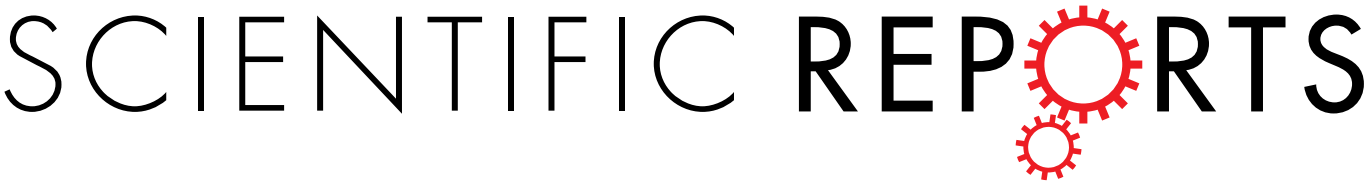

OPEN

Efficient and flexible implementation of Langevin simulation for gene burst production

Received: 24 August 2017
Accepted: 16 November 2017
Published online: 04 December 2017

Ching-Cher Sanders Yan[1], Surendhar Reddy Chepyala[1,2,3], Chao-Ming Yen[1,4,5] & Chao-Ping Hsu[1,6]

Gene expression involves bursts of production of both mRNA and protein, and the fluctuations in their number are increased due to such bursts. The Langevin equation is an efficient and versatile means to simulate such number fluctuation. However, how to include these mRNA and protein bursts in the Langevin equation is not intuitively clear. In this work, we estimated the variance in burst production from a general gene expression model and introduced such variation in the Langevin equation. Our approach offers different Langevin expressions for either or both transcriptional and translational bursts considered and saves computer time by including many production events at once in a short burst time. The errors can be controlled to be rather precise (<2%) for the mean and <10% for the standard deviation of the steady-state distribution. Our scheme allows for high-quality stochastic simulations with the Langevin equation for gene expression, which is useful in analysis of biological networks.

Gene expression is a series of biochemical reactions that produce proteins for various biological functions. For cells with identical genes, gene expression noise is observed in both prokaryotes[1,2] and eukaryotes[3,4]. One general source of such noise is from the probabilistic nature of chemical reactions, because the biological components involved in such reactions are in small copy numbers. In addition, as observed experimentally, both mRNAs[5] and proteins[6] are produced in discontinuous bursts of multiple copies in a short time, and thus, the corresponding fluctuation is increased[7]. Noise propagates through the biochemical networks[8] and may further contribute to the heterogeneity in the phenotypes[9–11]. With the noise, fluctuation-dissipation theorem allows us to derive the dynamic response and infer dynamic properties in a cell[12]. When a precise control is needed, it may be necessary to reduce or buffer such noises[13–15]. Therefore, to gain insights into general biological processes by modeling, a good description for the fluctuation in gene expression is needed.

A complete accounting for the fluctuation in chemical reactions can be obtained by simulations with the Gillespie algorithm[16]. The Gillespie algorithm is a scheme that simulates every reaction event with a proper probability. Without imposing any additional approximations[17], it generates trajectories that follow the exact probability distribution. Since each reaction involves only a small set of changes in molecular numbers, the process is time-consuming for a large system. To accelerate the simulation, a long leaping-time step can be used to account for several reaction events together. With slightly changed reaction propensities, a chemical Langevin equation can be derived[18]. Simulation is more efficient with the Langevin equation than the Gillespie algorithm. Moreover, the Langevin equation allows for a direct dissection and analysis of different noise sources[8,11]. It is therefore highly desirable to develop the Langevin equation for various biochemical processes.

To formulate a Langevin equation for gene expression, the burst properties need to be properly accounted for. Experiments found that for both mRNA and protein, the burst event can be described as a Poisson distribution, with the burst size as an exponential (or geometric) distribution. A general gene expression model[4,19,20] shown in

[1]Institute of Chemistry, Academia Sinica, Taipei, 115, Taiwan. [2]Bioinformatics Program, Taiwan International Graduate Program, Institute of Information Science, Academia Sinica, Taipei, 115, Taiwan. [3]Institute of Biomedical Informatics, National Yang-Ming University, Taipei, 112, Taiwan. [4]Institute of Biochemical Sciences, College of Life Science, National Taiwan University, Taipei, 106, Taiwan. [5]Institute of Biological Chemistry, Academia Sinica, Taipei, 115, Taiwan. [6]Genome and Systems Biology Degree Program, National Taiwan University, Taipei, 106, Taiwan. Ching-Cher Sanders Yan and Surendhar Reddy Chepyala contributed equally to this work. Correspondence and requests for materials should be addressed to C.-P.H. (email: cherri@sinica.edu.tw)

Fig. 1 allows us to define the burst frequency and the burst size in transcription and translation with fundamental rate constants[4,20–22]. Furthermore, the distributions of burst events and sizes derived from this model have the same features as those observed in experiments. The gene expression model shown in Fig. 1a can be written as:

$$\frac{dg}{dt} = k_g(1 - g) - \gamma_g g \tag{1}$$

$$\frac{dm}{dt} = k_m g - \gamma_m m \tag{2}$$

$$\frac{dp}{dt} = k_p m - \gamma_p p, \tag{3}$$

where g is the fraction of active gene for transcription and (m, p) are the amount of mRNA and protein, respectively; k_g and γ_g are the gene's activation and deactivation rates; k_m and k_p are the production rates for mRNA and protein; and γ_m and γ_p are the corresponding degradation rates. Following previous works[21,22], when $\gamma_g \gg (\gamma_m, k_g)$, mRNA production can be considered as occurring in bursts. Because the gene activation time $(1/\gamma_g)$ is rather short, the average amount of mRNAs produced in such short time interval is the mean burst size[22]:

$$\bar{b}_m = \frac{k_m}{\gamma_g}. \tag{4}$$

The low gene activation rate (k_g) leads to well-separated burst events. The k_g is considered the mRNA burst frequency. Similar limiting set $(\gamma_m \gg \gamma_p)$[7,20] applies to protein production, leading to an average burst size of protein as

$$\bar{b}_p = \frac{k_p}{\gamma_m}, \tag{5}$$

and burst frequency as the rate of mRNA production $(g(t)k_m)$. In Fig. 1b, we include a stochastic trajectory under the limit of burst-like production. In this work, we aimed to derive a Langevin equation that includes burst production effects and offers good number fluctuation for gene expression.

In the burst regime, when the upstream component is rarely-produced and fast-degraded, the slowly-degraded downstream component would be produced in bursts. Such difference in rates poses a difficulty for simulations with both the Gillespie algorithm and the standard Langevin equation. For the Gillespie algorithm, the slow reactions are sampled rarely, which leads to poor statistics. The Langevin simulation efficiency is also reduced, because the time step size has to be adjusted for the fast changes of the gene switching or mRNA number fluctuation. Therefore, we need a Langevin equation for the protein fluctuation that does not have to track the fast changes of a gene's state or mRNA's number[23,24].

Starting from the general model, we develop analytical expressions for the mean and variance in the production with the burst effect, and such expression is included in the Langevin equation. Our approach allows for the flexibility to include either or both of the mRNA's and protein's burst effects. We also found that our burst Langevin expression has a large applicable region, which is not limited by the case of burst production. Our algorithm can produce an accurate steady-state mean and similar distribution as that with Gillespie simulation. When a gene switches dynamically, our simulation also can produce accurate dynamics of average protein number. The burst Langevin equation we derived is effective in minimizing the computational time and memory in stochastic simulations. Our simulation scheme with the burst Langevin equation is useful in stochastic simulation for biological networks.

Theory

Langevin equation for burst production. To simplify the derivation of burst Langevin equation, we first consider a two-component model for the burst of either mRNA or protein. In this model, a short-lived x results in a burst event of y:

$$\frac{dx}{dt} = k_x - \gamma_x x \tag{6}$$

$$\frac{dy}{dt} = k_y x - \gamma_y y. \tag{7}$$

For the mRNA's burst production in equations (1) and (2), we assign x as the state of the gene and y as the mRNA. In this case, we can combine the terms $k_g + \gamma_g$ and set it to γ_x. Similarly, for the protein's burst production, x is mRNA and y is protein. We treat g as a constant in equation (2) for a constant mRNA production rate and set $k_m g$ as k_x. Thus, both the mRNA's and protein's production can be described by equations (6) and (7).

To develop an efficient stochastic simulation, we select a time interval τ that is longer than x's lifetime $(1/\gamma_x)$. When there are e_y burst events and each burst size is denoted as b_{yl}, the change in y is:

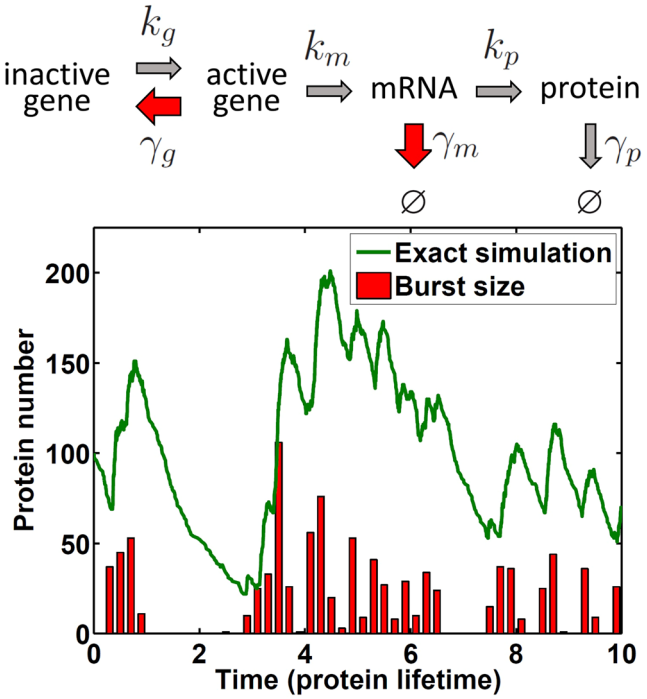

Figure 1. A general model of gene expression with burst productions and its stochastic dynamics of protein number. (a) The scheme of reactions for gene expression. (b) Shown are a stochastic trajectory (green) from the Gillespie algorithm, with the protein's intermittent burst production indicated by red bars in time steps of 0.2 protein lifetime ($1/\gamma_p$). Under the conditions applied, $\gamma_g \gg (\gamma_m, k_g)$ and $\gamma_m \gg \gamma_p$, rapid rises in the trajectory are seen, and protein production can be described as in bursts. Parameters used are $k_g = 5$, $\gamma_g = 95$, $k_m = 200$, $\gamma_m = 10$, $k_p = 100$ and $\gamma_p = 1$, which correspond to $\bar{p} = 100$, average mRNA burst size $\bar{b}_m = k_m/(k_g + \gamma_g) = 2$ and protein average burst size $\bar{b}_p = k_p/\gamma_m = 10$.

