Supplementary Text

A. Cell signaling cascade and mature protein production can be described with delayed birth-death process.

In cell signaling cascades, the activated signal is transduced throughout a chain of intermediate reaction steps and triggers a response whose intensity decays (Fig. 2A (i)). The dynamics of this process is mainly governed by the intermediate slowest reaction steps, namely rate-limiting steps, which determine the speed of signaling. These rate-limiting steps can be described with a chain of reversible activation and deactivation of signaling molecules (19, 35) (Fig. 2A (ii)). Upon initial signal activation with a rate of \( \lambda_a \), the first inactive intermediate becomes its active form, \( X_0 \), which activates the next inactive intermediate. Throughout this sequential activation with a rate of \( a_i \), the signal is transduced and, finally, the last inactive response molecule becomes its active form, \( X_n \). Each active intermediate is deactivated with a rate of \( r \) and the last response molecule decays with a rate of \( \lambda_d \). This signaling cascade can be modeled as done in previous work (19, 35):

\[
\begin{align*}
\frac{d[X_0]}{dt} &= \lambda_a - r[X_0], \\
\frac{d[X_k]}{dt} &= a_k[X_{k-1}] - r[X_k], \quad \forall k \in \{1,2,3,\ldots,n-1\}, \\
\frac{d[X_n]}{dt} &= a_n[X_{n-1}] - \lambda_d[X_n], \\
[X_k][0] &= 0 \quad \forall k \in \{1,2,3,\ldots,n\}.
\end{align*}
\]

These coupled ordinary differential equations can be simplified by taking the Laplace transform with auxiliary variable \( s \) as

\[
L([X_0]) = \frac{\lambda_a}{s(s+r)},
\]
\[ L([X_k]) = \frac{\lambda_a \prod_{i=1}^{k} a_i}{s(s+r)^{k+1}} \quad \forall k \in \{1, 2, 3, ..., n - 1\}, \]
\[ L([X_n]) = \frac{\lambda_a \prod_{i=1}^{n} a_i}{s(s+r)^{n(s+\lambda_d)}}. \]

Taking the inverse transform, we can solve the differential equations as follows:
\[ [X_0][t] = \left( \frac{\lambda_a}{r} \right) \cdot \frac{\gamma(1,rt)}{\Gamma(1)}, \]
\[ [X_k][t] = \left( \frac{\lambda_a}{r} \right) \cdot \prod_{i=1}^k \left( \frac{a_i}{r} \right) \cdot \frac{\gamma(k+1,rt)}{\Gamma(k+1)} \quad \forall k \in \{1, 2, 3, ..., n - 1\}, \]
\[ [X_n][t] = \int_0^t \left( \frac{\lambda_a}{r} \right) \cdot \prod_{i=1}^n \left( \frac{a_i}{r} \right) \cdot \frac{\gamma(n,rt)}{\Gamma(n)} \cdot \exp(-\lambda_d(t-\tau)) \, d\tau, \quad (S1) \]

where \( \Gamma(\alpha) \) and \( \gamma(\alpha, \beta) \) denote the gamma function and the lower incomplete gamma function, respectively. Note that \( \frac{\gamma(n,rt)}{\Gamma(n)} \) is the cumulative distribution function of the gamma distribution \( \Gamma(n, t_r) \) with a shape parameter of \( n \) and a scale parameter of \( t_r = r^{-1} \). Then, by taking the derivative of Eq. (S1) with respect to \( t \), we obtain
\[ \frac{d[X_n]}{dt} = \lambda_b \frac{\gamma(n,rt)}{\Gamma(n)} - \lambda_d [X_n], \quad (S2) \]

where \( \lambda_b = \lambda_a \cdot \prod_{i=1}^n \left( \frac{a_i}{r} \right) \). Eq. (S2) indicates that the signal amplified throughout the \( n \) intermediate rate-limiting steps triggers the activation of the response molecule \( X_n \) with a rate of \( \lambda_b = \lambda_a \cdot \prod_{i=1}^n \left( \frac{a_i}{r} \right) \). This activation is delayed due to the intermediate rate-limiting steps (Fig. 2A (iii)) with the signal transduction delay, \( \tau \), which follows the gamma distribution \( \Gamma(n, t_r) \) where \( n \) and \( t_r \) represent the number of intermediate rate-limiting steps and the time for each step, which is given by \( r^{-1} \), respectively.

Similarly, mature protein production (Fig. 2B (i)) can also be illustrated with a gamma-distributed delayed one-variable model. Specifically, the protein maturation with multiple intermediate steps (Fig. 2B (ii)) can be modeled as follows:
\[ \frac{d[X_0]}{dt} = \lambda_a - r[X_0], \]
\[ \frac{d[X_k]}{dt} = r([X_{k-1}] - [X_k]) \quad \forall k \in \{1, 2, 3, ..., n - 1\}, \]
\[ \frac{d[X_n]}{dt} = r[X_{n-1}] - \lambda_d [X_n], \]
\[ [X_k][0] = 0 \forall k \in \{0, 1, 2, 3, ..., n\}, \]
where \( \lambda_a \) and \( \lambda_d \) represent the synthesis rate of premature protein, \( X_0 \), and the decay rate of mature protein, \( X_n \), respectively, and \( r \) represents the rate of intermediate reactions for protein maturation (e.g., protein folding and protein oligomerization).
As described above, it can be shown that

$$\frac{d[X_n]}{dt} = \lambda_b \cdot \frac{r(n \tau)}{\Gamma(n)} - \lambda_d [X_n],$$

where $\lambda_b = \lambda_a$. Accordingly, mature protein synthesis also can be described with a birth-death process with a gamma-distributed delay, $\tau \sim \Gamma(n, t_r)$ (Fig. 2B (iii)).

**B. Derivation of the exact mean formula for delayed birth-death process (Eq. (1)).**

The mean number of molecules in the stochastic delayed birth-death process (Fig. 2D) at time $t$, $\mu(t)$, can be formulated using the transient Little’s law (39) as follows:

$$\mu(t) = \int_0^t \langle R_t \rangle \cdot S(t - \tau) d\tau.$$  \hspace{1cm} (S3)

Here, $S(t - \tau)$ denotes the probability that a molecule created at time 0 survives at least until time $t - \tau$, which is $\exp(-\lambda_d (t - \tau))$ since a death event follows a Poisson arrival process with mean death rate, $\lambda_d$. $\langle R_t \rangle$ denotes the mean creation rate of molecules at time $\tau$, $\langle \frac{dX^\xi_t}{dt} \rangle$, where $X^\xi_t$ is the stochastic process representing the number of birth events which have been completed until time $\tau$. In the following paragraphs, we will show that $X^\xi_t$ follows a time-inhomogeneous Poisson process with mean rate $\lambda_b \int_0^\tau g(t) dt$ where $g(t)$ is the probability density function of $\tau \sim \Gamma(n, t_r)$ so that $\langle \frac{dX^\xi_t}{dt} \rangle = \lambda_b \int_0^\tau g(t) dt$, and finally we will derive the exact mean formula $\mu(t)$.

Let $P_k(\tau)$ be the probability that $X^\xi_t = k$. Then,

$$P_k(\tau) = P[X^\xi_t = k] = \sum_{n=0}^{\infty} P[X^\xi_t = k | X^\xi_t = n] P[X^\xi_t = n]$$

where $X^\xi_t$ is the stochastic process representing the number of birth events which have been initiated until time $\tau$. Since $X^\xi_t$ follows a homogeneous Poisson process with mean rate $\lambda_b$ (i.e.,

$$P[X^\xi_t = n] = \frac{e^{-(\lambda_b \tau)} (\lambda_b \tau)^n}{n!},$$

$$P_k(\tau) = \sum_{n=0}^{\infty} P[X^\xi_t = k | X^\xi_t = n] \frac{e^{-(\lambda_b \tau)} (\lambda_b \tau)^n}{n!}. \hspace{1cm} (S4)$$

$P[X^\xi_t = k | X^\xi_t = n]$ can be calculated by using the fact that the probability that a birth event initiated at $\hat{t} \in (0, \tau)$ has not been completed until time $\tau$ is $1 - \int_0^{\tau-\hat{t}} g(t) dt$. Specifically, for any number $n$ of birth events initiated in $(0, \tau)$, the probability that any birth event has not been completed until time $\tau$ is

$$\frac{1}{\tau} \int_0^\tau \left(1 - \int_0^{\tau-\hat{t}} g(t) dt\right) d\hat{t} = \frac{1}{\tau} \int_0^\tau \left(1 - \int_0^{\hat{t}} g(t) dt\right) d\hat{t}$$

because the joint distribution of the initiation times of birth events over $(0, \tau)$ given that $X^\xi_t = n$ is the same as the distribution of $n$ points uniformly distributed over $(0, \tau)$ (40). Hence, the probability that a birth event initiated in $(0, \tau)$ will have been completed until time $\tau$ is

$$1 - \frac{1}{\tau} \int_0^\tau \left(1 - \int_0^{\hat{t}} g(t) dt\right) d\hat{t} = \frac{1}{\tau} \int_0^\tau \left(1 - \int_0^{\hat{t}} g(t) dt\right) d\hat{t}.$$
\[ P[X_T^c = k | X_T^i = n] = \binom{n}{k} \left[ \frac{1}{\tau} \int_0^\tau \int_0^\tau g(t) dt \, d\hat{t} \right]^k \left[ \frac{1}{\tau} \int_0^\tau \left( 1 - \int_0^\tau g(t) dt \right) d\hat{t} \right]^{n-k} \]

for \( k \leq n \), and \( P[X_T^c = k | X_T^i = n] = 0 \) for \( k > n \). Substituting this into Eq. (S4),

\[
P_k(\tau) = \sum_{n=k}^{\infty} \binom{n}{k} \left[ \frac{1}{\tau} \int_0^\tau \int_0^\tau g(t) dt \, d\hat{t} \right]^k \left[ \frac{1}{\tau} \int_0^\tau \left( 1 - \int_0^\tau g(t) dt \right) d\hat{t} \right]^{n-k} \frac{e^{-(\lambda_b\tau)(\lambda_b\tau)^n}}{n!} \\
= e^{-\lambda_b\tau} \left[ \frac{1}{\tau} \int_0^\tau \int_0^\tau g(t) dt \, d\hat{t} \right]^k \sum_{n=k}^{\infty} \frac{(\lambda_b\tau)^{n-k}}{(n-k)!} \left[ \frac{1}{\tau} \int_0^\tau \left( 1 - \int_0^\tau g(t) dt \right) d\hat{t} \right]^{n-k} \\
= e^{-\lambda_b\tau} \left[ \frac{1}{\tau} \int_0^\tau \int_0^\tau g(t) dt \, d\hat{t} \right]^k \sum_{m=0}^{\infty} \frac{1}{m!} \lambda_b \int_0^\tau \left( 1 - \int_0^\tau g(t) dt \right) d\hat{t} \right]^m \\
= e^{-\lambda_b\tau} \left[ \frac{1}{\tau} \int_0^\tau \int_0^\tau g(t) dt \, d\hat{t} \right]^k \left[ \lambda_b \int_0^\tau \left( 1 - \int_0^\tau g(t) dt \right) d\hat{t} \right]^{\tau \int_0^\tau g(t) dt} \exp \left( \lambda_b \int_0^\tau g(t) dt \right). \\
\]

Accordingly, the form of \( P_k(\tau)(= P[X_T^c = k]) \) is \( e^{-\int f(\tau)} f^k(\tau) \) where \( f(\tau) = \lambda_b \int_0^\tau \int_0^\tau g(t) dt \, d\hat{t} \)

and \( \frac{df(\tau)}{d\tau} = \lambda_b \int_0^\tau g(t) dt \). Thus, \( X_T^c \) follows a time-inhomogeneous Poisson process with mean rate \( \langle R_\tau \rangle = \lambda_b \int_0^\tau g(t) dt \).

