high overabundance; whereas, high-expression essential genes are expressed with low overabundance. Lab supply analogy: Low-cost items that are used stochastically (*e.g.* pipette tips) are purchased in great excess, while the higher cost items that are less stochastic (*e.g.* pipette) are purchased as needed. **Panel B: One-message rule.** Robust expression of essential genes requires them to be transcribed above a threshold of one message per cell cycle. **Panel C: Load balancing.** In eukaryotic cells, optimal fitness is achieved by balancing transcription and translation: The optimal message number is proportional to the optimal translation efficiency. High (low) expression levels are achieved by high (low) levels of transcription followed by high (low) levels of translation per message.

more consistent with experiments exploring essential-protein depletion [4].

Implications for nonessential genes. In our analysis, we have focused on essential genes in order to motivate the growth-threshold in the RLTO model. To what extent do nonessential genes share the same optimization? In support of the proposal that RLTO optima describe nonessential genes is the success of the model in predicting the translation efficiency for all genes, not just essential genes. (See Fig. 4 and Fig. 5.) Furthermore, the definition of a gene as *essential* depends on context: For instance, in the context of *E. coli* growth on lactose, the gene *lacZ* is essential, although it is nonessential on other carbon sources [27]. Under growth conditions where the *lacZ* gene is essential, we predict that LacZ should be overabundant, consistent with observation [26]. Finally, our modeling suggested that RLTO model phenomenology is the results of asymmetry of the cost of under versus overabundance. For nonessential genes whose activity significantly increases fitness, we still expect fitness asymmetry due to the low relative metabolic cost of increased expression. We therefore expect all gene products, most especially those with low expression, to be overabundant, under conditions where their activity increases fitness.

Implications of overabundance for inhibitors. The generic nature of overabundance, especially for low-expression proteins, has important potential implications for the targeting of these proteins with small-molecule inhibitors (*e.g.* drugs). For the highest expression proteins, like the constituents of the ribosome, relatively small decreases in the active fraction (*e.g.* a three-fold reduction) are expected to lead to growth arrest [3]. This may help explain why inhibitors targeting translation make such effective antimicrobial drugs. However, we predict that the lowest expression proteins require a much higher fraction of the protein to be inactivated, with the lowest-expression proteins expected to need more than a 100-fold depletion. This predicted robustness makes these proteins much less attractive drug targets [28].

The principles that govern central dogma regulation. We propose that robustness to noise fundamentally shapes the central dogma regulatory program for all genes and predicts a number of key regulatory principles. (See Fig. 6.) For high-expression genes, load balancing implies that gene expression consists of both high-amplification translation and transcription. The resulting expression level has low overabundance relative to the threshold required for function. In contrast, for essential low-expression genes, a three-fold strategy is implemented: (i) overabundance raises the mean protein levels far above the threshold required for function, (ii) load balancing, and (iii) the one-message rule ensure that message number is sufficiently large to lower the noise of inherently-noisy, low-expression genes. We anticipate that these regulatory principles, in particular protein overabundance, will have important implications, not only for our understanding of central dogma regulation specifically, but for understanding the rationale for protein expression level and function in many biological processes.

MATERIALS AND METHODS

RLTO model. The effect of stochastic cell arrest can be implemented analytically as follows: The probability of growth is the probability that all essential proteins are

above threshold, P_+. The population growth rate k is [29]:

$$\frac{k}{k_0} = 1 + \frac{1}{\ln 2} \ln P_+, \tag{6}$$

for a population of cells subject to stochastic arrest with probability $1 - P_+$ per cell cycle where k_0 is the growth rate of the non-arrested cells. For each gene i, the Paulsson model predicts the protein number CDF in terms of message number μ_m and translation efficiency ε [12]. Assuming the below-threshold probability is small, the probability that the cell is below threshold for gene i is:

$$\ln P_{+,i} = -\gamma(\frac{\mu_m}{\ln 2}, \frac{n_p}{\varepsilon \ln 2}), \tag{7}$$

where γ is the regularized incomplete gamma function and the CDF of the gamma distribution. (See Supplementary Material Sec. A 2.)

While protein underabundance slows cell growth by the arrest of essential processes, protein overabundance slows growth by increasing the metabolic load. To implement the metabolic-load contribution to cell fitness, we use a minimal model that realizes the metabolic cost of both transcription and translation that is analogous to those previously used in the context of resource allocation (*e.g.* [14]). The metabolic load of transcription and translation of gene i is:

$$\frac{k}{k_0} = 1 - \frac{\lambda + \varepsilon}{N_0} \mu_m, \tag{8}$$

where k_0 is the growth rate in the absence of the metabolic load of gene i, N_0 is the total cellular metabolic load, and λ is the metabolic message cost. (See the Supplementary Material Sec. A 2 for a detailed development of the model.) The λ-term represents the metabolic cost of transcription and the ε-term represents the metabolic cost of translation of gene i. We define the relative load as $\Lambda \equiv \lambda/N_0$ as the ratio of the metabolic load of a single message to the total metabolic cost of the cell. In *E. coli*, we estimate that Λ is roughly 10^{-5} and it is smaller still for eukaryotic cells.

Data analysis. We provide a detailed description of the data analysis for the one message rule, loading balancing, and the noise analysis in the Supplementary Material.

Acknowledgments. The authors would like to thank B. Traxler, A. Nourmohammad, J. Mougous, K. Cutler, M. Cosentino-Lagomarsino, S. van Teeffelen, C. Manoil, L. Gallagher, J. Bailey and S. Murray.
Funding: National Institutes of Health grant R01-GM128191
Author contributions: T.W.L., H.J.C., D.H. and P.A.W. conceived the research, performed the analysis, and wrote the paper.
Competing interests: The authors declare no competing interests.

Data availability. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Supplementary Materials
Supplementary Text
Figs. S1 to S12
Tables S1 to S3
Data Tables S1 to S7

[1] E. Dekel, U. Alon, Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436**, 588-92 (2005).
[2] L. Keren, J. Hausser, M. Lotan-Pompan, I. Vainberg Slutskin, H. Alisar, S. Kaminski, A. Weinberger, U. Alon, R. Milo, E. Segal, Massively parallel interrogation of the effects of gene expression levels on fitness. *Cell* **166**, 1282-1294.e18 (2016).
[3] J. M. Peters, A. Colavin, H. Shi, T. L. Czarny, M. H. Larson, S. Wong, J. S. Hawkins, C. H. S. Lu, B.-M. Koo, E. Marta, A. L. Shiver, E. H. Whitehead, J. S. Weissman, E. D. Brown, L. S. Qi, K. C. Huang, C. A. Gross, A comprehensive, CRISPR-based functional analysis of essential genes in bacteria. *Cell* **165**, 1493-1506 (2016).
[4] L. A. Gallagher, J. Bailey, C. Manoil, Ranking essential bacterial processes by speed of mutant death. *Proc Natl Acad Sci U S A* **117**, 18010-18017 (2020).
[5] H. G. S. Joseph W. Lengeler, Gerhart Drews, ed., *Biology of the Prokaryotes* (Georg Thieme Verlag, Rüdigerstrasse 14, D-70469 Stuttgart, Germany, 1998).
[6] N. M. Belliveau, G. Chure, C. L. Hueschen, H. G. Garcia, J. Kondev, D. S. Fisher, J. A. Theriot, R. Phillips, Fundamental limits on the rate of bacterial growth and their influence on proteomic composition. *Cell Syst* **12**, 924-944.e2 (2021).
[7] J. M. Raser, E. K. O'Shea, Noise in gene expression: origins, consequences, and control. *Science* **309**, 2010-3 (2005).
[8] J. R. S. Newman, S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble, J. L. DeRisi, J. S. Weissman, Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* **441**, 840-6 (2006).
[9] Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, X. S. Xie, Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533-8 (2010).
[10] J. Hausser, A. Mayo, L. Keren, U. Alon, Central dogma rates and the trade-off between precision and economy in gene expression. *Nat Commun* **10**, 68 (2019).
[11] F. Crick, Central dogma of molecular biology. *Nature* **227**, 561-3 (1970).
[12] J. Paulsson, M. Ehrenberg, Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks. *Phys Rev Lett* **84**, 5447-50 (2000).
[13] N. Friedman, L. Cai, X. S. Xie, Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys Rev Lett* **97**, 168302 (2006).
[14] M. Scott, C. W. Gunderson, E. M. Mateescu, Z. Zhang,

T. Hwa, Interdependence of cell growth and gene expression: origins and consequences. *Science* **330**, 1099-102 (2010).

[15] J. I. Steinfeld, J. S. Francisco, W. L. Hase, *Chemical Kinetics and Dynamics* (Prentice-Hall, 1999), second edn.

[16] D. A. Charlebois, N. Abdennur, M. Kaern, Gene expression noise facilitates adaptation and drug resistance independently of mutation. *Phys. Rev. Lett.* **107**, 218101 (2011).

[17] M. R. Silvis, M. Rajendram, H. Shi, H. Osadnik, A. N. Gray, S. Cesar, J. M. Peters, C. C. Hearne, P. Kumar, H. Todor, K. C. Huang, C. A. Gross, Morphological and transcriptional responses to CRISPRi knockdown of essential genes in *Escherichia coli*. *mBio* **12**, e0256121 (2021).

[18] S. Donati, M. Kuntz, V. Pahl, N. Farke, D. Beuter, T. Glatter, J. V. Gomes-Filho, L. Randau, C.-Y. Wang, H. Link, Multi-omics analysis of CRISPRi-knockdowns identifies mechanisms that buffer decreases of enzymes in *E. coli* metabolism. *Cell Syst* **12**, 56-67.e6 (2021).

[19] J. van Leeuwen, C. Pons, G. Tan, J. Z. Wang, J. Hou, J. Weile, M. Gebbia, W. Liang, E. Shuteriqi, Z. Li, M. Lopes, M. Ušaj, A. Dos Santos Lopes, N. van Lieshout, C. L. Myers, F. P. Roth, P. Aloy, B. J. Andrews, C. Boone, Systematic analysis of bypass suppression of essential genes. *Mol Syst Biol* **16**, e9828 (2020).

[20] T. Wang, K. Birsoy, N. W. Hughes, K. M. Krupczak, Y. Post, J. J. Wei, E. S. Lander, D. M. Sabatini, Identification and characterization of essential genes in the human genome. *Science* **350**, 1096-101 (2015).

[21] S. Ghaemmaghami, W.-K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, J. S. Weissman, Global analysis of protein expression in yeast. *Nature* **425**, 737-41 (2003).

[22] R. Balakrishnan, M. Mori, I. Segota, Z. Zhang, R. Aebersold, C. Ludwig, T. Hwa, Principles of gene regulation quantitatively connect DNA to RNA and proteins in bacteria. *Science* **378**, eabk2066 (2022).

[23] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, M. Selbach, Global quantification of mammalian gene expression control. *Nature* **473**, 337-42 (2011).

[24] H. Dourado, M. Mori, T. Hwa, M. J. Lercher, On the optimality of the enzyme-substrate relationship in bacteria. *PLoS Biol.* **19**, e3001416 (2021).

[25] M. Mori, S. Schink, D. W. Erickson, U. Gerland, T. Hwa, Quantifying the benefit of a proteome reserve in fluctuating environments. *Nat Commun* **8**, 1225 (2017).

[26] G. Lambert, E. Kussell, Memory and fitness optimization of bacteria under fluctuating environments. *PLoS Genet* **10**, e1004556 (2014).

[27] J. Monod, A. M. Pappenheimer, Jr, G. Cohen-Bazire, The kinetics of the biosynthesis of beta-galactosidase in *Escherichia coli* as a function of growth. *Biochim Biophys Acta* **9**, 648-60 (1952).

[28] B. Bosch, M. A. DeJesus, N. C. Poulton, W. Zhang, C. A. Engelhart, A. Zaveri, S. Lavalette, N. Ruecker, C. Trujillo, J. B. Wallach, S. Li, S. Ehrt, B. T. Chait, D. Schnappinger, J. M. Rock, Genome-wide gene expression tuning reveals diverse vulnerabilities of *M. tuberculosis*. *Cell* **184**, 4579-4592.e24 (2021).

[29] D. Huang, T. Lo, H. Merrikh, P. A. Wiggins, Characterizing stochastic cell-cycle dynamics in exponential growth. *Phys Rev E* **105**, 014420 (2022).

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

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

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

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

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

[35] J. A. Bernstein, A. B. Khodursky, P.-H. Lin, S. Lin-Chao, S. N. Cohen, 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).

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

[37] (2023).

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

[39] M. Kafri, E. Metzl-Raz, F. Jonas, N. Barkai, Rethinking cell growth models. *FEMS Yeast Res* **16** (2016).

[40] D. Huang, A. E. Johnson, B. S. Sim, T. W. Lo, H. Merrikh, P. A. Wiggins, The in vivo measurement of replication fork velocity and pausing by lag-time analysis. *Nat Commun* **14**, 1762 (2023).

[41] M. Kafri, E. Metzl-Raz, G. Jona, N. Barkai, The cost of protein production. *Cell Rep* **14**, 22-31 (2016).

[42] E. Metzl-Raz, M. Kafri, G. Yaakov, N. Barkai, Gene transcription as a limiting factor in protein production and cell growth. *G3 (Bethesda)* **10**, 3229-3242 (2020).

[43] G. Arfken, *Mathematical Methods for Physicists* (Academic Press, Inc., San Diego, 1985), third edn.

[44] A. Bartholomäus, I. Fedyunin, P. Feist, C. Sin, G. Zhang, A. Valleriani, Z. Ignatova, Bacteria differently regulate mRNA abundance to specifically respond to various stresses. *Philos Trans A Math Phys Eng Sci* **374** (2016).

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

[46] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J. Watson, *Molecular Biology of the Cell* (Garland, 2002), fourth edn.

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

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

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

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

[51] E. Yang, E. van Nimwegen, M. Zavolan, N. Rajewsky, M. Schroeder, M. Magnasco, J. E. Darnell, Jr, Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res* **13**, 1863-72 (2003).

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

[53] T. Baba, H.-C. Huan, K. Datsenko, B. L. Wanner, H. Mori, The applications of systematic in-frame, single-gene knockout mutant collection of *Escherichia coli* K-12. *Meth-*

ods Mol Biol **416**, 183-94 (2008).

[54] W. R. Blevins, T. Tavella, S. G. Moro, B. Blasco-Moreno, A. Closa-Mosquera, J. Díez, L. B. Carey, M. M. Albà, Extensive post-transcriptional buffering of gene expression in the response to severe oxidative stress in baker's yeast. *Sci Rep* **9**, 11005 (2019).

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

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

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

[58] F. Miura, N. Kawaguchi, M. Yoshida, C. Uematsu, K. Kito, Y. Sakaki, T. Ito, Absolute quantification of the budding yeast transcriptome by means of competitive PCR between genomic and complementary DNAs. *BMC Genomics* **9**, 574 (2008).

[59] L. M. F. de Godoy, J. V. Olsen, J. Cox, M. L. Nielsen, N. C. Hubner, F. Fröhlich, T. C. Walther, M. Mann, Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251-4 (2008).

[60] M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, F. Pontén, Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

[61] V. E. Velculescu, S. L. Madden, L. Zhang, A. E. Lash, J. Yu, C. Rago, A. Lal, C. J. Wang, G. A. Beaudry, K. M. Ciriello, B. P. Cook, M. R. Dufault, A. T. Ferguson, Y. Gao, T. C. He, H. Hermeking, S. K. Hiraldo, P. M. Hwang, M. A. Lopez, H. F. Luderer, B. Mathews, J. M. Petroziello, K. Polyak, L. Zawel, K. W. Kinzler, Analysis of human transcriptomes. *Nat Genet* **23**, 387-8 (1999).

