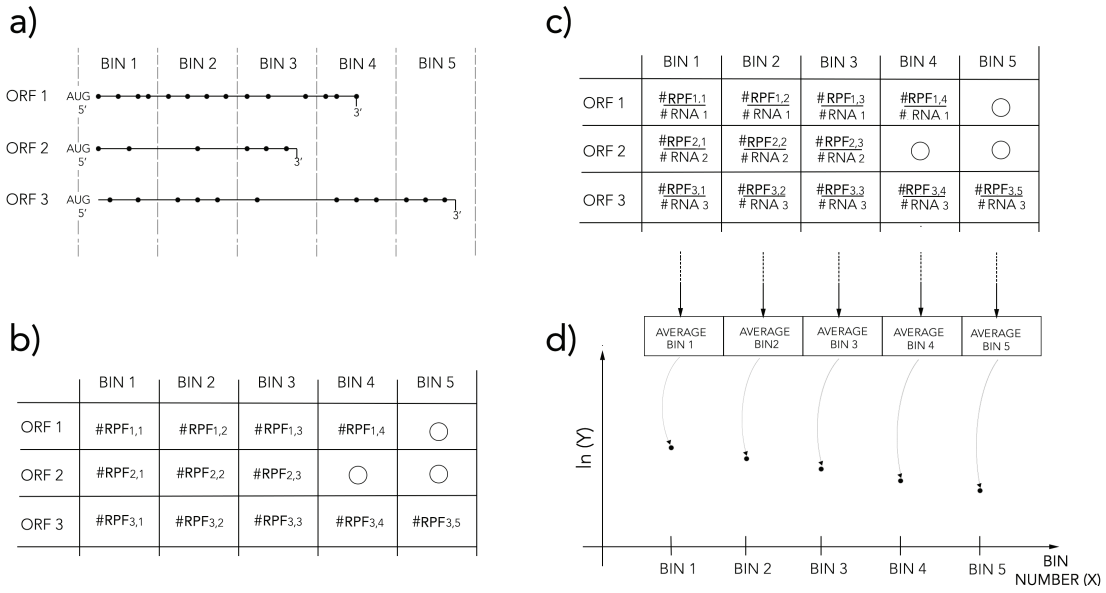
Supplementary materials

Quantitative assessment of ribosome drop-off in *E. coli*

Celine Sin, Davide Chiarugi, Angelo Valleriani

1 Downstream Analysis



Supplementary Figure 1: Illustration of the core steps of our analysis method to measure ribosome drop-off rates in Ribo-seq data. **a)** The ORFs are divided in bins of equal length; **b)** For each bin, the total number of RPFs mapping on it is reported in the RPF matrix; **c)** The elements in each line of the RPF matrix are divided by the average amount of RNA-seq reads mapping in each bin of the corresponding ORF, thus obtaining the NRPF matrix; **d)** The logarithm of the average value of each column of the NRPF matrix, $\ln(Y)$, is plotted against the bin number X.

1.1 Bootstrap approach

To obtain an accurate estimate of R_{BS} and of its associated error without making any assumptions on the distribution of the number of RPFs, we implemented a bootstrap

approach:

(i) We counted the number of elements E_X in each column of the NRPF matrix, *i.e.* the number of elements contributing to each bin (equal to the number of ORFs with at least X bins). As suggested by Figure 1 of the main text, the number E_X is not constant for all the columns, due to heterogeneous ORF length of the genes in the study.

(ii) From each column X of the NRPF matrix, we sampled a combination of E_X elements randomly with replacement, thus obtaining a matrix (call it BootStrap - BS - Matrix) that has the same dimensions of the NRPF matrix and contains the sampled elements in each column;

(iii) We computed the average for each column of the BS matrix, obtaining the vector Y_i

(iv) Given that the exponential relationship:

$$Y = A\,\mathrm{e}^{-RX} \tag{1}$$

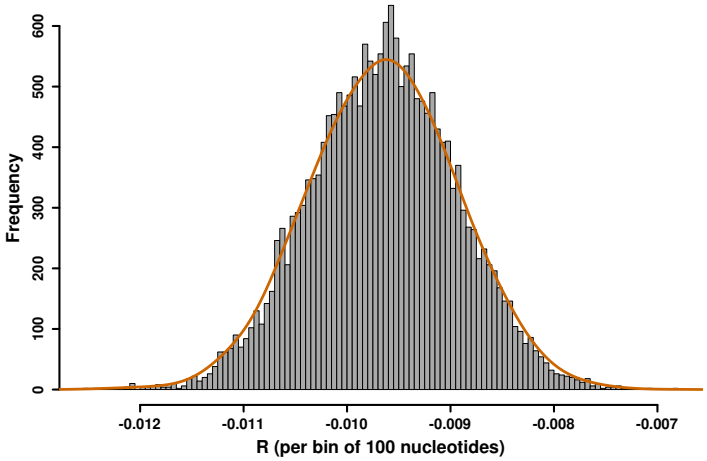
holds also for Y_i and X, we computed the rate R_i as the slope of the weighted linear regression of $\ln(Y_i)$ against X.

