doi:10.1038/nature22998

Table of contents

1 Minimal synthesis time of a set of r-proteins in *E. coli*

In *E. coli* there are 7,536 amino acids per ribosome [36], including S1, and the maximal peptide chain elongation rate is 22 amino acids per second [10]. These numbers give a minimum of 5.7 minutes for the time it takes to synthesize a complete set of r-proteins. In addition, it is estimated that in order to support elongation roughly 5.4 molecules of EF-Tu (394 amino acids), 0.8 molecules of EF-G (704 amino acids), and 0.18 molecules of EF-Ts (283 amino acids) are required per ribosome; and that in order to support initiation roughly 0.25 molecules of IF1 (72 amino acids), 0.3 molecules of IF2 (890 amino acids), and 0.2 molecules of IF3 (180 amino acids) are required per ribosome [10]. Taking these numbers into account gives a minimum of 10,598 amino acids that need to be synthesized per ribosome, thus amounting to no less than 8 minutes of ribosome time. Note that this is a minimum time, and that many other proteins are indirectly needed to ensure high elongation rates of the ribosome. However, for our purposes it is sufficient to note that ribosomes spend a significant fraction of their time making themselves and that there is a selective pressure to reduce that fraction. In what follows we neglect the turn-over of ribosomes, but taking this into account would only strengthen our conclusions further since ribosomes would then need to spend even more time reproducing themselves.

2 Deriving a simple bound on the fraction of time ribosomes must spend on r-protein production

In balanced growth cells must, on average, double all components each generation. Thus, if doubling the number of ribosomes takes some minimal amount of time, the average generation time, T_{gen}, could not drop below this number. Here we first consider the idealized scenario in which all r-proteins (including nascent peptides) are in the form of active ribosomes. Denoting the total amount of r-protein in the cell (measured in amino acids) by $N_{rp}^{total}(t)$, then

$$\frac{dN_{rp}^{total}(t)}{dt} = k\varphi\frac{N_{rp}^{total}(t)}{N} , \tag{1}$$

where k is the translation rate in amino acids per second, φ is the time fraction ribosomes spend elongating r-proteins, N is the number of amino acids in the ribosome, and $N_{rp}^{total}(t)/N$ is the total number of ribosomes in the cell. The doubling time in this idealized scenario serves as a lower bound on T_{gen} and is given by $\tau ln(2)/\varphi$, where $\tau = N/k$. Thus

$$T_{gen} \geq \tau ln(2)/\varphi , \tag{2}$$

and rearrangement gives

$$\varphi \geq \tau ln(2)/T_{gen} . \tag{3}$$

The factor $ln(2)$ reflects the idealized assumption that the investment in r-proteins immediately produces a return in terms of increased translation rate, i.e., that the delay between initiating the production of r-proteins and their eventual incorporation into ribosomes is negligible. That is a good approximation given the high number n of r-proteins, but we will relieve it in subsequent sections. Given the maximal translation rates observed, this means that in fast growth *E. coli* ribosomes must spend at least 20% of their time elongating r-proteins rather than making the rest of the proteome.

Eukaryotic cells take more time to double, but their ribosomes are also larger and slower. For example, cytosolic ribosomes in *S. cerevisiae* contain almost 13,000 amino acids which are roughly 75% more than the amount found in *E. coli*'s ribosome. In addition, these ribosomes have a maximal elongation rate of ~10 amino acids per second, i.e., roughly half that of *E. coli* ribosomes. The observed generation times of 75-90 minutes would thus require yeast ribosomes to spend at least 16-20% of their time on r-protein elongation, similarly to what is observed for fast growing *E. coli*. Such high demands are also expected in cells of higher organisms during critical points in their life cycle which require fast growth and proliferation, e.g. during development. The purpose of the main text is to point out that if not for many unusual features of ribosomes this requirement would be even more taxing. The unique features of ribosomes thus allow them to spend a shorter fraction of their time on self-production.

There could also be an efficiency advantage to producing r-proteins in small pieces due to the limited processivity of ribosomes, i.e., the fact that ribosomes sometimes terminate prematurely. It has been reported that as much as 10% of ribosomes in E. coli fail to finish translation [37]. However, even if those numbers were accurate, that does not mean that 10% are wasted, since premature termination can happen anywhere along the mRNA. If the premature termination occurred close to the middle of the mRNA on average (and due to the processive nature of the process, the average position for premature termination should be closer to the start site), the waste would be roughly 5% for a typical protein, and much smaller yet for a typical r-protein. This type of waste would also be reduced to some extent by making many small r-proteins, since everything up to the point of the premature termination is wasted, suggesting that the optimal number of separate r-proteins should even be slightly higher than predicted above. Accounting for premature termination thus supports qualitatively similar conclusions and for a similar reason: ribosomes should be made of many small r-proteins because it increases the efficiency of ribosome biogenesis.

However, here we do not explicitly take this effect into account, partly because it supports similar conclusions, but primarily because we simply do not believe the drop-off rates are that high in natural systems. Specifically, the frequency of premature termination is expected to depend strongly on amino acid composition and the availability of the various charged tRNAs. Perturbations of gene expression, which easily can create shortages of some charged tRNAs, can then easily lead to an over-estimate of the drop-off rate, e.g. if any proteins are over-expressed, as was the case for many studies reporting high drop-off rates. From an evolutionary perspective, we would also find it very peculiar if the translation machinery – which can achieve such a high accuracy that most protein molecules do not contain a single incorrect amino acid, even though most such proteins should still be functional – could not avoid a 10% risk of terminating prematurely. For the very long proteins in *E. coli*, which can be >1,500 amino acids long, the great majority of protein production events would then create a protein of incorrect size. It is therefore our belief that the actual premature termination rate in nature is substantially lower than estimated and that the effects are even smaller. However, our main conclusion here is still unchanged even if this is not the case.