Finally, by substituting \( \langle R_\tau \rangle \) and \( S(t - \tau) \) into Eq. (S3), we can obtain the mean number of molecules at time \( t \):

\[
\mu(t) = \langle R_\tau \rangle \cdot S(t - \tau) d\tau = \langle R_\tau \rangle \int_0^t \exp(-\lambda_d(t-\tau)) \exp(-\lambda_d t) \, d\tau \\
= \lambda_b \exp(-\lambda_d t) \int_0^t \int_0^\tau g(s) ds \, d\tau \exp(\lambda_d t) \, d\tau \\
= \lambda_b \exp(-\lambda_d t) \left[ \int_0^t \left( \int_0^\tau g(s) ds \right) \exp(\lambda_d t) \, d\tau \right] \left[ \int_0^t \exp(\lambda_d t) \exp(\lambda_d t) \, d\tau \right] \\
= \lambda_b \frac{\lambda_d}{\lambda_d} \left( \int_0^\tau g(t) dt - \exp(-\lambda_d t) \int_0^\tau g(t) \exp(\lambda_d t) \, d\tau \right). \tag{S5}
\]

Note that, in the above derivation, we did not use the condition that \( g(\tau) \) is the gamma probability density function. Thus, even if \( g(\tau) \) is an arbitrary probability density function, \( X_T^c \) follows a time-inhomogeneous Poisson process with the mean rate \( \lambda_b \int_0^\tau g(t) dt \), and thus Eq. (S5) still holds. In the special case that \( g(\tau) \) is an Erlang distribution with the shape parameter of \( n \), Eq. (S5) can be more easily derived by solving moment equations of a Markovian version of our model, the birth-death process having \( n \) intermediate rate-limiting steps with exponential distribution (19).

### C. Derivation of the exact variance formula for delayed birth-death process (Eq. (1)).

The variance of the number of molecules at time \( t \), \( \sigma^2(t) \), can be formulated using the Chemical Fluctuation Theorem (21) as follows:

\[
\sigma^2(t) = \mu(t) + \int_0^t \int_0^t R_{TCF}(\tau_1, \tau_2) S(t - \tau_1) S(t - \tau_2) d\tau_1 \, d\tau_2. \tag{S6}
\]
Here, \( S(t - \tau_i) \) is the probability that a molecule created at time 0 survives at least until time \( t - \tau_i \), which is \( \exp(-\lambda_d(t - \tau_i)) \), since a death event follows a Poisson arrival process with the mean death rate \( \lambda_d \). \( R_{TCF}(\tau_1, \tau_2) \) denotes the time correlation function of the creation rate of molecules, which is defined by \( \langle R_{\tau_1}R_{\tau_2} \rangle - \langle R_{\tau_1} \rangle \langle R_{\tau_2} \rangle - \delta(\tau_1 - \tau_2) \langle R_{\tau_2} \rangle \), where \( \delta(t) \) denotes the Dirac delta function.

Since the number of birth events which have been completed until time \( t \), \( X^\xi_t \), follows a time-inhomogeneous Poisson process (see above), \( \mu_{X^\xi_t} = \sigma^2_{X^\xi_t} \). Thus, by applying the Chemical Fluctuation Theorem to \( X^\xi_t \), we have

\[
\mu_{X^\xi_t} = \sigma^2_{X^\xi_t} = \mu_{X^\xi_t} + \int_0^t \int_0^t R_{TCF}(\tau_1, \tau_2) S_{X^\xi_t}(t - \tau_1) S_{X^\xi_t}(t - \tau_2) d\tau_1 d\tau_2 \quad (S7)
\]

where \( S_{X^\xi_t}(t - \tau_i) \) is the probability that a molecule created at time 0 survives at least until time \( t - \tau_i \) in the stochastic process \( X^\xi_t \). Since a death event does not occur in this process, \( S_{X^\xi_t}(t - \tau_i) = 1 \). Substituting this into Eq. (S7), we have

\[
\int_0^t \int_0^t R_{TCF}(\tau_1, \tau_2) d\tau_1 d\tau_2 = 0. \quad (S8)
\]

Since Eq. (S8) holds for every \( t \),

\[
R_{TCF}(\tau_1, \tau_2) = 0 \text{ for almost every } \tau_1 \text{ and } \tau_2 \in \mathbb{R} \quad (S9).
\]

Then, Eq. (S9) implies that

\[
\int_0^t \int_0^t R_{TCF}(\tau_1, \tau_2) S(t - \tau_1) S(t - \tau_2) d\tau_1 d\tau_2 = 0.
\]

Substituting this into Eq. (S6), we finally get the variance formula:

\[
\sigma^2(t) = \mu(t). \quad (S10)
\]

**D. Moment-based Bayesian inference method (MBI) for a single cell.**

Based on the formulae representing the mean (Eq. (S5)) and variance (Eq. (S10)) of the number of molecules, we developed a Bayesian Markov Chain Monte Carlo (MCMC) method, called the moment-based Bayesian inference method (MBI), that allows one to estimate the birth and death rates, \( \lambda_b \) and \( \lambda_d \), and delay distribution parameters, \( n \) and \( t_r \). First, we constructed a likelihood function for the parameters \( L_{MBI} \) by adopting a Gaussian noise assumption: the likelihood function is proportional to the product of normal distributions centered at \( \bar{\mu}(t) \) with the variance \( \bar{\sigma}^2(t) \), following previous studies (69-71). Specifically, for given data \( x = \left[ x^{(j)}_{t_l} \right]_{l=1}^K \), where \( x^{(j)}_{t_l} \) is the data point measured at time \( t_l \) of the \( j \)-th experiment with a single cell, the likelihood function is constructed as follows:

\[
L_{MBI}(\lambda_b, \lambda_d, n, t_r | x) \propto \prod_{i=1}^K p \left( \bar{x}_{t_l} | \bar{\mu}(t_l), \bar{\sigma}^2(t_l) \right)
\]

where \( \bar{x}_{t_l} = \frac{1}{R} \left( x^{(1)}_{t_l} + \cdots + x^{(R)}_{t_l} \right) \) and \( p(x|\mu, \sigma^2) \) is the probability density function of the normal distribution with mean \( \mu \) and variance \( \sigma^2 \). Here, \( \bar{\mu}(t) \) is given by \( \mu(t) \) (Eq. (S5)), which we
derived with the transient Little’s law (39). \( \bar{\sigma}^2(t) \) is given by \( \frac{1}{R} \bar{x}_{t_i} + \epsilon \), whose first term, \( \frac{1}{R} \bar{x}_{t_i} \), represents the unbiased estimator of the variance of the sample mean, \( \frac{1}{R} \sigma^2(t) \), since it is the same as \( \frac{1}{R} \mu(t_i) \approx \frac{1}{R} \bar{x}_{t_i} \) as proved in Eq. (S10), and the convergence of the sample mean to the true statistics (i.e., true mean and variance) is faster than that of the sample variance (41).
Moreover, the use of the unbiased estimator \( \frac{1}{R} \bar{x}_{t_i} \) instead of \( \frac{1}{R} \mu(t_i) \) allows for explicit derivation of the conditional posterior distribution of \( \lambda_b \), which leads to efficient MCMC sampling (see below for details). The second term (\( \epsilon = 10^{-3} \)) is added to avoid zero variance when the mean of the data is zero. According to Bayes’ Theorem, the joint posterior distribution over the parameters is expressed as:
\[
\pi(\lambda_b, \lambda_d, n, t_r | \mathbf{x}) \propto \pi(\lambda_b) \pi(\lambda_d) \pi(n) \pi(t_r) L_{MBI}(\lambda_b, \lambda_d, n, t_r | \mathbf{x})
\]
where \( \pi(\cdot) \)'s on the right-hand side are the prior distributions of the parameters.

Since direct sampling from the multi-dimensional posterior distribution (Eq. (S11)) is impossible, we used the Gibbs sampling algorithm that sequentially samples a parameter from the conditional posterior distribution of each parameter (72). In the Gibbs sampling algorithm, we can directly sample \( \lambda_b \) from its conditional posterior distribution as it is explicitly derived by setting the normal conjugate prior. First, we can factorize the formula for the mean number of molecules (Eq. (S5)) as:
\[
\mu(t) = \frac{\lambda_b}{\lambda_d} \left( \int_0^t g(\tau) d\tau - \exp(-\lambda_d t) \int_0^t g(\tau) \exp(\lambda_d \tau) d\tau \right) =: \lambda_b h(t; \lambda_d, n, t_r).
\]
Note that the function \( h(t; \lambda_d, n, t_r) \) is independent of \( \lambda_b \). Then the conditional posterior distribution of \( \lambda_b \) with a non-informative normal prior (\( \mu_{pri} = 1 \) and \( \sigma^2_{pri} = 10^6 \)) is given by:
\[
\pi(\lambda_b | \mathbf{x}, \lambda_d, n, t_r) \propto \pi(\lambda_b) L_{MBI}(\lambda_b | \mathbf{x}, \lambda_d, n, t_r)
\]
\[
\propto p(\lambda_b | \mu_{pri}, \sigma^2_{pri}) \prod_{i=1}^K p\left( \bar{x}_{t_i} | \mu = \lambda_b h(t_i; \lambda_d, n, t_r), \sigma^2 = \frac{1}{R} \bar{x}_{t_i} + \epsilon \right)
\]
\[
\propto p(\lambda_b | \mu_{pri}, \sigma^2_{pri}) \prod_{i=1}^K p\left( \lambda_b | \mu = \frac{x_{t_i}}{h(t_i; \lambda_d, n, t_r)}, \sigma^2 = \frac{1}{h(t_i; \lambda_d, n, t_r)} \right)
\]
\[
\sim N\left( \mu_{\lambda b}, \sigma^2_{\lambda b} \right)
\]
(S12)
where \( \sigma^2_{\lambda b} = \left( \sigma^2_{pri} + \sum_{i=1}^K \left( \frac{1}{R} \bar{x}_{t_i} + \epsilon \right) \right)^{-1} \) and
\[
\mu_{\lambda b} = \sigma^2_{\lambda b} \left( \sum_{i=1}^K \left( \frac{1}{R} \bar{x}_{t_i} + \epsilon \right)^{-1} x_{t_i} \right) / h(t_i; \lambda_d, n, t_r).
\]

For the other three parameters, \( \lambda_d, n, \) and \( t_r \), we used Metropolis-Hastings within Gibbs sampling (73) because their conditional posterior distributions cannot be derived explicitly. For each of these parameters, a non-informative gamma prior was assigned (mean = 1, var = 10^6).
Taken together, the MCMC algorithm to sample the parameters, $\lambda_b$, $\lambda_d$, $n$, and $t_r$ from their posterior distributions can be described as follows:

**Step 1.** Initialize values for the parameters.

**Step 2.** Sample the birth rate $\lambda_b$ from its conditional posterior distribution (Eq. (S12)).

**Step 3.** Sample $\lambda_b$, $n$, and $t_r$ from normal proposal distributions and accept the samples with the Metropolis-Hastings acceptance rate.

**Step 4.** Repeat steps 2-3 until convergence of the posterior samples is achieved.

We performed total 110,000 iterations, discarded the first 10,000 iteration as a burn-in period, and obtained the posterior distributions with 1,000 samples by selecting every 100 samples from the remaining 100,000 iterations.

**E. MBI with the mixed-effects model for a heterogeneous cell population.**

To infer parameters for a heterogeneous cell population, we extended MBI by using a mixed-effects modeling approach that works if the collected data can be assumed to be exchangeable (i.e., the data are collected from subjects with the same underlying identity) (25) like our data collected from isogenic cells in the same population (Fig. 5A and G). Specifically, the birth rates (i.e., the data are collected from subjects with the same underlying identity) (28, 60). Based on this assumption, the posterior distribution over the single-cell parameter sets, $\theta = [\theta^{(j)}]_{j=1,...,D} = [(\lambda_b^{(j)}, \lambda_d^{(j)}, n^{(j)}, t_r^{(j)})]_{j=1,...,D}$, and the hyperparameters, $\omega = (\alpha_{\lambda_b}, \beta_{\lambda_b}, \alpha_{\lambda_d}, \beta_{\lambda_d}, \alpha_n, \beta_n, \alpha_{t_r}, \beta_{t_r})$, was constructed according to Bayes’ Theorem as follows:

$$p(\theta, \omega | x) = \pi(\omega)\pi(\theta | \omega) \prod_{j=1}^{D} L_{\text{MBI}}(\theta^{(j)} | x^{(j)}) \quad (S13)$$

where $\pi(\omega)$ and $\pi(\theta | \omega)$ are the prior distribution of $\omega$ and that of $\theta$ for a given $\omega$, and $x^{(j)} = [x_{t_i}^{(j)}]_{i=1,...,K_{j=1,...,D}}$ is a collection of a single time-lapse measurements from each single cell ($x^{(j)} = [x_{t_i}^{(j)}]_{i=1,...,K_{j=1,...,D}}$). In other words, the posterior distribution is the product of $D$ posterior distributions in $D$ single cells, which are constructed using Eq. (S11) (i.e., $\pi(\theta | \omega) \prod_{j=1}^{D} L_{\text{MBI}}(\theta^{(j)} | x^{(j)})$), and the prior distribution of hyperparameters (i.e., $\pi(\omega)$). For each of the hyperparameters, a non-informative gamma prior (mean = 1, var = $10^6$) was used.