[62] S. Y. Gerdes, M. D. Scholle, J. W. Campbell, G. Balázsi, E. Ravasz, M. D. Daugherty, A. L. Somera, N. C. Kyrpides, I. Anderson, M. S. Gelfand, A. Bhattacharya, V. Kapatral, M. D'Souza, M. V. Baev, Y. Grechkin, F. Mseeh, M. Y. Fonstein, R. Overbeek, A.-L. Barabási, Z. N. Oltvai, A. L. Osterman, Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J Bacteriol* **185**, 5673-84 (2003).

[63] E. C. A. Goodall, A. Robinson, I. G. Johnston, S. Jabbari, K. A. Turner, A. F. Cunningham, P. A. Lund, J. A. Cole, I. R. Henderson, The essential genome of *Escherichia coli* K-12. *mBio* **9** (2018).

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

[65] A. Bar-Even, J. Paulsson, N. Maheshri, M. Carmi, E. O'Shea, Y. Pilpel, N. Barkai, Noise in protein expression scales with natural protein abundance. *Nat Genet* **38**, 636-43 (2006).

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

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

[68] C. R. Rao, Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of Calcutta Mathematical Society* **37**, 81-91. (1945).

[69] H. Cramer, *Mathematical methods of statistics.* (Princeton University Press., 1946).

[70] P.-J. Lahtvee, B. J. Sánchez, A. Smialowska, S. Kasvandik, I. E. Elsemman, F. Gatto, J. Nielsen, Absolute quantification of protein and mRNA abundances demonstrate variability in gene-specific translation efficiency in yeast. *Cell Syst* **4**, 495-504.e5 (2017).

[71] K. H. Hellton, M. Thoresen, The impact of measurement error on principal component analysis. *Scandinavian Journal of Statistics* **41**, 1051–1063 (2014).

Supplementary Material: Noise robustness and metabolic load determine the principles of central dogma regulation

Authors: Teresa W. Lo,[1,*] Han 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*

CONTENTS

A. Detailed development of the RLTO model 2
1. Methods: Detailed description of the noise model 3
a. Stochastic kinetic model for the central dogma. 3
b. Statistical model for protein abundance. 3
c. Gillespie simulation of stochastic kinetic scheme 3
d. Gamma function and distribution conventions 4
2. Methods: The derivation of the RLTO growth rate 4
a. The metabolic load of protein and the Hwa model 4
b. The metabolic load of mRNA 5
c. Metabolic load in the RLTO model 5
d. Growth rate with stochastic arrest 6
e. RLTO growth rate 6
f. When the ln approximation is avoided... 6
g. Single-gene equation 7
h. Summary of RLTO parameter values for figures 7
3. Discussion: The fitness landscape of a trade-off. 7
4. Methods: Central dogma optimization 7
a. Optimization of transcription and translation (eukaryotes) 7
b. Optimization of message number only 9
c. Inclusion of the noise floor 9
5. Discussion: Understanding the rationale for overabundance 10
6. Discussion: RLTO predicts larger overabundance in bacteria. 11
7. Discussion: RLTO predicts proteins are buffered to depletion. 11
8. Results: Detailed development of load balancing 11
a. Prediction of the optimal load ratio. 11
b. Measurements of the load ratio 11
9. Results: Translation efficiency is predicted to increase with transcription. 12
10. Results: RLTO predicts that message number responds to message cost 12
11. Results: RLTO predicts the yeast global regulatory response 12
12. Methods: Prediction of the proteome fraction 12
a. Parameter-free prediction of proteome fraction. 12
b. RLTO: proteome fraction 12
c. Constant-translation-efficiency model: proteome fraction 13
d. Sources of experimental data for proteome fraction analysis 13
13. Results: Load balancing is observed in eukaryotic cells 13
14. Discussion: Relation between load balancing and previous results 14
15. Methods: Analysis of translational limits & gene-specific load analysis 14
16. Results: Increased protein-specific cost reduces the optimal translation efficiency. 15
17. Discussion: Translation limits in *E. coli* 15

* These two authors contributed equally

† pwiggins@uw.edu

18. Methods: Estimate of the message cost and metabolic load 16
a. Detailed protocol 16
b. Estimate the message cost and metabolic load in yeast
c. Estimate the message cost and metabolic load in human cells 16
d. Estimate of the modified relative load in bacterial cells 17

B. Model robustness & exploring alternatives to RLTO 17
1. Methods: Defining alternative models 17
a. Model 1: Arrest (RLTO) model 17
b. Model 2: Slow-Growth model 17
c. Model 3: Symmetric model 17
2. Results: Overabundance is a robust prediction 18

C. Quantitation of central dogma parameters for one-message-rule 18
1. Methods: Selection of central dogma parameter estimates 18
a. *E. coli* data 18
b. Yeast data 18
c. Human data 19
2. Methods: Quantitative estimates of central dogma parameters 19
a. Estimating the *cellular message number*: $\mu_{m/c}$ 19
b. Estimating the *transcription rate*: β_m 19
c. Estimating the *message number*: μ_m 20
3. Results: Histograms of central dogma transcriptional statistics 21
4. Discussion: *E. coli* essential genes below the one-message-rule threshold 21

D. Analysis of gene-expression noise 21
1. Results: RLTO model predicts non-canonical noise scaling 21
2. Methods: Analysis of gene expression noise 21
a. Hypothesis test I 22
b. Hypothesis test II 23
c. Maximum likelihood estimates of the parameters 23
d. Details: Statistical details MLE estimate of the variance 23
3. Results: Non-canonical noise scaling is observed in yeast. 23
4. Methods: Parameter-free prediction of noise from protein-message relation. 24
a. Estimating *protein number* (μ_p) for the noise analysis 24
b. Empirical models for yeast gene expression 24
c. Empirical model for protein number 24
d. Empirical model for message number 24
e. Empirical model for translation efficiency 25
f. Empirical model for the coefficient of variation 25
5. Results: Parameter-free prediction of noise-abundance in yeast 25
6. Discussion: Implications of noise 26

E. Data Tables 26
1. Datasets for one-message-rule analysis. 26
2. Datasets for load-balancing analysis. 26
3. Datasets for noise analysis. 26

Appendix A: Detailed development of the RLTO model

In this section, we provide a detailed development of the RLTO model. First, we describe the stochastic kinetic model for the central dogma, which introduces key quantities for the RLTO model (Sec. A 1). Next, we provide a derivation of the growth rate as a function of the model parameters (Sec. A 2) as well as other methods (Secs. A 4, A 12, A 15, and A 18). For each of the results discussed in the main paper, we provide more detailed analyses, which include both supplemental results (Secs. A 8, A 9, A 11, A 13, and A 16) that support the story described in the main paper, as well as supplemental discussions (Secs. A 3, A 5, A 6, A 7, A 14, and A 17).

1. Methods: Detailed description of the noise model

a. *Stochastic kinetic model for the central dogma.*

The equilibrium Paulsson model for the central dogma describes multiple steps in the gene expression process [9, 12, 13]: Transcription generates mRNA messages. These messages are then translated to synthesize the protein gene products [11]. Both mRNA and protein are subject to degradation and dilution [30]. At the single cell level, each of these processes are stochastic. We will model these processes with the stochastic kinetic scheme [11]:

$$\begin{array}{ccccc} \mathrm{DNA} & \xrightarrow{\beta_m} & \mathrm{mRNA} & \xrightarrow{\beta_p} & \mathrm{Protein} \\ & & \gamma_m \big\downarrow & & \gamma_p \big\downarrow \\ & & \varnothing & & \varnothing, \end{array} \tag{A1}$$

where β_m is the transcription rate (s^{-1}), β_p is the translation rate (s^{-1}), γ_m is the message degradation rate (s^{-1}), and γ_p is the protein effective degradation rate (s^{-1}). The message lifetime is $T_m \equiv \gamma_m^{-1}$. For most proteins in the context of rapid growth, dilution is the dominant mechanism of protein depletion and therefore γ_p is approximately the growth rate [9, 31, 32]: $\gamma_p = T^{-1} \ln 2$, where T is the doubling time.

b. *Statistical model for protein abundance.*

To study the stochastic dynamics of gene expression, we used a stochastic Gillespie simulation [33, 34]. (See Sec. A 1 c.) 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 [12, 13], we find that the distribution of protein number per cell (at cell birth) was described by a gamma distribution: $N_p \sim \Gamma(a, \theta)$, where N_p is the protein number at cell birth and Γ is the gamma distribution which is parameterized by a scale parameter θ and a shape parameter a. (See Sec. A 1 d.) We refer to this distribution as the *Paulsson model*; The relation between the four kinetic parameters and these two statistical parameters has already been reported, and have clear biological interpretations [13]: The scale parameter:

$$\theta = \varepsilon \ln 2, \tag{A2}$$

is proportional to the translation efficiency:

$$\varepsilon \equiv \frac{\beta_p}{\gamma_m}, \tag{A3}$$

where β_p is the translation rate and γ_m is the message degradation rate. ε is understood as the mean number of proteins translated from each message transcribed. The shape parameter a can also be expressed in terms of the kinetic parameters [13]:

$$a = \frac{\beta_m}{\gamma_p}; \tag{A4}$$

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

$$\mu_m \equiv \beta_m T = a \ln 2, \tag{A5}$$

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.

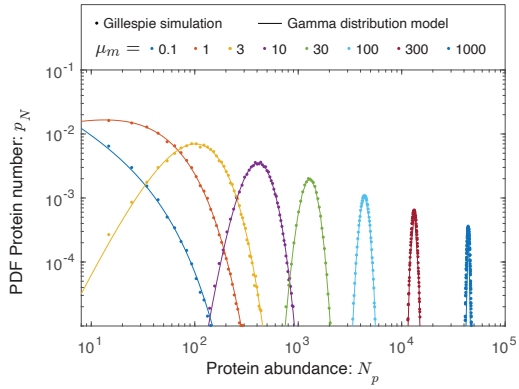

FIG. S1. **The protein abundance is approximately gamma distributed.** Protein abundance was modeled for eight different transcription rates using a Gillespie simulation, including the stochastic partitioning of the proteins between daughter cells at cell division. The range in abundance matches the observed range of expression levels in the cell. We observed that the simulated protein abundances were well fit by gamma distributions.

c. *Gillespie simulation of stochastic kinetic scheme*

Protein distributions based on the kinetic scheme defined in Sec. A 1 a were simulated with a Gillespie algorithm, with specific parameter values for *E. coli*. Assuming the lifetime of the cell cycle ($T = 30$ min) [35], mRNA lifetime ($T_m = 2.5$ min) [36], 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 ($\theta = \varepsilon \ln 2$, $a = \frac{\mu_m}{\ln 2}$) as used for the Gillespie simulation were also plotted with the protein distributions. We observe an excellent match between these

Gillespie simulations and the Paulsson statistical noise model (*i.e.* gamma function) as shown in Fig. S1.

d. Gamma function and distribution conventions

There are a number of conflicting conventions for the gamma function and distribution arguments. We will use those defined on Wikipedia [37]. The gamma distributed random variable X will be written:

$$X \sim \Gamma(a, \theta), \tag{A6}$$

where a is the shape parameter and θ is the scale parameter. The PDF of the distribution is:

$$p_\Gamma(x|a,\theta) \equiv \frac{x^{a-1}}{\theta^a \Gamma(a)} e^{-x/\theta}, \tag{A7}$$

where $\Gamma(a)$ is the gamma function. The CDF is therefore:

$$\begin{aligned} P_\Gamma(x|a,\theta) &= \int_0^x \mathrm{d}x' \; p_\Gamma(x'|a,\theta), && \text{(A8)} \\ &= P_\Gamma(\tfrac{x}{\theta}|a,1), && \text{(A9)} \\ &= \int_0^{x/\theta} \mathrm{d}x' \; \tfrac{x^{a-1}}{\Gamma(a)} e^{-x}, && \text{(A10)} \\ &= \gamma(a, x/\theta), && \text{(A11)} \end{aligned}$$

where γ is the regularized incomplete gamma function.

2. Methods: The derivation of the RLTO growth rate

a. The metabolic load of protein and the Hwa model

To model the effect of the metabolic load on cell growth, we will expand on a model used by Hwa and co-workers [14]. For conciseness, we will call the original model the *Hwa model*.

Consider a cell where the total number of proteins in the cell is N. The synthesis of these proteins requires two sets of processes: (i) the metabolic processes responsible for synthesizing the precursors (*i.e.* amino acids, *etc*) and (ii) the translation process. Proteins involved in the metabolic processes be referred to as the P sector and number N_P. The proteins involved in the translational process will be referred to as the R sector and number N_R. In addition to the P and R sectors, there is a third Q sector with protein number N_Q. The total protein number per cell is therefore:

$$N = N_R + N_P + N_Q. \tag{A12}$$

The proteome fractions are defined $\Phi_X \equiv N_X/N$ and have the normalization condition:

$$1 = \Phi_R + \Phi_P + \Phi_Q. \tag{A13}$$

The key assumption in the Hwa model is that the abundances of the R and P sectors can change in size to accommodate changes in the nutrient quality and translation load associated with a particular growth condition [14]. In contrast, the Q sector has a fixed proteome fraction, Φ_Q, irrespective of growth conditions. In the Hwa model, the size of these adjustable R and P sectors are chosen to optimize the growth rate k.

The condition for balanced growth requires that the overall protein output of the translation process match the growth rate:

$$kN = k_R N_{*R}, \tag{A14}$$

where k_R is the effective translation rate per protein and N_{*R} is the number of productive R sector proteins, which is subset of the total number N_R:

$$N_R = N_{*R} + N_{0R}, \tag{A15}$$

and N_{0R} represents unproductive R sector protein. We can rewrite Eq. A14 in terms of the proteome fraction:

$$k = k_R(\Phi_R - \Phi_{0R}). \tag{A16}$$

In the expression above, we will assume that the parameters k_R and Φ_{0R} are fixed, but the total fraction Φ_R is chosen to optimize the growth rate k.

For any productive sectors i, we will write analogous equations to Eq. A16 linking sector fraction size Φ_i to function:

$$k = k_i(\Phi_i - \Phi_{0i}), \tag{A17}$$

where, as before, Φ_{0i} represents a fixed-size fraction of unproductive protein.

To determine the unknown optimum growth rate and sector sizes, Eq. A17 can be re-written:

$$\Phi_{*i} = \Phi_i - \Phi_{0i} = \tfrac{k}{k_i}, \tag{A18}$$

and then summed over all sectors (excluding Q):

$$\Phi_* = k \sum_{i \neq Q} k_i^{-1}, \tag{A19}$$

where we define the total productive fraction of the proteome:

$$\Phi_* \equiv \sum_{i \neq Q} \Phi_{*i} = 1 - \Phi_0, \tag{A20}$$

and Φ_0 represents the total fraction of unproductive protein:

$$\Phi_0 = \Phi_Q + \sum_{i \neq Q} \Phi_{0i}, \tag{A21}$$

including the entire Q sector. Since Φ_Q and the Φ_{0i} are all assumed to be fixed, Eq. A19 determines the growth

rate. (In *E. coli*, Hwa and coworkers estimate that $\Phi_* \approx 0.55$.)