For each of the studied datasets (listed in Table 1 of the main text), we repeated the sequence of steps described above 10^5 times, thus obtaining 10^5 values for R_i $(R_1, \ldots, R_{10^5})$ for each dataset. Each distribution of the obtained values for R_i seems to follow a Gaussian distribution (Pearson's χ^2 test with P-value < 0.01); therefore, we assumed that the average value of each R_i distribution is representative of R_{BS} for the corresponding dataset. Supplementary Figure 2 provides an example of an R_i distribution.

1.2 Evaluation of the error

Here we review the technical details of how we estimate the errors for R. Due to the heterogeneous distribution of *E. coli* ORF lengths, the number of elements contribution to each bin is not constant through the bins; in particular, in both the NRPF and BS matrices, E_X becomes progressively small as X increases (see Figure 1 of the main text). Thus, the variance associated to the average of the E_X elements becomes progressively large and the bin average becomes a bad estimator of the correspondent element of the vector Y. On one hand, these arguments motivate us to compute the linear regression of $\ln(Y)$ vs. X by weighting each average by its associated variance (weighted linear regression). On the other hand, from the same arguments, it turns out that the reliability of our estimate of R depends both on the number of bins (*i.e.* the length of the Y vector) we consider for the linear regression and on the bin size.

To select the optimal bin size and the optimal Y length needed to obtain the best estimate of R and the associated error, we analyzed a set of simulated data for each of the 17 databases we considered.



Supplementary Figure 2: Sample of a R_i distribution obtained from 10^5 iterations of the bootstrapping process. The superimposed curve represents the best fitting Gaussian function. The data used for this plot come from the analysis of dataset 17 (see Table 1 of the main text).

In each case, we generated a simulated dataset by redistributing the reads of each ORF according to an exponential distribution with parameter equal to a preselected nominal value. This value was chosen equal to the rate R that, for each dataset, can be obtained by applying the bootstrap approach described above. Thus, in the simulations we performed, we preserved some of the features of the original datasets, namely the total number of reads per ORF and the gene lengths.

For each one of the original datasets we considered different bin sizes ranging from 10 to 130 nucleotides and, for each bin size, we generated 1000 simulated sets of ribosome positions. For each simulated data set, we estimated the value of R 47 times, taking into account a different number of bins each time. The minimum number of bins was 2 (the minimum required for linear regression), and the maximum number of bins is 49. In each case, we obtained the best estimate of R for a bin size of 100 nucleotides and a length of the Y vector of 39.

This data, combined with the results of the bootstrap procedure, allowed us to evaluate the systematic error associated to our estimate of R. From these simulations, we conclude that the correct value of R can be obtained by adding an offset Δ from the estimate R_{BS} provided by the bootstrap procedure. The standard deviation associated to the value of R (S_R) can be computed from the square root of the sum of the variance associated to the bootstrap process (S_{BS}) and the variance associated to the offset (S_Δ) by the formula:

$$S_R = \sqrt{S_{BS}^2 + S_\Delta^2} \tag{2}$$

The values of Δ and S_Δ we obtained for each dataset, are reported in Supplementary Table 1.

As described in the text, the value of R is obtained from the R_{BS} after a correction by Δ, according to the equation:

$$R = R_{BS} - \Delta \tag{3}$$

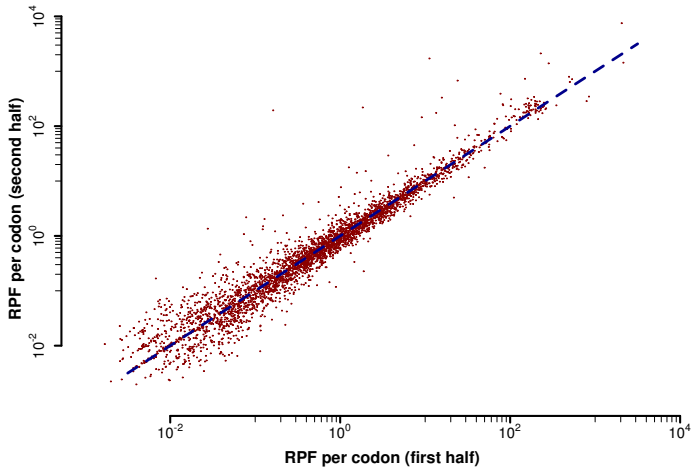
1.3 Comparison with other methods

Even though strikingly simple in principle, our binning strategy represents the Columbus' egg that allowed us to detect the signal of ribosome drop-off in Ribo-seq data. The other analytical approaches reported so far in the literature failed to reach this goal, essentially because the proposed binning strategy was not sensitive enough. As outlined in the main text, the usual way of evaluating ribosome drop-off in Ribo-seq data (for an example, see [1]) is by looking for a difference between the number of reads that map to two subsequent halves of each ORF. A significant difference between the two halves (fewer reads in the second half) reveals that a certain number of ribosomes has not successfully completed translation. The results of this analysis are typically illustrated through scatterplots similar to the one reported in Supplementary Figure 3, where the number of reads (expressed in terms of ribosome density, i.e. number of reads per nucleotide or per codon) mapping in the first half is plotted against the number of reads mapping in the second half. If there is no significant difference between the quantities reported in the two axes, the plotted points will cluster around a straight line having the slope equal to 1, as it happens in the case reported in Supplementary Figure 3. If fewer reads map in the second half with respect to the first half, the point corresponding to that ORF will plot below the line mentioned above.

