Central dogma rates and the trade-off between precision and economy in gene expression

Jean Hausser1, Avi Mayo1, Leeat Keren1 & Uri Alon1

Steady-state protein abundance is set by four rates: transcription, translation, mRNA decay and protein decay. A given protein abundance can be obtained from infinitely many combinations of these rates. This raises the question of whether the natural rates for each gene result from historical accidents, or are there rules that give certain combinations a selective advantage? We address this question using high-throughput measurements in rapidly growing cells from diverse organisms to find that about half of the rate combinations do not exist: genes that combine high transcription with low translation are strongly depleted. This depletion is due to a trade-off between precision and economy: high transcription decreases stochastic fluctuations but increases transcription costs. Our theory quantitatively explains which rate combinations are missing, and predicts the curvature of the fitness function for each gene. It may guide the design of gene circuits with desired expression levels and noise.
To function well in a given environment, cells need to express genes at the right protein copy number\(^1\). Steady-state protein abundance is set by two reactions of synthesis—transcription and translation—balanced by two processes of decay—dilution and degradation of mRNAs and proteins\(^1\). Together, these make up the four basic rates of the central dogma\(^2\).

The rates of these four central dogma reactions are controlled by diverse regulators. Transcription rate is set by transcription factors and chromatin remodelers\(^6\). Translation is modulated by RNA-binding proteins and non-coding RNAs\(^7\), and so on. The effects of these molecular controls can be summarized by the central dogma rates, such that each protein is a point in a four dimensional space whose axes are the four rates. In this study, we name this the Crick space, in honor of Francis Crick who proposed the central dogma\(^5\).

One important property of Crick space is that the same steady-state protein abundance can be achieved by many combinations of rates. For example, consider a protein made at 1000 copies per hour (Fig. 1). This can be achieved by transcribing 100 mRNAs and translating 10 proteins from each mRNA every hour. Alternatively, the 1000 proteins could be made from one mRNA translated into 1000 proteins per hour (in this example, we fixed mRNA and protein decay). There is an infinite number of ways to combine transcription and translation rates, \(\beta_m\) [mRNA h\(^{-1}\)] and \(\beta_p\) [protein mRNA\(^{-1}\) h\(^{-1}\)], in order to supply a given steady-state number of proteins \(p\), namely \(\beta_m p/\beta_p a_m = p\) where \(a_m\) [h\(^{-1}\)] and \(a_p\) [h\(^{-1}\)] are the rates of mRNA and protein decay by dilution and degradation (Methods).

Here we ask whether such combinations occur randomly, as expected if they are equally beneficial and historical accident or genetic drift is at play, or whether there are rules based on specific translation/transcription ratios that have selective advantage. If such rules exist, we might expect to see patterns in the way that genes occupy Crick space.

While there has not been systematic evidence for rules so far, previous work described how different combinations of central dogma rates can differ in their biological impact. One line of work shows that intrinsic noise\(^9,10\), the stochastic variation in protein number due to small-number effects, is largest when there are few mRNAs translated into many proteins\(^13\). Also in \(S.\) cerevisiae\(^17\), Newman et al.\(^17\) observed that a trade-off between cost and noise of gene expression. This theory accurately predicts the boundary of the empty region which varies by 2 orders of magnitude between the model organisms we considered. This approach might be of use to design synthetic gene expression circuits, and suggests rules for central dogma rates that seem to apply from bacteria to humans.

Results

Genes combining high transcription and low translation are depleted. We estimated transcription \(\beta_m\) and translation \(\beta_p\) rates, as well as mRNA and protein decay rates, for thousands of genes from previous mRNAseq and ribosome profiling (RP) experiments in \(S.\) cerevisiae\(^29\), \(M.\) musculus, \(H.\) sapiens\(^30\), and \(E.\) coli\(^3\) (Methods). All data were collected under conditions of rapid growth.

We find, in accordance with previous studies\(^5,26,32,33\), that transcription and translation rates in rapidly growing cells vary much more from gene to gene than mRNA and protein decay rates. Transcription and translation rates vary over a 1000-fold range compared to a 10-fold range for decay rates of mRNA and protein (Supplementary Fig. 1a–d). Taking into account gene-specific mRNA and protein decay rates has only a small impact on the position of genes in 2D Crick space (Supplementary Fig. 1e–g, Methods). We therefore simplify our discussion by considering a 2D Crick space, formed by transcription and translation rates.

Reducing the 4-dimensional Crick space to two dimensions neglects aspects of cell biology such as the dynamics of gene regulation in response to environmental perturbations\(^34,35\), but it allows us to focus on the most variable rates in setting steady-state protein abundance in growing cells, transcription and

\[^{1}\text{Crick space.}\]

\[^{2}\text{Central dogma.}\]

\[^{3}\text{Methods.}\]

\[^{4}\text{Crick dogma.}\]

\[^{5}\text{Arkin13.}\]

\[^{6}\text{McAdams13, and others14 proposed that}\]

\[^{7}\text{McAdams & Arkin}\]

\[^{8}\text{Arkin13.}\]

\[^{9}\text{expression by analyzing comprehensive data on central dogma rates from several model organisms11-17,26,29-38 and by theory on evolutionary trade-offs. We find that about half of the Crick space is empty: genes do not seem to combine high transcription with low translation. This depleted region is accessible by synthetic constructs, and hence its emptiness is not based on mechanistic constraints. We explain the empty Crick space by a trade-off between cost and noise of gene expression. This theory accurately predicts the boundary of the empty region which varies by 2 orders of magnitude between the model organisms we considered. This approach might be of use to design synthetic gene expression circuits, and suggests rules for central dogma rates that seem to apply from bacteria to humans.}\]

\[^{10}\text{McAdams & Arkin13, based on theoretical arguments. The prediction that a given protein abundance can be reached with the least noise when transcription is high and translation is low was validated by Ozbudak et al.14 using synthetic constructs with defined transcription and translation rates. Bar-Even et al.16 measured the noise and abundance of 43 S. cerevisiae proteins and found that noise scaled with protein abundance in a way consistent with the predictions of McAdams & Arkin13. Also in S. cerevisiae, Newman et al.17 observed a correlation between noise and mRNA abundance for 2500+ genes.}\]

\[^{11}\text{other14 proposed that}\]

\[^{12}\text{protein abundance. Testing this hypothesis has been difficult, partly because the central dogma rates could not be measured genome-wide until recently28. Another hurdle that prevented testing the hypothesis of a precision-economy trade-off in gene expression is that it is unclear how the interplay of precision and economy should affect the distribution of genes in the Crick space.}\]

\[^{13}\text{noise9-12, the stochastic variation in protein number due to small-number effects, is largest when there are few mRNAs translated into many proteins13-16. This large noise occurs because the relative fluctuations in the number of a few mRNAs are large, and are amplified by strong translation. This idea was first proposed by McAdams & Arkin13 based on theoretical arguments. The prediction that a given protein abundance can be reached with the least noise when transcription is high and translation is low was validated by Ozbudak et al.14 using synthetic constructs with defined transcription and translation rates. Bar-Even et al.16 measured the noise and abundance of 43 S. cerevisiae proteins and found that noise scaled with protein abundance in a way consistent with the predictions of McAdams & Arkin13. Also in S. cerevisiae, Newman et al.17 observed a correlation between noise and mRNA abundance for 2500+ genes.}\]

\[^{14}\text{we fixed mRNA and protein decay to simplify the visualization. See also Supplementary Fig. 1}\]

\[^{15}\text{The prediction that a given protein abundance can be achieved by many combinations of rates. For example, consider a protein made at 1000 copies per hour (Fig. 1). This can be achieved by transcribing 100 mRNAs and translating 10 proteins from each mRNA every hour. Alternatively, the 1000 proteins could be made from one mRNA translated into 1000 proteins per hour (in this example, we fixed mRNA and protein decay). There is an infinite number of ways to combine transcription and translation rates, \(\beta_m\) [mRNA h\(^{-1}\)] and \(\beta_p\) [protein mRNA\(^{-1}\) h\(^{-1}\)], in order to supply a given steady-state number of proteins \(p\), namely \(\beta_m p/\beta_p a_m = p\) where \(a_m\) [h\(^{-1}\)] and \(a_p\) [h\(^{-1}\)] are the rates of mRNA and protein decay by dilution and degradation (Methods).}\]

\[^{16}\text{Bar-Even et al.16 measured the noise and abundance of 43 S. cerevisiae proteins and found that noise scaled with protein abundance in a way consistent with the predictions of McAdams & Arkin13. Also in S. cerevisiae, Newman et al.17 observed a correlation between noise and mRNA abundance for 2500+ genes.}\]

\[^{17}\text{and translation rates can achieve a given protein abundance. For example, to obtain 1000 proteins per hour, one possibility is to translate 1000 proteins per hour from a single mRNA. Another option is to translate 10 proteins per hour from 100 mRNAs. We assumed fixed mRNA and protein decay to simplify the visualization. See also Supplementary Fig. 1}\]

\[^{18}\text{and translation rates can achieve a given protein abundance. For example, to obtain 1000 proteins per hour, one possibility is to translate 1000 proteins per hour from a single mRNA. Another option is to translate 10 proteins per hour from 100 mRNAs. We assumed fixed mRNA and protein decay to simplify the visualization. See also Supplementary Fig. 1}\]

\[^{19}\text{many proteins13, namely \(\beta_m p/\beta_p a_m = p\) where \(a_m\) [h\(^{-1}\)] and \(a_p\) [h\(^{-1}\)] are the rates of mRNA and protein decay by dilution and degradation (Methods).}\]

\[^{20}\text{expression by analyzing comprehensive data on central dogma rates from several model organisms11-17,26,29-38 and by theory on evolutionary trade-offs. We find that about half of the Crick space is empty: genes do not seem to combine high transcription with low translation. This depleted region is accessible by synthetic constructs, and hence its emptiness is not based on mechanistic constraints. We explain the empty Crick space by a trade-off between cost and noise of gene expression. This theory accurately predicts the boundary of the empty region which varies by 2 orders of magnitude between the model organisms we considered. This approach might be of use to design synthetic gene expression circuits, and suggests rules for central dogma rates that seem to apply from bacteria to humans.}\]
translating ribosomes. In those cases, the autoregulatory protein represses its own expression, explaining why their expression is low. In other cases, there are other genes co-transcribed with the target that are more highly expressed. Since these genes are transcribed at higher rates, their translation is more likely to be repressed. Additionally, the competitive binding of other ribosomal proteins to the mRNA may also play a role. For example, the expression of genes encoding ribosomal proteins is tightly controlled, and this could be reflected in the autorepression observed in the figure.