To understand the meaning of Eq. A19, we first define an ideal growth rate as

$$k_{\text{ideal}}^{-1} = \sum_{i \neq Q} k_i^{-1}, \tag{A22}$$

which would be the growth rate in the absence of unproductive protein; however, due to the presence of the unproductive protein, the growth rate is proportional to the productive fraction:

$$k = k_{\text{ideal}} \Phi_*. \tag{A23}$$

The optimal protein fractions can be determined using Eq. A18:

$$\Phi_i = \tfrac{k_{\text{ideal}}}{k_i} \Phi_* + \Phi_{0i}. \tag{A24}$$

How does the growth rate change when the unproductive protein fraction is changed by $\delta\Phi$? The productive fraction is reduced:

$$\Phi_* \to \Phi_*' = \Phi_* - \delta\Phi. \tag{A25}$$

The ratio of the new growth rate k' to the original is therefore:

$$\tfrac{k'}{k} = 1 - \tfrac{\delta\Phi}{\Phi_*.} \tag{A26}$$

Note that our generalized Hwa model is written for arbitrary number of functional sectors $i \neq Q$ and the key determinant of the change in growth rate is the fraction of productive protein Φ_*. This equation will be used to model the fitness cost of the metabolic load.

b. *The metabolic load of mRNA*

What is the cost of transcription? It is perhaps useful to first consider the estimates of biosynthetic cost of macromolecules in the cell in descending order in *E. coli* [38]:

Macromolecule	**Biosynthetic cost** (10^9 ATP)
Protein	4.5
Phospholidid	3.2
RNA	1.6
Lipopolysacchride	3.8
DNA	0.35
Peptidoglycan	0.17
Glycogen	0.03

So clearly the cost of RNA is itself not insignificant. Although a significant fraction of the RNA is rRNA rather than mRNA, the mRNA itself in *E. coli* undergoes multiple rounds of transcription due to its short lifetime, increasing its cost to what is required to synthesize the molecules observed in the *E. coli* cell at any time t. Furthermore, transcription is dependent on protein enzymes, which themselves must be synthesized. We therefore conclude from this estimate that the cost of transcription is likely a significant determinant of the metabolic load.

Experimentally, Kafri and coworkers have measured the fitness cost of transcription and translation independently using the DAmP (Decreased Abundance by mRNA Perturbation) system in *Saccharomyces cerevisiae* [39]. As expected, they report that the metabolic cost of transcription is comparable to translation and that the reduction in growth rate is linear in transcription, in close analogy to Eq. A26.

c. *Metabolic load in the RLTO model*

To produce a minimal model to study the trade-off between robustness and metabolic load, we must consider both the metabolic cost of transcription and translation. We will write that the metabolic load (in protein equivalents) associated with gene i is:

$$\delta N_i = \lambda \mu_{m,i} + \mu_{p,i}, \tag{A27}$$

where λ is the message cost, the metabolic load associated with an mRNA molecule relative to a single protein molecule of the gene product. $\mu_{m,i}$ is the mean number of messages transcribed per cell cycle (mRNA molecules per cell cycle) for gene i. $\mu_{p,i}$ is the mean number of protein translated per cell cycle for gene i. We will describe the mean protein number in terms of the translation efficiency ε_i, the number of proteins translated per message:

$$\mu_{p,i} = \varepsilon_i \mu_{m,i}. \tag{A28}$$

How does the cell growth rate change due to the metabolic load associated with the expression of gene i? The change in the metabolic load is:

$$\delta\Phi_i = \tfrac{\delta N_i}{N}, \tag{A29}$$

where N represents the total metabolic load of all components of the cell, in units of protein equivalents. Using Eq. A26, the resulting change in growth rate is:

$$\tfrac{k}{k_0} = 1 - \tfrac{(\lambda + \varepsilon_i)\mu_{m,i}}{\Phi_* N}, \tag{A30}$$

where k_0 is the growth rate in the absence of the metabolic load of gene i.

In our analysis, the exact size of the total metabolic load N will not be important. In the interest of simplicity we will therefore adsorb the productive fraction Φ_* into an effective total metabolic load:

$$N_0 \equiv \Phi_* N, \tag{A31}$$

and write a concise relation between the load from gene i and the growth rate:

$$\frac{k}{k_0} = 1 - \frac{(\lambda+\varepsilon_i)\mu_{m,i}}{N_0}. \tag{A32}$$

This equation has an intuitive interpretation: growth slows in proportion to the relative added metabolic load. Since Φ_* is order unity, we will ignore the distinction between the N and N_0 quantities hence forth.

Although the global parameters N_0 and λ provide an intuitive representation of the model, the relative growth rate depends on fewer parameters. Let k and k_0 be the growth rates in the presence and absence of the metabolic load of gene i. The relative growth rate is:

$$\frac{k}{k_0} = 1 - (\Lambda + E_i)\mu_{m,i}, \tag{A33}$$

where we have introduced two new reduced parameters: the relative load, defined as $\Lambda \equiv \lambda/N_0$, represents the ratio of the metabolic load of a single message to the total load and the relative translation efficiency, defined $E_i \equiv \varepsilon_i/N_0$, which is the ratio of the number of proteins translated per message to the total metabolic load N_0. (Note that due to the high multiplicity $N_0 \gg (\lambda+\varepsilon_i)\mu_{m,i}$, we can ignore the distinction between N_0 and N_0' in the denominator.) If we neglect the difference between the total metabolic load and the number of proteins, the proteome fraction for gene i is:

$$\Phi_i = E_i\mu_{m,i}. \tag{A34}$$

Both reduced parameters, Λ and E_i are extremely small. In *E. coli*, we estimate that both Λ and E_i are roughly 10^{-5} and they are smaller still for eukaryotic cells. (See Sec. A 18.)

d. Growth rate with stochastic arrest

For completeness, we provide a derivation of the growth rate with stochastic cell-cycle arrest that we have previously described [29]. Starting from the exponential mean expression for the population growth rate [29]:

$$k = \frac{\ln 2}{\overline{T}}, \tag{A35}$$

where k is the population growth rate and

$$\overline{T} \equiv -\frac{1}{k}\ln \mathbb{E}_T \exp(-kT), \tag{A36}$$

is the exponential mean, where T is the stochastic cell cycle duration and $\mathbb{E}$ is the expectation operator [29, 40].

Let P_+ be the probability of growth. When the cells are growing, the cell cycle duration τ is determined by the metabolic load predictions (Eq. A32). The probability mass function is therefore:

$$p_T(t) = \begin{cases} P_+, & t = \tau \\ (1-P_+), & t \to \infty \end{cases}. \tag{A37}$$

Evaluating the expectation in Eq. A36 gives:

$$\overline{T} = -k\ln P_+ + \tau. \tag{A38}$$

Using Eq. A35, we can solve for the growth rate k:

$$k = \tau^{-1}\ln(2P_+). \tag{A39}$$

As expected, the growth rate goes down as the probability of growth P_+ decreases, stopping completely at $P_+ = \frac{1}{2}$. We can then compute the ratio of the growth with (k) and without arrest (k_0):

$$\frac{k}{k_0} = 1 + \frac{1}{\ln 2}\ln P_+. \tag{A40}$$

where k_0 is computed by evaluating Eq. A39 at $P_+ = 1$.

e. RLTO growth rate

In the RLTO model, we will assume the probability of growth is the probability that all essential protein numbers are above threshold. We will further assume that each protein number is independent, and therefore:

$$P_+ = \prod_{i\in\mathscr{E}} \Pr\{N_{p,i} > n_{p,i}\}, \tag{A41}$$

where $\mathscr{E}$ is the set of essential genes. Clearly, this assumption of independence fails in the context of polycistronic messages. We will discuss the significance of this feature of bacterial cells elsewhere, but we will ignore it in the current context.

As we will discuss, the probability of arrest of any protein i to be above threshold is extremely small. It is therefore convenient to work in terms of the CDFs which are very close to zero:

$$\begin{aligned} \ln P_+ &= \sum_{i\in\mathscr{E}} \ln\left(1 - \Pr\{N_{p,i} < n_{p,i}\}\right), && \text{(A42)} \\ &\approx -\sum_{i\in\mathscr{E}} \Pr\{N_{p,i} < n_{p,i}\}, && \text{(A43)} \\ &= -\sum_{i\in\mathscr{E}} \gamma\left(\frac{\mu_{m,i}}{\ln 2}, \frac{n_{p,i}}{\varepsilon_i \ln 2}\right), && \text{(A44)} \end{aligned}$$

where γ is the regularized incomplete gamma function and the CDF of the gamma distribution (see A 1 b and A 1 d).

f. When the ln approximation is avoided...

The approximation discussed in the previous section is extremely well justified at the optimal central dogma parameters; however, there are a set of figures where we cannot use it. In the fitness landscape figures (Fig. S2), we compute the fitness not just at the optimal values but far from them. When cell arrest has a large effect on the growth rate, we cannot approximate the natural log with a series expansion, and we must use the full expression in Eq. A42.

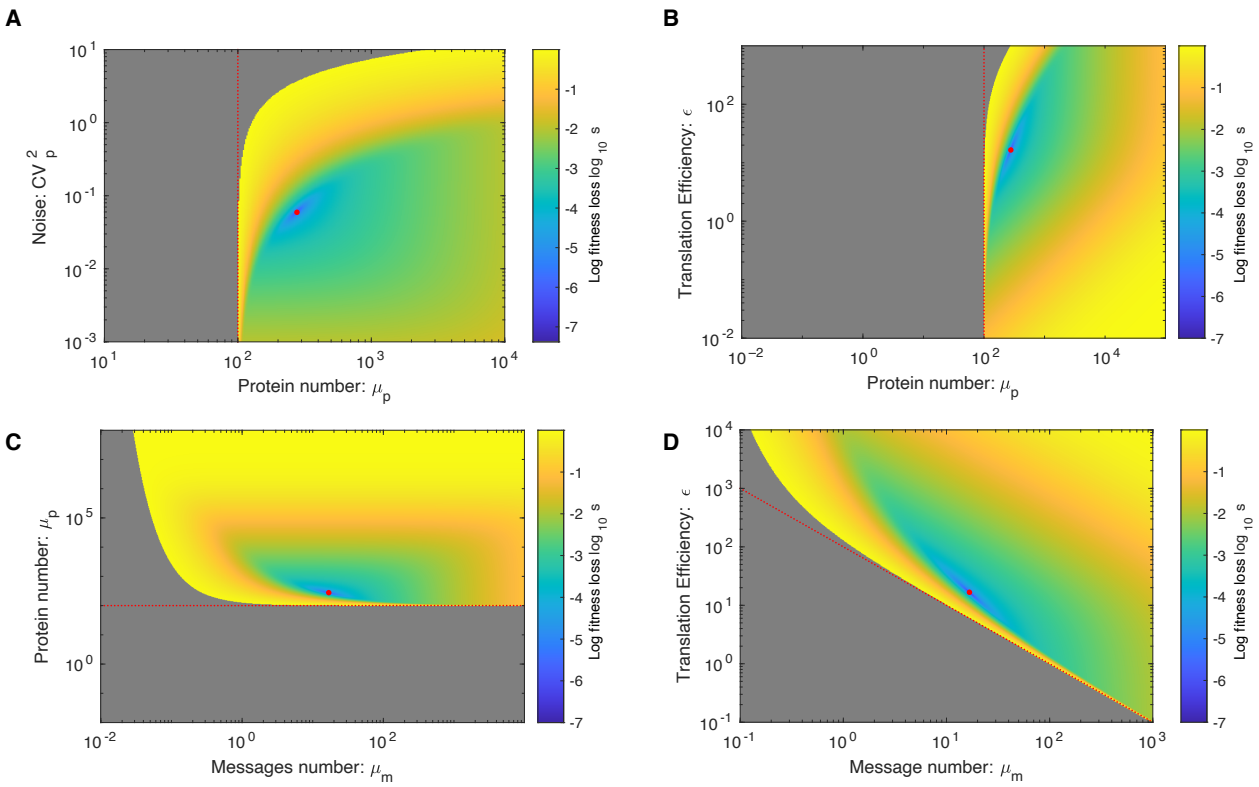

FIG. S2. **Four perspectives on the fitness landscape.** In each landscape, the red circle represents the fitness optimum. The red dotted line represents the mean protein number equal to the protein threshold $n_p = 10^2$. Here, fitness is quantified by the log growth rate: $s = \ln k_{\max}/k$. **Panel A: Mean protein number versus noise. Panel B: Mean protein number versus translation efficiency. Panel C: Mean message versus protein numbers. Panel D: Message number versus translation efficiency.**

g. *Single-gene equation*

By summing the fitness losses from the metabolic load and cell arrest (Eqs. A32, A40, and A44), we can write an expression for the growth rate including contributions from essential gene i:

$$\begin{aligned}\frac{k}{k_0} &= 1 - \frac{1}{N_0}(\lambda + \varepsilon_i)\mu_{m,i} + \ldots \\ &\quad - \frac{1}{\ln 2}\gamma\left(\frac{\mu_{m,i}}{\ln 2}, \frac{n_{p,i}}{\varepsilon \ln 2}\right),\end{aligned} \tag{A45}$$

where the second term on the RHS represents the fitness loss due to the metabolic load and the third term represents the fitness loss due to stochastic cell arrest due to protein i falling below threshold. The fitness landscape for different gene expression parameters is shown from four different perspectives in Fig. S2. From this point forward, we will drop the subscript i for the sake of brevity unless otherwise noted.

h. *Summary of RLTO parameter values for figures*

The parameter values for the RLTO model used for each figure in the paper are shown in Tab. S1.

3. Discussion: The fitness landscape of a trade-off.

The fitness landscape predicted by the RLTO model for representative parameters is shown in Fig. S2. The figure displays a number of important model phenomena: There is no growth for mean protein number μ_p below the threshold number n_p, and for high noise, μ_p must be in significant excess of n_p. Rapid growth can be achieved by the two mechanisms: (i) high expression levels (μ_p) are required for high noise amplitude (CV^2_p) or (ii) lower expression levels coupled with lower noise. This trade-off leads to a ridge-like feature of nearly optimal models. The optimal fitness corresponds to a balance between increasing the mean protein number (μ_p) and decreasing noise (CV^2_p). This optimal central dogma program strategy leads to significant overabundance.

4. Methods: Central dogma optimization

a. *Optimization of transcription and translation (eukaryotes)*

The relative growth rate is:

$$\frac{\Delta k}{k_0} = -(\Lambda + \frac{\varepsilon}{N_0})\mu_m - \frac{1}{\ln 2}\gamma\left(\frac{\mu_m}{\ln 2}, \frac{n_p}{\varepsilon \ln 2}\right), \tag{A46}$$

Figure number and panel:	Relative load: Λ (No units)	Protein cost: λ (molecules^{-1})	Protein threshold: n_p (molecules)	Message number: μ_m (molecules/cell cycle)	Translation efficiency: ε (No units)	Noise floor C_0 (No units)
2A	10^{-5}	10	10^2	Range	Range	0
2C	Range	NA	Range	Local optimum	Local optimum	0
4A	Weak dependence (10^{-5})	NA	Range	Local optimum	Local optimum	0
4B	Weak dependence (10^{-5})	NA	Range	Local optimum	Local optimum	0
4C	Weak dependence (10^{-5})	NA	Range	Local optimum	Local optimum	0
4D	Weak dependence (10^{-5})	NA	Range	Local optimum	Local optimum	0
5	Weak dependence (10^{-5})	NA	Range	Local optimum	Local optimum	0
S2A	10^{-5}	10	10^2	Range	Range	0
S2B	10^{-5}	10	10^2	Range	Range	0
S2C	10^{-5}	10	10^2	Range	Range	0
S2D	10^{-5}	10	10^2	Range	Range	0
S3A	Range	10^2	Range	Local optimum	Local optimum	0
S3B	Range	10^2	Range	Local optimum	Local optimum	0
S3C	Range	10^2	Range	Local optimum	Local optimum	0
S3D	Range	10^2	Range	Local optimum	Local optimum	0
S4A	Range	10^2	Range	Local optimum	Local optimum	0
S4B	Range	10^2	Range	Local optimum	Local optimum	0
S4C	Range	10^2	Range	Local optimum	Local optimum	0
S4D	Range	10^2	Range	Local optimum	Local optimum	0
S5	10^{-5}	NA	Range	Local optimum	30	0.1
S6A	Range	NA	Range	Local optimum	Local optimum	0
S7	Weak dependence (10^{-5})	NA	Range	Local optimum	Local optimum	0
S9A	Range	NA	Range	Local optimum	Local optimum	0
S12B	Weak dependence (10^{-5})	NA	Range	Local optimum	Local optimum	0

TABLE S1. **RLTO model parameters by figure.** *Range* appears if a range of parameters is used. *NA* appears if the parameter value is irrelevant. *Local optimum* appears if the parameter values in optimized to maximize the growth rate.

where γ is the regularized incomplete gamma function, which is the CDF of the gamma distribution and represents the probability of arrest due to gene i. (Note that this equation is identical to Eq. A45 but with the gene subscript i implicit.) We set the partial derivative of the growth rate with respect to message number equal to zero:

$$0 = -\frac{\lambda+\hat{\varepsilon}}{N_0} - \frac{1}{(\ln 2)^2}\gamma_{,1}\left(\frac{\hat{\mu}_m}{\ln 2}, \frac{n_p}{\hat{\varepsilon}\ln 2}\right), \tag{A47}$$

where we use the canonical comma notation to show which argument of γ has been differentiated. Next we differentiate with respect to the translation efficiency to generate a second optimization condition:

$$0 = -\frac{\hat{\mu}_m}{N_0} + \frac{n_p}{\hat{\varepsilon}^2(\ln 2)^2}\gamma_{,2}\left(\frac{\hat{\mu}_m}{\ln 2}, \frac{n_p}{\hat{\varepsilon}\ln 2}\right). \tag{A48}$$

We will work in the large multiplicity limit where the overall metabolic load is much smaller than the metabolic load associated with any single gene: $N_0 \gg (\lambda+\hat{\varepsilon})\hat{\mu}_m$. Next, we eliminate the threshold n_p in favor of the optimal overabundance:

$$\hat{o} \equiv \frac{\hat{\mu}_p}{n_p} = \frac{\hat{\varepsilon}\hat{\mu}_m}{n_p}, \tag{A49}$$

in both Eqs. A47 and A48. Eq. A48 can now be solved for the optimal translation efficiency:

$$\hat{\varepsilon} = \frac{N_0}{\hat{o}(\ln 2)^2}\gamma_{,2}\left(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o}\ln 2}\right). \tag{A50}$$

If we reinterpret γ as the CDF of the gamma distribution, we can rewrite this equation in terms of the gamma distribution PDF:

$$\hat{\varepsilon} = \frac{N_0}{\ln 2}p_\Gamma\left(\mu_m \middle| \frac{\hat{\mu}_m}{\ln 2}, \hat{o}\ln 2\right), \tag{A51}$$

which will be the optimization equation for the translation efficiency.