While this method is mathematically sound, it has a major drawback – the sensitivity of this approach depends critically on the ORF length. When the frequency of drop-off events is not large enough with respect to length of the ORFs, the difference in ribosome density between the two halves of the ORF are too small to be detected by eye and cannot influence the correlation coefficient in a log-log scatterplot. Supplementary Figure 4 provides an illustration of this phenomenon.

As a consequence, if the genome of interest prevalently contains short genes, the method may not be sensitive enough to detect the drop-off in the shorter genes. Then, the conclusion would necessarily be that, at the genome scale, the ribosome drop-off rate is not measurable. It turns out that the sensitivity of this method is too low to measure ribosome drop off on a global level if the drop off is occurring in a biologically viable cell; we discuss the specifics of this argument later.

Conversely, our analysis is not affected by the length of the ORFs. We can illustrate this scenario at the genome-wide level: Supplementary Figure 5 reports the results of a simulation where we spatially redistributed the RPFs associated to each ORF in the dataset of Ref. [1]. The RPFs are distributed according to an exponential distribution with parameter r equal to 1.40×10^{-4}, which corresponds to the drop-off rate per codon that we estimated for this dataset. In this way, we generated an artificial dataset very

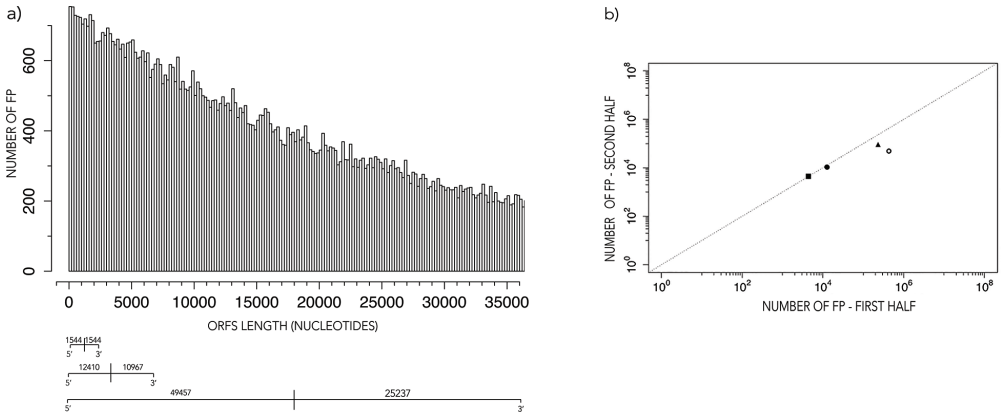


Supplementary Figure 3: Typical scatterplot obtained from the drop-off analysis method proposed in [1]. The ribosome density (number of RPFs per codon) of the first half of each gene is plotted against the ribosome density of the second half. The clustering of the plotted points along the dashed line indicates that the ribosome drop-off rate is negligible. Data taken from [1].

similar to the one from [1] except we have tailored the distribution of the RPFs on the ORFs to explicitly mimic the presence of ribosome drop-off.

We then used this dataset as a benchmark to test the capabilities of our method and the one proposed in [1] to detect the drop-off. As shown in Supplementary Figure 5a, the scatterplot reporting on the differences of the number of RPFs in the two halves of each ORFs shows a clustering around the straight line with slope equal to 1, which would deceptively suggest that no drop-off events occurred. The plot resulting from our analysis (Supplementary Figure 5b) is correctly consistent with the presence of drop-off, even when it occurs at a low rate. Interestingly, if we repeat the simulation described above in the case of an hypothetical genome with genes whose length is markedly longer than those in *E. coli*, the method proposed in [1] successfully detects the signals of drop off.

Thus, it turns out that the sensitivity of the method proposed in [1] suffices only when the average gene length goes beyond a biologically reasonable threshold. Indeed it is important to notice that for a given drop-off rate, the translation process remains reliable only if the gene lengths remain bounded within certain limits, roughly by $1/r$. If we assume a drop-off rate (or, analogously, a drop-off probability) which is constant along the whole length of the various mRNAs, the distribution of the RPFs density will decrease along the messengers according to an exponential distribution. In this case, the probability P_{S} that a ribosome will reach the stop codon located L codons away from