**Figure 2**

Genes combining high transcription and low translation are depleted from the Crick space across organisms. **a-d** There is a depleted region in the Crick space of four model organisms. Transcription and translation rates were estimated from ribosome profiling and mRNA sequencing data. The top percentile of translation rates ($\beta_{\text{p}}^{\text{max}}$) is represented as a horizontal dashed line. The observed boundary of the depleted region (diagonal dashed line) has slope 1 and is such that 99% of the genes have a larger translation/transcription ratio. Excluding 1% of genes in this way makes the boundary line less sensitive to measurement errors and to outlier genes, some of which are highlighted. The predicted boundary of the depleted region (red line) is according to the theory introduced later in this article. Technical constraints explain the absence of genes at low transcription and translation rates (region marked “Not sampled”). *S. cerevisiae* data (a) from Weinberg et al.29. *M. musculus* (b) and *H. sapiens* (c) data from Eichhorn et al.30. *E. coli* data (d) from Li et al.3. **e** The boundary of the depleted region has additional structure. Departing from the main distribution of genes are a set of 59 genes with very high transcription and translation rates ($\beta_{\text{p}} > 80 \text{ h}^{-1}$ and $\beta_{\text{m}} > 1000 \text{ mRNA}^{-1} \text{ h}^{-1}$), indicated by an arrow in Fig. 2d). About 90% of these genes are ribosomal proteins. The rest are high abundance proteins such as the glycolysis enzyme gapA (the *E. coli* equivalent of Gapdh), the ATP synthase c subunit, and outer-membrane proteins (OMPs). Most genes in

**Table 1**

The intercept $k$ of the boundary of the depleted region varies over two orders of magnitude across the studied organisms.

| Organism          | Measured $k$       | Predicted $k$       | $\beta_{\text{p}}^{\text{max}} \times 10^3$ | $\alpha_p$ | $\sum \beta_{\text{m}} \times 10^4$ | $\varepsilon_{\text{df}}$ |
|-------------------|--------------------|--------------------|-------------------------------------------|------------|-----------------------------------|---------------------|
| *S. cerevisiae*   | $1.1 \pm 0.1$     | $1.1 \pm 0.3$     | $7.1 \pm 0.7$                            | $1.3 \pm 0.2$ | $30 \pm 15$                     | $0.10 \pm 0.01$     |
| *E. coli*         | $14 \pm 3$        | $12 \pm 4$        | $11 \pm 1$                               | $19 \pm 3$  | $2.0 \pm 1.1$                   | $0.25 \pm 0.01$     |
| *M. musculus* (3T3) | $44 \pm 3$       | $29 \pm 8$        | $3.0 \pm 0.3$                            | $0.04 \pm 0.01$ | $2.5 \pm 1.2$                 | $0.3 \pm 0.03$     |
| *H. sapiens* (HeLa) | $66 \pm 4$    | $60 \pm 17$       | $5.1 \pm 0.5$                            | $0.05 \pm 0.01$ | $1.4 \pm 0.7$                  | $0.3 \pm 0.03$     |

$k$ can be predicted from the maximal translation rate $\beta_{\text{p}}^{\text{max}} \text{ [protein mRNA}^{-1} \text{ h}^{-1}]$, the protein decay rate $\alpha_p \text{ [h}^{-1}]$, the total transcriptional output $\sum \beta_{\text{m}} \text{ [mRNA h}^{-1}]$, and the noise floor $\varepsilon_{\text{df}}$ using Eq. (5). The measured $k$ is defined by having 99% of genes with $\beta_{\text{p}} / \beta_{\text{m}} = k$. Uncertainties represent standard errors determined from 3744 to 9770 genes depending on the organism.
this set are essential to cellular viability, as indicated by knockout experiments (Supplementary Fig. 2b). If one removes essential genes from the data, the boundary of the depleted region shifts up and tightly fits the rest of the genes with a higher intercept, $k = 44 \pm 9$. This higher intercept for non-essential genes is predicted from the theory introduced below (Supplementary Fig. 2b, Methods).

A depleted region is also found when we estimate transcription and translation from two proteomic and mRNAseq datasets in *H. sapiens* and *M. musculus* (Supplementary Fig. 2c, d). Finally, we observe the depleted region when plotting transcription burst rate against translational burst size inferred from single cell protein abundance measurements (Supplementary Fig. 2e).

We assessed the statistical significance of the depleted region by shuffling transcription and translation rates while conserving the distributions of protein abundance and translation rates (Supplementary Fig. 2f–h). We find that none of the $10^4$ shuffled datasets show a comparable depleted region in Crick space (equal or smaller number of genes with $\beta_p/\beta_m < k$, $p < 10^{-4}$).

Genes can mechanistically achieve high transcription and low translation. A possible explanation for the depleted region is that a (possibly yet unknown) biochemical constraint prevents high transcription combined with low translation. To test for this possibility, we re-analyzed measurements by Kosuri et al. on synthetic genes that provided a wide range of transcription and translation rates. In that study, GFP was expressed in *E. coli* under the control of 114 promoters and 111 Ribosomal Binding Sites (RBSSs) of varying strengths. Relative abundance of the GFP mRNA and protein was then quantified by mRNAseq and flow cytometry.

The transcription and translation rates of the synthetic constructs largely overlap with those of *E. coli* genes (Fig. 2e, Supplementary Fig. 2i). However, in contrast to *E. coli* genes, a large fraction of the synthetic constructs achieve a combination of high transcription and low translation rates. 25% of the synthetic genes fall in the depleted region seen for endogenous *E. coli* genes. In *S. cerevisiae* as well, 32% of synthetic promoters from the library of Sharon et al. fall in the depleted region (Supplementary Fig. 2j).

This observation supports the conclusion that the biochemistry of gene expression can achieve high transcription and low translation in principle. In support of this argument, there are indeed examples of such genes in the depleted region for all four organisms. These include the ribosome maturation factor rimM in *E. coli*, the amino-acid response regulator GCN4 in *S. cerevisiae*, and the iron homeostasis protein Fih1 in *M. musculus* and *H. sapiens* (Fig. 2a–d). Thus, the results in this paper concern ~99% of the genes, with the remaining ~1% requiring additional analysis (see discussion for suggested effects for these genes).

At constant protein abundance, increasing transcription increases both precision and cost of gene expression. Because biochemical constraints do not seem to explain the lack of genes combining high transcription with low translation, we asked whether evolutionary trade-offs might explain it.

One could hypothesize that cells avoid combining high transcription with low translation in order to minimize the cost of mRNA synthesis. In *S. cerevisiae* growing in rich medium, the fitness cost of mRNA is $c_m \sim 10^{-9}$ per transcribed nucleotide (Methods). Synthesizing a non-beneficial mRNA of length $l_m$ leads to a growth rate penalty $\Delta \mu_m$ that is linear with the transcription rate $\beta_m$. For a typical mRNA of length $l_m = 1300$ nucleotides transcribed at a rate $\beta_m = 30$ mRNA h$^{-1}$, the fitness cost of transcription is thus $c_m \mu_m \sim 4 \times 10^{-5}$ per hour (Fig. 3a, Methods), which is selectable. The cost of mRNA is also selectable in *E. coli*.

In addition to their cost, high transcription rates also have benefits in reducing the noise (Fig. 3b) and increasing the transcription rate while keeping protein abundance fixed should therefore decrease stochastic fluctuations in protein abundance.

To test if this prediction holds genome-wide across the diversity of chromosomal context and promoters, we use measurements of cell-to-cell variations in protein abundance in *S. cerevisiae* and *E. coli*. Cell-to-cell variations in protein abundance can be quantified by the coefficient of variation (CV). We determine contours of the CV as a function of transcription and translation rate using Gaussian smoothing (Methods), and compare these to contours of protein abundance. Both in *S. cerevisiae* (Fig. 3b) and *E. coli* (Supplementary Fig. 3a), the CV decreases with increasing transcription and decreasing translation on each equi-protein line. The CV mainly scales with transcription, as predicted by theory (Methods).

![Fig. 3](https://example.com/fig3.png) Increasing transcription at constant protein abundance increases transcription cost and decreases stochastic fluctuations in protein abundance. *S. cerevisiae* rates from Weinberg et al. Diagonal dotted lines are lines of constant protein abundance from $10^4$ to $10^5$ proteins per cell. a. The loss of fitness $c_m \mu_m$, due to transcription (black lines) is linear in the transcription rates $\beta_m$ and the mRNA length $l_m$. The linear factor $c_m$ (introduced in the next section) rescales transcription fluxes into fitness loss. b. In *S. cerevisiae*, $l_m = 1300$ nt, $c_m = 10^{-9}$ nt$^{-1}$. c. Precision in gene expression increases with transcription whereas protein abundance depends both on transcription and translation. Thus, at a given protein abundance, increasing transcription increases the precision of gene expression at the expense of higher transcription costs. See also Supplementary Fig. 3.
Hence, transcription and translation rates impact both gene expression precision and mRNA economy. At a given protein abundance, high translation/transcription ratios lead to economy but also higher gene expression noise, whereas low translation/transcription ratios yield high precision at the expense of higher mRNA cost (Fig. 3c). The lack of genes combining high transcription and low translation could be explained by this trade-off: for genes located in the depleted region, the benefits of increased precision may be smaller than transcriptional costs.