3 rRNA production requires little ribosome involvement

The rRNA is synthesized by RNA polymerases. Assuming that most RNA polymerases are active and that most protein allocated to RNA polymerase mass is in the form of assembled RNA polymerases – both of which are imperfect but have little impact on the main result – one can approximate

$$\frac{dN_{rRNA}^{total}(t)}{dt} = \gamma \frac{N_{pol}^{ribo}(t)}{N_{pol}} , \tag{4}$$

for the total amount, $N_{rRNA}^{total}(t)$, of rRNA in the cell (measured in nucleotides). Here, γ is the transcription rate, $N_{pol}^{ribo}(t)$ is the total amount of amino acids in the form of RNA polymerases dedicated to the synthesis of rRNA, N_{pol} is the number of amino acids per RNA polymerase, and $N_{pol}^{ribo}(t)/N_{pol}$ is the number of RNA polymerases dedicated to the synthesis of rRNA.

RNA polymerases are in turn synthesized by ribosomes and we can thus write

$$\frac{dN_{pol}^{ribo}(t)}{dt} = k\psi \frac{N_{rp}^{total}(t)}{N} , \tag{5}$$

where ψ is the fraction of their time ribosomes spend synthesizing RNA polymerases that would in turn be dedicated to the synthesis of rRNA. Combining Eq. (4) and (5), it immediately follows that

$$\frac{d^2N_{rRNA}^{total}(t)}{dt^2} = \frac{\gamma k\psi}{N_{pol}} \frac{N_{rp}^{total}(t)}{N} = \frac{\gamma k\psi}{N_{pol}} \frac{N_{rRNA}^{total}(t)}{N_{rRNA}} , \tag{6}$$

where we have defined N_{rRNA} to be the total number of RNA nucleotides per ribosome, and noted that the number of ribosomes in the cell can then also be written as $N_{rRNA}^{total}(t)/N_{rRNA}$. The solution to Eq. (6) is once again exponential growth for which the doubling time is given by $\sqrt{\tau_{pol}\tau_{rRNA}/\psi} ln(2)$, where $\tau_{pol} = N_{pol}/k$ and $\tau_{rRNA} = N_{rRNA}/\gamma$. Since this is an idealized scenario, we have the following bound on the generation time

$$T_{gen} \geq \sqrt{\tau_{pol}\tau_{rRNA}/\psi} ln(2) . \tag{7}$$

3.1 Derivation of Eq. (4) in the paper

In balanced growth doubling times of the r-protein and rRNA mass must match in order to avoid imbalances. In the idealized scenario described above the equalities in Eqs. (2) and (7) hold and this in turn implies that

$$\frac{\tau ln(2)}{\varphi} = \sqrt{\frac{\tau_{pol}\tau_{rRNA}}{\psi}} ln(2) , \tag{8}$$

from which it follows that

$$\psi = \varphi^2 \frac{\tau_{pol}\tau_{rRNA}}{\tau^2} . \tag{9}$$

Taking the equality in (3) and using it to eliminate one power of φ we recover Eq. (4) in the text

$$\frac{\psi}{\varphi} = \frac{[T_{gen}/\tau] ln(2)}{[T_{gen}/\tau_{pol}][T_{gen}/\tau_{rRNA}]} . \tag{10}$$

In *E. coli* the transcription rate for rRNA is estimated at ~85 nucleotides per second and there are 4,566 RNA nucleotides per ribosome. It follows that τ_{rRNA} is ~54 seconds. In addition, there are 7,536 amino acids per ribosome and 4,320 amino acids per RNA polymerase (including σ^{70}). Since translation rates range from 13 amino acids per second at slow growth (T_{gen} = 100 minutes) to 22 amino acids per second at fast growth (T_{gen} = 20 minutes) we find

$$\frac{1}{280} \leq \frac{[T_{gen}/\tau]ln(2)}{[T_{gen}/\tau_{pol}][T_{gen}/\tau_{rRNA}]} \leq \frac{1}{56} \,. \tag{11}$$

This means that in an ideal setting the fraction of time ribosomes spend on producing r-proteins should be roughly two orders of magnitude higher than that spent on the production of rRNA. The argument in the main text that rRNA requires a much smaller ribosomal time-investment per mass is thus highly robust to imperfections in the estimated numbers.

Depending on growth conditions, only about a sixth (slow growth) to a third (fast growth) of RNA polymerases in *E. coli* are active [10] and ψ above should thus be correspondingly higher in order to compensate. This can be corrected for in Eq. (9) but is hardly enough to reverse the trend. Moreover, accounting for auxiliary production costs further solidifies the conclusion that r-proteins are much more costly than rRNAs. The additional cost coming from the need to produce polymerases to support transcription of mRNAs coding for r-proteins adds to the r-protein burden, but probably not too much because of the strong amplification of transcriptional output by translation [38]; And this cost may also be further reduced by coupling between transcription and translation as this could prevent back-tracking of polymerases [39,40]. However, recent proteomics studies suggest that the amount of protein required in order to produce the nucleotides, charged tRNAs, amino acids, initiation and elongation factors that are needed for the r-protein production chain is much greater than that required for the synthesis of the nucleotides needed for rRNA [5,41].