In Fig. 4H, I and 5, we do not set the hyperparameters for $n$ (i.e., $\alpha_n$ and $\beta_n$) because $n$ is assumed to be the same among cells (i.e., no population distribution for $n$). Then, similar to Eq. (S13), the posterior distribution over the single-cell parameter sets $\theta = [\theta^{(j)}]_{j=1,...,D}$ =
\[
\left[ (\lambda^*(j), \lambda^d(j), n, t^r(j)) \right]_{j=1, \ldots, D}, \text{ and the hyperparameters, } \omega = (\alpha_{\lambda_b}, \beta_{\lambda_b}, \alpha_{\lambda_d}, \beta_{\lambda_d}, \alpha_{t_r}, \beta_{t_r}), \text{ was constructed as follows:}
\]
\[
p(\theta, \omega | \mathbf{x}) = \pi(\omega) \pi(\theta | \omega) \prod_{j=1}^{D} L_{\text{MBI}}(\theta^{(j)} | \mathbf{x}^{(j)})
\]

where \( \pi(\omega) \) and \( \pi(\theta | \omega) \) are the prior distributions of \( \omega \) and that of \( \theta \) for a given \( \omega \), respectively. For each of the hyperparameters and \( n \), a non-informative gamma prior with \( \mu_{\text{prior}} = 1 \) and \( \sigma_{\text{prior}}^2 = 10^6 \) was used.

One time-lapse measurement from each single cell, which is used for inference of heterogeneous parameters, is usually noisy and shows large short-term fluctuations, unlike the average of multiple measurements (e.g., red line, Fig. 2E). This may lead to biases in the variance estimate. \( \frac{1}{R} \bar{x}_t + \epsilon \). To handle this, the variance estimate is calculated by using the smoothed measurement with a moving average filter of window size five. Note that for the first and second variance estimates, the first three and four measurements were used for the average as there is no measurement before the first one.

Since direct sampling of \( \theta \) and \( \omega \) from the multi-dimensional posterior distribution (Eq. (S13)) is impossible, we use a Metropolis-Hastings within Gibbs algorithm (73) as follows:

**Step 1.** Initialize values for the parameters.

**Step 2.** Sample the parameters \( \lambda^*(j), \lambda^d(j), n^{(j)}, \) and \( t^r(j) \) for all \( j = 1, \ldots, D \), from normal proposal distributions and accept the samples with the Metropolis-Hastings acceptance rate.

**Step 3.** Sample the hyper parameters \( (\alpha_{\lambda_b}, \beta_{\lambda_b}, \alpha_{\lambda_d}, \beta_{\lambda_d}, \alpha_n, \beta_n, \alpha_{t_r}, \beta_{t_r}) \) from normal proposal distributions and accept the samples with the Metropolis-Hastings acceptance rate.

**Step 4.** Repeat steps 2–3 until convergence of the posterior samples is achieved.

Note that if \( n \) is shared among cells, steps 2 and 3 are adjusted accordingly. We performed total 110,000 iterations, discarded the first 10,000 iterations as a burn-in period, and obtained the posterior distributions with 100 samples by selecting every 100 samples from the remaining 100,000 iterations.

**F. Derivation of low-order moments of delayed processes with feedback regulation (Fig. 6A) and MBI with the derived moments.**

In the delayed birth-death process (Fig. 2D), birth initiation rate, \( \lambda_b \), is assumed to be constant over time. However, in the presence of feedback regulation, \( \lambda_b \) is affected by the past dynamics of final output \( X, X^* (t, w) = X(t - \tau_2, w) \), where \( X(t, w): [0, \infty) \times \Omega \to \mathbb{Z}_+ \) is a stochastic process and \( \tau_2 \) is a fixed time delay for feedback regulation. Specifically, the Poisson process describing the birth initiation reaction has stochastic intensity \( \lambda_b (X^*) \). Then, the time-varying exact mean, \( \mu_X (t) \), and variance, \( \sigma_X^2 (t) \), of \( X \) can be formulated as follows:

\[
\mu_X(t) = \int_0^t \int_0^{t_2} \mu_{\lambda_b(X^*)}(t_2 - t_1) g(t_1) S(t - t_2) d t_1 d t_2
\]
\[
\sigma_X^2(t) = \mu_X(t) + \int_0^t \int_0^t f_c(t_1, t_2) S(t - t_1) S(t - t_2) dt_1 dt_2
\]

where

\[
S(t) = \exp(-\lambda_d t)
\]

\[
X^*(t, w) = \begin{cases} 
0 & \text{if } t < \tau_2 \\
X(t - \tau_2, w) & \text{otherwise}
\end{cases}
\]

\[
f_c(t_1, t_2) = \int_0^{t_1} \int_0^{t_2} \left( E[\lambda_b(X^*(t_1 - x_1))\lambda_b(X^*(t_2 - x_2))] \\
- E[\lambda_b(X^*(t_1 - x_1))] E[\lambda_b(X^*(t_2 - x_2))] \right) g(x_1)g(x_2) dx_1 dx_2,
\]

and \( g(t) \) is a time delay distribution of the delayed birth reaction.

**Proof.** Because the Poisson process describing the birth initiation reaction, affected by the feedback regulation, has a stochastic intensity \( \lambda_b(X^*) \), we cannot derive the low-order moments as done in Supplementary Text B and C. To address this, we fix an element \( w \) in the sample so that we can define a subsample space in which the past dynamic \( X^* \) is now a deterministic function of \( t \). In other words, we can define the following subsample space:

\[
S_w = \{ w' : X(\tau - \tau_2, w') = X(\tau - \tau_2, w) \text{ for } \tau_2 \leq \tau \leq t \}.
\]

On \( S_w \), \( X^*(t, w) \), defined below, is just a deterministic function of \( t \).

\[
X^*(t, w) = \begin{cases} 
0 & \text{if } t < \tau_2 \\
X(t - \tau_2, w) & \text{otherwise}.
\end{cases}
\]

Accordingly, on \( S_w \), the Poisson process of the initial reaction has a time-varying deterministic intensity, \( \lambda_b(X^*(t, w)) \), not a stochastic intensity. Then, the creation process of \( X \) is a Poisson process with an intensity \( \int_0^t \lambda_b(X^*(t - t_1, w)) g(t_1) dt_1 \) (74). We can now derive the conditional expectation of \( X \) given \( w \) by using the transient Little’s law as follows:

\[
E[X \mid w] = \int_0^t \left( \int_0^{t_2} \lambda_b(X^*(t_2 - t_1, w)) g(t_1) dt_1 \right) \exp(-\lambda_d(t - t_2)) dt_2
\]

\[
= \int_0^t \int_0^{t_2} \lambda_b(X^*(t_2 - t_1, w)) g(t_1) \exp(-\lambda_d(t - t_2)) dt_1 dt_2. \tag{S14}
\]

Then, using the law of total expectation, we can derive the analytic formula for the mean of \( X \), \( \mu_X(t) \), as follows:

\[
\mu_X(t) = E[E[X \mid w]] = \int_\Omega \int_0^t \int_0^{t_2} \lambda_b(X^*(t_2 - t_1, w)) g(t_1) \exp(-\lambda_d(t - t_2)) dt_1 dt_2 dP(w)
\]

\[
= \int_0^t \int_\Omega \lambda_b(X^*(t_2 - t_1, w)) dP(w) g(t_1) \exp(-\lambda_d(t - t_2)) dt_1 dt_2.
\]
$$\int_0^t \int_0^{t_2} \mu_b(x^\ast)(t_2-t_1)g(t_1) \exp(-\lambda_d(t-t_2)) dt_1 dt_2. \quad (S15)$$

Next, we will derive an analytic formula of $\sigma_X^2$. For this, we will use the law of total variance:

$$\sigma_X^2 = E[\sigma_X^2] + \sigma_{E[X|w]}^2. \quad (S16)$$

First, let us show that $E[\sigma_X^2] = \mu_X$ by applying the Chemical Fluctuation Theorem (21) to the creation process of $X, X^c$, as done in Supplementary Text C. Specifically, since $X^c$ is Poissonian when $w$ is given, $E[X^c|w] = \sigma_{X^c}^2|w$. Then, by applying the Chemical Fluctuation Theorem to $X^c|w$, we can show that the time correlation function of the creation rate of $X|w, R_w(t_1, t_2)$, is zero for almost every $\tau_1$ and $\tau_2 \in \mathbb{R}$. Accordingly, by applying the Chemical Fluctuation Theorem to $X|w$,

$$\sigma_{X|w}^2 = E[X|w] + \int_0^t \int_0^t R_w(t_1, t_2) \exp(-\lambda_d(t-t_1)) \exp(-\lambda_d(t-t_2)) dt_1 dt_2 = E[X|w].$$

Thus, $\sigma_{X|w}^2 = E[X|w]$ and $E[\sigma_{X|w}^2] = E[E[X|w]] = \mu_X$. Substituting this into Eq. (S16), we get $\sigma_X^2 = \mu_X + \sigma_{E[X|w]}^2$. Since $\sigma_{E[X|w]}^2 = E[E[X|w]^2] - \mu_X^2$,

$$\sigma_X^2 = \mu_X + E[E[X|w]^2] - \mu_X^2 = \mu_X + \int_0^t \int_0^t (E[\lambda_b(X^\ast(t_2-t_1), w)]g(t_1) \exp(-\lambda_d(t-t_2)) dt_1 dt_2) dP(w) - \mu_X^2$$

$$= \mu_X + \int_0^t \int_0^t f_c(t_1, t_2) \exp(-\lambda_d(t-t_1)) \exp(-\lambda_d(t-t_2)) dt_1 dt_2 \quad (S17)$$

where

$$f_c(t_1, t_2) = \int_0^{t_1} \int_0^{t_2} (E[\lambda_b(X^\ast(t_1-x_1), \lambda_b(X^\ast(t_2-x_2))])$$

$$- E[\lambda_b(X^\ast(t_1-x_1)]E[\lambda_b(X^\ast(t_2-x_2)]]) g(x_1)g(x_2) dx_1 dx_2,$$

and the second equality holds based on Eq. (S14).

Based on the formulae (Eq. S15 and S17), we constructed the likelihood function as done in Supplementary Text D:

$$L(\lambda_b, \lambda_d, n, t_r, R, \tau_2|X) \propto \prod_{i=1}^K p(\tilde{x}_{i}, |b(t_i), \tilde{\sigma}^2(t_i))$$

where $p(x|\mu, \sigma^2)$ is the probability density function of the normal distribution with mean $\mu$ and variance $\sigma^2$ and $\tilde{\mu}(t)$ is given by $\mu_X(t)$ (Eq. (S15)). For $\tilde{\sigma}^2(t_i)$, the first term of the newly derived variance formula (i.e., $\mu_X$ in Eq. (S17)) was only exploited because solving a delayed integro-differential equation corresponding to the second term of Eq. (S17) is challenging due to the lack of well-established numerical algorithms. Specifically, the sample mean of the data, the unbiased estimator of $\mu_X$, was used as done in Supplementary Text D. Despite this, the inference
results are accurate (Fig. 6C and fig. S7). Finally, the joint posterior distribution for the parameters is expressed as

$$
\pi(\lambda_b, \lambda_d, n, t, R, \tau_2 | x) \propto \pi(\lambda_b)\pi(\lambda_d)\pi(n)\pi(t)\pi(R)\pi(\tau_2) L(\lambda_b, \lambda_d, n, t, R, \tau_2 | x)
$$

where $$\pi(\cdot)$$’s are the prior distributions of the parameters. To infer the parameters for a heterogeneous cell population, we extended this by using a mixed-effects modeling approach as done in Supplementary Text E.