The precision-economy trade-off and the noise floor explain the depleted region of Crick space. To quantitatively test whether a trade-off between precision and economy can explain the depleted region, we developed a minimal mathematical model of the fitness cost and benefit of transcription and translation (Fig. 4). The model has two main predictions: first that the optimal ratio between translation and transcription rates $\beta_p/\beta_m$ is set by the ratio of transcription cost per mRNA molecule $C$ and the gene’s sensitivity to noise $Q$ (defined below). The second prediction is an analytical formula for the boundary of the

**Fig. 4** A trade-off between precision and economy can explain the depletion of genes combining high transcription with low translation. **a** The noise load $\Delta f_{\text{noise}}$ is the loss of fitness due to stochastic fluctuations in protein abundance. **b** The optimal $\beta_p/\beta_m$ depends on the transcription cost per mRNA $C$ and on how noise sensitive ($Q - |f'(p')|$) the gene is. Noise-sensitive genes have narrower fitness functions ($|f'(p')|$ large). For a given $p'$, higher transcription $\beta_m$ decreases fluctuations in protein abundance. Genes that are sensitive to noise ($Q$ large) should thus have low translation/transcription ratio. On the other hand, precision is less critical for genes with flat fitness functions ($Q$ small). These genes should have higher $\beta_p/\beta_m$ to keep transcription costs low. **c** The precision of gene expression is limited by the noise floor $c_o$. Protein abundance and CV data re-plotted from Taniguchi et al.\(^40\). The noise floor is also found in the *E. coli* measurements of Silander et al.\(^41\) (Supplementary Fig. 4b). **d** Genes with fitness functions narrower than the noise floor $c_o$ have negative average fitness ($< f < 0$). Genes with negative fitness are not selectable. Thus, the noise floor $c_o$ prevents the selection of narrow, noise-sensitive fitness functions. **e** The maximal, selectable noise sensitivity $Q_{\text{max}}$ determines the position $k$ of the boundary of the depleted region. Proteins lying on the boundary have maximal noise sensitivity $Q = Q_{\text{max}}$. Feasible combinations of transcription and translation correspond to genes that are less sensitive to noise $Q < Q_{\text{max}}$. **f** The precision-economy trade-off theory predicts the position $k$ of the depleted region in organisms from bacteria to mammals. Error bars represent 95% confidence intervals. The $p$-value is computed using Pearson’s test on 4 organisms (two-sided). The correlation between measurements and theory is robust to varying the fraction of genes used in defining the depleted region (Supplementary Fig. 4e). See also Supplementary Fig. 4.
the transcription cost \(\Delta f_m\) and of the noise load \(\Delta f_{\text{noise}}\) (Methods),

\[
\begin{align*}
\frac{d(c_m \beta_m l_m m + f^* (p^*) | \sigma^2 | / 2)}{\delta \beta_m} &= 0 \Rightarrow \beta_m = \sqrt{\frac{\alpha_m \alpha_n}{C Q p}} \Rightarrow \beta_m = \frac{C}{Q}
\end{align*}
\]

where \(C = c_m \alpha_m \alpha_n = c_n \beta_m l_m m\) quantifies the cost of transcription per mRNA molecule \(m\) for this gene, and \(Q = \frac{1}{2} f^* (p^*) | p^*\) is the gene’s sensitivity to noise. Genes with narrower fitness functions (larger \(f^* (p^*)\)) are more sensitive to noise because the noise kicks protein abundance farther from optimum. The sensitivity of a gene to noise also depends on protein abundance \(p^*\) because the noise generally scales with protein abundance \(\sigma^2 \sim p^*\).

The model therefore predicts a relationship between \(\beta_p / \beta_m\) ratios and the shape of fitness functions (Fig. 4b). Broad fitness functions should have large \(\beta_p / \beta_m\) because genes with broad fitness functions are not sensitive to noise. For those, high precision provides little benefit, and it is best to maximize economy by lowering transcription. On the other hand, genes with narrow fitness functions are sensitive to noise. For those, requirements of high precision to keep the noise load low dominate the cost of transcription. Genes with narrow fitness functions should therefore have small \(\beta_p / \beta_m\) ratios.

We compared the model prediction for the curvature of the fitness function near its peak to recent measurement of fitness functions of 21 genes in \(S.\ ceriseivae\)45. We find that the curvatures predicted from \(\beta_p / \beta_m\) are within an order of magnitude of the measured curvatures without any fitting parameters (Supplementary Fig. 4a). Predictions and measurements correlate positively \((r = 0.39)\). A shuffling test suggests that the agreement between the predictions and measurements is unlikely due to chance \((p = 0.04,\ Methods)\). However, variability in the experimental measurements precludes a conclusive comparison.

To find a lower bound \(k\) on \(\beta_p / \beta_m\) and explain the boundary of the depleted region, we note that there is a limit to how small the noise in gene expression can be. This limit, called the noise floor \(c_{\text{lf}}\), is revealed by measurements of cell-to-cell variation of protein abundance in clonal populations17,40–42 (Fig. 4c). It is the reason for the noise floor \(c_{\text{lf}}\) and it has been proposed that it is due to extrinsic noise40 or larger transcriptional burst size of high abundance proteins42. The noise floor puts an upper bound \(Q_{\text{max}}\) on how noise-sensitive genes can be: if a gene had a fitness function narrower than this limit \((Q > Q_{\text{max}})\), the noise load would dominate the benefit of expressing the gene (Fig. 4d), leading to negative fitness. Because genes with \(Q > Q_{\text{max}}\) cannot be selected for, all expressed genes must satisfy \(Q < Q_{\text{max}}\). The boundary of the depleted region \(\beta_p / \beta_m = k\) hence corresponds to genes with highest sensitivity to noise \(Q_{\text{max}}\) (Fig. 4e). The depleted region \(\beta_p / \beta_m < k\) corresponds to transcription and translation rates that are optimal for genes with fitness functions too narrow given the noise floor.

To find \(k\), we substitute \(Q = Q_{\text{max}}\) in Eq. (4) for the optimal \(\beta_p / \beta_m\) ratio and rewrite \(k\) in terms of the noise floor and other cell biology constants (Methods):

\[
k = \frac{C}{Q_{\text{max}}} = \frac{\beta_p^{\text{max}} c_{\text{lf}}}{\sqrt{\alpha_p \sum \beta_m}}
\]

where \(\sum \beta_m\) is the combined transcriptional output of all genes.

This expression provides an intuition for the cellular parameters which set the boundary of the depleted region of the Crick space. For example, a larger noise floor \(c_{\text{lf}}\) raises the boundary because there is less benefit in having high
transcription. An increased transcripational output $\sum \beta_m$ lowers the boundary because individual mRNAs are less costly, allowing increased precision in gene expression.

This formula for $k$ accurately predicts the boundary of the depleted region of Crick space (Fig. 4f; red lines on Fig. 2) despite the fact that $k$ varies by nearly two orders of magnitude between organisms (Table 1). Thus, the depleted region can be explained in terms of fundamental parameters such as the noise floor, maximal translation rate, total transcription output, and mean protein decay rate. Individual cellular constants alone cannot accurately predict $k$ (Supplementary Fig. 4d). Neither can $k$ be predicted by noise alone without considering economy, such as hypothesizing that the depleted region is made of all $\beta_m$ and $\beta_p$ for which increasing transcription provides little extra precision relative to the the noise floor (Methods).

Discussion

We find that the distribution of genes in Crick space is not random; genes combining high transcription and low translation are depleted. Such combinations of high transcription and low translation can be achieved with synthetic gene constructs.\(^{38}\) Therefore, mechanistic constraints cannot explain this depletion. We explain the depletion by a trade-off between precision and economy: increasing transcription at constant protein abundance diminishes stochastic fluctuations, but at a fitness penalty due to the cost of transcription. High transcription rates are therefore optimal for genes that are sensitive to noise whereas low transcription rates are well suited for genes that can tolerate high noise (Fig. 5a). A quantitative model of this trade-off predicts the curvature of the fitness function for each gene and quantitatively explains the boundary of the depleted region.

Combinations of high transcription and low translation are achievable by the gene expression machinery, as evidenced by the thousands of synthetic reporter constructs in the depleted region, and the ~1% natural genes in the depleted region. The depleted region is also easily reachable by mutations from naturally occurring genes—for example, it typically takes a single mutation to turn a strong RBS into a weak RBS\(^{44}\), and single mutations in promoters can strongly affect transcription.\(^{45}\) This strengthens the hypothesis that the depleted region is due to selection. We evaluate other hypotheses to explain the depleted region in the Supplementary Discussion.

Ninety nine percent of the genes lie above the predicted boundary of the depleted region. The remaining 1% fall in the depleted region. For some of these genes, evidence from focused studies indicates low translation rates and/or translation control, such as the E. coli outliers rimM and trmD\(^{46}\), HAC1 in S. cerevisiae\(^{47}\), and Fhl1 in M. musculus and H. sapiens\(^{48}\). One explanation for such outlier genes is that they face constraints beyond the precision-economy trade-off. For example, the amino-acid response regulator GCN4 in S. cerevisiae is strongly transcribed and poorly translated in rich medium. But under amino-acid imbalance conditions, a general decrease in protein translation triggers derepression of GCN4 translation through a translation reinitiation mechanism involving short upstream open reading frames.\(^{49}\) This regulatory mechanism couples GCN4 synthesis to translation stress. It also bypasses transcription, which could allow for rapid upregulation. Such considerations of regulatory couplings or speed might overshadow precision-economy-based limits for certain genes, especially genes responsible for survival upon changes in the environment.\(^{50}\) We consider other aspects of gene regulation beyond precision and economy in the Supplementary Discussion.