To derive the optimization condition for the message number μ_m, we substitute Eq. A50 into Eq. A47:

$$\frac{\lambda\ln 2}{N_0} = -\frac{1}{\hat{o}\ln 2}\gamma_{,2}\left(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o}\ln 2}\right) - \frac{1}{\ln 2}\gamma_{,1}\left(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o}\ln 2}\right). \tag{A52}$$

The two terms on the RHS can now be collected as the single partial derivative of message number μ_m:

$$\Lambda\ln 2 = -\partial_{\hat{\mu}_m}\gamma\left(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o}\ln 2}\right), \tag{A53}$$

where the relative load is $\Lambda \equiv \lambda/N_0$.

The two optimization equations are summarized below:

$$\Lambda\ln 2 = -\partial_{\hat{\mu}_m}\gamma\left(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o}\ln 2}\right), \tag{A54}$$

$$\frac{\hat{E}}{\Lambda} = \frac{\hat{\varepsilon}}{\lambda} = \frac{1}{\Lambda\ln 2}p_\Gamma\left(\hat{\mu}_m \middle| \frac{\hat{\mu}_m}{\ln 2}, \hat{o}\ln 2\right). \tag{A55}$$

The optimal overabundance is shown for a range of relative loads in Fig. S3. The optimal translation efficiency and scaled translation efficiency are shown for a range of relative loads in Fig. S4.

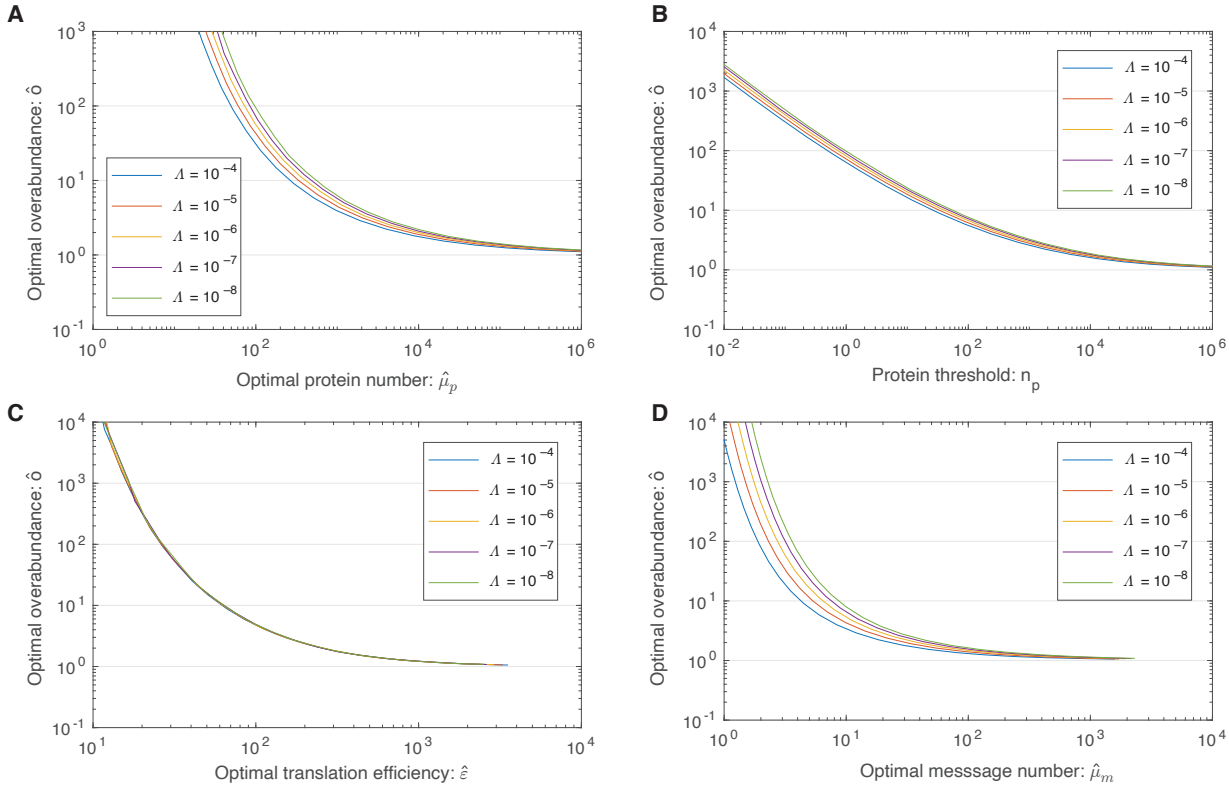

FIG. S3. **Four perspectives on the dependence of optimal overabundance on relative load Λ.** All these calculations are performed for protein cost $\lambda = 100$ in order to give real numbers in molecules per cell. **Panel A: Overabundance as a function of protein number.** Overabundance decreases as protein number increases. These calculations are λ *dependent*. **Panel B: Overabundance as a function of the protein threshold.** Overabundance decreases as protein threshold increases. These calculations are λ *dependent*. **Panel C: Overabundance as a function of translation efficiency.** Overabundance decreases as translation efficiency increases. These calculations are λ *dependent*. **Panel D: Overabundance as a function of message number.** Overabundance decreases as message number increases. These calculations are λ *independent*.

b. *Optimization of message number only*

Consider the special case of optimizing the message number only at fixed translation efficiency. Eq. A47 is the condition; however, in this case it makes sense to adsorb both the message and protein metabolic load into a single metabolic load. The optimum message number satisfies the equation:

$$\frac{(\lambda+\varepsilon)\ln 2}{N_0} = -[\partial_{\hat{\mu}_m}\gamma(\hat{\mu}_m, \hat{n}_m)]_{\hat{n}_m = \frac{\hat{\mu}_m}{\hat{o}}}. \tag{A56}$$

We define a modified relative load:

$$\Lambda' \equiv \frac{(\lambda+\varepsilon)}{N_0}, \tag{A57}$$

and substitute this into the optimum message number equation:

$$\Lambda' \ln 2 = -[\partial_{\hat{\mu}_m}\gamma(\hat{\mu}_m, \hat{n}_m)]_{\hat{n}_m = \frac{\hat{\mu}_m}{\hat{o}}}. \tag{A58}$$

which is clearly closely related to Eq. A53.

We compare this modified expression to the original for optimum overabundance as a function of message number in Fig. 2C and demonstrate that the two make nearly identical predictions.

c. *Inclusion of the noise floor*

In bacterial cells, the noise is dominated by the noise floor for high expression genes. Including the noise floor, the coefficient of variation squared is ([9] and Sec. D):

$$\mathrm{CV}_p^2 = \tilde{a}(\mu_m)^{-1} = \frac{\ln 2}{\mu_m} + C_0, \tag{A59}$$

where $C_0 = 0.1$ for bacterial cells [9]. In spite of the addition of noise from the noise floor, the observed distribution of protein number is still well described by the gamma distribution [9]; however, we need to modify the statistical parameters to account for the noise floor. (See the definition of the *statistical noise model* in Sec. A 1 b.) The modified gamma parameters are:

$$a = \tilde{a}, \tag{A60}$$

$$\theta = \varepsilon\frac{\mu_m}{\tilde{a}}, \tag{A61}$$

chosen such that the noise is determined by Eq. A59 but the protein number remains:

$$\mu_p = \varepsilon\mu_m, \tag{A62}$$

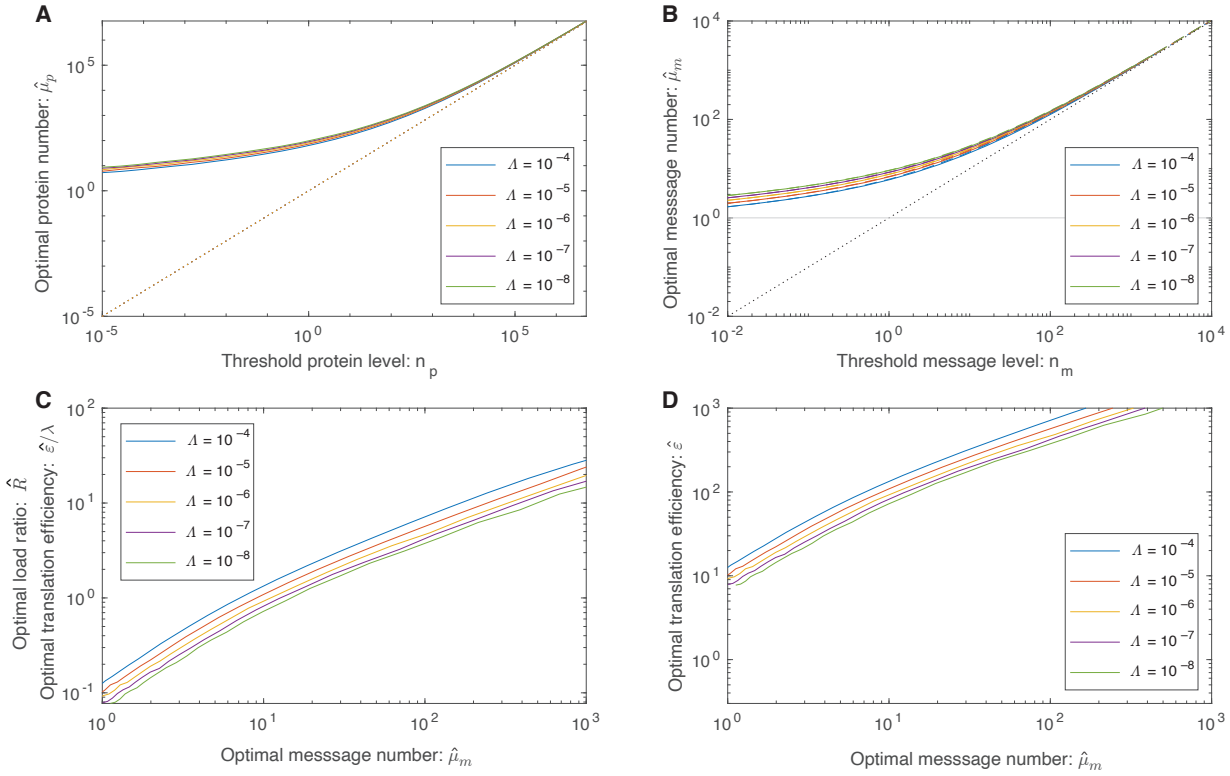

FIG. S4. **Four perspectives on load balancing.** All these calculations are performed for protein cost $\lambda = 100$ in order to give real numbers in molecules per cell. **Panel A: Protein number versus protein threshold.** At high expression levels, the protein number tracks the protein threshold; however, the one message rule forces the protein number to threshold for low expression levels. These calculations are λ *dependent*. **Panel B: Message number versus message threshold.** At high expression levels, the message number tracks the message threshold; however, the one message rule forces the message number to a threshold close to $\mu_m = 1$ for low expression levels. These calculations are λ *independent*. **Panel C & D: Message number versus translation efficiency.** The optimal translation efficiency grows almost linearly with the optimal message number. The scaled translation efficiency ($\hat{\varepsilon}/\lambda$) is independent of λ while the translation efficiency ($\hat{\varepsilon}$) is dependent on λ. The ratio ε/λ has a second interpretation: the load ratio R. R is defined as the metabolic cost of translation over transcription of the gene.

the product of the message number and translation efficiency.

The qualitative effect of the noise floor is to increase the noise, especially for low-copy messages. Above $\mu_m = 7$ messages, the noise is dominated by the noise floor. Increases in transcription above this point have little effect on reducing the noise. As a consequence, the overabundance stays high, even for high copy messages. We compare this modified expression to the original for optimum overabundance as a function of message number in Fig. 2C and demonstrate that bacterial cells are predicted to have much higher overabundance at high expression levels.

5. Discussion: Understanding the rationale for overabundance

Essential protein overabundance is the signature prediction of the RLTO model. Its mathematical rationale is the highly-asymmetric fitness landscape. To understand why we expect this rationale to be generic, consider the form of the optimization condition for message number:

$$\Lambda \ln 2 = -\frac{\partial}{\partial \hat{\mu}_m} \gamma\left(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{o \ln 2}\right). \tag{A63}$$

The growth rate is maximized when the probability of slow-growth (*e.g.* arrest) is roughly equal to the relative load of adding one more message. Since the cell makes roughly 10^5 messages per cell cycle, the relative load is extremely small and therefore the probability of slow growth must be as well. Making this probability very small requires vast overabundance for the inherently-noisy, low-expression proteins.

The reason we expect the RLTO model protein abundance predictions to be robust is that we generically expect the fitness cost of overabundance to be small due to the high multiplicity (*i.e.* total number of genes); whereas, the fitness cost of arrest of essential processes is very high.

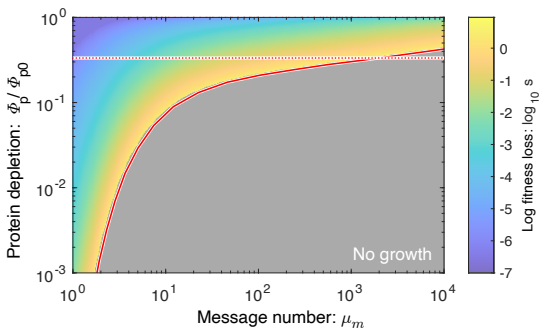

FIG. S5. **Optimal expression levels are buffered.** The predicted fitness loss as a function of protein depletion level and message number for bacterial cells (including the noise floor). Due to the overabundance phenomenon, all proteins are buffered against depletion, but low-expression genes are particularly robust due to higher overabundance. The solid red line represents $1/o$, and predicts the range of depletion values for which cell growth is predicted. The dotted red line represents a three-fold depletion.

