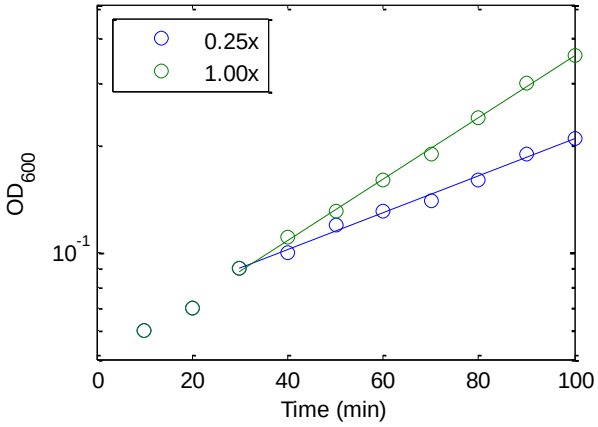
Supplement to "Dissecting the stochastic transcription initiation process in live *Escherichia coli*"

Jason Lloyd-Price, Sofia Startceva, Vinodh Kandavalli, Jerome G. Chandraseelan, Nadia Goncalves, Samuel M. D. Oliveira, Antti Häkkinen and Andre S. Ribeiro

I. Growth Curves



Supplementary Figure S1: Growth curves (OD_{600}, measured with an Ultraspec 10 cell density meter) of cells in 1x and 0.25x media (circles) at 37 °C. DH5α-PRO cells were grown overnight in 1x media at 30 °C with aeration of 250 rpm, and diluted into fresh 1x media to an initial OD_{600} of 0.05. Cells were incubated at 37 °C at 250 rpm until reaching the mid-log phase (~2 h), and re-diluted into the appropriate medium to an OD_{600} of 0.05. Their OD_{600} was measured every 10 minutes thereafter. At ~30 min, the cells in 0.25x media adjusted their growth rate (before this, the measurements overlap). Thus, growth rates were measured by least-squares fits (lines) from the data from 30 min onward. The slopes of the fits correspond to doubling times of 34.4 min (1.00x) and 57.9 min (0.25x).

II. Models of transcription initiation

To evaluate the cumulative distribution function (CDF) of the distribution of time intervals between production events from the full model of transcription initiation for a given value of R, we first translate this model into an observationally equivalent model of the form in equation 3. For the full model, this translation is given in the first row of Supplementary Table S1. The translated model's CDF can be evaluated using [1]. This CDF, when there are n steps after S_0, is referred to here as $F_{\mathrm{ON/OFF}+n}$. This distribution has a mean and variance of:

$$\mu_{\mathrm{ON/OFF}+n} = \frac{\lambda_{\mathrm{OFF}}}{\lambda_1 \lambda_{\mathrm{ON}}} + \sum_{i=1}^{n} \lambda_i^{-1} \tag{S1}$$

$$\sigma^2_{\mathrm{ON/OFF}+n} = \frac{\lambda_{\mathrm{OFF}}}{\lambda_1^2 \lambda_{\mathrm{ON}}} \left(2 + \frac{2\lambda_1 + \lambda_{\mathrm{OFF}}}{\lambda_{\mathrm{ON}}} \right) + \sum_{i=1}^{n} \lambda_i^{-2} \tag{S2}$$

Assumptions	CDF	λ_{ON}	λ_{OFF}	λ_1	λ_2	λ_3
	$F_{\mathrm{ON/OFF+3}}$	k_{ON}	$\frac{-k_{\mathrm{ON}}}{(Q_0-k_{\mathrm{ON}})(Q_2-k_{\mathrm{ON}})}$	$\frac{Q_1}{k_1k_2}$	Q_1^{-1}	k_3
$k_{-1} \gg k_2$, $k_1 \gg k_{\mathrm{OFF}}$	$F_{\mathrm{ON/OFF+2}}$	k_{ON}	$\frac{k_{\mathrm{OFF}}}{RK_a+1}$	$\frac{k_2RK_a}{RK_a+1}$	k_3	
$k_2 \gg k_{-1}$	$F_{\mathrm{ON/OFF+3}}$	k_{ON}	k_{OFF}	Rk_1	k_2	k_3
$k_{\mathrm{ON}} \gg k_1$	$F_{\mathrm{Hypo(3)}}$			$\frac{u+v}{2}$	$\frac{u-v}{2}$	k_3
$k_{\mathrm{ON}} \gg k_1$, $k_{-1} \gg k_2$	$F_{\mathrm{Hypo(2)}}$			$\frac{k_2RK_a}{RK_a+1}$	k_3	
$k_{\mathrm{ON}} \gg k_1$, $k_2 \gg k_{-1}$	$F_{\mathrm{Hypo(3)}}$			Rk_1	k_2	k_3