The distribution of genes in Crick space is bounded above by the maximal translation rate ($10^{3.6}-10^{4}$ proteins h\(^{-1}\)), and below by the boundary of the depleted region. The position of each gene in this space is determined, in the present picture, by the curvature of its fitness function. Genes with a narrow fitness function are most sensitive to noise, and are predicted to lie near the boundary of the depleted region. Genes with a broader fitness function are predicted to lie farther above this boundary. This prediction suggests an experimental test by measuring the curvature of the fitness function and comparing to the prediction. Accurate measurements of fitness functions can be performed by titrating protein concentration experimentally and measuring fitness. A recent experiment measured such fitness functions for 85 genes in S. cerevisiae.\(^{51}\) While we found a statistical agreement between the predicted and measured curvatures of fitness functions, the measurement errors were too large to permit a meaningful comparison to the present predictions (Supplementary Fig. 4a). Further experiments to measure fitness functions can test whether transcription and translation rates predict fitness curvature near the peak. Alternatively, transcription and translation could be manipulated to quantify the effect of precision and economy on fitness. Such experiments are challenging because state-of-the art assays can detect fitness changes of ~1% which is bigger than the fitness changes visible to natural selection ($10^{-8}$ to $10^{-4}$ h\(^{-1}\) depending on the species).

![Fig. 5](https://doi.org/10.1038/s41467-018-07391-8) The precision-economy trade-off suggests rules for choosing transcription and translation rates and for selecting regulatory strategies. a Due to the precision-economy trade-off, low $\beta_p/\beta_m$ is preferred for genes that are sensitive to noise. In log-log scale, low $\beta_p/\beta_m$ corresponds to a diagonal line. On the other hand, high $\beta_p/\beta_m$ is preferred for proteins that tolerate noise. Because translation rates have an upper limit $\beta_m^\text{max}$, genes with highest $\beta_p/\beta_m$ are found on a horizontal line. b Regulatory strategies that lead to the same protein abundance differ in how they impact precision. Upregulating transcription simultaneously increases protein abundance and precision. On the other hand, upregulating translation increases protein abundance while decreasing precision. Transcription control is thus advantageous assuming that precision is desirable. c When protein abundance changes by a large amount, pure transcription regulation can put the gene in the sub-optimal, depleted region. This can be avoided by co-regulating transcription and translation. See also Supplementary Fig. 5
Beyond optimizing transcription and translation under a precision-economy trade-off, an additional reason why genes may lie farther above the boundary of the depleted region is the possibility that noise is beneficial for some genes. This occurs for example in cases of bet hedging where a gene product brings little or no fitness advantage at present, but expression is maintained in case conditions change so that the gene product becomes important. In such cases, theory and experiments have shown that a wide cell–cell variation in protein level can be beneficial. Such genes expressed for possible future needs are predicted to lie far above the boundary of the depleted region. This prediction is in agreement with the finding of stress genes at relatively high positions above the boundary, and of essential genes closer to the boundary (Supplementary Fig. 5a, b, Supplementary Table S2, Supplementary Data 1).

The theory has further testable predictions. By expressing a protein from different synthetically produced combinations of translation and transcription rates, one should find that there is an optimal translation/transcription ratio. Upon changing growth conditions such that the protein becomes more important for growth, the optimal transcription rate should increase whereas the optimal translation rate should show little change.

The finding that essential, high-precision genes are located close to the boundary of the depleted region has implications for synthetic circuit design. If a protein needs to have a specific abundance for the circuit to function properly, that protein should be expressed using a promoter and ribosomal binding site that puts it close to the boundary of the depleted region. In other words, the optimal design should have $\beta_p/\beta_m = k$ with a value of $k$ appropriate to the organism (Table 1). On the other hand, if the circuit is insensitive to the exact concentration of that protein, the protein should be expressed using a weaker promoter and stronger RBS to save transcriptional resources. Combinations of strong transcription (promoters, enhancers) with weak translation (RBS and so on) should be avoided because they incur high transcriptional cost with no extra precision benefit.

The present approach can also help interpret the mode of regulation when the abundance of a protein needs to change. Increasing a protein level can be done by increasing transcription, translation, or both. Studies in several organisms indicate that transcription regulation is more prevalent and strong than translation regulation for most genes. The present theory provides a possible explanation for this observation (Fig. 5b). Transcription regulation increases protein abundance and at the same time decreases noise. Translation regulation will increase noise. Thus transcription control is advantageous assuming that precision is more desirable than economy. Precision could be more desirable than economy for genes which become key to growth upon a change in condition, such as amino-acids biosynthesis genes upon a sudden amino-acid depletion (Supplementary Fig. 5j). For these genes, fluctuations leading to low expression would be penalizing. This penalty can be mitigated by increasing transcription. The relatively rare cases of strong translation regulation may be due to considerations of faster response time, or to cases where it is beneficial to reduce precision, such as in bet hedging. One interesting case is when proteins need to be upregulated from a very-low to a very-high level. Geometric considerations rule out a purely transcriptional regulation, because this will put the gene into the depleted region; instead, a combined transcription and translation upregulation is predicted (Fig. 5c).

The present findings suggest that translation/transcription ratios are determined to some extent by rules, such as precision-economy trade-offs. Rules have been proposed in the past to explain features of the complexity of biological systems. Patterns such as empty regions of the Crick space can help define the rules, similar to the way empty regions of morphospace in animal morphology or single-cell gene expression can be used to infer potential evolutionary tasks and trade-offs. It would be interesting to discover additional rules for gene expression in order to interpret the evolved design of cells and to improve engineering of synthetic circuits.

**Methods**

**Central dogma rates and steady-state mRNA & protein abundance.** Using a commonly used formalism, gene expression can be modeled as a dynamic system. The system has two variables, $m$ and $p$ which represent the number of mRNAs and proteins per cell.

- mRNAs are transcribed at a rate $\beta_m$ [mRNA/h] and decay at a rate $\alpha_m$. Proteins are translated at a rate $\beta_p$ [protein mRNA $-1$ h$^{-1}$]) and decay at a rate $\alpha_p$.

- The protein decay rate $\alpha_p$ [h$^{-1}$] combines the effect of 1. degradation $\alpha_{d,m}$ [h$^{-1}$] and 2. dilution by cell growth and division which takes place at rate $\mu$ [h$^{-1}$]. Thus $\alpha_p = \alpha_{d,m} + \mu$.

- This consideration also applies to mRNA decay rates. However, mRNA decay is typically much faster than the cell cycle time (Supplementary Table 1). As a result, mRNA decay rate is set by the degradation rate.

Using these reaction rates, we can write the dynamics of $m$ and $p$ as

$$\frac{dm}{dt} = \beta_m - \alpha_m m$$

$$\frac{dp}{dt} = \beta_p m - \alpha_p p$$

To compute steady-state mRNA and protein abundance, we set $dm/dt = dp/dt = 0$. Solving for $m$ and $p$ yields expressions for steady-state mRNA and protein abundance as function of the central dogma reaction rates

$$m = \frac{\beta_m}{\alpha_m}$$

$$p = \frac{\mu \beta_p}{\alpha_m \alpha_p}$$

**Data sources and cellular constants in four organisms.** *S. cerevisiae:* We obtained the processed Reads Per Kilobase per Million (RPKM)s of the mRNAseq and RP experiments of Weinberg et al. from GEO (GSE75897). We used the RiboZero mRNAseq experimental. Experimental measurements estimate $N_m = 60,000$ mRNA copies per cell at cellular volume at 37°C. Given a protein concentration of $3 \times 10^6$ molecules per cell, we estimate that there are $N_p = 1.1 \times 10^9$ proteins per cell. We used a cell division time of 99 min (BNID100430). Given that the median protein half-life (excluding the dilution factor) is 45 min, the typical protein decay rate is $\alpha_p = 60 \log(2)/45 + 1/99 = 1.34 \text{ h}^{-1}$. Eser et al. estimated the typical mRNA decay rate at $\alpha_m = 5.1 \text{ h}^{-1}$. Multiple experiments found a noise floor $c_m \approx 0.15^{16,17,19}$.

E. coli: From the sequence reads obtained, we downloaded the mRNAseq and RP reads from the experiments Li et al. performed in rich medium (mRNAseq: SRR1067773, SRR1067774, RP: SRR1067765, SRR1067766, SRR1067767, SRR1067768). We obtained the E. coli genome sequence and transcriptome annotation from NCBI (accession NC_000913.3).

All genome mappings were performed using Bowtie2 in local alignment mode. We discarded all technical reads as well as reads that mapped against non-coding RNAs, defined as transcripts marked as ncRNA, tRNA, or tRNA in the genome annotation. Remaining reads were mapped to transcripts marked as CDS in the genome annotation. RP reads were mapped to coding transcripts after trimming the first and last 5 codons to remove the effect of translation initiation and termination. Reads that mapped equally well to multiple loci were assigned to one of the loci at random. We then computed RPMKPs per gene. Because reproducibility between runs was high, we combined reads from all runs for subsequent analyses.

The resulting mRNA abundances and protein synthesis rates estimates were highly correlated with those computed by Li et al. ($r^2 > 0.99$). Differences could be due to the updated genome version we used (NC_000913.3 vs NC_000913.3 in the original analysis of Li et al.), differences in the aligner (we used Bowtie2 while Li et al. used Bowtie), and other differences in the implementation of the bioinformatics pipeline. Repeating all the E. coli analyses using the protein synthesis rates from Supplementary Table 1 and mRNA abundance Supplementary Table S4 of Li et al. leads to minor changes in the exact position of genes in Crick space and supports all conclusions presented in this article.
There are 1380 mRNAs per cell (BNID1000064). Cell volume is \(\pm 1 \mu m^3\) (BNID100004). Assuming a protein concentration\(^{67}\) of \(3 \times 10^6 \text{ proteins per cell}\). Doubling time was measured by Li et al.\(^{21}\) at 21.5 min, which puts the growth rate \(\mu = 60 \log(2)/21.5 = 1.93 \pm 1\). RP RPKMs correlate well with protein abundance\(^{3}\). This suggests that protein decay mainly takes place through dilution. Thus, protein degradation can be neglected in estimating protein decay for experiments of 30 min\(^{32}\) from GEO (GSE60426). Assuming a protein concentration\(^{67}\) of \(3 \times 10^6 \text{ proteins per cell}\). Doubling time was measured by Li et al.\(^{3}\) at 21.5 min, which puts the growth rate \(\mu = 60 \log(2)/21.5 = 1.93 \pm 1\). RP RPKMs correlate well with protein abundance\(^{3}\). This suggests that protein decay mainly takes place through dilution. Thus, protein degradation can be neglected in estimating protein decay for experiments of 30 min\(^{32}\) from GEO (GSE60426).

