Ribosome biogenesis in replicating cells: Supporting Information

Tyler M. Earnest, John A. Cole, Joseph R. Peterson, Michael J. Hallock, Thomas E. Kuhlman, and Zaida Luthey-Schulten

S1 Semi-analytical Modeling

S1.1 Ribosomal Protein Operon mRNA Statistics

We consider the mRNA statistics for the ribosomal protein operons. From Equation 30, we can write out the CME for our system as:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}P(m,n|t) = k_t(t)\Big[P(m-1,n|t) - P(m,n|t)\Big] \\ + k_b\Big[(m+1)P(m+1,n-1|t) - mP(m,n|t)\Big] \\ + k_u\Big[(n+1)P(m-1,n+1|t) - nP(m,n|t)\Big] \\ + k_d\Big[(m+1)P(m+1,n|t) - mP(m,n|t)\Big],\end{aligned} \tag{S1}$$

where $k_t(t)$ represents the effective transcription rate as a function of time, with $k_t(t < t_r) = k_t$ and $k_t(t > t_r) = 2k_t$ (where t_r is the gene replication time, itself a function of the timing of chromosome replication and the gene's position on the chromosome). We note that we have implicitly assumed that transcription from both gene copies after duplication is independent and occurs at the same rate, which may not in general be true[1], but simplifies the model considerably. From this we can derive the system of ODEs to describe the time evolution of the mean counts of m and n, their mean squared counts, and the mean product of m and n:

$$\begin{aligned}\frac{d}{dt}\langle m\rangle(t) &= k_t(t) - k_d\langle m\rangle(t) - k_b\langle m\rangle(t) + k_u\langle n\rangle(t) \\ \frac{d}{dt}\langle n\rangle(t) &= k_b\langle m\rangle(t) - k_u\langle n\rangle(t) \\ \frac{d}{dt}\langle m^2\rangle(t) &= 2k_t(t)\langle m\rangle(t) + k_t(t) - 2k_b\langle m^2\rangle(t) + k_b\langle m\rangle(t) \\ &\quad + 2k_u\langle mn\rangle(t) + k_u\langle n\rangle(t) - 2k_d\langle m^2\rangle(t) + k_d\langle m\rangle(t) \\ \frac{d}{dt}\langle n^2\rangle(t) &= 2k_b\langle mn\rangle(t) + k_b\langle m\rangle(t) - 2k_u\langle n^2\rangle(t) + k_u\langle n\rangle(t) \\ \frac{d}{dt}\langle mn\rangle(t) &= k_t(t)\langle n\rangle(t) - k_b\langle mn\rangle(t) + k_b\langle m^2\rangle(t) - k_b\langle m\rangle(t) \\ &\quad + k_u\langle n^2\rangle(t) - k_u\langle mn\rangle(t) - k_u\langle n\rangle(t) - k_d\langle mn\rangle(t).\end{aligned} \tag{S2}$$

We expect that at cell division all components are distributed to the daughter cells according to an unbiased binomial distribution. This can be used to derive constraints for our system of ODEs, namely:

$$\begin{aligned}
\langle m\rangle(0) &= \frac{1}{2}\langle m\rangle(t_D) \\
\langle n\rangle(0) &= \frac{1}{2}\langle n\rangle(t_D) \\
\langle m^2\rangle(0) &= \frac{1}{4}\Big[\langle m\rangle(t_D) + \langle m^2\rangle(t_D)\Big] \\
\langle n^2\rangle(0) &= \frac{1}{4}\Big[\langle n\rangle(t_D) + \langle n^2\rangle(t_D)\Big] \\
\langle mn\rangle(0) &= \frac{1}{4}\langle mn\rangle(t_D).
\end{aligned} \tag{S3}$$

This system can be solved numerically, but parameters must be chosen carefully. Specifically we are concerned with the ribosome binding and unbinding rates. The binding rate, k_b, is clearly a function of the concentration of free ribosomes as well as other mRNA in the cell; as a first approximation, we might expect:

$$k_b \approx k_{b,0}[r_{\text{free}}] = k_{b,0}[r - C_r - n], \tag{S4}$$

where $k_{b,0}$ represents the binding rate of a single messenger to a single ribosome, r represents the ribosome copy number in the cell, C_r represents the number of competing mRNA that are bound to ribosomes, n is the number of ribosome-bound versions of the messenger we are interested in, and square brackets e.g. $[x] \approx x \cdot 2^{-t/t_D}/2\ln(2))$ denotes a per-cell concentration. We might assume that the competing mRNA are in equilibrium with respect the ribosomes, meaning:

$$k_u[C_r] = k_{b,0}[C - C_r][r - C_r - n] \tag{S5}$$

where C represents the total number of competing mRNA. Solving this for C_r and inserting the result into Equation S4 then yields:

$$k_b \approx k_{b,0}\left([r] - [n] - \frac{1}{2}\left(-\sqrt{(\frac{k_u}{k_{b,0}} - [n] + [r] + [C])^2 + 4([n][C] - [r][C])} + \frac{k_u}{k_{b,0}} - [n] + [r] + [C]\right)\right). \tag{S6}$$

Inserting Equation S6, a value for k_u chosen such that the ribosome-bound messengers have an appropriate mean lifetime, the mean value of $[r] = 3000$[6], and $C = c(m+n)$ (where c denotes the number of competing genes, assuming that the competing mRNA production roughly keeps pace with that of the messenger we are interested in) into Equation S2 and solving the system numerically (using the NDSolve function in Mathematica) yields traces for the mRNA statistics over the cell cycle. We can then perform the appropriate time-averaging over the cell cycle (see Peterson et al.[3] for details) in order to calculate the mean and variance of our mRNA:

$$\begin{aligned}
\mathrm{E}[m] &= \int_0^{t_D} \frac{2\ln(2)}{t_D} 2^{-t/t_D}\langle m\rangle(t)\,\mathrm{d}t \\
\mathrm{E}[n] &= \int_0^{t_D} \frac{2\ln(2)}{t_D} 2^{-t/t_D}\langle n\rangle(t)\,\mathrm{d}t \\
\mathrm{Var}[m] &= \int_0^{t_D} \frac{2\ln(2)}{t_D} 2^{-t/t_D}\langle m^2\rangle(t)\,\mathrm{d}t - \mathrm{E}[m]^2 \\
\mathrm{Var}[n] &= \int_0^{t_D} \frac{2\ln(2)}{t_D} 2^{-t/t_D}\langle n^2\rangle(t)\,\mathrm{d}t - \mathrm{E}[n]^2 \\
\mathrm{Cov}[m,n] &= \int_0^{t_D} \frac{2\ln(2)}{t_D} 2^{-t/t_D}\langle mn\rangle(t)\,\mathrm{d}t - \mathrm{E}[m]\,\mathrm{E}[n].
\end{aligned} \tag{S7}$$

From these we can compute the statistics of our total mRNA count:

$$\begin{aligned} \mathrm{E}[m+n] &= \mathrm{E}[m]+\mathrm{E}[n] \\ \mathrm{Var}[m+n] &= \mathrm{Var}[m]+\mathrm{Var}[n]+2\,\mathrm{Cov}[m,n]. \end{aligned} \tag{S8}$$

S1.2 Estimating Rate Parameters for an "Average mRNA"

Because the nine ribosomal protein operons have varying rates of production, translation, and degradation, we attempted, for the sake of simplicity, to investigate the behavior of an "average mRNA". We first computed the harmonic mean of the operons' transcription and degradation rates (yielding 0.0042 s^{-1} and 9.84×10^{-4} s^{-1}, respectively). Then, in order to estimate the ribosome unbinding rate for each operon we computed the mean lifetime of each mRNA-ribosome complex. Each operon has a different number of genes to be translated, each of which in turn has a different translation rate, meaning that each operon will be bound to a ribosome for a different amount of time. We can compute the mean mRNA-ribosome complex lifetime for each operon, and from that determine each operon's effective unbinding rate:

$$k_{u,i} = \left(\frac{1}{k_{\text{su, dissoc}}^{-1} + \sum_j k_{t1,i,j}^{-1}} \right)^{-1} \tag{S9}$$

where $k_{u,i}$ represents the unbinding rate for the i^{th} operon's messengers, $k_{\text{su, dissoc}}$ represents the rate at which a translated ribosome dissociates from the messenger, and $k_{tl,i,j}$ represents the translation rate of the j^{th} gene in operon i. The results of these computations are summarized in Table V.