4 Nascent ribosomal peptides are idle forcing more ribosome involvement in r-protein production

The calculation above assumed that all r-proteins mass is in the form of active ribosomes, thus neglecting the fact that some r-protein mass is nascent, i.e., in the process of being translated. To take this into account we let n denote the number of r-proteins and let L_i denote the length in codons of the mRNA transcript coding for the i-th ribosomal protein. Note that L_i is also the length in amino acids of the i-th ribosomal protein. In what follows we keep the total number of amino acids in the ribosome, $N = \sum_{i=1}^{n} L_i$, fixed and compute how the minimal fraction of ribosomes dedicated to the synthesis of ribosomal proteins depends on n and the set $\{L_1, \cdots, L_n\}$. This is again an idealized case because we ignore other sources for ribosome idleness that will be included in subsequent sections. We further note that the effect can be approximately understood from the simple argument in the main text: the length of the nascent peptides is roughly half the length of the full protein because the average position of the ribosome is close to the middle of the mRNA. However, because of the autocatalytic nature of the process, the average position is in fact not exactly in the middle. To intuitively understand this, for example imagine that the ribosome consisted of a single large r-protein and that ribosomes only made themselves. At balanced growth, there would then be twice as many ribosomes initiating rather than terminating the translation of the r-protein, much like populations of exponentially growing cells have twice as many newborn as dividing cells. Thus the average position of the ribosome is closer to the initiation site. This effect is minor when the ribosome consists of many r-proteins or when ribosomes also translate other proteins, but should still be accounted for. We further account for the fact that the r-proteins are all of different lengths. This section thus derives a more complete but mathematically more detailed description, and in subsequent sections we analyze the results with further emphasis on intuition.

Let $\rho_i(x,t)$ denote the r-protein mass density (coming from elongating ribosomes) at position x and time t on mRNAs of type i, and observe that this density obeys a simple transport equation

$$\frac{\partial \rho_i(x,t)}{\partial t} = -k\frac{\partial \rho_i(x,t)}{\partial x}\,, \tag{12}$$

$(1 \le i \le n,\ 0 \le x \le L_i)$. This approximates elongation as being continuous rather than taking discrete jumps between codons, but that approximation is virtually perfect even for the shortest r-proteins. The probability density to locate a ribosome at position x and time t on an mRNA of type i is then found by normalizing by the total number

$$p_i(x,t) = \rho_i(x,t)/N_{rp}^{total}(t)\,. \tag{13}$$

where $N_{rp}^{total}(t)$ is once again the total amount of r-protein in the cell. Combing Eqs. (12) and (13) we see that the probability densities $p_i(x,t)$ obey the following differential equation

$$\frac{\partial p_i(x,t)}{\partial t} = \frac{1}{N_{rp}^{total}(t)}\frac{\partial \rho_i(x,t)}{\partial t} - \frac{dN_{rp}^{total}(t)}{dt}\frac{\rho_i(x,t)}{\left[N_{rp}^{total}(t)\right]^2} = -k\frac{\partial p_i(x,t)}{\partial x} - \frac{dN_{rp}^{total}(t)}{dt}\frac{p_i(x,t)}{N_{rp}^{total}(t)}\ . \tag{14}$$

At steady state $\frac{\partial p_i(x,t)}{\partial t} = 0$. Setting $\pi_i(x) = \underset{t\to\infty}{lim} p_i(x,t)$ to be the steady state probability densities, we have

$$\frac{\partial \pi_i(x)}{\partial x} = -\frac{\pi_i(x)}{kN_{rp}^{total}(t)}\frac{dN_{rp}^{total}(t)}{dt} \tag{15}$$

which further implies that at steady state $\frac{1}{N_{rp}^{total}(t)}\frac{dN_{rp}^{total}(t)}{dt}$ must equal some unknown constant which does not depend on time. Moreover, since this constant determines the doubling time of r-protein mass it sets a lower bound for the cell generation time. This means that in the best case scenario

$$T_{gen} = \lim_{t\to\infty}\left\{\frac{ln(2)N_{rp}^{total}(t)}{dN_{rp}^{total}(t)/dt}\right\}, \tag{16}$$

which in turn gives

$$\frac{\partial \pi_i(x)}{\partial x} = -\frac{ln(2)}{kT_{gen}}\pi_i(x)\,. \tag{17}$$

The solution to Equation (17) is given by

$$\pi_i(x) = A_i e^{-xln(2)/kT_{gen}}, \tag{18}$$

where $\{A_1, \cdots, A_n\}$ are unknown constants to be determined.

In order to determine the unknown constants above we impose the boundary condition

$$k\rho_1(L_1, t) = ... = k\rho_n(L_n, t), \tag{19}$$

which asserts coordinated production of ribosomal proteins, and implies that

$$A_i = Ce^{\frac{L_i ln(2)}{kT_{gen}}}, \tag{20}$$

where C is some constant that does not depend on the index i. To determine it we observe that

$$\frac{dN_{rp}^{total}(t)}{dt} = \frac{k}{N}\sum_{i=1}^{n}\int_0^{L_i}\rho_i(x,t)dx\ . \tag{21}$$

Dividing both sides by $N_{rp}^{total}(t)$, taking the limit $t \to \infty$, and utilizing Eqs. (13) and (16), we see that at steady state

$$T_{gen}^{-1} ln(2) = \frac{k}{N}\sum_{i=1}^{n}\int_0^{L_i}\pi_i(x,t)dx = \frac{CkT_{gen}}{\tau ln(2)}\sum_{i=1}^{n}\left[e^{\frac{L_i ln(2)}{kT_{gen}}} - 1\right], \tag{22}$$

which gives

$$C = \frac{\tau ln^2(2)}{kT_{gen}^2\sum_{i=1}^{n}\left[e^{\frac{L_i ln(2)}{kT_{gen}}} - 1\right]}, \tag{23}$$

and

$$A_i = \frac{\tau ln^2(2) e^{\frac{L_i ln(2)}{kT_{gen}}}}{kT_{gen}^2\sum_{i=1}^{n}\left[e^{\frac{L_i ln(2)}{kT_{gen}}} - 1\right]}\ . \tag{24}$$