6. Discussion: RLTO predicts larger overabundance in bacteria.

There are two distinctive features of bacterial cells that could affect the model predictions: (i) the translation efficiency is constant [22] and bacterial gene expression is subject to a large-magnitude noise floor that increases the noise for high-expression genes [9]. The optimization of message number at fixed translation efficiency does result in a slightly modified optimization condition for the message number (Sec. A 4 b); however, the predicted overabundance is only subtly perturbed (Fig. 2C). In contrast, the noise floor increases the predicted overabundance, especially for high-expression proteins. As a result, the RLTO model predicts that the vast majority of bacterial proteins are expressed in significant overabundance. (See Fig. 2C.)

7. Discussion: RLTO predicts proteins are buffered to depletion.

A principle motivation for our analysis is the observation that many protein levels appear to be buffered. To explore the prediction of the RLTO model for protein depletion, we first computed the optimal message numbers and translation efficiencies for a range of protein thresholds. To model the effect of protein depletion, we computed the change in growth rate as function of protein depletion (equivalent to a reduction of the translation efficiency relative to the optimum.) The growth rate is shown in Fig. S5 for the RLTO model with parameters representative of a bacterial cell. (See Sec. A 4 c.)

In general, the RLTO model predicts that protein numbers have very significant robustness (*i.e.* buffering) to protein depletion. This is especially true for low expression proteins that are predicted to have the largest overabundance. For these genes, even a ten-fold depletion leads to very subtle reductions in the growth rate. For a three-fold reduction in the growth rate, only the very highest-expression genes (*e.g.* ribosomal genes) are expected to lead to qualitative phenotypes.

8. Results: Detailed development of load balancing

a. *Prediction of the optimal load ratio.*

The two-stage amplification of the central dogma implies that the expression and noise levels can be controlled independently by the balance of transcription to translation. How does the cell achieve high and low gene expression optimally, and how does this strategy depend on the message cost?

To understand the optimization, we first define the load ratio R for a gene as the metabolic cost of translation relative to transcription:

$$R \equiv \frac{\mu_p}{\lambda \mu_m} = \frac{\varepsilon}{\lambda}. \tag{A64}$$

In Sec. A 4, we show that the optimal load ratio is:

$$\hat{R} \equiv \frac{1}{\Lambda \ln 2}\, p_\Gamma(\hat{\mu}_m | \tfrac{\hat{\mu}_m}{\ln 2}, \tfrac{1}{\delta \ln 2}), \tag{A65}$$

where p_Γ is the PDF of the gamma distribution. The optimal load ratio is shown in Fig. S4C.

The dependence of the optimal load ratio $\hat{R}$ on Λ is extremely weak, but it is strongly dependent on message number. As a result, for low transcription genes ($\mu_m \ll 10$), the metabolic load is predicted to be dominated by transcription; whereas, for highly transcribed genes ($\mu_m \gg 10$), the metabolic load is dominated by translation. These predictions are robust since they are independent of the relative load Λ.

b. *Measurements of the load ratio*

Unfortunately, there is somewhat limited data to which to compare the model. The best source we found was Kafri *et al.* [41] who analyzed the differences in fitness between transcription and transcriptional-and-translation of a fluorescent protein driven by the pTDH3 promoter in yeast. This promoter is one of the strongest in yeast. Based on the RLTO model, we would predict this promoter to have a very high translation efficiency and therefore a large load ratio; however, the translation efficiency is much lower than one would predict based on a global analysis and likewise its load ratio is roughly unity, which based on the smaller than expected translation efficiency is broadly consistent with our expectations. A satisfactory test of this prediction will require

larger-scale measurements that probe more representative genes.

9. Results: Translation efficiency is predicted to increase with transcription.

Now that we have defined the optimal load ratio (Eq. A65), the equation for optimal translation efficiency can be written concisely:

$$\hat{\varepsilon} = \lambda \hat{R} \quad \text{or} \quad \hat{E} = \Lambda \hat{R}, \tag{A66}$$

where $\hat{R}$ depends weakly on the relative load Λ. The RLTO model predicts that optimal partitioning of amplification between transcription (gain μ_m) and translation (gain ε) has two important qualitative features: (i) As the message cost (λ) rises, the optimal translation efficiency increases in proportion. (ii) The optimal translation efficiency is also approximately proportional to message number ($\hat{\varepsilon} \propto \mu_m$). (See Fig. 4A.) Therefore, the RLTO model predicts that low expression levels should be achieved with low levels of transcription and translation, whereas high expression genes are achieved with high levels of both. We call this relation between optimal transcription and translation *load balancing*.

10. Results: RLTO predicts that message number responds to message cost

We will first focus on analyzing the implications of the message cost dependence in Eq. A66. At a fixed load ratio, Eq. A66 clearly implies that the translation efficiency increases as the message cost λ increases; however, the message number (and load ratio) also respond to compensate to changes in λ. To probe the dependence on message cost in an experimentally relevant context, consider optimal message numbers in a reference condition (relative load Λ_0) relative to a second perturbed condition (relative load Λ). The predicted relation between the optimal messages numbers is shown in Fig. S6A. The resulting relation between the optimal message numbers is roughly linear on a log-log plot, predicting the approximate power-law relation:

$$\hat{\mu}_m(\Lambda) \propto \hat{\mu}_m(\Lambda_0)^{\alpha}, \tag{A67}$$

describing a non-trivial global change in the regulatory program.

11. Results: RLTO predicts the yeast global regulatory response

To test the RLTO predictions, we compared the relative message numbers for yeast growing under phosphate depletion, which increases the message cost [41], to a reference condition [42]. As predicted, the relative transcriptome data was well described by a power law (Eq. A67) and the observed slope was smaller than one: $\hat{\alpha} = 0.837 \pm 0.001$, as predicted by the increased message cost. See Fig. S6B.

The observation of this large-scale regulatory change has an important implication: This response supports a nontrivial hypothesis that the RLTO model not only can predict how the cell is optimized in an evolutionary sense, but can predict global regulatory responses as well.

Note that Metzl-Raz *et al.* [42] also explored conditions that increased the cost of protein; however, here the predictions of the model are ambiguous. The complication arises due to the observed decrease in size of the cells in the experimental condition, which decreases the total metabolic load N_0. As discussed above, the relative load $\Lambda = \lambda/N_0$ is the key determinant in Eq. A67; however, even as the relative cost of transcription λ decreases in the experimental condition, the total metabolic load N_0 also decreases, making no clear prediction about how the relative load Λ changes.

12. Methods: Prediction of the proteome fraction

a. *Parameter-free prediction of proteome fraction.*

We now turn our focus to an analysis of the implications of *load balancing*: the message number dependence of the optimal translation efficiency (Eq. A66). The most direct test of this prediction is measuring the relation between proteome fraction and message number. The RLTO model predicts proteome fraction (Eq. A34):

$$\hat{\Phi}_p = \hat{E}\hat{\mu}_m \propto \hat{\mu}_m^2, \tag{A68}$$

where μ_m is the observed message number and the optimal relative translation efficiency is predicted by Eq. A66. The proportionality is only approximate but gives important intuition for how protein number depends on message number in the RLTO model, in contrast to a constant-translation-efficiency model: $\Phi_p \propto \mu_m$. To compare these predictions to protein abundance measurements, we will renormalize the protein fraction to be defined relative to total protein number rather than N_0. This renormalization eliminates the Λ dependence to result in a parameter-free prediction of the proteome fraction.

b. *RLTO: proteome fraction*

Starting from Eq. A96, clearly:

$$\hat{\mu}_p \propto \hat{\mu}_m \, p_\Gamma(\hat{\mu}_m | \tfrac{\hat{\mu}_m}{\ln 2}, \hat{o} \ln 2), \tag{A69}$$

which can be used to predict the proteome fraction (where we have restored the explicit gene i subscript):

$$\hat{\Phi}_{p,i} \equiv \tfrac{\hat{\mu}_{p,i}}{\sum_j \hat{\mu}_{p,j}}, \tag{A70}$$

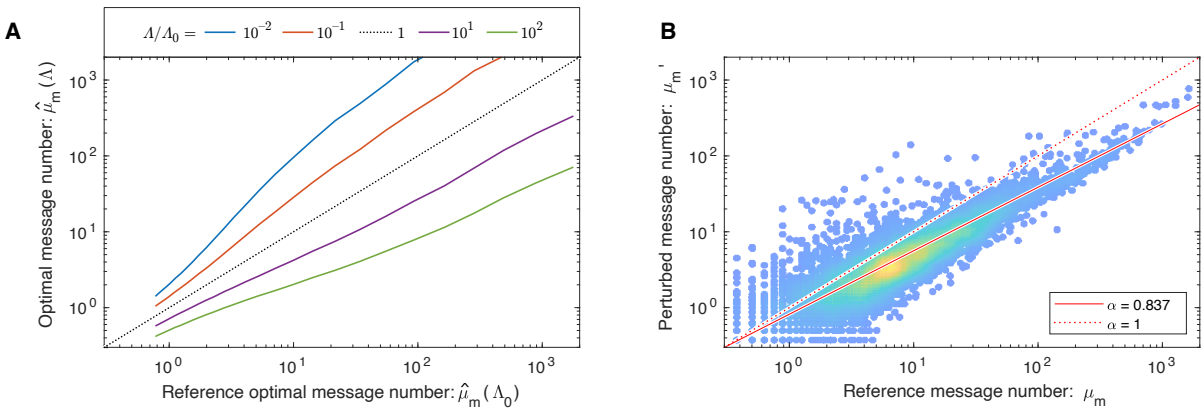

FIG. S6. **The message cost affects transciption genome wide. Panel A: Message number decreases with increased relative load Λ.** The optimal message number responds to changes in the message cost. The RLTO model predicts an approximate power-law relation (linear on a log-log plot) between message numbers. **Panel B: A power-law relation is observed.** To test whether central dogma regulation would adapt dynamically as predicted, we analyzed the relation between the yeast transcriptome under reference conditions and phosphate depletion (perturbed), which increases the message cost [41]. (Data from Ref. [42].) As predicted by the RLTO model, a global change in regulation is observed, which generates a power-law relation with scaling exponent $\alpha = 0.837 \pm 0.01$. The observed exponent is smaller than one, as predicted by an increased relative load Λ.

where the second subscript is the gene index. To predict the proteome fraction, we computed the proportionality constant C:

$$C \equiv \sum_i \left[\mu_{m,i} p_\Gamma(\mu_m | \tfrac{\mu_{m,i}}{\ln 2}, o \ln 2)|_{o=\hat{o}(\mu_{m,i})}\right], \quad \text{(A71)}$$

where the message numbers μ_{mi} for gene i are the experimentally observed message numbers, the implicit o_i values are predicted by the RLTO model (Eq. A53) for message number μ_{mi} and the sum index i runs over all genes. The predicted optimal proteome fraction is:

$$\hat{\Phi}_{p,i} = C^{-1}\left[\mu_{m,i} p_\Gamma(\mu_{m,i} | \frac{\mu_{m,i}}{\ln 2}, o \ln 2)|_{o=\hat{o}(\mu_{m,i})}\right], \quad \text{(A72)}$$

which generates the predicted solid curves shown in Fig. 4BCD.

c. *Constant-translation-efficiency model: proteome fraction*

For the constant translation efficiency model, we define the normalization:

$$C' \equiv \sum_i \mu_{m,i}, \quad \text{(A73)}$$

and the predicted proteome fraction is:

$$\Phi'_{p,i} = C'^{-1} \mu_{m,i}, \quad \text{(A74)}$$

which generates the predicted dotted curves shown in Fig. 4BCD.

d. *Sources of experimental data for proteome fraction analysis*

For *E. coli* data, the protein abundance data was generated by mass spec measurements and the message abundance data was from [22]. For the yeast data, the protein abundance data is measured by mass spec and message abundances are determined by [21]. For the mammalian data, we used mouse data. The protein abundance data is measured by mass-spec and message abundances are determined by [23].

We estimated the message number μ_m as described in Sec. C 2 c. For the mouse data, the study provided message lifetimes, the cell cycle duration and abundances in molecules per cell [23]. For the *E. coli* and yeast, the total number of proteins, messages *etc*, cell cycle duration and message lifetimes for each organism and their sources are described in Tab. S2.

13. Results: Load balancing is observed in eukaryotic cells

A non-trivial prediction of the RLTO Model is that translation efficiency and message number should be roughly proportional. Qualitatively, this strategy allows expression levels to be increased while distributing the added metabolic load between transcription, which reduces noise, and translation, which does not affect the noise. We predict the optimal translation efficiency versus message number which matches the observations in eukaryotic cells (Fig. 4BC). However, in *E. coli*, the translation efficiency and message number are *not* strongly correlated (Fig. 4D). Why does this organism appear not to load balance? We demonstrate that the observed translation efficiency is consistent with the RLTO model, augmented by a ribosome-per-message limit. (See Sec. A 15.) Hausser *et al.* have proposed just such a limit, based on the ribosome footprint of mRNA molecules [10]. (See Sec. A 17.) Although this augmented model is consistent with central dogma reg-

ulation in *E. coli*, it is not a complete rationale. This proposed translation-rate limit could be circumvented by increasing the lifetime of *E. coli* messages which would increase the translation efficiency. Why the message lifetime is as short as observed will require a more detailed *E. coli*-specific analysis.

14. Discussion: Relation between load balancing and previous results

Hausser *et al.* have previously performed a more limited analysis of the trade-off between metabolic load and gene-expression noise [10]. In this section, we will provide some more context into the differences between the two approaches.

Hausser *et al.* assume a symmetric (not an asymmetric) fitness landscape and consider only the metabolic cost of transcription (but not translation). Their model depends on two (not one) gene-specific parameters: an optimal protein number and a sensitivity, which defines the curvature of the fitness [10]. The authors maximize fitness with respect to the transcription rate (but not the translation rate) and the condition they derive depends on the two (not one) unknown, gene-specific parameters. As a result, this condition is not predictive of global regulatory trends without non-trivial, gene-specific measurements or assumptions about the unknown sensitivity.

The Hausser model assumes that the growth rate has the form:

$$k = k_0 - \tfrac{1}{2}|k''|(N_p - \mu_p)^2 - k_0\Lambda\mu_m, \tag{A75}$$

where k is the growth rate, and we have rewritten the form of the fitness to better match our own definitions. Here k'' is the second derivative of the growth rate at the optimal protein number μ_p and N_p is the stochastic protein number. If we take the expectation with respect to the protein number, we get:

$$k = k_0 - \tfrac{1}{2}|k''|\sigma_p^2 - k_0\Lambda\mu_m, \tag{A76}$$

and substituting the noise model for the variance of the protein number gives:

$$k = k_0 - \tfrac{1}{2}|k''|\mu_p^2(\tfrac{\ln 2}{\mu_m} + C_0) - k_0\Lambda\mu_m, \tag{A77}$$

where C_0 is the noise floor, and we have assumed the mean protein number is optimal (μ_p). If we maximize the growth rate with respect to μ_m, we get the following condition on the optimal message number:

$$\hat{\mu}_m^2 = \tfrac{1}{2}\tfrac{|k''|}{k_0}\mu_p^2\tfrac{\ln 2}{\Lambda}, \tag{A78}$$

which depends on the unknown curvature k''. To make global predictions about how transcription and translation are related, some added assumptions are necessary to describe how k'' scales with protein abundance.

To illustrate how this expression does not make explicit global predictions, let's consider a number of plausible possibilities. First, we will assume that k'' is independent of μ_p and on average all proteins are equally sensitive to changes in protein number. In this case, we find:

$$\mu_p \propto \hat{\mu}_m\sqrt{\Lambda}, \tag{A79}$$
$$\hat{\varepsilon} \propto \sqrt{\Lambda}, \tag{A80}$$

implying a constant translation efficiency which is inversely proportional to the square root of the relative load.