S1.3 mRNA statistics in the Limit when $k_d \to 0$

Peterson et al.[3] derived expressions for mRNA statistics that accounted for gene duplication due to chromosome replication; specifically, Equation S36 is given as:

$$\begin{aligned} \mathrm{E}[r] &= \frac{k_t}{k_d} 2^f \left[1 + \beta \frac{\mathrm{e}^{-k_d t_D (1-f)} - 2^{1-f}}{1 + \frac{k_d t_D}{\ln(2)}} + \gamma \frac{2^{-f} \mathrm{e}^{-k_d t_D f} - 1}{1 + \frac{k_d t_D}{\ln(2)}} \right] \\ \mathrm{Var}[r] &= \mathrm{E}[r] - \mathrm{E}[r]^2 \\ &\quad + \ln(2) \left(\frac{k_t}{k_d} \right)^2 \left[2\beta^2 \frac{1 - 2^{f-1} \mathrm{e}^{-2k_d t_D (1-f)}}{\ln(2) + 2k_d t_D} - 4\beta \frac{1 - 2^{f-1} \mathrm{e}^{-k_d t_D (1-f)}}{\ln(2) + k_d t_D} + \frac{2}{\ln(2)} (1 - 2^{f-1}) \right. \\ &\quad \left. + \gamma^2 \frac{2^f - \mathrm{e}^{-2k_d t_D f}}{\ln(2) + 2k_d t_D} - 4\gamma \frac{2^f - \mathrm{e}^{-k_d t_D f}}{\ln(2) + k_d t_D} - \frac{4}{\ln(2)} (1 - 2^f) \right], \end{aligned} \tag{S10}$$

where

$$\begin{aligned} \beta &= \frac{\mathrm{e}^{-k_d t_D f}}{2 - \mathrm{e}^{-k_d t_D}} \\ \gamma &= \left(1 + \frac{\mathrm{e}^{-k_d t_D}}{2 - \mathrm{e}^{-k_d t_D}} \right). \end{aligned} \tag{S11}$$

In these equations, k_t and k_d are the RNA transcription and degradation rates, respectively, t_D is the cell doubling time, and f represents the fraction of the cell cycle after gene replication ($f = 1 - t_r/t_D$). For the purposes of the present

work, we note that the nucleation and assembly of the ribosome occurs significantly faster than measured rates of mRNA degradation; as a result, we expect little rRNA to be lost, and essentially all of it to be found in the form of ribosomes in the cell. Similarly, because the mRNA-ribosome dissociation constant is small ($\sim 10^{-10}$ M), when the pool of ribosomes is large compared to the pool of available messengers, essentially all mRNA will remain bound to ribosomes and few will be degraded. We therefore consider the limit of the expressions in Equation S10 as k_d approaches zero:

$$\begin{aligned}\lim_{k_d\to 0} \mathrm{E}[r] &= \frac{k_t t_D 2^f}{\ln(2)} \\ \lim_{k_d\to 0} \mathrm{Var}[r] &= \frac{k_t t_D}{\ln^2(2)}\Big[2^f\Big(\ln(2)+2k_t t_D(3+\ln(4))\Big) - 4^f k_t t_D \\ &\quad - k_t t_D\Big(4+2(1+f)^2\ln^2(2)+(1+f)\ln(16)\Big)\Big].\end{aligned} \tag{S12}$$

S2 Estimating Cell Cycle Parameters from Copy Number Distributions

We consider the well known age distribution of exponentially growing cells[8]:

$$\phi(a) = 2\nu_m e^{-\nu_m a}\int_a^\infty f(\tau)\,\mathrm{d}\tau, \tag{S13}$$

where $\phi(a)$ is the probability that a cell is of age a, ν_m is the growth rate of the population, and $f(\tau)$ is the probability of a cell dividing at age τ. As per Powell[8], ν_m can be determined from the constraint

$$2\int_0^\infty e^{-\nu_m\tau} f(\tau)\,\mathrm{d}\tau = 1. \tag{S14}$$

Taylor expanding the LHS of Equation S14 about the mean division time:

$$\begin{aligned}1 &= 2\int_0^\infty e^{-\nu_m\tau} f(\tau)\,\mathrm{d}\tau \\ &= 2\langle e^{-\nu_m\tau}\rangle \\ &\approx 2\left[e^{-\nu_m\langle\tau\rangle} + \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}\tau^2}e^{-\nu_m\tau}|_{\langle\tau\rangle}\sigma_\tau^2\right] \\ &= 2\left[e^{-\nu_m t_D} + \frac{1}{2}\nu_m^2 e^{-\nu_m t_D}\sigma_\tau^2\right] \\ &= (2+\nu_m^2\sigma_\tau^2)e^{-\nu_m t_D}\end{aligned} \tag{S15}$$

where we assume that the mean age at division is t_D, and the division ages have some variance σ_τ^2. This can then be easily solved numerically for ν_m.

We now consider the probability that a cell has a single copy of a given gene. If t_r is the age at which the gene is replicated, we can write:

$$P_{\text{single copy}} = \int_0^{t_r}\phi(a)\,\mathrm{d}a. \tag{S16}$$

For simplicity, we can assume the division times are normally distributed,

$$f(\tau) = N(\tau; t_D, \sigma_\tau). \tag{S17}$$

and so,

$$\begin{aligned} P_{\text{single copy}}(t_r) &= \int_0^{t_r} 2\nu_m e^{-\nu_m a} \int_a^{\infty} f(\tau)\,\mathrm{d}\tau\,\mathrm{d}a \\ &= \int_0^{t_r} 2\nu_m e^{-\nu_m a} \frac{1}{2}\operatorname{erfc}\left(\frac{a-t_D}{\sqrt{2}\sigma_\tau}\right)\mathrm{d}a. \end{aligned} \tag{S18}$$

Promoting t_r to a random variable distributed according to some probability function $P(t_r;\langle t_r\rangle,\sigma_{t_r})$ where $\langle t_r\rangle$ and σ_{t_r} are the mean and standard deviation of the replication time, respectively, we can write

$$\langle P_{\text{single copy}}\rangle = \int_0^{\infty} \mathrm{d}t_r P_{\text{single copy}}(t_r) P(t_r;\langle t_r\rangle,\sigma_{t_r}). \tag{S19}$$