$$y(t+\tau) = y(t) + \sum_{l=1}^{e_y} b_{yl} - [\gamma_y y\tau + (\gamma_y y\tau)^{1/2}\mathcal{N}_2(0,1)]. \tag{8}$$

The burst production of y is the consequence of short-lived x. The number of burst events (e_y) is determined by the number of x produced in τ and each burst size (b_{yl}) is determined by the survival time of each x. Simulation for the production in equation (8) can be performed with a random number for e_y, followed by several random numbers for various burst sizes b_{yl}. For the degradation in τ, a Poisson distribution can be used, with both mean and variance being $\gamma_y y\tau$[18]. A Gaussian random number with zero mean and unit variance $\mathcal{N}_2(0, 1)$ is scaled by the standard deviation $(\gamma_y y\tau)^{1/2}$ for the noise part of degradation. An alternative approach is to reformulate the production of y in τ as:

$$y(t+\tau) = y(t) + [\Delta_y(\tau) + \sigma_{\Delta_y}(\tau)\mathcal{N}_1(0,1)] - [\gamma_y y\tau + (\gamma_y y\tau)^{1/2}\mathcal{N}_2(0,1)], \tag{9}$$

where $\Delta_y(\tau)$ and $\sigma_{\Delta y}(\tau)$ are the mean and standard deviation of y's production within time τ. In this way, the simulation steps are simplified, and the computation is more efficient. To estimate $\Delta_y(\tau)$, the average production of y in time τ, we assumed that burst events and burst sizes are independent random processes. Therefore, we can take their average separately:

$$\begin{aligned}\Delta_y(\tau) &= \left\langle \sum_{l=1}^{e_y} b_{yl} \right\rangle = \left\langle \sum_{l=1}^{e_y} \langle b_{yl} \rangle \right\rangle = \left\langle \sum_{l=1}^{e_y} \bar{b}_y \right\rangle \\ &= \langle e_y \rangle \bar{b}_y = \bar{e}_y \bar{b}_y,\end{aligned} \tag{10}$$

which is the product of average burst event ($\bar{e}_y$) and average burst size ($\bar{b}_y$).

The variance of y's production distribution in time τ, $\sigma^2_{\Delta_y}(\tau)$, was derived from the characteristic function of $P(y)$, the probability distribution of y's number, in the supplementary material of ref.[25]:

$$\sigma^2_{\Delta_y}(\tau) = \bar{e}_y \sigma^2_{by} + \sigma^2_{ey} \bar{b}^2_y, \tag{11}$$

where σ^2_{by} is the variance of burst size and σ^2_{ey} is that of burst event number in time τ. We found that it can also be derived directly,

$$\begin{aligned}\sigma^2_{\Delta_y}(\tau) &= \left\langle \Delta_y^2 \right\rangle - \left\langle \Delta_y \right\rangle^2 \\ &= \left\langle \left(\sum_{l=1}^{e_y} b_{yl} \right)^2 \right\rangle - \left\langle \sum_{l=1}^{e_y} b_{yl} \right\rangle^2 \\ &= \left\langle \sum_{l=1}^{e_y} b_{yl}^2 + 2 \sum_{1 \le l < l'}^{e_y} b_{yl} b_{yl'} \right\rangle - \bar{e}_y^2 \bar{b}_y^2 \\ &= \bar{e}_y \left\langle b_{yl}^2 \right\rangle + 2 \frac{\left\langle \bar{e}_y (\bar{e}_y - 1) \right\rangle}{2} \left\langle b_{yl} b_{yl'} \right\rangle - \bar{e}_y^2 \bar{b}_y^2. \end{aligned} \tag{12}$$

With the same assumption that different processes are independent, $\bar{e}_y$ and $\left\langle \bar{e}_y(\bar{e}_y - 1) \right\rangle$ can be separated from $\left\langle b_{yl}^2 \right\rangle$ and $\left\langle b_{yl} b_{yl'} \right\rangle$, respectively. We also replaced $\left\langle b_{yl} b_{yl'} \right\rangle$ with $\bar{b}_y^2$ by assuming different bursts are independent. With the definition of variance, we also replaced $\left\langle b_{yl}^2 \right\rangle$ with $\sigma_{by}^2 + \bar{b}_y^2$ and $\left\langle \bar{e}_y^2 \right\rangle - \bar{e}_y^2$ with σ_{ey}^2. Therefore, we obtain the same variance expression for y's burst production as in ref.[25] by direct estimation.

To simulate the downstream y's fluctuation with burst production, we can follow the Langevin equation as in equation (9) including the mean propagation $\Delta_y(\tau)$ as given in equation (10) and variance $\sigma^2_{\Delta_y}(\tau)$ as in equation (11). The expressions derived in this section can be applied to either or both the mRNA's and protein's burst production.

Langevin equations for either or both mRNA and protein bursts. Generally, different genes may have different dynamic behaviors depending on their degradation rates. Some genes in mammalian cells have only obvious mRNA burst production ($\gamma_g \gg \gamma_m \sim \gamma_p$)[26,27], whereas some genes in yeast have only protein burst production ($\gamma_m \gg \gamma_g \sim \gamma_p$)[4]. Furthermore, some genes in bacteria have both mRNA and protein bursts ($\gamma_g \gg \gamma_m \gg \gamma_p$) and some do not have any obvious burst production ($\gamma_g \sim \gamma_m \sim \gamma_p$)[28]. We further explored the criteria of γ_g and γ_m comparing to γ_p with and without burst production, as shown in Fig. 2a. For all these different burst cases, we show that the burst production variance in equation (11) has the flexibility to describe all of them.

mRNA burst. With the condition ($\gamma_g \gg \gamma_m \sim \gamma_p$), only mRNA is generated in bursts. We rearranged the expression in equation (1) as:

$$\frac{dg}{dt} = k_g - (k_g + \gamma_g) g, \tag{13}$$

which becomes identical to equation (6). For a single-copy gene, the activity fraction g in equation (13) is considered 0 or 1 for off or on state, respectively. Without loss of generality, we considered a single-copy gene in the present work. If the gene has n-copies, $(1 - g)$ in equation (1) can be replaced by $(n - g)$; thus, the first k_g in equation (13) needs to be replaced by nk_g.

mRNA burst frequency is equal to k_g (or nk_g for n-copy gene case). Assuming that each mRNA burst event is independent, we can approximate burst event distribution by a Poisson distribution, as observed in several experiments[5,22], where the mean of burst events ($\bar{e}_m = k_g \tau$) equals the variance (σ^2_{em}):

$$\bar{e}_m = k_g \tau = \sigma^2_{em}. \tag{14}$$

On the other hand, possible mRNA burst size (b_m) can be described by a geometric distribution[29],

$$P(b_m) = q(1 - q)^{b_m}, \tag{15}$$

where q is the probability of no mRNA produced from this activation period, thus, q is proportional to $k_g + \gamma_g$. One mRNA is produced with the probability of $(1 - q)$, which is proportional to k_m, the transcription rate constant in equation (2). The mean and variance of mRNA burst size are

$$\bar{b}_m = \frac{1 - q}{q} = \frac{k_m}{k_g + \gamma_g} \tag{16}$$

$$\sigma^2_{bm} = \bar{b}_m^2 + \bar{b}_m. \tag{17}$$

We note that the mRNA burst size definition is modified as in equation (16), instead of $\bar{b}_m = k_m/\gamma_g$ in the literature which is obtained with very small k_g[21,22]. This new definition for $\bar{b}_m$ yields accurate kinetic expression for the average amount of mRNA, as production (burst frequency (k_g) multiplied by size ($\bar{b}_m$)) divided by degradation rate constant (γ_m):