Estimating transcription and translation from omics datasets. For each gene \(i\), we estimated the number of mRNAs per cell \(m_i\) from the total number of mRNAs per cell \(N_m\) from the per-gene mRNAseq RPKM \(r_j\) data as

\[
m_i = N_m \frac{r_j}{s_j}
\]

At steady-state, mRNA abundance \(m\) is the ratio of the transcription rate \(\beta_m\) to the mRNA decay rate \(\alpha_m\). We thus estimated the transcription rates of each gene as

\[
\beta_m = m \alpha_m
\]

where \(\alpha_m\) is the median mRNA decay rate (Supplementary Table 1). Finally, we estimated translation rates \(\beta_p\) by combining three numbers: the total number of proteins per cell \(N_p\), the protein decay rate \(\alpha_p\), and the gene’s RP RPKM \(s_i\) of gene \(i\). The number of proteins synthesized per time unit is \(N_p \alpha_p\). A fraction \(s_i / N_p\) of this protein synthesis flux is translated from mRNAs \(m_i\) of gene \(i\). We estimate the translation rate \(\beta_p\) (expressed per mRNA copy per cell) by dividing the protein synthesis of each gene by the mRNA copy number \(m_i\):

\[
\beta_p = \frac{N_p \alpha_p s_i}{m_i}.
\]

Estimates of \(N_p\) and \(\alpha_p\) are provided in Supplementary Table 1.

To address the concern that gene-to-gene variations in the translation elongation rate may bias our estimates, we compared protein synthesis rates computed from protein abundance and decay data \(\rho\) measured by mass-spectrometry\(^{39}\) to the protein synthesis rates estimated from ribosomal density \(\beta_p\) on the same cell line (Mouse 3T3). We find that protein synthesis rates correlate well (Supplementary Fig. 1e, \(r = 0.70, p < 10^{-5}\) at Pearson’s test on more than 3000 genes). This agreement is probably due to the fact that although elongation rates vary from gene to gene, the initial translation is more than 3-fold (3 folding)\(^{-1}\) fold)\(^{-1}\) fold. Thus the latter tend to determine most of the variation.

To evaluate the effect of this simplification, we plot genes in 2D Crick space taking gene-specific mRNA and protein decay rates into account (Supplementary Fig. 1f). We then assign the same median mRNA and protein decay rates to all genes to re-estimate transcription and translation rates (Supplementary Fig. 1g).

The gene positions in the two resulting 2D Crick spaces differ by less than 0.3 (root mean square deviation in log10 rates), which is \(\pm 1\%\) of the total variation (about 3 log10 units in transcription and translation rates). We conclude that taking into account gene-specific mRNA and protein decay rates has only a small impact on the position of genes in 2D Crick space and thus on present conclusions.

Sequencing data processing. We consider only genes for which the measurement error was small enough to allow accurate estimation of transcription and translation rates. Accurate estimation of these rates is difficult for low abundance mRNAs because they may only collect a handful of reads. This leads to a large uncertainty on the mRNA copy number \(m\), and thus on the transcription rate \(\beta_m = m \alpha_m\). How many reads per gene should we require to be confident about our estimates of transcription rates?

Estimates of mRNA abundance \(m\) scale with the number of reads \(r\) mapping to a given gene (relative to the gene length). We thus compute the minimal number of reads per gene needed to keep the sampling noise on log10 mRNA abundance below a certain threshold \(\varepsilon\):

\[
\log_{10}(\frac{n + \sigma}{n}) < \varepsilon
\]

where \(\sigma\) is the standard deviation on \(n\) due to the sampling error. We model sequencing as a Poisson process, and thus \(\sigma = \sqrt{n}\). Substituting this expression for \(\sigma\) into Eq. (13), we compute the minimal number of reads necessary to control for a given error \(\varepsilon\) on log10 mRNA abundances:

\[
n > \frac{\ln(10^{-\varepsilon})}{\varepsilon^2}
\]

A minimum of 10 reads per mRNA is needed to keep the sampling error on log10 transcription rates in the \(\pm 0.1\) range (Supplementary Fig. 2o). We therefore discard genes with less than 10 reads per gene, a procedure which keeps the sampling error as low while keeping as many genes as possible in the analysis. Similarly, we require at least ribosomal profiling 20 reads per gene. We applied the same criteria to all four organisms. We repeated the analyses keeping only genes with at least 100 RP reads and reached the same conclusions as the one presented in the article.

In M. musculus and H. sapiens, we discarded canonical histone genes from the analysis because their mRNAs lack a polyA-tail. The polyA- selection step of mRNAseq discriminates against these mRNAs. Consequently, the abundance of canonical histone mRNAs is underestimated by mRNAseq RPKMs, leading to aberrant (high) translation rate estimates.

Estimating maximal translation rates. To estimate the maximal translation rate, we ask how fast proteins can be translated \(\beta_p\) from a single mRNA in the limit where translation initiation is no longer limiting. In this regime, ribosomes follow each other closely along the mRNA. A given ribosome needs to move forward before the next one can advance. The speed at which ribosome elongate the peptide chain and how many codons each ribosome occupies on the mRNA determine how fast proteins can be synthesized.

If a ribosome occupies \(L\) codons on the mRNA and \(n\) codons are translated into amino-acids per second, it takes \(L/n\) seconds for a ribosome to free space for the next ribosome. The maximal flux of ribosomes at given codon is thus \(\pi L\). This flux sets an upper bound on the translation rate: \(\beta_p \leq \pi L\).

In S. cerevisiae, the elongation rate \(\pi\) in favorable growth conditions is 10 amino-acids per second (BNID107871). Each ribosome occupies 28 nucleotides (BNID107874), so \(L = 93\) codons. This puts the maximal translation rate \(\beta_p\) at 10^5 proteins per hour. This bound is in good agreement with estimated translation rate in S. cerevisiae and other eukaryotes.

In E. coli, translation rate can be up to 10^9 proteins per hour (Fig. 2). While the size of ribosomal footprints have not been determined, prokaryotes have smaller ribosomes (21 nm, BNID102320) than eukaryotes (26.5 nm, BNID111542) and so ribosomal footprints should be smaller. Assuming that ribosomal footprints are proportional to the size of the ribosome, we estimate that prokaryotic ribosomes cover 22 nm or \(L = 73\) codons. In favorable conditions, E. coli can elongate up to \(n = 21\) amino-acids per second (BNID100059). This leads to a maximal translation rate of \(\beta_p = 10^4\) proteins per hour.

Testing the statistical depletion of genes in Crick space. To test for a statistical significant depletion of genes combining high transcription with low translation, we used a randomization strategy.
and the data in all four organisms, using a uniform criteria. We then asked if this figure of 1% was high or low compared to chance. To find out, we randomized transcription and translation rates. Because the main goal of gene expression is to express proteins at the right abundance, we required that randomized datasets have the same distribution of protein abundance as the original dataset. Also, we enforced the observed upper bound \( \beta_{\text{max}} \) on translation rates. To do so, we randomly sampled a protein abundance \( p \) and a translation rate \( \beta \) for each gene. We then computed the corresponding transcription rates \( \beta_i = \mu_i \alpha_i \beta \) with \( \alpha_i \) and \( \mu_i \), the mRNA and protein decay rates reported in Supplementary Table 1. Finally, we determined what fraction of genes in the randomized dataset were found below the line of slope 1 and leaving 1% of the genes of the original dataset below it.

We repeated the procedure \( 10^4 \) times to determine the distribution of the fraction of genes in the depleted region expected by chance. We finally estimated the \( p \)-value that genes avoid combining high transcription with low translation from the fraction of randomized datasets with more genes in the depleted region than the original dataset.

### Estimating transcription and translation for synthetic genes

We compared the distribution of transcription and translation rates of *E. coli* genes to that of the synthetic constructs library of Kosuri et al.\(^{38}\). This study quantified the mRNA and protein abundance of each construct. The constructs only differed in their ribosomal binding sites and promoters. We thus assumed that they shared the same mRNA decay rates and protein decay rates. As a result, the transcription rate of each construct is proportional to mRNA abundance. Translation rates are proportional to the ratio of protein abundance to mRNA abundance.

To compare these measurements to our absolute transcription and translation rates estimates of *E. coli* genes, we assumed that the strongest promoters and RBPs of *E. coli*’s strongest promoters and RBPs were limiting. We did so by aligning the 99th percentiles of the transcription and translation rates of the synthetic constructs to those of *E. coli* genes.

The conclusions of the comparison are robust to this assumption. For instance, even if we assume that the strongest Kosuri promoters and RBPs achieve transcription rates 10 times higher or lower than the strongest *E. coli* promoters (i.e., shifting the red cloud of Fig. 2e to the right or to the left by one unit), the high translation–low translation region would still be covered by a sizable fraction of synthetic constructs.