Now simply Taylor expanding about $\langle t_r\rangle$ yields

$$\begin{aligned} \langle P_{\text{single copy}}\rangle(\langle t_r\rangle,\sigma_{t_r}) \approx &\int_0^{\langle t_r\rangle} 2\nu_m e^{-\nu_m a}\frac{1}{2}\operatorname{erfc}\left(\frac{a-t_D}{\sqrt{2}\sigma_\tau}\right)\mathrm{d}a \\ &+ \frac{\sigma_{t_r}^2}{2}\left[-2\nu_m^2 e^{-\nu_m\langle t_r\rangle}\frac{1}{2}\operatorname{erfc}\left(\frac{\langle t_r\rangle - t_D}{\sqrt{2}\sigma_\tau}\right) - 2\nu_m e^{-\nu_m\langle t_r\rangle} f(\langle t_r\rangle)\right], \end{aligned} \tag{S20}$$

subject to Equation S14. Now, for each gene locus, we can estimate the fraction of cells we expect to see with a single gene copy. Inserting forms for the mean replication time, $\langle t_{r,i}\rangle = \mu_{\text{trep}} + \chi_i T_{\text{rep}}$, for a gene given its location, χ_i on the chromosome, as well as the standard deviation in the replication timing, $\sigma_{t_r} = \sigma_{\text{trep}}$ into Equation S20, we can construct a measure for for goodness of fit:

$$\Theta = \sum_{i\in\text{genes}} \left(\frac{\langle P_{\text{single copy}}\rangle(\langle t_{r,i}\rangle,\sigma_{t_r}) - \frac{n_i}{m_i}}{\sqrt{\langle P_{\text{single copy}}\rangle(\langle t_{r,i}\rangle,\sigma_{t_r})\left(1-\langle P_{\text{single copy}}\rangle(\langle t_{r,i}\rangle,\sigma_{t_r})\right)/m_i}}\right)^2 m \tag{S21}$$

and vary μ_{trep}, T_{rep}, and σ_{trep} in order to minimize it. Importantly, here n_i denotes the number of cells (with gene i labeled) observed with a single copy, m_i denotes the total number of cells (with gene i labeled) observed, and the term $\sqrt{\langle P_{\text{single copy}}\rangle(\langle t_{r,i}\rangle,\sigma_{t_r})\left(1-\langle P_{\text{single copy}}\rangle(\langle t_{r,i}\rangle,\sigma_{t_r})\right)/m_i}$ denotes an estimate for the error in the experimentally observed fraction of cells with one gene copy. This estimate is based in the assumption that a cell has probability $\langle P_{\text{single copy}}\rangle(\langle t_{r,i}\rangle,\sigma_{t_r})$ of being in a one-copy state, and that each measured cell represents an independent Bernoulli trial. This error estimate was introduced in order to give greater weight during fitting to the genes for which we have greater numbers of experimental images.

We assumed a value for σ_τ of 12 minutes, and, because Θ was found to change little with variations in σ_{t_r}, we initially required $20.2 < \sigma_{t_r} < 24.0$ (such that its value stays within the error bounds found by the more complete fitting method presented in the main manuscript, see Table I). Θ was then minimized using the Minimize routine in Mathematica. This resulted in estimates for μ_{trep} and T_{rep} of 34.4 and 45.9 minutes, respectively. We note this value for T_{rep} is well within the standard error reported in Table I but the value for μ_{trep} differs from the results of the main text by approximately 2.6 standard deviations. Comparison of the fit and experimental single-gene fractions shows similar qualitative agreement as was obtained using the method presented in the main text (see Figure S6). For the sake of comparison, releasing the bounds on σ_{t_r} had only a minor effect on Θ, μ_{trep}, and T_{rep} (their values changed by about 0.05%, 3%, and 0.2%, respectively), but the fit value of σ_{t_r} fell to an unreasonably low value of approximately 0.25 seconds.

S3 Fitting mRNA Distributions

The exact analytical theory set out in Peterson et al. describes the noise of idealized constitutively expressed genes which undergo duplication during the cell cycle[3]. As no experimental data is available to compare to the distributions of messengers computed in the RBM at the time of writing, the results were compared to this theory. Due to the fact that the ribosomal protein operon messengers are the only transcripts competing for the ribosomes, they are bound up

and effectively prevented from degradation. In the future, all cellular transcripts will be considered, thus the RBM will need to be reparameterized. By fitting the theory of Peterson et al. to the RBM simulations, we can estimate what the new parameters would need to be to give the same results as the RBM.

Distributions computed using Eq. S22 of[3] were computed with varying k_t and k_d and compared to the RBM simulated distributions. The mean squared deviation was computed between these simulations. The k_t and k_d associated with the distribution that has the minimum deviation from the simulated distribution represent the "effective" transcription and degradation rates that will be applicable in future simulations that include realistic counts of competing mRNAs. The fitted rates and fits can be seen in Table V and Figure 6. Fitted k_t and k_d are essentially scaled versions of the rates used in the RBM (as demonstrated in Figure S9). Fit k_d are about four times smaller than experimental values, while k_t are about four times as large.

S4 Varying Numbers of Non-ribosomal Genes in the SAM

We used the SAM to study the simultaneous effect of varying gene loci (which effects the timing of gene replication) and c on the mRNA copy number statistics for an "average gene" (see Section S1.2). Interestingly, we found that with increasing numbers of actively-expressed genes (while holding k_t and k_d constant), gene expression became significantly less noisy (see Figure S10b). This was found to be due largely to the fact that the messenger-ribosome dissociation constant was very small ($k_u/k_{\text{mrna\_assoc}} \sim 10^{-10}$ M). By comparison, the concentration of a single mRNA in a bacterial cell of volume $\sim$1 fl is approximately 1.7×10^{-9} M. This means that every messenger produced should have a high probability of being bound to a ribosome, provided a ribosome is available for binding. When small numbers of non-ribosomal genes are expressed (the small-c regime), the total number of messengers does not exceed the total number of ribosomes, and so every messenger is likely to be quickly bound and thereafter protected from degradation. The statistics of a given gene's mRNA in this regime are then essentially the same as those of a model in which *already bound* messengers are produced (at some transcription rate k_t) and only lost through cell division (with roughly half going to each daughter). The model of Peterson et al.[3] in the limit where $k_d \to 0$ (see Figure S10b, left-most dots, and see Equation S12 in the Supporting Information) gives exactly these statistics. At values of c between 30 and 50, the total mRNA content of the cell approaches and then surpasses the total number of ribosomes. From there on, with increasing values of c, the probability of a given messenger binding a ribosome becomes increasingly small. For any specific gene of interest, this results in an increase in the fraction of unbound mRNAs (relative to ribosome-bound mRNAs), and in turn an increase in the messenger's effective degradation rate, a decrease in its mean copy number, and a decrease in its Fano factor. In the limit where $c \to \infty$, the probability that any specific messenger is ever bound by a ribosome approaches zero; the total messenger count is then dominated by the unbound messengers and the statistics converge again to those of a model in which ribosome interactions are entirely neglected, e.g. that of Peterson et al.[3] (see Figure S10b right-most dots). These findings are in good agreement with additional explicit stochastic simulations (see Figure S10b, diamonds, and see Section 4.3).

We note that the SAM described here remains somewhat incomplete. It neglects, for example, the potential for transient non-specific mRNA-ribosome interactions[5] which have been shown to significantly impact ribosome diffusivity. Such interactions have been estimated to last on the order of a few seconds and may play a role in ribosomal 50S-30S subunit search and association[5]. Assuming that the transiently-bound mRNA are, like the specifically-bound mRNA, protected from degradation, then this type of interactions should have the net effect of lowering the free messenger counts, and in turn, lowering the effective degradation rate. When c in the SAM is small (e.g. $c = 8$), the effective degradation rate is already approximately zero, and so non-specific ribosome binding can not significantly affect the mRNA statistics. When c is in the biologically realistic regime ($c \sim 1000$), an upper-bound for the possible changes in the mRNA statistics can be roughly estimated by considering the effect of increasing the ribosome concentration in the SAM. For example, if we assume that only one messenger can transiently bind a ribosome at a given time, and that the transient mRNA-ribosome association rate is fast (e.g. occurring at a diffusion-limited rate on the order of 10^9 M^{-1} s^{-1}), then we can expect that the available non-specific binding sites should essentially always be occupied, and the mRNA statistics can be approximated simply by doubling the ribosome concentration in the SAM. For a gene situated on the chromosome halfway between the origin and terminus, this has the effect of increasing mean messenger count from 8.6 to 11.3 per cell, and only modestly changes the Fano factor from 1.6 to 1.7.