$$\bar{m} = \frac{k_g \bar{b}_m}{\gamma_m}. \tag{18}$$

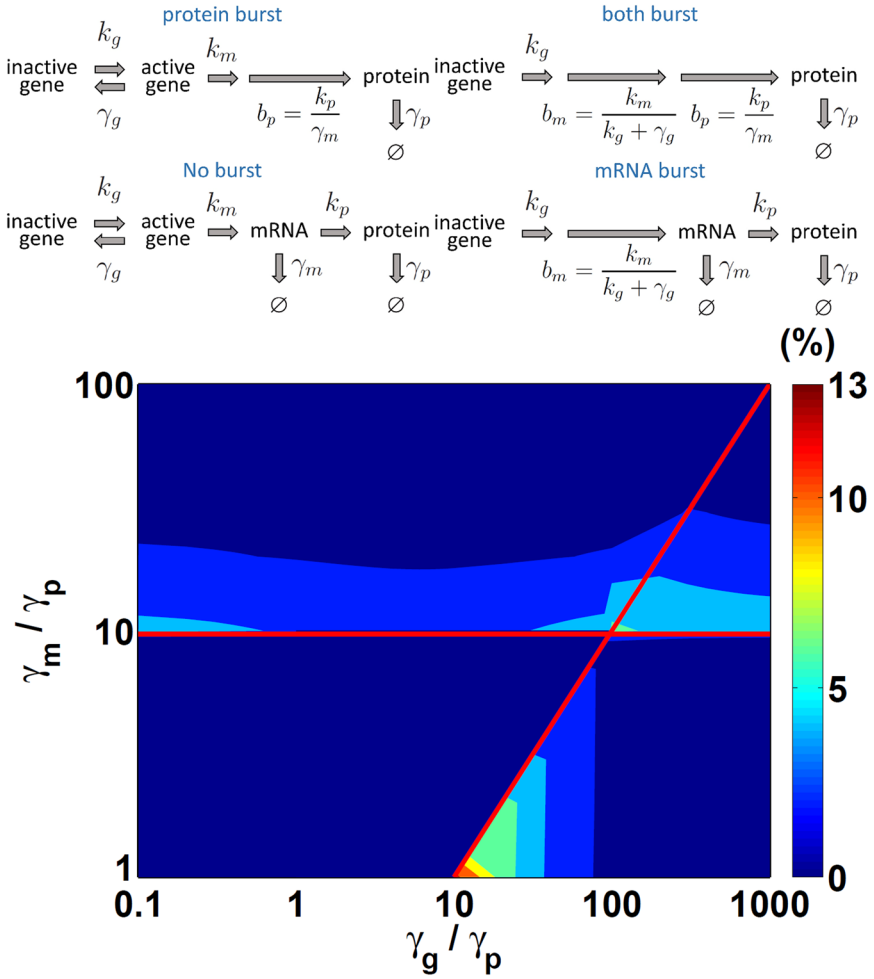

Figure 2. Four cases of gene expression dynamics and errors from different Langevin equations. (a) Four possible cases of gene expression. When an activated gene state is short-lived, the Langevin equation skips the tracking for the gene state, and a burst production following the statistics is used for mRNA. Similarly, when the mRNA's lifetime is short, burst production of protein is introduced, instead of tracking the mRNA. (b) Shown are normalized errors (%) of a steady-state protein's standard deviation ($\sigma_{p,ss}$) from the burst Langevin equations compared to the squared root of exact variance expression as in equation (26), as a function of gene deactivation rate γ_g and mRNA degradation rates γ_m. Red lines are the boundaries for the four different cases in the burst models. Other parameters are $\bar{p} = 100$, $k_g = 5$, mRNA burst size $\bar{b}_m = 2$, protein burst size $\bar{b}_p = 10$, and $\gamma_p = 1$.

These are the statistical features of burst distributions that needs to be included in the Langevin equation of mRNA burst.

The mean of mRNA production with bursts (with large γ_g) can be expressed as:

$$\Delta_m(\tau) = \bar{e}_m \bar{b}_m = k_g \tau \bar{b}_m. \tag{19}$$

by following equation (10). The variance for mRNA is

$$\begin{aligned} \sigma^2_{\Delta_m}(\tau) &= \bar{e}_m \sigma^2_{bm} + \sigma^2_{em} \bar{b}^2_m \\ &= k_g \tau (\bar{b}^2_m + \bar{b}_m) + k_g \tau \bar{b}^2_m \\ &= k_g \tau \bar{b}_m (2\bar{b}_m + 1), \end{aligned} \tag{20}$$

by following equation (11). With equations (19) and (20), the Langevin equation for mRNA is then

$$\begin{aligned} m(t+\tau) &= m(t) + (k_g \tau \bar{b}_m + (k_g \tau \bar{b}_m (2\bar{b}_m + 1))^{1/2} \mathcal{N}_1(0,1)) \\ &\quad - (\gamma_m m \tau + (\gamma_m m \tau)^{1/2} \mathcal{N}_2(0,1)), \end{aligned} \tag{21}$$

and following the amount of mRNA, the Langevin equation for protein is

$$\begin{aligned} p(t+\tau) &= p(t) + (k_p m\tau + (k_p m\tau)^{1/2}\mathcal{N}_3(0,1)) \\ &\quad - (\gamma_p p\tau + (\gamma_p p\tau)^{1/2}\mathcal{N}_4(0,1)), \end{aligned} \tag{22}$$

where the gene state is skipped. From the mRNA's burst Langevin equation in equation (21), we derived the mRNA's steady-state variance as:

$$\sigma^2_{m,ss} \approx \overline{m}(\overline{b}_m + 1). \tag{23}$$

by following the supplementary material of ref.[8]. The steps are transforming $m(t)$ in equation (21) to the Fourier space first and then squaring, averaging, and finally inverse Fourier transforming. The detailed derivation is in the supplementary information of this work. By comparing the production variance in equation (20) and steady-state variance in equation (23), we can see that the τ in equation (20) is replaced by $1/\gamma_m$, leading to $k_g\overline{b}_m/\gamma_m = \overline{m}$ in equation (23). Also, the $2\overline{b}_m$ in equation (20) becomes $\overline{b}_m$ in equation (23). However, the mRNA's exact variance expression in the steady state from linear noise approximation (LNA)[30–32] is

$$\sigma^2_{m,ss} = \overline{m}\left(\frac{\gamma_g}{k_g+\gamma_g+\gamma_m}\frac{k_m}{k_g+\gamma_g}+1\right) = \overline{m}\left(\frac{\gamma_g}{k_g+\gamma_g+\gamma_m}\overline{b}_m+1\right), \tag{24}$$

with detailed derivation given in our supplementary information. By comparing equations (23) and (24), we can see that the burst Langevin approximation can be achieved by assuming $\gamma_g/(k_g+\gamma_g+\gamma_m)\approx 1$ in the LNA's result, which is true that a large γ_g leads to mRNA bursts. Therefore, with equations (21) and (22), there is no need to track the fast-changing gene state $g(t)$ in the simulation, and a modest error is introduced in the mRNA's variance as in equation (23).

To further calculate the protein's steady-state variance, because of $\gamma_m \sim \gamma_p$, we can propagate $\sigma^2_{m,ss}$ from equation (23) by the variance propagation equation[33] to obtain $\sigma^2_{p,ss}$:

$$\begin{aligned} \sigma^2_{p,ss} &= \sigma^2_{m,ss}\frac{k_p}{\gamma_p}\frac{k_p}{\gamma_m+\gamma_p} + \overline{p} \\ &= \overline{p}\left(\overline{b}_m\frac{k_p}{\gamma_m+\gamma_p} + \frac{k_p}{\gamma_m+\gamma_p} + 1\right). \end{aligned} \tag{25}$$

This expression is also slightly different from the exact expression derived from LNA (details in the supporting information), which is given below:

$$\sigma^2_{p,ss} = \overline{p}\left(\frac{\gamma_g(\gamma_g+\gamma_m+\gamma_p+k_g)}{(\gamma_g+\gamma_m+k_g)(\gamma_g+\gamma_p+k_g)}\frac{k_m}{k_g+\gamma_g}\frac{k_p}{\gamma_m+\gamma_p} + \frac{k_p}{\gamma_m+\gamma_p} + 1\right). \tag{26}$$

In general, $\sigma^2_{p,ss}$ obtained by LNA as in equation (26) includes the overall intrinsic noise of a gene following equations (1) to (3). So it is desirable to compare the $\sigma^2_{p,ss}$ in equation (25) from the burst Langevin equation to the exact variance in equation (26). The difference between equations (25) and (26) gives us an indication of the burst Langevin equation's accuracy. Such difference is shown in the lower right region in Fig. 2b, where $\gamma_g \geq 10\,\gamma m$. The largest error of the bursting Langevin equation with mRNA burst alone is 12.5% that occurs at the lower left boundary of the region, which is still acceptable.

Both mRNA and protein bursts. In the condition $\gamma_g \gg \gamma_m \gg \gamma_p$, both the mRNA's and protein's production are produced in bursts. We can combine both bursts and derive one Langevin equation for the protein's fluctuation, thereby greatly simplifying the simulation. Because each mRNA corresponds to a protein burst event, the number of protein burst events in τ is

$$\overline{e}_p = k_g\overline{b}_m\tau, \tag{27}$$

leading to the protein production as

$$\Delta_p = k_g\overline{b}_m\overline{b}_p\tau. \tag{28}$$

Since mRNA is also produced in bursts, the variance of the protein's burst event is

$$\sigma^2_{ep} = \sigma^2_{\Delta_m} = k_g\overline{b}_m\tau(2\overline{b}_m+1), \tag{29}$$

which is identical to that in equation (20). The variance of the protein's production following equation (11) is

$$\begin{aligned}\sigma^2_{\Delta_p} &= \bar{e}_p\sigma^2_{bp} + \sigma^2_{ep}\bar{b}^2_p \\ &= k_g\bar{b}_m\tau(\bar{b}^2_p + \bar{b}_p) + k_g\bar{b}_m\tau(2\bar{b}_m + 1)\bar{b}^2_p \\ &= k_g\bar{b}_m\tau\bar{b}_p(2\bar{b}_m\bar{b}_p + 2\bar{b}_p + 1).\end{aligned} \tag{30}$$

We note that σ^2_{bp} is equal to $\bar{b}^2_p + \bar{b}_p$, by following the same assumption of geometric distribution as $\bar{b}_m$ in equation (15). We also note that similar results with both bursts were derived using the generation function of $P(p)$, the probability distribution of p's number in the supplementary material of ref.[34].

With equations (28) and (30), the Langevin equation for protein fluctuation with both bursts is

$$\begin{aligned}p(t+\tau) &= p(t) + (k_g\bar{b}_m\bar{b}_p\tau + (k_g\bar{b}_m\bar{b}_p\tau(2\bar{b}_m\bar{b}_p + 2\bar{b}_p + 1))^{1/2}\mathcal{N}_1(0,1)) \\ &\quad -(\gamma_p p\tau + (\gamma_p p\tau)^{1/2}\mathcal{N}_2(0,1)).\end{aligned} \tag{31}$$

It allows us to efficiently simulate protein's fluctuation, because we can skip tracking the gene state and the mRNA in the simulation.

