Effects of DNA replication on mRNA noise

Joseph R. Peterson[a,1], John A. Cole[b,1], Jingyi Fei[c,d], Taekjip Ha[a,b,c,d,e,f], and Zaida A. Luthey-Schulten[a,b,c,d,2]

[a]Department of Chemistry, University of Illinois at Urbana–Champaign, Urbana, IL 61801; [b]Department of Physics, University of Illinois at Urbana–Champaign, Urbana, IL 61801; [c]Carl Woese Institute for Genomic Biology, University of Illinois at Urbana–Champaign, Urbana, IL 61801; [d]Center for the Physics of Living Cells, University of Illinois at Urbana–Champaign, Urbana, IL 61801; [e]Howard Hughes Medical Institute, University of Illinois at Urbana–Champaign, Urbana, IL 61801; and [f]Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, Baltimore, MD 21205

Edited by José N. Onuchic, Rice University, Houston, TX, and approved November 11, 2015 (received for review August 14, 2015)

There are several sources of fluctuations in gene expression. Here we study the effects of time-dependent DNA replication, itself a tightly controlled process, on noise in mRNA levels. Stochastic simulations of constitutive and regulated gene expression are used to analyze the time-averaged mean and variation in each case. The simulations demonstrate that to capture mRNA distributions correctly, chromosome replication must be realistically modeled. Slow relaxation of mRNA from the low copy number steady state before gene replication to the high steady state after replication is set by the transcript's half-life and contributes significantly to the shape of the mRNA distribution. Consequently both the intrinsic kinetics and the gene location play an important role in accounting for the mRNA average and variance. Exact analytic expressions for moments of the mRNA distributions that depend on the DNA copy number, gene location, cell doubling time, and the rates of transcription and degradation are derived for the case of constitutive expression and subsequently extended to provide approximate corrections for regulated expression and RNA polymerase variability. Comparisons of the simulated models and analytical expressions to experimentally measured mRNA distributions show that they better capture the physics of the system than previous theories.

stochastic gene expression | stochastic simulation | analytical solutions | chromosome replication | master equation

Every step in the process of gene expression includes some inherent randomness. This may stem from the intrinsically stochastic nature of chemical reactions, chance differences between cells in the numbers of available reactants, intracellular crowding, or any of a number of other sources of biological variability (1–5). All told, noisy gene expression has profound effects on cellular behavior at both the individual and population levels, enabling switching between phenotypes by individual cells (6–12) as well as the potential for entire populations to divergently adapt to multiple niches within their environment (13). As a result, a great deal of work over the last decade has focused on understanding and quantifying the various sources of biological stochasticity.

In a series of now-classic papers, theorists and experimentalists alike have shown that the equations governing stochastic gene expression elicit steady-state distributions of proteins and mRNA in good agreement with observations (6, 14–19). Many of these works have also considered forms of transcriptional regulation wherein a gene can switch between active and inactive transcriptional states [either through the binding of a transcription factor (8, 13, 14, 19, 20) or through structural changes to the DNA that may occlude transcription start sites (21, 22)]. More recently, researchers have begun to venture beyond the steady-state approximation to address sources of noise that are tied to cell cycle-dependent processes. By considering mixtures of steady-state mRNA distributions associated with one and two copies of the DNA, Jones et al. (23) was the first, to our knowledge, to show that the duplication of a gene during replication can directly contribute to the observable noise in mRNA copy number. They used these results to partition experimentally observed mRNA noise into contributions associated with gene duplication, variability in RNA polymerase copy numbers, and experimental error (23).

In this paper we perform stochastic simulations, exactly sampling chemical master equations (CME) that explicitly account for chromosome replication, in order to study how gene duplication contributes to variability in mRNA expression. We find that our simulated results differ consistently and often significantly from the predictions of Jones et al. (23). We show that after gene duplication, a cell's mRNA count relaxes slowly from a low state (associated with the initial gene copy number) to a high state (associated with the copy number after replication) at a rate proportional to the mRNA half-life, a transition that can take several minutes and account for a significant portion of the overall cell cycle (Fig. 1). This seemingly minor effect can lead to divergence between the predicted and simulated mRNA Fano factors (a measure of the "noisiness" of the transcribed mRNA and equal to the variance over the mean) of 20% to greater than 80%, depending on the cell doubling time, the location of the gene on the chromosome, and the mRNA degradation rate. Such errors can easily lead to misattribution of observed mRNA variability to spurious sources and cloud the interpretation of experimental results.

Our findings motivated a time-dependent analytical treatment of the noise contribution originating from gene duplication, as well as several corrections to account for transcriptional regulation and variability in RNA polymerase (RNAP) and transcription factor copy numbers. The expressions are nearly exact for the case of constitutive transcription, even when including RNAP noise, and show extremely good agreement with both simulations and experiments when accounting for regulation. These results demonstrate that the explicit treatment of gene

Significance

Gene expression noise affects a cell's biological state and contributes to such phenomena as phenotype switching and cell fate determination. By examining chromosome replication—which is tightly controlled and thus exhibits little noise—we show that variability in mRNA levels across a population is less than previously expected. This noise is due to the transient relaxation of the mRNA from a low- to a high-copy steady state after a gene replication event. Gene location, mRNA degradation rate, and cell doubling time all contribute to observed noise. Our results demonstrate that it is essential to account for gene replication when modelling gene expression or when interpreting experimental results.

Author contributions: J.R.P., J.A.C., J.F., T.H., and Z.L.S. designed research; J.R.P., J.A.C., and J.F. performed research; J.R.P., J.A.C., J.F., and Z.L.S. analyzed data; and J.R.P., J.A.C., and Z.L.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]J.R.P. and J.A.C. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: zan@illinois.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1516246112/-/DCSupplemental.

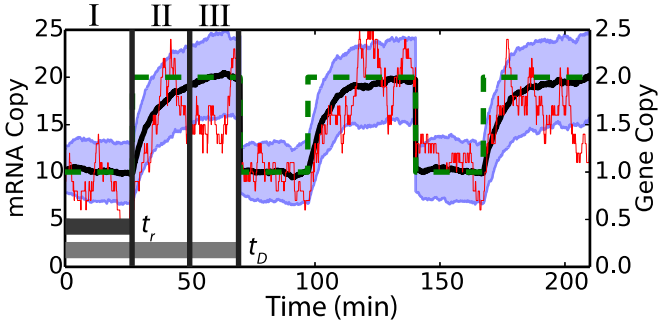


Fig. 1. Simulation schematic composed of 200 simulation replicates shows the progress of the average mRNA count (black line) before and after a gene duplication event (traced by a green dashed line). The area encompassing the average $\pm 1\sigma$ (blue) is shown along with a representative simulation trace (red). Gene duplication is followed by a transient period where the mRNA relaxes from an initially low to a high count at a rate proportional to the degradation rate of the mRNA. Three regions exist and are delineated by vertical lines: a preduplication state (I), wherein the mRNA is in a low copy number steady state, and a relaxation period just after duplication (II), where the mRNA relaxes up to a new equilibrium steady state (III). In these simulations the doubling time (t_D) was taken to be 70 min, the total DNA replication time was taken to be 45 min, the gene was positioned 55% of the way from the origin to the terminus ($t_r \sim 27$ min), the transcription rate k_t was 1.26 molecules/min, and the degradation rate k_d was 0.126/min.

replication and careful accounting for the subsequent product copy number relaxation time are necessary for accurately describing mRNA—and in turn protein—variability.

Results

Explicit Simulation of Gene Duplication for Constitutive Expression. Stochastic simulations of gene expression were used to determine the effect of chromosome replication on mRNA noise. A constitutive model of gene expression (Eq. **1**) wherein mRNA is transcribed from its gene at rate k_t and degraded at rate k_d is considered first, as the majority of genes are under no regulatory control under physiological conditions:

$$D \xrightarrow{k_t} D + m$$
$$m \xrightarrow{k_d} \emptyset. \qquad [1]$$

Simulations were performed using Gillespie's stochastic simulation algorithm (SSA) (24) as implemented in our Lattice Microbe software (25). For each simulation replicate, the mRNA copy number was tracked within a single lineage spanning 10 full cell cycles. Starting from a defined initial state, the copy number was allowed to evolve until t_r (the replication time for the gene) at which time the gene copy number was doubled to model the effect of replication. The simulations then continued until t_D (the division time) at which time the intracellular components were halved to account for cell division, and the next generation in the lineage begins. An example of this process is shown in Fig. 1. The cell cycle length, replication time, and transcription rate were all varied, but the mRNA degradation rate was held fixed at 0.126 min^{-1} to maintain the average mRNA half-life in *Escherichia coli* of 5.5 min (26). Two different cell doubling times were studied—70 min and 40 min. Genome replication in *E. coli* requires $\sim$45 min and is relatively insensitive to changes in growth rate or culturing conditions (27–29). As such, cells doubling in less than this amount of time must maintain multiple chromosome replication forks at different stages of completion. This means that depending on their location along the genome, some genes in our fast-growing cells ($t_D = 40$ min) exist with either two or four copies [we ignore the short-lived three-copy state that arises when one replication fork briefly outpaces the other (30)] whereas others exist with either one or two copies (*SI Appendix*, Fig. S1*B* and Table S5*C*). In the slow-growing cells ($t_D = 70$ min) either one or two copies of all genes exist. For both doubling times, simulations were performed across a series of replication times (t_r) corresponding to genes located across the genome (spanning from the origin of replication to the replication terminus in 5% increments).

Our simulations show that after gene duplication the mean mRNA count relaxes to twice its prior value on a timescale that is set by the mRNA's half-life. This, it turns out, can constitute a significant portion of the cell cycle (in *E. coli* the average mRNA half-life is $\sim$14% of a 40-min cell cycle, and the total relaxation takes about 40% of the cell cycle) and significantly impact the statistics of observable mRNA copy numbers (*SI Appendix*, Fig. S2).

Analytical Time-Dependent mRNA Statistics for Constitutive Expression. We derived expressions for the mean, variance, and Fano factor of an mRNA being constitutively expressed from a gene that is duplicated during the cell cycle (a detailed description can be found in *SI Appendix*, section 1.1). This work hinged on the fact that the mean mRNA copy number, $\langle m \rangle$, over an ensemble of cells can be written as a time average over the instantaneous mean copy number, $\overline{m}(t)$, and likewise, the variance, $\text{Var}[m]$, can be written in terms of a time average over the instantaneous variance, $\sigma_m^2(t)$, and the square of the mean copy number (*SI Appendix*, Eqs. **S23** and **S25**). Differential equations for the instantaneous mean and variance were derived from the chemical master equation (*SI Appendix*, Eqs. **S1**–**S20**) and solved to yield

$$\sigma_m^2(t) = \overline{m}(t) = \begin{cases} \dfrac{k_t}{k_d} & 0 < t < t_r \\ \dfrac{k_t}{k_d}\left(2 - e^{k_d(t_r - t)}\right) & t_r < t < t_D. \end{cases} \qquad [2]$$

Interestingly, the mRNA remains Poisson distributed after gene duplication as it relaxes to its new steady state [i.e., the probability of measuring m mRNA in a cell at time t after the start of its cell cycle can be written $P(m|t) = \text{Pois}(\overline{m}(t))$]. Time averaging over the cell cycle yielded the expressions

$$\langle m \rangle = \langle m \rangle_1 \left[1 + f + \frac{e^{-fk_d t_D} - 1}{k_d t_D}\right] \qquad [3]$$

$$\text{Var}[m] = \langle m \rangle - \langle m \rangle^2 + \langle m \rangle_1^2 \left[1 + 3f + \frac{8e^{-fk_d t_D} - e^{-2fk_d t_D} - 7}{2k_d t_D}\right] \qquad [4]$$

$$\text{Fano}[m] = 1 - \langle m \rangle + \frac{\langle m \rangle_1^2}{\langle m \rangle}\left[1 + 3f + \frac{8e^{-fk_d t_D} - e^{-2fk_d t_D} - 7}{2k_d t_D}\right], \qquad [5]$$

where $\langle m \rangle_1 = k_t/k_d$ represents the mean mRNA copy number before gene duplication, and $f = (t_D - t_r)/t_D$ represents the fraction of the cell cycle after the gene duplication event. Although these results were derived assuming that the ages of cells in a population are uniformly distributed, log-phase populations are in fact known to have exponentially distributed ages (29, 31). This can be easily accounted for analytically (*SI Appendix*, Eqs. **S22**–**S29**), but it amounts to a fairly small correction (< 10%; *SI Appendix*, Fig. S3) and significantly complicates the expressions. It is worth noting that in the limit where the mRNA degradation rate, k_d, becomes large, relaxation after gene duplication becomes instantaneous, and our "time-dependent" (TD) theory reduces to the "time-independent" (TI) theory of Jones et al. (23). In the limit of slow mRNA degradation, the mRNA

distribution never relaxes to the high state, and cells remain in the low copy number state until division.

Comparison of Eqs. **3** and **5** with simulations demonstrates the accuracy of the time-dependent theory (Fig. 2 *A* and *C* and *SI Appendix*, Figs. S4 and S5). For both doubling times our expressions for $\langle m \rangle$ and the Fano factor prove nearly exact, whereas the time-independent theory tends to overestimate both values. Comparing the shape of the mRNA distributions proves equally impressive. Numerically time averaging $P(m|t)$ yields distributions that strongly agree with histograms of our simulated mRNA counts (Fig. 2 *B* and *D* and *SI Appendix*, Figs. S6 and S7, orange lines). To quantify the agreement of the time-dependent and time-independent models, we computed the Kullback–Leibler divergence (*SI Appendix*, Eq. **S68**) between simulated and theoretical distributions. The divergence from our simulated distributions is ∼10-fold smaller when using the time-dependent theory, but we note that this improvement breaks down within a narrow range of gene loci in the fast-growing cells (*SI Appendix*, Fig. S8). This disagreement occurs among genes located between about 50% and 70% of the way from the origin to the terminus and is due to the fact that these genes are duplicated very late in the 40-min cell cycle. The associated mRNA counts have insufficient time to relax to their postduplication steady states, and upon division, they drop well below their preduplication steady state (*SI Appendix*, Fig. S9*A*). As a result, the dynamics of these mRNAs are better modeled assuming both early and late relaxations (*SI Appendix*, section 1.2, Eqs. **S37** and **S38** and Fig. S9 *B* and *C*).

Extending our comparisons to experimental data proves similarly fruitful. Theoretical distributions computed using measured k_d, f, and t_D were compared with 26 previously reported experimental data sets (23, 32). A few representative distributions for genes with different values of f, k_d, and $\langle m \rangle$ are shown in Fig. 3. The time-dependent theory outperforms the time-independent theory in all cases, clearly demonstrating its utility (*SI Appendix*, section 1.9 and Figs. S10 and S11). We note, however, that neither theory performs well when fitting mRNA distributions for strongly regulated genes. Fig. 3*D* shows the distribution of *ptsG* mRNA counts in single *E. coli* cells obtained via superresolution imaging and modeling (32). This gene is known to be regulated via transcription factors and small RNA (32–34). The orange and blue lines in Fig. 3*D* show theoretical distributions computed according to the time-dependent and time-independent treatments. Both curves are underdispersed, indicating the need for a model that directly accounts for transcriptional regulation

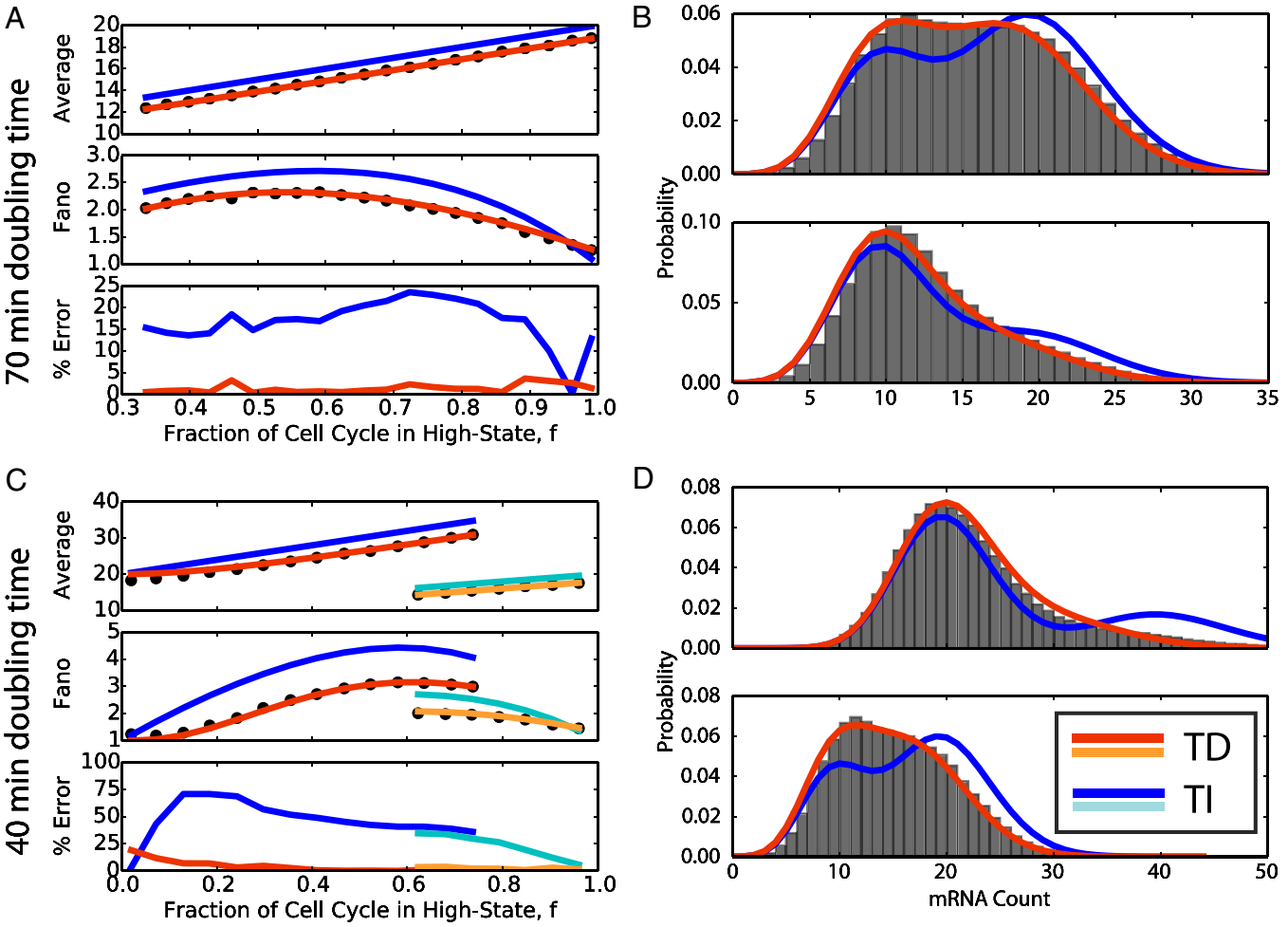


Fig. 2. Comparison of the time-dependent (TD) and time-independent (TI) theories for constitutively expressed genes. (*A*) The mean, Fano factor, and relative error in the Fano factor for slow-growing cells (70 min doubling time). Black circles represent the results of 200 simulated replicates, and orange and blue lines represent the TD and TI theories, respectively. (*B*) Comparison of mRNA distributions for slow-growing cells. The gray histogram represents the results of simulations, and the orange and blue lines again represent the TD and TI theories. *Upper* distribution is that of a gene copied half-way through the cell cycle ($f = 0.5$) and *Lower* distribution is that of a gene copied at the beginning of the cell cycle ($f \sim 1.0$). (*C* and *D*) Statistics (mean, Fano factor, and relative error) and distributions for fast-growing cells (40-min doubling time). Genes can exist in either two or four copies (dark orange and blue) or one or two copies (light orange and blue), depending on their location along the chromosome. In all cases, the time-dependent theory better captures simulation data.

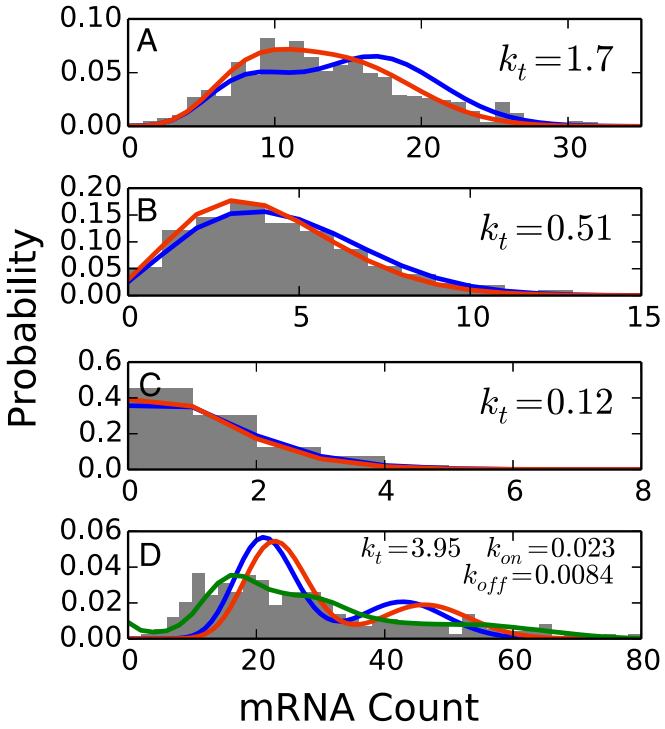


Fig. 3. Comparison with experiments. (*A–D*) A comparison of predicted distributions computed assuming constitutively expressed genes with (orange lines) and without (blue lines) accounting for the time dependence of the mRNA relaxation to experimental data for (*A–C*) various *lac* promoter mutants (23) and (*D*) *ptsG* (32). The *lac* mRNA has a half-life of 5.5 min and spends two-thirds of the cell cycle after gene replication, whereas the *ptsG* transcript has a half-life of 2.8 min and spends about one-third of the cell cycle after gene replication. A comparison to the regulated model shows much better agreement for *ptsG* (green line). All rates are given in units min^{-1}. See *SI Appendix*, section 1.9 and Figs. S11–S13 for further comparisons and details.