S5 Algorithms used in RBM

Algorithm 1 details the process of pruning the intermediate species graph. This method is used by Earnest et al.[2] to reduce the approximately 1600 potential intermediates down to a number of species that can be simulated with RDME.

The RDME trajectory generation process is detailed in Algorithm 2. The cell growth process is implemented using pyLM's "hybrid solver" interface[4] which allows for user-defined processes to occur during regular intervals when simulation data is saved. This is when cell geometry is updated to account for growth, and to add new DNA operons to the simulation to reflect the replication process.

The process for building the diving cell geometry is shown in Algorithm 3. Algorithm 4 is the process of discerning which discrete lattice sites fall inside a spherocylinder described via two points at the cylinder ends and a radius. Algorithm 5 detects boundaries between different site types and changes those sites to a third type. It is used to construct the membrane between the cytoplasm and extracellular space.

Algorithm 1: Ribosome SSU intermediate species network pruning

Data: graph $G = (V, E)$ with intermediate species as vertices and directed edges indicating reactions
Result: $G' = (V', E')$ representing pruned network
$V' = V$
$E' = E$
repeat
  Compute fluxes through G'
  $D \longleftarrow$ vertex in G' with lowest flux
  while $|D| > 0$ **do**
    for $v \in D$ **do**
      for $e \in E'$ **do** `// Remove edges to and from v`
        if $v \in e$ **then**
          $E' = E' - e$
        end
      end
      $V' = V' - v$ `// Remove v`
    end
    $D = \emptyset$
    `// Locate dead-end vertices`
    for $v \in V'$ **do**
      if *number of edges in or out of* $v = 0$ **then**
        $D = D + v$
      end
    end
  end
until $|V'| <$ *maxSpecies*

Algorithm 2: RDME trajectory process for RBG and cell division

Data: Initial particle lattice, time step τ, DNA Operons and their replication time
Result: particle lattice at simulation end time T
$t = 0$
$latticeWriteTime = writeInterval$
while $t < T$ **do**
    `x_axis_diffusion(`*lattice*`)`
    `y_axis_diffusion(`*lattice*`)`
    `z_axis_diffusion(`*lattice*`)`
    `perform_reactions(`*lattice*`)`
    $t = t + \tau$
    if $t > latticeWriteTime$ **then**
        if $t < doublingTime$ **then**
            `buildGrowingCells(`*cell dimensions, t/doublingTime*`)`
            for $op \in Operons$ **do**
                if $t >= op_{birth}$ **then**
                    Replicate Operon op
                end
            end
        end
        $latticeWriteTime = latticeWriteTime + writeInterval$
    end
end

Algorithm 3: buildGrowingCells: Construct lattice sites for growing cells

Data: length, width, nucWidthFraction, nucLengthFraction, gp
Result: lattice with sites matching cell geometry
$r = width/2$ `// Cell radius`
$h = (length - width)/2$ `// Capsule cylinder height`
`/* Determine total growth length, gp is a value in [0,1] to denote time in the cell cycle. The linear term ensures at least one lattice space of separation between daughter cells */`
$growth = length * (2^{gp} - 1) + gp$
`/* Find the capsule center points; one at each end of the inner cylinder */`
$mother1 = \{x_{center}, y_{center}, z_{center} - h/2 - growth\}$
$mother2 = \{x_{center}, y_{center}, z_{center} + h/2 - growth\}$
$daughter1 = \{x_{center}, y_{center}, z_{center} - h/2 + growth\}$
$daughter2 = \{x_{center}, y_{center}, z_{center} + h/2 + growth\}$
Mark every site in *sites* as extracellular
`buildCapsule(`*mother1, mother2, r, cytoplasm*`)`
`buildCapsule(`*daughter1, daughter2, r, cytoplasm*`)`
`buildMembrane(`*cytoplasm, extracellular*`)`
`/* Determine nuclei dimensions. Cell width and length are scaled by nucWidthFraction and nucLengthFraction, respectively. */`
$nr = r * nucWidthFraction$ `// Nucleus radius`
$nh = h - (1 - nucLengthFraction) * length/2$ `// Nucleus cylinder height`
$mother1 = \{x_{center}, y_{center}, z_{center} - nh/2 - growth\}$
$mother2 = \{x_{center}, y_{center}, z_{center} + nh/2 - growth\}$
$daughter1 = \{x_{center}, y_{center}, z_{center} - nh/2 + growth\}$
$daughter2 = \{x_{center}, y_{center}, z_{center} + nh/2 + growth\}$
`buildCapsule(`*mother1, mother2, nr, nucleus*`)`
`buildCapsule(`*daughter1, daughter2, nr, nucleus*`)`

Algorithm 4: buildCapsule

```
def buildCapsule(c1, c2, r, type): /* Construct capsule by building two spheres of radius
 r, centered at points c1 and c2, and a cylinder of radius r from c1 to c2.        */
    for s ∈ sites do
        if distance(s, c1) < r^2 or distance(s, c2) < r^2 or (s_x^2 + s_y^2 < r^2 and c1_z < s_z < c2_z) then
            s_type = type
        end
    end
```

Algorithm 5: buildMembrane

```
def buildMembrane(inside,outside): /* Locate boundary sites of an inside type that are
 touching a site marked as an outside type.  Convert these boundary sites to membrane
 sites.                                                                            */
    for s ∈ sites do
        if s_type = inside then
            N ⟵ all sites adjacent to s
            if ∃n ∈ N : n_type = outside then
                s_type = membrane
            end
        end
    end
```