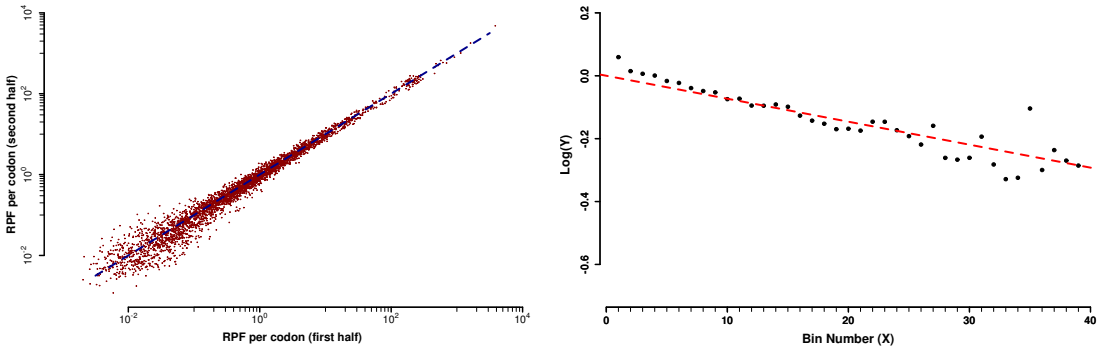


Supplementary Figure 4: Simulation for illustrating how a relatively low drop-off rate is not detected by the analysis method proposed in [1]. **(a)**: histogram describing the simulated decrease in the number of RPFs mapping on the ORFs. This decrease is generated as an exponential decay with rate 4×10^{-4} per codon, corresponding to the *E. coli* drop-off rate estimated in [2]. The drawing below the histogram depicts three ORFs of different lengths: 1000 nucleotides (close to the average ORF length in E. *E. coli*), 7077 nucleotides (the maximum gene length in *E. coli*) and 3500 nucleotides, a length about 5 times longer than the maximum ORF length in *E. coli*. The numbers above the three depicted ORFs report the number of RPFs mapping on the two halves according to the distribution above. **(b)**: Scatterplot obtained plotting the the number of RPFs mapping on the two halves of the sample ORFs depicted in Figure (a); square: ORF length = 1000; solid circle: ORF length = 7077; triangle ORF length = 35000; empty circle: ORF length = 100000 (not depicted in (a)). Note that a significant deviation from the dashed line is obtained only when non-biologically possible ORFs lengths are considered .

the start codon (technically the survival probability) is:

$$P_{\mathrm{S}} \;=\; (1-r)^{L} \sim \exp(-rL)\,, \tag{4}$$

where r is the drop-off rate per codon, with $r \ll 1$. Thus, the probability for a ribosome to successfully complete the translation process sharply decreases with the gene length. If we set r at the value of $4 \cdot 10^{-4}$ per codon, as estimated in [2], this probability falls under 50% for genes longer than 1700 codons, meaning that, on average one ribosome over two will drop-off the mRNA. In other words, the magnitude of the drop-off rate represents an important constraint for the possible genes lengths in living organisms, because exceptionally long genes would never be reliably translated. This scenario is consistent with the experimental results presented in [3], where the translation efficiency was observed to markedly decrease when progressively longer β-galactosidase gene constructs were assayed to detect ribosome drop-off.



(a) (b)

Supplementary Figure 5: Results of a simulation in which the distribution of the RPFs on the *E. coli* ORFs was artificially set with a drop-off rate of 1.40×10^{-4} per codon. **(a)** using the method proposed in [1] the drop-off is not detected (the plotted points cluster along the dashed line); **(b)** our method allows the measurement of the drop-off rate corresponding to the slope of the dashed line, obtained through the approach described in the text. The ORFs length is measured in number of bins of 100 nucleotides. The plot includes only the first 39 bins that we considered in our analysis. To facilitate the comparison with the similar graphs present in the paper we shifted the plot so that the y-intercepts of all plots will match.

2 Statistical tests

2.1 Computing the Confidence Interval

The 99% confidence interval (CI) was computed through the equation:

$$\mathrm{CI} = r \pm \mathrm{Z}_{0.05} \cdot S_r \tag{5}$$

where S_r is the standard deviation associated to the estimation of r and $\mathrm{Z}_{0.05}$ is the Z-score corresponding to $\frac{1-0.999}{2}$. The values we obtained for the confidence intervals provide us with two important clues about the accuracy of our estimate and the features of r. First, the relatively small value of $\mathrm{Z}_{0.05} \times S_r$, often referred to as the margin of error, indicates that our approach yields accurate estimate of r. Moreover, the definition of CI tells us that the "true value" of r (i.e. the average of the population of all possible r) lies between the boundaries of the CI with a probability of 99%; there is only a 1% probability that the values outside of the CI are a reliable estimate of r. The 4^{th} column of Table 2 of the main text shows that in all the cases but one (dataset 16) the values close to 0 are not in the range of the CI. This suggests that, in these cases, the drop-off rate r is significantly different from 0.

2.2 Z-test for the mean

To check whether the values we obtained for the drop-off rate per codon (r) were significantly different from 0, we performed a series of Z-tests for the mean. In particular, we evaluated the probability that r belongs to the normal distribution H_0 having the average $\mu_0 = 0$ and the same standard deviation S_r associated to r. In other words, we verified whether we can reject the null hypothesis ($\mathrm{H}_0 : r = 0$), which would indicate that the alternative hypothesis ($\mathrm{H}_1 : r \neq 0$) is more likely to be true. To do this, we computed the Z-score for each r hypothesizing that it belongs to H_0 ($\mathrm{Z}_{r|\mathrm{H}_0}$) through the equation:

$$\mathrm{Z}_{r|\mathrm{H}_0} = \frac{r - 0}{S_r} \tag{6}$$

and we compared it to $Z_{0.01}$ which is the Z-score corresponding to the significance level 0.01 (probability of type 1 error or false positive rate). To check whether the significance level we chosen was meaningful for our purposes, we evaluated the type 2 error (false negative rate) and the power of the test, considering the alternative hypothesis in which r belongs to a normal distribution with average $\mu_a = r$ and S_r. The detailed results of these tests are reported in Table 2.