Following the same process as we obtained the mRNA's steady-state variance $\sigma^2_{m,ss}$ as in equation (23), here we obtained the protein's steady-state variance from equation (31) as

$$\sigma^2_{p,ss} \approx \bar{p}(\bar{b}_m\bar{b}_p + \bar{b}_p + 1). \tag{32}$$

In the upper right region of Fig. 2b, we show the normalized error in $\sigma_{p,ss}$ for equation (32) to that from LNA as equation (26) with the condition of both bursts as following:

$$\gamma_g \geq 10\,\gamma_m \geq 100\,\gamma_p. \tag{33}$$

The largest possible error in the standard deviation is <6.5%, which is quite acceptable.

Protein burst. When the gene's active state is long-lived ($\gamma_g \sim \gamma_p$), the mRNA is not produced in bursts. For some genes ($\gamma_m \gg \gamma_g \sim \gamma_p$) reported in yeast[4], short-lived mRNA leads to the protein's burst production. Partial simplification of the Langevin equations for the gene expression is still possible if the protein is produced in bursts. For such case, we keep track the gene's activity, skip the short-lived mRNA, and develop the protein's Langevin equation with bursts:

$$g(t+\tau) = g(t) + k_g(1-g)\tau - \gamma_g g\tau \tag{34}$$

$$p(t+\tau) = p(t) + \left(gk_m\tau\bar{b}_p + \sigma_{\Delta p}\mathcal{N}_1(0,1)\right) - \left(\gamma_p p\tau + (\gamma_p p\tau)^{1/2}\mathcal{N}_2(0,1)\right) \tag{35}$$

The gene-switching probability is $k_g\tau$ with $g=0$ and $\gamma_g\tau$ with $g=1$. The mean production of the protein number in time τ is $gk_m\tau\bar{b}_p$, which is $gk_m\tau$, the number of protein burst events (same as the number of mRNA molecules) produced in τ, multiplied by $\bar{b}_p$, the mean protein burst size. For the noise strength ($\sigma_{\Delta p}$) in equation (35), $g(t)k_m$ can be considered a constant in τ because the state of the gene does not switch frequently in τ. Therefore, the mRNA produced or the protein burst event from equation (35) is a Poisson distribution, with the variance being the same as the mean:

$$\sigma^2_{ep} = \sigma^2_{\Delta m} = g(t)k_m\tau. \tag{36}$$

Following equation (11), the variance of protein production in equation (35)can be written as

$$\begin{aligned}\sigma^2_{\Delta_p} &= \bar{e}_p\sigma^2_{bp} + \sigma^2_{ep}\bar{b}^2_p \\ &= g(t)k_m\tau(\bar{b}^2_p + \bar{b}_p) + g(t)k_m\tau\bar{b}^2_p \\ &= g(t)k_m\tau\bar{b}_p(2\bar{b}_p + 1).\end{aligned} \tag{37}$$

In a simulation trial, the noise strength of protein's production (σ_{Δ_p}) follows the state of $g(t)$. Also, because $g(t) = 1$ in some τ steps and $g(t) = 0$ in others, the average gene state is $\bar{g} = k_g/(k_g + \gamma_g)$.

With only protein bursts, based on the condition of $\gamma_m \geq 10\gamma_p$, we can simplify $\sigma^2_{p,ss}$ in equation (26) from LNA to

$$\sigma^2_{p,ss} \approx \bar{p}\left(\frac{\gamma_g}{k_g + \gamma_g + \gamma_p}\bar{b}_m\bar{b}_p + \bar{b}_p + 1\right), \tag{38}$$

by reducing the first fraction and using $\bar{b}_p = k_p/\gamma_m$. In the upper left region of Fig. 2b, comparison of $\sigma_{p,ss}$ from equation (38) to that from the exact variance in equation (26) shows that the largest possible error is <5%.

In a special condition, where k_g and γ_g are significantly smaller than other four kinetic parameters in the gene expression model (equations (1) to (3)), bimodal distribution of the protein number can be obtained from the

	mRNA burst ($y=m$)	both bursts ($y=p$)	protein burst ($y=p$)
condition	$\gamma_g \geq 10\,\gamma_m$	$\gamma_g \geq 10\,\gamma_m \geq 100\,\gamma_p$	$\gamma_m \geq 10\,\gamma_p$
simulated subjects	$m(t), p(t)$	$p(t)$	$g(t), p(t)$
burst event distribution:			
$\bar{e}_y(\tau)$	$k_g\tau$	$k_g\tau\bar{b}_m$ ‡	$g(t)k_m\tau$
$\sigma^2_{ey}(\tau)$	$k_g\tau$	$k_g\tau\bar{b}_m(2\bar{b}_m+1)$ ‡	$g(t)k_m\tau$
burst size† distribution:			
$\bar{b}_y$	$\bar{b}_m$	$\bar{b}_p$	$\bar{b}_p$
σ^2_{by}	$\bar{b}_m^2+\bar{b}_m$	$\bar{b}_p^2+\bar{b}_p$	$\bar{b}_p^2+\bar{b}_p$
burst production in Langevin equation:			
$\Delta_y(\tau)$ by equation (10)	$k_g\tau\bar{b}_m$ ‡	$k_g\tau\bar{b}_m\bar{b}_p$	$g(t)k_m\tau\bar{b}_p$
$\sigma^2_{\Delta y}(\tau)$ by equation (11)	$k_g\tau\bar{b}_m(2\bar{b}_m+1)$ ‡	$k_g\tau\bar{b}_m\bar{b}_p(2\bar{b}_m\bar{b}_p+2\bar{b}_p+1)$	$g(t)k_m\tau\bar{b}_p(2\bar{b}_p+1)$
steady-state distribution: exact expression*			
$\bar{m}=\bar{g}\frac{k_m}{\gamma_m}$	same as exact	—	—
$\sigma^2_{m,ss}=\bar{m}(F_1\bar{b}_m+1)$	$\bar{m}(\bar{b}_m+1)$ §	—	—
$\bar{p}=\bar{m}\frac{k_p}{\gamma_p}$	same as exact	same as exact	same as exact
$\sigma^2_{p,ss}=\bar{p}\,(F_0\bar{b}_m\bar{b}_{p0}+\bar{b}_{p0}+1)$	$\bar{p}\,(\bar{b}_m\bar{b}_{p0}+\bar{b}_{p0}+1)$ ¶	$\bar{p}\,(\bar{b}_m\bar{b}_p+\bar{b}_p+1)$ ‖	$\bar{p}\,(F_2\bar{b}_m\bar{b}_p+\bar{b}_p+1)$ #

Table 1. Summary of statistics for three different cases of burst in gene expression. †Definition for burst sizes are: $\bar{b}_m=\frac{k_m}{\gamma_g+k_g}$, $\bar{b}_{p0}=\frac{k_p}{\gamma_m+\gamma_p}$, $\bar{b}_p=\frac{k_p}{\gamma_m}$. ‡The mean and variance of mRNA production with mRNA burst are the mean and variance of the burst events of protein in the case of both bursts. *Definitions for the fractions are: $F_0=\frac{\gamma_g(\gamma_g+\gamma_m+\gamma_p+k_g)}{(\gamma_g+\gamma_m+k_g)(\gamma_g+\gamma_p+k_g)}$, $F_1=\frac{\gamma_g}{(\gamma_g+\gamma_m+k_g)}$, $F_2=\frac{\gamma_g}{(\gamma_g+\gamma_p+k_g)}$. §With the conditions of $\gamma_g \gg \gamma_m$ and $\gamma_g \gg k_g$, F_1 replaced by 1 with mRNA burst. ¶With the conditions of $\gamma_g \gg \gamma_m$ and $\gamma_g \gg k_g$, F_0 replaced by 1 with mRNA burst. ‖With the conditions of $\gamma_g \gg \gamma_m$ and $\gamma_g \gg k_g$, F_0 replaced by 1 with both bursts; with $\gamma_m \gg \gamma_p$, $\bar{b}_{p0}$ replaced by $\bar{b}_p$. #With the conditions of $\gamma_m \gg \gamma_p$, F_0 replaced by F_2 with protein burst; and with $\gamma_m \gg \gamma_p$, $\bar{b}_{p0}$ replaced by $\bar{b}_p$.

numerical simulation. Such parameter sets lie on the upper and most-left region of Fig. 2b, where the error of $\sigma_{p,ss}$ from the burst Langevin is small as <5%. Considering $k_g=\gamma_g=0.1\gamma_p$ with $\gamma_m=10\gamma_p$, which leads to only protein bursts, the burst Langevin algorithm can fairly reproduce the bimodal distributions of the protein number with various combinations of k_m and $\bar{b}_p$. The comparison of distributions between the protein burst Langevin simulation and Gillespie algorithm in this special condition are shown in the supplementary information.

Overall, our comparison shows that the burst Langevin equation can provide reliable estimations of $\sigma_{p,ss}$ for all three cases, where bursts are observed in mRNA, protein or both mRNA and protein. For the three cases, we organized the statistical expressions including burst events, burst sizes, variance of production and steady-state variance in Table 1. For three cases of bursts, the variance of the burst event as in equation (11) needs to be modified accordingly.

Neither mRNA nor protein in bursts. For the case that γ_g and γ_m are close to γ_p, simulations with the Gillespie algorithm or the τ-leaping algorithm[35] would work well. The problems of inefficient simulation and poor statistics of rare events due to greatly different reaction rates do not exist in this case. Because mRNA and protein production are not in bursts, all three species in equations (1) to (3) need to be tracked in the simulation to fully account for intrinsic noise of gene expression. Simulation with the Gillespie algorithm has no imposed approximation, and thus the steady-state variance it produces is close to LNA in equation (26). Therefore, the lower left region of Fig. 2b indicates zero error.

Results

Single gene expression. The gene expression model as described in equations (1) to (3) is tested to see how the one-component burst Langevin equation in equation (31) can be used to replace a three-component model. We use the Gillespie algorithm to simulate the model as in equations (1) to (3) to obtain the exact numerical simulation results. We compared the normalized error of a protein's mean ($\bar{p}$) in the steady state from the burst Langevin simulation to that from the Gillespie simulation.