(Fig. 3*D*, green line; discussed in the next section and *SI Appendix*, section 1.9 and Figs. S12–S14).

Corrections to the Analytical Model for Regulation, as Well as RNAP and Transcription Factor Variability. Several corrections to our time-dependent analytical model were derived to account for other sources of noise. The first and most important is a correction that approximates the noise stemming from transcriptional regulation (*SI Appendix*, section 1.3). A gene is modeled as being in either an "off" or an "on" state; the on state is capable of producing mRNA at rate k_t, whereas the off state is silenced. Genes are allowed to switch between states at rates k_{on} and k_{off}; schematically,

$$\begin{aligned} &D_{\text{off}} \underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftharpoons}} D_{\text{on}} \\ &D_{\text{on}} \xrightarrow{k_t} D_{\text{on}} + m \\ &m \xrightarrow{k_d} \emptyset. \end{aligned} \qquad [6]$$

In the limit where the transcriptional state switching is fast compared with the mRNA degradation rate (i.e., $k_{\text{on}}, k_{\text{off}} \gg k_d$, meaning the average number of on genes relaxes quickly to its new steady state after gene replication) and k_{on} is greater than or at least of similar order as k_{off}, the Fano factor can be approximated with the addition of a single term,

$$\text{Fano}[m] \approx 1 + \frac{k_t k_{\text{off}}}{(k_{\text{on}} + k_{\text{off}})(k_{\text{on}} + k_{\text{off}} + k_d)} - \langle m \rangle + \frac{\langle m \rangle_1^2}{\langle m \rangle}\left[1 + 3f + \frac{8e^{-fk_d t_D} - e^{-2fk_d t_D} - 7}{2k_d t_D}\right], \qquad [7]$$

where, again, $\langle m \rangle$ is given by Eq. **3**, but now $\langle m \rangle_1 = (k_{\text{on}}/(k_{\text{on}} + k_{\text{off}}))(k_t/k_d)$. It is important to note that this result is based on the assumption that the relaxations of the instantaneous mRNA mean ($\overline{m}(t)$) and variance ($\sigma_m^2(t)$) occur on a similar timescale. Our own simulations indicate that this approximation may not in general be true, but it drastically simplifies the analysis and keeps the resulting expressions for the mean, variance, and Fano factor tractable. Within appropriate parameter ranges, we find good agreement with simulation (*SI Appendix*, Fig. S14), but we note that when the gene switching rates are slow or significantly favor the off state (meaning the mRNA is especially "bursty"), Eq. **7** shows poorer agreement. As a result, further corrections were derived for cases in which $k_{\text{on}}, k_{\text{off}} \lesssim k_d$ (*SI Appendix*, section 1.4). This refined analysis treats the mean number of on genes as a dynamic variable after gene duplication (rather than assuming rapid relaxation) and yields somewhat unwieldy expressions for $\langle m \rangle$ and Var[m], which themselves depend on whether the regulation is controlled by a repressor- or activator-type transcription factor (*SI Appendix*, Eqs. **S54–S57**).

Because gene transcription depends on the activity of a number of proteins including RNAP and any of several transcription factors (TFs), variability in these proteins' copy numbers can naturally impact mRNA levels within the cell. We considered how our time-dependent theory's results change when the numbers of either RNAP or an activator-type TF were assumed to vary (leading to variation in the effective transcription and gene activation rates; *SI Appendix*, sections 1.5 and 1.6).

The Fano factor correction derived for RNAP-associated noise resulted in a simple additive term,

$$\text{Fano}[m] \approx \text{Fano}[m]_{\overline{k}_t} + \frac{\langle m \rangle_{1,\overline{k}_t}^2}{\langle m \rangle_{\overline{k}_t}} \frac{\text{Var}[k_t]}{\overline{k}_t^2}\left[1 + 3f + \frac{8e^{-fk_d t_D} - e^{-2fk_d t_D} - 7}{2k_d t_D}\right]. \qquad [8]$$

Here, k_t is assumed to be given by the product of the (random) RNAP copy number, R, and $k_{t,0}$, the specific transcription rate for a single RNAP. $\overline{k}_t$ then represents the mean value of k_t, $\langle m \rangle_{\overline{k}_t}$ and $\text{Fano}[m]_{\overline{k}_t}$ are the mean mRNA number and Fano factor evaluated according to Eqs. **3** and **5** assuming $k_t = \overline{k}_t$, and $\langle m \rangle_{1,\overline{k}_t} = \overline{k}_t/k_d$ is the mean mRNA count before gene duplication. If R is Γ distributed, then $\text{Var}[k_t] \approx k_{t,0}^2(\langle R \rangle/\beta)$, where β represents the "rate" parameter of the distribution. For an *E. coli* doubling in ~40 min, the mean RNAP copy number has been measured to be ~3,000 per cell (35), placing it well into the "extrinsic noise limit" [for which $\sigma^2/\mu^2 \approx 0.1$ (16)], implying that β can be approximated as 1/300. Inserting this into Eq. **8**, we find that the contribution to the Fano factor from RNAP copy number variability can be roughly approximated as $\langle m \rangle/10$, in accordance with ref. 23 (*SI Appendix*, Fig. S15).

In contrast, the correction derived for TF-associated noise resulted in a cumbersome expression, which, when evaluated across a range of k_{off} and $\overline{k}_{\text{on}}$ values (where $\overline{k}_{\text{on}}$ represents the mean value of k_{on}), tended to be relatively small. We found it approached $\langle m \rangle/10$ only when $k_{\text{off}} \gg \overline{k}_{\text{on}}$, and in cases where $k_{\text{off}} \lesssim \overline{k}_{\text{on}}$ we found this correction remained well below ~ 3% of $\langle m \rangle$. Importantly, these results indicate that TF-associated variability generally imparts less mRNA noise than does RNAP-associated variability.

The corrections for RNAP- and TF-associated noise resulted from the promotion of certain rates—k_t and k_{on}, respectively—to random variables and Taylor expanding about their means. Similar analyses can be performed for other potential sources of noise, including variability in t_r or t_D; in both cases, however, experiments show that the variance of these parameters is generally much less than 10% of their mean (30), and thus they are not likely to significantly impact measurable mRNA noise.

The analytic expressions derived here can be leveraged to greatly simplify the determination of kinetic parameters. The fitting of the *ptsG* mRNA distribution in Fig. 3*D* exemplifies this; it cannot be fitted without accounting for transcriptional regulation but this requires the simultaneous varying of k_{on}, k_{off}, and k_t. Eqs. **5**, **7**, and **8** can be used to solve for k_{on} and k_{off} as functions of k_t (*SI Appendix*, section 1.9), meaning that the fitting problem can be reduced to a simple 1D scan over possible values for the transcription rate (assuming fixed k_d and $\langle m \rangle$). This significantly simpler problem was then performed numerically and resulted in the values $k_t = 3.95$ min^{-1}, $k_{on} = 0.023$ min^{-1}, and $k_{off} = 0.0084$ min^{-1}, all of which are physiologically reasonable (36).

Discussion

We have computationally and analytically studied the effects of DNA replication on mRNA noise. By formulating the process in terms of a CME, we were able to determine the time-dependent mean, variance, and distribution of the mRNA as a function of its degradation and transcription rates, the cell's doubling time, and the gene's position on the chromosome. We have found that failure to account for the slow relaxation of the messenger distribution to its postduplication steady state results in overestimation of the associated noise. Importantly, this overestimation can have a profound impact on the interpretation of both experimental and theoretical results.

As a hypothetical example, consider a single-cell mRNA counting experiment in which the cell doubling time is measured to be 40 min and the mean count to be 10 messengers per cell each with the average degradation rate in *E. coli* of 0.126 min^{-1}, with the gene of interest being located roughly one-third of the way between the origin and the terminus of replication. Assume the Fano factor for the population was measured to be 3.25. These reasonable values can lead to very different interpretations of experiential data, depending on how gene duplication is treated. Before the study by Jones et al. (23), the entirety of the noise larger than 1 might have been attributed to transcriptional regulation and extrinsic factors like RNAP variability. In that case, after accounting for the RNAP noise contribution, it would have been concluded that the gene was quite strongly regulated (*SI Appendix*, Fig. S16, *Left Bar*). After ref. 23, exactly the opposite conclusion could have been reached—essentially all of the observed noise could be attributed to RNAP variability and gene duplication. It would have appeared that there was no evidence of transcriptional regulation (*SI Appendix*, Fig. S16, *Center Bar*). In fact, our analysis shows that both gene duplication and regulation contribute similar but modest amounts to the overall noise level (*SI Appendix*, Fig. S16, *Right Bar*). We note that this example is a special case, and in general the different models will likely not yield such starkly divergent interpretations, but it nevertheless illustrates why accurately resolving the different noise contributions requires the time-dependent model developed here.

The misattribution of noise is particularly problematic in the development of kinetic models and analysis of experiments. Countless articles have presented stochastic simulations of noise in complex genetic circuits, and many appear to show strong quantitative agreement with experiments, but to our knowledge almost none have included duplication of the genes involved. One early study that did consider gene replication concluded that mRNA relaxation contributed little to the overall noise, but this was based in part on an assumed mRNA half-life of 1 min—considerably shorter than the mean for bacteria (17) and in the regime where the corrections are predicted to be small. Returning to the hypothetical experiment described above, if a simple model of transcriptional regulation (such as Eq. **6**) that did not account for gene duplication were used to fit the data, a modeler could arrive at estimates of k_{on} and k_{off}, for example, that deviate from the correct value by as much as 100%.

Because gene duplication-associated mRNA noise scales proportionally with the (mean) messenger expression level, the potential for its misattribution is greatest among highly expressed genes. In *E. coli*, these include a number of genes involved in key cellular processes like translation (including those encoding the ribosomal proteins), ATP synthesis (including the ATP synthase genes), transcriptional regulation, and central metabolism (including the glycolytic genes *gapA* and *eno*) (16, 37). The potential for noise misattribution is also related to f (the fraction of the cell cycle after gene duplication) and the messenger decay rate, k_d. Fig. 4 shows the relative error between our time-dependent Fano factor expression and that of the time-independent theory [computed as $(F_{TI} - F_{TD})/F_{TI}$] for a cell doubling in 70 min. We see for highly expressed, long-lived transcripts the error can easily be >100% whereas even in moderate cases the error can be in the range 20 − 50% (most of this divergence comes from deviation of the TI model, as the TD model agrees well with

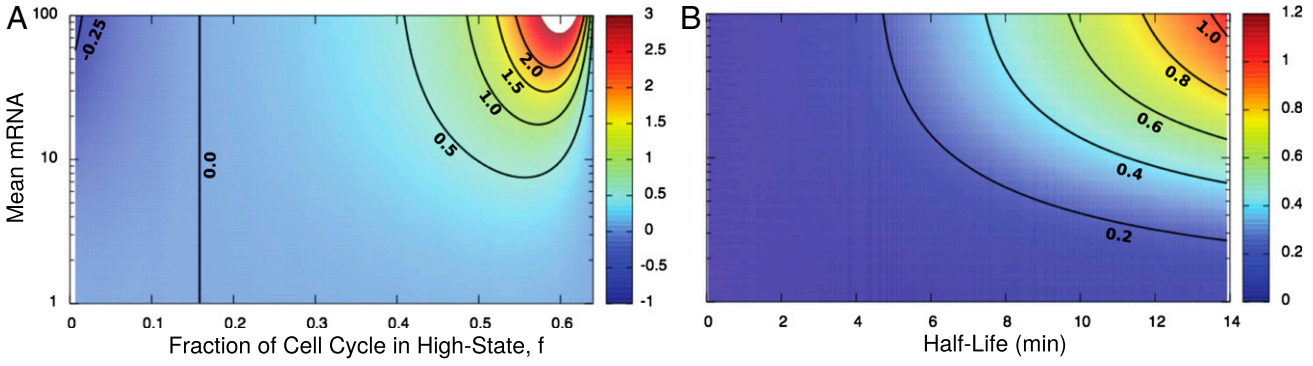


Fig. 4. (*A* and *B*) Deviation of the time-dependent from time-independent theory of the estimated Fano factor $((F_{TI} - F_{TD})/F_{TI})$ when neglecting time dependence of the mRNA relaxation as a function of (*A*) the mean mRNA count and fraction of cell cycle after gene replication and (*B*) mean mRNA and messenger half-life. Here a slow-growing cell was considered ($t_d \sim 70$ min). In *A* the mRNA half-life was the average in *E. coli* of 5.5 min. In *B* the fraction of the cell cycle after replication was taken to be 0.7. Scale bars indicate the value of the deviation. Contours are indicated with lines and the value along the contour is denoted.

simulation; *SI Appendix*, Fig. S17). Interestingly, because this error can change dramatically over a narrow range of values of f (i.e., $0.4 < f < 0.7$; Fig. 4*A*), and because f itself is a function of the cell's growth rate, small differences in cell doubling times can have a profound effect on the interpretation of mRNA noise. Taken altogether, these results indicate that the time dependence of gene duplication and mRNA relaxation should not be ignored when either modeling stochastic gene expression or analyzing experimental data.

Materials and Methods

Simulations were performed using the Gillespie stochastic simulation algorithm (24) as implemented in the Lattice Microbes software version 2.2 (25, 38). All simulations were performed using NVIDIA GPUs and analysis was written in Python, using the PyLM interface to Lattice Microbes version 1.0 (39). Input files can found in *SI Appendix*.

Both a constitutive model of gene expression and a two-state model of gene expression were considered (Eqs. **1** and **6**). Doubling times of 40 min or 70 min were examined and cell division was implemented by dividing the gene counts in half and binomially distributing the mRNA count between the cells with equal probability. The replication time (t_r) as well as the numbers of genes and replication forks at the start of the cell cycle are based on the theory of Cooper and Helmstetter (27). The DNA replication time was taken as 45 min, a value close to the average measured (28). When simulating regulation, the gene states were randomized at division time with probability to be active $P_{on} = k_{on}/(k_{on} + k_{off})$. The transcription rate constants k_t and k_d were varied as described in the main text. For each set of rate parameters, three technical replicate simulations were run, each of which included independent trajectories of 200 cell lineages growing for 10 generations.

ACKNOWLEDGMENTS. We thank Dr. Rob Phillips and his laboratory for supplying the experimental mRNA data from ref. 23. We acknowledge the Texas Advanced Computing Center at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper. This work is supported by the National Science Foundation (NSF) Graduate Research Fellowship Program under Grant DGE-1144245 (to J.R.P.), by the National Institutes of Health Grants GM112659 and 9 P41 GM104601-23, and by NSF (Center for the Physics of Living Cells) Grant PHY-1430124. T.H. is an investigator with the Howard Hughes Medical Institute. We acknowledge computing allocation (TG-MCA03S027) provided by Extreme Science and Engineering Discovery Environment, which is supported by NSF Grant ACI-1053575.

1. Thattai M, van Oudenaarden A (2001) Intrinsic noise in gene regulatory networks. *Proc Natl Acad Sci USA* 98(15):8614–8619.
2. Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. *Science* 297(5584):1183–1186.
3. Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A (2002) Regulation of noise in the expression of a single gene. *Nat Genet* 31(1):69–73.
4. Kaern M, Elston TC, Blake WJ, Collins JJ (2005) Stochasticity in gene expression: From theories to phenotypes. *Nat Rev Genet* 6(6):451–464.
5. Roberts E, Magis A, Ortiz JO, Baumeister W, Luthey-Schulten Z (2011) Noise contributions in an inducible genetic switch: A whole-cell simulation study. *PLoS Comput Biol* 7(3):e1002010.
6. Friedman N, Cai L, Xie XS (2006) Linking stochastic dynamics to population distribution: An analytical framework of gene expression. *Phys Rev Lett* 97(16):168302.
7. Schultz D, Ben Jacob E, Onuchic JN, Wolynes PG (2007) Molecular level stochastic model for competence cycles in Bacillus subtilis. *Proc Natl Acad Sci USA* 104(45): 17582–17587.
8. Choi PJ, Cai L, Frieda K, Xie XS (2008) A stochastic single-molecule event triggers phenotype switching of a bacterial cell. *Science* 322(5900):442–446.
9. Acar M, Mettetal JT, van Oudenaarden A (2008) Stochastic switching as a survival strategy in fluctuating environments. *Nat Genet* 40(4):471–475.
10. Assaf M, Roberts E, Luthey-Schulten Z, Goldenfeld N (2013) Extrinsic noise driven phenotype switching in a self-regulating gene. *Phys Rev Lett* 111(5):058102.
11. Lu M, Onuchic J, Ben-Jacob E (2014) Construction of an effective landscape for multistate genetic switches. *Phys Rev Lett* 113(7):078102.
12. Labhsetwar P, Cole JA, Roberts E, Price ND, Luthey-Schulten ZA (2013) Heterogeneity in protein expression induces metabolic variability in a modeled Escherichia coli population. *Proc Natl Acad Sci USA* 110(34):14006–14011.
13. Macneil LT, Walhout AJM (2011) Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Res* 21(5):645–657.
14. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S (2006) Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* 4(10):e309.
15. Shahrezaei V, Swain PS (2008) Analytical distributions for stochastic gene expression. *Proc Natl Acad Sci USA* 105(45):17256–17261.
16. Taniguchi Y, et al. (2010) Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* 329(5991):533–538.
17. Swain PS, Elowitz MB, Siggia ED (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Natl Acad Sci USA* 99(20):12795–12800.
18. So LH, et al. (2011) General properties of transcriptional time series in Escherichia coli. *Nat Genet* 43(6):554–560.
19. Peccoud J, Ycart B (1995) Markovian modeling of gene-product synthesis. *Theor Popul Biol* 48(2):222–234.
20. Golding I, Paulsson J, Zawilski SM, Cox EC (2005) Real-time kinetics of gene activity in individual bacteria. *Cell* 123(6):1025–1036.
21. Boeger H, Griesenbeck J, Kornberg RD (2008) Nucleosome retention and the stochastic nature of promoter chromatin remodeling for transcription. *Cell* 133(4): 716–726.
22. Earnest TM, Roberts E, Assaf M, Dahmen K, Luthey-Schulten Z (2013) DNA looping increases the range of bistability in a stochastic model of the lac genetic switch. *Phys Biol* 10(2):026002.
23. Jones DL, Brewster RC, Phillips R (2014) Promoter architecture dictates cell-to-cell variability in gene expression. *Science* 346(6216):1533–1536.
24. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* 81(25):2340–2361.
25. Roberts E, Stone JE, Luthey-Schulten Z (2013) Lattice Microbes: High-performance stochastic simulation method for the reaction-diffusion master equation. *J Comput Chem* 34(3):245–255.
26. Bernstein JA, Khodursky AB, Lin PH, Lin-Chao S, Cohen SN (2002) Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proc Natl Acad Sci USA* 99(15):9697–9702.
27. Cooper S, Helmstetter CE (1968) Chromosome replication and the division cycle of Escherichia coli B/r. *J Mol Biol* 31(3):519–540.
28. Kubitschek HE, Freedman ML (1971) Chromosome replication and the division cycle of Escherichia coli B-r. *J Bacteriol* 107(1):95–99.
29. Ho PY, Amir A (2015) Simultaneous regulation of cell size and chromosome replication in bacteria. *Front Microbiol* 6:662.
30. Skarstad K, Boye E, Steen HB (1986) Timing of initiation of chromosome replication in individual Escherichia coli cells. *EMBO J* 5(7):1711–1717.
31. Powell EO (1956) Growth rate and generation time of bacteria, with special reference to continuous culture. *J Gen Microbiol* 15(3):492–511.
32. Fei J, et al. (2015) Determination of in vivo target search kinetics of regulatory noncoding RNA. *Science* 347(6228):1371–1374.
33. Kimata K, Inada T, Tagami H, Aiba H (1998) A global repressor (Mlc) is involved in glucose induction of the ptsG gene encoding major glucose transporter in Escherichia coli. *Mol Microbiol* 29(6):1509–1519.
34. El Qaidi S, Plumbridge J (2008) Switching control of expression of ptsG from the Mlc regulon to the NagC regulon. *J Bacteriol* 190(13):4677–4686.
35. Klumpp S, Hwa T (2008) Growth-rate-dependent partitioning of RNA polymerases in bacteria. *Proc Natl Acad Sci USA* 105(51):20245–20250.
36. Vogel U, Jensen KF (1994) The RNA chain elongation rate in Escherichia coli depends on the growth rate. *J Bacteriol* 176(10):2807–2813.
37. Earnest TM, et al. (2015) Towards a whole-cell model of ribosome biogenesis: Kinetic modeling of SSU assembly. *Biophys J* 109(6):1117–1135.
38. Hallock MJ, Stone JE, Roberts E, Fry C, Luthey-Schulten Z (2014) Simulation of reaction diffusion processes over biologically relevant size and time scales using multi-GPU workstations. *Parallel Comput* 40(5-6):86–99.
39. Peterson JR, Hallock MJ, Cole JA, Luthey-Schulten Z (2013) A problem solving environment for stochastic biological simulations. *PyHPC '13: Proceedings of the 3rd Workshop on Python for High-Performance and Scientific Computing* (Association for Computing Machinery, New York).