2.3 The ANOVA test

To detect any possible significant difference between the values of r we measured, we considered all the values of r significantly different from 0 and, through the ANOVA test, we checked the null hypothesis that all of them are approximately equal, against the alternative hypothesis that there are at least two values of r that are significantly different. For the ANOVA test, we considered 13 groups (one for each estimated r significantly different from 0) each composed by 10^5 elements, i.e. , the number of elements composing each r_i distribution. Thus, the degrees of freedom "between" turned out to be 12 while the degrees of freedom "within" resulted to be 12999987. We set the significance level to 0.001 and the test indicates that we should reject the null hypothesis.

3 Drop-off rate and growth medium

According to [4], the kinetic properties of ribosomes are influenced by the growth medium. In particular, cells cultured in the same media should be characterized by ribosomes with very similar features. To check whether this holds for ribosome drop-off rate, we used a two-tailed Z-test to compare the values of r we obtained from samples coming from different series (*i.e.* from different laboratories) referring to experiments in which the bacteria were grown in the same medium in non-stressed conditions (control cultures). For the Z tests, we choose a significance level of 0.005. Table 3 reports the results of the comparisons between samples referring to the rich medium (MOPS) culturing conditions.

By comparing columns 2, 3 and 4, in two of three cases the Z-score falls into the rejection area (*i.e.* the Z-score is out of the boundaries delimited by the Bonferroni-

corrected significance levels). Hence, in spite of the fact that the cell cultures are grown in the same medium, the basal drop-off rate turns out to be different sometimes, implying that the experimental variability or some differences in the experimental protocols may have affected the value of the ribosome drop-off rate, at least in some cases.

In table 4 we report the results of the comparison between samples referring to bacteria grown in the minimal medium.

In this case the Z-score is consistent with the hypothesis of significantly equal drop-off rates.

Summing up, the outcomes of our tests do not provide us a clear response about possible correlations between the ribosome drop-off rate and the growth medium. Nevertheless, given the variability we observed, our analysis reveals an important information: even though cells are cultured on the same medium, the data coming from different laboratories might unpredictably be significantly different in terms of ribosome drop-off rates, possibly due to differences in the experimental protocols.

4 Complete plots

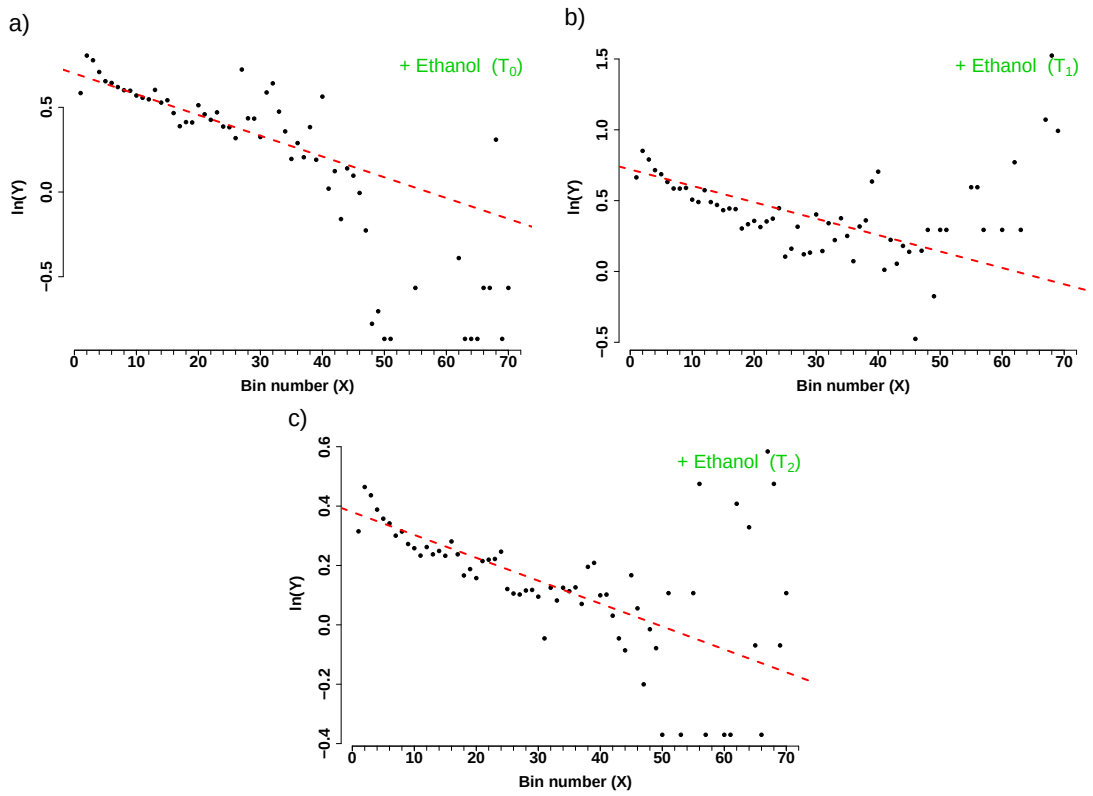
In this Section we report all the plots referring to the databases we analyzed. For the sake of readability, the plots reported in the main text are cut at the 39th bin and shifted vertically by a value corresponding to the intercept of the fitting line. These modifications are not present in the plots reported hereafter.