#### Expression for the fitness cost of transcription

Experiments and theory suggest that the fitness cost of transcription \( \Delta \text{fit} \), scales with the transcription rates\(^{1,21,22}\) and (pre-)mRNA length\(^{19,27}\). For a pre-mRNA of length \( l_m \) and transcription rate \( \beta_m \), we thus write the fitness cost of transcription \( \Delta \text{fit} \) as

\[
\Delta \text{fit} = c_m \alpha_m \beta_m
\]

(15)

where the constant \( c_m \) scales transcription fluxes (nucleotides per hour) into fitness units (per hour).

The exact biochemical process responsible for the cost of transcription costs is still under debate. The main candidates include the limited availability of the RNA polymerase which limits the initiation step\(^{21}\) and the availability of nucleotides\(^{18}\).

To estimate the proportionality constant \( c_m \), we hypothesize that transcriptional resources are limiting. In this case, making one non-benefit of \( m \) and \( \beta_m \) with \( \alpha_m \).

By identifying the terms in Eqs. (15) and (16), we see that \( c_m \) can be estimated from the growth rate \( \mu \), the typical pre-mRNA length \( l_m \) and the total transcriptional capacity \( \sum \beta_m \).

\[
c_m = \frac{\mu}{l_m \alpha_m N_m} = \frac{\mu}{l_m \sum \beta_i}
\]

(17)

\( c_m \) has units of m\(^{-1} \). Alternatively, \( c_m \) can also be expressed per mRNA copy (which we will use in the next section):

\[
c_m = \frac{\mu}{m} = \frac{\mu}{N_m}
\]

(18)

where \( N_m \) is the number of mRNAs per cell.

#### Predicting mRNA cost from cellular constants in *E. coli*

In this section, we show that the expression for \( c_m \) derived in the previous section predicts the fitness cost of synthesizing mRNA in *E. coli*. For this, we use the data of Kosuri et al.\(^{38}\) who quantified the abundance of 10,000 different clones that express a fluorescent protein under the control of different promoters and RBSSs. As a result, different clones express the fluorescent protein at a different abundance \( p \), from a different number of mRNAs \( m \).

We model mRNA and protein cost \( c_m \) and \( c_p \) as linear penalties on the growth rate \( \mu \)

\[
\mu = \mu_0 - c_p \mu_m
\]

(19)

where \( \mu_0 \) is the rate at which cells would grow in the absence of a fluorescent protein construct. If cells are growing exponentially for \( t \) hours, the concentration \( x_m(t) \) of a clone that expresses \( m \) non-beneficial mRNAs and \( p \) proteins is

\[
x_m(t) = x_0m(t)e^{\mu t}
\]

(20)

We can normalize \( x_m(t) \) to the concentration of clones \( x_0m(t) \) that express \( m \) per low and hence don’t experience a growth penalty \( \mu = \mu_0 \):

\[
\log x_m(t) = -\mu_0 x_0m(t)\mu
\]

(21)

Here, we have assumed that the transformation efficiency of clones is independent of \( m \) and \( p \). We \( x_0m(t) \) or \( x_0m(t) \) above for fitness units (per hour).

The conclusions of the comparison are robust to this assumption. For instance, even if we assume that the strongest Kosuri promoters and RBPs achieve transcription rates 10 times higher or lower than the strongest *E. coli* promoters (i.e., shifting the red cloud of Fig. 2e to the right or to the left by one unit), the high translation–low translation region would still be covered by a sizable fraction of synthetic constructs.

#### Estimating the fitness cost of mRNA in *S. cerevisiae*

Here we estimate the growth penalty of transcription in *S. cerevisiae* from the measurements of Kafri et al.\(^{21}\). This study introduced a fluorescent protein construct of pre-mRNA length \( l_m \), at different copy numbers \( n \) in the *S. cerevisiae* genome. The protein abundance \( \beta_m \) of the fluorescent protein depended on the genomic copy number of the construct, as did the transcription rate \( \beta_m \).

This study found that the growth rate \( \mu \) decreases linearly with the genomic copy number \( n \) of the fluorescent protein construct,

\[
\mu = \mu_0 - n \left( c_p + c_m \beta_m \right)
\]

(24)

where \( \mu_0 \) is the growth rate of the WT strain, \( c_m \) is the growth penalty of transcription (expressed per transcribed nucleotide) and \( c_p \) is the protein burden (growth penalty per protein). To compare cost across growth conditions, the study normalized the growth rate \( \mu \) of strains with genomic insertions of the construct to that of WT \( \mu_0 \) (e.g., Fig. 4b of Kafri et al.\(^{21}\)):

\[
\mu = \frac{1}{\mu_0} \left( \mu_0 - n \left( c_p + c_m \beta_m \right) \right)
\]

(25)
Following the notation of Kafri et al.21, we call $x_\beta$ the slope of the relative growth rate $\mu_{fl}$ as function of the genomic copy number of the construct $n$. We can write $x_\beta$ in terms of the fitness cost parameters $c_\beta$ and $c_{\phi}$:

$$x_N = \frac{c_\beta p + c_{\phi} \mu_{fl}}{\mu_0} \quad (26)$$

To distinguish between the cost of protein and transcription, Kafri et al.21 designed a second construct (DAmP) lacking a terminator, which makes the mRNA unstable. This reduces transcription synergy between 10 fold (according to qPCR) and 30 fold (according to qPCR) with minimal effect on transcription21. Similar to $x_\beta$, Kafri et al.21 defined $x_\phi$ as the slope of $\mu_{fl}$ as a function of the genomic copy number $n$ of the DAmP construct. We can also write $x_\phi$ as a function of fitness cost parameters $c_\beta$ and $c_{\phi}$:

$$x_\phi = \frac{c_\beta \phi + c_{\phi} \mu_{fl}}{\mu_0} \quad (27)$$

where the factor $\phi$ accounts for the decrease in protein synthesis of the DAmP strains.

From experimental measurements of $x_N$, $x_\phi$, $p$, $l_\beta$, and $\beta_m$, one can estimate the translation cost per nucleotide $c_\beta$:

$$\phi x_N - x_\phi = \frac{c_\beta \mu_{fl} (\phi - 1)}{\mu_0} \Rightarrow c_\beta = \frac{x_\phi - x_N (\phi - 1)}{\mu_0} \frac{1}{\phi} \quad (28)$$

Experiments in YPD medium21 measured the following values: $\phi x_N \approx 9.0 \times 10^{-5}$, $x_\phi \approx 4.1 \times 10^{-3}$, $l_\beta \approx 1000 \text{ mRNAs}, \phi \approx 0.06$, $\beta_m \approx 1300 \text{ mRNAs} \cdot \text{h}^{-1}$, and $\mu_0 \approx 0.42 \text{ h}^{-1}$. Plugging in these values in Eq. (28) estimates the fitness cost $c_\beta$ per transcribed nucleotide,

$$c_\beta \approx 1.2 \times 10^{-9} \text{ nt}^{-1} \quad (29)$$

All parameters are known up to two significant digits, except for $\phi (10 \leq \phi < 30)$, which is known up to one significant digit. This puts the measurement uncertainty at $\pm 1 \times 10^{-9} \text{ nt}^{-1}$.

### Predicting mRNA cost from cellular constants in yeast.

Plugging the cellular constants of S. cerevisiae from Table 1 into the formula for $c_\beta$ (Eq. (17)) estimates the transcription cost at $c_\mu = \mu_0 / \sum \mu_{fl} \sim 1.2 \times 10^{-9} \text{ nt}^{-1}$. This value is in excellent agreement with the experimental measurements of Kafri et al.21. Such an agreement between experimental fitness parameters in the order of the 10^{-9} may appear surprising given the typical accuracy of biological measurements. The reason for such a good agreement is that the parameters are expressed per transcribed nucleotide whereas actual measurements were performed for full-length mRNAs of a highly transcribed S. cerevisiae gene. Because both the transcript length $l_\beta$ and the transcription rates $\beta_m$ are in the order of 10^3 in the experiments of Kafri et al.21, we are actually comparing fitness parameters in the order of 10^{-9}, which are accessible experimentally.

### Plotting the coefficient of variation over Crick space.

We obtained measurements of the coefficients of variation $c_\beta$ of S. cerevisiae and E. coli genes from the studies of Newman et al.17 and Taniguchi et al.40. To determine the contours of $c_\beta$ as a function of transcription and translation, we first applied a Gaussian smoother. The smoother estimates the coefficient of variation $c_\beta$ for given transcription $\beta_m$ and translation $\beta_p$ rates from a weighted average of genes with similar $\beta_p$ and $\beta_m$. Genes with comparable $\beta_p$ and $\beta_m$ weight more in the average than genes with very different $\beta_p$ and $\beta_m$.

Formally, we estimated $c_\beta(\beta_m, \beta_p)$ as a weighted average,

$$c_\beta(\beta_m, \beta_p) = \frac{1}{\sqrt{2\pi} \sigma} \sum_i \sum_j w_i \sigma_j c_{ij} \quad (30)$$

where the Gaussian weights $w_i$ are defined as:

$$w_i = e^{-\frac{\left(\frac{\beta_m - \mu}{\sigma}\right)^2}{2}} \left(\frac{\beta_m - \mu}{\sigma}\right)^3 \quad (31)$$

We set the smoothing width $\sigma$ to one fifth of the data range. We only plotted contours of $c_\beta(\beta_m, \beta_p)$ for densely populated regions of the Crick space ($\sum w_i > 200$).

### Protein fluctuations as a function of the central dogma rates.

In this section, we derive an expression for the coefficient of variation $c_\beta$ as a function of the transcription rate $\beta_m$ and the protein abundance $p$ and the protein decay rate $\alpha_p$. Following a extensive line of theoretical and experimental research52-58, we model gene activation and inactivation as a telegraph process (Supplementary Fig. 3c). Genes are activated at a rate $k_{on}$ and inactivated at a rate $k_{off}$. Active genes synthesize mRNAs at a rate $\delta$. Messenger RNAs are translated into proteins at a rate $\beta_p$ and degrade at a rate $\alpha_p$. At steady-state, the coefficient of variation $c_\beta$ on protein abundance of this stochastic process can be computed analytically59:

$$c_\beta^2 = \left(\frac{\beta_p}{\beta_m}\right)^2 = \frac{1}{\beta_p} + \frac{\alpha_p}{\beta_m} c_\beta^2 \quad (32)$$

The first term of the equation accounts for the Poisson noise on protein abundance stemming from the protein birth-death process. The second term accounts for the noise caused by translating proteins from mRNAs of low copy number. The last term models the noise caused by gene activation and inactivation and transcriptional bursting.