We would now like to determine the time fraction φ ribosomes spend on r-protein translation. To do this, we first observe that not all r-protein mass is found in the form of assembled ribosomes. Indeed, a quantity

$$N_{rp}^{nascent}(t) = \frac{1}{N}\sum_{i=1}^{n}\int_0^{L_i}x\rho_i(x,t)dx\ . \tag{25}$$

is found in the form nascent, or semi-translated, ribosomal peptides and we thus have

$$\varphi = \frac{\sum_{i=1}^{n}\int_0^{L_i}\rho_i(x,t)dx}{N_{rp}^{total}(t) - N_{rp}^{nascent}(t)}\ . \tag{26}$$

Dividing throughout by $N_{rp}^{total}(t)$, going to steady state, utilizing Eq. (22), and recalling that we still neglect some other sources for ribosome idleness (such as noisy expression or the time spent initiating) we find

$$\varphi \geq \frac{\sum_{i=1}^{n}\int_{0}^{L_i}\pi_i(x)dx}{1-\frac{1}{N}\sum_{i=1}^{n}\int_{0}^{L_i}x\pi_i(x)dx} = \frac{\tau ln(2)/T_{gen}}{1-\frac{1}{N}\sum_{i=1}^{n}\int_{0}^{L_i}x\pi_i(x)dx}\,, \tag{27}$$

and comparison to Eq. (3) makes it clear that idleness coming from nascent ribosomal peptides results in higher demand for ribosome involvement in r-protein production. Working out the sum and integrals in the denominator and rearranging we conclude that

$$\varphi \geq \frac{\tau ln(2)/T_{gen}}{1+\frac{kT_{gen}C}{ln(2)}-1} = \sum_{i=1}^{n}\left[e^{\frac{L_i ln(2)}{kT_{gen}}}-1\right] = \sum_{i=1}^{n}\left[e^{\frac{\tau_i ln(2)}{T_{gen}}}-1\right]\,, \tag{28}$$

with $\tau_i = L_i/k$. In the subsequent sections we consider more intuitive formulations and special cases of this result.

4.1 Derivation of Eq. (1) in the paper — smaller r-proteins reduce idleness

Eq. (1) in the paper could be obtained from Eq. (28) by looking at the case where all r-proteins have equal lengths. In this case $\tau_i = N/nk = \tau/n$, and

$$\varphi \geq n(2^{\tau/nT_{gen}}-1)\,. \tag{29}$$

Rearrangement gives the equivalent formulation in terms of a bound on the generation time

$$T_{gen} \geq \tau/[nLog_2(1+\varphi/n)]\,. \tag{30}$$

This result could be further interpreted and understood by recalling that

$$Log_2(1+x) = \frac{x}{ln(2)} - \frac{x^2}{2ln(2)} + o(x^2)\,. \tag{31}$$

Equation (30) could thus be written approximately as

$$T_{gen} \geq \frac{\tau ln(2)}{\varphi(1-\varphi/2n)} \simeq \frac{\tau(1+\varphi/2n)ln(2)}{\varphi}\,, \tag{32}$$

which in turn means that to a good approximation Eq. (32) could have been obtained directly from Eq. (2) by replacing τ there with $\tau_{eff} = \tau(1+\varphi/2n)$. This should make intuitive sense because from every ribosome which is engaged in the translation of r-proteins there hangs a nascent, or semi-translated, ribosomal peptide which is roughly $N/2n$ amino acids long, i.e., half the length of the mature protein on average. This could be seen as an effective increase in r-protein mass but since it is only a fraction φ of all ribosomes that are engaged in the translation r-proteins, the effective number of amino acids added, per ribosome, is $\varphi N/2n$. Replacing the number of amino acids in the ribosome, N, with $N_{eff} = N(1+\varphi/2n)$, and $\tau = N/k$ with $\tau_{eff} = N_{eff}/k$ then produces the same result. This makes it easy to understand why having more and smaller r-proteins will reduce the penalty coming from nascent ribosomal peptides that cannot contribute to production, and that in the limit $n \to \infty$ Eqs. (29) & (30) reduce to Eqs. (3) & (2) as expected.

4.2 Derivation of Eq. (3) in the paper — r-proteins with similar lengths reduce idleness

To derive Eq. (3) we note that when r-proteins are not equally sized we can always write

$$\tau_i = \tau/n + \delta(\tau_i) \,, \tag{33}$$

where $\delta(\tau_i)$ is defined as the difference between τ_i and τ/n. Combining Eqs. (28) and (33) we have

$$\varphi \geq \sum_{i=1}^{n} \left[e^{\frac{\tau ln(2)}{nT_{gen}}} e^{\frac{\delta(\tau_i) ln(2)}{T_{gen}}} - 1 \right] . \tag{34}$$