Alternatively, we can assume that $k'' \propto \mu_p^{-2}$ and, on average, the cell is equally sensitive to changes in the relative number of proteins (*i.e.* $\Delta p/\mu_p$), regardless of expression level. In this case,

$$\hat{\mu}_m \propto 1/\sqrt{\Lambda}, \tag{A81}$$
$$\hat{\varepsilon} \propto \mu_p\sqrt{\Lambda}, \tag{A82}$$

implying a constant message number, irrespective of expression level, and a translation efficiency that is proportional to expression level.

Finally, we will assume that $k'' \propto \mu_p^{-1}$, which is the intermediate case. Here:

$$\mu_p \propto \hat{\mu}_m^2\Lambda, \tag{A83}$$
$$\hat{\varepsilon} \propto \hat{\mu}_m\Lambda, \tag{A84}$$

implying that translation efficiency should increase with message number, analogous to our prediction. It appears that Hausser *et al.* implicitly also favor this model, since they define their sensitivity parameter to include a power of protein number μ_p. They justify this assumption by arguing that since $\sigma_p^2 \propto \mu_p$, it makes sense to define the *sensitivity to noise* to include a factor of μ_p [10]. At best, this is somewhat fuzzy logic since, as we demonstrate in the paper, Eq. A84 implies that the protein variance does not scale $\sigma_p^2 \propto \mu_p$!

The authors also propose a lower limit on the translation-transcription ratio; however, their limit is dependent on the noise floor, which only affects genes with the highest transcription rates in eukaryotic cells. The implementation of a more appropriate estimate of the noise, relevant for the vast majority of genes, does not lead to the same limit.

15. Methods: Analysis of translational limits & gene-specific load analysis

To explore the consequences of a protein-specific load, we can modify the metabolic load term in the growth rate equation (Eq. A46):

$$E\mu_m \to \lambda_p E\mu_m, \tag{A85}$$

which includes an additional parameter: the protein cost λ_p, which is 1 if the fitness cost is equal to the

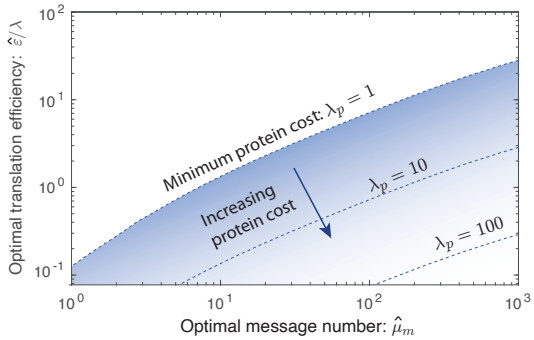

FIG. S7. **Increased protein cost decreases optimal translation efficiency.** A protein cost of $\lambda_p = 1$ corresponds to the metabolic cost of protein synthesis only, and is the minimum protein cost. For larger protein costs, the optimal translation efficiency is lower. As a result, the $\lambda_p = 1$ curve represents an upper bound of the optimal translation efficiency.

metabolic load and greater than one if the cost is higher. We will also treat the metabolic load per message λ as a gene-specific parameter in this section only. The optimization can be repeated for this augmented model.

To analyze the effect of increased protein load, we modify Eq. A46:

$$\ln \frac{k}{k_0} = -(\Lambda + \lambda_p E)\mu_m - \frac{1}{\ln 2}\gamma(\frac{\mu_m}{\ln 2}, \frac{n_p}{\varepsilon \ln 2}), \tag{A86}$$

to contain the supplemental load factor λ_p which is unity if the only protein load is metabolic and $\lambda_p > 1$ if there is additional load (*e.g.* toxicity). The optimization conditions (Eqs. A47 and A48) become:

$$0 = -\frac{\lambda + \lambda_p \hat{\varepsilon}}{N_0} - \frac{1}{(\ln 2)^2}\gamma_{,1}(\frac{\hat{\mu}_m}{\ln 2}, \frac{n_p}{\hat{\varepsilon} \ln 2}), \tag{A87}$$

$$0 = -\frac{\lambda_p \hat{\mu}_m}{N_0} + \frac{n_p}{\hat{\varepsilon}^2 (\ln 2)^2}\gamma_{,2}(\frac{\hat{\mu}_m}{\ln 2}, \frac{n_p}{\hat{\varepsilon} \ln 2}). \tag{A88}$$

Using the same algebraic approach as before, we can derive the same optimal overabundance and load equations:

$$\Lambda \ln 2 = -\partial_{\hat{\mu}_m}\gamma(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o} \ln 2}), \tag{A89}$$

$$\hat{R} = \frac{1}{\Lambda \ln 2} p_\Gamma(\hat{\mu}_m | \frac{\hat{\mu}_m}{\ln 2}, \hat{o} \ln 2); \tag{A90}$$

however, the relation between the load and the translation efficiency now has an extra factor: λ_p:

$$R = \frac{\lambda_p \varepsilon \mu_m}{\lambda \mu_m} = \frac{\lambda_p \varepsilon}{\lambda}, \tag{A91}$$

representing the modified total load ratio.

16. Results: Increased protein-specific cost reduces the optimal translation efficiency.

The relation between the overabundance and message number is unchanged (Eq. A63). This result can be rationalized in the following way: The optimal overabundance is determined by the noise which is determined by message number only. This relation is unaffected by the added parameter λ_p. However, the optimal translation efficiency is affected:

$$\hat{\varepsilon} = \frac{\lambda}{\lambda_p}\hat{R}, \tag{A92}$$

where $\hat{R}$ is the optimal load ratio, defined by Eq. A66. The optimal curves are shown in Fig. S7.

How do these added considerations affect the RLTO predictions? First, we consider message and protein length. What are the optimal translation efficiencies for two proteins, one ten times the length of the other, at fixed protein number? In this case, we will assume that both the transcriptional cost (λ) as well as the translational cost (λ_p) increase tenfold. These increases cancel, resulting in the same optimal translation efficiency since it is only the relative cost of transcription to translation that is determinative of the translation efficiency.

Now consider a tenfold protein-specific increase in protein cost at fixed message cost and fixed protein number. The message number and translation efficiency would change by compensatory factors of 10:

$$\hat{\mu}_m \rightarrow 10 \cdot \hat{\mu}_m, \tag{A93}$$

$$\hat{\varepsilon}_m \rightarrow \frac{1}{10} \cdot \hat{\varepsilon}_m, \tag{A94}$$

to maintain the protein number.

Returning to our original motivation, we can understand how genes with a higher protein-to-message cost migrate downwards and rightwards off the optimal $\lambda_p = 1$ curve, predicting a cloud versus a narrow strip in proteome fraction measurements shown in Fig. 4. If the relative load Λ were directly measured, we would expect the predicted optimal translation efficiency curve for $\lambda_p = 1$ to lie at the top edge of the observed data cloud rather than the bisecting it. This bisection is the consequence of fitting an effective relative load parameter to the abundance data in the unaugmented RLTO model.

17. Discussion: Translation limits in *E. coli*

A critical assumption in the RLTO model to this point has been that the optimal central dogma parameters are realizable in the cell; however, translation can be limited by a number of different mechanisms. The superior performance of the constant- over the optimal-translation-efficiency model in *E. coli* (Fig. 4D) suggests that this assumption may not be satisfied for bacteria. How do translation limits affect the model phenomenology?

When considering possible limits on translation, there are two natural mechanisms: (i) ribosome-number limit, where the number of ribosomes in the cell limits translation and (ii) a ribosome-per-message limit, where the number of ribosomes per message is limiting. Assuming

the ribosome-number-limit mechanism, the original unconstrained optimization problem can be recast as a constrained optimization problem where the protein cost λ_p is reinterpreted as a Lagrange multiplier to constrain the number of proteins translated (*e.g.* [43]). In spite of this reformulation, we would still predict the same functional form for the coupling between the optimal translation efficiency and message number. (*I.e.* it is still optimal to have a higher translation efficiency for highly-expressed genes even if the total number of proteins is fixed.) Therefore, the ribosome-number-limit mechanism cannot be the rationale for the constant translation efficiency observed in *E. coli*.

Assuming the ribosome-per-message-limit mechanism, we limit the translation efficiency to a restricted range of values. If the unconstrained optimum lies above this range, the optimum is at the maximum limiting value. If the unconstrained optima for all genes lie above the realizable range, the model predicts a translation efficiency uncoupled from the message number, as observed. These predictions are consistent with the observed central dogma regulatory program in *E. coli*. In added support of this hypothesis, Hausser *et al.* have argued that *E. coli* translates close to just such a ribosome-per-message limit as a consequence of the finite ribosome complex footprint on a message [10].

18. Methods: Estimate of the message cost and metabolic load

We can estimate the message cost λ from the known total protein number for yeast and mammalian cells. (For *E. coli* this estimate is not possible since the protein cost in not determinative of the translation efficiency.)

The optimal translation efficiency for gene i is (Eq. A51):

$$\hat{\varepsilon} = \frac{\lambda}{\Lambda \ln 2} p_\Gamma(\hat{\mu}_m | \frac{\hat{\mu}_m}{\ln 2}, \hat{o} \ln 2), \tag{A95}$$

and therefore the optimal protein number for gene i is:

$$\hat{\mu}_p = \hat{\mu}_m \hat{\varepsilon} = \frac{\lambda}{\Lambda \ln 2} \hat{\mu}_m p_\Gamma(\hat{\mu}_m | \frac{\hat{\mu}_m}{\ln 2}, \hat{o} \ln 2). \tag{A96}$$

We define the normalization constant A (where we restore the explicit gene i subscript):

$$A = \sum_i \hat{\mu}_{m,i} \cdot \frac{1}{\Lambda \ln 2} p_\Gamma(\hat{\mu}_{m,i} | \frac{\hat{\mu}_{m,i}}{\ln 2}, \hat{o}_i \ln 2), \tag{A97}$$

where we have restored the explicit gene i index running over all genes. Now, by summing Eq. A96, over all genes, we derive an expression for the total protein number N_p^{tot} in terms of the message cost λ and the normalization constant A:

$$N_p^{\text{tot}} = \lambda A. \tag{A98}$$

Solving for the protein cost results in the estimate:

$$\hat{\lambda} = \frac{N_p^{\text{tot}}}{A}. \tag{A99}$$

This message cost estimate $\hat{\lambda}$ can then be plugged into the metabolic load definition:

$$N_0 \equiv L_0 + \sum_i (\lambda + \varepsilon_i) \mu_{m,i}, \tag{A100}$$

to estimate its size:

$$\hat{N}_0 \equiv L_0 + \hat{\lambda} N_m^{\text{tot}} + N_p^{\text{tot}}, \tag{A101}$$

where we have ignored the non-protein and non-message contributions to the load ($L_0 = 0$).

a. Detailed protocol

We first estimate the message numbers, as described in Sec. C 2 c, from data. For each gene i, we set the optimal message number equal to the observed message number and then compute the optimal overabundance from the message number using Eq. A54. (Since the result is independent of the assumed Λ value, we set an arbitrary initial value of $\Lambda = 10^{-5}$.) We then use these single gene optimal message number and overabundances to compute A using Eq. A97. In Eqs. A99 and A101, we use the N_p^{tot} from Tab. S2. N_m^{tot} is computed by summing the estimated message numbers.

b. Estimate the message cost and metabolic load in yeast

In yeast, the estimates are:

$$\begin{aligned} A &= 4.8 \times 10^5, && \text{(A102)} \\ \hat{\lambda} &= 1.0 \times 10^2, && \text{(A103)} \\ \hat{N}_0 &= 6.2 \times 10^7, && \text{(A104)} \\ \hat{\Lambda} &= 1.6 \times 10^{-6}, && \text{(A105)} \end{aligned}$$

where the data sources are described in detail in Sec. C 1 b.

c. Estimate the message cost and metabolic load in human cells

In human cells, the estimates are:

$$\begin{aligned} A &= 4.3 \times 10^6, && \text{(A106)} \\ \hat{\lambda} &= 7.1 \times 10^2, && \text{(A107)} \\ \hat{N}_0 &= 2.4 \times 10^9, && \text{(A108)} \\ \hat{\Lambda} &= 2.9 \times 10^{-7}, && \text{(A109)} \end{aligned}$$

where the data sources are described in detail in Sec. C 1 c.

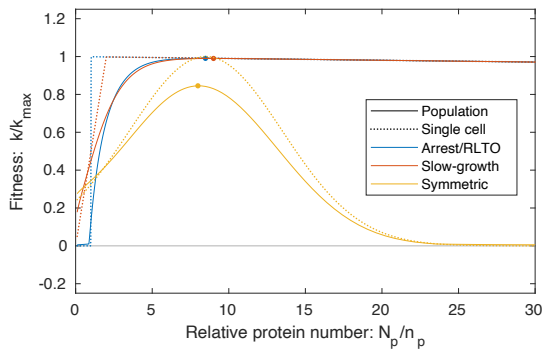

FIG. S8. **Exploring the mathematical mechanism of overabundance.** Single-cell and population growth rate are compared for three different models: arrest (RLTO), slow-growth, and symmetric models. In the arrest model (RLTO), the growth rate goes to zero below threshold protein level n_p. In the slow-growth model, the growth rate transitions continuously to zero as the N_p is depleted below n_p. In both the arrest and slow-growth models, there is a small negative slope above the threshold corresponding to the metabolic load. In the symmetric model, the fitness cost is symmetric about the optimum. Both the threshold-like and slow-growth models are optimized at mean expression levels μ_p far exceeding the threshold level n_p. This is a consequence of the highly-asymmetric dependence of the fitness on protein number N_p. This leads to the phenomenon of protein overabundance. In contrast, the symmetric model is optimized in close proximity to its single-cell optimum.

d. *Estimate of the modified relative load in bacterial cells*

In bacterial cells, we will assume a constant translation efficiency model. We therefore use the modified relative load formula (Eq. A57) to estimate Λ'. We will assume that the load is dominated by proteins and messages:

$$N_0 = \sum_i (\lambda+\varepsilon)\mu_{m,i} = (\lambda+\varepsilon)N_m, \quad \text{(A110)}$$

where N_m is the total number of messages. We can then solve this equation for Λ':

$$\hat{\Lambda}' = \frac{\lambda+\varepsilon}{N_0} = \frac{1}{N_m} \approx 10^{-5}, \quad \text{(A111)}$$

based on the total message number estimate for *E. coli*. (See Tab. S2.)

Appendix B: Model robustness & exploring alternatives to RLTO

In this section, we investigate the phenomenology of three different single-cell growth rate functions to determine what model features result in overabundance. We consider an *arrest model* (the RLTO model), a *slow-growth model*, and a *symmetric model*.

1. Methods: Defining alternative models

In each case, we will assume that the protein number is described by a gamma distribution:

$$N_p \sim \Gamma(\tfrac{\mu_m}{\ln 2}, \varepsilon \ln 2). \quad \text{(B1)}$$

We will assume the cell-cycle duration T is determined by this stochastic protein number N_p and then compute the population growth rate using Eq. A36 for a range of different message numbers μ_m. In each case, $\tau_0 = 1/N_0$, $N_0 = 10^5$, $\varepsilon = 30$, $n_p = \varepsilon \ln 2$. The mean expression level is $\mu_p = \mu_m \varepsilon$.

a. *Model 1: Arrest (RLTO) model*

The arrest (RLTO) model has cell cycle duration:

$$T = \tau_0 \begin{cases} \infty, & N_p < n_p \\ N_0 + N_p, & N_p > n_p \end{cases}, \quad \text{(B2)}$$

where protein expression below threshold n_p results in growth arrest.

b. *Model 2: Slow-Growth model*

In the slow-growth model, we imagine two processes: (i) checkpoint process X and (ii) other processes. The cell will divide after whichever process finishes last. Other processes will finish after time predicted by the metabolic load, identical to the threshold model defined above. However, we model checkpoint process X as the completion of a fixed amount of activity in an irreversible process. We will therefore assume it will take a time inversely proportional to the amount of enzyme X (N_p). The amount of activity is set by effective threshold n_p:

$$T = \tau_0 \max\{N_0 + N_p, \tfrac{2n_p N_0}{N_p}\} \quad \text{(B3)}$$

such that n_p defines the level of protein required to make the growth rate half the metabolic limit.