Supplementary Table S1: Relation between kinetic parameters from equations (1) and (2) of the main manuscript with the parameters of the model from equation (3), for a given value of R. Here, $K_a = k_1k_{-1}^{-1}$, $u = Rk_1 + k_{-1} + k_2$, $v = \sqrt{(k_{-1}+k_2-Rk_1)^2 + 4Rk_1k_{-1}}$, and Q_n are the roots of $-x^3 + bx^2 - cx + d$ *, where $b = u + k_{\mathrm{ON}} + k_{\mathrm{OFF}}$, $c = uk_{\mathrm{ON}} + k_{\mathrm{OFF}}(k_{-1}+k_2) + Rk_1k_2$ $d = Rk_1k_2k_{\mathrm{ON}}$, ordered such that $\lambda_{\mathrm{OFF}} \geq 0$.

,

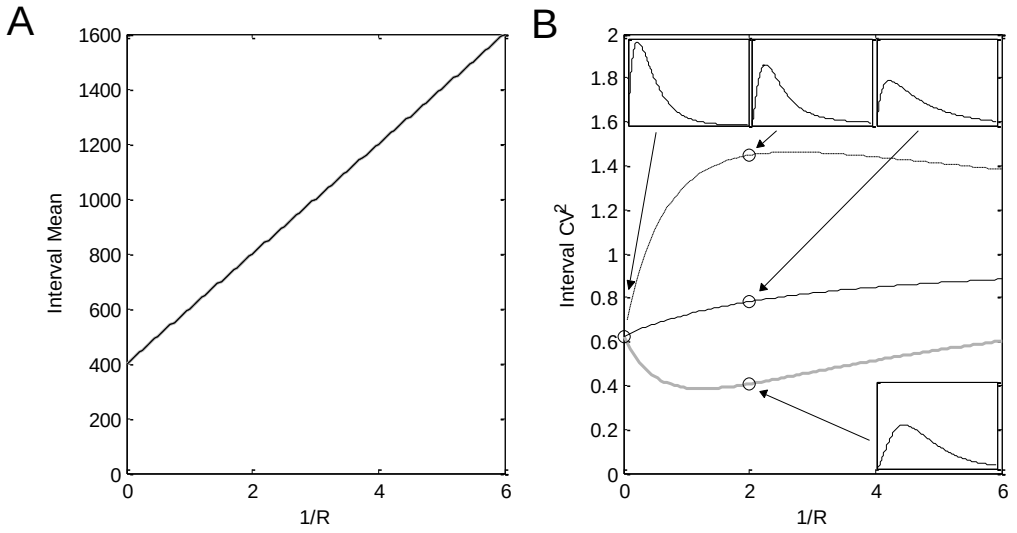
In the manuscript, several limiting cases of this model are considered. The first is that the ON/OFF mechanism is fast relative to initiation, i.e. $k_{\mathrm{ON}} \gg k_1$. In this case, the model's CDF simplifies to that of a hypoexponential distribution with three exponentials with rates λ_1, λ_2 and λ_3, which relate to the parameters of 0 as shown in the fourth row of Supplementary Table S1. The hypoexponential CDF with n exponentials is referred to here as $F_{\mathrm{Hypo}(n)}$.

Two further simplifications are considered, referred to in the manuscript as Limiting Mechanisms I and II. Both of these result in models with CDFs that are equivalent to either $F_{\mathrm{ON/OFF}+n}$ or $F_{\mathrm{Hypo}(n)}$. The parameters of the CDFs of the models derived from these three simplifying assumptions are presented in Supplementary Table S1. The final model simplification considered in the manuscript is when $k_3 = \infty$, i.e. when there is no rate-limiting third step in initiation, which removes the step parameterized by k_3 from the model.