Supporting Information

J. R. Peterson, J. A. Cole, J. Fei, T. Ha and Z. Luthey-Schulten
Effects of DNA Replication on mRNA Noise, 2015

1 SI Text

1.1 Derivation of the Fano Factor in the case of constitutive mRNA expression

We consider the case where a single gene is present within a cell from time 0 to t_r—the gene replication time—and thereafter until the cell divides (at t_D) there are two copies. We are interested in computing how this gene-doubling event impacts mRNA expression. We can begin by writing out the reaction network:

$$\begin{aligned} D_1 &\xrightarrow{k_{t,1}} D_1 + m \\ m &\xrightarrow{k_d} \emptyset \end{aligned} \tag{1}$$

for $t < t_r$, and:

$$\begin{aligned} D_1 &\xrightarrow{k_{t,1}} D_1 + m \\ D_2 &\xrightarrow{k_{t,2}} D_2 + m \\ m &\xrightarrow{k_d} \emptyset \end{aligned} \tag{2}$$

for $t > t_r$. These yield the chemical master equations (CMEs):

$$\begin{aligned} \partial_t P\left(m|t < t_r\right) &= k_{t,1} P\left(m-1|t\right) + k_d\left(m+1\right) P\left(m+1|t\right) \\ &\quad - k_{t,1} P\left(m|t\right) - k_d m P\left(m|t\right) \\ \partial_t P\left(m|t > t_r\right) &= \left(k_{t,1} + k_{t,2}\right) P\left(m-1|t\right) \\ &\quad + k_d\left(m+1\right) P\left(m+1|t\right) \\ &\quad - \left(k_{t,1} + k_{t,2}\right) P\left(m|t\right) - k_d m P\left(m|t\right) \end{aligned} \tag{3}$$

In the case where $k_{t,1} = k_{t,2} = k_t$, then the equation for $t > t_r$ above can be simplified as:

$$\begin{aligned} \partial_t P\left(m|t > t_r\right) &= 2k_t P\left(m-1|t\right) + k_d\left(m+1\right) P\left(m+1|t\right) \\ &\quad - 2k_t P\left(m|t\right) - k_d m P\left(m|t\right) \end{aligned} \tag{4}$$

To compute the time evolution of the mean $\bar{m}$ and variance of the mRNA distribution after the gene duplication event. We substituted the RHS of Eq. S4 into the definitions:

$$\begin{aligned} \frac{d\bar{m}}{dt} &= \frac{d}{dt}\sum_{m=0}^{\infty} m P\left(m|t\right) \\ &= \sum_{m=0}^{\infty} m \frac{dP\left(m|t\right)}{dt} \\ &= \sum_{m=0}^{\infty} m\big[2k_t P\left(m-1|t\right) + k_d\left(m+1\right) P\left(m+1|t\right) \\ &\qquad - 2k_t P\left(m|t\right) - k_d m P\left(m|t\right)\big] \end{aligned} \tag{5}$$

Evaluating each term individually:

$$\begin{aligned}\sum_{m=0}^{\infty} m2k_tP\left(m-1|t\right) &= \sum_{y=-1}^{\infty}\left(y+1\right)2k_tP\left(y|t\right)\\ &= 0+\sum_{y=0}^{\infty}\left(y+1\right)2k_tP\left(y|t\right)\\ &= 2k_t\left(1+\bar{y}\right)\\ &= 2k_t\left(1+\bar{m}\right)\end{aligned} \tag{6}$$

$$\begin{aligned}\sum_{m=0}^{\infty} mk_d\left(m+1\right)P\left(m+1|t\right) &= \sum_{y=1}^{\infty}\left(y-1\right)yk_dP\left(y|t\right)\\ &= 0+\sum_{y=0}^{\infty}\left(y-1\right)yk_dP\left(y|t\right)\\ &= k_dE\left[\left(y-1\right)y\right]\\ &= k_dE\left[\left(m-1\right)m\right]\\ &= k_d\left(E\left[m^2\right]-\bar{m}\right)\end{aligned} \tag{7}$$

$$\sum_{m=0}^{\infty} m2k_tP\left(m|t\right) = 2k_t\bar{m} \tag{8}$$

and finally:

$$\sum_{m=0}^{\infty} m^2k_dP\left(m|t\right) = k_dE\left[m^2\right] \tag{9}$$

Summing these with their appropriate signs yields:

$$\begin{aligned}\frac{d\bar{m}(t>t_r)}{dt} &= 2k_t+2k_t\bar{m}+k_dE\left[m^2\right]-k_d\bar{m}-2k_t\bar{m}-k_dE\left[m^2\right]\\ &= 2k_t-k_d\bar{m}\end{aligned} \tag{10}$$

while an identical argument gives:

$$\frac{d\bar{m}(t<t_r)}{dt} = k_t-k_d\bar{m} \tag{11}$$

The solution to these ODEs are straightforward. If gene duplication occurs relatively early in the cell cycle, such that the mRNA count has time to equilibrate (to a value of $\frac{2k_t}{k_d}$) before the cell divides, then we can assume the mean mRNA count at the beginning of the cell cycle is approximately half this value, or $\frac{k_t}{k_d}$. Using this as an initial condition we can write down the mean as a function of time.

$$\bar{m}(t) = \begin{cases}\frac{k_t}{k_d} & 0<t<t_r\\ \frac{k_t}{k_d}\left(2-e^{k_d(t_r-t)}\right) & t_r<t<t_D\end{cases} \tag{12}$$

It should be noted that the assumption that the mean mRNA is approximately equilibrated at the end of the cell cycle is not strictly necessary (although it simplifies the resulting expressions enormously). In section S1.2, we give a more exact derivation that does not rely on this assumption.

We can now consider the evolution of the variance:

$$
\begin{aligned}
\frac{d\sigma_m^2}{dt} &= \frac{d}{dt}\left(E\left[m^2\right] - \bar{m}^2\right) \\
&= \frac{dE\left[m^2\right]}{dt} - 2\bar{m}\frac{d\bar{m}}{dt} \\
&= \frac{dE\left[m^2\right]}{dt} - 2\bar{m}\left(2k_t - k_d\bar{m}\right) \\
&= \left\{\sum_{m=0}^{\infty} m^2 \frac{dP\left(m|t\right)}{dt}\right\} - 2\bar{m}\left(2k_t - k_d\bar{m}\right) \\
&= \left\{\sum_{m=0}^{\infty} m^2 \big[2k_t P\left(m-1|t\right) + k_d\left(m+1\right)P\left(m+1|t\right) \right. \\
&\qquad \left. - 2k_t P\left(m|t\right) - k_d m P\left(m|t\right)\big]\right\} - 2\bar{m}\left(2k_t - k_d\bar{m}\right)
\end{aligned}
\tag{13}
$$

As before, the terms in the summation are evaluated independently:

$$
\begin{aligned}
\sum_{m=0}^{\infty} m^2 2k_t P\left(m-1|t\right) &= \sum_{y=-1}^{\infty}\left(y+1\right)^2 2k_t P\left(y|t\right) \\
&= \sum_{y=0}^{\infty}\left(y+1\right)^2 2k_t P\left(y|t\right) \\
&= 2k_t\left(E\left[y^2\right] + 2\bar{y} + 1\right) \\
&= 2k_t\left(E\left[m^2\right] + 2\bar{m} + 1\right)
\end{aligned}
\tag{14}
$$

$$
\begin{aligned}
\sum_{m=0}^{\infty} m^2\left(m+1\right)k_d P\left(m+1|t\right) &= \sum_{y=1}^{\infty} y\left(y-1\right)^2 k_d P\left(y|t\right) \\
&= \sum_{y=0}^{\infty} y\left(y-1\right)^2 k_d P\left(y|t\right) \\
&= k_d\left(E\left[y^3\right] - 2E\left[y^2\right] + \bar{y}\right) \\
&= k_d\left(E\left[m^3\right] - 2E\left[m^2\right] + \bar{m}\right)
\end{aligned}
\tag{15}
$$

$$
\sum_{m=0}^{\infty} m^2 2k_t P\left(m|t\right) = 2k_t E\left[m^2\right] \tag{16}
$$

and finally:

$$
\sum_{m=0}^{\infty} m^3 k_d P\left(m|t\right) = k_d E\left[m^3\right] \tag{17}
$$

Summing all these expressions together and simplifying gives:

$$
\begin{aligned}
\frac{d\sigma_m^2}{dt} &= 2k_t - 2k_d E\left[m^2\right] + k_d\bar{m} + 2k_d\bar{m}^2 \\
&= 2k_t - 2k_d\left(E\left[m^2\right] - \bar{m}^2\right) + k_d\bar{m} \\
&= 2k_t + k_d\bar{m} - 2k_d\sigma_m^2
\end{aligned}
\tag{18}
$$

Substituting Eq. S12 and solving for $\sigma_m^2\left(t\right)$ when $t > t_r$ yields:

$$
\sigma_m^2\left(t\right) = \frac{k_t}{k_d}\left(2 - e^{k_d\left(t_r - t\right)}\right) + ce^{-2k_d t} \tag{19}
$$

where c is an arbitrary integration constant. Noting that we are considering constitutively expressed mRNA for which we expect $m\left(t < t_r\right) \sim \mathrm{Pois}\left(k_t/k_d\right)$, we can expect $\sigma_m^2(t_r) = k_t/k_d$. Using this in above as an initial condition yields:

$$\sigma_m^2(t) = \bar{m}(t) = \begin{cases} \frac{k_t}{k_d} & 0 < t < t_r \\ \frac{k_t}{k_d}\left(2 - e^{k_d(t_r - t)}\right) & t_r < t < t_D \end{cases} \tag{20}$$

Interestingly, the mean and variance of the mRNA remain equal after the gene duplication event, indicating that the the mRNA remains Poisson–distributed. Although not necessary for the derivation at hand, substituting $P(m|t) = \text{Pois}(\bar{m}(t))$ into Eq. S3 shows that this is indeed the case.

Armed with these results, we can consider sampling the per-cell mRNA copy number of a population of cells. Assuming cells are sampled from across the cell cycle, we can write out the joint probability distribution for a randomly picked cell to have a given mRNA copy number:

$$P(m,t) = P(m|t)\,P(t) \tag{21}$$

where $P(m|t)$ is the distribution of m at a given time t along the cell cycle, and $P(t)$ represents the age distribution of cells in the population. For log-phase cells it has been shown that this distribution decays exponentially with age [1, 2]. Ignoring cell-to-cell variability growth rate (which can be substantial [3]) and cell cycle duration, $P(t)$ can be given approximately by $\frac{2\,ln(2)}{t_D}\,2^{-t/t_D}$. From here, we can compute the probability that a cell will have m mRNA by simply marginalizing against the time variable:

$$P(m) = \int_0^{t_D} P(m,t)\,dt \tag{22}$$

Let's consider the expectation value of this distribution:

$$\begin{aligned} E[m] &= \sum_0^\infty mP(m) \\ &= \sum_0^\infty m \int_0^{t_D} P(m,t)dt \\ &= \sum_0^\infty m \int_0^{t_D} P(m|t)P(t)dt \\ &= \int_0^{t_D} \sum_0^\infty mP(m|t)P(t)dt \\ &= \int_0^{t_D} \bar{m}(t)\frac{2\,ln(2)}{t_D}\,2^{-t/t_D}dt \end{aligned} \tag{23}$$

Evaluating this yields:

$$E[m] = \frac{k_t}{k_d}\frac{1}{ln(2) + k_d t_D}[k_d t_D 2^{1-t_r/t_D} + ln(2)e^{-k_d(t_D - t_r)}] \tag{24}$$

Likewise, we can compute:

$$\begin{aligned} \text{Var}[m] &= \sum_0^\infty m^2P(m) - E[m]^2 \\ &= \sum_0^\infty m^2 \int_0^{t_D} P(m|t)P(t)dt - E[m]^2 \\ &= \int_0^{t_D} \sum_0^\infty m^2P(m|t)P(t)dt - E[m]^2 \\ &= \int_0^{t_D} (\sigma_m^2(t) + \bar{m}(t)^2)P(t)dt - E[m]^2 \\ &= \int_0^{t_D} \sigma_m^2(t)P(t)dt + \int_0^{t_D} \bar{m}(t)^2P(t)dt - E[m]^2 \\ &= \int_0^{t_D} \bar{m}(t)P(t)dt + \int_0^{t_D} \bar{m}(t)^2P(t)dt - E[m]^2 \\ &= E[m] + \int_0^{t_D} \bar{m}(t)^2\frac{2\,ln(2)}{t_D}\,2^{-t/t_D}dt - E[m]^2 \end{aligned} \tag{25}$$

We have already computed the functional form of $E[m]$, so all we have to do is evaluate the above integral:

$$\int_0^{t_D} \bar{m}(t)^2 \frac{2\,ln(2)}{t_D}\, 2^{-t/t_D} dt = (\frac{k_t}{k_d})^2 \Bigg[(2 - 2^{1-t_r/t_D}) - 4(1 - 2^{1-t_r/t_D}) + 4\,ln(2)\frac{e^{-k_d(t_D-t_r)} - 2^{1-t_r/t_D}}{ln(2) + k_d t_D} - ln(2)\frac{e^{-2k_d(t_D-t_r)} - 2^{1-t_r/t_D}}{ln(2) + 2k_d t_D}\Bigg] \tag{26}$$

Packing this all up yields $\mathrm{Var}[m]$:

$$\begin{aligned}\mathrm{Var}[m] &= E[m] - E[m]^2 \\ &+ (\frac{k_t}{k_d})^2 \Bigg[(2 - 2^{1-t_r/t_D}) - 4(1 - 2^{1-t_r/t_D}) \\ &+ 4\,ln(2)\frac{e^{-k_d(t_D-t_r)} - 2^{1-t_r/t_D}}{ln(2) + k_d t_D} - ln(2)\frac{e^{-2k_d(t_D-t_r)} - 2^{1-t_r/t_D}}{ln(2) + 2k_d t_D}\Bigg]\end{aligned} \tag{27}$$

and hence:

$$\begin{aligned}\mathrm{Fano}[m] &= \frac{\mathrm{Var}[m]}{E[m]} \\ &= 1 - E[m] + \frac{1}{E[m]}(\frac{k_t}{k_d})^2 \Bigg[(2 - 2^{1-t_r/t_D}) - 4(1 - 2^{1-t_r/t_D}) \\ &+ 4\,ln(2)\frac{e^{-k_d(t_D-t_r)} - 2^{1-t_r/t_D}}{ln(2) + k_d t_D} - ln(2)\frac{e^{-2k_d(t_D-t_r)} - 2^{1-t_r/t_D}}{ln(2) + 2k_d t_D}\Bigg]\end{aligned} \tag{28}$$

From here it is straightforward to cast these results in terms of the parameters $f = (t_D - t_r)/t_D$ (the fraction of cell cycle after the gene duplication event), and $\langle m \rangle_1 = k_t/k_d$ (the steady-state mean copy number prior to the gene duplication event):

$$\begin{aligned}E[m] &= \frac{\langle m \rangle_1}{1 + \frac{k_d t_D}{ln(2)}}\left[\frac{k_d t_D}{ln(2)} 2^f + e^{-k_d t_D f}\right] \\ \mathrm{Var}[m] &= E[m] - E[m]^2 \\ &+ \langle m \rangle_1^2 \Bigg[(2 - 2^f) - 4(1 - 2^f) + 4\,ln(2)\frac{e^{-k_d t_D f} - 2^f}{ln(2) + k_d t_D} - ln(2)\frac{e^{-2k_d t_D f} - 2^f}{ln(2) + 2k_d t_D}\Bigg] \\ \mathrm{Fano}[m] &= 1 - E[m] \\ &+ \frac{\langle m \rangle_1^2}{E[m]} \Bigg[(2 - 2^f) - 4(1 - 2^f) + 4\,ln(2)\frac{e^{-k_d t_D f} - 2^f}{ln(2) + k_d t_D} - ln(2)\frac{e^{-2k_d t_D f} - 2^f}{ln(2) + 2k_d t_D}\Bigg]\end{aligned} \tag{29}$$

If one were simply to assume a uniform distribution of ages, rather than the exponential distribution used above, the errors in $E[m]$ and $\mathrm{Fano}[m]$ would be within approximately 6% for cells doubling in 40 min (and less than 8% for cells doubling in 70 min). The expressions that result, however, are considerably simpler (and thereby potentially more useful to a broad audience):

$$\begin{aligned}E[m] &= \langle m \rangle_1 \left[1 + f + \frac{e^{-f k_d t_D} - 1}{k_d t_D}\right] \\ \mathrm{Var}\,[m] &= E[m] - E[m]^2 \\ &+ \langle m \rangle_1^2 \left[1 + 3f + \frac{8e^{-f k_d t_D} - e^{-2f k_d t_D} - 7}{2k_d t_D}\right] \\ \mathrm{Fano}\,[m] &= 1 - E[m] \\ &+ \frac{\langle m \rangle_1^2}{E[m]} \left[1 + 3f + \frac{8e^{-f k_d t_D} - e^{-2f k_d t_D} - 7}{2k_d t_D}\right]\end{aligned} \tag{30}$$

For this reason, we ultimately chose to include Eq. S30 in the main manuscript.

1.2 Relaxing the assumption that the mRNA counts equilibrate prior to cell division

Returning to Eqs. S10 & S11; in the prior section we solved for the mean copy number under the assumption that the mRNA counts have ample time to relax to the post-gene duplication steady state. For fast-growing cells with short doubling times, or when the gene of interest is duplicated near the end of a cell cycle, this assumption can prove untrue and can lead to disagreement between the analytical and simulated results (see for example Fig. SS9). We can correct for this straightforwardly by introducing some initial mean mRNA count, $\bar{m}_0$, and solving for it by imposing the requirement that $\bar{m}_0 = \bar{m}(0) = \frac{1}{2}\bar{m}(t_D)$ (meaning the mean mRNA count after division is half what it was before division). We begin by writing down the mean before and after gene duplication:

$$\begin{aligned}\bar{m}(0 < t < t_r) &= (\bar{m}_0 - \frac{k_t}{k_d})e^{-k_d t} + \frac{k_t}{k_d} \\ \bar{m}(t_r < t < t_D) &= (\bar{m}(t_r) - 2\frac{k_t}{k_d})e^{-k_d(t-t_r)} + 2\frac{k_t}{k_d}\end{aligned} \tag{31}$$

Evaluating $\bar{m}(t_r)$ yields $(\bar{m}_0 - \frac{k_t}{k_d})e^{-k_d t_r} + \frac{k_t}{k_d}$ and, in turn, evaluating $\bar{m}(t_D)$ yields $[(\bar{m}_0 - \frac{k_t}{k_d})e^{-k_d t_r} - \frac{k_t}{k_d}]e^{-k_d(t_D - t_r)} + 2\frac{k_t}{k_d}$. Now simply imposing our boundary condition yields:

$$\begin{aligned}2\bar{m}_0 &= [(\bar{m}_0 - \frac{k_t}{k_d})e^{-k_d t_r} - \frac{k_t}{k_d}]e^{-k_d(t_D - t_r)} + 2\frac{k_t}{k_d} \\ \rightarrow \bar{m}_0 &= \frac{k_t}{k_d}[1 - \frac{e^{-k_d(t_D - t_r)}}{2 - e^{-k_d t_D}}]\end{aligned} \tag{32}$$

This yields the exact solution for the mean:

$$\bar{m}(t) = \begin{cases} \frac{k_t}{k_d}[1 - \frac{e^{-k_d(t_D - t_r)}}{2 - e^{-k_d t_D}}e^{-k_d t}] & 0 < t < t_r \\ \frac{k_t}{k_d}[2 - (1 + \frac{e^{-k_d t_D}}{2 - e^{-k_d t_D}})e^{-k_d(t - t_r)}] & t_r < t < t_D \end{cases} \tag{33}$$

Because, as noted in the previous section, $\text{Pois}(\bar{m}(t))$ solves the master equations for this problem, we can simply write $\sigma_m^2(t) = \bar{m}(t)$, although this could also be derived from Eq. S18 and similar arguments to those appearing above. Implicit in this, of course, is the assumption that the mRNA is Poisson–distributed after cell division; at least in the case of a perfectly unbiased division process this is easy to check. The probability that a daughter cell will contain m mRNAs immediately after division can be computed as:

$$P_{\text{daughter}}(m) = \sum_{n=m}^{\infty} P(m|n) P_{\text{mother}}(n) \tag{34}$$

where $P_{\text{mother}}(n)$ represents the probability that the mother cell contains n mRNAs at division time, and $P(m|n)$ represents the probability that the daughter will contain m mRNA given that its mother contains n. If $P_{\text{mother}}(n) = \text{Pois}(\bar{n}(t_D))$ and cell division distributes mRNA with equal probabilities between the daughters we can write:

$$\begin{aligned}P_{\text{daughter}}(m) &= \sum_{n=m}^{\infty} \binom{n}{m} (\frac{1}{2})^n \frac{e^{-\bar{n}(t_D)}\bar{n}(t_D)^n}{n!} \\ &= \frac{e^{-\bar{n}(t_D)}}{m!} \sum_{n=m}^{\infty} \frac{(\frac{\bar{n}(t_D)}{2})^n}{(n-m)!} \\ &= \frac{e^{-\bar{n}(t_D)}}{m!} \sum_{k=0}^{\infty} \frac{(\frac{\bar{n}(t_D)}{2})^{k+m}}{k!} \\ &= \frac{e^{-\bar{n}(t_D)}}{m!} e^{\bar{n}(t_D)/2} (\frac{\bar{n}(t_D)}{2})^m \\ &= \text{Pois}(\frac{\bar{n}(t_D)}{2})\end{aligned} \tag{35}$$

From this we see that the unbiased division of mRNA does indeed result in Poisson–distributed mRNA counts in the daughters.