There are six parameters in the model as in equations (1) to (3). We first chose the unit for time as $1/\gamma_p$. In other words, the protein degradation rate was set to 1. We further set $\gamma_m=10$ and $\gamma_g=100$, for a fast degradation rate in mRNA and an even faster DNA deactivation rate, respectively. This is at the margin of treating both mRNA and protein production with bursts (equation (32)), where the largest error (<6.5%) could be produced, as shown in the upper right region of Fig. 2b. To test the applicable range of the burst Langevin simulation, we scanned the gene activation constant ($k_g=1-100$), which covers the mRNA burst frequency value of 5 to 45 as observed in the experiment[22]. The other parameter we scanned is the protein burst size ($\bar{b}_p=1-100$), or equivalently protein production rate ($k_p=10-1000$), which also covers the values observed in the experiment[28]. We first fixed the

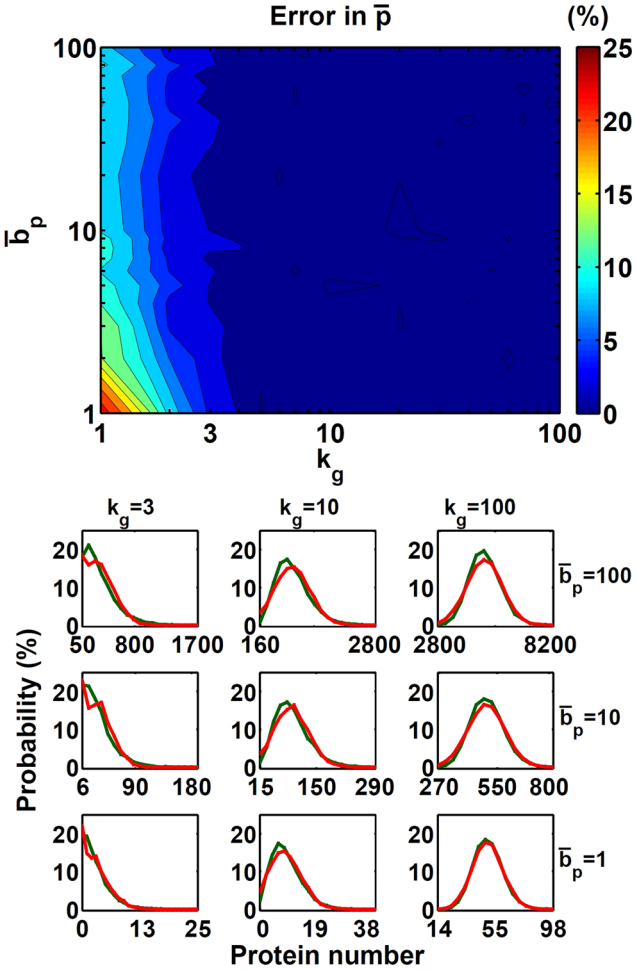

Figure 3. Comparison of $\overline{p}$ and steady-state distributions with the burst Langevin simulation and Gillespie simulation. Shown in (a) are the $\overline{p}$ difference (in %) with the burst Langevin simulation and Gillespie simulation and in (b) steady-state distributions with the burst Langevin simulation (red) and Gillespie simulation (green) with different gene activation rates $k_g = 3,10,100$ and burst size $\overline{b}_p = 1,10,100$. Statistics were taken at the steady state of 10,000 independent points for the model as defined in equations (1) to (3) with parameters $k_m = 100$, $\gamma_g = 100$, $\gamma_m = 10$ and $\gamma_p = 1$.

parameter k_m as 100, which approximately yields $\overline{b}_m = 1$ by equation (16), similar to the value generally observed in bacteria[28]. The average amount of protein in the steady state from the parameter set we scanned can be calculated as

$$\overline{p} = \frac{k_g}{k_g + \gamma_g}\frac{k_m}{\gamma_m}\frac{k_p}{\gamma_p} = \frac{k_g}{k_g + 100}\frac{100}{10}\frac{k_p}{1}, \tag{39}$$

which covers $0.9 < \overline{p} \leq 5000$, the range of observed protein copy number in an *E. coli* cell[28].

We compared the average protein number ($\overline{p}$) of the steady-state distribution from the burst Langevin simulation to that from the Gillespie simulation and shown in Fig. 3a. When $k_g \geq 3$ and $\overline{b}_p \geq 1$, corresponding to $\overline{p} \geq 3$, our algorithm's error is <5%. The largest error of $\overline{p}$ is found at the lower left corner, which is caused by the Gaussian function in the Langevin simulation deviating from the Poisson distribution. Such deviation affects all kinds of Langevin simulations, including our burst Langevin scheme.

Figure 3b compares the protein's steady-state distribution with the burst Langevin simulation and Gillespie simulation. The burst Langevin simulation can reproduce the distributions with different combinations of burst frequency (k_g) and burst size ($\overline{b}_p$).The normalized error in standard deviation for these cases ranges from −13% to 14% (details included in the supplementary information). Although all the steady-state distributions have some error in $\sigma_{p,ss}$, they are sufficiently good for further applications. We further analyzed the sources of such error and discussed them in the supplementary information for interested readers.

Figure 4 compares the computational time percentage with the burst Langevin simulation to that with the Gillespie simulation. The burst Langevin simulation always uses less time than the corresponding Gillespie simulation. When particle number is ≤100, the Gillespie simulation is already efficient; thus, the time usage with

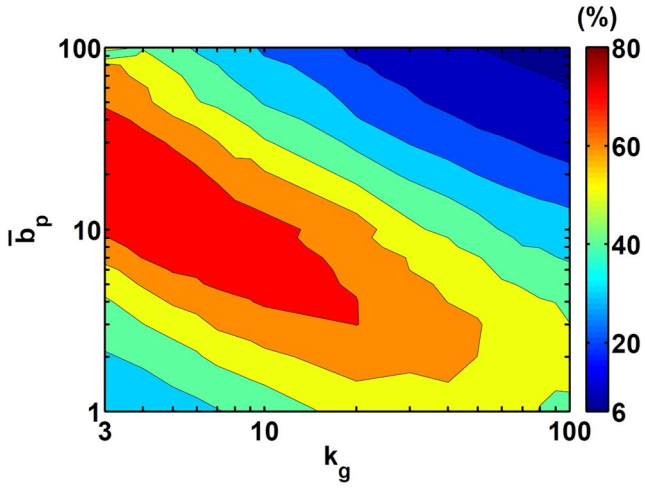

Figure 4. Comparison of simulation time with the burst Langevin simulation and Gillespie simulation. Shown are the percentage of computer time used by the burst Langevin simulation compared to that by the Gillespie simulation for different k_g and burst size $\bar{b}_p$.

the burst Langevin simulation is 40% to 80% of that with the Gillespie simulation. However, when the particle number is large, the burst Langevin uses only <10% simulation time as compared with the Gillespie simulation. Therefore, the burst Langevin simulation is efficient.

We also checked the accuracy of the burst Langevin simulation comparing to the Gillespie simulation by varying mRNA burst size. In this test, with a fixed $k_g = 5$, we scanned the other parameter pair: mRNA mean burst size, $\bar{b}_m = 1 - 30$ and protein mean burst size, $\bar{b}_p = 1 - 100$. The parameter $\bar{b}_m = 1 - 30$ corresponds to the mRNA burst size observed in mammalian cells[22]. The parameter region tested corresponds to $\bar{p} = 4 - 15{,}000$. As shown in Fig. 5a, the errors in $\bar{p}$ are within ±2%. The steady-state distributions between two methods shows a good agreement (Fig. 5b). Comparison of the standard deviation in the steady state (from −8% to 2.5%) is included in the supplementary information. These results indicate that our burst Langevin algorithm is applicable for a wide range of biological systems.

Burst Langevin for non-linear regulation. We further tested a gene's expression under regulation to show how steady-state distribution errors of the upstream can affect the downstream mean number, especially with a non-linear regulation. Here we chose repressing regulation as an example. We use the gene expression model shown in equations (1) to (3) as an upstream protein, p_1 with varying $\bar{b}_{m1}$ and $\bar{b}_{p1}$ (as in Fig. 5a). The downstream gene's transcription is repressed by p_1 through the Hill function with threshold (K) as in the following equations:

$$\frac{dp_1}{dt} = k_{g1}\bar{b}_{m1}\bar{b}_{p1} - \gamma_{p1}p_1, \tag{40}$$

$$\frac{dp_2}{dt} = k_{g2}\left(\bar{b}_{m2}\frac{K^{n_H}}{K^{n_H} + p_1^{n_H}} + k_l\right)\bar{b}_{p2} - \gamma_{p2}p_2. \tag{41}$$

We compared the difference in $\bar{p}_2$ between the burst Langevin simulation and Gillespie simulation. The simulation result for an activation regulation with negative n_H (equivalent to a positive regulation) can be found in the supplementary information. We introduced k_l for p_2's possible leaking of mRNA, so that the expression of an repressed gene may remain at a low level but not zero[36]. In this way, a basal production for p_2 is introduced, and thus, the problem of the Langevin simulation with very low p_2 can be mostly avoided.

In the lower-left corner of Fig. 6a, the downstream gene expression level is $\bar{p}_2 = 125$, which means that p_2 is fully activated and there are only a few p_1. With increasing p_1, p_2 is reduced to $\bar{p}_2 = 25$ as seen in the upper-right corner of Fig. 6a. The errors in $\bar{p}_2$ in Fig. 6b are from −4% to 8%. The red line in Fig. 6b indicates $\bar{p}_1 = K$, the threshold value of the repression. Within the region close to the threshold, the production of p_2 is sensitive to fluctuations in p_1. However, even in this region nearby, the error at most is only −4%. Therefore, even with a non-linear regulation in this system, the burst Langevin simulation can produce accurate results.