The model of transcription initiation predicts the same linear change in the mean interval duration with $1/R$, regardless of the model simplifications (Figure S2A). However, the different simplifications result in different distributions of intervals as a function of $1/R$, which will differ in, e.g., noise (Figure S2B).

---

* Q_n can be evaluated with $Q_n = -2\sqrt{p}\cos\left[\frac{1}{3}\left(\cos^{-1}\left(\frac{-q}{2p^{3/2}}\right) - 2\pi n\right)\right] + \frac{b}{3}$, where $p = \frac{b^2-3c}{9}$ and $q = \frac{b}{3}\left(\frac{9b^2}{2} - c\right) + d$.



Supplementary Figure S2: Model prediction for (A) mean and (B) CV^2 of intervals as a function of 1/R with assumptions $k_2 \gg k_{-1}$ (dashed black line, $k_{\mathrm{ON}}^{-1} = 1000$, $k_{\mathrm{OFF}}^{-1} = 200$, $k_1^{-1} = 200 k_{\mathrm{ON}} \left(k_{\mathrm{ON}} + k_{\mathrm{OFF}}\right)^{-1}$, $k_2^{-1} = 300$, $k_3^{-1} = 100$), $k_{\mathrm{ON}} \gg k_1$, $k_{-1} \gg k_2$ (black lines, $K_a = 1.5$, $k_2^{-1} = 300$, $k_3^{-1} = 100$), and $k_{\mathrm{ON}} \gg k_1$, $k_2 \gg k_{-1}$ (grey lines, $k_1^{-1} = 200$, $k_2^{-1} = 300$, $k_3^{-1} = 100$). Note that in (A), all three lines overlap. Interval distributions for several parameter sets are shown in the insets of (B) (the axes of the insets are the same).

III. Parameter Estimation

Model parameter estimation was performed using a censored log-likelihood objective function as in [1], which accounts for uncertainty in the measurement of R, and for the uncertainty in the interval durations that arises from the limited framerate of the measurements and from the limited observation time:

$$\log L(\boldsymbol{\theta}) = \sum_m \mathbb{E} \log L_m\left(\boldsymbol{\theta}; R^{-1}\right) \tag{S3}$$

where $\mathbb{E}$ is the expectation over R^{-1}, and the conditional log-likelihood for condition m at relative RNAp concentration R is:

$$\log L_m(\boldsymbol{\theta}; R^{-1}) = \sum_i \log\left[F_{\mathcal{M}}(t_{m,i} + T_M; \boldsymbol{\theta}, R^{-1}) - F_{\mathcal{M}}(\max(0, t_{m,i} - T_M); \boldsymbol{\theta}, R^{-1})\right] \tag{S4}$$
$$+ \sum_i \log\left[1 - F_{\mathcal{M}}(c_{m,i}; \boldsymbol{\theta}, R^{-1})\right]$$

where $F_{\mathcal{M}}\left(x; \boldsymbol{\theta}, R^{-1}\right)$ is the CDF of the model being fit (either $F_{\mathrm{ON/OFF}+n}$ or $F_{\mathrm{Hypo}(n)}$) with parameters translated as appropriate using Supplementary Table S1, $\boldsymbol{\theta}$ is the parameter vector, $t_{m,i}$ are measured intervals in condition m, T_M is the time between frames, and $c_{m,i}$ are the right-censored intervals.

The expectation of $\log L_m\left(\boldsymbol{\theta}; R^{-1}\right)$ over R in equation (S3) accounts for the uncertainty in the measurement of R. This was performed with $R^{-1} \sim \mathcal{N}\left(\hat{R}_m^{-1}, \sigma^2\left(\hat{R}_m^{-1}\right)\right)$, which was approximated by

evaluating the conditional log likelihood at 21 equally-spaced points in the interval $\left[\hat{R}_m^{-1}-3\sigma\left(\hat{R}_m^{-1}\right),\hat{R}_m^{-1}+3\sigma\left(\hat{R}_m^{-1}\right)\right]$.

Fitting was performed using the 'fminsearch' function in Matlab, with multiple restarts, to ensure that a local minimum was not selected. Each restart was started randomly in the parameter subspace where the model's mean interval at $R=1$ matched the corresponding measured mean interval.