Now we can compute the expectation value and variance for the messengers in a population of cells. Assuming cells are exponentially distributed yields:

$$
\begin{aligned}
E[m] &= \langle m\rangle_1 2^f\left[1+\beta\frac{e^{-k_d t_D(1-f)}-2^{1-f}}{1+\frac{k_d t_D}{ln(2)}}+\gamma\frac{2^{-f}e^{-k_d t_D f}-1}{1+\frac{k_d t_D}{ln(2)}}\right]\\
\mathrm{Var}[m] &= E[m]-E[m]^2\\
&\quad + ln(2)\langle m\rangle_1^2\left[2\beta^2\frac{1-2^{f-1}e^{-2k_d t_D(1-f)}}{ln(2)+2k_d t_D}-4\beta\frac{1-2^{f-1}e^{-k_d t_D(1-f)}}{ln(2)+k_d t_D}+\frac{2}{ln(2)}(1-2^{f-1})\right.\\
&\quad \left.+\gamma^2\frac{2^f-e^{-2k_d t_D f}}{ln(2)+2k_d t_D}-4\gamma\frac{2^f-e^{-k_d t_D f}}{ln(2)+k_d t_D}-\frac{4}{ln(2)}(1-2^f)\right]\\
\mathrm{Fano}[m] &= 1-E[m]\\
&\quad + ln(2)\frac{\langle m\rangle_1^2}{E[m]}\left[2\beta^2\frac{1-2^{f-1}e^{-2k_d t_D(1-f)}}{ln(2)+2k_d t_D}-4\beta\frac{1-2^{f-1}e^{-k_d t_D(1-f)}}{ln(2)+k_d t_D}+\frac{2}{ln(2)}(1-2^{f-1})\right.\\
&\quad \left.+\gamma^2\frac{2^f-e^{-2k_d t_D f}}{ln(2)+2k_d t_D}-4\gamma\frac{2^f-e^{-k_d t_D f}}{ln(2)+k_d t_D}-\frac{4}{ln(2)}(1-2^f)\right]
\end{aligned}
\tag{36}
$$

while assuming cells to be uniformly distributed yields:

$$
\begin{aligned}
E[m] &= \langle m\rangle_1\left[1+f+\frac{\beta}{k_d t_D}(e^{-k_d t_D(1-f)}-1)+\frac{\gamma}{k_d t_D}(e^{-k_d t_D f}-1)\right]\\
\mathrm{Var}[m] &= E[m]-E[m]^2\\
&\quad + \langle m\rangle_1^2\left[1+3f-\frac{4\beta}{2k_d t_D}(1-e^{-k_d t_D(1-f)})-\frac{8\gamma}{2k_d t_D}(1-e^{-k_d t_D f})\right.\\
&\quad \left.+\frac{\beta^2}{2k_d t_D}(1-e^{-2k_d t_D(1-f)})+\frac{\gamma^2}{2k_d t_D}(1-e^{-2k_d t_D f})\right]\\
\mathrm{Fano}[m] &= 1-E[m]\\
&\quad + \frac{\langle m\rangle_1^2}{E[m]}\left[1+3f-\frac{4\beta}{2k_d t_D}(1-e^{-k_d t_D(1-f)})-\frac{8\gamma}{2k_d t_D}(1-e^{-k_d t_D f})\right.\\
&\quad \left.+\frac{\beta^2}{2k_d t_D}(1-e^{-2k_d t_D(1-f)})+\frac{\gamma^2}{2k_d t_D}(1-e^{-2k_d t_D f})\right]
\end{aligned}
\tag{37}
$$

where:

$$
\begin{aligned}
\beta &= \frac{e^{-k_d t_D f}}{2-e^{-k_d t_D}}\\
\gamma &= (1+\frac{e^{-k_d t_D}}{2-e^{-k_d t_D}})
\end{aligned}
\tag{38}
$$

These results show nearly exact agreement with simulation (see Figure SS9). In the limit where f approaches 1 and t_D is large, these expressions reduce to those of Eqs. S29 & S30; the maximum difference ($\approx 17\%$) occurs when f is small, but we note that for values of f greater than 0.1 the difference between the above expressions and those of Eqs. S29 & S30 remains less than 8%. Because the expressions in equation S37 are considerably more complicated than those appearing in equation S30, and because they generally amount to a fairly small correction, we have not included them in the main manuscript.

1.3 Corrections to the Fano Factor for the case of regulated mRNA expression

When mRNA production is regulated (*e.g.* by a transcription factor) the dynamics of the system can be significantly more complicated. The behaviour of a single gene switching between "off" (dormant) and "on" (active, capable of producing mRNA) states has been studied on multiple occasions [4, 5, 6, 7, 8]. These analyses have shown that the mean and variance of the mRNA distribution at steady-state approach:

$$E_{ss}[m] = \frac{k_t}{k_d}\frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}} \tag{39}$$

and:

$$\text{Var}_{ss}[m] = E_{ss}[m]\left(1 + \frac{k_t k_{\text{off}}}{(k_{\text{on}}+k_{\text{off}})(k_{\text{on}}+k_{\text{off}}+k_d)}\right) \tag{40}$$

respectively.

As before, we are interested in computing the Fano factor in the case where a gene replication event occurs during the cell cycle. We saw previously that this quantity can be derived directly by integrating over time-dependent expressions for the mean and variance of the mRNA copy number. The mean is most easily studied in the continuum limit; we can write:

$$\frac{d\bar{m}(t)}{dt} = k_t\bar{g}(t) - k_d\bar{m}(t) \tag{41}$$

where $\bar{g}(t)$ represents the instantaneous mean number of "on" genes after the gene duplication event. As a first approximation, $\bar{g}$ might be assumed constant, and equal to $2k_{\text{on}}/(k_{\text{on}}+k_{\text{off}})$. This would be reasonable in the limit where the gene switching is fast such that $\bar{g}$ relaxes to its new steady state quickly compared to the time required for $\bar{m}(t)$ to relax to its new steady state. This assumption immediately yields:

$$\bar{m}(t) = \begin{cases} \frac{k_t}{k_d}\frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}} & 0 < t < t_r \\ \frac{k_t}{k_d}\frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}}\left(2 - e^{k_d(t_r-t)}\right) & t_r < t < t_D \end{cases} \tag{42}$$

Not surprisingly, the mean expression is identical to the unregulated case, except that it is scaled by $\frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}}$.

We can now turn our attention to the estimating $\sigma_m^2(t)$. This is non-trivial, and we will not attempt a complete derivation here. But Eq. S40 indicates that the variance before and long after the duplication event should be proportional to the mean; if, as a first approximation, we were to simply assume that the dynamics of the variance occur on a similar time-scale as the dynamics of the mean, then we could immediately write:

$$\sigma_m^2(t) \approx \bar{m}(t)\left(1 + \frac{k_t k_{\text{off}}}{(k_{\text{on}}+k_{\text{off}})(k_{\text{on}}+k_{\text{off}}+k_d)}\right) \tag{43}$$

which yields:

$$\text{Var}[m] = E[m]\left(1 + \frac{k_t k_{\text{off}}}{(k_{\text{on}}+k_{\text{off}})(k_{\text{on}}+k_{\text{off}}+k_d)}\right) - E[m]^2 + \int_0^{t_D}\bar{m}(t)^2 P(t)dt \tag{44}$$

Evaluating this assuming an exponential age distribution yields:

$$\begin{aligned} E[m] &= \frac{\langle m\rangle_1}{1+\frac{k_d t_D}{ln(2)}}\left[\frac{k_d t_D}{ln(2)}2^f + e^{-k_d t_D f}\right] \\ \text{Var}[m] &= E[m]\left(1 + \frac{k_t k_{\text{off}}}{(k_{\text{on}}+k_{\text{off}})(k_{\text{on}}+k_{\text{off}}+k_d)}\right) - E[m]^2 \\ &\quad + \langle m\rangle_1^2\left[(2-2^f) - 4(1-2^f) + 4\,ln(2)\frac{e^{-k_d t_D f}-2^f}{ln(2)+k_d t_D} - ln(2)\frac{e^{-2k_d t_D f}-2^f}{ln(2)+2k_d t_D}\right] \\ \text{Fano}[m] &= \left(1 + \frac{k_t k_{\text{off}}}{(k_{\text{on}}+k_{\text{off}})(k_{\text{on}}+k_{\text{off}}+k_d)}\right) - E[m] \\ &\quad + \frac{\langle m\rangle_1^2}{E[m]}\left[(2-2^f) - 4(1-2^f) + 4\,ln(2)\frac{e^{-k_d t_D f}-2^f}{ln(2)+k_d t_D} - ln(2)\frac{e^{-2k_d t_D f}-2^f}{ln(2)+2k_d t_D}\right] \end{aligned} \tag{45}$$

where

$$\langle m\rangle_1 = \frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}}\frac{k_t}{k_d} \tag{46}$$

Evaluating assuming a uniform distribution yields:

$$
\begin{aligned}
E[m] &= \langle m \rangle_1 \left[1 + f + \frac{e^{-f k_d t_D} - 1}{k_d t_D} \right] \\
\text{Var}\,[m] &= E[m] \left(1 + \frac{k_t k_{\text{off}}}{(k_{\text{on}} + k_{\text{off}})(k_{\text{on}} + k_{\text{off}} + k_d)} \right) - E[m]^2 \\
&\quad + \langle m \rangle_1^2 \left[1 + 3f + \frac{8e^{-f k_d t_D} - e^{-2 f k_d t_D} - 7}{2 k_d t_D} \right] \\
\text{Fano}\,[m] &= \left(1 + \frac{k_t k_{\text{off}}}{(k_{\text{on}} + k_{\text{off}})(k_{\text{on}} + k_{\text{off}} + k_d)} \right) - E[m] \\
&\quad + \frac{\langle m \rangle_1^2}{E[m]} \left[1 + 3f + \frac{8e^{-f k_d t_D} - e^{-2 f k_d t_D} - 7}{2 k_d t_D} \right]
\end{aligned}
\tag{47}
$$

Comparison with Eqs. S29 & S30 shows that these are functionally very similar to the unregulated case but with an additional term, $\frac{k_t k_{\text{off}}}{(k_{\text{on}}+k_{\text{off}})(k_{\text{on}}+k_{\text{off}}+k_d)}$, and, as noted above, the replacement $\langle m \rangle_1 \rightarrow \frac{k_t}{k_d} \frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}}$.

1.4 Further corrections for cases in which $k_{\text{on}}, k_{\text{off}} \lesssim k_d$

When the gene state switching rates, k_{on} and k_{off}, are not faster than k_d, we might expect that some additional refinements are in order. The first of which would be that the dynamics of $\bar{g}(t)$ appearing in Eq. S41 ought not be ignored.

We can write down a differential equation for the $\bar{g}(t)$ after the gene duplication:

$$
\frac{d\bar{g}(t)}{dt} = (2 - \bar{g}(t)) k_{\text{on}} - \bar{g}(t) k_{\text{off}} \tag{48}
$$

for which the solution is:

$$
\bar{g}(t) = \frac{2k_{\text{on}}}{k_{\text{on}} + k_{\text{off}}} + c e^{-(k_{\text{on}}+k_{\text{off}})t} \tag{49}
$$

where c is an arbitrary integration constant. There are a few things to consider before we decide on an initial condition. Genes are replicated by a large protein complex that sweeps along the DNA, unzipping it and replicating both strands as it goes. Any transcription factors bound to the original gene copy's promoter region would have been unbound by the replication complex, and so both genes start off in an unbound state at time t_r. This can mean one of two things—if the transcription factor was a repressor, then both genes would begin "on" ($\bar{g}(t_r) = 2$), while if it were an activator, both genes would begin "off" ($\bar{g}(t_r) = 0$). This yields:

$$
\bar{g}(t) = \begin{cases} \frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}} & 0 < t < t_r \\ \frac{2k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}} \left(1 - e^{(k_{\text{on}}+k_{\text{off}})(t_r - t)}\right) & t_r < t < t_D \end{cases} \tag{50}
$$

if the regulator is an activator, and:

$$
\bar{g}(t) = \begin{cases} \frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}} & 0 < t < t_r \\ \frac{2}{k_{\text{on}}+k_{\text{off}}} \left(k_{\text{on}} + k_{\text{off}} e^{(k_{\text{on}}+k_{\text{off}})(t_r - t)}\right) & t_r < t < t_D \end{cases} \tag{51}
$$

if the regulator is a repressor.

We can insert these into Eq. S41 and solve it yielding:

$$
\bar{m}(t) = \begin{cases} \frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}} \frac{k_t}{k_d} & 0 < t < t_r \\ \frac{k_{\text{on}}}{k_{\text{on}} + k_{\text{off}}} \frac{k_t}{k_d} \left[\left(2 - e^{k_d(t_r - t)}\right) + \frac{2k_d}{k_{\text{on}} + k_{\text{off}} - k_d} \left(e^{(k_{\text{on}}+k_{\text{off}})(t_r - t)} - e^{k_d(t_r - t)} \right) \right] & t_r < t < t_D \end{cases} \tag{52}
$$

if the regulator is an activator, and:

$$\bar{m}(t) = \begin{cases} \frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}}\frac{k_t}{k_d} & 0 < t < t_r \\ \frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}}\frac{k_t}{k_d}\left[\left(2 - e^{k_d(t_r-t)}\right) - \frac{2k_d}{k_{\text{on}}+k_{\text{off}}-k_d}\frac{k_{\text{off}}}{k_{\text{on}}}\left(e^{(k_{\text{on}}+k_{\text{off}})(t_r-t)} - e^{k_d(t_r-t)}\right)\right] & t_r < t < t_D \end{cases} \tag{53}$$

if the regulator is a repressor.

From here, evaluating $E[m]$ and $\text{Var}[m]$ is straightforward, if somewhat laborious. In the end, the functional forms they take are not particularly illuminating, but we have included them here for the sake of completeness:

$$\begin{aligned} E_{\text{act}}[m] = &\frac{1}{t_D}\frac{k_t}{k_d}\frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}} \\ &\times \left(t_r + \frac{2k_d}{(k_{\text{on}}+k_{\text{off}})(k_{\text{on}}+k_{\text{off}}-k_d)}\left(1 - e^{-(k_{\text{on}}+k_{\text{off}})(t_D-t_r)}\right)\right. \\ &\left. + \frac{k_{\text{on}}+k_{\text{off}}+k_d}{k_{\text{on}}+k_{\text{off}}-k_d}\frac{1}{k_d}\left(e^{-k_d(t_D-t_r)} - 1\right) + 2(t_D - t_r)\right) \end{aligned} \tag{54}$$

and:

$$\begin{aligned} \text{Var}_{\text{act}}[m] = &\left(1 + \frac{k_t k_{\text{off}}}{(k_{\text{on}}+k_{\text{off}})(k_{\text{on}}+k_{\text{off}}+k_d)}\right)E_{\text{act}}[m] - E_{\text{act}}[m]^2 \\ &+ eta\frac{t_r}{t_D}\left(\frac{k_t}{k_d}\frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}}\right)^2 + \frac{k_t^2 k_{\text{on}}^2}{2t_D k_d^2(k_{\text{on}}+k_{\text{off}})^2(k_{\text{on}}+k_{\text{off}}-k_d)^2} \\ &\times \left(\frac{4k_d^2}{(k_{\text{on}}+k_{\text{off}})}\left(1 - e^{-2(k_{\text{on}}+k_{\text{off}})(t_D-t_r)}\right)\right. \\ &+ \frac{(k_{\text{on}}+k_{\text{off}}+k_d)^2}{k_d}\left(1 - e^{-2k_d(t_D-t_r)}\right) \\ &+ \frac{k_d^2 - (k_{\text{on}}+k_{\text{off}})^2}{k_d}\left(1 - e^{-k_d(t_D-t_r)}\right) \\ &- \frac{8k_d(2k_d - k_{\text{on}} - k_{\text{off}})}{k_{\text{on}}+k_{\text{off}}} \\ &+ 8k_d\left(e^{-(k_{\text{on}}+k_{\text{off}}+k_d)(t_D-t_r)} + 2\frac{k_d - k_{\text{on}} - k_{\text{off}}}{k_{\text{on}}+k_{\text{off}}}e^{-(k_{\text{on}}+k_{\text{off}})(t_D-t_r)}\right) \\ &\left. + 8(k_{\text{on}}+k_{\text{off}}-k_d)^2(t_D - t_r)\right) \end{aligned} \tag{55}$$

for activation-type regulation, and:

$$\begin{aligned} E_{\text{rep}}[m] = &\frac{1}{t_D}\frac{k_t}{k_d}\frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}} \\ &\times \left(t_r + \frac{2k_d k_{\text{off}}\left(e^{-(k_{\text{on}}+k_{\text{off}})(t_D-t_r)} - 1\right)}{(k_{\text{on}}+k_{\text{off}})(k_{\text{on}}+k_{\text{off}}-k_d)k_{\text{on}}}\right. \\ &\left. + \frac{k_d(2k_{\text{off}}+k_{\text{on}}) - k_{\text{on}}(k_{\text{off}}+k_{\text{on}})}{k_d k_{\text{on}}(k_{\text{off}}+k_{\text{off}}-k_d)}\left(1 - e^{-k_d(t_D-t_r)}\right) + 2(t_D - t_r)\right) \end{aligned} \tag{56}$$

and:

$$
\begin{aligned}
\mathrm{Var}_{\mathrm{rep}}[m] = &\left(1+\frac{k_t k_{\mathrm{off}}}{(k_{\mathrm{on}}+k_{\mathrm{off}})(k_{\mathrm{on}}+k_{\mathrm{off}}+k_d)}\right)E_{\mathrm{act}}[m]-E_{\mathrm{act}}[m]^2\\
&+\frac{t_r}{t_D}\left(\frac{k_t}{k_d}\frac{k_{\mathrm{on}}}{k_{\mathrm{on}}+k_{\mathrm{off}}}\right)^2+\frac{k_t^2}{2t_D k_d^2(k_{\mathrm{on}}+k_{\mathrm{off}})^2(k_{\mathrm{on}}+k_{\mathrm{off}}-k_d)^2}\\
&\times\left(\frac{4k_d^2k_{\mathrm{off}}^2}{(k_{\mathrm{on}}+k_{\mathrm{off}})}\left(1-e^{-2(k_{\mathrm{on}}+k_{\mathrm{off}})(t_D-t_r)}\right)\right.\\
&+\frac{(k_{\mathrm{on}}(k_{\mathrm{on}}+k_{\mathrm{off}})-k_d(2k_{\mathrm{off}}+k_{\mathrm{on}}))^2}{k_d}\left(1-e^{-2k_d(t_D-t_r)}\right)\\
&+\frac{8k_{\mathrm{on}}\left(k_d^2\left(2k_{\mathrm{off}}+k_{\mathrm{on}}\right)-2k_d\left(k_{\mathrm{off}}+k_{\mathrm{on}}\right)^2+k_{\mathrm{on}}\left(k_{\mathrm{off}}+k_{\mathrm{on}}\right)^2\right)}{k_d}\left(e^{-k_d(t_D-t_r)}-1\right)\\
&+\frac{8k_dk_{\mathrm{off}}\left(-k_d\left(3k_{\mathrm{off}}k_{\mathrm{on}}+2k_{\mathrm{off}}^2+k_{\mathrm{on}}^2\right)+2k_d^2k_{\mathrm{on}}-k_{\mathrm{on}}\left(k_{\mathrm{off}}+k_{\mathrm{on}}\right)^2\right)}{\left(k_{\mathrm{off}}+k_{\mathrm{on}}\right)\left(k_d+k_{\mathrm{off}}+k_{\mathrm{on}}\right)}\\
&-\frac{8k_dk_{\mathrm{off}}e^{-(k_{\mathrm{on}}+k_{\mathrm{off}}+k_d)(t_D-t_r)}}{(k_{\mathrm{on}}+k_{\mathrm{off}})(k_{\mathrm{on}}+k_{\mathrm{off}}+k_d)}\left[2k_d^2k_{\mathrm{on}}e^{k_d(t_D-t_r)}\right.\\
&\left.-k_{\mathrm{on}}(k_{\mathrm{on}}+k_{\mathrm{off}})^2\left(2e^{k_d(t_D-t_r)}-1\right)-k_d\left(2k_{\mathrm{off}}^2+3k_{\mathrm{off}}k_{\mathrm{on}}+k_{\mathrm{on}}^2\right)\right]\\
&\left.+8k_{\mathrm{on}}^2(k_{\mathrm{on}}+k_{\mathrm{off}}-k_d)^2(t_D-t_r)\right)
\end{aligned}
\tag{57}
$$

for repression-type regulation.