We now observe that

$$e^x \geq 1 + x + x^2/2 \,, \tag{35}$$

for $x \geq 0$, from which it follows that

$$\varphi \geq n(2^{\tau/nT_{gen}} - 1) + \frac{2^{\tau/nT_{gen}}}{2} \sum_{i=1}^{n} \left(\frac{\delta(\tau_i) ln(2)}{T_{gen}} \right)^2 , \tag{36}$$

where we have further used the fact that $\sum_{i=1}^{n} \delta(\tau_i) = 0$ by definition. Denoting the mean and variance of the translation times $\{\tau_1, ..., \tau_n\}$ by $\langle \tau_i \rangle = \frac{1}{n}\sum_{i=1}^{n} \tau_i = \frac{\tau}{n}$ and $\sigma^2(\tau_i) = \frac{1}{n}\sum_{i=1}^{n} (\tau_i - \langle \tau_i \rangle)^2 = \frac{1}{n}\sum_{i=1}^{n} \delta^2(\tau_i)$ correspondingly, we find

$$\varphi \geq n(2^{\tau/nT_{gen}} - 1) + 2^{\tau/nT_{gen}} \frac{\tau^2 ln(2)^2}{2nT_{gen}^2} \frac{\sigma^2(\tau_i)}{\langle \tau_i \rangle^2} . \tag{37}$$

Equation (3) in the paper then follows by noting that $\tau_i = L_i/k$. Hence, the coefficient of variation (ratio between standard deviation and mean) of the translation times $\{\tau_1, ..., \tau_n\}$ is effectively the coefficient of variation for r-protein lengths $\{L_1, ..., L_n\}$. Slight rearrangement then gives

$$\varphi \geq n \left(2^{\tau/nT_{gen}} \left[1 + \frac{1}{2} \left(\frac{\tau ln(2)}{nT_{gen}} \right)^2 CV_L^2 \right] - 1 \right) , \tag{38}$$

and we see that having r-proteins which are more equally sized (smaller CV_L^2) reduces the idleness coming from nascent ribosomal peptides. Also, note that in the limit $CV_L^2 \to 0$ Eq. (38) reduces to Eq. (29) as expected.

Some intuition for the result in Eq. (37) could once again be gained by expanding its right hand side to first order in $1/n$. Recalling that

$$2^x = 1 + xln(2) + \frac{1}{2}\big(xln(2)\big)^2 + o(x^2) \,, \tag{39}$$

we see that

$$n(2^{\tau/nT_{gen}} - 1) + 2^{\tau/nT_{gen}} \frac{\tau^2 ln(2)^2}{2nT_{gen}^2} \frac{\sigma^2(\tau_i)}{\langle \tau_i \rangle^2} \simeq \frac{\tau ln(2)}{T_{gen}} + \frac{\tau^2 ln(2)^2}{2nT_{gen}^2} \left(1 + \frac{\sigma^2(\tau_i)}{\langle \tau_i \rangle^2} \right) , \tag{40}$$

which in turn means that to first order in $1/n$ Eq. (37) is equivalent to

$$T_{gen} \geq \frac{\tau\left[1+\frac{\varphi}{2n}\left(1+\frac{\sigma^2(\tau_i)}{\langle\tau_i\rangle^2}\right)\right]ln(2)}{\varphi}\,. \tag{41}$$

Equation (41) could have been obtained directly from Eq. (2) by replacing τ there with $\tau_{eff} = \tau\left[1+\frac{\varphi}{2n}\left(1+\frac{\sigma^2(\tau_i)}{\langle\tau_i\rangle^2}\right)\right]$. This makes sense because from every ribosome engaged in the translation of r-proteins there hangs a nascent, or semi-translated, ribosomal peptide; and while the length of the latter is roughly $N/2n$ amino acids—the fact that r-proteins are not exactly equally sized will slightly change this result. Indeed, the average length of the nascent peptide associated with ribosomes translating the i-th r-protein is roughly $\frac{1}{2}L_i$, but the frequency in which different nascent peptides occur is not equal throughout. For a given production rate, ribosomes will spend twice as much time on transcripts which are twice as long, and there would hence be roughly twice as many ribosomes translating these transcripts. In other words, the probability to find a ribosome on a transcript of a certain type is proportional to the length of this transcript, thus suggesting that the typical length of a nascent peptide is determined by the following weighed average

$$\text{Average length of nascent peptide} = \sum_{i=1}^{n}\frac{1}{2}L_i p_i\,, \tag{42}$$

where $p_i = L_i/N$. This in turn gives

$$\text{Average length of nascent peptide} = \frac{N}{2n}\left[1+CV_L^2\right], \tag{43}$$

where CV_L^2 is the coefficient of variation of the set $\{L_1, .., L_n\}$. Since only a fraction φ of all ribosomes are engaged in the translation r-proteins, we once again note that this amounts to an effective increase of $\frac{\varphi N}{2n}\left[1+CV_L^2\right]$ in the number of amino acids per ribosome. Replacing the number of amino acids in the ribosome, N, with $N_{eff} = N\left(1+\frac{\varphi}{2n}\left[1+CV_L^2\right]\right)$, and $\tau = N/k$ with $\tau_{eff} = N_{eff}/k$ then gives the desired result.

5 Ribosomal proteins are unusually small and similarly sized

As shown in the main text, all cytosolic r-proteins are on average much smaller than the corresponding genomic averages. However, because many other proteins are smaller yet, the question is if this difference is so unusual. That is, is the set of r-proteins that make up a ribosome significantly different (statistically) from a random sets of proteins taken from the genome? Here we perform the detailed calculations for *E. coli* that contains 56 r-proteins that average 132 amino acids. Indeed we find that the probability of averages of 132 amino acids or lower is exceedingly small. In fact, after picking 10^6 random sets of 56 proteins from the genome we did not generate a single set with remotely as small average as the actual r-proteins, i.e., the probability that the mean protein length in a randomly selected sample of 56 proteins is equal or smaller than 132 amino acids is too low to be determined by our brute force Monte Carlo simulations (Extended Data Figure 2, left). However, a bound on the probability can be analytically derived by utilizing the Chernoff bound from probability theory. This also allows for a systematic, multi-organism examination.