In Fig. 6b, the largest error in $\bar{p}_2$ is about 8%, found with $\bar{b}_{m1} = 30$ and $\bar{b}_{p1} \geq 5$. In this region, p_1's copy number is high, and its number fluctuation is also high with such large burst-size pairs. p_2 is fully repressed by high p_1 number and kept at its basal expression level, $\bar{p}_2 = 25$. Here the 8% error comes from a two- to three-particle difference in $\bar{p}_2$, and such error is quite acceptable in stochastic simulations.

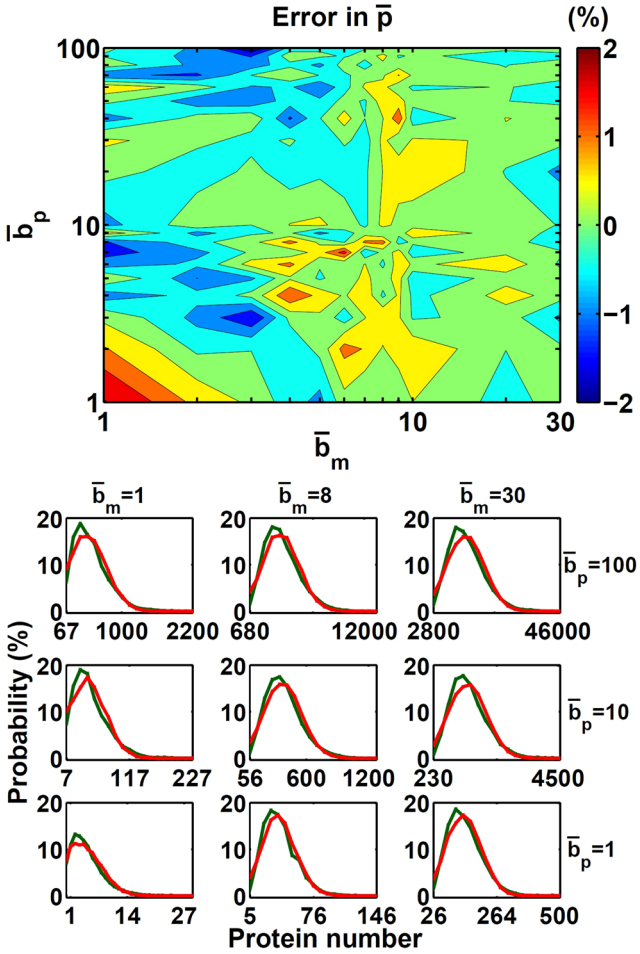

Figure 5. Comparison between the burst Langevin simulation and Gillespie simulation with different $\bar{b}_m$ and $\bar{b}_p$. Shown in (a) are the $\bar{p}$ difference (in %) with the burst Langevin simulation and Gillespie simulation and in (b) steady-state distributions with the burst Langevin simulation (red) and Gillespie simulation (green) with different $\bar{b}_m = 1,8,30$ and $\bar{b}_p = 1,10,100$. k_p and k_m were determined by equations (5) and (16) with given $\bar{b}_p$ and $\bar{b}_m$, respectively, with the other parameters $k_g = 5$, $\gamma_g = 100$, $\gamma_m = 10$ and $\gamma_p = 1$.

In the region that we scanned, p_1's $\sigma_{p1,ss}$ error is from −8.5% to 2.5%, which mainly follows the value of $\bar{b}_{m1}$ (as shown in supplementary information). Such error may propagate through the regulation and cause error in $\bar{p}_2$. However, as seen in Fig. 6b, the error in $\bar{p}_2$ has only a mild correlation with increasing $\bar{b}_{m1}$ alone. The overall trend of increasing error roughly follows inversely with increasing p_2 from the down-left corner to up-right corner in Fig. 6b, and thus, the errors in the standard deviation of the upstream do not affect the quality of the downstream.

Dynamics of average protein number. Besides steady-state behaviors, we demonstrate the accuracy of the burst Langevin simulation in dynamics. In Fig. 7, we include the mean protein number dynamics with the burst Langevin simulation and Gillespie simulation. In this model, the gene is activated at time $t = 7$ by setting $k_g = 30$ and deactivated at $t = 14$ by setting $k_g = 3$. From Fig. 7, we can see that our simulation algorithm can produce reasonably accurate dynamics in the mean and standard deviation as compared with the exact Gillespie simulation. Only a small deviation can be found in the standard deviation. In this test, we selected $\bar{b}_m = 2$ and $\bar{b}_p = 10$. However, similar results with reasonable dynamics are obtained by varying combinations of $\bar{b}_m$ and $\bar{b}_p$ (supplementary information).

Discussion

In this work, we have developed a Langevin equation that can account for the noise arising from gene expression bursts. We found a large range of parameters with which our burst Langevin simulation can well reproduce the statistics comparing to the Gillespie algorithm, and it covers the protein expression level for more than 4 orders of magnitude. For the case of mRNA (protein) burst production, the deactivation (degradation) rates of the gene (mRNA) should be 10 times faster than that of mRNA (protein). The burst Langevin equation has the flexibility to include only mRNA or protein burst or both bursts. In addition, the gene activation rate constant (k_g) has multiple effects in the mean and variance of the production distribution; thus, it is a critical parameter in the accuracy

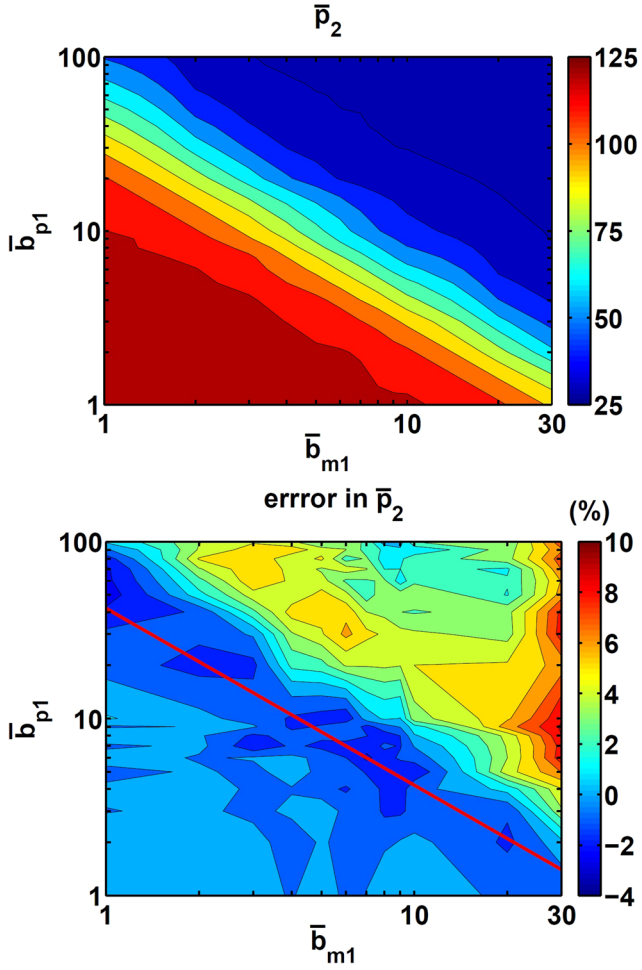

Figure 6. A test for simulation error of gene expression under non-linear repressive regulation. Shown in (a) is the steady-state $\bar{p}_2$ with the burst Langevin simulation and in (b) the error in $\bar{p}_2$ from the burst Langevin simulation comparing to that from the Gillespie simulation. Here the p_1's burst frequency, $\bar{b}_{m1}$, and burst size, $\bar{b}_{p1}$, are varied over a range. Other parameters for p_1 are $k_{g1}=5$, $\gamma_{g1}=100$, $\gamma_{m1}=10$ and $\gamma_{p1}=1$. For p_2, the parameters are $k_{g2}=5$, $\gamma_{g2}=100$, $k_{m2}=200$, $k_l=60$, $\gamma_{m2}=10$, $k_{p2}=100$ and $\gamma_{p2}=1$, $K=200$ and $n_H=3$. The red line in (b) corresponds to $\bar{p}_1 = K$.

of the burst Langevin simulation. When $k_g \geq 3$, which leads to $\bar{p} \geq 3$, the burst Langevin simulation can produce an accurate steady-state mean and standard deviation as compared with the Gillespie simulation. Furthermore, the burst Langevin simulation can produce accurate dynamics of genetic switching and genes under non-linear regulation. Therefore, the burst Langevin equation is applicable for a wide range of genetic regulation network.

To fully consider all intrinsic noises of a gene, with the Gillespie simulation, all of the three components including the gene state, mRNA and protein are simulated with a total of six reaction channels. However, with the burst Langevin simulation, with both mRNA and protein bursts, the model can be reduced to only protein with two reaction channels. Thus, the burst Langevin simulation uses less computational time and memory than the Gillespie simulation. Therefore, our algorithm is an efficient stochastic simulation method.

Besides efficiency, the Langevin equation allows for easily dissecting the contribution of noise from different sources, because the Gaussian random number for each reaction channel can be easily set as zero. Therefore, the burst Langevin simulation can be used to analyze the dynamics of gene expression noises propagating through the regulation network[8,11]. Moreover, one can introduce a desirable scaling parameter to the noise strength in the Langevin equation for mimicking other possible sources. Therefore, one can reproduce the noises close to that from the Gillespie simulation or that observed in various biological systems.