Unlike the arrest model, cell growth slows but does not stop for $N_p < n_p$. This model will test whether the results of the RLTO model are an artifact of the assumed arrest-based slow growth.

c. *Model 3: Symmetric model*

For the symmmetric model, we choose the model parameters such that the single-cell optimum was close to

the other models: $n_0 = 8.5$, $\sigma_n = 5$. The cell-cycle duration is:

$$T = \tau_0 N_0 \exp\left(\frac{(N_p - n_0)^2}{2\sigma_n^2}\right), \tag{B4}$$

such that the noise-free growth rate will be Gaussian is N_p.

2. Results: Overabundance is a robust prediction

The growth rates as a function of the mean expression level μ_p are shown in Fig. S8. The symmetric model has a population optimum in close proximity to its single-cell optimum, as we intuitively expect. However, both the arrest (RLTO) model and the slow-growth model have optima far above the threshold number n_p. We therefore conclude that it is fitness asymmetry rather than growth arrest that is responsible for the overabundance phenomenon.

Why doesn't growth arrest of a sub-population lead to a stronger effect than the same sub-population growing slowly? In Ref. [29], we showed that the population doubling time $\overline{T}$ can be understood as the exponential mean of the stochastic cell-cycle duration:

$$\overline{T} \equiv f^{-1}[\mathbb{E}_T f(T)], \tag{B5}$$

where $\mathbb{E}_T$ is the expectation over the stochastic duration T and $f(t) \equiv \exp(-kt)$, where $k = \overline{T}^{-1} \ln 2$ is the population growth rate. Due to the functional form of $f(t)$, any long cell cycles are exponentially suppressed in their contribution to the exponential mean. Therefore, low-probability extremely-long-duration cell cycles only contribute to the growth rate by reducing the fraction of growing cells.

Appendix C: Quantitation of central dogma parameters for one-message-rule

The RLTO model predicts the *one-message-rule* for the lower threshold on transcription for essential genes. In this section, we use transcriptome data from the literature to test this prediction. We first describe the sources of the data (Sec. C 1), how the estimates are computed (Sec. C 2), the results (Sec. C 3) and discussion (Sec. C 4).

1. Methods: 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 [52]. 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. S2, 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.* [10].

a. *E. coli data*

Message lifetimes: The message lifetimes (and median lifetime) were taken from a recent transcriptome-wide study by Chen *et al.* [36]. 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 [9]. 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 [44], 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 [52]).

Doubling time: The source of the doubling times for rapid (LB) and slow (M9) growth of *E. coli* comes from Bernstein [35].

Essential gene classification. The classification of essential genes in *E. coli* comes from the construction of the Keio knockout collection from Baba *et al.* [53].

Protein number. The total protein number in *E. coli* came from Milo's recent review of this subject [45].

b. *Yeast data*

Message lifetimes: The message lifetimes (and median lifetime) were taken from Chia *et al.* [47].

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^{tot}	Total number of proteins: N_p^{tot}	Average translation efficiency: ε	Average translation rate: β_p (h^{-1})
Escherichia coli (*E. coli*)	LB	30 min [35]	2.5 min [36]	12	7.8×10^3 [44]	9.4×10^4	3×10^6 [45]	22	530
	M9	90 min [35]	2.5 min [36]	36	2.4×10^3 [44]	8.6×10^4	3×10^6 [45]	24	580
Sacchromyces cerevisiae (Yeast–haploid)	YEPD	90 min [46]	22 min [47]	4	2.9×10^4 [48]	1×10^5	5×10^7 [49]	4×10^2	410
Mus musculus (Mammalian-mouse)	Tissue	27.5 h [23]	15 h [23]	1.8	1.7×10^5 [23]	3×10^5 [23]	3×10^9 [23]	1×10^4	660
Homo sapiens (Human)	Tissue	24 h [50]	14 h [51]	1.7	3.6×10^5 [46]	5×10^5	2×10^9 [45]	4×10^3	120

TABLE S2. **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.

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 [8].

mRNA abundance: The transcriptome data comes from the very recent Blevins *et al.* study [54].

Total cellular message number. There are a wide-range of estimates for the total cellular message number in yeast: 1.5×10^4 [55] (BNID 104312 [52]), 1.2×10^4 [56] (BNID 102988 [52]), 6.0×10^4 [57] (BNID 103023 [52]), 2.6×10^4 [48] (BNID 106763 [52]) and 3.0×10^4 [58]. We used the compromise value of 2.9×10^4.

Doubling time: The doubling time was taken from [46].

Protein number. The total protein number in yeast comes from Futcher *et al.* [49].

Essential gene classification. The classification of essential genes in yeast comes from van Leeuwen *et al.* [19].

Proteome abundance data: The proteome abundance data came from two sources: flow cytometry of fluorescent fusions from Newman *et al.* [8] as well as mass-spec data from de Godoy *et al.* [59].

c. *Human data*

Message lifetimes: The message lifetimes (and median lifetime) were taken from Yang *et al.* [51] 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 [60], which we averaged over tissue types.

Total cellular message number. The total cellular message number in human comes from Velculescu *et al.* [61] (BNID 104330 [52]).

Doubling time: The doubling time was taken from [50].

Protein number. The total protein number in human came from Milo's recent review of this subject [45].

Essential gene classification. The classification of essential genes in human comes from Wang *et al.* [20].

2. Methods: 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 $N_{m/c}^{\text{tot}}$. This number and its source are provided in Tab. S2. To estimate the number of messages corresponding to gene i, we re-scaled the un-normalized abundance level r_i:

$$N_{m/c,i} = N_{m/c}^{\text{tot}} \frac{r_i}{\sum_j r_j}, \tag{C1}$$

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

b. *Estimating the transcription rate: β_m*

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

$$N_{m/c,i} = \beta_{m,i}/\gamma_{m,i}, \tag{C2}$$

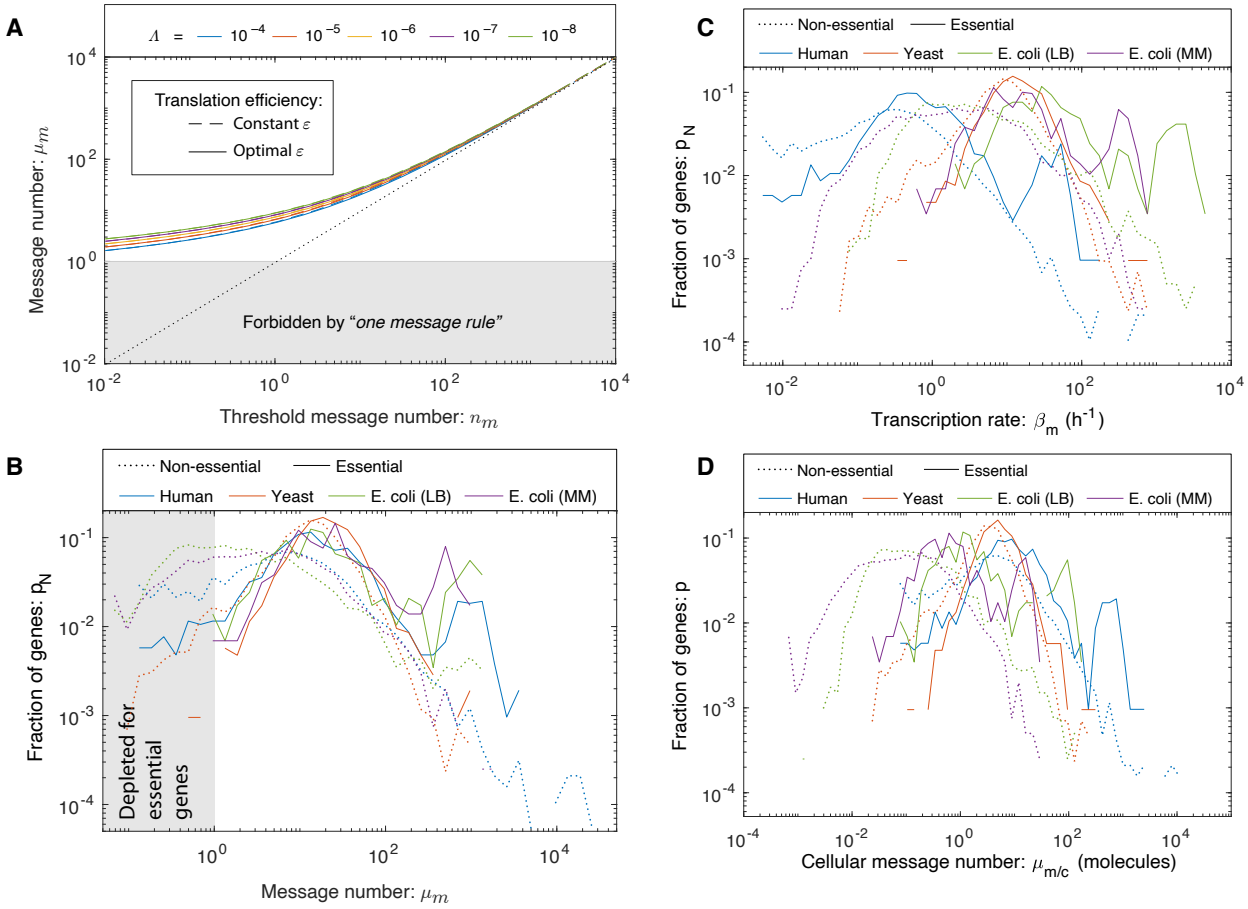

FIG. S9. **The one message rule. Panel A: One-message-rule for essential genes.** For highly transcription genes (high μ_m), little compensation for noise is required and the optimal message number tracks with the threshold message number n_m. However, as the threshold message number approaches one ($n_m \to 1$), the noise is comparable to the mean, and the optimal message number μ_m increases to compensate for the noise. As a result, a lower threshold of roughly one message per cell-cycle is required for essential genes. This threshold is predicted for both fixed (dashed) and optimized translation efficiency (solid). The threshold is weakly dependent on relative load Λ. **Panel B: A one message threshold is observed in three evolutionarily-divergent organisms.** As predicted by the RLTO model, essential, but not nonessential genes, are observed to be expressed above a one message per cell-cycle threshold. 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. **Panel C: The distribution of gene transcription rate.** The typical transcription rate varies by two orders-of-magnitude between organisms. **Panel D: The distribution of gene cellular message number.** There is also a two-order-of-magnitude variation between typical cellular message numbers. No consistent lower threshold is observed for either statistic.

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* [36]), we estimate the decay rate as the inverse gene-median message lifetime:

$$\gamma_{m,i} = \tau_m^{-1}, \tag{C3}$$

for which a consensus value was found from the literature. This number and its source are provided in Tab. S2. We then estimate the gene-specific transcription rate:

$$\beta_{m,i} = N_{m/c,i}/\tau_m. \tag{C4}$$

c. *Estimating the message number: μ_m*

To estimate the message number of gene i, we use the predicted value from the steady-state stochastic kinetic model (see A 1 a):

$$N_{m,i} = T\beta_{m,i} = \frac{T}{\tau_m} N_{m/c,i}, \tag{C5}$$

where T is the doubling time and $N_{m/c,i}$ is the cellular message number (Eq. C1).

3. Results: Histograms of central dogma transcriptional statistics

We generated histograms for each of the three transcriptional statistics: transcription rate β_m, cellular message number $\mu_{m/c}$, and message number μ_m. The histograms for transcription rate and cellular message number do not show a consistent lower limit (as predicted) and are shown in Fig. S9; however, the histogram for message number does show a consistent lower bound for the three model organisms and is shown in Fig. S9B.

4. Discussion: *E. coli* essential genes below the one-message-rule threshold

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 [53]. By this classification, 10 essential genes were below threshold. (See Tab. S3.) Our first step was to determine what fraction of these genes were also classified as essential using transposon-based mutagenesis [62, 63]. 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 [64]. In λ 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.

Appendix D: Analysis of gene-expression noise

This section provides a detailed development of gene expression noise. We continue the discussion of the model from Sec. A 1 that provided a self-contained development of the noise models developed by others which are the input to the RLTO model. Secs. D 1-D 6 describe the RLTO prediction of non-canonical noise scaling and the test of this model.

1. Results: RLTO model predicts non-canonical noise scaling

The predicted scaling of the optimal translation efficiency with message number has many important implications, including on the global characteristics of noise. Based both on theoretical and experimental evidence, it is widely claimed that gene-expression noise should be inversely proportional to protein abundance [9, 65]:

$$\mathrm{CV}_p^2 \propto \mu_p^{-1}, \tag{D1}$$

for low-expression proteins, as observed in *E. coli* [9]; however, the more fundamental prediction is that the noise is inversely proportional to the message number:

$$\mu_p = \mu_m \varepsilon, \tag{D2}$$

$$\mathrm{CV}_p^2 = \tfrac{\ln 2}{\mu_m}. \tag{D3}$$

In *E. coli*, the translation efficiency is roughly constant (*i.e.* $\hat{\mu}_p \propto \hat{\mu}_m$, Fig. 4D) and therefore Eq. D3 is consistent with the canonical noise model (Eq. D1). However, in eukaryotes, the translation efficiency grows with message number (*i.e.* $\hat{\mu}_p \propto \hat{\mu}_m^2$, Fig. 4BC). If we substitute this proportionality into Eq. D3, we predict the non-canonical noise scaling:

$$\mathrm{CV}_p^2 \propto \mu_p^{-1/2}, \tag{D4}$$

for eukaryotic cells.

2. Methods: Analysis of gene expression noise

The quantitative model for gene expression noise includes multiple contributions:

$$\mathrm{CV}_p^2 \approx \tfrac{1}{\mu_p} + \tfrac{\ln 2}{\mu_m} + c_0, \tag{D5}$$

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.* [66, 67].

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 [9, 66, 67], 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 μ_p^{-1} for low expression, we will consider two competing empirical models for the noise (Fig. S10). In the null hypothesis, we will consider a model:

$$\eta_0(\mu_p; b, c) = \tfrac{b}{\mu_p} + c, \tag{D6}$$

and an alternative hypothesis with an extra exponent parameter a:

$$\eta_1(\mu_p; a, b, c) = \tfrac{b}{\mu_p^a} + c. \tag{D7}$$

Gene name	Message number: μ_m	Annotated function from Ecocyc	Essential (E)/ Nonessential (N) Ref. [53], [62], [63]
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 *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* Δ*yhhQ* 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 S3. **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.

We will assume that CV_p^2 is normally distributed about η with unknown variance σ_η^2.

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

$$S_I(\boldsymbol{\theta}) \equiv \sum_i [\mathrm{CV}_{p,i}^2 - \eta_I(\mu_{p,i};\boldsymbol{\theta})]^2, \tag{D8}$$

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

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} S_I(\boldsymbol{\theta}), \tag{D9}$$

with residual norm:

$$\hat{S}_I = S_I(\hat{\boldsymbol{\theta}}). \tag{D10}$$

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

$$Z \equiv 2 \ln \frac{q_1}{q_0}, \tag{D11}$$

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

a. *Hypothesis test I*

In our first analysis, we will estimate the variance directly. We computed the mean-squared difference for successive CV_p^2 values, sorted by mean protein number

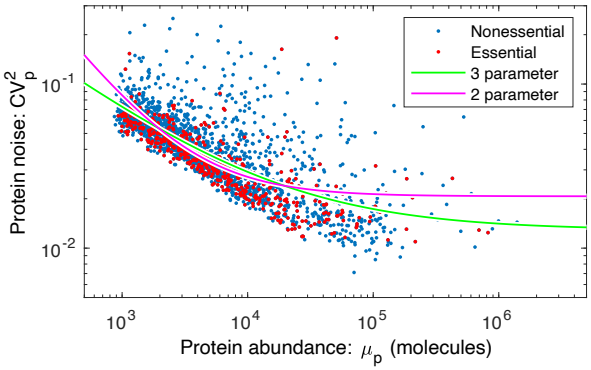

FIG. S10. **Yeast noise fit against canonical noise model, with a noise floor.** Yeast noise data fit with the 2- (null hypothesis with μ_p^{-1} dependence) and 3- parameter (μ_p^a) models. The two-parameter model corresponds to the canonical noise model (Eq. D1) and fails to quantitatively fit the data.