At present time, it is difficult to measure $k_{on}$ and $k_{off}$ genome-wide experimentally. We therefore seek a simplified, approximate expression for the coefficient of variation $c_\beta$ in which these parameters occur only implicitly through the transcription rate $\beta_m$. Note that $\beta_m$ and $k_{on}$ should be related to each other.

$$\beta_m = \delta k_{off} = \frac{\beta_m}{k_{on} + k_{off}} \quad (33)$$

The gene activation rate $k_{on} \approx 10^{-11}$ is largely constant across E. coli promoters, in contrast to $k_{off}$ which determines the transcription rate $\beta_m$. Using Eq. (33) which relates the transcription rate $\beta_m$ to gene (in-)activation rates $k_{on}$ and $k_{off}$, we can rewrite the coefficient of variation in terms of $\beta_m$.

The Poisson noise term $1/\beta$ is typically negligible compared to the other terms. Small mRNA copy number noise term $\alpha_p/\beta_m$ dominates at low $\beta_m$ (Supplementary Fig. 3d). This term is consistent with the observation that protein noise initially decreases with protein abundance (Fig. 4c, Supplementary Fig. 4b, c), and that protein noise decreases with transcription (Fig. 3b).

The third term (gene activation noise) becomes dominant for large $\beta_m$ (Supplementary Fig. 3d). Because the third term is almost a constant for physiological values of $\beta_m$ (Supplementary Fig. 3d), we ask whether it could explain the noise floor found in the single cell experiments of Taniguchi et al.40 and Silander et al.41.

In the growth conditions of Taniguchi et al.40, the doubling time was 150 min, which implies a protein decay rate $\alpha_p = 0.28 \text{ h}^{-1}$. Plugging this $\alpha_p$ in Eq. (35), the third term is about 0.03, which is two-fold below the noise floor $c_\beta^2 = 0.07$ observed in the measurements of Taniguchi et al.40 (Fig. 4c). In Silander et al.43 grew E. coli in M9 + 0.2% arabinose, a condition in which $\alpha_p \approx 0.45 \text{ h}^{-1}$. With this $\alpha_p$, the third term is about 0.04, which is comparable to the noise floor $c_\beta^2 \approx 0.05$ in the measurements of Silander et al.43 (Supplementary Fig. 4b).

We conclude that gene activation noise may explain the noise floor in the measurements of Silander et al.43. In the experiments of Taniguchi et al.40, gene activation noise is too small to explain the noise floor. There, the noise floor may be explained by other factors such as extrinsic noise. Independently of the specific cause of the noise floor, we model it as a phenomenological constant $c_\beta^2 \approx 0.06$ inferred from the measurements of Taniguchi et al.40 and Silander et al.43 (or $c_\beta^2 \approx 0.25$). This yields an expression for the noise that is accurate with both datasets,

$$c_\beta^2 = \frac{\alpha_p}{\beta_p} + \frac{\alpha_p}{\beta_m} c_\beta^2 \quad (36)$$

**S. cerevisiae, H. sapiens, M. musculus:** In eukaryotes, an approximate expression for the coefficient of variation can also be derived, but is slightly more complicated because the separation of time-scales is less clear than in E. coli: messenger RNAs decay typically faster than proteins, but not by a full order of magnitude. A more realistic, data-driven assumption is

$$\alpha_m = q \beta_p \quad (37)$$
with $q \approx 3$ (Supplementary Table 1). Measurements in *S. cerevisiae* and *H. sapiens* suggest that gene activation dynamics are much faster than protein decay,

$$\kappa_m + \kappa_{off} \gg \alpha_p. \quad (38)$$

Under these assumptions, we can approximate the analytical expression for the coefficient of variation (Eq. (32)) as

$$\left(\frac{\sigma}{\mu}\right)^2 \approx \frac{1}{p} + \frac{\alpha_p}{\beta_m} \frac{q}{p + \frac{1}{q}} \left(1 + \frac{\delta_{koff} \beta_m}{\kappa_m (\kappa_m + \kappa_{off})}\right). \quad (39)$$

Using Eq. (33) which expresses transcription $\beta_m$ as a function of gene (in)activation parameters $\kappa_m, \kappa_{off}$, we can eliminate the last term from (39), setting $\beta_m$ from the previous step.

$$\left(\frac{\sigma}{\mu}\right)^2 \approx \frac{1}{p} + \frac{\alpha_p}{\beta_m} \frac{q}{p + \frac{1}{q}} \left(1 + \frac{\delta_{koff} (1 - P_m)^{-1}}{C_0}\right) \quad (40)$$

We also eliminate $\delta$ by introducing the transcriptional burst size $b$, which is the average number of mRNAs that are synthesized each time the gene is activated:

$$b = \frac{\delta_{koff}}{C_0} \quad (41)$$

Plugging Eq. (41) for $b$ into Eq. (40) for $\sigma_p$,

$$\left(\frac{\sigma}{\mu}\right)^2 \approx \frac{1}{p} + \frac{\alpha_p}{\beta_m} \frac{q}{1 + \frac{1}{q}} \quad (42)$$

where

$$\phi = \frac{q}{1 + \frac{1}{q}} \left(1 + b (1 - P_m)^{-1}\right) \quad (43)$$

accounts for gene activation dynamics.

Except for highly expressed genes, the transcriptional burst size $b$ is typically small ($b \approx 1$). For $q = 3$ (Supplementary Table 1), varying $P_m$ across its full range causes $\phi$ to vary only between 0.75 and 3 (Supplementary Fig. 3e). Hence, in the worst case, neglecting gene activation dynamics by setting $\phi = 1$ would result in a 1.5-fold error on $\sigma_p$. We conclude that neglecting gene activation dynamics by setting $\phi = 1$ in Eq. (42) for $\sigma_p$ results in a reasonable approximation of the coefficient of variation for most genes, except for highly expressed genes.

For highly expressed genes, experiments in *S. cerevisiae* and *H. sapiens* found a noise floor. This noise floor might occur when the transcription rate exceeds the maximal rate of gene activation. In this scenario, high transcription rate can only be achieved by increasing the burst size $b^{12,42}$. Since $b < 1$ is a constant in the large $P_m$ regime, we rewrite $b$ as the term of the transcription rate $P_m$. To do so, we combine Eq. (33) which expresses $P_m$, in terms of the kinetic rates of gene activation ($\kappa_m, \kappa_{off}, \delta$) and Eq. (41) which defines the burst size $b$ in terms of $\delta$ and $\kappa_{off}$ to find

$$b = \frac{\delta_{koff}}{\kappa_m + \kappa_{off}}. \quad (44)$$

Plugging this expression for $b$ in Eq. (42) eliminates the burst size:

$$\left(\frac{\sigma}{\mu}\right)^2 \approx \frac{1}{p} + \frac{\alpha_p}{\beta_m} \frac{q}{1 + \frac{1}{q}} + \frac{q}{k_{off}} \frac{(1 - P_m)^{-1}}{C_0}. \quad (45)$$

Neglecting the $1/p$ term which is in the order of $10^{-3}$ or smaller, we can get an expression for the noise floor $\sigma_p$ by taking the limit of large $P_m$. In this limit, the second term in Eq. (45) vanishes and $\sigma_p$ approaches

$$\sigma_p^2 \approx \frac{q}{1 + \frac{1}{q}} \frac{(1 - P_m)^{-1}}{k_{off}}. \quad (46)$$

Plugging in gene activation parameter values typical for *H. sapiens* ($P_m = 0.18$, $k_{off} = 2$ h$^{-1}$) and setting $q = \alpha_p/\alpha_p + \delta_m = 3$, $\alpha_p = 0.05$ h$^{-1}$ (see Table 1) puts the noise floor $\sigma_p$ at 0.27, a value comparable to experimental observations$^{42}$ ($q \approx 0.3$).

Hence, the noise floor on protein abundance could occur when transcription rates saturate gene activation kinetics. In this case, substituting Eq. (46) for $\sigma_p$ into Eq. (45) leads to:

$$\sigma_p^2 \approx \frac{1}{p} + \frac{\alpha_p}{\beta_m} \frac{q}{1 + \frac{1}{q}} + \frac{\alpha_p}{\beta_m} \frac{q}{k_{off}}. \quad (47)$$

This final expression is identical to the one we previously derived for *E. coli* (Eq. (36)). It would also hold if the noise floor was caused by a mechanism different from the transcriptional saturation of gene activation dynamics, such as extrinsic noise. Note that we neglected the $q/(1+q)$ term of Eq. (45). This is because the resulting approximation error on $\sigma_p^2$ would be less than 50% (Supplementary Fig. 3e). Because $\beta_m$ scales inversely with $\sigma_p^2$ (Eq. (47)), neglecting gene activation kinetics implies at most a 50% error on $\beta_m$, or equivalently a 0.2 error on log$_2$$\beta_m$. This is small compared the dynamic range of transcription rates which vary over 2–3 order of magnitudes.