G. Computational package of MBI.

We developed and provide a user-friendly computational package of moment-based Bayesian inference method (MBI) implemented for R software. To apply the package to the experimentally measured time-lapse data, users just need to store the data as a csv file in the folder containing R codes of the package and run the codes. A more detailed manual for implementing the package is described below.

**Step 1.** Download the R codes of the package and example CSV files (e.g., ‘input_data_example.csv’ and ‘post_samples_MBI_single_cell_example.csv’) from the following Zenodo repository: https://doi.org/10.5281/zenodo.5904961.

**Step 2.** Store the time-lapse data as a CSV file named ‘input_data.csv’, of which format is the same as that of ‘input_data_example.csv’, in the folder including the R codes of the package. Set the folder as the working directory by using R function ‘setwd()’. The first column of the csv file represents the time points when the data were measured. The second, third, …, $D + 1$-th columns represent the single-cell time-lapse data measured in the first, second, …, $D$-th single cells, respectively. Note that if the initial values of the time-lapse data filled in the csv file are not zero, the preprocessed data, obtained by subtracting the initial values from the input data, are used for the estimation.

**Step 3.** If the time traces in the data monotonically increase and are saturated, run ‘MBI_single_cell_main.R’, ‘MBI_multiple_cells_main.R’ or ‘MBI_multiple_cells_same_n_main.R’. Specifically, if the data were obtained from a single cell, run ‘MBI_single_cell_main.R’. If the data were obtained from multiple cells and the number of their rate-limiting steps can be different, run ‘MBI_multiple_cells_main.R’. If the data were obtained from multiple cells and it is assumed that the number of rate-limiting steps is the same between cells, run ‘MBI_multiple_cells_same_n_main.R’. If the time traces are fluctuating, run ‘MBI_feedback_single_cell_main.R’, ‘MBI_feedback_multiple_cells_main.R’ or ‘MBI_feedback_multiple_cells_same_n_main.R’ depending on the user’s purpose. To obtain efficient convergence of the posterior samples, the iteration number of Gibbs sampling and the variance of the normal proposal distributions for the parameters might need to be tuned. To do this, please see ‘Tuning_strategy.pdf’ in the Zenodo repository.

Implementation of the computational package gives the following outputs containing the estimates of the parameters.
Outputs of ‘MBI_single_cell_main.R’ or ‘MBI_feedback_single_cell_main.R’
If ‘MBI_single_cell_main.R’ is run, the posterior samples of $\lambda_b$, $\lambda_d$, $n$, and $t_r$ are stored as a CSV file named ‘post_samples_MBI_single_cell.csv’, whose format is the same as that of ‘post_samples_MBI_single_cell_example.csv’. The means and the standard deviations of the samples are stored in a CSV file named ‘mean_std_post_samples_MBI_single_cell.csv’, whose format is the same as that of ‘mean_std_post_samples_MBI_single_cell_example.csv’. If ‘MBI_feedback_single_cell_main.R’ is run, all quantities from the parameters, $\lambda_b$, $\lambda_d$, $n$, $t_r$, $R$, and $\tau_2$, described above are saved with the similar file names including the suffix ‘feedback’.

Outputs of ‘MBI_multiple_cells_main.R’ or ‘MBI_feedback_multiple_cells_main.R’
If ‘MBI_multiple_cells_main.R’ is run, the posterior samples of the parameters of $j$-th cell, $\lambda_b^{(j)}$, $\lambda_d^{(j)}$, $n^{(j)}$, and $t_r^{(j)}$, are stored in a CSV file named ‘post_samples_jth_cell_MBI_multiple_cells.csv’. The means and the standard deviations of the samples of the $j$-th cell are stored in a CSV file named ‘mean_std_post_samples_jth_cell_MBI_multiple_cells.csv’. The posterior samples of the hyperparameters, $\alpha_{\lambda_b}, \beta_{\lambda_b}, \alpha_{\lambda_d}, \beta_{\lambda_d}, \alpha_{n}, \beta_{n}, \alpha_{t_r}, \beta_{t_r}$, and their means and standard deviations are stored in CSV files named ‘post_samples_hyperparam_MBI.csv’ and ‘mean_std_post_samples_hyperparam_MBI.csv’, respectively. If ‘MBI_feedback_multiple_cells_main.R’ is run, all quantities from the parameters, $\lambda_b$, $\lambda_d$, $n$, $t_r$, $R$, and $\tau_2$, described above are saved with the similar file names including the suffix ‘feedback’.

Outputs of ‘MBI_multiple_cells_same_n_main.R’ or ‘MBI_feedback_multiple_cells_same_n_main.R’
If ‘MBI_multiple_cells_same_n_main.R’ is run, the posterior samples of the parameters of $j$-th cell, $\lambda_b^{(j)}$, $\lambda_d^{(j)}$, $n^{(j)}$, and $t_r^{(j)}$, are stored in a CSV file named ‘post_samples_jth_cell_MBI_multiple_cells_same_n.csv’. Note that the posterior samples of $n$ are the same among all single cells. The means and the standard deviations of the samples of the $j$-th cell are stored in a CSV file named ‘mean_std_post_samples_jth_cell_MBI_multiple_cells_same_n.csv’. The posterior samples of the hyperparameters, $\alpha_{\lambda_b}, \beta_{\lambda_b}, \alpha_{\lambda_d}, \beta_{\lambda_d}, \alpha_{t_r}, \beta_{t_r}$, and their means and standard deviations are stored in CSV files named ‘post_samples_hyperparam_MBI_same_n.csv’ and ‘mean_std_post_samples_hyperparam_MBI_same_n.csv’, respectively. If ‘MBI_feedback_multiple_cells_same_n_main.R’ is run, all quantities from the parameters, $\lambda_b$, $\lambda_d$, $n$, $t_r$, $R$, and $\tau_2$, described above are saved with the similar file names including the suffix ‘feedback’.

H. Application of MBI to single-cell time-lapse microscopy data.
The time-lapse microscopy data used in Fig. 5 were previously collected (3) and provided by the authors of (3). In each microscopic experiment of the study, one of three antibiotic stresses, tetracycline (TET), trimethoprim (TMP), and nitrofurantoin (NIT), was given to colonies of *Escherichia coli* (*E. coli*) cells. In response to the stress, either yellow fluorescent protein (YFP) or cyan fluorescent protein (CFP), having similar maturation time (75), was expressed from a promoter. Specifically, time-lapse YFP expressions from eight promoters (*dnaK*, *cspA*, *ydiU*, *ahpC*, *iscR*, *ndrH*, *rpsA*, and *rpmE*) and time-lapse CFP expressions from one promoter (*iscR*) in response to TET were recorded. Moreover, time-lapse YFP expression from 12 promoters (*gadW*, *gadA*, *folA*, *recA*, *fpr*, *purT*, *purM*, *ldhA*, *guaB*, *gadB*, *osmC*, and *dps*) in response to TMP, and time-lapse YFP expression from four promoters (*fpr*, *ybjC*, *recA*, and *cysK*) in response to NIT were recorded. For each experiment, at most four replications were performed (see (3) for details).

Before applying MBI to the data, we conducted data preprocessing to prevent bias in the estimation. In this experimental data, cell division occurred during the experiment and thus progeny cells were derived. For instance, in the first experiment with *dnaK* promoter (red curves, Fig. 5A), at the time of TET addition (i.e., \( t = 0 \) hr), 14 cells were present. After adding the stress, cell division occurred and thus the number of cells increased to 160 during the experiment (3). This resulted in the tree structure of the data as observed in previous work (3, 13). If the data of all progeny cells are considered as independent trajectories, a bias can occur because fluorescence trajectories of the cells originating from the same mother cell overlap before cell division.

To circumvent this, we performed a weighted average on each cell lineage and obtained the mean time-lapse gene expression level in each single cell that was present at the stress addition time, \( t = 0 \) hr (i.e., 14 cells for *dnaK* colony 1). Specifically, if a single cell that was present at \( t = 0 \) hr gave rise to \( N \) progeny cells during the experiment, we obtained the collection of data points, \( \left[ x_{t_i}^{(j)} \right]_{i=1,...,K, j=1,...,N} \), where \( x_{t_i}^{(j)} \) is the data point measured at time \( t_i \) of a progeny cell \( j \). If the progeny cell \( j \) was formed by cell division at \( t_c \), its data points were the same as that of its mother cell before \( t_c \). For each junction where a cell division occurred, the mean of the two branches can be regarded as a mother cell’s trajectory, which would have existed unless the division occurred. Thus, we merged these two branches by taking their average, and we continued this merging until the single mean time-lapse gene expression level was obtained. Consequently, the mean gene expression level in each single cell that was present at \( t = 0 \) hr (Fig. 5A), denoted by \( x_{t_i} \), is given by

\[
x_{t_i} = \sum_{j=1}^{N} \frac{x_{t_i}^{(j)}}{n^{(j)}},
\]

where \( n^{(j)} \) is the number of cell divisions that occurred before the cell division forming the progeny cell \( j \) (i.e., the number of junctions until the progeny cell was formed). Finally, for each
mean time-lapse gene expression level, we subtracted the first data point from all the data points to exclude the basal gene expression level and focus on the expression level induced by the stress.

We next filtered the preprocessed data if the mean gene expression level induced by the stress was so low so that it could not be clearly distinguished from basal gene expression noise before the stress addition. Specifically, in a colony in which the experiment was performed, for each preprocessed gene expression level, we calculated the ratio between the standard deviation of the data before the stress addition (i.e., background noise) and the average of the data after the stress addition. Then, if the mean value of the ratios from all cells in the colony was larger than 0.1, we excluded the colony from data analysis. Through this filtering, among the genes described above, all colonies for cysK, fpr, gadA, gadB, gadW, guaB, ldhA, recA, rpmE, rpsA, and ydiU were excluded.

We further filtered the data if the time-lapse gene expression level largely decreased after it peaked as this indicated that the amplified stress signal activation rate ($\lambda_b$, Fig. 2A (iii)) changed over time, possibly due to autoregulation and the short half-life of an acid stress regulator (2), so that the data could not be described with our model (Fig. 2A and D). Specifically, if the time-lapse gene expression data such that the ratio of its level at the last time point to its maximum level was smaller than 0.8 occupied more than a half of the total data, we excluded the colony. After this filtering, all colonies for cspA, folA, osmC, and fpr and colony 1 for dps and nrdH were excluded.

To infer parameters from the preprocessed and filtered data (Fig. 5G), we assigned a strongly informative gamma prior (mean=0.35 hr$^{-1}$ and CV = 0.1) to $\lambda_d$, which was adopted based on experimentally measured cell growth rate after the stress addition (3). Specifically, we fixed the hyperparameters $\alpha_{\lambda_d}$ and $\beta_{\lambda_d}$ as 100 and 0.0035 hr$^{-1}$, respectively, so that the mean and CV of the gamma prior were 0.35 hr$^{-1}$ and 0.1, respectively. Then, with the informative prior, we applied MBI to the preprocessed and filtered data (Fig. 5G).

For the expression data from two colonies for fpr promoter in Fig. 6D and 6E, we applied MBI with a delayed process including feedback regulation (see Supplementary Text F) to the data after preprocessing it as described above. We again adopt the strongly informative gamma prior to $\lambda_d$ and exploited the mixed-effects model. On the other hand, we did not use a mixed-effects modeling approach for the expression data from various promoters in Fig. 6F and fig. S8, which were previously collected (2), because only a single time trace was given for each promoter. In this analysis, we fixed $\lambda_d$ for each promoter as it can be directly estimated from the experimentally measured cell growth rate after antibiotic addition (2).

I. A large number of rate-limiting steps leads to large cell-to-cell variability in stress response intensity.
In this section, we describe the relationship between the intermediate rate-limiting steps (n), the amplified activation rate of the final output ($\lambda_b$), and the heterogeneity of response intensity at steady state ($\lambda_b/\lambda_d$).