The Chernoff bound is attained by applying Markov's inequality to e^{-tX}, where X is a non-negative random variable and $t > 0$. Letting X denote the average length of a protein in a random sample $\{X_1, ..X_n\}$ of size n we have

$$X = \frac{1}{n}\sum_{i=1}^{n} X_i \,, \tag{44}$$

the Markov's inequality asserts that

$$Pr(X \leq a) = Pr(e^{-tX} \geq e^{-ta}) \leq \frac{E[e^{-tX}]}{e^{-ta}} = e^{ta}E\Big[\prod_{i=1}^{n} e^{-\frac{tX_i}{n}}\Big] \,. \tag{45}$$

Sampling such that $\{X_1, ..X_n\}$ are independent and identically distributed, this formula further reduces to

$$Pr(X \leq a) \leq e^{ta}\prod_{i=1}^{n} E\Big[e^{-\frac{tX_i}{n}}\Big] = e^{ta}\Big(E\Big[e^{-\frac{tX_1}{n}}\Big]\Big)^n \,. \tag{46}$$

Finally, noting that the above equation holds for every $t > 0$ we have

$$Pr(X \leq a) \leq \min_{t>0}\Big\{e^{ta}\Big(E\Big[e^{-\frac{tX_1}{n}}\Big]\Big)^n\Big\} \,. \tag{47}$$

Given *E. coli's* genome the expectation value on the right hand side of Eq. (47) can be numerically computed for $n = 56$, $a = 132$, and various values of t. Minimizing over $t > 0$ we then find

$$Pr(X \leq 132) \leq 10^{-17} \,. \tag{48}$$

The ribosomal proteins in *E. coli* are thus indeed unusually small, and unlikely to arise without selection on the lengths of ribosomal proteins. Utilizing the same method to the analysis of more than a thousand different organisms we arrive at similar conclusions for all organisms considered (Extended Data Figure 2, right).

As mentioned in the main text, we also find that in addition to being unusually small, ribosomal proteins are much more similar to each other in length than proteins in the genome overall. Again the fact that there are so many r-proteins in the ribosome means that this effect is highly significant statistically. Specifically, the coefficients of variation observed for the length distributions of r-proteins in various different organisms are extremely unusual.

Intuitively it may seem that the unusually low CV_L could be a by product of the unusually low averages, since the latter means that proteins are largely confined to the left tail of the distribution, i.e., it could have been the case that r-proteins are similar to one another in length because they are small on average. However, since the coefficient of variation is defined as the ratio between the standard deviation and mean, random samples in which the average protein length has dropped two or three fold in comparison to the overall genomic mean would need to show an even sharper drop in their standard deviation to start and display some reduction in the coefficient of variation—let alone explain the CVs observed for ribosomes. This could in principle be tested by generating many random sets of the expected average length (allowing some small deviations from the exact average length observed), and analyzing the corresponding width of the length distribution in those sets. However, because the low averages are so exceedingly rare to begin with, this is not feasible. We therefore used two separate approaches.

To generate random complexes with $n \leq 20$, we used a simple brute force sampling method. We then derive the CV_L in random samples that have approximately the correct average r-protein lengh (the exact cut-offs were not important), as a function of n. This allows us to directly estimate the probability of finding a CV_L as low or lower than the actual CV_L for the r-proteins. The actual CV_L is indeed lower than expected by random sampling for all n, and with increasing n the effect becomes more and more significant. In this way we can reach $n = 20$ where the probability of achieving such low CV_L is already exceedingly small. However, using this method to generate random complexes with average protein lengths as low as those seen in ribosomes becomes computationally changeling for higher n as the probability to observe these type of events becomes exceedingly small. And yet, as the need to sample rare events has arisen time and again in various different areas of research, different approaches have been developed to address this reoccurring problem [42], and we have utilized one such approach for our purposes here. In particular, to obtain the data used to draw curves for $n > 20$, we first use brute force sampling to generate a "seed" of size $n = 2$ proteins and an average protein length that is within the desired range (average length of an r-protein ±5 amino acids). This seed is then made into a complex with $n = 3$ proteins by adding to it a randomly chosen protein for which the average protein length of the joint complex is still within the desired range (several trials may be required). Additional randomness is then introduced by replacing all proteins in the newly generated complex with randomly drawn proteins, one by one, while keeping the average protein length in the desired range. This procedure is repeated many times. The end result is a randomly generated complex with $n = 3$ proteins and an average protein length that is within the desired range. This could then be used as a seed to iteratively obtain complexes of size $n = 4, 5, 6, ...$ in the very much the same manner described above. As it turn out, this sampling method is considerably faster than the brute force one for $n > 20$, and we have verified that for $n \leq 20$ the results generated by the two are practically equivalent. Results coming from combining these two approaches are described in Extended Data Figure 3, which shows that the observed coefficients of variation are indeed highly unusual, even when conditioning on the small average length of the r-proteins.

6 Deriving the initiation penalty for increasing n and predicting the optimum number of r-proteins

To take into account the time ribosomes are sequestered from elongation due to initiation (and similar processes), we first consider a fixed overhead time τ_{oh} for every translation event. We let $\rho_i(x, t)$ denote the r-protein mass density (coming from ribosomes) at position x and time t on mRNAs of type i, and again observe that this density obeys a simple transport equation at the elongation phase

$$\frac{\partial \rho_i(x, t)}{\partial t} = -k \frac{\partial \rho_i(x, t)}{\partial x} , \tag{49}$$

($1 \leq i \leq n$, $0 \leq x \leq L_i$), and that its evolution at the overhead phase could be effectively modeled in the same way

$$\frac{\partial \rho_i(x,t)}{\partial t} = -k\frac{\partial \rho_i(x,t)}{\partial x}\,, \tag{50}$$

($1 \leq i \leq n$, $-L_{oh} \leq x \leq 0$), where L_{oh} is a fictitious number of "overhead codons" chosen to assure $\tau_{oh} = L_{oh}/k$. Just as the description of elongation does not account for the statistical variation in the movement of the ribosome from codon to codon, modeling initiation this way ignores the (unknown) statistical variation in the time required for initiation. However, we only consider the steady state average where such variation is of little importance.