S6 Supplementary figures

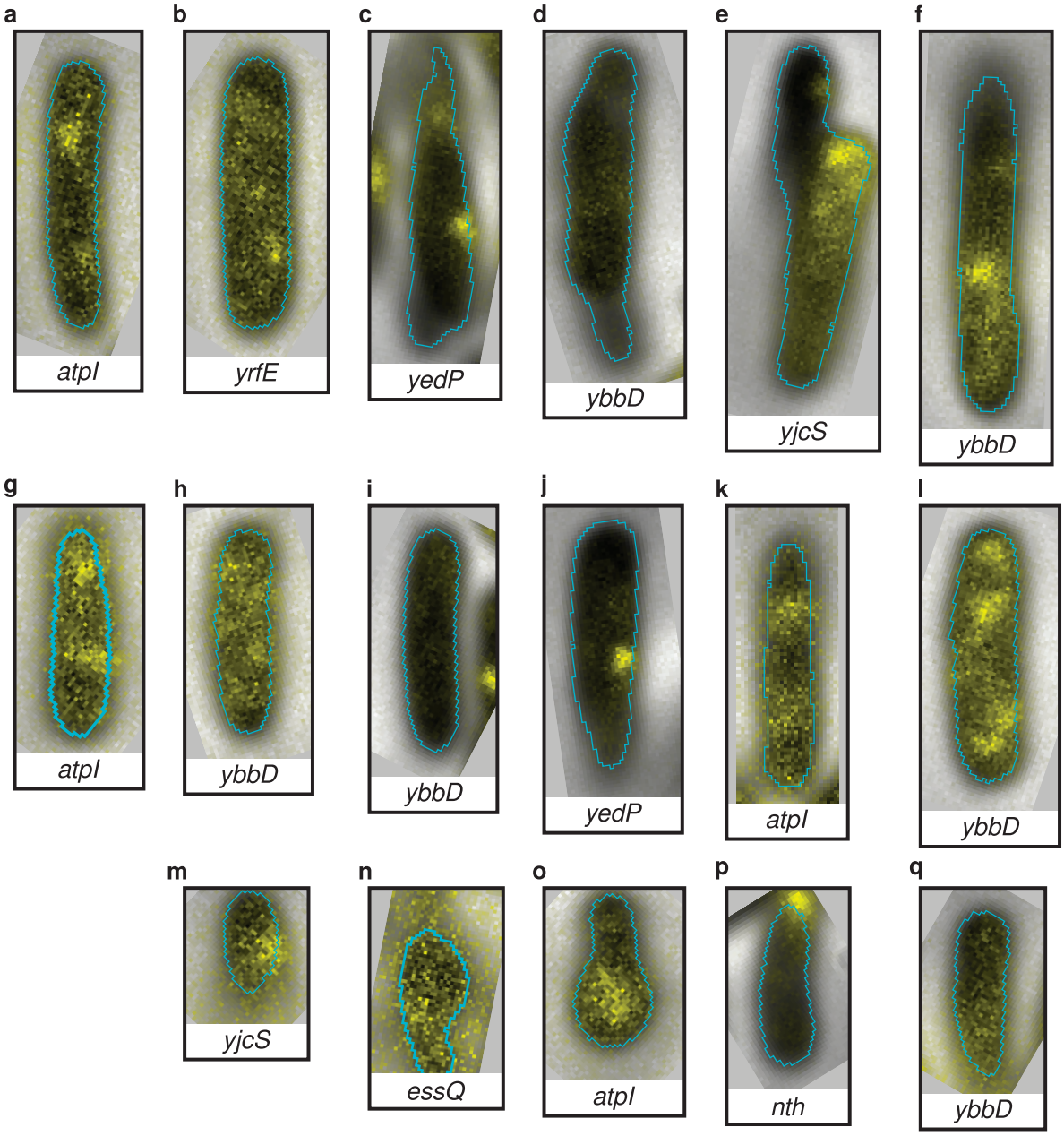

FIGURE S1 Examples of rejected regions. (**f**, **j**, **q**) Cell partially out of focal plane. (**d**, **i**, **p**) No fluorescence. (**a**, **b**, **e**, **f**, **g**, **h**, **k**, **l**, **m**) Poor dynamic range in fluorescence channel. (**c**, **d**, **e**, **m**, **n**, **o**) Bad morphology of thresholded brightfield image. (**n**) Debris on surface.

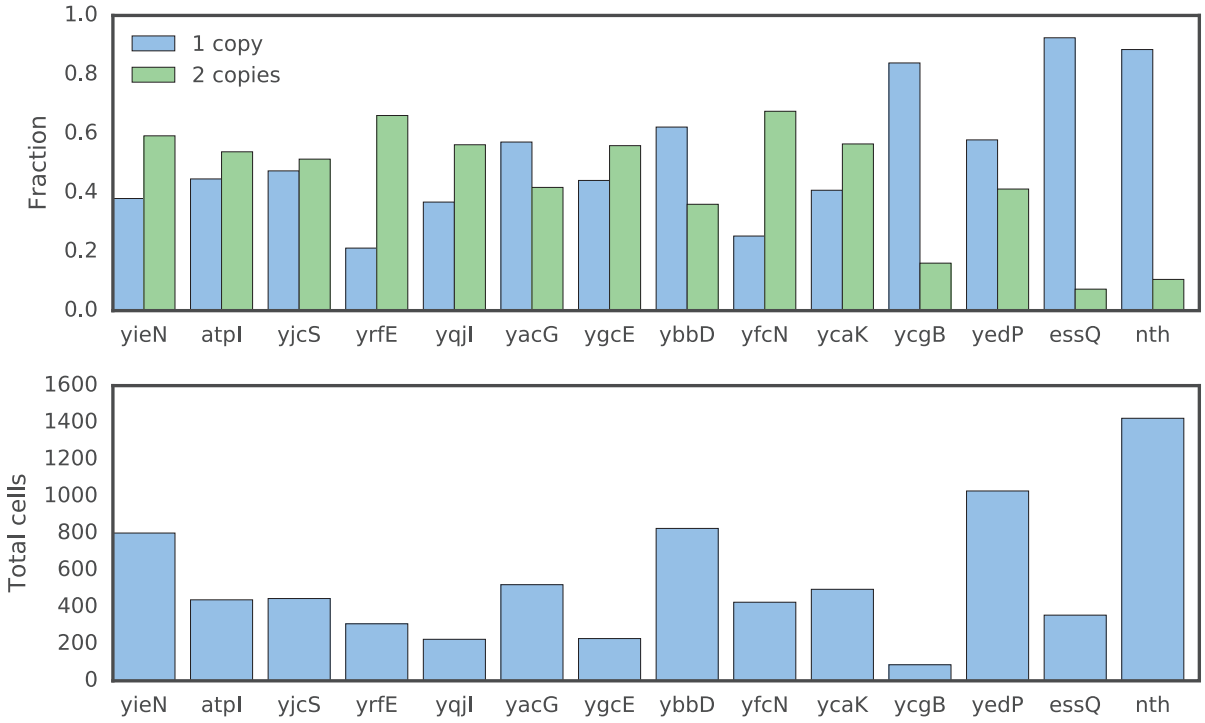

FIGURE S2 Fraction of cells observed with one or two operon copies and total count of cells with one or two copies.

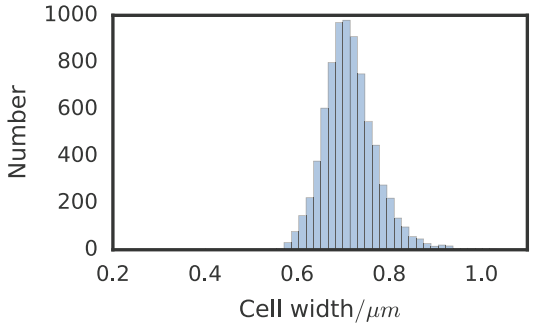

FIGURE S3 Distribution of cell widths. The cells were measured to have a mean of 0.715 μm and standard deviation of 0.059 μm.