The initial activation rate ($\lambda_a$) is amplified throughout the intermediate rate-limiting steps and the amplified activation rate of the final output ($\lambda_b$) can be formulated as $\lambda_b = \lambda_a \prod_{i=1}^{n} a_t^i$ (see Supplementary Text A). Then, the coefficient of variation of $\lambda_b$ in a heterogeneous cell population can be expressed as follows:

$$CV_{\lambda_b,n} = \frac{\sqrt{E\left[\left(\lambda_a \prod_{i=1}^{n} a_t^i\right)^2\right]} - E\left[\lambda_a \prod_{i=1}^{n} a_t^i\right]}{E\left[\lambda_a \prod_{i=1}^{n} a_t^i\right]} = \frac{\sqrt{E\left[\lambda_a \prod_{i=1}^{n} a_t^i\right]^2} - 1}{\sqrt{E\left[\lambda_a \prod_{i=1}^{n} a_t^i\right]^2} - 1}$$

where $E[\cdot]$ denotes the expectation of the random variable.

Then, $\lambda_a$, $a_i$, and $r^{-1}(= t_r)$ are assumed to follow independent gamma distributions (i.e., $\lambda_a \sim \Gamma(\alpha_{\lambda_a}, \beta_{\lambda_a})$, $a_i \sim \Gamma(\alpha_{a_i}, \beta_{a_i})$, and $r^{-1} = t_r \sim \Gamma(\alpha_{t_r}, \beta_{t_r})$) following previous work (28, 60) and Supplementary Text E. Accordingly,

$$CV_{\lambda_b,n} = \frac{\sqrt{E\left[\lambda_a \prod_{i=1}^{n} a_t^i\right]^2} - 1}{\sqrt{E\left[\lambda_a \prod_{i=1}^{n} a_t^i\right]^2} - 1} = \frac{\sqrt{E\left[\lambda_a \prod_{i=1}^{n} a_t^i\right]^2} - 1}{\sqrt{E\left[\lambda_a \prod_{i=1}^{n} a_t^i\right]^2} - 1}$$

When $X \sim \Gamma(\alpha_X, \beta_X)$, $E[X^n] = \beta_X^n \Gamma(n + \alpha_X)/\Gamma(\alpha_X)$ where $\Gamma(\cdot)$ is the gamma function, so

$$CV_{\lambda_b,n} = \frac{\sqrt{\left(\frac{1}{\alpha_a} + 1\right) \prod_{i=1}^{n} \left(\frac{1}{\alpha_{a_i}} + 1\right) \frac{E[t_r^{2n}]}{E[t_r^{2n}]} - 1}}{\sqrt{\left(\frac{1}{\alpha_a} + 1\right) \prod_{i=1}^{n} \left(\frac{1}{\alpha_{a_i}} + 1\right) \left(\frac{1}{\alpha_{t_r}}\right)^2 \frac{\Gamma(2n + \alpha_{t_r})}{\Gamma(\alpha_{t_r})} - 1}}$$

Then,

$$CV_{\lambda_b,n+1} = \frac{\frac{1}{\alpha_{a_n+1}} + 1}{\alpha_{a_n+1}} \prod_{i=1}^{n+1} \left(\frac{1}{\alpha_{a_i}} + 1\right) \left(\frac{1}{\alpha_{t_r}}\right)^2 \frac{\Gamma(2n + 2 + \alpha_{t_r})}{\Gamma(\alpha_{t_r})} - 1$$

$$= \frac{\frac{1}{\alpha_{a_n+1}} + 1}{\alpha_{a_n+1}} \prod_{i=1}^{n+1} \left(\frac{1}{\alpha_{a_i}} + 1\right) \left(\frac{1}{\alpha_{t_r}}\right)^2 \frac{\Gamma(2n + 2 + \alpha_{t_r})}{\Gamma(\alpha_{t_r})} - 1$$
\[
\begin{align*}
\sqrt{\frac{1}{\alpha_a + 1} \prod_{i=1}^{n} \left( \frac{1}{\alpha_{ai}} + 1 \right) \left( \frac{1}{\alpha_{tr}} \right)^{2n} \frac{r(2n+\alpha_{tr})}{r(\alpha_{tr})} } - 1 &= CV_{\lambda_b,n} \\
\text{The inequality holds because } \frac{(2n+\alpha_{tr})(2n+\alpha_{tr})}{\alpha_{tr}^2} > 1 \text{ and } \left( \frac{1}{\alpha_{a,n+1}} + 1 \right) > 1. \text{ Thus, when the number of intermediate rate-limiting steps (i.e., } n \text{) increases, the magnitude of the heterogeneity of } \lambda_b \text{ (i.e., } CV_{\lambda_b,n} \text{) increases.}
\end{align*}
\]

Even if the underlying distribution of the parameter values is arbitrary (i.e., not gamma), we can show that the increase in the rate-limiting steps increases the magnitude of cell-to-cell variability in response intensity as follows. From the independence assumption among \( \lambda_a, a_i, \) and \( r^{-1}(= t_r), \) we have

\[
CV_{\lambda_b,n} = \frac{E[\left(\lambda_a \prod_{i=1}^{n} a_i^2\right)]^2}{E[\left(\lambda_a \prod_{i=1}^{n} a_i^2\right)]} - 1 = \frac{E[\lambda_a^n(\prod_{i=1}^{n} a_i^2)]E[t_r^{2n}]}{E[\lambda_a]E[\prod_{i=1}^{n} a_i]E[t_r^{2n}]} - 1, \quad \text{and}
\]

\[
CV_{\lambda_b,n+1} = \frac{E[\lambda_a^n(\prod_{i=1}^{n+1} a_i^2)]E[t_r^{2n+2}]}{E[\lambda_a]E[\prod_{i=1}^{n+1} a_i]E[t_r^{2n+2}]} - 1.
\]

It is clear that \( CV_{\lambda_b,n+1} \geq CV_{\lambda_b,n} \) if and only if \( \frac{E[a_{n+1}^2]}{E[a_{n+1}]^2} \leq \frac{E[t_r^{2n+2}]}{E[t_r^{2n}]E[t_r^2]} \geq 1. \) According to the moment monotonicity (i.e., \( E[|X|^c] \leq E[|X|^d] \) for \( c \leq d, \) \( E[a_{n+1}] \leq E[a_n] \leq E[a_{n+1}] \) and \( E[t_r^{2n}]E[t_r^2] \leq E[t_r^{2n}]E[t_r^{2n+1}] = E[t_r^{2n}] \frac{1}{n} = E[t_r^{2n}] \frac{n+1}{n} \leq E[t_r^{2n+2}] \). Therefore, \( \frac{E[a_{n+1}^2]}{E[a_{n+1}]^2} \frac{E[t_r^{2n+2}]}{E[t_r^{2n}]E[t_r^2]} \geq 1 \) so that \( CV_{\lambda_b,n+1} \geq CV_{\lambda_b,n} \). In other words, when \( n \) increases, the magnitude of the heterogeneity of \( \lambda_b \) (i.e., \( CV_{\lambda_b,n} \)) increases.

In the above, we have shown that the CV of \( \lambda_b \) increases as \( n \) increases. Based on this, we can show that the heterogeneity of response intensity at steady state (\( \lambda_b/\lambda_d \)) also increases as \( n \) increases if the distribution of \( \lambda_d \) is independent of that of \( \lambda_b \) as follows. The CV of response intensity at steady state (\( \lambda_b/\lambda_d \)) can be expressed as

\[
CV\left( \frac{\lambda_b}{\lambda_d} \right) = \frac{\sqrt{E\left[ \left( \frac{\lambda_b}{\lambda_d} \right)^2 \right]}}{\sqrt{E\left[ \frac{\lambda_b}{\lambda_d} \right]^2}} - 1 = \frac{E[\lambda_b^2]E[\lambda_d^{-2}]}{E[\lambda_b]^2E[\lambda_d^{-1}]^2} - 1 = \sqrt{\frac{E[\lambda_b^2]}{E[\lambda_b]^2} - 1} \left( \frac{E[\lambda_d^{-2}]}{E[\lambda_b]^2} + \frac{E[\lambda_b^2]}{E[\lambda_d^{-1}]^2} \right) - 1.
\]

This shows that \( CV(\lambda_b/\lambda_d) \) increases as \( CV(\lambda_b) \) increases. Thus, since \( CV_{\lambda_b,n+1} \geq CV_{\lambda_b,n} \) as shown above, \( CV(\lambda_b/\lambda_d) \) also increases as \( n \) increases (Fig. 5M).
J. Derivation of low-order moments or probability generating functions of non-Markovian delayed processes with downstream regulation, cell division, or bursting.

i) Downstream regulation (fig. S10A)

In the delayed birth-death process (Fig. 2D), an initial signal activation rate, \( \lambda_b \), is assumed to be constant over time. However, it can vary due to influence of upstream molecules, whose abundance randomly fluctuates, which potentially leads to cell-to-cell variability (13). This can be incorporated to the model by assuming that \( \lambda_b \) is a function of the number of upstream molecules that is the stochastic process denoted as \( U \) (i.e., \( \lambda_b(U) \)). Specifically, let the Poisson process describing initial activation have stochastic intensity, \( \lambda_b(U) \), where \( U(t,w) \): \([0,\infty) \times \Omega \to \mathbb{Z}_+\) is an arbitrary stochastic process with a sample space \( \Omega \). In addition, let the random time delay \( \tau \) follow an arbitrary distribution whose density function is \( g(\tau) \). Then, the time-varying exact low-order moments, mean, \( \mu_X(t) \), and variance, \( \sigma^2_X(t) \), of \( X \) can be formulated as follows:

\[
\mu_X(t) = \int_0^t \int_0^{t_2} \mu_{\lambda_b(U)}(t_2-t_1) g(t_1) S(t-t_2) dt_1 dt_2
\]

\[
\sigma^2_X(t) = \mu_X(t) + \int_0^t \int_0^t f_c(t_1,t_2) S(t-t_1) S(t-t_2) dt_1 dt_2
\]

where

\[
S(t) = \exp(-\lambda_d t)
\]

\[
f_c(t_1,t_2) = \int_0^{t_1} \int_0^{t_2} \left( E[\lambda_b(U(t_1-x_1)) \lambda_b(U(t_2-x_2))] - E[\lambda_b(U(t_1-x_1))] E[\lambda_b(U(t_2-x_2))] \right) g(x_1) g(x_2) dx_1 dx_2.
\]

Proof. Because the Poisson process describing the birth initiation reaction has a stochastic intensity \( \lambda_b(U(t,w)) \), the low-order moments cannot be derived as done in Supplementary Text B and C. To tackle this, we fix an element \( w \) in the sample so that \( U(t,w) \) is now a deterministic function of \( t \), and thus the Poisson process of the initial reaction has a time-varying deterministic intensity, \( \lambda_b(U(t,w)) \), not a stochastic intensity. Then, the creation process of \( X \) is a Poisson process with an intensity \( \int_0^t \lambda_b(U(t-t_1,w)) g(t_1) dt_1 \) (74). We can now derive the conditional expectation of \( X \) given \( w \) by using the transient Little’s law as follows:

\[
E[X|w] = \int_0^t \left( \int_0^{t_2} \lambda_b(U(t_2-t_1,w)) g(t_1) dt_1 \right) \exp(-\lambda_d(t-t_2)) dt_2
\]

\[
= \int_0^t \int_0^{t_2} \lambda_b(U(t_2-t_1,w)) g(t_1) \exp(-\lambda_d(t-t_2)) dt_1 dt_2.
\]