The solution to Eq. (49) is once again given by

$$\pi_i(x) = A_i e^{-x ln(2)/kT_{gen}}, \tag{51}$$

($1 \leq i \leq n$, $0 \leq x \leq L_i$), where the constants A_i are given in Eq. (24) above. For Eq. (50), we similarly find

$$\pi_i(x) = B_i e^{-x ln(2)/kT_{gen}}, \tag{52}$$

($1 \leq i \leq n$, $-L_{oh} \leq x \leq 0$), and $B_i = A_i$ for continuity at $x = 0$.

We then determine the time fraction φ ribosomes spend on the elongation of r-proteins. Denoting the totality of r-protein engaged in overhead processes $N^{oh}_{rp}(t)$, we have

$$\varphi = \frac{\sum\limits_{i=1}^{n}\int\limits_{0}^{L_i} \rho_i(x,t)dx}{N^{total}_{rp}(t) - N^{nascent}_{rp}(t) - N^{oh}_{rp}(t)}\,. \tag{53}$$

Dividing by $N^{total}_{rp}(t)$ throughout, considering the steady state, calculating the integrals, and remembering that this calculation neglects additional sources for ribosome idleness, we find

$$\varphi \geq \frac{\sum\limits_{i=1}^{n}\int\limits_{0}^{L_i} \pi_i(x)dx}{1 - \frac{1}{N}\sum\limits_{i=1}^{n}\int\limits_{0}^{L_i} x\pi_i(x)dx - \sum\limits_{i=1}^{n}\int\limits_{-L_{oh}}^{0} \pi_i(x)dx} = \frac{\sum\limits_{i=1}^{n}\left[e^{\frac{\tau_i ln(2)}{T_{gen}}} - 1\right]}{1 - \left[e^{\frac{\tau_{oh} ln(2)}{T_{gen}}} - 1\right]\sum\limits_{i=1}^{n}\left[e^{\frac{\tau_i ln(2)}{T_{gen}}}\right]}\,. \tag{54}$$

A comparison to Eq. (28) makes it clear that idleness coming from overhead processes such as initiation results in higher demand for ribosome involvement in r-protein production.

6.1 Derivation of Eq. (2) in the paper

To further derive Eq. (2) in the paper we note that in the relevant range of parameters $T_{gen} \gg \tau_i \gg \tau_{oh}$. Thus, by expanding to leading orders, we find that Eq. (54) could be written to an excellent approximation as

$$\varphi \geq \frac{\tau ln(2)}{T_{gen}}\left[1 + \frac{\tau ln(2)}{T_{gen}}\left(\frac{1 + CV_L^2}{2n} + \frac{n\tau_{oh}}{\tau}\right)\right], \tag{55}$$

where we have again used the fact that CV_L^2, the coefficient of variation of the set $\{L_1, .., L_n\}$, is also the coefficient of variation of the set $\{\tau_1, .., \tau_n\}$ as these are related by $\tau_i = L_i/k$. Minimizing the expression on the right hand side of Eq. (55) with respect to the number of r-proteins n we find

$$n^* = \sqrt{\frac{1 + CV_L^2}{2}\frac{\tau}{\tau_{oh}}}\ , \tag{56}$$

for the optimal number of proteins in the ribosome. In particular, note that n^* does not depend on the generation time T_{gen} or on the fraction φ, and is hence robust with respect to changes in growth conditions. Finally, we note that accounting for cases where during the overhead processes only one subunit of the ribosome is bound, and hence idle, can be done by introducing a factor $\alpha \leq 1$ to multiply τ_{oh} in Eq. (56). In *E. coli* the large subunit of the ribosome does not participate in certain stages of the regular initiation process but is thought to participate in the "scanning" mode of initiation that seems to be prevalent in r-protein translation and in translation from operons in general. Since this subunit contains about 60% of the total amount of protein in the ribosome this suggests a range of $0.4 \leq \alpha \leq 1$, thus leading to some uncertainty in the determination of n_{opt}. We also note that this is a generous range since even without scanning initiation there is no direct evidence that cells exploit the fact that only the small sub-unit is bound for the initial part of the initiation process.

7 The effect of noisy production of r-proteins

Even if production rates for all r-proteins were perfectly matched on average, chance would inevitably create more of some r-proteins and less of others. The assembly of complete ribosomes is then limited by the r-protein that by chance is present in the lowest number, creating an idle surplus pool of all other r-proteins (Extended Data Figure 1A). To quantify the severity of this effect, we assume a commonly observed statistical distribution of gene expression for each individual r-protein, and consider a worst case scenario in which there is no coordination between the expression of different r-proteins. More concretely, we consider a case where the abundance of each r-protein follows a negative binomial – a right-skewed discrete distribution that has been both frequently predicted and observed in single cells for several organisms. The negative binomial distribution can be parametrized in terms of its mean, μ, and variance, σ^2, to give the probability