The Bayesian Information Criterion (BIC) was used to compare models. We selected it over other candidates, such as the Akaike Information Criterion (AIC), due to its consistency. That is, as the number of samples $n\to\infty$, the probability that the BIC will select the true model (assuming it is among the candidate models) approaches 1, while the AIC will tend to over-fit the data[2]. We note, however, that in the case of all model comparisons in the manuscript, none of the conclusions are altered by utilizing the AIC over the BIC.

The BIC is calculated as follows:

$$\mathrm{BIC}=-2\log L(\boldsymbol{\theta}_{\max})+\log n \tag{S5}$$

where $\boldsymbol{\theta}_{\max}$ is the parameter set which maximizes $\log L(\boldsymbol{\theta})$.

IV. Number of transitions into the OFF state per RNA production event

In this section, we estimate the number of times that, on average, a promoter will transit into the OFF state for each time it commits to transcription. This estimation is made for the best fitting model (see Table 2 in the main manuscript).

For the best-fitting model (Limiting Mechanism I), the back-and-forward transitions between P_{ON} and RP_c states can be considered to be fast (since $k_{-1}>>k_2$ and $k_1>>k_{\mathrm{OFF}}$). We can therefore apply the slow-scale SSA to merge these two states[3]. In this limit, the probabilities $\mathrm{P}(P_{ON})$ and $\mathrm{P}(P_c)$ of being in P_{ON} and P_c states, respectively, are:

$$\mathrm{P}\left(P_{ON}\right)=\frac{1}{K_a+1}\ \text{and}\ \mathrm{P}\left(P_c\right)=\frac{K_a}{K_a+1},\ \text{with}\ K_a=k_1k_{-1}^{-1} \tag{S6}$$

The propensity of changing from the merged state to RP_o is then $\left[\mathrm{P}\left(P_c\right)k_2\right]$, while the propensity to move from the merged state to P_{OFF} equals $\left[\mathrm{P}\left(P_{ON}\right)k_{OFF}\right]$. The probability of moving into P_c instead of P_{OFF} is therefore given by:

$$\mathrm{P}_{c/OFF}=\frac{\mathrm{P}\left(P_c\right)k_2}{\mathrm{P}\left(P_{ON}\right)k_{OFF}+\mathrm{P}\left(P_c\right)k_2} \tag{S7}$$

Since each attempt at transcription is independent in the model, and has a constant probability of committing at each attempt, the number of times that the systems changes into the OFF state prior to committing to transcription follows a geometric distribution with a probability of success of $\mathrm{P}_{c/OFF}$. The mean of this distribution is:

$$\mu = \frac{1 - \mathrm{P}_{c/OFF}}{\mathrm{P}_{c/OFF}} \tag{S8}$$

Converting this in terms of model parameters (and given $k_1 k_2 k_{-1}^{-1} k_{\mathrm{OFF}}^{-1} = 0.11$ from Table 2) one obtains:

$$\mu = \frac{k_{OFF}}{k_2 K_a} = \frac{1}{0.11} \tag{S9}$$

V. Minimum samples required for a given precision

To estimate the number of samples required to obtain a given precision in the estimates of $\tau_{\overline{\mathrm{CC}}}$ and τ_{CC}, consider the following alternate method of measuring these values if we could sample the uncensored interval distribution between transcription events.

Let these measurements be at two RNAp concentrations $\hat{R}_m$, where $m = \{1,2\}$ such that $D = \hat{R}_1 / \hat{R}_2 > 1$. Let I_m be the population mean of the inter-transcription intervals in medium m, with corresponding standard deviation σ_m, and that we have n_m samples of this distribution (we assume, without significant loss of generality, that $n_1 = n_2 = n$). For sufficient n, estimates of the population means $\hat{I}_m$ will follow Normal distributions with $\sigma^2\left(\hat{I}_m\right) = \sigma_m^2 / n$. The least-squares fit of a line to these points will thus result in:

$$\hat{\tau}_{\overline{\mathrm{CC}}} = \frac{\hat{I}_1 D - \hat{I}_2}{D-1},\ \sigma\left(\hat{\tau}_{\overline{\mathrm{CC}}}\right) = \frac{D^2\sigma_1^2 + \sigma_2^2}{n(D-1)^2} \tag{S10}$$