4.1 Datasets 9, 11 and 13 : Ethanol-induced stress

4.2 Datasets 5, 6, 7, 8: Amino acids starvation

4.3 Datasets 15 and 16: a novel σE -induced sRNA

4.4 Datasets 1, 2, 3 and 4: Heat and Osmotic Stress.



Supplementary Figure 6: Plot of the vector Y vs. the number of bins (X). The slopes of the dashed lines correspond to the drop-off rate r reported in Table 1 of the main text. **a)**: Dataset 9 - Control (T_0). **b)**: Dataset 11 - T_1, after 10' of ethanol stress. **c)**: Dataset 13 - T_2, after 70' of ethanol stress.

Supplementary Table 1: Parameters R_{BS}, S_{BS}, Δ and S_Δ used for the evaluation of the error in the estimation of the drop-off rate. Column 1: Dataset ID. Column 2: the value R_{BS} estimated through the bootstrap approach. Column 3: the standard deviation S_{BS} associated to R_{BS}. Columns 4 and 5: results of the simulations performed to evaluate the error Δ and the associated standard deviation S_Δ. Column 4: Offset Δ. Column 5: Standard deviation S_Δ. Column 6: the drop-off rate R per bin of 100 nucleotides. Columns 7: the standard deviation associated to R.

Dataset ID	R_{BS}	S_{BS}	Δ	S_Δ	R	S_R
	(10^{-4})	(10^{-4})	(10^{-4})	(10^{-4})	(10^{-4})	(10^{-4})
1	112.9	8.7	16.2	3.98	96.7	9.6
2	88.5	11.7	16.2	3.84	72.3	12.3
3	96.3	7.3	16.7	3.78	79.6	8.2
4	80.2	10.0	16.1	3.61	64.1	10.6
5	39.6	8.0	16.7	3.80	22.9	8.9
6	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
7	79.2	6.9	16.0	3.64	63.2	7.8
8	*n.a.*	*n.a.*	*n.a.*	*n.a*	*n.a.*	*n.a.*
9	97.6	17.0	16.7	3.81	80.9	17.4
10	89.1	22.1	16.7	3.80	72.4	22.4
11	184.8	12.1	16.9	3.65	167.9	12.6
12	201.5	10.0	16.9	3.89	184.6	10.7
13	94.2	7.8	16.2	3.76	78.0	8.7
14	91.3	8.0	16.5	3.63	74.8	8.8
15	116.7	9.2	16.8	3.70	99.9	9.9
16	0.00	12.7	16.2	3.72	0	13.2
17	63.0	7.0	16.5	3.75	46.5	7.9

Supplementary Table 2: Results of the (right tail) Z-tests to assess whether the drop-off rates are significantly different from 0. Columns 1 and 2: GEO coordinates of the datasets (Series, Ribo-seq sample, RNA sample). Column 3: Drop-off rate per codon. Column 4: percentile of the standard normal distribution corresponding to a rejection area (right) of 0.01. Column 5: Z-score associated to the comparison between the distribution with average r and standard deviation S_r versus the null distribution with average 0 and the same standard deviation S_r. Column 6: power of the corresponding Z-test.

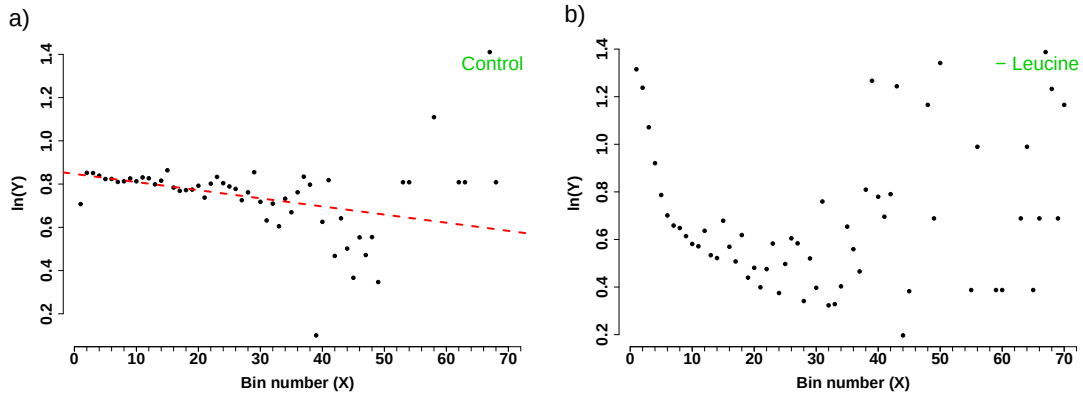
Dataset ID	**Drop-off rate**	$Z_{0.01}$	$Z_{r\|H_0}$	**Power**
(Ref. Table 1)	Per codon ($\times 10^{-4}$)			π
1	2.9	2.33	9.67	0.97
2	2.2	2.33	5.50	0.96
3	2.4	2.33	12.0	0.98
4	1.9	2.33	6.30	0.93
5	0.7	2.33	2.41	0.83
6	*n.a.*	2.33	*n.a.*	*n.a.*
7	1.9	2.33	9.50	0.90
8	*n.a.*	2.33	*n.a*	*n.a.*
9	2.4	2.33	4.80	0.96
10	2.2	2.33	3.14	0.91
11	5.1	2.33	12.8	0.98
12	5.6	2.33	18.7	0.99
13	2.3	2.33	7.67	0.98
14	2.3	2.33	7.67	0.96
15	3.0	2.33	10.2	0.94
16	0.0	2.33	1.22	0.82
17	1.4	2.33	7.00	0.98