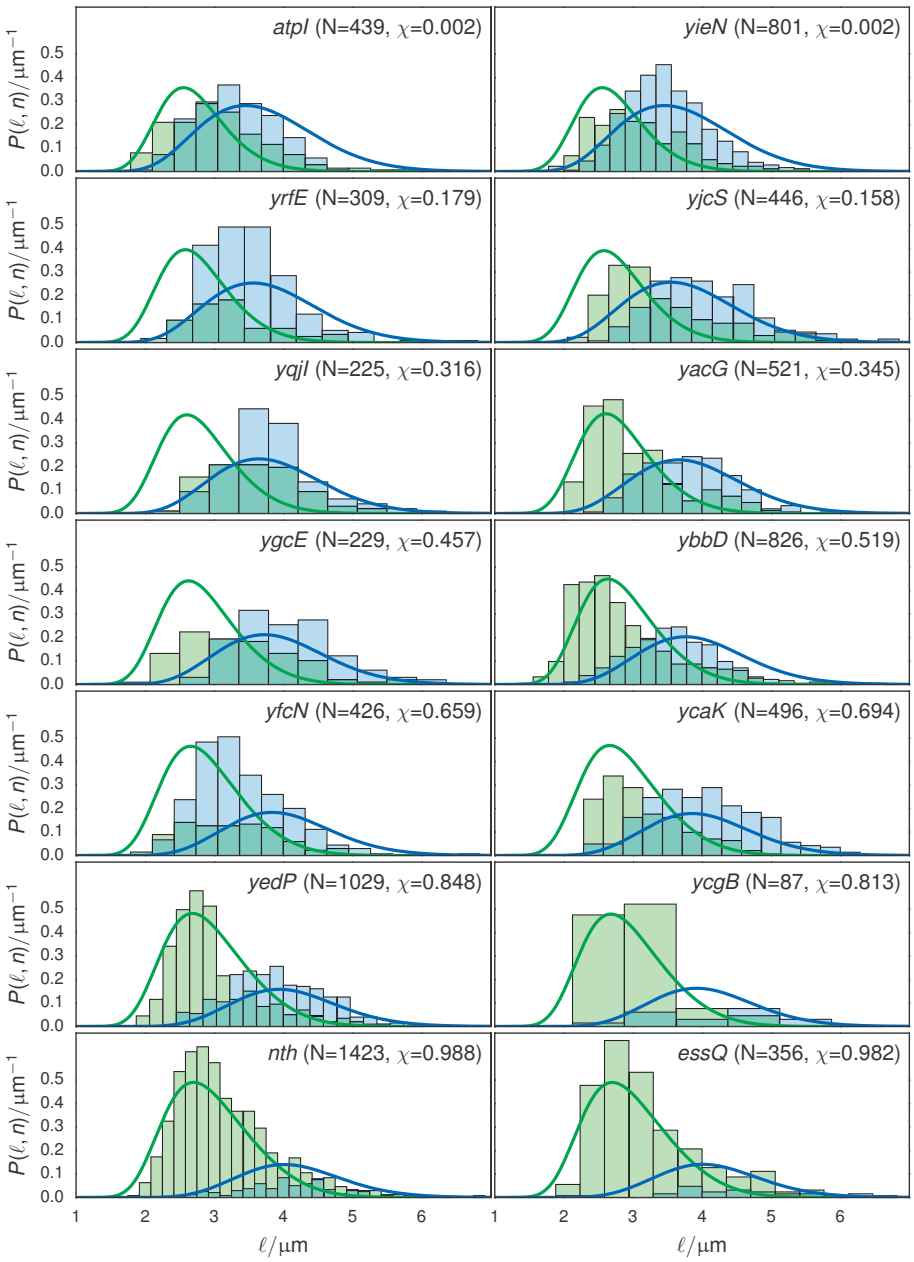

FIGURE S4 Comparison of experimental length distributions to fitting of growth/replication model. The number of cells binned and fractional position of the gene along its replichore are specified in each subplot as N and χ respectively. Blue denotes cells with two copies, green with one copy.

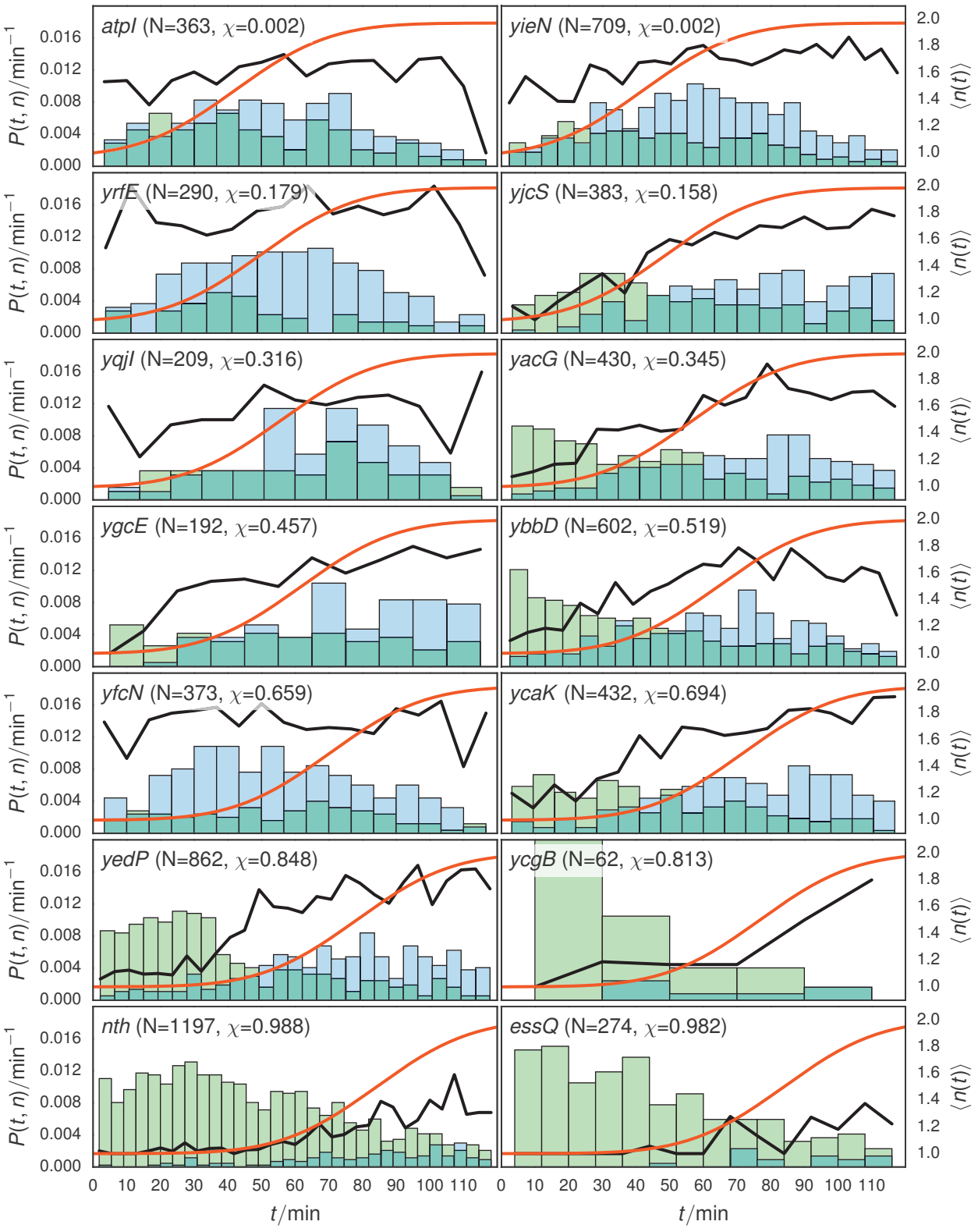

FIGURE S5 Comparison of inferred cell age distributions to fitting of growth/replication model. The number of cells binned and fractional position of the gene along its replichore are specified in each subplot as N and χ respectively. Blue denotes cells with two copies, green with one copy, the orange and black curves are the predicted and observed average copy numbers respectively.