With the variance expression in equation (11), our burst Langevin equation can be flexible to include various cellular factors. With the assumption that burst events and burst sizes are determined by independent processes, we use two different distributions to estimate the variance of burst production in the burst Langevin equation. For the cases of mRNA or protein burst alone, we use the Poisson distribution for burst events and geometric distribution for burst sizes, whereas for the case of both bursts, the protein burst event's variance is enhanced by the mRNA burst, and the overall variance is obtained by the same expression in equation (11). In general, gene expression in the model (equations (1) to (3)) may be influenced by other factors in the cell[37], such as chromatin

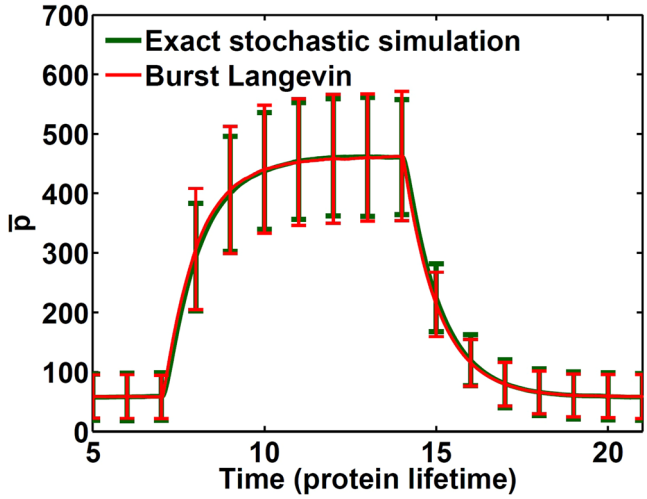

Figure 7. Comparison of different algorithms for genetic switching dynamics. Shown are average protein numbers with the standard deviation of the distribution at different times from 10,000 independent stochastic trajectories with the burst Langevin algorithm (red) and Gillespie simulation (green) for the model defined in equations (1) to (3) with parameters $k_g = 30$ for $t = 7$ to 14; otherwise $k_g = 3$ and other parameters $\gamma_g = 100$, $k_m = 200$, $\gamma_m = 10$, $k_p = 100$ and $\gamma_p = 1$.

template and promoter structure[38,39], which leads to different mRNA and protein production distributions other than Poisson or geometric distributions[40]. Also, post-translational modification introduces an additional step after protein production, which can modify the overall protein production rate, k_p. Thus, protein burst size $\bar{b}_p$ and variance σ^2_{bp} may also be modified from the geometric distribution. For a more detailed model[40], if burst event and size are determined independently, their distributions can be introduced in equation (11) to estimate consequent burst production variance. Therefore, the burst Langevin equation in equation (31) can be modified accordingly to include other factors or more detailed steps in the gene expression model.

Different steps in gene expression are implemented by different molecular machineries. A gene is activated by chromatin remodeling[4,21], whereas mRNA is produced by RNA polymerase and protein is produced by ribosome. There is no machinery competition between different steps. Therefore, we assumed that different steps in gene expression are independent processes and derived the variance of burst production. However, under different physiological conditions in bacteria[41,42], negative correlations are reported between transcription and translation. And thus, regulation or competition for resources may exist between transcription and translation. Yet, once a gene is expressing, protein requires more biomass than transcripts. Protein synthesis also consumes most of the energy, but other processes in gene expression consume a non-relevant amount (<10%) of energy[43,44]. Therefore, energy competition may only be possible in some extreme conditions and assuming different steps in gene expression as independence processes is valid.

Methods

Burst Langevin Simulation Settings. *τ selection for burst events.* In propagating the Langevin equation, the Euler-Maruyama scheme[45] is technically rather simple to implement. While using this scheme, we need to select a proper step size τ, such that all reactant's expected changes are within a small proportion, ε. For a general biological model, besides the protein's burst production and degradation in equation (31), the protein may involve other jth reaction with reaction propensity a_j and number change ν_{pj}:

$$\begin{aligned} p(t+\tau) &= p(t) + (k_g \bar{b}_m \bar{b}_p \tau + (k_g \bar{b}_m \bar{b}_p \tau (2\bar{b}_m \bar{b}_p + 2\bar{b}_p + 1))^{1/2} \mathcal{N}_1(0,1)) \\ &\quad - (\gamma_p p\tau + (\gamma_p p\tau)^{1/2} \mathcal{N}_2(0,1)) \\ &\quad + \sum_j (\nu_{pj} a_j \tau + (\nu_{pj} a_j \tau)^{1/2} \mathcal{N}_j(0,1)). \end{aligned} \tag{42}$$

Following the τ-leaping scheme[35], the step size τ is determined by:

$$\tau = \min_i \left\{ \frac{\max\{\varepsilon p, 1\}}{|k_g \bar{b}_m \bar{b}_p + (-1)\gamma_p p + \sum_j \nu_{pj} a_j|}, \frac{\max\{\varepsilon p, 1\}^2}{k_g \bar{b}_m \bar{b}_p + (-1)^2 \gamma_p p + \sum_j \nu_{pj}^2 a_j}, \tau_{i \neq p} \right\}, \tag{43}$$

where the first denominator multiplied by τ is p's expected change in τ and the second denominator multiplied by τ is the variance of expected change. The τ is selected as the minimum τ_i among all reacting species i, including p. The setting in the numerator is for the efficiency of simulation. When the amount of protein is large, the term εp in the numerator includes as many reactions as possible in τ and accelerates the simulation being faster than Gillespie algorithm. When there are only a few proteins, with the second term, 1 in the numerator, the τ is large

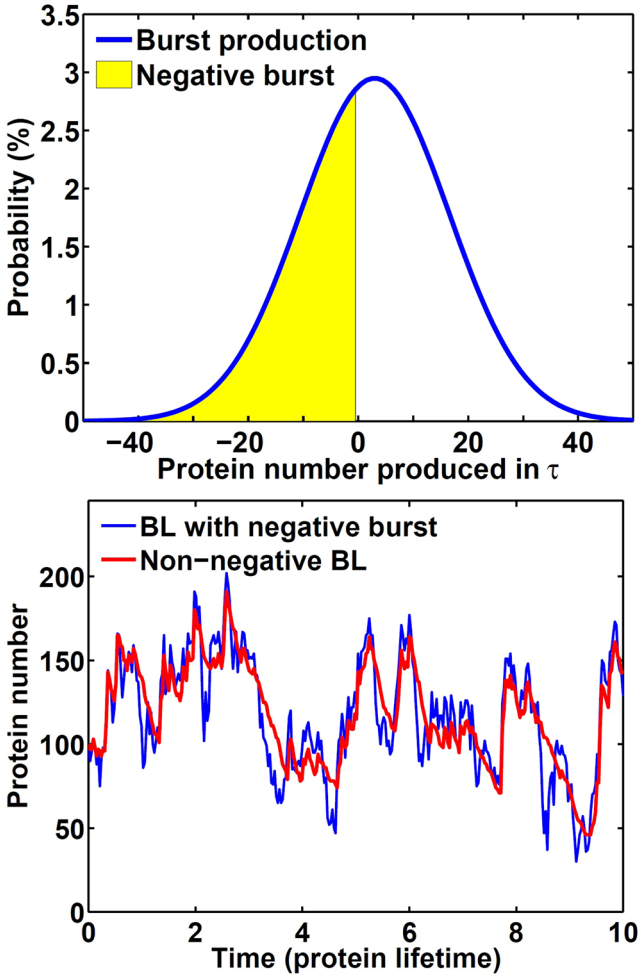

Figure 8. A representative burst production distribution and the effect of negative burst on a protein's fluctuation. Shown in (a) is a typical burst production distribution as defined in equation (31) with $\tau = 0.03$ and colored area as the negative production and in (b) are two stochastic trajectories from the algorithm following equation (31), with (blue) and removing (red) negative burst production. Parameters are $k_g = 5$, $\gamma_g = 95$, $k_m = 200$, $\gamma_m = 10$, $k_p = 100$ and $\gamma_p = 1$, corresponding to $\overline{p} = 100$ with $\overline{b}_m \overline{b}_p = 20$.

enough for some reactions such that reactant numbers are changed at least by one particle. In the situation of few particles, the τ-leaping scheme is close to the Gillespie algorithm, which tracks every reaction.

There are some reactions whose propensity changes drastically even with one reaction event. These are classified as critical reactions in the system[35]. Examples are the switching steps between the two gene states in equation (34) or the protein degradation reaction with protein number <10, where the number 10 is suggested in the literature[35]. Besides the τ selected from equation (43), when there are critical reactions in the current state, another τ_c is randomly selected from an exponential distribution function with $\overline{\tau}_c$, which is the average time for one critical reaction event. The $\overline{\tau}_c$ is defined as the reciprocal of the sum of all critical reaction propensities. If τ_c is smaller than the τ from equation (43), the system is propagated with τ_c with one critical reaction event. Otherwise, the system is propagated with τ without any critical reactions.

However, in the Langevin simulation with bursts, a large-size burst causes an additional problem in the τ-leaping scheme[35] when the amount of protein is low. In equation (43), large $k_g \overline{b}_m \overline{b}_p$ value in the denominator leads to a very small τ, and only one protein is produced. And such small τ will be selected consecutively for a complete burst event. No reactions occur in such small τ other than one protein produced, and thus, the simulation time is wasted. To overcome this situation, we reformulated equation (43) such that one burst event is allowed in one τ step. With this modification in mind, we consider two different τ's, one estimated from the burst production and the other ($\tau_{pj \in nb}$) from other non-burst reactions (including protein degradation) following the original scheme as in equation (43). The smaller τ between the two is then chosen. Therefore, our τ selection scheme is modified as:

$$\tau = \min_i \left\{ \min_{i=p} \left\{ \max \left\{ \frac{\varepsilon p}{k_g \overline{b}_m \overline{b}_p}, \frac{\overline{b}_p}{k_g \overline{b}_m \overline{b}_p} \right\}, \tau_{pj \in nb} \right\}, \tau_{i \neq p} \right\}. \quad (44)$$

The first fraction is still the same as that from equation (43) for only burst production. When $\varepsilon p \leq 1$, we multiply the original selection of $\tau = 1/k_g \bar{b}_m \bar{b}_p$ by $\bar{b}_p$ to include a whole burst event. Further detailed models included effects of chromatin template and promoter structure[38,39], which change the waiting time distribution for next burst event. And thus, the term $\tau = 1/k_g \bar{b}_m$, which is derived from an exponential distribution, needs to be modified accordingly if the chromatin structural change is considered. Other detailed considerations are included in the supplementary information accompanying this work.