1.5 Corrections to the Fano Factor arising from variability in RNAP copy number

We consider how cell-to-cell variability in RNA polymerase (RNAP) copy numbers can impact the Fano factor of a gene that doubles during the cell cycle. In our analysis thus far, we have considered a single constant transcription rate, k_t. If we assume that this rate is proportional to the number of RNAPs available to transcribe a gene, then we can simply promote k_t to a random variable and analyse its effect on our earlier results. For simplicity, we return to the case of constitutive expression, and specifically to our considerations of $P(m)$ (see Eqs. S21 and S22). In that case we had assumed k_t was fixed and derived the mean and variance of $P(m|t)$ at every t; now we assume k_t is random and realize our expressions actually give the mean and variance of $P(m|t,k_t)$. From here we simply write:

$$
\begin{aligned}
P(m) &= \int_0^\infty \int_0^{t_D} P(m,t,k_t)\,dt\,dk_t\\
&= \int_0^\infty \int_0^{t_D} P(m|t,k_t)P(t)P(k_t)\,dt\,dk_t
\end{aligned}
\tag{58}
$$

Now, evaluating $E[m]$ follows the same logic as before:

$$
\begin{aligned}
E[m] = \sum_0^\infty mP(m) &= \sum_0^\infty m\int_0^\infty dk_t\int_0^{t_D} dtP(m|t,k_t)P(t)P(k_t)\\
&= \int_0^\infty\int_0^{t_D}\sum_0^\infty mP(m|t,k_t)P(t)P(k_t)\,dt\,dk_t\\
&= \int_0^\infty\int_0^{t_D}\bar{m}(t,k_t)P(t)P(k_t)\,dt\,dk_t\\
&= \int_0^\infty E[m|k_t]P(k_t)\,dk_t\\
&= E_{k_t}[E[m|k_t]]\\
&\approx E[m|\bar{k_t}]+\frac{1}{2}\left(\frac{\partial^2E[m|k_t]}{\partial k_t^2}\right)\Big|_{\bar{k_t}}\mathrm{Var}[k_t]\\
&= E[m|\bar{k_t}]
\end{aligned}
\tag{59}
$$

where $E[m|k_t]$ represents $E[m]$ (given by Eq. S24) evaluated at a specific value of k_t, $E_{k_t}[f(k_t)]$ represents the expectation value of $f(k_t)$ over k_t, and $\bar{k_t}$ represents the mean value of k_t. Note that the second to last line follows from a Taylor expansion of $E[m|k_t]$ about $E[m|\bar{k_t}]$.

Likewise we can do the same type of analysis for $\mathrm{Var}[m]$:

$$\begin{aligned}
\mathrm{Var}[m] &= \sum_0^\infty m^2 P(m) - E[m]^2 \\
&= \sum_0^\infty m^2 \int_0^\infty dk_t \int_0^{t_D} dt P(m|t,k_t) P(t) P(k_t)\, dt\, dk_t - E[m]^2 \\
&= \int_0^\infty \int_0^{t_D} \sum_0^\infty m^2 P(m|t,k_t) P(t) P(k_t)\, dt\, dk_t - E[m]^2 \\
&= \int_0^\infty \int_0^{t_D} \left(\sigma_m^2(t,k_t) + \bar{m}^2(t,k_t)\right) P(t) P(k_t)\, dt\, dk_t - E[m]^2 \\
&= \int_0^\infty \left(\mathrm{Var}[m|k_t] + E[m|k_t]^2\right) P(k_t)\, dk_t - E[m]^2 \\
&= E_{k_t}\left[\mathrm{Var}[m|k_t] + E[m|k_t]^2\right] - E[m]^2 \\
&= E_{k_t}\left[\mathrm{Var}[m|k_t]\right] + E_{k_t}[E[m|k_t]^2] - E[m]^2 \\
&\approx \mathrm{Var}[m|\bar{k_t}] + \frac{1}{2}\left(\frac{\partial^2 \mathrm{Var}[m|k_t]}{\partial k_t^2}\right)\Big|_{\bar{k_t}} \mathrm{Var}[k_t] \\
&+ E[m|\bar{k_t}]^2 + \frac{1}{2}\left(\frac{\partial^2 E[m|k_t]^2}{\partial k_t^2}\right)\Big|_{\bar{k_t}} \mathrm{Var}[k_t] - E[m]^2
\end{aligned} \tag{60}$$

which ultimately yields:

$$\mathrm{Var}[m] \approx \mathrm{Var}[m|\bar{k_t}] + \frac{\nu \mathrm{Var}[k_t]}{k_d^2} \tag{61}$$

where

$$\nu = \begin{cases} (2-2^f) - 4(1-2^f) + 4\, ln(2) \frac{e^{-k_d t_D f} - 2^f}{ln(2) + k_d t_D} - ln(2) \frac{e^{-2k_d t_D f} - 2^f}{ln(2) + 2k_d t_D} & \text{when } P(t) \text{ is exponential} \\ 1 + 3f + \frac{8e^{-f k_d t_D} - e^{-2f k_d t_D} - 7}{2k_d t_D} & \text{when } P(t) \text{ is uniform} \end{cases} \tag{62}$$

and therefore the Fano factor simply picks up an additive correction:

$$\mathrm{Fano}[m] \approx \mathrm{Fano}[m|\bar{k_t}] + \frac{\nu \mathrm{Var}[k_t]}{E[m] k_d^2} \tag{63}$$

for the case of constitutive mRNA expression.

It is worth noting that nothing about this analysis is specific to k_t; the same types of arguments can be made in order to account for variability in any parameters, including the mRNA degradation rate (k_d), the cell cycle length (t_D), or the gene replication time (t_r). In the next section we will use a similar argument to consider how variability in transcription factor copy number affects regulated mRNA expression.

1.6 Corrections to the Fano Factor arising from variability in transcription factor copy number

We return to our earlier consideration of regulated mRNA expression (see Section S1.3). We had previously considered a model wherein the gene can be in either an "on" state capable of being transcribed or an "off" state which is incapable of being transcribed. We might assume that the transition between states is mediated by the binding of a transcription factor (TF). As we did previously for k_t, we can promote k_{on} (if the TF is an activator) or k_{off} (if the TF is a repressor) to a random variable which can be assumed to vary from cell-to-cell. We can then compute how this effects the mean mRNA copy number. If, for example, we assume the TF is an activator we can write:

$$
\begin{aligned}
E[m] &\approx E[m|\bar{k_{\text{on}}}] + \frac{1}{2}\left(\frac{\partial^2 E[m|k_{\text{on}}]}{\partial k_{\text{on}}^2}\right)\Bigg|_{\bar{k_{\text{on}}}} \text{Var}[k_{\text{on}}] \\
&= E[m|\bar{k_{\text{on}}}]\left[1 - \frac{k_{\text{off}}\text{Var}[k_{\text{on}}]}{\bar{k_{\text{on}}}(\bar{k_{\text{on}}} + k_{\text{off}})^2}\right]
\end{aligned}
\tag{64}
$$

Evaluating the variance is tedious, and likely best performed using a computer algebra system:

$$
\begin{aligned}
\text{Var}[m] \approx\ & \text{Var}[m|\bar{k_{\text{on}}}] + \frac{1}{2}\left(\frac{\partial^2 \text{Var}[m|k_{\text{on}}]}{\partial k_{\text{on}}^2}\right)\Bigg|_{\bar{k_{\text{on}}}} \text{Var}[k_{\text{on}}] \\
&+ E[m|\bar{k_{\text{on}}}]^2 + \frac{1}{2}\left(\frac{\partial^2 E[m|k_{\text{on}}]^2}{\partial k_{\text{on}}^2}\right)\Bigg|_{\bar{k_{\text{on}}}} \text{Var}[k_{\text{on}}] - E[m]^2 \\
&= \text{Var}[m|\bar{k_{\text{on}}}] + E[m|\bar{k_{\text{on}}}]^2\left[1 - \left(1 - \frac{k_{\text{off}}\text{Var}[k_{\text{on}}]}{\bar{k_{\text{on}}}(\bar{k_{\text{on}}} + k_{\text{off}})^2}\right)^2\right] \\
&- \frac{\text{Var}[k_{\text{on}}]k_{\text{off}}k_t}{k_d^2(\bar{k_{\text{on}}} + k_{\text{off}})^4(k_d + \bar{k_{\text{on}}} + k_{\text{off}})^3} \\
&\times \Big[2\bar{k_{\text{on}}}^4\nu k_t - k_{\text{off}}^4\nu k_t + 3k_d^2\eta k_{\text{off}}^3 \\
&+ 3k_d^3\eta k_{\text{off}}^2 + 3k_d^2\eta\bar{k_{\text{on}}}^3 + 3k_d^3\eta\bar{k_{\text{on}}}^2 \\
&+ k_d\eta k_{\text{off}}^4 + k_d^4\eta k_{\text{off}} + k_d\eta\bar{k_{\text{on}}}^4 \\
&+ k_d^4\eta\bar{k_{\text{on}}} + 4k_d\eta k_{\text{off}}\bar{k_{\text{on}}}^3 + 4k_d\eta k_{\text{off}}^3\bar{k_{\text{on}}} \\
&+ 6k_d^3\eta k_{\text{off}}\bar{k_{\text{on}}} + 3k_d\eta k_{\text{off}}^3 k_t + 2k_d^3\eta k_{\text{off}}k_t \\
&- 3k_d\eta\bar{k_{\text{on}}}^3 k_t - k_d^3\eta\bar{k_{\text{on}}}k_t - 3k_d k_{\text{off}}^3\nu k_t \\
&- k_d^3 k_{\text{off}}\nu k_t + 6k_d\bar{k_{\text{on}}}^3\nu k_t + 2k_d^3\bar{k_{\text{on}}}\nu k_t \\
&+ 5k_{\text{off}}\bar{k_{\text{on}}}^3\nu k_t - k_{\text{off}}^3\bar{k_{\text{on}}}\nu k_t + 6k_d\eta k_{\text{off}}^2\bar{k_{\text{on}}}^2 \\
&+ 9k_d^2\eta k_{\text{off}}\bar{k_{\text{on}}}^2 + 9k_d^2\eta k_{\text{off}}^2\bar{k_{\text{on}}} + 5k_d^2\eta k_{\text{off}}^2 k_t \\
&- 3k_d^2\eta\bar{k_{\text{on}}}^2 k_t - 3k_d^2 k_{\text{off}}^2\nu k_t + 6k_d^2\bar{k_{\text{on}}}^2\nu k_t \\
&+ 3k_{\text{off}}^2\bar{k_{\text{on}}}^2\nu k_t - 3k_d\eta k_{\text{off}}\bar{k_{\text{on}}}^2 k_t + 3k_d\eta k_{\text{off}}^2\bar{k_{\text{on}}}k_t \\
&+ 2k_d^2\eta k_{\text{off}}\bar{k_{\text{on}}}k_t + 9k_d k_{\text{off}}\bar{k_{\text{on}}}^2\nu k_t + 3k_d^2 k_{\text{off}}\bar{k_{\text{on}}}\nu k_t\Big]
\end{aligned}
\tag{65}
$$

where:

$$
\nu = \begin{cases} (2-2^f) - 4(1-2^f) + 4\,ln(2)\frac{e^{-k_d t_D f} - 2^f}{ln(2)+k_d t_D} - ln(2)\frac{e^{-2k_d t_D f} - 2^f}{ln(2)+2k_d t_D} & \text{when } P(t) \text{ is exponential} \\ 1 + 3f + \frac{8e^{-fk_d t_D} - e^{-2fk_d t_D} - 7}{2k_d t_D} & \text{when } P(t) \text{ is uniform} \end{cases}
\tag{66}
$$

and:

$$
\eta = \begin{cases} \frac{1}{1+\frac{k_d t_D}{ln(2)}}\left[\frac{k_d t_D}{ln(2)}2^f + e^{-k_d t_D f}\right] & \text{when } P(t) \text{ is exponential} \\ 1 + f + \frac{e^{-fk_d t_D} - 1}{k_d t_D} & \text{when } P(t) \text{ is uniform} \end{cases}
\tag{67}
$$

This is, obviously, a somewhat long and cumbersome expression, but we can nonetheless estimate the size of the corrections. If we assume $k_d = 0.126$ min^{-1}, $k_t = 10k_d$, and that there are on average 10 copies of the TF per cell (which, assuming the TF is in the extrinsic noise limit where $\text{Var}[\text{TF}]/E[\text{TF}]^2 \approx 0.1$ [9], yields an estimate of $\text{Var}[k_{\text{on}}] \approx 0.1 \times \bar{k_{\text{on}}}^2$), we can compute the corrections to the mean and Fano factor spanning a range of values of $\bar{k_{\text{on}}}$ and k_{off} (from 10^{-3} to 10^2 min^{-1} in both rates). Accounting for TF variability generally resulted in a small decrease in the value of $E[m]$. Over the range of kinetic parameters studied, the largest change to the mean mRNA copy number computed was only around 3%. TF variability made a marginally larger impact on the Fano factor values, but these changes were highly dependent on the values of $\bar{k_{\text{on}}}$ and k_{off}. We found that when $\bar{k_{\text{on}}} << k_{\text{off}}$, the Fano factor increased by

approximately $E[m]/10$—a contribution of similar magnitude to that stemming from RNAP variability—but when $k_{\text{on}}^{-} \sim k_{\text{off}}$ we find that this contribution drops to below 3% of $E[m]$. When $k_{\text{on}}^{-} >> k_{\text{off}}$ the correction becomes vanishingly small. Importantly, these results indicate that TF variability generally imparts less mRNA noise than does RNAP variability.

1.7 Comparison between different models considering gene copy number variation

In order to show that mRNA relaxation dynamics play an important role and that gene replication should be handle explicitly, we compared to previous treatments of gene copy number effect. Two studies have examined the contributions to transcriptional noise arising from of variations in gene copy number [10, 11]. In one model, the "constant DNA model", the average number of gene copies across a population of cells is computed and each simulated cell is assumed to have this copy number over all time [10]. In this case, the mean mRNA copy number scales linearly with gene copy number and for constitutively expressed genes the Fano factor remains unitary. This is due to the fact that the addition of multiple Poisson variables (*i.e.* the mRNA produced from each gene), yields a Poisson variable. Therefore, simulating a population with an average gene copy of 1 or 3 does not change the observed noise (see Fig. SS2, blue bars).

In the second model, the "weighted DNA model" [11], each replicating cell's gene copy number is constant over the cell cycle; however, the copy number for each cell is drawn from the distribution of copy numbers observed in a population of cells. In this case, the theory of Cooper and Helmstetter [12] can be used to calculate the probability of having a particular gene copy number from the doubling time and gene location by taking the fraction of the cell cycle in which that number count exists (see Tables in Figs. SS4C and SS5C). We simulated cells in slow growth (70 min doubling time) and a fast growth (40 min doubling time) conditions, with genes located 10% and 90% from replication origin. Consistent with reported by Jones *et al.*, under all four conditions, Fano factors are increased to different extend, demonstrating approaches using the constant DNA model are qualitatively wrong (Fig. SS2, yellow bars). When we compared the results obtained when simulating DNA replication explicitly, it became apparent that the weighted model overestimates the noise from gene replication (compare red to yellow bars in Fig. SS2B). Including explicit replication and mRNA relaxation dynamics in fact results in noise that is consistently lower than the weighted model by a significant amount (Fig. SS2, red bars). Simulated mRNA distributions for a wide range of mRNA locations at two different doubling times demonstrate that the weighted DNA model is consistently quantitatively, and even qualitatively incapable of capturing the observed noise from these more realistic simulations (see, for example, Fig. 2 in the main manuscript, Figs. SS4, SS5, and SS7 blue lines). On the other hand, the mean and Fano factor as well as mRNA distributions computed via the time-dependent theory developed in this paper are both qualitatively and quantitatively more correct when compared to exact simulations in almost every scenario studied, demonstrating that mRNA relaxation dynamics, which constitute a significant portion of the overall cell cycle, impact the statistics of observable mRNA copy numbers. As a concrete example, to reach the new steady–state mRNA level after DNA replication (defined by relaxation to within 1σ of mean), it takes ∼6.4 min, ∼16% of a 40-min cell cycle.

1.8 Simulated and Analytical Distributions for Constitutively Expressed Genes

Expanded versions of Fig. 2 are shown in Figs. SS5 and SS4 that include values the average gene count as a function of the distance from *ori* to *ter*. In addition to the distributions shown in these figures, we have computed the exact distributions for constitutively expressed genes located at positions spaced every 5% of the way from origin to terminus for fast- and slow-growing cells. Distributions were computed by integrating Eq. S22 assuming a Poisson-distributed mRNA number at each point along the cell cycle, where the mean and variance of the distribution is taken to be Eq. S12. Resulting distributions are shown in Figs. SS6 and SS7.

The TD theory assuming the mRNA relaxes to steady–state before cell division is not exact for all values of f as demonstrated by the relatively poor agreement for the 40 minute doubling time case for genes that duplicate near cell division (see, for instance, Fig. SS8 and Fig. SS6). When a gene duplicates very close to cell division, the mRNA has insufficient time to relax to the high steady-state, and therefore after cell division, the average level does not represent the low steady-state. In fact, the gene must relax after cell division up to the low steady-state, prior to gene duplication. This phenomenon can be seen for a gene located 60% of the way from origin to terminus in Fig. SS9 wherein there are two clear relaxations.

Our model does not capture this behaviour, however, we derived slightly more involved equations that take this into account (Eqs. S36-38).

$$D_{KL}\left(P||Q\right) = \sum_m P\left(m\right) \ln\left(\frac{P\left(m\right)}{Q\left(m\right)}\right) \tag{68}$$

1.9 Comparison of Numerical and Experimental Distributions

To assess the quality of the theory on real world data, it was applied to the various mutation studies reported in Jones et al. and compared against their experimental data [11]. Their data is associated with a gene that spends 1/3 of the cell cycle before gene duplication (f=0.67). The mean of the experimental data was taken to be the time-averaged mean $\langle m\rangle$ of Eq. S29, which was used to compute the mRNA for the low state, $\langle m\rangle_1$ as:

$$\langle m\rangle_1 = \frac{\langle m\rangle}{1 + f + \frac{e^{-fk_d t_D} - 1}{k_d t_D}} \tag{69}$$

The mRNA half-live and the cell doubling time must also be defined in order to compute the theoretical distributions. We took the mRNA half-life to be 5.5 minutes as done in the main text. Jones et al. grew their *E. coli* in M9 minimal salts media supplemented with 0.5% glucose so t_D was taken as 40 minutes (Reshes *et al.* reported an *E. coli* doubling time of 38±1 min for cells grown in M9 salts+0.4% glucose [13]). Using these parameters for k_d and t_D, $\langle m\rangle_1$ was computed via Eq. S69, and the exact distributions of mRNA from the time-dependent theory were computed by integrating Eq. S22 considering only the case of constitutively expressed mRNA.

In the case of the time-independent theory, Eq. S69 reduces to:

$$\langle m\rangle_1 = \frac{\langle m\rangle}{1 + f} \tag{70}$$

as found in Jones et al. [11]. We used this mean to compute the time–independent distribution as:

$$P\left(m\right) = f\text{Pois}\left(2\langle m\rangle_1\right) + \left(1 - f\right)\text{Pois}\left(\langle m\rangle_1\right) \tag{71}$$

Figure SS10 shows the comparison of these distributions to experiments. The resulting time-dependent distributions better represent the data than do those of the time-independent theory; capturing the shape both qualitatively and quantitatively. As discussed in the main text (see Figure 4), the time-dependent theory becomes more important as the mean mRNA becomes large ($\langle m\rangle_1 < 1.0$), and this holds true when comparing to experimental data. However, the time-dependent theory is even quantitatively better for experiments with mean mRNA smaller than 1 (see Figure SS10).

In order to quantify the agreement, we compute the mean-squared deviation (MSD) between the computed and experimental distributions as:

$$MSD = \frac{1}{N}\sum_{m=0}^{N}\left(P_{\text{theory}}\left(m\right) - P_{\text{exp}}\left(m\right)\right)^2 \tag{72}$$

The results are shown in Fig. SS11A. Indeed the MSD was smaller using the time-dependent theory in every case, sometimes up to a factor of 10x, verifying that the theory is appropriate. A KL-divergence, computed neglecting contributions where the experimental probability is zero, qualitatively shows the same picture (Fig. SS11B).

As an independent validation, we compared the experimental distribution, computed by pooling the data from two independent experiments, of the messenger RNA *ptsG* acquired via smFISH of *E. coli* cells growing in 0.2% glucose (see [14] for experimental methods). In this case, the gene is located 25% of the way form the origin to the terminus and the cells had doubling times ∼40 minutes; therefore f=0.34375. The degradation rate k_d was previously measured to be 0.246±0.049 min^{-1} [14]. This time, the theory based on constitutive expression cannot capture the mRNA distribution (see Figure SS12C). This was attributed to the fact that *ptsG* is known to be under regulatory control [15]. To demonstrate the utility of the regulated theory (Eqs. S42–47), we used the mean and Fano factor calculated from the experimental distribution to constrain k_{on} and k_{off}.