Then, using the law of total expectation and the law of total variance as in Supplementary Text F, we can derive \( \mu_X(t) \) and \( \sigma^2_X(t) \) as follows:

\[
\mu_X(t) = E[E[X|w]] = \int_\Omega \int_0^t \int_0^{t_2} \lambda_b(U(t_2-t_1,w)) g(t_1) \exp(-\lambda_d(t-t_2)) dt_1 dt_2 dP(w)
\]

\[
\sigma^2_X(t) = E[E[X^2|w]] - \mu_X(t) = \int_\Omega \int_0^t \int_0^{t_2} \lambda_b(U(t_2-t_1,w)) g(t_1) \exp(-\lambda_d(t-t_2)) dt_1 dt_2 dP(w) - \mu_X(t)
\]
\[
\begin{align*}
\sigma_X^2 &= E[E[X|w]] + \sigma_{E[X|w]} = \mu_X + \sigma_{E[X|w]} \\
&= \mu_X + \int_0^\infty \left( \int_0^\infty \int_{t_2}^{t_2} \lambda_b(U(t_2 - t_1, w))g(t_1) \exp(-\lambda_d(t - t_2)) dt_1 dt_2 \right)^2 dP(w) - \mu_X^2. \\
&= \mu_X + \int_0^\infty \int_0^\infty \int_{t_2}^{t_2} f_c(t_1, t_2) \exp(-\lambda_d(t - t_1)) \exp(-\lambda_d(t - t_2)) dt_1 dt_2
\end{align*}
\]

where

\[
f_c(t_1, t_2) = \int_0^{t_1} \int_{t_2}^{t_2} (E[\lambda_b(U(t_1 - x_1))]\lambda_b(U(t_2 - x_2))] \\
- E[\lambda_b(U(t_1 - x_1))]E[\lambda_b(U(t_2 - x_2))] \quad g(x_1)g(x_2) dx_1 dx_2.
\]

ii) Cell division (fig. S10B)

Cell division-induced dilution can lead to cell-to-cell variability (56, 76). While we used the simple first-order decay reaction to implicitly model cell division-induced dilution, it can underestimate the noise (55). To address this, the delayed model can be extended by incorporating binomial partitioning of proteins at cell division following previous work (55, 56) (fig. S10B). For the extended model, the probability generating function (PGF) can be derived as follows:

\[
G(z, t) = \left[ \prod_{s=1}^{\left[ \frac{t}{T} \right]} \exp \left( \psi \left( \frac{2^{s-1}+z}{2^s}, T \right) \right) \right] \exp \left( \psi \left( z, t - T \times \left[ \frac{t}{T} \right] \right) \right),
\]

where \( \left[ x \right] \) is the greatest integer less than or equal to \( x \), \( T \) is the period of cell division, \( \psi(z, t) = -\lambda_b \int_0^t g(t_1) dt_1 \int_{t_2}^{t_2} (1 - z) \), and \( g(\tau) \) is an arbitrary time delay distribution of the birth event. Note that time \( t \) is equivalent to time \( t - T \times \left[ \frac{t}{T} \right] \) at generation \( \left[ \frac{t}{T} \right] + 1 \) (i.e., the number of completed cell divisions is \( \left[ \frac{t}{T} \right] \)). Then, using the derived PGF, one can calculate the \( k \)-th moment (i.e., \( \left( z \frac{\partial}{\partial z} \right)^k G(z, t) \bigg|_{z=1^-} \)). Here, we used the convention that

\[
\prod_{s=1}^{\left[ \frac{t}{T} \right]} \exp \left( \psi \left( \frac{2^{s-1}+z}{2^s}, T \right) \right) = 1 \text{ if } \left[ \frac{t}{T} \right] < 1.
\]

Proof. Following a previous study (55), we assume that cell division occurs at regular time intervals of length \( T \), which has been observed in several cell types (e.g., early frog embryos) (77). Then, the chemical master equation (CME) describing our system at time \( t = \hat{t} + (j - 1) \times T \) is given by

\[
\frac{dP(n, \hat{t})}{dt} = \left( \lambda_b \int_0^\hat{t} g(\tau) d\tau \right) \left( P^{j-1}(n, \hat{t}) - P^j(n, \hat{t}) \right)
\]
where $P^j(n, \hat{t})$ is the probability that the number of molecule $X$ is $n$ at cell age $\hat{t}$ ($0 \leq \hat{t} \leq T$) in generation $j = [t/T] + 1$. Here, $\lambda_b \int_0^{\hat{t}} g(\tau) d\tau$ is the propensity for the delayed birth reaction (Supplementary Text B). Since the PGF $G^j$ of $P^j$ is defined by $G^j(n, \hat{t}) = \sum_n \left( z^n \cdot P^j(n, \hat{t}) \right)$, the differential equation associated with the PGF corresponding to the CME is given by

$$\frac{\partial G^j(z, \hat{t})}{\partial \hat{t}} = - \left( \lambda_b \int_0^{\hat{t}} g(\tau) d\tau \right) G^j(z, \hat{t})(1 - z).$$

This equation is a linear ODE with respect to $G^j(z, \hat{t})$, so its time-dependent solution is easily obtained as

$$G^j(z, \hat{t}) = G^j(z, 0) \exp \left( \psi(z, \hat{t}) \right) \quad (S18)$$

where $\psi(z, \hat{t}) = -\lambda_b \int_0^{\hat{t}} \int_{\tau_1}^{\tau_2} g(\tau_1) d\tau_1 d\tau_2 (1 - z)$. Note that $G^j(z, 0) = \sum_n \left( z^n \cdot P^j(n, 0) \right)$ is the PGF corresponding to the protein distribution at cell division (i.e., $\hat{t} = 0$) in generation $j$. Then, by introducing the binomial partitioning with the probability $1/2$, we can obtain a relationship between the protein distribution at cell division of a cell in generation $j$ and the distribution at the birth of the progeny cell in generation $j + 1$,

$$P^{j+1}(n, 0) = \sum_{i=n}^{\infty} \binom{i}{n} 2^{-i} P^j(i, T).$$

By summing both sides, we get the relation for the PGF:

$$G^{j+1}(z, 0) = G^j \left( \frac{1+z}{2}, T \right). \quad (S19)$$

Here, substituting Eq. (S19) into Eq. (S18) with index $j + 1$, we get

$$G^{j+1}(z, \hat{t}) = G^j \left( \frac{1+z}{2}, T \right) \exp(\psi(z, \hat{t})).$$

Then, PGF at time $t$, $G(z, t)$, can be represented as follows:

$$G(z, t) = G^{[t/T]} \left( \frac{1+z}{2}, T \right) \exp \left( \psi \left( z, t - T \times [t/T] \right) \right)$$

$$= G^1 \left( \frac{2^{[t/T]} - 1 + z}{2^{[t/T]}}, 0 \right) \prod_{z=1}^{[t/T]} \exp \left( \psi \left( 2^{z-1+z} / 2^z, T \right) \right) \exp \left( \psi \left( z, t - T \times \left[ t/T \right] \right) \right)$$

because time $t$ is equivalent to time $t - T \times [t/T]$ at generation $[t/T] + 1$, i.e., the number of completed cell divisions is $[t/T]$. 
If no proteins exist at time $t = 0$, $G^1 \left( \frac{2^{|T|} - 1 + z}{2^{|T|}}, 0 \right) = \sum_n (z^n \cdot P^1(n, 0)) = P(0, 0) = 1$. Then,

$$ G(z, t) = \left[ \prod_{s=1}^{[T]} \exp \left( \psi \left( \frac{2^s - 1 + z}{2^s}, T \right) \right) \right] \exp \left( \psi \left( z, t - T \times \left\lfloor \frac{t}{T} \right\rfloor \right) \right). $$

### iii) Bursting (fig. S10C)

In the simple delayed birth-death process (Fig. 2D), consecutive production was assumed, which is only valid under some circumstances (20, 38, 59). For instance, if the plasmid copy number is high and thus the relevant gene copy number is also high (e.g., > 20 per cell) (38), total protein production would be consecutive while individual genes might exhibit bursty expression (78-80). When bursting has a large contribution to cell-to-cell variability, it is needed to be explicitly included in the model (Fig. 2D) as in previous work (55, 57, 81). For this, we can incorporate bursting whose frequency and size following arbitrary probability distributions into the model as in previous work (78). Then, its integral equation of PGF can be derived, enabling the calculation of moments.

Specifically, let a molecule $U$ (e.g., premature mRNA or protein) be formed in bursts. Furthermore, let the inter-burst period and burst size follow arbitrary probability distributions whose density and mass functions are $f(\tau)$ and $b(n)$, respectively. Then, $U$ is transformed into another molecule $X$ after a random time delay (e.g., mRNA or protein maturation) following an arbitrary probability distribution whose density function is $h(\tau)$. For this process, the integral equation of the PGF of the number of $U$ at time $t$ and that of $X$ at time $t$ can be formulated (82) as follows:

$$ G_U(z, t) = 1 - F(t) + \int_0^t G_U(z, t - \tau) G_B(z + (1 - z)H(t - \tau)) dF(\tau) $$

$$ G_X(z, t) = 1 - F(t) + \int_0^t G_X(z, t - \tau) G_B(1 + (z - 1)H(t - \tau)) dF(\tau) $$

where $F(t) = \int_0^t f(\tau) d\tau$, $G_B(z) = \sum_{n=1}^{\infty} b(n) \cdot z^n$, and $H(t) = \int_0^t h(\tau) d\tau$.

Proof. We will first derive the PGF equation for $X$. The PGF equation for $U$ can be derived similarly. Let $T_n$ be the time at which the $n$-th burst occurs, $B_n$ be the $n$-th burst size, and $H_{n,i}$ be the time for $i$-th transformation of $U$ molecules formed in the $n$-th burst. Then, the number of $X$ at time $t$, $X(t)$, is given by the following process:

$$ X(t) = \sum_{0 \leq T_n \leq t} g(t - T_n, B_n) $$

where

$$ g(t - T_n, B_n) = \sum_{i=1}^{B_n} I(t - T_n, H_{n,i}), $$

and

$$ I(t - T_n, H_{n,i}) = \begin{cases} 0 & \text{if } t - T_n \leq H_{n,i} \\ 1 & \text{if } t - T_n > H_{n,i}. \end{cases} $$
Note that function $I$ is independent to $i$ because the molecules $U$ being formed in the bursts are independently transformed into $X$ following $h(\tau)$.

Then, under the condition that $T_1 = \tau$,

$$X(t) = \begin{cases} g(t - \tau, B_1) + X^*(t - \tau) & \text{if } \tau \leq t \\ 0 & \text{if } \tau > t \end{cases} \quad (S20)$$

where $X^*(t)$ has the same distribution to $X(t)$.

Since $T_1$ is the time at the first burst occurs, the distribution of $\tau$ is the same as that of the inter-burst period; from Eq. $(S20)$ we get

$$G_X(z, t) = E[z^{X(t)}] = \int_{\tau=0}^{\infty} z^0 dF(\tau) + \int_{0}^{t} E[z^{g(t-\tau, B_1) + X^*(t-\tau)}]dF(\tau)$$

$$= 1 - F(t) + \int_{0}^{t} E[z^{g(t-\tau, B_1) + X^*(t-\tau)}]dF(\tau). \quad (S21)$$

Then, since $g(t - \tau, B_1)$ and $X^*(t - \tau)$ are independent of each other,

$$\int_{0}^{t} E[z^{g(t-\tau, B_1) + X^*(t-\tau)}]dF(\tau) = \int_{0}^{t} E[z^{g(t-\tau, B_1)}]E[z^{X^*(t-\tau)}]dF(\tau). \quad (S22)$$

Using the law of total expectation,

$$E[z^{g(t-\tau, B_1)}] = E \left[ E[z^{g(t-\tau, B_1)} | B_1] \right].$$

Then,

$$E \left[ E[z^{g(t-\tau, B_1)} | B_1] \right] = \sum_{n=1}^{\infty} b(n)E[z^{\sum_{i=1}^{n} I(t-\tau, H_{1,i})}] = G_B(1 + (z - 1)H(t - \tau)) \quad (S23)$$

since $I(t-\tau, H_{1,i})$ are independent and identically distributed and the PGF of $I(t-\tau, H_{1,i})$ is $1 - H(t - \tau) + z \cdot H(t - \tau) = 1 + (z - 1)H(t - \tau)$ from its definition. Using Eq. $(S21-S23)$,

$$G_X(z, t) = 1 - F(t) + \int_{0}^{t} G_B(1 + (z - 1)H(t - \tau))E[z^{X^*(t-\tau)}]dF(\tau). \quad (S24)$$

