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}

^aDepartment of Chemistry, University of Illinois at Urbana–Champaign, Urbana, IL 61801; ^bDepartment of Physics, University of Illinois at Urbana–Champaign, Urbana, IL 61801; ^cCarl Woese Institute for Genomic Biology, University of Illinois at Urbana–Champaign, Urbana, IL 61801; ^dCenter for the Physics of Living Cells, University of Illinois at Urbana–Champaign, Urbana, IL 61801; ^eHoward Hughes Medical Institute, University of Illinois at Urbana–Champaign, Urbana, IL 61801; and ^fDepartment 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.

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

²To 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:

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 ~ 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. S1B and Table SSC). 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, $\bar{m}(t)$, 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

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

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 $P(m|t) = \text{Pois}(\bar{m}(t))$]. Time averaging over the cell cycle yielded the expressions

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

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

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

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. 3D, 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_r , whereas the off state is silenced. Genes are allowed to switch between states at rates k_{on} and k_{off} ; schematically,

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,

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

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 ($\bar{m}(t)$) 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 $\text{Var}[m]$, 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,

$$\begin{aligned} \text{Fano}[m] &\approx \text{Fano}[m]_{\bar{k}_t} \\ &+ \frac{\langle m \rangle_{1, \bar{k}_t}^2}{\langle m \rangle_{\bar{k}_t}} \frac{\text{Var}[k_t]}{\bar{k}_t^2} \left[1 + 3f + \frac{8e^{-fk_{atD}} - e^{-2fk_{atD}} - 7}{2k_{atD}} \right]. \end{aligned} \quad [8]$$

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. $\bar{k}_t$ then represents the mean value of k_t , $\langle m \rangle_{\bar{k}_t}$ and $\text{Fano}[m]_{\bar{k}_t}$ are the mean mRNA number and Fano factor evaluated according to Eqs. 3 and 5 assuming $k_t = \bar{k}_t$, and $\langle m \rangle_{1, \bar{k}_t} = \bar{k}_t / k_d$ 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 $\sim 3,000$ 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 $\bar{k}_{\text{on}}$ values (where $\bar{k}_{\text{on}}$ represents the mean value of k_{on}), tended to be relatively small. We found it approached $\langle m \rangle / 10$ only when $k_{\text{off}} \gg \bar{k}_{\text{on}}$, and in cases where $k_{\text{off}} \lesssim \bar{k}_{\text{on}}$ we found this correction remained well below $\sim 3\%$ 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_i 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_i . Eqs. 5, 7, and 8 can be used to solve for k_{on} and k_{off} as functions of k_i (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 $k_i = 3.95 \text{ min}^{-1}$, $k_{on} = 0.023 \text{ min}^{-1}$, and $k_{off} = 0.0084 \text{ min}^{-1}$, 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 experimental 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 ($t_d \sim 70 \text{ min}$). 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. 4A), 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 be 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-MCA035027) provided by Extreme Science and Engineering Discovery Environment, which is supported by NSF Grant ACI-1053575.

- Thattai M, van Oudenaarden A (2001) Intrinsic noise in gene regulatory networks. *Proc Natl Acad Sci USA* 98(15):8614–8619.
- Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. *Science* 297(5584):1183–1186.
- 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.
- Kaern M, Elston TC, Blake WJ, Collins JJ (2005) Stochasticity in gene expression: From theories to phenotypes. *Nat Rev Genet* 6(6):451–464.
- 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.
- 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.
- 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.
- 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.
- Acar M, Mettetal JT, van Oudenaarden A (2008) Stochastic switching as a survival strategy in fluctuating environments. *Nat Genet* 40(4):471–475.
- 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.
- Lu M, Onuchic J, Ben-Jacob E (2014) Construction of an effective landscape for multistate genetic switches. *Phys Rev Lett* 113(7):078102.
- 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.
- 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.
- Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S (2006) Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* 4(10):e309.
- Shahrezaei V, Swain PS (2008) Analytical distributions for stochastic gene expression. *Proc Natl Acad Sci USA* 105(45):17256–17261.
- Taniguchi Y, et al. (2010) Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* 329(5991):533–538.
- 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.
- So LH, et al. (2011) General properties of transcriptional time series in *Escherichia coli*. *Nat Genet* 43(6):554–560.
- Peccoud J, Ycart B (1995) Markovian modeling of gene-product synthesis. *Theor Popul Biol* 48(2):222–234.
- Golding I, Paulsson J, Zawilski SM, Cox EC (2005) Real-time kinetics of gene activity in individual bacteria. *Cell* 123(6):1025–1036.
- Boeger H, Griesenbeck J, Kornberg RD (2008) Nucleosome retention and the stochastic nature of promoter chromatin remodeling for transcription. *Cell* 133(4):716–726.
- 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.
- Jones DL, Brewster RC, Phillips R (2014) Promoter architecture dictates cell-to-cell variability in gene expression. *Science* 346(6216):1533–1536.
- Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* 81(25):2340–2361.
- 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.
- 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.
- Cooper S, Helmstetter CE (1968) Chromosome replication and the division cycle of *Escherichia coli* B/r. *J Mol Biol* 31(3):519–540.
- Kubitschek HE, Freedman ML (1971) Chromosome replication and the division cycle of *Escherichia coli* B-r. *J Bacteriol* 107(1):95–99.
- Ho PY, Amir A (2015) Simultaneous regulation of cell size and chromosome replication in bacteria. *Front Microbiol* 6:662.
- Skarstad K, Boye E, Steen HB (1986) Timing of initiation of chromosome replication in individual *Escherichia coli* cells. *EMBO J* 5(7):1711–1717.
- Powell EO (1956) Growth rate and generation time of bacteria, with special reference to continuous culture. *J Gen Microbiol* 15(3):492–511.
- Fei J, et al. (2015) Determination of in vivo target search kinetics of regulatory noncoding RNA. *Science* 347(6228):1371–1374.
- 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.
- 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.
- Klumpp S, Hwa T (2008) Growth-rate-dependent partitioning of RNA polymerases in bacteria. *Proc Natl Acad Sci USA* 105(51):20245–20250.
- Vogel U, Jensen KF (1994) The RNA chain elongation rate in *Escherichia coli* depends on the growth rate. *J Bacteriol* 176(10):2807–2813.
- Earnest TM, et al. (2015) Towards a whole-cell model of ribosome biogenesis: Kinetic modeling of SSU assembly. *Biophys J* 109(6):1117–1135.
- 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.
- 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).