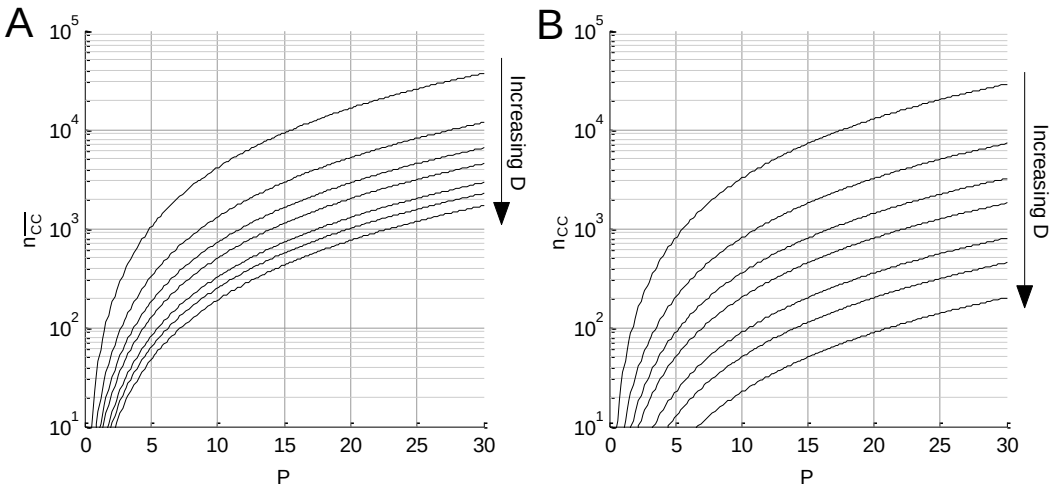
$$\hat{k}_{\mathrm{CC}}^{-1} = \frac{\hat{I}_2 - \hat{I}_1}{D-1},\ \sigma\left(\hat{k}_{\mathrm{CC}}^{-1}\right) = \frac{\sigma_1^2 + \sigma_2^2}{n(D-1)^2} \tag{S11}$$

Note that this method will overestimate the uncertainty in $\hat{\tau}_{\overline{\mathrm{CC}}}$ and $\hat{k}_{\mathrm{CC}}^{-1}$ since these estimates are highly anti-correlated. We define the precision of the measurement as $P = I_1 / \sigma\left(\hat{\tau}_x\right)$, where $\hat{\tau}_x$ is $\hat{\tau}_{\overline{\mathrm{CC}}}$ or $\hat{k}_{\mathrm{CC}}^{-1}$. Intuitively, this definition relates the uncertainty in the estimate with the mean timescale of the intervals. For example, if the intervals are on a timescale of ~500 s, to achieve a precision of 10 in $\hat{\tau}_{\overline{\mathrm{CC}}}$, we must know it to within 50 s. Assuming that $\sigma_1^2 I_1^{-2} \approx \sigma_2^2 I_2^{-2} = \eta^2$, i.e. that the CV^2 of the interval distribution is similar between the two RNAp concentrations, the number of samples required to achieve a given precisions in $\hat{\tau}_{\overline{\mathrm{CC}}}$ and $\hat{k}_{\mathrm{CC}}^{-1}$ is:

$$n_{\overline{\mathrm{CC}}} = \eta^2 P^2 \frac{D^2 + 1}{(D-1)^2} \tag{S12}$$

$$n_{\mathrm{CC}} = \eta^2 P^2 \frac{2}{(D-1)^2} \tag{S13}$$

Note that the above assumes that there is no variance in the estimate of the RNAp concentration, and that all n samples are uncensored. Equations (S12) and (S13) should therefore be considered as only a rough guide for the number of samples required. The number of samples required for a range of precisions and possible dynamic ranges in RNAp concentrations is shown in Supplementary Figure S3.



Supplementary Figure S3: Number of samples required in two conditions to achieve a given precision in (A) $\hat{\tau}_{\overline{CC}}$ and (B) $\hat{k}_{CC}^{-1}$, with production interval measurements at only two RNAp concentrations with ratio D and assuming $\eta^2 = 1$. Lines are shown for values of D of 1.25, 1.5, 1.75, 2, 2.5, 3, and 4 (from top to bottom).