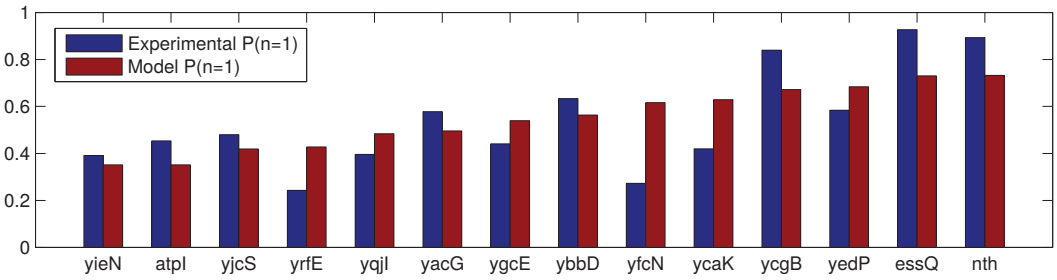

FIGURE S6 Fractions of cells with a single gene copy. Blue bars indicate the fraction of experimental cells observed with a single copy, while red bars indicate the fraction predicted by minimizing Θ in Equation S21 (while requiring $20.2 < \sigma_{t_r} < 24.0$).

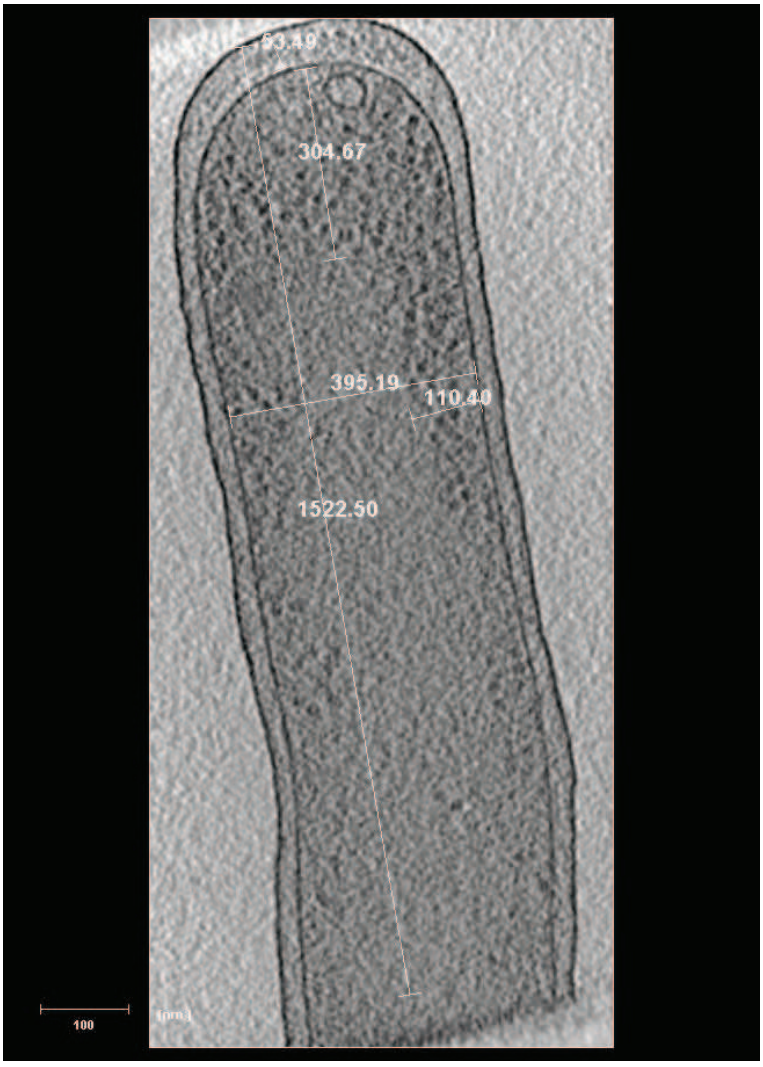

FIGURE S7 Cryo-electron tomogram of slow-growing *E. coli*[6] used to measure the nucleoid geometry for the whole-cell simulations. Units are in nanometers.

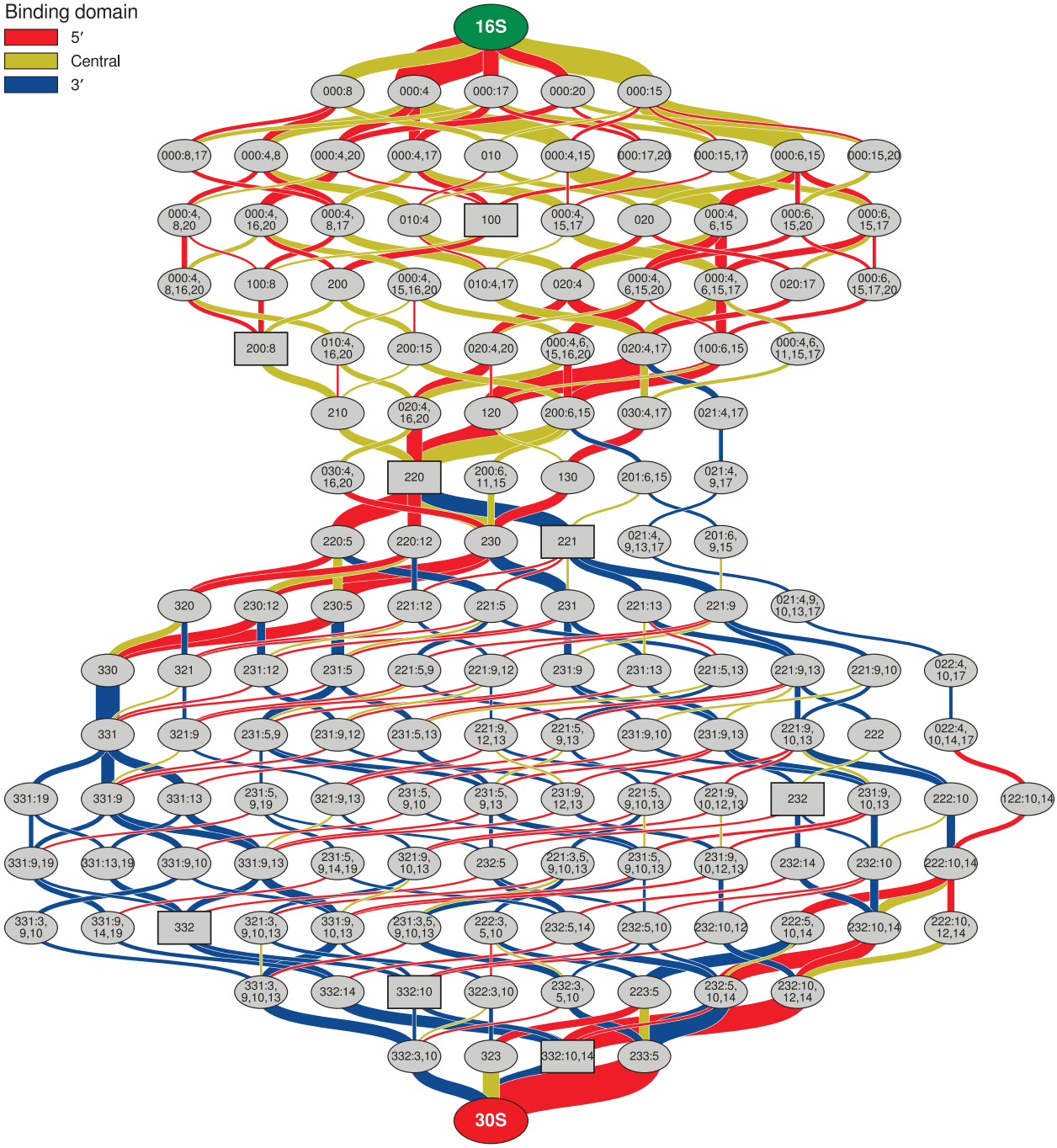

FIGURE S8 Reduced assembly network for SSU assembly at 40 °C. There are 147 SSU intermediates used in the biogenesis model, excluding the completed SSU and bare 16S. Each node is an assembly intermediate, labeled according to which proteins are bound. A three digit number describes the set of r-proteins bound to each domain (5′-, central-, and 3′- respectively), All remaining r-proteins are listed after the three digit number. The edges connecting the intermediates represent the r-protein binding reactions. The width represents the total amount of intermediate converted by that reaction, and the color indicates the binding domain of that protein (5′-red, central-yellow, and 5′-blue.) Predicted assembly intermediates from pulse/chase qMS and cryoEM[7] are represented using rectangles.