We begin by computing $\langle m\rangle_1$ (according to Eqn. 69) and noting:

$$\langle m \rangle_1 = \frac{k_{\text{on}}}{k_{\text{on}} + k_{\text{off}}} \frac{k_t}{k_d} \tag{73}$$

We can solve this for α, the ratio of k_{off} to k_{on}, as:

$$\alpha = \frac{k_t}{k_d \langle m \rangle_1} - 1 \tag{74}$$

From the experimental distribution we can also compute the Fano factor; by subtracting off the contributions associated with gene duplication and RNAP variability we can arrive at the Fano factor contribution associated with regulation (see Eqns. 7 and 8 in the main manuscript):

$$\begin{aligned} \text{Fano}_{\text{reg}} \approx & \frac{\text{Var}[m]}{\langle m \rangle} - 1 + \langle m \rangle - \frac{\langle m \rangle}{10} \\ & - \frac{\langle m \rangle_1^2}{\langle m \rangle} \left[1 + 3f + \frac{8e^{-fk_d t_D} - e^{-2fk_d t_D} - 7}{2k_d t_D} \right] \\ = & \frac{k_t k_{\text{off}}}{(k_{\text{on}} + k_{\text{off}})(k_{\text{on}} + k_{\text{off}} + k_d)} \end{aligned} \tag{75}$$

which we can then solve for k_{on} and k_{off}:

$$\begin{aligned} k_{\text{on}} &= \frac{1}{1+\alpha} \left[\frac{k_t \alpha}{(1+\alpha)\text{Fano}_{\text{reg}}} - k_d \right] \\ k_{\text{off}} &= \alpha k_{\text{on}} \end{aligned} \tag{76}$$

Because of the constraints on k_{on} and k_{off}, the fitting problem reduced to a single dimensional scan over values for k_t, and therefore the computational time required to fit the distribution was greatly reduced. Simulations demonstrate that $k_t = 3.95 \pm 0.1 min^{-1}$ best fit the distribution (Figs. SS12A–C). This parameter was found to be robust to bin size for the data (Fig. SS12C).

Uncertainty in the input data will make the solution to the fitting problem non-unique. In order to test this effect on the fit, we varied k_d within $\pm 1\sigma$ of the mean value reported and ran linear scans over reasonable k_t value to identify the optimal fit transcription rate and regulation parameters k_{on} and k_{off}. Various metrics comparing simulated distributions of mRNA compared to experimental distributions in Fig. SS13. All metrics demonstrate that the optimal solution falls on a line in $k_t - k_d$ plane that corresponds to values for k_{on} and k_{off} that are indistinguishable, within uncertainty, of the values obtained from fitting to the average k_d (p=0.256 for k_{on} and p=0.892 for k_{off} by t-test; see Fig. S13).

1.10 Results of Simulations Including Regulation and RNAP Variability

Comparisons of simulations including regulated gene expression with $k_{\text{on}} = k_{\text{off}} = 0.2/min$ with the theories are shown in Fig. SS14. In these simulations k_t was taken to be $2.52 min^{-1}$ to maintain the same averages seen as in the constitutive expression, so that resulting noise can be compared. Again, agreement is nearly exact.

Contributions of extrinsic noise to the total noise was approximated by including RNAP variability in simulations. The average RNAP were taken to be 2500 and 5500 for cells doubling in 70 and 40 minutes, respectively [16]. RNAP distributions were modelled as Γ-distributions with the shape parameter of 10 and scale parameters of 250 (40 min) or 550 (70 min). A total of 2000 cells were simulated in each case and for each cell a single RNAP count was sampled from the respective Γ-distribution and held constant for ten cell cycles. Simulation results for the average and Fano factor are compared to the analytical theory assuming now that the variation affects k_t (Eqn. 8 in the main manuscript or Eqn. S63) in Figure SS15. This type of noise is accurately captured by the theory.

2 SI Figures

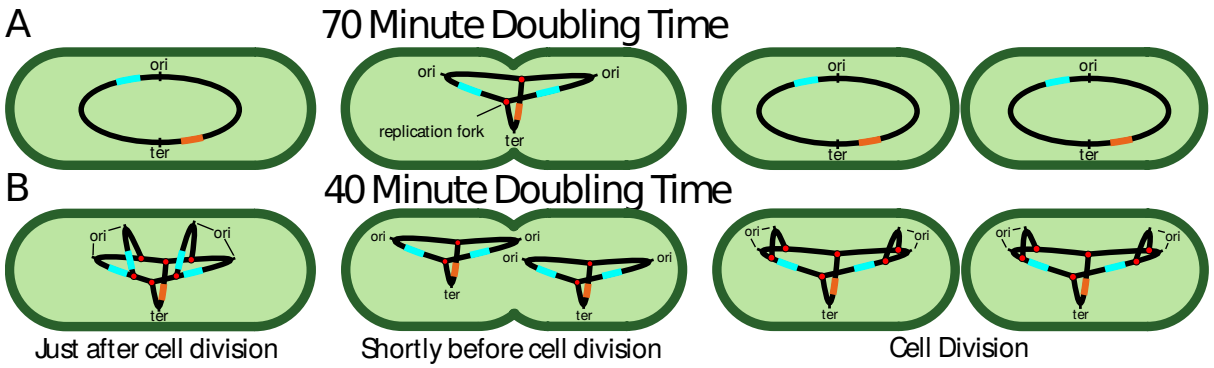


Fig. S1: **Replication Schematics** A schematic showing the replication of DNA containing one gene close to the origin (blue) and one close to the terminus (orange) at various timepoints in the cell cycle. Replication proceeds from the origin (*ori*) to the terminus (*ter*) and multiple replication forks (red dots) can exist simultaneously. Snapshots through the cell cycle from cells with doubling times (A) slower (t_D = 70 minutes) and (B) faster (t_D = 40 minutes) than the DNA replication time (45 minutes) are shown. For slow growing cells the initiation of replication occurs shortly after cell division and completes before the cell divides. For cells growing faster than the replication time, multiple copies of the genome must exist and therefore the number of replication forks can change dramatically throughout the cell cycle. The effect on gene count depends on the gene location; for instance a gene close to the origin is duplicated durign the same cell cycel that the replication is initiated, resulting in 2 or 4 copiles of the gene (A, middle). Conversely, a gene close to the terminus is replicated in the next cell cycle and only 1 or 2 copies can exist (B, right)

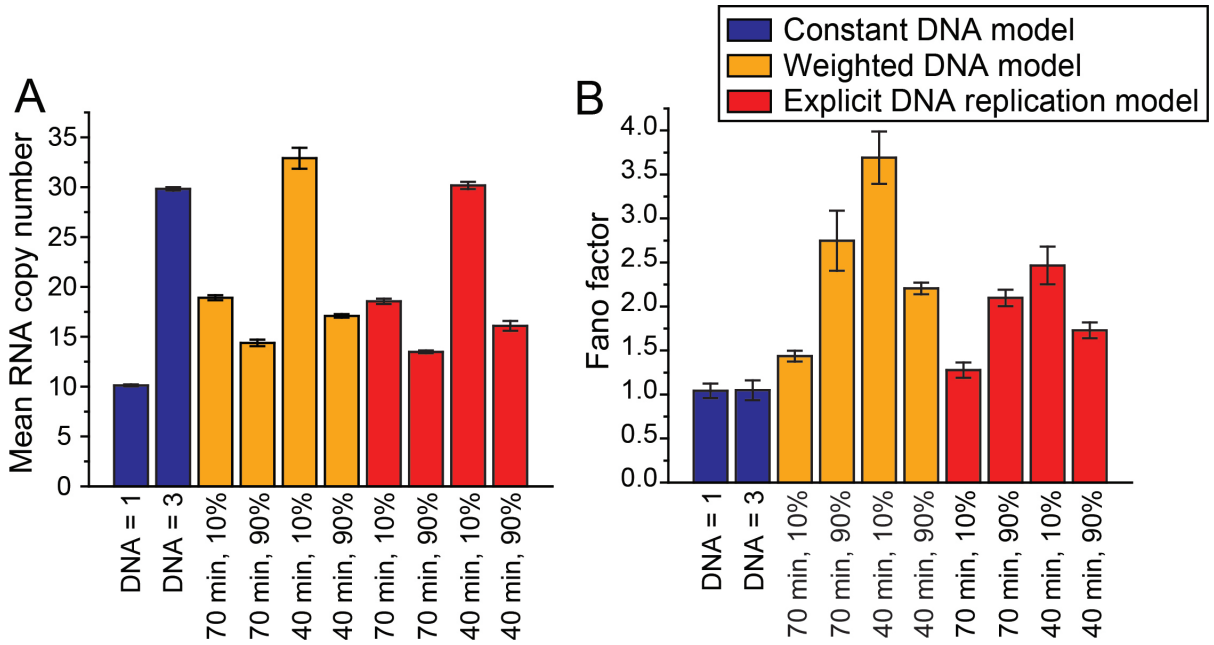


Fig. S2: **Division Time Contribution** Comparison of various average mRNA levels (A) and their associated Fano factors (B) for different treatments of gene copy number in stochastic simulations. The "Constant DNA model" assumes that there is only one gene copy number and all cells in the population have that number over all time. The "Weighted DNA model" is equivalent to the time-independent theory, in that each cell is considered to have either a high or a low count of the gene based on the fraction of time after gene replication, f, and assumed to have that copy nubmer for all time. The "Explicit DNA replication model" is that of the time-dependent theory, where the gene is duplicated during the simulation and the mRNA is allowed to relax to the new steady-state. Simulations with genes at different locations (10 or 90% of the distance from the origin to terminus) at two doubling times are considered. The noise observed in the explicit replication model is consistently lower than that in the weighted DNA model.

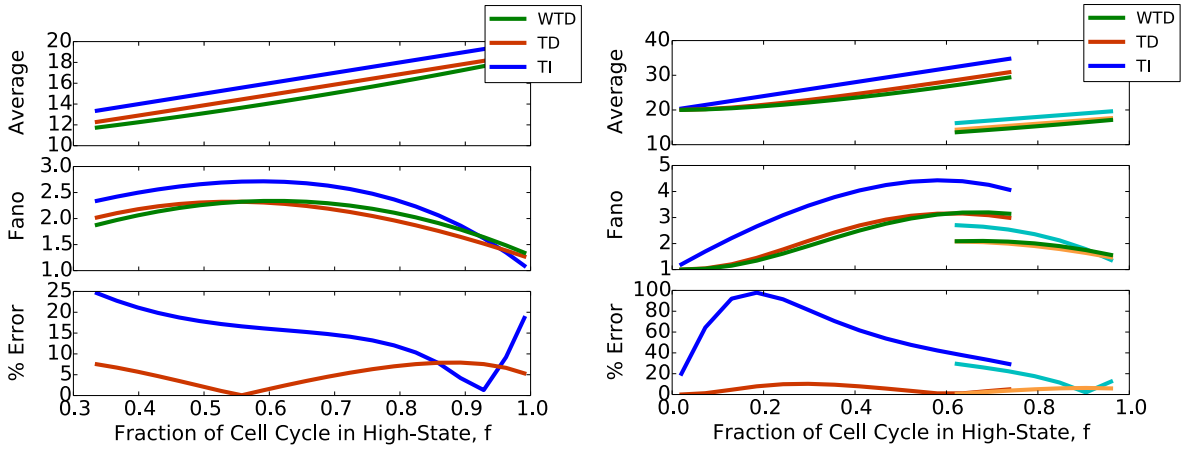
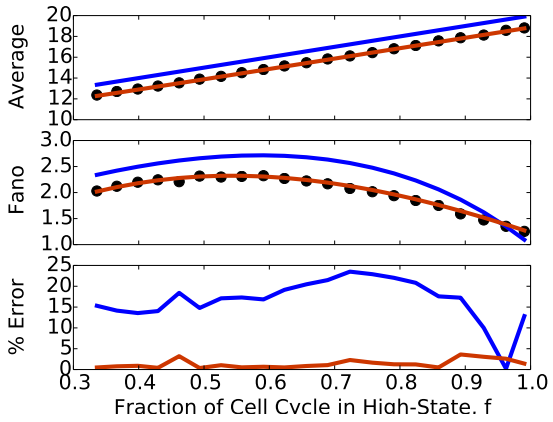
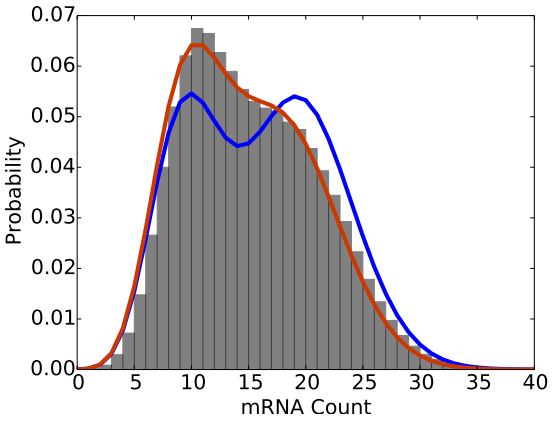


Fig. S3: **Cell–Age Weighted Results** Comparison of average mRNA and Fano factor predicted by theories with (orange/light orange lines) and without (blue/cyan lines) accounting for time dependent mRNA to the theory that weights the results with exponentially distributed cell ages (green) for cells doubling in (a) 70 minute and (b) 40 minutes. The form of the weighted time-dependent theory (WTD) is based on Eq. S28. The bottom plot shows errors of the TD and TI theories relative to the exponentially weighted TD theory (green lines), which are generally below 8%.



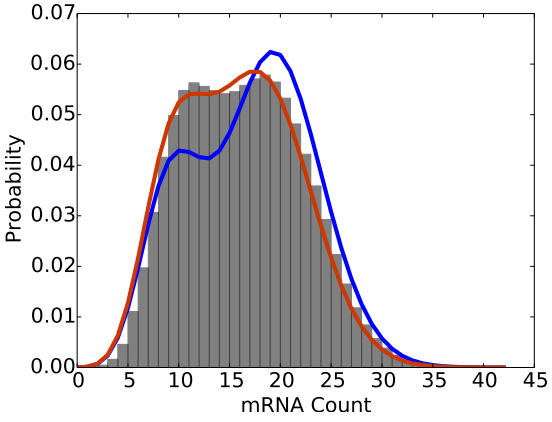
(a)



(b)

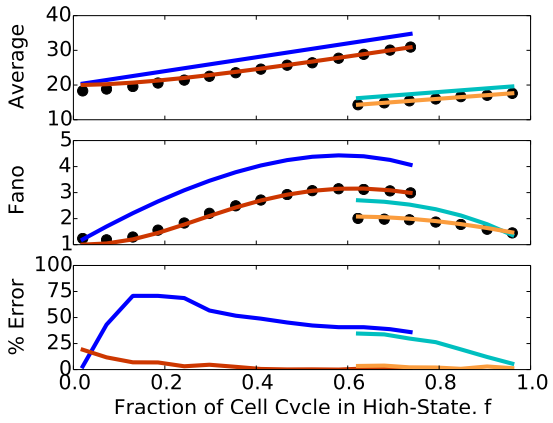
Gene locus (% from *ori*)	70 minute cell cycle		
	N = 1	N = 2	Mean
1	0.0064	0.9936	1.9936
10	0.064	0.936	1.936
20	0.129	0.871	1.871
30	0.193	0.807	1.807
40	0.257	0.743	1.743
50	0.312	0.679	1.679
60	0.386	0.614	1.614
70	0.45	0.55	1.55
80	0.514	0.486	1.486
90	0.579	0.421	1.421
100	0.643	0.357	1.357

(c)

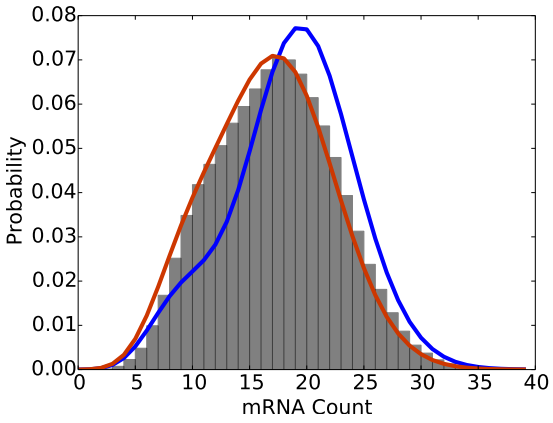


(d)

Fig. S4: **70 minute doubling time** A) Comparison of average mRNA and Fano factor predicted with (orange lines) and without (blue lines) accounting for time-dependent mRNA relaxation to results from exact simulations (points). The time-dependent theory shows nearly exact agreement in all cases. When comparing numerically computed distributions for genes that spend (B) 61% and (D) 74.3% of the cell cycle in the high state, to simulated distributions (gray histograms) it becomes apparant that including time-dependence (orange lines) better captures both qualitatively and quantitatively the data than does the time-independent theory (blue lines).



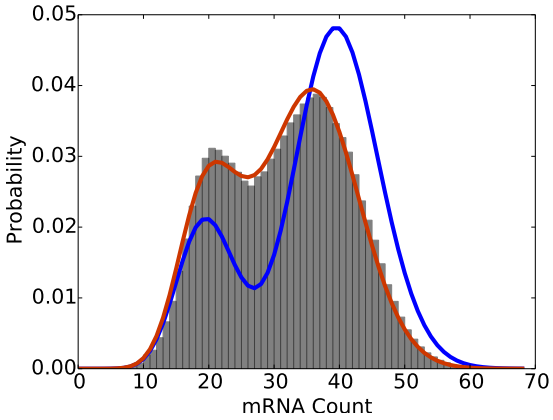
(a)



(b)

Gene locus (% from *ori*)	40 minute cell cycle			
	N = 1	N = 2	N = 4	Mean
1		0.261	0.739	3.478
10		0.3625	0.6375	3.275
20		0.475	0.525	3.05
30		0.5875	0.4125	2.825
40		0.7	0.3	2.6
50		0.8125	0.1875	2.375
60		0.925	0.075	2.15
70	0.0375	0.9625		1.9625
80	0.15	0.85		1.85
90	0.2625	0.7375		1.7375
100	0.375	0.625		1.625

(c)



(d)

Fig. S5: **40 minute doubling time** A) Comparison of average mRNA and Fano factor predicted with (orange/light orange lines) and without (blue/cyan lines) accounting for time-dependent mRNA relaxation to results from exact simulations (points). Darker orange and blue lines represent genes that are duplicated during the cell cycle when the replication of that genome is initiated, and therefore have either 2 or 4 gene copies. Lighter orange and cyan lines represent genes that are duplicated in the cell cycle following the one in which the replication was initiated, and therefore either 1 or 2 gene copies exist. The time dependent theory shows nearly exact agreement. Comparisons of the mRNA distribution from simulation (gray histogram) to theories with (orange lines) and without (blue lines) time-dependence demonstrates the advantange of considering the mRNA relaxation for genes that spend (B) 27.5% and (D) 62.5% of their time in the high state.

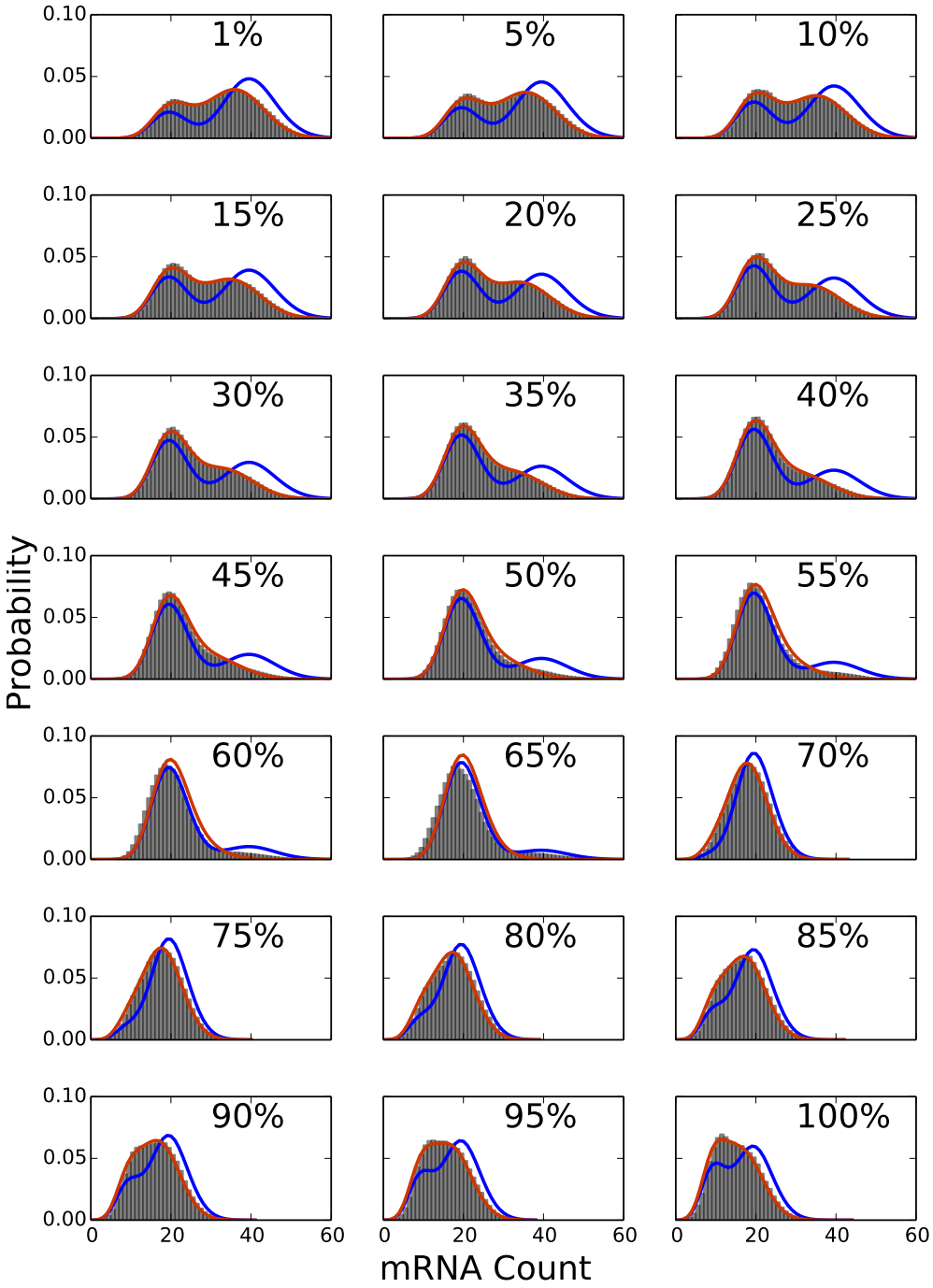


Fig. S6: **40 min Doubling Time Distributions** Results for distributions computed via theory for a cell doubling every 40 minutes for a genes located at the indicated positions between the origin and the terminus. Distributions computed with the analytical theory (orange lines) nearly exactly represent simulations (gray distributions) whereas distributions computed via the time-indepenent model (blue lines) often qualitatively predict strong bimodal behavior, where none should exist. In all cases, the time-dependent theory is superior as demonstrated in Figure S8. However, as discussed in the main text, the comparison becomes worst between 50-60% of the way along the genome, due to inadequite time to relaxation to the high-state.