Supplementary Table 3: Results of the Z-tests for comparing the drop-off rates of samples coming from different GEO Series (different laboratories), referring to cultures in Rich (MOPS) Medium. Column 1: Samples ID, referring to Table 1 of the main text. Column 2: Z-score computed from the comparison of the drop-off rates. Column 3: percentiles of the standard normal distribution corresponding to a total rejection area of 0.005. Column 4: percentiles of the standard normal distribution corresponding to a total rejection area of 0.005, corrected according to the Bonferroni method.

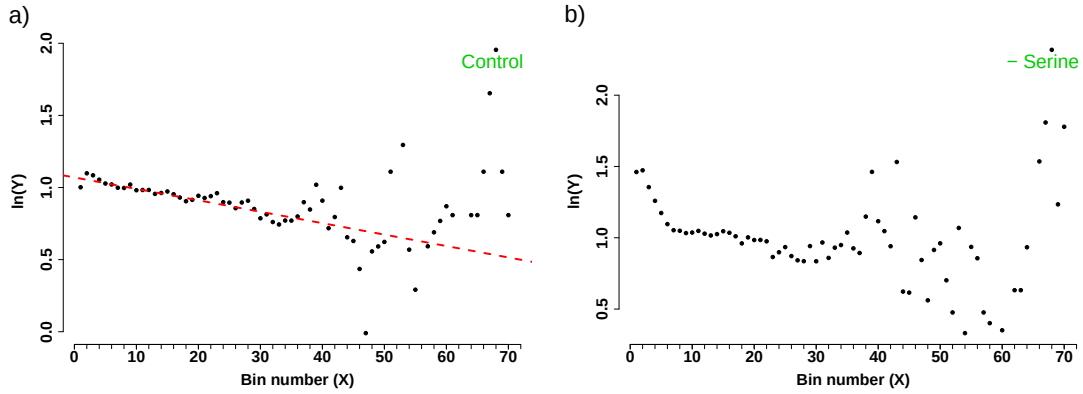
Compared samples	**Z score**	**Significance level**	Z_B
(Sample ID)		($\pm Z_{0.0025}$)	
15 vs. 17	4.24	± 2.81	± 3.14
5 vs. 17	2.00	± 2.81	± 3.14
5 vs.15	5.85	± 2.81	± 3.14

Supplementary Table 4: Results of the Z-tests for comparing the drop-off rates of samples coming from different GEO Series, referring to cultures in Minimal Medium. Column 1: Samples ID, referring to Table 1 of the main text. Column 2: Z-score computed from the comparison of the drop-off rates. Column 3: percentiles of the standard normal distribution corresponding to a total rejection area of 0.005. Column 4: percentiles of the standard normal distribution corresponding to a total rejection area of 0.005, corrected according to the Bonferroni method.

Compared samples	**Z score**	**Significance level**	Z_B
(Sample ID)		($\pm Z_{0.0025}$)	
9 vs. 3	0.05	± 2.81	*n.a.*

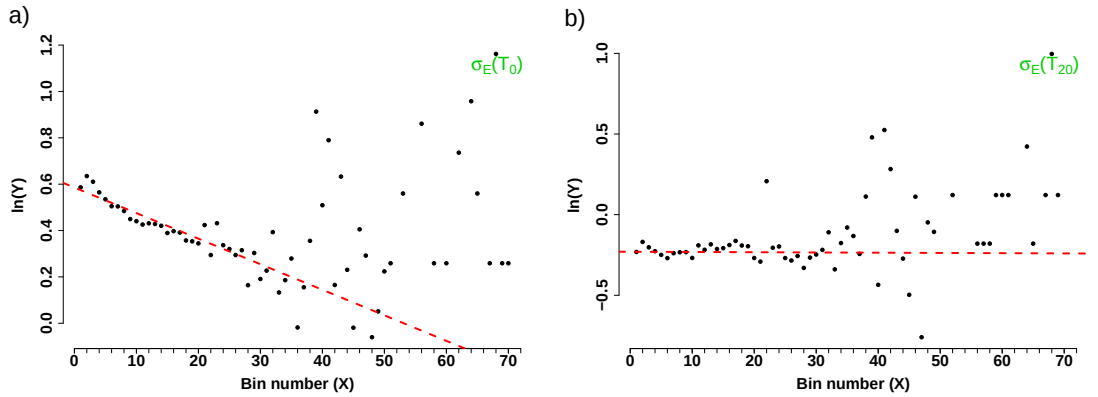


Supplementary Figure 7: Plot of the vector Y vs. the number of bins (X). The slope of the dashed line corresponds to the drop-off rate r reported in Table 1 of the main text. **a)** Dataset 5 - Control (MOPS - Rich medium) **b)** Dataset 6 - Leucine starvation. In this case, due to the poor fit with a single exponential model, we could not compute r. Thus, the regression line is not represented here.