Since the distribution of $X(t)$ is the same as that of $X^*(t)$,

$$E[z^{X^*(t-\tau)}] = G_X(z, t - \tau).$$

By substituting this into Eq. $(S24)$, we can obtain the PGF equation for $X$ as follows:

$$G_X(z, t) = 1 - F(t) + \int_{0}^{t} G_X(z, t - \tau)G_B(1 + (z - 1)H(t - \tau))dF(\tau).$$

In a similar way, we can derive the PGF equation for $U$ as follows:
\[ G_U(z, t) = 1 - F(t) + \int_0^t G_U(z, t - \tau) G_B(z + (1 - z)H(t - \tau)) dF(\tau). \]

With the above PGF equations, the transient Little’s law, and the Chemical Fluctuation Theorem, the mean and variation of \( X \) at time \( t \) could be expressed in closed form when \( X \) degrades although its form would be complex.
Fig. S1. MBI is still accurate and precise when only the partial information of sample time traces is given. (A) Despite sparser measurements, the estimates are accurate and precise. Box plots of $10^2$ posterior means that were obtained with measurements at intervals of 2 min, 4 min and 6 min. For each posterior mean, we used 40 sample time traces randomly and repeatedly selected from the $10^2$ time traces in Fig. 2E. Here, non-informative priors were used (see Supplementary Text D). (B, C) The delay distributions are accurately estimated even with only the measurement of relative molecular levels. The mean (circle) and variance (triangle) of the scaled time traces obtained by multiplying $2^{-1}$, $2^0$ and $2^1$ to the sample time traces in Fig. 2E (B). Even with scaled data, estimates of death rate, mean time delay, and variance of time delay are still accurate and precise (C). Moreover, the estimate of birth rate is accurate up to scale. Box plots are based on $10^2$ posterior means, each estimated using 40 sample time traces randomly and repeatedly selected from the $10^2$ time traces in Fig. 2E. Here, non-informative priors were used (see Supplementary Text D).
**Fig. S2.** Estimation of heterogeneous parameter values of $\lambda_b$, $\lambda_d$, and $\mu_\tau$ is still accurate and precise even if the magnitude of heterogeneity in the parameters is large. Box plots of $10^2$ Hellinger distances, which were calculated from the $10^2$ independent estimation trials. In each trial, we used $10^2$ time traces from $10^2$ model with parameter sets randomly sampled from gamma distributions whose means are the values in Fig. 2E and coefficients of variation (CV) are either 0.3, 0.4 or 0.5. In this estimation, non-informative priors were used (see Supplementary Text E).
Fig. S3. Strong correlation between \( n \) and \( t_r \). The posterior samples of \( n \) and \( t_r \) for four representative single cells, which were inferred with MBI from the 10² time traces used in Fig. 4B-G. In this estimation, non-informative priors were used (see Supplementary Text E).
Fig. S4. The heterogeneous values of $\lambda_b$, $\lambda_d$, $\mu_r$, and $\sigma^2_r$ can be accurately estimated under the assumption that $n$ is the same among single cells. Posterior means of $\lambda_b$, $\lambda_d$, $\mu_r$, and $\sigma^2_r$ were estimated from the $10^2$ time traces used in Fig. 4H and I under the assumption that the number of rate-limiting steps, $n$, is the same among single cells as in Fig. 4H and I. In this estimation, non-informative priors were used (see Supplementary Text E).
Fig. S5. MBI accurately estimates the magnitude of cell-to-cell heterogeneity of $\lambda_b$ even with relatively measured data. (A) The estimated posterior means of $\lambda_b$ from scaled timeseries data. Here, $10^2$ time traces in Fig. 4H-I scaled with $2^{-1}$, $2^0$, and $2^1$ were analyzed using MBI with the mixed-effects model. The estimated $\lambda_b$ values are accurate up to scale. (B, C) Box plots of $10^2$ Hellinger distances that were calculated from the $10^2$ independent estimation trials (B). As well as $\lambda_b$ (A), the other parameters were accurately estimated (i.e., median of $10^2$ Hellinger distances < ~0.15). For $\lambda_b$, the distances were calculated after the estimated $\lambda_b$ values were scaled to check whether they were accurate up to scale. Box plots of $10^2$ posterior means of $\lambda_b$ that were obtained from the $10^2$ trials (C). (D) Box plots of $10^2$ population CVs of $\lambda_b$ estimated from the scaled time trace data. The estimated CV of $\lambda_b$ was scale invariant. The estimates in (C) and (D) were normalized by dividing by the true values.
Fig. S6. When the number of intermediate rate-limiting steps increases, the magnitude of heterogeneity of response intensity increases. The signaling cascade model described in Supplementary Text A and Fig. 2A (ii) was simulated using the Gillespie algorithm. Specifically, $10^2$ time traces were simulated using $10^2$ heterogeneous models. The parameter values of the $10^2$ models were randomly sampled from gamma distributions where $\lambda_a = 20 \text{ min}^{-1}$, $\lambda_d = 0.5 \text{ min}^{-1}$, $r = 1 \text{ min}^{-1}$, $a_i = 1 \text{ min}^{-1}$ on average and the CV is 0.3. As the number of rate-limiting steps ($n$) increased from 1 to 5, the stationary distribution becomes broader (histogram).
Fig. S7. MBI estimates the parameters of the cell signaling process with feedback inhibition. (A) The posterior samples of parameters obtained with MBI using the mean of the fluctuating sample time traces in Fig. 6B. All the parameters were accurately estimated. The sample values were normalized as in Fig. 3B. (B) Box plots of $10^2$ posterior means obtained as in Fig. 3B. (C) Even if time traces generated with a random feedback delay ($\tau_2 \sim \Gamma(n = 3, t_r = 2$ min)) were used for estimation, the estimates were accurate.
Fig. S8. Analysis of single-cell time-lapse fluorescent protein expression from various promoters in response to TET, TMP, NIT, and CHL. (A) The posterior means of parameters, which were inferred as in fig. S7A. The dilution rate ($\lambda_d$) was given as it can be directly estimated from the experimentally measured cell growth rate. Here, promoters stimulated by TET, TMP, NIT, and CHL are colored in red, yellow, green, and blue, respectively. Each estimate was normalized as in Fig. 5H. (B) Dependence of the dynamics of output response on $\mu_{r_1+r_2}$ and $\lambda_b/R$. The signal response intensity shows the adaptions with a longer duration as $\mu_{r_1+r_2}$ becomes larger. The intensity shows oscillations as $\lambda_b/R$ becomes large, and its period increases with larger $\mu_{r_1+r_2}$. 
The cell signaling cascade model (Fig. 2A (ii)) with a feedback inhibition via protein sequestration was simulated using the Gillespie algorithm as in fig. S6. For the simulation, the parameter values were randomly sampled from gamma distributions where $\lambda_a = 30 \text{ min}^{-1}$, $\lambda_d = 0.05 \text{ min}^{-1}$, $R = 50$, $r = 1 \text{ min}^{-1}$, $a_t = 1 \text{ min}^{-1}$ on average and the CV is 0.3. As the number of rate-limiting steps ($n$) increased from 1 to 5, the distribution of molecular number after a transient state becomes broader (histogram).
Fig. S10. Extensions of the simple delayed birth-death process (Fig. 2D). (A) Delayed birth-death process with downstream regulation. Downstream regulation of an upstream molecule, $U$, was incorporated into the delayed birth-death process with a stochastic intensity, $\lambda_3(U)$, because the number of upstream molecules can also be a stochastic process (see Supplementary Text J (i)). (B) Delayed process with cell division-induced dilution. To capture the cell division-induced dilution, binomial partitioning of proteins at cell division was incorporated into the delayed birth-death process as in previous work (55) (see Supplementary Text J (ii)). (C) Delayed bursty birth-death process. A signal activation occurring in bursts, whose frequency and size were modeled with an arbitrary probability distribution, was incorporated into the delayed birth-death process as in previous work (78) (see Supplementary Text J (iii)). The derived formulae describing the dynamics of these processes (Supplementary Text J) can be incorporated into MBI.
REFERENCES AND NOTES

1. A. Raj, A. Van Oudenaarden, Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell* **135**, 216–226 (2008).

2. K. Mitosch, G. Rieckh, T. Bollenbach, Noisy response to antibiotic stress predicts subsequent single-cell survival in an acidic environment. *Cell Syst.* **4**, 393–403.e5 (2017).

3. K. Mitosch, G. Rieckh, T. Bollenbach, Temporal order and precision of complex stress responses in individual bacteria. *Mol. Syst. Biol.* **15**, e8470 (2019).

4. M. Niepel, S. L. Spencer, P. K. Sorger, Non-genetic cell-to-cell variability and the consequences for pharmacology. *Curr. Opin. Chem. Biol.* **13**, 556–561 (2009).

5. B. R. Levin, D. E. Rozen, Non-inherited antibiotic resistance. *Nat. Rev. Microbiol.* **4**, 556–562 (2006).

6. Y. Wakamoto, N. Dhar, R. Chait, K. Schneider, F. Signorino-Gelo, S. Leibler, J. D. McKinney, Dynamic persistence of antibiotic-stressed mycobacteria. *Science* **339**, 91–95 (2013).

7. A. Raj, S. A. Rifkin, E. Andersen, A. Van Oudenaarden, Variability in gene expression underlies incomplete penetrance. *Nature* **463**, 913–918 (2010).

8. P. B. Gupta, I. Pastushenko, A. Skibinski, C. Blanpain, C. Kuperwasser, Phenotypic plasticity: Driver of cancer initiation, progression, and therapy resistance. *Cell Stem Cell* **24**, 65–78 (2019).

9. S. Di Talia, J. M. Skotheim, J. M. Bean, E. D. Siggia, F. R. Cross, The effects of molecular noise and size control on variability in the budding yeast cell cycle. *Nature* **448**, 947–951 (2007).

10. M. Wallden, D. Fange, E. G. Lundius, O. Baltekin, J. Elf, The synchronization of replication and division cycles in individual *E. coli* cells. *Cell* **166**, 729–739 (2016).

11. B. B. Aldridge, M. Fernandez-Suarez, D. Heller, V. Ambravaneswaran, D. Irimia, M. Toner, S. M. Fortune, Asymmetry and aging of mycobacterial cells lead to variable growth and antibiotic susceptibility. *Science* **335**, 100–104 (2012).
12. J. Yang, M. Reth, Oligomeric organization of the B-cell antigen receptor on resting cells. *Nature* **467**, 465–469 (2010).

13. S. Yang, S. Kim, Y. R. Lim, C. Kim, H. J. An, H.-H. Kim, J. Sung, N. K. Lee, Contribution of RNA polymerase concentration variation to protein expression noise. *Nat. Commun.* **5**, 4761 (2014).

14. P. Pozarowski, Z. Darzynkiewicz, Analysis of Cell Cycle by Flow Cytometry, in *Checkpoint controls and cancer*, A. H. Schönthal, Ed. (Springer, 2004), vol. 281, chap. 22. pp. 301–311.

15. A. Sarrion-Perdigones, L. Chang, Y. Gonzalez, T. Gallego-Flores, D. W. Young, K. J. T. Venken, Examining multiple cellular pathways at once using multiplex hextuple luciferase assaying. *Nat. Commun.* **10**, 1710 (2019).

16. I. C. Macaulay, C. P. Ponting, T. Voet, Single-cell multiomics: Multiple measurements from single cells. *Trends Genet.* **33**, 155–168 (2017).

17. J. Yu, J. Zhou, A. Sutherland, W. Wei, Y. S. Shin, M. Xue, J. R. Heath, Microfluidics-based single-cell functional proteomics for fundamental and applied biomedical applications. *Ann. Rev. Anal. Chem.* **7**, 275–295 (2014).

18. M. Barrio, A. Leier, T. T. Marquez-Lago, Reduction of chemical reaction networks through delay distributions. *J. Chem. Phys.* **138**, 104114 (2013).

19. M. Beguerisse-Díaz, R. Desikan, M. Barahona, Linear models of activation cascades: Analytical solutions and coarse-graining of delayed signal transduction. *J. R. Soc. Interface* **13**, 20160409 (2016).

20. B. Choi, Y. Y. Cheng, S. Cinar, W. Ott, M. R. Bennett, K. Josić, J. K. Kim, Bayesian inference of distributed time delay in transcriptional and translational regulation. *Bioinformatics* **36**, 586–593 (2020).