**Expression for the optimal translation/transcription ratio.** We consider a protein of abundance $p$. The protein contributes a quantity $f(p)$ to the organism’s fitness. The cost of transcription $\Delta_m$ is linear in the transcription rate $\beta_m$ and the fitness cost of transcription per nucleotide $c_m$.

$$\Delta_m = c_m \ln \beta_m. \quad (48)$$

The overall fitness is thus $F = f(p) - \Delta_m$. The fitness function $f(p)$ reaches its maximum $f_{max}$ at $p = p^*$. Expanding $f(p)$ around the optimum $p = p^*$ to second order, the overall fitness becomes

$$F(p, \beta_m) \approx f_{max} + f' (p^*) (p - p^*) + \frac{1}{2} f'' (p^*) (p - p^*)^2 - \Delta_m. \quad (49)$$

Plugging $f(p) = 0$ since $f(p)$ reaches its optimum $f_{max}$ at $p = p^*$, $f'' (p^*)$ is the curvature of the fitness function at its maximum. It is a negative number which characterizes how narrow the fitness function is.

Because protein abundance fluctuates around $p^*$, the cell doesn’t experience the maximum fitness $f_{max}$ but rather a lower average fitness $f$. Averaging $F$ over fluctuations in protein abundance, we obtain

$$\langle F(p, \beta_m) \rangle = f_{max} + f' (p^*) \alpha^2 - c_m \ln \beta_m. \quad (50)$$

where $\alpha^2$ is the variance of protein abundance fluctuations (Fig. 4a). The curvature times this variance, $-f'' (p^*) / \alpha^2$, is the noise load $\Delta_{noise}$, the fitness lost due to the stochastic fluctuations in protein abundance$^{26,25}$ (Fig. 4a). It is a positive quantity because the curvature $f''(p)$ is negative at the maximum $p = p^*$.

We seek the transcription rate $\beta_m$ that maximizes fitness. For this purpose we note that $\beta_m$ affects the noise level $\sigma^2$ in a well-characterized way. Theory and experiments of gene expression noise$^{15,17,40}$ indicate that variance of protein noise is given by

$$\sigma^2 \approx p^* \left(\frac{1}{\beta_m} + \frac{\alpha_p}{\beta_m} + \frac{\alpha_p}{\beta_m}\right) \quad (51)$$

where $\alpha_p$ is the protein decay rate and $\alpha_p$ is the noise floor due to extrinsic noise$^4$ or the larger transcriptional burst size of high abundance proteins$^{42}$ (see previous section). We can now solve for the $\beta_m$ that maximizes fitness, by finding $\frac{\partial F}{\partial \beta_m} = 0$.

The optimal translation/transcription ratio rate $\beta_m$ satisfies

$$\beta_m = \frac{2c_m \ln \beta_m}{C_0} = \frac{C}{Q} \quad (52)$$

where we have used $p^* = \frac{\beta_m \alpha_p}{\alpha_p}$. Note that we introduced two new variables $C$ and $Q$.

$$C := c_m \ln \beta_m - c_m \ln \beta_m / m = \Delta_{noise} / m \quad (53)$$

represents the fitness cost of transcription per mRNA molecule for a gene with pre-mRNA length $l_m$.

$$Q := -

\frac{1}{f'' (p^*)} \quad (54)$$

is the gene’s sensitivity to noise. Genes with narrower fitness functions ($f'' (p^*) < 0$) are more noise sensitive. A gene’s sensitivity to noise also depends on $p^*$ because stochastic fluctuations $\alpha^2$ scale with $p^*$ (Eq. (51)).

From the expression for the optimal $\beta_m$ ($\beta_m$) (Eq. (32)), we see that genes dominated by comparable requirements of precision and economy are predicted to share the same translation/transcription ratio. Genes with narrow fitness function require more precision and thus have lower translation/transcription ratios (Fig. 4b). Higher translation cost per mRNA molecule—due to longer mRNAs $l_m$ or rapid mRNA turnover/rna or a scarcity of nucleotides leading to increased cost $c_m$—shifts the balance towards higher ratios.

Note that the translation/transcription ratio does not depend on translation cost, although this cost is typically larger than transcriptional cost$^{16,38,31}$. This is because we assumed that, for a given protein abundance, translation cost are the same if the proteins are synthesized from few or many mRNAs.

**Predicting the boundary of the depleted region of Crick space.** In this section, we ask what is the predicted offset of the line that forms the boundary of the depleted region, namely the constant $k$ such that $\beta_m/k > k$ for all genes. We provide estimates based on known fundamental parameters of cell biology suggested by the theory.
To estimate k, we note that the precision-economy theory predicts that low $\beta_m/\beta_n$ occur for genes with narrow fitness functions (Eq. (52), Fig. 4b). But the noise floor $c_n$ (Fig. 4c) sets a limit on how narrow fitness functions can be: for fitness functions that are too narrow given the noise floor, average fitness is negative (Fig. 4d). Such fitness functions cannot be selected for in evolution. We can therefore estimate the maximal transcription rate by determining the largest, selectable fitness function curvature, and then compute the optimal transcription rate for that function. The curvature should be small enough that the smallest, unavoidable protein fluctuations (set by the organism’s noise floor $c_n$) do not dominate the fitness benefit of expressing the protein. In other words, the fitness benefit of expressing the protein $f_{\text{max}}$ should be larger than the noise load $\Delta_{\text{noise}}$:

$$f_{\text{max}} - \Delta_{\text{noise}} > 0 \Rightarrow f_{\text{max}} > \frac{1}{2} \left[ f'(p) \right] \sigma^2 > 0. \quad (55)$$

Here we neglected mRNA cost because for proteins with narrow fitness functions, it is small compared to the noise load $\Delta_{\text{noise}}$.

If fitness is mainly set by the growth rate $\mu$, the fitness contribution $f_{\text{max}}$ of a gene cannot be larger than the growth rate: $f_{\text{max}} < \mu$. In addition, fluctuations cannot be smaller than the noise floor, $a(p) > c_n$. From these two considerations, we can compute an upper bound on the noise sensitivity $Q$:

$$Q = \frac{1}{2} \left[ f'(p) \right] p < \frac{f_{\text{max}}}{p} < \frac{\mu}{p} c_n. \quad (56)$$

By substituting this upper bound into Eq. (52) for the optimal $\beta_m/\beta_n$, we find an upper bound on $p_{\text{max}}$ on transcription rates

$$\beta_m < p_{\text{max}} \leq \frac{1}{c_n} \sqrt{\frac{\mu}{\beta_n} \frac{\sigma_{\text{nuc}}}{C}}. \quad (57)$$

We now consider the (hypothetical) protein expressed at maximal transcription $p_{\text{max}}$ and maximal translation $\beta_{\text{max}}$. This protein has highest protein abundance $p$. It also has narrowest fitness function (narrower fitness are not selectable due to the noise floor), and thus highest noise sensitivity $Q_{\text{max}}$. We can plug Eq. (57) for $p_{\text{max}}$ into Eq. (52) for the optimal $\beta_m/\beta_n$ to find $Q_{\text{max}}$:

$$\frac{\beta_{\text{max}}}{\beta_m} = \frac{C}{Q_{\text{max}}} \Rightarrow Q_{\text{max}} = \frac{\sqrt{\alpha_p \sigma_{\text{nuc}} C}}{p_{\text{max}} c_n}. \quad (58)$$

In this expression, we can see that from a higher noise floor implies that genes need to be less sensitive to noise. We now use $Q_{\text{max}}$ in the Eq. (52) for the optimal $\beta_m/\beta_n$ to find $k$:

$$\frac{\beta_m}{\beta_{\text{max}}} > \frac{C}{Q_{\text{max}}} \Rightarrow k \rightarrow k = \frac{\sqrt{p_{\text{max}} c_n}}{\frac{\sqrt{\alpha_p \sigma_{\text{nuc}} C}}{p_{\text{max}} c_n} \sqrt{\alpha_P \sum P_m}} \quad (59)$$

where we have used $C = c_n \alpha_P \sigma_{\text{nuc}} \alpha_P$ (Eq. (33)). The fitness cost per transcribed nucleotide $c_n$ can be estimated from the average contribution of each nucleotide of each mRNA to the organism’s fitness, $c_n = \mu \sum_\beta \beta_n p_n$ (Eq. (17)). Neglecting differences in mRNA length between genes, we finally find:

$$k \approx \frac{\rho \sigma_{\text{nuc}}}{\sqrt{\alpha_P \sum P_m}} \quad (60)$$

Thus, the lowest translation/transcription ratio can be predicted from measurable, fundamental parameters of each organism.

In this derivation, we have assumed that the gene’s contribution to fitness $f_{\text{max}}$ is smaller than the growth rate $\mu$, $f_{\text{max}} < \mu$ (Eq. (56)). For essential genes, $f_{\text{max}} \approx \mu$. For non-essential genes, we can use a tighter upper bound on $f_{\text{max}}$:

$$f_{\text{max}} < \rho \mu \quad (61)$$

with $0 < \rho < 1$. For example, if deleting a gene decreases fitness by 1% or less, we have $\rho = 0.01$. By repeating the derivation, we find

$$k \approx \frac{p_{\text{max}} c_n}{\sqrt{\alpha_P \sum P_m}} \quad (62)$$

Thus, for non-essential genes ($\rho < 1$), the predicted boundary of the depleted region has higher intercep.t

In the Supplementary Methods, we consider alternative theories to explain the boundary of the depleted region, such as the existence of a power-law scaling of mRNA noise with mRNA abundance. None of the alternative theories can explain the depleted region.

Code availability. We deposited the code used to produce the figures of the present manuscript at Mendeley data, https://doi.org/10.17632/2vbrg3w4p3.1 (https://data.mendeley.com/datasets/2vbrg3w4p3/draft?a=955cbd6-9f26-4fbb-970b-eeb0681c1f5e).
**Additional information**

*Supplementary Information* accompanies this paper at [https://doi.org/10.1038/s41467-018-07391-8](https://doi.org/10.1038/s41467-018-07391-8).

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

**Reprints and permission** information is available online at [http://npg.nature.com/reprintsandpermissions/](http://npg.nature.com/reprintsandpermissions/)

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

---

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit [http://creativecommons.org/licenses/by/4.0/](http://creativecommons.org/licenses/by/4.0/).

© The Author(s) 2019