μ_p. The variance estimator is

$$\hat{\sigma}_\eta^2 = \tfrac{1}{2}\left\langle (\mathrm{CV}_{p,i}^2 - \mathrm{CV}_{p,i+1}^2)^2 \right\rangle_i = 6.3 \times 10^{-4}, \quad \text{(D12)}$$

where the brackets represent a standard empirical average over gene i for the μ_p-ordered gene CV_p^2 values. The test statistic can now be expressed in terms of the residual norms:

$$\begin{aligned} Z &= (\hat{S}_1 - \hat{S}_2)/\hat{\sigma}_\eta^2, & \text{(D13)} \\ &= 3.3 \times 10^4, & \text{(D14)} \end{aligned}$$

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

b. 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:

$$\begin{aligned} Z &= N \ln \tfrac{\hat{S}_1}{\hat{S}_2}, & \text{(D15)} \\ &= 1.6 \times 10^2, & \text{(D16)} \end{aligned}$$

where N is the number of data points. (Details of derivation are in Sec. D 2 d.) In this case, the p-value can be computed assuming the Wilks' theorem (*i.e.* the chi-squared test):

$$p = 6 \times 10^{-36}, \quad \text{(D17)}$$

again, strongly rejecting the null hypothesis.

c. Maximum likelihood estimates of the parameters

In the alternative hypothesis, the maximum likelihood estimate (MLE) of the empirical noise model (Eq. D7) parameters are (Fig. S10):

$$\begin{aligned} a &= 0.57 \pm 0.02, & \text{(D18)} \\ b &= 3.0 \pm 0.5, & \text{(D19)} \\ c &= 0.013 \pm 0.001, & \text{(D20)} \end{aligned}$$

where the parameter uncertainty has been estimated using the Fisher Information in the usual way using the MLE estimate of the variance [68, 69].

d. Details: Statistical details MLE estimate of the variance

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

$$h_I(\hat{\boldsymbol{\theta}}, \sigma^2) = \tfrac{N}{2} \ln 2\pi\sigma^2 + \tfrac{1}{2\sigma^2}\hat{S}_I, \quad \text{(D21)}$$

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

$$\partial_{\sigma^2} h|_{\hat{\sigma}^2} = 0, \quad \text{(D22)}$$

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

$$\hat{\sigma}^2 = \tfrac{1}{N}\hat{S}_I. \quad \text{(D23)}$$

Next we evaluate h at the variance estimator:

$$h_I(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2) = \tfrac{N}{2}\left[\ln 2\pi \tfrac{\hat{S}_I}{N} + 1\right]. \quad \text{(D24)}$$

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

$$\begin{aligned} Z &= 2h_0(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2) - 2h_1(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2), & \text{(D25)} \\ &= N \ln \tfrac{\hat{S}_0}{\hat{S}_1}, & \text{(D26)} \end{aligned}$$

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

3. Results: Non-canonical noise scaling is observed in yeast.

To test the RLTO model predictions for noise scaling, we reanalyze the dataset collected by Newman *et al.*, who performed a single-cell proteomic analysis of yeast by measuring the abundance of fluorescent fusions by flow cytometry [8]. Since the competing models (Eqs. D1 and D4) make different scaling predictions, we first apply a statistical test to determine whether the observed scaling is consistent with the canonical model (Eq. D1). We consider the null hypothesis of the canonical model (Eq. D6) and the alternative hypothesis with an unknown scaling exponent (Eq. D7). To test the models, we perform a null hypothesis test. (A detailed description of the statistical analysis, which includes the

contribution of the noise floor, is given in the Sec. D.) We reject the null hypothesis with a p-value of $p = 6 \times 10^{-36}$. The observed scaling exponent is $\hat{a} = -0.57 \pm 0.02$, which is close to our predicted estimated exponent from the RLTO model ($-\frac{1}{2}$).

4. Methods: Parameter-free prediction of noise from protein-message relation.

By combining the noise model (Eq. D3) with a protein-message abundance relation, the relation between protein abundance and noise can be predicted without additional fitting parameters. (We call this prediction *parameter-free* since, although a parameter is fit when determining the protein-message abundance, once this relation has been established, no new parameters are fit in order to predict the noise.) To test this prediction, we will compare three competing models: (i) the RLTO model, (ii) an empirical protein-message abundance model, and (iii) the constant-translation-efficiency model.

a. *Estimating protein number (μ_p) for the noise analysis*

The protein abundance data for yeast grown in YEPD media and measured with flow cytometry fluorescence [8] 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 [59]. 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, 5×10^7. (See Sec. C 1 b.) 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. S11) which corresponds to a multiplicative scaling factor. 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 [8] and found a similar offset result.

b. *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.* [8], re-scaled to estimate protein number (Sec. D 4 a) and transcriptome data from Lahtvee *et al.* [70], re-scaled to estimate message number (Sec. C 2 c).

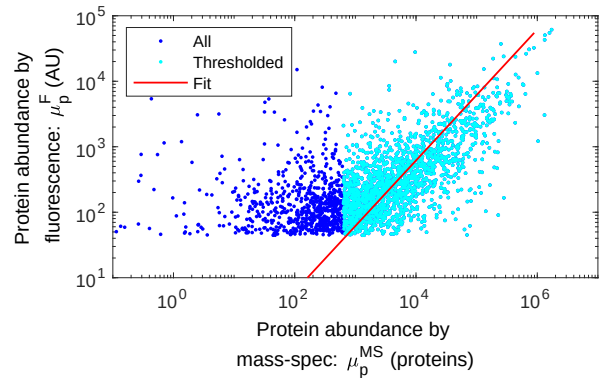

FIG. S11. **Fit to rescale fluorescence intensity to protein number.** Protein abundance from flow cytometry fluorescence [8] as a function of mass-spectrometry scaled abundance [59]. The mass-spectrometry data was thresholded at 10 proteins, and then a linear fit was performed to find the multiplicative offset of 3.9, which was used to convert protein fluorescence AU to number.

c. *Empirical model for protein number*

We initially fit the empirical model for protein number,

$$\mu_p = C_0 \mu_m^{\alpha_0}, \tag{D27}$$

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 [71], which assumes comparable error in both variables. The model parameters are:

$$\alpha_0 = 2.1 \pm 0.04, \tag{D28}$$
$$C_0 = 8.0 \pm 1.0, \tag{D29}$$

where the uncertainties are the estimated standard errors. The result of the empirical model fit is shown in Fig. S12A, along with the constant-translation-efficiency model, and the RLTO model.

d. *Empirical model for message number*

For the prediction of the coefficient of variation, it is useful to invert Eq. D27 to generate a model for message number as a function of protein number:

$$\mu_m = C_0^{-1/\alpha_0} \mu_p^{1/\alpha 0}, \tag{D30}$$
$$= C_1 \mu_p^{\alpha_1}, \tag{D31}$$

where the last line defines two new parameters: a coefficient C_1 and an exponent α_1. The resulting parameters

and uncertainties are:

$$\alpha_1 \equiv 1/\alpha_0, \tag{D32}$$
$$= 0.48 \pm 0.01, \tag{D33}$$
$$C_1 \equiv C_0^{-1/\alpha_0}, \tag{D34}$$
$$= 0.37 \pm 0.02, \tag{D35}$$

where the uncertainties are the estimated standard errors.

e. *Empirical model for translation efficiency*

To generate an empirical model for translation efficiency, we started from the empirical model for protein number (Eq. D27), and then use Eq. D2 to relate protein number, message number, and translation efficiency:

$$\varepsilon = \frac{\mu_p}{\mu_m}, \tag{D36}$$
$$= C_0 \mu_m^{\alpha_0 - 1}, \tag{D37}$$
$$= C_2 \mu_m^{\alpha_2}, \tag{D38}$$

where the last line defines two new parameters: a coefficient C_2 and an exponent α_2. The resulting parameters and uncertainties are:

$$\alpha_2 = \alpha_0 - 1, \tag{D39}$$
$$= 1.07 \pm 0.04, \tag{D40}$$
$$C_2 = C_0, \tag{D41}$$
$$= 8.0 \pm 1.0, \tag{D42}$$

where the uncertainties are the estimated standard errors.

f. *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. D31), and then substitute this into the statistical model prediction for CV_p^2 (Eq. D3):

$$\mathrm{CV}_p^2 = \frac{\ln 2}{\mu_m}, \tag{D43}$$
$$= C_0^{1/\alpha_0} \ln 2 \cdot \mu_p^{-1/\alpha_0}, \tag{D44}$$
$$= C_3 \mu_p^{\alpha_3}, \tag{D45}$$

where the last line defines two new parameters: a coefficient C_3 and an exponent α_3. The resulting parameters and uncertainties are:

$$\alpha_3 \equiv -1/\alpha_0, \tag{D46}$$
$$= -0.48 \pm 0.01, \tag{D47}$$
$$C_3 \equiv C_0^{1/\alpha_0} \ln 2, \tag{D48}$$
$$= 1.9 \pm 0.1, \tag{D49}$$

where the uncertainties are the estimated standard errors.

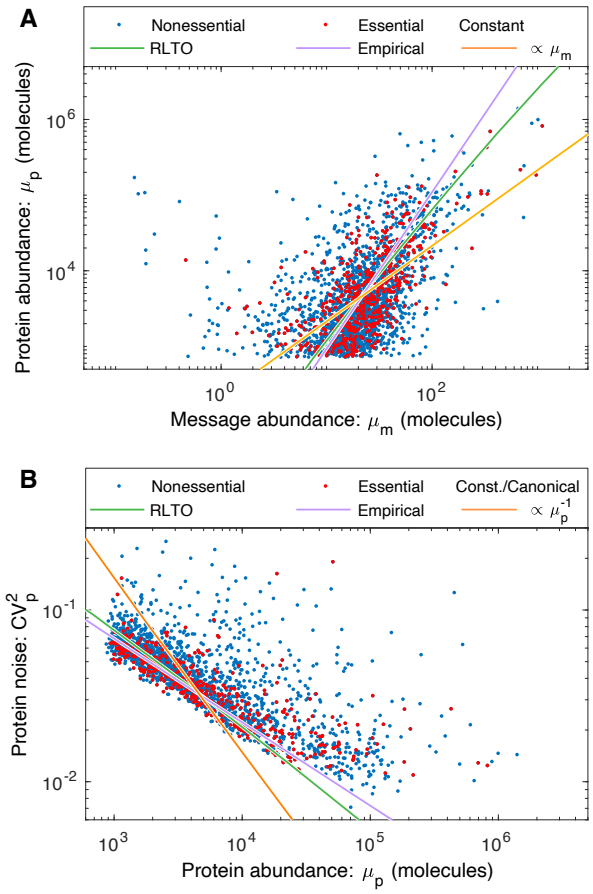

FIG. S12. **Load balancing predicts the scaling of noise. Panel A: Three competing models for protein abundance in yeast.** The empirical model (purple) fits the slope and the y offset. The RLTO (green) and constant-translation-efficiency (orange) models fit a parameter corresponding to the y offset only. As discussed in the analysis of the proteome fraction, the RLTO model qualitatively captures the scaling of the protein abundance with message number better than the constant translation efficiency model; however, the predicted fit does not correspond to the optimal power law, which is represented by the empirical model. The protein abundance has a cutoff near 10^1 due to autofluorescence [8]. **Panel B: Predictions of the noise-protein abundance relation.** Using each competing protein abundance model, the noise-protein abundance relation can be predicted using Eq. D3. The canonical noise model (Eq. D1) fails to capture even the scaling of the noise. In contrast, both the RLTO and empirical models quantitatively predict both the scaling and magnitude of the noise. The empirical model has the highest performance, presumably due to its two-parameter fit to the protein abundance in Panel A. A fit accounting for the noise floor is shown in Fig. S10.

5. Results: Parameter-free prediction of noise-abundance in yeast

The fit of the competing protein-message abundance models are shown in Fig. 5A. Using each model, we can now predict the relation between protein abundance

and noise without additional fitting parameters. The predictions of the three competing models are compared to the experimental data in Fig. 5B.

In both its ability to capture the protein abundance and predict the noise, the RLTO model vastly outperforms the constant-translation-efficiency model. The purely empirical model that best captures the protein abundance data, due to directly fitting both the y-offset and slope, also performs best in predicting the noise. It is important to emphasize that the prediction of the noise in all models is non-trivial since there are no free parameters fit, once the protein abundance relation is determined. We therefore conclude that the noise model (Eq. D3) quantitatively predicts the observed noise from the message number and that eukaryotic noise has non-canonical scaling due to load balancing.

6. Discussion: 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.* [7]. Our model suggests that robustness to noise fundamentally shapes the central dogma regulatory program. With respect to message number, the one-message-rule sets a lower bound on the transcription rate of essential genes. (See Fig. 6B.) With respect to protein expression, robustness to noise has two important implications: Protein overabundance significantly increases protein levels above what would be required in the absence of noise and therefore reshapes the metabolic budget. (See Fig. 6A.) Robustness to noise also gives rise to load balancing, the proportionality of the optimal transcription and translation rates. (See Fig. 6C.) Not only does robustness to noise affect central dogma regulation, but there is an important reciprocal effect: Load balancing changes the global scaling relation between noise and protein abundance. (See Fig. S12B.)

Appendix E: Data Tables

1. Datasets for one-message-rule analysis.

Data S1: Organism: *S. cerevisiae* (yeast). Original source: [70], [54], [59] & [8]. Essential gene classification: [19]. Processing: We merged these datasets. We added message number (messages per cell-cycle) as described in Sec. C.

Data S2: Organism: *Homo sapiens* (human). Original sources: mRNA abundances: Human Protein Atlas [60]. Essential gene classification: [51]. Processing: We merged these datasets. We added message number (messages per cell-cycle) as described in Sec. C.

Data S3: Organism: *E. coli* grown in rich media (LB). Original sources: mRNA abundance: [44]. Essential gene classification: [53]. Processing: We merged these datasets. We added message number (messages per cell-cycle) as described in Sec. C.

Data S4: Organism: *E. coli* grown in minimal media (MM), Original sources: mRNA abundance: [44]. Essential gene classification: [53]. Processing: We merged these datasets. We added message number (messages per cell-cycle) as described in Sec. C.

2. Datasets for load-balancing analysis.

Data S5: Organism: *S. cerevisiae* (yeast). Original sources: Protein abundance [21]. mRNA abundance: [70]. Processing: We merged these datasets. We added protein number and message number (messages per cell-cycle) as described in Sec. C.

Data S6: Organism: *Mus musculus* (mammalian). Original source: [23]. Processing: We added protein number and message number (messages per cell-cycle) as described in Sec. C.

Data S7: Organism: *E. coli*. Original source: [22]. Processing: We added protein number and message number (messages per cell-cycle) as described in Sec. C.

3. Datasets for noise analysis.

Data S8: Organism: *S. cerevisiae* (yeast). Original sources: Noise measurements: [8]. Essential gene classifications: [19]. Protein fluorescence data, rescaled by mass-spec data as described in Sec. D 4 a: [8]. Processing: We merged these datasets. We added protein number as described in Sec. C.