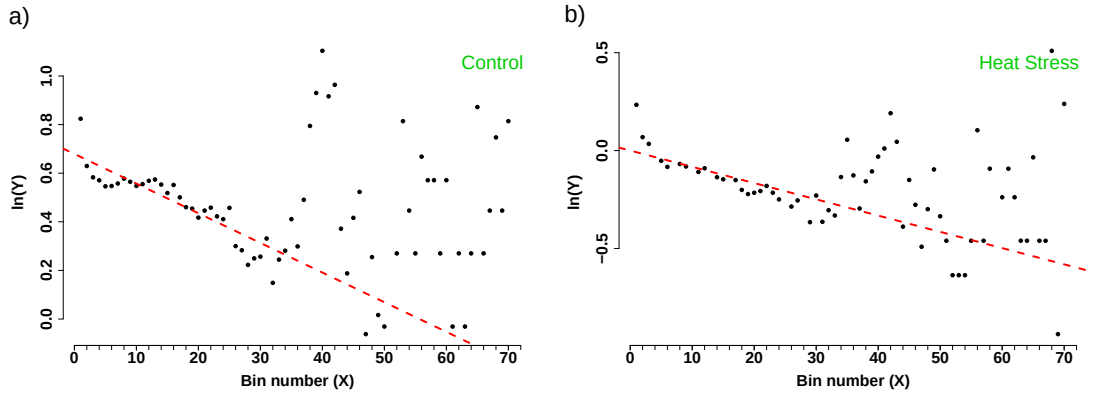


Supplementary Figure 8: Plot of the vector Y vs. the number of bins (X). The slope of the dashed line corresponds to the drop-off rate r reported in Table 1 of the main text. **a)** Dataset 7 - Control (MOPS - Rich medium) **b)** Dataset 8 - Serine starvation. In this case, due to the poor fit with a single exponential model, we could not compute r. Thus, the regression line is not represented here.

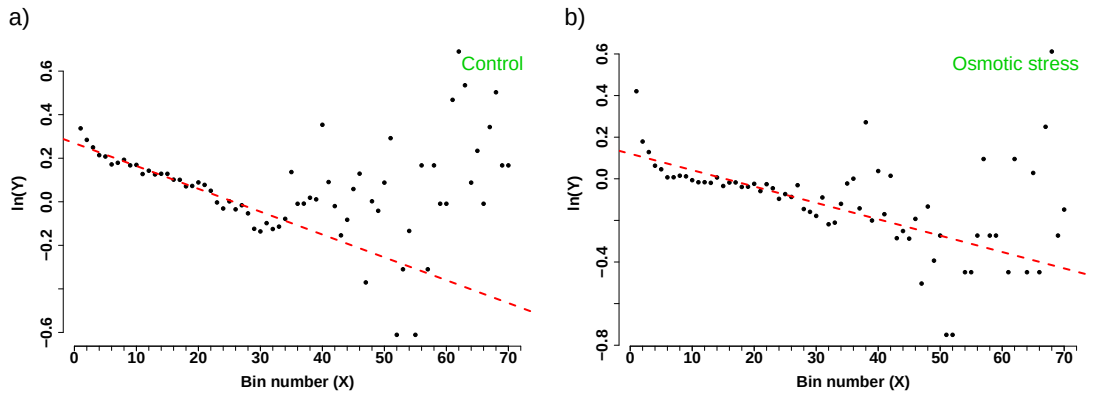
th!]



Supplementary Figure 9: Plot of the vector Y vs. the number of bins (X). The slope of the dashed lines correspond to the drop-off rates r reported in TTable 1 of the main text. **a)**: Dataset 15 - Control (T_0). **b)**: Dataset 16 - T_1, after 20 minutes of σ^E over expression induction.



Supplementary Figure 10: Plot of the vector Y vs. the number of bins (X). The slope of the dashed line corresponds to the drop-off rate r reported in Table 1 of the main text. **a)** Dataset 1 - Control (MOPS - Rich medium) **b)** Dataset 2 - Acute heat Stress ($47°C$ for 7')



Supplementary Figure 11: Plot of the vector Y vs. the number of bins (X). The slope of the dashed line corresponds to the drop-off rate r reported in Table 1 of the main text. **a)** Dataset 3 - Control (Minimal medium) **b)** Dataset 4 - Osmotic Stress (NaCl 0.3M for 20' at 37°C).

References

[1] Li, G.W., Burkhardt, D., Gross, C., and Weissman, J. S. (2014) Quantifying absolute protein synthsis rates reveals principles underlying allocation of cellular resources. *Cell*, **157**, 624–634.

[2] Menninger, J.R. (1976) Peptidyl transfer RNA dissociates during protein synthesis from ribosomes of *Escherichia coli*. *J. Biol. Chem.*, **251**, 3392–3398.

[3] Manley, J.L.. (1978). Synthesis and degradation of termination and premature-termination fragments of beta-galactosidase in vitro and in vivo. *J. Mol. Biol.*, **125**, 407–432.

[4] Kurland, C.G. and Mikkola, R. (1993). The impact of nutritional state on the microevolution of ribosomes. In Starvation in Bacteria. In Kjelleberg, S. (eds), *Starvation in Bacteria*, Plenum Press, New York, pp. 225–238.