VI. Photo-toxicity measurements

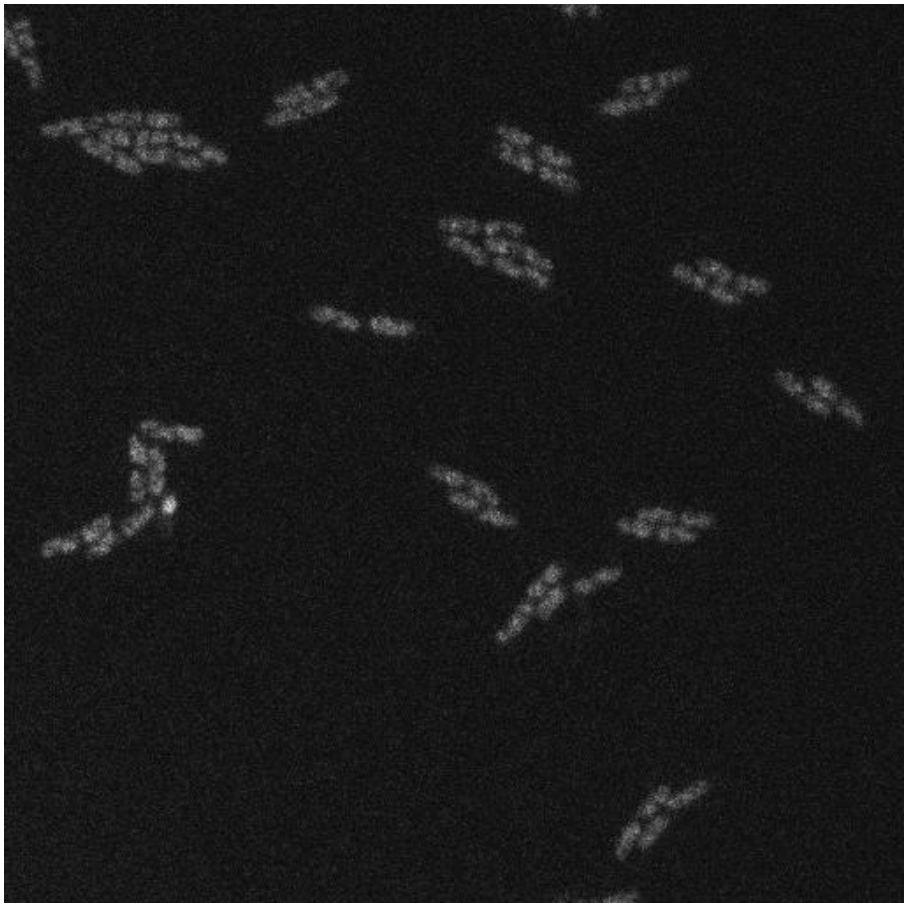
To assess the level of phototoxicity from the imaging procedure under the microscope, we took the measurements in the 1.00x case (Table 1, main manuscript), and estimated the cells' doubling time under the microscope by counting the number of cells at the start and end of the two hour measurement period (first row of Table S2). In this case, cells were imaged by phase contrast every 5 minutes, and confocal microscopy every minute for two hours. We then imaged two new populations of cells, but in the first, we only imaged the cells with phase contrast (i.e. no confocal, row 2 of Table S2), while in the second, only two images were taken in total, one at the start and one at the end (row 3 of Table S2).

Phase Contrast	Confocal	Cells at start	Cells at end	Doubling Time
5 min	1 min	206	468	52.8 min
5 min	Not used	399	962	49.8 min
2 h	Not used	480	1189	48.4 min

Supplementary Table S2: Phototoxicity under the microscope for different imaging intervals and channels. All measurements took 2 hours. The first two columns of the table show the intervals at which images were taken. The subsequent columns show the number of cells at the start and end of the measurements, obtained from single phase contrast images. Finally, it is shown the estimated doubling time of the cells, which was determined from the fold change.

From Supplementary Table S2, the estimated doubling time while taking images with both channels is only 4.4 minutes longer than in the case with minimal imaging. Thus, while there is an observable effect on the doubling time, it is not expected to cause significant differences in the transcription initiation dynamics. In any case, any changes would affect all conditions similarly, and will not affect *relative* RNAp concentrations. Finally, we note that the effect from phase contrast imaging appears to be negligible.

VII. Cell-to-cell variability in RNAp concentrations



Supplementary Figure S4: Confocal image of RL1314 cells expressing fluorescently-tagged RpoC in 1x media, one hour after being placed in the thermal imaging chamber at 37 ^{o}C. Contrast was enhanced for easier visualization.