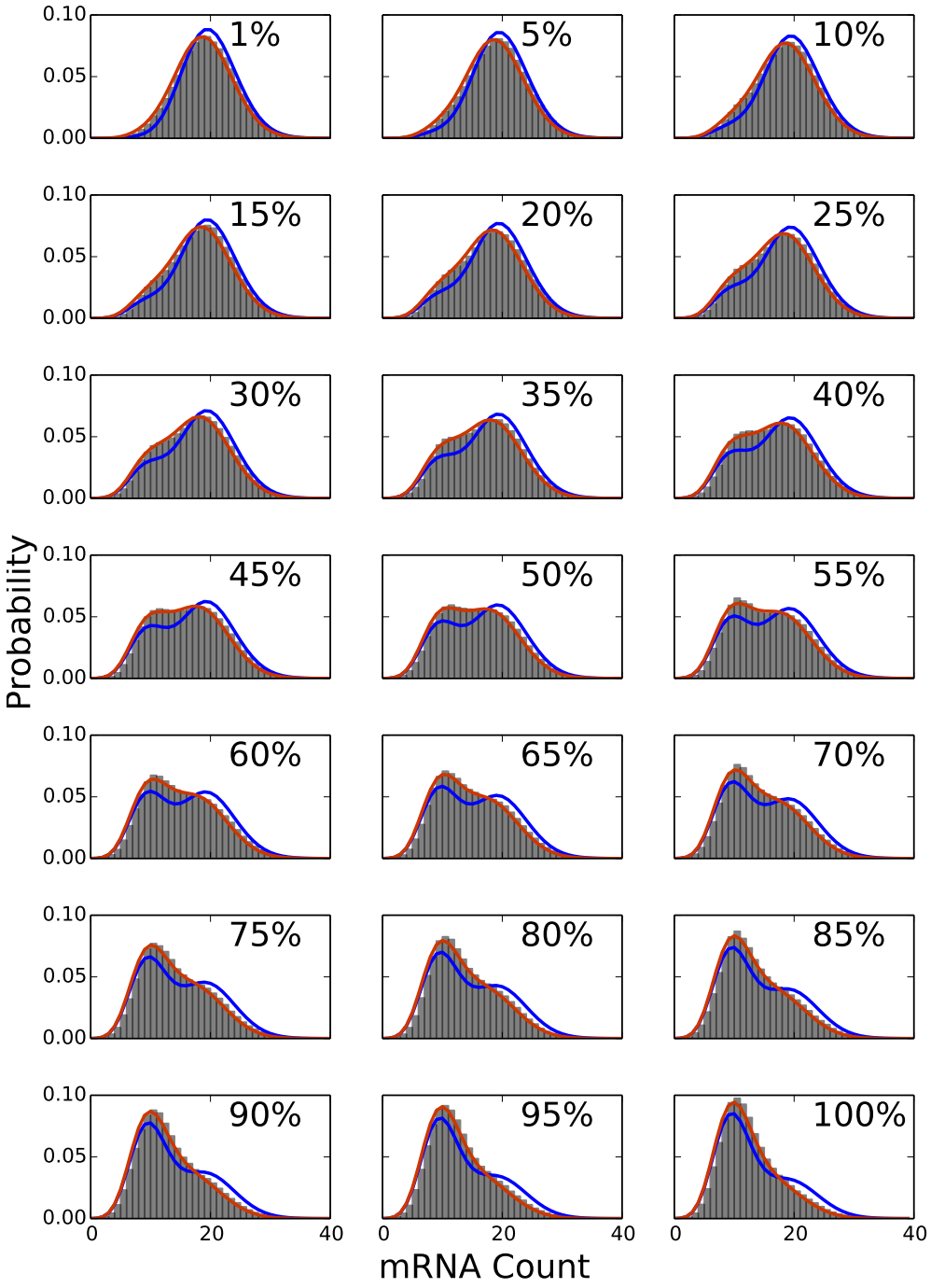
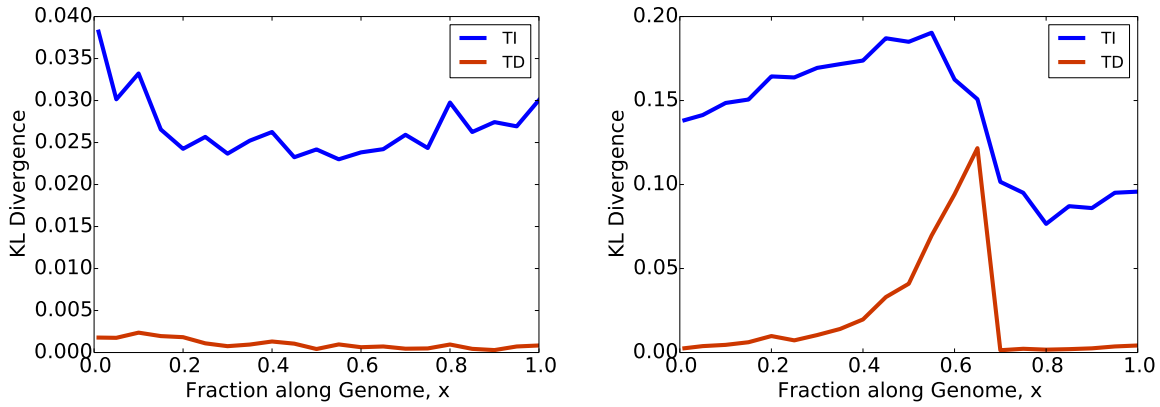


Fig. S7: **70 min Doubling Time Distributions** Results for distributions computed via theory for a cell doubling every 70 minutes for a genes located at the indicated positions between the origin and the terminus. Distributions computed with the analytical theory (orange lines) nearly exactly represent simulations (gray distributions) whereas distributions computed via the time-indepenent model (blue lines) often qualitatively predict strong bimodal behavior, where none should exist. In all cases, the time-dependent theory is superior as demonstrated in Figure S8.



(a) (b)

Fig. S8: **Goodness of Fit** Kullback-Leibler divergence of theory from simulation computed for doubling times of (a) 70 and (b) 40 minutes. The theory incorporating time-dependent mRNA dynamics (orange lines) better captures the mRNA distribution than does the static theory (blue lines). The rapid rise in divergence in the 40 minute doubling time comparison is due to the fact that mRNA from genes that duplicate close to division time does not have enough time to relax to the new steady-state prior to division, thus starting the next cell cycle away from steady-state (see Fig. S9 description in the main text).

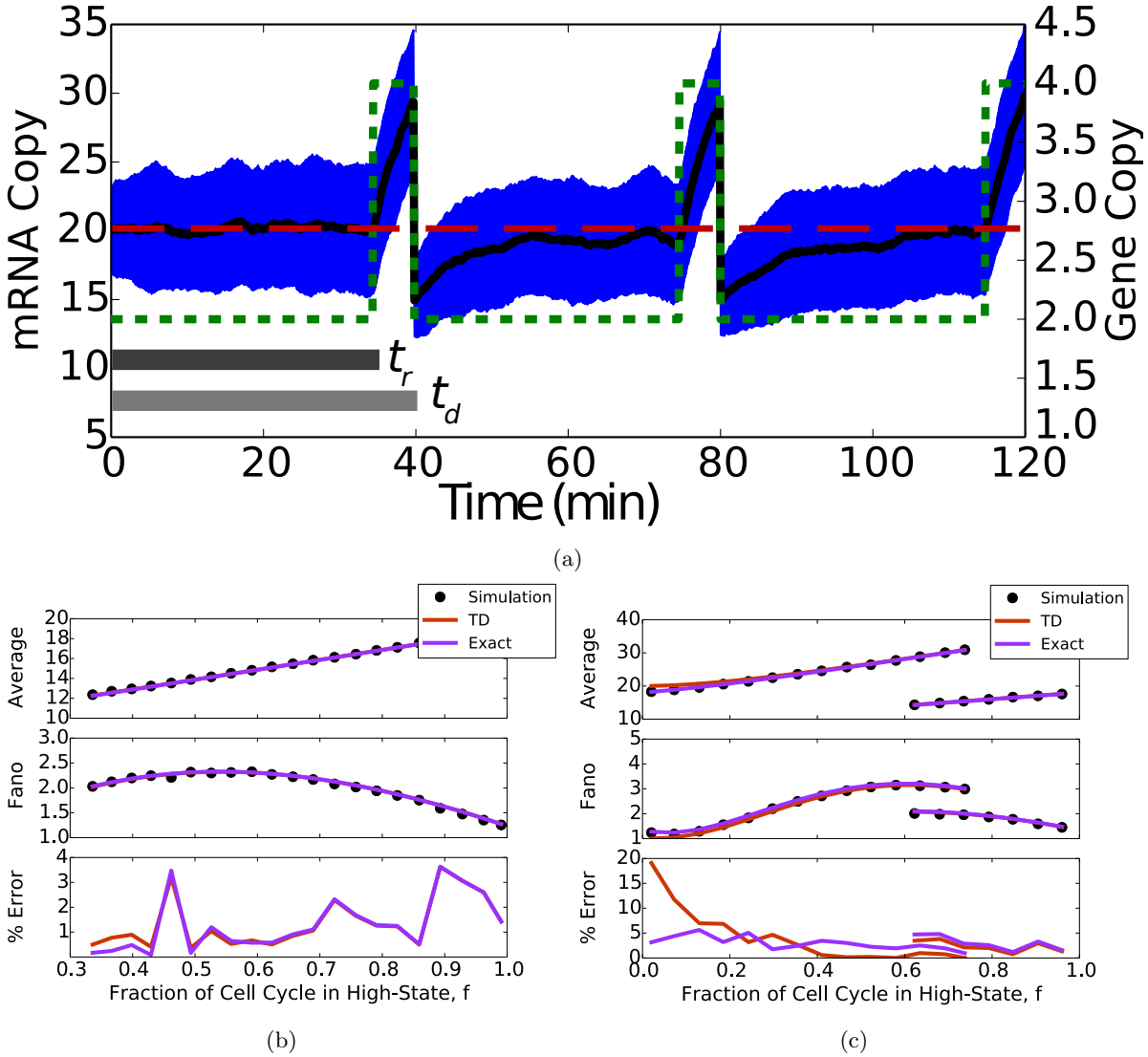


Fig. S9: **Deviation Near Division** A) A schematic composed of 200 simulation replicates showing the progress of the average mRNA (black line) levels before and after a gene duplication event (green dotted line). The area encompassing the average $\pm 1\sigma$ (blue). As can be seen, replication is followed by relaxation of the mRNA from an initially low level but does not relax to a new steady-state level prior to cell division, and therefore the cells begin their next cell cycle with a non-steady-state mRNA distribution. Only about half-way through the next cell cycle does the mRNA approach steady-state (red dashed line). In a case such as this, the TD theory put out in the paper will deviate, as it was derived with the assumption that the mRNA reaches steady state before division. A modified TD theory lifts this assumption allowing for more accurate estimation of the average, variance, Fano factor, and mRNA distributions (Eqs. S35-37). The doubling time (t_D) was taken to be 40 minutes, the total DNA replication time was taken to be 45 minutes, the gene was positioned 55% of the way from the origin to the terminus ($t_r \approx 35$ minutes), the transcription rate k_t was 1.26 min^{-1} and the degradation rate k_d was $0.126 min^{-1}$. The assumption that the mRNA level relaxes to steady-state prior to cell division can be lifted and "Exact" equations can be derived (purple lines; Eqs. S37-38) that better capture the Fano factor, especially for genes that replicate close to cell division (f <0.1), as demonstrated for the 70 minute (B) and 40 minute (C) cases.

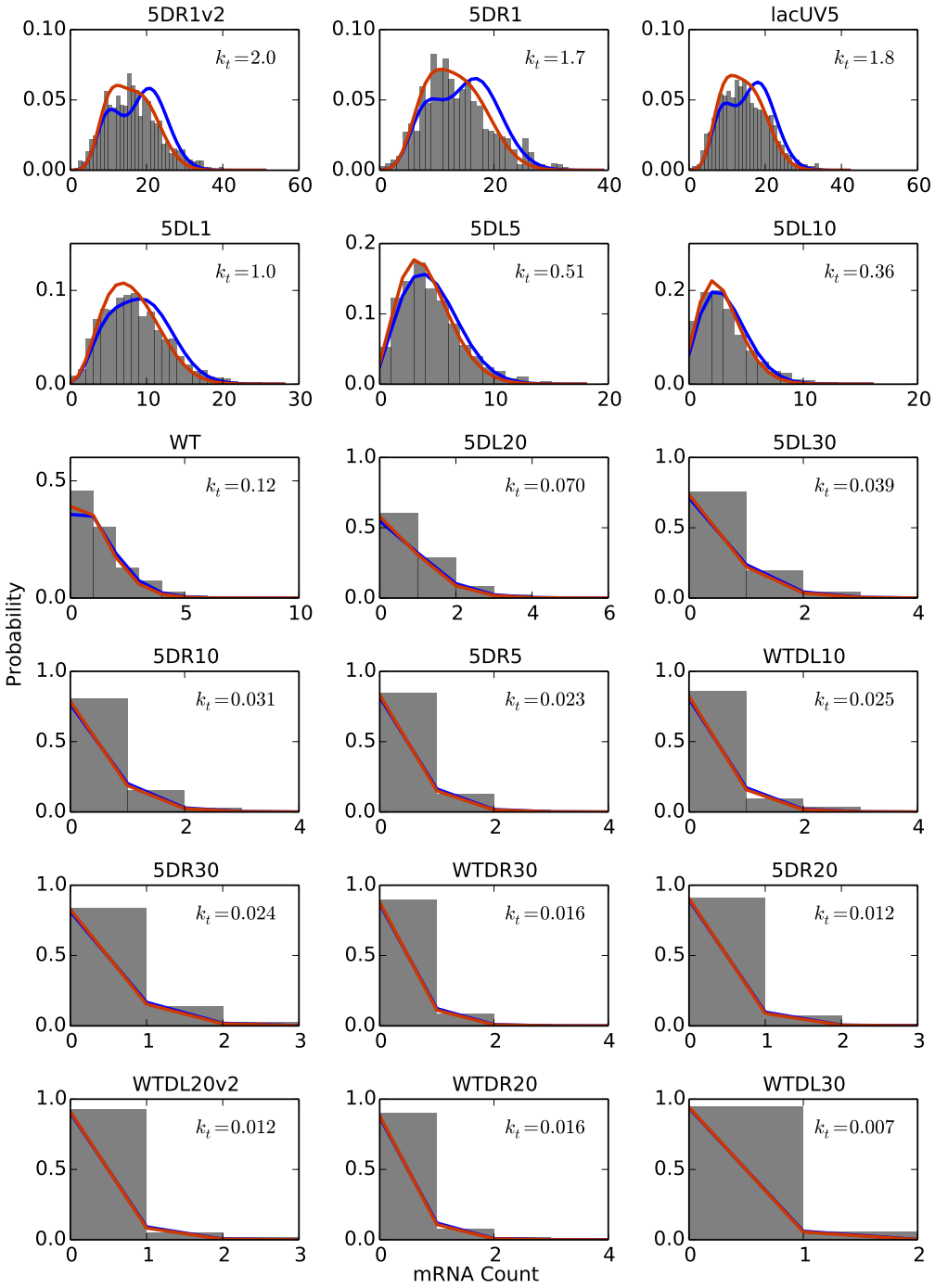
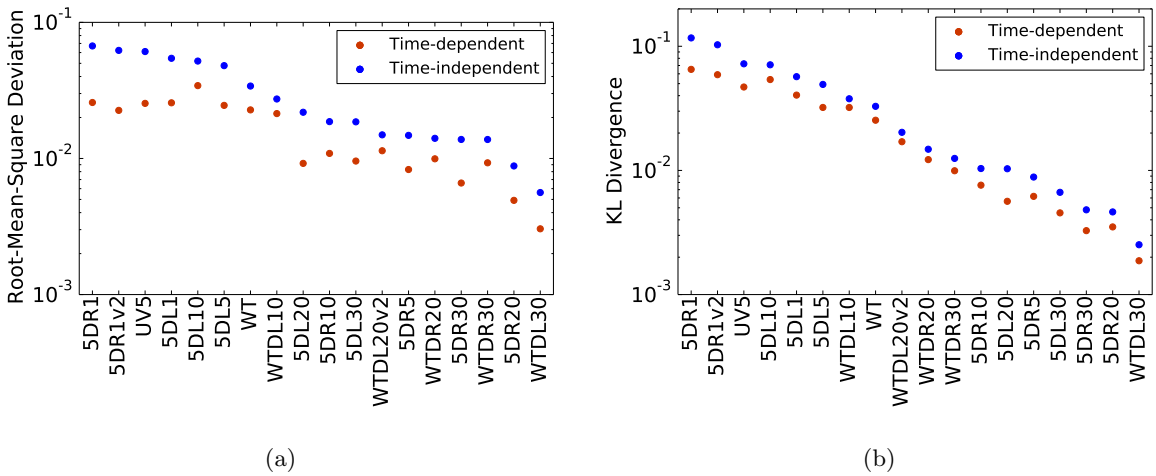


Fig. S10: **Comparison to Experimental Distributions** Comparison of time-dependent (orange) and time-independent (blue) theories for the mRNA distribution to experimental data of Jones et al. [11]. Theoretical curves are computed taking half-life and doubling times of 5.5 and 40 minutes, respectively, for a gene that is in the high-state for 2/3 of the cell cycle. The mean from a single gene copy was computed as disscussed in Section S1.9; the associated transcription rate is shown in the figure in units of per-minute.



(a) (b)

Fig. S11: **Distribution Comparisons** A) Mean squared deviation (MSD) computed (via Eq. S72) between the predicted and experimental mRNA distributions of Fig. SS10. B) KL-divergence computed (via Eq. S68) on the same data. Both metrics demonstrate that the TD theory better represents the data than does the TI theory. Smaller MSD or KL indicates better agreement.