$$Pr(X_i = k) = \binom{k + \mu^2/(\sigma^2 - \mu) - 1}{k}\left(\frac{\mu}{\sigma^2}\right)^{\mu^2/(\sigma^2-\mu)}\left(\frac{\sigma^2 - \mu}{\sigma^2}\right)^k\ , \tag{57}$$

that the number of proteins, X_i, of any given r-protein of type $i = \{1, ..., n\}$ is $k = \{0, 1, 2, ...\}$. We are then interested in the r-protein found in least numbers

$$X_{min} = min\{X_1, ..., X_n\}\ , \tag{58}$$

as only X_{min} complete ribosomes can be assembled. We further consider the average "noisy" surplus

$$\langle\Delta\rangle = \langle X_i - X_{min}\rangle = \langle X_i\rangle - \langle X_{min}\rangle\ , \tag{59}$$

noting that the average size of the r-protein surplus pool (measured in amino acids) is given by $N\langle\Delta\rangle$, where N again is the total number of amino acids in the ribosome. Sampling from the negative binomial distribution for various values of its parameters and various values of the number of r-proteins, n, within the relevant range, we observe that for $n = 1$, $\langle\Delta\rangle = 0$ by definition, and that for $1 < n < 200$ (Extended Data Figure 1B, inset)

$$\frac{\langle\Delta\rangle}{\sigma} \simeq a + b\sqrt{ln(n)}\ , \tag{60}$$

with $a \simeq -3/4$ and $b \simeq 3/2$, which in turn means that the idle fraction of r-protein mass due to noise obeys (Extended Data Figure 1B, main)

$$\frac{\langle\Delta\rangle}{\mu} \simeq \frac{\sigma}{\mu}\left(a + b\sqrt{ln(n)}\right)\ . \tag{61}$$

Thus, the *change* in the idle fraction for a change in n is set by the square root of the logarithm of n. One intuition for this damped response is that with each independent draw from a distribution it is less likely that the next draw will be lower than any of the preceding draws. If it does happen to be lower, it is likely not by much, because the probability in the tails of many distributions decrease so sharply when moving far from the average. For example, a classical result from extreme value theory asserts that when $\{X_1, ...X_n\}$ are independent, and identically distributed, Gaussian random variables with mean μ and variance σ^2

$$\langle X_{min}\rangle = \mu - \sigma\sqrt{2ln(n)} + o(1)\ , \tag{62}$$

which gives the same type of behaviour for $\langle\Delta\rangle$ as the negative binomial.

To double-check that similar principles hold for more complete kinetic models of stochastic gene expression (again in the worst-case scenario of no coordinated production), we further simulated the perhaps most commonly used model of stochastic gene expresion for each of the r-proteins, considering a Poisson process for production of mRNA, a Poisson of process for the production of proteins for each mRNA, exponentially distributed lifetimes of mRNAs and proteins, and explicitly accounting for growth and division. This created very similar $\sim\sqrt{ln(n)}$ scaling for $\langle\Delta\rangle$, as expected because those models produce distributions that are very close to negative binomials. This has two interesting consequences for the combined protein mass in the form of surplus pools and nascent peptide for r-proteins. First, the optimal number of r-proteins that minimizes the total idle fraction, n_{opt}, can be very high even with passive noise control without coordination between genes, particularly in eukaryotes where abundances are higher and spontaneous noise may be smaller compared to the average. Second, the total idle fraction is almost constant around and above n_{opt} (Extended Data Figure 1C), meaning that any higher n is almost optimal as well. This means that noise does not necessarily limit the number of r-proteins n, despite the fact that we approximated incomplete ribosomes as useless, when in reality some r-proteins are not essential. Gene expression is also subject to 'extrinsic' noise, but any differences that are shared by the r-proteins would not create wasteful differences between them. Though such effects can make gene expression in general seem very noisy, it would thus have very little relevance for this problem.

Though noise is not a problem, it may still seem like it would be hard for cells to ensure the same *average* expression from all genes. However, a similar argument applies to that problem:

if the average expression rates followed some distribution, this should increase the wasted surplus pools of r-proteins, but for many types distributions, the problem of unmatched averages should only become marginally worse with increasing n, particularly at high n. In addition, as described in the main text, *E. coli* uses transcriptional coupling, translational coupling, and negative feedback control to ensure that they shut off production of any r-proteins that accumulate in free form, whether due to noise or different average expression rates. Finally, this system is also under extreme selective pressure to maximize efficiency, as opposed to many synthetically modified genes studied in the field of stochastic gene expression.

References

[36] See BNID 110218: `http://bionumbers.hms.harvard.edu//bionumber.aspx?id=110218&ver=7`

[37] Sin, C., Chiarugi, D. and Valleriani, A., Quantitative assessment of ribosome drop-off in E. coli. *Nucleic acids research*, **44**(6), pp.2528-2537, (2016).

[38] Klumpp, S. and Hwa, T. Traffic patrol in the transcription of ribosomal RNA. *RNA biology*, **6**(4), pp.392-394 (2009).

[39] Epshtein, V. and Nudler, E. Cooperation between RNA polymerase molecules in transcription elongation. *Science*, **300**(5620), pp.801-805, (2003).

[40] Proshkin, S., Rahmouni, A.R., Mironov, A. and Nudler, E. Cooperation between translating ribosomes and RNA polymerase in transcription elongation. *Science*, **328**(5977), pp.504-508, (2010).

[41] Nagarjuna, N. *et al.* System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Mol. Cell. Proteomics* **11**, M111-013722 (2012).

[42] Morio, J., Balesdent, M., Jacquemart, D. and Verge, C. A survey of rare event simulation methods for static input-output models. Simulation Modelling Practice and Theory, **49**, 287-304 (2014).