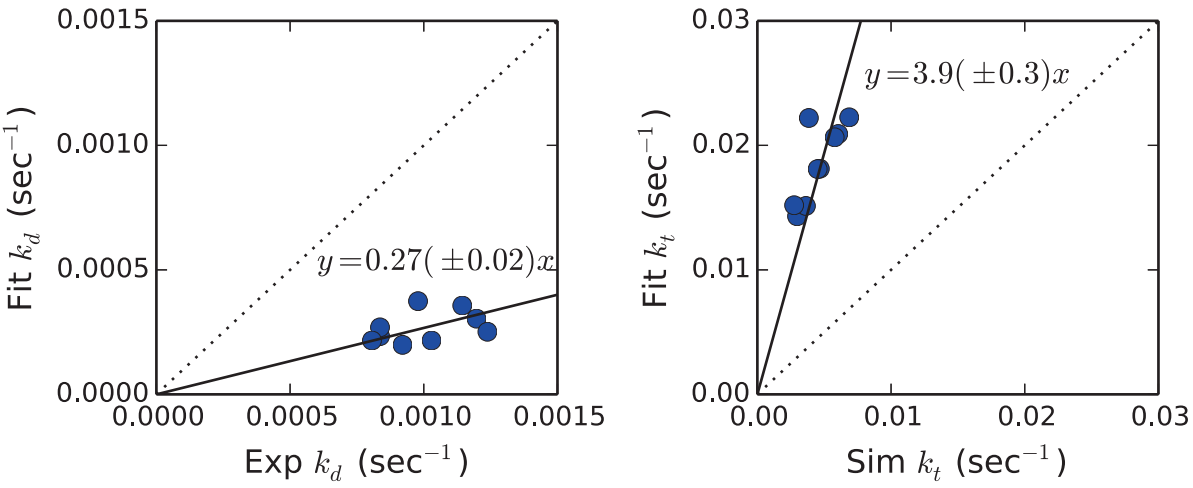

FIGURE S9 Comparison of fitted k_d and k_t estimated by minimizing the mean square deviation between the distributions calculated in the RBM and those predicted by Peterson et al.[3]. The dotted lines indicate $y = x$. The solid line indicates the linear trend between the fitted k_t and the value used in the RBM.

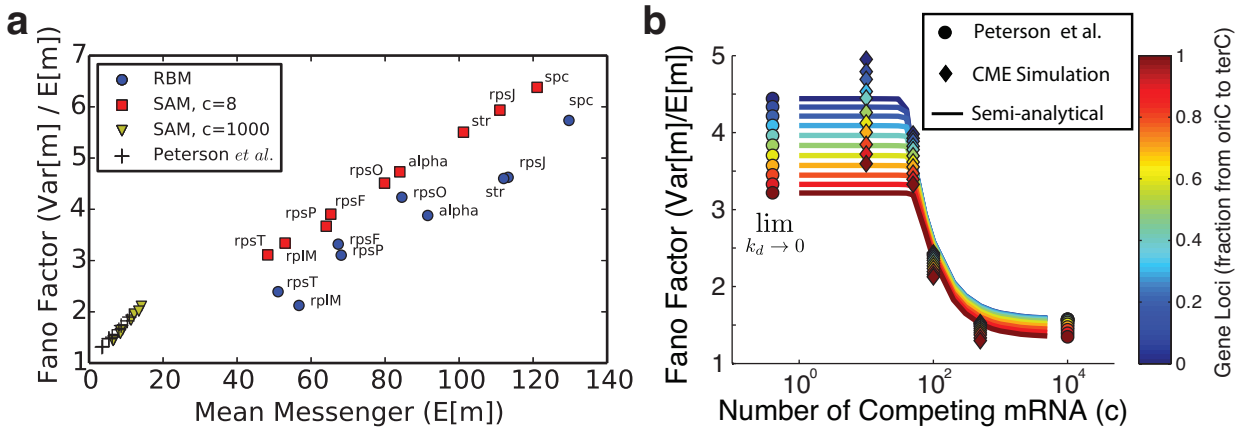

FIGURE S10 (a) Fano factor plotted versus mean messenger for ribosomal protein operon messengers from the biogenesis simulations. Statistics from ribosome biogenesis simulations (blue dots) are poorly represented by the analytical theory of Peterson et al.[3] (plus symbols). This prompts a new theory (SAM) that includes the effect of mRNA sequestration by unbound ribosomes (red squares) which better captures the simulated data. Additionally, it is shown that in the limit where the number of competing mRNAs is large, the SAM theory converges nearly to the theory of Peterson et al. (yellow triangles). (b) Fano factors for an average ribosomal protein mRNA as a function of gene loci and number of competing genes. Color indicates fraction along genome from *oriC* to *terC*. Circles indicate predictions from theory[3] either without modification (right-most) or in the limit where $k_d \to 0$ (left-most). Diamonds indicate the results of CME simulations of a models that explicitly account for gene duplication, cell division, interactions with ribosomes, and varying numbers of genes being expressed. Lines indicate the semi-analytical model developed herein.

References

1. Skinner SO, Xu H, Nagarkar-Jaiswal S, Freire PR, Zwaka TP, and Golding I (2016) "Single-cell analysis of transcription kinetics across the cell cycle". *eLife*, 5. doi:10.7554/elife.12175.

2. Earnest TM, Lai J, Chen K, Hallock MJ, Williamson JR, and Luthey-Schulten Z (2015) "Toward a whole-cell model of ribosome biogenesis: Kinetic modeling of SSU assembly". *Biophys. J.*, 109(6): 1117–1135. doi:10.1016/j.bpj.2015.07.030.
3. Peterson JR, Cole JA, Fei J, Ha T, and Luthey-Schulten ZA (2015) "Effects of DNA replication on mRNA noise". *Proc. Natl. Acad. Sci. USA*, 112(52): 15886–15891. doi:10.1073/pnas.1516246112.
4. Peterson JR, Hallock MJ, Cole JA, and Luthey-Schulten ZA (2013) "A problem solving environment for stochastic biological simulations". In "PyHPC 2013", Supercomputing 2013.
5. Llopis PM, Sliusarenko O, Heinritz J, and Jacobs-Wagner C (2012) "In vivo biochemistry in bacterial cells using FRAP: Insight into the translation cycle". 103(9): 1848–1859. doi:10.1016/j.bpj.2012.09.035.
6. Roberts E, Magis A, Ortiz JO, Baumeister W, and Luthey-Schulten Z (2011) "Noise contributions in an inducible genetic switch: A whole-cell simulation study". *PLoS Comput. Biol.*, 7(3): e1002010. doi:10.1371/journal.pcbi.1002010.
7. Mulder AM, Yoshioka C, Beck AH, Bunner AE, Milligan RA, Potter CS, Carragher B, and Williamson JR (2010) "Visualizing ribosome biogenesis: Parallel assembly pathways for the 30s subunit". *Science*, 330(6004): 673–677. doi:10.1126/science.1193220.
8. Powell EO (1956) "Growth rate and generation time of bacteria, with special reference to continuous culture". *J. Gen. Microbiol.*, 15(3): 492–511.