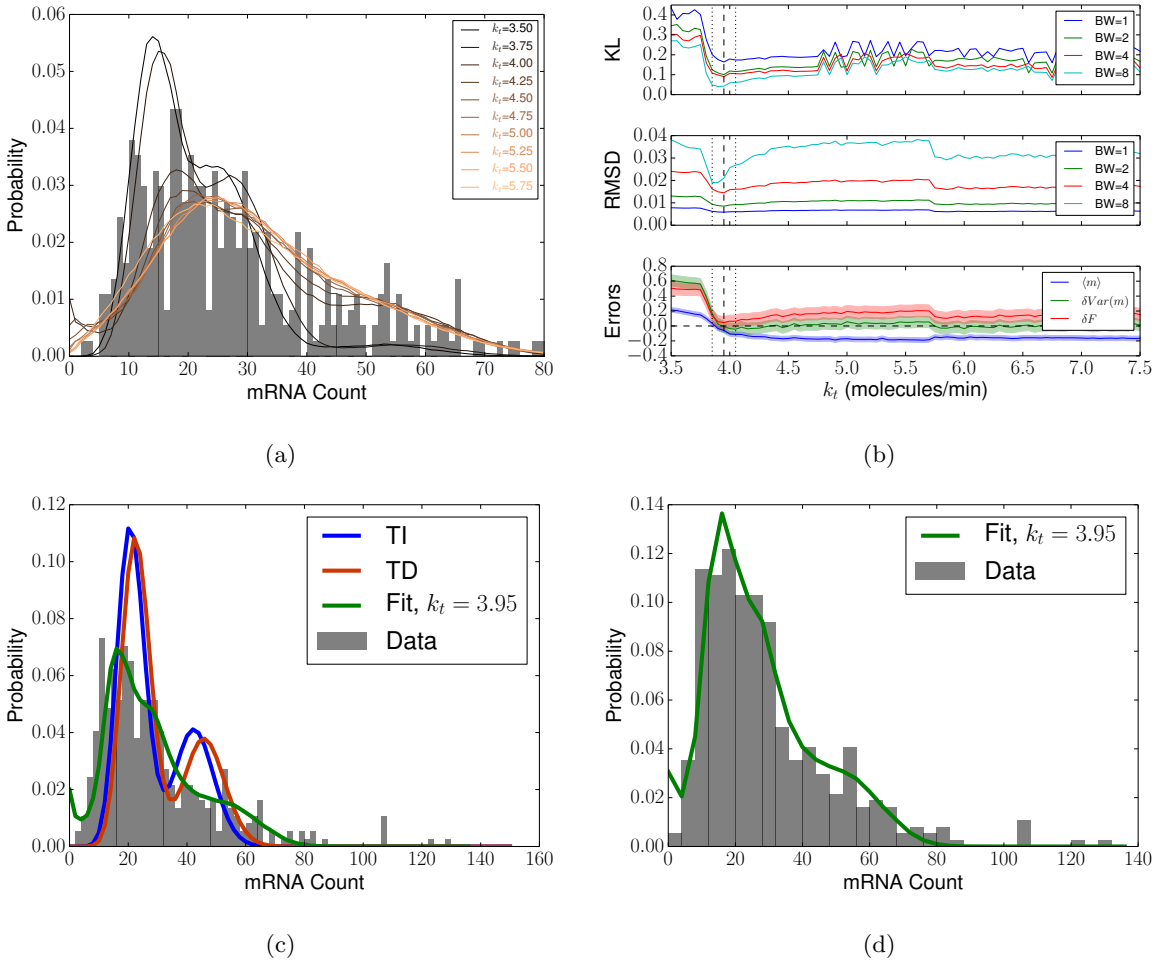


Fig. S12: **Fit to *ptsG* Distribution** Fitting of the experimental distribution (of two pooled replicates) for *ptsG* via simulations while constraining k_{on} and k_{off} via the regulated theory (Eqs. S42–47). A) By varying k_t distributions were simulated. B) Distributions were compared with experimental data using various metrics such as KL divergence (top) and RMSD (mid) at various binning widths (BW), from which the most optimal k_t was chosen (dashed vertical line; ± 1 SEM dotted vertical lines), that also minimizes the error in the mean, variance and Fano factor. The corresponding regulation rates are $k_{\text{on}} = 0.023 min^{-1}$ and $k_{\text{off}} = 0.0084 min^{-1}$. This model significantly outperforms the constitutive theory as shown for comparisons with bin widths of C) 2 and D) 4. The noise of the observed distribution can only be caputed with this model as *ptsG* is a highly regulated gene [15]. Using the analytical theory, only 1 parameter was varied; therefore the effort expended on fitting was significantly reduced. Other model parameters include a half-life of 2.8 minutes for a gene located 25% from the *ori*, corresponding to $f \approx 0.35$ [14].

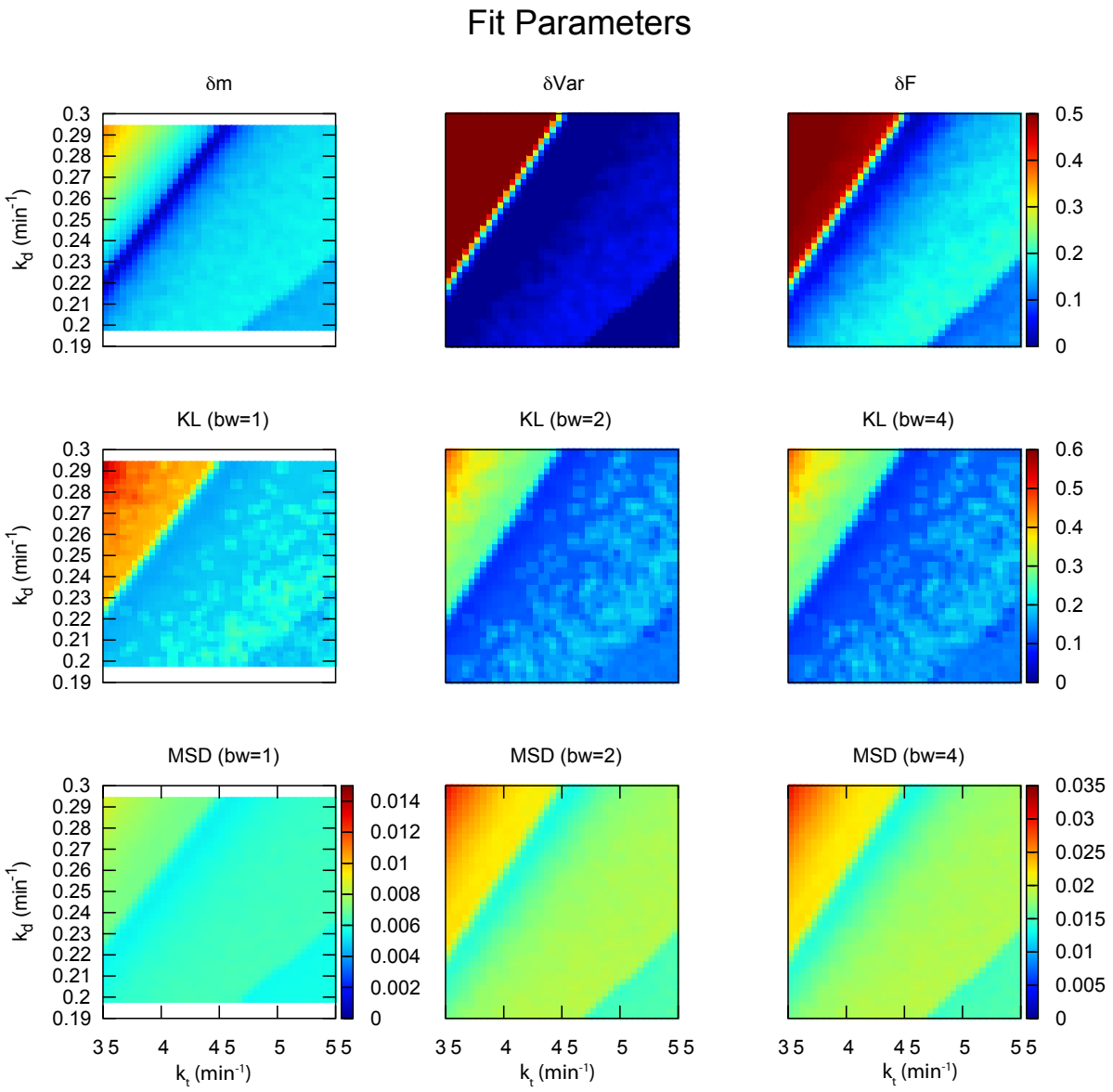
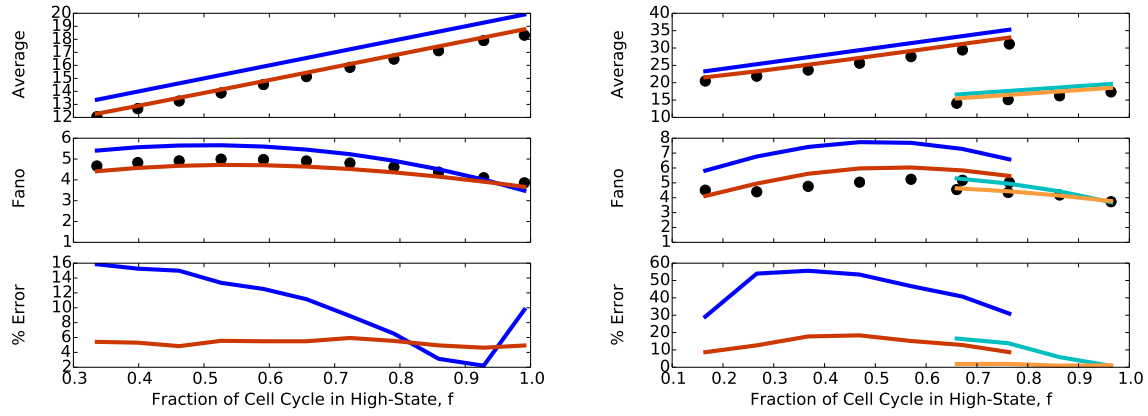


Fig. S13: **Fitting Statistics** Various measures of fit agreement between simulated and experimental distributions as k_d is varied within 1σ of the average plotted versus the free parameter k_t. (Top Row) Absolute value of the relative error in the mean δm, variance δVar and Fano factor δF for the simulated parameter sets. (Middle Row) Computed Kullback Leibler-divergence for various histogram binning widths. (Bottom Row) Computed mean–squared–deviation for various histogram binning widths.



(a) (b)

Fig. S14: **Approximated Regulated Noise** Comparison of average mRNA and Fano factor predicted by theories with (orange/light orange lines) and without (blue/cyan lines) accounting for time dependent mRNA to simulation results (points) for (a) 70 minute and (b) 40 minute doubling times. The form of the time-dependent theory is based on the approximate corrections of Eq. S47.

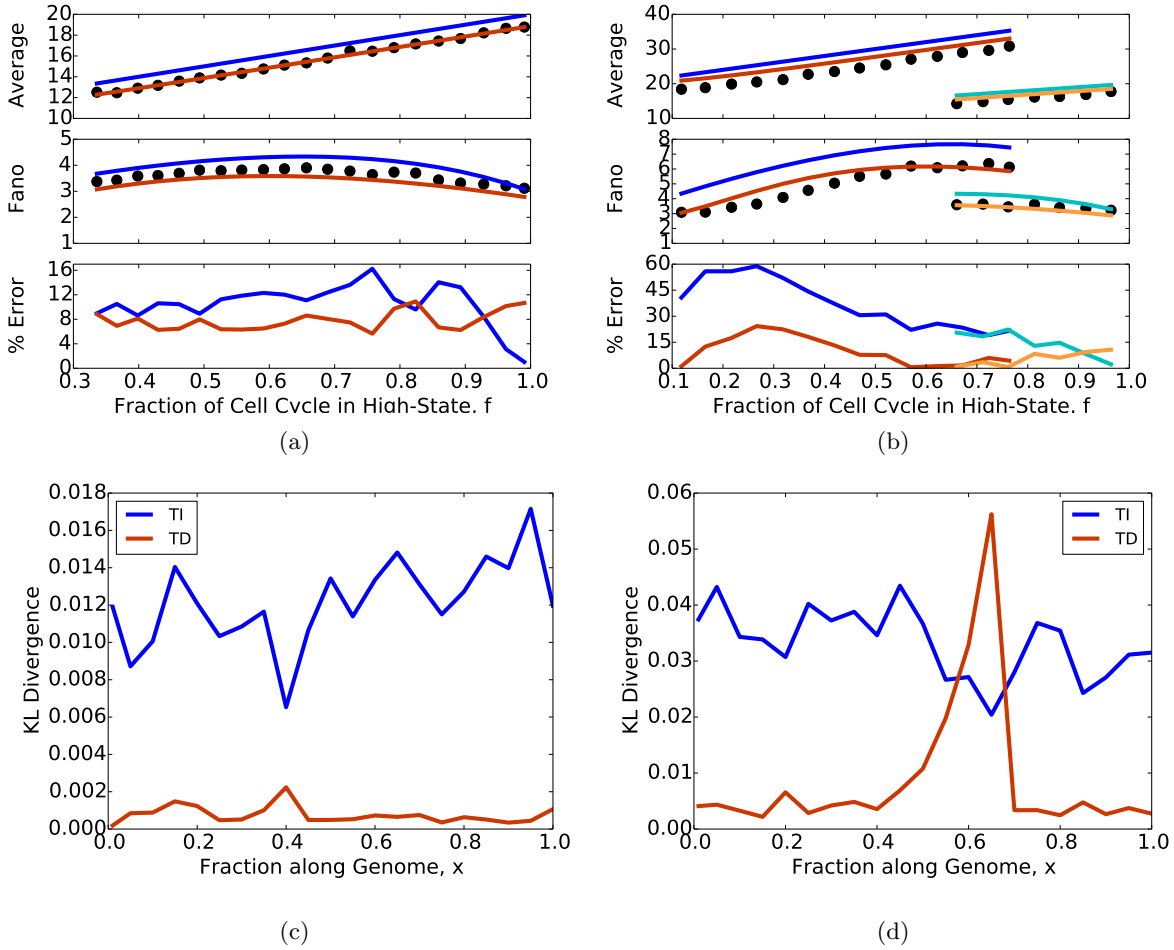


Fig. S15: **RNAP Noise** A comparison of theory with simulations where extrinsic noise is entirely approximated by variations in the RNAP number, and consequently variation in the apparent transcription rate k_t. (A) 70 minute and (B) 40 minute doubling times. The Kullback-Leibler divergence comparing numerically computed distributions with simulated distributions are shown for (B) 70 minute and (D) 40 minute doubling times demonstrating that incorporation of the time-dependent mRNA relaxation is required even when considering extrinsic effects such as RNAP fluctuations. The time-dependent theory is calculated using the approximate solution Eq. S63. The time-independent theory is based on Eqs. S33 and S54 of Jones et al. [11].

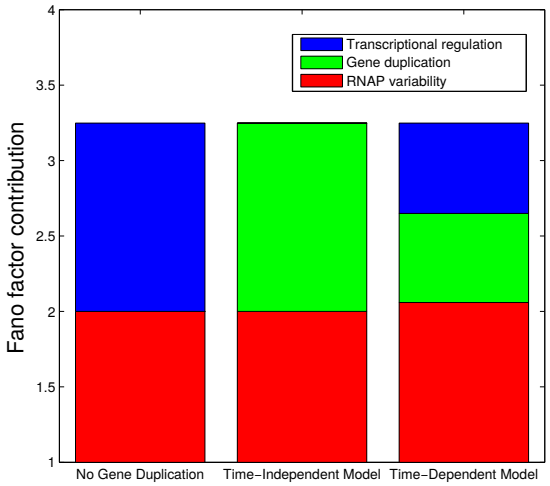


Fig. S16: Noise contributions assessed using different models of mRNA noise. In each case it is assumed $t_D = 40$ min, $\langle m \rangle = 10$, $f = 0.35$, $k_d = 0.126$ min^{-1}, and Fano$[m] \approx 3.25$. In the left-most bar, the noise is assumed to originate entirely from RNAP variability or transcriptional regulation. In the central bar, noise contributions are computed according to the time-independent theory. In this case RNAP variability and gene duplication alone completely account for the observed Fano factor; in turn, transcriptional regulation appears not to contribute. In the right-most bar, noise contributions are computed according to the time-dependent theory. In this case we see that the time-independent theory overestimates gene duplication-associated noise, obscurring the fact that some transcriptional regulation is taking place, although not as much as might have been suspected had gene replication not been accounted for.

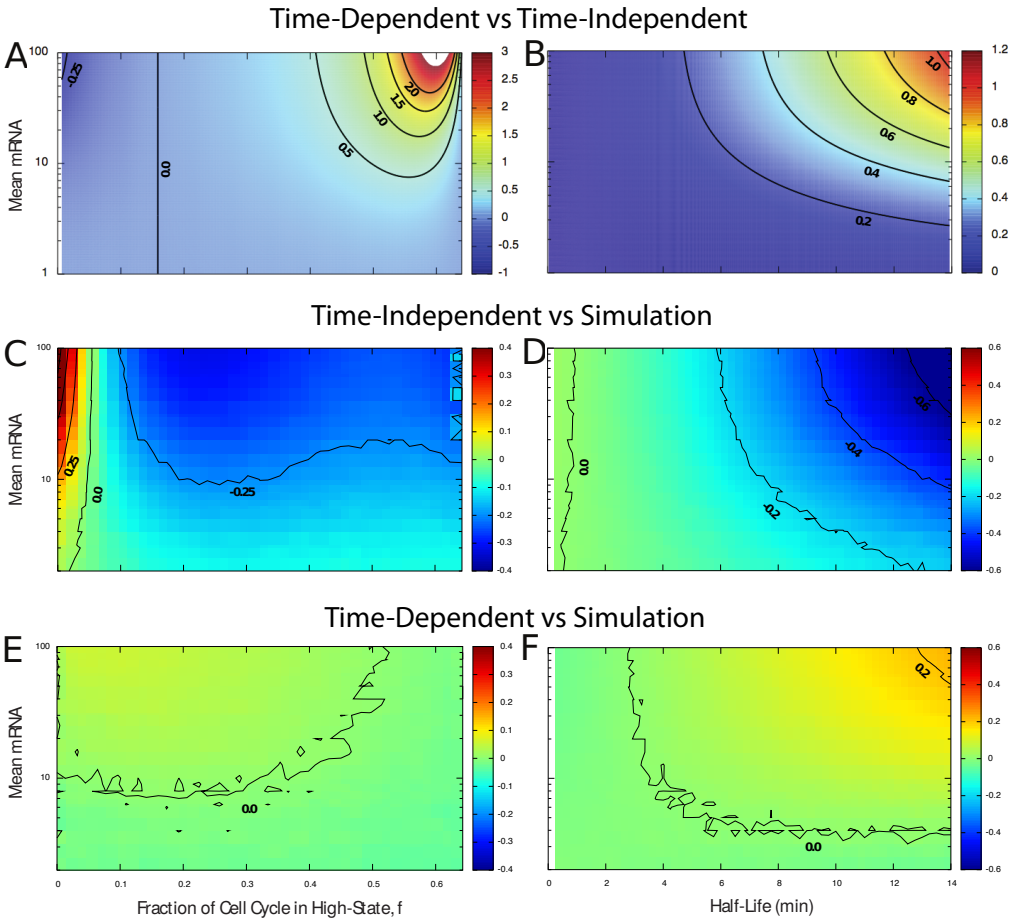


Fig. S17: **Deviation of Theories and Simulations** (A) and (B) as in Fig. 4 where the error is $(F_{TI} - F_{TD})/F_{TI}$. Comparison of the Fano factor computed via theory to stochastic simulations can be seen for the TI $((F_{Sim} - F_{TI})/F_{Sim}$; C & D) and TD $((F_{Sim} - F_{TD})/F_{Sim}$; E & F) expressions. Simulations were averages of 1000 indpendent cell lineages each growing for 10 generations. Contour lines are for the indicated values, and are not smooth due to the variation in the data due to limited sampling. The average deviation of the TD theory is generally less than 20% over the ranges studied, while the TI can deviate by over 60% in the same ranges. The error in E & F can be reduced to zero within numerical uncertainty and sampling error by using Eqs. S35-37 as opposed to Eq. S28.

Supporting References

[1] Powell, E (1956) Growth rate and generation time of bacteria, with special reference to continuous culture. *J Gen Microbiol* 15(3):492–511.

[2] Ho, PY, Amir, A (2015) Simultaneous regulation of cell size and chromosome replication in bacteria. *Front Microbiol* 6:662.

[3] Labhsetwar, P, Cole, JA, Roberts, E, Price, ND, Luthey-Schulten, ZA (2013) Heterogeneity in protein expression induces metabolic variability in a modeled *Escherichia coli* population. *Proc Natl Acad Sci USA* 110(34):14006–14011.

[4] Peccoud, J, Ycart, B (1995) Markovian Modeling of Gene-Product Synthesis. *Theor Popul Biol* 48(2):222–234.

[5] Ozbudak, EM, Thattai, M, Kurtser, I, Grossman, AD, van Oudenaarden, A (2002) Regulation of noise in the expression of a single gene. *Nat Genet* 31(1):69–73.

[6] Raj, A, Peskin, CS, Tranchina, D, Vargas, DY, Tyagi, S (2006) Stochastic mRNA Synthesis in Mammalian Cells. *Plos Biol* 4(10):e309.

[7] Friedman, N, Cai, L, Xie, XS (2006) Linking Stochastic Dynamics to Population Distribution: An Analytical Framework of Gene Expression. *Phys Rev Lett* 97(16):168302.

[8] Shahrezaei, V, Swain, PS (2008) Analytical distributions for stochastic gene expression. *Proc Natl Acad Sci USA* 105(45):17256–17261.

[9] Taniguchi, Y, Choi, PJ, Li, GW, Chen, H, Babu, M, Hearn, J, Emili, A, Xie, XS (2010) Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* 329(5991):533–538.

[10] So, L, Ghosh, A, Zong, C, Sepúlveda, LA, Segev, R, Golding, I (2011) General properties of transcriptional time series in *Escherichia coli*. *Nat Genet* 43(6):554–560.

[11] Jones, DL, Brewster, RC, Phillips, R (2014) Promoter architecture dictates cell-to-cell variability in gene expression. *Science* 346(6216):1533–1536.

[12] Cooper, S, Helmstetter, CE (1968) Chromosome replication and the division cycle of Escherichia coli B/r. *J Mol Biol* 31(3):519–540.

[13] Reshes, G, Vanounou, S, Fishov, I, Feingold, M (2008) Timing the start of division in *E. coli*: a single-cell study. *Phys Biol* 5(4):046001.

[14] Fei, J, Singh, D, Zhang, Q, Park, S, Balasubramanian, D, Golding, I, Vanderpool, CK, Ha, T (2015) Determination of *in* vivo target search kinetics of regulatory noncoding RNA. *Science* 347(6228):1371–1374.

[15] Qaidi, SE, Plumbridge, J (2008) Switching Control of Expression of ptsG from the Mlc Regulon to the NagC Regulon. *J Bacteriol* 190(13):4677–4686.

[16] Klumpp, S, Hwa, T (2008) Growth-rate-dependent partitioning of RNA polymerases in bacteria. *Proc Natl Acad Sci USA* 105(51):20245–20250.