21. S. J. Park, S. Song, G. S. Yang, P. M. Kim, S. Yoon, J. H. Kim, J. Sung, The chemical fluctuation theorem governing gene expression. *Nat. Commun.* **9**, 297 (2018).
22. J. Zhang, T. Zhou, Markovian approaches to modeling intracellular reaction processes with molecular memory. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 23542–23550 (2019).

23. C. Gupta, J. M. López, R. Azencott, M. R. Bennett, K. Josić, W. Ott, Modeling delay in genetic networks: From delay birth-death processes to delay stochastic differential equations. *J. Chem. Phys.* **140**, 204108 (2014).

24. C. Gardiner, *Stochastic Methods* (Springer, 2009), vol. 4.

25. M. Karlsson, D. L. I. Janzén, L. Durrieu, A. Colman-Lerner, M. C. Kjellsson, G. Cedersund, Nonlinear mixed-effects modelling for single cell estimation: When, why, and how to use it. *BMC Syst. Biol.* **9**, 52 (2015).

26. J. Almquist, L. Bendrioua, C. B. Adiels, M. Goksör, S. Hohmann, M. Jirstrand, A nonlinear mixed effects approach for modeling the cell-to-cell variability of Mig1 dynamics in yeast. *PLOS ONE* **10**, e0124050 (2015).

27. D. Schnoerr, G. Sanguinetti, R. Grima, Approximation and inference methods for stochastic biochemical kinetics—A tutorial review. *J. Phys. A: Math. Theor.* **50**, 093001 (2017).

28. C. Zechner, J. Ruess, P. Krenn, S. Pelet, M. Peter, J. Lygeros, H. Koeppl, Moment-based inference predicts bimodality in transient gene expression. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 8340–8345 (2012).

29. A. Leier, T. T. Marquez-Lago, Delay chemical master equation: Direct and closed-form solutions. *Proc. R. Soc. A: Math. Phys. Eng. Sci.* **471**, 20150049 (2015).

30. S. Calderazzo, M. Brancaccio, B. Finkenstadt, Filtering and inference for stochastic oscillators with distributed delays. *Bioinformatics* **35**, 1380–1387 (2019).

31. J. R. Moffitt, C. Bustamante, Extracting signal from noise: Kinetic mechanisms from a Michaelis–Menten-like expression for enzymatic fluctuations. *FEBS J.* **281**, 498–517 (2014).
32. J. Roux, M. Hafner, S. Bandara, J. J. Sims, H. Hudson, D. Chai, P. K. Sorger, Fractional killing arises from cell-to-cell variability in overcoming a caspase activity threshold. *Mol. Syst. Biol.* **11**, 803 (2015).

33. B. N. Kholodenko, Cell-signalling dynamics in time and space. *Nat. Rev. Mol. Cell Biol.* **7**, 165–176 (2006).

34. R. A. Bradshaw, E. A. Dennis, *Handbook of Cell Signaling* (Academic press, 2009).

35. N. Korsbo, H. Jönsson, It’s about time: Analysing simplifying assumptions for modelling multi-step pathways in systems biology. *PLoS Comput. Biol.* **16**, e1007982 (2020).

36. E. Balleza, J. M. Kim, P. Cluzel, Systematic characterization of maturation time of fluorescent proteins in living cells. *Nat. Methods* **15**, 47–51 (2018).

37. A. Leier, M. Barrio, T. T. Marquez-Lago, Exact model reduction with delays: Closed-form distributions and extensions to fully bi-directional monomolecular reactions. *J. R. Soc. Interface* **11**, 20140108 (2014).

38. Y. Y. Cheng, A. J. Hirning, K. Josic, M. R. Bennett, The timing of transcriptional regulation in synthetic gene circuits. *ACS Synth. Biol.* **6**, 1996–2002 (2017).

39. D. Bertsimas, G. Mourtzinou, Transient laws of non-stationary queueing systems and their applications. *Queueing Syst.* **25**, 115–155 (1997).

40. L. Kleinrock, *Queueing Systems* (Wiley, 1975).

41. W. J. Dixon, F. J. Massey Jr., *Introduction to Statistical Analysis* (1951).

42. S. Tanase-Nicola, P. B. Warren, P. R. ten Wolde, Signal detection, modularity, and the correlation between extrinsic and intrinsic noise in biochemical networks. *Phys. Rev. Lett.* **97**, 068102 (2006).

43. J. Selimkhanov, B. Taylor, J. Yao, A. Pilko, J. Albeck, A. Hoffmann, L. Tsimring, R. Wollman, Systems biology. Accurate information transmission through dynamic biochemical signaling networks. *Science* **346**, 1370–1373 (2014).
44. A. Cersini, M. C. Martino, I. Martini, G. Rossi, M. L. Bernardini, Analysis of virulence and inflammatory potential of Shigella flexneri purine biosynthesis mutants. *Infect. Immun.* **71**, 7002–7013 (2003).

45. D. W. Bryant, D. R. McCalla, Nitrofuran induced mutagenesis and error prone repair in Escherichia coli. *Chem. Biol. Interact.* **31**, 151–166 (1980).

46. S. Sonneveld, B. M. Verhagen, M. E. Tanenbaum, Heterogeneity in mRNA translation. *Trends Cell Biol.* **30**, 606–618 (2020).

47. O. Brandman, T. Meyer, Feedback loops shape cellular signals in space and time. *Science* **322**, 390–395 (2008).

48. V. Shimoga, J. T. White, Y. Li, E. Sontag, L. Bleris, Synthetic mammalian transgene negative autoregulation. *Mol. Syst. Biol.* **9**, 670 (2013).

49. J. K. Kim, J. J. Tyson, Misuse of the Michaelis–Menten rate law for protein interaction networks and its remedy. *PLoS Comput. Biol.* **16**, e1008258 (2020).

50. J. K. Kim, Protein sequestration versus Hill-type repression in circadian clock models. *IET Syst. Biol.* **10**, 125–135 (2016).

51. Y. M. Song, H. Hong, J. K. Kim, Universally valid reduction of multiscale stochastic biochemical systems using simple non-elementary propensities. *PLoS Comput. Biol.* **17**, e1008952 (2021).

52. S. Beesley, D. W. Kim, M. D’Alessandro, Y. Jin, K. Lee, H. Joo, Y. Young, R. J. Tomko Jr, J. Faulkner, J. Gamsby, J. K. Kim, C. Lee, Wake-sleep cycles are severely disrupted by diseases affecting cytoplasmic homeostasis. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 28402–28411 (2020).

53. T. Zhou, T. Liu, Quantitative analysis of gene expression systems. *Quant. Biol.* **3**, 168–181 (2015).

54. H. Hong, J. Kim, M. A. Al-Radhawi, E. D. Sontag, J. K. Kim, Derivation of stationary distributions of biochemical reaction networks via structure transformation. *Commun. Biol.* **4**, 1–10 (2021).
55. C. H. Beentjes, R. Perez-Carrasco, R. Grima, Exact solution of stochastic gene expression models with bursting, cell cycle and replication dynamics. *Phys. Rev. E* **101**, 032403 (2020).

56. C. Jia, R. Grima, Frequency domain analysis of fluctuations of mRNA and protein copy numbers within a cell lineage: Theory and experimental validation. *Phys. Rev. X* **11**, 021032 (2021).

57. Z. Cao, R. Grima, Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 4682–4692 (2020).

58. Q. Jiang, X. Fu, S. Yan, R. Li, W. Du, Z. Cao, F. Qian, R. Grima, Neural network aided approximation and parameter inference of non-Markovian models of gene expression. *Nat. Commun.* **12**, 2618 (2021).

59. M. J. Cortez, H. Hong, B. Choi, J. K. Kim, K. Josić, Hierarchical Bayesian models of transcriptional and translational regulation processes with delays. *Bioinformatics* **38**, 187–195 (2021).

60. C. Zechner, M. Unger, S. Pelet, M. Peter, H. Koeppl, Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. *Nat. Methods* **11**, 197–202 (2014).

61. A. P. Frei, F. A. Bava, E. R. Zunder, E. W. Y. Hsieh, S. Y. Chen, G. P. Nolan, P. F. Gherardini, Highly multiplexed simultaneous detection of RNAs and proteins in single cells. *Nat. Methods* **13**, 269–275 (2016).

62. P. D. Dixit, E. Lyashenko, M. Niepel, D. Vitkup, Maximum entropy framework for predictive inference of cell population heterogeneity and responses in signaling networks. *Cell Syst.* **10**, 204–212.e8 (2020).

63. K. Sachs, O. Perez, D. Pe’er, D. A. Lauffenburger, G. P. Nolan, Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**, 523–529 (2005).

64. D. Pe’er, A. Regev, G. Elidan, N. Friedman, Inferring subnetworks from perturbed expression profiles. *Bioinformatics* **17**, S215–S224 (2001).
65. J. Tyler, D. Forger, J. K. Kim, Inferring causality in biological oscillators. *Bioinformatics* **38**, 196–203 (2021).

66. L. U. Haberbeck, X. Wang, C. Michiels, F. Devlieghere, M. Uyttendaele, A. H. Geeraerd, Cross-protection between controlled acid-adaptation and thermal inactivation for 48 *Escherichia coli* strains. *Int. J. Food Microbiol.* **241**, 206–214 (2017).

67. R. Wang, N. Gao, Y. Li, D. Fan, Z. Zhen, K. Feng, H. Chen, J. An, Cross-protection against four serotypes of dengue virus in mice conferred by a Zika DNA vaccine. *Front. Cell. Infect. Microbiol.* **9**, 147 (2019).

68. J. S. Gewandter, M. P. McDermott, H. He, S. Gao, X. Cai, J. T. Farrar, N. P. Katz, J. D. Markman, S. Senn, D. C. Turk, R. H. Dworkin, Demonstrating heterogeneity of treatment effects among patients: An overlooked but important step toward precision medicine. *Clin. Pharmacol. Ther.* **106**, 204–210 (2019).

69. J. Liu, D. Hansen, E. Eck, Y. J. Kim, M. Turner, S. Alamos, H. G. Garcia, Real-time single-cell characterization of the eukaryotic transcription cycle reveals correlations between RNA initiation, elongation, and cleavage. *PLoS Comput. Biol.* **17**, e1008999 (2021).

70. C. J. Geyer, Practical markov chain monte carlo. *Stat. Sci.* **7**, 473–483 (1992).

71. D. Barber, Y. Wang, in *International Conference on Machine Learning* (PMLR, 2014), 1485–1493.

72. A. F. Smith, G. O. Roberts, Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Stat. Soc. Series. B Stat. Methodol.* **55**, 3–23 (1993).

73. W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**, 97–109 (1970)

74. S. G. Eick, W. A. Massey, W. Whitt, The physics of the *M/G/∞* queue. *Oper. Res.* **41**, 731–742 (1993).

75. R. Milo, R. Phillips, *Cell Biology by the Numbers* (Garland Science, 2015).
76. R. Perez-Carrasco, C. Beentjes, R. Grima, Effects of cell cycle variability on lineage and population measurements of messenger RNA abundance. J. R. Soc. Interface 17, 20200360 (2020).

77. P. Wang, S. Hayden, Y. Masui, Transition of the blastomere cell cycle from cell size-independent to size-dependent control at the midblastula stage in *Xenopus laevis*. J. Exp. Zool. 287, 128–144 (2000).

78. J. M. Pedraza, J. Paulsson, Effects of molecular memory and bursting on fluctuations in gene expression. Science 319, 339–343 (2008).

79. C. R. Bartman, N. Hamagami, C. A. Keller, B. Giardine, R. C. Hardison, G. A. Blobel, A. Raj, Transcriptional burst initiation and polymerase pause release are key control points of transcriptional regulation. Mol. Cell 73, 519–532.e4 (2019).

80. Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, X. S. Xie, Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. Science 329, 533–538 (2010).

81. V. Shahrezaei, P. S. Swain, Analytical distributions for stochastic gene expression. Proc. Natl. Acad. Sci. U.S.A. 105, 17256–17261 (2008).

82. L. Liu, B. Kashyap, J. Templeton, On the $G^X/G/\infty$ system. J. Appl. Prob. 27, 671–683 (1990).