With a selected τ, we can determine the protein's burst production number and degradation number according to equation(31). When $p(t)$ is small, a randomly selected degradation number may be so large that negative $p(t+\tau)$ is obtained. If $p(t+\tau)$ becomes negative, we take half of the originally selected τ, such that particle changes become small, and repeat this procedure if necessary, until $p(t+\tau) \geq 0$. Such half-τ scheme is suggested in the work of Cao *et al.*[35].

Removing negative production. The protein's production number in each τ is calculated by the second term in the right-hand side of equation (31) and then rounded to the nearest integers. Shown in Fig. 8a is the Gaussian distribution we used to approximate the burst production. With $\tau = 0.03$, the mean production number is 3 and the standard deviation is about 13.5. From such Gaussian distribution, there is a nearly 40% chance to obtain a negative random number, which is then rounded to a negative production number when it is ≤ -0.5. Even with a high expression level, with $k_g = 100$ and $\bar{b}_p = 100$, the negative part still can be >15% of burst production. A more complete profile for the negative burst percentages is included in the supplementary information. Such negative production also reduces the protein number as the degradation process and leads to many sudden drops in the blue trajectory in Fig. 8b. Negative production is an artifact due to the Gaussian distribution used in the Langevin simulation.

We used a rather simple approach to remove such negative productions and keep the mean of the distribution at the same time. When a negative production number is selected, the negative number is temporarily stored for accumulation with the next production. The production in the current time step is zero, as the silent moment between two bursts. Only when the accumulated protein number becomes positive is there a burst with such a positive number, and the protein number is increased by production. As seen in Fig. 8b, the unrealistic sudden drop is removed in the red trajectory. We note that the red trajectory has a similar shape due to the rapid production and slow degradation as in Fig. 1b.

Data Availability. All data generated or analysed during this study are included in this published article (and its supplementary information file).

References

1. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).
2. Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & van Oudenaarden, A. Regulation of noise in the expression of a single gene. *Nature Genet.* **31**, 69–73 (2002).
3. Blake, W. J., Kaern, M., Cantor, C. R. & Collins, J. J. Noise in eukaryotic gene expression. *Nature* **422**, 633–637 (2003).
4. Raser, J. M. & O'Shea, E. K. Control of stochasticity in eukaryotic gene expression. *Science* **304**, 1811–1814 (2004).
5. Golding, I., Paulsson, J., Zawilski, S. M. & Cox, E. C. Real-time kinetics of gene activity in individual bacteria. *Cell* **123**, 1025–1036 (2005).
6. Yu, J., Xiao, J., Ren, X. J., Lao, K. Q. & Xie, X. S. Probing gene expression in live cells, one protein molecule at a time. *Science* **311**, 1600–1603 (2006).
7. Friedman, N., Cai, L. & Xie, X. S. Linking stochastic dynamics to population distribution: An analytical framework of gene expression. *Phys. Rev. Lett.* **97**, 168302, https://doi.org/10.1103/PhysRevLett.97.168302 (2006).
8. Pedraza, J. M. & van Oudenaarden, A. Noise propagation in gene networks. *Science* **307**, 1965–1969 (2005).
9. Eldar, A. & Elowitz, M. B. Functional roles for noise in genetic circuits. *Nature* **467**, 167–173 (2010).
10. Norman, T. M., Lord, N. D., Paulsson, J. & Losick, R. Memory and modularity in cell-fate decision making. *Nature* **503**, 481–486 (2013).
11. Chepyala, S. R. *et al*. Noise propagation with interlinked feed-forward pathways. *Sci. Rep.* **6**, 23607, https://doi.org/10.1038/srep23607 (2016).
12. Yan, C.-C. S. & Hsu, C.-P. The fluctuation-dissipation theorem for stochastic kinetics-implications on genetic regulations. *Journal of Chemical Physics* **139**, 224109, https://doi.org/10.1063/1.4837235 (2013).
13. Wang, L., Xin, J. & Nie, Q. A critical quantity for noise attenuation in feedback systems. *PLOS Comput. Biol.* **6**, e1000764, https://doi.org/10.1371/journal.pcbi.1000764 (2010).
14. Chen, M., Wang, L., Liu, C. C. & Nie, Q. Noise attenuation in the on and off states of biological switches. *ACS Synthetic Biology* **2**, 587–593 (2013).
15. Ji, N. *et al*. Feedback control of gene expression variability in the caenorhabditis elegans wnt pathway. *Cell* **155**, 869–880 (2013).
16. Gillespie, D. T. General method for numerically simulating stochastic time evolution of coupled chemical-reactions. *J. Comput. Phys.* **22**, 403–434 (1976).
17. Gillespie, D. T. Exact stochastic simulation of coupled chemical-reactions. *J. Phys. Chem.* **81**, 2340–2361 (1977).
18. Gillespie, D. T. The chemical langevin equation. *J. Chem. Phys.* **113**, 297–306 (2000).
19. Peccoud, J. & Ycart, B. Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.* **48**, 222–234 (1995).
20. Thattai, M. & van Oudenaarden, A. Intrinsic noise in gene regulatory networks. *Proc. Natl. Acad. Sci. USA* **98**, 8614–8619 (2001).
21. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mrna synthesis in mammalian cells. *PLOS Biol.* **4**, e309, https://doi.org/10.1371/journal.pbio.0040309 (2006).
22. Dey, S. S., Foley, J. E., Limsirichai, P., Schaffer, D. V. & Arkin, A. P. Orthogonal control of expression mean and variance by epigenetic features at different genomic loci. *Mol. Syst. Biol.* **11**, 806, https://doi.org/10.15252/msb.20145704 (2015).
23. Lin, Y. T. & Doering, C. R. Gene expression dynamics with stochastic bursts: Construction and exact results for a coarse-grained model. *Phys. Rev. E* **93**, 022409, https://doi.org/10.1103/PhysRevE.93.022409 (2016).
24. Lin, Y. T. & Galla, T. Bursting noise in gene expression dynamics: linking microscopic and mesoscopic models. *J. R. Soc. Interface* **13**, 20150772, https://doi.org/10.1098/rsif.2015.0772 (2016).
25. Pedraza, J. M. & Paulsson, J. Effects of molecular memory and bursting on fluctuations in gene expression. *Science* **319**, 339–343 (2008).
26. Schwanhausser, B. *et al*. Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).

27. Albayrak, C. *et al*. Digital quantification of proteins and mrna in single mammalian cells. *Mol. Cell* **61**, 914–924 (2016).
28. Taniguchi, Y. *et al*. Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–538 (2010).
29. Paulsson, J. & Ehrenberg, M. Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks. *Phys. Rev. Lett.* **84**, 5447–5450 (2000).
30. van Kampen, N. G. *Stochastic processes in physics and chemistry*, 3rd edn (Elsevier, Amsterdam, 2007).
31. Elf, J. & Ehrenberg, M. Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome Res.* **13**, 2475–2484 (2003).
32. Grima, R. Linear-noise approximation and the chemical master equation agree up to second-order moments for a class of chemical systems. *Phys. Rev. E* **92**, 042124, https://doi.org/10.1103/PhysRevE.92.042124 (2015).
33. Paulsson, J. Summing up the noise in gene networks. *Nature* **427**, 415–418 (2004).
34. Hensel, Z. *et al*. Stochastic expression dynamics of a transcription factor revealed by single-molecule noise analysis. *Nat. Struct. Mol. Biol.* **19**, 797–802 (2012).
35. Cao, Y., Gillespie, D. T. & Petzold, L. R. Efficient step size selection for the tau-leaping simulation method. *J. Chem. Phys.* **124**, 044109, https://doi.org/10.1063/1.2159468 (2006).
36. Choi, P. J., Cai, L., Frieda, K. & Xie, S. A stochastic single-molecule event triggers phenotype switching of a bacterial cell. *Science* **322**, 442–446 (2008).
37. McManus, J., Cheng, Z. & Vogel, C. Next-generation analysis of gene expression regulation - comparing the roles of synthesis and degradation. *Mol. Biosyst.* **11**, 2680–2698 (2015).
38. Zhang, J., Chen, L. & Zhou, T. Analytical distribution and tunability of noise in a model of promoter progress. *Biophys. J.* **102**, 1247–1257 (2012).
39. Zhang, J. & Zhou, T. Promoter-mediated transcriptional dynamics. *Biophys. J.* **106**, 479–488 (2014).
40. Kumar, N., Singh, A. & Kulkarni, R. V. Transcriptional bursting in gene expression: Analytical results for general stochastic models. *PLOS Comput Biol* **11**, e1004292, https://doi.org/10.1371/journal.pcbi.1004292 (2015).
41. Berthoumieux, S. *et al*. Shared control of gene expression in bacteria by transcription factors and global physiology of the cell. *Mol. Syst. Biol.* **9**, 634, https://doi.org/10.1038/msb.2012.70 (2013).
42. Iyer, S., Park, B. R. & Kim, M. Absolute quantitative measurement of transcriptional kinetic parameters *in vivo*. *Nucleic Acids Res.* **44**, e142, https://doi.org/10.1093/nar/gkw596 (2016).
43. Russell, J. B. & Cook, G. M. Energetics of bacterial-growth - balance of anabolic and catabolic reactions. *Microbiol. Rev.* **59**, 48–62 (1995).
44. Wagner, A. Energy constraints on the evolution of gene expression. *Molecular Biology and Evolution* **22**, 1365–1374 (2005).
45. Higham, D. J. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Review* **43**, 525–546 (2001).

Acknowledgements

We acknowledge the financial support from Academia Sinica and Ministry of Science and Technology of Taiwan (Project 104-2627-M-001-003- and 105-2113-M-001-009-MY4). We thank Shui-Tein Chen for suggestions to this work.

Author Contributions

C.C.S.Y., C.M.Y. and C.P.H. designed the research, analyzed the data. C.C.S.Y., S.R.C. and C.P.H. wrote the article. C.C.S.Y. and C.M.Y. developed the simulation method. C.C.S.Y., S.R.C. and C.M.Y. performed the simulation. C.P.H. supervised the research project. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-16835-y.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.