VIII. Number of promoter copies during the cell lifetime

The model fitting procedure employed in the main text assumes that there is only one copy of the target promoter in a cell at all times. To determine to what extent this assumption is not true in our experimental system, we measured the fraction of time cells contain two chromosomes. Since the F-plasmid replicates at the same time[4] or shortly after[5] the chromosome, this provides an upper bound for the fraction of time the cells spend with more than one promoter of interest (it is worth noting that, in our measurements, we did not observe cells with more than 2 nucleoids at any given point).

For this, *E. coli* DH5α-PRO cells (see main text) were transformed with the pAB332 plasmid carrying the gene *hupA-mcherry* that encodes a fluorescent protein tag under the control of the *hupA* constitutive promoter[6]. This tagging protein, composed of a nucleoid-associated protein (HupA) fused with a red fluorescent protein (mCherry), can be used to assess the location and size of nucleoids in live cells[7] (see Methods).

Cells were diluted from overnight culture to an OD_{600} of 0.05 in fresh 1x media, supplemented with appropriate antibiotics, and kept at 37°C in a shaker at 250 rpm, until reaching an OD_{600} of 0.3. Cells were then placed in a thermal chamber (FCS2, Bioptechs, USA), set to 37°C, and imaged once every minute for 1 hour (the red signal was too weak to continue after 1 hour) using a Nikon Eclipse (Ti-E, Nikon) inverted microscope equipped with C2+ (Nikon) confocal laser-scanning system. To visualise HupA-mCherry-tagged nucleoids, we used a 543 nm HeNe laser (Melles-Griot) and an emission filter (HQ585/65, Nikon). Phase contrast images of cells were captured every 5 minutes by a CCD camera (DS-Fi2, Nikon).

Cells were segmented from phase contrast images using CellAging[8]. Fluorescent nucleoids were segmented and quantified from confocal images as in [7,9]. Of the cells that were born and divided during the time series (124 cells), we found that the mean fraction of time points in which cells had two nucleoids was 0.114 ± 0.010.

Thus, we estimate the fraction of time spent with multiple target promoters to be at most 11.4 ± 1.0% in 1x media. As this was the most nutrient-rich condition tested, other conditions should have even lower fractions[5].

References

1. Häkkinen, A., and Ribeiro, A. S. 2015, Characterizing rate limiting steps in transcription from RNA production times in live cells. *Bioinformatics*, in press. DOI: 10.1093/bioinformatics/btv744.

2. Burnham, K. P., and Anderson, D. R. 2004, Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociol. Methods Res.*, **33**, 261–304.

3. Cao, Y., Gillespie, D. T., and Petzold, L. R. 2005, The slow-scale stochastic simulation algorithm. *J. Chem. Phys.*, **122**, 14116.

4. Cooper, S., and Keasling, J. D. 1998, Cycle-specific replication of chromosomal and F plasmid origins. *FEMS Microbiol. Lett.*, **163**, 217–22.

5. Keasling, J. D., Palsson, B. Ø., and Cooper, S. 1991, Cell-cycle-specific F plasmid replication: Regulation by cell size control of initiation. *J. Bacteriol.*, **173**, 2673–80.

6. Fisher, J. K., Bourniquel, A., Witz, G., Weiner, B., Prentiss, M., and Kleckner, N. 2013, Four-dimensional imaging of *E. coli* nucleoid organization and dynamics in living cells. *Cell*, **153**, 882–95.

7. Oliveira, S. M. D., Neeli-Venkata, R., Goncalves, N. S. M., et al. 2016, Increased cytoplasm viscosity hampers aggregate polar segregation in *Escherichia coli*. *Mol. Microbiol.*, **99**, 686–99.

8. Häkkinen, A., Muthukrishnan, A.-B., Mora, A., Fonseca, J. M., and Ribeiro, A. S. 2013, CellAging: A tool to study segregation and partitioning in division in cell lineages of *Escherichia coli*. *Bioinformatics*, **29**, 1708–9.

9. Mora, A. D., Vieira, P. M., Manivannan, A., and Fonseca, J. M. 2011, Automated drusen detection in retinal images using analytical modelling algorithms. *Biomed. Eng. Online*, **10**, 59.